Project Report for Principles of Data Science project

Created by: Karunakar Kotagaram

1. Introduction

London as a city vary greatly as we move between the different boroughs and wards. The purpose of this report is to present a picture of the crime statistics in London. Our objective is to investigate the socio-demographic, census, economic, temporal and spatial factors that influence the crime rate in London at borough level.

One of the problems associated with crime data is that it paints a biased picture and may lead to misinterpretations. Data doesn't represent all the crimes occurring in the society as all the crimes are not reported. Crimes like 'rape' is commonly not reported for many reasons. So the crime rate for number of rapes in a certain area may not portray the true figures. Domestic violence is more common among certain races and is mainly not reported. Another factor is that some crimes are not reported instantly. Burglaries taking place during the holidays like Christmas period or Easter break are reported when people get back from holidays and find out what happened in their absence. Therefore, reporting date differs from the date when the incident actually happened.

Crime analysis is defined as:

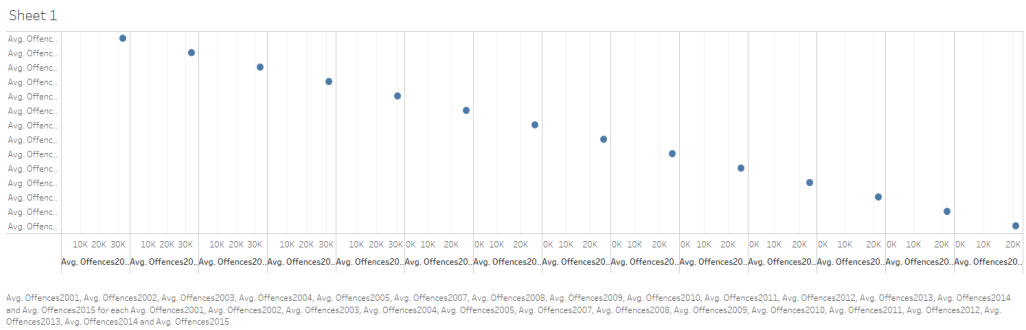
"*The systematic study of Crime and disorder problems as well as other police-related issues including socio-demographic, spatial and temporal factors-to assist the police in criminal apprehension, crime and disorder reduction, crime prevention and evaluation"*.

DR. Rachel Boba (Florida Atlantic University).

We will focus on the factors contributing to the crime rate, rather than just the numbers and figures of different types crimes in Boroughs of London. First we will identify the factors that have an impact on the higher/lower crime rates. Employment/Un-employment rate, Youth un-employment rate (18-24), Working-age population who claim out-of-work benefits, Gross annual income, Number of cars per household, Average age, Number of ambulance incidents, Average Public Transport Accessibility score.

We have following research question: **Does crime rate increase with the increase in the ease of public transport accessibility?**

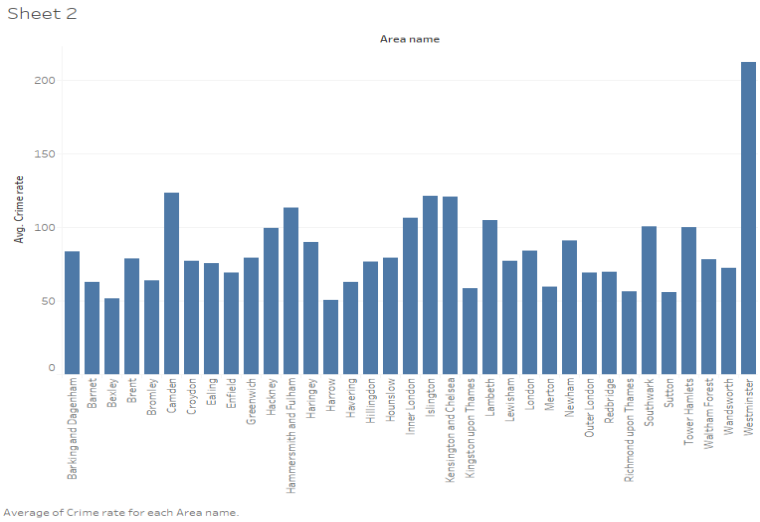
AS we can see in the figure 1 below, crime rate (average offences) has dropped in the past fifteen years in London. This figure was obtained by merging Police record data set with London Borough Profiles data set in Tableau. We plotted average offences recorded from 2001-2015 and we can see the drop in crime rate.



2.Initial Data Discovery

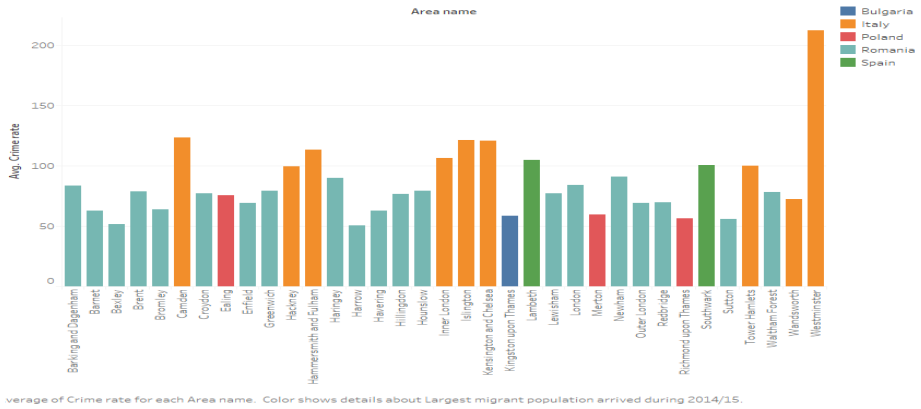
The data set under consideration is collected from London Datastore. 'London-borough-profiles' is a CSV file and it has 32 rows and 86 columns ,76 of which are numeric. We have performed the analysis in Python using Pandas and Numpy. Seaborn , Python library for statistical data visulaisations has been used for as well. For initial data discovery we have taken advantage of Tableau to extract valuable insights using data visulaizations.

We start with data discovery in Tableau. In the figure(2) below Average crime rate in each borough is plotted. We spot the boroughs with high crime rate. Wesminster stands out as the borough with the highest crime rate in London. This gives us an early information about the outliers in the data. We will investigate the outliers further in section().tourist area and lots of people visit it through out the year. Public transport is also widely used in this area. These two reasons make it prone to crimes like theft(wallet thefts), snatch and run, violence against person, and traffic crimes.



Figure(2) average crime rate for each Borough of London

Below is a visualization produced in Tableau figure() that shows average crime rates for each borough. Another interesting feature in the figure is the colours of the bars that is done according to the largest migrant population arrived in 2014/2015. This shows that the boroughs with the highest crime rate have biggest number of Italian migrants(arrived in 2014/2015). Migrant population and Ethnic origins of the population is another interesting way to analyse crime data but that is out of the scope of this report as we focus on public transport accessibility and crime rate.



Figure(3) Average Crime rate for each area name, colours show the migrant population entering London in 2014/15

It is believed that the *"crime rates tend to flow with the number of youth at a given time in the area"*. Figure() below is inspired from the above stated 'age distribution theory', average crime rate for each area name and

colours indicate the proportion of workin age population.It is evident from the figure(4) that areas like Camden, Islington, Lambeth, Hackney, Hammersmith and Fulham, Haringey, Kensington and Chelsea, Newham, Southwark, Tower Hamlets and Westminster have high crime rates and higher proportion of working age population. Another point is what proportion of this working age population is employed, there is a high crime rate in areas with higher proportion of people claiming out of work benefits.

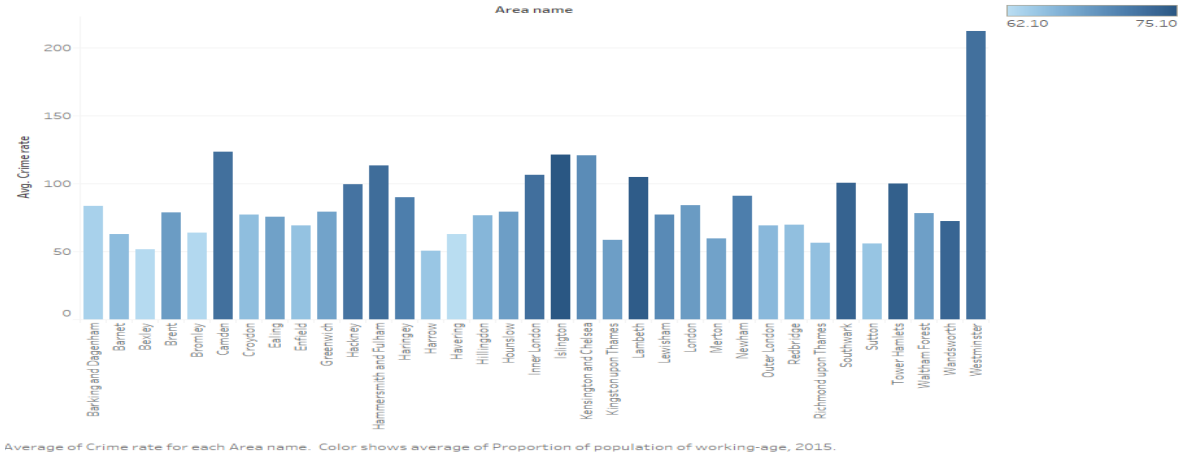
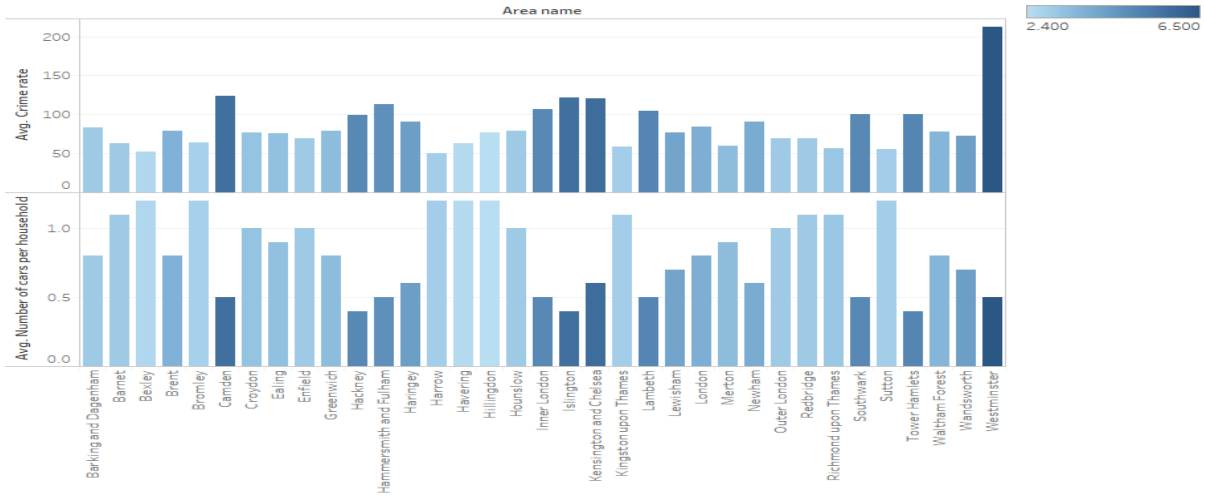


Figure (4) Average crime rate for each area names, colours indicate the proportion of working age population.

A very interseting way to look at the relatoinship between average crime rate and public transport accessibilty score is to look at the average number of cars per household in the Boroughs of London. Average crime rate is inversely related to the number of cars per household.



Figure(5) average crime rate and average number of cars per houshold for area names.Colours indicate the public transport accessibility score.

2. Research Question:

Does the crime rate increase in an area if the public transport is easily accesible?

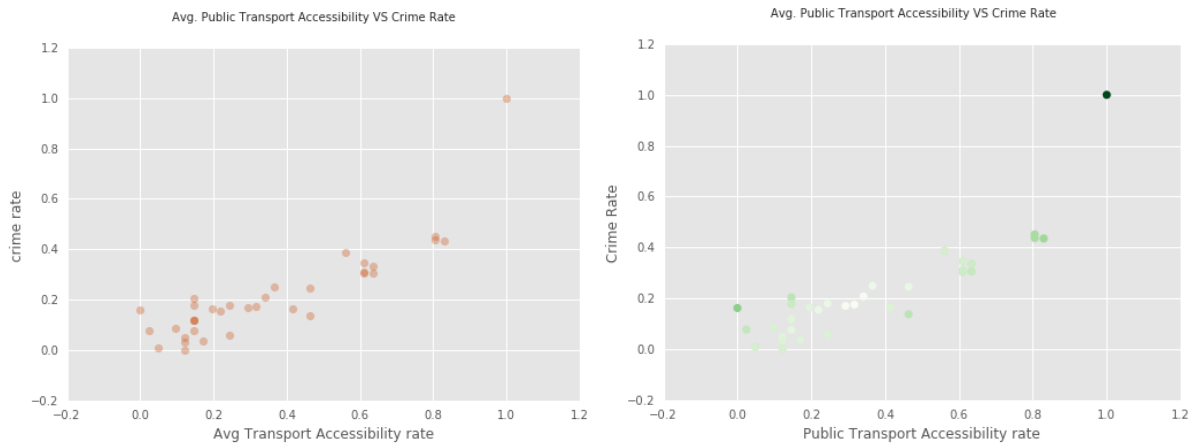
2.1 Pre-processing

Missing values in the numerical data always cause problems for data analysts. Firstly, we fill the missing values with the mean of that particular column. Secondly, the values in the dataset vary widely so we apply **standard normalization** to bring it to a common scale.Now the data has a zero mean and unit variance. When calculating Euclidean distance if one of the features has a broad range of values, it governs the distance matrix for MDS(Multi Dimensional Scaling). When applying clustering techniques it is easier to compare the data if it is normalized.

2.2 Detecting Outliers

Some areas have properties that make them more attractive for specific types of crimes and therefore, we get skewed data and outliers. Tourist area, Westminster has the highest crime rate and is the outlier in our dataset. Some areas have properties that make them more attractive for specific types of crimes and therefore, we get skewed data and outliers. Tourist area, Westminster has the highest crime rate and is the outlier in our dataset. When we investigated further into this problem found out that theft rate (wallet thefts), snatch and run, violence against person and traffic crimes are highest in this area.

We applied **Mahalabonis distance** to measure outliers. As stated earlier, our data has been scaled to unit variance and zero mean so Mahalabonis distance corresponds to Euclidean distance in transfromed space. Dark green spot in the figure(6) is our outlier Westminster. Westmintser stands out as an outlier because of it's higest score for ease in accesibility of public transport and crime rate.



Figure(6) Scatter plot of average public transport accessibility and crime rate. Green colour mapping shows the outliers.

2.3 Correlation

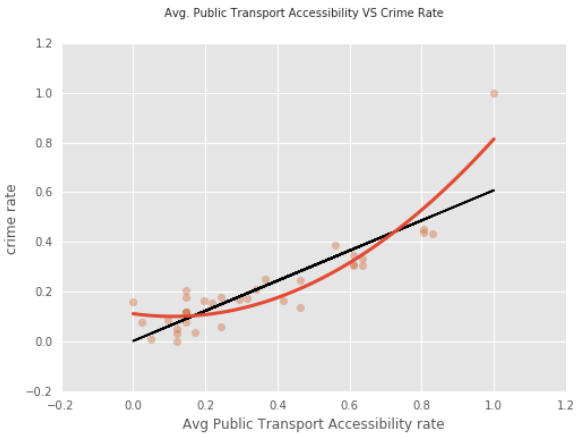
To investigate the relationship between Average public transport accesibility and crime rate we find the correlation between them. We form a hypothesis that the public transport accessibility and crime rate are positively correlated. Results prove that our hypothesis is right. Crime rate increases with the increase in public transport accessibility because there are more chances of mobile phones or other valueables snatch and run, wallet thefts, violence against person in or around the train stations and at the bus stops.

High **Pearson Correlation** value of 0.866220756494 shows the strong linear relationship between the two variables. The scatter plot in figure() below shows the increase in crime rate with the increase in accesibility of public transport in London.

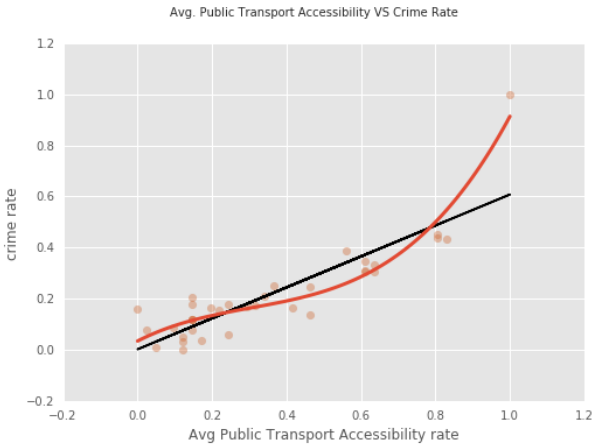
2.4 Linear Regression Model

In simple linear regression we predict the crime rate based on the scores of public transport accessibilty. Now we apply **Linear regression** from Scipy library on two variables. Crime rate is our dependant variable and public transport accesibility is our independent variable. We fit the model with second order polynomial in figure(7.1),third order polynomial in figure(7.2) and fourth order polynomial in figure(7.3). Here we notice that third order polynomial captures the relation better and fourth order polynomial overfits the model. Black line is the simple y=mx+b function for all points within the minimum and maximum range of x-axis.

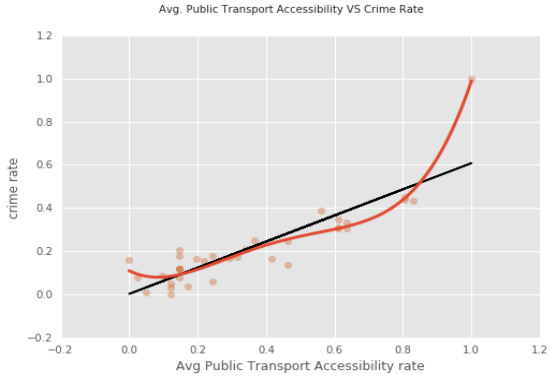
How well the model fits the data?To find out how close the data is to our fitted regression line we check R-squared value, which in our case is high value of 0.866. So, we can say that **86.6%** of the variation in the data can be explained by our model.



Figure(7.1) Second order polynomial



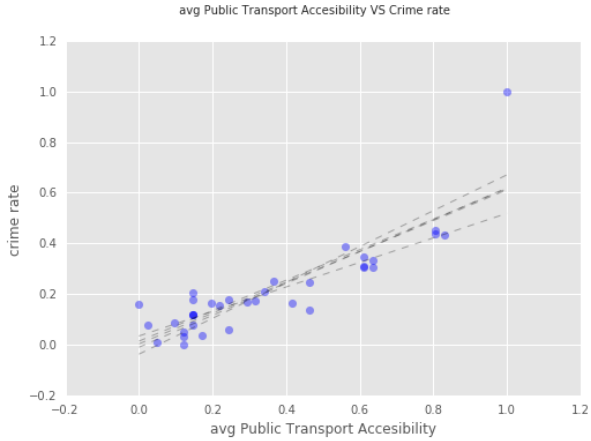
Figure(7.2) Third order polynomial



Figure(7.3) Fourth order polynomial

2.5 Cross-Validation

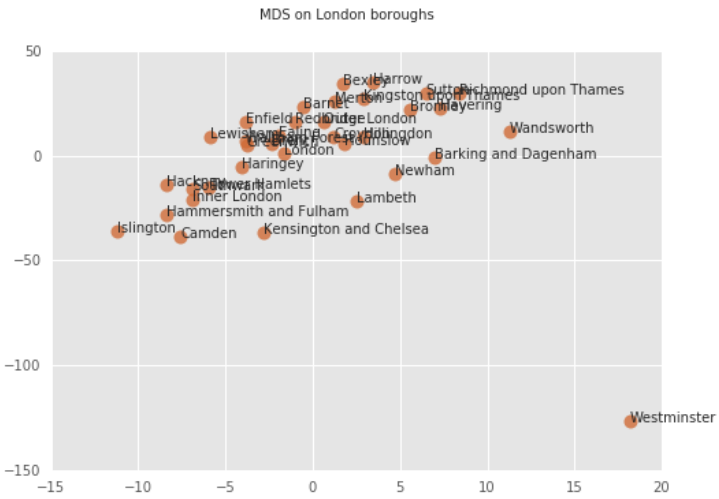
We used scikit-learn library to run K-Fold cross validation on our regression analysis model for 5 folds to check how well our model generalizes to new data. We compare the real values of the data to the estimated values, the small differences between them show that this is a good model. We chose K-Fold validation method over the hold-out method because every data point gets to be in test-set once and in training-set k-1 times, it doesn't matter how the data set is divided. The evaluation variance is not as high as it is with hold-out method.



Figure(8) scatter plot and overlaid regression lines

2.6 Multi dimensional Scaling(MDS)

We applied Multi-Dimensional Scaling to reveal the structure of the data set by plotting points in two dimensions. Human eye can spot patterns easily in two dimensional plots. Distances computed in higher dimensions do not capture the true relationships between the variables. We had 85 columns in the data set and 85*85 distance matrix is very high dimensional, the curse of dimensionality makes it really hard to interpret the plot. So, we select the columns that have strong correlation with crime rate (Number of cars per household, Average Age, Proportion of population of working-age, Employment rate, Youth Unemployment, working-age population who claim out-of-work benefits, Ambulance incidents, Number of cars per household, Average Public Transport Accessibility score, Crime rate). In the figure (9) below we notice that Westminster is different from rest of the boroughs. Here, we see Westminster is different from rest of the boroughs of London but this difference depends on the distance function we defined. Westminster has the highest crime rate so it is far from other boroughs. We also notice Lambeth, Newham, Kensington and Chelsea, Hammersmith and Fulham, Hackney, Tower Hamlets, Camden, Islington Boroughs are showing a similar pattern and rest of the boroughs are in a different category and the share common features. One perspective to look at this is that boroughs have been categorized according to similarities in crime rates and factors contributing to it.



Figure(9) Multi Dimensional Scaling on London Boroughs

2.7 Cluster analysis

We applied K-Means Clustering algorithm (with k=3, initial seeds= 10) on the boroughs of London, visualised the results on principle components and we found three distinct clusters i.e. red, blue and green. We notice that only Westminster (obvious outlier) is in red cluster. Boroughs in blue cluster share similar characteristics to other boroughs in blue cluster and same holds true for the boroughs in green cluster (figure (10.1)).

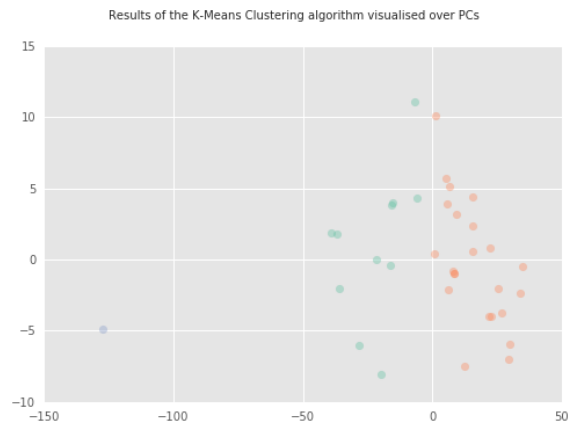


Figure (10.1) Results of clustering with K=3

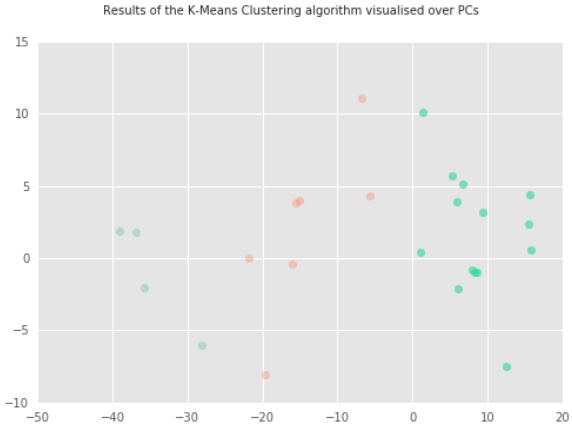


Figure (10.2) Results of clustering with K=5

We see that even with different values of K our clusters are nicely separated when viewed on principal component axis.

3. Conclusion

Anne Milgram (57th Attorney General of New Jersey, 2007-2010) said "*Bringing data analytics and statistical analysis to the US criminal justice system helped in transformation to better decision making which helped in significantly reducing the crime rate.*"

Among many other factors that contribute towards the crime rates 'proportion of working age people' has a big impact on crime rates. A lot of research has been done in the direction of analysing the relationship between crime rate and un-employment/employment rate, poverty rate etc. We focused on the relationship between crime rate and average public transport accessibility score within the boroughs of London. We also discussed in detail the properties that make Westminster (outlier in our data set) an attractive place for certain theft crime, snatch and run crime, violence again person and traffic crimes. Our results support the hypothesis that areas with better public transport accessibility have higher crime rates and areas with higher number of cars per household have lower crime rates. These variables are highly correlated and have a linear relationship as indicated by Pearson correlation of 0.866. Linear regression model is a good fit and it can explain 86.6% of variation in the data. Boroughs close together in MDS analysis share common features like crime rates, employment rates, average age, ambulance incidents, median house prices, average income, job density, availability of public transport scores etc.

Place-based policing significantly decreases crime rates. Deployment of more police in and around the train/tube stations, bus and tram stops may reduce the crime rates. Finding crime hotspots is an interesting research area and it can help people feel safe in their own neighbourhood.