Title: Exploratory spatial analysis of the Crime Data from the Metropolitan Police

1. Introduction

This is an exploratory analysis of what implies to map crime data. The Metropolitan Police publishes data of crimes reported in csv files with location of events (LSOA codes, longitude, and latitude), organised by geographical areas and months. This data, when compared with other sources facilitates research on various topics like crime trends across neighbourhoods or relationships between crime types and locations. However, it's crucial to process the data properly to avoid misleading conclusions. This analysis will explore the considerations necessary to make proper sense of this valuable dataset.

The report is organised, as recommended, in Methodology, Implementation, Results and Discussion, and Conclusions. In the Methodology the issues encountered with the data collection and data selection are presented. In the Implementation, the work is carried out in 3 sections: data exploration in QGIS, data analysis in R Studio and spatial visualisation in QGIS. The Results and Discussion summarize the main decisions taken during the exploration and the justification for these. Finally, the conclusions round up the report.

1. Methodology

As any exploration, the workflow wasn't lineal but in cycles. It started with the selection of an area of study and which crime file to download. The London metropolitan area was chosen, for December 2023, which coincided with the tenth year of data from the Police archives.

It also involved to download a suitable map (shapefile) of London to visualise the spatial data. The Police's Street crime data had LSOA codes, so, the map had to include those codes. The last map of London LSOAs found was from 2011. After checking that some LSOAs had changed since then, it required to download an England map with the LSOAs boundaries for 2021 and edit it to get only Greater London.

Table 1. Summary of methodology

Data collection	Data selection	Interpretation and visualisation
- Police Data csv files for street crime reported - Open Geography Portal (ONS) for the map with LSOAs codes	London Metropolitan area for street crimes reported for December 2023	- Excel - QGIS - RStudio

The basic exploratory data analysis in excel and python and continued in QGIS for the spatial visualisation and RStudio for further statistical analysis.

The selection of LSOAs was not only chosen for practicality, because the police data had this geographic code. But it was also suitable because is a relatively small area. The Lower Layer Output Areas have an average population of 1500 people or 650 households, whereas the MSOAs, or Middle Layer Super Output Areas, have an average population of 7500 people or 4000 households. Any of those may be suitable to represent this data because the population is relatively similar. It also explains why some areas are bigger than others, considering the density of the population is not even in the Metropolitan Area or that certain areas contain geographic

features that affect its size and shape. In contrast, to work in Boroughs would show the distribution of the data less clearly, because these are big areas with very different populations. For instance, the City of London has around 8500 people when other boroughs have 140,000 or 390,000 people.

2. Implementation

Data exploration in QGIS

The first exploration showed that the data not only contained information of London events, because the Metropolitan Police also reported crimes in other areas of England. Figure 1 shows this with a point in Cornwall selected.

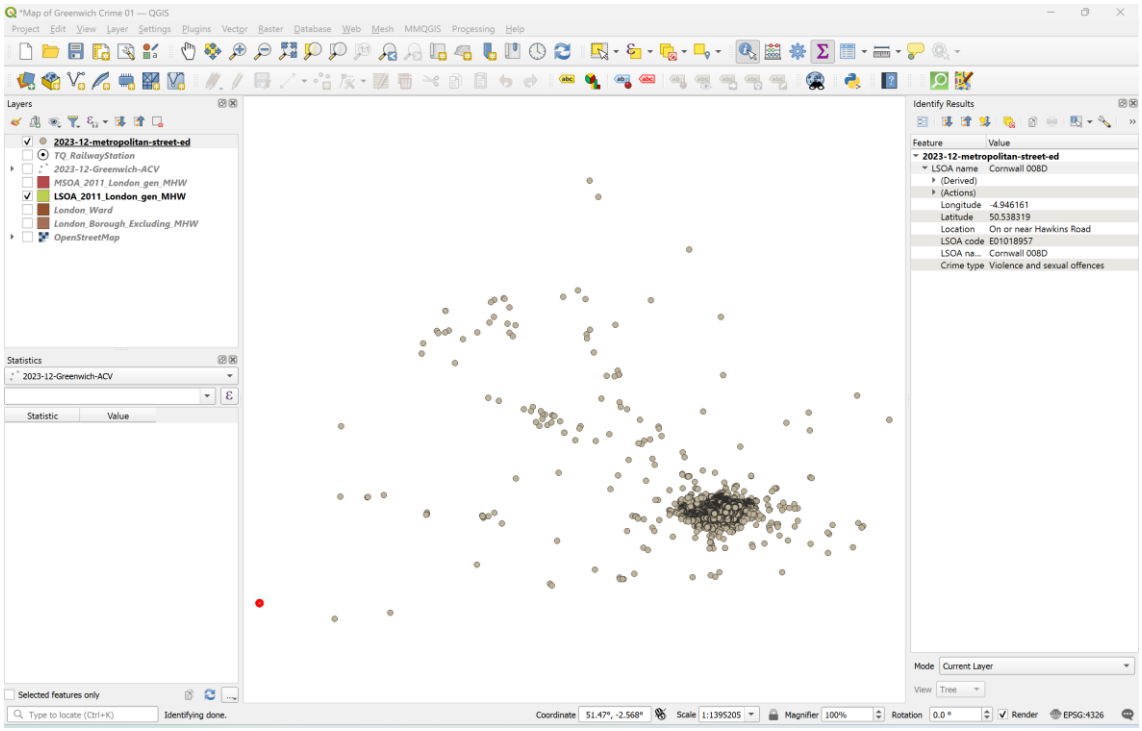


Figure 1. Map of crimes reported without a map.

In the same way, it was found that the LSOAs codes didn't coincide with the boundaries of the map from 2011. For certain cases it may not be necessary to know this, as shown in Figure 2, the map shows that most crimes reported were made in the category of Violence and sexual offences, in comparison with the other categories. But there are a few issues with this map. For start, when trying to see the areas with the higher number of events, the points are overlapping when they are close to each other, not to mention when they are reported in the same location. This hides a huge amount of information.

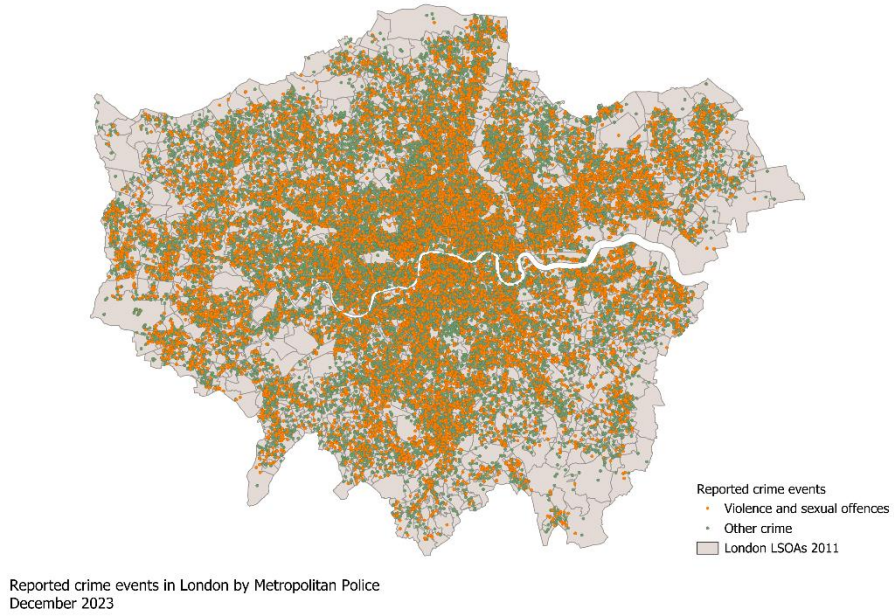


Figure 2. Map of reported crime events in London in December 2023, on map of London 2011.

So, a thematic visualization was tried. In Figure 3, the map in the left tries to show that most crime types reported by area are "Violence and sexual offences" shown in orange, but this is misleading in the way that when in numbers, it doesn't matter if the majority in that LSOA is by one event, the shading will be orange. And it gives the impression that the number of crimes reported is the same in all the areas shaded in the same colour.

In Figure 3, the map of the right, it tries to show that the number of crimes reported is not the same in all areas. In a quick look it shows that, but it also seems that the number of crimes is relatively spread in all of London. If we take a closer look, the legend tells us that the last category varies hugely from the others, so we need a better way to separate the categories and the shading of the areas.

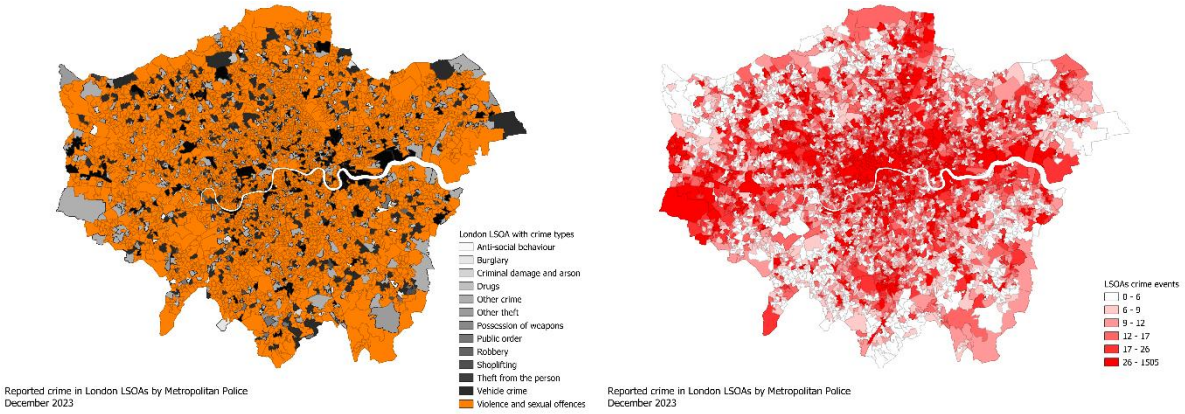


Figure 3. Thematic maps of crimes reported in London, December 2023, on a map of 2011.

Exploring further, we cannot really see which are the outliers in the map, because some of the LSOA names have changed, so we need to update the map with the correct name of LSOAs. The map of England LSOAs 2021had to be clipped to visualise just Greater London.

Figure 4 shows the process of map exploration in QGIS. The preparation of the map always ended in the question of what we are really seeing and what information we want to show.

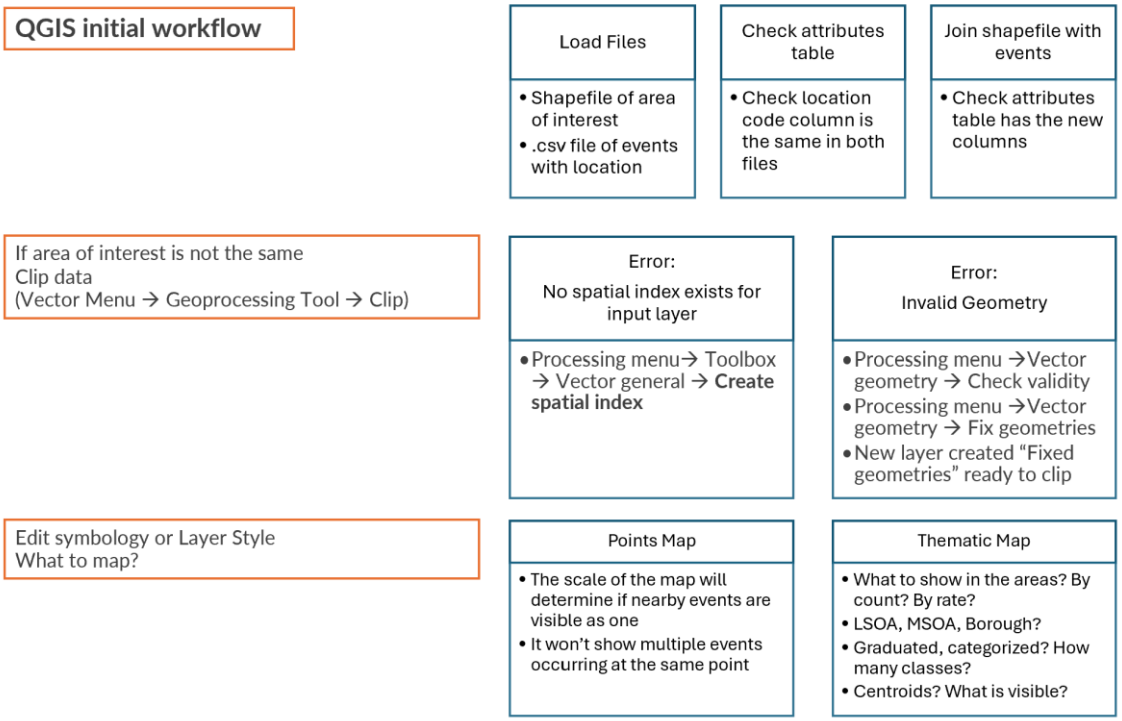


Figure 4. QGIS initial workflow

Data analysis in R Studio

With the obstacles found in QGIS, it was clear that a statistical distribution of the data was needed to show a better representation of the crimes reported.

The preparation of the data was made in Excel. This involved erasing the columns we don't need like the "Crime ID", "Month" (all data in the file is for December 2023), "Reported by" (all the events were reported to Metropolitan Police Force), "Falls within", "Context" and "Last outcome". Then, a quick check showed some events reported with no location, so these rows were erased. The file changed from 12 columns and 93,396 rows to 6 columns and 92,668 rows. This was the file loaded in R Studio for analysis.

Figure 5 shows the diagram followed in this part of the analysis.

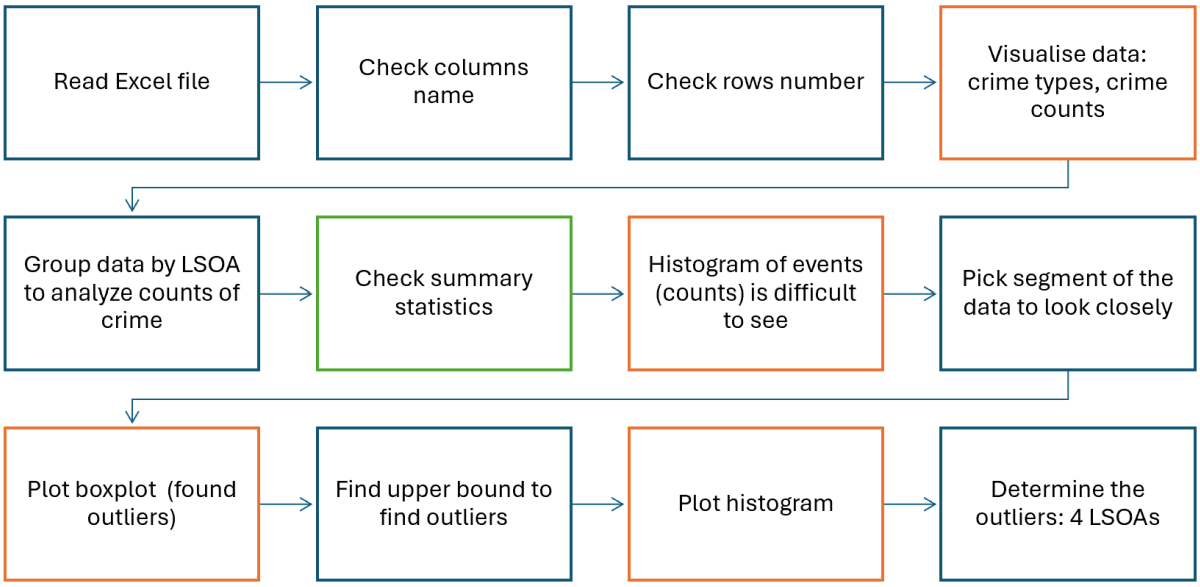


Figure 5. Workflow of analysis in R Studio

In the first visualisation, the "Crime types" bar graph confirms that the "Violence and sexual offences" is the most reported crime. However, the histogram of crime counts by LSOA is more difficult to read.

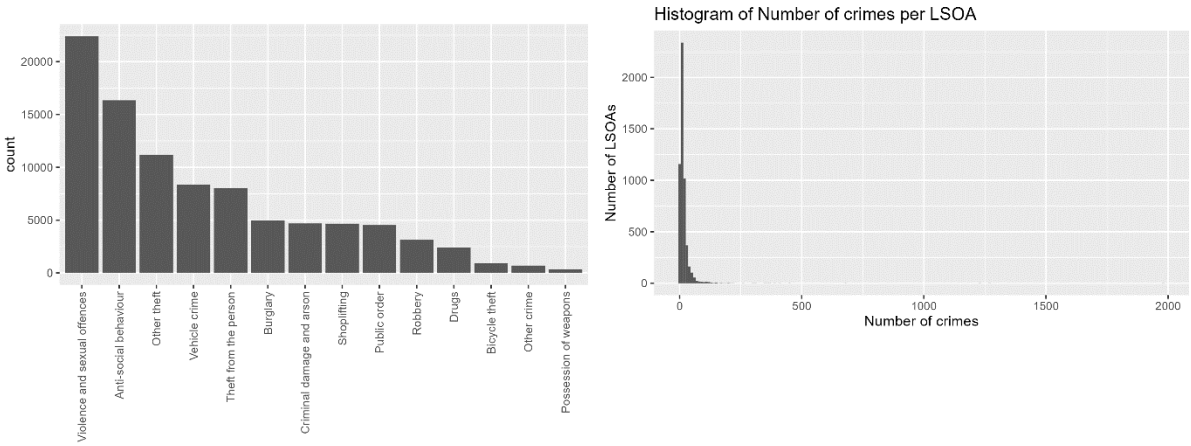


Figure 6. Count of the "Crime types" attribute (Left). Number of crimes per LSOA (Right).

A separate data frame was created to summarize the number of reported events by LSOA, and the statistics were checked (Table 2). Here is visible that between the 3rd quantile and the maximum number there is a great difference. So, the higher 25% of the data needs to be looked at closely. The histogram of this was still hard to read, but a boxplot showed that many outliers were interfering, especially the values larger than 500.

Table 2. Summary statistics of the number of crimes per LSOA

Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
1	6	11	17.4	19	1981

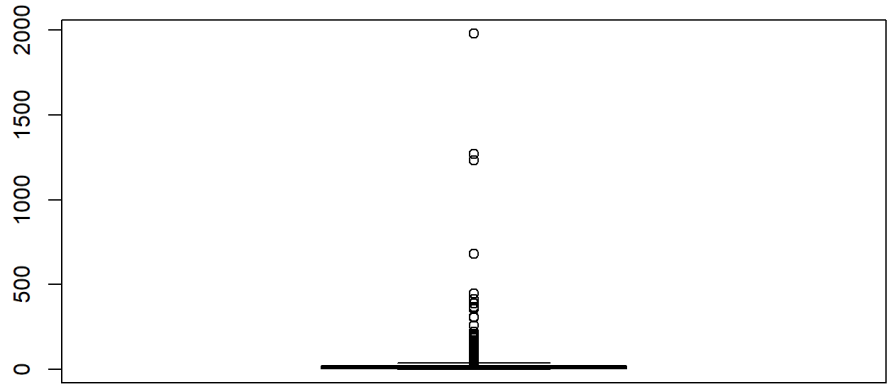


Figure 7. Boxplot of crimes reported by LSOA.

To find the outliers, histograms are produced with different sections of the data: the histogram of the top 2.5% higher values and then again, the top 2.5% without values larger than 500. In the latter the data distribution was clearer (Figure 8), so the 4 LSOAs with highest number of crimes as outliers were determined.

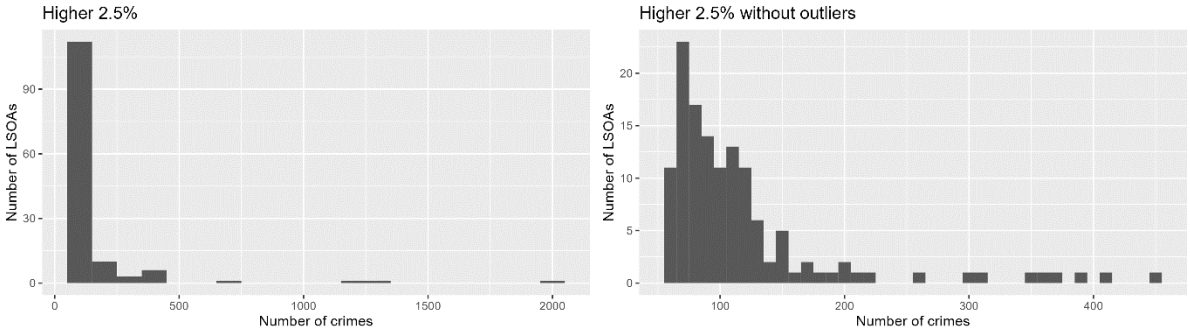


Figure 8. Histograms of higher values.

The data distribution for this analysis is shown in Table 3. The 4 LSOAs with the highest number of crimes are the outliers, then the group of 131 LSOAs, that represent the top 2.5% of LSOAs, with very high crime numbers, followed by 1,149 LSOAs, that represent the top 25% without the highest 2.5%, with high numbers of crimes reported. The first 75% of LSOAs with medium or lower numbers of crimes reported will be considered as a single category.

Table 3. Data distribution

Number of LSOAs	Number of crimes	Statistical significance
381	1	Minimum value (first 7.15%)
245	11	Median
3660	2 - 19	Between the 2.5% and 75%
1149	20 - 63	Between the 75% and 97.5%
135	64-1981	Above the 97.5%
4	679 - 1981	Outliers (larger than 500)
5325	1 - 1981	Total

Spatial visualisation in QGIS

Applying the graduated symbology to the map of crime events in QGIS we find that there is an option by quantiles. But as we know, it doesn't show how the data is really distributed (Figure 9 left). There is another option "Natural breaks" that represent better our statistical results, but the algorithm is not entirely clear, and the numbers don't match exactly in the legend (Figure 9 right), so it was decided to manually style the map. The outliers are highlighted in red to differentiate better from the other areas with very high crime.

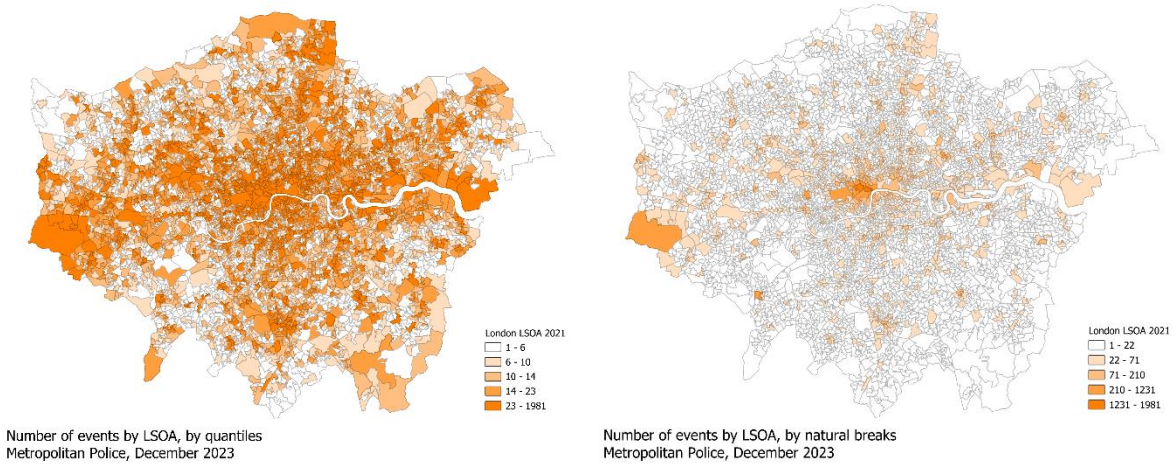


Figure 91. Map of London with number of crimes reported. Graduated symbology in QGIS.

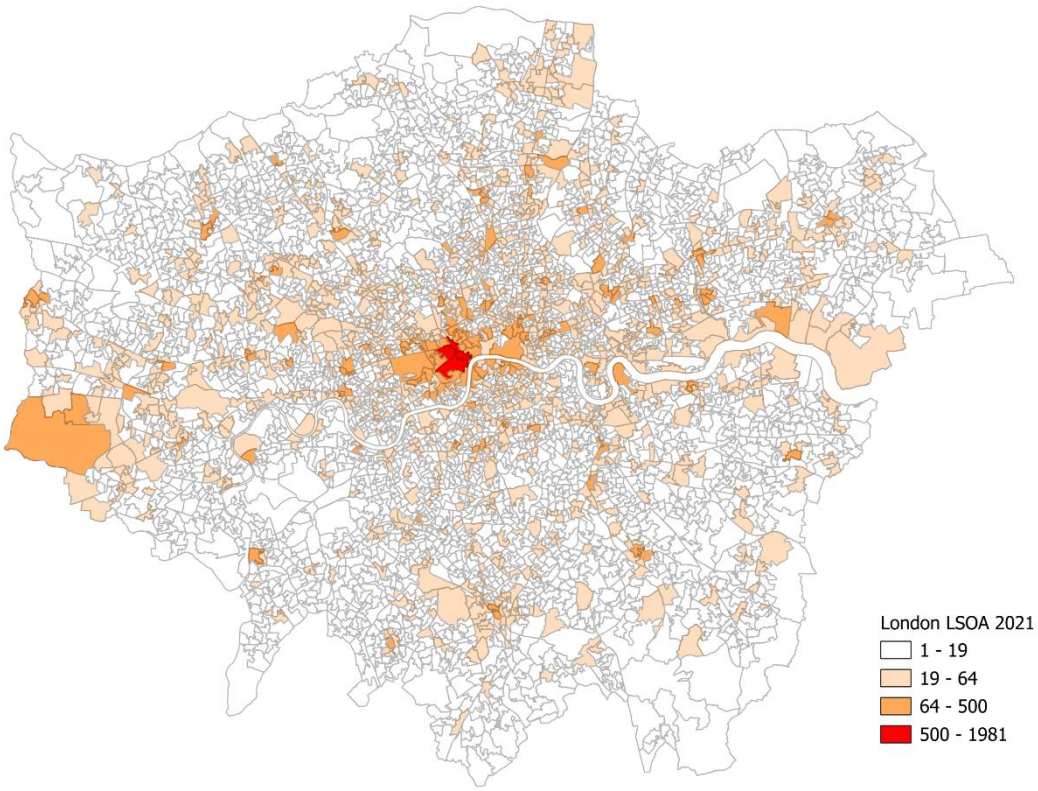
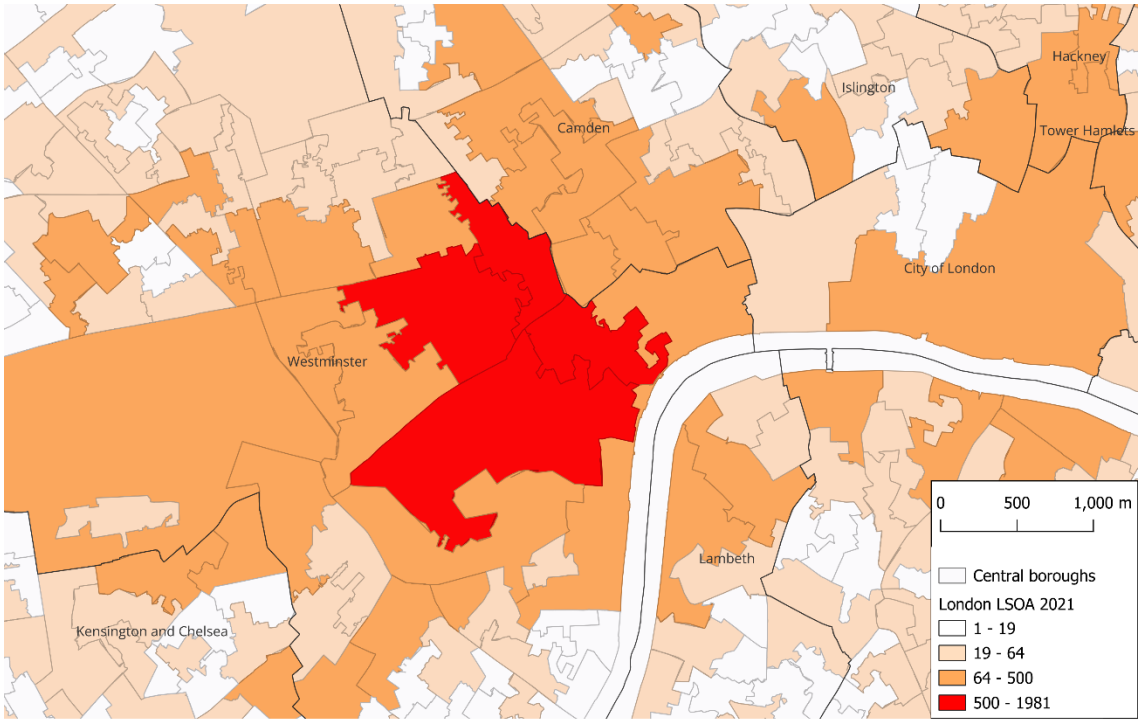


Figure 10. Map of London with number of crimes reported in December 2023, by LSOA. Graduated showing the distribution of the data.

With this map (Figure 10) we can see that the outliers we found are together in central London. There are other areas with a very high number of crimes reported, mostly in central London, and others dispersed in other boroughs.

Looking closely at the LSOAs with the highest numbers of crimes reported, we can add a layer of boroughs now, to see they belonged to Westminster, but neighbouring Camden and in some extent Lambeth (Figure 11).



Visualisation of the LSOAs in London with the largest number of reported crimes.
Metropolitan Police, December 2023

Figure 21. Area with the highest number of crimes reported.

In Figure 11 we can see the difference in LSOAs even in central London. As we know, Westminster has many commercial and institutional premises, with a lower number of inhabitants, this makes it look larger in size. Other nearby LSOAs look much smaller, but they concentrate a higher number of inhabitants. The number of crimes also varies greatly in LSOAs even in the same borough, which we would not have been able to appreciate analysing the data in boroughs.

If we compare the LSOAs we must be careful of the ecological fallacy, that is to think that all areas in red have the same level of risk. If we take a closer look at the area, mapping the events in clusters, as in Figure 12, we can see that actually the crimes are concentrated in the upper area in red, and not so much in the southern part of the red area.

3. Results and Discussions

This analysis highlighted many of the issues stated in Johnson (2017). Here is a summary of the results:

- Points map or choropleth/thematic map: The points map was discarded because the first area of analysis was the whole of London, and the points were too many to show the distribution of points. After the analysis, it is recommended that some areas of Westminster can be studied at a closer scale, in which the location of the points would give a better understanding of why there are so many crimes reported in this area. A further analysis of spatial clustering could be used as well.
- In a thematic map, how to divide the areas: by borough, MSOA or LSOA? As this is an exploratory study, the LSOAs were chosen as the unit of analysis. Their relatively small size allowed for a more granular view of the data's variations. LSOAs also have similar populations, providing a degree of comparability despite potential differences in shape. After the analysis is also visible that the variation of crimes reported between LSOAs is greater than anticipated so a further analysis, could involve locating the LSOAs in neighbourhoods or other geographical features. For instance, in the west there is a big LSOA with a very high number of crimes reported. We know that this area contains Heathrow airport, which explains its size and the high amount of people passing there. So, another study could cross the locations of transportation hubs with crime data and identify if there is any correlation.
- By count or by rate? In this analysis the data is shown by counts, instead of by rate, because it's a basic exploration. This showed that the counts vary in great measure, even when the chosen areas to visualise (the LSOAs) are similar in population. A more useful study would be to consider the type of crime reported divided by the population at risk (called a rate). For instance, if we are going to investigate house burglaries we can divide the reported burglaries by the number of residences in each area. Johnson (2017) also states the difficulty of working in city centres where there is a great number of visitors compared to the number of inhabitants, so we cannot use the rate of crime with the resident's population, we need the visitors or floating population data.
- Centroids or graduated map? The centroid map was difficult to implement in LSOAs even in the smaller scale, because the shapes are too irregular, and the geometric centroids appear in places difficult to associate. This would be more useful in the borough level, but then again, to have more significance, it would need to be crossed with other information or just show one type of crime. The graduated map was more useful but to choose how to graduate it, a further look into the dataset had to be done, so the message is visible and has statistical significance.
- One obstacle of working with LSOAs was that for the size of Greater London, QGIS was slow to process, so it took some time to produce the final big map (Figure 10). To work with the data for one year and see how it changes, for instance, the large amount of data would need to be processed in another program first to summarize it or to change the software.

4. Conclusion

The exploratory analysis arises from the challenges of working with large datasets like the monthly crimes reported by the Metropolitan Police. This volume of data raises numerous questions and presents researchers with a multitude of decisions before they can effectively visualize a problem or tell a story through an image, graph, or map.

In this report, the analysis was made in the position of a data scientist, which goal is to report anomalies in the data, not to explain the underlying issues in these anomalies. To further analysis we would need theory, or literature review, to take informed decisions about which other data sources to use to corroborate our hypotheses.

A good amount of time was spent figuring out how to truthfully represent the dataset, and how to work with the different software (R Studio and QGIS). As stated, the work wasn't lineal often requiring me to start again. However, these iterations led me to develop the workflow diagrams, as self-help guides for future projects.

5. References

Johnson, S. D. (2017). Crime mapping and spatial analysis. In R. Wortley, & M. Townsley, *Environmental Criminology and Crime Analysis* (pp. 199-223). New York: Routledge.

London Datastore (no date). *Statistical GIS Boundary Files for London*. Available at https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london (Accessed: February 2024)

Metropolitan Police (no date). *Data Downloads.* Available at https://data.police.uk/data/archive/ (Accessed: February 2024)

Open Geography Portal (no date). *Lower layer Super Output Areas (December 2021) Boundaries EW BFC V8*. Available at https://geoportal.statistics.gov.uk/datasets/bb427d36197443959de8a1462c8f1c55_0/explore?location=51.473249%2C-0.048998%2C10.00 (Accessed: February 2024)