WQD7003 – Data Analytic

Group Assignment

Analysis and Prediction of Crime Statistic in London.

By

Gunasegarran Magadevan - WQD170002

Mathavan Chandrasegaram - WQD17007

4W 1H Question

- **What** :

  This is an analytical study to categories the crime and also to produce insight info to prevent or reduce the crime in future.

- **Why** :

  To produce valuable and useable insight from historical raw data which may helps to prevent crimes which mostly people get involved

- **Where** :

  Raw data of criminal report in London City (UK)

- **When** :

  Criminal record from Jan 2008 - Dec 2016

- **How** :

  Use few analytical tools to process the raw data and come out with useful insight and prediction.

Overview

- We have use **data of criminal records** happening in **London city** during Ja**n 2008 until 2016** to analyze and produce some useful insight from the raw data.
- We have categories the crime by **month**, **year**, **severity of crime**, type of most happening **crime** in city, etc.
- This help to focus on the dangerous crime and prevent it moreover it could bring awareness among people around the city.
- We also produce **visualization** of crime based on **boroughs** (district) of the city and the population against it.
- Hence, **graph** has been **plotted** on **minor crimes against major crimes** to show each major category crime among its minor category crime.

Motivation

- Currently there are few bodies official and nonofficial teams are working to streamline the crime management processes and to improve current method of crime management.
- As such, the outcome of this analysis will enable them to focus on specific problems or area and formulate the targeted solutions

Objective

- Aim to produce useful insight by using **clustering data analytical.**
- To **visualize** crime rate by **borough**, **major category**, **minor category** and **populations.**
- To summaries the top categories of **crime by borough** and **identify top problem spots**.

Dataset

The dataset is consisting by 7 variables:

- **lsoa_code**: LSOA in London (United Kingdom)
- **borough**: borough (district) names of in London (United Kingdom)
- **major_category**: categorization of high level crime
  minor_category: categorization of low level crime
- **value**: monthly reported count of categorical crime in given borough (district)
- **year**: year of reported counts, 2008-2016
- **month**: month of reported counts, 1-12 (January-December)
- *The variables **lsoa_code**, **borough**, **major_category**, **minor_category**, **year** and **month** are **categorical** variables, while value is a **discrete numerical** variable.*

3. Data Insight

```
print("Counts of data : ",ds.count())
print("\n\n")
print("Top 5 data  : ",ds.head())
```

```
Counts of data :  lsoa_code         13490604
borough           13490604
major_category    13490604
minor_category    13490604
value             13490604
year              13490604
month             13490604
dtype: int64



Top 5 data  :     lsoa_code     borough               major_category  \
0  E01001116     Croydon                        Burglary
1  E01001646   Greenwich  Violence Against the Person
2  E01000677     Bromley  Violence Against the Person
3  E01003774   Redbridge                      Burglary
4  E01004563  Wandsworth                       Robbery

                minor_category  value  year month
0  Burglary in Other Buildings      0  2016    11
1              Other violence       0  2016    11
2              Other violence       0  2015     5
3  Burglary in Other Buildings      0  2016     3
4            Personal Property      0  2008     6
```

Data Cleaning

From the dataset, we used the used the fastest way to identify determine if **ANY** value in a series is missing. Therefore the finding is the data are clean.

```
In [6]: ds.isnull().values.any()
Out[6]: False
```

Quantitative Variables Analysis :

1. Since **247** unique values of the dataset's samples have the variable **value** equals to **0**.
2. To conclude, the window of time from **2008** to **2016** was not too compact of criminal activities.

Quantitative Variables Analysis :

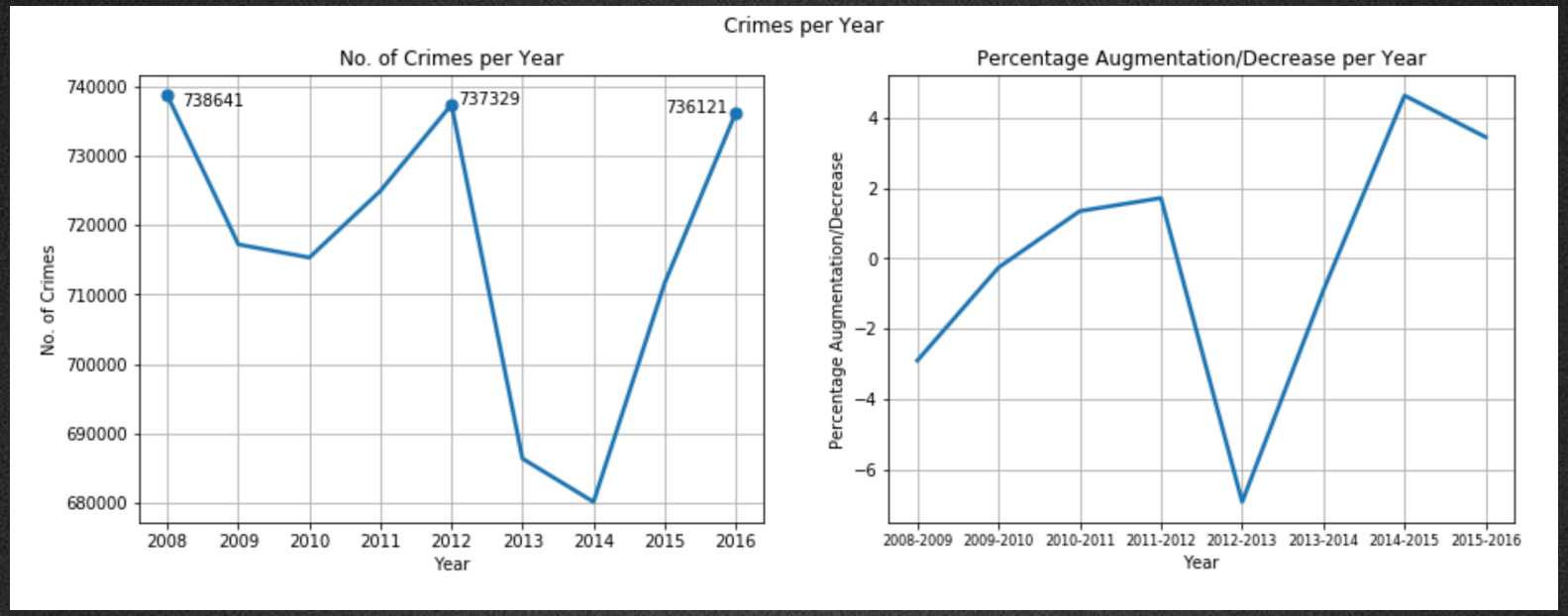
```
In [7]: num_summary = ds.describe(include=np.number)

In [8]: print('MIN: {}, MAX: {}, UNIQUE VALUES: {}, MODE: {}'.
              format(int(num_summary['value']['min']),
                     int(num_summary['value']['max']),
                     ds['value'].unique().shape[0],
                     stats.mode(ds['value'])[0][0]))

        MIN: 0, MAX: 309, UNIQUE VALUES: 247, MODE: 0
```

1. Since **247** unique values of the dataset's samples have the variable **value** eguals to **0**.
2. To conclude, the window of time from **2008** to **2016** was not too compact of criminal activities.

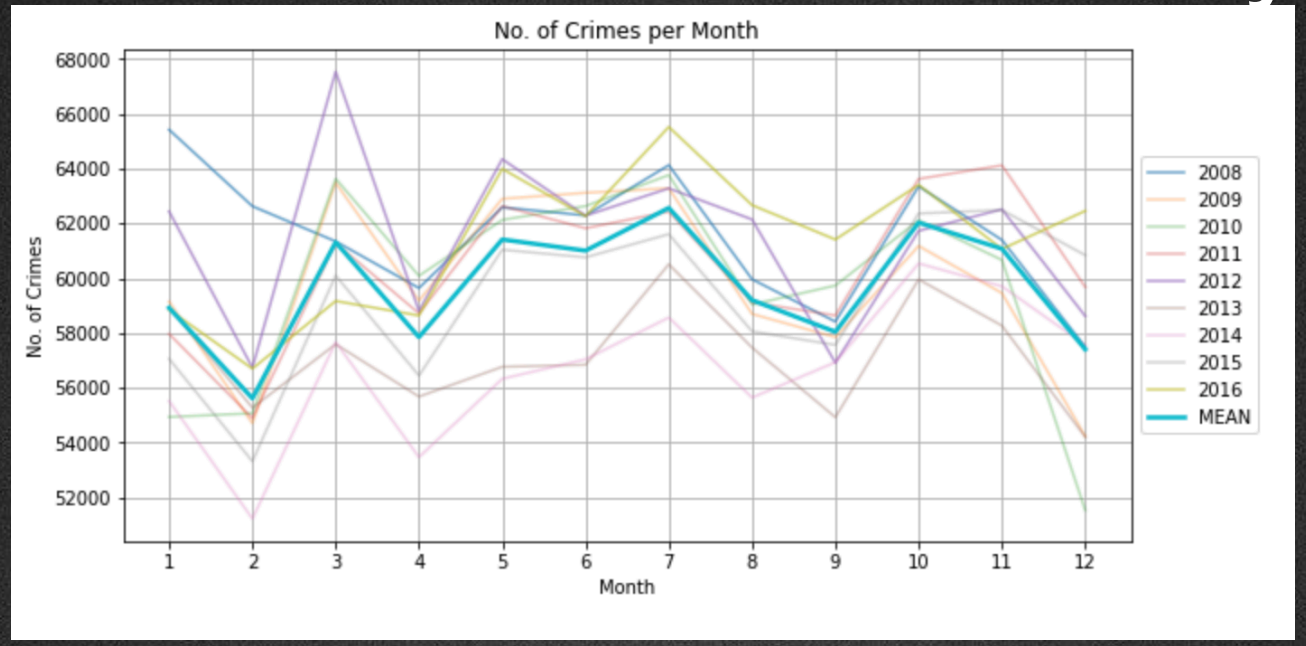
Quantitative Variables Analysis :



- The figure above represents the flow of criminal activities on a by yearly basis:

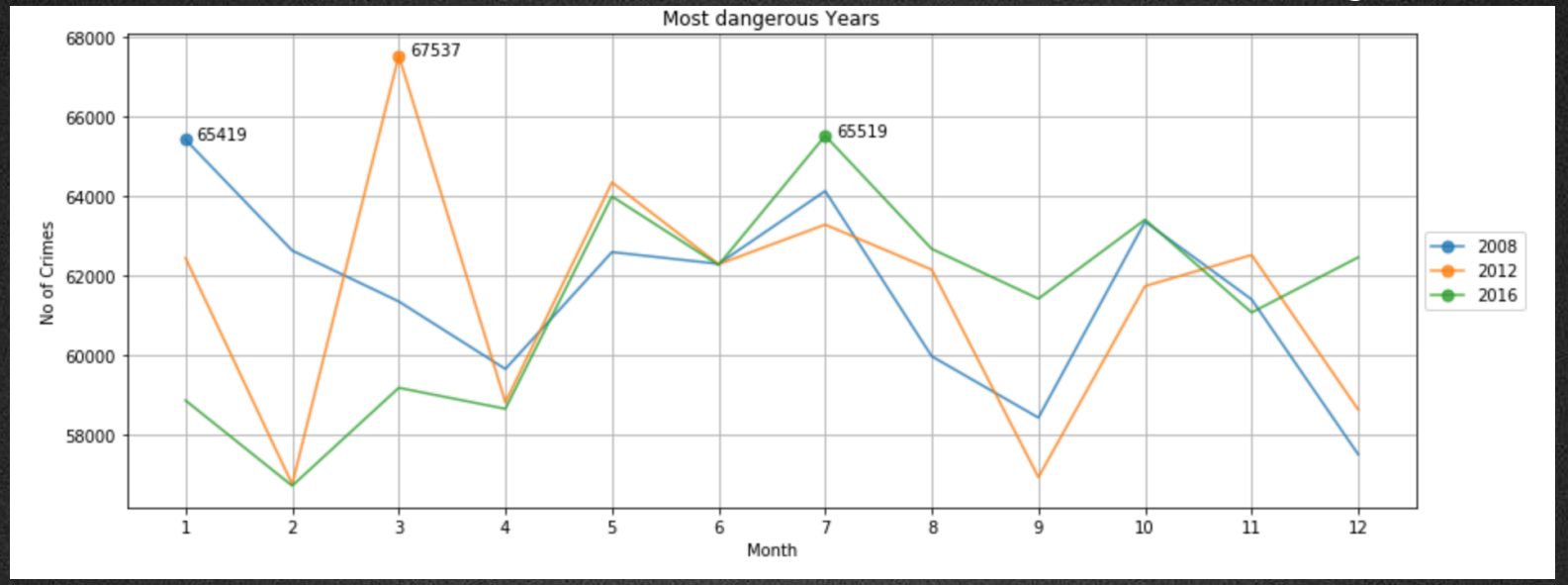
1. The most criminally decrease year are by year **2008**, **2012** and **2016**.
2. The most peaceful year are **2014** and **2013**.

Quantitative Variables Analysis :



- The figure above represents the flow of criminal activities on a by month basis:
  1. Observing a behaviour that remains coherent with the flow of criminal activities on a by yearly basis.

Quantitative Variables Analysis :

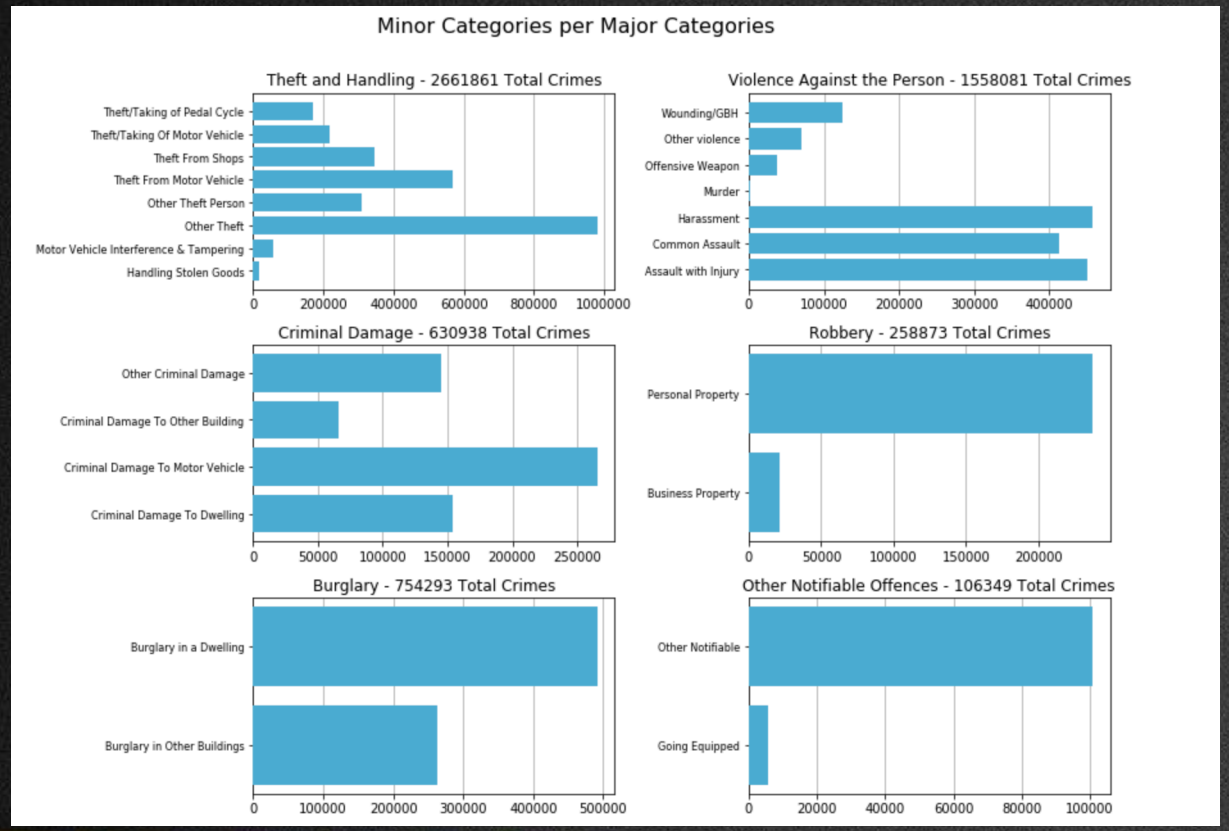


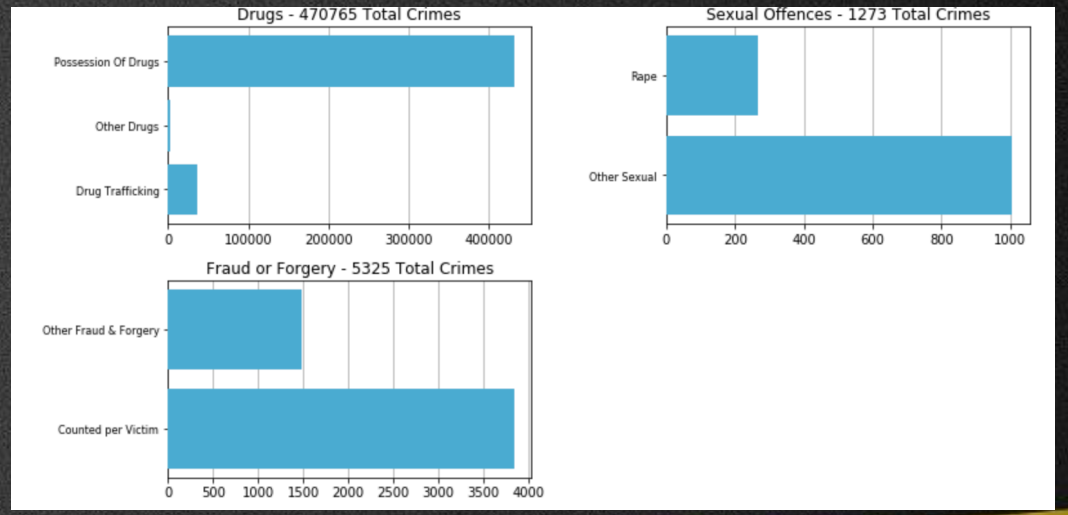
- The figures above shows the flow of criminal activities on a by month basis for the most decrease years:
  1. By looking at the criminal activities represented in this graphs like a flow, it is unique that the amount of criminal reports have the tendency to increase once every four years.

Categorical Variables Analysis :

1. The year **2016** rise aside from numerical analysis.
2. Despite being the least decrease of criminal activities in the top three represented by the years, in descending order, **2008**, **2012** and **2016**, is the one that owns the majority of the records in the cropped dataset.
3. It means that, remaining coherent with what rise in the numeric variable's analysis, it has the lower crime per month ratio among the three.

Categorical Variables Analysis :

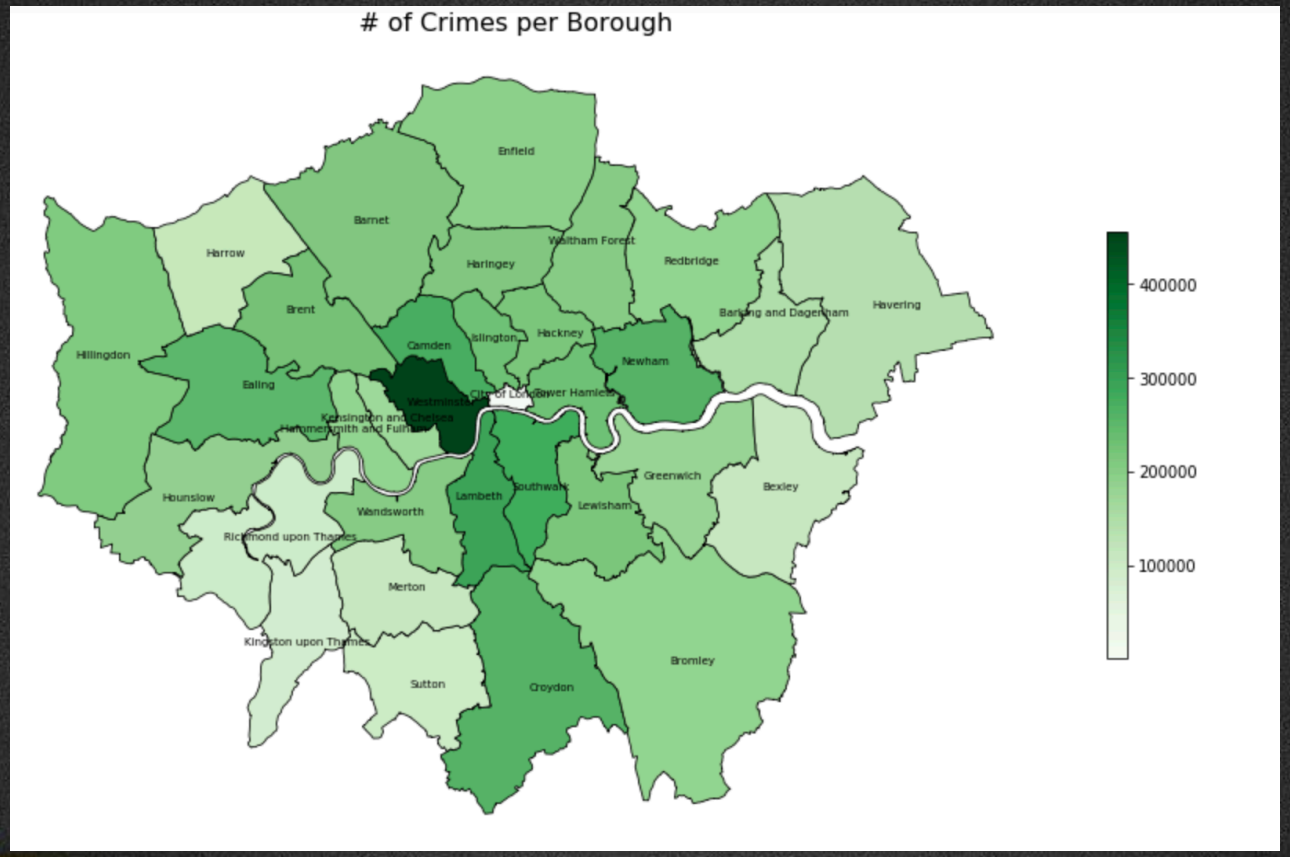




Categorical Variables Analysis :

1. The **minor category crimes** classification is very rich, with Theft and Handling being the most diversified with eight minor categories.
2. The subclass of the total number of criminal activities for each major category crime among its minor categories.
3. By observing the graphs it is possible to extract the most frequent minor category for each major category:
   I. Theft and Handling -> Other Theft
   II. Violence Against the Person -> Harrasment
   III. Criminal Damage -> Criminal Damage To Motor Vehicle
   IV. Robbery -> Personal Robbery
   V. Burglary -> Burglary in a Dwellingv
   VI. Other Notifiable Offences -> Other Notifiable
   VII. Drugs -> Possession Of Drugs
   VIII. Sexual Offences -> Other Sexual
   IX. Fraud or Forgery -> Counted per Victim

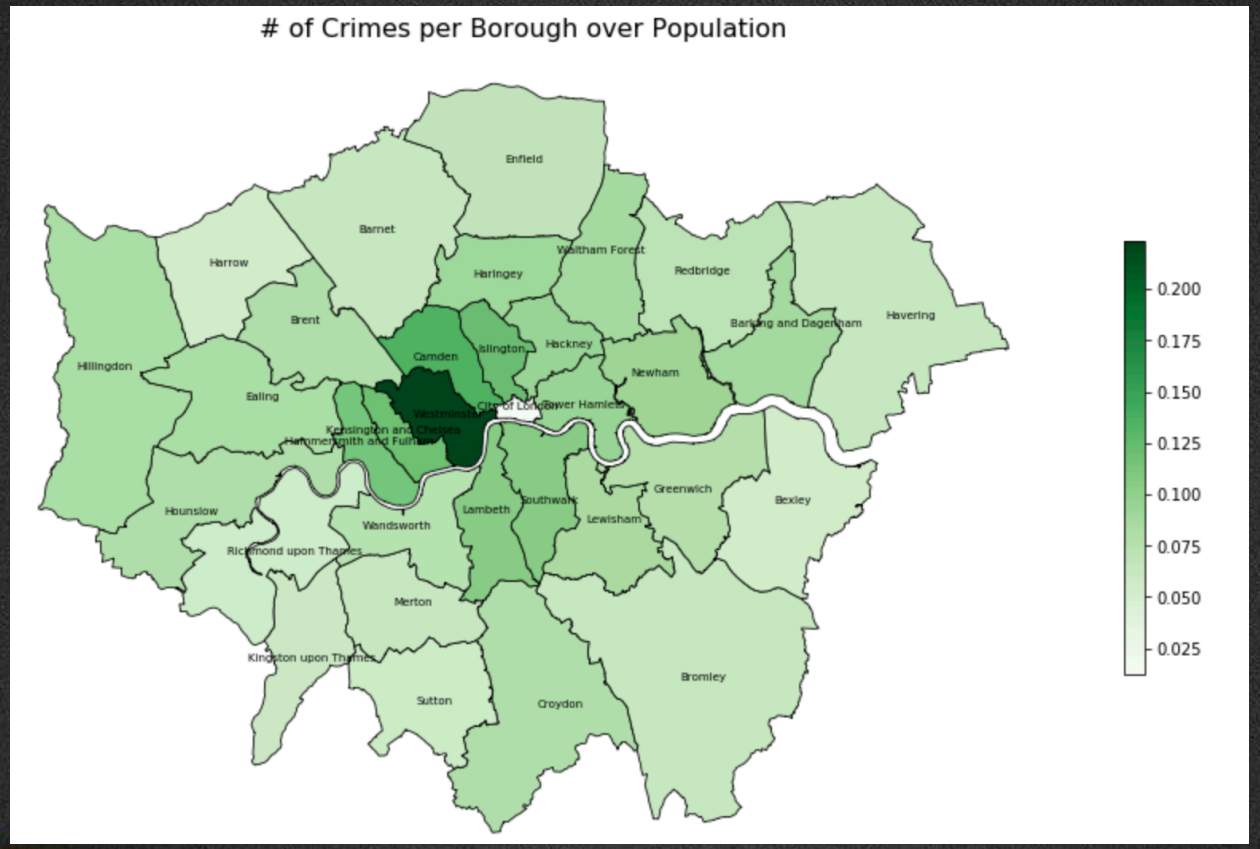
Categorical Variables Analysis :



Categorical Variables Analysis :

1. From the geographic visualization proves what has been discovered so far.
2. Westminster is confirmed as the most decrease of criminal activities among the boroughs, while City of London is confirmed as the least dense of criminal activities.
3. The naive assumption that there is no correlation between the number of crimes committed during the window of time proposed by the dataset and the **boroughs** territorial extension.

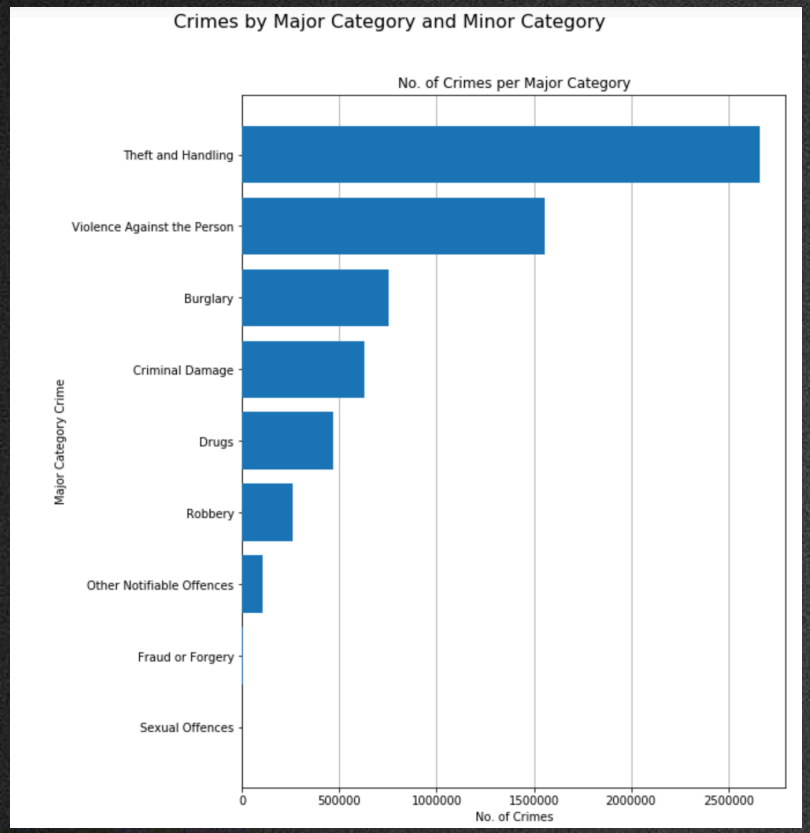
Categorical Variables Analysis :

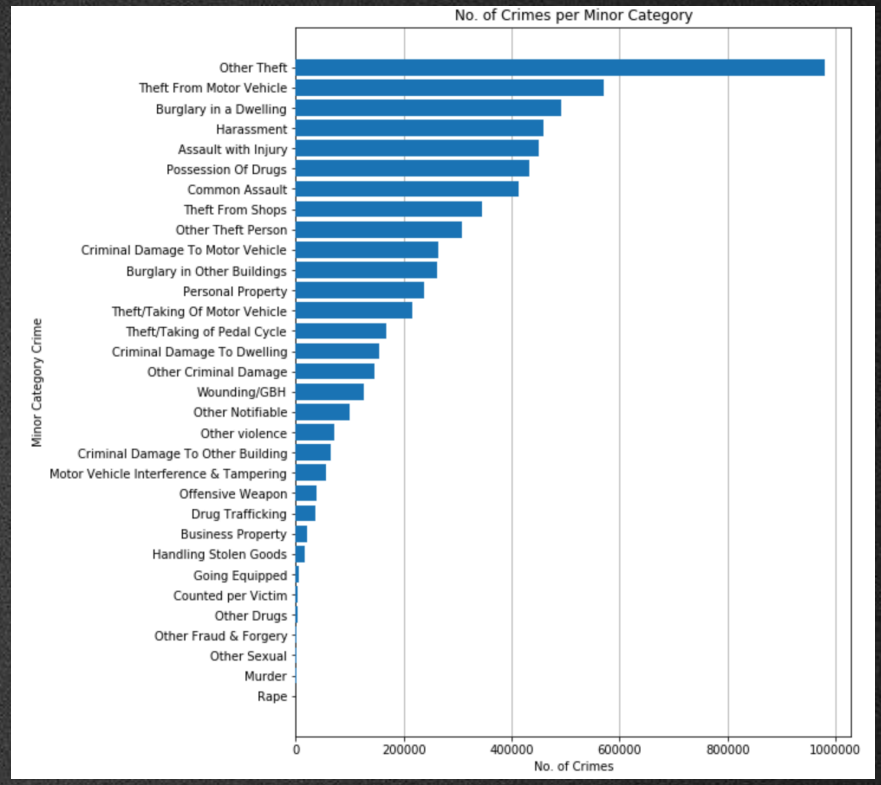


Categorical Variables Analysis :

1. This visualization shows a general very low score for the ratio between the number of crimes committed in a district and its population.
2. This means that, for a certain window of time, the number of criminal activities are fewer than the population density - in a way confirms the fact that the period of time investigated by the dataset is a quite safe window of time.

Categorical Variables Analysis :





Categorical Variables Analysis :

1. Despite being Lambeth the most popular borough among the cropped dataset's records, the most dangerous is actually Westminster, as depicted in the visualizations.
2. Theft and Handling is the most frequent major category crime and Other Theft is the most frequent minor category crime.

Correlation

	lsoa_code	borough	major_category	minor_category	value	year	month
lsoa_code	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Independent
borough	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
major_category	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
minor_category	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
value	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
year	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent
month	Independent	Dependent	Dependent	Dependent	Dependent	Dependent	Dependent

1. The results returned by the correlation analysis are not surprising as expected.
2. The dataset is composed by a set of variables that are all **depending** on each other.
3. In the correlation table above, the majority of variables have a relation with the other variables that can be classified as **dependent**, while the variables **lsoa_code** and **month** are classified as **independent**.

Conclusion

1. Lambeth the most popular borough among the cropped dataset's records.
2. The most dangerous is actually Westminster, as depicted in the visualizations.
3. Theft and Handling is the most frequent major category crime and Other Theft is the most frequent minor category crime.
4. The variables in datasets are all depending on each other, the majority of variables have a relation with the other variables that can be classified as dependent, while the variables lsoa_code and month are classified as independent.