London Crime: Data Understanding report

Maddalena Amendola Giuliano Cornacchia Mario Leonardo Salinas

Contents

1 Dataset descritption

The original dataset consists of 13.490.604 records, each record is described by 7 variables:

- The `lsoa_code` is a code that identifies the 4835 different LSOAs (Lower Layer Super Output Area), namely a geographic area inside London.
- The `borough` is a nominal variable that identifies one of the 33 boroughs present in London, each of them contains a variable number of LSOAs.
- The `major_category` is a categorical attribute that specifies the category of the crime committed; there are 9 major categories: `Theft and Handling`, `Violence Against the Person`, `Criminal Damage`, `Robbery`, `Burglary`, `Other Notifiable Offences`, `Drugs`, `Sexual Offences`, `Fraud or Forgery`.
- The `minor_category` is a categorical attribute that specifies precisely the crime committed; there are 32 minor categories.
- The `year` is a numerical attribute that indicates the year when the crime was committed; year varies in [2008,2016].
- The `month` is a numerical attribute that defines in which month the crime was committed.
- The `value` is a numerical attribute that indicates how many crimes of a certain type in a certain period in a LSOA were committed.

Reviewing the data, there aren't null values or missing ones. Furthermore we choose to keep all columns in order to be able to analyze from different perspectives the crime distribution; but comes out that approximately the 70% of the records has a value equal to zero.

Considering a generic line of the dataset, for each lsoa_code, minor_category, month and year the value attribute specifies how many crimes of that kind were committed, even if there isnt any crime that was committed.

Since a `value` = 0 gives no contribution for the computation of the standard statistical measures, the dataset was filtered deleting all the rows where the value was equal to 0. This operation leads to an important memory space saving: we started with about 13M records and, after the filtering, we ended up with 3M non-zero records (the 23% of the real size). This will also allowed a faster of computations.

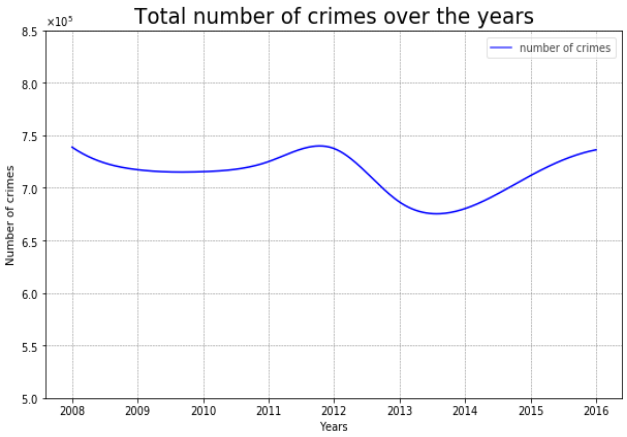
2 Exploratory Analysis

2.1 Features Distributions

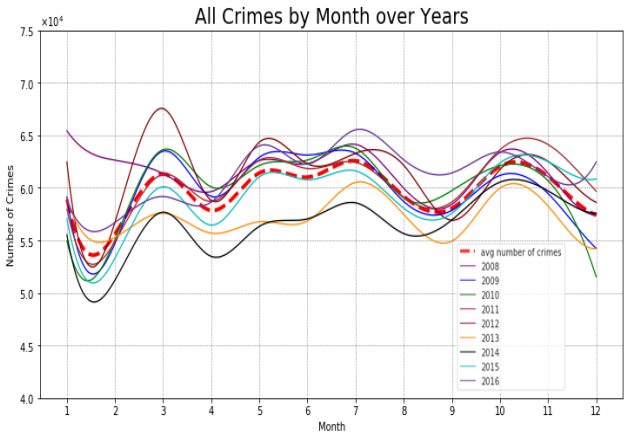
We start our analysis by having a look at the distribution of the number of crimes over the years and months.

The line graph 1a shows that the number of crimes is equally distributed between the years with the peaks in 2008, 2012 and 2016; while in 2013 and 2014 there were fewer crimes. In particular, the maximum number of crimes is in 2008 with a total of 738k crimes, while the lowest value is in 2014 with 680k crimes. The trend can be defined as constant except for small fluctuations of $\pm 0.1\%$.

In figure 1b it is possible to compare the monthly trend of each crime in each "year" with the average trend. The observations from figure 1a are confirmed, the years 2013 and 2014 are below the average while 2008, 2012(with a critical peak in March), 2016 are above. The total number of crimes varies each month compared to its average at most by 5k / 10k. Finally, both the average number of crimes line and the ones referring to the years show some common behavior:there are peaks in March, May, July and October, while February is the month with less crimes. Up to now we know that the number of crimes is



(a) Total number of crimes during the years.



(b) Total number of crimes over months.

almost constant over the years and months less than small fluctuations; the next step is to analyze how the crimes are distributed over the years in terms of major category.

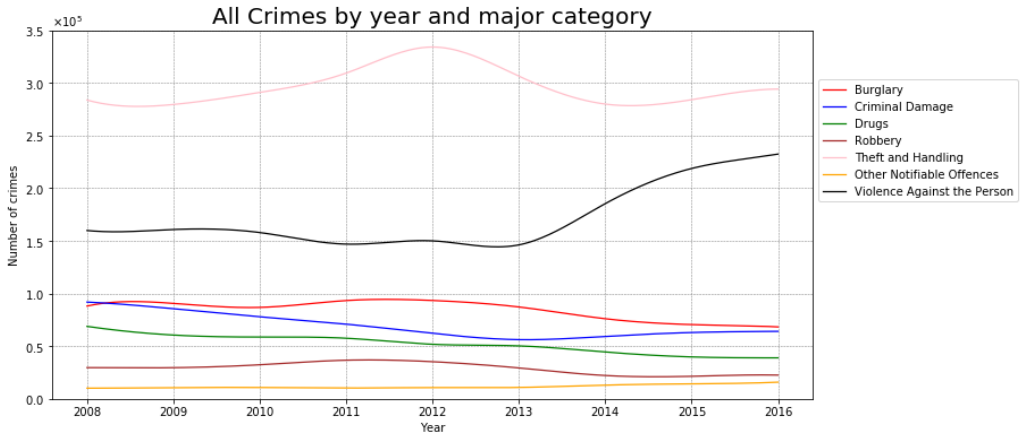


Figure 1: Total number of crimes during the years.

We decided not to show the lines related to Fraud and Forgery and Sexual Offences since the dataset contains only 2008 data for these categories. However we can observe that lines related to different crimes almost never cross each other; this means that the fraction of felony with a specific category over the total number of crimes is constant in time. Other Notifiable Offences, Robbery, Criminal Damage, Burglary don't show particular fluctuations, while Drugs crime are decreasing linearly. Theft and Handling has a peak in the year 2012; maybe due to the XXX Olympic Games hosted by the capital. More concerning is the Violence Against the Person rise since 2013; its value doubles in two years.

We can now visualize the proportions between the major crime's categories without distinguishing by year, since we've already seen that these fractions remain constant during the years.

The donut chart in figure 2 shows that the two most frequent major categories are Theft and Handling and Violence Against the Person, which respectively represents the 41% and 24% of the total crimes; Burglary, Criminal Damage and Drugs are the second most frequent crimes with percentages between, while the other categories have percentages ≤ 4, thus really rare.

We have also analyzed what are the most influential minor category in each major category and discovered that each major category has at most two dominant minor categories.

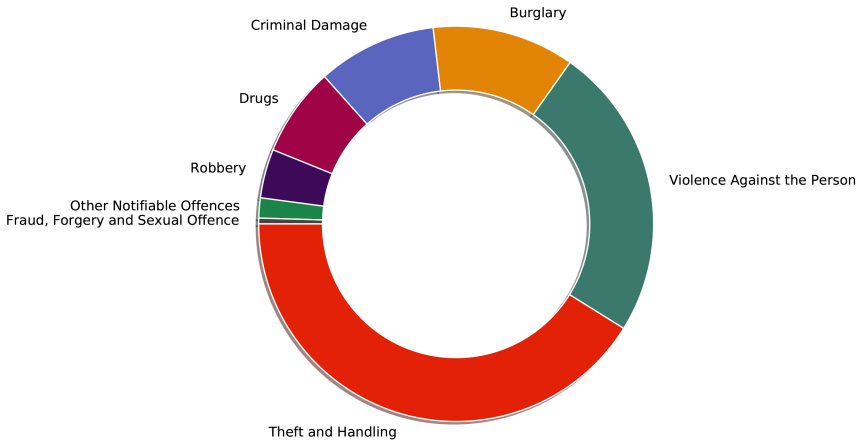


Figure 2: Major crime fractions over the total.

2.2 Geolocalized Features Distribution

We already looked at data from a statistical perspective, but since we're dealing with geographic entities such as LSOAs and boroughs a further geolocalized crime inspection is needed. In order to do so we enriched the filtered dataset with geographical attributes, such as the shapes of boroughs and LSOA.

To proceed with a time-series-like approach, we split the geo-dataset by year. At this point we have 9 datasets, one for each year between $[2008, 2016]$, sharing the same schema and semantics.

We also computed a quantitative attribute `crime/month`, equals to $\frac{value}{12}$. This measure is more understandable since its referred to a smaller time interval and so it gives a stronger impression.

Since here the purpose is to better understand the trends highlighted in the exploratory analysis section, we choose to reproduce the same plots in this geolocalized analysis. The quantitative information that we choose to represent is the crime/month attribute. We modeled and visualized this dimension at LSOA-level with the help of a color-map, while we still show boroughs edges.

In each choropleth map we also show:

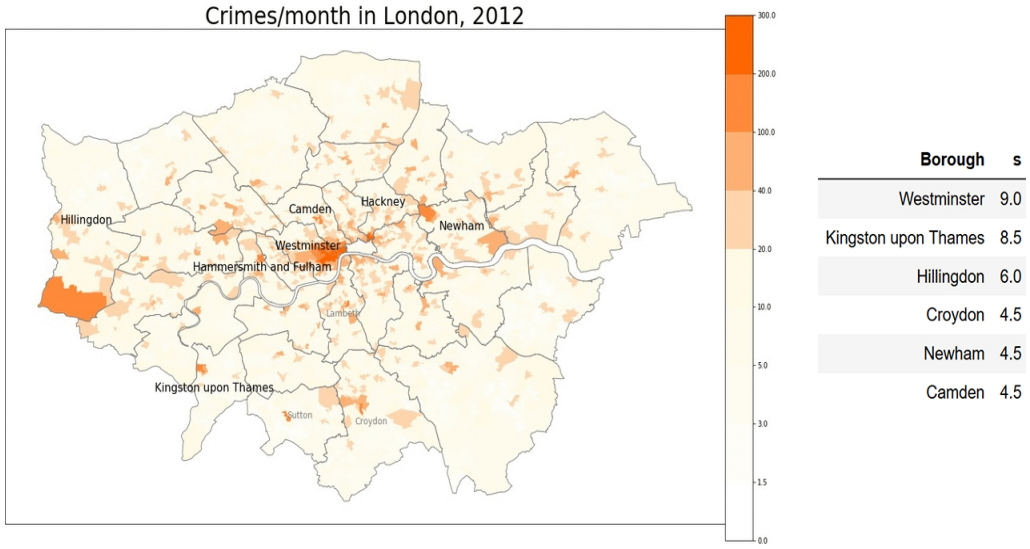
- in *black* the names of the borough with at least one LSOA that falls in the two highest ranges of the color-map. We will call such LSOA an **hotspot**;
- in *gray* the names of the boroughs with at least one LSOA that falls in the third highest ranges of the color-map; We will call such LSOA a **concerning** one;

This additional information will also provide us an implicit borough classification, indeed we can now compute a **colormap-induced-score** s and rank boroughs in each plot. For each borough b, the score is computed as:

$$s(b) = n_b + \frac{n_g}{2}, \quad n_b, n_g \in [0, 9]$$

where n_b and n_g are respectively the number of times b appears in black and gray in the map. To assign a lower weight to gray-appearance-events we divided n_g by 2. This score will provide us a way to better understand the crime spatial distribution. We found that there's a **spatial locality** in crime, meaning that dangerous boroughs and hotspots don't change too much over the years. This behavior has been observed not only for the general crime value, but also for all major categories. As an example Figure 3 shows the map related to the total number of crimes in London in 2012.

We notice that hotspots inside dangerous boroughs are usually isolated, indeed in a dangerous borough there is at most one critical LSOA, and this LSOA often confines only with not-critical ones - an example is Kingston upon Thames; Westminister and, in general the LSOAs above the Thames tend to be the most dangerous zones.



Borough	s
Westminster	9.0
Kingston upon Thames	8.5
Hillingdon	6.0
Croydon	4.5
Newham	4.5
Camden	4.5

Figure 3: All crimes distribution over London, 2012.

Finally the central white LSOA corresponds to the City of London area and there are no crimes associated with this zone.

The table on the right of the map shows the boroughs ranked by our score s computed for each year. The first three dangerous boroughs, resulting from the table, have a score really close to the maximum. This means that they have been the most dangerous ones almost every year.

We repeated the very same analysis for each major category c_i, computing every time the related score s_{c_i}. Once we have collected all the scores, for each borough we can now define a **weighed score**:

$$\forall b \in \{Boroughs\}, ws(b) = \sum_{c \in MC} s_c(b) \times w_c$$

where $ws(b) \in [0, 9]$, MC is the set of major categories of crime considered, $s_c(b)$ is the score of the borough b in the category c and w_c is the weight assigned to the crime c. The weights are the percentage of crimes with major category equals to c over the total crime number.

This score represent a first, naive, **danger indicator**: a score close to 9 means that inside that borough, since 2008, there was at least one hotspot for each crime category. The table below shows a normalized version of ws.

Borough	*ws*
Westminster	0.98
Kingston upon Thames	0.77
Camden	0.75
Hillingdon	0.63
Croydon	0.56
Hammersmith and Fulham	0.53
Newham	0.52

Looking at the boroughs names we have that:

- 3 boroughs that are in the city centre and on the northern bank of the Thams: Westminster, Camden, Hammersmith and Fulham; this confirms the hypotized dangerousness the involved zones
- Hillingdon, even if appears with an high score, shows almost always only one hotspot, that coincides with the LSOA that contains the airport
- all the other boroughs in the list are like Hillingdon, and this high crime density in just few LSOA is certainly a phenomenon that deserves further analysis.

3 Correlations

Also in the correlation analysis we choose to study the problem with a time-series-like approach by means of the previously splitted datasets. The main problem in this phase was the lack of numerical quantitative attributes in the original dataset; indeed we only have value. The first thing to do is to reshape the datasets in order to have more independent variables.

3.1 Dataset Reshaping

The first idea was to create one numerical attribute for each major category that contains the sum of the values of that kind of crime in the current year - since we're dealing with a dataset split by year - for a specific LSOA. Thus these new datasets will contain 4835 rows, namely the number of LSOA, and 9 new attributes, one for each major category.Furthermore we extend our dataset with economic and demographic indicators and points of interest (Subways, Train Stations, Taxis, Airports, Bus Stations, Parkings, Monuments, Hospitals, Police Stations, Stadiums).

3.2 Correlations Analysis

We first compute the correlation between crimes both at LSOA and borough level.

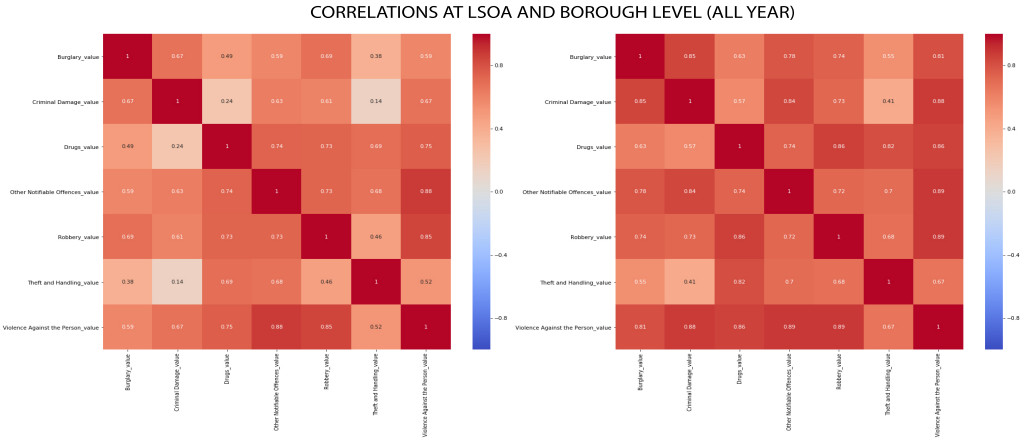


Figure 4: On the left the LSOA-level crime's correlation matrix, on the right the borough-level one.

Since the mutual correlations between different categories of crimes are strong both at LSOA and borough level we can say that if a borough has an high number of crimes, and thus can be considered dangerous with our score, it's likely to have an high number of crimes in all categories; this observation its related to the score defined in the previous section because the crime levels seem to be higher in some recurring zones as the northern bank of the Thames and the city center in general, resulting in very high scores in the final rank table.

Then we analyzed the correlation among all the attributes of the enriched dataset at LSOA level, but the only strongly correlated subset of attributes was again the one of crime values. All the new features

seemed not correlated with the values of the crimes, the only notable correlation is between the number of Airports in an LSOA and Other Notifiable Offences. We now try to see the same correlations at borough level. This time the correlation matrix showed better results. We analyze the most significant ones by looking how the correlation varies over the years.

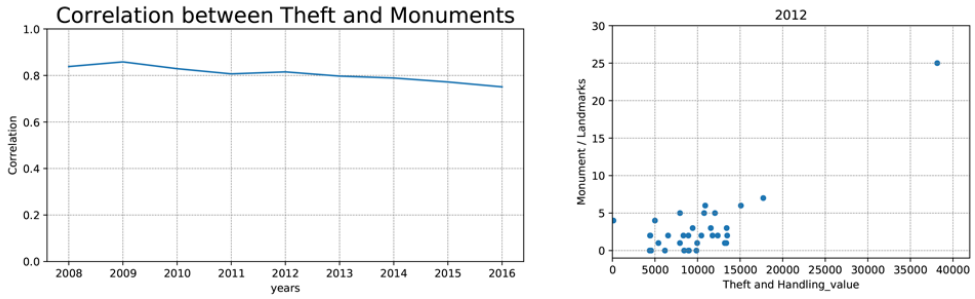


Figure 5: Correlation between Theft and Handling and Monuments, with scatter plot

The line-plot 4 confirms our first impression: the theft crime is focused in the center of London; the (Touristic) Center can be individuated by the position of the monuments and assuming that a Borough is as central as the number of monuments that falls in that Borough we can conclude that places near monuments have an high risk of theft and handling crimes (for example because they are crowded by thousand of tourists). We also noticed that this correlations is decreasing linearly during years.

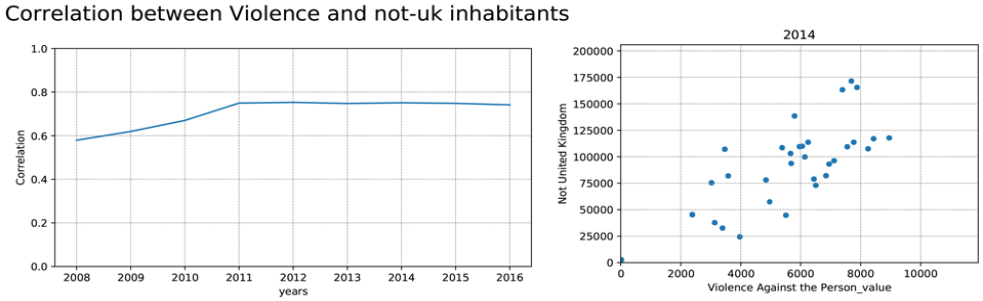


Figure 6: Correlation between Violence and non UK inhabitants, with scatter plot

In the plot above we found strong correlations between a Violence and not UK inhabitants

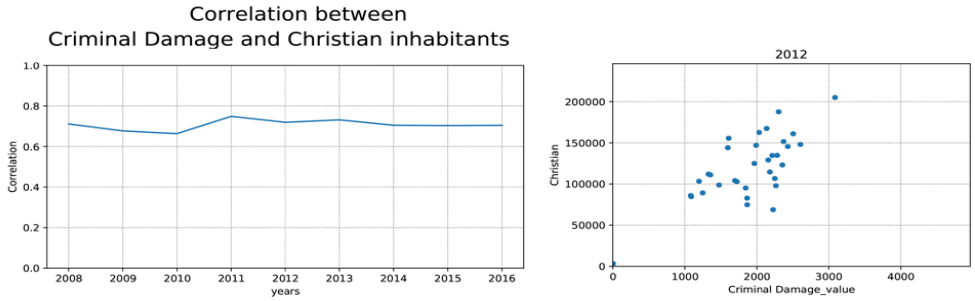


Figure 7: Correlation between Criminal Damage and non Christian inhabitants, with scatter plot

Finally these last plots gives us the opportunity to highlight that even if a crime is happened in a place it doesn't mean that their population is the active part, but can also be the victim. So in this case we can't know if the Christian/not UK population plays an active or passive role in this crime category.

4 Analytic Project Proposal

In this section we propose three analytic projects based upon our analysis:

1. Model that predicts the number and the kind of crimes that will occur in a specific LSOA given the month and year.

2. Model that predicts the number and the kind of crimes that will occur in a specific BOROUGH given the month and year.

3. Given the available resources of London Police provide a *smart* way to arrange them in order to reduce the number of crimes according to the prediction of our model.

In each case we will consider this problem as a time-series and use relative clustering algorithms and modeling techniques.

After the data understanding presentation, we discussed points 1 and 2 of the list and, together with the PhD researcher Luca Pappalardo, we decided to merge them in one unique parametric model. In this way the user will be able tho choose the level of prediction (borough or LSOA).

The last proposal remains still valid and can be seen as an extension of our predictive model.