Predicting Serious Crime Trends in South Africa Using Data Analytic Techniques

Submitted in fulfilment of the requirements for the Degree of

Master of Information and Communications Technology

in the Faculty of

Accounting and Informatics

Durban University of Technology

by

Olayemi Success Falope

(21960223)

Date Submitted: April 2024

Supervisor:
Prof. S. Thakur (D.Tech. Information Technology)

6 April 2024
Date

Abstract

This dissertation aims to investigate the application of data analytics in forecasting serious crime trends in South Africa. The escalating rates of serious crimes, including homicide, robbery, and sexual assault, present significant challenges to the country's economic growth and the safety of its citizens. Recent South African crime statistics indicate a notable increase of over 9.6% in serious crimes, rising from 444,452 incidents in December 2021 to 486,960 in December 2022. This upward trajectory underscores the urgency to predict future serious crimes preemptively, facilitating the development of proactive strategies by law enforcement agencies, policymakers, and community organizations to prevent and mitigate criminal activities.

To achieve this objective, this study employs a comprehensive dataset comprising historical crime records and spatial data to analyse serious crime trends across South Africa's nine provinces from 2005 to 2020. Data pre-processing techniques are applied to clean and normalize the data, ensuring its suitability for subsequent analysis. Exploratory data analysis is conducted using Python (Anaconda) and the Flourish studio environment to identify patterns, relationships, and potentially influential factors associated with serious crimes in South Africa. Various data analytics techniques, including machine learning algorithms, time series analysis, and spatial analysis, are utilized to construct models for predicting serious crime trends. These predictive models are trained using historical crime data and relevant contextual features, facilitating the identification of patterns and correlations that could inform future crime trends.

The evaluation of these predictive models involves rigorous performance metrics and validation techniques to assess their predictive power, stability, and generalizability. The results reveal an increase in serious crime across South Africa, with certain provinces emerging as hotspots for specific serious crimes, such as Gauteng with a 21% increase in sexual crimes, KwaZulu-Natal with a 23.1% increase in murders, and the Western Cape with a 38% increase in drug-related crimes.

This dissertation contributes to the field of crime analysis by presenting a comprehensive approach to predicting serious crime trends in South Africa. The insights gained from this research can inform the development of proactive strategies and resource allocation by law enforcement agencies, policymakers, and community organizations to address serious crimes effectively. Furthermore, this study lays the groundwork for future research in crime prediction and prevention, highlighting the potential of data analytics techniques in tackling complex societal issues. Future research may explore advanced techniques such as ensemble learning and deep learning to enhance the accuracy and robustness of predictive models.

Keywords

Predictive Data Analytics, Ordinary Least Square Regression, Machine Learning, Time Series Analysis, Spatial Analysis, Serious Crimes.

Declaration

I, Olayemi Success Falope, hereby declare that the content within this thesis is my own work. All sources that I have used or quoted have been acknowledged in the text by the means of completed references. This study has not been previously submitted in any form to the Durban University of Technology or to any other institution for assessment or for any other purpose.

Student: 6 April, 2024

Olayemi Success Falope Date

Approved for final submission

Supervisor: 6 April, 2024

Professor Surendra Thakur Date

Dedication

I dedicate this dissertation to the Almighty God, who has been so helpful to me throughout this research, and to the undying spirit of South Africa's vulnerable victims of crimes seeking justice and social reintegration. I hope this work will help fight their cause. To my supportive husband, my sister and her husband, my little brothers, and other family members and friends, whose unwavering support has inspired me to keep pursuing excellence.

Acknowledgements

This dissertation would not have been possible without the help of the Durban University of Technology (DUT), my supervisor and my mentor, who supported me throughout this journey. I wish to thank the following people for their much-valued support and contributions:

- My supervisor, Prof. Surendra Thakur: Thank you so much for your commitment and coaching during this study. Your guidance inspired me and made working with you a pleasure.
- My mentor, Prof. Annelie Jordan: Thank you very much for your support and encouragement. Working with you and sharing ideas made the work easily understandable. I appreciate your mentorship efforts towards my development.
- Durban University of Technology (DUT): The support from DUT is highly appreciated. I want to thank DUT for the opportunity to pursue this qualification.
- University of Valladolid (UVa), Spain: The help from UVa is highly appreciated. I want to thank UVa for the mobility program opportunity offered to study for a semester at her institution.
- I want to thank all other individuals not listed here but who directly or indirectly contributed to the successful completion of this study.

Financial Acknowledgement

The financial assistance of the Durban University of Technology (DUT) scholarship scheme and DUT Research and Postgraduate Support towards this research is well acknowledged. The financial support of the University of Valladolid towards the Erasmus mobility program is well appreciated.

Table of Contents

List of Figures

List of Tables

List of Acronyms

CRISP-DM: Cross-Industry Standard Process for Data Mining

GBV: Gender-based Violence

KDD: Knowledge Discovery in Databases

KNN: k-nearest Neighbours algorithm

OLS: Ordinary Least Square Regression

PII: Personally Identifiable Information

SAPS: South African Police Service

SEMMA: Sample, Explore, Modify, Model, Assess

SVM: Support Vector Machine

UNODC: United Nations Office on Drugs and Crime

WEKA: Waikato environment for knowledge analysis

WHO: World Health Organization

Output

Conferences

1. Falope, O.S. and Thakur, C., 2022a. Data analytics for South Africa's sexual crime landscape. In Nemisa Summit and Colloquium, 15--17 February 2022.

2. Falope, O.S. and Thakur, C., 2022b. Sexual Crime Prediction in an African Context. *2022 International Conference on Intelligent and Innovative Computing Applications, (ICONIC2022)*, 8--9 December 2022.

Workshops/Seminars Attended

1. Towards Systematising Postgraduate research proposal writing by Professor SO Ojo, 11 June 2021.
2. PG Seminar: Postgraduate research as a journey by Professor SO Ojo, 18 June 2021.
3. PG lecture: Research methods for machine learning by Dr T Adeliyi, 24 June 2021
4. PG lecture: An introduction to Big Data Analytics by Dr Israel Edem Agbehadji and Prof Millham, 01 July 2021.
5. Research methods for Unstructured Data by Mr Yasin Khan and Prof Thakur, 08 July 2021.
6. Hackathon: Smart Global Ecosystems, Palencia campus, Spain, 24-25 February 2022.
7. Writing the Literature Review by Dr J Wing, 04 August 2022.
8. PG Lecture Series: Research Methods by Professor SO Ojo, 11 August 2022.
9. PG Lecture Series: Big Data by Prof Richard Millham, 01 September 2022.
10. PG Lecture Series: Systematic literature review by Dr Adeliyi and Mrs Joseph, 08 September 2022.

Conflict of Interest

No conflict of interest declared.

Chapter 1 Introduction

1.1 Introduction

This study explores serious crime prediction, encompassing sexual, drug-related, and murder crimes, through data analytic techniques. Leveraging historical crime data and socio-economic indicators enables proactive measures to create safer communities and address root causes, benefiting policymakers, law enforcement, and communities.

High youth unemployment exacerbates social discontent and fosters criminal networks (Monyeki, 2021; Statistics South Africa, 2021). Organized crime, gang violence, and drug trafficking thrive amid socio-economic challenges (Gastrow, 1998). Addressing these complexities requires a multidimensional approach, integrating data analytics for prediction and prevention (Ahishakiye et al., 2017).

Crimes present significant global challenges, affecting developed and developing regions (Chang et al., 2019; Mathews & Collin-Vézina, 2019). South Africa, a developing nation, grapples with high crime, including homicide, robbery, and sexual assault (UNODC, 2020). Socio-economic inequality, stemming from the apartheid legacy, persists despite progress (Triegaardt, 2021). Persistent poverty, unemployment, and limited opportunities create fertile ground for criminal activity.

The rising serious crime rates impacts governance significantly (Chauhan & Sehgal, 2017) which triggers socioeconomic and political instability, The apartheid era contributed to a culture of violence (Kempen, 2019). African crime rates exceed those of Europe and Asia, suggesting further data analytics (Africa Times Editor, 2019).

Law enforcement faces challenges in analyzing vast crime evidence without assistive technologies (Chauhan & Sehgal, 2017). Crime data analysis aids in detecting and predicting crime trends (Gahalot, Dhiman, & Chouhan, 2020). This study aims to evaluate state-of-the-art data analytic techniques for serious crime prediction in the South African context, offering feasible solutions to prevailing challenges.

1.2 Serious Crime

Serious crime pertains to criminal offences that are deemed especially severe or grave, given their potential impact on individuals, communities, and society at large. Such crimes generally involve significant harm, violence, or a threat to public safety. Examples of serious crimes include homicide, sexual assault, armed robbery, kidnapping, human trafficking, terrorism, and drug trafficking. The severity of serious crimes leads to stringent legal consequences upon conviction (UNODC, 2020).

1.2.1 Sexual crimes

According to the World Health Organization (WHO) (2012, pp. 2), sexual violence is

> any sexual act, attempt to obtain a sexual act, unwanted sexual comments or advances, or acts to traffic or otherwise directed against a person's sexuality using coercion, by any person regardless of their relationship to the victim, in any setting.

Sexual crime encompasses unwelcome sexual comments, sexual acts, sexual advances, and acts targeting an individual's sexuality.

WHO (2012) Edwards, O'Mahoney and Vincent (2014) and Oyasor (2020) classified sexual crime as: incest and rape, workplace or school sexual assault, sexual harassment against detained or imprisoned women, violence against women who have been displaced, women trafficking, crime in the home.

Sexual crimes include sexual assault, exploitation, harassment, trafficking, and rape (Falope and Thakur, 2022 b). Sexual crimes are most often committed in secret. Typically, the criminals threaten their victims with harm if they tell someone. This can impede the disclosure of the true incident; the danger of becoming accused or disbelieved, and likewise hinder the reporting of violence. Sexual victims are scared or ashamed to reveal cases of abuse due to the fear of being harmed further by the culprit, perception of shame and anxiety of not being believed, leading sexual victims to suffer in isolation (Falope and Thakur, 2022 a). The harassment does not end with a single occurrence all the time; victims of sexual crimes may be exposed to other

similar side-line abuses (Abrahams *et al.* 2014; Thomas, 2014; Landström, Strömwall and Alfredsson, 2016; Meinck *et al.* 2017).

The sexual crime of rape is considered an act of contact without the victim's consent. It is a global public health issue that impacts the victims' physical and emotional health and could lead to depression, suicide, post-traumatic stress disorder (PTSD), the human immunodeficiency virus, or other sexually transmitted diseases (Diehl *et al.* 2020). The results of the public security poll by the Brazilian Forum are alarming (Diehl *et al.* 2020). In 2015, an estimated 45 460 people were raped—one every 11 minutes are raped in Brazil every hour (Diehl *et al.* 2020).

Sexual crimes can also include the harassment of married partners as a result of several irreversible detrimental sexual effects related to childhood sexual violence, and the negative psychological implications of the psychological perpetrator can be explained by a combination of painful factors that may leave the survivor stigmatised, humiliated, displeased, and powerless (Vaillancourt-Morel *et al.* 2016). Sexual crimes have long-lasting adverse psychological effects and cause harm to the victims, who endure both physical and emotional pain. Survivors may sustain horrific physical injuries that endanger and harm their internal perceptions (Clark, 2014; Landström, Strömwall and Alfredsson, 2016).

In addition to causing substantial injury to physical health, sexual crimes can also result in death, unwanted pregnancy, high-risk miscarriages, excessive pubic discomfort, and sexually transmissible diseases (Altinyelken and Le Mat, 2018). If the victim is raped as a virgin, it may have a significant effect on her personality by generating a false sacramental platform that leaves her with a sense of societal uncertainty and a misunderstanding about her significance in society. In addition, an act of rape could rob a woman of the possibility of having a husband or her own home (Clark, 2014). The psychological effects of sexual harassment can affect the rationality and efficacy of sexual victims, destabilise their status, and undermine their ability to comprehend and be attentive (Altinyelken and Le Mat, 2018).

Sexual predators contact their future victims through social media, e-mails, and online gaming. Chassiakos *et al.* (2016) cite an instance of online priming that contributes to

the creation of a make-believe relationship, usually with the offender falsifying himself or herself as a caring and innocent individual. A developing online relationship could lead to sexting or tricking the victim into meeting the culprit at a negotiated location before the victim is sexually abused. However, the perpetrator leaves a digital footprint behind, from which data mining can capture the identity, motive, location, and potential attacks by the perpetrator. As criminals improve their technological skills, it is essential that the crime squad keeps up with the latest technology. (Chauhan and Sehgal, 2017).

Sexual crimes and gender-based violence against women and girls are widespread in South Africa (Enaifoghe *et al.* 2021) with the authorities, appearing unable of effectively responding. Sexual crimes, makes it more difficult for women to fight them. In most cases, such attempts result in disastrous outcomes for the women involved Anxiety, humiliation, re-traumatisation, and mistrust in the criminal justice system result in only a few sexual crimes being reported to the police (Jewkes and Abrahams, 2002; Knox and Monaghan, 2005; Naidoo, 2013; Kelly, 2013; Holland and Cortina, 2017; Moletsane, 2018. Therefore, the number of sexual assaults is likely far higher (Naidoo, 2013; Moletsane, 2018). In all South African provinces, violence against women considerably outnumbers other types of gender-based violence (Jewkes, Levin and Penn-Kekana, 2002). Violence against women is “a serious problem for public health, social policy, and human rights” (Sullivan, 1994, pp.152-167). It is challenging to assess how widespread violence against women is and to develop reliable comparison statistics that can be used as a starting point for tracking progress on this issue (Jewkes, Levin and Penn-Kekana, 2002).

The prevalence of various sexual crimes against females in South Africa, combined with the country’s ineffective response, has created “a *rape culture in which sexual crime has become normalised*” (Gray, 2019, pp.35-51). “*Rape culture*” encompasses more than just the act of sexual abuse or rape (Gray, 2019, pp.35-51). It is a patriarchal way of thinking that gives men more power than women and keeps “women and men within certain boundaries and categories" (Moffett, 2006). Rape culture effectively undermines “the citizenship of women and girls” (Du Toit, 2004), emphasising their

assigned inferiority and fostering “a culture of silence, fear, and shame” to control their behaviour.

1.2.2 Drug-related crime

Drug-related crimes are associated with “the use, possession, manufacture, or distribution of drugs such as cocaine, heroin, marijuana, and amphetamines which are classified as having a potential for abuse” (Craddock, Collins and Timrots, 1994, pp. 1). Drug-related crimes are also related to crime through the influences they have on the user’s behaviour and by generating violence and other illegal activities in connection with drug trafficking (Zafarghandi *et al.* 2022). Drug-related crimes are unacceptable as they hurt people's health, society, businesses, and public and recreational spaces (Zafarghandi *et al.* 2022). Factors influencing drug-related crime include restricted access to employment and education, poor neighbourhoods, housing characteristics, and poor living conditions (Marco, Gracia and López-Quílez, 2017; Manhica *et al.* 2021).

1.2.3 Murder

Any act that causes the unlawful death of another person without justification or a legal justification is considered murder, especially when it is done with malice aforethought (Britannica Academic, 2022). Innumerable lives are lost every year as a result of murder, making it one of the most common causes of death globally (Santos and Testa, 2018). Murder is a severe public health problem globally (Lindegaard, 2017; Santos and Testa, 2018). The rate of unemployment is a catalyst for murder in society (Mazorodze, 2020). Murder, assassination, sexual crimes, common robbery, robbery of property, industrial crime, armed robbery and drug-related crimes are some of the crimes committed in South Africa (Africa Check, 2019; Mike and Paul, 2019).

1.2.4 Data analytics for crime prediction

Data analytics refers to a set of tools, techniques, models and systems for analysing collected data and extracting useful and unknown trends, relationships, and information. For instance, it can be used to extract previously unknown, valuable, valid and hidden patterns and information from massive datasets, as well as to find

significant relationships between the variables stored (Elgendy and Elragal, 2014). It is also applicable for conducting predictive analysis.

Basically, predictive analytics is a term that refers to a set of methods for predicting future events using past and present data. Among the well-known predictive analytic methods are the support vector machine (SVM) regressor, the decision tree regressor, linear regression, clustering, and association rules (Vashisht and Gupta, 2015). Predictive analytics has many benefits; for example, a security or law enforcement agency may look at its internal data to see if there are any serious crime trends and then take the required steps to mitigate such trends in the future. It is also a beneficial approach for effective policing and crime control because it can help to figure out which crime control approaches have been successful or have failed over time (Pradhan, 2018). Disadvantages of predictive analytics include incomplete data and varying quality and format of data collected from various sources (Vassakis, Petrakis and Kopanakis, 2018), as well as privacy and time, but these disadvantages would not limit the serious crime analysis and prediction of this study because the cross-industry standard process for data mining (CRISP-DM) methodology provides step-by-step guidance (the steps to be taken) to make the data reliable. Also, the CRISP-DM methodology is the most widely used, which means other studies were satisfied with this data cleaning methodology (Berwind *et al.* 2016; Edoka, 2020; Razali, 2020; Obagbuwa and Abidoye, 2021).

1.3 Research Rationale

The study was motivated by the serious crime rate in South Africa which poses significant challenges. South Africa experiences high levels of serious crimes, including homicide, robbery and sexual assault, which have profound societal and economic implications. In retrospect, predicting the serious crime rate in South Africa is of paramount importance. By developing accurate and reliable predictive models for crime rates, several key benefits can be achieved, which are presented as follows:

i. *Crime Prevention and Resource Allocation*: Predictive models identifies high-risk areas and trends which supports effective resource allocation.

ii. *Policy Development and Intervention Strategies*: Accurate serious crime provide policymakers with essential information to develop evidence-based policies and targeted intervention strategies (Sri *et al.* 2020).

iii. *Community Engagement and Public Awareness*: Predictive models enhance community engagement enabling enforcement agencies to collaborate with communities. By sharing crime predictions public awareness is raised about potential hotspots, and the community can take necessary precautions, report suspicious activities, and participate in crime prevention initiatives.

iv. *Resource Optimisation*: Predictive models identify specific crime types, locations and resources required for targeted intervention. This is a cost savings and efficient use of limited resources, such as police personnel, and surveillance equipment.

v. *Early Warning Systems*: Developing reliable crime prediction models lead to early warning systems (Kim *et al.* 2018) which alert law enforcement agencies about spikes in serious crimes, enabling them to respond swiftly and effectively.

vi. *Evaluation of Policy Effectiveness:* By comparing predicted crime with actual data, the policy effectiveness can be evaluated. This feedback loop allows data-driven decisions to refine policies and strategies to reduce crime rates.

vii. *International Collaboration*: Predictive crime modelling facilitates knowledge sharing and collaboration between South Africa and countries facing similar challenges. The insights inform and improves, policies, and interventions.

1.4 Research Problem

South Africa has one of the highest crime rates in the world, and serious crimes such as murder, robbery and assault are prevalent. The high crime rate in South Africa has been attributed to factors such as a lack of vocational and social skills, poor housing and living conditions, unemployment and inequality, heavy alcohol usage, a large immigrant population and ineffective parenting (Crush and Peberdy, 2018; Mazorodze, 2020; Monyeki, Naicker and Obagbuwa, 2020). Financially, physically, and psychologically, both wealthy and impoverished civilizations are negatively impacted by crime (Monyeki, 2021). Despite the efforts of law enforcement agencies,

serious crimes continue to occur, and there is a need for more effective crime prevention strategies. There have been limited studies that analyse trends in serious crimes and employ exhaustive data analytic approaches to predict new serious crime hotspots in the South African context. Hence, the aim of this research is to apply data analytic techniques to predict serious crime trends in South Africa.

A key challenge is to devise ways of leveraging historical crime data, socio-economic indicators, demographic information, and spatial data to create a framework that can be used to identify patterns, correlations, and influential factors associated with serious crimes in South Africa. The predictive model is expected to enable law enforcement agencies and policymakers to anticipate crime trends, allocate resources effectively, and develop proactive prevention and intervention strategies. Additionally, the model is expected to consider the dynamic nature of crime with respect to certain influential factors like the socio-economic inequalities, unemployment rates, and organised crime networks in the South Africa context. This research problem statement spurred some research questions which are presented in the next section.

1.5 Research Questions

The following research questions have all been taken into consideration to help understand and analyse the problems more constructively.

1.5.1 Research question 1

What factors have influenced changes in serious crime rates over time in South Africa (for example, socioeconomic status, demographics, environmental factors and education level)? Is the data associated with these identified factors useful to classify areas and provinces based on their risk of serious crime?

1.5.2 Research question 2

Are data related to demographics, socioeconomic status and education level relevant to identify areas and provinces based on their risk of serious crime?

1.5.3 Research question 3

What data analytic techniques could be applied to analyse data to develop an effective predictive model in a bid to forecast future serious crime trends in South Africa?

1.6 Research Objectives

As serious crimes are still happening around the world, it is better for law enforcement authorities and the government to always deal with such issues with all seriousness. It would not only give an upper hand to the law enforcement officer protecting the neighbourhoods but also build awareness of why serious crimes are committed in the first place. This research aimed to use data analytics techniques to analyse and predict serious crime trends in South Africa in to provide appropriate solutions or recommendations to security agencies.

The following objectives have been taken into consideration to help achieve the research aim and address the research question.

Objective 1: Identify the factors that contribute to serious crime in South Africa.

Objective 2: Collect and analyse crime data from various sources based on objective (1) above to identify trends and patterns using data analytics techniques.

Objective 3: Develop a predictive model and evaluate its effectiveness in predicting serious crime trends in South Africa.

1.7 Research Methodology

The research paradigm, research design and research methodologies to address the research question, aims, and objectives are described in this section.

1.7.1 Research paradigm

Post-positivist acknowledges the fallibility of observations and advocate for revisable theories. In contrast, positivists persistently strive for objective truth, even if absolute knowledge remains unattainable through social science alone (Apuke, 2017; Chih-Pei & Chang, 2017; Creswell & Creswell, 2017; Khaldi, 2017).

1.7.2 Research design

The research context is provided to support the research design choices. There is a country-specific need to develop effective predictive models in understanding, mitigating, and preventing criminal activities. By harnessing data science, this

research investigates techniques into the dynamics of serious crime and contribute to the development of proactive strategies for enhancing public safety in South Africa. This study therefore adopts a quantitative methodology using data science.

An overview of the various research methodologies is provided below:

1.7.2.1 Quantitative research methodology

Quantitative research evaluates objective concepts by examining the relationship between variables. The purpose statement for quantitative research stresses the measurable relationships between variables allowing numerical data to be analysed with statistical procedures. The two types of quantitative research are experimental and non-experimental (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017). Quantitative analysis therefore has relevance with respect to crime detection by using this numerical data and interpreting the statistical and analytical knowledge such as numbers, frequency and locality. (Apuke, 2017; Joshi, Sabitha and Choudhury, 2017). A quantitative methodology analysis tool may also be used to analyse collected data to predict occurrences of future serious crimes (Shamsuddin, Ali and Alwee, 2017).

1.7.2.2 Qualitative research methodology

Qualitative research investigates and analyses the significance that individuals or groups attach to a human or social situation. Emerging questions and methods are components of the research procedure. Qualitative research is a purpose statement concerning the point of attention, while understanding and interpreting social interactions is its primary goal. This type of research can be either interactive or non-interactive (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017).

1.7.2.3 Mixed-methods research methodology

The use of mixed methods combines or links qualitative and quantitative approaches to investigation. It involves utilising qualitative and quantitative methodologies as well as combining the two in a study. Mixed-methods research has the advantage of combining the strengths of quantitative and qualitative approaches while compensating for the flaws of each (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017).

1.7.3 Research approach

This study adopts an experimental research approach, as data analytics tools would be used to analyse serious crime trends. There are several methodologies used in research, including knowledge discovery in databases (KDD); cross-industry standard process for data mining (CRISP-DM); and sample, explore, modify, model, assess (SEMMA), among others (Wirth and Hipp, 2000; Debuse *et al.* 2001; Azevedo and Santos, 2008; Berwind *et al.* 2016; Edoka, 2020; Obagbuwa and Abidoye, 2021).

In this study, CRISP-DM, a quantitative methodology was adopted, as this methodology is the most widely used (Berwind *et al.* 2016; Edoka, 2020; Razali, 2020; Obagbuwa and Abidoye, 2021). CRISP-DM methodology is incredibly efficient, well-suited for data mining and predictive data analytics projects, and is widely used in data mining research (Obagbuwa and Abidoye, 2021). Tasks can be completed in any order, and it is often necessary to go back and repeat actions since this is an iterative method. The CRISP-DM (Berwind *et al.* 2016) methodology would be used for data collection and analysis of serious crime datasets in order to predict serious crime trends in all provinces of South Africa and mitigate serious crimes in hotspot areas. The CRISP-DM data mining life cycle is divided into six phases (Razali, 2020), which the researcher would use for the data extraction and analysis, including business understanding, data understanding, data preparation, modelling, evaluation and deployment (Razali, 2020). A more in-depth explanation is provided in Figure 3.1.

1.7.4 Instruments

The datasets used in this research were obtained from the Crime Statistics for South Africa on the Kaggle website[1] (Kaggle, 2021) and the SAPS website[2] (SAPS, 2022). Various data analytic techniques, including machine learning algorithms, time series analysis and spatial analysis, are employed to build models for predicting serious crime trends using the CRISP-DM methodology. Some of the techniques are pre-trained tools that deployed in Python packages and libraries.

[1] Crime Statistics for South Africa | Kaggle

[2] https://www.saps.gov.za/services/crimestats.php

1.7.5 Data collection

No primary data collection was carried out in this research. However, a secondary data collection was done by gathering available serious crime data related to the South African context from both the Kaggle and the SAPS websites as presented in *Instruments* (1.7.4). The serious crime dataset from 2005 to 2020 was extracted from the South African crime statistics database. For ethical purposes, personally identifiable information (PII) was anonymised before the analyses were conducted as per best practices. This study chose popular data repositories Kaggle and SAPS because they contain vita information regarding serious crimes in South Africa that is publicly available to researchers without cost (Kaggle, 2021; SAPS, 2022).

1.7.6 Prediction

For serious crime trends and prediction purposes, ordinary least squares (OLS) regression was employed due to the ability of OLS to fit the function well to the data.

1.8 Significance of the Study

This research analyses serious crime trends and identifies serious crime hotspots in South Africa's nine provinces. The insights obtained from the crime trends prediction provide helpful information that can be disseminated to individuals or organisations seeking to build new strategies to combat serious crime. It would also strengthen the knowledge capacity of the SAPS and relevant agencies. Moreover, analysis and prediction of serious crime trends using data science tools and techniques would aid authorities in combating serious crimes and benefit South Africa. This is a gap. Specifically, by developing accurate reliable predictive models for crime rates, several key benefits are achieved, including crime prevention, resource optimisation, policy development and intervention strategies, community engagement and public awareness, and early warning systems to alert law enforcement agencies and relevant stakeholders about potential spikes in serious crimes, evaluation of policy effectiveness and aiding international collaboration and knowledge sharing. These potential gains have been presented in the section on *Research Rationale.*

1.9 Contributions of the Study

The primary contributions from this research work are as follows:

i) The key factors that contribute to serious crimes in South Africa were identified.
ii) Data analytic techniques, including machine learning algorithms, time series analysis, and spatial analysis, were employed to build accurate models for predicting serious crime trends following the CRISP-DM methodology steps.
iii) The evaluation of the developed predictive model revealed its effectiveness in predicting serious crime trends in South Africa.

1.10 Delimitations

This study does not cover all crime. It focuses on serious crime because the rates are notably higher than those of other crime (Mathews *et al.* 2019). This study applied a fifteen-year dataset from 2005 to 2020 of South African crime statistics (Kaggle, 2021; SAPS, 2022). Linear regression was used to predict the model in this study.

1.11 Chapter Synopsis

The remaining chapters of this study are arranged as follows:

Chapter 2 evaluates previous and recent literature relevant to the research aim, and objectives of this study. The literature review describes the country's extensive criminal history and outlines past strategies used to address the crime problem.

Chapter 3 presents the research methodology used. It shows how data analytics techniques like machine learning, time series analysis, and spatial analysis, are employed to build accurate models for predicting serious crime trends following the CRISP-DM methodology steps. The chapter details the data sets and tools used.

Chapter 4 presents and discusses the results and visualisations obtained from the application of some data analytics techniques, including machine learning algorithms, time series analysis, and spatial analysis, in developing a predictive model for the serious crimes in South Africa. This includes the results of the evaluation of the model developed. The results of the predictions were compared to current crime trends.

Chapter 5 summarises the results and concludes with a summary of the contributions and limitations of the study, and future research suggestions.

1.12 Chapter Summary

This chapter has explained the study's background, introduction, research rationale, research problem, significance, aim, objectives and contributions. In the next chapter, a detailed literature review is presented that examines the theoretical underpinnings and literature backgrounds regarding serious crimes in South Africa and methods that have been developed in recent times to address or mitigate such crimes from becoming widespread in the country.

Chapter 2 Literature Review

2.1 Introduction

Serious crime is a significant issue that needs to be resolved before it becomes unsolvable (Obagbuwa and Abidoye, 2021). There is no doubt that South Africa is one of the affected countries in Africa with a high incidence of serious crimes (Bhorat, Thornton and Van der Zee, 2017; Monyeki, Naicker and Obagbuwa, 2020; Obagbuwa and Abidoye, 2021). Serious crimes can be committed by either a male or female perpetrator (Almond *et al.* 2017; Velopulos *et al.* 2019). This study is being conducted with this assumption in mind. Previous works on serious crime analysis have identified unemployment, inadequate educational attainment, a lack of social and occupational skills, alcohol misuse, substandard housing and living conditions, and poor parenting abilities as potential factors for serious crimes in South Africa (Mazorodze, 2020; Monyeki, Naicker and Obagbuwa, 2020; Monyeki, 2021). These factors, which have had a negative impact on people's lives, are believed to be the major causes of serious crimes such as murder and sexual and drug-related crimes (Mazorodze, 2020; Obagbuwa and Abidoye, 2021). Cheteni, Mah and Yohane (2018) argue that hard times have hurt young people in poor areas the most, making them angry and leading to more crime. The authors further argue that the neglect of poverty among jobless youth has contributed to increased crime (Cheteni, Mah and Yohane, 2018).

Serious crime is a global problem that keeps getting worse (Altinyelken and Le Mat, 2018; Chang *et al.* 2019). Few countries have been able to control this problem, which hurts whole countries as much as it hurts individuals (Monyeki, 2021). This study is essential because it provides information about serious crime trends that authorities can use to combat crime in a way that benefits the public. Table 2.1 depicts a snapshot of the country’s yearly serious crime rates for the three selected crimes from 2005 to 2020 (Kaggle, 2021; SAPS, 2022). The dataset consists of 18 columns and 3495 rows. Table 2.1 consists of the first five rows and the last five rows of the dataset.

Table 2.1 A Cross section of the South African serious crimes statistics (2005-2020) (Source: Kaggle, 2021; SAPS, 2022)

	Province	Station	Category	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020
0	Western Cape	Mitchells Plain	Drug-related crime	3064	3683	4792	5699	6571	6260	5850	6310	6044	4768	4609	4914	4930	3475	3783
1	KwaZulu-Natal	Phoenix	Drug-related crime	1950	1809	1270	622	1085	1180	1452	1785	2043	2323	2653	3224	2184	1712	1672
2	Western Cape	Cape Town Central	Drug-related crime	1171	1303	1457	1393	1474	1539	1832	1963	2149	2360	2712	2796	2894	2313	2052
3	Western Cape	Bishop Lavis	Drug-related crime	1091	1333	1348	1430	1759	2459	2753	2977	2577	2738	2472	2898	3432	1855	2009
4	Western Cape	Steenberg	Drug-related crime	989	1174	1234	1588	1656	1711	1352	1630	1273	1456	1831	1698	2444	1847	1897
	....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
3490	Western Cape	Samora Machel	Murder	0	0	0	0	0	0	0	0	0	0	0	0	0	30	106
3491	Western Cape	Samora Machel	Sexual Offences	0	0	0	0	0	0	0	0	0	0	0	0	0	32	108
3492	Western Cape	Simon's Town	Drug-related crime	0	0	0	0	0	0	0	0	0	0	0	48	21	34	5
3493	Western Cape	Simon's Town	Murder	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1
3494	Western Cape	Simon's Town	Sexual Offences	0	0	0	0	0	0	0	0	0	0	0	13	4	3	10

2.2 Global Crimes

Crimes are evolving today in different more advanced ways, even for reasons previously unknown to humans (Arifin, 2020). Globally, governments face critical issues every day in the form of crime (Saeed *et al.* 2015). Along with the evolution of technology and techniques that are being used to allow highly complex criminal activities, crime continues to be a serious challenge to all societies and nations around the world (Hassani *et al.* 2016). Crime is one of the most prevalent and troubling facets of our society, as described by Pednekar *et al.* (2018), and preventing it is a critical challenge. Arifin (2020) emphasised in his criminology article that different variables, ranging from the surroundings to the economy, schooling and even family home

issues, cause different behaviours in adolescents. A violent crime has been described by the Federal Bureau of Investigation as an offence involving violence or threat, which they classified in four categories: murder, rape, robbery and worsened abuse (McClendon and Meghanathan, 2015; Prabakaran and Mitra, 2018). The authors also emphasised four types of crime: fraud detection, road abuse, violent crime, and cybercrime and sexual offences. Arifin (2020) emphasised that variety of crimes, especially on the streets, such as robbery, pickpocketing, assault, combat, drugs and substance abuse, or even sexual assault usually occurs as result of broken homes and dropouts in the Ngaliyan area and a few other regions in Indonesia.

These types of crimes seem to have become ubiquitous in society; for example, Los Cabos, Mexico, had the world's highest murder rate in 2020, with 111.3 murders per 100 000 people (Statista Research Department, 2021). Crime has gained greater exposure in recent years, thereby requiring deeper scrutiny of crimes and violent crimes by police and the public around the world (Justice, 2011; Hossain *et al.* 2020; Statista Research Department, 2021).

According to the Bureau of Justice Statistics 2018 survey discussed by Akpinar and Chouldechova (2021), due to fear of retaliation, the victims' perceptions that their victimisation was minor or may be viewed as such by police, and personal relationships with the perpetrator, only 25% of rapes/sexual assaults, 61% of aggravated assaults, 63% of robberies, 38% of simple assaults, and 61% of simple assaults are reported to the police (Akpinar and Chouldechova, 2021).

With the increasing population and rapid growth of towns and cities, Mittal *et al.* (2019) examined crime trends in India that are constantly evolving, and the major worrying factor for the Indian Government is the rise in crimes in any area, particularly crimes against females, kids, and other vulnerable members of society (Mittal *et al.* 2019).

Justice (2011) asserted that the purpose of researching crime is to deter it, and the collected crime data is to understand the reasons for crime occurrence. Hence, this knowledge is needed to inform policy responses to minimise potential crime and its harmful repercussions. Hossain *et al.* (2020) described crime as an activity that constitutes a criminal offence under the law, and the prevention of crime has been a

great challenge to society and the nation at large. Crime analysis using data is vital for the prevention and resolution of crime. Law enforcement is faced with the challenge of interpreting the growing amount of crime data correctly and effectively. Without the help of computerised machines, the analysis of crime data could put human beings under pressure because the strength of humans cannot withstand millions of data points (Shamsuddin, Ali and Alwee, 2017).

Ratul and Rab (2020) examined a 478 578-incident real-world crime and accident dataset from Denver County, USA, between January 2014 and May 2019. Based on the prediction rates, the project aims to predict and highlight occurrence trends, which would assist law enforcement agencies and the government in identifying preventive actions (Ratul and Rab, 2020).

Violence against girls and women violates human rights and is a significant health concern worldwide; the overall findings of Öberg, Heimer and Lucas (2020) indicate that different forms of brutality are prevalent among females in Sweden during their early years, teenage years, and maturity.

Obagbuwa and Abidoye (2021) employed machine learning and linear regression techniques to extract meaningful information from crime data from South Africa's nine provinces to develop a predictive model to assess crime data trends and predict future crime, and their results showed a good prediction.

2.3 A Brief Overview of Serious Crime Trends

The serious crime trend theory explains why people commit crimes in specific areas (Santos, 2016). Serious crimes can include but are not limited to murder, aggravated and simple assault, rape and sexual assault, and robbery (Teker *et al.* 2017; Maluleke and Dlamini, 2019). According to the theory, crime occurs when a victim's or target's activity space intersects with an offender's activity space (Santos, 2016). A person's activity space is made up of places they visit daily, such as their home, work, school, shopping and entertainment (Ronald and John, 2014). These private locations are referred to as nodes (Brantingham and Brantingham, 2013). Personal paths are the paths or routes that people take to and from these nodes (Brantingham and

Brantingham, 2013). A perimeter is formed by personal paths connecting to various nodes. The awareness space of a person is defined by this perimeter (Brantingham and Brantingham, 2013; Wortley and Townsley, 2016). By discovering where, when, and why specific crimes are likely to occur, crime trend analysis reveals the underlying interactive process between crime events (Wang and Zhang, 2020). According to crime trend theory, a crime between an offender and a victim or target can only happen when their activity spaces cross (Wang and Zhang, 2020). Simply put, if an area provides the opportunity for crime and it is within an offender's awareness space, then crime would occur (Andresen and Felson, 2010; Brantingham and Brantingham, 2013; Santos, 2016). Crime trend theory provides analysts with an organised way to explore trends in behaviour.

It is likely that offenders who committed these crimes in the location where they contacted their victims first assessed the place as suitable for the crime. Other offenders deemed the contact location unfit for crime completion, most likely due to the presence of others, and chose to commit the offence elsewhere later (Wang and Zhang, 2020). In the next section, an overview of serious crimes related to sex, murder and drugs in the South African context is discussed.

2.3.1 South Africa's sexual crime landscape

Sexual crimes include but are not limited to sexual assault, sexual exploitation, sexual harassment, sexual trafficking, and rape (Mathews and Collin-Vézina, 2019). Unfortunately, in South Africa, women are the main targets of sexual crime and rarely have a protective option to negotiate safer sex; males hold the power, and women are disadvantaged mainly due to the patriarchal system (Pitcher and Bowley, 2002; Wojcicki, 2002). Due to their deep roots in the oppressive patriarchies of colonialism, apartheid, and the Cold War, these deeply ingrained patterns of sexual crime did not vanish with the transition to democracy (Britton, 2006).

It is important to note that South Africa has a high incidence of critical rape risk factors, with many men currently enduring childhood trauma and a strong gang ideology (Falope and Thakur, 2022a). These issues serve to legitimise women's maltreatment in many countries. However, sexual assault is not unique to South Africa and is a result of the post-apartheid era. As a result, Soweto and certain informal townships in

Johannesburg's south became regarded as the world's rape capital (Mabasa, 2009). Race, on the other hand, should not be used as a legitimate indicator because rape occurs in all racial groupings and enclaves (Holzman, 1996). The high rate of sexual crime in this area reflects the harmful culture instilled in these communities by apartheid (Holzman, 1996; Wojcicki, 2002).

Numerous incidents of sexual crime are not reported or spoken about, not only to law enforcement authorities but also to victims' families and colleagues. According to the Medical Research Council (MRC), the number of sexual crimes may be up to nine times higher than the number of cases reported (Jewkes and Morrell, 2010). According to the South African Police Service (SAPS), 7 071 500 sexual crime cases were reported as complaints by victims or families of victims in 2017/2018. It is reasonable to encourage critical thinking about the incidences of sexual crimes in South Africa since, if they are closer to 643 500 a year, it would mean that someone is sexually violated or indecently attacked every minute in South Africa. In any event, with so many sexual violations not being reported and recorded cases of violations being classified as sexual crimes in South Africa, determining the true nature and extent of sexual violence is almost impossible (Oyasor, 2020).

Police Minister Bheki Cele said at a media briefing that according to the crime statistics in 2020, the number of sexual crimes had increased by 5% and that there would be more GBV-related crimes in the fourth quarter of 2020. Statistics showed that there were 12 218 reports of rape between October and December 2020, which is a 1.5% increase from July to September. Most cases were reported in Inanda and Umlazi in KwaZulu-Natal and Lusikisiki in the Eastern Cape. The location of occurrence for the rape cases shows that over 4 900 occurred at the victim's or rapist's home, 570 were domestic violence-related, and 547 of the rape cases in this category involved female victims, and 23 were males (Health-e News, 2021; Nganga, 2021).

Rape made up the majority of the 12 133 sexual crime cases reported from January to March 2021. In the crime statistics, Gauteng, KwaZulu-Natal, and the Eastern Cape reported 2,031, 1,722, and 1,660 cases respectively, with Gauteng having the highest incidence rate. One gender activist highlights that the number of rape cases recorded during that period underscores the inadequacy of efforts to combat GBV in the country,

while a researcher emphasizes that the police should not only release statistics but also devise strategies to actively reduce rape cases (Ismail, 2021). Online sexual predators must stop hurting more people, including children. (Chassiakos *et al.* 2016; D'Angelo and Moreno, 2019) said that most parents do not realise how dangerous it is for their kids to talk to strangers online. They said parents need to start protecting their kids online at a young age, set rules for Internet use, and use software to keep an eye on what their kids do online. Concerns about sexual harassment should be reported to healthcare workers right away.

Figure 2.1 presents the South African trends on sexual crimes per province from 2005 to 2020 (Kaggle, 2021; SAPS, 2022). On the graph, the y-axis shows the total number of sexual crimes, and the x-axis shows the time period. It is evident from the figure that Gauteng has the highest number of sexual crime cases, followed by KwaZulu-Natal and the Eastern Cape.

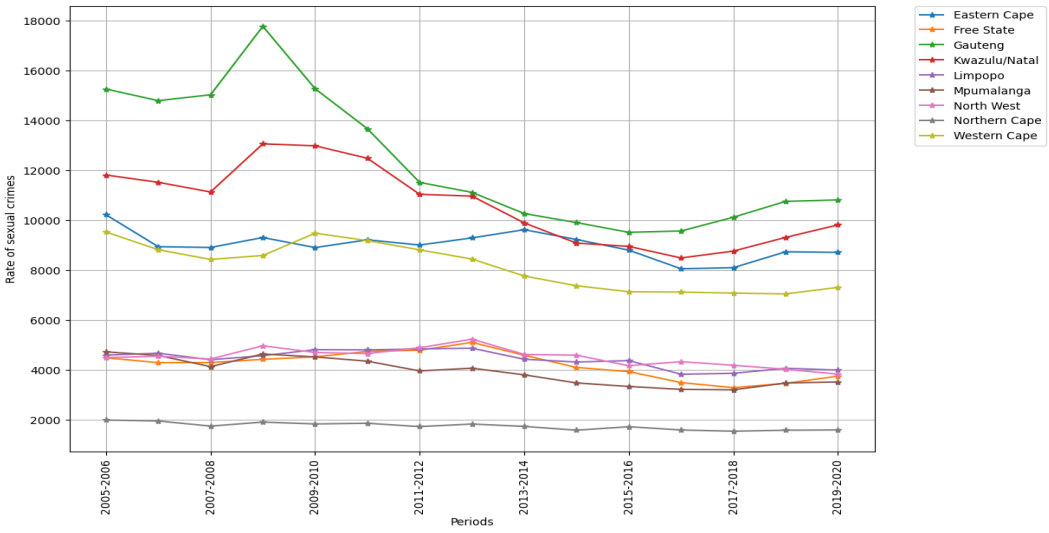


Figure 2.1 South African sexual crimes trends per province (2005 – 2020) (Source: Kaggle, 2021; SAPS, 2022)

2.3.2 South Africa drug-related crime

The problem of drug-related crime affects almost every nation on earth (Thamrin and Liao, 2018) kratom, cocaine, methamphetamine crystal, heroin, and methadone are among the most popular illicit substances (Rigoni, Breeksema and Woods, 2018). Some claim that these substances' effects are comparable to those of cocaine; however, they are more prone to severe health problems, such as aggressive behaviours, high blood pressure, palpitations, paranoia, and chest pain (Monyeki, 2021). Drug-related crimes can facilitate all other types of crimes and put a lot of stress on societies (de Bont *et al.* 2018). Psychopharmacological, economic-compulsive, drug law, and systemic crimes are all classified as types of drug-related crimes (European Monitoring Centre for Drugs and Drug Addiction, 2007; de Bont *et al.* 2018). Individual personality and the environmental, biological, social and cultural background in which drug use takes place, all have an impact on drug-related issues (de Bont *et al.* 2018). Adolescent and adult illicit drug users in the US increased from 27 million in 2014 to 53.2 million in 2018 (Xia, Stewart and Fan, 2021). Coronado and Saucedo (2019) researched the effects of drug-related crimes on employment in Mexico, and their results indicated that drug-related crimes negatively impacted employment. The results indicated that a 10% increase in drug-related crimes reduces total employment by up to 0.9%. They also found that skilled employment responds at an increasing rate when drug-related crimes skyrocket (Coronado and Saucedo, 2019).

South Africa has experienced a huge rise in the amount and types of illegal drugs being made, sold, and used in the past few years, which has led to more crime and health risks in society (Machethe and Mofokeng, 2022). South Africans sometimes feel that most crime waves in the country, like drug use, violence, murder, sexual crimes, and break-ins, are caused by immigrants (Kollamparambil, 2019; Singh, 2020). The crime statistics released by Police Minister Bheki Cele in 2020 showed that the misuse of alcohol and drugs is primarily responsible for murders (Mahlakoana, 2020). Cheteni, Mah and Yohane (2018) used the autoregressive-distributed lag error correction model to analyse data a study in South Africa to find the link between poverty and drug use. The model found a strong link between crime and poverty in the

short and long term. The study also showed that poverty is a factor in many drug-related crimes in the country (Cheteni, Mah and Yohane, 2018).

Figure 2.2 presents the South African trends on drug-related crimes per province from 2005 to 2020 (Kaggle, 2021; SAPS, 2022). The y-axis shows the total number of drug-related crimes, and the x-axis shows when the crimes happened per province.

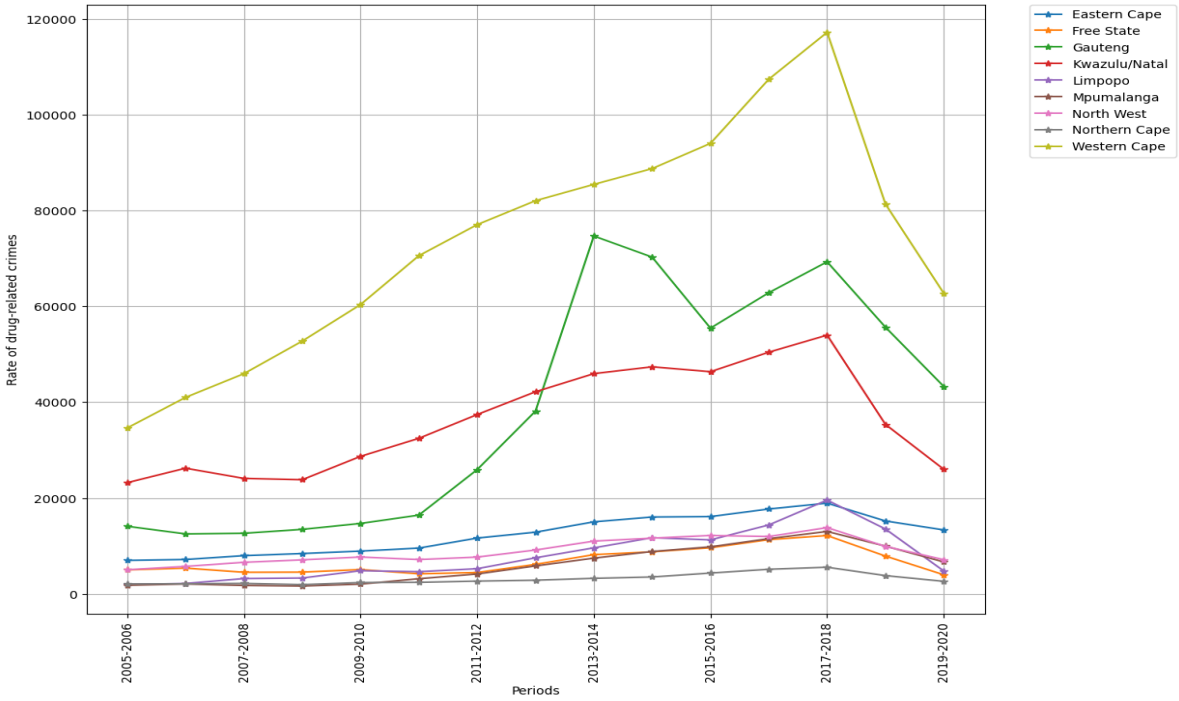


Figure 2.2 South African drug-related crimes trends per province (2005 – 2020) (Source: Kaggle, 2021; SAPS, 2022)

The Figure 2.2 shows that the Western Cape has the highest incidences of drug-related criminal cases, followed by Gauteng and KwaZulu-Natal. The literature has also shown that the Western Cape has had the highest drug-related crime rate in South Africa (Nyabadza and Coetzee, 2017; Kaggle, 2021; SAPS, 2022).

2.3.3 South Africa murder crime

Murder crime in South Africa has emerged as one of the highest in Africa (Otieno *et al.* 2015; Osuafor and Okoli, 2019; Monyeki, Naicker and Obagbuwa, 2020). People often see murder as a serious mental health and criminal justice problem that puts a huge amount of stress on people and society at large (Moen, 2020). In South Africa,

between 2012 and 2013, there were 31.1 murders and 355.6 assaults with the intent to cause great bodily harm per 100 000 people (South African Police Service, 2013). From 2017 to 2018, 167 352 assaults with the intent to cause great bodily harm were reported (South African Police Service, 2018). Murder is a severe form of violence that reduces the number of years that a society or a nation can expect its citizens to live (Otieno *et al.* 2015). Studies on the rise in murder in South Africa have previously been conducted and published in the literature (Otieno *et al.* 2015; Matzopoulos *et al.* 2018; Mathews *et al.* 2019; Monyeki, Naicker and Obagbuwa, 2020).

Murders in South Africa are predominantly the result of domestic violence and sexual harassment (Lindegaard, 2017). Women are typically murdered in their homes, whereas men are mostly murdered in public. Arguments frequently result in the murder of males, whereas sexual harassment frequently results in the murder of women (Lindegaard, 2017). The number of crimes increased after apartheid, and serious crimes like assault, robbery, and murder occurred more often in urban areas as people go there in search of job opportunities. Many studies have found that poverty and inequality are the main reasons people commit crimes (Ghani, 2017; Sampson and Wilson, 2020). Murders have also been caused by police brutality. In one case, the Northwest Province SAPS killed 34 striking miners and injured 78 others in Marikana. The police said they did this in self-defence. Figure 2.3 presents the South African trends in murder per province from 2005 to 2020 (Kaggle, 2021; SAPS, 2022). The y-axis shows the total number of murders, and the x-axis shows when they happened.

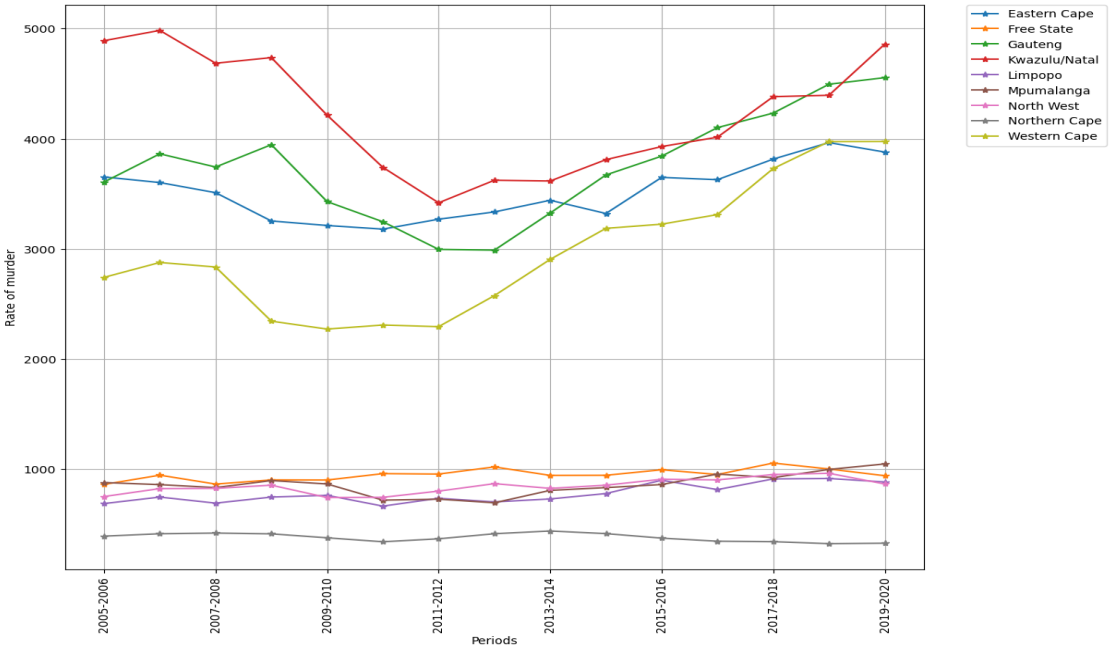


Figure 2.3 South African murder trends per province (2005 – 2020) (Source: Kaggle, 2021; SAPS, 2022)

The crime of murder is one of the most common crimes in South Africa (Bhorat, Thornton and Van der Zee, 2017; Maluleke and Dlamini, 2019; Kaggle, 2021; SAPS, 2022). As shown in Figure 2.3, murder crime in provinces such as KwaZulu-Natal, Gauteng, and the Eastern Cape has been on the rise, and researchers have come up with several reasons for these changes, such as inequality and unemployment, bad environmental conditions, and a low standard of education in general (Kollamparambil, 2019; Mazorodze, 2020; Monyeki, Naicker and Obagbuwa, 2020; Obagbuwa and Abidoye, 2021).

2.4 Data Analytics

Data analytics is the general idea of improving data mining, knowledge discovery, and machine learning (Sarker *et al.* 2020). This means making algorithms and programs that can learn on their own, along with the original data analysis and descriptive analytics from a statistical point of view (Sarker *et al.* 2020). Data collection, organisation, pre-processing, transformation, modelling, and interpretation are all included in this process (Moreira, Carvalho and Horvath, 2018). Data analytics is the process of analysing data sets with the help of specialised software and systems to

derive conclusions about the information they contain (Choi, Wallace and Wang, 2018). Data analytics is used by researchers and scientists to confirm or refute scientific hypotheses, models, and theories (Karpatne *et al.* 2017). Commercial businesses also use data analytics techniques and technologies exclusively to help them make better business decisions (Insights Desk, 2020b). Data analytics is related to data science, business analytics, and business intelligence (Vassakis, Petrakis and Kopanakis, 2018).

2.4.1 Tools for Analyzing Data

Data analytics is indeed a useful technology that allows criminal investigators who lack considerable data analysis experience to investigate enormous databases quickly and efficiently (Cao, 2017). Computers can process thousands of instructions in seconds, saving valuable time (Kelly *et al.* 2018). Furthermore, the cost of installing and running software is sometimes cheaper than the cost of hiring and training new employees (Lamberton, Brigo and Hoy, 2017). Human investigators, especially those who work long hours, are more prone to mistakes than computers. Crime data analytics tools can help police operations by enhancing efficiency and minimising errors, allowing investigators to focus on more important tasks (Eck and Rossmo, 2019). The correct and effective analysis of increasing volumes of crime data is a major challenge for all law enforcement and intelligence-gathering operations (Wang and Siau, 2019). Complex conspiracies, for example, might be difficult to decipher since suspect information is often dispersed geographically and over long periods (Wang and Siau, 2019).

Data analytical tools such as RapidMiner, Weka, KNIME. Xplenty, Octoparse, Splunk, R-programming, Python, Microsoft Excel, and Power BI are widely used by many organisations for data processing (Vassakis, Petrakis and Kopanakis, 2018; Insights Desk, 2020a). Data analytics tools have been applied to security and criminal investigation during the last three decades, demonstrating criminal analytics' creation and growth (Pramanik *et al.* 2017). In short, data analytics has the potential to change how law enforcement and security intelligence agencies get important information (like criminal networks) from many data sources in real time to help their investigations (Pramanik *et al.* 2017).

Data analytics can be a useful tool for crime reduction and prevention. This is attributable to three variables, according to several scholars (Lamberton, Brigo and Hoy, 2017; Wang and Siau, 2019; Wainana *et al.* 2020). First, vast amounts of data exist since security organisations and agencies collect as much information as possible while investigating crimes (Wang and Siau, 2019). This allows them to monitor and investigate potential crimes, as well as prevent crimes from occurring again. Second, they have the potential to find trends in enormous data sets, making data analytics a more dependable technique than manual data processing, which is time-intensive and ineffective for large data sets (Brown, 1998; Wainana *et al.* 2020). Third, the deteriorating economy has left security agencies with a tight budget that prevents them from paying their employees and conducting adequate investigations into crimes. Within their limited budget, data analytics makes it easier and more efficient for them to uncover crime trends that are useful to them (Lamberton, Brigo and Hoy, 2017). To study the process and discover trends in computer crime, Abbasabadei *et al.* (2020) used data analytics and MATLAB software for data modelling. Their research demonstrated that the system could be considered one of the most successful and low-cost techniques to identify cyber-criminal conduct; thus, computer crime experts can effectively run this model on their systems (Abbasabadei *et al.* 2020).

Al-Hashedi and Magalingam (2021) analysed state-of-the-art research in financial fraud detection from 2009 to 2019 and classified it based on the forms of fraud and data analytics technology used to detect financial fraud. The review provided a sample of 75 relevant publications, which were divided into four categories (cryptocurrency fraud, bank fraud, financial statement fraud, and insurance fraud). According to the report, 34 data analytics tools were used to detect fraud across a variety of financial applications. The SVM is one of the most extensively used financial fraud detection approaches, accounting for 23% of the total study, followed by both naive Bayes and random forest, which account for 15% each. The review found that most data analytics tools are widely used in bank and insurance fraud, with 61 research publications out of 75 accounting for 81.33 per cent of the total number of articles. The review also provided a valuable reference source for both academic and practical businesses in directing the identification of financial fraud, with relevant information on the most

important data analytics tools utilised and a list of nations that are vulnerable to financial fraud (Al-Hashedi and Magalingam, 2021).

AlJanabi and Hayda (2010) proposed a crime and criminal analysis model using the K-means clustering method and the Apriori algorithm. The model was based on a comprehensive collection of over 350 crime raw datasets collected from the police departments of Libyan cities including Benghazi, Tripoli, and Al-Jafara Supreme Security Committee. They analysed the data with the help of applications such as Google App Engine, WEKA, and Microsoft Excel. The major goal of this research is to assist Libyan security officials in detecting criminal activity and determining the relationship between criminal age and the type of crime committed in Libya (AlJanabi and Hayda, 2010).

Sharma and Kumar (2013) conducted a survey and discovered that previous crime records could be useful in predicting future crime occurrences. They discovered that the most widely used data mining techniques were classification and clustering. They concluded that advancements in clustering could improve classifier evaluation.

The perception of crime trends and suspect predictions were reviewed by Das and Nayak (2021), who found a variety of limitations to consider when predicting suspects based on the type of crime. During chain snatching, for example, the limitations could include location, time, criminal height, appearance, type of vehicle, weapon used, colour, and so on. As a result, the article provided information regarding the strategies used to predict crime patterns and suspects.

David and Suruliandi (2017) conducted numerous surveys on supervised and unsupervised learning approaches often used for criminal identification. They were able to learn about many data mining approaches for assessing and predicting future crime from the survey.

A new method, ‘Series Finder’ developed by Wang *et al.* (2013) for pattern detection. The identification of crime trends by the same perpetrator proved useful. It was able to match crimes with patterns that experts could miss. According to the functioning mechanism, the algorithm looked for commonalities between crimes and sought to identify a particular offender’s modus operandi. The modus operandi could be

increasingly defined as the number of offences increases. This series finder is flagged in particular for South Africa due to the unusually high number of serial rapists (Chi *et al.* 2017).

Zhang *et al.* (2011) used a Bayesian network technique to develop a whole new framework for detecting phishing websites. Their system was capable of distinguishing between the original website and the suspect website. They used classifiers such as text, pictures, or a mix of the two to discriminate between the actual and fake web pages. They were able to estimate the matching threshold.

In a comparison of machine learning regression-based algorithms, Gonzalez and Leboulluec (2019) identified and examined crime trends based on social variables, such as per capita income and educational attainment. They did this by using Python and the statistical analysis system. The crime dataset was subjected to four machine learning techniques (multiple linear regression, regression using random forests, regression using neural networks, and regression using Bayesian inference), with random forest outperforming the others ($R^2 = 0.971$) (Gonzalez and Leboulluec, 2019). SPSS, R Development Core Team (2012), Python, and Perl are just a few examples of commonly used data analytics tools (Wainana *et al.* 2020). The goal of this research was to analyse and predict serious crime cases in South Africa using some data analytics tools.

2.4.2 Exploring Methods of Data Mining

Data mining is the process of extracting or mining information from enormous amounts of data (Osman, 2019). It entails analysing and reviewing vast pre-existing datasets to produce new knowledge that may inform more relevant applications in practice (Alasadi and Bhaya, 2017). Time series analysis, machine learning techniques, and spatial analysis are examples of data mining approaches (Bhowmik, 2008; Osman, 2019). Although data mining approaches have been applied to many research domains, limited effort has been directed towards their application in the study of crime (David and Suruliandi, 2017). Crime data analytics uses data processing techniques in crime investigations, following the rising crime rate that is the product of high-tech advances as well as an increase in population. Crime detection is a structured

evaluation that distinguishes and defines the paradigm of illegal activity, which has become one of the most important law enforcement tasks in the real world. The usage of the criminal report is a fascinating way to deal with criminal physiognomies to help people maintain a healthy lifestyle (Jayaweera *et al.* 2015; Prabakaran and Mitra, 2018).

Data mining techniques can be used to detect and deter illegal activity. Compared to old-style data mining techniques, novel approaches focus on structured and unstructured data for pattern detection. Many of the existing solutions are now using variations of data mining techniques to achieve simpler and more precise mining with the introduction of big data. The high prevalence of physical, mental, and sexual exploitation of children in South Africa is attributed to acute or even long-term negative health effects. Global relief organisations, such as charitable programs like the United Nations International Children's Emergency Fund (UNICEF), in partnership with the national government, have acknowledged the compelling need for coordinated hospitals, police, and civil and criminal justice activities to assist victims of sexual abuse. To ensure the welfare of the general population, the protection of crime is one of the key responsibilities of every government. To support the government and the public justice agencies discover ways to deter violence by using predictive statistics, detect criminals in advance, disperse state budgets, and detect problems that cause illegal conduct. An accurate model for detecting and predicting crime can be used (Bello-Orgaz, Jung and Camacho, 2016; Hassani *et al.* 2016; Meinck *et al.* 2017; Ratul and Rab, 2020).

Data mining is an important technique that provides a wide variety of ways to support felony investigators in focusing more on the major information hidden inside the Big Data report. In addition to increasing the use of technological systems in crime detection, law enforcement agents and police are also being aided by machine learning specialists to fasten the protocols for apprehending the suspect and to develop a statistical algorithm for predicting future crimes (Hassani *et al.* 2016; Farsi *et al.* 2018; Feng *et al.* 2019). Association rule mining, classification, prediction, and clustering are the four major areas of data mining (Farsi *et al.* 2018). In recent years, several data mining techniques have been proposed for overcoming the problem of

extracting knowledge from explosive data using various algorithms. One of these applications is the detection of criminal activity from existing data by examining the frequency of occurrences (Awal *et al.* 2016; David and Suruliandi, 2017; Farsi *et al.* 2018; Yerpude, 2020).

2.5 Related Works on Data Analytics in Crime Trend Prediction

In this section, several related works that have adopted data analytics techniques in their analyses of crime trends and predictions are presented and discussed. Machine learning is a part of artificial intelligence and a key component of digitalisation solutions that has attracted a great deal of attention in the internet world. It deals with mathematical techniques and allows machines the opportunity to learn from previous experience. Machine learning can be classified as supervised, unsupervised, or reinforcement learning (Kim *et al.* 2018; Ray 2019). This research uses supervised learning due to the existence of input data and output goals required. Supervised learning can be classified as classification and regression (Edoka, 2020). Classification predicts a discrete class name, while regression is a function of predicting a continuous quantity. Data collection, classification of that data, identification of patterns, prediction of events, and visualisation are the five steps in crime prediction using machine learning (Vural and Gök, 2017; Waduge and Ranathunga, 2017). Researchers use different types of machine learning algorithms to evaluate crime trends by comparing how successfully they work (Edoka, 2020).

Saeed and Abdulmohsin (2023) explores crime analysis and prediction using various machine learning and data mining techniques. The aim is to provide a concise review of these algorithms for crime prediction. The study supports future research by presenting crime definitions, system challenges, and classifications. Notably, supervised learning approaches, particularly logistic regression, have proven effective in predicting crime.

In the research study by Obagbuwa and Abidoye (2021), the authors explained that data mining tools were used to detect and analyse trends in crime data for South Africa through the extraction of crime data from the country's crime statistics database and

build a crime predictive model using a linear regression algorithm, which resulted in an accurate prediction with 84.7% accuracy using R-squared accuracy metrics.

Sukhija, Singh and Kumar (2020) used linear regression to perform a study on regression analysis to assess the correlation of characteristics linked with rape crime in Haryana and discovered relevant variables that can help police officials prevent crime more effectively

KNN and Naive Bayes were used to forecast crime in India by Kalsi (2019); their goal was to compare the two models in terms of accuracy to find out which performs better in crime prediction. The suggested approaches were implemented in Python; both techniques were 77% accurate and 96% accurate, respectively, with Naive Bayes proving to be better.

On the other hand, Ingilevich and Ivanov (2018) used three machine learning algorithms to predict the types of crimes that occur frequently. The proposed models for the work were linear regression, gradient boosting, and logistic regression. They aimed to compare the three models to see which one worked better in terms of accuracy. The gradient booster model performed better, while the linear regression forecast negative values.

A decision tree algorithm was used by Shi et al. (2018) to predict who commits crimes. They have also introduced other algorithms: Bayes Network, Logistic Regression, and Naive Bayes, and compared the accuracy of the results to check how they were working. The decision tree was better with 80% accuracy than other algorithms used.

Kim et al. (2018) analysed crime by using KNN and a boosted decision tree. The results obtained indicate that the accuracy was 39% and 44%, respectively, for both algorithms, establishing the boosted decision tree as having the highest accuracy. The drawback of their work shows that the accuracy of the KNN model has a poor predictive rate and that more research is required. Crime analysis was also carried out using time-series analysis and clustering techniques. Kiran and Kaishveen (2018), in their work, compared the KNN classifier with the Naive Bayesian classifier to see how well the model worked in the accuracy and execution of time analysis of crimes in

India. Both models had an accuracy of 77% and 87%, with timely execution of 0.5 seconds and 0.2 seconds, respectively, with Naive Bayesian's classifier doing better.

KNN and naive Bayes were also used by Abdulrahman and Abedalkhader (2017) to predict crime in San Francisco. Their approach was based on a comparison of the two classifiers. Two different techniques, uniform and inverse were used for the KNN classification system. Gaussian, Bernoulli, and multinomial methods were therefore deployed in naive Bayes. The obtained result indicates that the performance of KNN was poor, as it exhibited prolonged execution times during the classification and regression procedures. The results obtained from applying the naive Bayes-Gaussian model were inconclusive, indicating that the data utilised in the analysis may not possess continuous attributes but rather discrete characteristics. Conversely, the implementation of the naive Bayes Bernoulli and multinomial models proved to be more straightforward, as the data was directly incorporated into the training dataset without undergoing any error detection or outlier identification procedures.

Hassani *et al.* (2016) in their study, analysed data mining approaches for detecting illegal operations consisting of entity mining, clustering, association rule extraction, and classification approaches such as decision trees, supporting vector machines, Naive Bayes law, neural networks, and social network processing, with a proof focused on the number of applications for classification. With the growing mass of data, the biggest challenge facing law enforcement and intelligence agencies is examining huge amounts of data accurately and efficiently due to a lack of professionals with sufficient knowledge to apply data mining techniques. The increased use of data mining methods, together with the rise of Big Data, further underlines the need for further learning and ventures to educate and provide young people with information on the appropriate, increasing and applicable use of data mining methods.

Bello-Orgaz, Jung and Camacho (2016) focused on crime prediction concerning mining locations through social network connections, which enable data to be easily obtained using spatial experiences and data groupings by people. The hotspot mapping technique is often used to forecast when crime can occur using data from the past to alert potential actions. Crime detection accuracy issues include the

modelling of crimes for the development of effective crime detection techniques, reliable identification, data collection and storage, and processing time. Complex and proliferating crime results in questions about understanding the nature of criminals, predicting crime, effective detection, and managing vast quantities of data collected from a variety of sources.

Meanwhile, Hamdy et al. (2015) analysed a classification model for the detection of cynical activity based on social network posts that can predict and represent people's actions from reality-based data sources.

Meanwhile, Kiani, Mahdavi and Keshavarzi (2015), the authors completed a criminal investigation based on clustering and classification methods consisting of criminal paradigm mining using criminal investigation based on available criminal data, crime prediction based on the spatial distribution of existing data and crime recognition. The k-means clustering algorithm was used to classify crime datasets. By optimising the operator's external detection parameters with the genetic algorithm, the model they used for investigations and crime prediction was able to show important problems that needed to be fixed in the models. Among these problems are finding the best number of clusters and the best way to use the system in the prediction stage (Kiani, Mahdavi and Keshavarzi, 2015).

Bharti and Mishra (2015) used the hidden link algorithm technique to find secret connections between the suspect networks in India, which exposed a likely future criminal partner and another network outside the actual network. They used prognostic analysis in the investigation of a crime that aims to deter criminals before they occur.

Naive Bayesian, Logistic Regression and Support Vector Machine (SVM) classifiers were used to identify crime patterns and variables daily using a pattern identification process technique that uses the Apriori algorithm to quickly recognise trends and paradigms in the crime. The Decision Tree algorithm was used to predict and detect possible locations and patterns of crime (Awal *et al.* 2016; Thongsatapornwatana, 2016; David and Suruliandi, 2017; Vaidya *et al.* 2018; Kumar and Nagpal, 2019). Predicting crime based on the type of space and time can aid law enforcement officers in the estimation of offences at a given location. The decision tree classification and

Naive Bayes' classification were used to predict the forms of crime in various places. Then, an Apriori algorithm was used to classify the trends in regular crime. The study found that both the decision tree classification and the Naive Bayesian classification were 51% and 54%, respectively, in predicting crimes across various locations (Edoka, 2020).

Furthermore, McClendon and Meghanathan (2015) used WEKA, an open-source data mining program, to run a comparative study using linear regression, additive regression and decisions stump algorithms using the same collection of features to demonstrate how effectively and accurately the data mining machine learning algorithm was used to forecast violent crime patterns. Linear regression was considered most impressive among others and can accommodate unpredictability to an expected degree in the experiment samples as well. They added that data mining and machine learning have become important in detecting and preventing crime. Prabakaran and Mitra (2018) discussed how kernel density estimation, logistic regression and random forest simulation were used to conduct spatial and temporal analyses of sexual assaults in their literature. The kernel density approximation was used to assess the probability density function of sexual assault over a period.

Finally, a variety of classification algorithms have been implemented by Ratul and Rab (2020) including Random Forest, Decision Tree, AdaBoost Classifier, Extra Tree Classifier, Linear Discriminant Analysis, K-Neighbours Classifier, and four Ensemble Models to categorise various forms of illegal behaviour. Train-test split and k-fold cross-validation test methods were used to obtain the findings for each assessment based on which Ensemble Model obtained greater outcomes. The two identified shortcomings were the decision tree that gave them an odd yield in each evaluation stage, for which they tried different mixtures of parameters, but it was always over-fitted to the dataset used at the end. Secondly, the various algorithms generated a tremendous amount of time complexity to accomplish their projections, which is unsuitable for any practical application in life.

2.6 Comparisons of Similar Research and Results

A comparative study of the relevant work carried out by researchers in terms of models (algorithms), criteria, and evaluation metrics used (Edoka, 2020), and the best-performing model that was embraced by various authors in the analysis of the crimes carried out is shown in Table 2.2 as follows:

Table 2.2 Comparative Study of Related Works

Criteria	Model used	Evaluation metrics	Best-performing model	Outcomes	Authors
Prediction of crime	Simple linear regression, Multiple linear regression, decision tree regression, support vector regression, Random forest regression	Accuracy	Random Forest regression	96%	(Aziz *et al.* 2022)
Prediction of crime	Linear regression	Accuracy	Linear regression	84%	(Obagbuwa and Abidoye, 2021)
Prediction analysis	Multiple linear regression	Accuracy	Linear regression	97%	(Rath *et al.* 2020)
Prediction of crime	KNN and Naive Bayes	Accuracy	Naive Bayes	96%	(Kalsi, 2019)
crime patterns analysis	Multiple linear regression, Random Forest regression, Neural network regression, and Bayesian regression	Accuracy	Random Forest regression	97%	(Gonzalez and Leboulluec, 2019)
Crime offender's prediction	Decision tree, Bayes network, Logistic regression and Naive Bayes	Accuracy and precision	Decision tree	80%	(Shi *et al.* 2018)
Crime analysis	KNN and Boosted Decision Tree	Accuracy	Boosted Decision Tree	44%	(Kim *et al.* 2018)
Crime analysis	KNN classifier and Naive Bayesian classifier	Accuracy	Naive Bayesian's classifier	87%	(Kiran and Kaishveen, 2018)

2.7 Gaps Identified in Current Research

In most of the works reviewed, a more detailed analysis of crime trends in the South African context is missing. Approaches that have been successfully applied to analyse crime include supervised machine learning algorithms, such as Random Forest, Logistic Regression, Naive Bayes and KNN, Decision-Making Tree, Support Vector Machine, and Linear Regression (Shi *et al.* 2018; Gonzalez and Leboulluec, 2019; Obagbuwa and Abidoye, 2021). Therefore, in this research, various data analytics techniques, including machine learning algorithms, time series analysis, and spatial analysis, are employed to build models for predicting serious crime trends following the CRISP-DM methodology steps.

2.8 Chapter Summary

This chapter presented the literature review results on general crimes and the different types of serious crimes, like murder, sex, and drug-related crimes. This study presented published crime-related data for South Africa's nine provinces. Appropriate studies were sourced to identify how crimes are reduced both worldwide and in South Africa. Most studies used different methods to help reduce crime, but few used data analytics to analyse and predict serious crimes. The research methodology is presented in the next chapter.

Chapter 3 Research Methodology

3.1 Introduction

This section covers the selection of suitable research methodologies to meet the study's objectives and provide meaningful responses to the research questions. This chapter explains how to solve South Africa's serious crime problem using data analytics techniques.

3.2 Population Size

Population size is the set of components on which the research focuses (Monyeki, 2021). The study used South African crime statistics from 2005 to 2020 (15 years) (Kaggle, 2021; SAPS, 2022). Crime statistics, including all documented serious crimes, were collected from rural areas, townships, city centres, and suburbs in all nine provinces in South Africa. The population size here is the aggregated data sets from this period.

3.3 Research Design

The research questions and the research's execution or implementation are connected by the research design, which is a strategic framework for action. Research designs are plans that direct how to set up the data collection and analysis parameters to balance process economy with relevance to the research purpose (Kankam, 2019). The research design is like an architectural plan for a research project. It links design, data collection, and analysis to the research questions and ensures that everything on the study agenda is done. A research study's credibility, usefulness, and ability to be done depend on the design used (Rezigalla, 2020). A quantitative method based on the post-positivist research paradigm was used to collect and analyse data on serious crime trends in South Africa’s nine provinces for this study.

3.4 Research Paradigm

The post-positivist research paradigm was adopted for this study. Positivism and post-positivism are not the same. Positivists believe the universe is inevitable and that the

scientific method is the sole way to determine cause and effect. The post-positivist asserts that the methods of operation, thought and activity of both scientists and non-scientists are no different. Post-positivism contends that not all observations are free of fallibility and inaccuracy, and that the theory is revisable and changeable, whereas the positivist believes that the duty of science is to adhere persistently to the objective of getting it right about reality, even if no goal is ever attained, and that absolute knowledge of reality cannot be obtained through social science alone (Apuke, 2017; Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017).

3.5 Research Design Choice

To meet the research aims and objectives, this study adopted a quantitative research methodology. Data analytics based on a scientific approach whereby measurements and statistical analysis to evaluate hypotheses were used. Importantly, the study did not apply qualitative research, as the aim was to analytically analyse and predict serious crime trends using quantitative data. Below is a brief overview of the various research methodologies:

3.5.1 Quantitative research methodology

Quantitative research examines the correlation between variables in order to evaluate objective concepts. Then, these variables can be measured on instruments so that statistical methods can be used to look at the numbers. The focus of the purpose statement, which is based on quantitative research, is the relationship between the variables and the unit of analysis. The two types of quantitative research are experimental and non-experimental (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017).

3.5.2 Qualitative research methodology

The qualitative research methodology explores and gains a comprehensive understanding of the significance that individuals or groups attribute to a given social or human situation. The research process encompasses the inclusion of emerging inquiries and methodologies. Qualitative research is a purpose statement concerning the point of attention, while understanding and interpreting social interactions is the

primary goal of qualitative research. This type of research can be either interactive or non-interactive (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017).

3.5.3 Mixed-methods research methodology

The use of mixed methods combines or links qualitative and quantitative approaches to investigation. It involves utilising qualitative and quantitative methodologies as well as combining the two in a study. It also involves making philosophical assumptions (Chih-Pei and Chang, 2017). In order to make a study's overall strength stronger than either qualitative or quantitative research alone, it therefore takes more than merely collecting and evaluating both forms of data (Creswell and Creswell, 2017). As a result, mixed-methods research adheres to scientific rigour, and it has the advantage of combining the strengths of quantitative and qualitative approaches while compensating for the flaws of each (Chih-Pei and Chang, 2017; Creswell and Creswell, 2017; Khaldi, 2017). However, this study is not using a qualitative or mixed method because the data used is open-access data.

3.6 Research Strategy

This study adopted an experimental research strategy, as data analytics tools would be used to analyse serious crime trends. The data analytics tool would be informed by information gathered through a thorough literature review and the cross-industry standard process for data mining (CRISP-DM) methodology. CRISP-DM is a well-known methodology for doing data mining operations (Berwind *et al.* 2016).

3.7 Data Collection

Data-collection techniques make it possible to gather details on research subjects (i.e., people, things, and occurrences) as well as the environments in which they take place (Wei *et al.* 2022).

Data must be collected in a precise manner. If data are collected carelessly, it would be impossible to conclusively respond to our study's questions (Chaleunvong, 2009). Kabir (2016) defines data collection as the systematic process of gathering and

measuring information on variables of interest, enabling a researcher to answer the research questions, test hypotheses, and evaluate results.

3.7.1 Secondary data collection

Secondary data is a term used to describe information that has already been compiled and is easily accessible from other sources (Juneja, 2015). Such data is less expensive and faster to gather than primary data; it saves time and money and aids in a better analysis of the problem (Juneja, 2015). Another benefit of using secondary data is that it already has a high level of reliability and validity, reducing the need for the researcher to go through it again (Kabir, 2016). Secondary data also has drawbacks such as the fact that the accuracy of secondary data is unknown. It is also possible that the information is outdated (Juneja, 2015), and the data is being collected for objectives that may not correspond to current research needs, are difficult or expensive to obtain, and there is no real control over quality criteria (Greener, 2008; Saunders, Lewis and Thornhill, 2009).

A secondary data collection method was used to gather existing data from trusted online data repositories. For this research, the serious crime dataset from 2005 to 2020 was extracted from the South African crime statistics database (Kaggle, 2021; SAPS, 2022). The data was collected based on the study's aim and objectives to avoid ethical problems during the data collection phase. Kaggle.com and saps.gov.za were chosen for this study because they contain all South African crime datasets and are publicly available to researchers without cost.

3.7.2 Data collection instrument

Python software tools were employed in this research to collect, process, and analyse the serious crime data following the CRISP-DM methodology steps.

3.7.3 Data collected

For this study, South African crime statistics were used. The dataset included all types of crimes that have been classified and all nine provinces in South Africa. This dataset tracked crime rates over 15 years, from 2005 to 2020. The three focus areas of this

study are sexual crimes, murder, and drug-related crimes. (Kaggle, 2021; SAPS, 2022).

3.8 The Cross-industry Standard Process for Data Mining (Crisp-Dm) Methodology

This study enhances the performance of data analytics algorithms in analysing and predicting serious crime, allowing law enforcement to gain a better understanding of the problem and take the necessary steps to reduce the serious crime rate. Data mining is a subfield of knowledge discovery in databases, which is a multidisciplinary field (Awal *et al.* 2016; Majeed and Naaz, 2018). Data mining is the process of gathering raw data and extracting information that can be used to create predictions in a variety of real-world scenarios (Majeed and Naaz, 2018). The cross-industry standard process for data mining (CRISP-DM), a quantitative methodology, was adapted for this research project because it fulfils the set objectives. CRISP-DM is a data mining model that depicts the life cycle of a data mining project, according to Pete Chapman *et al.* (2000) and Edoka (2020). The CRISP-DM is divided into six phases, each with its own set of tasks and relationships between them (Berwind *et al.* 2016). Figure 3.1 depicts the six phases of CRISP-DM, which are conceptual understanding, description of data used in the study, data preparation, modelling, evaluation, and deployment.

Figure 3.1 Cross Industry Standard Process for Data Mining (CRISP-DM) (Source: Berwind et al. 2016)

3.8.1 Conceptual understanding

This section, as shown in Figure 3.1, focuses on the research problem and objectives. Past work done to solve data analytics problems associated with crime detection led to the development of the research objectives. This research also focuses on utilising existing data analytics tools to solve the problem of serious crime occurrences and determining how accurate the techniques are in aiding serious crime reduction. The purpose of this research was to use the existing South African serious crime data to analyse the crime trends and spawn useful information that can be conveyed to the government and security agencies for them to make timely decisions on how to reduce serious crime in the country.

3.8.2 Description of data used in the study

The need for a suitable dataset was taken into thoughtful consideration to achieve the set objectives. This work made use of existing datasets obtained from Kaggle.com and saps.gov.za. The datasets spanned 15 years (2005-2020) and contained an

extracted serious crime csv (comma-separated values) file with 3495 rows and 18 columns, which was sufficient for the research objectives. The file was uploaded to a Jupyter Notebook, where the data cleansing and preparation procedures were executed to ensure uniformity within the data before conducting experimental analysis on the cleaned dataset. At this stage, activities such as data description, data exploration, and data quality verification are conducted.

3.9 Data Preparation

The dataset for serious crime was arranged and ready for statistical analysis. The Python library, Scikit-learn (Sklearn) was used for data selection, cleaning, data building and data incorporation.

3.9.1 Data pre-processing

This process involves removing null or infinite values that could affect the system's accuracy. The cleaning process is used to remove or correct some missing data as well as incomplete data. Many missing null values were found in the supplied dataset. In order to cut down on noise, the data had to be cleaned, processed, and combined. Data preparation should be performed several times in random order (Monyeki, 2021). Python's Jupyter Notebook was used to perform the pre-processing. Python was used to complete tasks like data cleansing, data transformation for modelling tools, and the breaking up of a large dataset into smaller pieces. Pre-processing was done in order to prepare the data for analysis, such as analysing and visualising trends and making predictions.

3.9.2 Data coding

According to Greener (2008), coding refers to using computer-assisted analytical tools to analyse data. The author emphasises the significance of maintaining a codebook that contains precise, lucid, and easily accessible recordings, serving as a connection between the participant and the data (Rashid *et al.* 2019). Except for a few kinds of data, it is necessary to record all data using numerical codes. This practice allows faster data collection with reduced errors and enables data analysis using software (Saunders, Lewis and Thornhill, 2009). Data coding is the initial step in the data

preparation process (Sekaran and Bougie, 2019). All implementation within this study was carried out on a Python Jupyter notebook environment within the Anaconda distribution. Some of the Python libraries used include Pandas, Numpy and Scikit-learn.

3.9.2.1 Python pandas

The open-source Python Pandas package provides data structures and tools that are both high-performance and user-friendly for the purpose of data analysis (McKinney, 2010). Data frames in Python are utilised to store tabular and matrix data in a structured format, organised into rows and columns. This facilitates the ability to perform dynamic data processing operations. In this project, it was mostly utilised to prepare data for use by machine learning algorithms. The csv files were loaded and converted into a data frame using the Python Pandas object usually denoted as 'pd'. Python Pandas allow the importing and analyses of data, usually in real-time (Nagpal and Gabrani, 2019). The general Python syntax for this process is:

import pandas as pd

df = pd.read_csv('Serious crimes data.csv')

df # To show the serious crimes dataset as shown in Table 2.1*.*

3.9.2.2 NumPy

NumPy is a Python toolkit for dealing with multidimensional data and performing scientific and mathematical operations on it (Nagpal and Gabrani, 2019). NumPy was utilised as an add-on to the Pandas library to conduct some fundamental mathematical operations in this project. NumPy is a potent package for Python-based scientific computation (Nagpal and Gabrani, 2019). It can be used by creating an N-dimensional object array, which is typically denoted by np. A NumPy object can be created using the following generic syntax:

import NumPy as np

3.9.2.3 Scikit-learn

Scikit-learn is a Python machine learning package that offers a variety of classification, regression, and clustering techniques (Buitinck *et al.* 2013). This library is utilised to do the real task of building the algorithm and prediction in this project. It includes several evaluation metrics to validate the algorithm's performance, making it a useful tool.

3.10 Modelling

This is a major step in the data mining process because it allows you to use predictive data analytics algorithms to evaluate the data and produce predictive models that can be used to make future predictions based on the rich information generated by undetected data trends. Selecting suitable modelling techniques, such as an appropriate machine learning algorithm, is necessary to build a predictive model. Additionally, a test design should be created to assess the quality and validity of the model. Subsequently, the model should be built and executed using the prepared dataset, creating one or more models. Finally, the model's interpretation should be based on domain expertise, predetermined success criteria, and the desired test. This phase involved the implementation of data analytics algorithms. The most common method for predicting and determining cause-and-effect correlations between variables is linear regression. There are many other methods that can be used for prediction, like decision tree classifier, naive bayes, gradient boosting, support vector machine (SVM), and KNN, but they can all not function well in the same categories of prediction (Lalwani *et al.* 2022). Some are exceptionally good in association rule mining, classification, regression, and clustering. Classification predicts a discrete class name, while regression is a function of predicting a continuous quantity. Linear regression was employed in this study due to the nature of the dataset, which is a continuous quantity.

3.11 Evaluation

In this step, the model was evaluated to see how well it met the project's objectives. Linear regression accuracy was evaluated using the mean square error and R-squared score.

3.12 Deployment

During this phase, which also includes the final report, strategies for the evaluation results were decided. In Python, Matplotlib was used to visualise the project's output. A comprehensive review of the project was also conducted to ensure that the project's objectives were met. The first and last phases are similar for the purposes of this experiment. The first phase is concerned with the goal of predicting serious crime; as a result, the goal was to study different predictive algorithms. The experiment's findings were reported in the final phase.

This section provides an overview of the statistical modelling techniques used in this study: the linear regression algorithm. The linear regression model is a well-known statistical model for studying correlational relationships among variables in educational research (Norouzian and Plonsky, 2018).

3.13 Data Analysis

Serious crime data, population statistics, density, and area were used to complete the data analysis, although the type of statistical data used was determined by the type of data extracted.

3.13.1 Predictive data analytics algorithm

A predictive data analytics algorithm is a method of building a predictive algorithm. To make those predictions, the process includes a data analytics algorithm that learns specific properties from a training dataset. Regression and feature classification are two types of algorithms used in predictive data analytics. Regression models are used to analyse relationships between variables and identify trends in order to make predictions about continuous variables (Edoka, 2020). Feature classification, unlike

regression algorithms, involves assigning discrete class labels to data values as a prediction's output (Edoka, 2020).

3.13.1.1 Naive Bayes

These classifiers are based on the Bayes theorem with independence assumptions between features, and they are commonly used in machine learning. When the class labels are selected from a finite set, the approach creates classifier models that assign class labels to issue occurrences represented as vectors of feature values (Lalwani *et al.* 2022). This is mentioned but delimited for this study and recommened for future work.

3.13.1.2 Decision tree classifier

In a decision tree classifier, a tree-shaped graph or algorithm of decisions is used in decision making. It is a way to show an algorithm in a visual fashion.

3.13.1.3 Logistic regression

Logistic regression is a statistical regression algorithm that uses classification or binary dependent variables.

3.13.1.4 Linear regression

Linear regression can be described as a statistical analysis technique that can be used to determine how variables relate to one another. Using linear regression, one may model the relationship between one or more predictors (X) and a scalar dependent variable (Y). Simple linear regression occurs when there is just one independent variable. To predict the status of serious crimes, a linear regression approach was used. The simple linear regression algorithm is a straightforward algorithm with only one independent and dependent variable. The algorithm becomes a multiple linear regression (MLR) algorithm if there is more than one independent variable (Sukhija, Singh and Kumar, 2020). As a result, the multiple linear regression algorithm was utilised to predict serious crime status since more than one parameter that affects crime negatively or positively was considered. The regression algorithm gives sufficient insight into how independent or input variables X influence the output or dependent variable Y. The aim is to utilise an appropriate predictive model to predict

the factors influencing a high incidence of serious crime. This analysis aims to support law enforcement agencies in making informed decisions and implementing appropriate interventions. To that end, the linear regression predictive data analytics algorithm was accurately implemented.

The use of a data mining tool (Scikit Learn Modules) based on the Python modules NumPy, SciPy, and Matplotlib was used in this work to build a model that can predict sexual crime with the use of linear regression and decision tree algorithms using Python. Some of the key terminologies in linear regression include correlation, dependent, and independent variables which are explained below.

i. Correlation

The measurement of the relationship between two quantitative variables is called a correlation (Cherry, 2020). These variables are already present in the density or population, and the researcher has no control over them. A positive correlation refers to a linear relationship between two quantitative variables, wherein an increase in one variable also translates to an increase in the other variable. In the context of a negative correlation, it can be observed that a decrease in the quantity of a second variable accompanies an increase in the quantity of one variable (Cherry, 2020).

Generally, there are always several independent variables that might be employed in a linear regression; choosing which of these to employ is one of the most challenging tasks (Shingleton, 2012). The relationship between the dependent and independent variables must be established and examined before the variables to employ in a regression can be determined. Correlation, which is defined as the linear relationship between two variables, is an important relationship for this study. This relationship is measured between pairs of observed variables. To draw the correct conclusions, it is important to understand the relationship between variables (Kalla, 2011).

ii. Dependent and independent variables transformation

A straight line would not always be the best fit for the dependent variables as a function of the independent variables. As a result, both dependent and independent variables must be transformed or adapted. Variable transposition, multiplying several variables

(interactions), taking the variable to a power, and using the log function on the variable are some common variable transformations. The variables would be re-inserted into the regression after transformation. With any transformation, the original variable would be left in the regression as a standard of practice.

iii. Linear regression problem formulation

It is usual to label outputs with y and inputs with x. When there are two or more independent variables, they are represented by the vector $\mathrm{x} = (x_1, \ldots, x_r)$, where r is the number of inputs (predictors), and equation **(3.1)** assumes a linear relationship between y and x:

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon \qquad \textbf{(3.1)}$$

where $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients, and ε is the random error (James *et al.* 2013). One of the most important and extensively used regression approaches for predicting quantitative values is linear regression, and one of its key merits is the ease with which the results may be interpreted. Linear regression calculates the estimators of the regression coefficients or simply the predicted weights, denoted with $\beta_0, \beta_1, \ldots, \beta_r$. The estimated regression function is expressed in equation **(3.2)**

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \qquad \textbf{(3.2)}$$

This function identifies and analyses the relationships and dependencies between the inputs and outputs. The goal is to have the estimated or predicted response, $f(x_i)$, for each observation $\mathrm{I} = 1, \ldots, \mathrm{n}$, be as close as possible to the corresponding actual response $\mathrm{y_i}$ (James *et al.* 2013). The differences $y_i - f(x_i)$ for all observations $\mathrm{I} = 1, \ldots, \mathrm{n}$, are called the residuals. The goal of regression is to find the best predicted weights, namely, the weights that correspond to the smallest residuals.

The sum of squared residuals (SSR) for all observations is often minimised to acquire the optimal weights $\mathrm{I} = 1, \ldots, \mathrm{n}$: $\mathrm{SSR} = \sum_i \big(y_i - f(x_i)\big)^2$. This approach is known as the ordinary least squares method (James *et al.* 2013; Obagbuwa and Abidoye, 2021; Stojiljković, 2021).

iv. Performance of linear regression

The actual response variation y_i, $\mathrm{i} = 1, \ldots, n$, occurs partly because of the dependencies on the predictors x_i. There is, however, an additional inherent variance in the output. The coefficient of determination, represented as R^2, indicates how much variance in y can be explained by the dependence on x using the regression model. A higher R^2 implies a better fit, implying that the model can explain the variation of the output with different inputs better. The value R^2 = 1 corresponds to SSR = 0, that is, to the perfect fit since the values of predicted and actual responses fit completely to each other (James *et al.* 2013; Stojiljković, 2021).

In Scikit (Pedregosa *et al.* 2011; Buitinck *et al.* 2013), Sklearn linear regression enables the investigation of relationships between two continuous (quantitative) variables: one variable, represented by X, is the predictor (population, density, and so on). The other variable, denoted y, is regarded as the response (high sexual crime) variable. The equation for a linear regression line is presented in equation **(3.3)**

$$\mathrm{y} = \mathrm{a} + \mathrm{bX} \qquad \textbf{(3.3)}$$

Where y is the dependent variable and X is the predictor (independent) variable. Therefore, the regression Python syntax can be written as:

```
from sklearn.linear_model import LinearRegression
from sklearn import linear_model
regr = linear_model.LinearRegression()  #Object of regression
regr.fit(X, y) #Fitting the data in the regression object
```

3.14 Performance Measures for Data analytics Techniques

Different metrics can be used to evaluate a linear regression, including the following metrics:

3.14.1 Mean squared error

The concept of mean squared error (MSE) is discussed by Glen (2021) as a measure of proximity between a regression line and a given set of points. This is achieved by

calculating the squared distances between the points and the regression line, commonly called "errors". The process of squaring is essential in order to eliminate any negative signs. Furthermore, it assigns greater significance to larger disparities than smaller ones in the outcome. The term "mean squared error" is computed from the average value of a set of errors. It is the average of the squares of all the data points in the given dataset. The lower the MSE, the better the prediction; it is one of the most widely used metrics (Glen, 2021).

$$\text{MSE} = \left(\frac{1}{n}\right) * \sum(actual - predicted)^2 \quad \textbf{(3.4)}$$

where:

“n = number of items,

Σ = summation notation,

Actual = original or observed y-value,

Predicted = y-value from regression”.

3.14.2 Mean absolute error

The mean absolute error metric is the average value of the absolute errors observed across all data points within a given dataset.

3.14.3 Median absolute error

The median absolute error within a given dataset is calculated as the average numerical value of all individual errors. One of the advantages of this metric is its robustness against outliers. In contrast to a mean error metric, the presence of a single outlier in the test dataset does not significantly influence the overall error metric.

3.14.4 Explained variance score

The metric quantifies the extent to which the model is capable of explaining the variability observed within the dataset. The model achieves an optimal performance if it attains a score of 1.0.

3.14.5 R-squared

The coefficient of determination is measured by R-squared. This indicates how effectively the model would predict the data. The highest possible score is 1.0, but it can also be negative. An R-squared close to 1 indicates that the model can accurately predict the data. One or two metrics was chosen to evaluate the model because keeping track of all of them can be difficult. Making sure the R-squared is high is a good practice.

$$\mathcal{R}^2 = 1 - \frac{SSres}{SStot} \qquad \textbf{(3.5)}$$

where:

SSres = the sum of squares of the residuals.

SStot = the total sum of squares.

The coefficient of determination, commonly referred to as “R-square”, holds a significant value derived from regression analysis. The coefficient of determination, often denoted as "R-square," is a crucial metric obtained from regression analysis. The coefficient of determination measures the extent to which the independent variable can accurately predicts the variance in the dependent variable. The coefficient of determination, ranging from 0 to 1, is the square of the correlation between the predicted and actual y values. If the R-square value is zero, it indicates no linear relationship between the independent and dependent variables, making it impossible to predict the dependent variable based on the independent variable. If the R-square value is 1, the dependent variable can be accurately predicted from the independent variable without any error.

When the scale of the independent variable is an interval or ratio, linear regression models are commonly used. The errors are assumed to be independent, normally distributed random variables with common variance in linear regression models. The least squares method is used to estimate the parameters in linear regression models. Linear regression algorithms can only be used to capture the linear relationships between independent variables and the dependent variable (Jin, 2013). In this

situation, the coefficient estimates may be very unstable in response to slight changes in the data; thus, the relationship among the variables may be hard to interpret. But the predictability of the model may not necessarily be diminished. Linear regression is used to predict serious crime trends in Chapter 4. Linear regression fits in this problem well. The grades were normally distributed on nearly a continuous scale.

3.15 Data Validity and Reliability

Although validity and reliability are intricately connected, the terms have different meanings. A measurement can be accurate but not valid. A reliable measurement, on the other hand, is typically valid (Middleton, 2019). A valid research design, in most cases, maximises validity by providing a coherent explanation of the phenomenon under study and minimising all possible biases or confounds that could confuse or distort the findings (Bickman, Rog and Hedrick, 2009).

3.15.1 Data validity

Data validity is the accuracy with which a method measures what it is supposed to measure (Middleton, 2019). Research with a high level of validity gives results that match traits, characteristics, and differences in the real world (Middleton, 2019).

3.15.2 Data reliability

Reliability is the term used to describe the level of consistency with which a particular method measures data. The measurement is reliable if it always gives the same result when the same steps are taken in the same conditions (Middleton, 2019). The degree to which data collection techniques and processing procedures would produce consistent results is referred to as reliability (Saunders, Lewis and Thornhill, 2009). The goal of research reliability is to give confidence that the results were not falsified in any way to get a specific result (Greener, 2008).

3.16 Data Ethical Considerations

Secondary data is often made available for free on the Internet, in books, or on other public forums, with the implied permission to use and analyse it further. Nevertheless, it is important to acknowledge the ownership of the original data (Tripathy, 2013). As

a result, it is important to analyse the data leveraging diverse criteria, including the methodology employed for data collection, the accuracy achieved, the duration of the data collection period, the underlying objective of the data collection, and the nature of the data itself. The maintenance of the data should be limited to the duration of its necessity. Protecting against unauthorised access, potential loss, or deliberate destruction is imperative (Tripathy, 2013).

In the case of this study, data that is publicly accessible on reliable website repositories is obtained and used according to the South African Protection of Personal Information Act (POPIA) (Netshakhuma, 2020). The data used in this study remains anonymous.

3.17 Plagiarism and Copyright

Plagiarism is when you take someone else's work and claim ownership of the ideas without giving proper acknowledgment to the author (Helgesson and Eriksson, 2015). The ability to contribute to the existing body of knowledge genuinely and uniquely, which is undercut by plagiarism, gives scholarly work its integrity (Bretag, 2013).

Individuals, institutions, research, and public perceptions of education are all impacted by academic integrity. Plagiarism and copyright are just two of the many factors that can damage or promote the integrity of research and education. In the case of this study, the author made sure to reference all the materials used properly.

3.18 Chapter Summary

In this chapter, the researcher described the dataset, population size, design, and methodology. This chapter discussed in detail how the tools and datasets were used to achieve the research objectives. The chapter provided a detailed explanation of how the technologies and ordinary least squares regression were used to analyse the dataset in a Jupyter notebook environment.

Chapter 4 Results and Discussion

4.1 Introduction

This section presents the results, discussions, and interpretations of the findings. The visualisation of serious crime trends as well as the results of the predictive analytics are presented to achieve the research objectives. The implementation entails extracting meaningful features from the dataset used in this research, and the steps taken are detailed in this section. The model is developed using the Python programming language in a Jupyter notebook integrated development environment within the open-source Anaconda distribution. The Jupyter notebook was chosen because of its flexibility and ability to manage large scale implementations (Nagpal and Gabrani, 2019). The visualisation was done using Python (Anaconda) environment and Flourish Studio.

Table 2.1 in Chapter 2 provided a cross-section of South African serious crime statistics (2005-2020). The table depicts a snapshot of the nine provinces, including the station's annual rates for the three serious crimes selected for this study (sexual, murder, and drug related). The dataset consists of 18 columns and 3495 rows. According to the dataset, some stations had no entries in some years, possibly due to poor data collection or data unavailability. However, as time passed, data availability has improved.

The dataset for each province was summed up in Table 4.1. The total number of serious crimes per province in South Africa between 2005 and 2020 was presented, with Gauteng being the most affected and the Northern Cape being the least affected with respect to sexual crimes.

Table 4.1 South African serious crimes summed totals per province (2005 – 2020)

	PROVINCE	MURDER	DRUG-RELATED CRIMES	SEXUAL CRIMES
0	Gauteng	56 039	579 307	185 316
1	KwaZulu/Natal	63 298	543 509	159 245
2	Mpumalanga	12 875	89 745	58 935
3	Western Cape	44 545	1 101 009	122 045
4	Limpopo	11 647	117 764	66 367
5	Eastern Cape	52 711	185 962	134 977
6	Northwest	12 659	133 818	67 604
7	Free State	14 220	101 261	63 199
8	Northern Cape	5 677	46 867	26 135

In addition, the population, density, and area dataset from Kaggle (2021) is shown in Table 4.2. These variables were essential in this study since they are the independent variables and serious crimes (sexual, murder, and drug-related crimes) are the dependent variables. The serious crime data was then summed across all nine provinces, as shown in Table 4.1. The dataset contains three independent variables (population, density, and area) and three dependent variables (sex, murder, and drug-related crimes). The data in each instance belongs to the nine provinces of South Africa.

Model Selection: A variety of data mining techniques (Awal *et al.* 2016) was employed to statistically predict the status of serious crime. One of such methods is the ordinary least square regression model which was used in this study. This is because the regression model is straightforward and describes how the input influences the outcome. It calculates a linear function of another variable X (input variable/features) to predict a variable Y (target variable).

4.2 Serious Crimes Data Analysis and Visualisations

Serious crime data analysis and visualisation were done using Python (Anaconda3) and the Flourish Studio tool. Flourish Studio[3] is a tool that allows the upload of data and creates various data visualisations, ranging from simple bar graphs and charts to interactive story maps without coding. The researcher ensured that the data was well prepared before the analysis. The discussion on the visualisation trends of the three serious crimes selected for the study follows.

4.2.1 South Africa murder case statistics

In this section, the statistics of South African murder crimes committed between 2005 and 2020 are presented based on the analysis. The results of the analysis are represented in Figures 4.1 and 4.2.

[3] https://flourish.studio

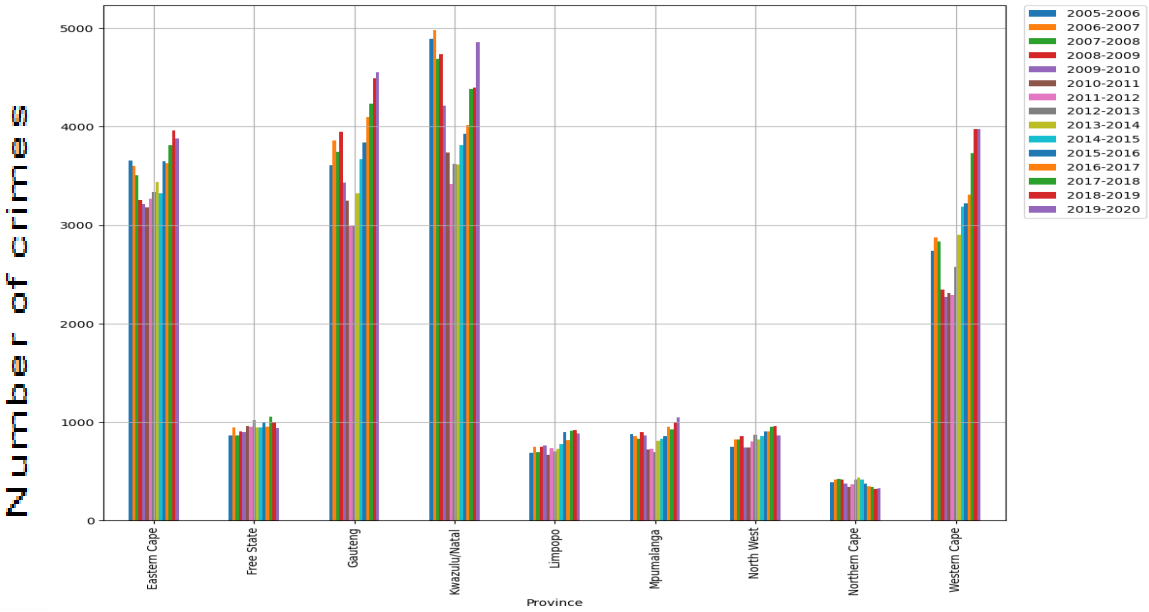


Figure 4.1 South African murder crime statistics (2005-2020)

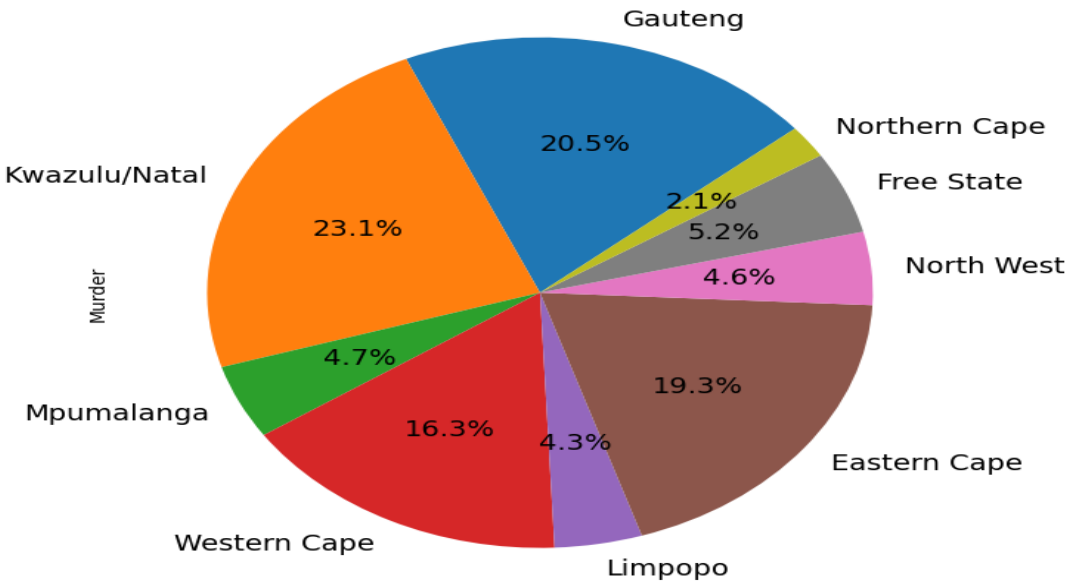


Figure 4.2 South African murder percentage per province

Figure 4.1 displays the statistics of total murder cases in each South African province from 2005 to 2020. Kwazulu-Natal and Gauteng are the provinces with the highest number of murder cases. While the Northern Cape and Limpopo have the lowest number of murder cases in the country. Figure 4.2 also displays the percentage of murder cases for each province on a pie chart for greater understanding, displaying the percentage of murder cases for each province. It is obvious that Kwazulu-Natal has the highest percentage rate of murder crime at 23.1%, while the Northern Cape has the lowest percentage rate of murder crime at 2.1%.

4.2.2 South Africa Drug-related crimes statistics

In this section, the statistics of South African drug-related crimes are presented. Figure 4.3 shows the trends of drug-related crimes in each province of South Africa from 2005 to 2020. From the result of the analysis obtained, the Western Cape, Gauteng and Kwazulu-Natal have the highest rates of drug-related crimes while the Northern Cape and Mpumalanga have the lowest drug-related crime rates in the country.

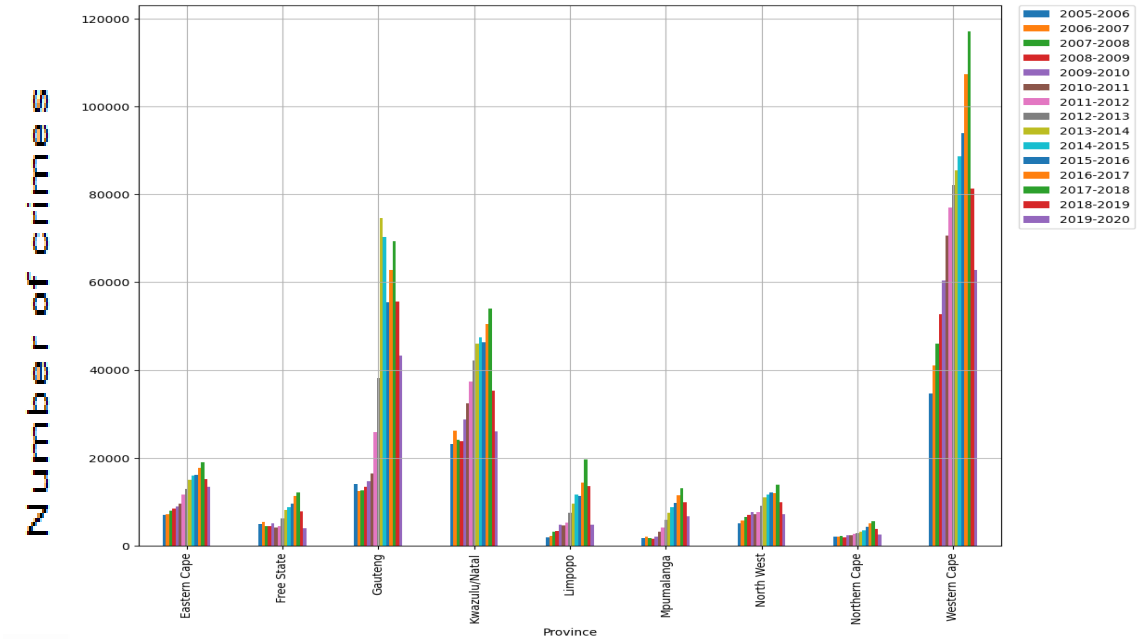


Figure 4.3 South African drug-related crimes statistics (2005-2020)

Also, the percentage of drug-related crimes for each of the provinces is presented in Figure 4.4 as a pie chart visualisation for greater understanding. The pie chart displays the percentage of each province's drug-related crime rates; it was glaring that the Western Cape (38%), Gauteng (20%), and Kwazulu-Natal (18.7%) have the highest percentages of drug-related crimes, while the Northern Cape (1.6%) has the lowest percentage of drug-related crimes.

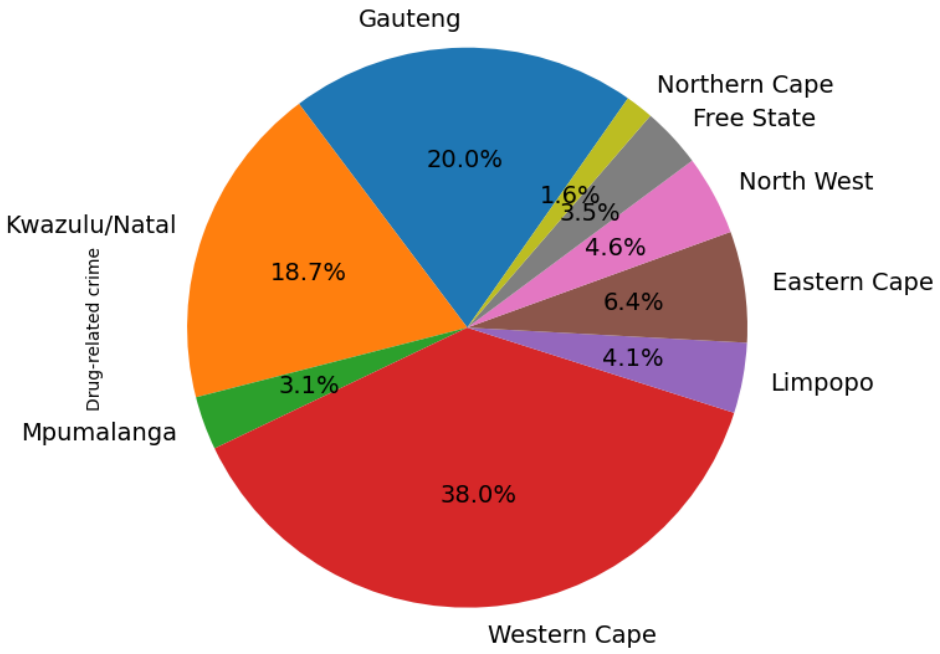


Figure 4.4 South African drug-related crimes percentage per province

4.2.3 South Africa sexual crimes statistics

In this section, the statistics of South African sex-related crimes are presented. Figure 4.5 displays the trends of total sexual crime in each province of South Africa from 2005 to 2020. Results of the analysis reveal that Gauteng, Kwazulu-Natal, the Eastern Cape, and the Western Cape are the provinces with the highest rates of sexual crime, while the Northern Cape and Mpumalanga have the lowest sexual crime rates in the country.

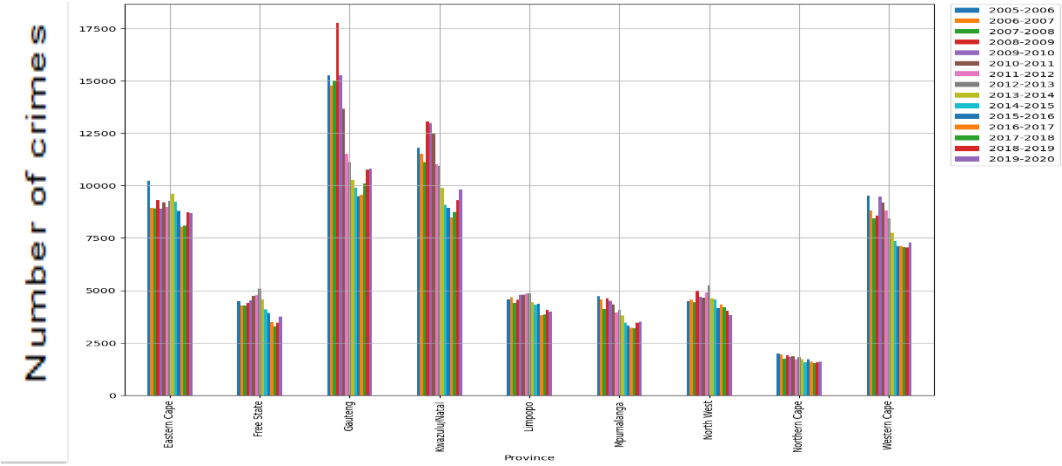


Figure 4.5 South African Sexual crimes statistics (2005-2020)

Also, the visualisation of the percentage of sexual crimes for each of the provinces is presented in Figure 4.6 as a pie chart for greater understanding. The pie chart displays the percentage of each province's sexual crime rate; it is glaring that Gauteng has the highest percentage of sexual crime occurrences (21%), while the Northern Cape have the lowest percentage of sexual crime (3%).

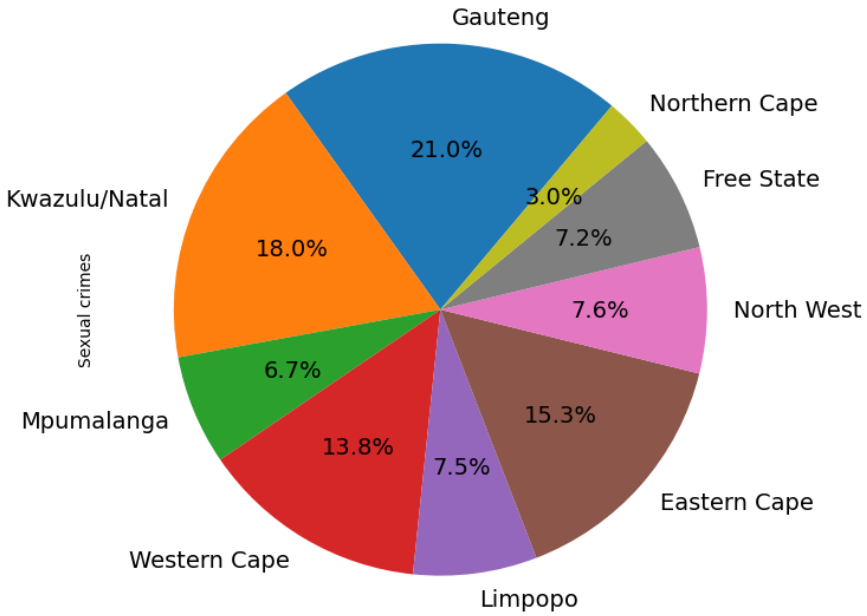


Figure 4.6 South African sexual crime percentage per province

4.2.4 South Africa serious crimes category statistics

In this section, the statistics of the South African serious crime categories are presented. Figure 4.7 and Figure 4.8 displayed the three selected serious crimes (drug-related, murder, and sexual crimes) with a Flourish tool adapted for Figure 4.8. The most prevalent of the three crimes is drug-related crime, which is dominant in the Western Cape Province. Next is sexual crime, which is prevalent in Gauteng Province, and the last crime, murder, has the highest occurrences in the province of KwaZulu-Natal, South Africa.

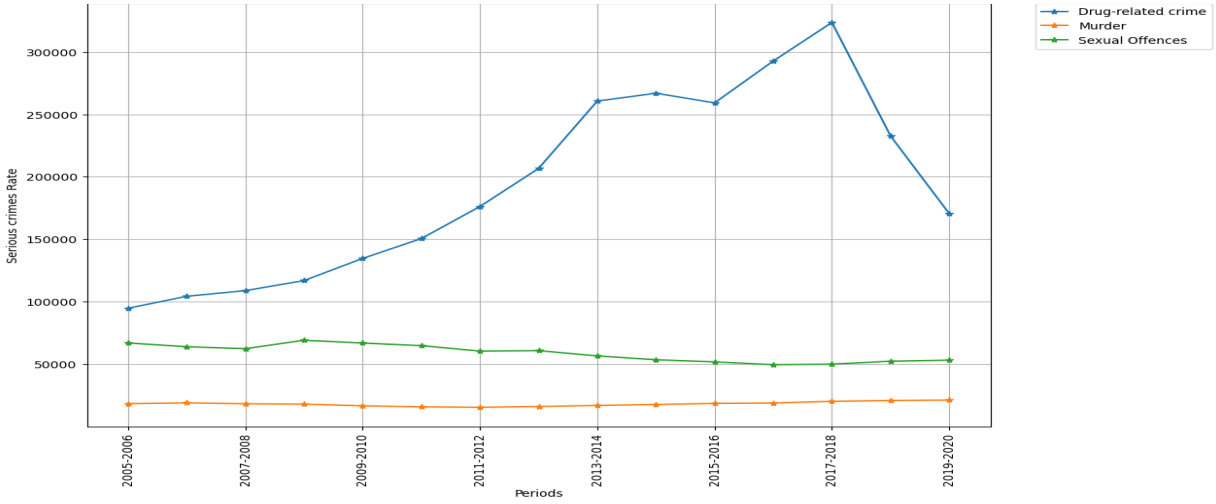


Figure 4.7 South African selected serious crimes category

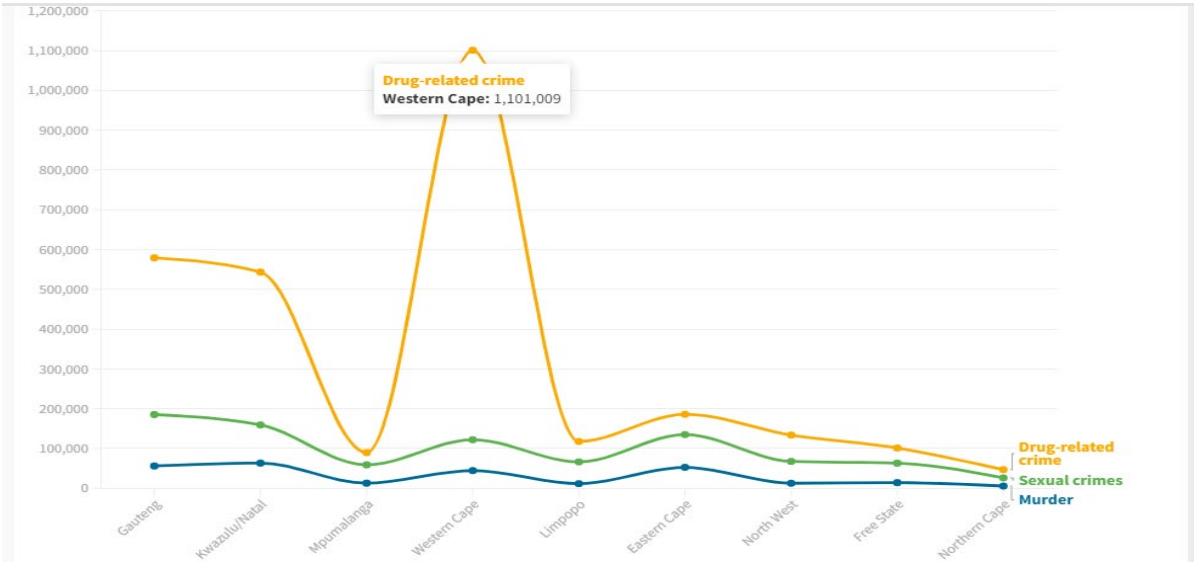


Figure 4.8 South African selected serious crimes category using flourish tools

4.3 South Africa Population Statistics

In this section, the statistics of the South African population are presented. The pie chart in Figure 4.9 was used to visualise the percentage of the province's population statistics. Gauteng has the highest population followed by KwaZulu-Natal and the Eastern Cape, while the least populated province is the Northern Cape.

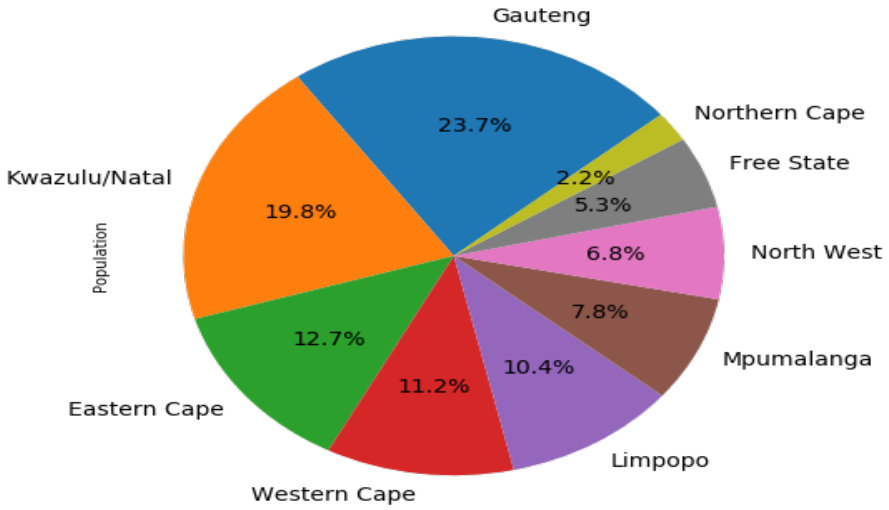


Figure 4.9 South African population statistics by Province

4.3.1 Population, area, density

For the purpose of prediction, the population, area, and density are shown in Table 4.2. Serious crime numbers for each province were summed up using Python. The population, density, and area are the independent variables, while serious crime is the dependent variable. It is important to make sure that a linear relationship exists between the dependent variable (serious crimes) and the independent variables (population, density,and area) before executing a linear regression model.

Table 4.2 Population, area and density of South Africa by Province

	Province	Population	Area	Density
0	Gauteng	12 272 263	18 178	675.1
1	KwaZulu/Natal	10 267 300	94 361	108.8
2	Mpumalanga	4 039 939	76 495	52.8
3	Western Cape	5 822 734	129 462	45.0
4	Limpopo	5 404 868	125 755	43.0
5	Eastern Cape	6 562 053	168 966	38.8
6	Northwest	3 509 953	104 882	33.5
7	Free State	2 745 590	129 825	21.1
8	Northern Cape	1 145 861	372 889	3.1

4.3.1.1 Density percentage

The pie chart in Figure 4.10 displays the percentage of each province's density. It is noticed that Gauteng has the highest density (66.1%) percentage while the Northern Cape has the lowest density (0.3%) percentage, respectively.

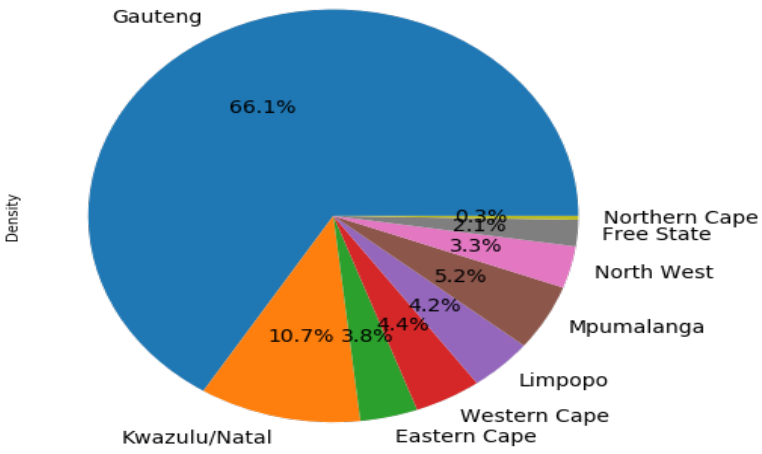


Figure 4.10 Population Density Percentage of South Africa by Province

4.3.1.2 Area percentage

The pie chart in Figure 4.11 displays the percentage of each province's area. It is noticed that Gauteng has the lowest area percentage (1.5%), while the Northern Cape has the highest area percentage (30.5%) as shown, respectively. The area can be said to be the opposite of population and density. This means that it is possible to have a highly populated province with a small area and vice versa.

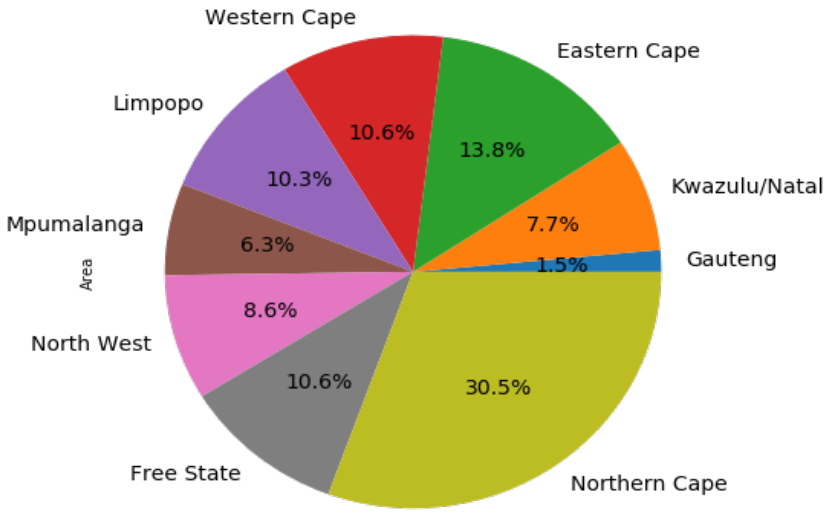


Figure 4.11 South African provinces by Area percentage

4.4 Word Cloud of the Three Highly Prone Serious Crime Provinces

The word cloud is a data visualisation technique that can represent text data in which the size of each word denotes its frequency of occurrence or impact. The more often a specific word appears in the text, the bolder it appears in the word cloud. For instance, looking at the Gauteng sexual crimes hotspot areas shown in Figure 4.12, the word "Pretoria" is noticeably the biggest and boldest among others, which shows that sexual crimes occur more often in this part of the Gauteng Province than in other stations of the province.

Each word size in the word cloud in Figure 4.12, Figure 4.13, and Figure 4.14 show frequent occurrence of serious crimes in these areas. For example, to view the most

notable stations where murder occurs frequently in KwaZulu-Natal Province; Durban, Ntuzuma, Port Road were all easily noticeable in Figure 4.13. For the word cloud visualisation, the researcher chose the three highly prone provinces of each of the three serious crimes for the study, respectively.

Figure 4.12 Gauteng sexual crimes hotspot areas

Figure 4.13 Kwa-Zulu/Natal murder hotspot areas

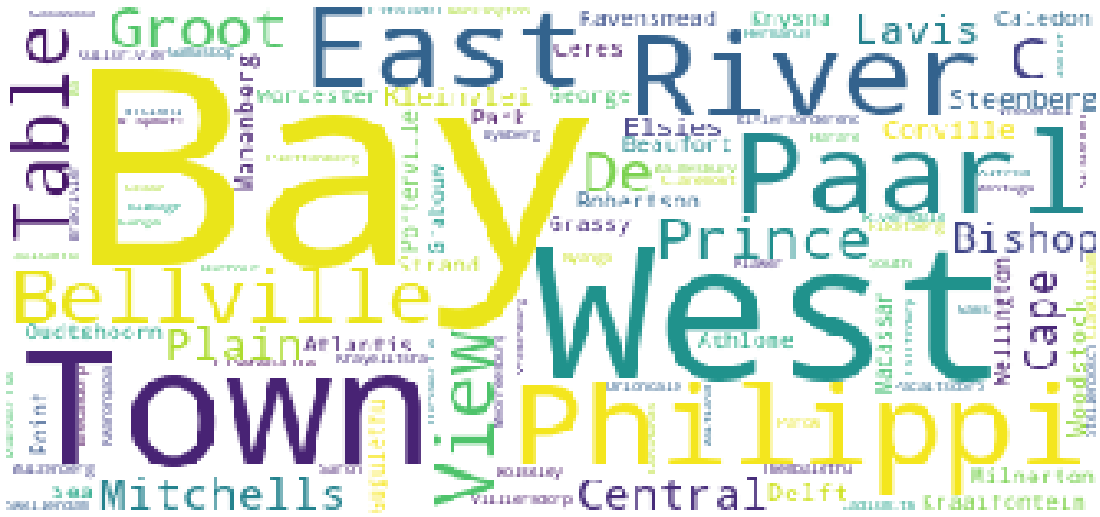


Figure 4.14 The Western Cape drug-related crimes hotspot areas

4.4.1 Interactive visuals

Recognising the audience with whom one works is integral to information dissemination. One must be mindful that numeric literacy is low in Africa and exceptionally low in South Africa (Khuluvhe, 2021); given this scenario, one needs to communicate the urgency and importance of the dire state of serious crimes in the country. Visualisations are an ideal tool to impress upon the reader the scope, depth, and range of these terrible crimes. Therefore, the researcher decided to use visualisations; however, the researcher also realised that animating the visualisations would have an even more profound impact. For these reasons, the researcher created the Flourish visualisations as a recommendation for the South African Police Service (SAPS) to have an animated dashboard rather than a simple static dashboard for all crime-related data.

The following interactive maps are plotted based on the total number of sexual, murder, and drug-related crimes reported per station in each province.

4.4.2 Sexual crimes

The interactive maps for the total number of sexual crimes reported per station in each province in the study area are presented in Figures 4.15 to 4.23.

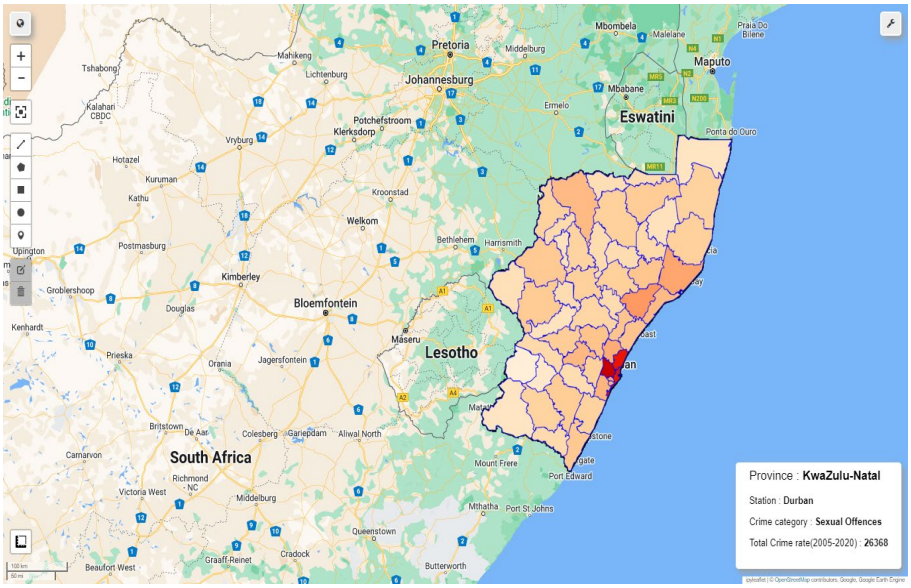


Figure 4.15 Total sexual crime rates in Durban, KZN Pronvince (2005-2020)

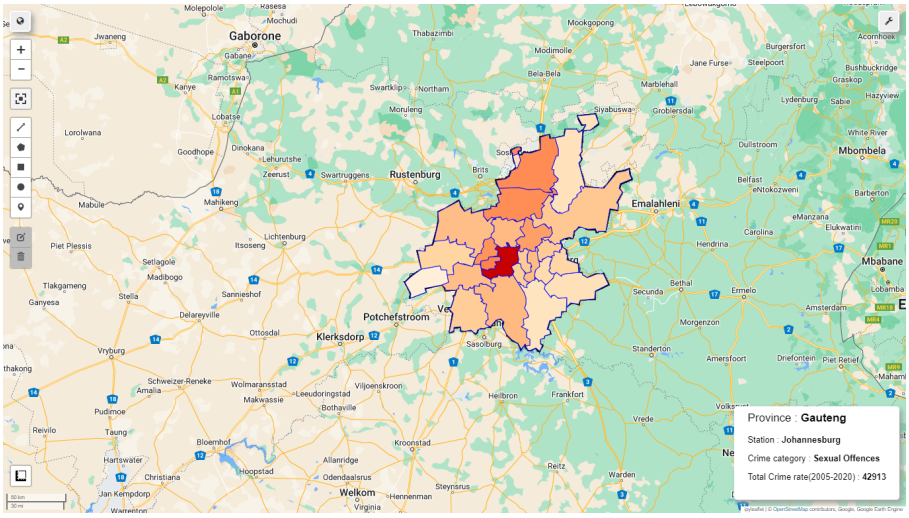


Figure 4.16 Total sexual crime rates in Johannesburg, Gauteng Province (2005-2020)

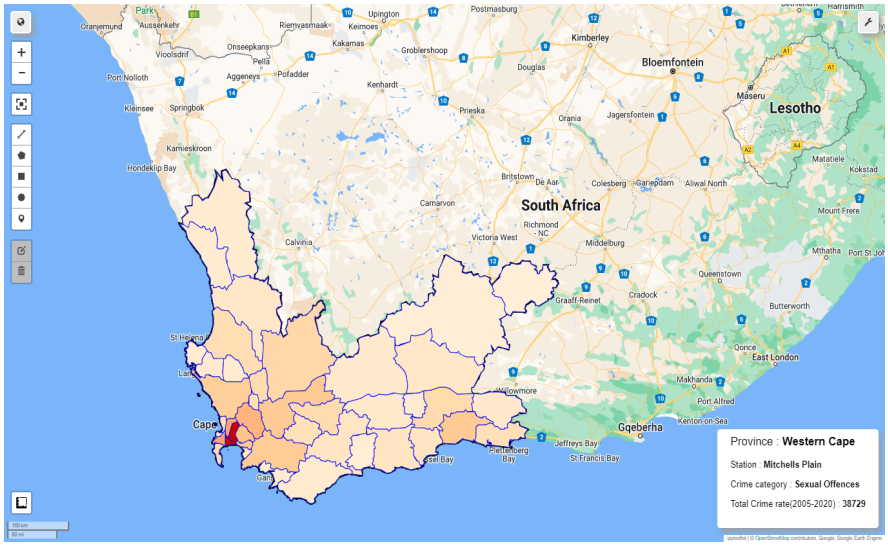


Figure 4.17 Total sexual crime rates in Mitchells Plain, Western Cape Province (2005-2020)

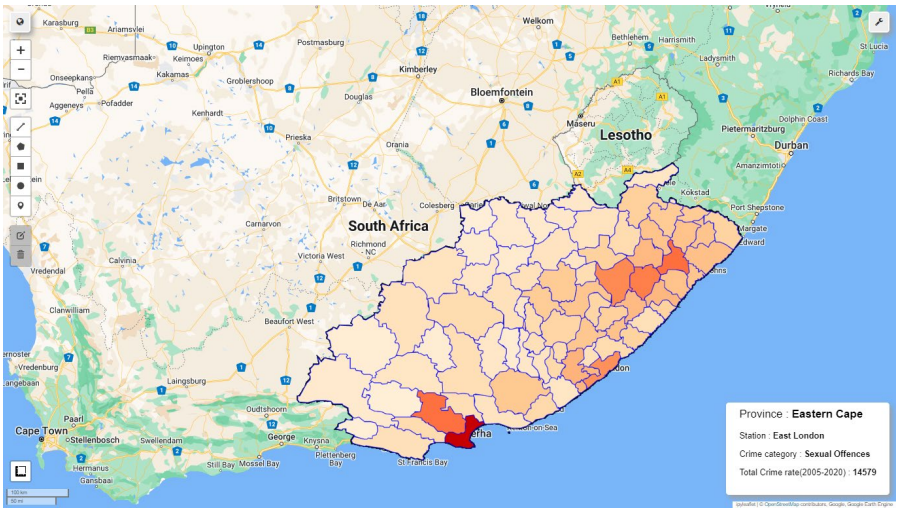


Figure 4.18 Total sexual crime rates in East London, Eastern Cape Province (2005-2020)

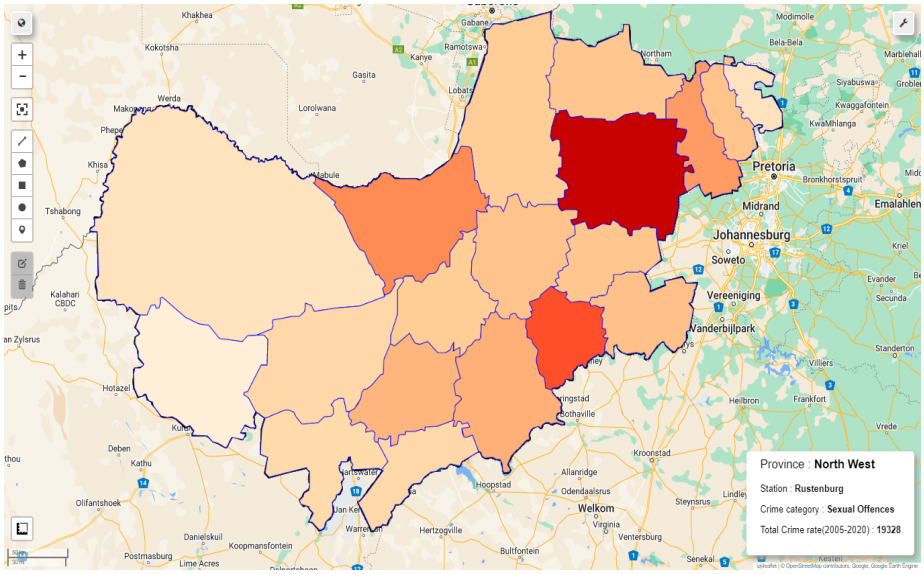


Figure 4.19 Total sexual crime rates in Rustenburg, North West Province (2005-2020)

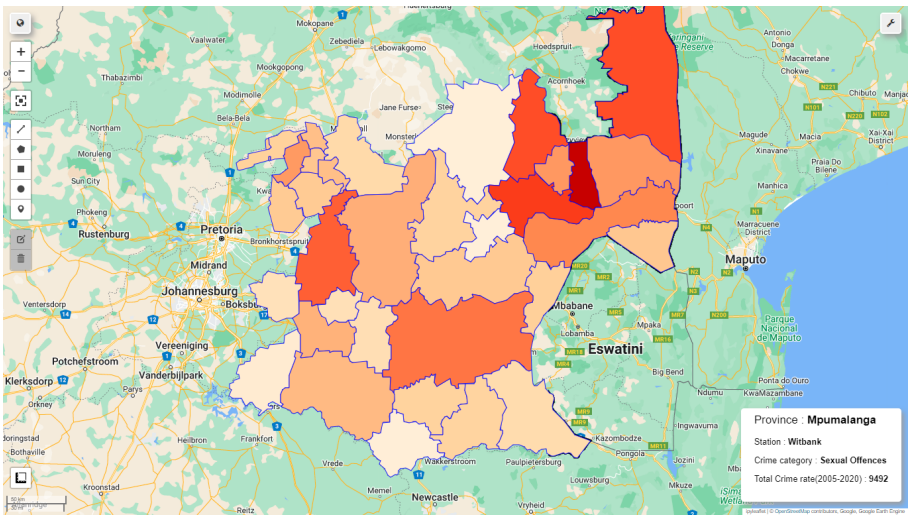


Figure 4.20 Total sexual crime rates in Witbank, Mpumalanga Province (2005-2020)

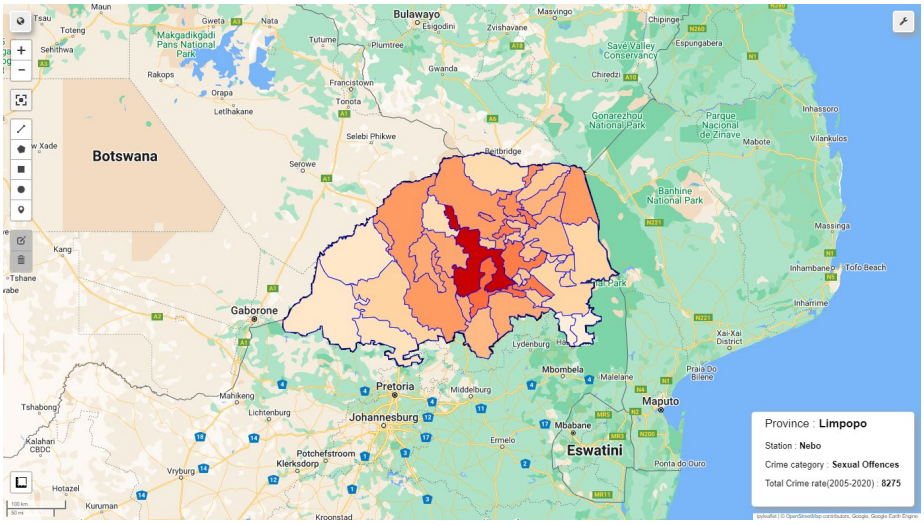


Figure 4.21 Total sexual crime rates in Nebo, Limpopo Province (2005-2020)

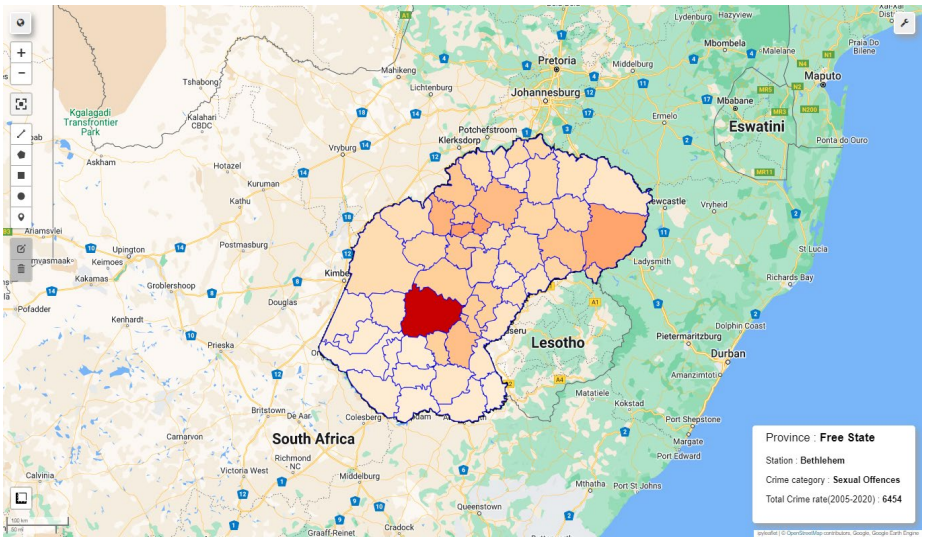


Figure 4.22 Total sexual crime rates in Bethlehem, Free State Province (2005-2020)

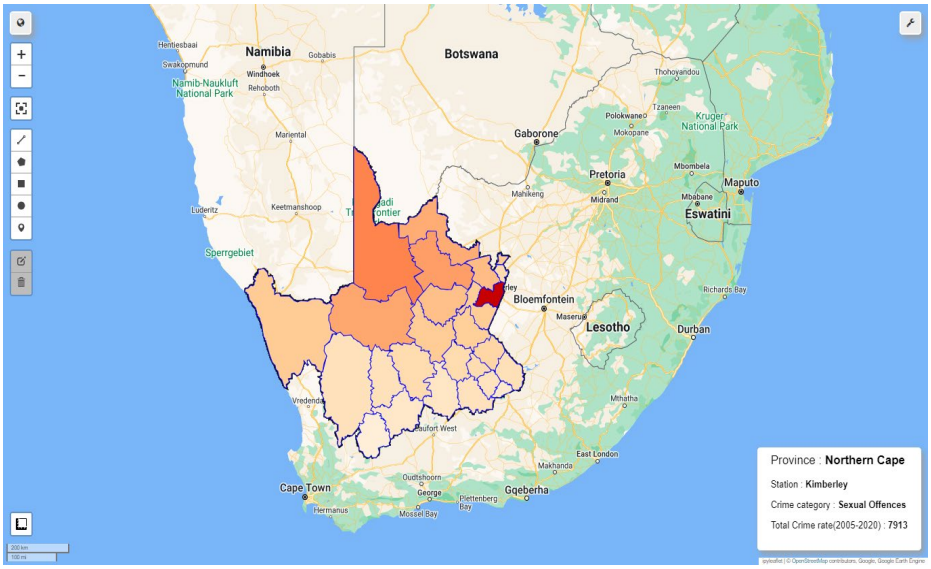


Figure 4.23 Total sexual crime rates in Kimberley, Northern Cape Province (2005-2020)

4.4.3 Murder

The interactive maps for the total number of murder crimes reported per station in each province in the study area are presented in Figures 4.24 to 4.32.

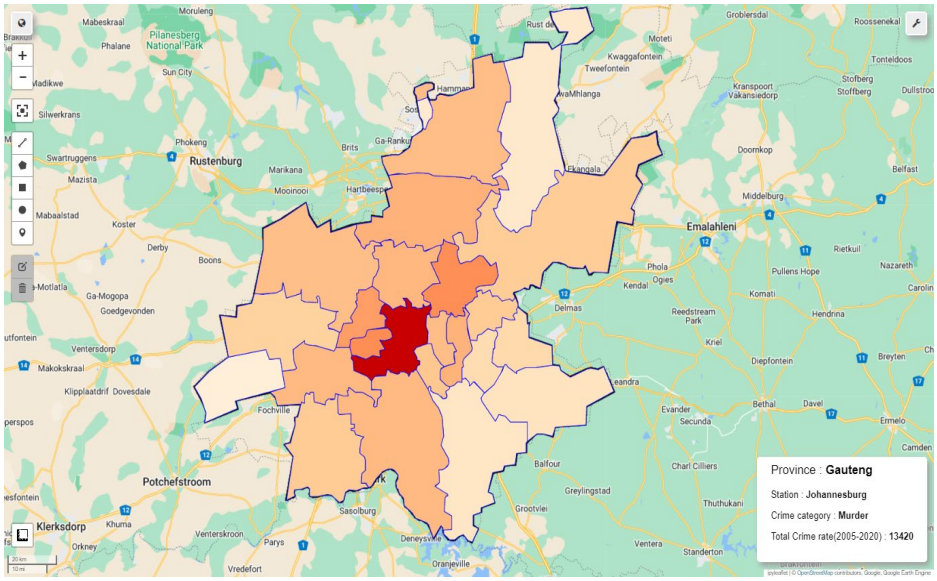


Figure 4.24 Total murder rates in Johannesburg, Gauteng Province (2005-2020)

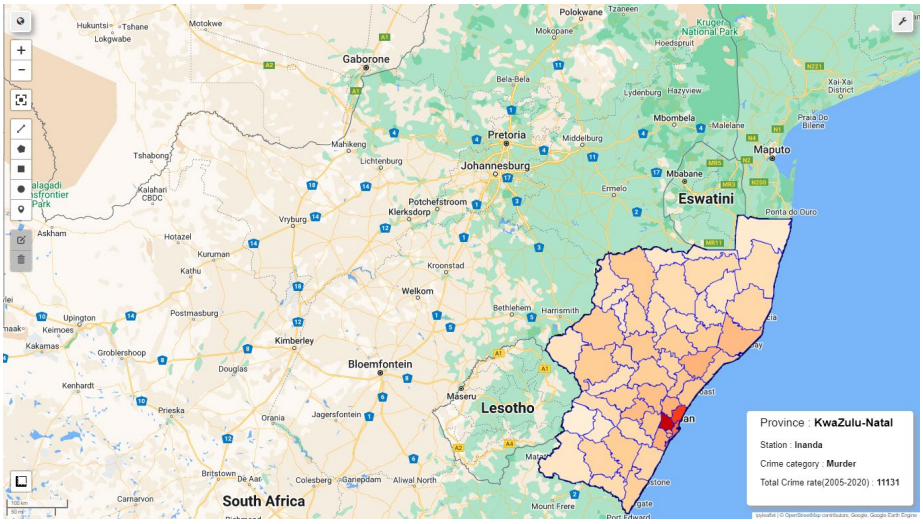


Figure 4.25 Total murder rates in Inanda, KZN Provine (2005-2020)

Figure 4.26 Total murder rates in Mitchells Plain, Western Cape Province (2005-2020)

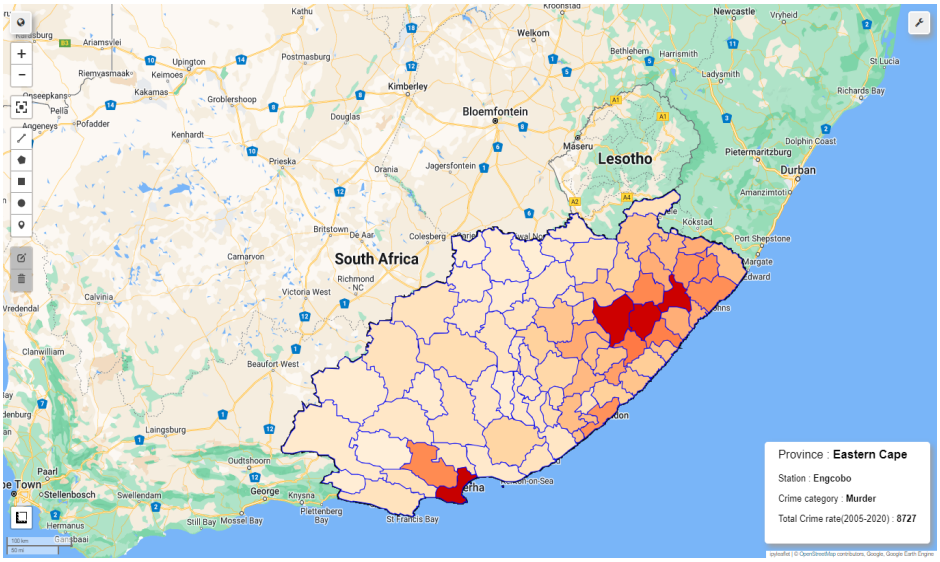


Figure 4.27 Total murder rates in Engcobo, Eastern Cape Province (2005-2020)

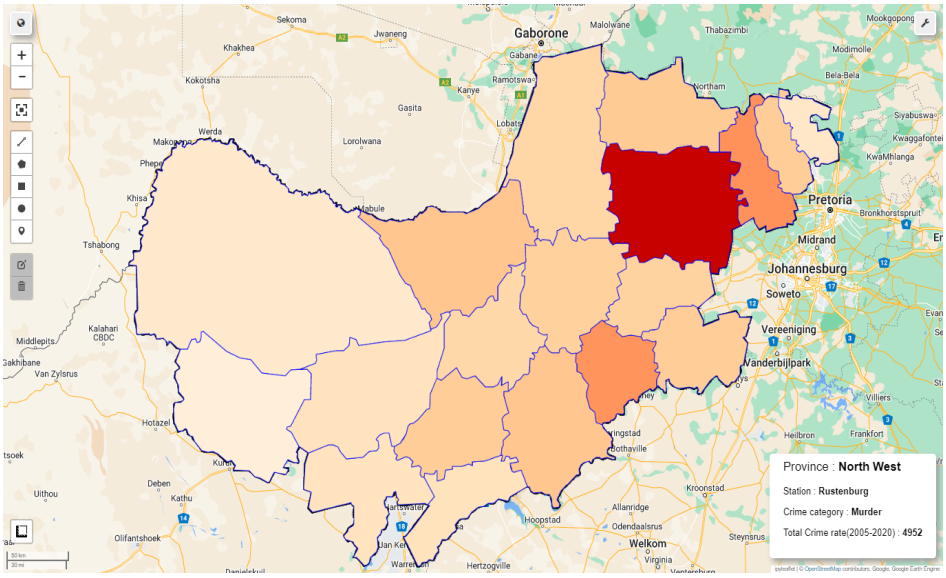


Figure 4.28 Total murder rates in Rustenburg, North West Province (2005-2020)

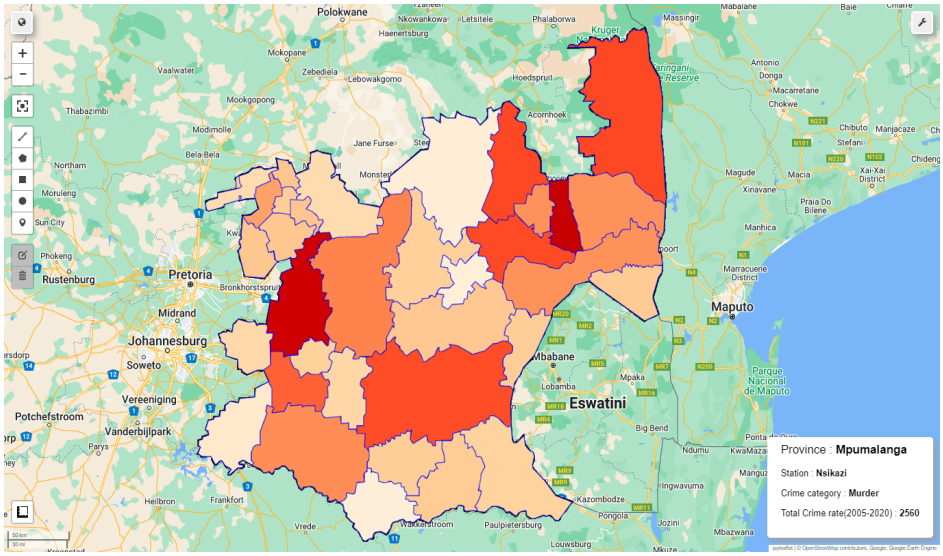


Figure 4.29 Total murder rates in Nkwazi, Mpumalanga Province (2005-2020)

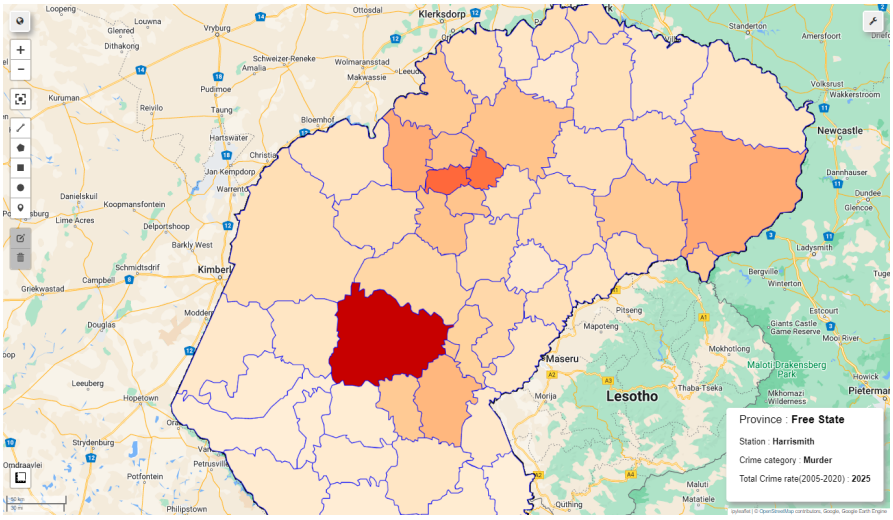


Figure 4.30 Total murder rates in Harrismith, Free State Province (2005-2020)

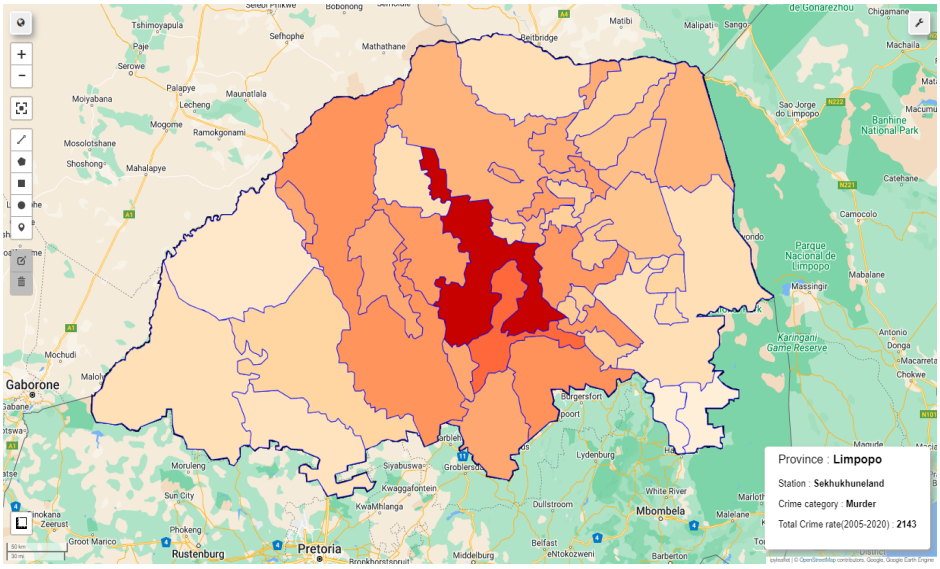


Figure 4.31 Total murder rates in Sekhukhuneland, Limpopo Province (2005-2020)

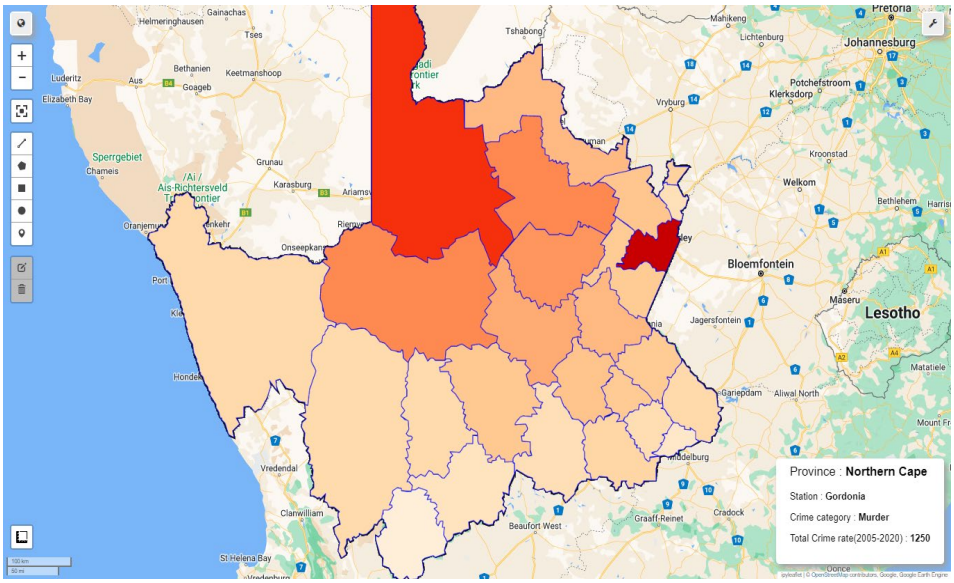


Figure 4.32 Total murder rates in Gordonia, Northern Cape Province (2005-2020)

4.4.4 Drug-related crimes

The interactive maps for the total number of drug-related crimes reported per station in each province in the study area are presented in Figures 4.33 to 4.41.

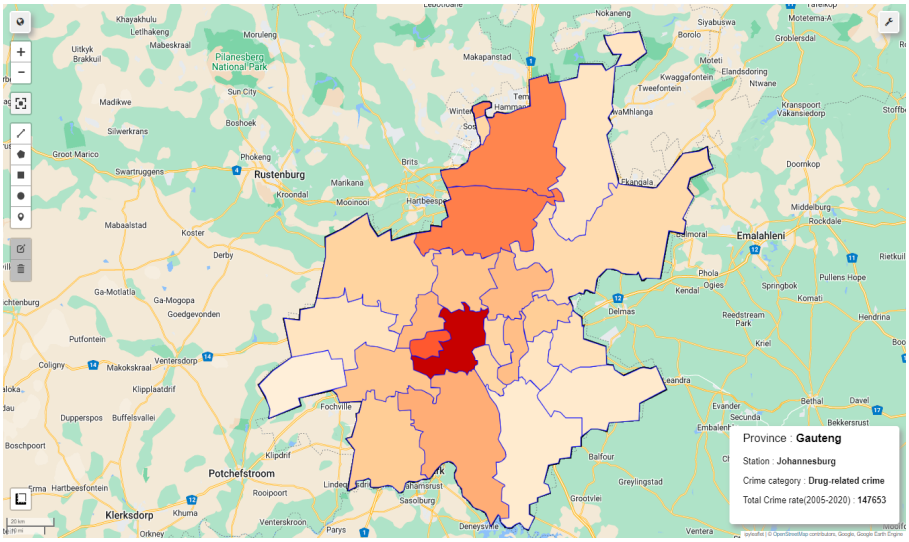


Figure 4.33 Total drug-related crime rates in Johannesburg, Gauteng Province (2005-2020)

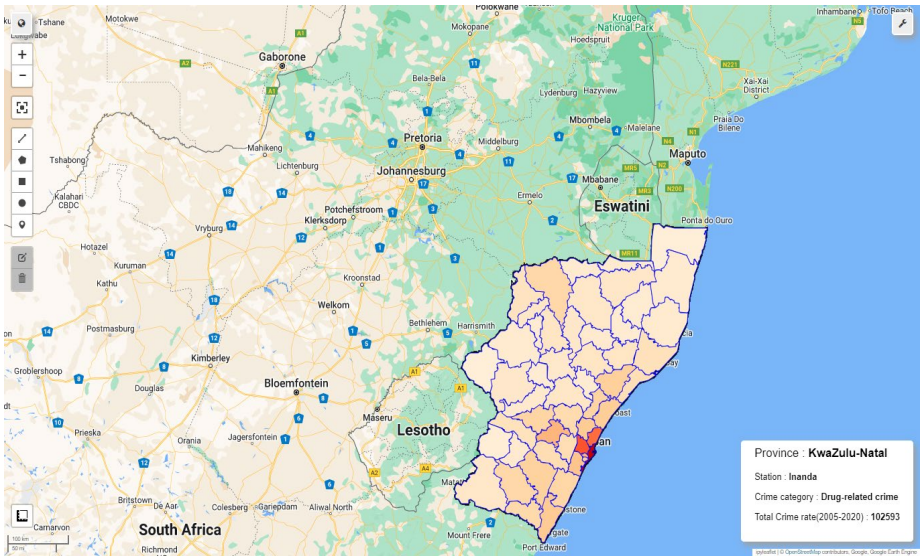


Figure 4.34 Total drug-related crime rates in Inanda, KZN Province (2005-2020)

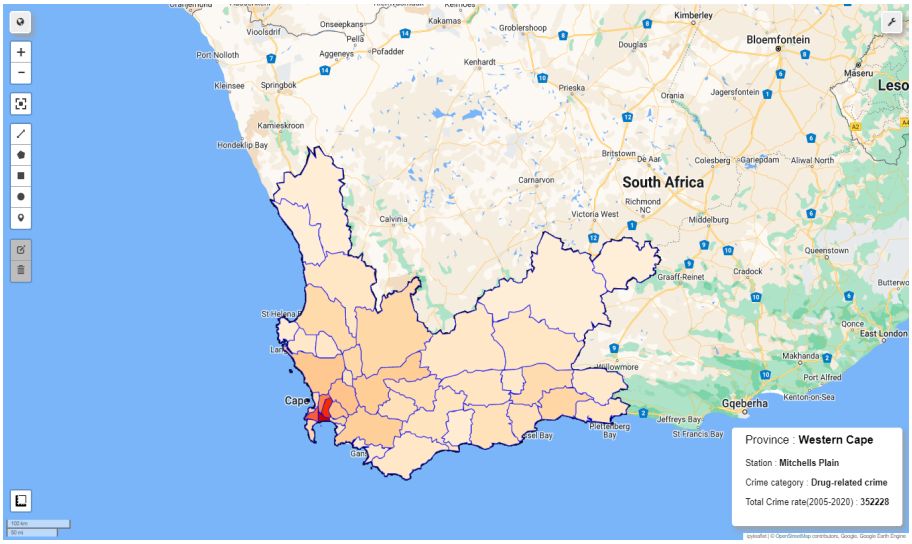


Figure 4.35 Total drug-related crime rates in Mitchells Plain, Western Cape Province (2005-2020)

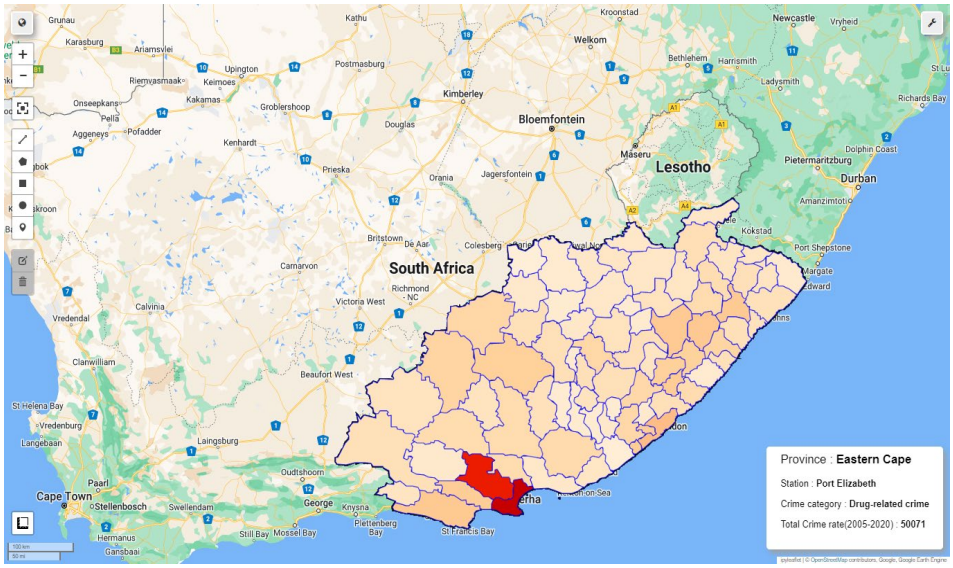


Figure 4.36 Total drug-related crime rates in Port Elisabeth, Eastern Cape Province (2005-2020)

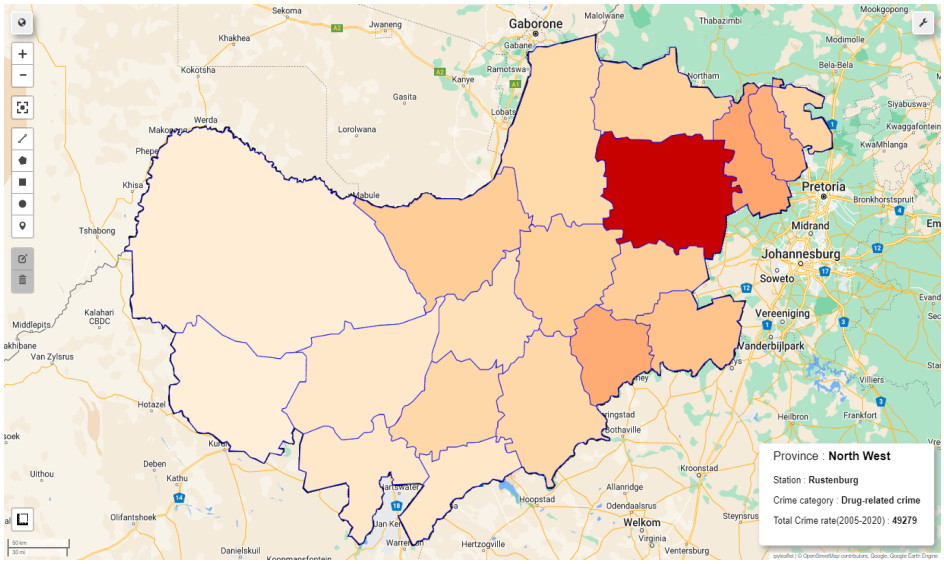


Figure 4.37 Total drug-related crime rates in Rustenburg, North West (2005-2020)

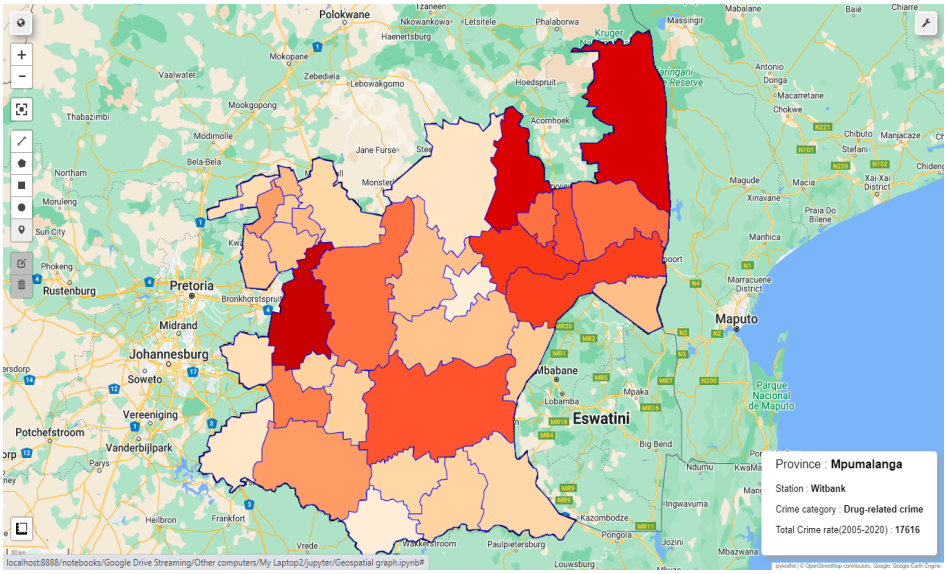


Figure 4.38 Total drug-related crime rates in Witbank, Mpumalanga Province (2005-2020)

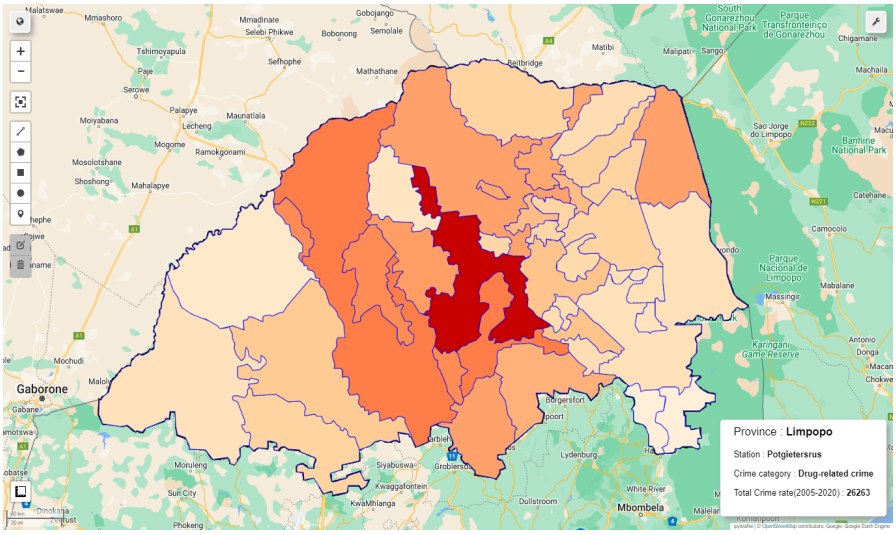


Figure 4.39 Total drug-related crime rates in Potgietersrus, Limpopo Province (2005-2020)

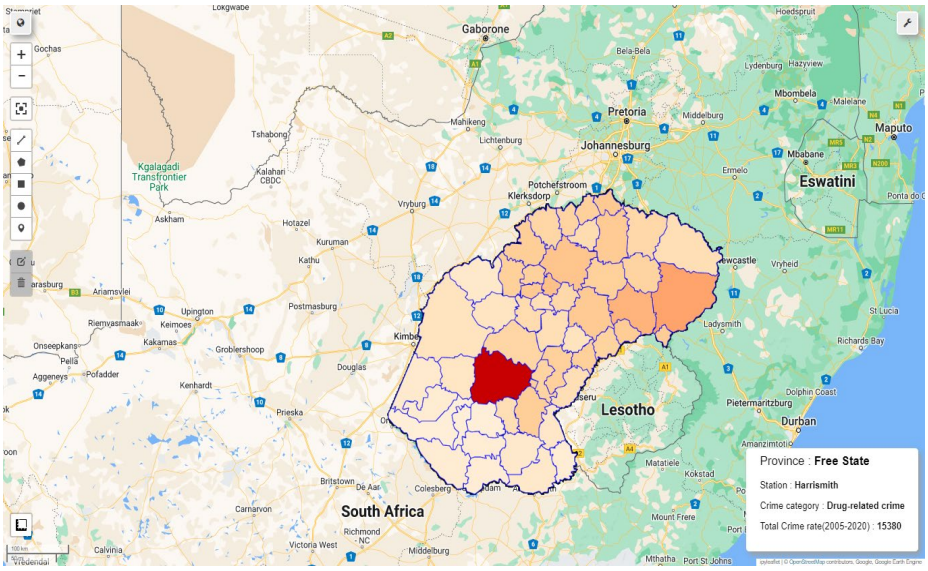


Figure 4.40 Total drug-related crime rates in Harrismith, Free State Province (2005-2020)

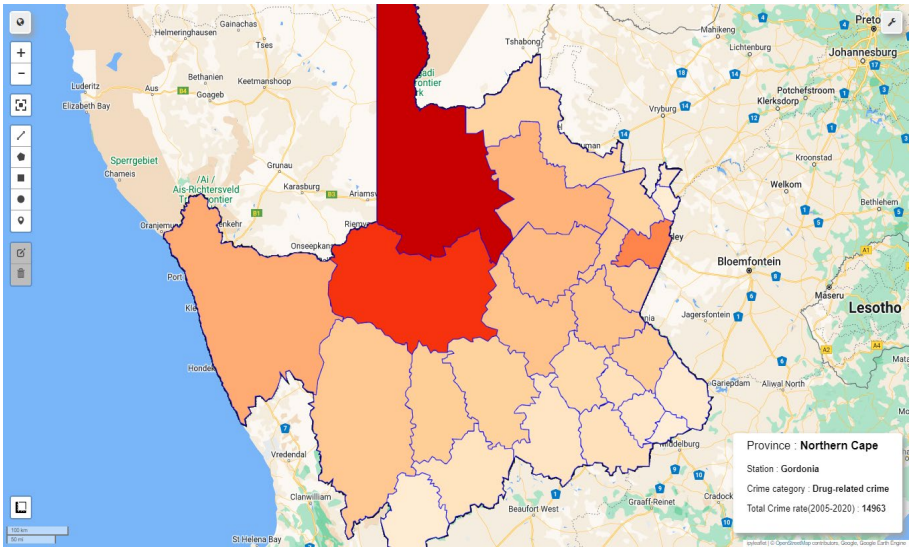


Figure 4.41 Total drug-related crime rates in Gordonia, Northern Cape Province (2005–2020)

The interactive maps presented in Figure 4.15 to Figure 4.41 show the total reported crimes in each station per province. This visualisation of all the station's crime numbers can be obtained by hovering the mouse over the map in the Python (Anaconda) environment. The stations with darker colours indicate higher crime rates in that province.

4.4.5 Flourish animated visualisation

The animated visualisation[4] in Figure 4.42 depicts the number of cases reported in each station for a specific crime category and year within a specific province. A click on the province displays the annual rate of the three crimes (murder, drug-related, and sexual crimes) in each of the stations.

[4] https://public.flourish.studio/story/1746396

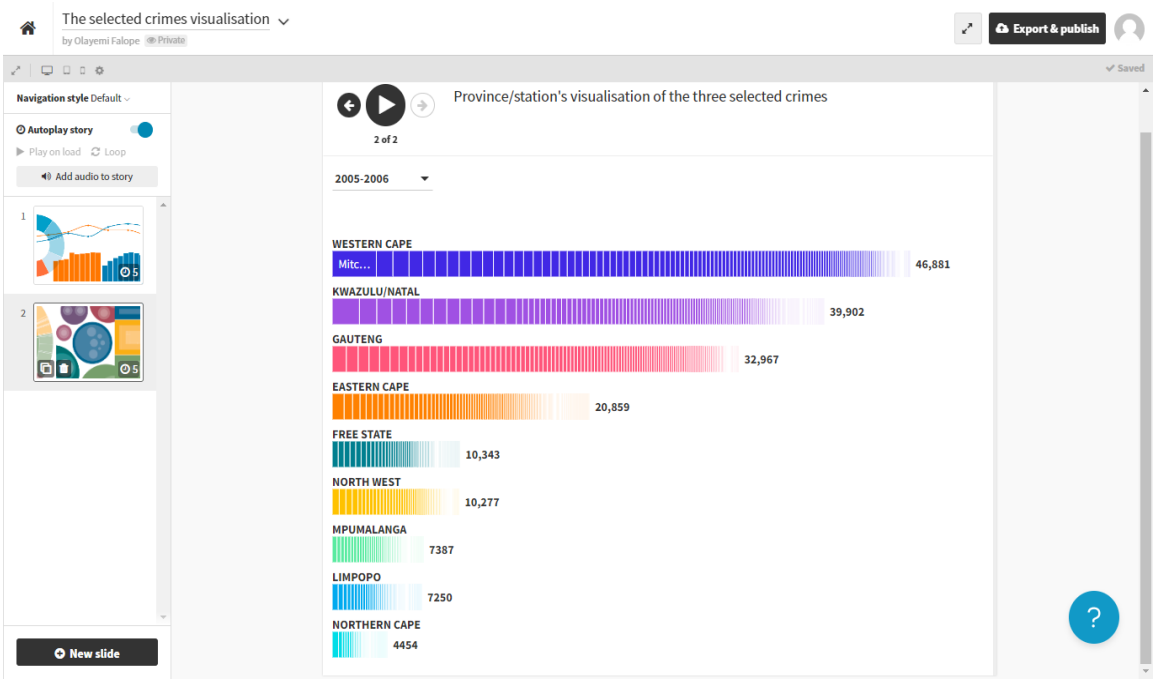


Figure 4.42 Reported cases in Station per the Province

4.5 Linear Correlation Analysis using Heatmaps

Correlation is a statistical method that can be employed for testing hypotheses about causal relationships between variables. Correlation is frequently employed in practical applications for predicting trends. For example, in the case where a province exhibits a robust positive correlation between its population size and the frequency of crime incidents, it is reasonable to infer that an increase in population would correspondingly result in an elevated occurrence of crimes. The correlation coefficient can assume any value from -1 to 1 (Schober, Boer and Schwarte 2018). In the case that the value is equal to 1, a positive correlation would exist between the two variables, indicating that an increase in one variable would correspondingly result in an increase in the other variable. A value of -1 indicates a negative correlation between the two variables, implying that as one variable increases, the other variable decreases (Schober, Boer and Schwarte 2018). The value of 0 indicates the absence of correlation between the two variables, implying that the variables exhibit random changes independent of each other.

It is important to confirm that certain assumptions are met before executing a linear regression model. In this part of the study, a check is made whether a linear relationship exists between the dependent variable and the independent variables, which means if a linear relationship exists between the three selected serious crimes (dependent variables) and population, density, and area (independent variables). Heatmap correlation was carried out on the three selected serious crimes, population, density, and area in the following:

4.5.1 Sexual crimes heatmap

The correlation heatmap of the sexual crimes is presented in Figure 4.43. Based on the heatmap, a strong positive correlation is evident between the total number of sexual crimes and the population. However, there is also a correlation between the number of sexual crimes and the density of the province. A positive correlation exists between the total number of police stations in the province and the total number of sexual crimes. It is important to note that this correlation does not imply more sexual crime simply because there are more police stations. However, there may be more stations to combat more serious sexual crimes. A solid negative correlation can also be seen between the size of the province's land area and the number of sexual crimes. Figure 4.43 can be further illustrated with a correlation gradient for more clarity in Figure 4.44.

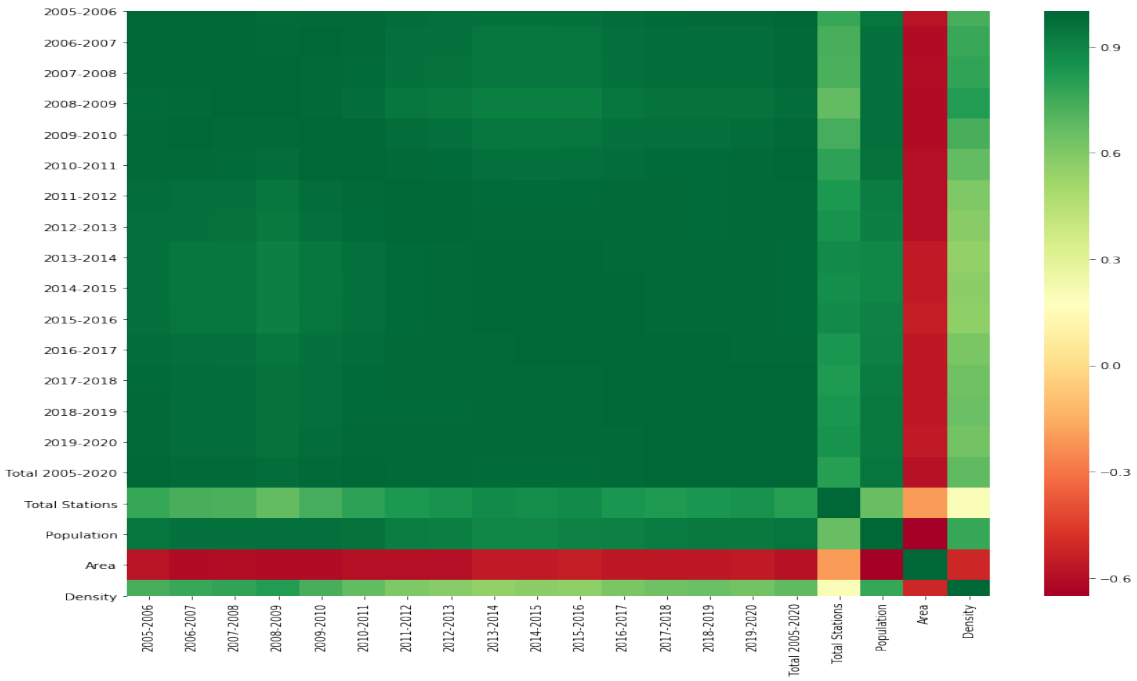
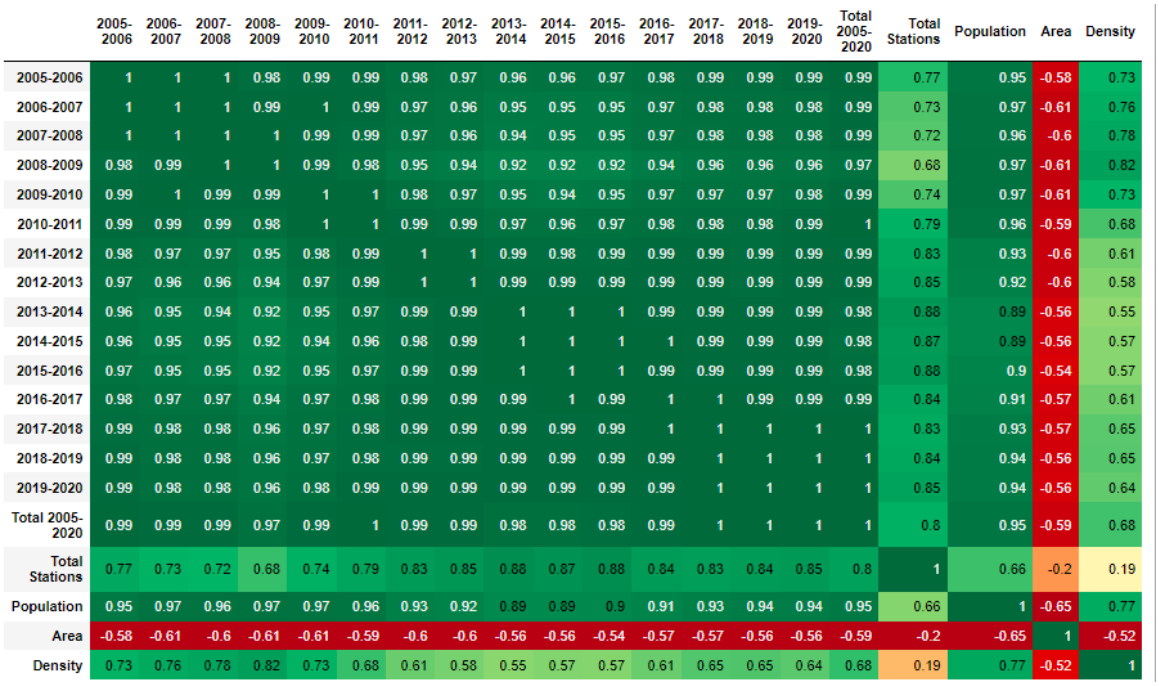


Figure 4.43 Sexual crimes correlation heatmap

Figure 4.44 presents a correlation gradient that shows a clear and understandable correlation of 0.95 between the total number of sexual crimes and the population. This means that there is a strong positive correlation between them. That is, an increase in population leads to an increase in sexual crimes, which means the higher the population, the higher the sexual crimes. The total number of sexual crimes and density was shown to be 0.68, which means a positive correlation occurs between them. Lastly, the total number of sexual crimes and area was shown to be -0.59, which denotes a strong negative correlation between them.



	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total 2005-2020	Total Stations	Population	Area	Density
2005-2006	1	1	1	0.98	0.99	0.99	0.98	0.97	0.96	0.96	0.97	0.98	0.99	0.99	0.99	0.99	0.77	0.95	-0.58	0.73
2006-2007	1	1	1	0.99	1	0.99	0.97	0.96	0.95	0.95	0.95	0.97	0.98	0.98	0.98	0.99	0.73	0.97	-0.61	0.76
2007-2008	1	1	1	1	0.99	0.99	0.97	0.96	0.94	0.95	0.95	0.97	0.98	0.98	0.98	0.99	0.72	0.96	-0.6	0.78
2008-2009	0.98	0.99	1	1	0.99	0.98	0.95	0.94	0.92	0.92	0.92	0.94	0.96	0.96	0.96	0.97	0.68	0.97	-0.61	0.82
2009-2010	0.99	1	0.99	0.99	1	1	0.98	0.97	0.95	0.94	0.95	0.97	0.97	0.97	0.98	0.99	0.74	0.97	-0.61	0.73
2010-2011	0.99	0.99	0.99	0.98	1	1	0.99	0.99	0.97	0.96	0.97	0.98	0.98	0.98	0.99	1	0.79	0.96	-0.59	0.68
2011-2012	0.98	0.97	0.97	0.95	0.98	0.99	1	1	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.83	0.93	-0.6	0.61
2012-2013	0.97	0.96	0.96	0.94	0.97	0.99	1	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.85	0.92	-0.6	0.58
2013-2014	0.96	0.95	0.94	0.92	0.95	0.97	0.99	0.99	1	1	1	0.99	0.99	0.99	0.99	0.98	0.88	0.89	-0.56	0.55
2014-2015	0.96	0.95	0.95	0.92	0.94	0.96	0.98	0.99	1	1	1	1	0.99	0.99	0.99	0.98	0.87	0.89	-0.56	0.57
2015-2016	0.97	0.95	0.95	0.92	0.95	0.97	0.99	0.99	1	1	1	0.99	0.99	0.99	0.99	0.98	0.88	0.9	-0.54	0.57
2016-2017	0.98	0.97	0.97	0.94	0.97	0.98	0.99	0.99	0.99	1	0.99	1	1	0.99	0.99	0.99	0.84	0.91	-0.57	0.61
2017-2018	0.99	0.98	0.98	0.96	0.97	0.98	0.99	0.99	0.99	0.99	0.99	1	1	1	1	1	0.83	0.93	-0.57	0.65
2018-2019	0.99	0.98	0.98	0.96	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1	0.84	0.94	-0.56	0.65
2019-2020	0.99	0.98	0.98	0.96	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1	1	0.85	0.94	-0.56	0.64
Total 2005-2020	0.99	0.99	0.99	0.97	0.99	1	0.99	0.99	0.98	0.98	0.98	0.99	1	1	1	1	0.8	0.95	-0.59	0.68
Total Stations	0.77	0.73	0.72	0.68	0.74	0.79	0.83	0.85	0.88	0.87	0.88	0.84	0.83	0.84	0.85	0.8	1	0.66	-0.2	0.19
Population	0.95	0.97	0.96	0.97	0.97	0.96	0.93	0.92	0.89	0.89	0.9	0.91	0.93	0.94	0.94	0.95	0.66	1	-0.65	0.77
Area	-0.58	-0.61	-0.6	-0.61	-0.61	-0.59	-0.6	-0.6	-0.56	-0.56	-0.54	-0.57	-0.57	-0.56	-0.56	-0.59	-0.2	-0.65	1	-0.52
Density	0.73	0.76	0.78	0.82	0.73	0.68	0.61	0.58	0.55	0.57	0.57	0.61	0.65	0.65	0.64	0.68	0.19	0.77	-0.52	1

Figure 4.44 Sexual crimes correlation gradient

4.5.2 Murder heatmap

As evident in Figure 4.45, there is a strong positive correlation between the total number of murder crimes and the population. There is also a strong positive correlation between the number of murder crimes and the number of police stations in the province. There is a weak positive correlation between the density and the total number of murder crimes. A strong negative correlation can also be seen between the area and the number of murder crimes. Figure 4.45 is further explained with a correlation gradient for more clarity in Figure 4.46.

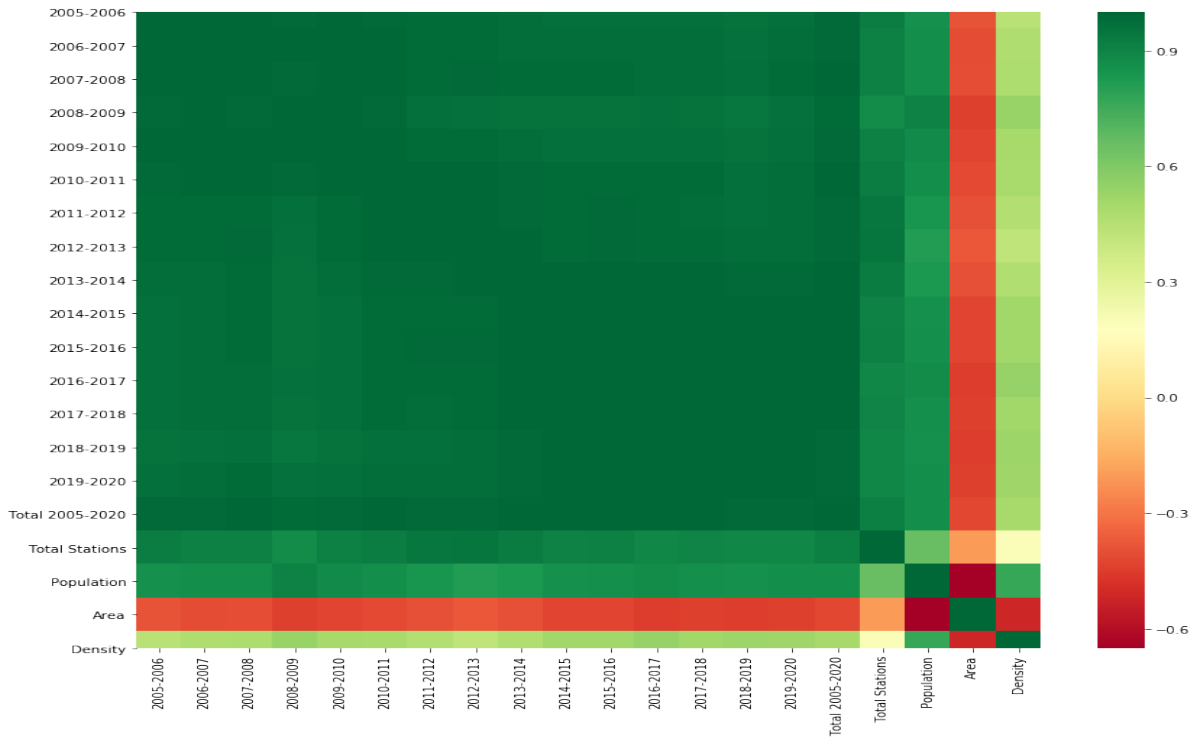


Figure 4.45 Murder correlation heatmap

	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total 2005-2020	Total Stations	Population	Area	Density
2005-2006	1	1	1	0.99	1	0.99	0.99	0.99	0.98	0.97	0.97	0.97	0.97	0.96	0.97	0.99	0.92	0.85	-0.39	0.44
2006-2007	1	1	1	0.99	1	0.99	0.98	0.99	0.98	0.98	0.98	0.97	0.98	0.97	0.98	0.99	0.91	0.87	-0.41	0.47
2007-2008	1	1	1	0.99	1	1	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.97	0.98	1	0.91	0.87	-0.4	0.48
2008-2009	0.99	0.99	0.99	1	1	0.99	0.97	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.98	0.87	0.9	-0.44	0.54
2009-2010	1	1	1	1	1	1	0.99	0.98	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.99	0.91	0.88	-0.42	0.5
2010-2011	0.99	0.99	1	0.99	1	1	1	0.99	0.99	0.98	0.99	0.98	0.98	0.97	0.98	1	0.93	0.87	-0.42	0.5
2011-2012	0.99	0.98	0.99	0.97	0.99	1	1	1	0.99	0.98	0.99	0.98	0.98	0.97	0.97	0.99	0.95	0.84	-0.4	0.46
2012-2013	0.99	0.99	0.99	0.97	0.98	0.99	1	1	1	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.95	0.82	-0.38	0.42
2013-2014	0.98	0.98	0.98	0.96	0.97	0.99	0.99	1	1	1	1	0.99	0.99	0.99	0.99	1	0.93	0.83	-0.39	0.47
2014-2015	0.97	0.98	0.98	0.96	0.97	0.98	0.98	0.99	1	1	1	1	1	1	1	0.99	0.9	0.86	-0.42	0.51
2015-2016	0.97	0.98	0.98	0.96	0.97	0.99	0.99	0.99	1	1	1	1	1	1	0.99	1	0.92	0.86	-0.43	0.51
2016-2017	0.97	0.97	0.98	0.96	0.97	0.98	0.98	0.98	0.99	1	1	1	1	1	1	0.99	0.9	0.87	-0.45	0.54
2017-2018	0.97	0.98	0.98	0.96	0.97	0.98	0.98	0.98	0.99	1	1	1	1	1	1	0.99	0.9	0.86	-0.44	0.51
2018-2019	0.96	0.97	0.97	0.95	0.96	0.97	0.97	0.98	0.99	1	1	1	1	1	1	0.99	0.89	0.86	-0.44	0.53
2019-2020	0.97	0.98	0.98	0.96	0.97	0.98	0.97	0.98	0.99	1	0.99	1	1	1	1	0.99	0.89	0.87	-0.44	0.52
Total 2005-2020	0.99	0.99	1	0.98	0.99	1	0.99	0.99	1	0.99	1	0.99	0.99	0.99	0.99	1	0.92	0.87	-0.42	0.5
Total Stations	0.92	0.91	0.91	0.87	0.91	0.93	0.95	0.95	0.93	0.9	0.92	0.9	0.9	0.89	0.89	0.92	1	0.66	-0.2	0.19
Population	0.85	0.87	0.87	0.9	0.88	0.87	0.84	0.82	0.83	0.86	0.86	0.87	0.86	0.86	0.87	0.87	0.66	1	-0.65	0.77
Area	-0.39	-0.41	-0.4	-0.44	-0.42	-0.42	-0.4	-0.38	-0.39	-0.42	-0.43	-0.45	-0.44	-0.44	-0.44	-0.42	-0.2	-0.65	1	-0.52
Density	0.44	0.47	0.48	0.54	0.5	0.5	0.46	0.42	0.47	0.51	0.51	0.54	0.51	0.53	0.52	0.5	0.19	0.77	-0.52	1

Figure 4.46 Murder correlation gradient

Figure 4.46 shows clearly that the correlation between the total number of murder crimes and population is 0.87, which means there is a strong positive correlation between them, meaning that an increase in population also leads to an increase in murder crimes, meaning that the higher the population, the higher the murder crimes.

The total number of murders and density was shown to be 0.5, which means the correlation between them is averagely positive. Lastly, the total number of murders and area was shown to be -0.42, which denotes a strong negative correlation between them.

4.5.3 Drug-related crimes heatmap

An average positive correlation between the total number of drug-related crimes and the population is shown in Figure 4.47. Also, there is an average positive correlation between the number of drug-related crimes and the police stations in the province. There is a weak positive correlation between the density and the total number of drug-related crimes. A strong negative correlation is seen between the area and the total number of drug-related crimes. Further explanation with a correlation gradient is displayed in Figure 4.48.

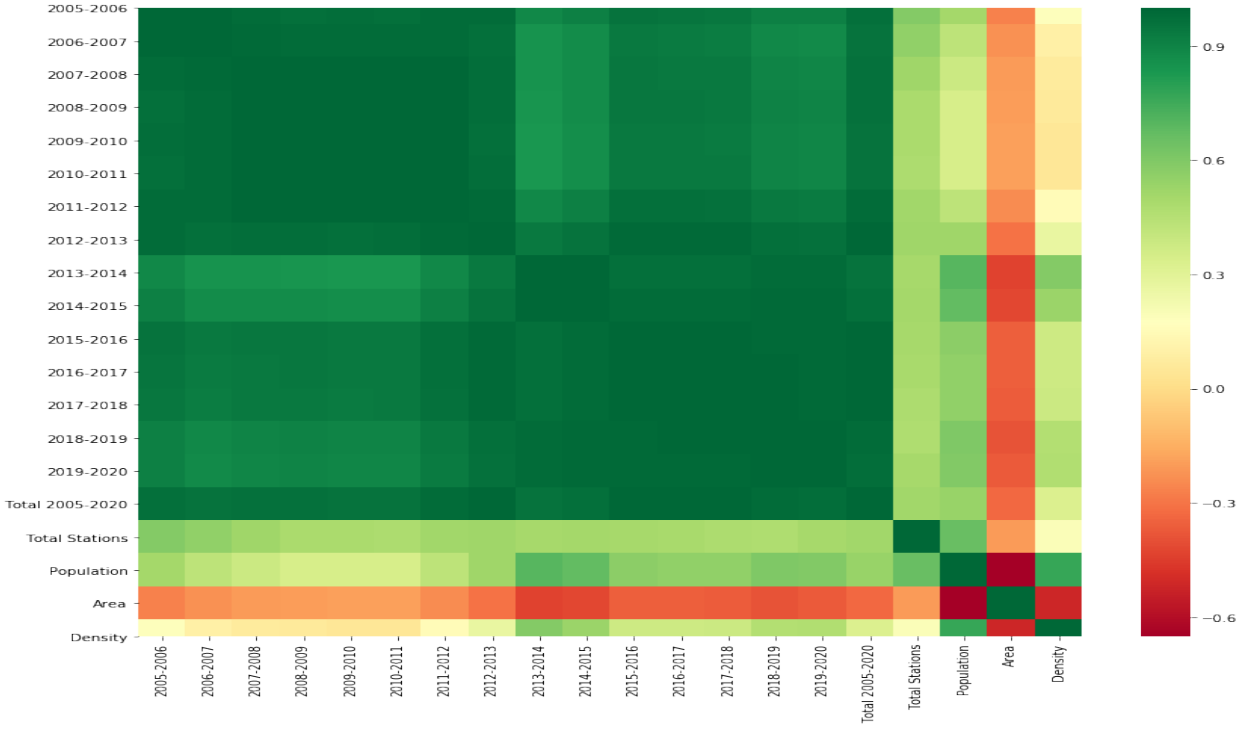
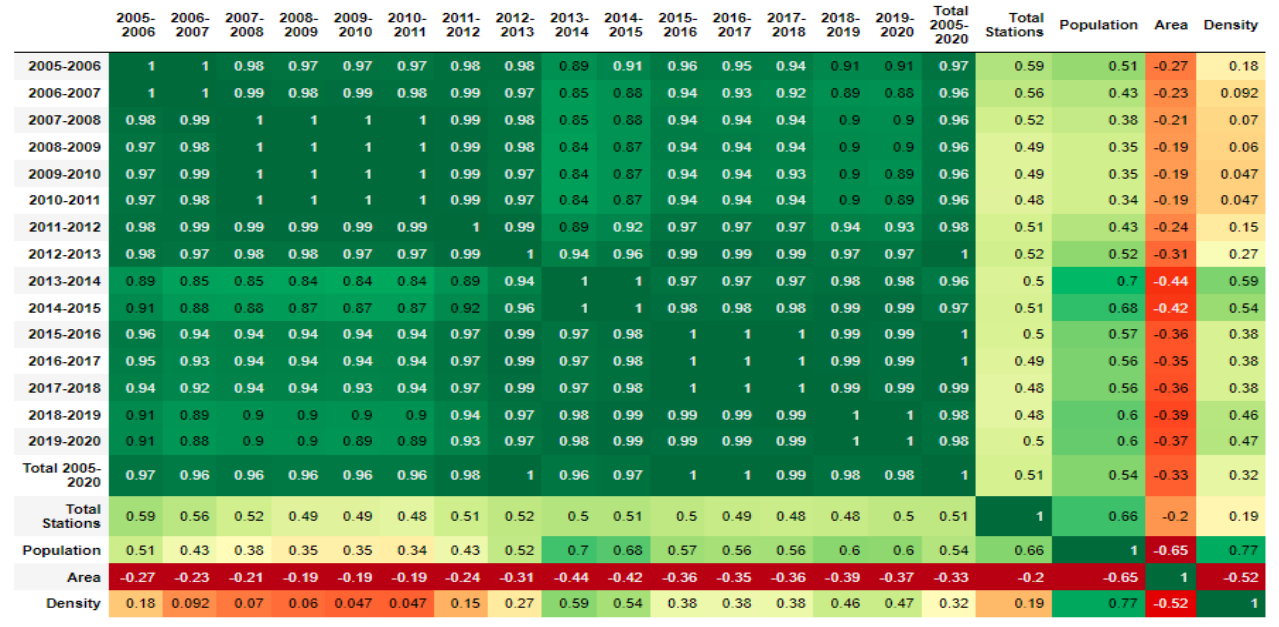


Figure 4.47 Drug-related crimes correlation heatmap



	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	2019-2020	Total 2005-2020	Total Stations	Population	Area	Density
2005-2006	1	1	0.98	0.97	0.97	0.97	0.98	0.98	0.89	0.91	0.96	0.95	0.94	0.91	0.91	0.97	0.59	0.51	-0.27	0.18
2006-2007	1	1	0.99	0.98	0.99	0.98	0.99	0.97	0.85	0.88	0.94	0.93	0.92	0.89	0.88	0.96	0.56	0.43	-0.23	0.092
2007-2008	0.98	0.99	1	1	1	1	0.99	0.98	0.85	0.88	0.94	0.94	0.94	0.9	0.9	0.96	0.52	0.38	-0.21	0.07
2008-2009	0.97	0.98	1	1	1	1	0.99	0.98	0.84	0.87	0.94	0.94	0.94	0.9	0.9	0.96	0.49	0.35	-0.19	0.06
2009-2010	0.97	0.99	1	1	1	1	0.99	0.97	0.84	0.87	0.94	0.94	0.93	0.9	0.89	0.96	0.49	0.35	-0.19	0.047
2010-2011	0.97	0.98	1	1	1	1	0.99	0.97	0.84	0.87	0.94	0.94	0.94	0.9	0.89	0.96	0.48	0.34	-0.19	0.047
2011-2012	0.98	0.99	0.99	0.99	0.99	0.99	1	0.99	0.89	0.92	0.97	0.97	0.97	0.94	0.93	0.98	0.51	0.43	-0.24	0.15
2012-2013	0.98	0.97	0.98	0.98	0.97	0.97	0.99	1	0.94	0.96	0.99	0.99	0.99	0.97	0.97	1	0.52	0.52	-0.31	0.27
2013-2014	0.89	0.85	0.85	0.84	0.84	0.84	0.89	0.94	1	1	0.97	0.97	0.97	0.98	0.98	0.96	0.5	0.7	-0.44	0.59
2014-2015	0.91	0.88	0.88	0.87	0.87	0.87	0.92	0.96	1	1	0.98	0.98	0.98	0.99	0.99	0.97	0.51	0.68	-0.42	0.54
2015-2016	0.96	0.94	0.94	0.94	0.94	0.94	0.97	0.99	0.97	0.98	1	1	1	0.99	0.99	1	0.5	0.57	-0.36	0.38
2016-2017	0.95	0.93	0.94	0.94	0.94	0.94	0.97	0.99	0.97	0.98	1	1	1	0.99	0.99	1	0.49	0.56	-0.35	0.38
2017-2018	0.94	0.92	0.94	0.94	0.93	0.94	0.97	0.99	0.97	0.98	1	1	1	0.99	0.99	0.99	0.48	0.56	-0.36	0.38
2018-2019	0.91	0.89	0.9	0.9	0.9	0.9	0.94	0.97	0.98	0.99	0.99	0.99	0.99	1	1	0.98	0.48	0.6	-0.39	0.46
2019-2020	0.91	0.88	0.9	0.9	0.89	0.89	0.93	0.97	0.98	0.99	0.99	0.99	0.99	1	1	0.98	0.5	0.6	-0.37	0.47
Total 2005-2020	0.97	0.96	0.96	0.96	0.96	0.96	0.98	1	0.96	0.97	1	1	0.99	0.98	0.98	1	0.51	0.54	-0.33	0.32
Total Stations	0.59	0.56	0.52	0.49	0.49	0.48	0.51	0.52	0.5	0.51	0.5	0.49	0.48	0.48	0.5	0.51	1	0.66	-0.2	0.19
Population	0.51	0.43	0.38	0.35	0.35	0.34	0.43	0.52	0.7	0.68	0.57	0.56	0.56	0.6	0.6	0.54	0.66	1	-0.65	0.77
Area	-0.27	-0.23	-0.21	-0.19	-0.19	-0.19	-0.24	-0.31	-0.44	-0.42	-0.36	-0.35	-0.36	-0.39	-0.37	-0.33	-0.2	-0.65	1	-0.52
Density	0.18	0.092	0.07	0.06	0.047	0.047	0.15	0.27	0.59	0.54	0.38	0.38	0.38	0.46	0.47	0.32	0.19	0.77	-0.52	1

Figure 4.48 Drug-related crimes correlation gradient

The correlation between the total number of drug-related crimes and the population, as depicted in Figure 4.48, is 0.54. This means that, there is an average positive correlation between them. This implies that an average increase in population would lead to an average increase in drug-related crimes. The total number of drug-related crimes and density is shown to be 0.32, which means the positive correlation between them is weak. Finally, the total number of drug-related crimes and the area was shown to be -0.33, which indicates a strong negative correlation between them.

4.5.4 Linear regression prediction results

Linear regression predicts the result of a response variable as an independent variable. Linear regression aims to represent the linear correlation between the predictor (independent) variables and the response (dependent) variables. The input variables in linear regression should not be dependent on one another (multicollinear). Multicollinearity among other features was detected easily by feature-feature correlation analysis in Figure 4.49 to Figure 4.54, respectively.

The simplest and most popular estimator is the ordinary least squares (OLS) method, which ensures that the square difference between the predicted and actual values is

as slight as possible. The technique used to figure out the relationship between one or more independent variables and a dependent variable is to minimise the sum of the squares in the difference between the actual and predicted values of the dependent set of variables as a straight line. It can be suitable to predict the values of a continuous response variable using one or more predictors and determine the strength of the relationships between them. The following predictions were carried out using the OLS regression method implemented using the Python programming language.

4.5.5 Sexual crimes prediction

The output results of the OLS regression are shown in Figure 4.49. The p-value (P>|t|) is associated with the model coefficients since the p-value for population (0.000) is less than .05. It can then be concluded that there is a statistically significant relationship between the population and sexual crimes. The R-squared value is the coefficient of determination; it is a statistic that shows how well the regression line approximates the actual data values.

OLS Regression Results

Dep. Variable:	Sexual crimes	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	64.99
Date:	Wed, 16 Nov 2022	Prob (F-statistic):	8.68e-05
Time:	12:06:59	Log-Likelihood:	-99.788
No. Observations:	9	AIC:	203.6
Df Residuals:	7	BIC:	204.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.583e+04	1.18e+04	1.337	0.223	-1.22e+04	4.38e+04
Population	0.0143	0.002	8.061	0.000	0.010	0.019

Omnibus:	0.074	Durbin-Watson:	1.762
Prob(Omnibus):	0.964	Jarque-Bera (JB):	0.262
Skew:	0.149	Prob(JB):	0.877
Kurtosis:	2.219	Cond. No.	1.32e+07

Figure 4.49 Sexual crimes linear regression prediction

The proportion of variance explained is used to evaluate the fitness of the relationship between the model and the response variable. This shows that the model explains a sizeable proportion of the variability in the observed data. The R-squared value ranges

from 0 to 1, with a high value showing that the model can explain more variance. In this case, the R-squared shown in Figure 4.49 is 0.903, which means that around 90% of the variation in the sexual crimes rate is explained by the population variable, which is generally considered a very good rate and not reaching the level of overfitting. Because the adjusted R-Squared (0.889) reflects model complexity (the number of variables) as it relates to the data, the value is always a little lower than the R-Squared values, adjusted R-Squared has a more correct measure of model performance.

The F-statistic (64.99) and the corresponding p-value (8.68e-05) indicate the overall significance of the regression model and whether the predictor variables in the model help explain the variation in the response variable. Since the p-value is less than .05, our model is statistically significant, and the population is thought to help explain why sexual crimes change over time. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are both used to compare models' efficacy in the linear regression process. AIC calculates how much relative information a given model loses; the less information a model loses, the higher the model's quality. As a result, the lower the AIC value, the better. The AIC and BIC values of 203.6 and 204.0 show that the model is of excellent quality. The graph plots of the actual and predicted sexual crime rates are both linear. As a result, the predicted sexual crime rate is like the actual sexual crime rate, showing that the linear regression is correct.

The Omnibus/Prob (Omnibus) is a test of the residual characteristic's skewness and kurtosis. The Prob (Omnibus) function is utilised to conduct a statistical test to determine the probability of the residuals conforming to a normal distribution. The fact that the Prob (Omnibus) value of 0.964 is close to one show that the data is normal. Skewness is a measure of data symmetry, and its value drives Omnibus; in this case, the low skew of 0.149 denotes a normal residual distribution.

The kurtosis, on the other hand, is a measure of data 'peakiness' or curvature; the 2.219 value's higher kurtosis shows tighter clustering of residuals around zero, implying a better model with fewer outliers. For the Durbin-Watson tests for homoscedasticity characteristics, the values must be between 1 and 2; the Durbin-Watson result of 1.762 shows that the data is within the limits.

The Jarque-Bera test is used to figure out whether an error has a normal distribution or not. Jarque-Bera (JB)/Prob (JB) tests both skew and kurtosis, like the Omnibus test; the Prob (JB) value of 0.877, which is close to the Prob (Omnibus) value of 0.964, is a confirmation of the Omnibus test. Figure 4.49 depicts the total results, which prove that the linear regression model is an effective model for predicting the sexual crime rate in South Africa. It can be used to predict sexual crimes in any of the country's nine provinces. The prediction algorithm was trained using sexual crime data and uses population input to predict the occurrence of sexual crimes in Figure 4.50.

Figure 4.50 is a graphical user interface (GUI) experimental prototype to show the prediction output result and to allow users to input independent variables to get the predicted outcomes of sexual crime occurrences.

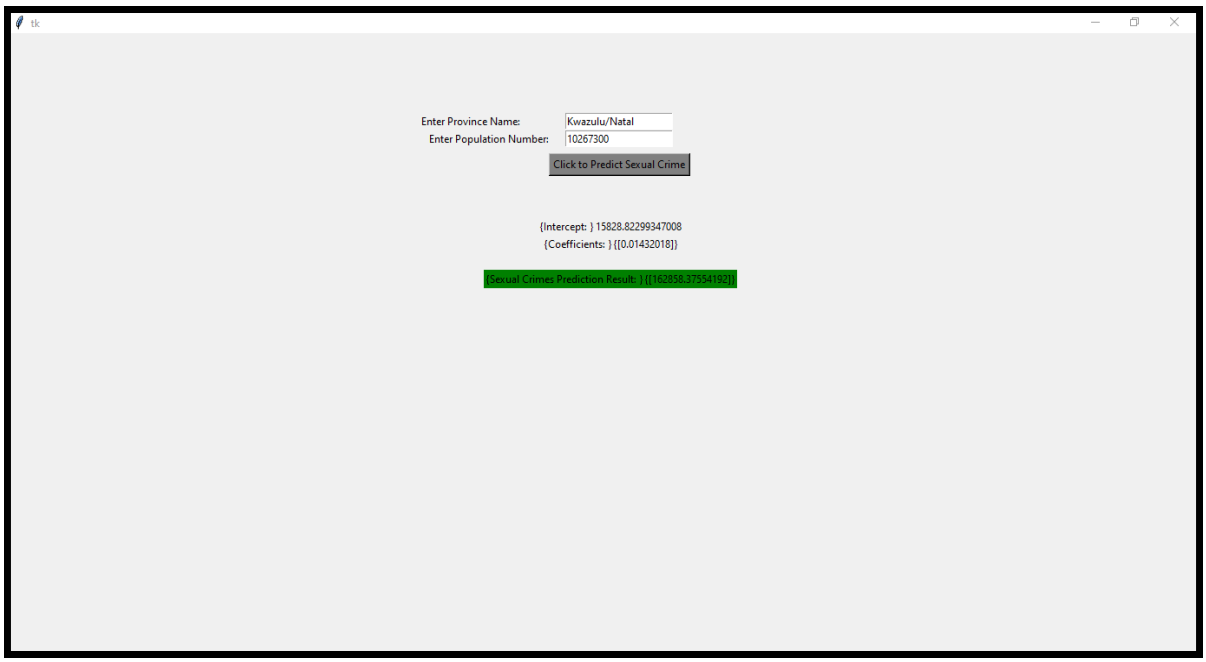


Figure 4.50 Sexual crimes prediction prototype

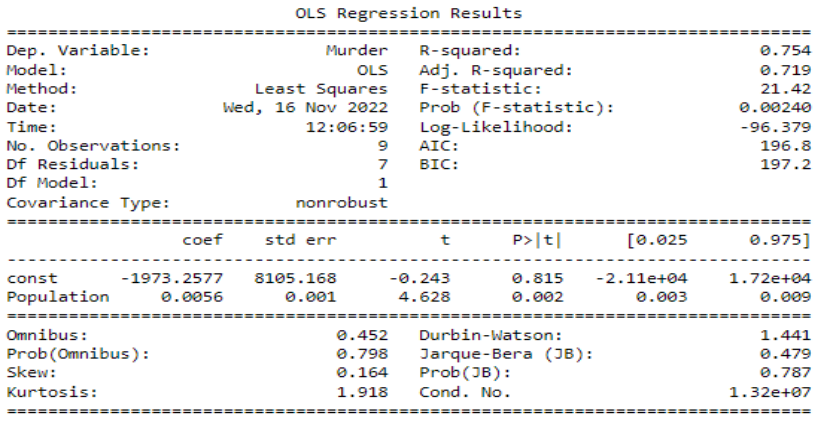
The actual value of sexual crimes and the predicted value results are displayed in Table 4.3.

Table 4.3 Sexual crimes actual and predicted values per province

	PROVINCE	ACTUAL	PREDICTED
0	Gauteng	185 316	191 570
1	Kwazulu-Natal	159 245	162 858
2	Mpumalanga	58 935	73 681
3	Western Cape	122 045	99 211
4	Limpopo	66 367	93 228
5	Eastern Cape	134 977	109 799
6	Northwest	67 604	66 092
7	Free State	63 199	55 146
8	Northern Cape	26 135	32 238

4.5.6 Murder crimes prediction

The p-value (P>|t|) for the population (0.002) as shown in Figure 4.51 is less than .05. This means that the relationship between population and murder crimes is statistically significant. The R-squared is 0.754, which means that around 75% of the variation in the murder crime rate is explained by the population variable, which is generally considered a good rate and not reaching the level of overfitting. The adjusted R-Squared 0.719 reflects model complexity (the number of variables) as it relates to the data; the value is always lower than the R-Squared values, and the adjusted R-Squared is a correct measure of model performance.



OLS Regression Results

Dep. Variable:	Murder	R-squared:	0.754
Model:	OLS	Adj. R-squared:	0.719
Method:	Least Squares	F-statistic:	21.42
Date:	Wed, 16 Nov 2022	Prob (F-statistic):	0.00240
Time:	12:06:59	Log-Likelihood:	-96.379
No. Observations:	9	AIC:	196.8
Df Residuals:	7	BIC:	197.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1973.2577	8105.168	-0.243	0.815	-2.11e+04	1.72e+04
Population	0.0056	0.001	4.628	0.002	0.003	0.009

Omnibus:	0.452	Durbin-Watson:	1.441
Prob(Omnibus):	0.798	Jarque-Bera (JB):	0.479
Skew:	0.164	Prob(JB):	0.787
Kurtosis:	1.918	Cond. No.	1.32e+07

Figure 4.51 Murder crimes linear regression prediction

The F-statistic (21.42) and the corresponding p-value (0.00240) provide information regarding the overall significance of the regression model. Specifically, they indicate whether the predictor variables included in the model effectively explain the variability observed in the response variable. Since the p-value is less than .05, our model is meaningful, and the population is thought to help explain why murder crimes vary over time. The AIC and BIC values of 196.8 and 197.2 showed the model to be of good quality. The Prob (Omnibus) value of 0.798 is close to 1 showing that the data is okay. The low skew of 0.164 denotes a normal residual distribution. The kurtosis 1.918 value is higher than kurtosis showing tighter clustering of residuals around zero, implying a better model with fewer outliers. The Durbin-Watson result of 1.441 shows that the data is within the limits of 1 to 2.

The Prob (JB) value of 0.787, which is close to the Prob (Omnibus) value of 0.798, is a confirmation of the Omnibus test. The results proved that the linear regression model is an effective model for predicting the murder crime rate in South Africa, as depicted in Figure 4.51. It can be used to predict murder crimes in any of the country's nine provinces. The prediction algorithm was trained using murder crime data and population input to predict the occurrence of murder crimes in the prototype in Table 4.4. This was corroborated by the results obtained and presented in Table 4.4.

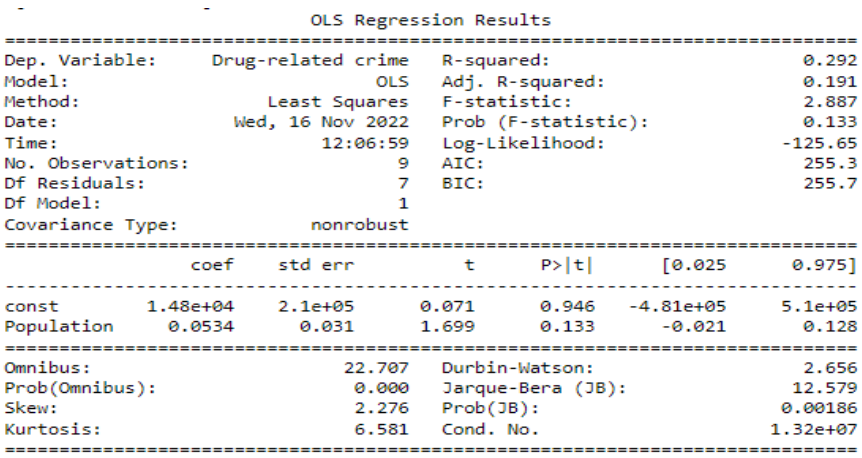
Table 4.4 Murder crimes actual and predicted values per province

	PROVINCE	ACTUAL	PREDICTED
0	Gauteng	56 039	67 111
1	Kwazulu-Natal	63 298	55 824
2	Mpumalanga	12 875	20 768
3	Western Cape	44 545	30 805
4	Limpopo	11 647	28 452
5	Eastern Cape	52 711	34 966
6	Northwest	12 659	17 785
7	Free State	14 220	13 482
8	Northern Cape	5 677	4 477

Table 4.4 showed the murder crimes actual values and the predicted results per province.

4.5.7 Drug-related crimes prediction

As presented in Figure 4.52, the p-value (P>|t|) for the population (0.133) is higher than .05. This means that the relationship between the population and drug-related crimes is very weak. The R-squared is 0.292, which means that around 29% of the variation in the drug-related crime rate is explained by the population variable, which is considered weak. The adjusted R-Squared of 0.191 reflects model complexity (the number of variables) as it relates to the data.



OLS Regression Results

Dep. Variable:	Drug-related crime	R-squared:	0.292
Model:	OLS	Adj. R-squared:	0.191
Method:	Least Squares	F-statistic:	2.887
Date:	Wed, 16 Nov 2022	Prob (F-statistic):	0.133
Time:	12:06:59	Log-Likelihood:	-125.65
No. Observations:	9	AIC:	255.3
Df Residuals:	7	BIC:	255.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.48e+04	2.1e+05	0.071	0.946	-4.81e+05	5.1e+05
Population	0.0534	0.031	1.699	0.133	-0.021	0.128

Omnibus:	22.707	Durbin-Watson:	2.656
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12.579
Skew:	2.276	Prob(JB):	0.00186
Kurtosis:	6.581	Cond. No.	1.32e+07

Figure 4.52 Drug-related crimes linear regression prediction

Linear regression was used to predict drug-related crimes using the population as the input, given the outcome of weak linearity. This showed that population is not the primary factor that causes the rise in drug-related crimes in the country. Other factors could include the rate of unemployment, the rate of illegal immigrants, and the rate of hooliganism. However, **Error! Reference source not found.** 4.5 shows how drug-related crimes can be predicted by the population in the GUI experimental prototype.

Table 4.5 Drug-related crimes actual and predicted values per province

	PROVINCE	ACTUAL	PREDICTED
0	Gauteng	579 307	670 491
1	Kwazulu-Natal	543 509	563 368
2	Mpumalanga	89 745	230 650
3	Western Cape	1 101 009	325 902
4	Limpopo	117 764	303 576
5	Eastern Cape	185 962	365 403
6	Northwest	133 818	202 334
7	Free State	101 261	161 495
8	Northern Cape	46 867	76 023

Table 4.5 showed the actual and predicted values of drug-related crimes.

4.6 Chapter Summary

In this chapter, the researcher analysed the annual rates of three serious crimes selected for this study. As described in this chapter, the researcher used heatmaps to analyse the linear correlation between the dependent and independent variables, as well as linear regression to predict future crime trends.

Chapter 5 Conclusion and Recommendations

5.1 Introduction

This chapter presents the summary, conclusions, and recommendations of the study. The dataset was pre-processed in this study using Python techniques, and the annual rates of three serious crimes were analysed to detect the trends of these crimes. A linear regression to predict future trends was performed. The purpose of using these approaches and procedures was to educate the masses and provide concrete, science-based information for consideration and use by the relevant authorities responsible for seeking serious crime reduction strategies in South Africa.

5.2 Summary of Conclusions

This section presents the summary of the conclusions and how the objectives were achieved.

Chapter 1

The research problem and research objectives were described in this chapter. A brief discussion of the research design was followed by an outline of the dissertation chapters.

Chapter 2

The section included the literature reviews on prevalent crimes worldwide, as well as a discussion on the three selected serious crimes. Data analytics tools used by earlier researchers were examined and a comparison of similar studies and their outcomes was discussed.

Chapter 3

This chapter discussed the research methodology. The CRISP-DM was thoroughly explored in the study through its six life-cycle phases. The nine research-focused provinces of South Africa were discussed. This chapter also expounded on the data

collection method used in this study. The importance of data integrity, validity, and reliability were discussed.

Chapter 4

The researcher analysed the annual rates of the three selected serious crimes. The dataset was visualised using Python software and flourish studio tools. This was conducted to detect trends in serious crimes from 2005 to 2020. These findings can provide law enforcement agencies and the South African government with important information. The graphical user interface experimental prototype was built using the linear regression method for prediction output results, which allows users to input independent variables to get the predicted outcome of the crimes per province.

Chapter 5

The findings of the literature review and the outcomes of the experimental analysis were used as a basis in this last chapter to develop answers to the serious crime problems that South Africans experience. The next sections expand on these points.

5.3 Objectives and Findings

The research aim was to use data analytics techniques to analyse and predict serious crime trends in South Africa, which may assist security agencies and the government in taking swift action to curb serious crimes. The research outlined six objectives to achieve this aim.

Objective 1

- To identify available data analytics algorithms for crime prediction.

The decision tree, naive Bayes, gradient boosting, linear regression, support vector machine (SVM), KNN, and other data analytics algorithms that have been employed by earlier researchers were identified in the literature. It was discovered that these algorithms do not perform well in all prediction categories.

Objective 2

- To use suitable data analytics algorithms to analyse and predict serious crime trends in South Africa.

With the use of data analytics tools, serious crime trends were analysed, and predicted using a linear regression algorithm, which can help mitigate and prevent serious crime occurrences in all provinces of South Africa.

Objective 3

- To evaluate the performance and accuracy of the data analytics algorithm.

The linear regression model's accuracy was evaluated using R-squared. Sexual crimes had 90%, murder had 75% and drug-related crimes had 29% R-squared accuracies. A model prototype was also built to evaluate the uniqueness and effectiveness of the model.

5.4 Implications of Study

The results showed that serious crimes in South Africa are increasing. This study aimed to use data analytics techniques to analyse and predict serious crime trends in South Africa and provide appropriate solutions or recommendations to security agencies. This was achieved by analysing the dataset, detecting trends, and predicting potential occurrences. Linear regression provided insight into the path and prediction of serious crime occurrences in the future. The output results of the visualisation and linear regression showed that serious crimes in South Africa are increasing. Gauteng, Kwazulu-Natal, the Eastern Cape, and the Western Cape are the major provinces prone to these serious crimes (sexual crimes, murder, and drug-related crimes).

5.5 Limitations

A limitation stated in Chapter 1 is the restriction of analysis to serious crime. Crime becomes habitual and is perceived to escalation from petty to serious. Analysing other crime data may present other patterns for analysis.

Another limitation of the study is that the analysis of serious crime trend can only serve as a help/tool to detectives rather than a complete substitute for their role. Data mining is susceptible to the quality of the input data, encompassing inaccuracies, information gaps, and potential data entry errors, among other factors. Mapping actual data to data mining characteristics is not always accessible, and it is often necessary to use expert data miners and criminal data analysts with extensive experience in the field. They must collaborate closely with the detectives during the earliest stages. While also bearing in mind that the results obtained and reported here are from secondary data sources that the researcher could lay hands on for analysis as at the time of the study.

5.6 Opportunities

Results of this study have shown the need to address serious crimes in South Africa, especially in the highly prone provinces such as Gauteng, Kwazulu-Natal, the Eastern Cape, and the Western Cape. Poverty, social-economic inequality, and unemployment sometimes lead people to commit crimes (Kempen, 2019). Most crimes are reported in urban areas because most people are moving from rural areas in search of job opportunities in the cities. The government could help in serious crime reduction by creating sustainable job opportunities in rural and urban areas.

5.7 Contributions

The results of this study shows that different provinces experience different types of crimes. For instance, the Western Cape is shown to be prone to drug-related crimes; Gauteng is an area prone to sexual crimes; and KwaZulu-Natal is an area prone to murder. Thus, more police stations should be built with the deployment of more police personnel in the serious crime-prone Provinces. The authorities should also employ technology to analyse data, detect trends, and predict the crime hotspots in each Province. Mounting surveillance cameras at suitable points is also recommended as this would enable the easy identification of suspects and capture of criminals. People should also work hand in hand with the authorities to reduce the spread of serious crime in the society as the police and other detective agents are required to actively protect the masses in the study area. This study recommends the law enforcement agencies, police, and government of South Africa to employ data analytics tools that

are capable of predicting serious crimes in South Africa to prevent or reduce these crimes in the Country.

5.8 Recommendations for Future work

Other machine learning techniques including deep learning would be considered in the future extension of this work to solve crime cases in South Africa. Thus the Bayesian classifier, decision tree classifier and Logistic regression is mentioned but delimited for this study.

Visualisations tools, such as Flourish and Tableau that do not require coding would be considered as well in future work.

5.9 Chapter Summary

This chapter outlines how each objective was met. The study confirmed that serious crimes are getting worse in South Africa, which shows that something needs to be done as the rate of serious crimes affects day-to-day living. Few studies have employed data analytics to solve serious crimes. Hence, this chapter recommends the use of technological tools to fight, mitigate, and prevent serious crimes. Sharing ideas and conducting such studies could minimise serious crime. This study is essential since it could assist to manage the rising trend of serious crimes and benefit the South African community immensely.

References

Abbasabadei, S., Karimi, A., Torkestani, J. A. and Zarafshan, F. 2020. Process Modeling and Extraction of Patterns of Computer Crimes Using Data Mining. *Computer Science Journal of Moldova*, 82 (1): 45-58.

Abdulrahman, N. and Abedalkhader, W. 2017. KNN classifier and naivebayse Classifier for crime prediction in san francisco context. *International Journal of Database Management Systems (IJDMS)*, 9 (4): 1-9.

Abrahams, N., Devries, K., Watts, C., Pallitto, C., Petzold, M., Shamu, S. and GarcÍa-Moreno, C. 2014. Worldwide prevalence of non-partner sexual violence: a systematic review. *The Lancet*, 383 (9929): 1648-1654.

Africa Check. 2019. *Fact sheet: South Africa's crime statistics for 2019/20*. Available: https://africacheck.org/factsheets/factsheet-south-africas-crime-statistics-for-2019-20/ (Accessed 9th October, 2021).

Africa Times Editor. 2019. Report: African homicide rate far higher than global average. *Africa Times.* Available: https://africatimes.com/2019/07/08/report-african-homicide-rate-far-higher-than-global-average/ (Accessed 30 October, 2022).

Ahishakiye, E., Taremwa, D., Omulo, E.O. and Niyonzima, I., 2017. Crime prediction using decision tree (J48) classification algorithm. *International Journal of Computer and Information Technology*, *6*(3), pp.188-195.

Akpinar, N. J. and Chouldechova, A. 2021. The effect of differential victim crime reporting on predictive policing systems. *arXiv preprint arXiv:2102.00128*, Article ID.

Al-Hashedi, K. G. and Magalingam, P. 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40: 100402.

Alasadi, S. A. and Bhaya, W. S. 2017. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12 (16): 4102-4107.

AlJanabi, K.B.S. and Haydar, K., 2010. Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures. ICIT.

Almond, L., McManus, M. A., Giles, S. and Houston, E. 2017. Female Sex Offenders: An Analysis of Crime Scene Behaviors. *J Interpers Violence*, 32 (24): 3839-3860.

Altinyelken, H. K. and Le Mat, M. 2018. Sexual violence, schooling and silence: teacher narratives from a secondary school in Ethiopia. *Compare: a journal of comparative and international education*, 48 (4): 648-664.

Andresen, M. A. and Felson, M. 2010. Situational crime prevention and co-offending. *Crime patterns and analysis*, 3 (1): 3-13.

Apuke, O. D. 2017. Quantitative research methods: A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, 33 (5471): 1-8.

Arifin, R. 2020. Crimes and Society, How Do the Law Respond to Disruptive Conditions? *Law Research Review Quarterly*, 6 (1): i-iv.

Awal, M. A., Rabbi, J., Hossain, S. I. and Hashem, M. 2016. Using linear regression to forecast future trends in crime of Bangladesh. In: Proceedings of *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, 333-338.

Azevedo, A. I. R. L. and Santos, M. F. 2008. KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*, Article ID.

Aziz, R.M., Hussain, A., Sharma, P. and Kumar, P., 2022. Machine learning-based soft computing regression analysis approach for crime data prediction. Karbala International Journal of Modern Science, 8(1), pp.1-19.

Bello-Orgaz, G., Jung, J. J. and Camacho, D. 2016. Social big data: Recent achievements and new challenges. *Information Fusion*, 28: 45-59.

Berwind, K., Bornschlegl, M., Hemmje, M. and Kaufmann, M. 2016. Towards a cross industry standard process to support Big Data applications in virtual research environments. *CERC2016*, Article ID: 82.

Bharti, S. and Mishra, A. 2015. Prediction of Future Possible Offender's Network and Role of Offenders. In: Proceedings of *2015 Fifth International Conference on Advances in Computing and Communications (ICACC)*. IEEE, 159-162.

Bhorat, H., Thornton, A. and Van der Zee, K. 2017. Socio-economic determinants of crime in South Africa: an empirical assessment. Article ID.

Bhowmik, R. 2008. Data mining techniques in fraud detection. *Journal of Digital Forensics, Security and Law*, 3 (2): 3.

Bickman, L., Rog, D. J. and Hedrick, T. E. 2009. Applied research design: A practical approach. *Handbook of applied social research methods*, 2: 3-43.

Brantingham, P. and Brantingham, P. 2013. Crime pattern theory. In: *Environmental criminology and crime analysis*. Willan, 100-116.

Brantingham, P. L. and Brantingham, P. J. 2017. Environment, routine, and situation: Toward a pattern theory of crime. In: *Routine activity and rational choice*. Routledge, 259-294.

Bretag, T. 2013. Challenges in addressing plagiarism in education. *PLoS medicine*, 10 (12): e1001574.

Britannica Academic. 2022. *Crime*. Available: https://academic-eb-com.eu1.proxy.openathens.net/levels/collegiate/article/crime/111023 (Accessed 14 October 2022).

Murder. 2022. *Encyclopedia Britannica*. Available: https://www.britannica.com/topic/murder-crime (Accessed 22 September 2022).

Britton, H. 2006. Organising against gender violence in South Africa. *Journal of Southern African Studies*, 32 (1): 145-163.

Brown, D. E. 1998. The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals. In: Proceedings of *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*. IEEE, 2848-2853.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A. and Grobler, J. 2013. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, Article ID.

Caldero, M., Dailey, J. and Withrow, B. 2018. *Police ethics: The corruption of noble cause*. Routledge.

Cao, L. 2017. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50 (3): 1-42.

Chaleunvong, K. 2009. Data collection techniques. *Training Course in Reproductive Health Research Vientine*, Article ID.

Chang, X., Jiang, X., Mkandarwire, T. and Shen, M. 2019. Associations between adverse childhood experiences and health outcomes in adults aged 18–59 years. *PloS one*, 14 (2): e0211850.

Chassiakos, Y. L. R., Radesky, J., Christakis, D., Moreno, M. A. and Cross, C. 2016. Children and adolescents and digital media. *Pediatrics*, 138 (5): e20162593.

Chauhan, C. and Sehgal, S. 2017. A review: crime analysis using data mining techniques and algorithms. In: Proceedings of *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 21-25.

Cherry, K. 2020. *Introduction to Psychology Research Methods*. Available: https://www.verywellmind.com/introduction-to-research-methods-2795793 (Accessed 31st August, 2021).

Cheteni, P., Mah, G. and Yohane, Y. K. 2018. Drug-related crime and poverty in South Africa. *Cogent Economics & Finance*, 6 (1): 1534528.

Chi, H., Lin, Z., Jin, H., Xu, B. and Qi, M. 2017. A decision support system for detecting serial crimes. *Knowledge-Based Systems*, 123: 88-101.

Chih-Pei, H. and Chang, Y. Y. 2017. John W. Creswell, research design: Qualitative, quantitative, and mixed methods approaches. *Journal of Social and Administrative Sciences*, 4 (2): 205-207.

Choi, T. M., Wallace, S. W. and Wang, Y. 2018. Big Data Analytics in Operations Management. *Production and Operations Management*, 27 (10): 1868-1883.

Clark, J. N. 2014. A crime of identity: Rape and its neglected victims. *Journal of Human Rights*, 13 (2): 146-169.

Coronado, R. and Saucedo, E. 2019. Drug-related violence in Mexico and its effects on employment. *Empirical Economics*, 57 (2): 653-681.

Craddock, A., Collins, J. J. and Timrots, A. D. 1994. *Fact sheet: Drug-related crime*. US Department of Justice, Office of Justice Programs, Bureau of Justice ….

Creswell, J. W. and Creswell, J. D. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Crush, J. and Peberdy, S. 2018. Criminal tendencies: immigrants and illegality in South Africa. Article ID.

D'Angelo, J. D. and Moreno, M. A. 2019. Not at the dinner table–take it to your room: adolescent reports of parental screen time rules. *Communication Research Reports*, 36 (5): 426-436.

Das, D. and Nayak, M. 2021. Crime Pattern Detection Using Data Mining. *Intelligent Data Analytics for Terror Threat Prediction: Architectures, Methodologies, Techniques and Applications*, Article ID: 221-236.

David, H. and Suruliandi, A. 2017. Survey On Crime Analysis and Prediction Using Data Mining Techniques. *ICTACT Journal on soft computing*, 7 (3).

De Bont, R., Groshkova, T., Cunningham, A. and Liem, M. 2018. Drug-related homicide in Europe—First review of data and sources. *International Journal of Drug Policy*, 56: 137-143.

Debuse, J., De la Iglesia, B., Howard, C. and Rayward-Smith, V. 2001. Building the kdd roadmap. In: *Industrial Knowledge Management*. Springer, 179-196.

Diehl, A., Molina de Souza, R., Madruga, C. S., Laranjeira, R., Wagstaff, C. and Pillon, S. C. 2020. Rape, child sexual abuse, and mental health in a Brazilian national sample. *Journal of interpersonal violence*, Article ID: 0886260520915546.

Du Toit, L. 2004. A phenomenology of rape: Forging a new vocabulary for action. Article ID.

Eck, J. E. and Rossmo, D. K. 2019. The new detective. *Criminology & Public Policy*, 18 (3): 601-622.

Edoka, N. O. 2020. Crime Incidents Classification Using Supervised Machine Learning Techniques: Chicago. Article IDDublin, National College of Ireland.

Edwards, P. K., O'Mahoney, J. and Vincent, S. 2014. *Studying organizations using critical realism: A practical guide*. OUP Oxford.

Elgendy, N. and Elragal, A. 2014. Big data analytics: a literature review paper. In: Proceedings of *Industrial conference on data mining*. Springer, 214-227.

Enaifoghe, A., Dlelana, M., Durokifa, A. A. and Dlamini, N. P. 2021. The Prevalence of Gender-Based Violence against Women in South Africa: A Call for Action. *African Journal of Gender, Society & Development*, 10 (1): 117.

European Monitoring Centre for Drugs and Drug Addiction. 2007. *Drugs in focus; Briefing of the European Monitoring Centre for Drugs and Drug Addiction*. Available:

https://www.emcdda.europa.eu/attachements.cfm/att_44774_EN_Dif16EN.pdf (Accessed 15th October, 2022).

Falope, O.S. and Thakur, C., 2022 (a), February. Data Analytics for South Africa's Sexual Crime Landscape. In *NEMISA Summit and Colloquium 2022*.

Falope, O. and Thakur, S., 2022 (b), December. Sexual Crime Prediction in an African Context. In *International Conference on Intelligent and Innovative Computing Applications*, 1-12.

Farsi, M., Daneshkhah, A., Far, A. H., Chatrabgoun, O. and Montasari, R. 2018. Crime data mining, threat analysis and prediction. In: *Cyber Criminology*. Springer, 183-202.

Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y. and Liu, Q. 2019. Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. *IEEE Access*, 7: 106111-106123.

Gahalot, A., Dhiman, S. and Chouhan, L. 2020. Crime Prediction and Analysis. In: Proceedings of *2nd International Conference on Data, Engineering and Applications (IDEA)*. IEEE, 1-6.

Ghani, Z. A. 2017. A comparative study of urban crime between Malaysia and Nigeria. *Journal of Urban Management*, 6 (1): 19-29.

Glen, S. 2021. *"Mean Squared Error: Definition and Example" From StatisticsHowTo.com: Elementary Statistics for the rest of us!* Available: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/ (Accessed 7 July, 2021).

Gonzalez, J. J. and Leboulluec, A. 2019. Crime Prediction and Socio-Demographic Factors: A Comparative Study of Machine Learning Regression-Based Algorithms. *Journal of Applied Computer Science & Mathematics*, 13 (27).

Gray, K. 2019. Evidence of rape culture in modern music. *CLA Journal*, 7: 35-51.

Greener, S. 2008. *Business research methods*. BookBoon.

Hamdy, E., Adl, A., Hassanien, A. E., Hegazy, O. and Kim, T. H. 2015. Criminal act detection and identification model. In: Proceedings of *2015 Seventh International Conference on Advanced Communication and Networking (ACN)*. IEEE, 79-83.

Hassani, H., Huang, X., Silva, E. S. and Ghodsi, M. 2016. A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9 (3): 139-154.

Haylock, S., Boshari, T., Alexander, E. C., Kumar, A., Manikam, L. and Pinder, R. 2020. Risk factors associated with knife-crime in United Kingdom among young people aged 10-24 years: a systematic review. *BMC Public Health*, 20 (1): 1451.

Health-e News. 2021. Sexual offences increased in last quarter of 2020 – crime statistics. *Health-e News* Available: https://health-e.org.za/2021/02/19/gbv-sexual-offences-rise-5-in-south-africa/ (Accessed 17th July, 2021).

Helgesson, G. and Eriksson, S. 2015. Plagiarism in research. *Medicine, Health Care and Philosophy*, 18 (1): 91-101.

Holland, K. J. and Cortina, L. M. 2017. “It Happens to Girls All the Time”: Examining Sexual Assault Survivors’ Reasons for Not Using Campus Supports. *American Journal of Community Psychology*, 59 (1-2): 50-64.

Holzman, C. G. 1996. Counseling adult women rape survivors: Issues of race, ethnicity, and class. *Women & Therapy*, 19 (2): 47-62.

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M. and Sarker, I. H. 2020. Crime Prediction Using Spatio-Temporal Data. In: Proceedings of *International Conference on Computing Science, Communication and Security*. Springer, 277-289.

Ingilevich, V. and Ivanov, S. 2018. Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*, 136: 472-478.

Insights Desk. 2020a. *10 Best Data Analytics Tools*. Available: https://www.demandtalk.com/insights/data/analytics/10-best-data-analytics-tools/ (Accessed 9th August, 2021).

Insights Desk. 2020b. *What is Data Analytics? Benefits, Challenges, and Use Cases*. Available: https://www.demandtalk.com/insights/data/analytics/what-is-data-analytics-benefits-challenges-and-use-cases/ (Accessed 9th August, 2021).

Ismail, A. 2021. Inside SA's rape stats: Gauteng records most reported rape cases. *news24* Available: https://www.news24.com/news24/southafrica/news/inside-sas-rape-stats-gauteng-records-most-reported-rape-cases-20210515 (Accessed 9th August, 2021).

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning*. Springer.

Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I. and Wijayasiri, A. 2015. Crime analytics: Analysis of crimes through newspaper articles. In: Proceedings of *2015 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 277-282.

Jewkes, R. and Abrahams, N. 2002. The epidemiology of rape and sexual coercion in South Africa: an overview. *Social science & medicine*, 55 (7): 1231-1244.

Jewkes, R., Levin, J. and Penn-Kekana, L. 2002. Risk factors for domestic violence: findings from a South African cross-sectional study. *Social science & medicine*, 55 (9): 1603-1617.

Jewkes, R. and Morrell, R. 2010. Gender and sexuality: emerging perspectives from the heterosexual epidemic in South Africa and implications for HIV risk and prevention. *Journal of the International AIDS society*, 13 (1): 1-11.

Jin, Q. 2013. Modeling student success in engineering education. Article IDPurdue University.

Joshi, A., Sabitha, A. S. and Choudhury, T. 2017. Crime analysis using K-means clustering. In: Proceedings of *2017 3rd International Conference on Computational Intelligence and Networks (CINE)*. IEEE, 33-39.

Juneja, P. 2015. *Management Study Guide: Secondary Data*. Available: https://www.managementstudyguide.com/secondary_data.htm (Accessed 5 October, 2021).

Justice, C. 2011. World crime trends and emerging issues and responses in the field of crime prevention and criminal justice. Article ID.

Kabir, S. M. S. 2016. Basic Guidelines for Research. *An Introductory Approach for All Disciplines*, Article ID: 168-180.

Kalla, S. 2011. *Relationship Between Variables*. Available: https://explorable.com/relationship-between-variables (Accessed August 31, 2021).

Kalsi, M. J. A. S. 2019. Naive Bayes Approach for the Crime Prediction in Data Mining. *International Journal of Computer Applications*, 178: 33-37.

Kankam, P. K. 2019. The use of paradigms in information research. *Library & Information Science Research*, 41 (2): 85-92.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V. 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29 (10): 2318-2331.

Kaufmann, M., Egbert, S. and Leese, M. 2019. Predictive policing and the politics of patterns. *The British Journal of Criminology*, 59 (3): 674-692.

Kelly, L. 2013. *Surviving sexual violence*. John Wiley & Sons.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M. and D'Mello, S. K. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47 (7): 451-464.

Kempen, A. 2019. Crime statistics 2018/2019. *Servamus Community-based Safety and Security Magazine*, 112 (11): 10-13.

Khaldi, K. 2017. Quantitative, Qualitative or Mixed Research: Which Research Paradigm to Use? *Journal of Educational and Social Research*, 7 (2): 15.

Khuluvhe, M. 2021. Adult illiteracy in South Africa. *Pretoria: South African Department of Higher Education and Training*, Article ID.

Kiani, R., Mahdavi, S. and Keshavarzi, A. 2015. Analysis and prediction of crimes by clustering and classification. Article ID.

Kim, S., Joshi, P., Kalsi, P. S. and Taheri, P. 2018. Crime analysis through machine learning. In: Proceedings of *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 415-420.

Kiran, J. and Kaishveen, K. 2018. Prediction analysis of crime in india using a hybrid clustering approach. In: Proceedings of *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on*. IEEE, 520-523.

Knox, C. and Monaghan, R. 2005. Violence in a changing political context: Northern Ireland: Northern Ireland and South Africa. In: *The meanings of violence*. Routledge, 202-220.

Kollamparambil, U. 2019. Immigration, internal migration and crime in South Africa: A multi-level model analysis. *Development Policy Review*, Article ID.

Kumar, R. and Nagpal, B. 2019. Analysis and prediction of crime patterns using big data. *International Journal of Information Technology*, 11 (4): 799-805.Lalwani, P., Mishra, M. K., Chadha, J. S. and Sethi, P. 2022. Customer churn prediction system: a machine learning approach. *Computing*, 104 (2): 271-294.

Lamberton, C., Brigo, D. and Hoy, D. 2017. Impact of Robotics, RPA and AI on the insurance industry: challenges and opportunities. *Journal of Financial Perspectives*, 4 (1).

Landström, S., Strömwall, L. A. and Alfredsson, H. 2016. Blame attributions in sexual crimes: Effects of belief in a just world and victim behavior. *Nordic Psychology*, 68 (1): 2-11.

Lindegaard, M. R. 2017. Homicide in South Africa. *The handbook of homicide*, Article ID: 499-514.

Mabasa, T. B. 2009. Understanding and preventing rape: perceptions of police officers in inner city Johannesburg. Article ID.

Machethe, P. and Mofokeng, J. T. 2022. The impact of illicit drug networks on the effectiveness of law enforcement in South Africa. *Technium Soc. Sci. J.*, 27: 338.

Mahlakoana, T. 2020. OVER 2 MILLION DRUG-RELATED CRIMES IN SA OVER PAST 10 YEARS. *Eyewitness News* Available: https://ewn.co.za/2020/08/01/over-2-million-drug-related-crimes-in-sa-over-past-10-years (Accessed 16 September 2022).

Majeed, I. and Naaz, S. 2018. Current State of Art of Academic Data Mining and Future Vision. *Indian Journal of Computer Science and Engineering (IJCSE)*, 9: 49-56.

Maluleke, W. and Dlamini, S. 2019. The prevalence of organised cross-border crimes in South Africa: A non-empirical statistical data analysis on stock theft and hijacking of motor vehicles. *International Journal of Social Sciences and Humanity Studies*, 11 (1): 116-145.

Manhica, H., Straatmann, V. S., Lundin, A., Agardh, E. and Danielsson, A. K. 2021. Association between poverty exposure during childhood and adolescence, and drug use disorders and drug-related crimes later in life. *Addiction*, 116 (7): 1747-1756.

Marco, M., Gracia, E. and López-Quílez, A. 2017. Linking Neighborhood Characteristics and Drug-Related Police Interventions: A Bayesian Spatial Analysis. *ISPRS International Journal of Geo-Information*, 6 (3): 65.

Mathews, B. and Collin-Vézina, D. 2019. Child sexual abuse: Toward a conceptual model and definition. *Trauma, Violence, & Abuse*, 20 (2): 131-148.

Mathews, S., Abrahams, N., Martin, L. J., Lombard, C. and Jewkes, R. 2019. Homicide pattern among adolescents: A national epidemiological study of child homicide in South Africa. *PLOS ONE*, 14 (8): e0221415.

Matzopoulos, R., Simonetti, J., Prinsloo, M., Neethling, I., Groenewald, P., Dempers, J., Martin, L., Rowhani-Rahbar, A., Myers, J. and Thompson, M. 2018. A retrospective time trend study of firearm and nonfirearm homicide in Cape Town from 1994 to 2013. *South African Medical Journal*, 108 (3): 197-204.

Mazorodze, B. T. 2020. Youth unemployment and murder crimes in KwaZulu-Natal, South Africa. *Cogent economics & finance*, 8 (1): 1799480.

McClendon, L. and Meghanathan, N. 2015. Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, 2 (1): 1-12.

McKinney, W. 2010. Data Structures for Statistical Computing in Python. In: Proceedings of *Proceedings of the 9th Python in Science Conference*. 56 - 61.

Meinck, F., Cluver, L., Loening-Voysey, H., Bray, R., Doubt, J., Casale, M. and Sherr, L. 2017. Disclosure of physical, emotional and sexual child abuse, help-seeking and access to abuse response services in two South African Provinces. *Psychology, Health & Medicine*, 22 (sup1): 94-106.

Middleton, F. 2019. *Reliability vs validity: what's the difference?* Available: https://www.scribbr.com/methodology/reliability-vs-validity/ (Accessed 6th September 2021).

Mike, C. and Paul, V. 2019. *Horror of gender-based violence revealed in South African report*. Available: https://www.biznews.com/undictated/2019/09/12/murder-rape-sexual-assault-crime-stats (Accessed 9 October, 2020).

Mittal, M., Goyal, L. M., Sethi, J. K. and Hemanth, D. J. 2019. Monitoring the impact of economic crisis on crime in India using machine learning. *Computational Economics*, 53 (4): 1467-1485.

Moen, M. C. 2020. Characteristics for the Identification of Children Who Commit Family Murder in South Africa. *Journal of Interpersonal Violence*, 35 (21-22): 4796-4813.

Moffett, H. 2006. 'These women, they force us to rape them': Rape as narrative of social control in post-apartheid South Africa. *Journal of Southern African Studies*, 32 (1): 129-144.

Moletsane, R. 2018. " Stop the War on Women's Bodies": Facilitating a Girl-Led March Against Sexual Violence in a Rural Community in South Africa. *Studies in Social Justice*, 12 (2).

Monyeki, P. 2021. Data mining to analyse recurrent crime in South Africa. Article ID.

Monyeki, P., Naicker, N. and Obagbuwa, I. C. 2020. Change-Point Analysis: An Effective Technique for Detecting Abrupt Change in the Homicide Trends in a Democratic South Africa. *The Scientific World Journal*, 2020: 1-10.

Moreira, J., Carvalho, A. and Horvath, T. 2018. *A general introduction to data analytics*. John Wiley & Sons.

Mythen, G. 2017. *Understanding the risk society: Crime, security and justice*. Bloomsbury Publishing.

Nagpal, A. and Gabrani, G. 2019. Python for Data Analytics, Scientific and Technical Applications. In: Proceedings of. 2019. IEEE,

Naidoo, K. 2013. Rape in South Africa-a call to action. *SAMJ: South African Medical Journal*, 103 (4): 210-211.

Netshakhuma, N.S., 2020. Assessment of a South Africa national consultative workshop on the Protection of Personal Information Act (POPIA). Global Knowledge, Memory and Communication, 69(1/2), pp.58-74.

Nganga, M. 2021. Fighting the scourge of sexual assault and violence. *iol News* Available: https://www.iol.co.za/weekend-argus/news/fighting-the-scourge-of-sexual-assault-and-violence-579bc558-fc3f-4d81-9999-1f250a63a625 (Accessed 11 March, 2022).

Norouzian, R. and Plonsky, L. 2018. Correlation and simple linear regression in applied linguistics. In: *The Palgrave handbook of applied linguistics research methodology*. Springer, 395-421.

Nyabadza, F. and Coetzee, L. 2017. A Systems Dynamic Model for Drug Abuse and Drug-Related Crime in the Western Cape Province of South Africa. *Computational and Mathematical Methods in Medicine*, 2017: 1-13.

Obagbuwa, I. C. and Abidoye, A. P. 2021. South Africa Crime Visualization, Trends Analysis, and Prediction Using Machine Learning Linear Regression Technique. *Applied Computational Intelligence and Soft Computing*, 2021.

Öberg, M., Heimer, G. and Lucas, S. 2020. Lifetime experiences of violence against women and men in Sweden. *Scandinavian journal of public health*, Article ID: 1403494820945072.

Ojedokun, U. A., Tade, O. and Aderinto, A. A. 2021. Trends and patterns of homicides arising from interpersonal violence in Nigeria (2006-2016). *Journal of interpersonal violence*, 36 (17-18): 8456-8470.

Osman, A. S. 2019. Data mining techniques. *International Journal of Data Science Research (IJDSR), Al-Madinah International University Malaysia*, 2 (1).

Osuafor, G. N. and Okoli, C. E. 2019. Alcohol consumption as a factor in gun or knife crimes in South Africa. *African Journal of Drug and Alcohol Studies*, 18 (2): 85-96.

Otieno, G., Marinda, E., Bärnighausen, T. and Tanser, F. 2015. High rates of homicide in a rural South African population (2000–2008): findings from a population-based cohort study. *Population Health Metrics*, 13 (1).

Oyasor, J. I. 2020. Mining tweets on sexual violence in South Africa. Article ID.

Pednekar, V., Mahale, T., Gadhve, P. and Gore, A. 2018. Crime rate prediction using KNN. *International Journal on Recent and Innovation Trends in Computing and Communication*, 6 (1): 124-127.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. and Perrot, M. a. D., E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz and Wirth, C. S. a. R. 2000. *CRISP-DM 1.0: Step-by-step data mining guide*. Available: http://www.statoo.com/CRISP-DM.pdf (Accessed 10 August, 2021).

Gastrow, P. (1998). Organized crime in southern Africa: An Assessment of Its Nature and Origins. Pp.1-75. https://www.ojp.gov/ncjrs/virtual-library/abstracts/organised-crime-south-africa-assessment-its-nature-and-originsPitcher, G. J. and Bowley, D. M. 2002. Infant rape in South Africa. *Lancet*, 359 (9303): 274-275.

Prabakaran, S. and Mitra, S. 2018. Survey of analysis of crime detection techniques using data mining and machine learning. In: Proceedings of *Journal of Physics: Conference Series*. IOP Publishing, 012046.

Pradhan, I. 2018. Exploratory data analysis and crime prediction in San Francisco. Article ID.

Pramanik, M. I., Lau, R. Y., Yue, W. T., Ye, Y. and Li, C. 2017. Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7 (4): e1208.

Rashid, Y., Rashid, A., Warraich, M. A., Sabir, S. S. and Waseem, A. 2019. Case Study Method: A Step-by-Step Guide for Business Researchers. *International Journal of Qualitative Methods*, 18: 160940691986242.

Rath, S., Tripathy, A. and Tripathy, A.R., 2020. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & metabolic syndrome: clinical research & reviews*, *14*(5), pp.1467-1474.

Ratul, M. and Rab, A. 2020. A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining. *arXiv preprint arXiv:2001.02802*, Article ID.

Ray, S. 2019. A quick review of machine learning algorithms. In: Proceedings of *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, 35-39.

Razali, M. N. 2020. A Classification Approach for Crime Prediction. In: Proceedings of *Applied Computing to Support Industry: Innovation and Technology: First International Conference, ACRIT 2019, Ramadi, Iraq, September 15–16, 2019, Revised Selected Papers*. Springer Nature, 68.

Rezigalla, A. A. 2020. Observational study designs: synopsis for selecting an appropriate study design. *Cureus*, 12 (1).

Rigoni, R., Breeksema, J. and Woods, S. 2018. Speed limits: Harm reduction for people who use stimulants. Article ID.

Ronald, V. C. and John, E. E. 2014. *Crime Analysis for Problem Solvers in 60 Small Steps*. Available: https://web.archive.org/web/20140809214029/http://www.popcenter.org/learning/60steps/index.cfm?stepNum=16 (Accessed 19 October, 2021).

Zafarghandi, S. M. B., Eshrati, S., Rashedi, V., Vameghi, M., Arezoomandan, R., Clausen, T. and Waal, H. 2022. Indicators of Drug-Related Community Impacts of Open Drug Scenes: A Scoping Review. *European Addiction Research*, 28 (2): 87-102.

Saeed, R.M. and Abdulmohsin, H.A., 2023. A study on predicting crime rates through machine learning and data mining using text. *Journal of Intelligent Systems*, *32*(1), p.20220223.

Saeed, U., Sarim, M., Usmani, A., Mukhtar, A., Shaikh, A. B. and Raffat, S. K. 2015. Application of machine learning algorithms in crime classification and classification rule mining. *Research Journal of Recent Sciences ISSN*, 2277: 2502.

Saltos, G. and Cocea, M. 2017. An exploration of crime prediction using data mining on open data. *International Journal of Information Technology & Decision Making*, 16 (05): 1155-1181.

Sampson, R. J. and Wilson, W. J. 2020. Toward a theory of race, crime, and urban inequality. In: *Crime, inequality and the state*. Routledge, 312-325.

Santos, M. R. and Testa, A. 2018. Global trends in homicide. In: *Homicide and Violent Crime*. Emerald Publishing Limited.

Santos, R. B. 2016. Crime analysis with crime mapping. 5th Edition. *Sage publications*, Inc. ISBN-13: 978-1506331034

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P. and Ng, A. 2020. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7 (1).

Saunders, M., Lewis, P. and Thornhill, A. 2009. *Research methods for business students*. Pearson education.

Schober, P., Boer, C. and Schwarte, L. A. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126 (5): 1763-1768.

Sekaran, U. and Bougie, R. 2019. *Research methods for business: A skill building approach*. john wiley & sons.

Shamsuddin, N. H. M., Ali, N. A. and Alwee, R. 2017. An overview on crime prediction methods. In: Proceedings of *2017 6th ICT International Student Project Conference (ICT-ISPC)*. IEEE, 1-5.

Sharma, A. and Kumar, R. 2013. The obligatory of an algorithm for matching and predicting crime-using data mining techniques. *International Journal of Scientific and Engineering Research*, 4 (2): 289-292.

Shi, S. C., Chen, P., Yuan, P. H., Hou, C. and Ming, H. X. 2018. The prediction of offender identity using decision-making tree algorithm. In: Proceedings of *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 405-409.

Shingleton, J. S. 2012. *Crime trend prediction using regression models for salinas, california*. NAVAL POSTGRADUATE SCHOOL MONTEREY CA.

Singh, S. B. 2020. Victimisation of African Foreign Nationals in Durban, South Africa. *International Journal of Criminology and Sociology*, 9: 686-694.

South African Police Service. 2013. An analysis of the national crime statistics: Addendum to the annual report 2012/13. Article ID.

South African Police Service. 2018. *Annual crime report 2017/2018: Addendum to the SAPS annual report*: SAPS.

South African Police Service. Police, D. O. 2022. *Crime statistics: integrity*. Available: https://www.saps.gov.za/services/crimestats.php (Accessed 02 September 2022).

Sri, L. A., Manvitha, K., Amulya, G., Sanjuna, I. S. and Pavani, V. 2020. FBI CRIME ANALYSIS AND PREDICTION USING MACHINE LEARNING. Article ID.

Statista Research Department. 2021. *Crime worldwide - Statistics & Facts*. Available: https://www.statista.com/topics/780/crime (Accessed 9th August, 2021).

Statistics South Africa. (2021). Quarterly labour force survey: Quarter 1, 2021. Retrieved from https://www.statssa.gov.za/publications/P0211/P02111stQuarter2021.pdf

Stojiljković, M. 2021. *Linear Regression in Python*. Available: https://realpython.com/linear-regression-in-python/ (Accessed 22nd June, 2021).

Sukhija, K., Singh, S. N. and Kumar, M. 2020. Using Linear Regression to investigate parameters associated with Rape crime in Haryana. In: Proceedings of *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 107-111.

Sullivan, D.J., 1994. Women's human rights and the 1993 World Conference on Human Rights. American Journal of International Law, 88(1), pp.152-167

Teker, K., Topçu, S., Başkan, S., Orhon, F. and Ulukol, B. 2017. The relationship between family functioning and the crime types in incarcerated children. *Minerva Pediatr*, 69 (3): 206-212.

Thamrin, H. and Liao, Y.-M. 2018. Drug-Related crimes and control in Indonesia and Taiwan: Cooperation regarding narcotics smuggling prevention and countermeasures from the point of view of international law. In: Proceedings of *International Conference on Knowledge Management in Organizations*. Springer, 312-323.

Thomas, S. P. 2014. The Deepest, Darkest Secret. *Issues in mental health nursing*, 35 (5): 321-322.

Thongsatapornwatana, U. 2016. A survey of data mining techniques for analyzing crime patterns. In: Proceedings of *2016 Second Asian Conference on Defence Technology (ACDT)*. IEEE, 123-128.

Triegaardt, J. D. (2021). Reflections on poverty and inequality in South Africa: policy considerations in an emerging democracy, 1-9. Retrieved from: https://www.dbsa.org/sites/default/files/media/documents/2021-

03/Poverty%20and%20inequality%20in%20South%20Africa%20Policy%20considerations%20in%20an%20emerging%20democracy.pdf

Tripathy, J. P. 2013. Secondary Data Analysis: Ethical Issues and Challenges. *Iranian journal of public health*, 42 (12): 1478-1479.

UNODC (United Nations Office on Drugs and Crime). (2020). Global study on homicide 2019. United Nations.

Vaidya, O., Mitra, S., Kumbhar, R., Chavan, S. and Patil, R. 2018. COMPREHENSIVE COMPARATIVE ANALYSIS OF METHODS FOR CRIME RATE PREDICTION. Article ID.

Vaillancourt-Morel, M. P., Godbout, N., Sabourin, S., Briere, J., Lussier, Y. and Runtz, M. 2016. Adult sexual outcomes of child sexual abuse vary according to relationship status. *Journal of Marital and Family Therapy*, 42 (2): 341-356.

Van Dijk, A. and Crofts, N. 2017. Law enforcement and public health as an emerging field. *Policing and society*, 27 (3): 261-275.

Vashisht, P. and Gupta, V. 2015. Big data analytics techniques: A survey. In: Proceedings of *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE, 264-269.

Vassakis, K., Petrakis, E. and Kopanakis, I. 2018. Big data analytics: applications, prospects and challenges. In: *Mobile big data*. Springer, 3-20.

Velopulos, C. G., Carmichael, H., Zakrison, T. L. and Crandall, M. 2019. Comparison of male and female victims of intimate partner homicide and bidirectionality-an analysis of the national violent death reporting system. *J Trauma Acute Care Surg*, 87 (2): 331-336.

Vural, M. S. and Gök, M. 2017. Criminal prediction using naive Bayes theory. *Neural Computing and Applications*, 28 (9): 2581-2592.

Waduge, N. and Ranathunga, D. L. 2017. Machine learning approaches for detect crime patterns. *Sri Lanka*, Article ID.

Wainana, S. M., Karomo, J. N., Kyalo, R. and Mutai, N. 2020. Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya. *International Journal of Data Science and Analysis*, 6 (1): 20.

Wang, T., Rudin, C., Wagner, D. and Sevieri, R. 2013. Learning to detect patterns of crime. In: Proceedings of *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 515-530.

Wang, W. and Siau, K. 2019. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management (JDM)*, 30 (1): 61-79.

Wang, Z. and Zhang, H. 2020. Construction, detection, and interpretation of crime patterns over space and time. *ISPRS International Journal of Geo-Information*, 9 (6): 339.

Wei, X., Guo, H., Wang, X., Wang, X. and Qiu, M. 2021. Reliable Data Collection Techniques in Underwater Wireless Sensor Networks: A Survey. *IEEE Communications Surveys & Tutorials*, 24 (1): 404-431.

Kaggle, 2021. *Crime statistics for South Africa*. Available: https://www.kaggle.com/slwessels/crime-statistics-for-south-africa. (Accessed March 4, 2021).

Wirth, R. and Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. In: Proceedings of *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK,

Wojcicki, J. M. 2002. " She drank his money": survival sex and the problem of violence in taverns in Gauteng Province, South Africa. *Medical anthropology quarterly*, 16 (3): 267-293.

World Health Organization. 2012. *Understanding and addressing violence against women: Sexual violence*. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/77434/WHO_RHR_12.37_eng.pdf.

Wortley, R. and Townsley, M. 2016. *Environmental criminology and crime analysis*. Taylor & Francis.

Xia, Z., Stewart, K. and Fan, J. 2021. Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major us metropolitan area. *Computers, environment and urban systems*, 87: 101599.

Yerpude, P. 2020. Predictive Modelling of Crime Data Set Using Data Mining. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, 7.

Zhang, H., Liu, G., Chow, T. W. and Liu, W. 2011. Textual and visual content-based anti-phishing: a Bayesian approach. *IEEE transactions on neural networks*, 22 (10): 1532-1546.