## CISTUP - IISc

## Summer Internship - Round 1

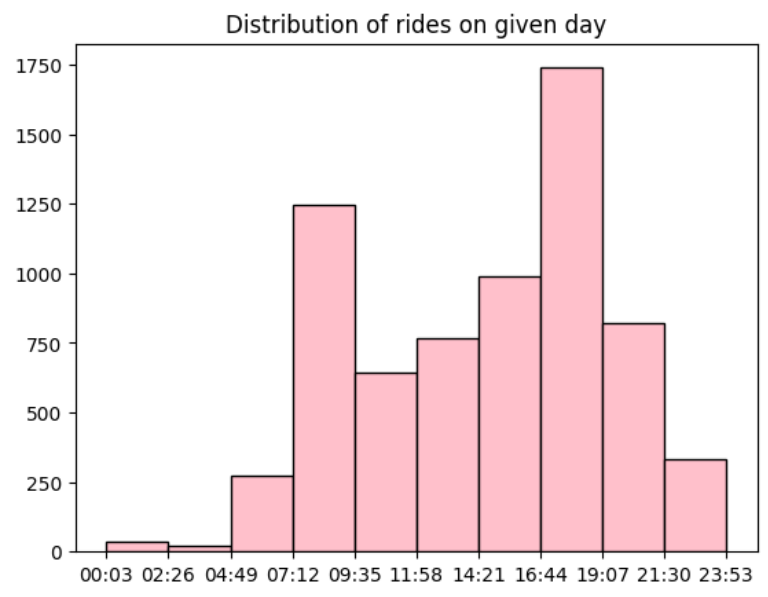
## Test 1(Python) - Report

Name : Renita Kurian

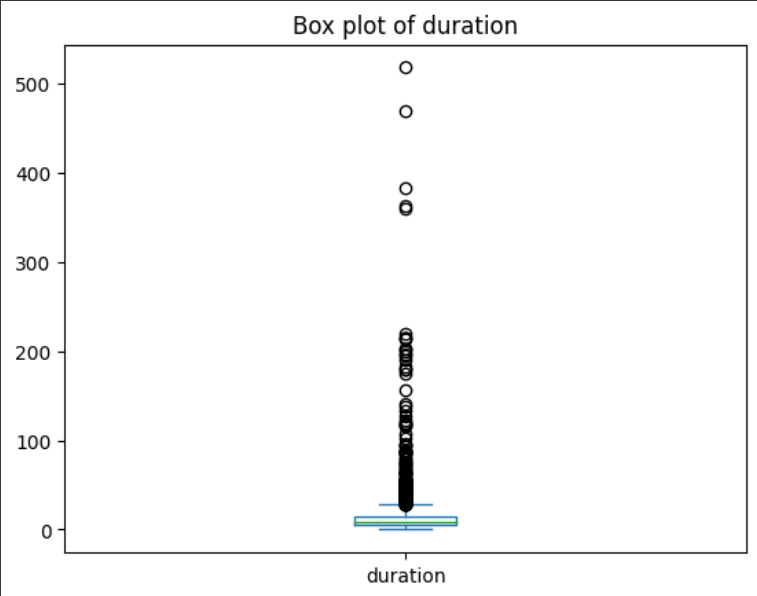
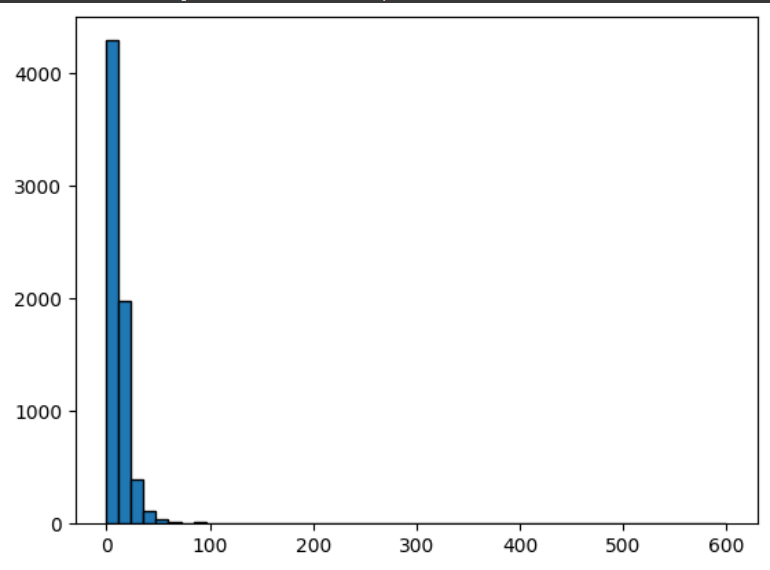
College : PES University

# Question 1

Shown below are a few visualizations obtained from the given bike data -



The above histogram shows the distribution of rides on the given day. It can be seen that there is a lot of demand between 7 and 9AM and also between 4:30 and 7PM.

The above graphs show the box plot and histogram distribution of the duration of rides. It is observed that most rides are between 0 to 50 minutes long and there are very few rides longer than 100 minutes.

1. The values mentioned below are obtained from the python file (attached in google form)

Maximum Duration: 518 minutes

Minimum Duration: 1 minute

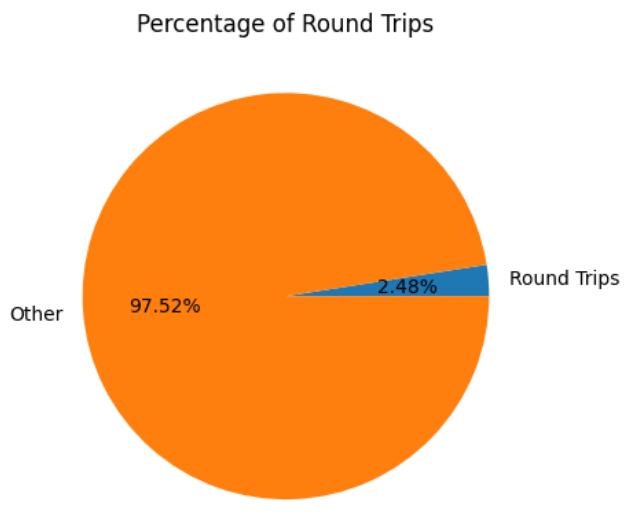
Number of trips with minimum duration: 89

Number of Round Trips: 169

Total trips: 6821

Percentage of Round Trips: 2.48

Runtime of Function for Part I : 194.998 ms



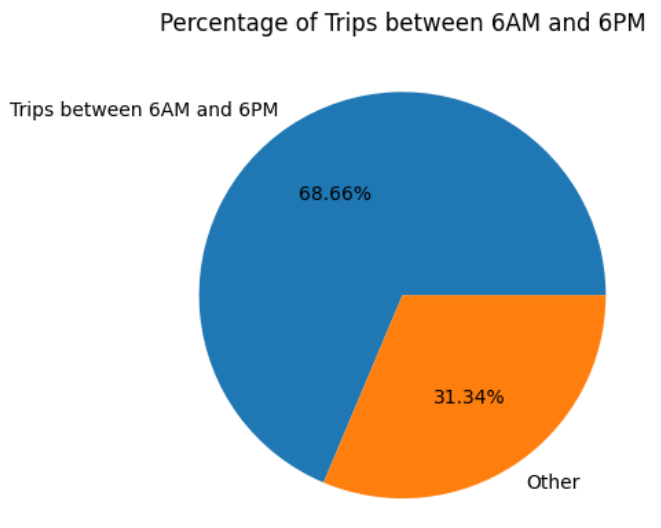
On removing the zero-minute duration trips, there was a total of 6821 records remaining. It was seen that the maximum and minimum duration of trips is 518 and 1 minute long respectively. In the dataset provided, there were 89 trips that were 1 minute long. It was also found that there were 169 round trips, which is approximately 2.5% of the total rides.

The total runtime of the function to calculate above stats is approximately 195 milliseconds.

2. The values mentioned below are obtained from the python file (attached in google form)

Number of Feasible Pairs: 41640

Runtime of Function for Part II: 91854.027 ms



It was observed that roughly 68% of the trips were between 6AM and 6PM.

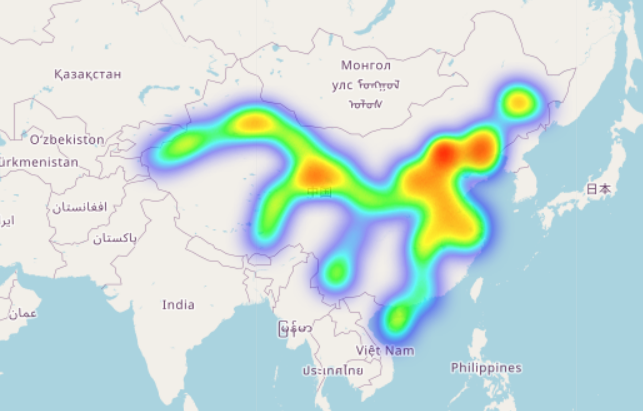
The number of feasible pairs in the filtered dataset was calculated to be 41640.

The total runtime for the filtering and calculation of number of feasible pairs was roughly 92 seconds.

# Question 2

1. The code and output for the first part (total distance for each user) can be obtained from the python file attached in the google form.

2. The pictures given below show the heatmap of the complete data(a) and the heatmap of data limited to Beijing city(b) respectively.

It can be seen from the above heatmaps that the central part of the city has a very high number of data points and the signature gets weaker as we go outward. There are relatively few data points in the outskirts of the city.

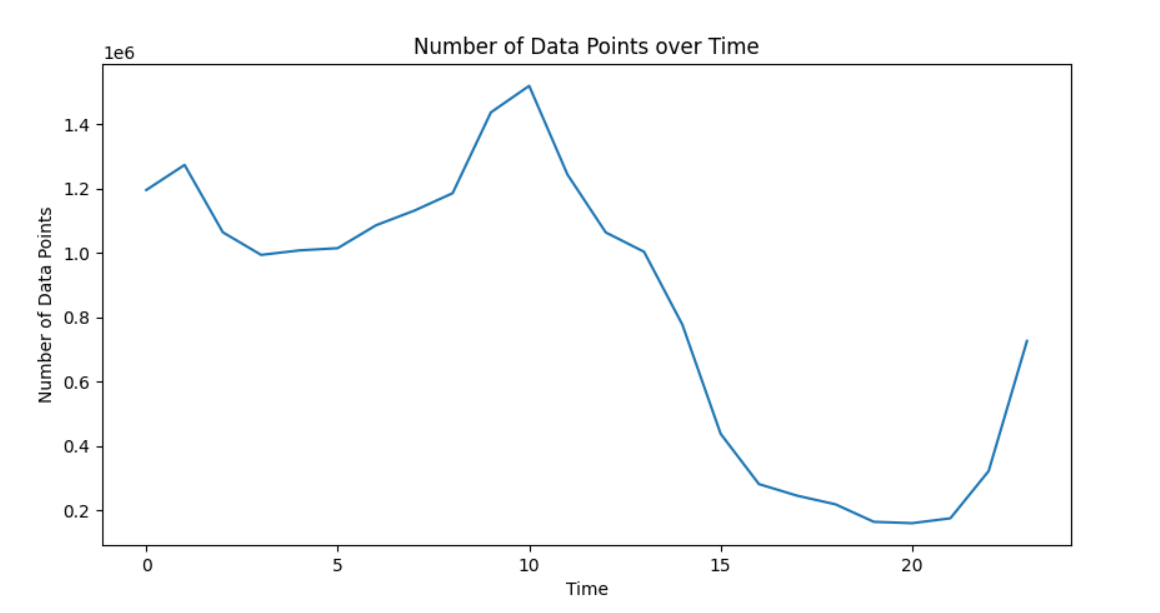
In order to filter the data, the below coordinates were taken as city boundaries –

lat\_min = 39.4428 #southmost lat

lat\_max = 41.0619 #northmost lat

lon\_min = 115.4204 #westernmost

lon\_max = 116.7815 #easternmost



The above graph shows the temporal distribution of data. It can be seen that a lot of GPS points were captured between around 7AM and 11AM. This can possibly correspond to the commute of working individuals.

3.

Given access to a dataset with GPS data, it would be possible to efficiently plan public transport systems or new cities. Finding patterns in the movements of people in a city could help pinpoint the important routes and roads. This could be used to figure the bottlenecks and other issues relating to traffic and connectivity.

GPS data could also be used to find the social hotspots in a given city. A cluster of data points at a given location for a certain minimum duration, consistently seen across an extended period of time could indicate that that location is a social hotspot. Finding the hotspots in different cities could give an insight to the similarities and differences between different cultures. It may also capture certain differences between urban and rural locales. For example, it might be seen that in one city, people tend to spend their evenings at home while in another people like to spend their time at the local parks, restaurants, etc.