The data is available from Los Angeles Metro Bike Share. There are basically two data available from the website. One data consist of information about the bike rides and other data consist of information about the bike station in Los Angeles areas. These data are basically in 4 regions of Los Angeles- Downtown LA, Pasedena, Port of LA, Venice

The bike rides data consist of information about:

* Trip\_id: the unique value of individual trip (no repetition in dataset)
* Bike\_id: the unique value of individual bike used for the trip ( total 1505 different bikes were used for the trip)
* Start\_station: the unique value of individual station from where trip start ( total 141 station were used actively for trip)
* End\_station: the unique value of individual station on which trip ends ( total 143 station were used actively for trip)
* Trip\_route\_category: the two type of trip used-round trip or one way trip
* Start\_time: Time at which the trip begins
* End\_time: Time at which the trip ends
* Start\_lat: Latitude geographical value from where trip start
* Start\_lon: Longitude geographical value from where trip start
* End\_lat: Latitude geographical value on which trip end
* End\_lon: Longitude geographical value on which trip end
* Passholder\_type: 4 type of passes available for the trip. Every pass has individual cost structure. Annual pass for 365 days, Monthly pass for 30 days, Flex pass for 1 day and walk up is not for any particular time
* Plan\_duration: The number of days pass is entitled

The station data consist of information about:

* Station\_id: the unique value of individual station from where trip start ( total 143 station)
* Station\_name: Name of location of station in Los Angeles as per unique station\_id
* Region: The region where these individual station is located in LA
* Go\_live\_data: Date from which the station got activated
* Status: Tells whether station is active or inactive

The bike ride data consist of total 639786 rows and 13 columns. After using isnull function on dataset, the sum of null values in individual column was found as below.

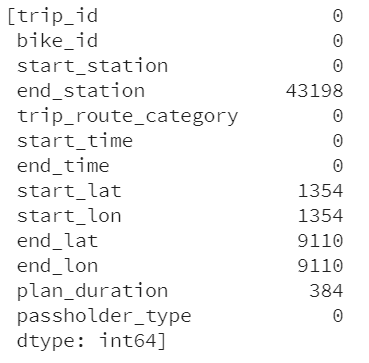


Fig: Null values sum in individual column

For predicting the number of cycles in each station the new variable were created using these column value provided and hence the cleaned data was required for further modelling. For example the distance variable can be defined using longitude and longitude column so that to get distance for individual trip and similarly the time duration variable can be introduced using tart time and end time column so that to get individual trip total duration.

So, all those values which were null in latitude and longitude column were removed. Total 9946 rows were found to be null and were removed. Even though end\_station and plan\_duration column were having null values but not removed because not later it didn’t played any major role in modelling.

After removing null values, the data set was divided into four cluster based on region wise so that to do further prediction or network expanding. This cluster was made in Tableau using world map and after visualizing data it was found that some data were showing point in china and equator. It means some uncertainty in latitude and longitude value and thus removed the rows which were having 0 in latitude and longitude. Those point which were showing in china were simply because of sign issue and then sign was reversed for those points. Then finally visualized the point as shown below.

After removing all the outliers and missing values total 577223 rows were left and then distance and time duration for individual trip was calculated.

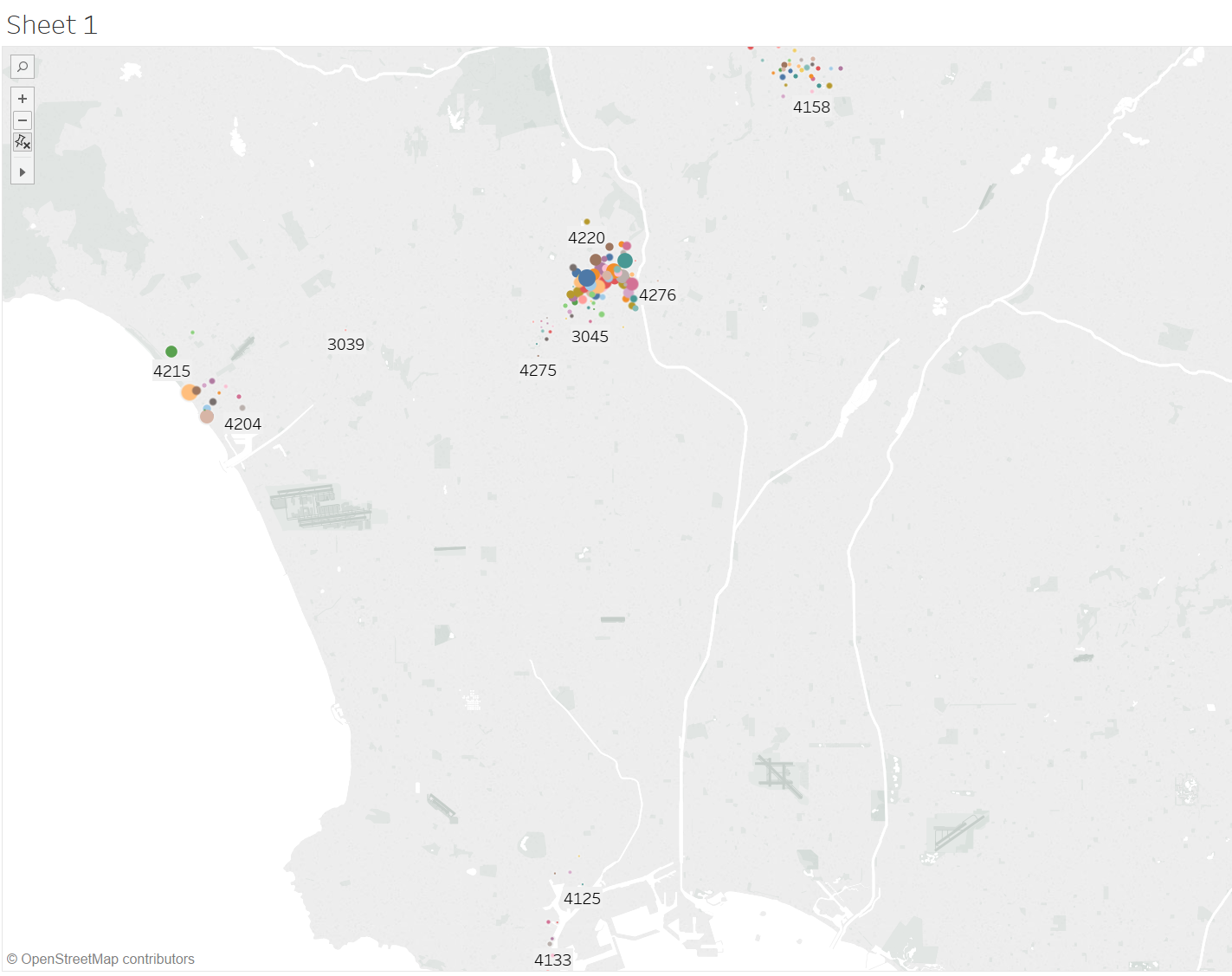


Fig: Cluster of data based on four region

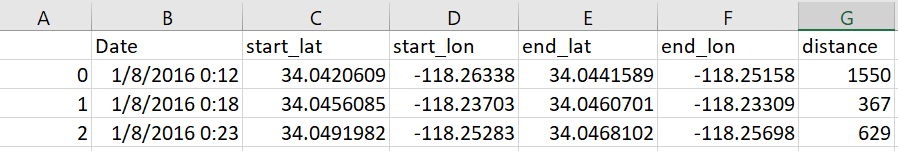
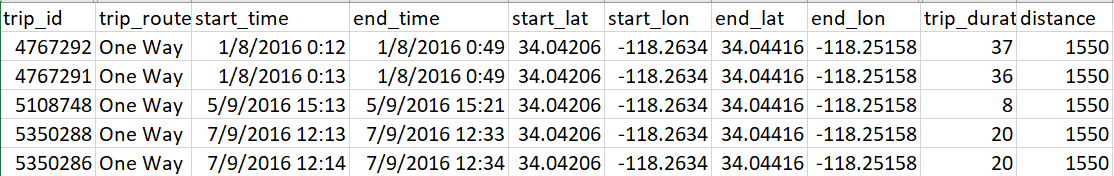
After removing rows which were irrelevant in latitude and longitude distance variable was calculated using Distance Matrix API available in google maps. Initially using distance function predefined in python based on latitude and longitude was used to calculate the distance but then realized that it would calculate the displacement instead of distance for individual trip so to do correction google map API was used. Distance for individual trip was calculated in meters as shown below and then merged to original file for modelling.

Table: Distance column added to each trip

Then trip duration in minutes was calculated using start time and end time data available. Using datetime function in python the time in minutes were calculated for individual trip as shown below so that to take that into consideration for modelling.

Table: Trip duration added to each trip

After calculating duration in minutes it was found that the some value were having negative minutes and then after looking upon data it was found that end time format in some rows were opposite to start time format and that’s why it showed negative minutes. So, all those negative duration were deleted and then found some minutes having large amount of duration. So, normalized the data separately based on one way trip and round trip and set 90 percentile value as upper limit and kept all those values which were less then this upper limit. The result is shown below

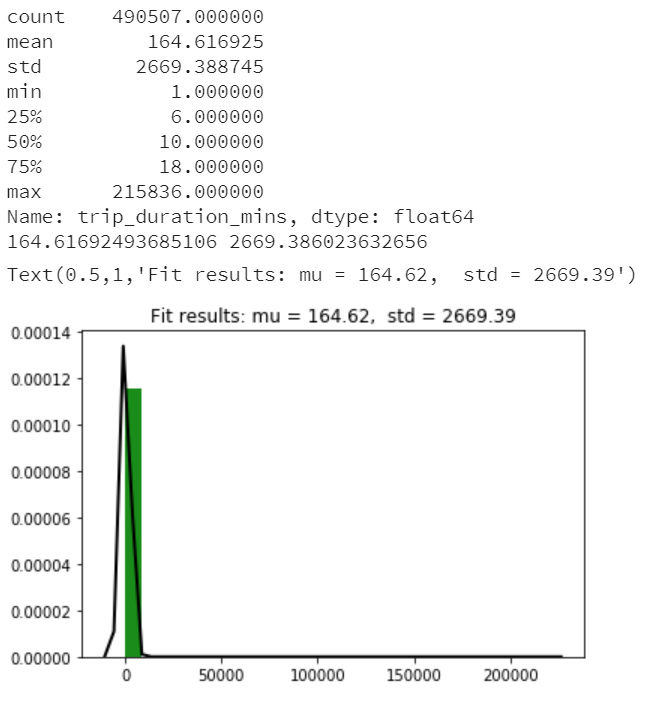


Fig: One way trip normalization

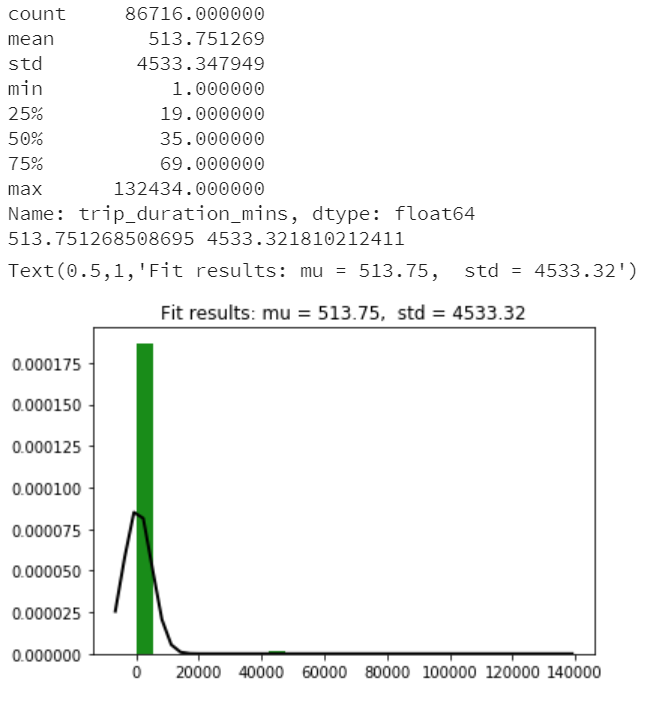


Fig: Round trip normalization

After removing all such upper limit values finally the dataset was made and found to be consist of 521140 rows with new variable distance and trip duration for individual trip.

After this as modelling was required to be done on data starting from July 2016 because after visualizing data based on month it was found that from January 2016 to June 2016 only for 2 days in individual month was present in the data so this showed inconsistency in the time series for further analysis so all rows less than time July 2016 were removed. Finally the dataset after preprocessing was found to be having 513685 rows with distance and duration as extra column. This data now was good to be used for further modelling or visualization.