TRAN5340M

Module Name: Transport Data Science

Assignment Title: Finding the safest place to live in London

Student ID: 201484484

Word Count: 3450

Lecturer: Dr Robin Lovelace

Submission Date: 28/05/2021

Semester: 1

Academic Year: 2020/21

UNIVERSITY OF LEEDS

Finding the safest place to live in London

1. Introduction

Because everyone is different, crime will affect them in different ways. Knowing that a crime was committed on purpose by another person is one of the most difficult aspects of dealing with it. People who commit crimes, unlike accidents or illnesses, want to damage others. The consequences of crime can continue a long time, regardless of how "severe" the act was. These days there are many crimes are made around the world. We can try to avoid the impact of crime by transport data science. In the United Kingdom (UK), the Government made a collection of data of the past crime with the geographical location of London.With the help of these historical data, we will find the safest place in London to live.

Using transportation data science techniques, this report aims to contribute to this research area. The increasing availability of data and sophisticated analytical and modelling techniques have made it possible to address such transportation problems using data science. Although the focus of this report is on London, the code has been designed to be reproducible and scalable for larger areas.

1.1 Scope

The primary goal of this report is to examine the UK police crime dataset in London between 2020-10 and 2021-03, with the goal of informing the safest place to live. We'll start by describing the methodologies for comprehending, preparing, cleaning, and visualizing data. The scope is to understand the data, and by using different methodologies, we will compare each and every place. Then, using faceted heatmaps, I'll look at crime frequency and type per location over time. Also, with the help of the map, we will plot the crime take placed in London using longitude and latitude in the dataset so that we can virtually see the location on the map.

1.2 Area of Study

London is the capital of the United Kingdom, is a twenty-first-century metropolis with a Roman past. The city is located on the River Thames in England's southeast, at the mouth of its 50-mile (80-kilometer) estuary that leads to the North Sea. London, which was initially known as Londinium, has been a major settlement for over two millennia. One of the important global cities in the world is London. Arts, commerce, education, entertainment, finance, healthcare, media, professional services, research and development, tourism, and transportation are all influenced. London is home to diverse cultures and people, with over 300 languages spoken in the area. It has a population of 9 million people in mid-2018 (equivalent to Greater London), making it Europe's third-most populous city. London is home to 13.4% of the population of the United Kingdom. The Greater London Built-up Area is Europe's fourth most populous.

1.3. Datasets

The primary data used in this report were the crime data from UK police available from "data.police.uk". I am using data for six months between October 2020 to March 2021. It provides crime, outcome, and stop and search data at the street level, with the 2011 lower layer super output area (LSOA). It contains a crime type column that will give information about what type of crime the crime id does. The main advantage of this dataset is that it has latitude and longitude with is very helpful in this report. This dataset contains

only LSOA, but we need Middle Layer Super Output Areas (MSOA) to analyze, so I download another dataset that contains LSOA codes dataset with MSOA codes. This will help to create a new column MSOA in the main table. Providers table provides the open street map for the analysis (data, n.d.).

2. Data Understanding and Pre-Processing

It was necessary to first understand the raw datasets and pre-process them into forms suitable for subsequent analyses before analyzing the data. In this part, we will understand the data and prepare the data for analysis.

2.1 Understanding the Datasets

We can understand the data by looking at the names of the column in the table, that will give a piece of brief information about the table. Here the crime data table (td) contains 11 columns. Also, the table has the location of the crime that takes place, type of crime, last outcome of the crime, LSOA code, and LSOA name. There are six separate datasets for each month. Each month has its own records. Crime type has information about the different types of crime made in the city. This table has a column called location, which tells the place where the crime takes place. There are approximately 5,40,000 rows in all six tables. Most of the columns in the dataset are character class.

2.2 Data Preparation

Before processing and analysis, raw data is cleaned and transformed, and this process is known as data preparation. It's a crucial step before processing that often entails reformatting data, making data corrections, and combining data sets to enrich data.

After understanding the data, the next step is to prepare the data for the analyses. We got six months of crime data from January 2020 to March 2021. First, we will import the CSV file from the file location where there are stored. For that, I am using read_csv function. This function will import all six individual tables. Once this was done, we need to combine all six tables into one so that we can analyze them all together. Rbind function will help to combine the tables with the respective column. And the new table is td. Also, we will import lsoa_msoa table because the main table has only the LSOA number and name, but we don't have the MSOA number and name. From lsoa_msoa table, we will get the MSOA name using the LSOA number.

Lsoa_msoa table has MSOA names with the suffix of some number. We don't need that number because it increases the unique count of the name. So we will remove the last three characters of the word using str_sub function.

The name of the column is more important because it is easy to pick a column when the column name is informative. So we will change the column name to informative and easy to type, which is a name without space. Also, there is an unwanted column called context. We remove this column by subset function.

NULL value makes the analysis more bug-prone and difficult. This will miss leads to wrong output. So we will remove all the null values from the table using the function na.omit. This will remove all the rows containing null values.

As we don't have MSOA in td table, we will add a new column from lsoa_msoa table by matching the lsoa number of both tables. So that we will get 33 unique values. That is 33 places in London.

3. Heat-mapping of total crimes in London

A heat map is a two-dimensional data visualization tool that displays the magnitude of phenomena as color. The color fluctuation might be via hue or intensity, providing the reader clear visual indications about how the occurrence is clustered or evolves over time (wikipedia, n.d.).

3.1 Methodology

The demographic information available in the crime dataset is crime type, location, and month. We use the function ggplot() function from ggplot2 package to plot the heatmap. This heatmap will provide insights of the crime rate of the msoa. To do that, we will create a new table by grouping the msoa and month of the td table and count the frequency of the crime made in the respective place for all six months. Using police_grid table we will plot the heatmap with the month as the x-axis and msoa area as the y-axis. The heat of the map will be based on the number of crimes (.r-graph-gallery, n.d.).

3.2 Results and Discussions

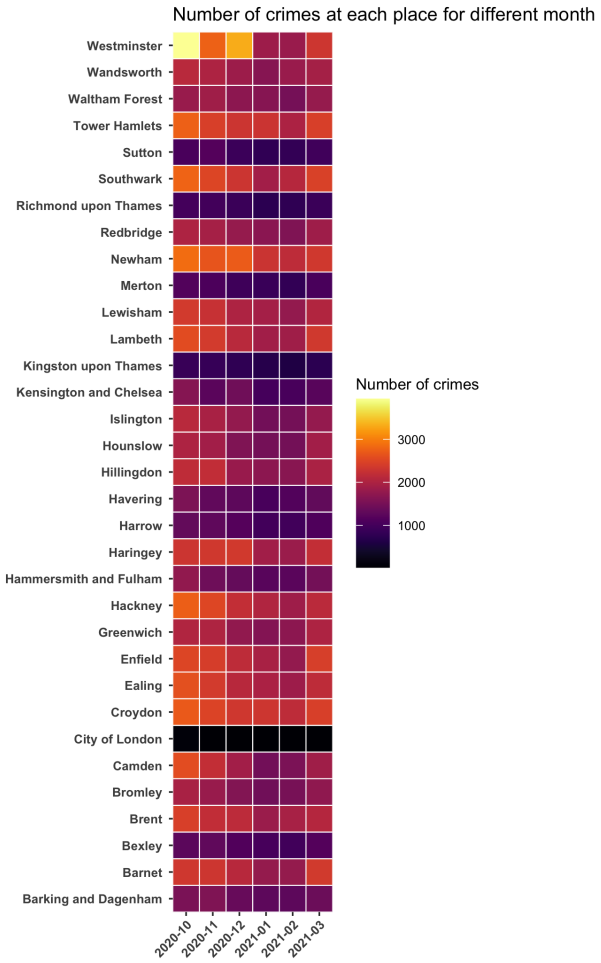

Figure 1: Number of crimes at each placr for different months

The resulting plot in the figure shows the number of total crimes of each place for all the months. The Colour of the map tells the number of crimes in that place. We can see that city of London row is entirely black this means it is the least affected place by crimes. In the month of October 2020, Westminster has the highest number of crimes, but in the following month's crime rate gradually decreases. So police force over there has taken some action to control the crime rate. Also, we can see that Newham, Southwark, and Tower Hamlets have a high crime rate in all six months. So, the police force over there is should need to take some action to control crimes. Kingston upon Thames, Richmond upon Thames, Sutton's police force is good that for the past six months compared to other places, crime take place is very less. In the month of October, we can see the crime rate in all the places is a bit high compare to other months.

4. Heat-map for type of crime in London

4.1 Methodology

Now with the help of msoa name and crime type in the crime dataset, we can find which place has which type of crime. This will helps to develop the program to control it. So we will group the table by msoa name and crime type using group by function. Also, count the frequency of crime and stored it in a new table called crime_grid. Now with the help of ggplot2 package, we will plot the crime_grid table using ggplot function. Here crime type will be the x-axis and msoa area will be the y-axis, and the frequency of the crime fills the color. This heat map will show how the places are affected based on the crime.

4.2 Results and Discussions

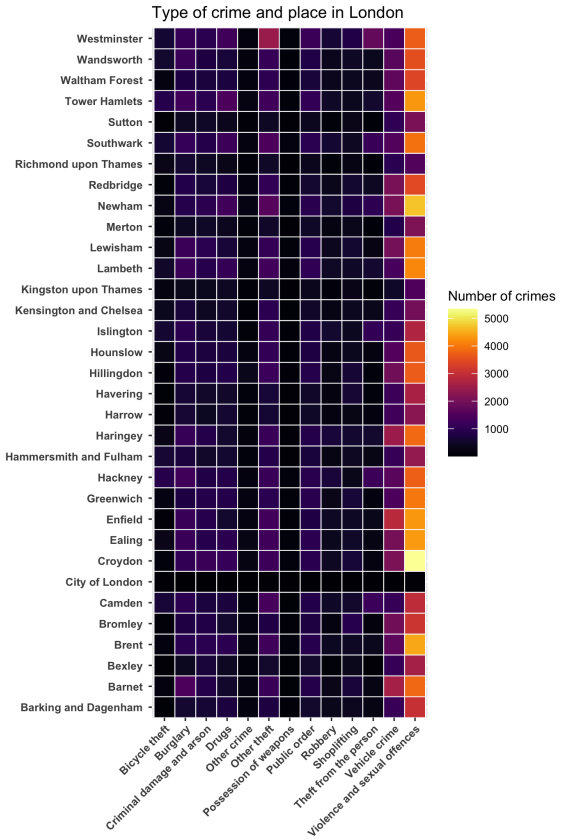

Figure 2: Type of crime and places in London

Heatmap showed above shows which areas tend to have more crime and which don't. Another informative plot. This shows clearly that violence and sexual offenses are high in every place in London. In it, croydon has the most rate of violence and sexual offenses. So police need to focus on this type of crime in every place. We can see Kingstone upon Thames row is completely black, which means the crime rate over there is very less. Bicycle theft is very less in all the city. Violence and sexual offenses is dominating other crimes in barking and Dagenham, which shows the violence and sexual offenses is two times higher than the other type of crime.

Robbery in these places is very less, where other crimes are dominating this. The robbery rate is less than 1000, which we can assume that people here are safer from robbers, and also the police have taken some action for this type of crime.

5. GEOGRAPHIC HEAT MAPS OF ALL CRIMES

A heat map is a geographical depiction of data that emphasizes the density of the data using a color-coded rubric or another type of map key. Warmer colors, such as red, are used to signify higher-density areas, while cooler colors, such as blue, are used to represent lower-density areas (espatial, n.d.).

5.1 Methodology

I am going to plot the location of each and every crime that is committed in London using longitude and latitude in the td table. Plotting on the map will give the best view that we can see the place of crime. First, we will convert longitude and latitude in td table into numerical and store the mean of both in center_lng and center_lat so that we can fix the map size according to the data. This heat map will show how close/far from each other. We will use the leaflet function to plot the heat map (studio.github, n.d.). This function is most popular for interactive maps. The Leaflet map widget is returned by the method leaflet(). It holds a list of items that may be edited or updated later. Most functions in this package take an argument map as their first argument, making it simple to utilize the Magritte package's pipe operator % > %. Providers table has different types of maps which can be selected using tab key after Doller sine. Size of the map is set by center_lng and centr_lat in setview(). Using the clusterOptions feature, we can provide information about the number of plotted crime points. The heat map is produced by a radius of 5.

5.2 Results and Discussions

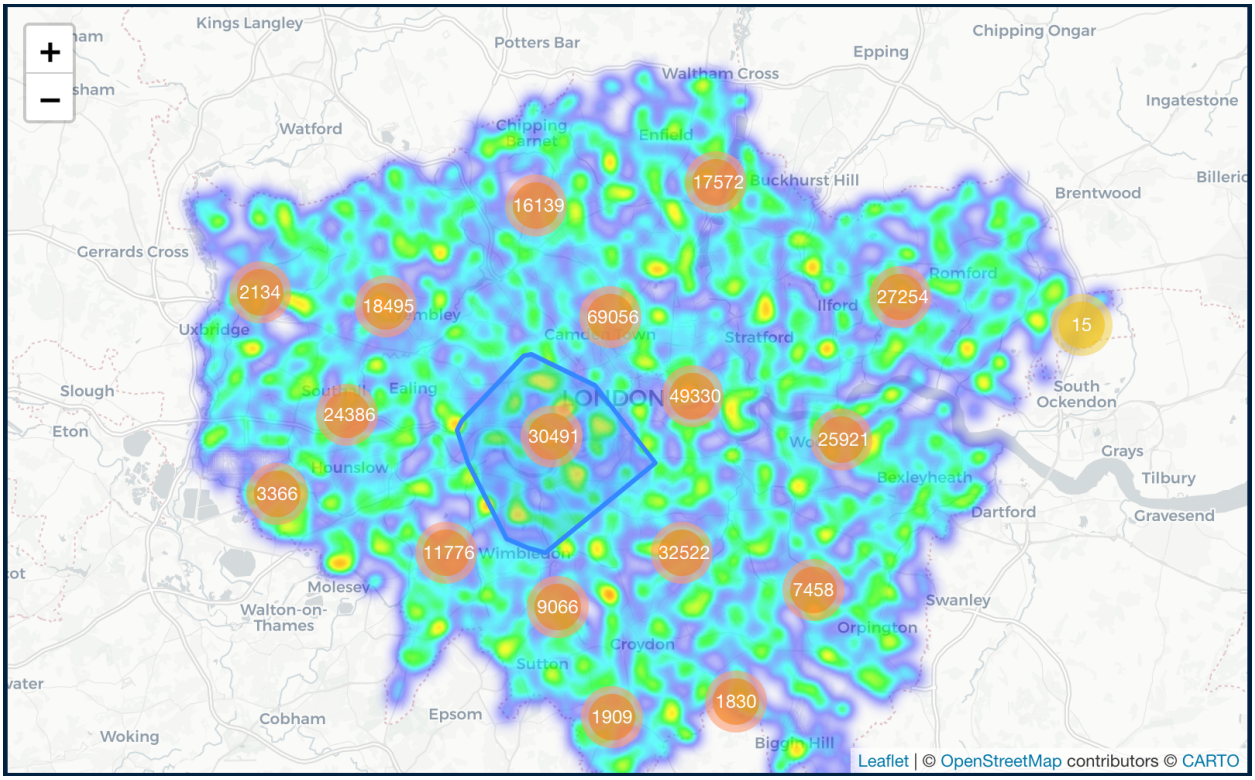

Figure 3: GEOGRAPHIC HEAT MAPS OF ALL CRIMES

When this function is run, we will get the heat map in an interactive map that we can zoom in and see where the crimes are committed. The map shows me how the crimes are spread over the area. The whole data of six months is plotted. We can see that almost every place is covered by heat, and the cluster in the middle has a more numerical count. It is an interactive map, so while zoom in, we can see the cluster gets split into many. Cluster in Camden town has the highest numerical number which shows this place is most affected by crime which is located very near to the city of London. Also, North Ockenden has less crime rate because it is located far from the center and located in the corner of London.

6. GEOGRAPHIC HEAT MAPS FOR ROBBERY

6.1 Methodology

Here we mainly focus on the robbery in march month. For that, we will filter the td table, which is equal to 2021-03 and also crime type is equal to robbery so we will get filtered table called march_robbert. as before out shows the result of all crime that is made in London but now heat map is displayed only for robbery show that we can find the best place to live in London without robbery. We will use leaflet() function to plot in the interactive map. The size of the view map is set using cernter_lng, and center_lat. These points will act as the center point of the map. To popup the number of the crime of each place in the map we use addcirclemarkers() which will show the location and number count in the location. Finally, the heat map is plotted using addheatmap() function with a radius of 5.

6.2 Results and Discussions

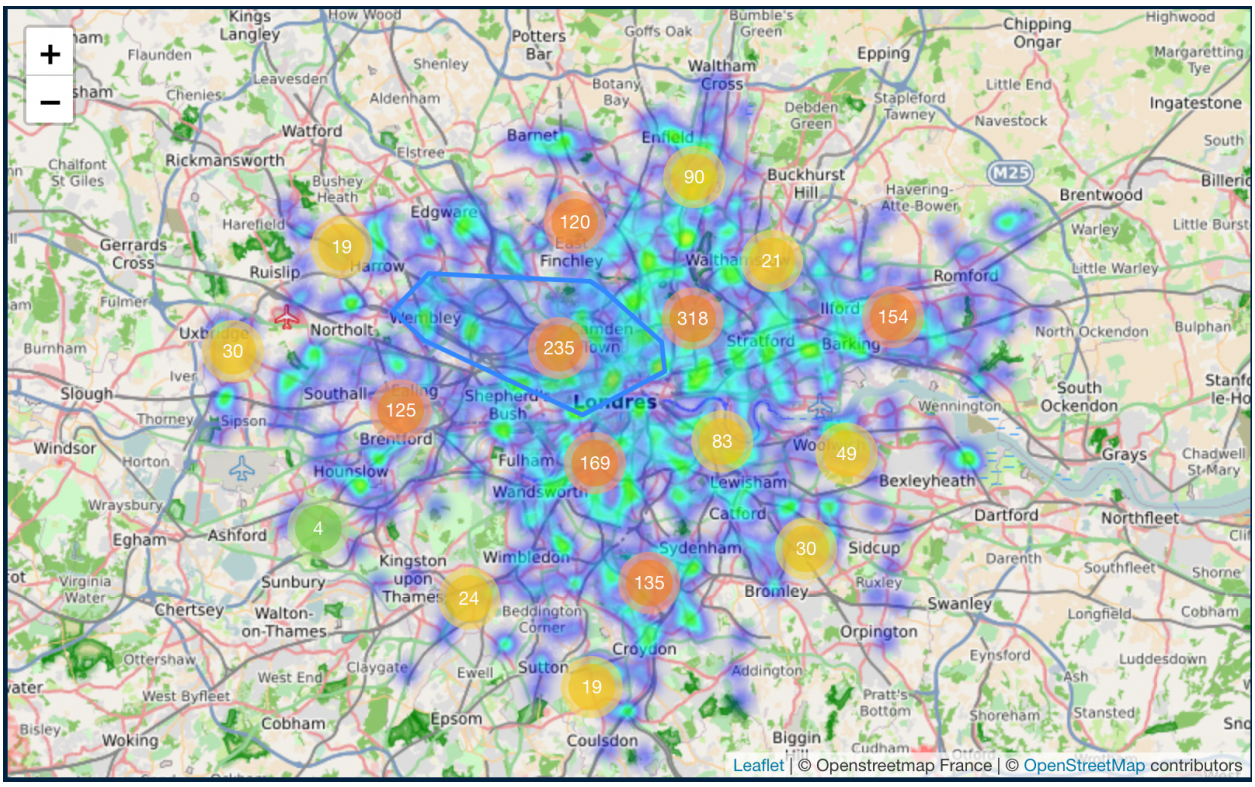

Figure 4: GEOGRAPHIC HEAT MAPS FOR ROBBERY

Robbery in march 2021 is plotted on an interactive map. Where we can zoom in and see where the robbery is committed. by seeing the result in march month, the center part of London has high crime rate than the outer part of London. The heat of the map is high around the city of London. The cluster near Sunbury has the lowest count, which means this place is safe from robbers. This will the best place to live who are sifting to London.

7. HISTOGRAM FOR TPYE OF CRIME IN LONDON

A histogram is a graphical representation of data that uses bars of varying heights. Each bar in a histogram divides numbers into ranges. More data falls inside the range as the bars get taller. The form and dispersion of continuous sample data are represented by a histogram (khanacademy, n.d.).

7.1 Methodology

From the total dataset td we will filter the msoa column and crime_type column and stored them in the td1 table. To plot a histogram for every place in London, first, we need to know how many places are there in London. For that, we use the unique function in msoa. This will give the unique places in London. Then to count the frequency of the number of crimes in each and every place, we use the table function on msoa crime type column. Finally, the data is ready to plot. We use ggplot function to plot the histogram from ggplot2 package (r-graph-gallery, n.d.). In ggplot, x-axis will be crime type, and the color is filled based on the crime type, and the bar is raised by the count of the crime in each place. To display all the places at the same time facet_wrap function is used. This will display all histogram in the same panel which makes easy to analyze.

7.2 Results and Discussions

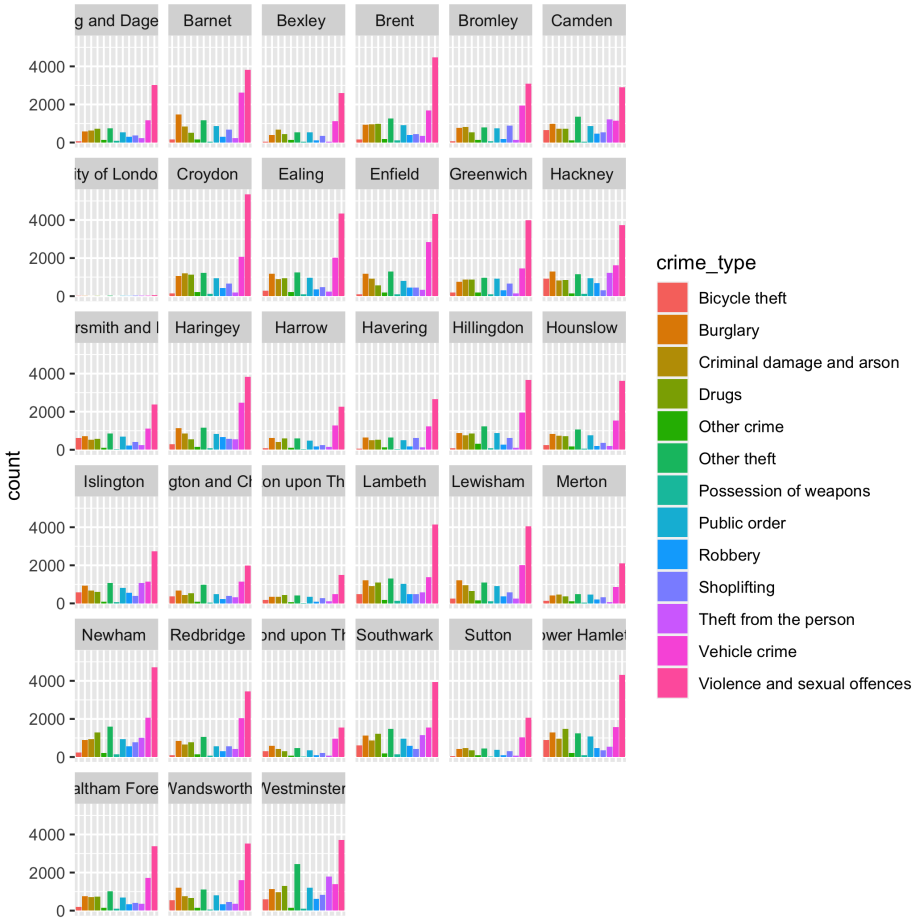

Figure 5: HISTOGRAM FOR TPYE OF CRIME IN LONDON

In this figure, the histogram is plotted for each place in London. Each place has different histogram plotting. Bars in the histogram refers to the type of crime in that particular place. Hight of the bar depends on the count of the crime in that place. Seeing the histogram data violence and sexual offenses is high in all the place in London. This shows police should consent to violence and sexual offenses activities and develop some laws to bring the count down after that vehicle crime is the next highest crime bar in most of the places. So that we can assume that in these places, there are a lot of cars are parked in the non-surveillance area, or this place has fewer police patrols. Bicycle theft is the lowest bar among other crime bars in every place. We can assume that bicycles may be less in number in these places or police force is more active for the bicycle in these places. Data displayed in the histogram is of six months.

8. HISTOGRAM FOR ROBBERY

The act of unlawfully stealing property from a person or place via force or fear of force is known as robbery. The number of robberies in the area tells how safe the site is. So to choose a safe place to live, we need to analyze the data of robbery separately.

8.1 Methodology

In this modeling, we will polt the number of robberies that take place in each and every place of London. We will filter one march month's robbery from the td table and store it in the march_robbery table. Using this data, we are going to plot the histogram. We will use ggplot function from ggplot2 package. Here let msoa be the x-axis, and the color of the bar is based on the msoa. each bar will have there own color, which refers to the unique msoa location. geom_histogram function is added to ggplot, which will plot the histogram with the count of the robbery.

8.2 Results and Discussions

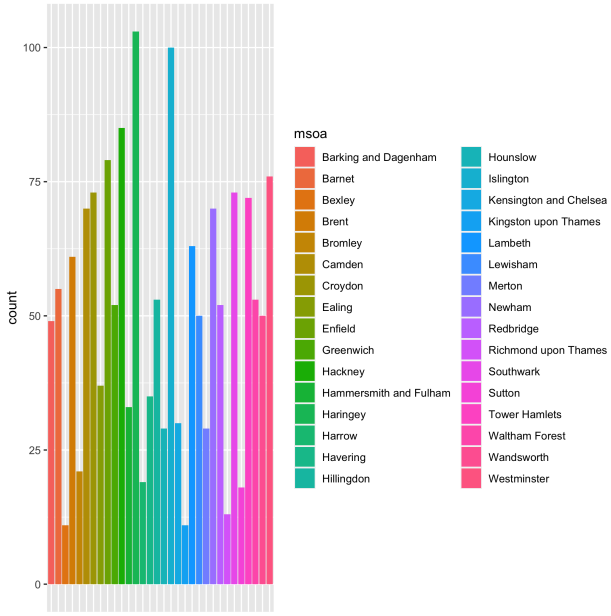

Figure 6: HISTOGRAM FOR ROBBERY

This model shows the histogram of the robberies count in the different areas (msoa) of march month. In this, we can see that bar of Haringey is tallest than other bars, which shows this area is recorded more than 100 robberies are made in this area. So police should take some action to control this crime. Similar to it bar of Islington also tall. Robbery rate in these two places dominating other places. from this chart we can see Kingston upon Thames and Richmond upon Thames are the safest place from the robbery because their bars are too small that means the count of crime there is significantly less.

9. Limitations of Analyses

The conclusions reached in this research were not without flaws. We'll focus on major drawbacks in this section. The first is about the primary dataset, which contains only LSOA numbers and names. There are 4830 unique values of losa, which is too high to analyze based on the place so that I found the MSOA table from google, which contains the lsoa number column and msoa name and number. msoa will cluster the lsoa into few areas. By matching the column lsoa number in both tables, we will add a new column msoa in the main table.

Another drawback is that shapefile. I tried to find the shapefile to analyze it on a map, but I can't find the perfect shapefile with longitude and latitude. Also, I tried to download a shapefile from an open street map which I can download only a small area. Because of this shapefile problem, I tried to plot it using the leaflet function, which contains a provider table and has different type maps inbuilt. So by giving the center point and distance, we can get the interactive map that can be used for analysis.

10. Conclusions

In their early twenties, a substantial number of people from all around the country go to London. This is due to the diversity and quantity of career options available in the capital. So they are looking for a safe place to stay in London. from this report, we can say that all over London crime is happening, but there are someplace with is has the lower crime rate. Geographical heatmap of crime data of October 2020 to March 2021, which helps a lot in this analysis which shows perfectly where the crime rate is low. All above modeling shows that Kingston upon Thames and Richmond upon Thames are the safest place compared to others.

References

.r-graph-gallery. (n.d.). .r-graph-gallery. Retrieved from .r-graph-gallery: https://www.r-graph-gallery.com/79-levelplot-with-ggplot2.html

data, U. p. (n.d.). UK police data. Retrieved from https://data.police.uk/data/open-data/

espatial. (n.d.). espatial. Retrieved from espatial: https://www.espatial.com/create-heat-maps#:~:text=A%20heat%20map%

khanacademy. (n.d.). khanacademy. Retrieved from khanacademy: https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/histograms/v/histograms-intro#:~:text=A%20histogram%20is%20a%20graphica

r-graph-gallery. (n.d.). r-graph-gallery. Retrieved from r-graph-gallery: https://www.r-graph-gallery.com/histogram_several_group.html

studio.github. (n.d.). studio.github. Retrieved from studio.github: https://rstudio.github.io/leaflet/markers.html

wikipedia. (n.d.). wikipedia. Retrieved from wikipedia: https://en.wikipedia.org/wiki/Heat_map#:~:text=A%20heat%20map%20(or%20heatmap,clustered%20or%20varies%20over%20space.