**Introduction:**

Crime analysis is the process of reviewing data collected and/or stored in order to identify trends of criminal activities so that law enforcement can effectively adjust to deter that crime and continue to protect the public’s safety. The process of crime analysis involves five basic steps: collection, categorization, analysis, dissemination, and evaluation of information (S.R Stiles 1981). For this project, we will focus on the analysis stage of the crime analysis process, using PySpark to analyze over 65 million dataset of United Kingdom crimes from the years 2010-2021.

**Approach:**

The dataset was analyzed by grouping across the month and year to determine the trend of the number of cases reported between 2010 and 2021. The dataset was filtered by crime types and the fall within (cities) to see the count of reported cases between these two instances. Additionally, the top 5 crime types and top 5 fall within (cities) with the highest crimes were identified. Each of the top 5 fall within was analyzed to establish the relatedness between the crime types and fall within (cities).

**Data overview:**

The data used in this project is from the United Kingdom, and includes data for crimes from the years 2010-2021. The first step was to understand the dataset and ensure that it includes all necessary columns of information and is correctly formatted and of good quality. The dataset and other necessary dependencies were imported into a PySpark DataFrame, and selected columns of interest were extracted for further analysis.

**Methods:**

To perform the analysis, PySpark's SQL module and functions were used to group and count the number of crimes committed for each case type and fall within. The data was also transformed to extract the month and year from the 'timestamp' column for trend analysis. Matplotlib was also used for data visualization.

**Results:**

The results of the analysis showed the top 5 crime types and top 5 fall within (cities) with the highest crimes. Additionally, the trend of the number of cases reported was analyzed by grouping across the month and year, and relatedness between the crime types and fall within (cities) was established.

From 2010 to 2021, the crime rate reduced towards 2021.

A picture containing text, clipart

Description automatically generated

4 out of the identified top 5 crime types were consistent with the top crimes in the top 5 cities with the highest crimes. This suggests that there is a correlation between crime types and specific locations, and law enforcement may want to focus their efforts in those areas where specific crimes are more prevalent.

Conclusion:

PySpark was successfully used to analyze a dataset of United Kingdom crimes from the years 2010-2013. The results of the analysis provide valuable insights for law enforcement to adjust their strategies for reducing crime and ensuring public safety. The correlation between crime types and specific locations can aid in the development of targeted crime prevention programs and increase the effectiveness of law enforcement efforts. Additionally, the decreasing trend in crime rate over the years can also be considered as a positive outcome, indicating that the efforts of law enforcement have been successful in reducing crime. Overall, this analysis can be used as a valuable tool for crime analysis and decision-making in the future