**LA Metro Bike Share Project**

**Introduction :**

The following project is focused on a analyzing the Los Angles Metro Bike Share data. Los Angles Metro Bike Share operation provides rental bicycles 365 days a year since the 2d quarter of 2016. The main objective of the project is to focus on predicting the number of bicycles needed at each station, evaluating the pricing system currently used to make recommendations for possible pricing changes and expanding the network. Forecasting the number of bicycles needed at each station is done using time series analysis and linear regression. An in-depth evaluation of pricing system was carried out and the linear as well as quadratic optimization algorithms have been used based on the number of riders and the price of each ride to find the optimal solution for maximizing revenue. Pricing model of the competitors have also been considered while evaluating and optimizing the total revenue. Based on the forecasting models, exploratory data analysis and the evaluation of pricing system, recommendations for expanding the network has been made.

**Data Collection :**

The data is collected from the official website of Los Angeles Metro Bike Share. Dataset consist of information about the bike rides and the bike stations in Los Angeles areas. The data represents 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 stations is located in LA
* Go\_live\_data: Date from which the station got activated
* Status: Tells whether station is active or inactive

**Data Preprocessing:**

The bike ride data consist of total 639786 rows and 13 columns. Missing values in the dataset is shown below:

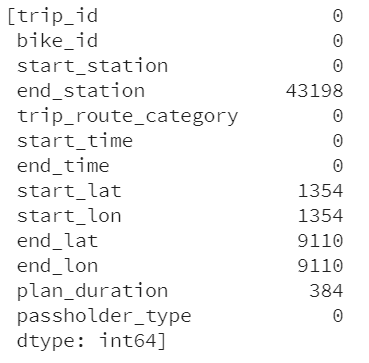


Fig: Null values sum in individual column

New variables were created for predicting the number of bikes in each station. The distance variable was created using longitude and longitude column to get the distance for individual trip and similarly the time duration variable was introduced using tart time and end time column to get individual trip duration.

All the null values in latitude and longitude column were removed. Total 9946 rows were found to be null and were removed.

After removing the missing values, the data set was divided into four cluster based on region to do further prediction or recommend network expansion. This cluster was made in Tableau using world map. After visualizing data, it was found that some data points have latitudes and longitudes in china and on the equator. It indicates some error in data collection so the corresponding rows which were having 0 in latitude and longitude were eliminated. The point which were located in china were simply because of the sign issue in the longitude so the sign was reversed for those points. Finally, the points were visualized as shown below:

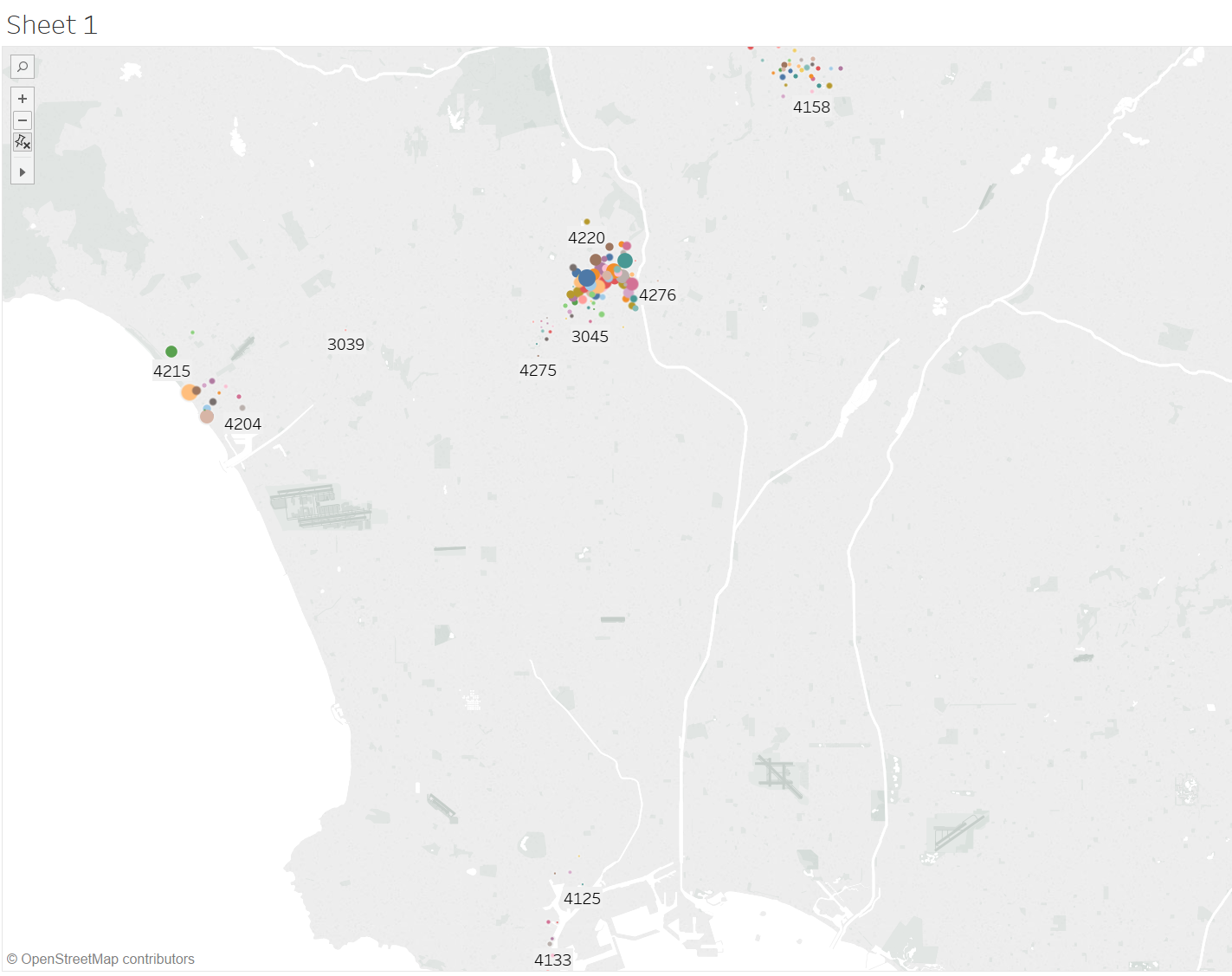


Fig: Cluster of data based on four regions

Outliers and missing values total were removed leaving behind 577223 rows. The distance and time duration for individual trip was then calculated.

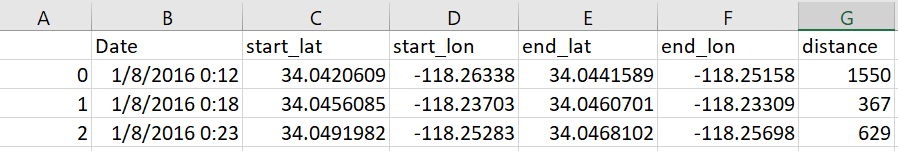
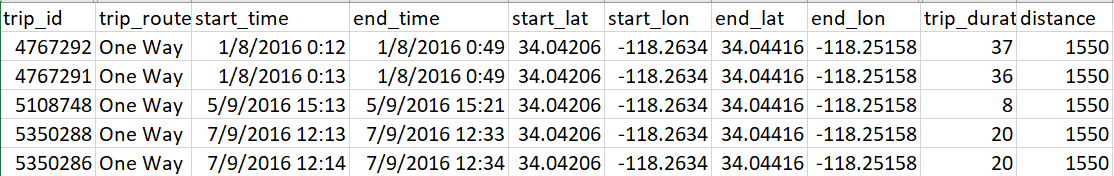
 Distance was calculated using Distance Matrix API available in google maps by setting the mode of transport to bike ride.

Table: Distance column added to each trip

Table: Trip duration added to each trip

Some of the trip duration calculation resulted in negative values. Upon further inspection it was found that end time format in some rows were opposite to start time format. So, all the negative trip durations were deleted. Some values of trip duration were extremely large. So, the durations were normalized separately based on one-way trip and round trip. Upper limit was set at 90 percentile value and the lower limit was 0 minutes. The result is shown below:

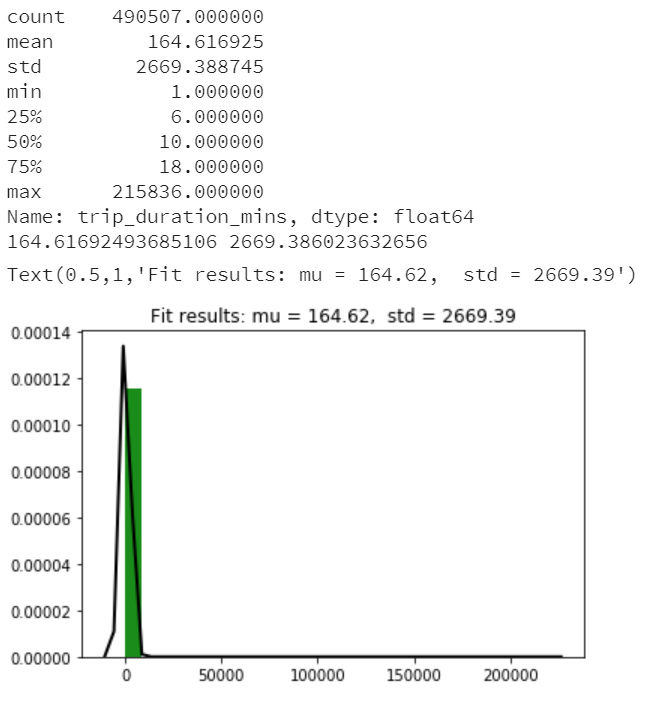


Fig: One way trip normalization

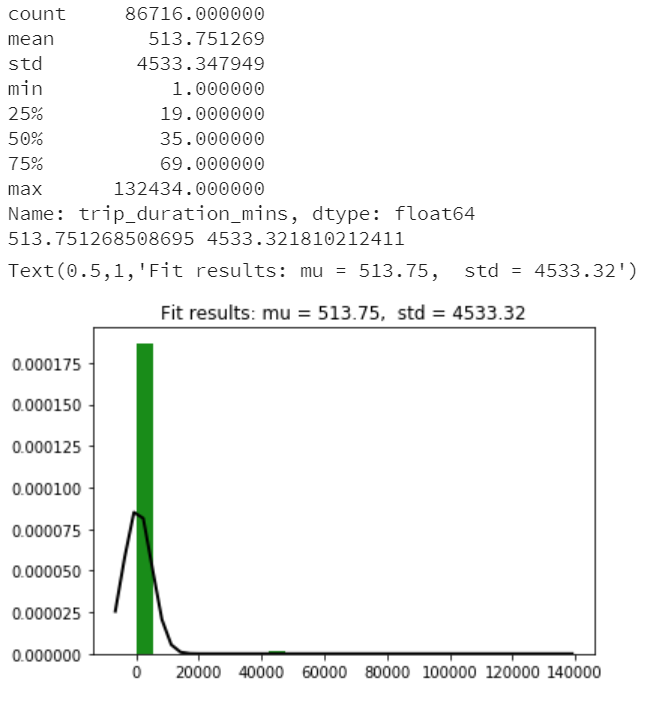


Fig: Round trip normalization

After removing all such upper limit values, the preprocessed dataset had 521140 rows with new variable distance and trip duration for individual trip.

It was also found that from January 2016 to June 2016, data for 2 days were present in each month. So, in order to maintain consistency for time series analysis rows corresponding to July 2016 and earlier were removed. Finally, the preprocessed dataset was having 513685 rows. The cleaned and preprocessed data was used for data visualization and analysis.

**Data Visualization:**

Sayeed

**Forecasting Bicycle Demand:**

**Time Series Analysis**

Time series analysis is a technique of analyzing a series of data collected at different points in time highly correlated to the adjacent points. This restricts the applicability of the many traditional statistical methods which are dependent on the assumption that adjacent data points are independent and identically distributed. Operating under this assumption, a linear regression model trained on time series data may fail to perform.

In this section, an attempt is made to analysis our dataset from the perspective of the Time Series Analysis and compare the results with the forecasting obtained by the regression method. On the onset, an appropriate time series was prepared using the 'trip\_id' and 'start\_time' variables. We have considered 'trip\_id' as a random variable 'X' and 'start\_time' as the timestamp of the event ‘X’. As the first step of all the time series investigation, we have plotted observed data over time.

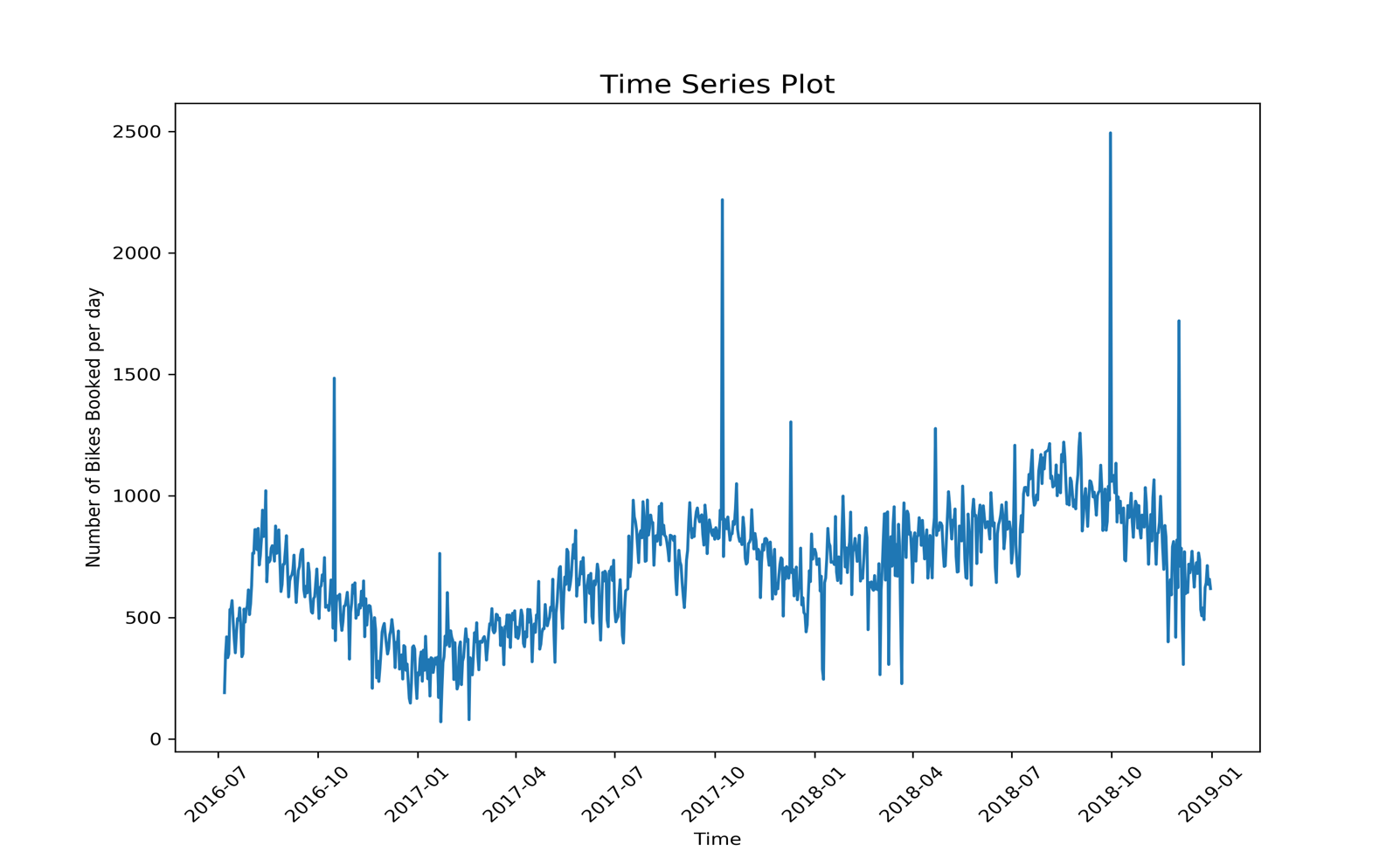


Figure 1 shows total numbers of bikes booked per day. We can see a gradually increasing underlying trend and a regular variation superimposed on the trend that seems to repeat yearly.

Looking at the plot, it is clear that the time series is a non-stationary as the mean of the series is changing with the time (the trend in the plot), the variance of the series is changing with the time (heteroscedasticity in the series), and the change in covariance with time. These are an evident proof that the data is time series which is not a stationary series.

Another way to check the stationarity of the series is the Dickey-Fuller Test. The null hypothesis is that the series is non-stationary. The test results include Test statistic and some critical values for different confidence level. If the 'Test Statistic' is less than the 'Critical Value,' we can reject the null hypothesis and can say that the series is stationary.

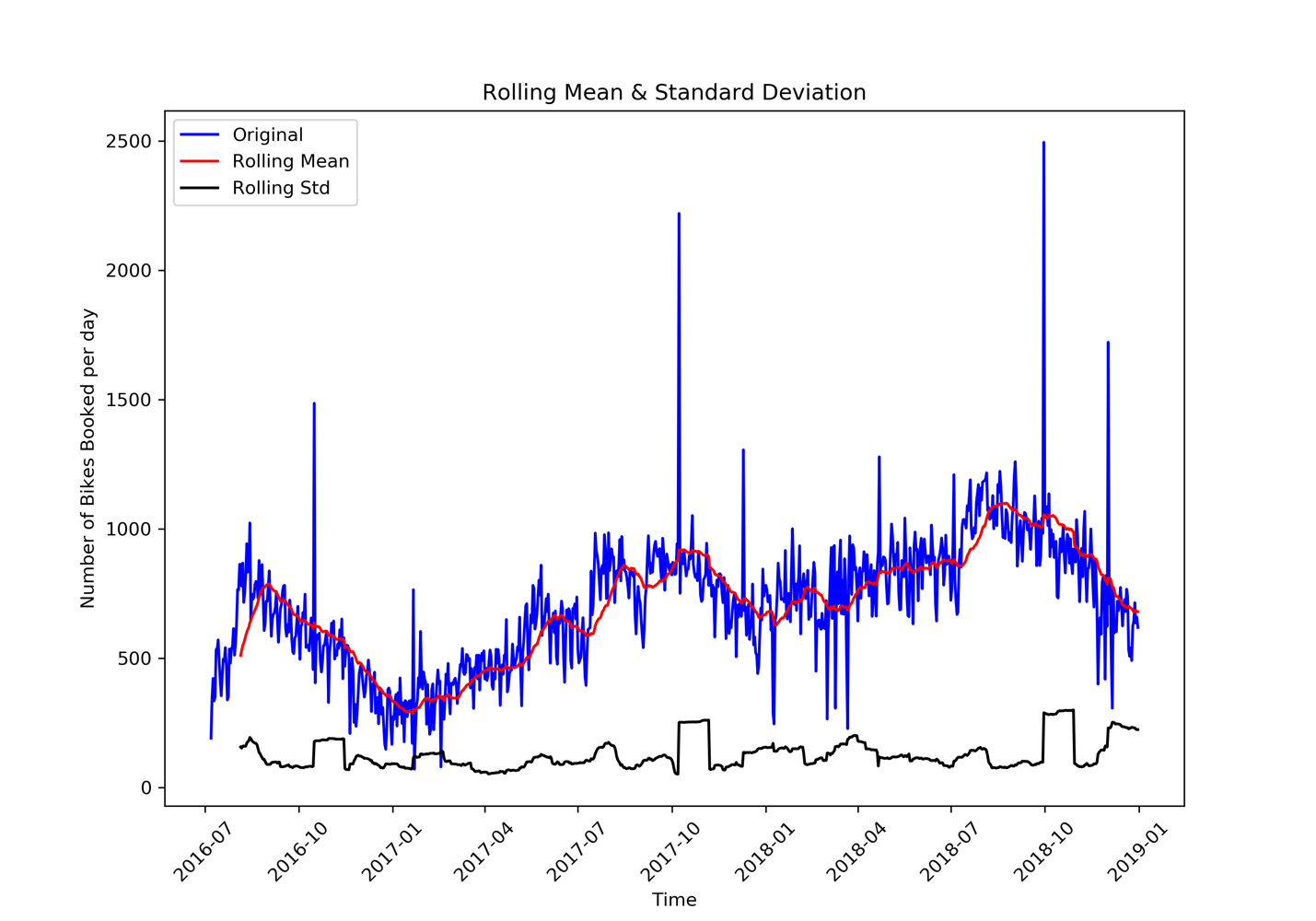


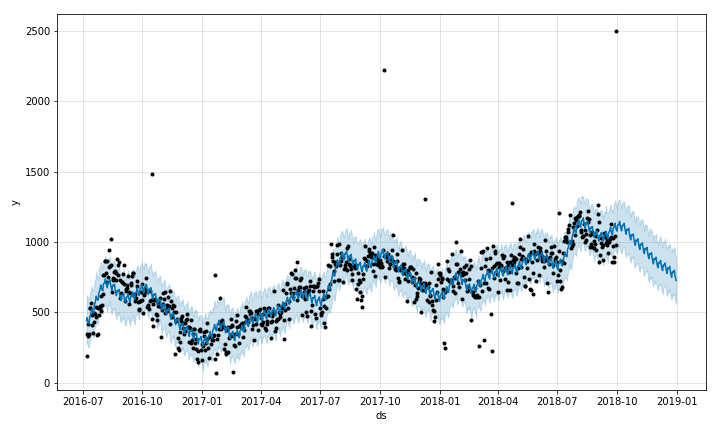
Figure 2 Dicky Fuller Test performed on Time series of demand of bikes per day.

|  |  |
| --- | --- |
| **Parameters** | **Values** |
| Test Statistics | -2.08776 |
| p-value | 0.24946 |
| Number of Lags Used | 13.00000 |
| Number of Observations Used | 894.00000 |
| Critical value (1%) | -3.43768 |
| Critical value (5%) | -2.86478 |
| Critical value (10%) | -2.56849 |

Figure 3 Results of Dickey-Fuller Test

From the Dickey Fuller test, we got the p value for the null hypothesis as 0.24946 which is higher than 0.05 but not sufficiently large to support the argument that the data is stationary. On the other hand, we are less than approx. 75% confident that the data is stationary. The most basic assumption of the time series models is that the series is stationary. In this case time series is not stationary, hence various methods should be used to stationaries the series. Some of the methods are elimination method (removal of trend and seasonality), Decomposition (decomposing series in trend, season and residual). Once the series is stationary, we can use ARIMA models to obtain the forecasting of the given timeseries data. We have implemented various transformation methods and applied ARIMA model. However, results obtained from ARIMA models were not significant and it failed to converges at high backlogs.

Apart from the traditional time series models, we have also used a python package named Prophet. It is developed by Facebook for forecasting time series data based on an additive model where non-linear trends are fit with different seasonality. It is robust to missing data, outliers and change in trend. We used this model to forecast the future demands. To check the performance of the model, Data is divided into training and testing data. Q4 of 2018 is kept for testing and Q3 2016 to Q3 2018 was used for training the model. Below is the graph for the fitted model along with the original dataset.



Prediction Trend

Figure 4 Prophet prediction graph

The graph below is to compare the predicted values to the original values. It can be seen that the prediction values is very close to the actual values indicating the better performance of the model.

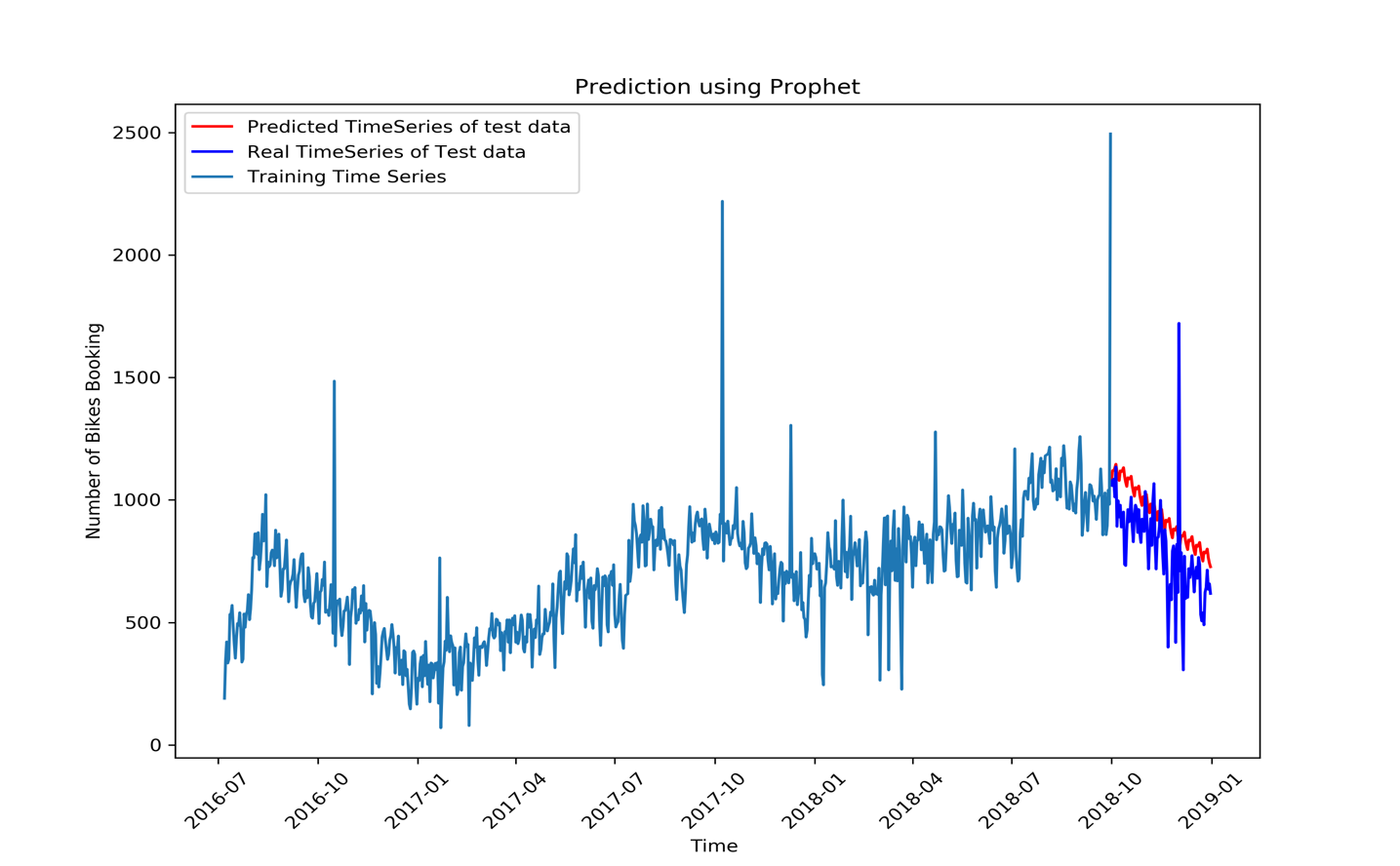


Figure 5 Predicted TimeSeries Against the Actual TimeSeries

The following is the forecast components of the model. The first graph is the trend, weekly seasonality and yearly seasonality.

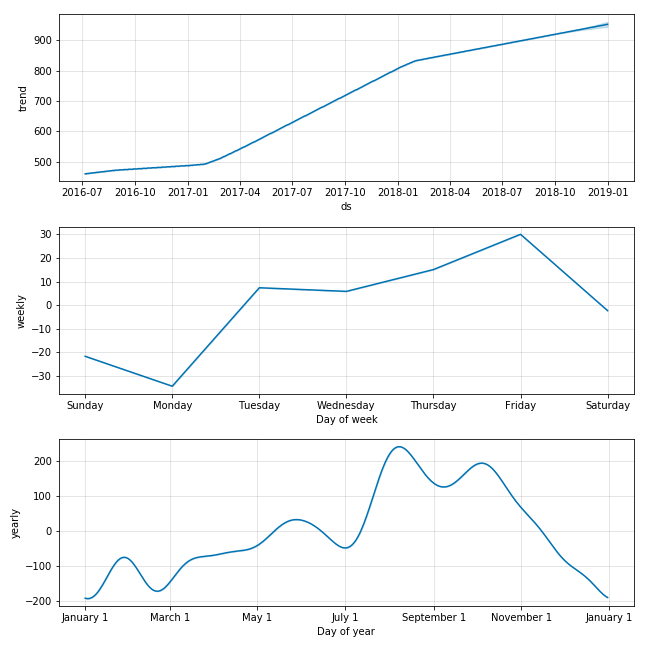


Figure 6 Forecast Components

To find the performance of the model, cross validation is performed on the dataset. The following are the results of the cross validation.

|  |  |
| --- | --- |
| **Parameters** | **Values** |
| RSME (Root Mean Square) | 193.5309 |
| MAE (Mean Absolute Error) | 156.9437 |
| MAPE (Mean Absolute Percentage Error) | 0.1949 |

**Recommendations for possible pricing modifications**

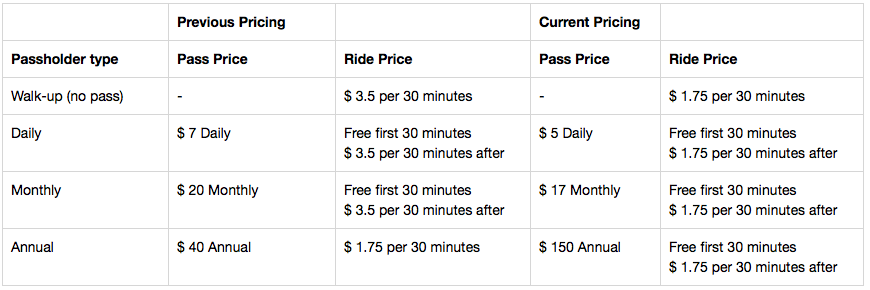
Recommendations for possible pricing modification can be made only after understanding the current pricing model. The pricing model then needs to be optimized to maximize total revenue.

The total revenue can be split into two main components viz Total fare of all rides and Total sales of passes (Daily, Monthly, Annual).

A demand model also needs to be estimated to derive the appropriate price to demand relationship. For this purpose, number of passes sold for each pass type is estimated as shown below:

Since the pricing system is based on 30 minutes blocks spent by a user on the bike, a column is coded to calculate the number of 30 minutes block spent by each rider on a trip. The pass holders would be charged for one less time block than the walk-up riders so the passholders get the first 30 minutes of their ride free. Therefore, two columns are added namely “time\_block\_count” and “time\_block\_count\_post\_free” for getting the 30 minutes block that the walk-up users have to pay for and the second column for 30 minutes blocks for passholders that excludes first 30 minutes block.

Moreover, the pricing model of Metro bike company was changed on July 12, 2018 so in order to analyze and optimize the pricing model correctly, the dataset was divided into two parts with one having data before the data July 12, 2018 and the second dataset having data after that date.



The total number of trips in each segment of the monthly pass, daily pass, annual pass and walk-up are calculated along with the total duration of the rides.

There is no unique identifier for passholders so the number of passes sold can’t be inferred from data itself. Total number of passes sold since January 2016 is 67,013 in 3 year period so it would be safe to expect roughly 20,000 passes sold during the year-long period for first two years and the remaining during the last year. The sales of each pass type would be estimated from a breakeven perspective i.e the average number of trips needed to cover the price of a pass.

**Breakeven analysis for estimating number of passes sold:**

The breakeven analysis is conducted for trips under 30 minutes since the median trip duration is 12 minutes. Number of rides to break even is calculated by dividing the pass price by the difference in the price to ride without pass and the price to ride with pass. Then, the number of passes sold is estimated by dividing the total number of rides by the number of rides to break even.

This calculation provided the number of rides per day required for daily pass: 2.0 Rides per month required for monthly pass: 5.714285714285714 Rides per year required for annual pass: 22.857142857142858 under the old pricing model. These numbers seems to be on the low side looking at the number of rides required to breakeven and assessing the numbers since Daily pass holders are likely to perform more than a round trip if they were purchasing a day pass and Monthly pass holders are likely to be the occasional commuters and even at a conservative 2 rides/week to work would tally 8 rides a month. Annual pass holders are the hardest to gauge due to the low $40 price point of the flex point from the previous pricing plan. They'll be scaled to the same factor as day and monthly pass numbers by a common factor of 40%. Under the old pricing model, the number of rides required per day for daily pass was scaled up to 3. Similarly, the number of rides required per month for monthly pass was scaled up to 8 and the number of rides required per year for annual pass was scaled up to 32. The total number of trips under different pass types were divided by the scaled-up value of the number of rides required for that particular passholder calculated above.

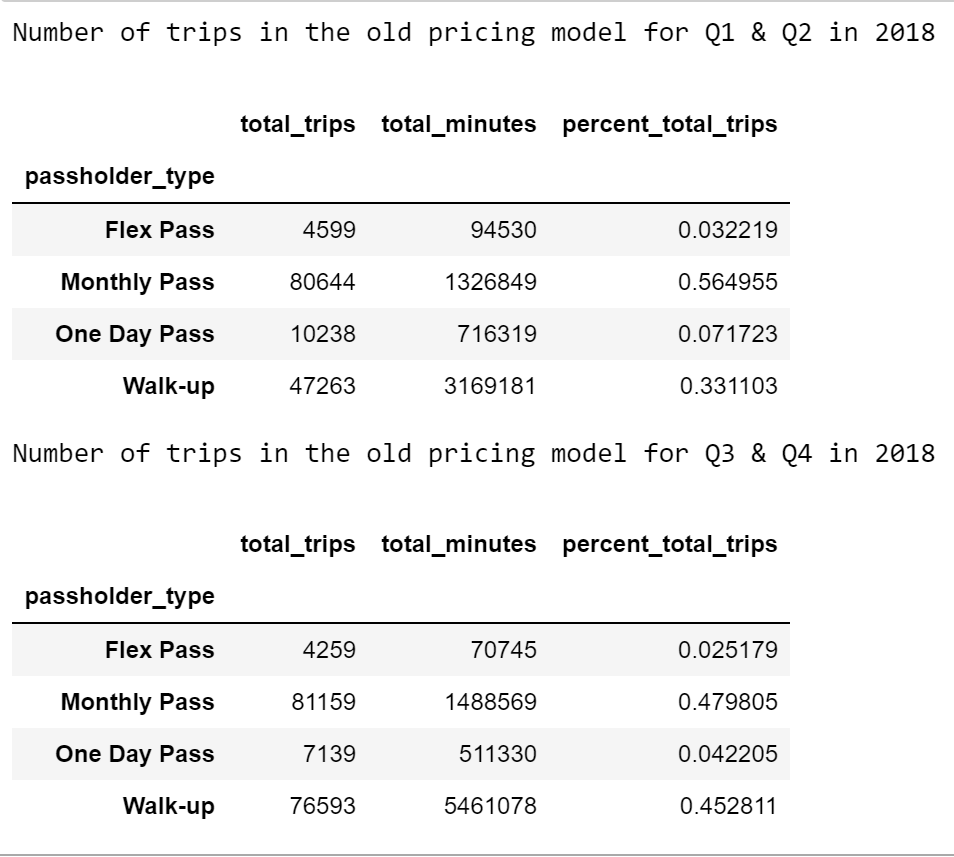
The estimated number of passes sold for daily pass, monthly pass and annual pass were 5299, 35507 and 711 respectively.

The total revenue was calculated by adding the revenue from 30 minutes blocks used by different types of riders, revenue from post 30 minutes usage and the revenue from selling passes. The total revenue calculation under the old pricing model came out to be $1479654.25 from July 2016 to July 2018.

**Linear Optimization:**

An objective function was formulated for this optimization model to maximize total revenue from pass sales and rides. Binary variables for each plan are created and a basic optimization is run over the previous price plans. Given that riders have already been complaining about the $3.50 per half hour block rate, any dropping of passes in the past pricing scheme would have led to a precipitous drop in ridership. Furthermore, given the fierce competition in the area of bike and scooter sharing, the dropping of certain passes could simply lead to riders moving to other competitors. A baseline attrition of 30% for each category is assumed and the parameters adjusted as the model is run. The optimal solution returned the following numbers: $1614687.75 for the old pricing model which is higher than it’s respective total revenue figure. The optimization results for the old pricing model favored the monthly pass and the annual pass but daily pass was not fruitful according to the results.

**Comparison of the number of rides in each category between the old and new pricing model**

****

A comparison is carried out between the old and new pricing model by considering the data of the same time frame for both old and new pricing model. The estimated number of passes sold for daily pass, monthly pass and annual pass was 10238, 80644 and 4599 respectively under the old pricing model for quarter 1 and quarte2 of 2018 whereas the estimated number of passes sold for daily pass, monthly pass and annual pass was 7139, 81159 and 4259 respectively. This steep fall in the number of daily passes sold indicate that the customers are not willing to invest in the daily pass under the new pricing model as the price of the walk-up has been reduced from $3.5 for 30 minutes to $1.75 for 30 minutes. Furthermore, it can be observed that the number of walk-up has also increased from 47623 to 76593. The increase in number of walk-up and the decrease in the number of passes sold suggests us that the bike riders prefer to use the walk-up option under the new pricing model.

**Optimization Using additional Information**

There are multiple ways in which the pricing scheme can be structured such as:

1) Variable rate by minute

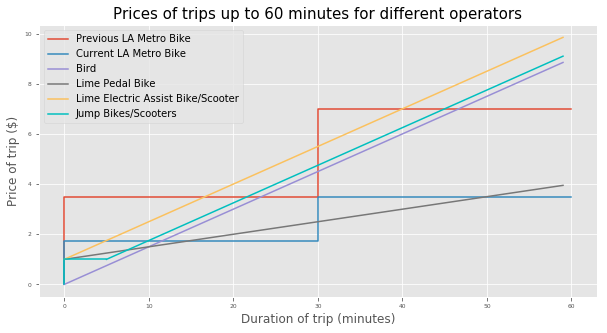
2) Variable rate by time intervals (30 minutes for LA Metro currently)

3) Fixed fee for any ride

4) Fixed fee for a time interval before implementing a variable rate

5) Weekly, monthly passes that cover all rides up to a particular duration. Variable rate thereafter

The current model by LA metro is examined and compared with other competitors in the market.



It can be observed from the above plot that the current pricing model enacted by LA Metro is one of cheapest being $3.5 for an hour-long bike ride. For rides shorter than 15 minutes and between 30 to 50 minutes, Lime pedal bike is the cheapest. These values are used to set up upper bound on prices to be used in the optimization model.

1. The variable rate per minute without a pass is set under the average of the two most common prices of 0.05 and 0.15:

Variable rate per minute without pass <= $0.10

1. The variable rate per minute with a plan is set under the average of the LA Metro’s current $0.06 and Jump’s $0.07:

Variable rate per minute with pass <= $0.065

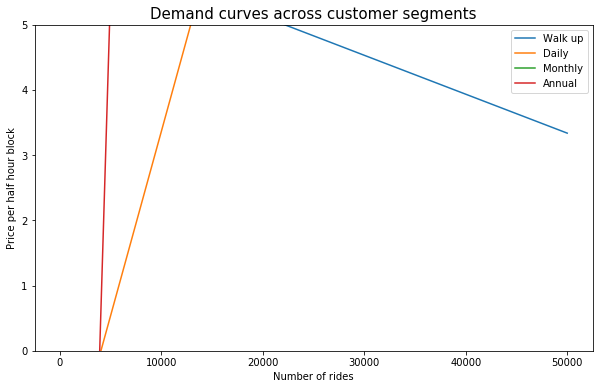
1. The variable rate per 30 minutes is set under the nearest competitor(Lime pedal):

Variable rate per 30 minutes <= $2.5

**Quadratic Optimization:**

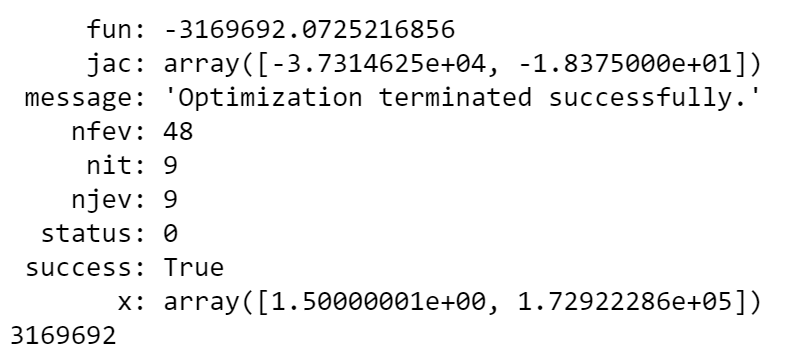
The total revenue would be optimized for two main factors: price and number of rides. If a simple linear optimization with an upper bound for price was run, it would intuitively move towards the upper bound for optimization so a demand curve that would vary total rides taken along with the prices charged for each hour block is taken into consideration.

For Demand curve estimation, 2018 Q1 and Q2 would be considered for data before the price change and Q3 and Q4 would be considered for data after the price change since 2018 Q2 was the last datapoint before the price change and 2018 Q3 was the first after. The slope for the demand curve was calculated by dividing the difference in price change with the difference in number of bike rides.



It can be noted from the above figure that the typical demand curve with a negative slope is observed only for the walk-up riders. This is because of the drastic rate cut in rates from $3.5 to $1.75 which motivated customers to start using walk-up option instead of purchasing the passes and the prices for passes did not fall proportionately for half hourly rates so it became more economical to forgo the passes.

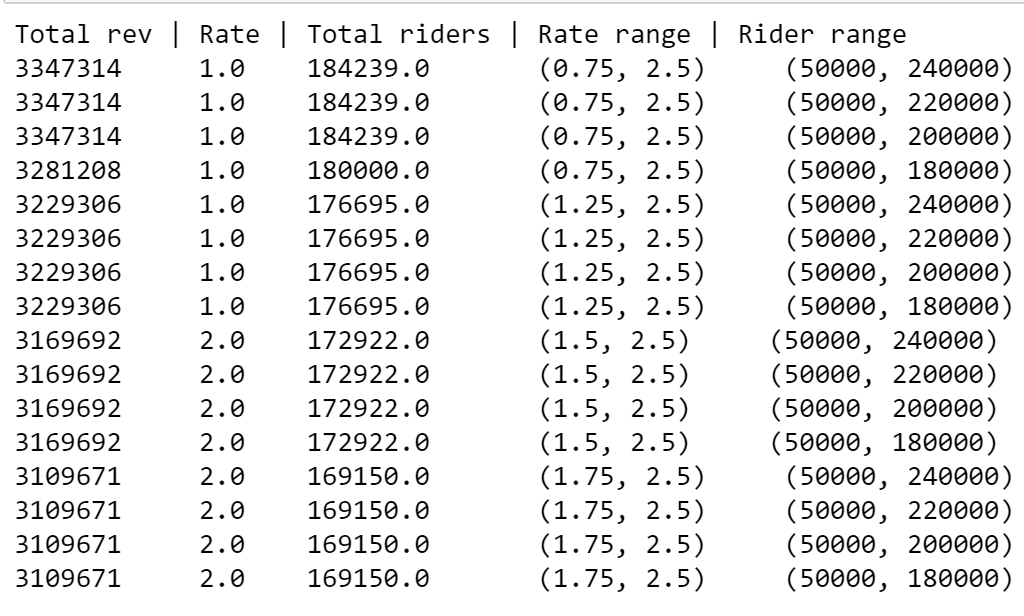
An objective function of total revenue based on the price and the number of rides was defined.



From the optimization results, it can be observed that the revenue is optimized when the price is set at $1.5 for every half hour. This would increase the total number of rides to 172,922 and the maximized revenue would be $3,169,692 which is much higher than what was observed in the previous pricing model.

**Sensitivity analysis :**

A further sensitivity analysis is carried out on both the factors viz price and the number of rides to observe the effect on the objective function i.e the total revenue.



It can be observed that adopting the lowest possible price of Dollar 1 results in a maximum of 184,239 rides. It also results in the highest possible revenue of Dollar 3,347,77314. Across the board, it seems that as long as LA Metro is able to attract more than 180,000 rides, the strategy should be to reduce prices. Given the results seen here, perhaps the drastic drop of 50% from the previous pricing scheme is justified.

**Summary :**

Shirish and Drupad

**Final Recommendation for optimizing revenue:**

The price should be dropped to $1.5 for every half hour for maximizing the total number of rides to 172,922 and the total revenue to $3,169,692. An additional surge pricing can be added during the peak hours which is 4 pm to 7 pm since this time period has the highest number of bike rides as seen from the data visualization part. The surge pricing should be for walk-up which will encourage the riders to buy more number of passes thus maximizing revenue by the sale of the passes as well as by the increase in the walk-up price during the peak hours. Another point to be considered while adding surge pricing is that it can lead to the downfall in the number of riders as they might tend to choose other competitors if available at that time. This experimentation can be carried out for a trial period of a month or so and the difference in total revenue calculations can justify whether or not to include the surge pricing during peak hours.