LOS ANGELES METRO BIKE SHARE DATA ANALYSIS

Table of Contents

INTRODUCTION

With the rising global warming and increased health consciousness among the public, there is rise in the use of bicycle for transportation. Bicycles or bikes are easy to ride and a fun way to commute at places with dense traffic or short distances.

Metro Bike Share is a Bike sharing company spread across LA. In this project, we analyze the bike data available at its website. The dataset includes two tables LABikeData and StationTable. The first table has 13 attributes which includes the unique trip id, start and end station, its location, bike id, start and end time of trip just to name a few. The dataset is from July 7, 2016, to December 31st, 2018. The second table has 4 attributes which includes station name, ID, status, region and Go Live Date. There are 4 regions viz, Downton LA, Port of LA, Venice, Pasadena across which the stations of Metro Bike share are located. (LA Metro Bike Share Data, n.d.)

Descriptive statistics is used to examine the dataset. Dataset is converted into time series data by converting Start time attribute to datettime type and used to pivot as index values. The outliers are identified by creating a trip duration column using start time and end time columns. Additionally, outliers such as discontinued station are identified with go live data column and removed. Missing values are imputed as discussed under Data Cleaning section. (LA Metro Bike Share Data, n.d.)

Bike IDs are used as a feature to forecast the bike staging part of the problem. Start station is an important feature as that forms the basis of forecast number of trips. Forecast model is built using Facebook's open source package- Prophet. The model predicts the number of trips for Q3 of 2018 to Q1 of 2019, thereby predicting the number of bikes at a region level. Additionally, based on the analysis and model results, pricing recommendations and network management and expansion suggestions are mentioned. The recommendations are visually shown using Tableau. (Prophet Algorithm, n.d.)

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. Additionally, another dataset was provided specifying the details of each station such as their go live dates, region they belong to, status. (LA Metro Bike Share Data, n.d.)

Following are the observations of the two datasets:

- The LAbike dataset had 639786 unique trip details with 13 attributes which includes start and end station IDs with their co-ordinates, start and end times, pass holder type and trip category. (LA Metro Bike Share Data, n.d.)
- 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 stations is shown in the image below (the analysis is done after cleaning the data, pre-processing steps explained below). (LA Metro Bike Share Data, n.d.)
- Currently, the company is operating in 3 regions- Downtown LA, Port of LA and Venice, with DTLA being the leading region accounting for 79% of total trips due to downtown crowd. (LA Metro Bike Share Data, n.d.)
- Venice has shown a significant rise in percentage ridership share (can be seen in the graph below). So, this region can be focused to expand the network of Bike stations to increase the number of trips.
- Monthly Pass is the most popular among the riders as compared to other pass types. Unfortunately, we don't have the user data to analyze the trips of individual users.

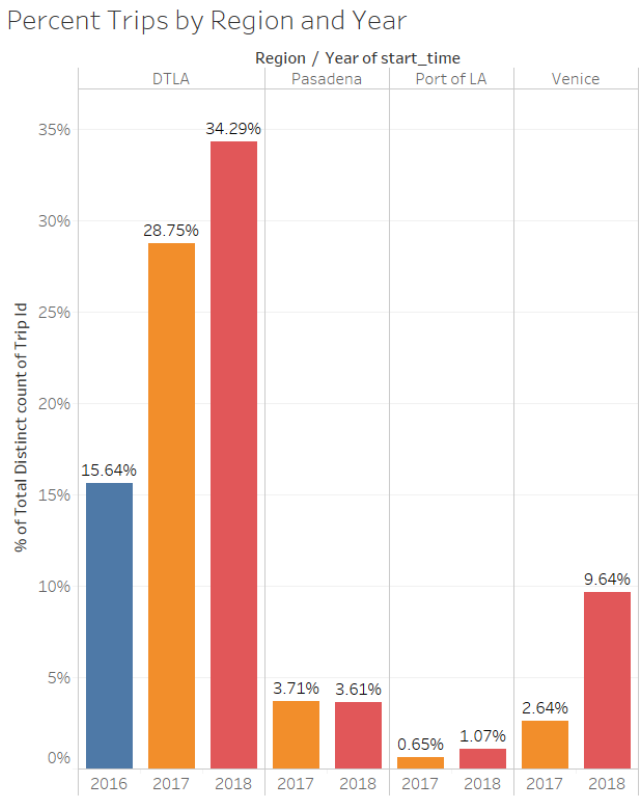

Figure 1- Percent Trips by Region and Year

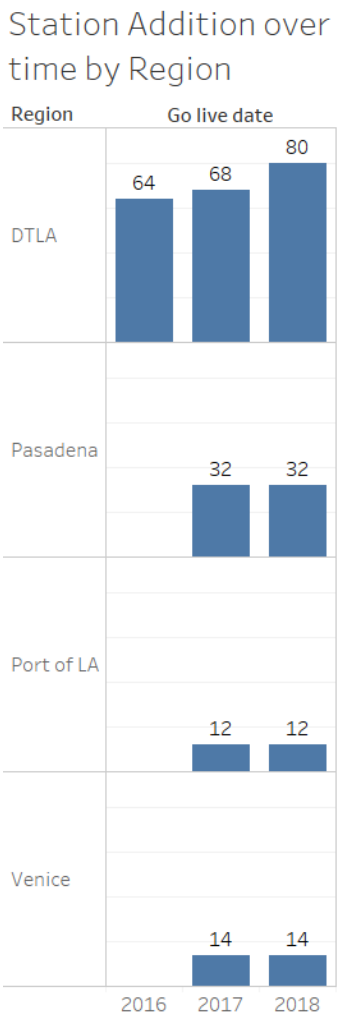

Figure 2- Station Addition by Region

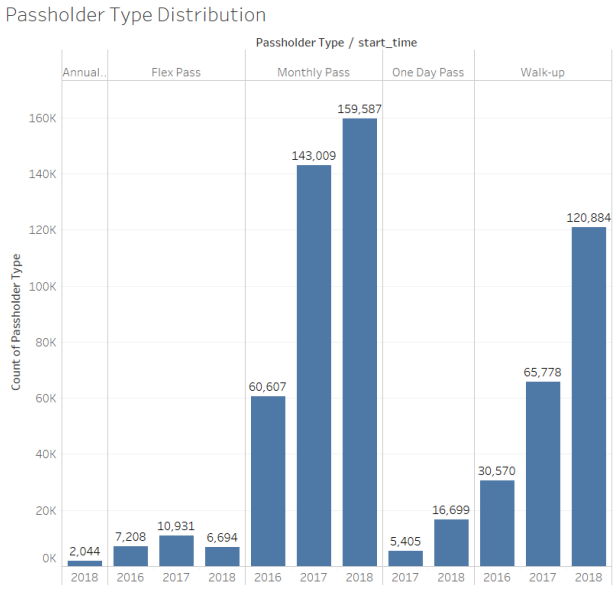

Figure 3- Passholder Type Distribution

Problem Statement

- Forecasting bicycle Demands for the Q3 2018 thru Q1 2019
- Revenue Recommendations
- Suggesting Potential new station locations

DATA PRE-PROCESSING

Pre-processing was performed in two steps.

- Data Cleaning – Imputation of Missing Values and removing outliers
- Feature Engineering – Creating new attributes for analysis from available attributes

Data Cleaning

The following steps were followed for data cleaning (used Structured Query Language):

- Created a new column 'trip_duration' with unit as minutes (part of feature engineering).
- Removed all trips which have a length less than a minute or more than 24 hours as they are outliers as per the competition website.
- Removed station ID 3000 as it's a virtual station (has majority of the rows amongst removed data).
- Removed stations with ID 4110,4118,4276 as they were not used in any trip and their station information such as region, go live date attributes are blank.
- Removed inactive stations (43 stations) that is as of Q2 2018.
- Encoded missing values in end_station attribute in LAbike dataset by using the condition start_station=end_station for all Round Trips. (LA Metro Bike Share Data, n.d.)
- Updated null values of end_station attribute using bike id attribute on where the bike was used prior and subsequently for all One-way Trips.
- Updated null values in start_lat,start_lon,end_lat,end_lon attribute using maximum occurrence of the latitude, longitude coordinates in the LAbike dataset for a given station. (LA Metro Bike Share Data, n.d.)

Initial data had 639786 rows. After cleaning the dataset, it has 629416 rows. So, removed 10,370 rows (1.6% data).

Feature Engineering

- **Count** – To forecast future trips, we aggregate the count (number of trips) on hourly, weekly, monthly quarterly basis as shown in the figure below. Used the trip start time attribute for analysis by converting it into a datetime datatype and extracting month, year etc. to calculate the count of trips.
- **Duration**- To analyze the dataset and to remove the outliers, we created a trip duration attribute by subtracting start time from end time attribute.

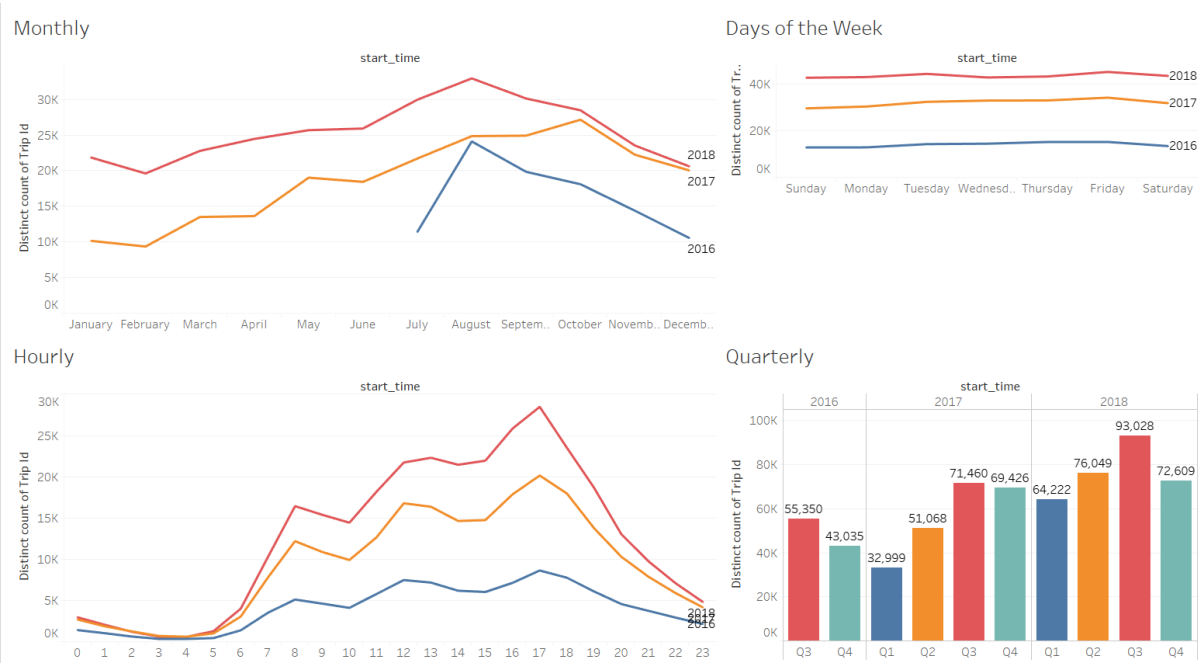

Figure 4 – Number of Trips during an hour, week, month and Quarter

FORECASTING TRIPS AND BICYCLE DEMAND

The initial step was to create a pivot between the start-time and bike stations. We created a column of date-time indices. From a date-time index, we rolled it at a daily level to start our proceedings. There were huge spikes in demand on certain dates. This was due to a regular cycling competition happening in LA. The following dates were also removed from the model - Oct 16,2016, Oct 08, 2016, Sep 30, 2018 and Dec 2, 2018. (Past Bike Ride Competition, n.d.)

For forecasting, Facebook's Prophet Library in python was used. The data was split into 80-20 train- test and prediction was done for Q1 2019. This was done at a bike station level by running a loop over all stations. (Prophet Algorithm, n.d.)

Dashboards are built in tableau that gave the trend, seasonality, variation and spikes in demand. There were similar patterns in quarterly, monthly levels with demand increasing in summer and spring and the least during winter and rainy seasons. The demand was correlated with the weather data by getting data from the API. There was no significant pattern at a weekly level. The demand is stationary irrespective of the day of the week. And, as expected at hourly level, there were peaks in demand at school hours.

After incorporating the above observations, fit the model to get the predictions. Used MSE as the loss function to evaluate the performance of the models. The data was aggregated for visualizing at all levels.

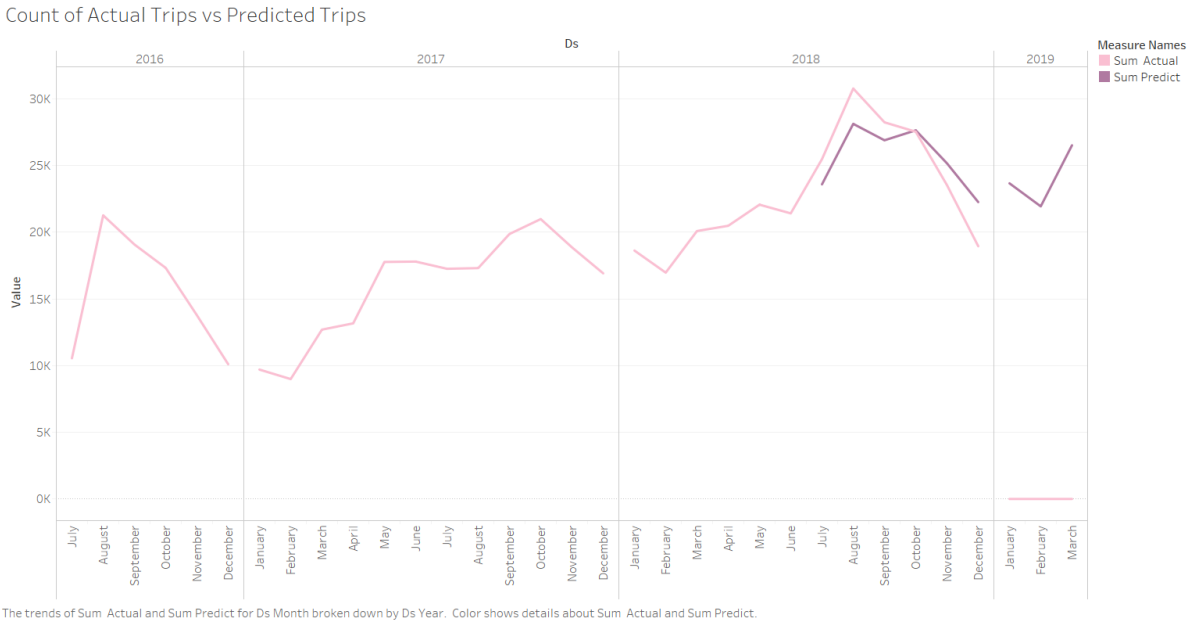

Figure 5 – Count of Actual Trips vs Predicted Trips

PRICING RECOMMENDATIONS

The revenue of the LA Bike Share program can be divided into two components-

- Cost of subscribing to a pass
- Cost of each ride when the duration exceeds thirty minutes

To calculate the revenue obtained in the bike share program when a ride exceeds thirty minutes, use trip duration attributes, adjusting based on the passholder type. Cost of each ride is calculated using the appropriate price (the price changes on July 12, 2018, the cost of a standard unit is reduced from $3.5 to $1.75).

A pivot table is used to aggregate the revenue earned for each date. For pricing recommendation, it is assumed that the number of trips for each passholder type is proportional to the actual number of users in the category. The following observations were made:

- Flex pass holders showed a decreasing trend, suggesting the presence of an alternative service which is more feasible and hence price reduction is required to increase the usage of the flex pass holders.
- The revenue generated through Monthly pass rides changes around and hence no price modification is required.
- One day pass shows an increasing trend however it can be clearly seen there is a major dip in its revenue during Q4. This could be due to the cold weather during December.
- Walk up pass shows an increasing trend which means the prices can be increased to generate more revenue as there is an increasing demand.
- For Annual Pass, it was used from September 2018 onwards, so it shows an increasing trend, however cannot comment with certainty about its trend due to insufficient data.

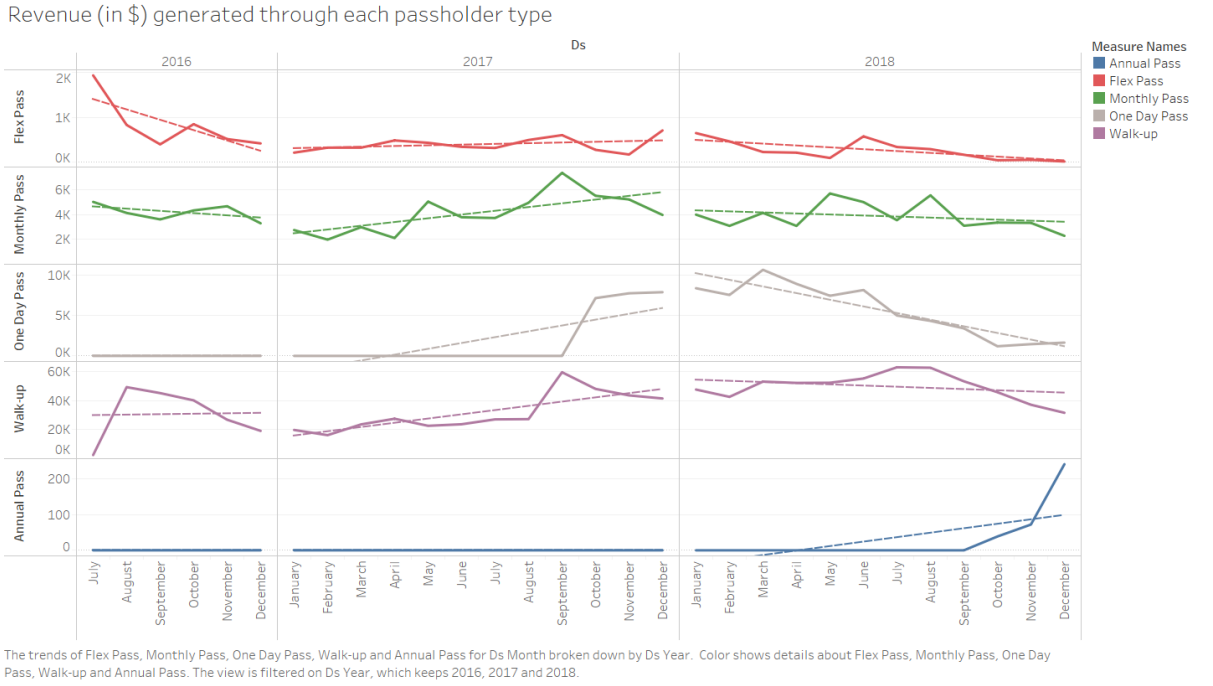

Figure 6 – Revenue (in $) generated through each passholder type

NETWORK MANAGEMENT

Analysis of Average Number of Rides per Bike

- Although the bike prices changed on 12th July 2018, Pasadena continued to go down in terms of average number of rides per bike and hence service was discontinued in that area.
- DTLA is fairly consistent in all quarters throughout the 3 years and service must be continued with the same trips per bike ratio.
- During the Q4 of each year, Port of LA and Venice may perform badly due to unsuitable weather conditions for biking. Hence, we keep the ratio of number of rides to bikes at around 13 for Port of LA and 58 for Venice.

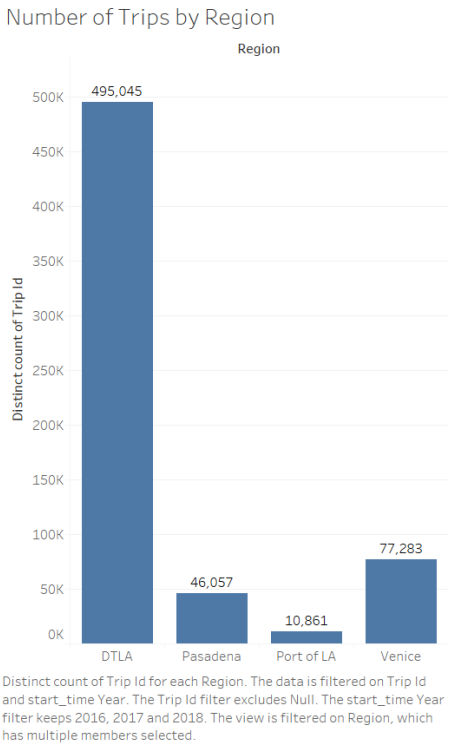

Figure 7 – Trips by Region

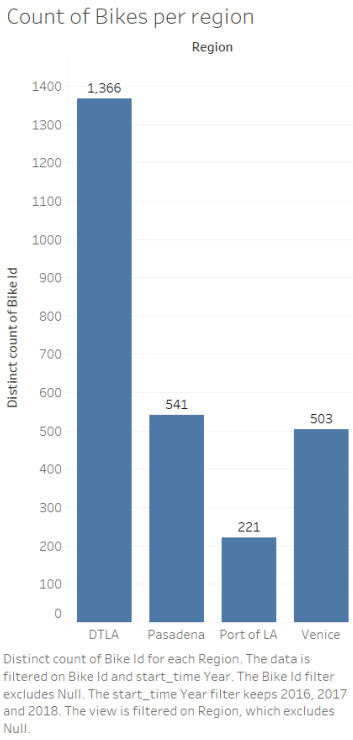

Figure 8 – Bikes per Region

Number of Trips per Bike

	start_time									
	2016		2017				2018			
Region	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
DTLA	72.73	59.11	44.06	69.13	62.01	56.05	52.43	58.77	58.18	59.37
Pasadena				25.00	35.38	22.57	21.00	26.83	17.76	
Port of LA					19.63	12.48	9.85	11.39	21.94	8.07
Venice					22.34	66.08	56.79	59.11	76.75	43.31

Figure 9 – Trips per Bike

- Now to predict the number of bikes, we use the method of Moving Average of past 4 periods:

$$Number\ of\ Bikes = \frac{Number\ of\ Trips}{Avg(Average\ Number\ of\ Rides\ per\ Bike\ in\ 2018)}$$

- Accordingly, **DTLA** will require **717**, **Port of LA** will require **212** and **Venice** will require **485** bikes in **Q1 of 2019**.

Estimated Number of bikes for Q1 of 2019			
	No of Trips Estimated	Average Number of Rides per Bike	No. of Bikes
DTLA	40787	56.93	717
Port of LA	2715	12.806	212
Venice	28621	59.106	485

Recommendations

- From the analysis, it is clear that places close to coast or beach and major downtowns expand quickly. Cities away from major business capitals (such as Pasadena) decline after a point of time and shouldn't be considered for future expansion.
- In **Port of LA,** all the stations are along the coast or near the beach. Recommendation is to expand into the interiors so that people and tourists can travel from their houses/hotels to the coast.
- Expand the company by installing stations at **Long Beach** and its interiors as it is the **39th most populous city** in the **United States of America.**

REFERENCES

Past Bike Ride Competition. (n.d.). Retrieved from Ciclavia: https://www.ciclavia.org/events_history

Prophet Algorithm. (n.d.). Retrieved from Facebook.github: https://facebook.github.io/prophet/docs/quick_start.html#python-api