DAT7015_2_2312602

by Oluwatosin Ojo

Submission date: 15-Jan-2024 11:52AM (UTC+0000)
Submission ID: 222042511
File name: 150319_Oluwatosin_Ojo_DAT7015_2_2312602_2208030_2035717712.docx (10.04M)
Word count: 5490
Character count: 29831

University of Bolton
Teaching Intensive, Research Informed

M.Sc. Data Analytics & Technologies

2023-2024

Big Data Technology in Crime Analysis: A Comprehensive Review of Criminal Datasets, Addressing Challenges and Proposing Solutions.

DAT7015 BIG DATA TECHNOLOGIES

Assignment 2

Student Id: 2312602

Modul Tutor: Anchal Argar | Pradeep

1

Abstract

The introduction of machine learning has resulted in a meteoric launch in the developments occurring in various sectors. There is a sudden technological shift in the use of artificial intelligence in the development of applications and solutions being used. There is an increase in the volume of data being generated in different formats (unstructured or unstructured), and the speed of generating and processing the data which is known as Big Data. Big data is primarily composed of 3 characteristics which are volume, variety, and velocity.

Big data just like the name says refers to a large set of data that cannot be processed using conventional methods. As a result, non-conventional methods had to be introduced to handle big data. This report highlights the two datasets given named criminal dataset and offense code dataset. Both datasets are regarded as big data.

The report studies the influence of big data in crime analysis, identifies issues before the introduction of big data, and the solutions proposed to solve the issues while referring to the datasets provided.

Keywords:

- Dataset
- Big Data,
- Machine Learning
- Crime Analysis
- Technology.

Table of Contents

List of Figures

List of Tables

Executive Summary

This report critically examines provided datasets which include criminal analysis across various financial years. The primary objective of the documentation is to explain the relationship within the datasets and also point out the relationship between big data and crime analysis. It aims to draft research questions after carefully reviewing the given dataset, perform some calculations on the dataset, and provide solutions to research questions.

In addition to addressing the research questions, the report focuses on the challenges associated with the use of big data solutions and the solutions provided in crime analysis. Upon highlighting the challenges, it offers strategic solutions adopted to mitigate against the known challenges. The approach used in the report ensures a thorough examination of the subject field, infusing both theoretical insights with practical knowledge.

Research Questions

The following research questions below have been carefully drafted after an extensive review of the datasets provided. The aim is to address the positive and negative effects of big data analytics in crime analysis.

1. What are the most common types of offenses recorded in the database? What are the top five offenses in the dataset and what relevance does that cover in the criminal analysis?
2. What force made the most arrests across all the financial years and in what financial year did they make the most arrest?
3. Are there any inconsistencies or anomalies in the data that need to be addressed to improve the data integrity? What technologies were adopted to ensure data reliability and robustness to provide meaningful contributions?
4. What analytical approach proves most effective in predictive modelling for anticipating the frequency and nature of offenses, and how can this method be optimally applied to enhance proactive law enforcement strategies?

Research Methodology

The methodology being adopted is the systematic review approach and the qualitative research method. The systematic review approach entails extracting already published journals, books, and articles. After extracting the necessary documents, I then select the documents that support my focus topic "Crime Analysis" and review the selections to answer the research questions provided (Laura, 2021).

Upon completion of the review process, we adopt the Triangulation approach, specifically the qualitative research process which includes working with multiple datasets and theories or technologies to address the research questions (Bhandari, 2022).

Some of the technologies adopted to answer the research questions include:

- the use of python,
- Pandas,
- Spark,
- Spark SQL,
- Hive
- Correlations,
- Map Reduce
- Machine Learning for predictive analytics,
- Tableau for Visualizations.

The technologies adopted and findings are further explained in SECTION A: Data Analysis.

PRISMA Diagram

The PRISMA flow diagram is a representation of the studies carried out, the phases involved, the selection process, and the findings.

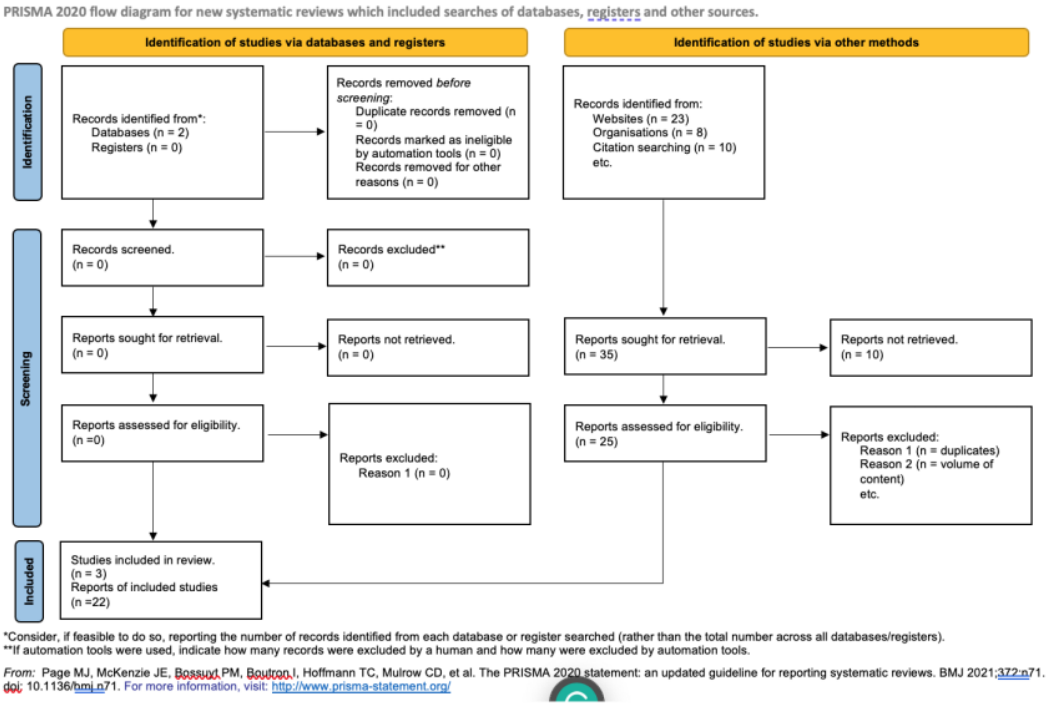


Figure 1: PRISMA Diagram

Source: (Page *et al.*, 2021)

Section A: Data Analysis

A1. My Setup

1.1 Installing Spark and its session.

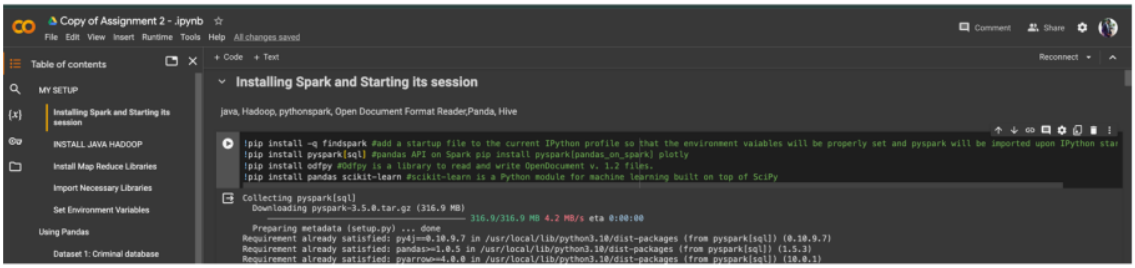


Figure 2: Installing Dependencies

1.2 Java and Hadoop

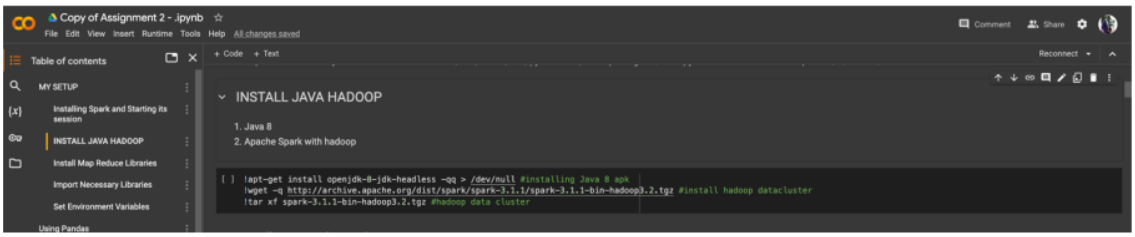


Figure 3: Install Java and Hadoop

1.3 Install Map Reduce

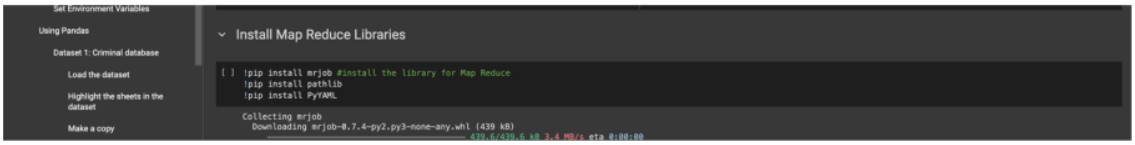


Figure 4: Install Map Reduce Libraries

1.4 Importing necessary libraries.

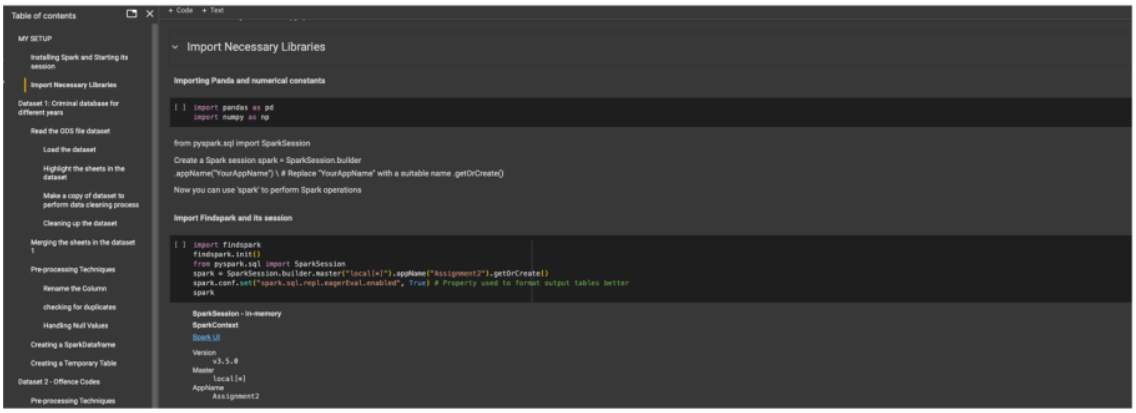


Figure 5: Import Pandas, Findspark

1.5 Set Environment Variables

Figure 6: Set Environment Variable and Enable Spark session.

A2. Using Pandas

2.1 Read the ODS file.

Given that the datasets given are both in ODS formats, I must first read the ODS file by using Pandas. Afterward, I will convert the panda file into a PySpark and create a database for the dataset.

2.2 Load Dataset 1:

Dataset 1 is the criminal workbook containing multiple worksheets. In this section, I endeavor to highlight the worksheets I have decided to work with, remove the worksheets I don't need, and merge all the worksheets highlighted into a single worksheet for easy access.

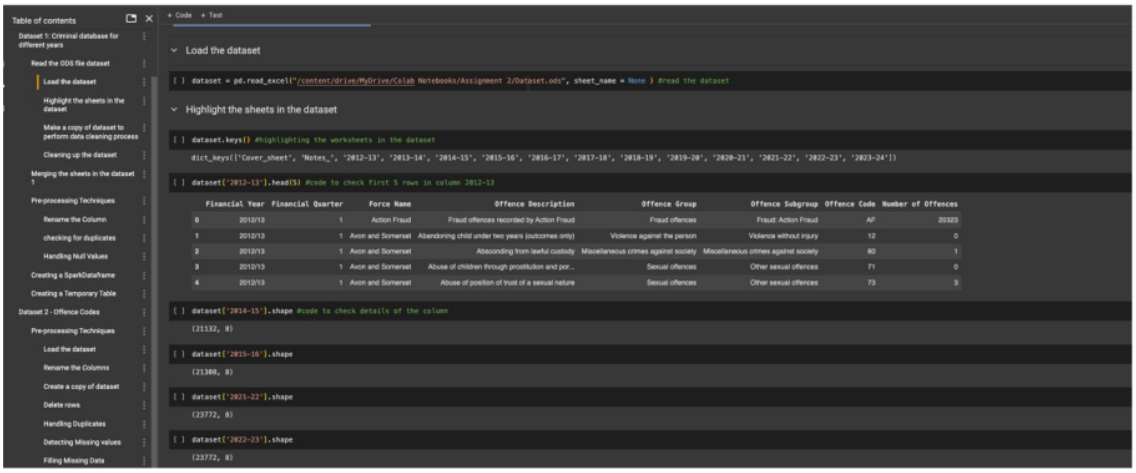


Figure 7: Load the dataset.

- To Read the dataset:
  Dataset = pd.read_excel("/content/drive/MyDrive/Colab Notebooks/Assignment 2/Dataset.ods", sheet_name = None)
- To know how many worksheets are in the workbook: dataset.keys()
- To check the first 5 rows of a particular worksheet: dataset ['2012-13'].head(5) #code to check first 5 rows in 2012-13 worksheet

- To know the details of a particular worksheet: dataset ['2014-15'].shape #code to check details of the column

2.3 Create a copy of dataset.

Make a copy of the dataset before performing operations so you have an original file to fall back to.

Figure 8: Create Dataset Copy

2.4 Remove worksheets from the workbook.

Figure 9: Removing worksheets.

Upon making a copy, I can now highlight worksheets I won't be using and remove them from the workbook.

In the first line of code, I attempt to delete the worksheet cover sheet. However, due to the volume of worksheets identified, I attempted to write a for-loop construct to delete the worksheets highlighted once.

Upon deletion, I therefore use .keys() to show the worksheets left in the workbook.

2.5 Merging the worksheets.

In this section, after removing unwanted worksheets, I proceed to merge the worksheets left into one worksheet for easy access so I can easily clean up and perform our pre-processing activities conveniently.

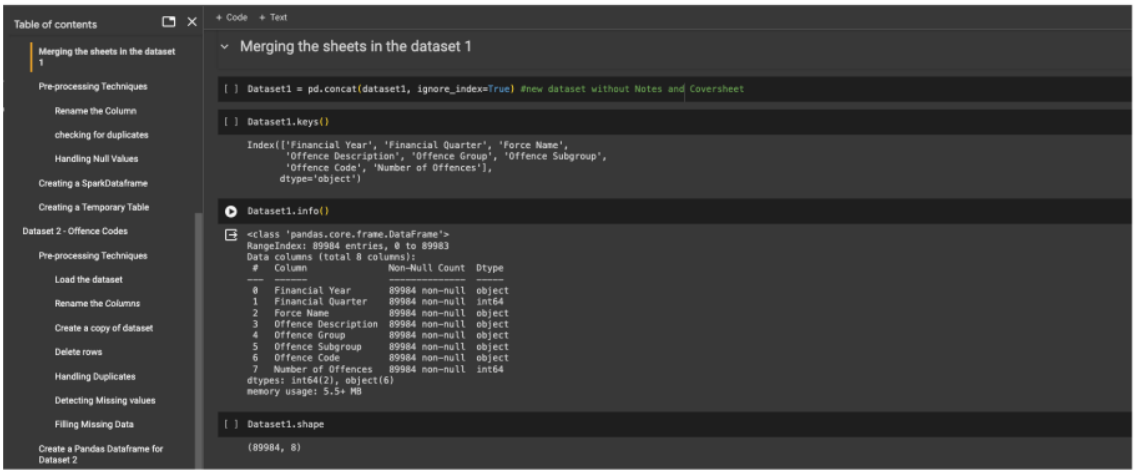


Figure 10: Merge Worksheets

2.6 Pre-Processing Activity

The pre-processing activities done in this section are to ensure the dataset is being worked on and prepared for final use. This is to remove redundancies and correct the errors identified.

2.6.1 Rename Columns

The first task is to rename the columns. This is done because having column names with spaces is difficult to read and work on.

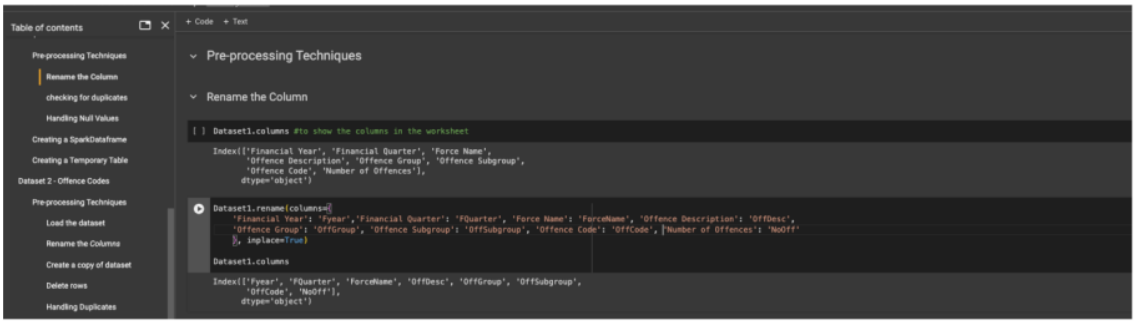


Figure 11: Rename Column

2.6.2 Drop Rows

I identified some unwanted rows at the end of our dataset and use pandas to remove them.

- dataset2.tail(5)
- dataset2.drop(dataset2.index[-2:], inplace=True) #delete last two rows of the dataset

Figure 12: Drop Rows

2.6.3 Handling duplicates.

In this dataset, I attempt to show a list of the duplicate rows.

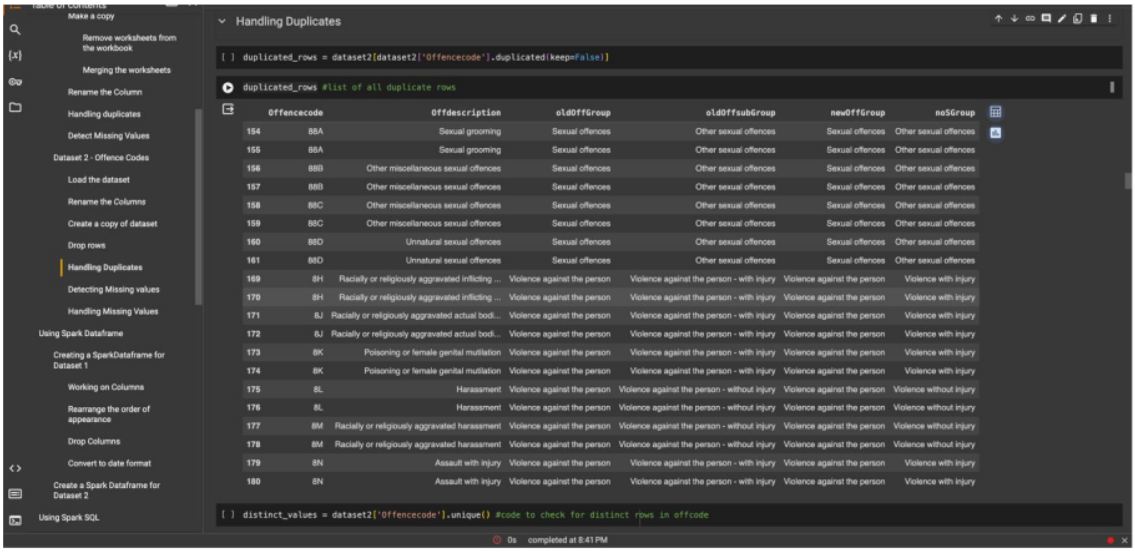


Figure 13: Show Duplicates

Some files could have been entered multiple times, therefore I use the .drop_duplicates() to remove duplicate entries.

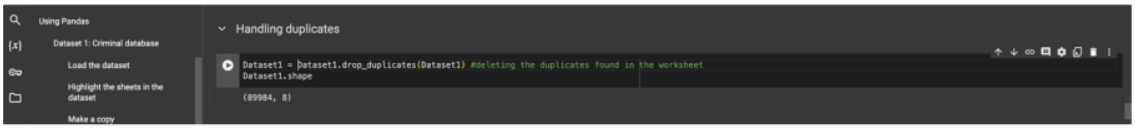


Figure 14: Handling duplicates in Dataset 1

There are no duplicates in dataset 1 as it still returns the initial dataset information.

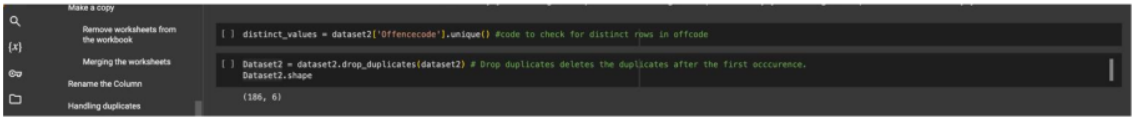


Figure 15: Handling Duplicates in Dataset2

2.6.4 Missing Values

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given data. Missing values in datasets are common in datasets and can sometimes be needed to ensure the accuracy of data. In any dataset, you must first

- Detect missing values.
- Handle missing values.

2.6.5 Detect Missing Values

The next step is to work with missing Null Values. To do this, I must first identify the null values in the dataset, upon identifying this, if any, I can then perform some activities on it.

As shown in the figure below,

- Isnull() will check the dataset and show true or false for where it finds null values.
- Isnull().sum() will check for sum of null values in each column.

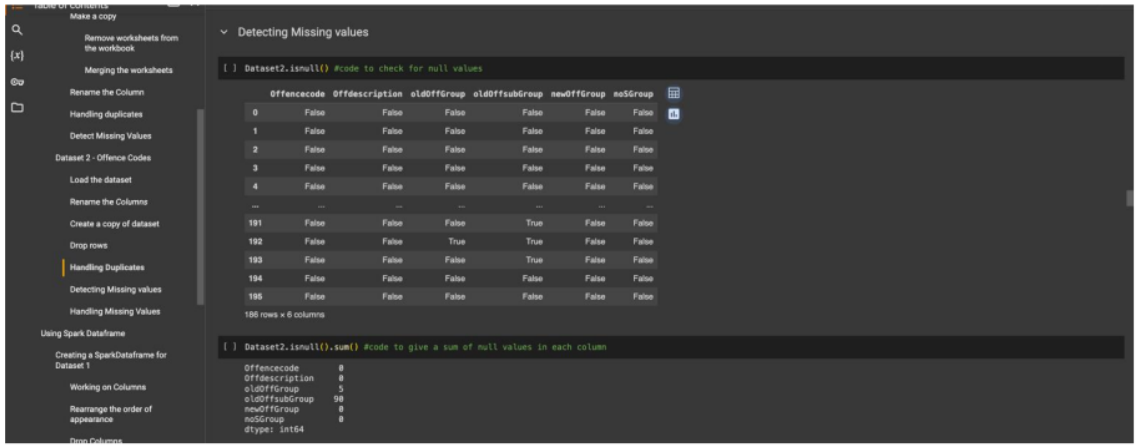


Figure 16: Detect Missing Values

2.6.6 Handle Missing Values

There are different methods to handle missing values some of which include using:

- Replace with mean, median, mode,
- Replace with constant values (0),
- Imputation (using nearby values),
- Delete rows/columns,
- Impute the data by yourself.

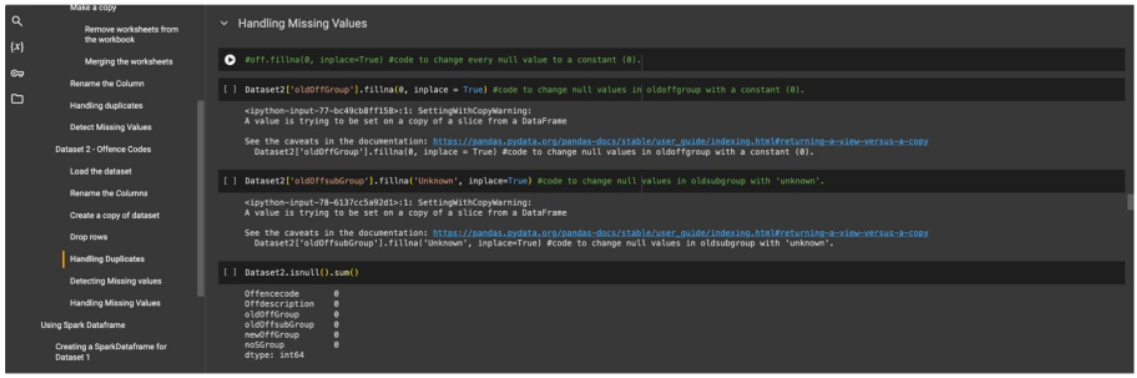


Figure 17: Handle Missing Values

A3. Using Spark SQL

3.1 Create a Spark data frame.

Upon completing the initial cleaning with pandas, I convert the padas dataset into a spark data frame.

Figure 18: Create a Spark Dataframe

3.2 Create a Temporary Table

For us to work with SQL, I have to save the spark dataframe on a SQL database by creating a temporary table to accept the values in datasets 1 and 2.

Figure 19: Create a temporary Table.

To perform SQL operations, I create a temporary table using createOrReplaceTempView(). This is done to accept the details of the dataset in the temporary table format in SQL.

3.3 Spark Queries

SQL Queries used in this code are:

3.3.1 SQL Join

SQL Join is used to merge different tables using similar values between both tables. It combines rows on different tables using the related columns. (W3Schools, no date)

The SQL join query can be done in different ways, some of which are:

- Inner Join
- Left Join
- Right Join
- Full join
- Union

Figure 20: SQL Left Join

3.3.2 SQL Select

SQL Select is used to specify the columns to view in the table.

- SQL Select Column_name is used to select a specific column in the table named column_name.

Figure 21: SQL Select

3.3.3 SQL Count

SQL Count is used to count the number of rows in a table

Figure 22: SQL Count

3.3.4 SQL Describe

SQL Describe gives a breakdown of the structure of a database.

Figure 23: SQL Describe

3.3.5 SQL Show Columns

Show Columns state is used to show the details of the columns in a database.

Figure 24: SQL Columns

3.3.6 SQL Distinct

SELECT DISTINCT statement is used to show distinct values in a column.

Figure 25: SQL Distinct

3.3.7 SQL Null Values

In SQL we first detect the null values using IS NULL.

After detecting Null Values, we can either handle them by:

- Deletion
- Imputation using Coalesce Function as shown below.

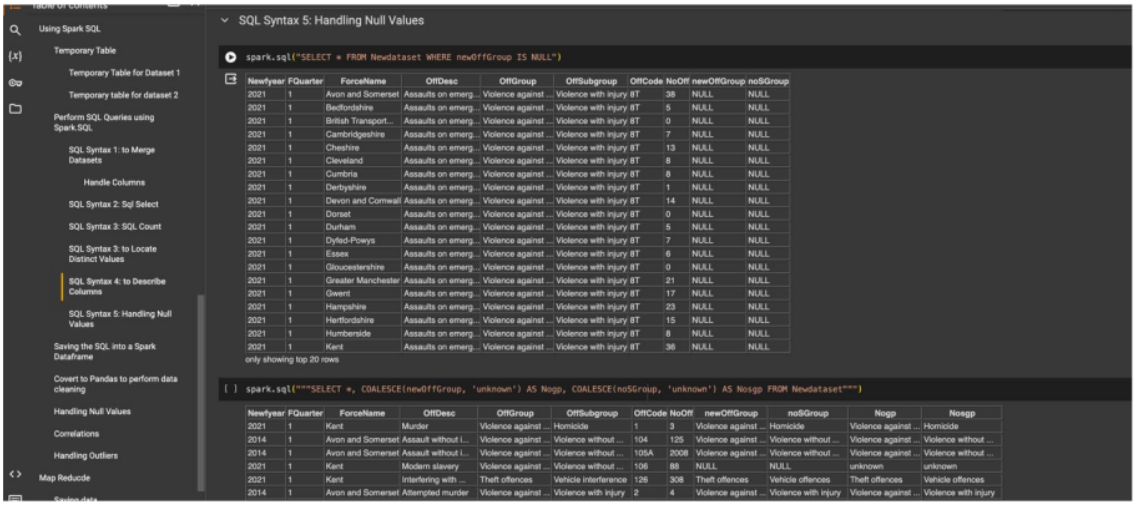


Figure 26: SQL Null Values

3.3.8 SQL Where

SQL where is used to specify a particular argument or position. It helps for filtering purposes.

Figure 27: SQL Where

3.3.9 SQL Order By

SQL Order by is used when trying to show a result in either ascending or descending order.

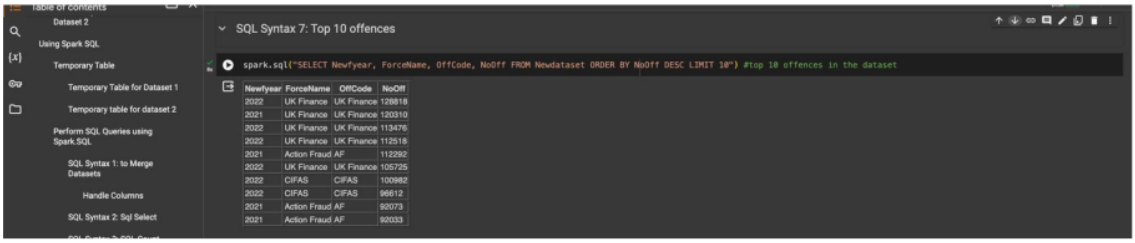


Figure 28: SQL Order By

A4. Using Hive

Upon completing the SQL queries and clean-up, I then attempt to write the SQL table into a Hive table to perform Hive Queries using Google Collab.

4.1 Convert Spark dataframe to Hive.

In this section, I convert the spark dataframe named Newdataset to Hive named NewHive.

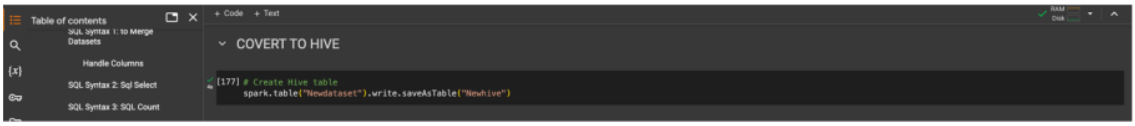


Figure 29: Create a Hive Table

4.2 Perform Hive Queries

4.2.1 Hive Query 1

Hive Query to show the hive table

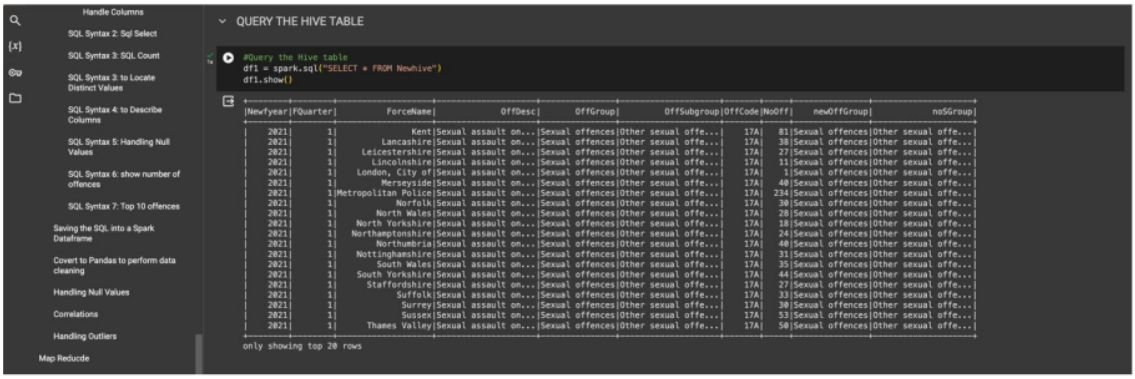


Figure 30: Hive Query 1

4.2.2 Hive Query 2

Hive query to the data of offenses in the year 2021

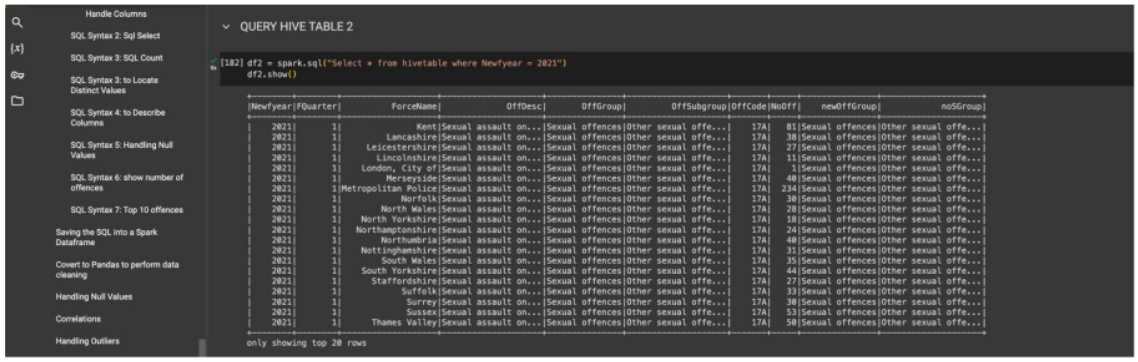


Figure 31: Hive Query 2

A5. Using Spark

5.1 Create a Spark Dataframe

Save the Hive table into a Spark dataframe to perform further analysis.

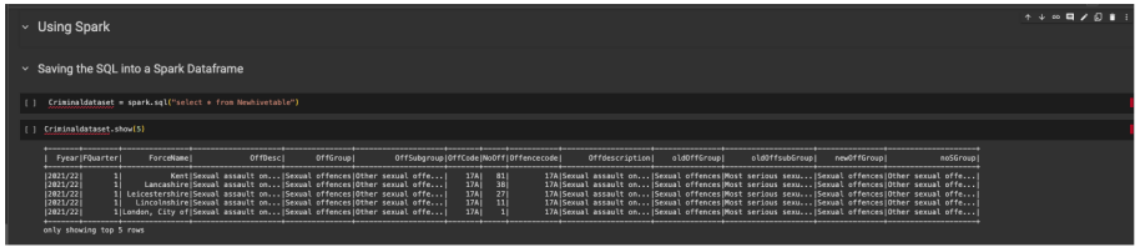


Figure 32: Create a New Spark Dataframe

5.2 Splitting on columns.

There are some inconsistencies with the data, and I attempt to clean it further using spark. In the figure below, I are splitting the Financial Year column by extracting values before and after \ into a separate column.

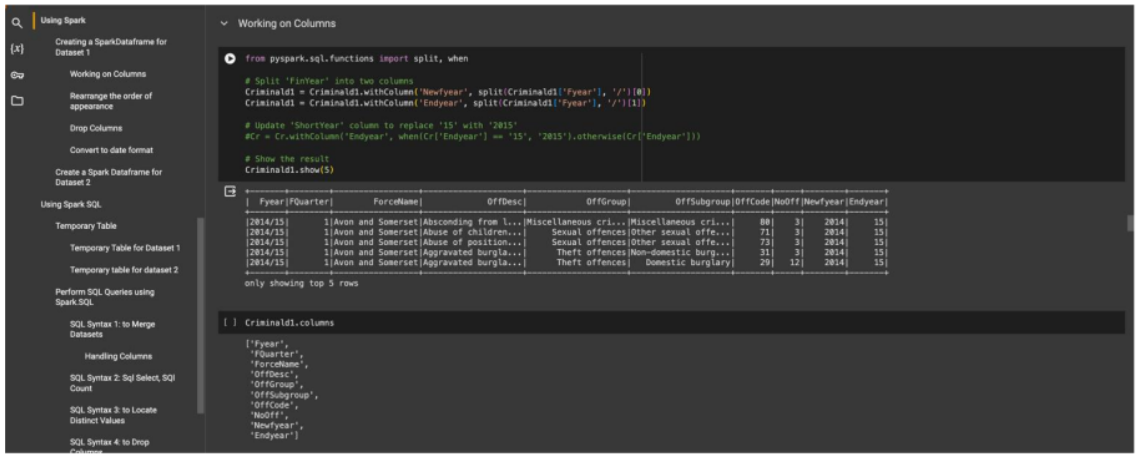


Figure 33: Split Column

5.3 Rearrange Columns

After splitting the column and creating new columns, I rearrange the order of appearance of the columns on the table.

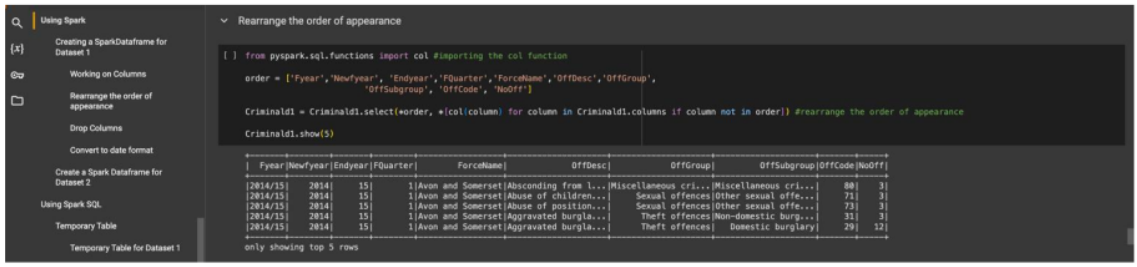


Figure 34: Reorder Column

5.4 Drop Columns

Now I attempt to drop the columns I don't need in the dataset. The columns not needed are dropped using .drop[] syntax.

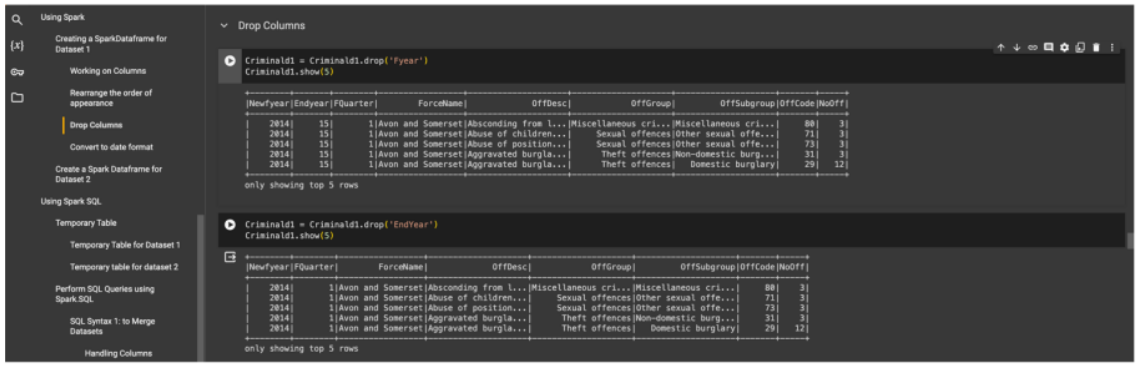


Figure 35: Drop Column

5.5 Correlation

Correlations is being used to find the similarities between different variables in a dataset.

I use cor() in the pyspark data frame to check the correlation between different columns with numerical values. The heatmap is being used to identify the relationship between different numerical variables while the pair plot is used to identify the relationship between different pairs of variables.

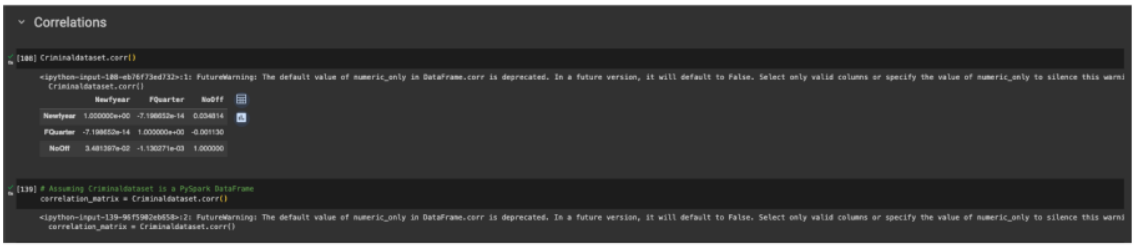


Figure 36: Correlations in the dataset

Figure 37: Numerical Correlation

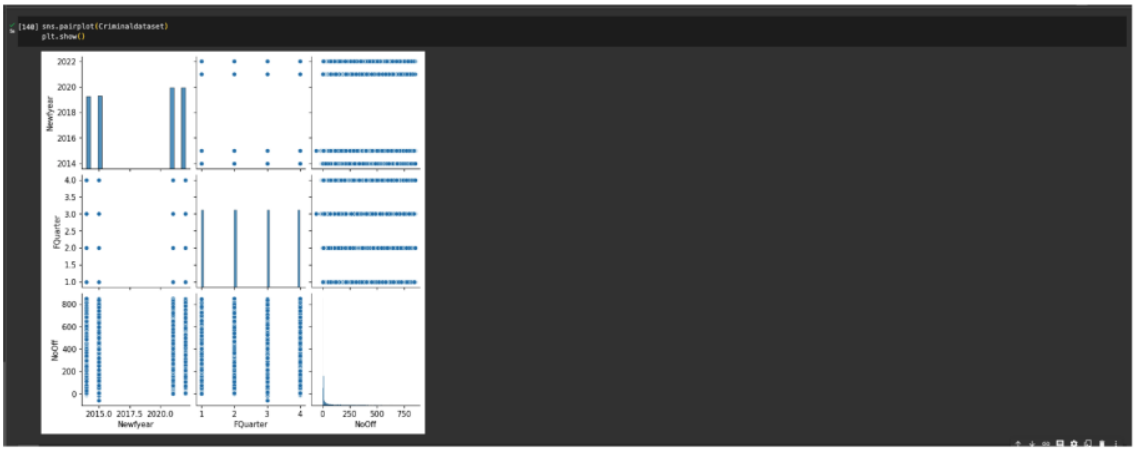


Figure 38: Bivariate Correlation

5.6 Outliers

Outliers are referred to as data points that significantly differ from other results.

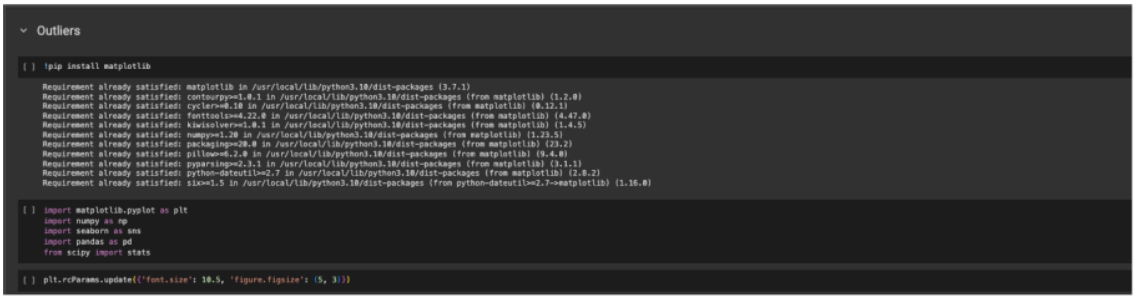


Figure 39: Installing Libraries for Outliers

5.6.1 Detect Outliers

To detect outliers, I plot a scatter plot diagram to plot a graph showing the outliers.

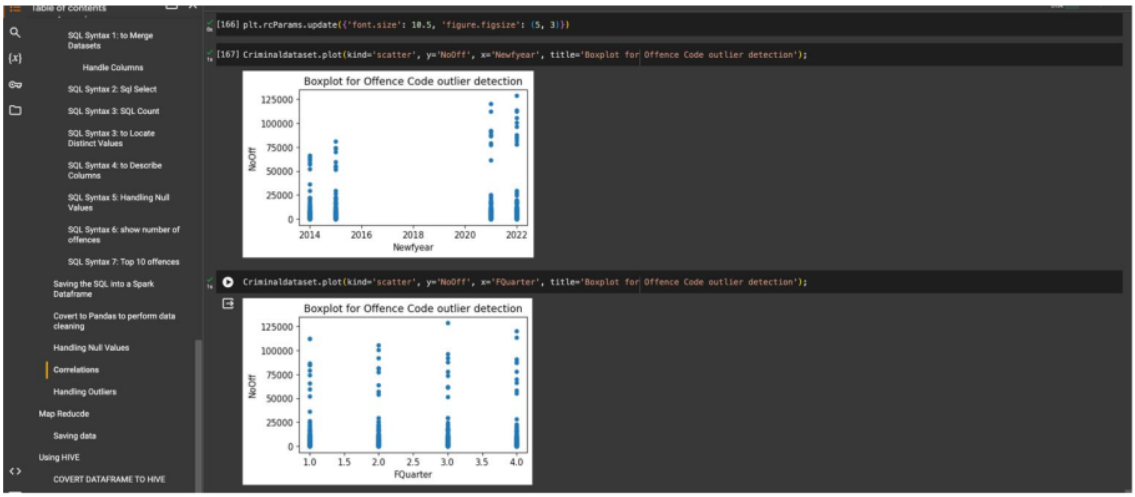


Figure 40: Detect Outliers

5.6.2 Handle Outliers

To handle outliers, I draw a graph showing the no of offenses against the financial years.

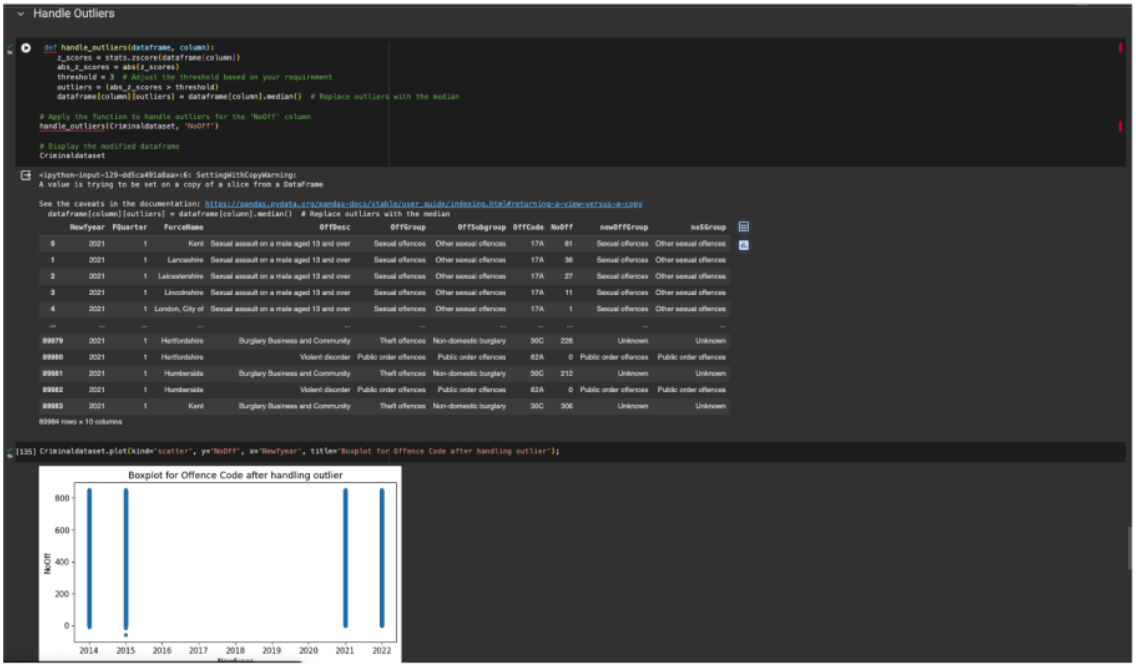


Figure 41: Handle Outliers

A6. Map Reduce

A MapReduce job usually splits the input data set into independent chunks which are processed by the map tasks in a completely parallel manner. Map reduce splits data into key-value pairs and reduces the data by counting the amount of time those key-value pairs appear.

To perform Map-Reduce for the dataset, I use the functools function and import the reduce functionality.

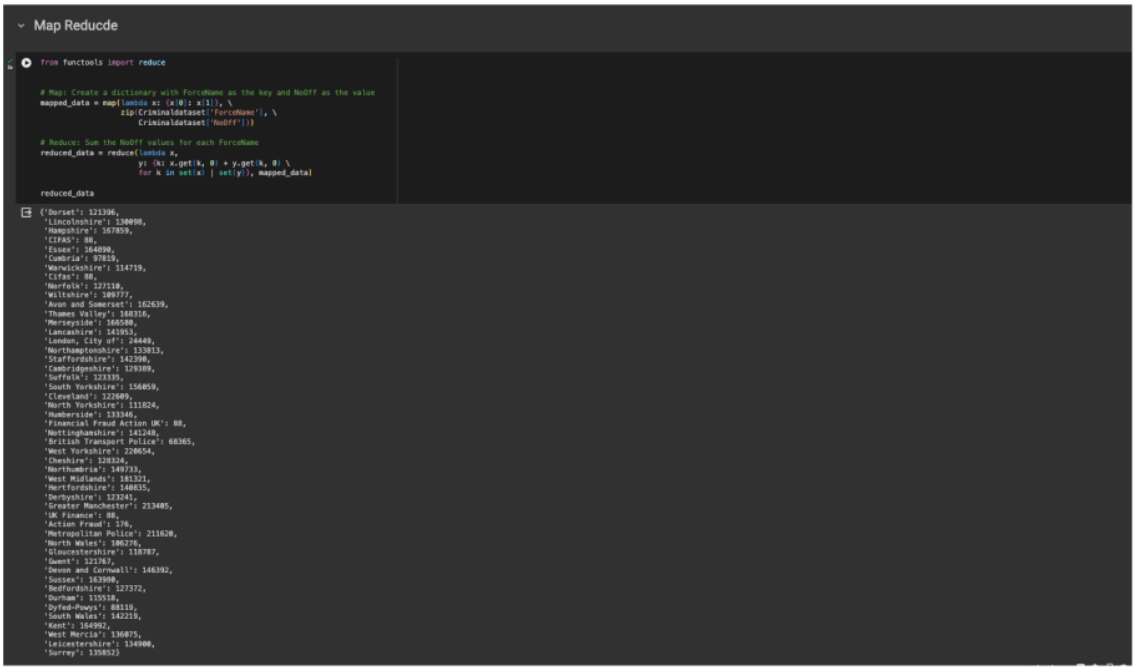


Figure 42: Reduce Functools

A7. Machine Learning

Installing all the necessary libraries needed to perform the predictions and all.

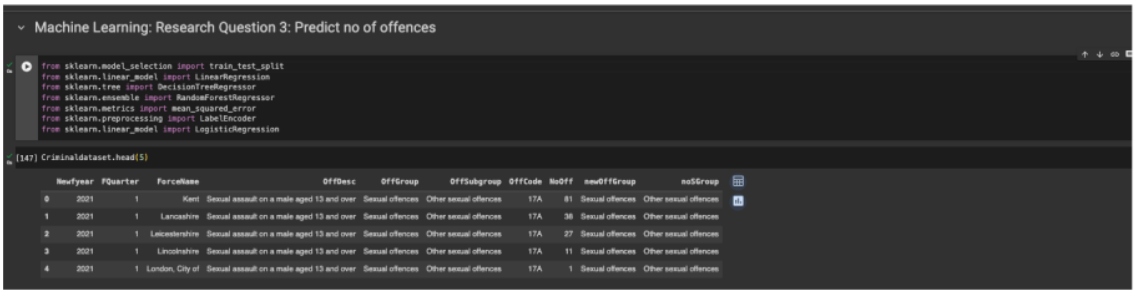


Figure 43: Importing Libraries

7.1 Extract the dataset.

Extract the data needed using .loc[:,[]]

Figure 44: Extract Dataset

7.2 Encoding the columns.

Encoding is done in ML for string columns. This is done by providing each column with a unique variable so I can perform ML queries on the specified columns.

Figure 45: Encode Categorical Variables

7.3 Split the dataset.

For us to perform ML queries I need to have a train and test dataset. Given the dataset provided is one large dataset, I have to write a code to split the dataset into a training dataset and a testing dataset.

To split the dataset, I use the train_test_split () as shown in the figure below.

Figure 46: Train_Test_Split

7.4 Features and Target

Here I set the features and target.

Figure 47: Features and Target

7.4.1 Highlight Features and targets for Train and testing data

The next step is to highlight the features and target for the test and training datasets.

Figure 48: Highlight Features and Target

7.5 Initiate Models

In the given dataset, I am trying to predict the number of offenses. Hence, the regression model is being used because I am working on getting continuous values.

Figure 49: Initiate Models

7.5.1 Linear Regression

Figure 50: Linear Regression

7.5.2 Decision Tree

Figure 51: Decision Tree

7.5.3 Random Forest

Figure 52: Random Forest

7.6 Mean Square Error

The Mean Square Error (MSE) calculation is being done to check for the preferred regression model to use with the least error.

Figure 53: Mean Square Error (MSE).

7.4 Plot Graph

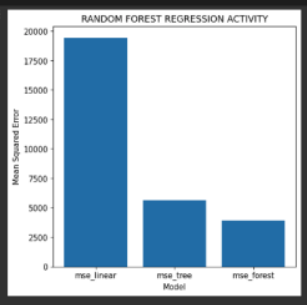
I attempt to plot a graph to capture the MSE calculation done in the figure above.

Plot the graph

```
from re import X
import matplotlib.pyplot as plt
plt.figure(figsize=(6, 6))

x_axis = ['mse_linear', 'mse_tree', 'mse_forest']
y_axis = [lr, tr, fr, ]

plt.bar(x_axis, y_axis)
plt.title('RANDOM FOREST REGRESSION ACTIVITY')
plt.xlabel('Model')
plt.ylabel('Mean Squared Error')
plt.show()
```




```
[162] # Determine the model with the least error
      min_error_model = min(mse_linear, mse_tree, mse_forest)

      if min_error_model == mse_linear:
          print("Linear Regression has the least error.")
      elif min_error_model == mse_tree:
          print("Decision Tree has the least error.")
      else:
          print("Random Forest has the least error.")

      Random Forest has the least error.
```

Figure 54: Least Error Graph

A8. Visualization

Data Visualization refers to the graphical representation of information and data with charts, graphs, maps, etc. In big data, data visualization refers to the graphical representation of data for an informed decision-making process (Kirk, 2019). The figure below shows a graphical representation of the datasets provided over some time.

8.1 Visualization with Power BI

In the figure below, Power BI is used to illustrate the research questions.

From the figure below, we can determine the total number of arrests to the nearest whole number, the total number of distinct forces involved, identify the financial year with the highest crime rate, and also identify the force with the highest number of arrests. Further analysis shows a detailed breakdown of the research questions, thereby enhancing comprehension and facilitating decision-making processes.

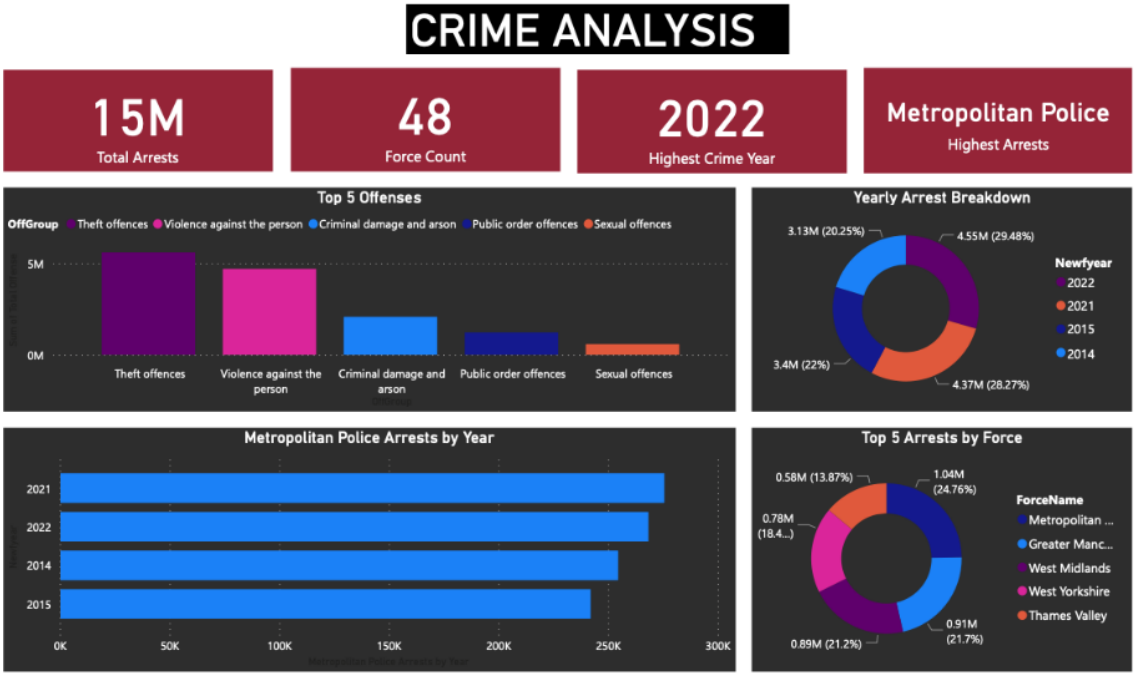


Figure 55: Crime analysis with Power Bi

8.2 Visualization with Tableau

Tableau is used for visualizing the dataset below using the story format of Tableau. Story 1 named Offense Analysis shows a breakdown of the total number of offenses against the financial year, offense group, and subgroup.

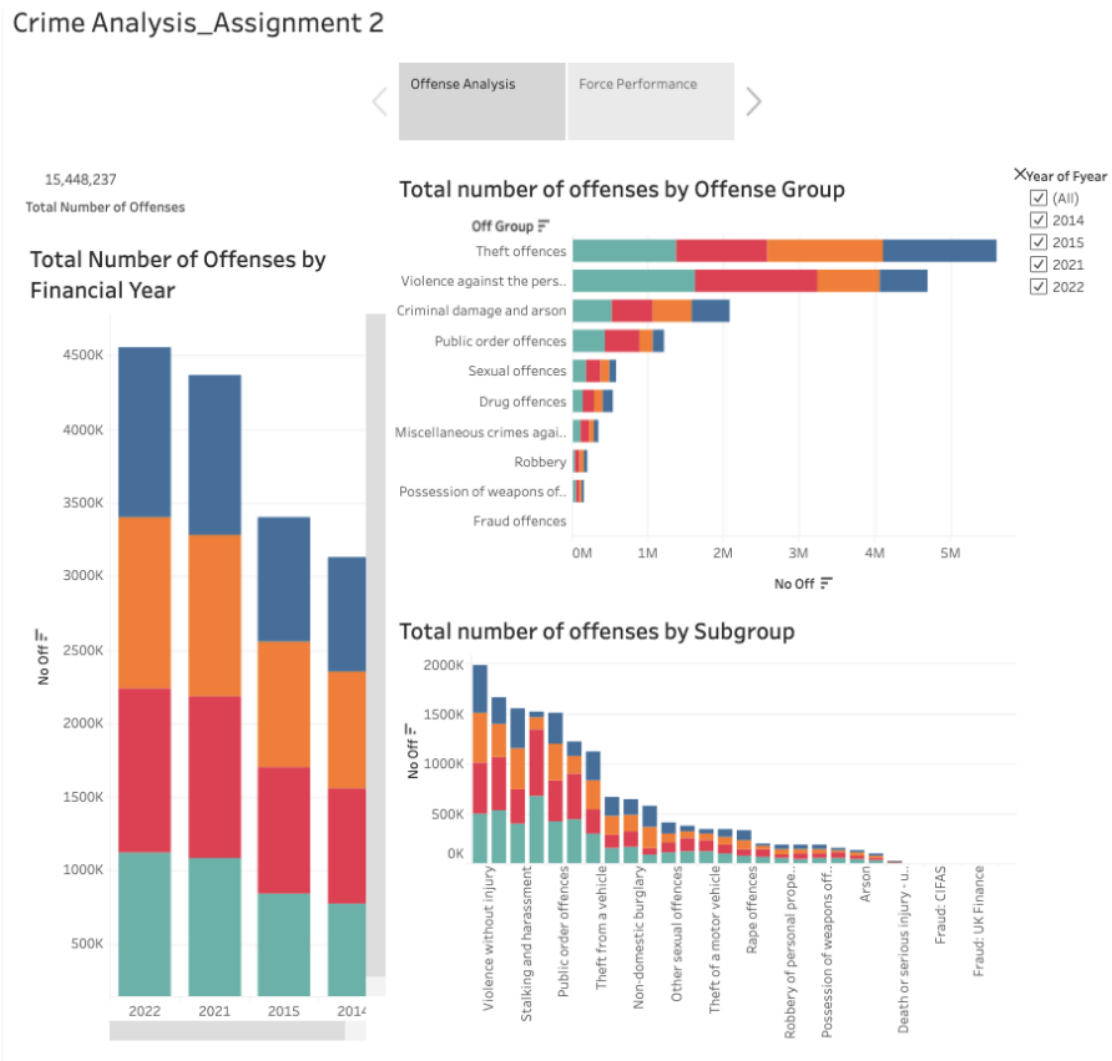


Figure 56: Offense Analysis

Story 2 named Force Performance shows a breakdown of how the forces performed before the changes made in 2017 and after the changes made in 2017.

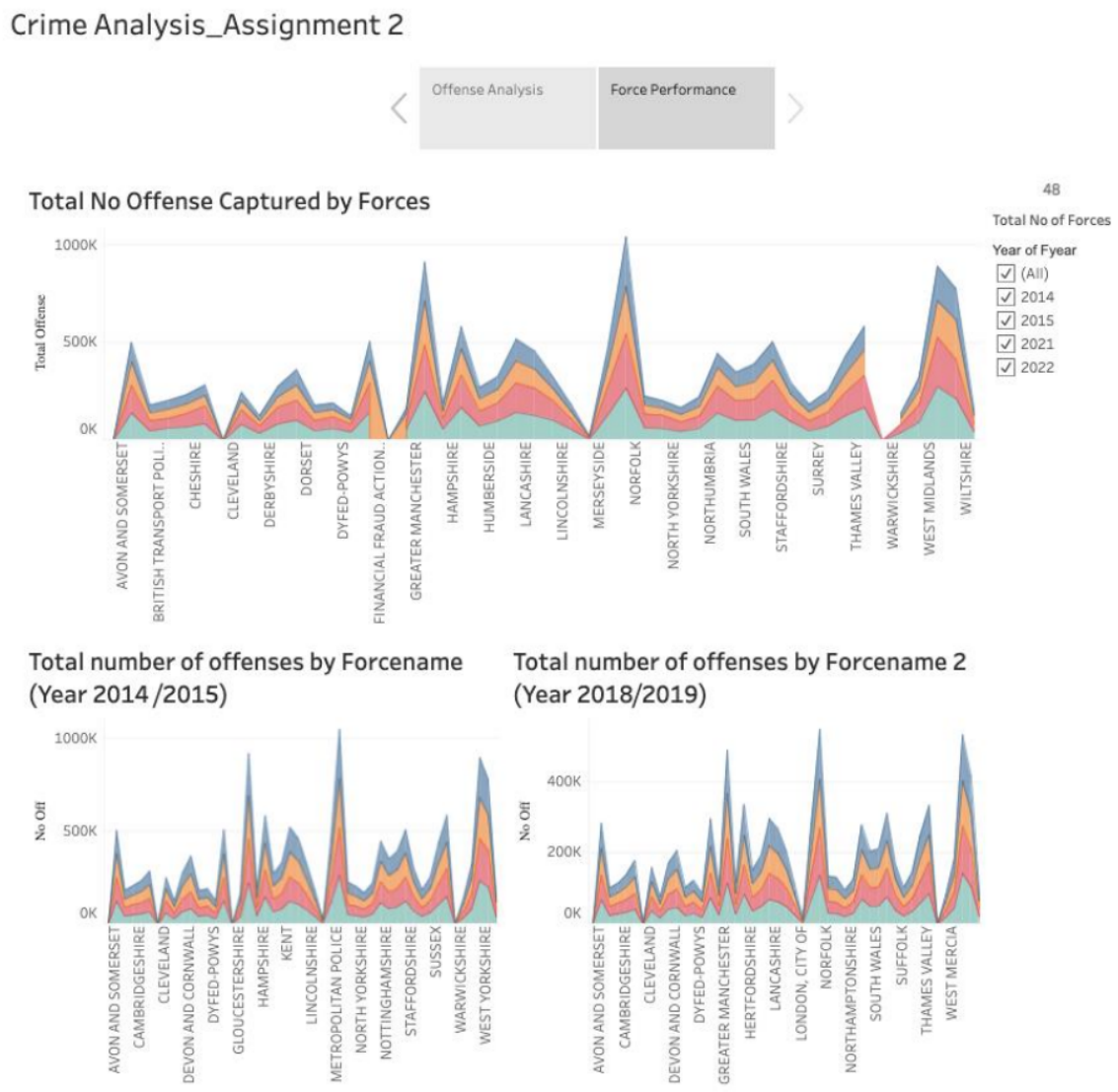


Figure 57: Force Analysis

Section B: Crime Analysis Report

B1. Analysis Answering Research Questions

Table 1: Analysis answering research questions.

Research Question	CHALLENGES	TECHNOLOGY USED	IMPACT
1. What are the most common types of offenses recorded in the database? What are the top five offenses in the dataset and what relevance does that cover in the criminal analysis?			
What are the most common types of offenses recorded in the database?	The extensive volume of data poses a challenge, as the substantial amount collected makes it time-intensive to meticulously review each entry and precisely identify the various offense types as well as the top ten offenses within the dataset.	• Spark • SQL Distinct • SQL Count • SQL Select • SQL Order By • DESC Limit	• Violence against the person • Theft offences • Sexual offences • Robbery • Public order offences
What are the top five offenses in the dataset and what relevance does that cover in the criminal analysis?			• Theft Offence • Violence against the person • Fraud Offences • Criminal damage offences • Public order offences
2. What force made the most arrests across all the financial years and in what financial year did they make the most arrest?			

What force made the most arrest?	Given the substantial volume of collected data, specialized tools are essential to efficiently identify all distinct law enforcement agencies and calculate the total number of offenses committed each year. This process is crucial for promptly discerning the law enforcement agency with the highest arrest rates.	• Hive SQL Syntax • SQL Sum • SQL Order By	• Metropolitan Police
What financial year did they make the most arrest?			• 2022/23

3. Are there any inconsistencies or anomalies in the data that need to be addressed to improve the data integrity? What technologies were adopted to ensure data reliability and robustness to provide meaningful contributions?

Are there any inconsistencies or anomalies in the data that need to be addressed to improve the data integrity? What technologies were adopted to ensure data reliability and robustness to provide meaningful contributions?	After scrutinizing the provided dataset, discrepancies were identified, introducing the potential for negative or inaccurate results. Additionally, the execution of calculations on the data may have led to erroneous decision-making.	• Pandas • Spark • SQL • Scatter plot • Heatmaps • Map Reduce	• Splitting columns • Drop columns/rows. • Renaming columns • Change data types. • Handle Missing Values • Handling Outlier • Reduce Functools 1.
4. What analytical approach proves most effective in predictive modelling for anticipating the frequency and nature of offenses, and how can this method be optimally applied to enhance proactive law enforcement strategies?			
What analytical approach proves most effective in predictive modelling for anticipating the frequency and nature of offenses, and how can this method be optimally applied to enhance proactive law enforcement strategies?	There was no way to accurately predict the frequency and nature of offenses.	• Regression Model/ Classification Model • Linear Regression • Decision Tree • Random Forest	Regression model is being used in this calculation because we are working on Number of offenses and are thus predicting a continuous variable.

B2. Discussions and Research Findings

In this section, the report addresses the syntax run to provide answers to the research questions.

2. **What are the most common types of offenses recorded in the database? What are the top five offenses in the dataset and what relevance does that cover in the criminal analysis?**

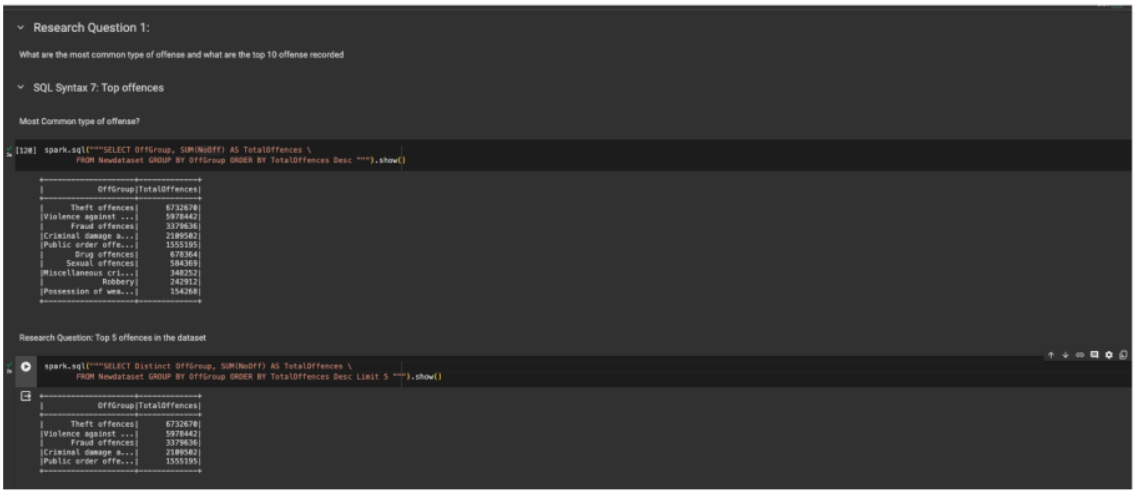


Figure 58: Research Question 1 Solution

The figure above provides a solution to the research question 1.

The most common types of offenses are:

- Violence against the person
- Theft offenses
- Sexual offenses
- Robbery
- Public order offense

The top five offenses are:

- Theft Offence
- Violence against the person

- Fraud Offenses
- Criminal damage offenses
- Public order offenses

Using this in crime analysis can help while performing predictive analysis to find the prevalent type of offense. It can also help government officials plan how to allocate resources and better tackle criminal organizations within an area.

3. **What force made the most arrests across all the financial years and in what financial year did they make the most arrest?**

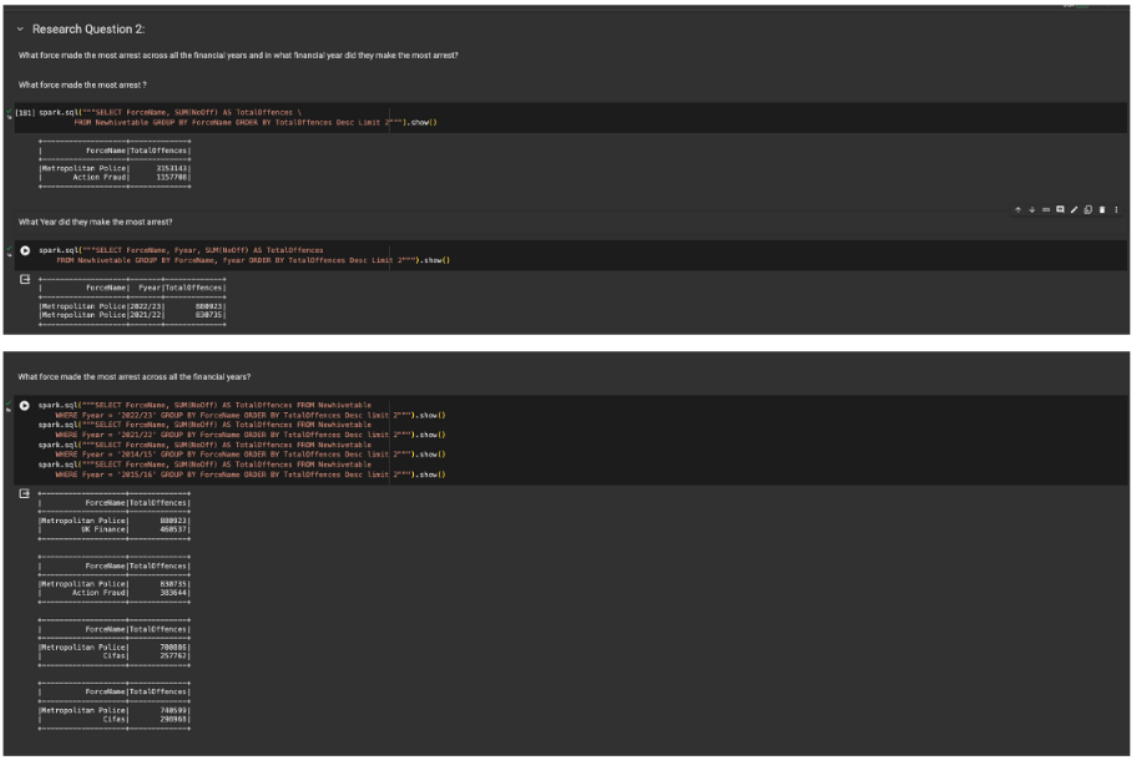


Figure 59: Research Question 2 Solution

The metropolitan police force made the most arrests in all the financial years where most of the arrest was made in the year 2022/23.

4. **Are there any inconsistencies or anomalies in the data that need to be addressed to improve the data integrity? What technologies were adopted to ensure data reliability and robustness to provide meaningful contributions?**

The inconsistencies or anomalies identified in the dataset were addressed using different solution methods across the notebook. Duplicates were handled using pandas, Null values were handled, and columns were dropped using spark. The report then proceeds to show how to handle outliers and perform map-reduce functionalities using the Functools function.

5. **What analytical approach proves most effective in predictive modeling for anticipating the frequency and nature of offenses, and how can this method be optimally applied to enhance proactive law enforcement strategies?**

Given the predictive analysis being done is for several offenses, the regression model is adopted to predict the number of offenses and also determine the best model to use with the least number of errors involved.

Figure 60: Regression Model Graph

B3. Issues In the Use of Big Data Analytics in Crime Analysis.

1. **Data privacy and security concerns:**

   The increased collection of data and use in crime analysis has raised several ethical issues surrounding the way the data is being collected and used. The constant use of big data solutions in crime analysis calls for uneasiness as it raises concerns about the volatility of data being collected imposing there has to be a that separates the collection of data and violation of data privacy (Mayer-Schönberger and Cukier, 2013).

2. **Bias and Discrimination Concerns:**

   Due to the historical crime data and algorithm prediction, there is an increased probability of bias and discriminatory behaviours towards a certain group. Bias can result from the group in charge of the algorithm or can also arise from the lack of diversity in the data collected (Diakopoulos, 2016).

3. **Data Quality and Accuracy:**

   In crime analysis, data quality and accuracy play a vital role. The probability of a flawed prediction arises from the quality of data collected. Incoherent, incomplete, or inaccurate data can lead to misleading insights and therefore result in the ineffectiveness of the police force and can also lead to bias (Manyika *et al.*, 2011).

4. **Ethical and Legal Implications:**

   The increased adoption of big data technologies in criminal analysis presents new opportunities to make law enforcement easier but also leads to different ethical considerations and legal implications of adopting the technologies. The registration of evidence to be used against offenders specifically evidence generated through the use of big data technologies can be challenged to be a breach of security and safety concerns and considered inadmissible (Hellman, 2020), and (Zarsky, 2016).

5. **Data Transparency with Public Perception:**

   Big data solutions in crime analysis are being accepted by various forces and it is imperative to keep the public aware of the intention and management of the data being collected. Involving the communities in the developments in the introduction of big data, the methodologies, and its implications has become vital as it reduces the resistance and eases doubts (Lum, 2021) , and (Berk, 2021).

6. **Security Risks:**

   Big data systems are prone to attacks and breaches which pose a significant threat to the validity and security of the data. As highlighted by (Dhillon and Torkzadeh, 2006), and (Schneier, 2016), there is an increased number of cybersecurity threats for big data systems. This can lead to tampering of data and inaccuracies of law enforcement in providing safety and justice when needed.

B4. Solution Using Various Big Data Technologies.

1. **Data Privacy and Security Concerns:**

   To mitigate the security concerns highlighted, some techniques can be adopted such as anonymization, pseudonymization, and data encryption. These techniques can be adopted to protect against misuse of information. Also, frameworks and policies should be put in place to effective and efficient data usage (Solove, 2013).

2. **Bias and Discrimination:**

   It should be made a framework for there to be regular checks on the algorithms used. To also ensure fairness, the data should be inclusive and there should be a responsible party managing the collection and use of the data (Selbst *et al.*, 2019).

3. **Data Quality and Accuracy:**

   (Redman, 1996) talks about implementing rigorous practices to ensure the quality of data being used in crime analysis. Emphasizing the need for regular security audits, validation of data collected, and scrutiny to ensure standard procedures are being followed when doing data collection. In addition, we must ensure proper data cleansing to remove irregularities and address concerns that might come up. Machine learning algorithms should be incorporated into crime analysis for anomaly detection and rectifying inaccuracies in the data collected.

4. **Ethical and Legal Implications:**

   Ethical guidelines that ensure data compliance with existing and new frameworks and principles should be put in place to ensure accountability and effectiveness of data thereby leading to accountability during the decision-making process. The frameworks adopted should ensure fairness and accountability across boards. It should incorporate collaboration between the respective stakeholders in law, technology, ethics, and community guidelines to ensure proper accountability.(Diakopoulos, 2016)

5. **Data Transparency with Public Perception:**

   Communication in any community is key to development. There should be inclusivity and effective communication in the decision-making process. This is to ensure that information is passed across within the community about how the data will be collected and how it will be used. Proper communication further eases the doubts, privacy, and ethical concerns the public might have.

6. **Security Risks:**

   Due to the increased number of cyberthreats in bid data, security measures should be put in place to ensure data protection. Cybersecurity measures should be introduced, the personnel should be educated, and informed on best practices to follow.(Schneier, 2016)

Conclusion

Big data utilization has become a major component of the crime analysis process. Different sectors have been adopting big data technologies which has led to an increase in efficiency and effectiveness and the crime sector is no different. Big data technology has been adopted in crime analysis, from crime patterns and predictions and automating the fraud detection process.

This report adopts a CRISP_DM methodology to answer the research questions and also a systematic review methodology to examine the challenges and solutions associated with the introduction of big data technologies.

The document examines the datasets, highlights the issues associated with working with the datasets, and provides solutions adopted to handle the issues. The report emphasizes the importance of big data technologies in crime analysis bringing us to a conclusion that there are more benefits to the introduction of big data solution.

References

Berk, R.A. (2021) 'Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement', *Annual Review of Criminology*, 4(1), pp. 209–237. Available at: https://doi.org/10.1146/annurev-criminol-051520-012342.

Bhandari, P. (2022) *Triangulation in Research | Guide, Types, Examples*, *Scribbr*. Available at: https://www.scribbr.co.uk/research-methods/triangulation-in-research/ (Accessed: 7 January 2024).

Dhillon, G. and Torkzadeh, G. (2006) 'Value-focused assessment of information system security in organizations', *Information Systems Journal*, 16(3), pp. 293–314. Available at: https://doi.org/10.1111/j.1365-2575.2006.00219.x.

Diakopoulos, N. (2016) 'Accountability in algorithmic decision making', *Communications of the ACM*, 59(2), pp. 56–62. Available at: https://doi.org/10.1145/2844110.

Hellman, D. (2020) 'MEASURING ALGORITHMIC FAIRNESS', *Virginia Law Review*, 106(4), p. 56.

Kirk, A. (2019) *Data visualisation: a handbook for data driven design*. 2nd edition. Los Angeles London New Delhi Singapore Washington DC Melbourne: SAGE.

Laura, M. (2021) *How to write the methods section of a systematic review*, *Covidence*. Available at: https://www.covidence.org/blog/how-to-write-the-methods-section-of-a-systematic-review/ (Accessed: 5 November 2023).

Lum, C. (2021) 'Perspectives on Policing: Cynthia Lum', *Annual Review of Criminology*, 4(1), pp. 19–25. Available at: https://doi.org/10.1146/annurev-criminol-093020-023221.

Manyika, J. *et al.* (2011) *Big data: The next frontier for innovation, competition, and productivity | McKinsey*. McKinsey Global Institute. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation (Accessed: 8 January 2024).

Mayer-Schönberger, V. and Cukier, K. (2013) *Big data: a revolution that will transform how we live, work and think*. 1. publ. London: Murray.

Page, M.J. *et al.* (2021) 'PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews', *BMJ*, p. n160. Available at: https://doi.org/10.1136/bmj.n160.

Redman, T.C. (1996) *Data quality for the information age*. Boston: Artech House.

Schneier, B. (2016) *Data and Goliath: the hidden battles to collect your data and control your world*. First published as a Norton paperback 2016. New York London: W.W. Norton & Company.

Selbst, A.D. *et al.* (2019) 'Fairness and Abstraction in Sociotechnical Systems', in *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19: Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, pp. 59–68. Available at: https://doi.org/10.1145/3287560.3287598.

Solove, D.J. (2013) *Introduction: Privacy Self-Management and the Consent Dilemma*, *Harvard Law Review*. Available at: https://harvardlawreview.org/print/vol-126/introduction-privacy-self-management-and-the-consent-dilemma/ (Accessed: 8 January 2024).

W3Schools (no date) *SQL Tutorial*, *W3Schools*. Available at: https://www.w3schools.com/sql/default.asp (Accessed: 8 January 2024).

Zarsky, T. (2016) 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making', *Science, Technology, & Human Values*, 41(1), pp. 118–132. Available at: https://doi.org/10.1177/0162243915605575.

Abbreviations

A List of Abbreviations in this document include:

- AI - Artificial Intelligence
- ML - Machine Learning
- pd - Panda
- np - numerical library
- plt - machine learning library for plotting graph
- sns - seaborn
- os - operating system

DAT7015_2_2312602

ORIGINALITY REPORT

12%	11%	6%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Bolton Student Paper	2%
2	e-space.mmu.ac.uk Internet Source	1%
3	Submitted to University College Birmingham Student Paper	1%
4	dspace.cuni.cz Internet Source	1%
5	Submitted to The University of Manchester Student Paper	1%
6	www.researchgate.net Internet Source	1%
7	Submitted to University of Liverpool Student Paper	1%
8	Submitted to University of Melbourne Student Paper	<1%
9	ouci.dntb.gov.ua Internet Source	<1%

10	www.ijirset.com Internet Source	<1 %
11	easychair-www.easychair.org Internet Source	<1 %
12	www.emerald.com Internet Source	<1 %
13	www.gov.uk Internet Source	<1 %
14	www.mdpi.com Internet Source	<1 %
15	fdotwww.blob.core.windows.net Internet Source	<1 %
16	justice.ie Internet Source	<1 %
17	Submitted to City University Student Paper	<1 %
18	Submitted to Victorian Institute of Technology Student Paper	<1 %
19	ndl.ethernet.edu.et Internet Source	<1 %
20	www.citethisforme.com Internet Source	<1 %
21	Submitted to University of North Texas Student Paper	<1 %

22	export.arxiv.org Internet Source	<1%
23	www.crystadigitalagency.com Internet Source	<1%
24	Submitted to Softwarica College Of IT & E-Commerce Student Paper	<1%
25	bds.sagepub.com Internet Source	<1%
26	docplayer.biz.tr Internet Source	<1%
27	edoc.unibas.ch Internet Source	<1%
28	fairmlbook.org Internet Source	<1%
29	repository.unsri.ac.id Internet Source	<1%
30	slashcry.com Internet Source	<1%
31	www.scribd.com Internet Source	<1%
32	etheses.whiterose.ac.uk Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

DAT7015_2_2312602

GRADEMARK REPORT

FINAL GRADE

85/100

GENERAL COMMENTS

Your submission demonstrates a strong performance, earning an excellent mark of 85. While your presentation was excellent, there were noted issues with time management that need improvement. Additionally, the lack of an introduction and background section hampers the overall structure of your work.

Your understanding of the problem scenario and current big data tools is commendable. However, to further enhance the quality of your work, consider organizing it with a logical flow and adhering to academic writing standards and professional report structure. These elements are essential for clarity and coherence in your presentation.

Furthermore, it's important to ensure that the best model is utilized in your analysis to optimize results. Paying attention to these aspects will help you achieve even greater success in future assignments. Keep up the good work, and continue refining your skills and techniques!

PAGE 1

Comment 1

???

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55

PAGE 56

PAGE 57