

**REPORT ON**

**UK Police Crime Data**

**at**

**Faculty of Computing, Engineering and Media**

**Name: Aathira Puthenpurayil**

**PNO: P2736418**

**Course: MSc Data Analytics**

|  |  |
| --- | --- |
| **Module name:** | **Big Data Analytics** |
| **Module code:** | **IMAT5322** |
| **Title:** | **Big Data Analytics – Assessment 2** |

Table of Contents

[Abstract 3](#_Toc135611544)

[Data Preparation and Exploration 3](#_Toc135611545)

[Data Gathering 3](#_Toc135611546)

[Data Loading 3](#_Toc135611547)

[Data Cleaning 5](#_Toc135611548)

[Data Modification 6](#_Toc135611549)

[Data Validation 6](#_Toc135611550)

[Data Analysis 7](#_Toc135611551)

[Analysis 1: Various crime types 7](#_Toc135611552)

[Analysis 2: Within Cities 9](#_Toc135611553)

[Analysis 3: Across cities 15](#_Toc135611554)

[Analysis 4: Trend across years 20](#_Toc135611555)

[Discussions and Findings 23](#_Toc135611556)

[Discoveries 23](#_Toc135611557)

[Conclusion 24](#_Toc135611558)

[References 25](#_Toc135611559)

# Abstract

The UK police dataset is retrieved from Data. police.UK website which contains street crime analysis for each city. The dataset gives valuable insights such as the month, reported by, falls within, latitude, longitude, location, LSOA name, LSOA code, crime type, last outcome and context.

In this project, the crime analysis is done with the West Midlands, the City of London and Leicestershire from the time period January 2022 to March 2023. Spark is used to perform various analytics to uncover key insights and trends. The analysis begins by examining the crime type counts for each city by identifying the crime type with the highest count, highlighting the most prevalent types of crimes in these regions. The analysis is done within the cities and across the cities for various comparison purposes. Furthermore, investigated the overall crime trends over time by analyzing the crime counts for different years. This analysis helps us understand variations in crime rates in the respective cities.

Analytics for crime in these three cities with Spark helped in gaining a deeper understanding of crime dynamics, aiding in the development of effective crime prevention and intervention strategies.

# Data Preparation and Exploration

## Data Gathering

The data is collected from [Data downloads | data.police.uk](https://data.police.uk/data/).[1] which contains relevant information such as the month, reported by, falls within, latitude, longitude, location, LSOA name, LSOA code, crime type, last outcome and context.

## Data Loading

Data is loaded as a single CSV file into Kaggle for analysis and the PySpark (Version 3.2) is installed in the platform for the analysis of data.

A picture containing text, screenshot, font, line

Description automatically generated

Secondly, implemented the code to handle warnings, start the spark session and configuration as well as create an instance for SQL Context [6].

A screen shot of a computer code

Description automatically generated with low confidence

Thirdly imported all the necessary libraries such as spark for various SQL operations, pandas [3] and NumPy for mathematical calculations and data frame creations, and matplotlib [4] and seaborn[5] for visualization of the analysed data.

A screenshot of a computer

Description automatically generated with medium confidence

The next step is the creation of a schema by mentioning the required datatypes for each field to ensure consistency and uniformity in the dataset, making it easier to understand, validate and work with the data.

A screenshot of a computer code

Description automatically generated with low confidence

As now the environment, libraries, and schema are ready, the final step is to read the imported data. Created a variable UKCrime to load the dataset. Using spark with options header as true, delimiter, and schema with created myschema loaded the CSV file into UKCrime.

A screen shot of a computer code

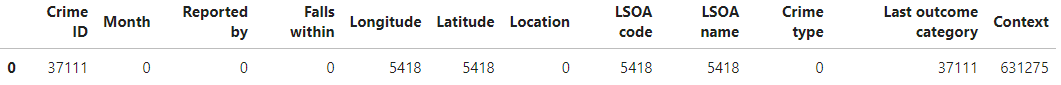
Description automatically generated with medium confidence

## Data Cleaning

Need to check if there are any Null values present in each column and convert the output to Pandas for better visualization. Execution of code gives the output with null values in crime ID and Last outcome category as 37111, LSOA Code, name, Longitude, and latitude has 5418 missing values, and finally with context having all rows missing with 631275.

A picture containing text, screenshot, font, line

Description automatically generated



As the same count is missing for crime ID and the Last outcome category needs to have further analysis to find its impact on other fields. Execution of code analysed the crime type Anti-social behavior has the crime ID and Last outcome category missing. But not removing the Last outcome category as it has crime type, month, and reported by fields which are required for the analyses. Replacing is also not done as its real-time data cannot replace with duplicates which can impact the analyses.

A screenshot of a computer code

Description automatically generated with medium confidence



Crime ID is automatic hexadecimal values generated and the Context column which has all fields blank, so both can be dropped.

A picture containing text, font, screenshot, white

Description automatically generated

The next analysis is done on the basis of missing counts in LSOA name and code and execution provides the result as the location field with No Location is having these fields missing which can lead to missing values in longitude and latitude and this is expected.

A close-up of a computer code

Description automatically generated with low confidenceA picture containing text, font, line, screenshot

Description automatically generated

## Data Modification

After cleaning, tidy up all necessary column names for better understanding.

A screenshot of a computer screen

Description automatically generated with low confidence

## Data Validation

Checking the final schema after cleaning and processing is good for analysis.

A screenshot of a computer code

Description automatically generated with medium confidence

Checking the first 3 rows to see if the data displays with the expected output.

A screenshot of a computer

Description automatically generated with low confidence

Checking the total row counts for the analysis.

A close-up of a website

Description automatically generated with low confidence 

# Data Analysis

For the analysis, purpose created a temporary view UKdata from UKCrime.

A close-up of a logo

Description automatically generated with low confidence

## Analysis 1: Various crime types

**Question 1:** **What crime categories does the Police data use in all three cities?**

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing font, text, white, graphics

Description automatically generated

Analysis: The various crime types in all three cities are displayed below.

A picture containing text, screenshot, font, number

Description automatically generated A screenshot of a phone

Description automatically generated with low confidence

**Question 2: Are the crime categories the same in all the cities?**

**Since the crime categories displayed above are for all three cities together for analysis within cities and across cities need to check if crime categories are the same in each of the three different cities.**

A screen shot of a computer code

Description automatically generated with low confidence

Analysis: Execution of code confirms that the crime categories are same in all three cities.



**Question 3: what is the prevalence of the top 5 crime types in certain locations in all three cities?**

**To check the top 5 crime types located in each city separately and also displaying its count will help to get better insights into the most happened crime types in each location and can help higher authorities to alert on the issue.**

A screenshot of a computer code

Description automatically generated with low confidence

Analysis: The below output suggests that in West Midlands and Leicestershire, most of the crime type reported is violence and sexual offenses with a count of 13505 cases in West Midlands which is much higher than in Leicestershire with 3114 cases reported in the location on or near whereas, In the city of London, most crimes reported is for theft from a person with a count of 238.

A picture containing text, screenshot, font, number

Description automatically generated A picture containing text, screenshot, font, number

Description automatically generated



A picture containing text, screenshot, font, number

Description automatically generated

## Analysis 2: Within Cities

**Question 1: What are the total counts for each type of crime recorded in West Midlands, the City of London, and Leicestershire?**

Through the analysis of the total count of each crime type, the investigation authorities can get better insights into different crime type occurrences in each city for future investigation.

A screenshot of a computer code

Description automatically generated with low confidence

Analysis: In West Midlands, the crime count for all locations is high for violence and sexual offenses followed by vehicle crime in second position and the least reported is Bicycle theft. In the city of London, the other theft category has the most reported crime followed by violence and sexual offenses and the least crime reported for possession of weapons. In Leicestershire violence and sexual offenses has more cases reported followed by public order and the least crime reported by robbery.

A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence

**Question 2: What is the total count of the most reported crime types in three locations considering all locations?**

This Analysis will help to get a view of the total number of highest crimes that happened in each of these three cities

A screen shot of a computer code

Description automatically generated with low confidence

From the result analysis, West Midlands police reported the highest crimes for violence and sexual offenses with a count of 202926 followed by the Leicestershire police with 54951 cases. City of London police reported the highest crime type for other theft with a count of 2024.

A picture containing text, screenshot, font, line

Description automatically generated

**Question 3: Are levels of violent crime constant, increasing, or decreasing?**

**From the above analysis, both Leicestershire and West Midlands had violence and sexual offenses with the highest count and the city of London has the second highest for this crime type therefore an analysis in common for all cities to check whether the violent crime is constant or increasing or decreasing across months.**

A picture containing text, screenshot, font

Description automatically generated

**Analysis: From the graph plotted, violent crime numbers gradually increase from April 2023, dropped again in May and from June again increased and reached their pick in the month of July 2022 and sudden drop is noticed for the upcoming months and the least crime is marked in 2022, December.**

A graph with red lines

Description automatically generated with low confidence

**Question 4: How are the Last Outcome Categories distributed in each of the three cities?**

**The last outcome categories analysis will help to attain the insights of each city’s final outcome obtained by combining all crime types by each city police.**

A screenshot of a computer

Description automatically generated with low confidence

Analysis: West Midlands recorded the highest number of cases with an investigation completed followed by Leicestershire and the city of London police. It is noticed that there are no cases with offenders given a penalty by Leicestershire and west midlands police. There is no case with an offender given a drug possession penalty in the west midlands.

A screenshot of a computer

Description automatically generated with low confidence

**Question 5: what is the Count of Police Service Who Reported the Crime?**

**This analysis will help to understand each city and how many cases have been recorded in total and can also compare between cities to find which city has the most crimes reported.**

A close-up of a computer screen

Description automatically generated with low confidence

A screen shot of a computer

Description automatically generated with medium confidence

**Analysis: West Midlands police reported the highest crime count with a total of 479044 crimes whereas Leicestershire police with 142239 and the city of London with 9992 crimes.**

A picture containing text, font, line, number

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

**Question 6: what the top 5 crimes per location for an ALL category in are each city?**

This will help to identify in each city what are the top five locations most crimes had occurred.

A screenshot of a computer code

Description automatically generated with medium confidence

**Analysis: In West Midlands on or near locations marked the highest number of crimes almost reaching a peak of 50000 in count whereas the next top three locations including on or near parking areas and shopping areas reported the same number of crimes and a little fewer in supermarkets and the least in a petrol station.**

A picture containing text, diagram, plot, rectangle

Description automatically generated

**In the city of London, the greatest number of crimes are reported in on or near the following with no location. On or near Bishopsgate marked more than 200 crimes whereas slightly less in Liverpool Street. On or near Victoria Avenue has least in the top 5 locations with a count of nearly 200.**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

**Leicestershire police have most crimes located on or near followed by no location. On or near the supermarket reported nearly 2000 crimes followed by the parking area and the least in petrol station**

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

## **Analysis 3: Across cities**

**Question 1: what are the top 10 Hot spots with the highest crime types across cities?**

A screenshot of a computer screen

Description automatically generated with low confidence

Analysis: In all three cities most of the crime type happened in on or near location type. West Midlands reported the top 10 crime hotspots in on or near parking areas with the most occurring crime types as violence followed by shoplifting, other theft, public order, vehicle crime, criminal damage, anti-social behaviour, burglary, and robbery whereas Leicestershire with highest crime spots in on or near followed by no location and parking area with all crime types in west midlands except burglary and robbery. City of London hotspots includes on or near, no location and bishop gate with all crime types in Leicestershire along with other theft and other crime as well.

A screenshot of a computer screen

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with low confidence

**Question 2: What are the crime rates, crime types, and outcomes reported across the city?**

Analyzing the crime rate will provide which city has the most crimes reported as per the data.

A screen shot of a computer code

Description automatically generated with low confidence

Analysis: West Midlands police has the highest crime report rate of 75% whereas Leicestershire has around 23% and the city of London with 1.5%. Crime types are the same reported across the cities. Types of outcomes also remain the same across the cities, but each city’s outcome count differs.

A picture containing text, font, screenshot, line

Description automatically generated

A close-up of a text

Description automatically generated with low confidence

A close-up of a document

Description automatically generated with low confidence

**Question 3: Across cities which city reported the highest crime type along with the crime category and count?**

This code will analyze the highest crime count reported city from the west midlands, London, and Leicestershire with the crime category and the count in numbers

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with medium confidence

Analysis: The city with the highest crime type count is reported in the west midlands with the crime type as violence and sexual offenses with a count of 202926.



**Question 4: what is the count of the last outcome strategy?**

This code will help to get the total count for each outcome category across cities through which can get an insight into which outcome has the highest and least distributions

A picture containing text, screenshot, font

Description automatically generated**Analysis: Across the cities, most of the crime was unable to prospect suspect categories, and the second highest with an investigation completed without identifying the suspect. The remaining categories of the last outcome are less than 50000 in the count.**

A picture containing text, screenshot, number, parallel

Description automatically generated

**Question 5: How are the last outcome categories distributed across cities?**

**This code will analyse the total percentage of distributions of the last outcome with the help of a pie chart.**

A screen shot of a computer code

Description automatically generated with low confidence

**Analysis: West Midlands police has most of the investigations completed with a record of 75.45 followed by Leicestershire police with 22.5% and the City of London police with 1.6%.**

A blue circle with orange and green text

Description automatically generated with low confidence

## **Analysis 4: Trend across years**

**Question 1: Which is the city with the highest crime reported in the year 2023?**

**This code provides which city has the highest crime reported in the year 2023.**

A screenshot of a computer code

Description automatically generated with low confidence

Analysis: West Midlands reported the highest crime among the other two cities.



**Question 2: Which month has the highest reported crimes in each city?**

This code provides each city with the highest crime counts reported per month.

A picture containing text, font, line, screenshot

Description automatically generated

Analysis: The city of London has the maximum crime count in December whereas Leicestershire reported maximum crime in July and the west midlands in April.

A picture containing text, font, line, screenshot

Description automatically generated

**Question 3: what is the trend in the number of crimes over time in all cities together?**

**Code analyses the number of crimes for each crime type across the year 01-2022 to 03-2023. This will give an overview of the trend across all crime types in each year.**

A screenshot of a computer code

Description automatically generated with medium confidence

Analysis: From the plot most crimes happened in the year 2022 during the months of May, July, and August, and the least crimes in December. In 2023, march month reported the highest crime count and the least in February. The chart also provides violent crime as the highest and least with possession of weapons and other crimes.

A picture containing text, screenshot, colorfulness, line

Description automatically generated

**Question 4: what is the forecasted number of crimes for the next 12 months?**

**This snippet provides the future 12 months’ forecast about whether crime will decrease, increase, or remain constant using the ARIMA model for prediction.**

A screenshot of a computer code

Description automatically generated with low confidence

**Analysis:** The following graph predicts that there will be an increase in count from April to June and a dip in July. Then there is a small increase in August and a gradual dip in September and the graph remains constant for the remaining time periods.

A picture containing text, plot, line, diagram

Description automatically generated

# Discussions and Findings

* The analysis provided information on various types of crimes reported, such as theft, burglary, assault, etc. to identify the most prevalent crimes in each city which can be useful for law enforcement agencies in understanding the nature of crimes in specific areas.
* The analysis identified the city with the highest count of a specific crime type which can assist law enforcement agencies and policymakers in targeting specific areas for crime prevention initiatives.
* By examining crime data across different cities, geographic variations in crime rates and patterns can be observed. Understanding these can aid in developing localized crime prevention strategies and allocating resources effectively.
* Analyzing the distribution of last outcome categories provides insights into the effectiveness of the criminal justice system, including the rates of arrests, charges, convictions, and other resolutions to highlight areas for improvement in the criminal justice process.

# Discoveries

* Unexpected Crime Hotspots: The analysis revealed certain areas within cities that experienced significantly higher crime rates compared to others. Identifying these unexpected crime hotspots can help focus law enforcement efforts and implement targeted interventions to reduce crime in those specific areas.
* Seasonal Variations: The analysis uncovered seasonal variations in crime rates. Certain types of crimes showed higher occurrences during specific seasons or months. Understanding patterns can assist in allocating resources and implementing proactive measures to address crime spikes during times of the year.
* Disproportionate Crime Types: The analysis may uncover certain crime types that are disproportionately high compared to others to challenge existing assumptions or priorities regarding resource allocation and crime prevention strategies that have a significant impact on public safety.
* Inefficiencies in the Criminal Justice System: Examining the last outcome categories of reported crimes can highlight potential inefficiencies or gaps in the criminal justice system. These discoveries can trigger discussions on improving the efficiency and effectiveness of the criminal justice system.

# Conclusion

Overall, the analysis of the UK police dataset provides valuable insights into crime patterns, trends, and outcomes through data exploration, analysis, and visualizations using Spark. These findings can inform law enforcement agencies, policymakers, and researchers in making data-driven decisions to enhance public safety, allocate resources effectively, and develop targeted crime prevention strategies.

By leveraging the power of big data analytics and Spark [2], we were able to process and analyze large volumes of crime data efficiently. The use of Spark's distributed computing capabilities allowed us to handle the dataset's size and complexity, enabling faster processing and analysis.

These insights will help the higher authorities and investigators to have more in-depth analysis criteria in the future, but analyses are done with historical data so need continuous monitoring and analysis as well to get better results.

With the continued advancement of big data [7], it opens the door for more realistic analysis in the future which can help society to have better insight.

# References

1. Home Office. (2021). UK Police Data. Retrieved from <https://www.gov.uk/government/statistics/police-recorded-crime-open-data-tables>
2. Apache Spark. (n.d.). Retrieved from <https://spark.apache.org/>
3. Pandas Documentation. (n.d.). Retrieved from <https://pandas.pydata.org/docs/>
4. Matplotlib Documentation. (n.d.). Retrieved from <https://matplotlib.org/stable/contents.html>
5. Seaborn Documentation. (n.d.). Retrieved from <https://seaborn.pydata.org/>
6. Python Software Foundation. (n.d.). Python Programming Language. Retrieved from <https://www.python.org/>
7. Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51-56). Retrieved from <https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf>