Executive summary of the problem

We analyzed the Los Angles Metro Bike Share data provided to us from Q3 of 2016 till Q4 of 2018. Our forecast was at the overall level and also at the individual region level since the metro serves 4 stations. Our forecast ultimately supports the proposal to expand into newer regions in LA. We came up with the potential candidates for expansion based on our analysis. Our models predict the number of trips for the last quarter of 2018 and Q1 of 2019, thereby predicting the number of bikes at a station level. We also give quantitative and qualitative recommendations for possible pricing changes by forecasting income from ticket sales and expanding the network. We evaluate characteristics for the region to be successful

General Approach

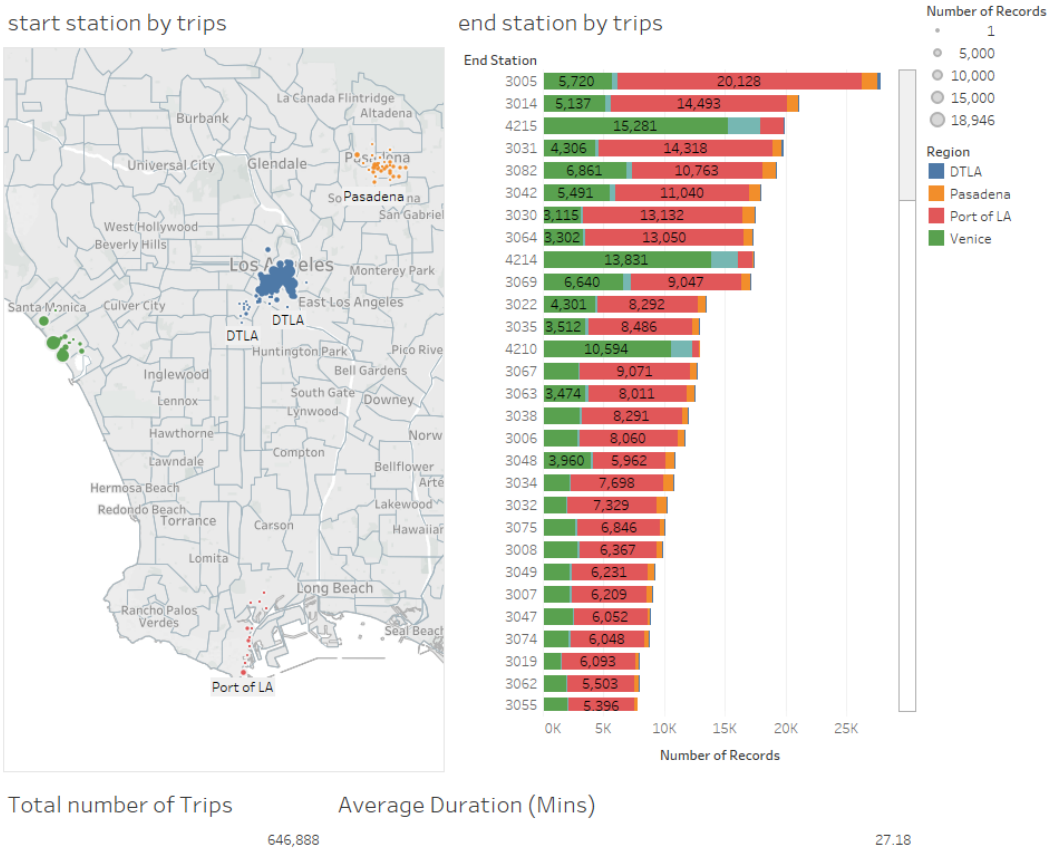
We used some descriptive statistics to examine the data. We followed this with a replace impute encode procedure to get rid of the outliers. Then, we trained and tested the model to make predictions until Q1 of 2019. We used an open source package by facebook – called prophet. The recommendations are visually shown in tableau.

Estimated Benefits

1. The revenue is sure to increase with the expansion.
2. Bike utilization is also a key parameter that we focused on, which is bound to increase at all regions.
3. The pricing recommendations we suggest in sure to increase the activity in winter months.

The problem and data collection

We were given with two data files – the bicycle trip data and the station data. Descriptive statistics for each feature was analyzed. Bike IDs were used as a feature to forecast the bike staging part of the problem. Start station was an important feature as that forms the base of our forecast number of trips. Start times were used to pivot as index values. Lat – long details were used to geo code and get the correlation from weather data. Passholder type and plan duration are related to each other. Also, using station ID as the key both tables were left-joined



Data Preprocessing

In Station Table - Updated the missing values in Region, go_live_date, status column

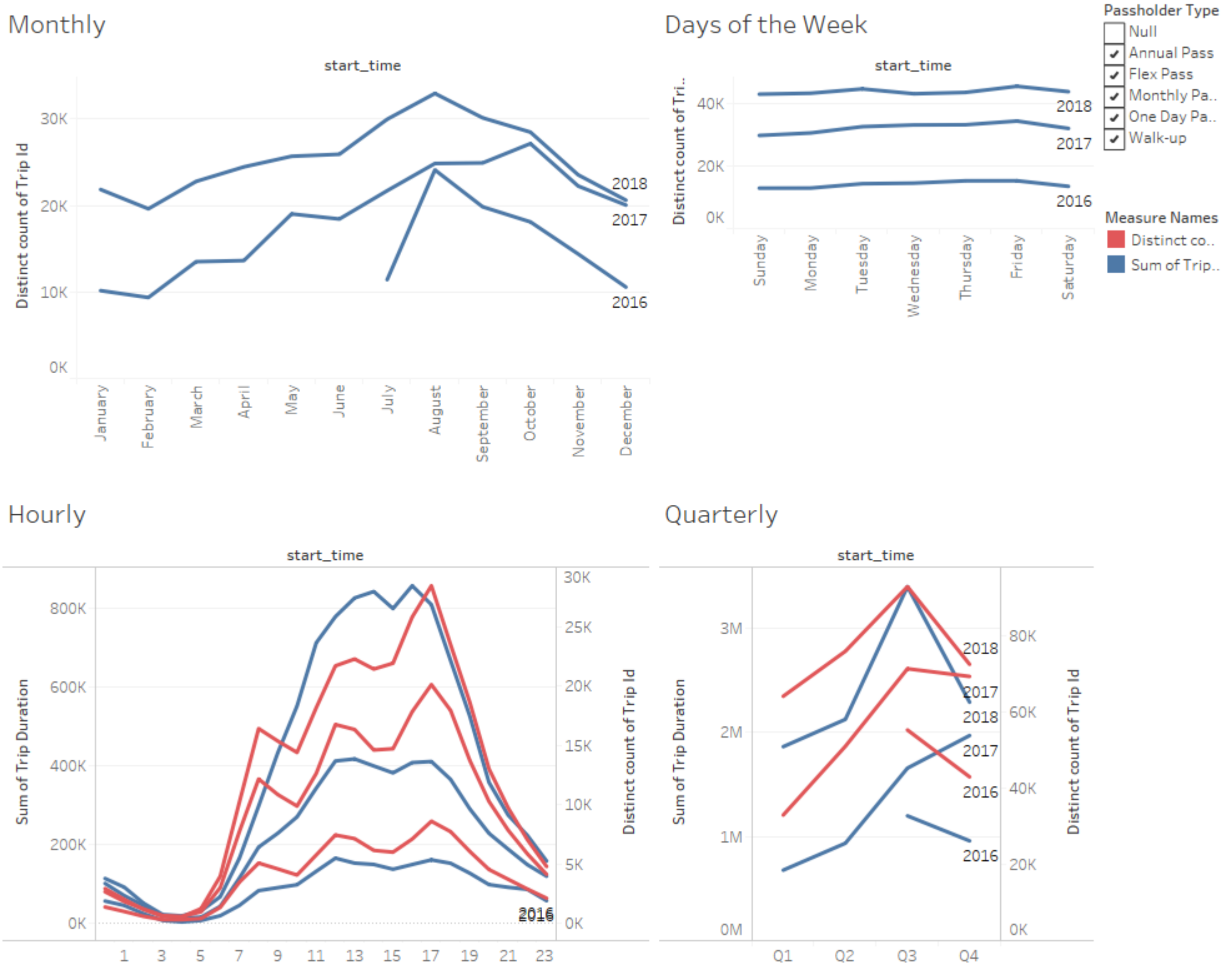
In Trip Table - Updated trip information.

The following steps were followed for data preprocessing:

- Created a new column ‘trip_duration’ (unit=minutes)
- Removed all trips which have a length less than a minute or more than 24 hours (outliers as mention in the competition website)
- Removed station ID 3000 as it’s a virtual station (has majority of the rows amongst removed data)
- Used the fact that Start_station=end_station for all Round Trips, and encoded missing values
- Updated null values of end_station using bike id – on where the bike was next used and previously used.
- Updated null/0 values in start_lat,start_lon,end_lat,end_lon

Initial data had 639786 rows. New data table has 629416 rows. So, removed 10,370 rows (1.6% data)

Few stations were removed from the analysis - 43 start stations are inactive right now + 3 stations (4110,4118,4276) which had station table rows blank - 4110 4118 were never in start_station.



Forecasting Trips and Bicycle Demand

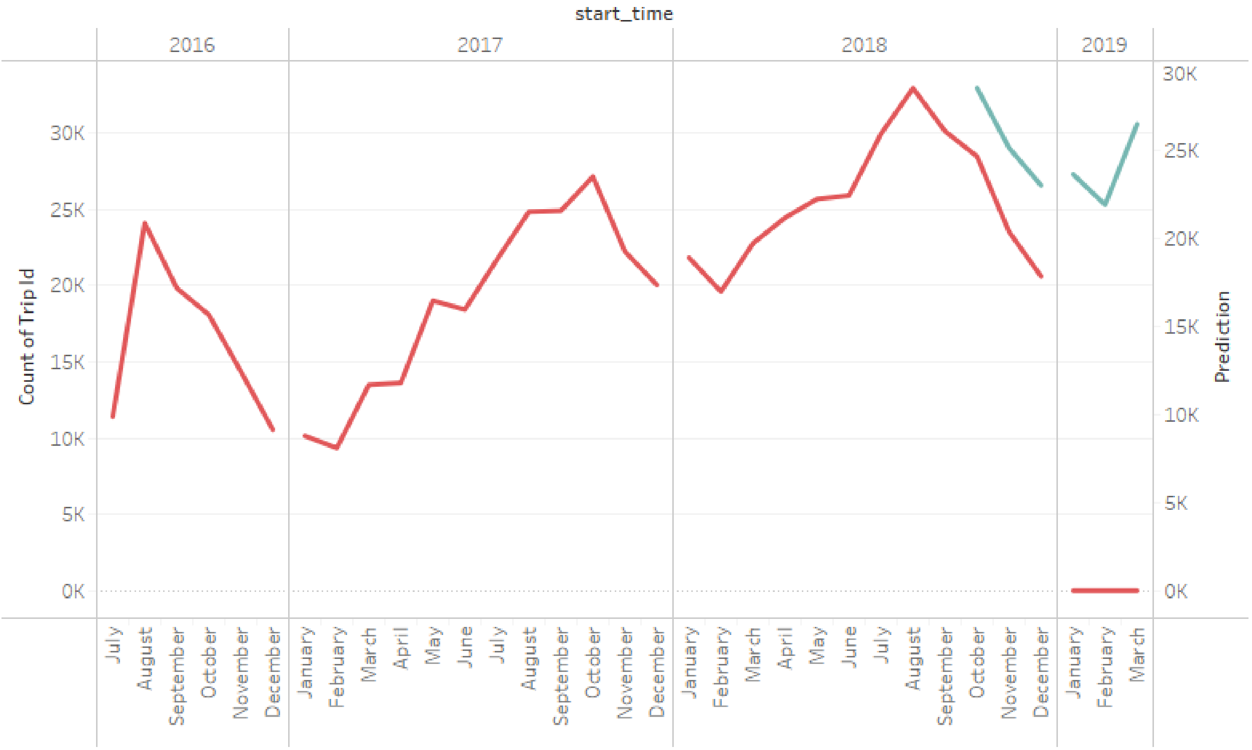
The initial step was to create a pivot between the start-time and bike stations. We created a column of date-time indices. From a date-time index, we rolled it at a daily level to start our proceedings. We removed the few dates that were from Q1 of 2019.

There were huge spikes in demand on certain dates. We figured out that this was due to a regular cycling competition happening in LA. https://www.ciclavia.org/events_history . These dates were also removed from the model.

We used the facebook's prophet library in python to do our forecasting. We did 80-20 train-test split and also predicted for Q1 of 2019. This was done at a bike station level by running a loop over all stations.

We built dashboards in tableau that gave us the trend, seasonality, variation and spikes in demand. We were able to get it at quarterly level, monthly level, weekly level and hourly level. We noticed similar patterns in quarterly, monthly levels with demand increasing in summer and spring and the least during winter and rainy seasons. We correlated this with the weather data by getting data from the API. We also noticed there were 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. We incorporated this into the model and got our predictions. Used MSE as the loss function to evaluate the performance of the models. We then aggregated this for visualizing at all levels to present our forecast in tableau.

Predict at a region level and at a station ID level



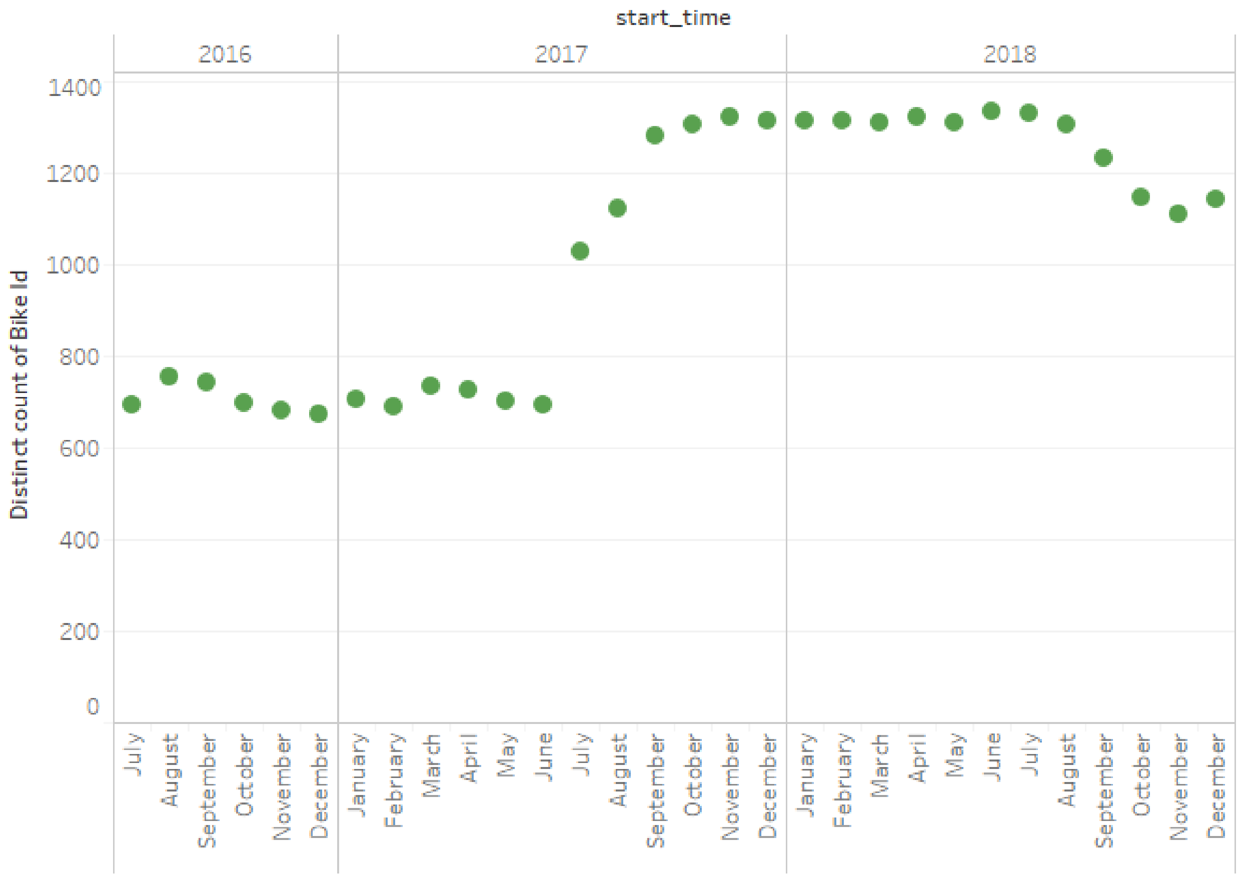
The trends of count of Trip Id and Prediction for start_time Month broken down by start_time Year. Color shows details about count of Trip Id and Prediction. The data is filtered on Region and Bike Id. The Region filter keeps DTLA, Pasadena, Port of LA and Venice. The Bike Id filter keeps 1,505 of 1,505 members.

Measure Names

🟥 Count of Trip Id

🟩 Prediction

Bike Data - Shows utilization and when previous expansion was made



Distinct count of Bike Id for each start_time Month broken down by start_time Year. Color shows details about distinct count of Bike Id. The data is filtered on Region, which keeps DTLA, Pasadena, Port of LA and Venice. The view is filtered on Exclusions (MONTH(start_time),YEAR(start_time)), which keeps 30 members.

Measure Names

🟩 Distinct count of Bike Id

Forecasting income and pricing recommendations

The revenue of the LA Bike Share program can be divided into two components, the cost of subscribing to a pass and the cost of each ride when the duration exceeds standard unit of thirty minutes.

We calculate the revenue obtained in the bike share program when a ride exceeds thirty minutes by finding the no of standard units of duration that the bike as been rented and we adjust this quantity based on the passholder type. We then obtain cost of each ride obtained by 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).

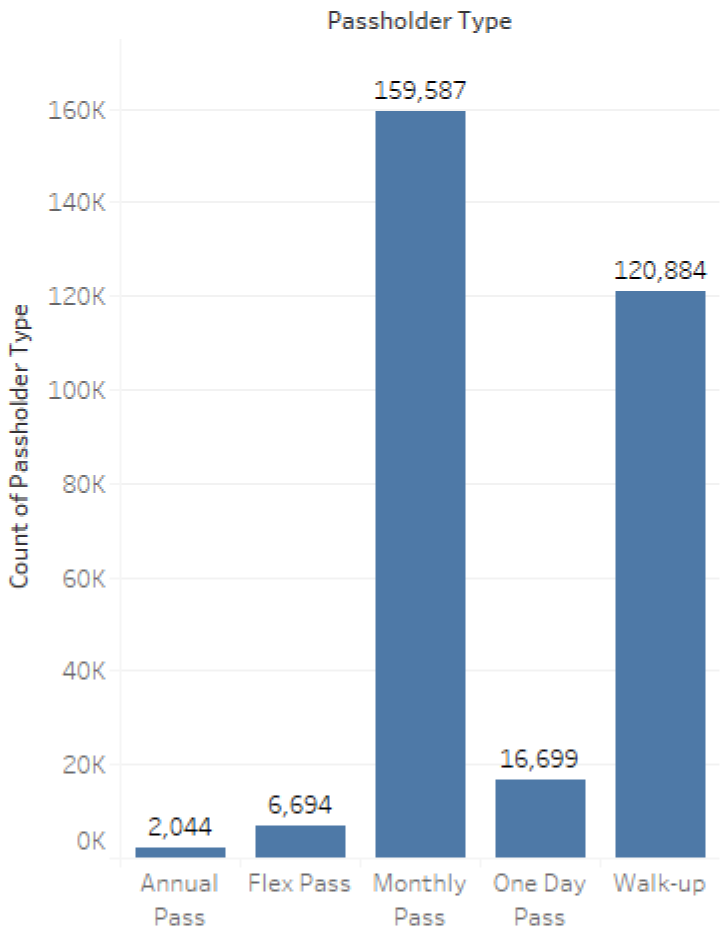
Once the revenue obtained from each ride is calculated we create a pivot table that aggregates the revenue earned for each date. We then use this data to estimate the future revenue obtained from every ride by using fakebook's prophet library. We use a 80-20 train test split and predict the revenue for Q1 of 2019. We used the weather data to identify the dips in demand. We use the events data to identify dips or peaks in the data. We use MSE as the loss function to evaluate the model.

A slightly higher test MSE is obtained for the training MSE due to the removal of service from Pasadena. An RMSE of lesser than one percent of the forecast is obtained.

To estimate the trend to subscription of the pass we forecast the number of rides each passholder type will undertake in the quarter 1 of 2019 we first create a pivot table from the data that aggregates the number of rides that passholder of each type has undertaken. We then use prophet to forecast the demand for quarter 1 in 2019.

We make the pricing recommendation based on the assumption that the number of trips for each passholder type is proportion to the actual number of users in the category. The following observations were made , the number of flex pass holders showed a decreasing a trend which suggests the presence of an alternative service which is a more feasible and hence we need to reduce the price to increase the usage of the number of flex pass holders. The forecast of the number of monthly pass rides changes around a mean and remains constant and hence we do not make changes to the price of monthly pass. One day pass show a similar trend in its forecast. The forecast for the number of people having walk-up prices increases which suggests the absence of alternative services. The price of this service can be increased.

Passholder Type by Numbers



Count of Passholder Type for each Passholder Type. The marks are labeled by count of Passholder Type. The data is filtered on Region and Start Time (copy) Year. The Region filter keeps Null, DTLA, Pasadena, Port of LA and Venice. The Start Time (copy) Year filter keeps 2018. The view is filtered on Passholder Type, which keeps Annual Pass, Flex Pass, Monthly Pass, One Day Pass and Walk-up.

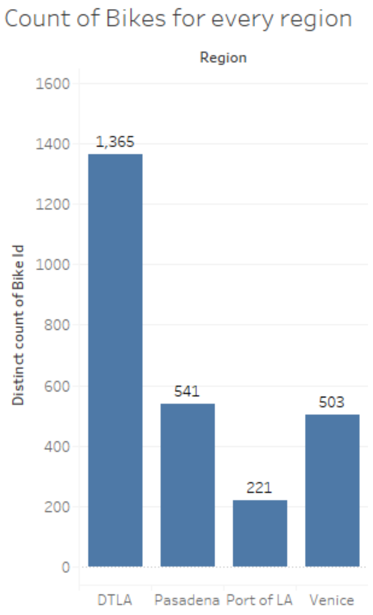
Characteristics that identify whether a region will be successful

- Whether a region is a metropolitan.
- Whether a region has a beach or it is close to the coast.
- Analysis is explained in Network Management.

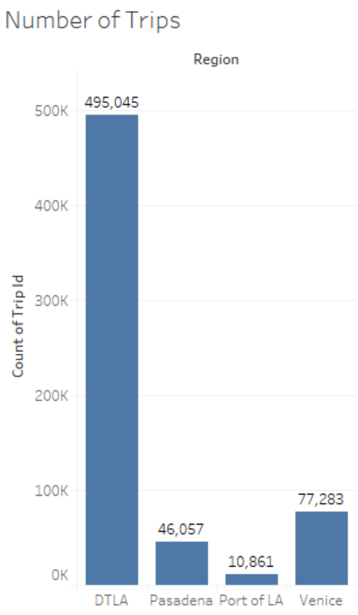
ANALYSIS FOR AVERAGE NUMBER OF RIDES PER BIKE:

Key Points:

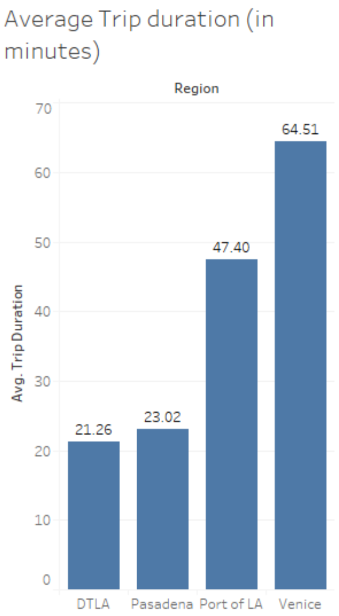
- Although the bike prices changed on 12th July 2018, **Pasadena** continued to **go down** in terms of average no. of rides per bike and hence **service was discontinued** in that area.
- **DTLA** is fairly **consistent** in all quarters throughout the 3 years and service must be continued with the 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, keep the ratio of number of rides to bikes at around 15 for Port of LA and 55 for Venice.



Distinct count of Bike Id for each Region. The marks are labeled by distinct count of Bike Id. The view is filtered on Region, which keeps DTLA, Pasadena, Port of LA and Venice.



Count of Trip Id for each Region. The marks are labeled by count of Trip Id. The view is filtered on Region, which keeps DTLA, Pasadena, Port of LA and Venice.



Average of Trip Duration for each Region. The marks are labeled by average of Trip Duration. The view is filtered on Region, which keeps DTLA, Pasadena, Port of LA and Venice.

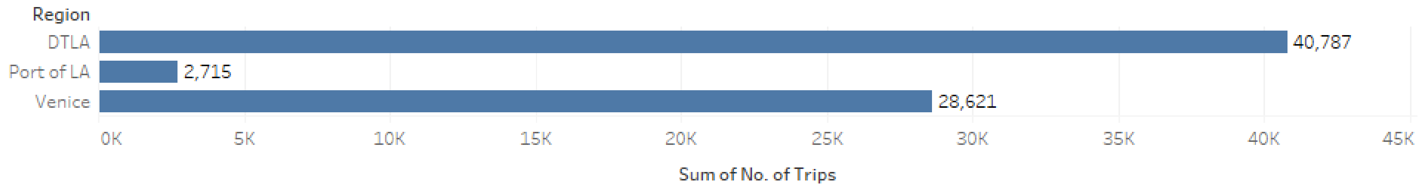
	Year 2016				**Year 2017**				**Year 2018**			
	Q1	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*
DTLA	N/A	N/A	72.733	59.183	44.11364	69.13008	62.01102	56.05212	52.42519	58.76636	58.17669	59.37095
Pasadena	N/A	N/A	N/A	N/A	N/A	25	35.38107	22.56995	20.99718	26.82571	17.75904	N/A
PLA	N/A	N/A	N/A	N/A	N/A	N/A	19.625	12.48031	9.848739	11.38926	21.93798	8.068182
Venice	N/A	N/A	N/A	N/A	N/A	N/A	22.33735	66.08205	56.78661	59.58943	76.7491	43.31373

- Now to predict the number of bikes, we use the method of Moving Averages 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.**

Q1 Trips Prediction



Sum of No. of Trips for each Region. The marks are labeled by sum of No. of Trips. The data is filtered on Ds Quarter, which keeps Q1.

Estimated No. of bikes for Q1 of 2019			
	No. of Trips Estimated	Yearly Average of 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

Summary

1. The Bike Share metro is doing good business overall which is evident from the trend.
2. With expansion comes along the cost of capital and purchasing more bikes. This may or may not be done due to the bike utilization statistics explained earlier. Bike staging is taken care under this.
3. Pricing according to pass type also demands a chance as discussed.
4. From the analysis, it was clear that places close to coast or beach and major downtowns expand quickly. Cities away from major business capitals decline after a point of time and shouldn’t be considered for future expansion.
5. In **Port of LA,** all the stations are along the coast or near the beach. Our recommendation is to expand into the interiors so that people and tourists can travel from their houses/hotels to the coast.
6. 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.**