Big Data Project

Analysing London Police Data

Zahrah Sharif

KUBRICK GROUP PROJECT

This project was done during the month of November.

First release, December 2016

Contents

1. Introduction

1.1 Motivation

The Home Office (HO) has become increasingly concerned by potential trends over the last few years of crime in the UK. Therefore by analysing key data sets readily available we may be able to draw some conclusions and gain important insight. This will help in making better decisions to better allocate resources more effectively.

1.2 Objective

The objective of this project is to cleanse and normalise data and find trends. By finding trends we are able to justify expenditure and better target resources. Some of the key objectives are outlined below:

- Use SQL to normalise datasets into third normal form and form a central table including all primary keys
- See what trends can be seen from the data and download outside datasets in order to see if there are outside factors that may contribute to what is observed
- identifying potential suggestions on how to make the UK a safer and better place

2. Understand your data

2.1 First ideas

Crime data is readily available on https://data.gov.uk/ for the UK and https://data.police.uk/. Given the time scale of the project it would be inefficient to analyse the entire UK crime dataset. This led me and my team to focusing purely on London particularly focusing on the London Metropolitan Police, Surrey and the City of London police.With London ranking top five in one of the cities with the highest rates of crime as well, it is in the public's interest to find out what exactly the crimes are both for safety and to better allocate police resources saving the government money.

Before we started the analysis the most important part was to select to the raw data to process. The next step was to then ingest and clean the data appropriately as well as profiling the data correctly. The correct normalised tables were also formed in a visually appealing way.

2.1.1 The Dataset

Many datasets were looked at in order to see any potential relationships. Crime data which listed the number of crimes per LSOA codes for the given police areas were downloaded for the years 2011-2015. These years were taken so that we could see if anything changed over time. This data was downloaded from https://data.police.uk/. Furthermore population data was also taken which noted the people per sq km per LSOA code. In addition to further see any patterns deprivation data was downloaded which noted what is called the IMD rank explained later.

2.2 Preprocessing your data

Visual studio is a component or tool of SSIS where we formed ETL packages which involved loops to transform the data. The data which was in the form of what we call a flat file source or csv was successfully uploaded to SQL database and transformed by a series of manipulations which were managed in visual.

Once ingested into SQL and before normalisation is even carried out an important step is data profiling. Data profiling is where we examine the data by SQL server data profile viewer where we collect statistics or small summaries about our data. For example, in our case the data based on number of crimes from the years 2011-2015 contained some nulls and data from other cities other than London was appearing. We performed some cleansing on the data for example where the LSOA code was blank (0.9401%) was removed. Reported by was correct with only Surrey, city of London and metropolitan police being used in the analysis, however most crimes do fall within the London metropolitan police that is 91%.

From the profiling on the crime data sets, crimes seemed to have been recorded outside of London yet were picked up by the three London constituents. Whilst there could be many reasons for this, and whilst they may constituent a very small percentage of the total crimes recorded these were subsequently removed by downloading a London specific LSOA codes table to make the necessary joins to filter. However no cleansing was required for the population data.

Below is an image of what the normalised tables looked like with the central table in the middle.

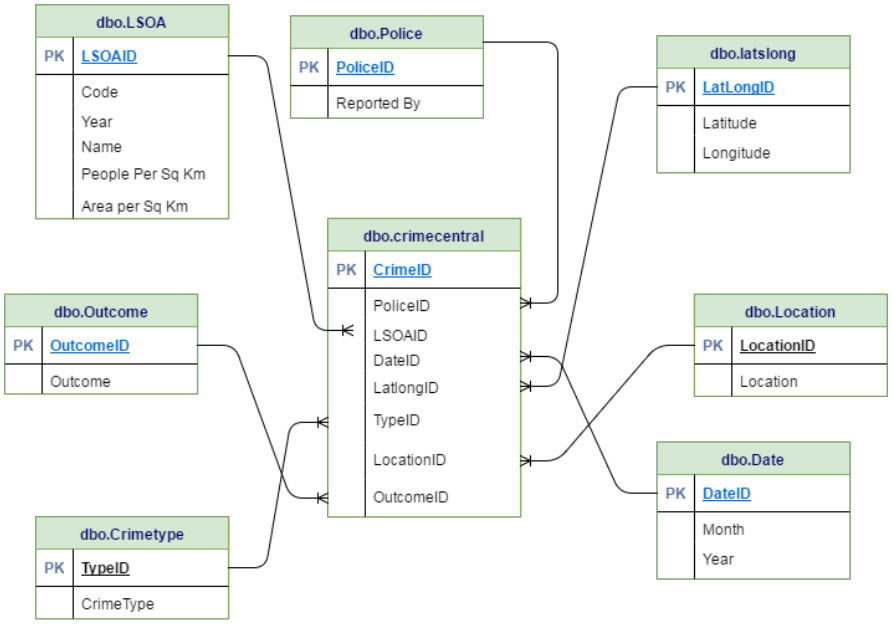

Figure 2.1: Normalised table

The code to develop this is as below:

```
select
    c.[Crime ID]
    ,p.[PoliceID]
    ,d.DateID
    ,lt.LatlongID
    ,ls.[LSOAID]
    ,ct.[CrimeTypeID]
    ,o.[OutcomeID]

into dbo.crimecentral
from dbo.crimedatabase1 c
join dbo.Dates d on c.Year=d.year and d.month=c.month
join dbo.Crimetype ct on c.[Crime type]=ct.[Crime type]
join dbo.Outcome o on c.[Last outcome category]=o.[Last outcome category]
join dbo.Police p on c.[Reported By]=p.[Reported by]
join dbo.latslongs lt on c.Latitude=lt.Latitude and c.Longitude=lt.Longitude
join dbo.LSOA ls on c.Year=ls.[Year] and c.[LSOA code]=ls.[Code]

select * from dbo.crimecentral

order by [Crime ID]
```

Figure 2.2: Code to form central table

2.3 Software available

The software used for pre-processing the data was Microsoft visual, excel and data profile viewer. SQL was used since it is best for dealing with large data sets where we had some 4 millions rows of data sets. Later on tableau was used to produce graphs since it is easy to use and diverse.

3. Analysis

The data, once in normalised form was then analysed through various techniques as can be seen in the SQL queries, then ran through tableau.

London Crimes

Through particular queries in SQL the change in crime types over the years 2011-2015 was then projected into tableau. This as done for the London regions all three police constituents were covering.

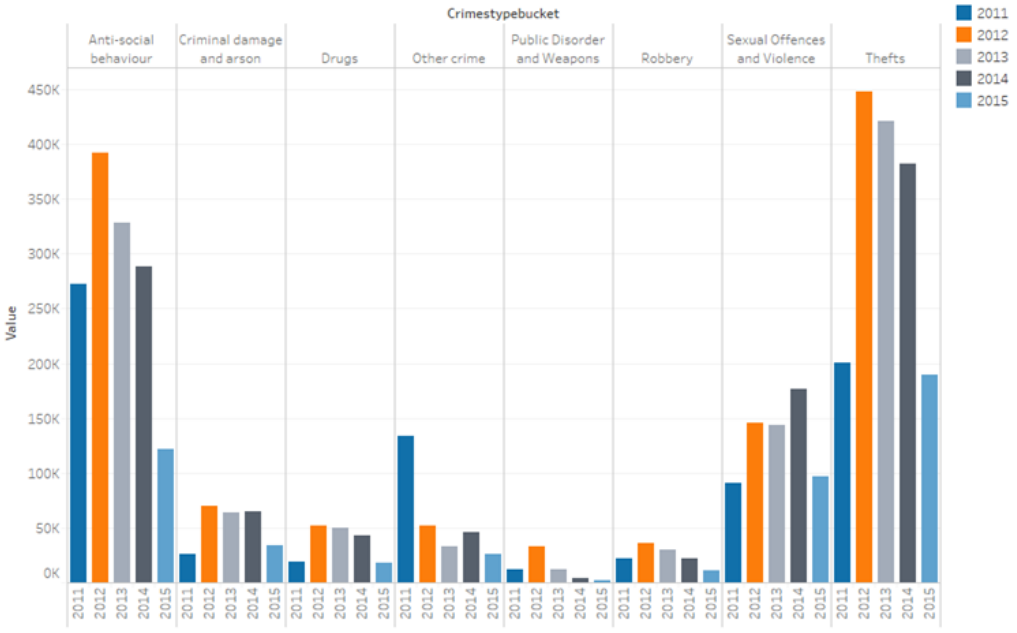

Figure 3.1: Crime Type over the Years

Note that when taken the crime types over the years there were a few problems in analysing this data. Some crime types had a zero count for some of the years due to the police force discontinuing some of the definitions of the crime types. I therefore bucketed some of the crime types through SQL, reducing the number of crime types from 16 to 8. For example, for 'Other crime' I grouped together other crime with public order to form a new other crime group therefore removing the 0s'. The above graph displays the eight crime types against count.

As we can see there are eight crime types and noticeably thefts seems to of had the record breaking value of the most crimes in 2012. In fact both anti-social behaviour and thefts seem to be the top two crimes with the highest number.

It would then be interesting to see which regions, by LSOA, have the highest number of crimes. By zooming in on particular regions, and having data on the entire population number per LSOA code I was then able to calculate the number of crimes per head so that it is easy to compare which areas have most crimes. The top five were looked at and in almost every case Westminster seemed to top the area with the highest crime rate over all years. To be more specific it is the area with LSOA code E01033595; Westminster 013E which tops the highest number of crimes every year. Below displays how the crimes per head has changed over the years:

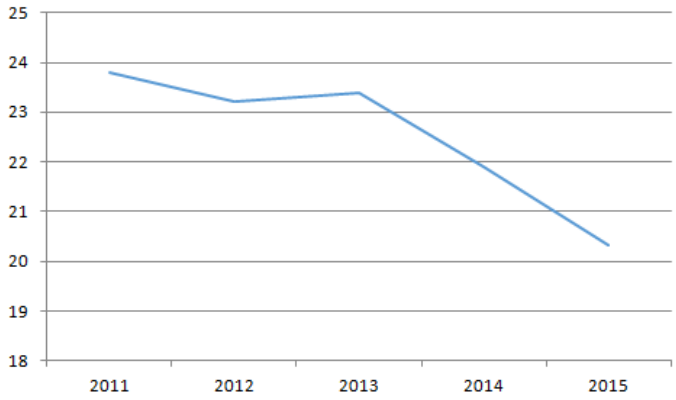

Figure 3.2: Crime per head in Westminster 2011-2015

Despite the fall in crimes per head it still remains the area with the greatest number of crimes in all of London. Over the five years the top 3 areas with the most crimes has not changed either.

The below map displays the areas, highlighted in red, with the highest crime rate. If we saw 2015 we can see that although there was a decrease the areas highlighted in red with highest rates do not seem to have changed much.

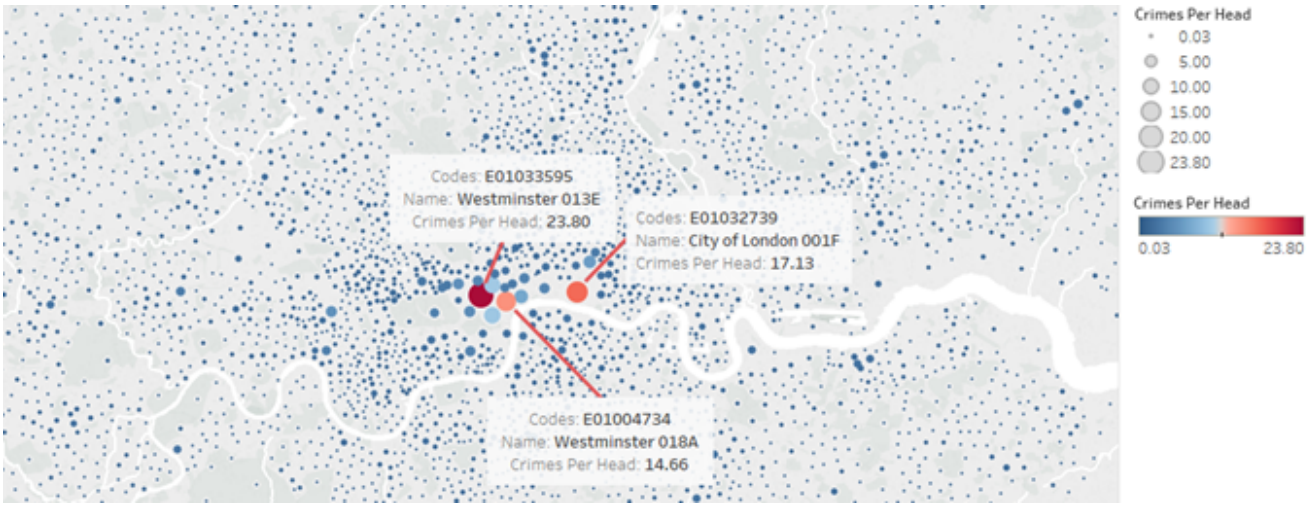

Figure 3.3: 2011 crimes per head

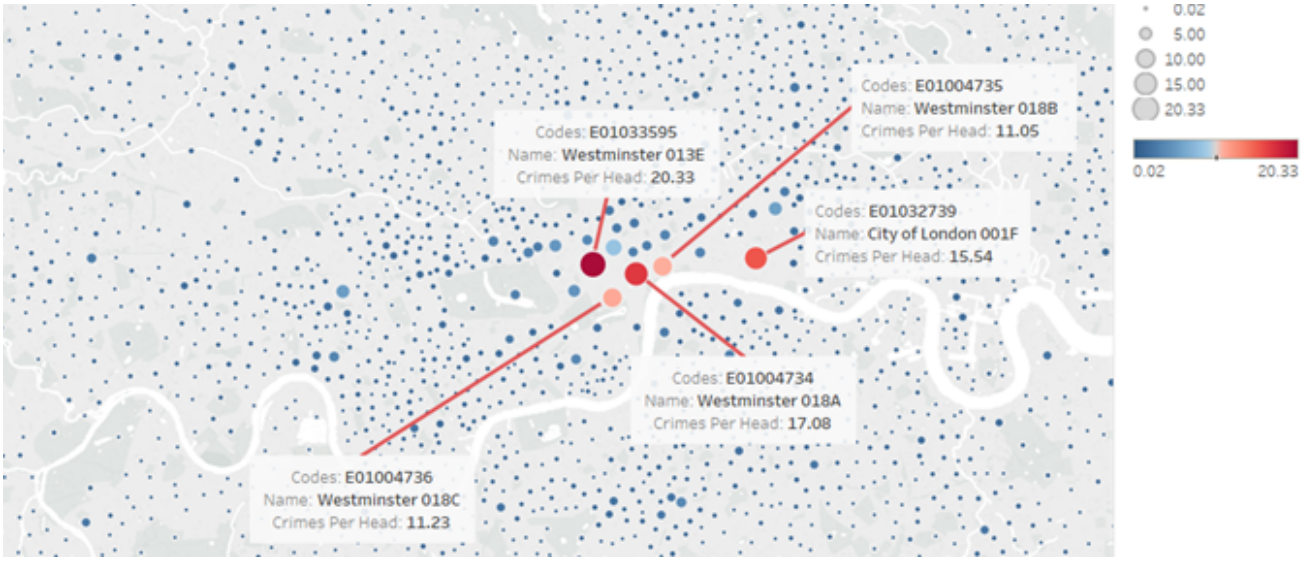

Figure 3.4: 2015 crimes per head

3.0.1 Westminster

Now that we have targeted Westminster (more specially the area with code E0103359 with specific name Westminster 013E, it would be useful to know which crime type is most prevalent here. By taking a count on the most common crime type in this area, three crimes are noticeably on top. Theft seems to be the most common crime type across the years 2012-2014. Shoplifting and anti-social behaviour also seem to be listed in the top five most common crime type across the five years. The following page presents these three common crimes in Westminster.

Comments to be made are that overall from 2013 onwards the crimes have decreased in all areas. It would be interesting to find out why it is that around 2012-2013 crimes for each of the three seemed to peak. Any past measures from 2013 onwards which have caused the decrease in the crimes should be reinforced and maintained to ensure the downward trend continues over the next few years.

Table 3.1: Crime Type Trends in Westminster

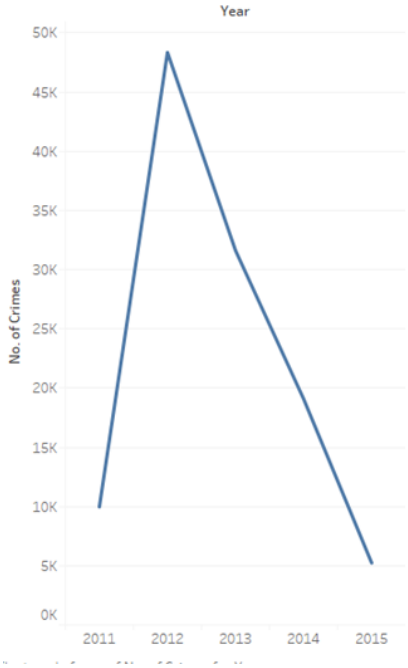

Figure 3.5: Theft Trend

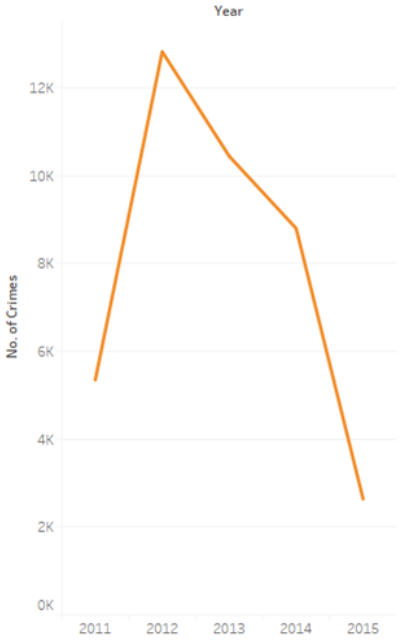

Figure 3.6: Anti-Social Behaviour

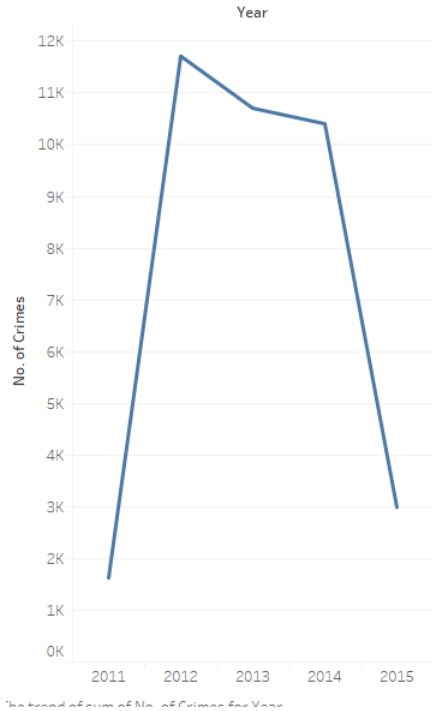

Figure 3.7: Shoplifting trend

3.1 Deprivation Rank

There could be many reasons as to why Westminster has the highest rates of crime. To try and explain this I extracted data for the English indices of deprivation. As stated in gov.uk it is the official measure of relative deprivation for small areas (or neighbourhoods) in England. The Index of multiple deprivation ranks every small areas in England (by LSOA code) from 1 being most deprived area to 32,844, the least deprived area. We can directly compare the ranks of different small areas in England, so if one areas rank is closer to say 10 than another area then it is more deprived.

There are seven domains taken to calculate the index of deprivation, to name a three such inputs are income deprivation and employment deprivation.Without going into much detail each of the seven variables are given a weighting based on academic literature. The map on the following page (fig 3.8) highlights the crime rates for the year 2011.

When compared against the areas of deprivation in the map below, by visual inspection we can see that there is no link between a low IMD level with a high rate of crime. When similar maps are produced for 2015 there is not much change. In-fact there doesn't seem to be a relationship between IMD and the number of crimes from the figures on the following page.

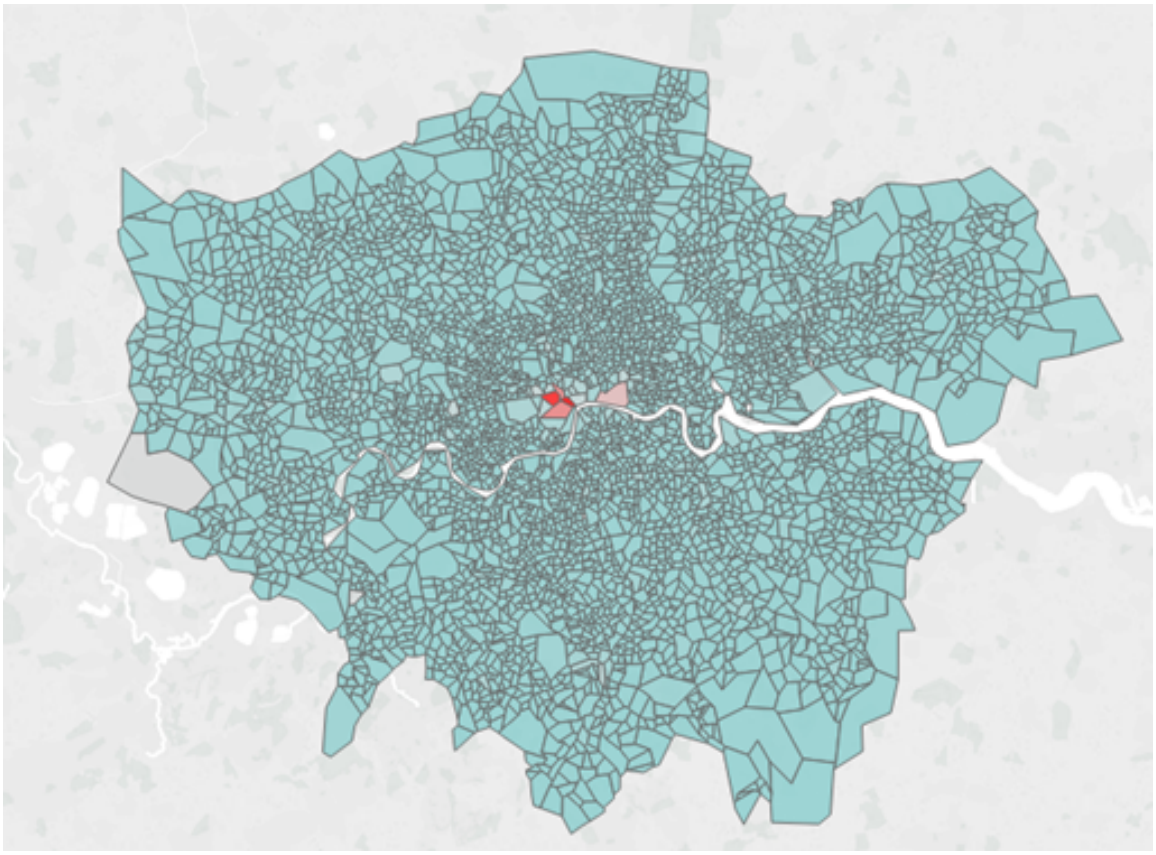

Figure 3.8: Crimes where red indicates high rate

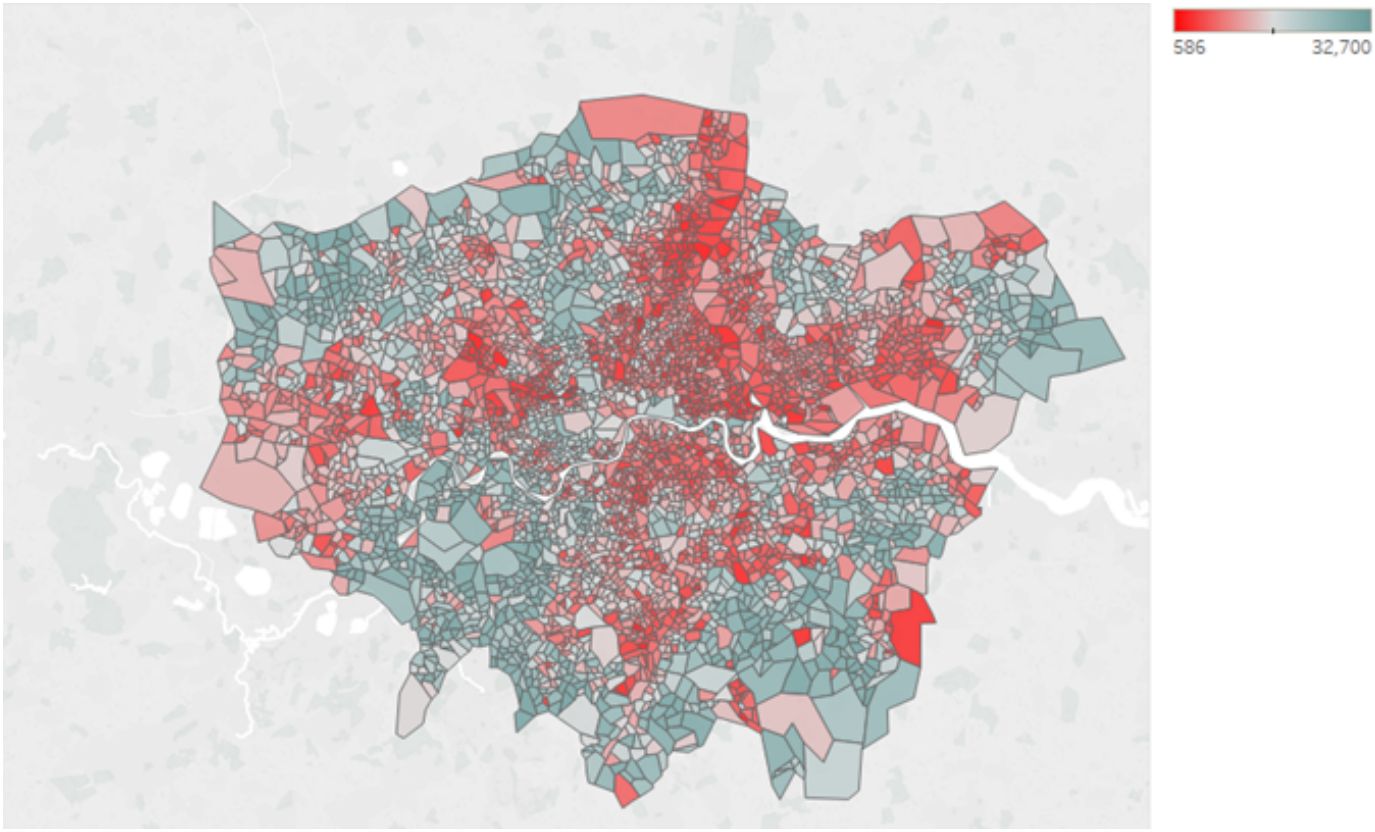

Figure 3.9: IMD where red indicates low IMD

Table 3.2: IMD vs No. of Crimes

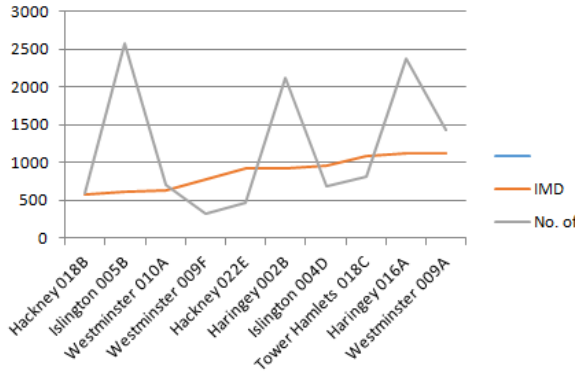

Figure 3.10: 2011

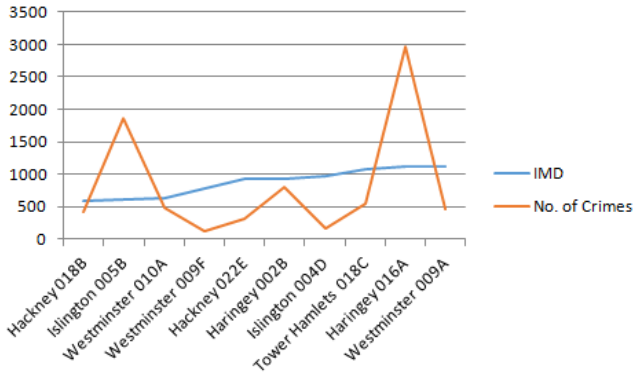

Figure 3.11: 2015

By taking the top 10 areas with lowest IMD (that is most deprived relative to one another) I plotted the number of crimes associated with each area against its IMD. The above figures show how the IMD changes with the number of crimes; as can be seen there does not seem to be a clear relationship between the two.

Whilst the IMD vs the number of crimes might not show a clear relationship there are many other factors to consider. Westminster is an areas which attracts a high level of tourists every year and is home to probably a large amount of hotels and bars. This however is an assumption, but common sense would lead one to say that because it's a hot tourist attraction this could be why there are lots of thefts. Data for the number of hotels or bars however could not be found by LSOA area so it is difficult to say.

3.2 Conclusion

From the results present I have found that Westminster has the highest rate of crime, not only for the one year but consistently over the five year period. I would suggest that the home office pay special attention to this part of the city and ensure that whatever measures were being taken by the police force between the years 2012 onwards to cause the decline in the most common types of crime, that they continue to be vigilant.

Furthermore improvements can be made on my analysis; with the first being to obtain data on hotels and bars per LSOA code. Then if a correlation is seen between rates of crime and the number of bars measures could be taken to have stricter laws or more police force situated in these areas. Overall however, since Westminster has a high rate of crime it would be advised that the police direct resources to this region with special advice to citizens to be wary of theft.