***Part 1: Executive Summary***

The general approach we used to solve the problem is the **SEMMA** approach. The abbreviation stands for: Sample Explore Modify Model Assess. This is a systematic approach to solving big data problems. A brief outline is listed below.

**Sampling** involves gathering data. We did not spend time gathering data because the data is readily available.

**Explore & Modify** is data preprocessing. We spent the majority of our time preprocessing the data in order to create our models. The core steps in this process are: identifying and correcting outliers, identifying attributes with missing values and correcting them if possible, identifying a target attribute for the models, examination of the input attributes with plots and tables, creation of new attributes, examination of the relationship between the input attributes and target attribute.

**Modeling** is the process of creating models to predict or explain the target attribute.

**Assessment** involves assessing how well our models perfom and selecting the best model for the posed question.

Provide an executive summary of the problem, the general approach used to address the problem, and the estimated benefits for adopting these recommendations.

2. The Problem & Data Collection:describe the data used for this analysis and the teams view of what problem(s) they tried to address using these data.

3.Data Preprocessing: describe how outliers were identified and addressed along with missing values and what addition parameters, if any, were created to better model these data.

4.Forecasting Bicycle Demand: describe how the model for forecasting bicycle demand was developed and validated. With this model, prepare forecasts for Q3 2018 thru Q1 2019.

5.Forecasting Ticket Sales:describe how the model for forecasting monthly ticket sales and revenue was developed and validated. With this model prepare forecasts for Q3 2018 thru Q1 2019.

6.Success Characteristics:Identify characteristics that are associated with a successful or unsuccessful region.

7.Network Management:based upon your analyses, describe recommendations for staging bicycles in 2019 and for expanding the network.

8.Summary:Final conclusions and recommendations

***Station Data:***

Three stations are missing information. We can look at the bike data to determine latitudes and longitudes. Then, use google to determine what the station is and how it was used.

**Station 4110** - lat = 34.04387, lon=-118.2576, region=DTLA, name=SoulCycle, status=Inactive

This station was active from September 7 to September 30, 2017. It is in front of a fitness place called SoulCycle. Google Maps does not show any bike racks here. Perhaps it was a temporary station.

**Station 4118**: lat = 34.02611, lon= -118.2383, region=DTLA, name=Channing Street, status=Active

This station appears as a smart bike on the Metro Bike Share website. It appears to be in a sketchy area of town. In one of the Google Maps pictures, there is a metro bike van parked outside.

**Station 4276**: lat = 34.04739, lon=-118.2188, region=DTLA, name=Mariachi Plaza, status=Inactive

This station is in Mariachi Plaza in downtown. All of the rides here were from 12/2/2017. There was some sort of festival going on during this time. <https://www.laworks.com/opportunity/a0C1N00000GHHzqUAH>

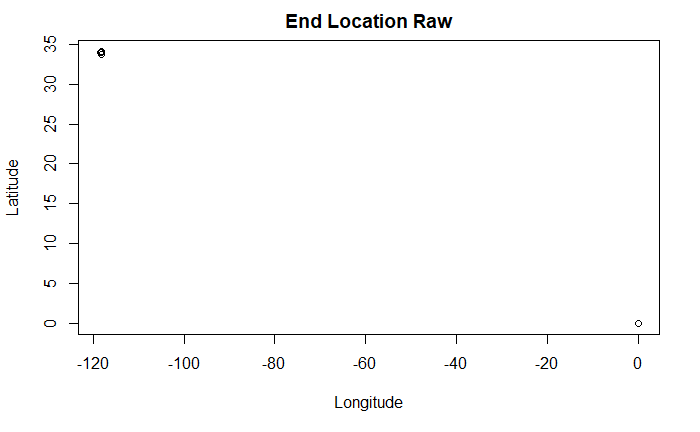
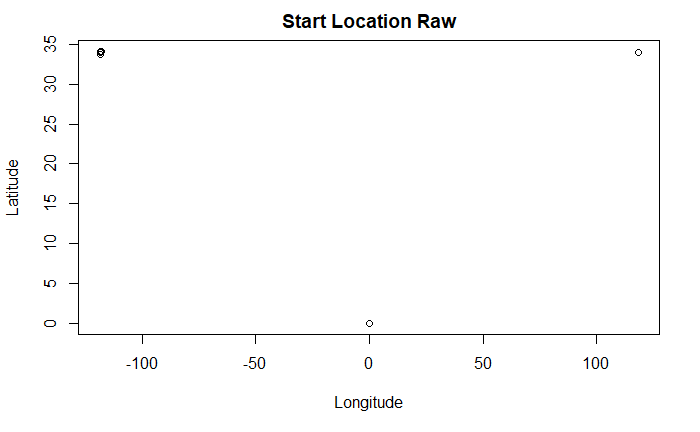
**Station 3000:** This station is not missing in the data. However, there is no location. From the metro bike website: A "Virtual Station" listed in the checkout and return kiosks, is used by staff to check in or check out a bike remotely for a special event or in a situation in which a bike could not otherwise be checked in or out to a station.

**Outliers in Latitude and Longitude:**

There was only one miscoded value identified. The longitude was missing a negative value. I corrected this value. The station location was wrong as well. It was an LA warehouse end locaiton.

There were a lot of zeros coded for the end location. These were all the LA warehouse location as well. Something is odd about this station.

Outlier Detection in the Bicycle Data:



**Start Latitude:** There appears to be some miscoded latitudes and longitudes. The values for Los Angeles should be something around -118 latitude and 34 longitude. There is one miskeyed starting longitude missing a negative sign, trip 28609706. That latitude and longitude (34.02596, 118.2383) corresponds to a starting station of 4118. The actual value coded is station 3066. This station is not in the station data. Use the values from station 4118.

There are 32 data values coded zero for the starting latitude. These values all come from station 4108. We can use the latitude and longitude of this station to replace the zero values. (34.02589, 118.2382)

**Start Longitude:** There are 32 values coded zero for the starting longitude. These are all the same as the 32 values coded zero for the starting latitude.

**End Latitude:** There are 48 data values coded zero for the end latitude. All of these are from station 4108. (34.02589, 118.2382)

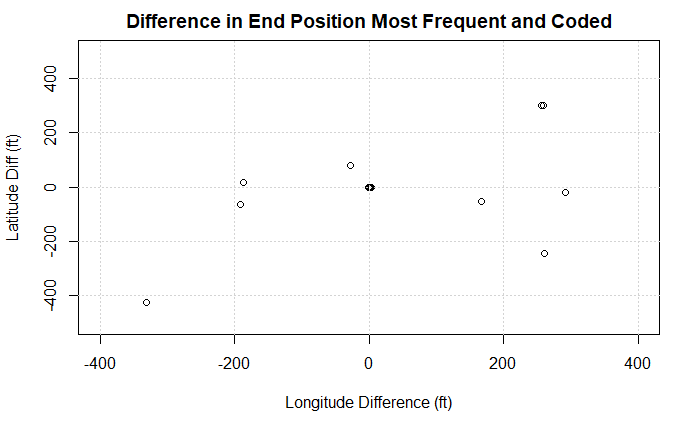
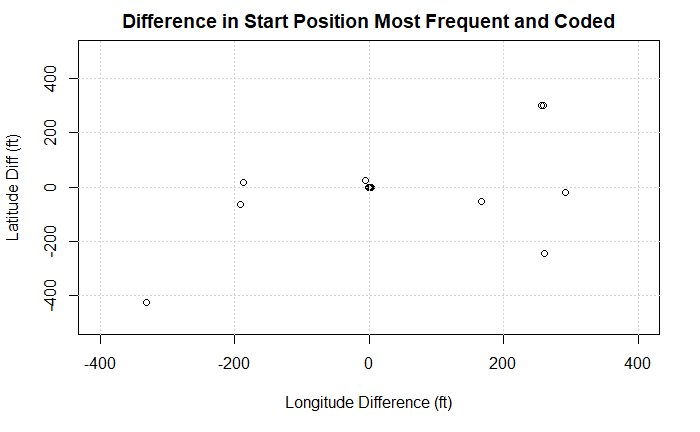
**End Longitude**: There are also 48 data values coded zero for the end longitude. These are all from station 4108 as well. They correspond to the missing end latitudes.

Station 4108 appeared to have some problems recording zeros for the latitudes and longitudes in the first three months of operation. The problem looks like it was corrected after 9/29. However, now there are several values recorded as NA in the longitudes and latitudes.

**New Variables Latitude/Longitude, More Outlier Detection in Latitude/Longitude**:

I added the latitude and longitude for each station to this data set. In order to do this, I took the most frequent latitude and longitude associated with each station in the bike data. The location data in the bike data associated with each station was not unique. This indicates that the stations may have moved at some point. The plots below helped discover which stations had location data coded away from their most frequently coded position.

Distances were calculated using 1 degree = 1 nautical mile = 1.151 miles. 1 mile = 5280 ft. This can be used here because we’re not worried about accuracy, just worrying about identifying differences.



New Station 9999: The plot above helped us determine a temporarily new station, station 9999, for the location at Olive & 5th. It appears that this location is currently at Olive & 6th or station 3036. This is in a location called Pershing square. It looks like the station moved across the square at some point.

The other locations in this plot look like they moved some down the street a couple hundred feet. They are :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Station | Station Name | Movement | n observations | Comments: |
| 3036 | Olive & 6th | Down block | ? | added new stn |
| 4227 | Grand/LATTC | Around Corner | ? |  |
| 4213 | 7th & Westminster | Down Street | 88 | maybe new stn needed? |
| 4148 | Pasadena Library | Around Corner | 4 | all pre live date |
| 3046 | 2nd & Hill | Around Corner | 2405 |  |
| 3005 | 7th & Flower | Around Corner | 7196 |  |
| 4146 | City Hall (Pasadena) | Down Street | ? |  |
| 4108 | LA Warehouse | Behind Fence |  |  |

Need to double check movement on these.

***Probably could label on the plots.***

Most of these stations look like they’ve just been moved a couple hundred feet and don’t need a new location. I felt like the Olive & 5th moved enough to justify a new station name. We can look at modeling to determine if we need to merge it back with Olive & 6th.

Station 4108 is the same as station 4118. Both of these are the LA warehouse. I changed any data point coded as 4118 to 4108. This particular location seems odd. The location is behind a fence and concertina wire. It does not look like it is the best of areas.

***Missing Bicycle Data by Attribute:***

**Start Station:** No missing values

**End Station:** 43198 missing. I determined the value for these using the ending location data. I looked up the latitude and longitude in the station table. Any stations that weren’t filled in were examined. This was another way to discover the Olive & 5th location. There were 486 left. All of these are the same as the missing end latitude and longitude values below. I encoded these as station 3000, virtual station.

**Trip Route Category:** No missing values

**Start Time:** None missing.

**End Time:** None missing.

**Start Latitude:** 1354 missing. All of these have a starting station of 3000, which is the virtual station.

**Start Longitude**: 1354 missing. All of these have missing start latitude and are from the virtual station.

**End Latitude**: 9110 missing. 8624 of these are from the virtual station. The remaining 486 coded as one-way trips with various starting stations. All of the end stations are missing as well. These are all in the period of October 2016 to the end of December 2016. Perhaps the users did not return the bicycles to the station.

**End Longitude:** 9110 missing. 8624 are from the virtual station. The remaining 486 are the same as the 486 missing from End Latitude.

**Plan Duration**: 384 missing. All of these have a passholder type of monthly pass. I substituted 30 in for all of these values.

**Passholder type:** No missing values.

Below is a table of passholder type and coded duration.

Duration\Type

Annual Pass Flex Pass Monthly Pass One Day Pass Walk-up

0 0 0 0 0 132566

1 0 0 0 23319 87262

30 0 0 365449 0 2276

150 0 0 269 0 0

365 2057 25160 1044 0 0

I coded all the 150 as 30 because they are all monthly passes.

Duration\Type

Annual Pass Flex Pass Monthly Pass One Day Pass Walk-up

0 0 0 0 0 132566

1 0 0 0 23319 87262

30 0 0 366102 0 2276

365 2057 25160 1044 0 0

From the table, we can see that there are a lot of walk-ups that buy 1 day passes and some that buy monthly passes. I’m not sure what the monthly pass coded as 365 duration are. They should either be thrown into the monthly pass or annual/flex pass at some point.

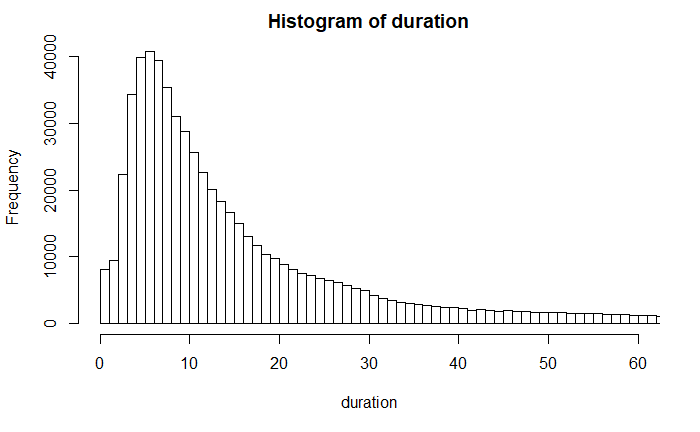
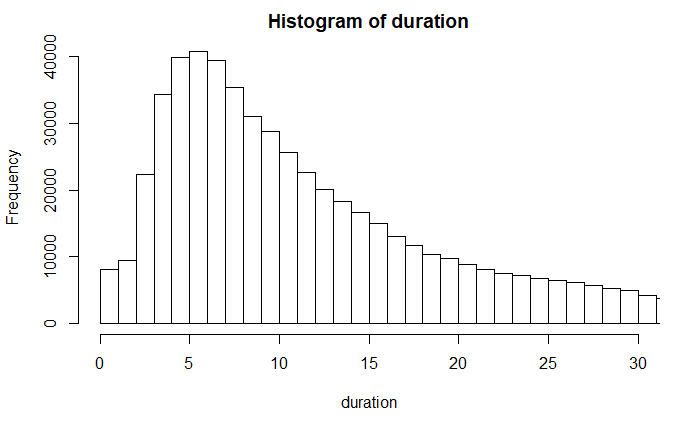
New Variables

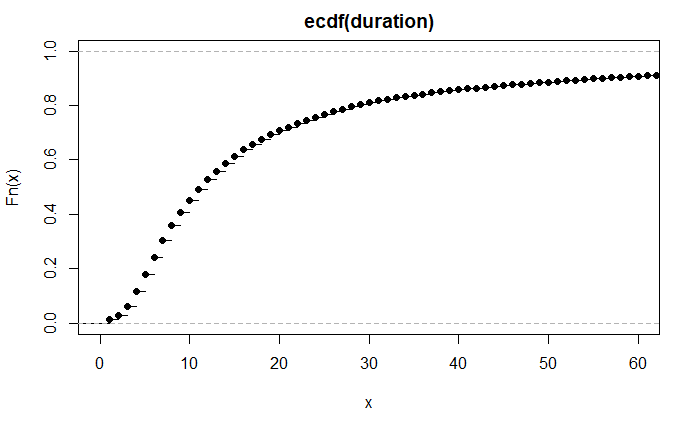
Duration, Distance, Velocity, time(mm-yy)(hh), revenue(estimated),

**Creating Duration:**

I created a duration in the data. I capped the duration at 1 day. There were 2123 observations like this. Many, 1322, had 3000 for an end station. I bet these were bikes that had to be fetched by staff. I ended up recoding the end time as one day and one minute longer. This gives us an opportunity to throw these out or encode a new variable later.

There were 7 observations where the duration was negative. These were from the time change on 11/5/2017. I changed the end time to adjust for the time change. I looked at other time changes, but there weren’t any problems. R calculates time differences and adjusts for the time change.



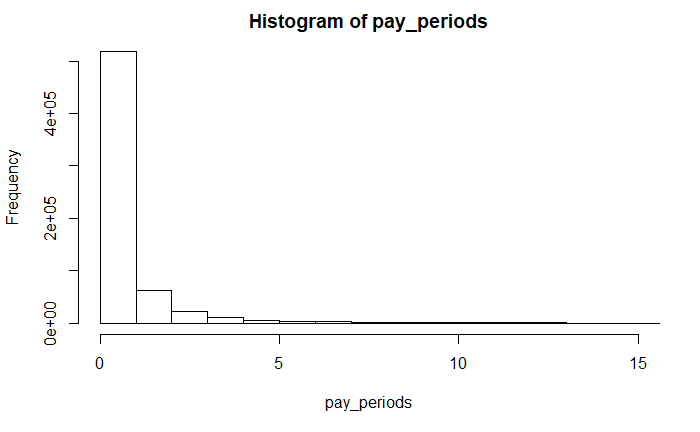
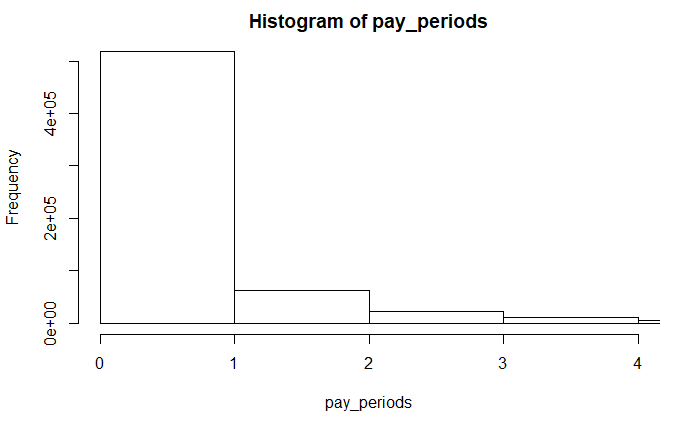


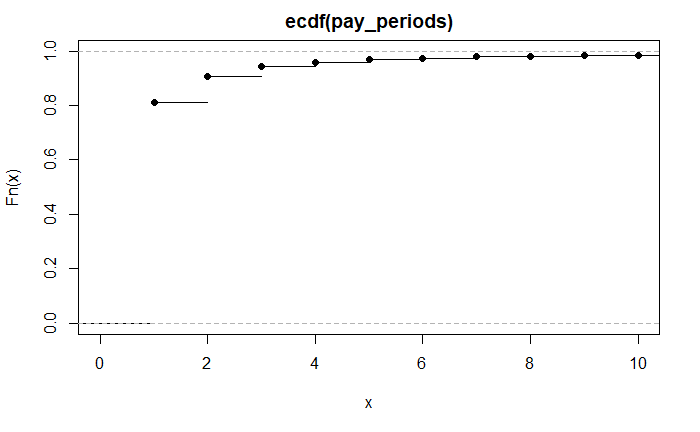
You can see that roughly 90 percent of all the observations are 60 minutes or less with about half of the rides 12 minutes or less.

**We should proably talk about which plots we want to include in our presentation. I’m not sure who our audience is and how well they would be able to interpret a CDF plot.**

**Creating Pay Periods:**

Billing occurs in 30-minute periods. I created a new variable to account for the number of 30 minute intervals the bike was out and about. The plot below shows basically the same thing as the duration, just in a different way.



Link to a map I created with some station data:

https://batchgeo.com/map/baf252fa0064e0636ee697893049661f

Batchgeo, you can create custom maps without having to figure out the google API. It’s kinda nifty.

Our data doesn’t have any stations in Culver city like the website does. It must be that they just launched these called Westside. Pasadena looks defunct. http://www.pasadenanow.com/main/pasadena-plans-to-pull-out-of-metro-bike-share-program-on-tuesday/

**Creating Revenue:**

From the website, the pricing is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Base Fee | First 30 Minutes | Add’l 30 minutes |
| Annual Pass | $150 / year | Free | $1.75 |
| Flex Pass | $10/ month\* | Free | $1.75 |
| Monthly Pass | $17/ month | Free | $1.75 |
| Daily Pass | $5 / day | Free | $1.75 |
| Walk Up | $0\* | $1.75 | $1.75 |

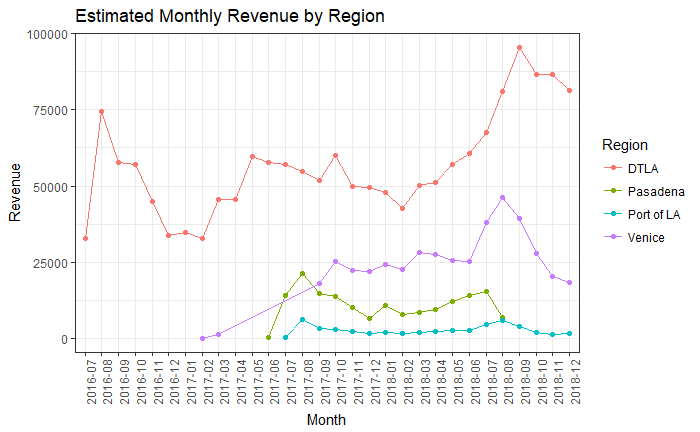
\*I coded flex passes as business passes. This might not be the case

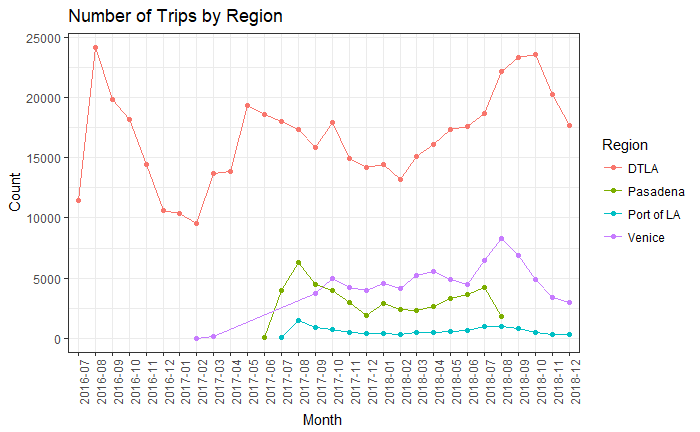
\*\*There is a bulk buy program where you can buy a bunch of passes at a discounted rate

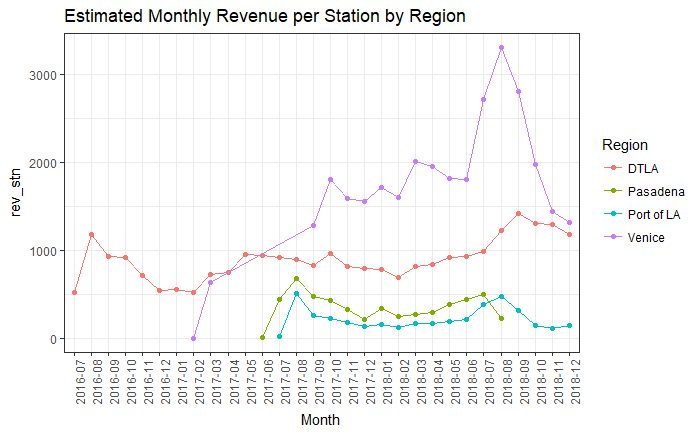
We must estimate revenue because we don’t have any information about passholders. The best we can do is estimate the number of passes from the data on the website. On 3/12/19, the number of trips recorded was 685,388 and the number of passes sold (annual, flex, and monthly) was 65,092. If we multiply the number of passes by the proportion of trips in our cleaned data set, we can estimate the number of passes at the end of 2018: 60,761.

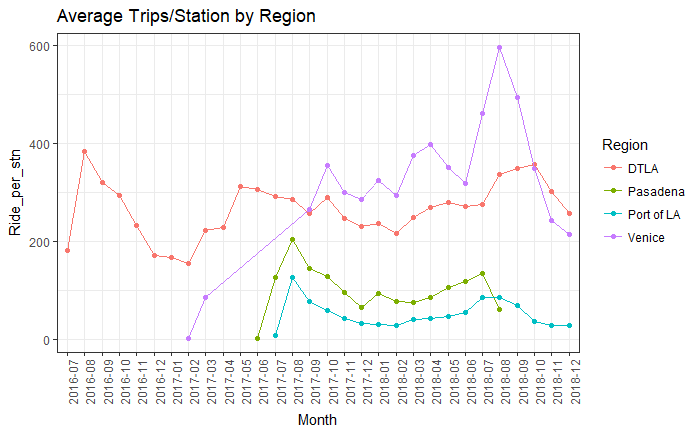
Now that we have the total number of passes, we can estimate the number of passes by passholder type by taking the ratio of trips for passholders in each category and multiplying that by the number of passes calculated above. We can then estimate the total revenue and average revenue per trip by multiplying the cost per pass times the estimated number of passes then dividing by the total number of trips.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Pass Type | Number of Trips | % Trips | Est. Passes | $/Pass | Est. Avg Revenue/Trip | Est. Total Revenue |
| Annual | 2,057 | 0.52% | 317 | $150 | $23.12 | $47,550 |
| Flex | 25,160 | 6.38% | 3,877 | $10 | $1.54 | $38,770 |
| Monthly | 367,146 | 93.10% | 56,568 | $17 | $2.62 | $961,656 |
| Total | 394,363 | 100% | 60,762 | -- | -- | $1,047,976 |

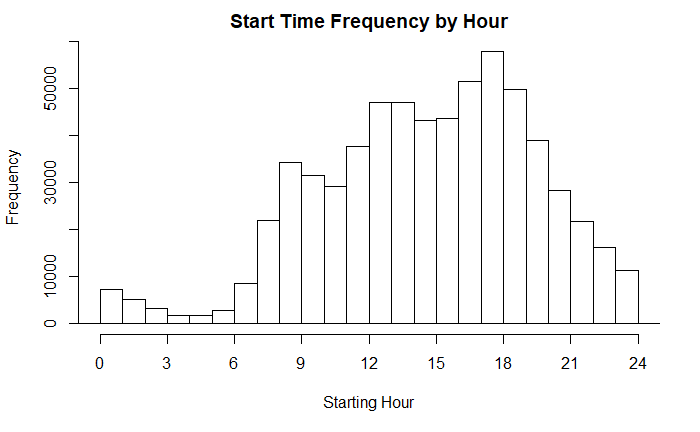




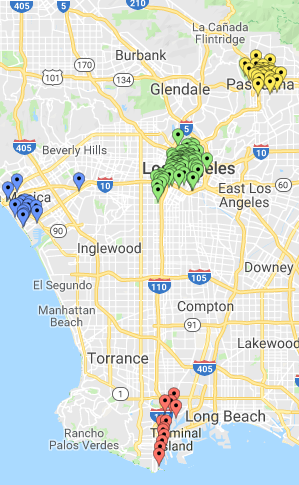




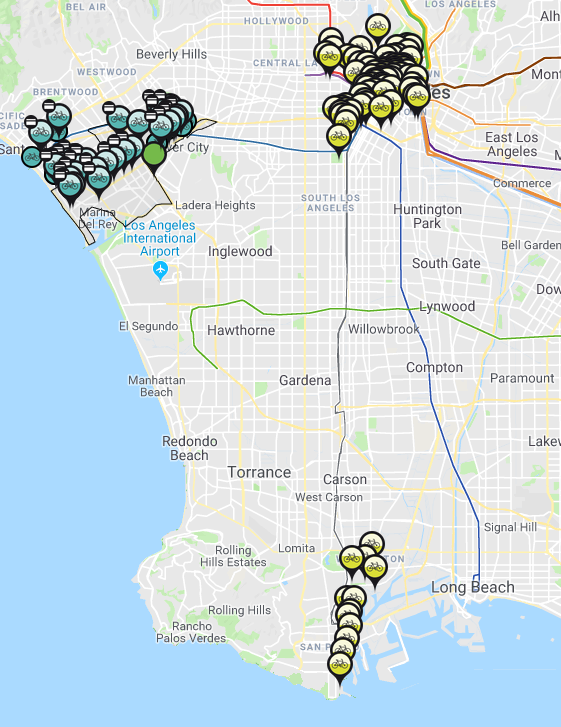
Start Hour: Create this one just to see when the busy times are. We can play with this to see a more granular level.



Custom Map:



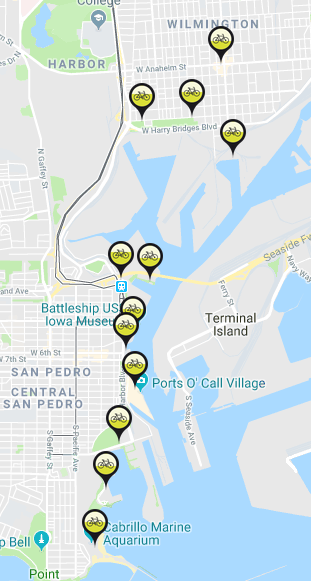
Station Map from website:



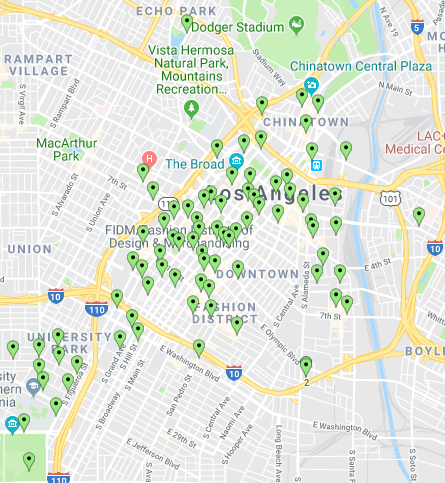
Custom Map for Port of LA:



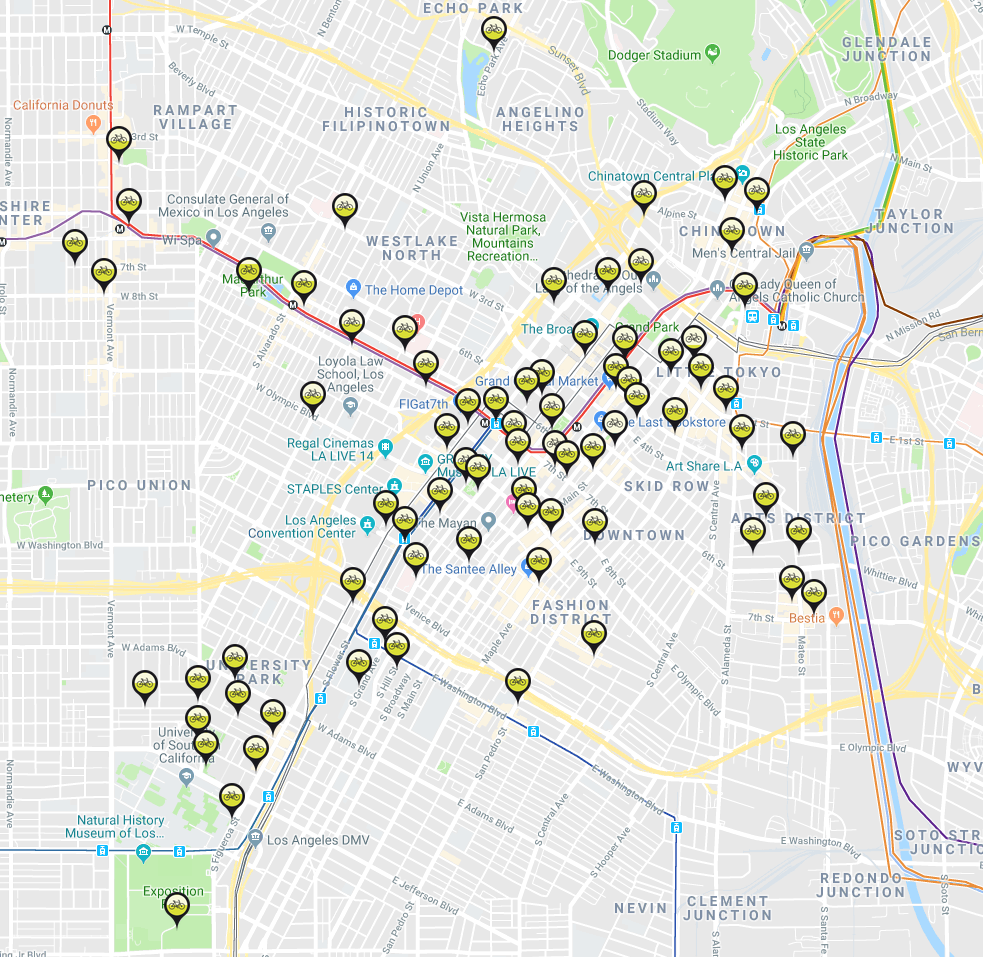
Port of LA



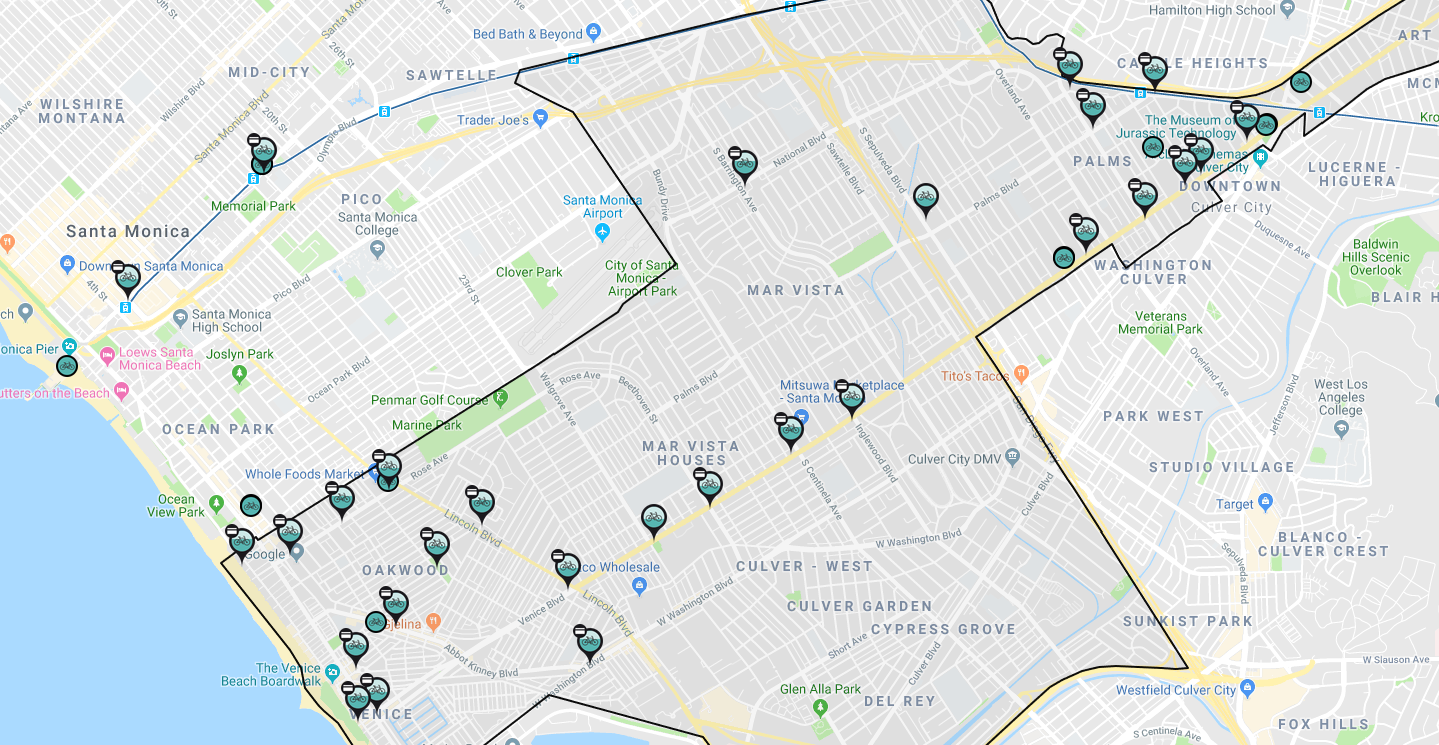
Custom Map for DTLA:



DTLA:



Venice Locations:



Custom Map Locations Venice:

Custom Map for Pasadena:



There is a location at (34.02448, -118.3939), station 3039 in the station data. It is not on the website either. . In addition, it looks like the westside stations are not there. The location is cluver and Washington in Venice Beach. There’s 115 data points here. The first 2 are on 3/17/17 and the bike id is “Dock Block 2”. The rest of the trips are on 3/26/17 between 7:30 AM and 3:06 PM. Looks like this was a special event.

<https://ktla.com/2017/03/26/ciclavia-bike-festival-from-culver-city-to-venice-underway/>