LOS ANGELES METRO BIKE SHARE DATA ANALYSIS

Texas A&M Data Science Competition 2019

Team Name – ENSEMBLE
Team Members: 1) Ayan Patel
2) Gaurav Burman
3) Prakhar Bajpai
4) Saksham Agrawal
Mentor- Dr. Rui Tuo

Contents

INTRODUCTION

Owing to the increased health consciousness among the public and the rising campaigns concerning global warming, bicycles have recently gained wide popularity as a form of transportation. Bicycles or, the trendier term, bikes, are a fast, easy and fun way to commute at places with dense traffic and short distances. Metro Bike Share is one of LA Metro's multiple public transportation options for Angelenos and visitors, which offers convenient round-the-clock access to a fleet of bicycles for short trips.

We have tried to analyze the data available on LA metro bike share website to derive insights and answer some critical business questions through Data Science principles. The data provides the details of each trip registered from July 7, 2016, to December 31st, 2018 for each of the regions viz, Downtown LA, Port of LA, Venice, and Pasadena. Another table with the details of status and the starting date of each 143 stations is also provided in another sheet. Both the datasets were merged and analyzed as a single dataset with 639785 rows and 17 attributes.

The dataset was cleaned by removing missing end stations, virtual stations, 0 Latitude and Longitude, outliers based on very long durations or distances, reducing the number of observations to 587633. Patterns were studied between different attributes. Dataset was also converted to time-series data with the count of trips/day as the target variable to perform forecasting operations. Several interesting correlations and patterns were discovered between the count of trips, their distance, and durations. For example, a sudden spike is seen on days of CicLAvia festival (https://www.ciclavia.org/), Downtown LA accounts for more than 78% of the total count of trips

The questions that we targeted through our approach concern forecasting the demand for bicycle, revenue generation through ticket sales and recommendations for Bicycle Staging in 2019. For forecasting purposes, performed feature engineering to obtain the count of bike trips for a given day and time-series analysis was done using the ARIMA algorithm to predict the demand for next quarter. This demand is distributed in the proportion of trips at each station as per given data and the bicycle demand is estimated using this ratio. For the pricing analysis, we have only used the data available for the walk-up category as the rider information is not provided in the dataset and it cannot be ascertained that how many times an individual rider used his pass in a given day. From the network management aspect, we have categorized the routes which have the highest traffic and the maximum distance and proposed new stations at these routes.

DATA COLLECTION AND PROBLEM STATEMENT

Data Collection

The data was provided on the LA Metro Bike share website with the details of each trip from July 7, 2016, to December 31st, 2018. Another dataset was provided for the details of individual stations with their Go live dates, region they belong to and their status as on date. Following points can be noted from the by analyzing the Data Sets coherently:

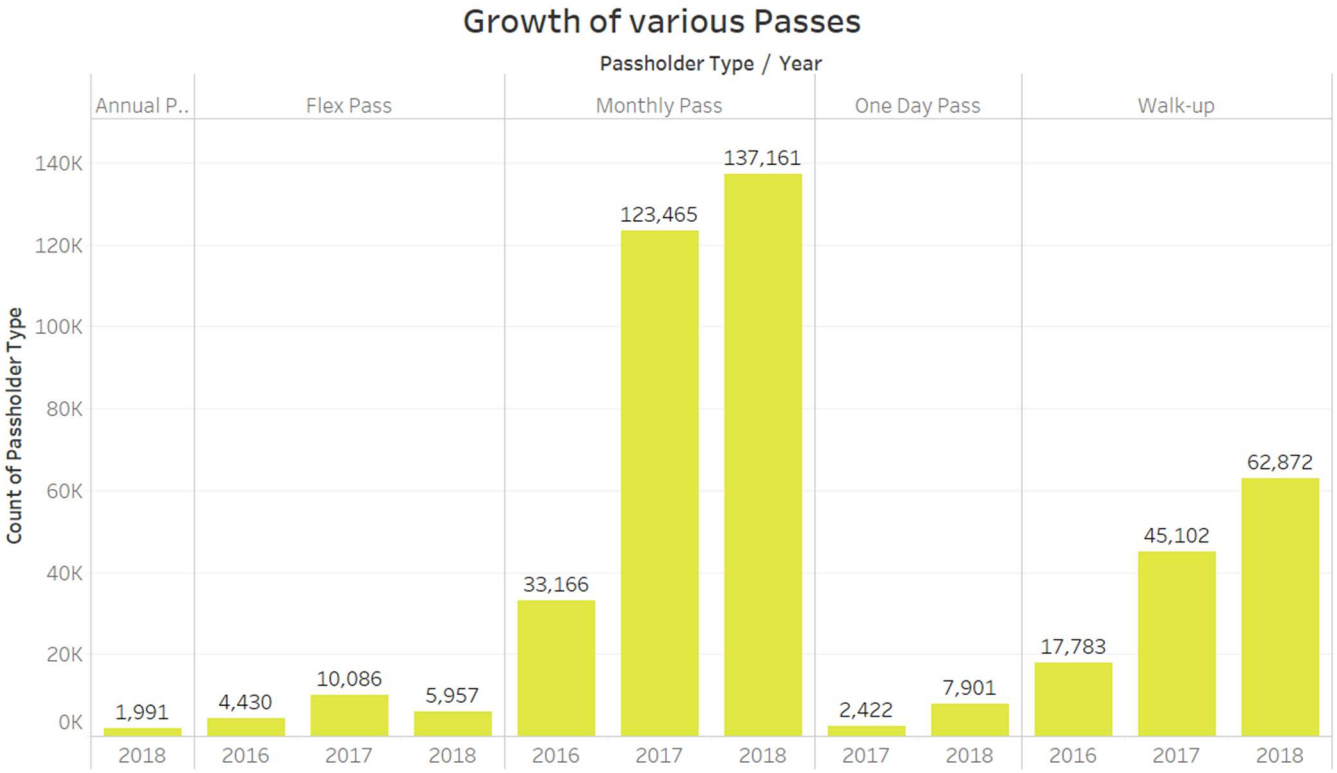
- The LAbike dataset had 639786 unique trip details with 13 attributes pertaining to start and end station IDs with their co-ordinates, start and end times, pass holder type and trip category whether one way or round trip.
- Stations_Table dataset provides the details of each 143 stations with their unique Station IDs, station names and their Go Live Date. The region wise growth of station can be seen below.
- The company started with 64 stations in Downtown LA including the 7th & Flower station which has been the busiest station in terms of trip start point till date.

- Currently, the company is operating in 3 regions viz. Downtown LA, Port of LA and Venice, with DTLA being the leading region accounting for 78% of total trips due to the obvious reason of downtown crowd.

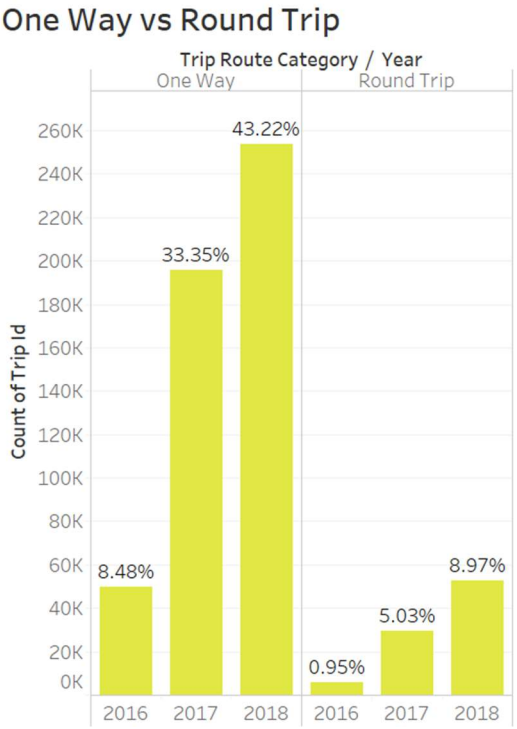
- Pasadena was started on June 21, 2017 with 375 bikes. However, due to poor ridership in this region with less than one ride per bike per day on average, the service was discontinued from August 16, 2018.

- Venice has shown a significant rise in percentage ridership share since its inception in Sep 01, 2017. This region can be focused to expand the network of Bike stations to increase the ridership as will be explained in detail in forthcoming sections.

- Monthly Pass has been found to be highly popular among the riders as compared to other pass types. However, we do not have the user data like UserID or UserNum so we cannot identify the count of distinct trips taken by individual user.



- Also, major proportion of trips are one-way trips as given alongside.



- As of Dec 2018, Status of 45 stations was inactive and that of 95 stations was inactive. No information is available for 3 station IDs namely 3009, 3039, 4276. Out of 45 inactive stations 33 belong to Pasadena itself.

Problem Statement -

We will work upon the following problems and provide our recommendations through data science models to improve the business model and provide inputs to better guide the operations of the Bike Share system.

- Forecasting bicycle Demands for the Q3 2018 thru Q1 2019.
- Forecasting monthly Ticket Sales and Revenue for Q3 2018 thru Q1 2019.
- Suggesting Potential new station locations and characteristics to look for while selecting the same.

DATA PREPROCESSING

This was the most crucial and the most time-consuming part of whole project as the quality of the data input plays a big factor in determining the accuracy of our model and subsequently the accuracy of our predictions. Corrupted values can corrupt our model misleading us to wrong conclusions. Data Preprocessing was performed in two steps.

1. Data Cleaning – Removal of Missing Values and outliers
2. Feature Engineering – Creation of new attributes from available information.

1. Data Cleaning

- **Missing Values:** The count of missing values found in each dataset is as follows.

  - For Forecasting the demand of bicycles, we only required the count of trips for a given day which could be obtained from the trip IDs and start time, so we did not remove any missing values for training the forecasting model.

  - However, for predicting new station locations and pricing model we needed the duration and distance covered for each trip, so we removed these missing values before our analysis.

  - After removing the missing observations, we were left with 587633 observations from 639786 observations initially.

trip_id	0
bike_id	0
start_station	0
end_station	43198
trip_route_category	0
start_time	0
end_time	0
start_lat	1354
start_lon	1354
end_lat	9110
end_lon	9110
plan_duration	384
passholder_type	0
dtype: int64	

Table 1: Missing Values in Trips dataset

  - In the Second dataset, no details were provided for 3 station IDs viz 3009, 3039, 4276. We have removed these 3 stations from our analysis.

Station_ID	0
Station_Name	3
Go_live_date	3
Region	4
Status	3
dtype: int64	

Table 2: Missing Values in Station Dataset

- **Outliers:**
  - Several Longitude and Latitude co-ordinates were found to be (0.000,0.0000) which made no sense as the Bike Share system only serves in the LA County which lies within the co-ordinates 34°3’N 118°15’W / 34.050°N 118.250°W / 34.050. We removed all values from our consideration for predicting new stations and pricing models.

  - Some trip durations (End time – Start Time) went beyond 24 hrs i.e either the user kept cycling continuously for whole day or forgot to end the trip after completing the ride. These led to a few abrupt values in duration of trips. We considered values within 1 - 99 percentile

for each region to remove vague values which could disturb the statistics of the data like mean and std deviation.

- There were several trips starting from 'Virtual Stations' which we assume were generated for the testing purposes as no region is assigned to these and as implied by their name.

2. Feature Engineering

- **Count -** We are provided with the details of individual trips, but to make the forecast of future trips we need to aggregate the count of demand on temporal basis like hourly, daily, weekly etc. For our analysis we have used the trip start time variable, converted it to *datetime* datatype and calculated the count of trips for a given day to create a *Count* variable for no. of trips per day.

- Using the datetime variable we have extracted daily, monthly and yearly trends in our data.

Day	Count of trips
7/7/2016	190
7/8/2016	344
7/9/2016	421
7/10/2016	334
7/11/2016	349
7/12/2016	531
7/13/2016	534
7/14/2016	570
7/15/2016	484

Table 3: Generated Count data

- **Duration-** To analyze the station locations and pricing model, we needed the time for which each trip runs, i.e., duration of the trip. We used a function to subtract start time from the end time to calculate that.

- **Distance-** To get the idea of each station location and the popular routes, we calculated distances between start and end location for each trip and added it to a new column. For round trip, this distance was zero. The distance was calculated by the calculating distance between end and start longitude/latitude points.

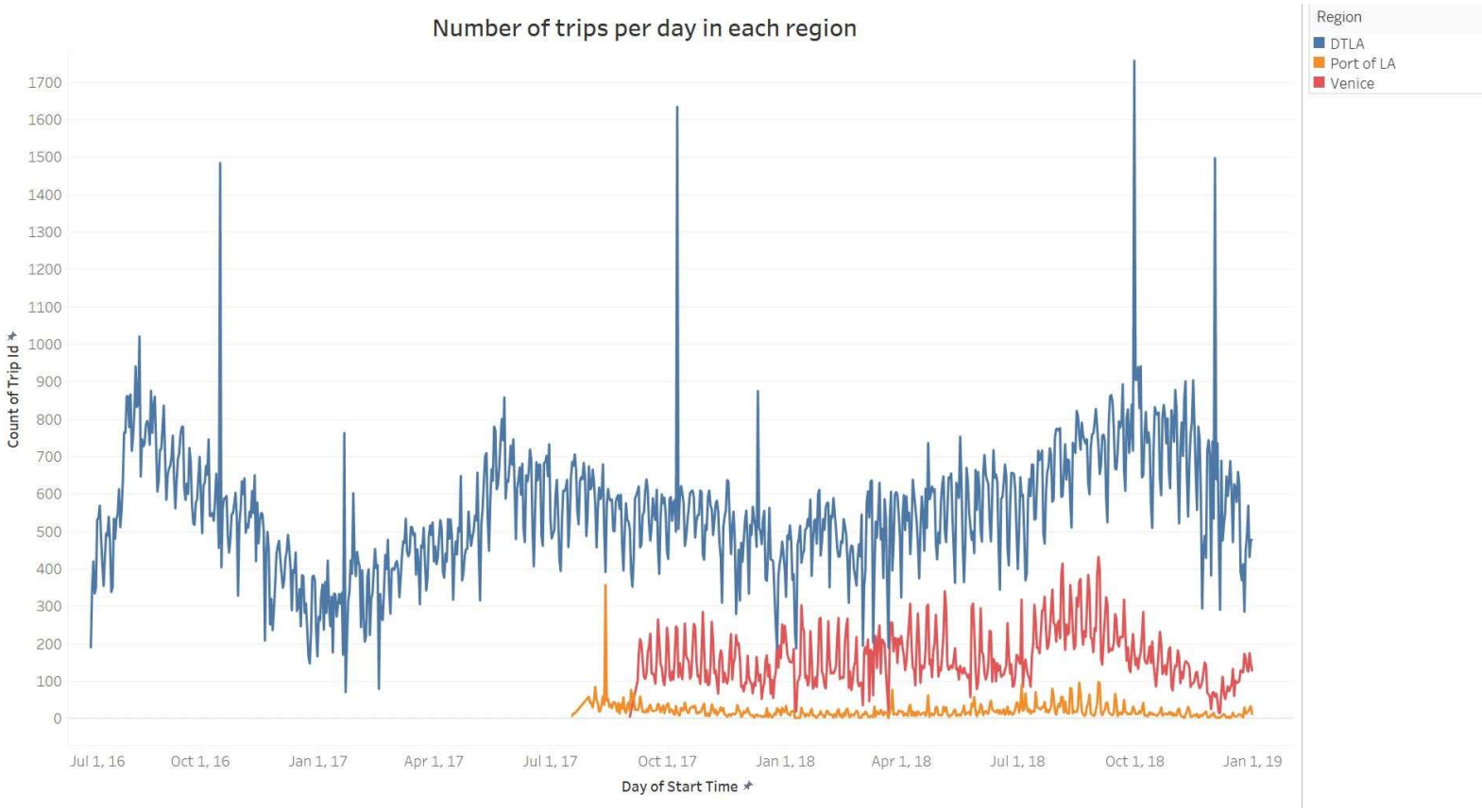
start_station	end_station	trip_route_category	Duration(Mins)	Distance(Miles)
3014	3014	Round Trip	3	0
3021	3054	One Way	13	0.442559759
3022	3014	One Way	10	0.765484117
3076	3005	One Way	10	0.62504642
3031	3031	Round Trip	48	0
3031	3078	One Way	16	1.558068376

Table 4: Duration and Distance Calculated for each trip

- **Categorical Variables**- One hot encoding was done for *Pass_Holder_type* and *Route_category* variables to make it ready for linear regression.

FORECASTING BICYCLE DEMAND

Using the time-series dataset generated as described in the feature engineering section above, we were able to capture the trend of number of trips covered per day. Since the stations at different regions were started at different dates with gaps of months, the count of trips was influenced with the Region variable. Therefore, we chose to create separate models for each Region to keep the prediction independent of the region variable. The trend for of count trips can be visualized below:



The spikes in the trend for the number of trips in the above graph can be attributed to the CicLAvia event which closes streets to car traffic and opens them to Los Angelenos to use as a public park. creating a safe place to bike, walk, skate, roll, and dance through Los Angeles.

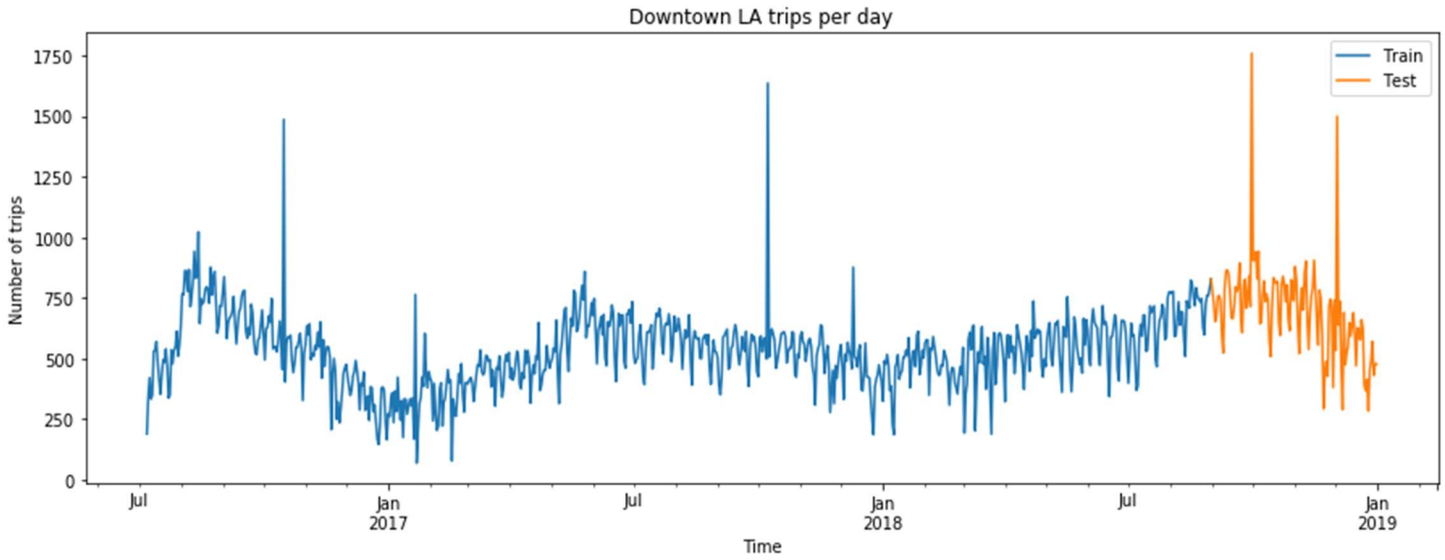
The spikes on the graph for DTLA match exactly with the dates of CicLAvia in that region, namely Oct 16, 2016, Oct 08, 2016, Sep 30, 2018 and Dec 2, 2018. This can be verified from https://www.ciclavia.org/events_history.

Seasonal variation can also be observed in the above trend with a steady increase from spring through fall and decrease from winter through spring.

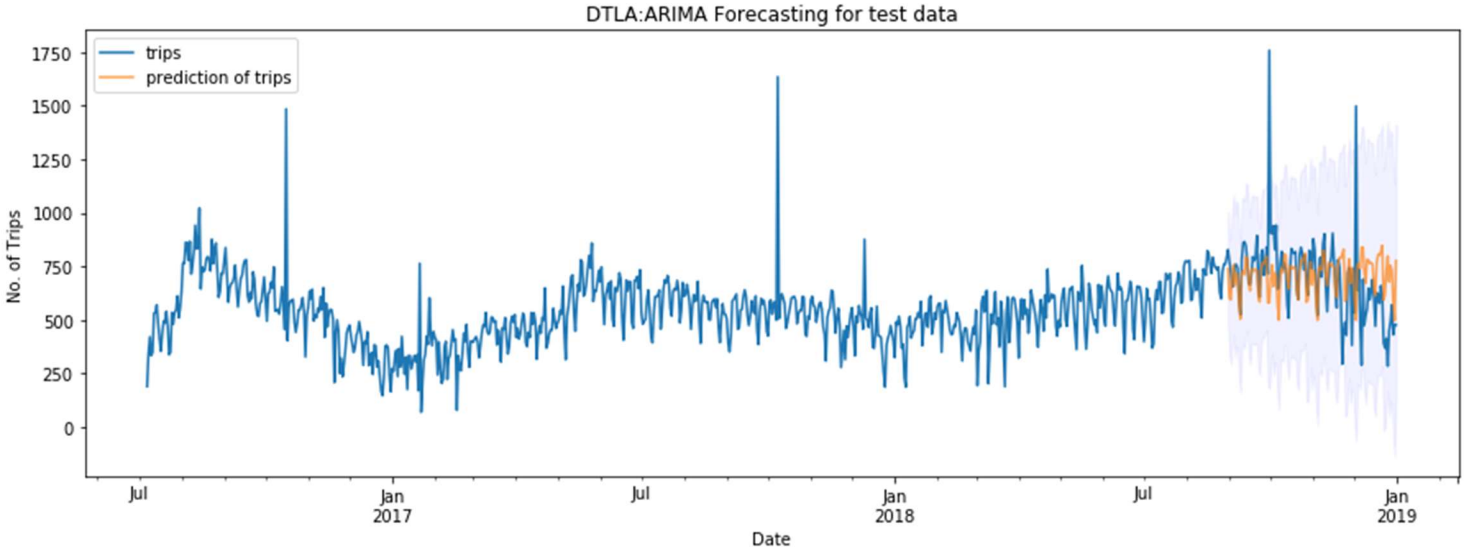
ARIMA Algorithm

Using the seasonality and trends pattern we used above time series dataset to train a Seasonal ARIMA algorithm. ARIMA stands for Seasonal Auto Regressive Integrated Moving Average and predicts the future values for the future time period. ARIMA Algorithm demands thee arguments to create forecasts i.e (p, q and d). Here p represents the auto-regressive term, d represents differencing to remove trend and seasonality and render a stationary and q represents lag or error component (part of time series not explained by the trend or seasonality.

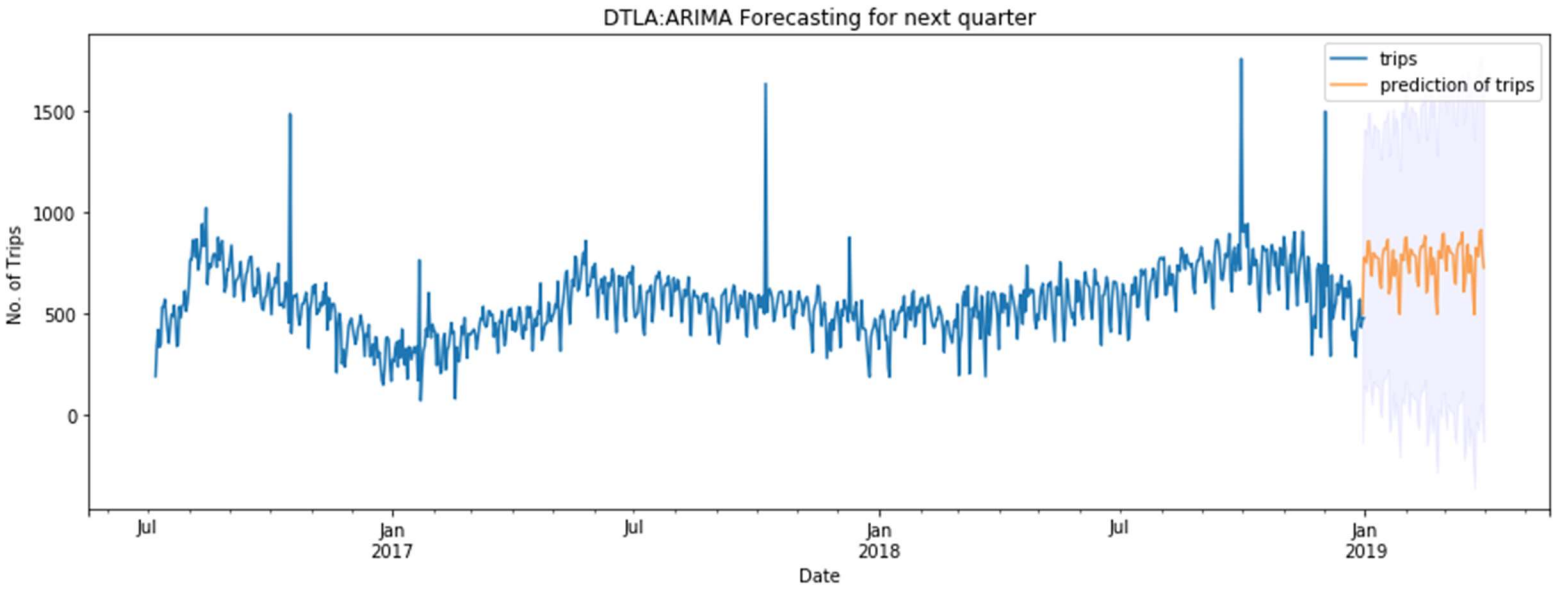
We divided the dataset into train and test datasets to train our model and validate the predictions respectively. The training set for DTLA belonged to the time period 07/07/2016 to 07/07/2018 and testing dataset belonged to time period 07/08/2018 to 12/31/2018 region can be seen below:



On training the ARIMA model as per the above trend and predicting the count of trips for the test data we obtained the following plot. The grey region represents 95% Confidence Interval for the prediction.



Extending the same model and predicting the count of trips for the Q1 2019, we get the following trend



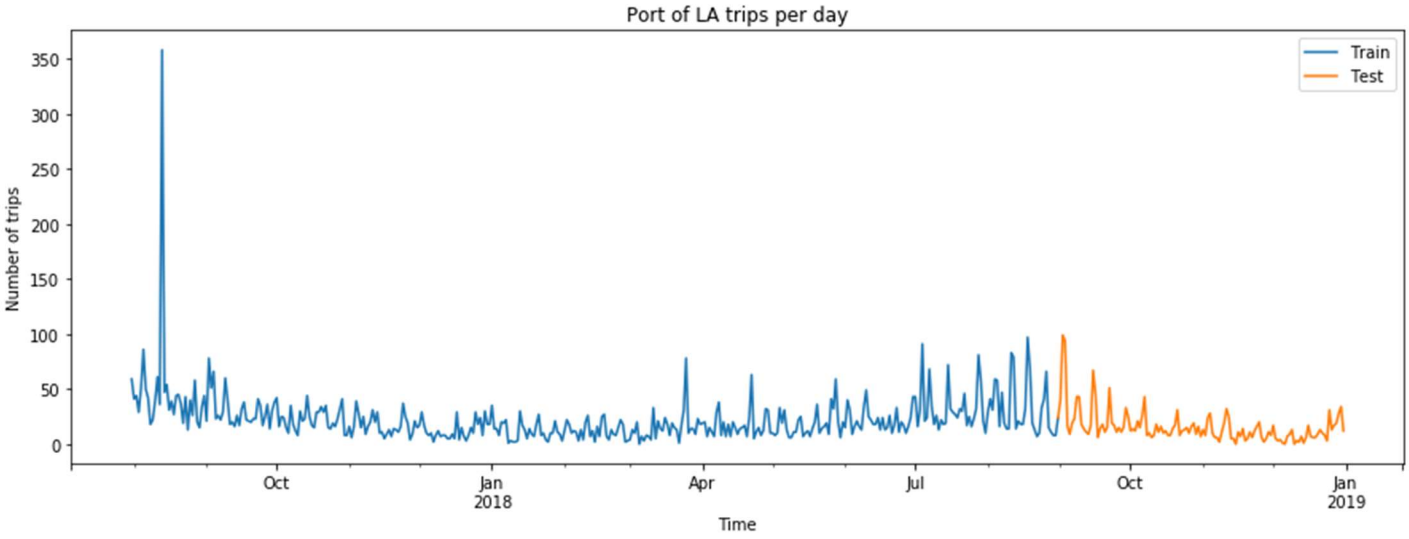
The Predictions on Test Datasets for three regions gave us the following Accuracy

Region	Mean Absolute Percentage error
Downtown LA	21.53%
Port of LA	55.50%
Venice	56.42%

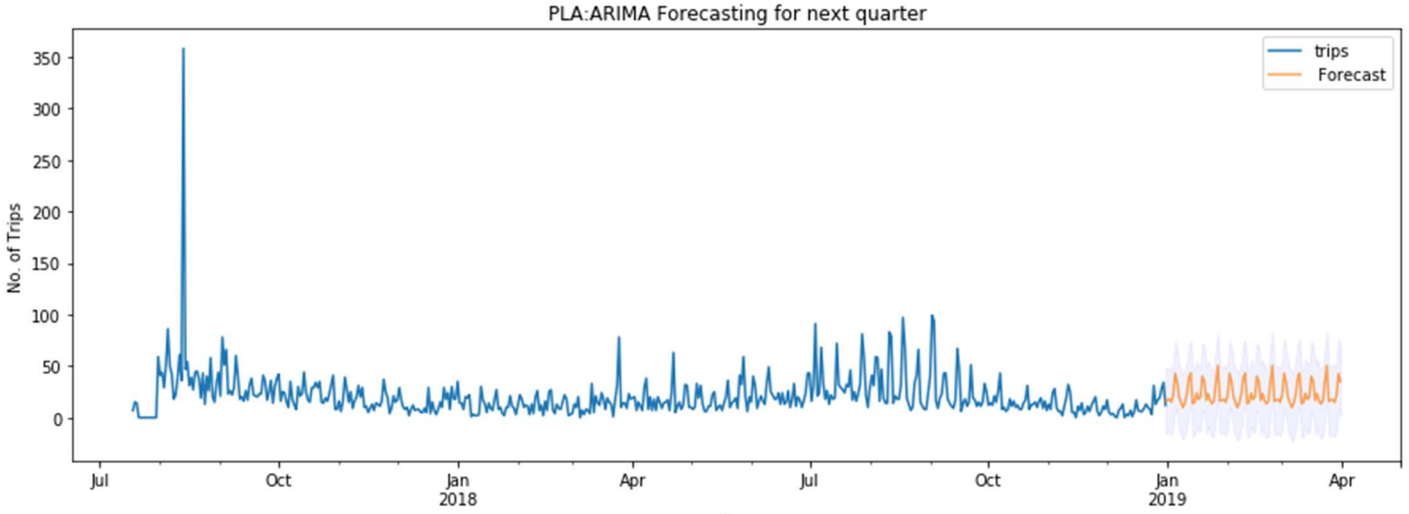
Table 4: Prediction Accuracy for Test Data

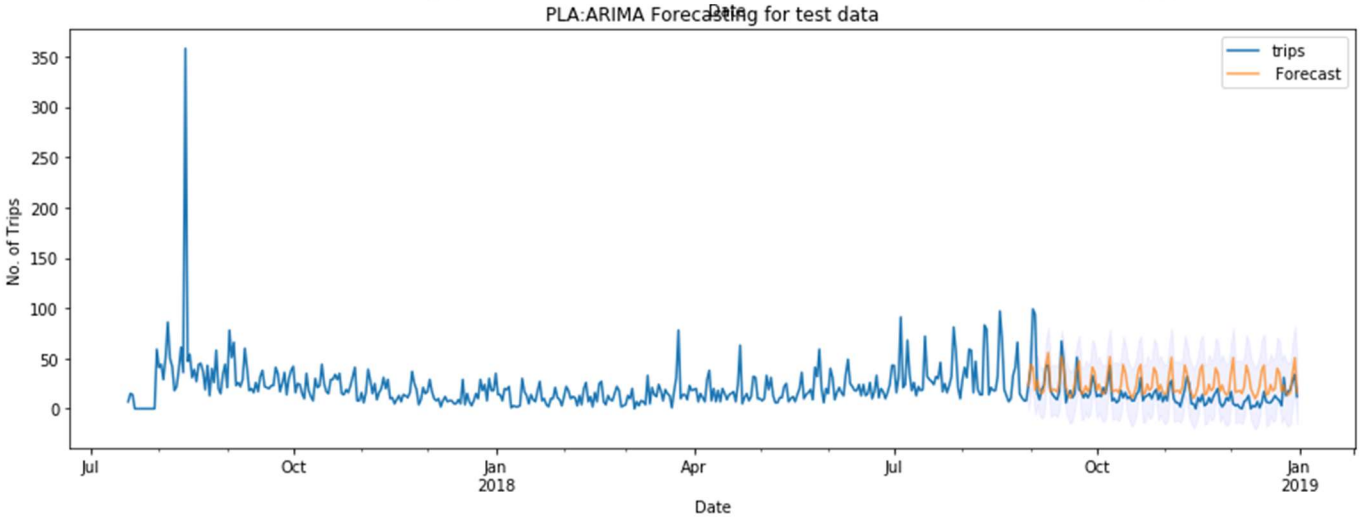
For Port of LA

Similarly, training the models for Port of LA region we get the following results and plots:



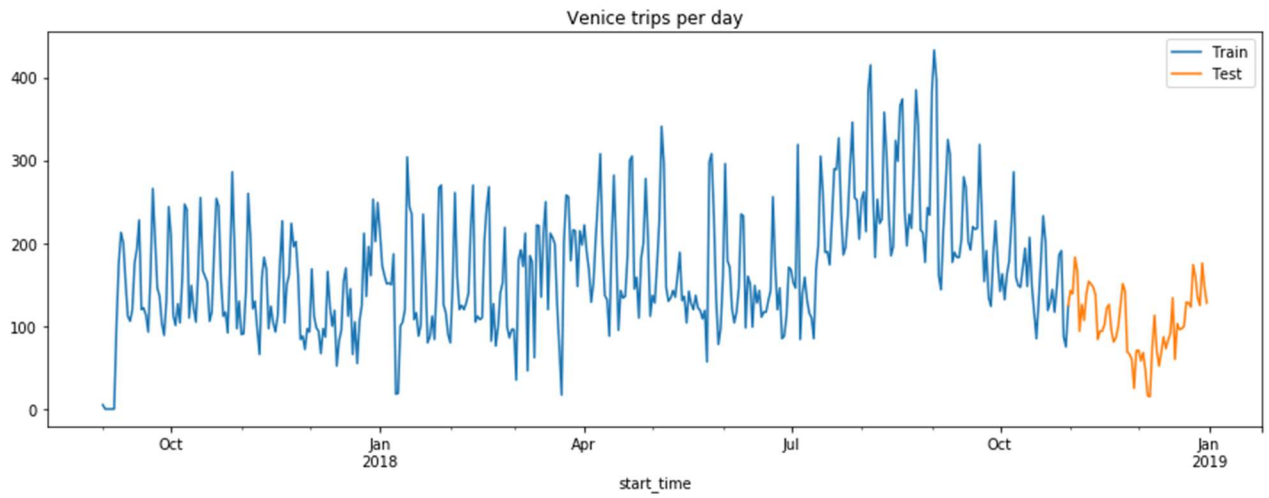
Forecasting the demand or count of trips for test data for Port of LA



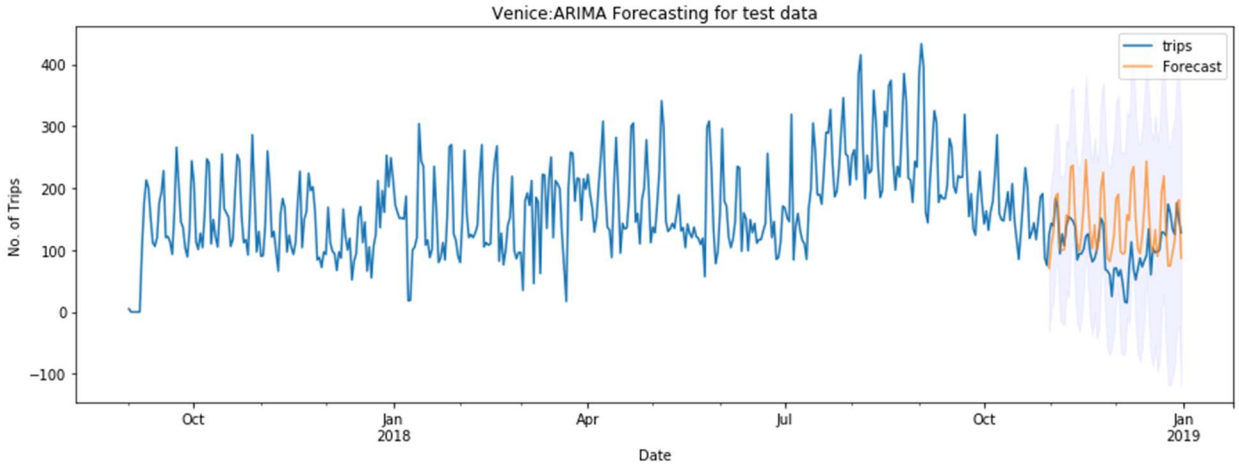


For Venice

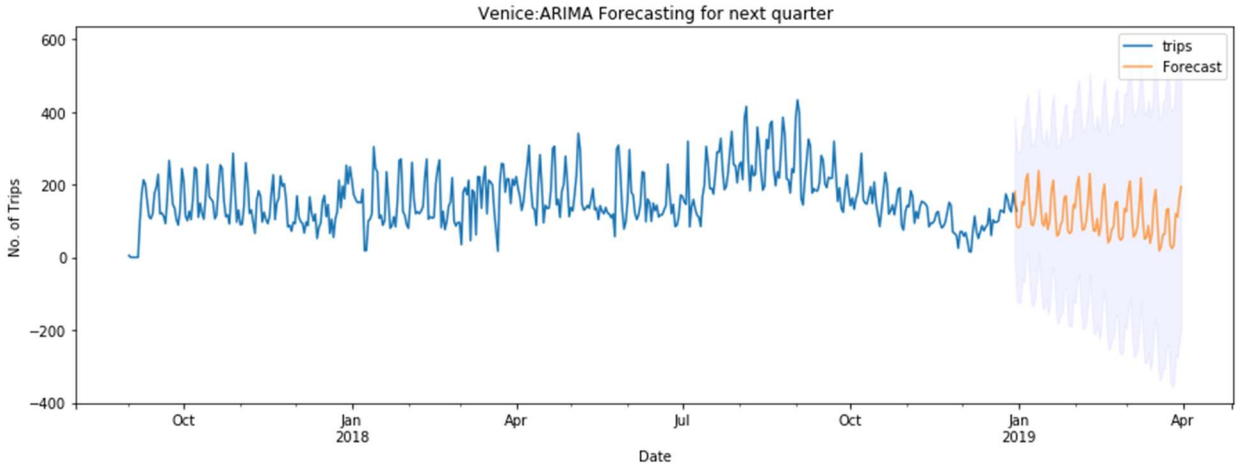
Similarly, training the models for Port of LA region we get the following results and plots:



Forecasting the demand or count of trips for test data for Venice:



Forecasting the demand or count of trips for Q1 2019 for Venice:



NETWORK MANAGEMENT

Potential New Routes

To model the optimum location of a bike station two factors, play key roles to define the profitability of a given point:

1. How many people want to go to a location?
2. How much are they willing to pay for that route?

Since this data requires to be extracted from various sources like tourist and employment location density for a given county which is not available to us, so we have used the features extracted from the available data. We have used the count of trips as the popularity of the point and the distance between them as the measure of the money people are willing to spend to cover the distance.

We have created a heat map of popular routes of all the regions as shown above. There are 6372 used route combinations within all the regions. In the popular routes map, we have selected top 45 routes which cover 15% of total one-way trips. We have identified these routes on basis of requirement of a new station between these routes. The size of the block in the map determines the number of trips between two station and color intensity identifies the distance between the two stations. From the map, we can infer following insights,

- The station 4215 (Downtown Santa Monica Expo Line Station) is most repetitive station in our analysis and a lot of long duration rides start and end at this station. So, a new station can be started anywhere between 4215,4210, 4211, 4212, depending on geographical feasibility.

- Similarly, the route 3082-3085 is also one of the busy routes and have an approximate distance of 1.221 miles.

The management can follow this map to identify the potential station points in case they decide to expand the network in future. This analysis can be further extended to include more routes and trips as and when needed.

Competitor Analysis

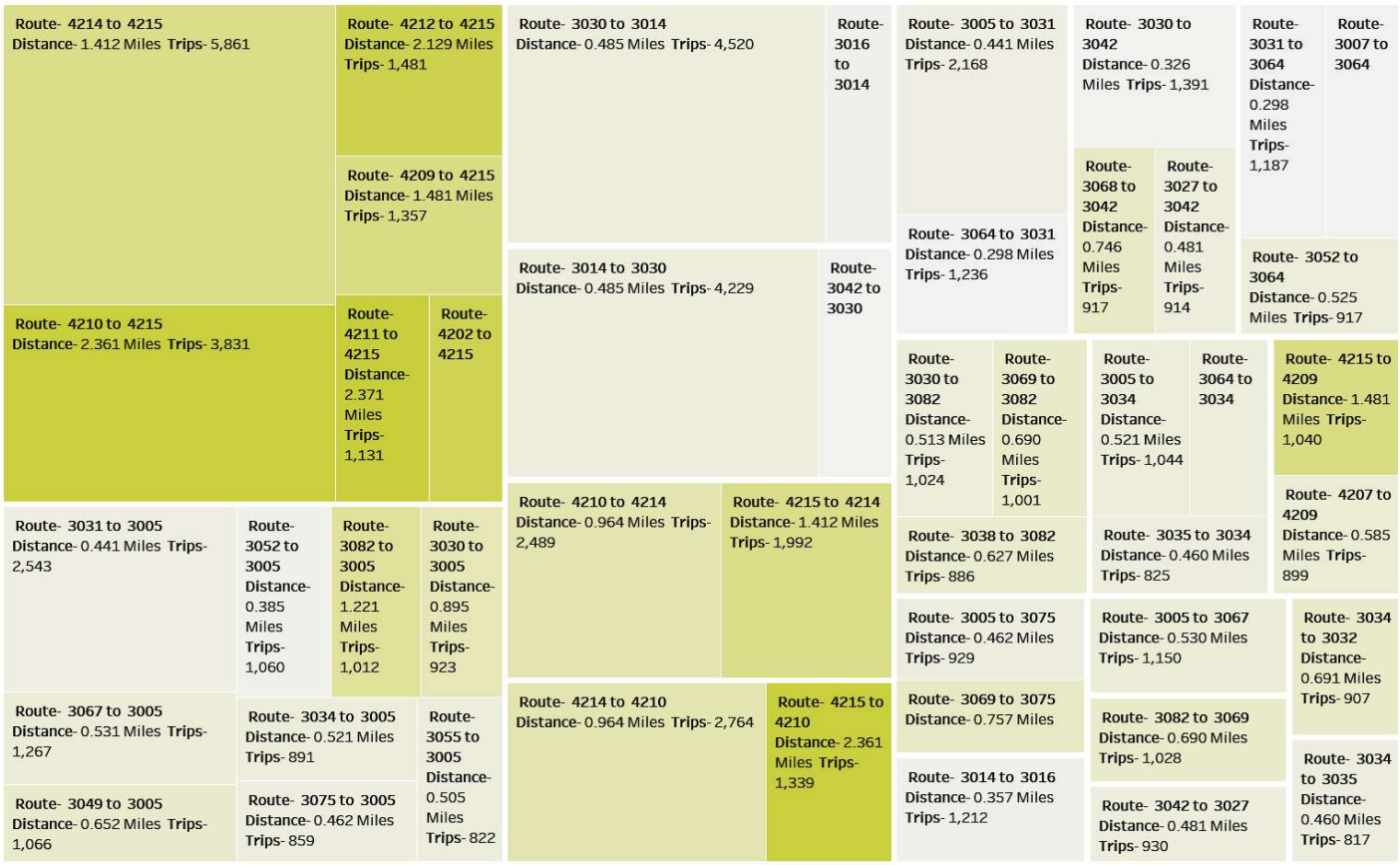
In order to analyze the existing price models in the region, we did a competitor analysis in the regions where Metro Bike Share operates. There were few interesting insights from it as given below.

The competitors in the region of Downtown LA are using dockless feature, which means that the user can leave his bike anywhere and not need to worry about the parking station, as in the case of Metro Bike Share.

Metro Bike share is also facing competiton form companies like Bird and Lime, which provide scooters on rent at competitive pricing.

In order to sustain, the Metro Bike share should either adapt the dockless feature or rethink of the current price modelling, untill they become well equipped with the new technology and features.

Popular Routes



RECOMMENDATIONS

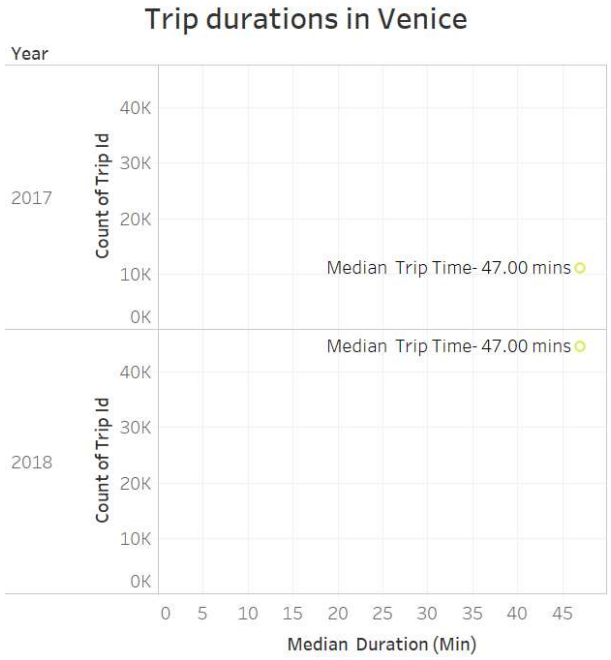
Price Modelling

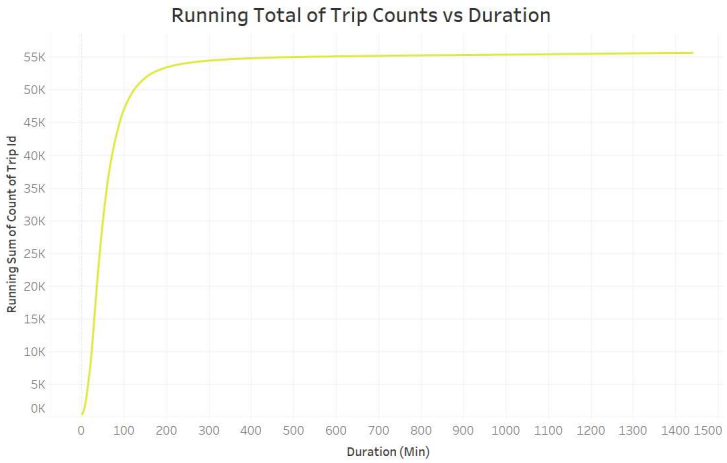
Metro Bike share had made changes to their payment plans in June 2018, which also reflected as approximately 2% increase in the ridership in the following month across the city. However, aligning with the vision of the company which focusses on the Longevity, there is a need to restructure the payment model. We propose the following payment plan in the walk-up category,

“All the walk-up rides should be charged as $1/20 minutes and subsequent charges should be $1 for every next 10 minutes.”

In order to run a hyposthesis test,

We have considered all the walk up trips in Venice region for testing. The top 1% of the duration data is removed since there were abrupt values in some of the trips which can hamper the analysis.





We can see that the 90% percentile value in the region lies around 300 minutes, i.e., almost 90% of the rides are less than 300 minutes of duration, and the median time is 47 minutes approximately.

Time Duration	# Trip	Old Revenue	New Revenue	Net Profit
0-20	7574	$13,254.50	$7,574.00	-$5,680.50
21-30	7709	$13,490.75	$15,418.00	$1,927.25
31-40	7940	$27,790.00	$23,820.00	-$3,970.00
41-50	6653	$23,285.50	$26,612.00	$3,326.50
51-60	5490	$19,215.00	$27,450.00	$8,235.00

Table 5: Cost- benefit analysis for sample data

The net profit from this table is $3838.25.

This cost-benefit analysis suggests that cutting down minimum cost will not do a lot of harm since we will lose approximately $0.75/ per ride in rides ranging from 0-20 minutes, and gain $.25/ride for rides ranging from 21-30 minutes. Similarly, we can get $0.5/ride more for all ride ranging from 41-50 minutes.

Additionally, if we decrease our minimum charges, it will attract more customers who need small trips. This goes in hand with the network proposal which shows the need of new stations in Venice area. The competitors are also less in the area and the region has a lot of scope in terms of tourists and students living in the region.

If this pilot program is well received, this program can be extended to other regions also, which will lead to more market penetration for the company.

Number of Bikes in Different Regions

The following number of bikes should be allotted to each region as per the demand forecast.

The number of bikes is obtained by dividing the predicted number of trips by the ratio od counts of trips/bike obtained from the available datset.

Region	Number of Bikes Projected in Q1, 2019
DTLA	438
Port of LA	120
Venice	121

REFERENCES

Time Series Forecasting- https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3

Metro Bike Share- https://bikeshare.metro.net/about/data/

ARIMA- https://otexts.com/fpp2/weekly.html

CICLAVIA- https://www.ciclavia.org/events_history

Python- https://pythonprogramming.net/resample-data-analysis-python-pandas-tutorial

Cost Benefit Analysis- http://www.sjsu.edu/faculty/watkins/cba.htm