# 

***Report of the Project Done in Partial Fulfilment of the Requirements for the Award of Master of Science in Data Analytics***

**Big Data Analytics** - **IMAT5322**

REPORT ON POLICE DATA UK



*Done by*

Juny Maria John – P2743266

*Under the guidance of*

**Ms Bangli Liu**

**Mr Bamba M**

*Professor at the Department of Computing, Engineering & Media*

Master In Science

Data Analytics

Faculty of Computing, Engineering & Media (CEM)

De Montfort University

Leicester – LE1 9BH

**MAY- 2023**

Contents

[Abstract 2](#_Toc135610136)

[Introduction 2](#_Toc135610137)

[Task Breakdown 2](#_Toc135610138)

[Methodology 3](#_Toc135610139)

[Result Analysis 3](#_Toc135610140)

[1.Initialization of Pyspark and Importing files and Libraries: 3](#_Toc135610141)

[2.Data Pre-processing 3](#_Toc135610142)

[3.Data cleaning 3](#_Toc135610143)

[3.1 Checking for missing values 3](#_Toc135610144)

[3.2 Handling missing values 4](#_Toc135610145)

[4. Exploratory Data Analysis: 4](#_Toc135610146)

[4.1 Crime Data Analysis: 5](#_Toc135610147)

[4.2 Crimes across and within cities 6](#_Toc135610148)

[4.3 Analysing crimes on the basis of location 10](#_Toc135610149)

[4.4 Identifying the trends across the years 11](#_Toc135610150)

[Highlighting discoveries and findings: 15](#_Toc135610151)

[Results 16](#_Toc135610152)

[Conclusions 19](#_Toc135610153)

[Appendix 20](#_Toc135610154)

[Appendix A: Features 20](#_Toc135610155)

# 

# Abstract

The Police open-source data is real-time valuable data collected by enforcement agency they provide valuable information related to crime such as crime types, incident locations, investigation outcomes and so on. These are essential information which can provide insights for researchers to analyses different trends and patterns which can help to promote transparency, accountability, and data-driven decision-making in the criminal justice system. In this report, we will utilize PySpark 3.2 to perform the data visualization of the dataset on the Kaggle-platform. We would be analysis the dataset and will be deriving key points necessary from the analysis.

Keywords: Pyspark, Dataframe, pandas

# Introduction

In order for the analysis, I have also considered the data from: **Cumbria Constabulary, Leicestershire Police, Nottinghamshire Police** for the data ranging from **April 2020 to March 2023** collected from [Data downloads | data.police.uk](https://data.police.uk/data/).In total, I had 108 files which was uploaded to Kaggle for visualization.

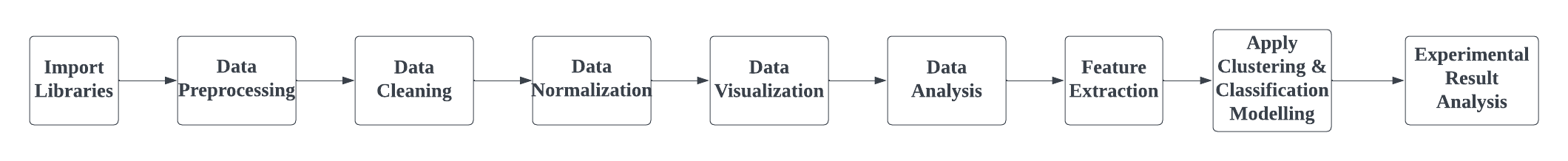
## Task Breakdown

1. Prepare the data – import and data cleaning in order to converting it into organized data.

2. Extract the characteristics *(Refer Appendix A: Features).*

3. Analyze information using various plots and pandas

# Methodology



• Import the dataset and comprehend the main libraries.

• Conduct an Exploratory Data Analysis (EDA) by utilizing Pyspark and Pandas.

• Identify and handling missing values (NaN)– Raw data contains noise.

• Gain insights from data by visualizing data characteristics and by reviewing them.

**Tools Used:** Kaggle, Pyspark3.2

**Programming language:** Python

**Analysis :** Bar, Pie, Violin and line chart and Pandas for the analysis

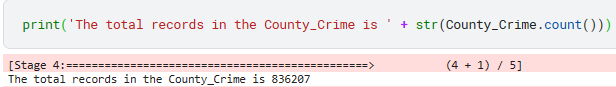
# Result Analysis

## 1.Initialization of Pyspark and Importing files and Libraries:

Pyspark3.2 is installed and initailized, after which libraries necessary for the analysis is imported.



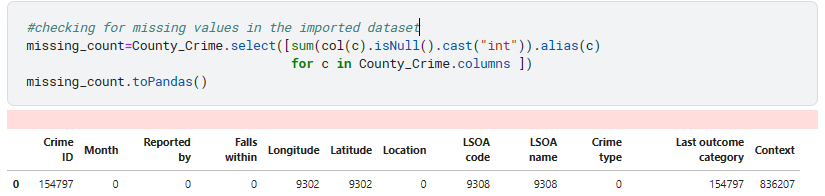
## 2.Data Pre-processing

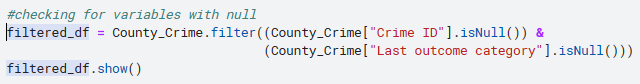
The police source-data files - Cumbria Constabulary, Leicestershire Police and Nottinghamshire Police for the year April 2020 to March 2023 are imported.It contains 108 files.

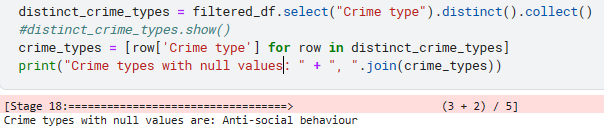
## 3.Data cleaning

### 3.1 Checking for missing values

Raw data contains noise and cannot be directly used for analysis. Therefore it is necessary to convert raw data before analysis.





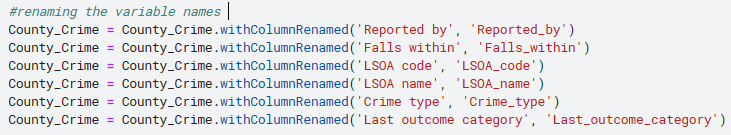


### 3.2 Handling missing values

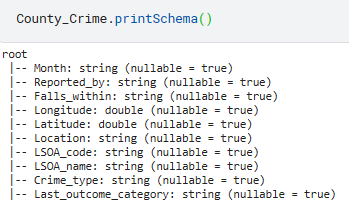
Removing variables with missing values and leaving the longitude and latitude as it is.

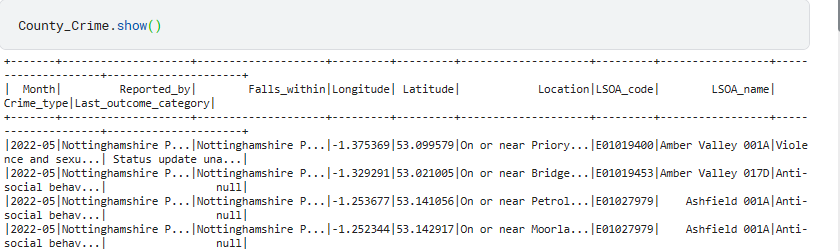


We are changing the column names of the dataset:



Final data output used for analysis contains:





## 4. Exploratory Data Analysis:

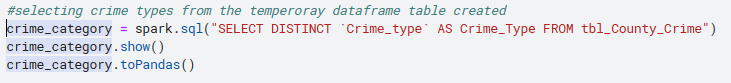
The main **research questions** focused while analysis of the public crime-data :

1. **Crime data analysis**
2. **Crimes across and within the cities**
3. **Analysing crimes on the basis of location.**
4. **Identifying the trends across the years**

## 4.1 Crime Data Analysis:

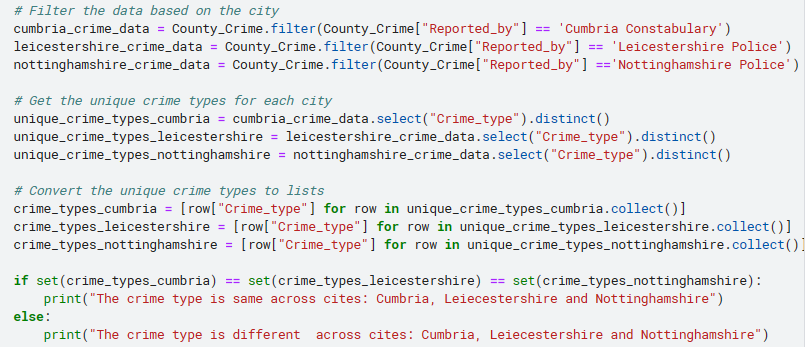
**What types of criminal offenses are included in the police data?**





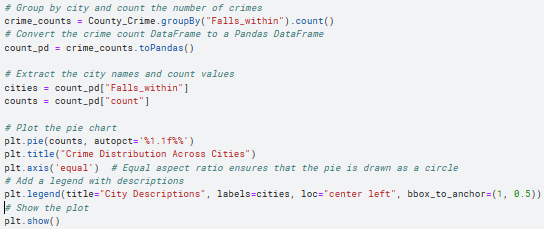
13 crime types are focused in the police data-source.

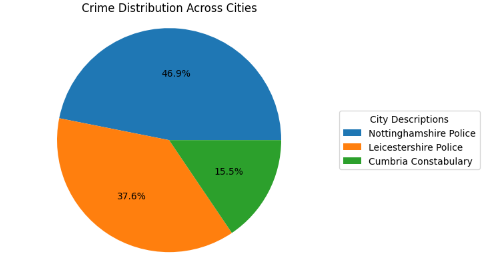
**To check if the crime categories are the same for each city?**





**Find percentage of crime rates across each area of jurisdiction and the highest crime rated city?**





Highest crime-rate city-Nottinghamshire

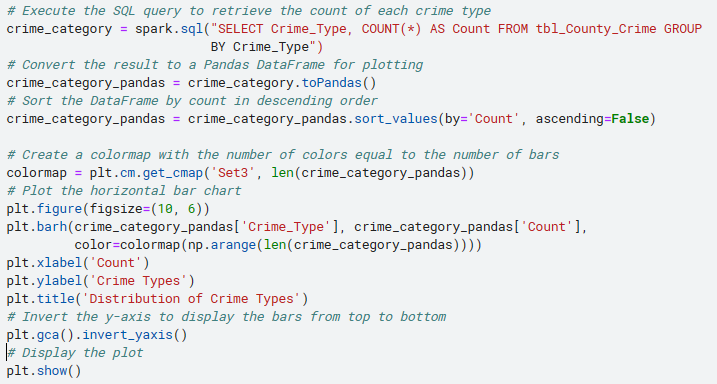
**How can we classify crimes based on severity and determine the total number of crimes in each severity level?**

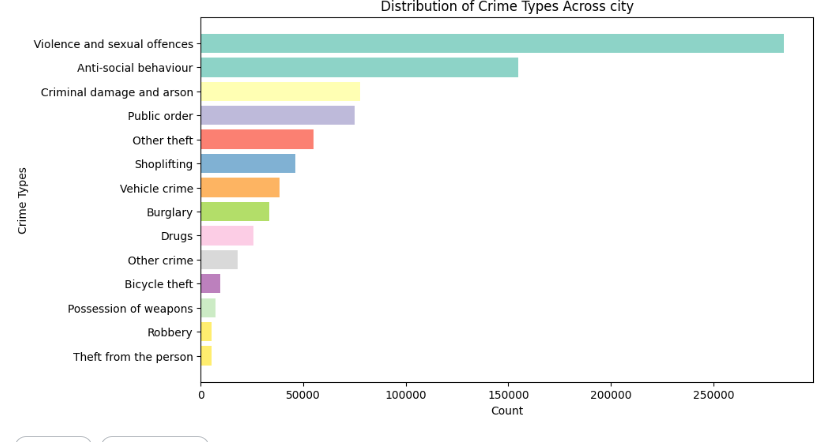


The number of severe crime 284564 were as the lowest severity crimes are 77846.

## 4.2 Crimes across and within cities

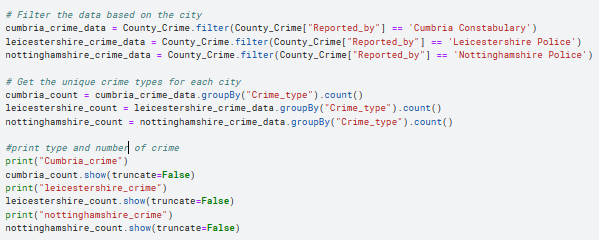
**Show the distribution of Crime across cities?**

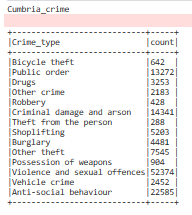
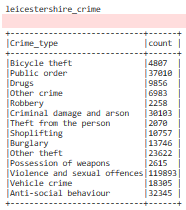
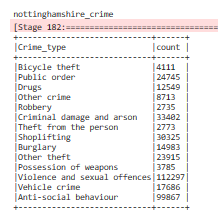




Here we can see that the highest crime type is **Violence and sexual offense**

**What is the total counts each type of crime recorded with cities?**

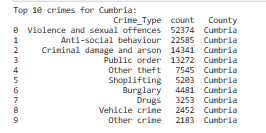
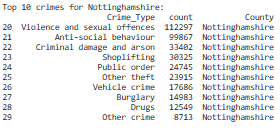
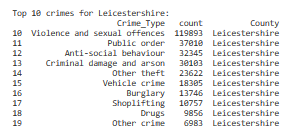


In cumbria highest crime type is fo Anti-social behaviour while for leicestershire and county nottinghamshire it is Violence and sexual offences and the lowest is Theft from the person, robbery,robbery for Cumbria, leicestershire and nottinghamshire respectively.

**Find Top 10 crimes across cities?**

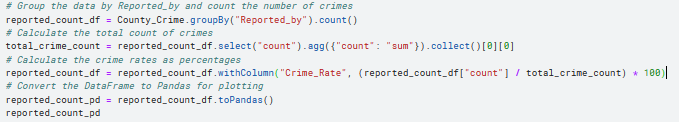
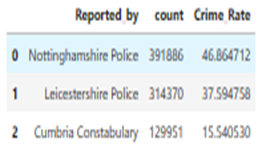


Highest crime is Violence and sexual offense which we have noticed earlier.

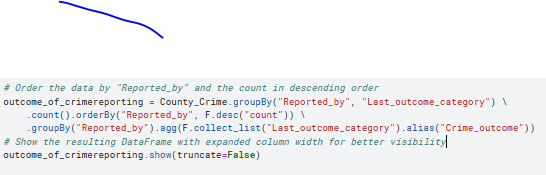
**Enforcement with highest reported crimes?**Highest crime-reported - Nottinghamshire Police

**Analyse crime rates across cities?**

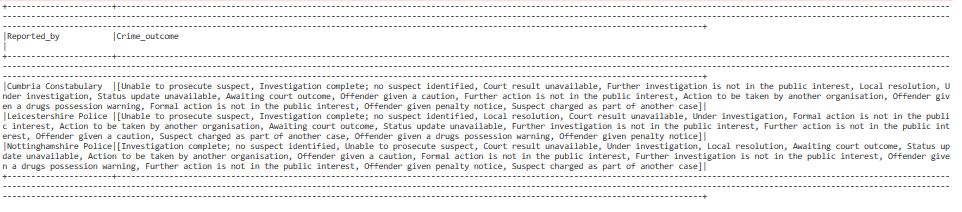
****

The reported crime rates are the highest for Nottinghamshire Police



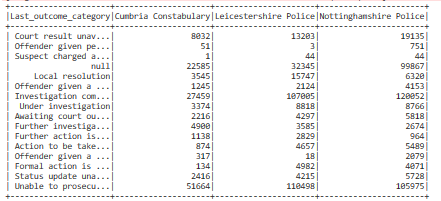
**Investigate outcomes across city?** 

From the result below we can see, that the outcome observations are almost same for all cities however the results vary considerably.



**How are the last outcome categories distributed in each city?**

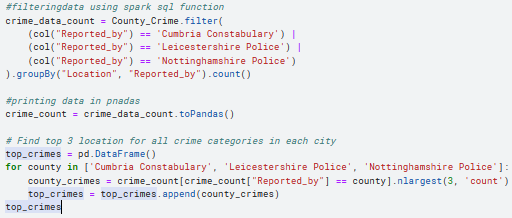




Distribution of the last outcomes is different for different cities. For Cumbria highest outcome category is ‘unable to prosecute suspect’ and least is ‘Suspect charged as part of another case’ whereas for Leicestershire, highest outcome is the same. However, the least outcome is ‘Offender given penalty notice’ and for Nottinghamshire Police highest is ‘Court results unavailable’ and lowest ‘Suspect charged as part of another case’.

## Analysing crimes on the basis of location

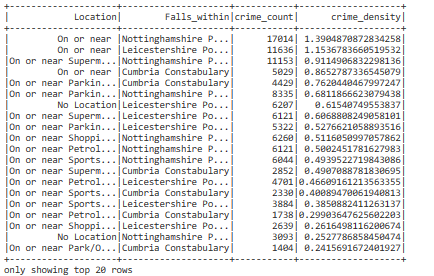
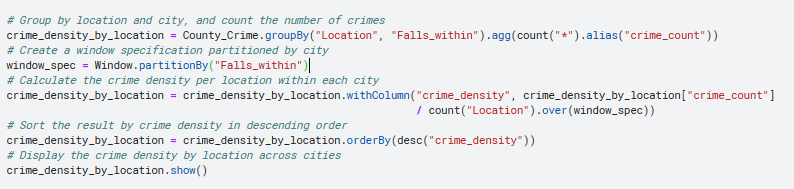
**Find the top 3 location of crime in each city?**



From graph we can understand, most crimes take place on or near streets with the Leicestershire leading by 11636 crimes followed by on or near supermarkets in Nottinghamshire by 11153 crimes and 3rd crime spot is on or near parking area in Nottinghamshire itself with 8335 crimes.

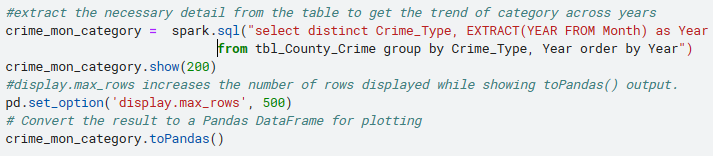
**Find the crime density across cities ?**



The highest crime density is for Nottinghamshire where the crime density is 1.39

## 4.4 Identifying the trends across the years

**Across the years have the same type of crime categories been used?**



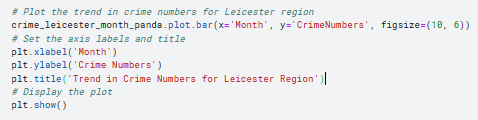
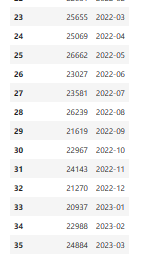
Same category of data used across all years.

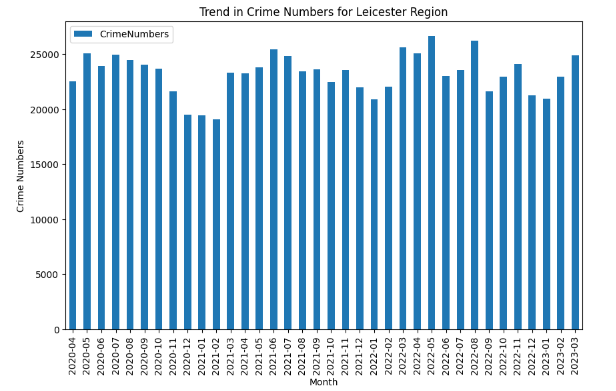
**The Leicestershire region was put to local lockdown in July 2020. Verify your findings (if any) with respect to that region only?**



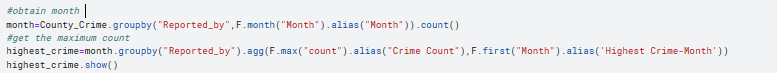


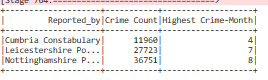
Extracting year inorder to plot the graph



 From the graph, it can be observed that there was a significant increase in crime in Leicester during the lockdown period in July 2020. The crime rate gradually decreased throughout the winter months and then started to increase again in April 2021. This indicates a fluctuation in crime rates during different periods, with a noticeable spike during the initial lockdown period and a subsequent increase in April 2021.

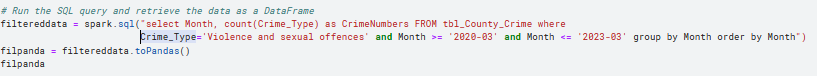
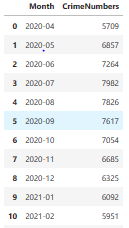
**Which month had highest crime reported across cities?**





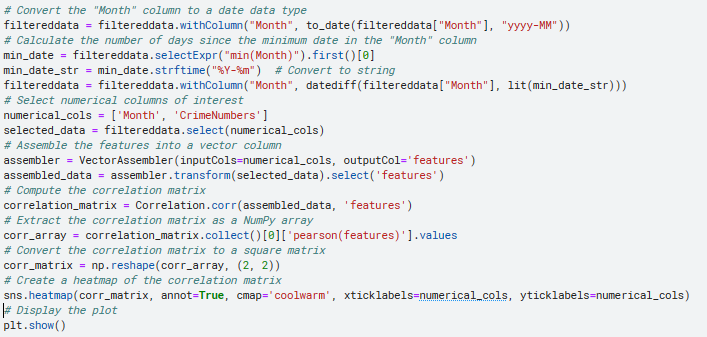
For Cumbria city the highest crime reported was in April while for Leicestershire it was July whereas for Nottinghamshire it was August.

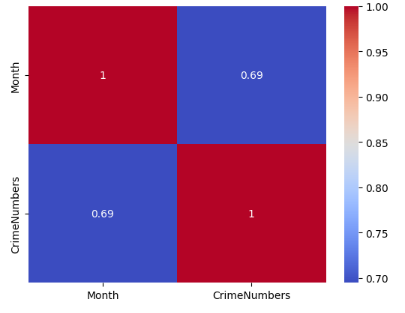
**Is there any correlation between months and crime?**



Filtering month and feature extracting crime number to get

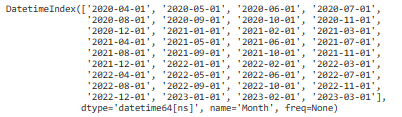
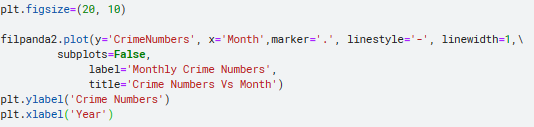
numerical for analysing purpose



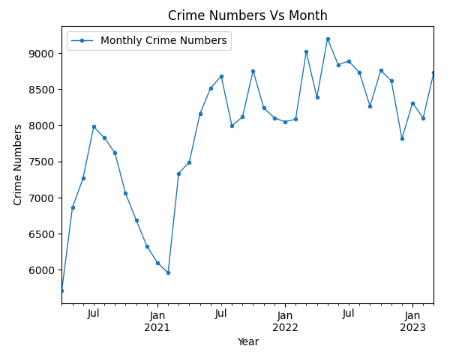
From the correlation matrix, the correlation coefficient appears to be a moderate positive linear relationship between the month since the minimum date and the count of violence and sexual offences. This suggests that as the month progresses there is a tendency for the count of crimes to increase as well. But this dependents on other factors or variables influencing relationship.



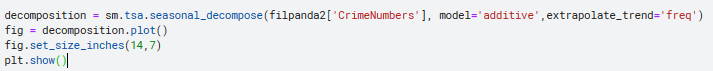
converting to date string for plotting

.

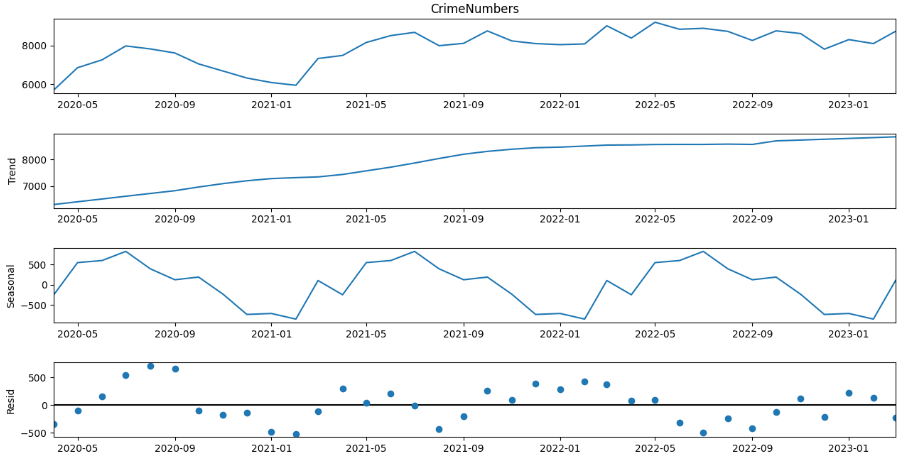
From the graph, it is evident that there has been a noticeable increase in crime over time. As we move from earlier months to later months, the crime count tends to rise consistently. This indicates a growing trend of criminal activities as the months progress.



**Are there any seasonal or temporal patterns in crime rates for each county? Analyze the crime rates over time to identify any recurring patterns or trends**.



Decomposition allows us to analyze the long-term direction, identify seasonal patterns, and examine any irregularities or anomalies in the data.



Based on the graph, decomposition analysis is used to break down the time series data into its underlying components: trend, seasonality, and residual.Tt is evident that the crime rates generally increase as the months progress. Additionally, there is a seasonal pattern where crime rates tend to decrease during the winter season and then show a steep increase during the summer season. This suggests a correlation between crime rates and the changing seasons, with higher crime rates observed during the warmer months compared to the colder months.

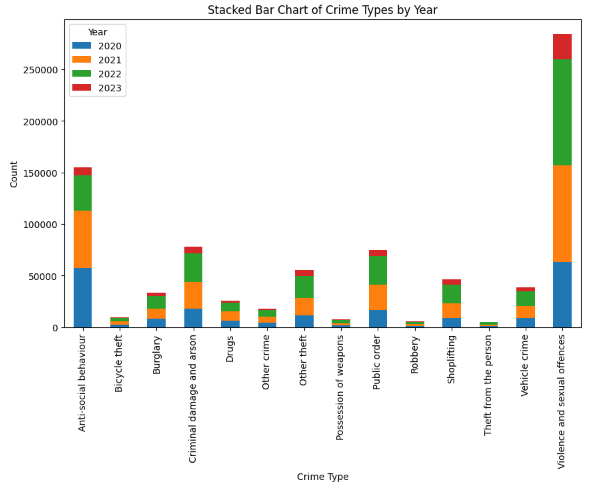
# Highlighting discoveries and findings:

* The highest crime type is violence and sexual offense, indicating the prevalence of such crimes in the dataset
* Nottinghamshire has the highest reported crime rate, suggesting a higher incidence of crimes compared to other cities in the dataset.
* There is a seasonal pattern observed in the crime rates, with a decrease during the winter season and a significant increase during the summer season. This indicates a correlation between crime rates and the changing seasons, with warmer months showing higher crime rates.
* There is a relationship between the months and crime, implying that certain months may have higher crime rates compared to others.
* The crime density is highest in Nottinghamshire, with a crime density value of 1.39, indicating a higher concentration of crimes in that area.
* During the lockdown period in Leicester, there was an increase in crimes in July 2020, suggesting a potential impact of the lockdown on crime rates in that city.
* Nottinghamshire has the highest reported crime count among all the cities, indicating a higher overall incidence of crimes in that jurisdiction.

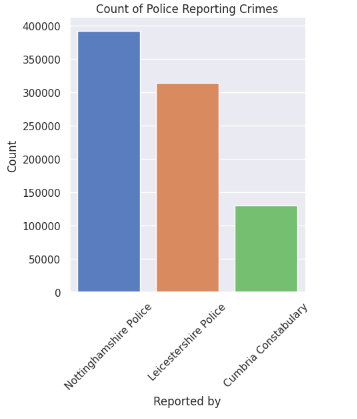
These highlighted points provide key insights into the crime trends, patterns, and characteristics observed in the given dataset.

# Results

1. **Across years same crime types used**

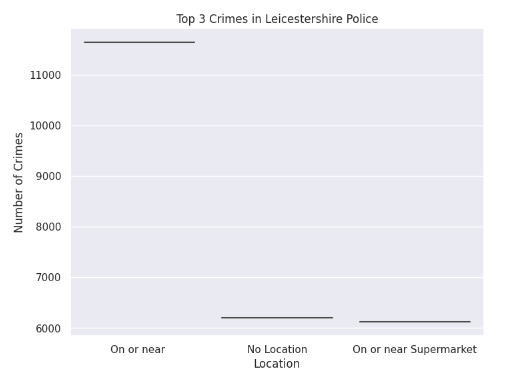
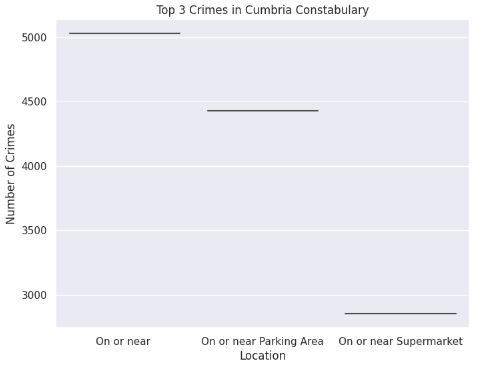


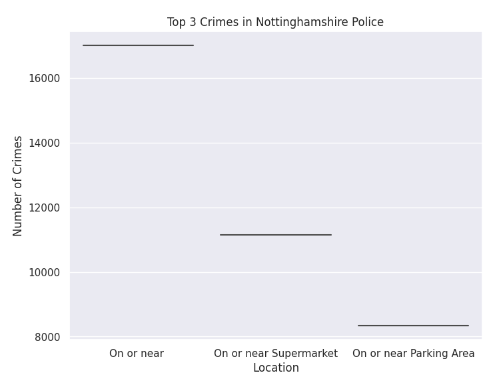
This indicates same category of crim types used with violence having the most reported across all years

1. **Enforcement trend with respect to crime reporting**

The highest crime reporting is by Nottinghamshire Police whereas the least by Cumbria constabulary

1. **Based on location top 3 crime spots within each city**

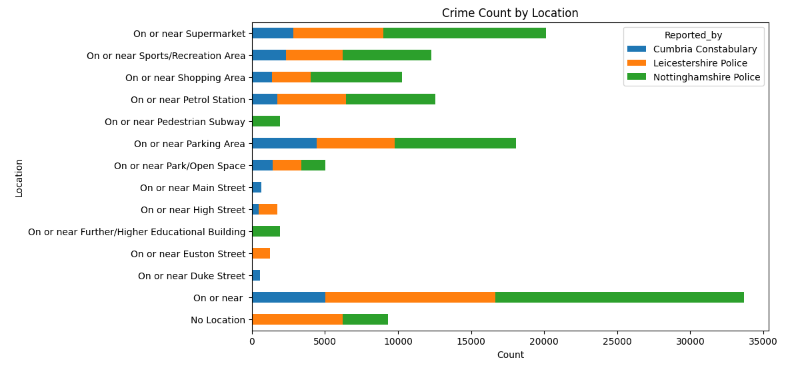




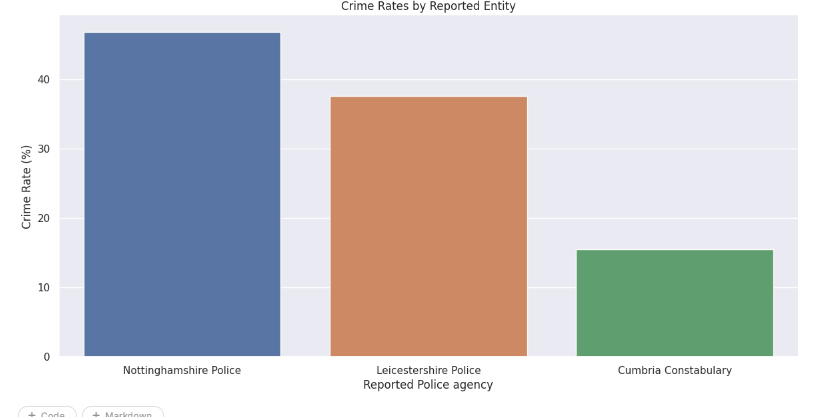
Top 3 crime locations for each city are:

**Leicestershire:** One or near, No location, On or near supermarket.

**Nottinghamshire**: On or near, One or near supermarket, On or near parking area.

**Cumbria:** On or near, On or near parking area, On or near supermarket.

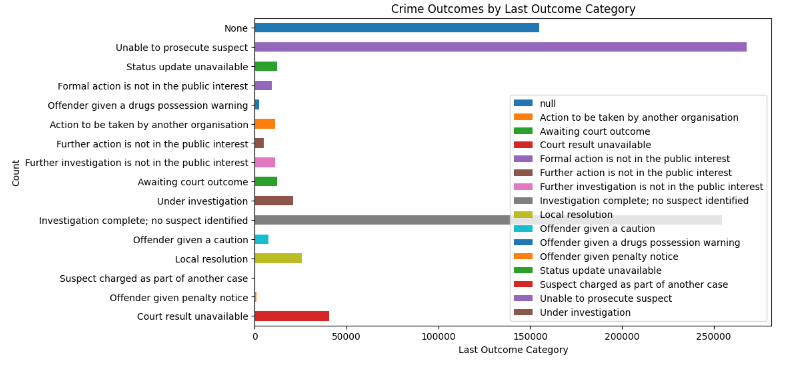
1. **The across city highest police reported crime- rates**



Nottinghamshire Police has highest crime-rate reporting % and Cumbria has the least % reporting.

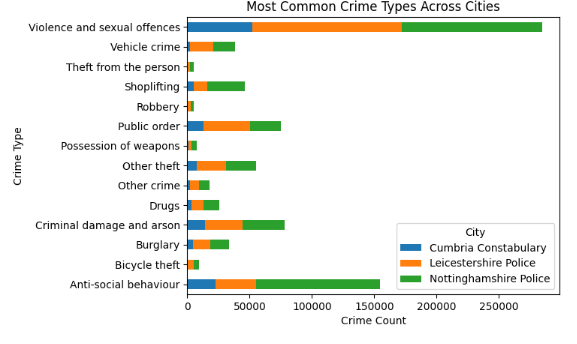
1. **Outcome distribution across city**





Across the cities, the highest outcome rate for reported crimes is "Investigation complete; no suspect identified."

1. **Common-crimes across cities**



From the graph, it is evident that the crime types are consistent across all cities, with the highest frequency observed for "Violence and sexual offences." The second and third highest crime types are "Anti-social behaviour" and "Public order" respectively.

# Conclusions

After the analysis of the police dataset records from Cumbria, Nottinghamshire, and Leicestershire, we can understand that: The dataset provides valuable insights into the types of crimes reported in the respective cities, with "Violence and sexual offenses" being the most prevalent category. The analysis reveals that the crime types are consistent across all cities, indicating a similar pattern of criminal activities. Nottinghamshire has the highest reported crime count, indicating a higher incidence of crimes compared to Cumbria and Leicestershire. After analysing crime on the basis of Location, we understand: top three crime spots within each city are a: "One or near," "No location," and "On or near supermarket" for Leicestershire; "On or near," "One or near supermarket," and "On or near parking area" for Nottinghamshire; and "On or near," "On or near parking area," and "On or near supermarket" for Cumbria. On Identifying Trends Across the Years, we notice that that the crime types remain consistent over the years, with "Violence and sexual offenses" being the most reported category throughout the dataset.

In conclusion, the analysis provides valuable insights into crime trends, prevalence, and patterns within the cities of Cumbria, Nottinghamshire, and Leicestershire. It improves our understanding of crime data analysis, crimes across and within cities, crimes based on location, and trends observed over the years. These findings contribute to a better understanding of crime patterns and can assist in developing effective strategies for crime prevention and law enforcement.

# Reference:

[1]. *User guide s*(no date) *User Guides - PySpark 3.4.0 documentation*. Available at: <https://spark.apache.org/docs/latest/api/python/user_guide/index.html>

[2] *Crime analysis and clustering using geoanalytics and pyspark.ml* (no date) *ArcGIS API for Python*. Available at: <https://developers.arcgis.com/python/samples/crime-analysis-and-clustering-using-geoanalytics-and-pyspark/>

[3.]Classroom lectures and notes

# Appendix

## Appendix A: Features

