Public Transport Behaviour in Italy

By- Mosammat Mehrin Fatima Hoque

Contents

Introduction

The paper studies the findings from urban mobility simulations from vehicles involved in public transport in the city of Rome, Italy, and provides useful knowledge for identifying and solving problems in a city's road system.

The dataset has 4 columns named DriveNo, Data_and_Time, Latitude, and Longitude. All the columns are useful for the analysis of the behavioural patterns of the taxis. The DriveNo column is the unique number assigned to every taxi driver- this helps in distinguishing individual taxis and finding their specific behaviour. The Latitude and Longitude column specifies the coordinates of the taxi driver at any given time. The time the coordinates are noted and is given by the column Time_and_Date.

To recognize the behaviours of the vehicles, their movement is first visualized. The data points in taxi.csv dataset are plotted on top of a world map by merging the values of longitude and latitude of the world map and the taxi dataset. R programming was used for the analysis. Some of the packages used to employ the world map are maps, mapdata, and rworldmap.

The vehicles being examined have only travelled within Rome, a city in Italy. To plot those on top of the entire world map would produce a diagram that has all the useful information in a very small section and a very large segment of unnecessary data. The dataset also has some outliers and indicated that some taxis have travelled outside of Rome. Hence, the entire Italy was chosen from the world map to plot the values to include the outliers, noise, and unexpected travels on the initial figure.

Visualization

The package ggplot2 was used to plot the data. The plot produced was 665.8 MB which is too large for R to produce with the memory allocated for a single vector. The memory limit for a vector was increased to be able to hold the plot.

The geom_polygon() function was used to show the movement of the taxis in form of polygons.

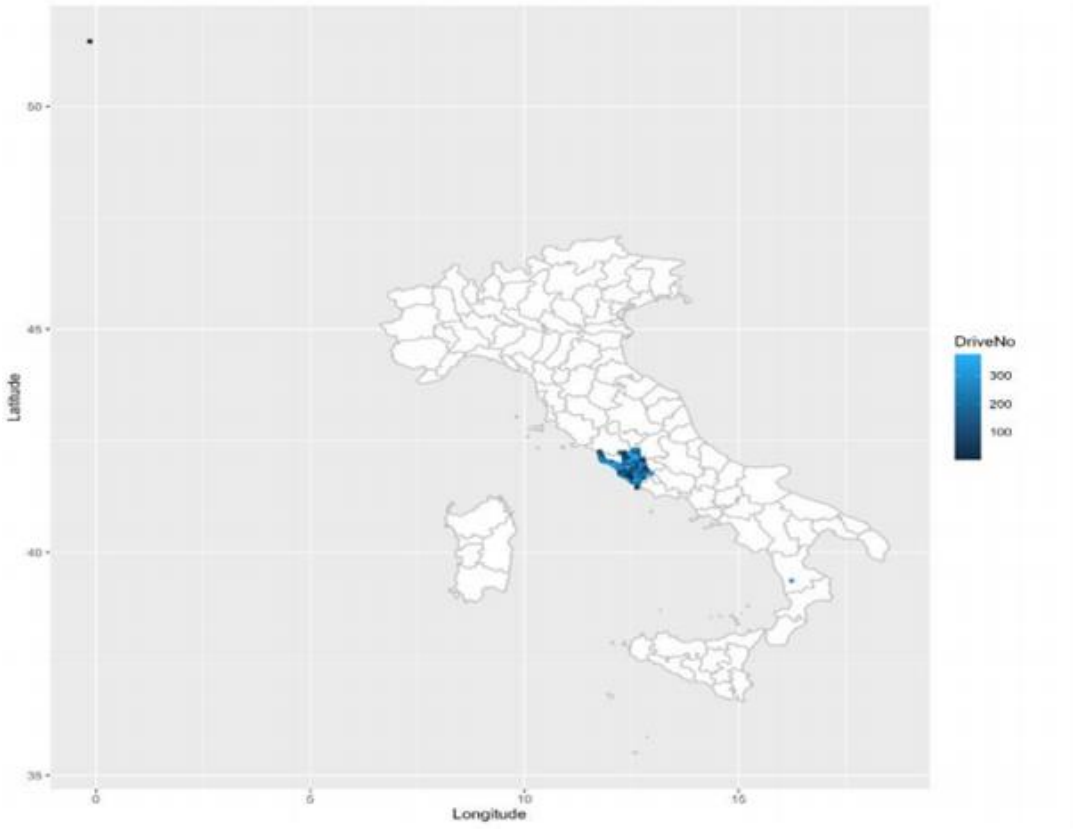

Figure 1- Movement of all taxis visualised with a ggplot

The Grammar of Graphical plot or ggplot above depicts the movement of the vehicles in form of polygons. The legend on the right corner can be used to identify each polygon with its individual driver.

Data Preparation

The latitude of Italy ranges from 36.71703 to 46.99623 and the longitude from 7.05809 to 18.37819, and the latitude of Rome, Italy is 41.902782, and the longitude is 12.496366.

The range of Latitude in the taxi dataset is 39.36232 to 51.45488. Hence, the maximum values of latitude are clearly beyond the borders of Rome. The range of Longitude is -0.1453681 to 16230832. Once again, the values are beyond the limits of the coordinates of Rome. This is indicative of the presence of outliers in the dataset.

Outliers and noise-

-All negative longitudes are noise as they are not in Italy- the latitude of Italy lies between 40 and 50 degrees.

- The longitude of Rome does not go below 11.5-degrees or exceed 13.5-degrees, any data points beyond that are outliers

- The latitude of Rome does not go below 41.4-degrees or exceed 42.4-degrees, any data point beyond that are outliers.

In figure 1, we can see two outlier points, one around coordinates (39, 16) and another around (51, 0). These outliers are apparent as the dataset is concerned with vehicles located in Rome. Some movements can also be noticed outside of the borders of Rome. But they are fairly close to Rome and within Italy, so they will not be removed. The initial dataset has 21817851 observations.

Removing outliers using Longitude values

- Removing outliers beyond 0-degree longitude removed 18 data points, and the dataset contains 21817833 rows.

- Removing outliers below 11.5-degree longitude and over 0-degree removed 0 datapoints

- Removing outliers beyond 13.5- degrees removed 3 points, this changed the total number of rows for the dataset to 21817830.

Removing outliers using Latitude values

- 18 rows were removed while filtering data with latitude greater than 42.4

- 5 rows were removed while filtering data with latitude less than 41.4

The total dataset after removing all outliers has 21817830 observations.

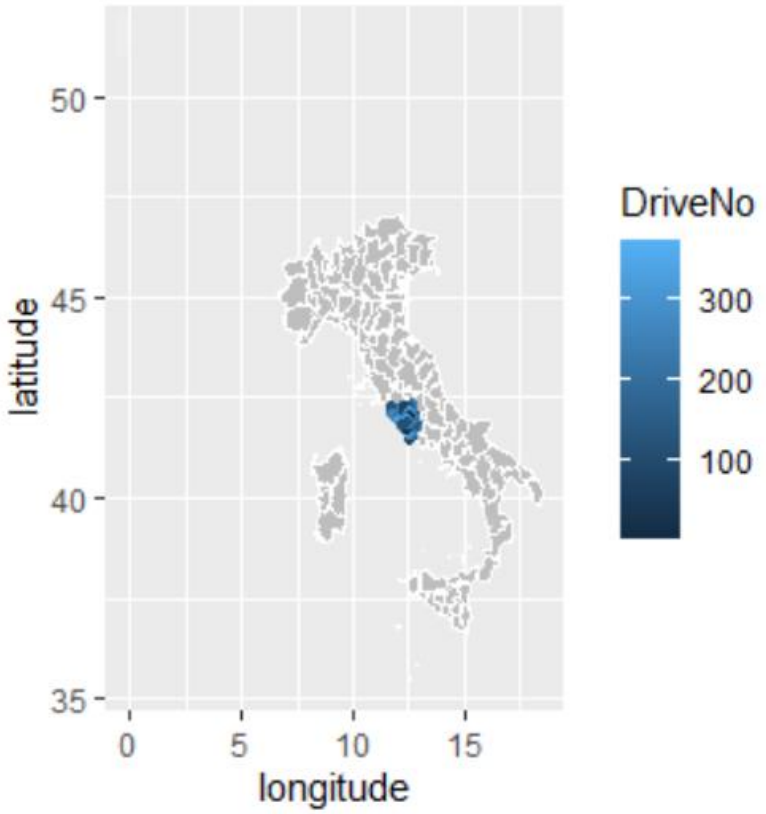

Figure 2 Taxi movement after outliers are removed

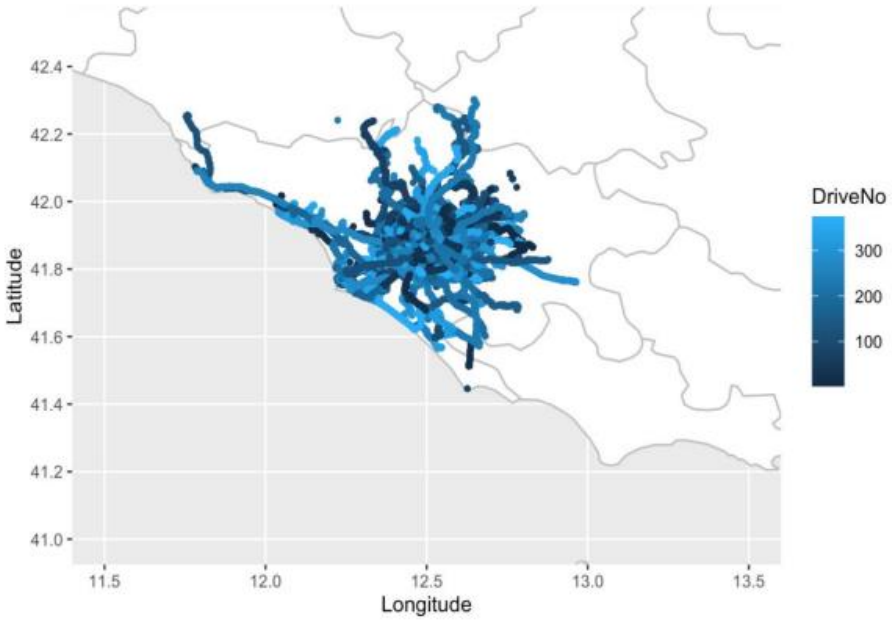

Figure 3 Taxi movement when zoomed in

Analysing Taxi Behaviour

Statistical Analysis of Location

The minimum, maximum, and mean position values have been determined after invalid points and outliers have been removed. The following is the result:

- The minimum location value of Latitude is 41.44596
- The maximum location value of Latitude is 42.30219
- The mean location value of Latitude is 41.89201
- The minimum location value of Longitude is 11.75315
- The maximum location value of Longitude is 12.96347
- The mean location value of Longitude is 12.47264

We can see that all the coordinates lie within the border of Italy and in or very close to Rome.

Statistical Analysis of Taxi Movements

Next, we analyze the time travelled by individual taxi drivers. We are interested in finding the drivers who spend the most time, average time, and least time driving. This will allow us to find drivers with the highest or lowest hours on the road. These stats can be used to predict the number of hours drivers usually spend driving.

To find the total time driven by a specific driver, we use the clean dataset with no outliers and noise and find all the unique DriveNo. This gives us a list of 316 driver IDs. Initially, we started with 320 unique driver IDs; cleaning the data removed 4 DriveNo.

We use the unique driver number to extract all the rows that hold information about the drivers' activities. The Date.and.Time column is used to find the duration of each journey for an individual driver. Summing the duration of all the travels gives the total time driven by the driven in a month's time.

If the duration of a trip is less than 5 minutes or 600 seconds, the trip is considered an outlier. Because semantically a taxi trip does not take less than 5 minutes. These outliers could be caused due to the driver taking a break, going to a different location that has more prospective customers, accidents, etc.

The total time driven by each driver is found. These values are then used to calculate the minimum, maximum and mean duration of the drivers drive for.

- The minimum duration of driving is 82.79806s (outlier)
- The mean duration of driving is 571890.1s or 158.9 hours
- The maximum duration of driving is 974891.9s or 270.8 hours

The result shows the minimum time duration of driving is around 83s which was disposed of as an outlier for reasons mentioned earlier.

The mean duration of travelling is approximately 160 hours a month, or 40 hours a week. This seems accurate as it is the number of hours full-time workers usually work in a week. This indicates that most taxi drivers are full-time workers and are actively working during work hours. One driver seems to have worked 270 hours in the month, which is 67.5 hours a week. We can assume the driver worked double shifts and were on the road more than the drivers working an average number of hours. We can hence conclude that most drivers of Italy work full-time and usual hours, i.e., 5 days a week and 8 hours a day.

Analysing A Specific Taxi

In this section, I will be analyzing the activities of DriveNo 124.

The location points are plotted for DriveNo 124. To do so, all the outliers are once again removed from the data set using the filter() function. A ggplot is developed for visualizing the movement of driver 124 using a polygon.

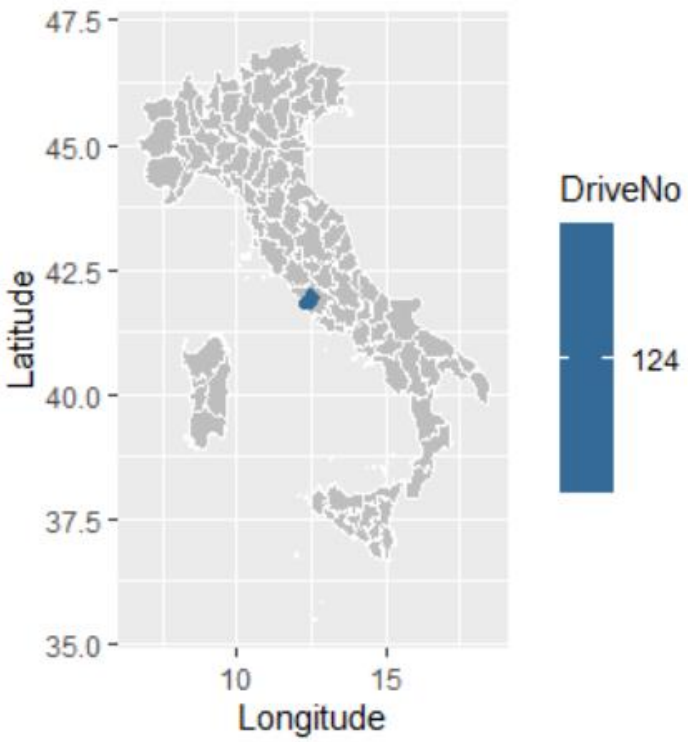

Figure 4 Movement of Driver No. 124 Visualized

From the visual, we can clearly see that driver number 124 is always travelling around central west of Rome. They tend to stay around central Rome and travels around almost half of the city. Also, we can see that the driver has never left Rome as there would be trail marks up to the borders of Rome which is included in the visual.

Comparing Specific Taxi movement to Global Values

	Driver number 124	Global values of drivers
Minimum Latitude	41.78525	41.44596
Maximum Latitude	41.99113	42.30219
Mean Latitude	41.89882	42.30219
Minimum Longitude	12.24892	11.75315
Maximum Longitude	12.61964	12.96347
Mean Longitude	12.48406	12.47264

The minimum and maximum values for both latitude and longitude of driver number 124 significantly differ from the global values. Because the global values contain drivers who travel all across Rome, whereas driver 124 travels around central Rome only.

The mean values for longitude and latitude for driver 124 correspond closely with that the global values. Certainly, the global mean values of longitude and latitude will depict central Rome as the drivers travel across the entirety of Rome and the average of all the coordinates is the centre of Rome city.

We then use a clean dataset with no outliers or noise to find the total time driven by driver number 124, and we split the rows corresponding to DriveNo 124. The filter() function is used to accomplish this.

The Date.and.Time column is used to determine the length of each trip for each driver. The cumulative time driven by DriveNo 124 in a month is calculated by adding the length of all travels.

A trip is considered an outlier if its length is less than 5 minutes or 600 seconds since logically, a taxi ride should not take less than 5 minutes. Outliers may occur as a result of the driver taking a break, moving to a different location with more potential customers, and so on.

The total time is driven by driver number 124 during the period from 1st Feb 2014 until 2nd Mach 2014 is 548911.9s or 152.5 hours. This indicates that the driver works around 38 hours a week.

Comparing these values with the global values of duration of driving shows that driver 124 drives a little less than the global average number of hours in a week, i.e. 40 hours. The difference is very little, and we can deduce that DriveNo 124 drives as much as the average drivers drive and work regular shifts, i.e. 5 days a week and 8 hours a day. There is a vast difference between driver 124's working hours and the minimum and maximum driving hours of global drivers.

Next, we analyse the distance driver number 124 travels in the time period. To do so, we use the difference in latitude and longitude. Mathematical formulae are applied, and the total distance travelled but the driver is found to be 1747153m or 1747.153 km.

As driver 124 travelled for 152.5 hours during the one-month period, his average speed was:

- ⇨ 1747.153km/152.5 hours = 11.46 km/h

This is a very low speed, so we can assume that the driver either travels on busy roads with lots of traffic or enjoys driving safely.