Final Project: Ford GoBike

Brianna Kincaid

May 7, 2018

Project Summary

Ford GoBike is a public bicycle sharing system based in the San Francisco Bay Area in California that began operation in August 2013. It is the first regional and large-scale bicycle sharing system deployed in California and on the West Coast of the United States. There are currently 2,500 bicycles across 260 stations that are available 24 hours a day, 7 days a week, to be rented and ridden. It is expected that the system will expand to around 7,000 bicycles across 540 stations in the Bay Area.

The Data

Ford GoBike makes available both historical data (since June 2017) as well as real-time data.

Trip History Data

There is data given for each trip that has been taken since June 2017. Each trip is anaonymized and the data given includes:

- Trip Duration (seconds)
- Start Time and Date
- End Time and Date
- Start Station ID
- Start Station Name
- Start Station Latitude
- Start Station Longitude
- End Station ID
- End Station Name
- End Station Latitude
- End Station Longitude
- Bike ID
- User Type (Subscriber or Customer)
- Member Year of Birth
- Member Gender

The following files were downloaded from www.fordgobike.com/system-data:

2017:

- `2017-fordgobike-tripdata.csv`

2018

- `201801-fordgobike-tripdata.csv`
- `201802-fordgobike-tripdata.csv`
- `201803-fordgobike-tripdata.csv`

Real-Time Data

Ford GoBike publishes real-time system data in General Bikeshare Feed Specification format. At any time, the following data about *each station* is given:

- Station ID
- Number of Bikes Available
- Number of Bikes Disabled
- Number of Docks Available
- Number of Docks Disabled
- Is the station on the street
- Is the station renting
- Is the station accepting returns
- Last Reported (POSIX timestamp)

The station status was saved from https://gbfs.fordgobike.com/gbfs/en/station_status.json at various points in time. The data was saved to separate files for each time point.

Exploratory Analysis and Descriptive Statistics

I begin my analysis with an exploratory examination of both the historical data and the real-time data. I will look at the distribution of the variables as well as averages of variables across other variables. I will use this basic analysis to help me determine what to explore further.

Trip History Data

First I look at the historical trip data. Firth, I imported the trip data for 2017 using read_csv. The data itself is very tidy and well organized with a limited number of missing (NA) values. I had to replace the NA values for gender and birth year with "Not Specified". I also separated the time column into separate columns for Year, Month, Day, Hour, Minute, and Second in order to make grouping easier later.

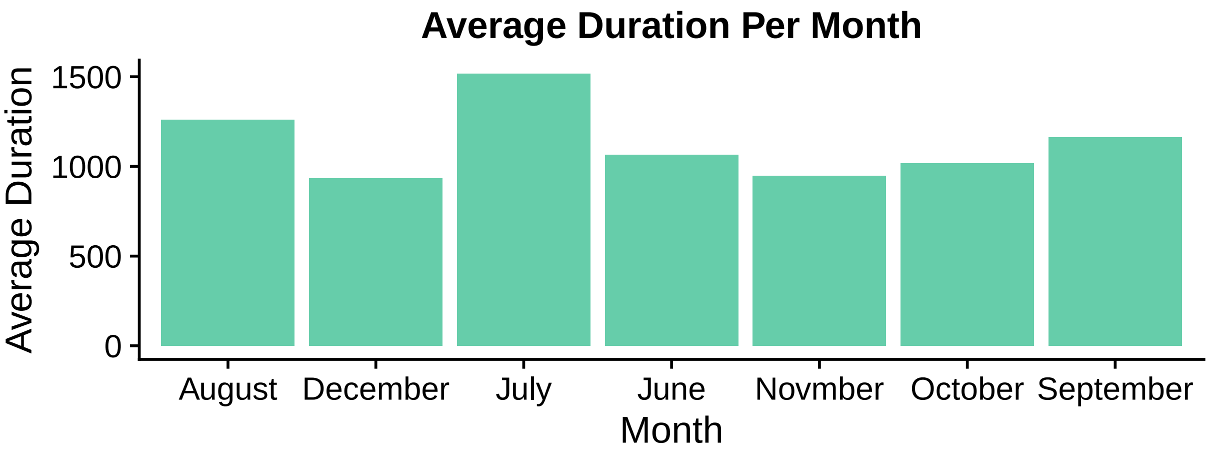
Trip Duration

The data for each trip includes the duration of the trip in **seconds**. I will examine here how the duration changes across other variables, such as month, origin station, and destination station.

First I begin by looking at the average duration for each month of 2017, or at least for each of the months given in the data.

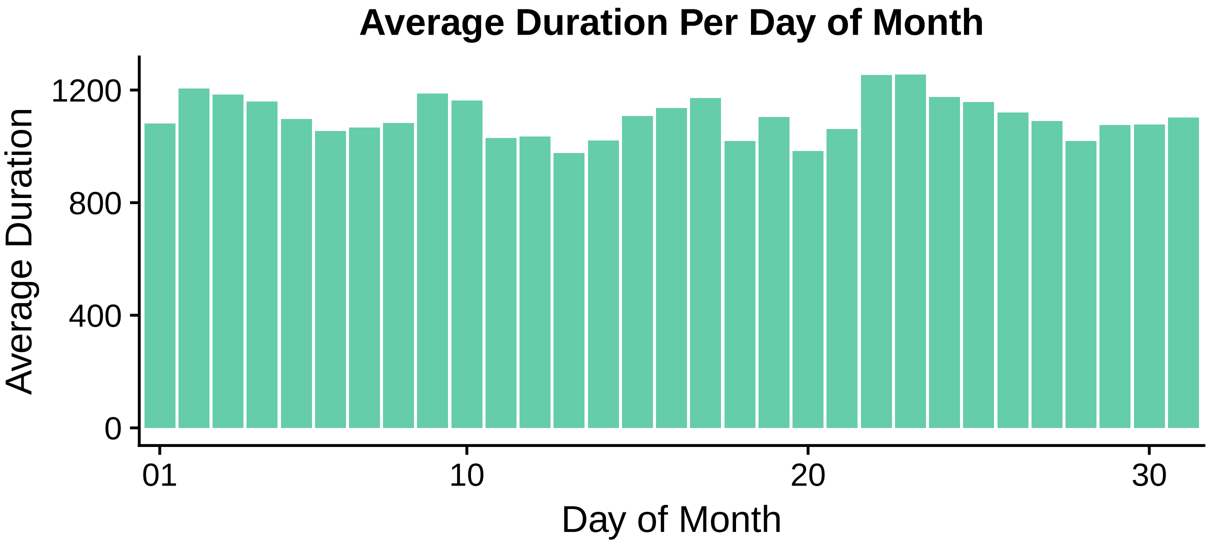
Table 1: Average Duration Per Month

Month	Average Duration
August	1260.4057
December	932.4337
July	1516.7567
June	1065.0852
Novmber	947.7423
October	1015.9937
September	1161.4966



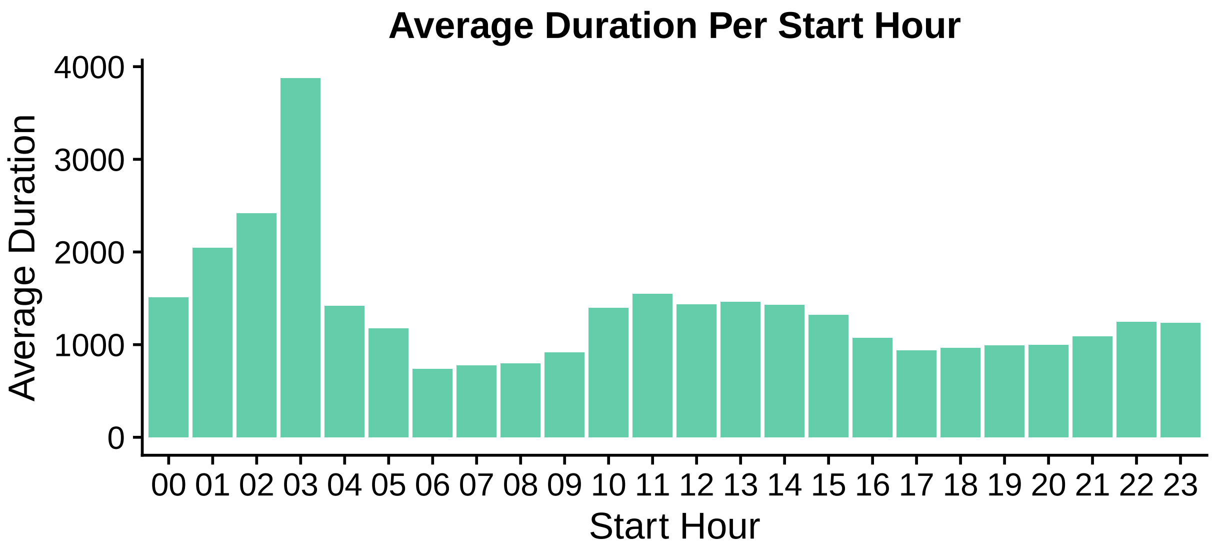
It can clearly be seen that July has the longest average duration. This makes sense, as July (and August) are typically the warmest months of the year, so it would be expected that the duration of the trips would be longer on average during these months. However, the difference between the average duration in July and the rest of the months is not significant, and I will not be examining this further.

Next, I look to see if the average duration changes during the time of the month. I was mostly looking here to see if trips tended to be longer at the start or end of the month.



It is clear that there really is no relation between the day of the month and the average duration of the trips. There is some minor cyclicaly variation in the average duration that may have something to do with the day of the week, but it is noth worth examining further.

Next I look at how the hour of the day affects the average duration. This was mostly to see if longer trips, on average, started earlier in the day. I expected that the trend across the Start Hour would be significant enough for further examination.



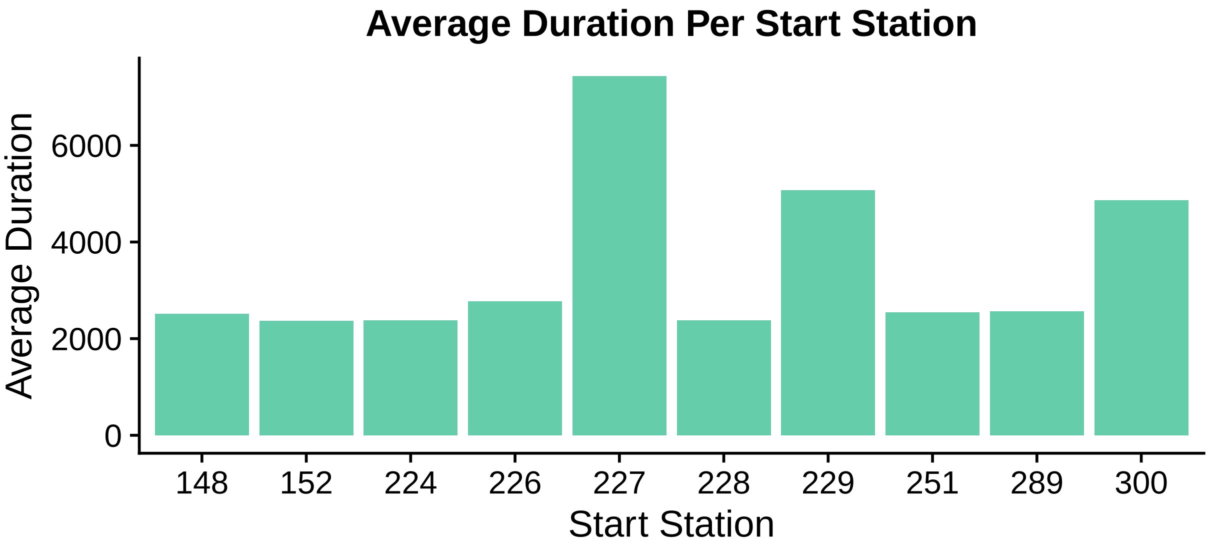
It is clear that trips that start earlier in the day have longer trip durations on average. This makes sense, because if you have a longer trip planned, you would leave earlier.

Next, I look at stations. This part of the exploratory process is important because station status and the most popular stations are a major point of interst of this project. Here, I just look at the general trends to set the basis for further analysis.

Here is a table with the average duration for the ten (start) stations with the highest average trip duration. It is unclear what any trends here actually mean, but later on I will examine fruther variables that may go into why these stations may be more popular or consistently are on one end of long trips.

Table 2: Top 10 Stations By Duration

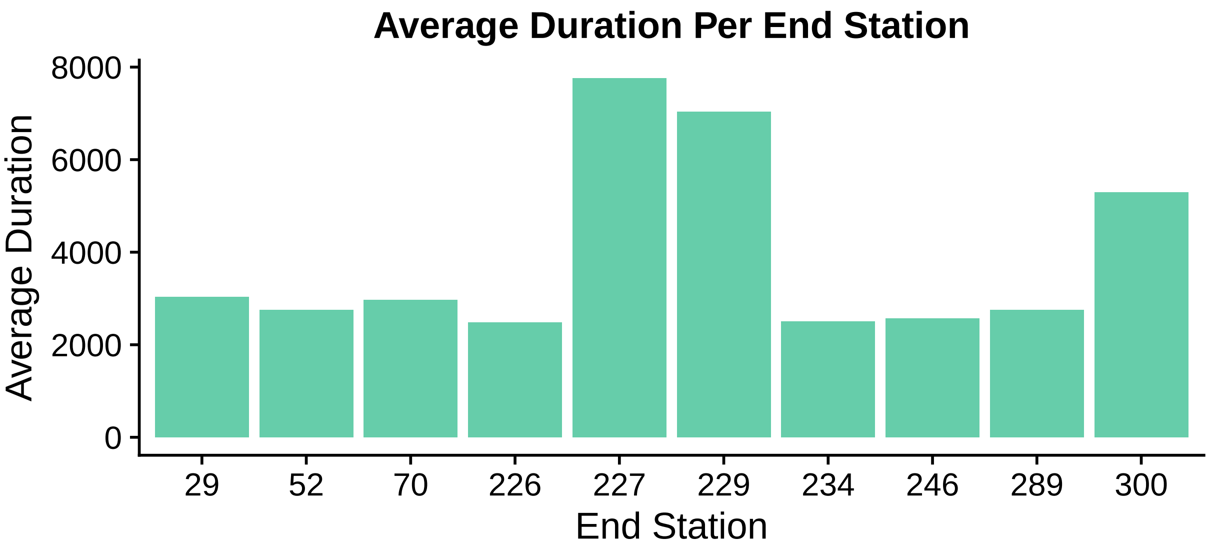
Start Station ID	Average Duration
227	7434.557
229	5070.179
300	4858.621
226	2774.293
289	2565.545
251	2545.694
148	2513.371
224	2380.778
228	2374.929
152	2369.016



Here is a table with the average duration for the ten (end) stations with the highest average trip duration. Again, it is unclear what any trends here actually mean, but later on I will examine fruther variables that may go into why these stations may be more popular or consistently are on the end of longer trips.

Table 3: Top 10 Stations By Duration

End Station ID	Average Duration
227	7763.313
229	7034.278
300	5291.432
29	3028.718
70	2970.270
289	2754.786
52	2752.389
246	2564.182
234	2503.747
226	2481.689

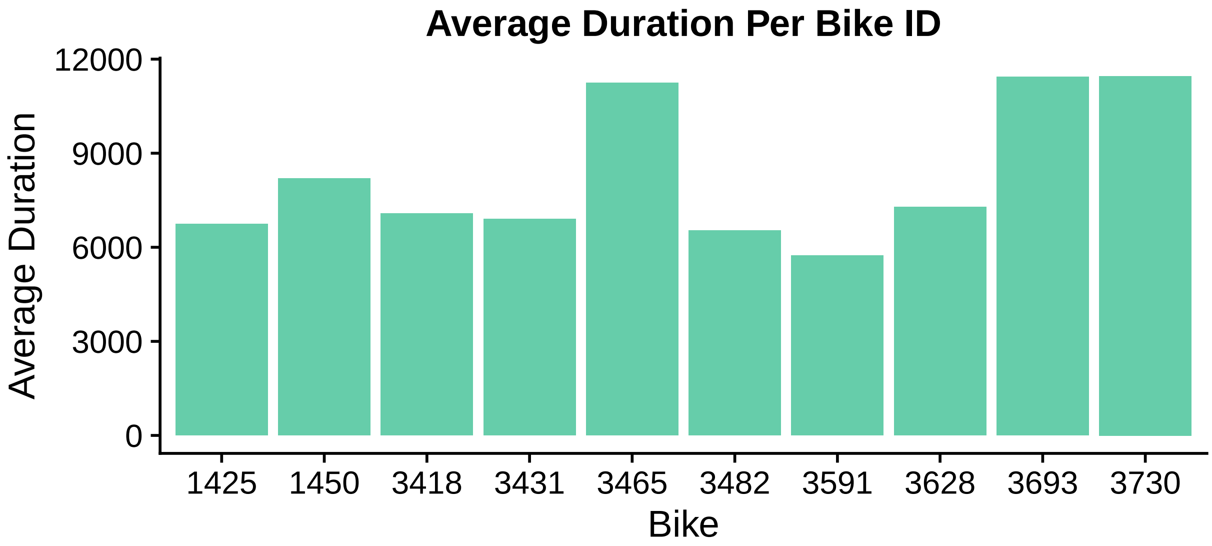


Stations 227 and 229 are clearly involved with longer trips. This may have something to do with their location or their proximity to other stations.

Next I'll look at the average trip duratios for each specific bike. I don't expect anything interesting to appear here, as the bikes all are the same and customers don't typically have much choice when it comes to the bike they chose.

Table 4: Top 10 Bikes By Duration

Bike ID	Average Duration
3730	11461.125
3693	11434.714
3465	11248.250
1450	8198.818
3628	7288.800
3418	7072.923
3431	6906.083
1425	6739.500
3482	6538.214
3591	5742.225



As I expected, there really isn't anything intersting here.

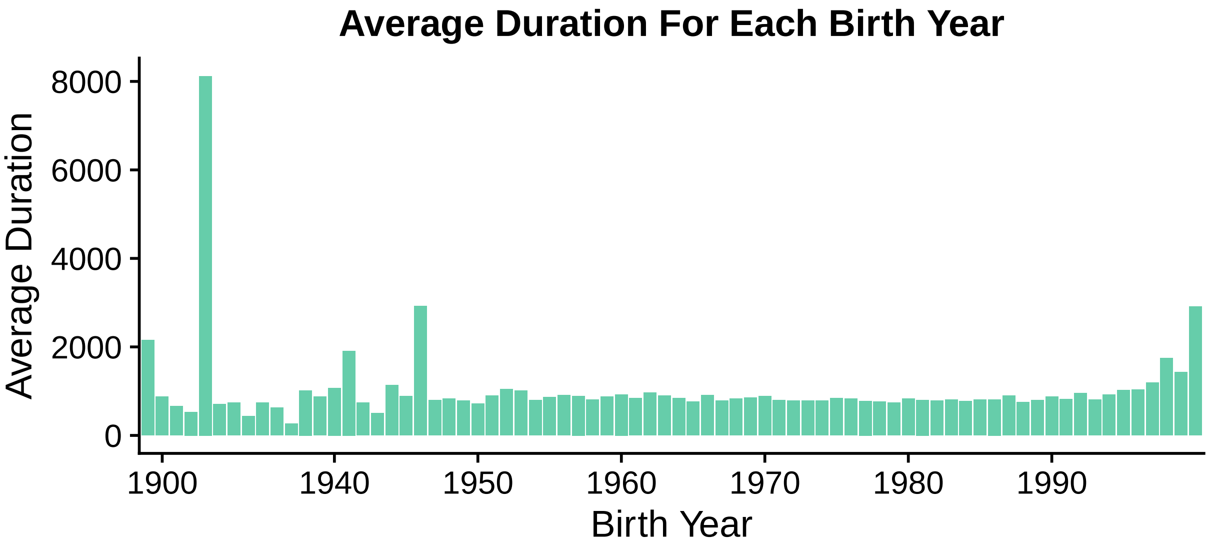
Next, I look at how the user type. There are two different types of users that use Ford GoBike: those who pay for a signle ride or day pass, and those who have a monthly membership. A single ride is $2 for 30 minutes, while a $15 monthly membership gives unlimited 45-minute trips. These users are broken up into two categories: "Customer" and "Member". Here is the average duration for the two user types:

Table 5: Average Duration By User Type

User Type	Average Duration
Customer	2557.4458
Subscriber	705.3105

Customers clearly have longer trips (with respect to time). This may be because customers (or, those who are not a member or subscribed), may chose to try to get more out of each trip by using it longer than subscribers, who pay a flat rate for unlimited 45-minute trips.

Next, I look at birth year. This data is difficult and hard to interpret ecause some inconsistent and out of place values occur. For example, there ar trips where the birth year of the member is listed as 1886, but this is impossible. Furthermore, there is an abnormally large jump in the average duration for those with a birth year around 1910. Therore, this data is not really reliable and I wil not draw any conclusions.



Finally, let's look at the average trip duration for each gender. This data is hardly conclusive, because there are more unspecified genders than there are of any other category. However, of those who did specify a gender, the data showed that the average trip duration was longer for both "Female" and "Other" over "Male". Furthermore, those who didn't specify gender had a significantly longer trip duration than any other category. It is hard to draw conclusions based on this data because not every customer or member chose to disclose their gender and therefore gender data is not available for every trip.

Table 6: Average Duration For Each Gender

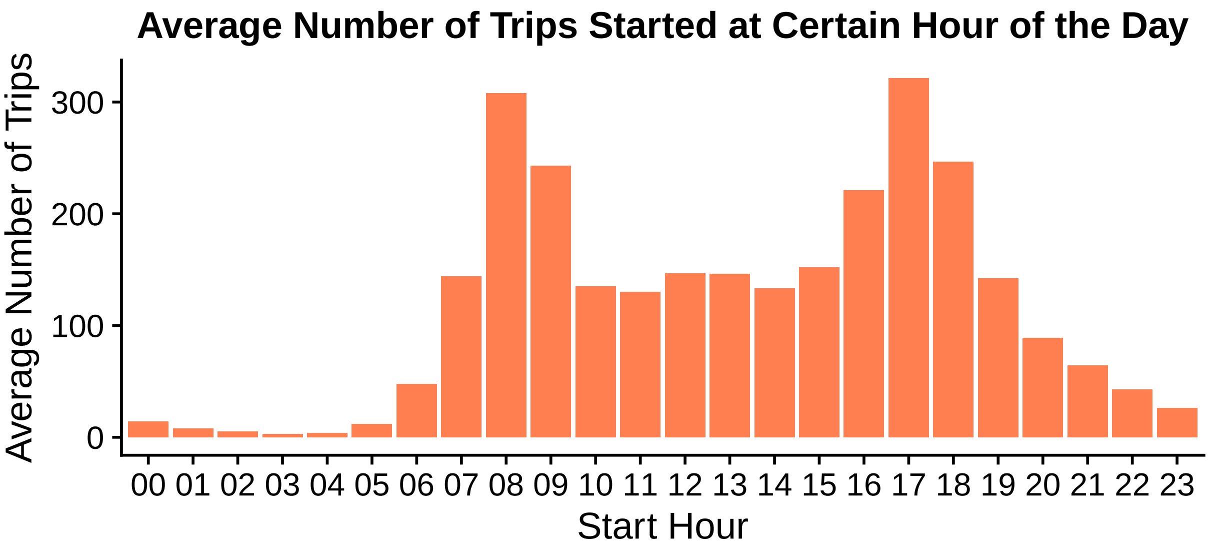
Gender	Average Duration
Female	1027.8911
Male	774.4409
Not Specified	2913.1171
Other	1019.2588

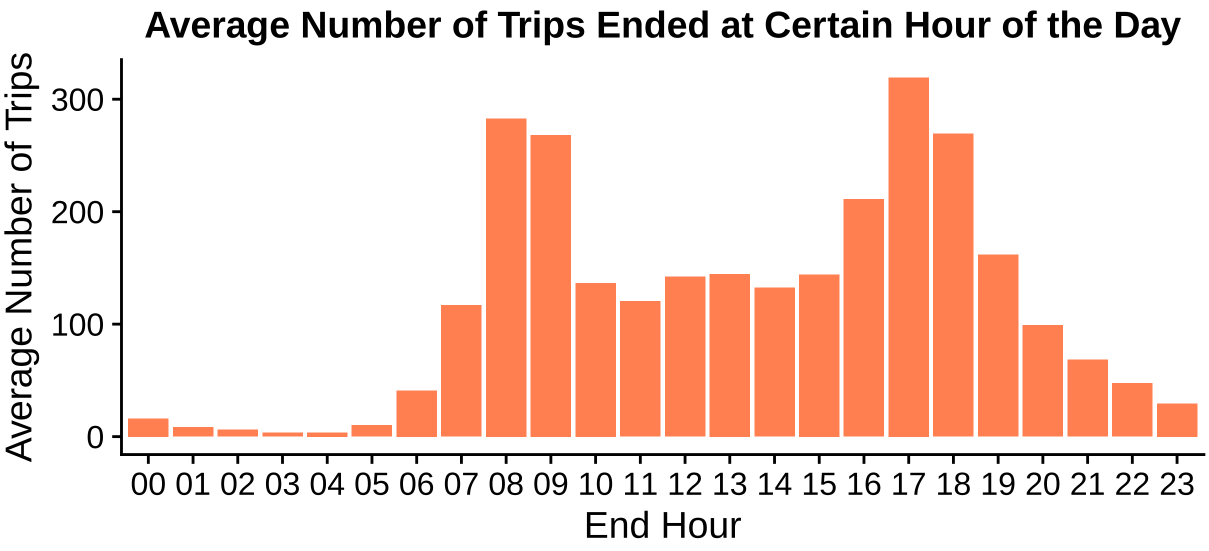
This concludes the exploration of trip duration as given in the historical data for 2017. It is clear that hour (both start and end), station (both start and end), and user type all affect the average trip duration.

Number of Trips

Because we have data for every trip, it is easy to group together the trips by certain variales and count the number of trips in each group. In doing so, we can look at how other variables influence the number of trips.

First, let's look at month, day, and hour. As was clear above, there is no significant difference in the variable between months and days of the month, and that is true here as well. Therefore, I will leave out the data for the number of trips per month and the number of trips per day of the month. What is interesting however is the distribution across the time of day. Below are the plots for the average number of trips that started or ended at a certain hour of the day. Both graphs look very similar, with peaks around 8 o'clock in the morning and 5 o'clock in the evening.



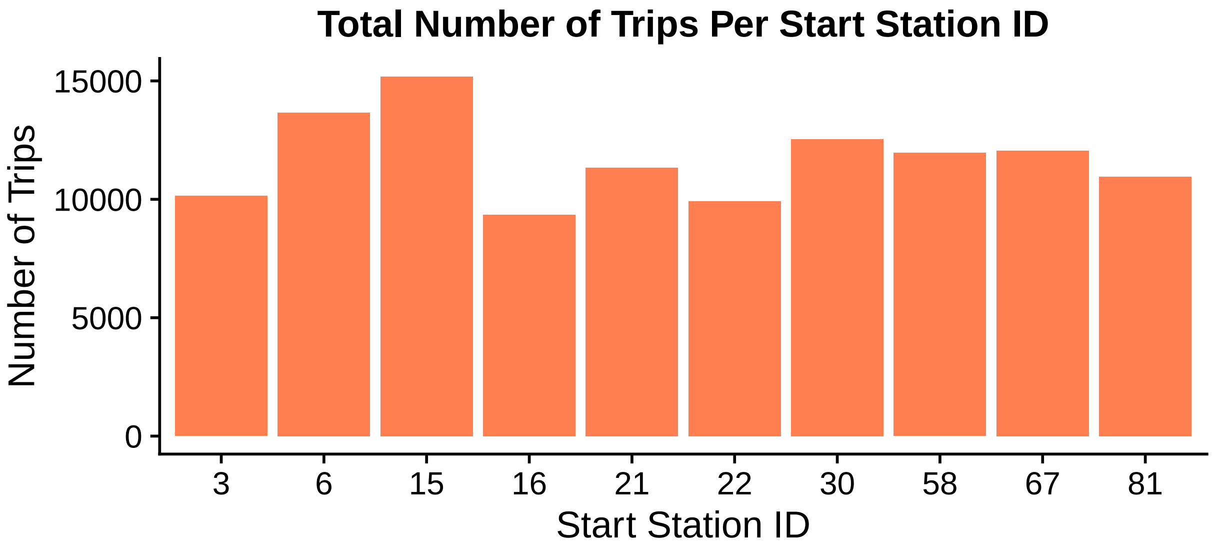


Now let's look at how the number of trips varies across the stations. Below is a table of the ten stations that were the origin for the most number of trips. We also plot this data in a histogram, although the histogram is not very revealing.

Table 7: Total Number of Trips Per Start Station ID

Station ID	Lon	Lat	Trips
15	-122.3942	37.79539	15187
6	-122.4032	37.80477	13664
30	-122.3953	37.77660	12546
67	-122.3955	37.77664	12055
58	-122.4174	37.77662	11960
21	-122.4008	37.78963	11334
81	-122.3932	37.77588	10956
3	-122.4049	37.78638	10142
22	-122.3946	37.78976	9926

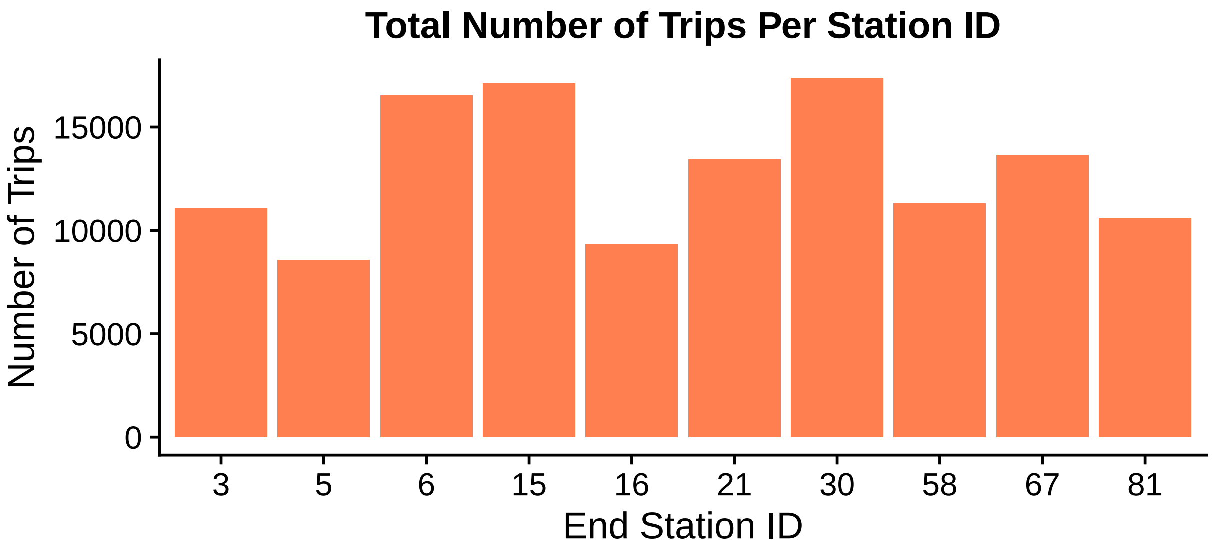
Station ID	Lon	Lat	Trips
16	-122.3944	37.79413	9347



Below is a table of the ten stations that were the destination for the most number of trips. It is interesting that the top three are the same, although not in the same order, as the top three for origin stations. There is obviously something about these stations that make them popular origins and destinations. I will explore this further later in the project.

Table 8: Total Number of Trips Per Station ID

Station ID	Lon	Lat	Trips
30	-122.3953	37.77660	17378
15	-122.3942	37.79539	17109
6	-122.4032	37.80477	16531
67	-122.3955	37.77664	13658
21	-122.4008	37.78963	13443
58	-122.4174	37.77662	11298
3	-122.4049	37.78638	11064
81	-122.3932	37.77588	10611
16	-122.3944	37.79413	9321
5	-122.4084	37.78390	8563



Now let's look at how the number of trips varies between the two types of users discussed above. Below is a table with the total number of trips for each customer type. Clearly most of the users are subscribers rather than (non-member) customers.

Table 9: Total Number of Trips Per User Type

User Type	Trips
Customer	110470
Subscriber	409230

Next, I could look at how the number of trips varies across birth year, but I will pass on this as I have already concluded that the birth year data is inconclusive, inconsistent, and not worth anything.

Finally, I look at which genders have the most umber of trips. As I discussed above, the data is not entirely revealing because there are a large number of unspecified genders for the trip data. However, as seen in the below table, men take significantly more trips than women, and this would still be true even if all of the unspecified genders were female.

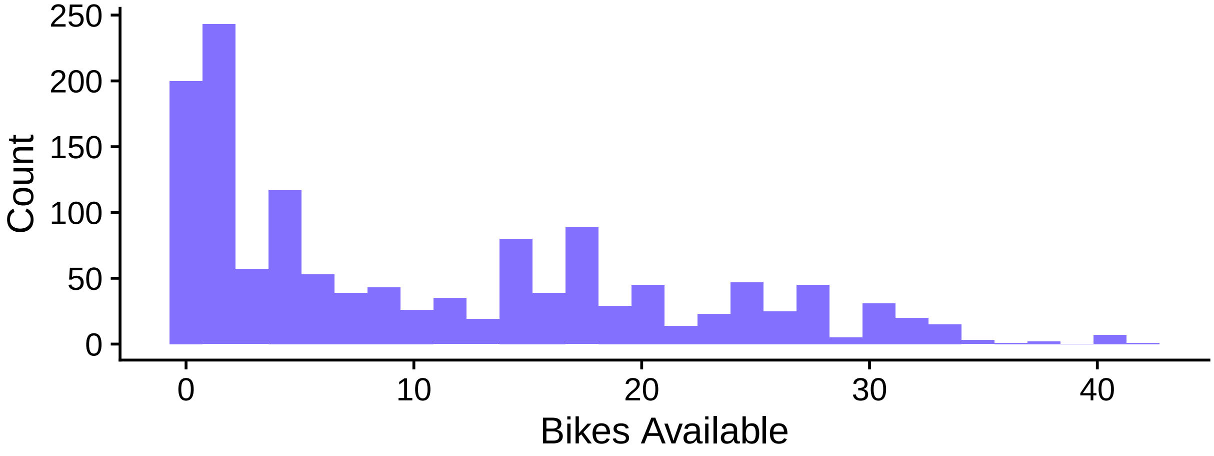
Table 10: Total Number of Trips Per Gender

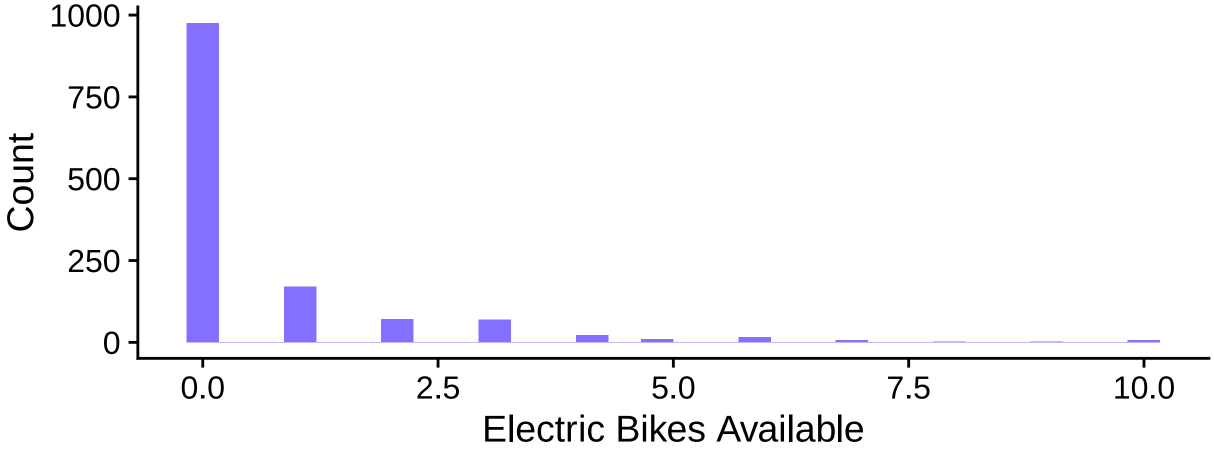
Gender	Trips
Female	98621
Male	348318
Not Specified	66462
Other	6299

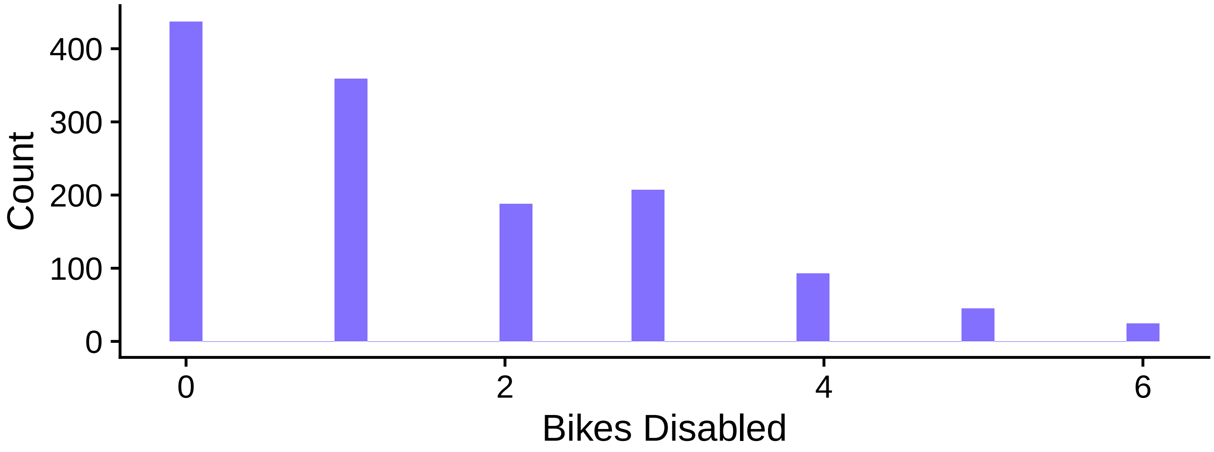
Real-Time Data

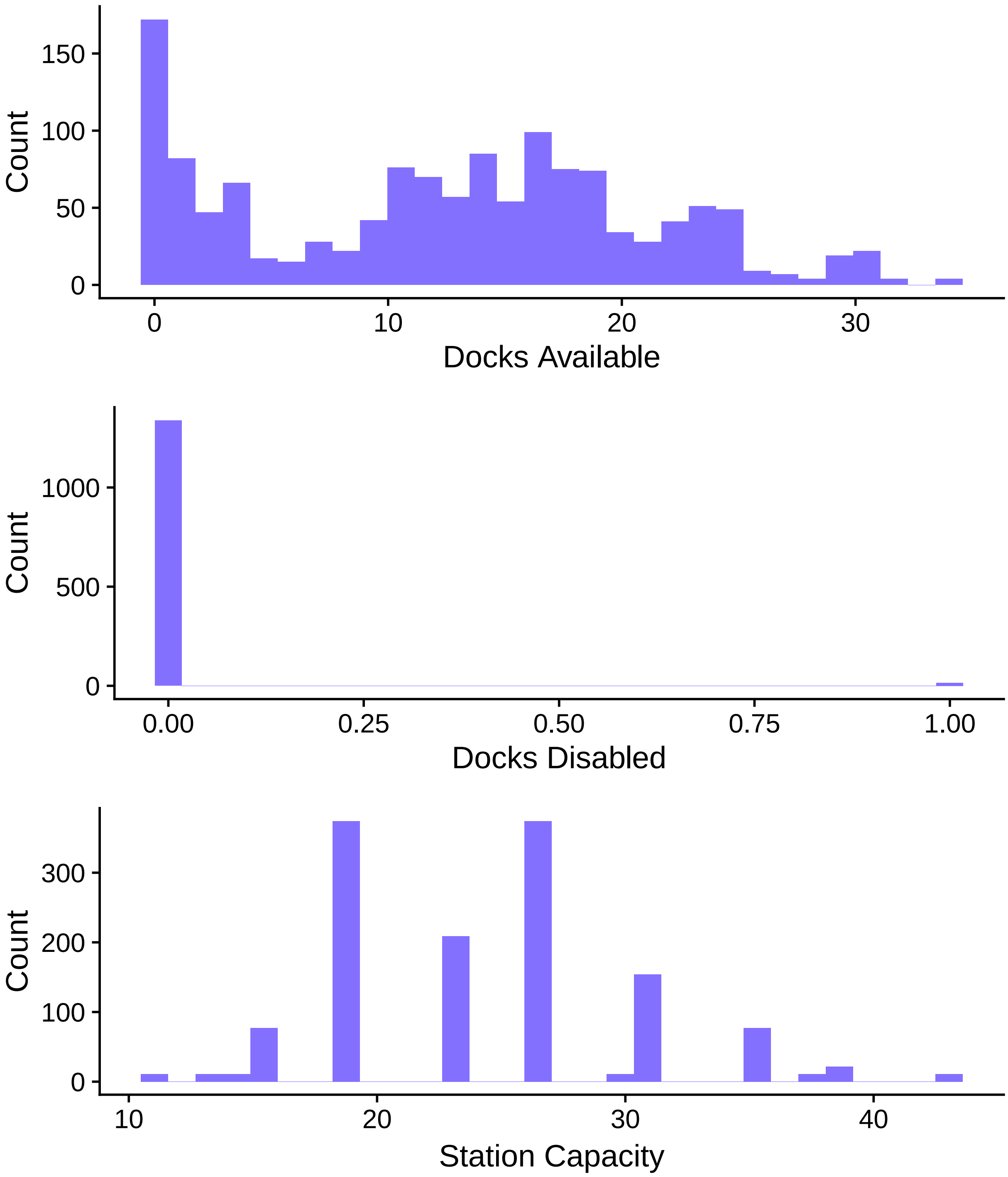
The real time data that Ford GoBike provides is different from the historical data. The historical data provides data for each trip while the real-time data provides real-time status updates on each of the stations. The data was saved and put together in real_time.R, where the data was saved to real_time_data.csv. We input that file here. The data can be updated with new real-time data and the same analyis can be applied.

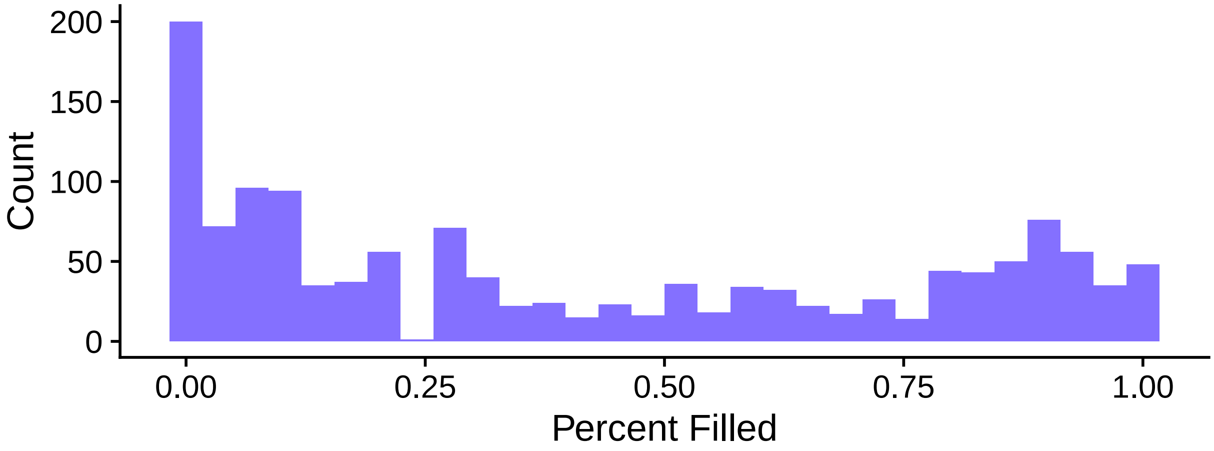
In this section, I will just do some basic exploratory analysis before expanding on it later. Let's first look at the histograms of all the variables.This will give us a general idea of how each variable is distributed.








Count
150
100
50
0
0
10
20
30
Docks Available
Count
1000
500
0
0.00
0.25
0.50
0.75
1.00
Docks Disabled
Count
300
200
100
0
10
20
30
40
Station Capacity



The only variable that seems to be normally distributed is the Station Capacity. This should not affect my future analysis too much.

Station Analysis

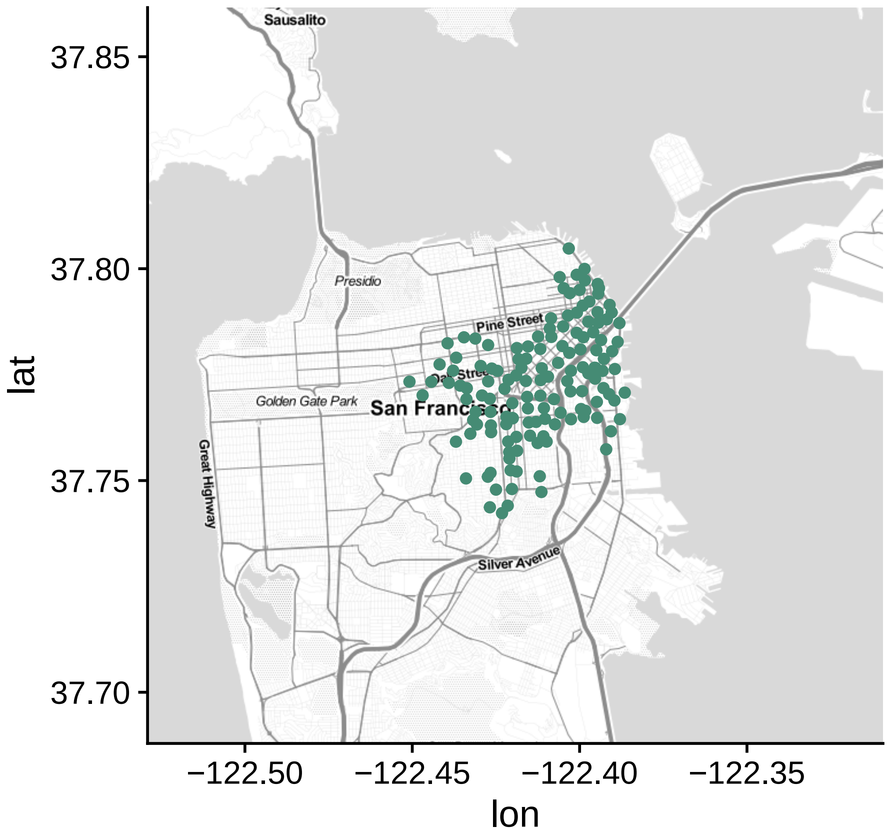
From my exploratory analysis, it seems that the most interesting information can be found in looking at variation across stattions. General station information can be found on the Ford GoBike site, downloaded from station_information.json and loaded into R. The tibble stationinfo contains the following columns:

- Station ID
- Name
- Short Name
- Latitude
- Longitude
- Region ID
- Capacity

Table 11: First 10 Rows of stationinfo

Station ID	Short Name	Lat	Lon	Region ID	Capacity
74	SF-J21	37.77643	-122.4262	3	27
3	SF-G27	37.78638	-122.4049	3	35
4	SF-G26	37.78588	-122.4089	3	35
5	SF-H26	37.78390	-122.4084	3	35
6	SF-A27	37.80477	-122.4032	3	23
7	OK-L5	37.80456	-122.2717	12	35
8	SF-C28-1	37.79995	-122.3985	3	23
9	SF-C28-2	37.79857	-122.4009	3	19
10	SF-D27	37.79539	-122.4048	3	31
11	SF-D28	37.79728	-122.3984	3	35

For further analysis, I will limit the data to the Region ID = 3, which is the ID for **San Francisco**. There are 140 stations within San Francisco, and all of the most popular origin and destination stations are within San Francisco, so this filtering of the data is warranted. Below is a map of all of the stationsin Region 3 (San Francisco).

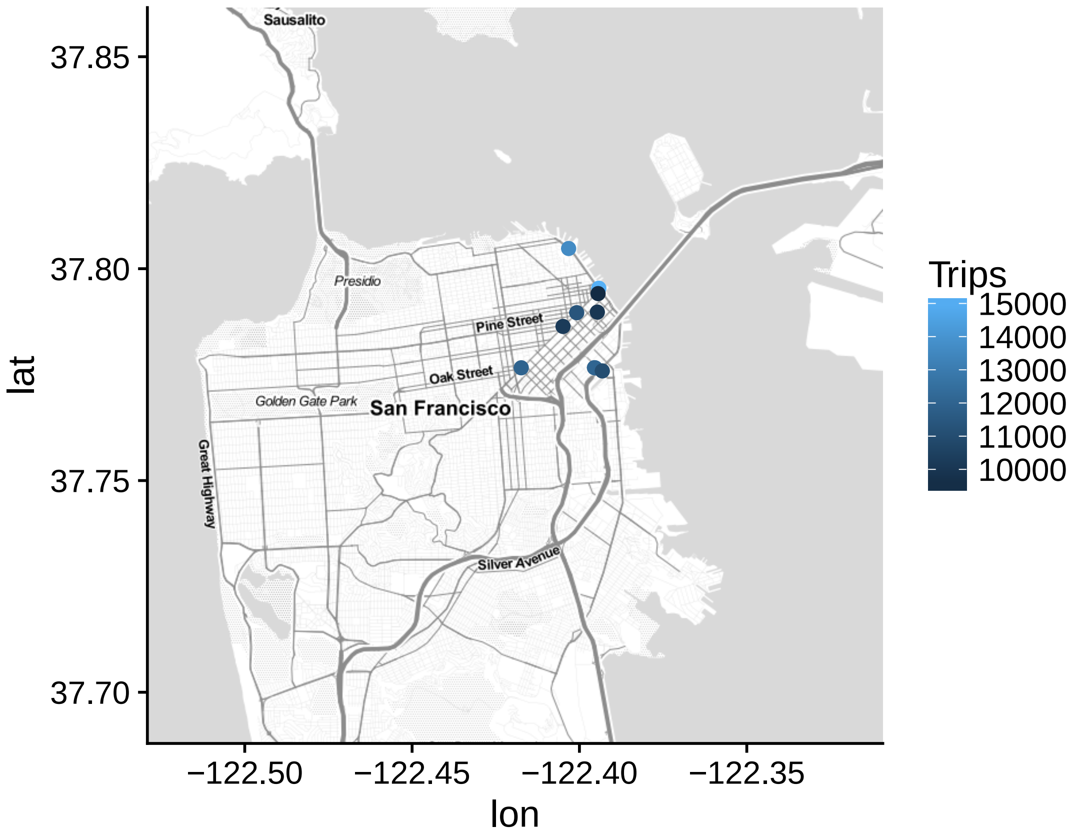


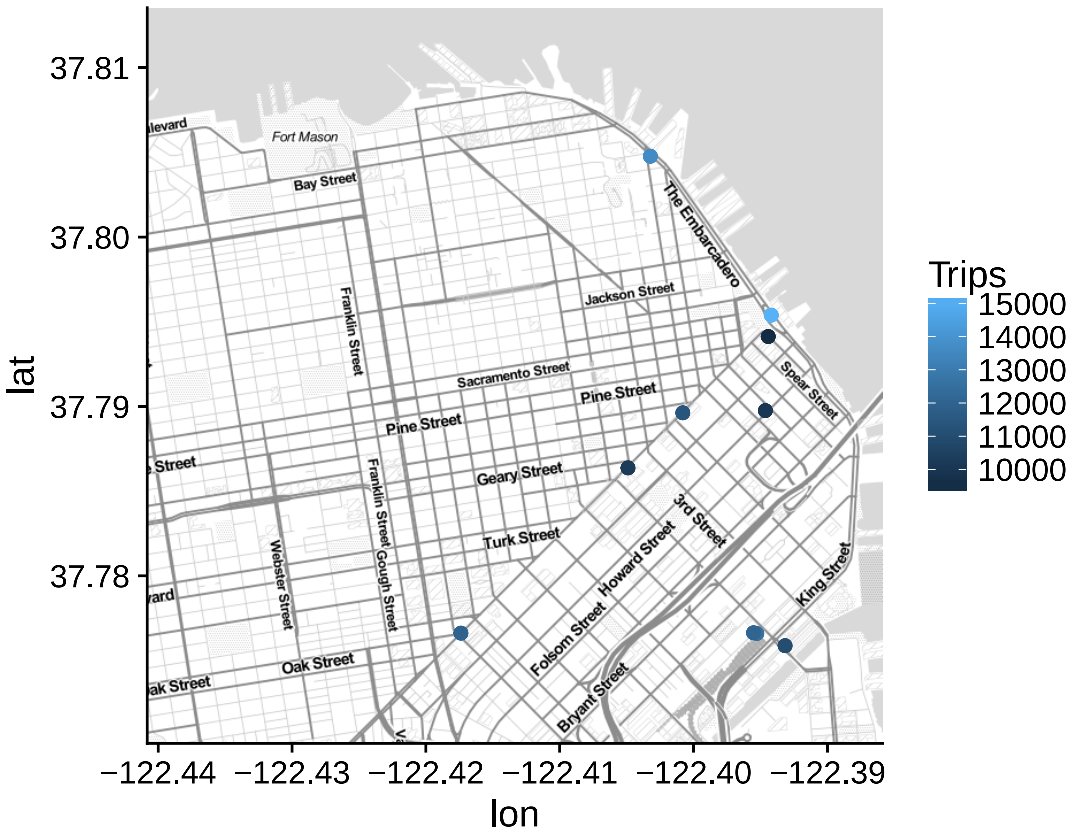
Let's look again at the ten stations that *produce* the most number of trips, meaning they are the starting station for the most trips. This data is given in the historical data that is available.

Table 12: Total Number of Trips Per Start Station ID

Station ID	Lon	Lat	Trips
15	-122.3942	37.79539	15187
6	-122.4032	37.80477	13664
30	-122.3953	37.77660	12546
67	-122.3955	37.77664	12055
58	-122.4174	37.77662	11960
21	-122.4008	37.78963	11334
81	-122.3932	37.77588	10956
3	-122.4049	37.78638	10142
22	-122.3946	37.78976	9926
16	-122.3944	37.79413	9347

Let's look at the top 10 stations on a map. It looks like the stations where the greatest number of trips originate are located along the water as well as along Market st, which can be seen if we zoom in.


Sausalito
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
37.85
37.80
37.75
37.70
lat
−122.50
−122.45
−122.40
−122.35
lon
Trips
15000
14000
13000
12000
11000
10000

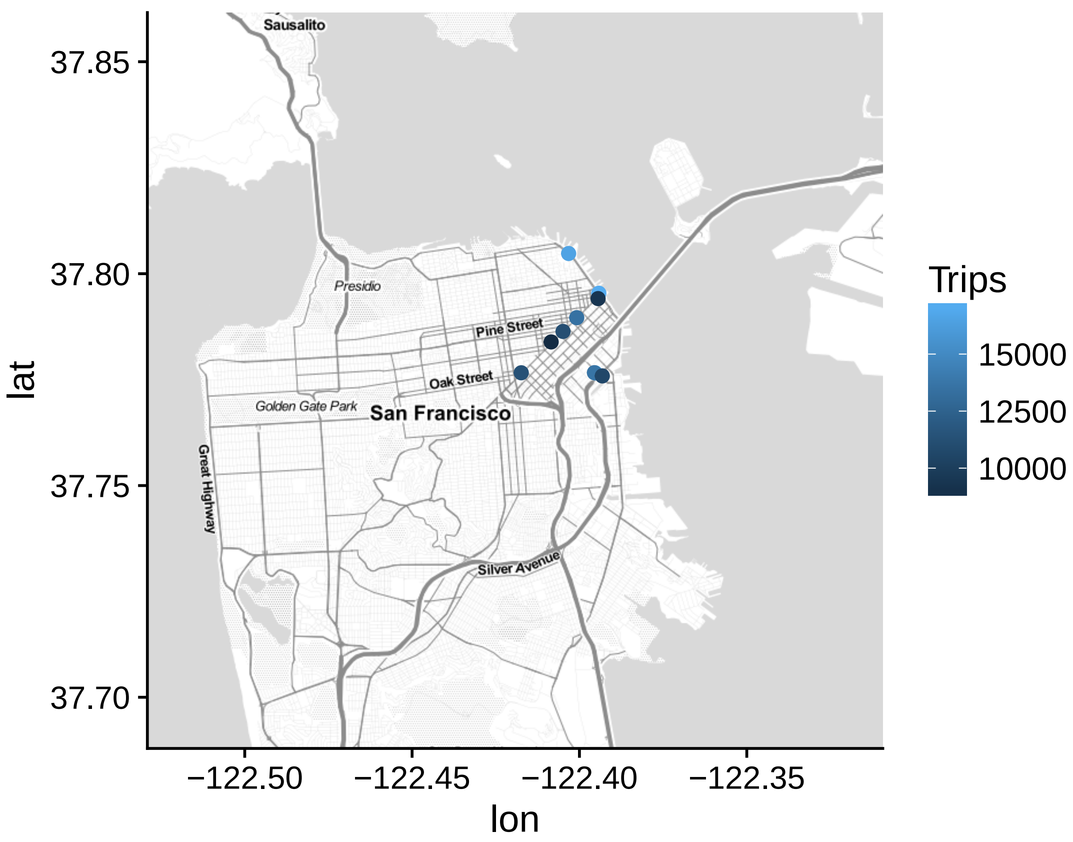


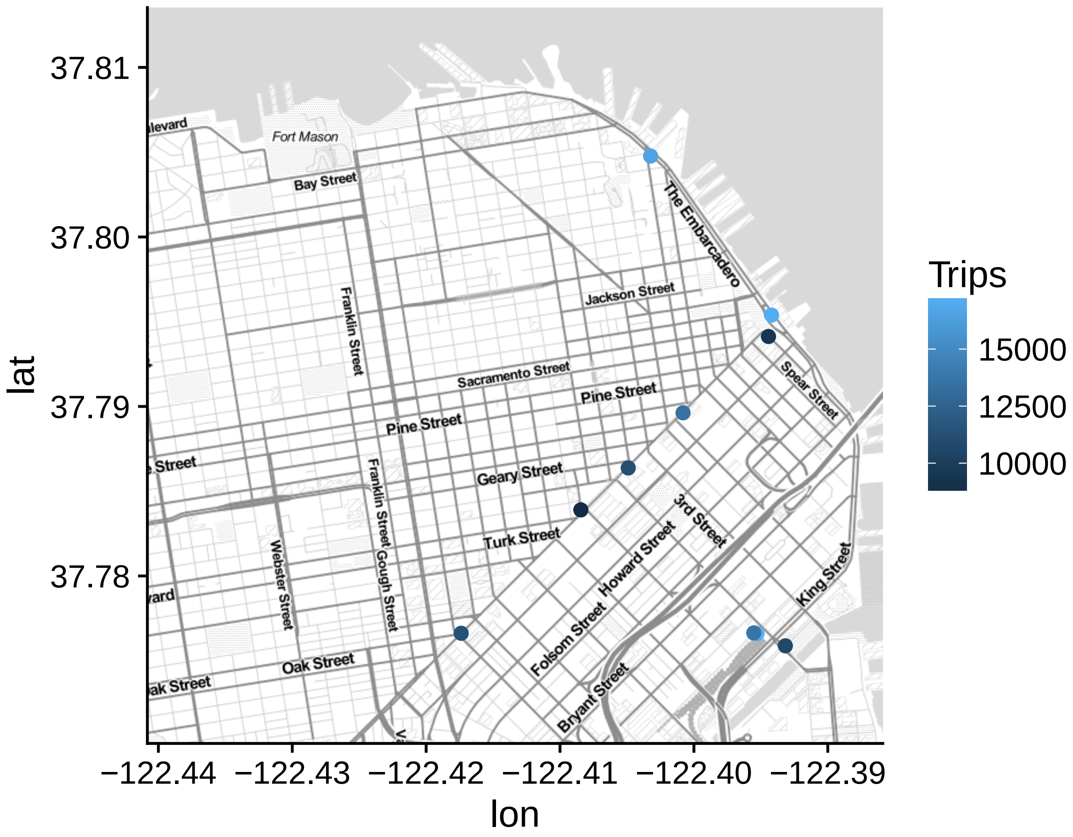
Now, let's look at the ten stations that *receive* the most number of trips (total from all the data from 2017), meaning they are the ending station for the most trips.

Table 13: Total Number of Trips Per Station ID

Station ID	Lon	Lat	Trips
30	-122.3953	37.77660	17378
15	-122.3942	37.79539	17109
6	-122.4032	37.80477	16531
67	-122.3955	37.77664	13658
21	-122.4008	37.78963	13443
58	-122.4174	37.77662	11298
3	-122.4049	37.78638	11064
81	-122.3932	37.77588	10611
16	-122.3944	37.79413	9321
5	-122.4084	37.78390	8563

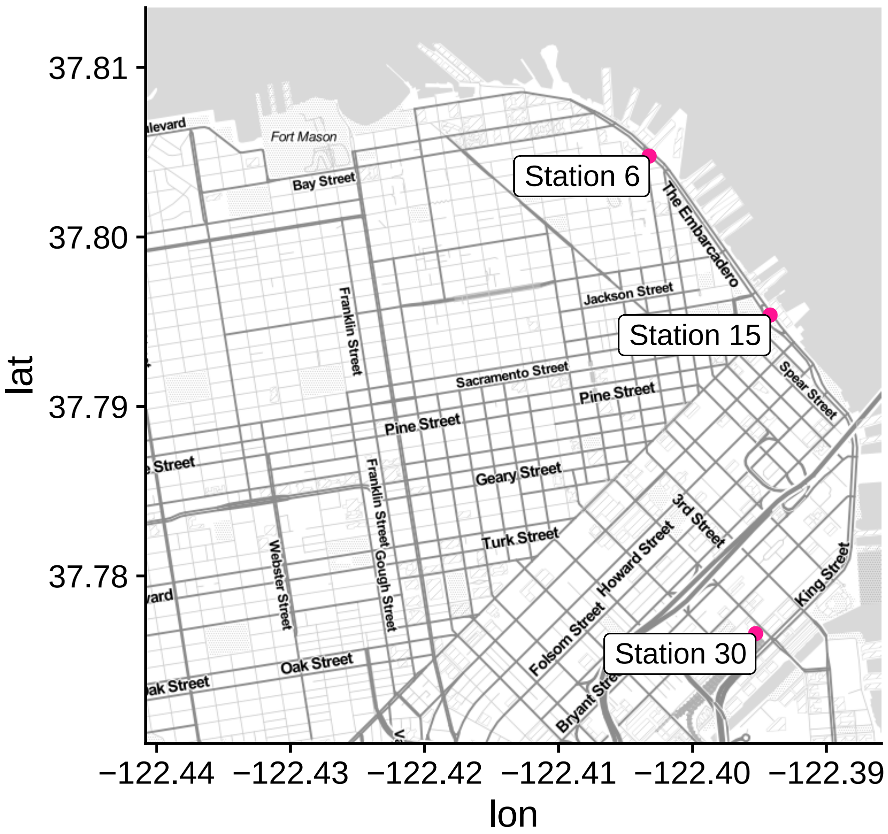
Let's look at the top ten stations on a map. It looks like the stations where the greatest number of trips originate are located along the water as well as along Market st, which can be seen if we zoom in.


Sausalito
Presidio
Pine Street
Oak Street
Golden Gate Park
San Francisco
Great Highway
Silver Avenue
37.85
37.80
37.75
37.70
lat
−122.50
−122.45
−122.40
−122.35
lon
Trips
15000
12500
10000



There are definitely similarities between the stations that produce the greatest umber of trips and the stations that receive the greatest number of trips. It seems interseting that the stations along that single street, Market St., appear in both “top 10” lists.

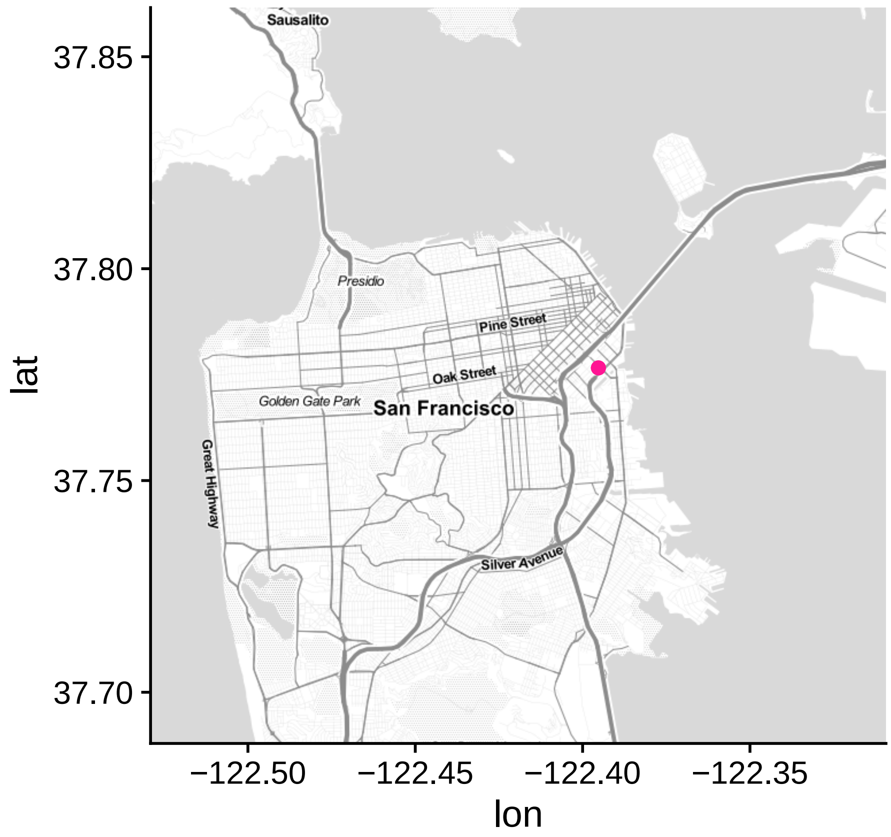
Before moving on to looking at the individual stations, let’s look at the location of all three stations on a map.



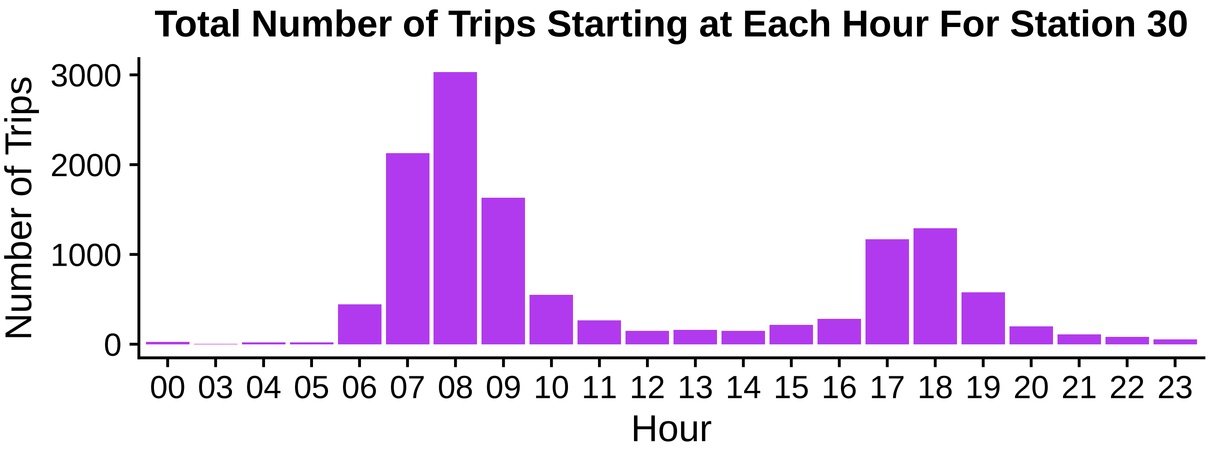
Station 30

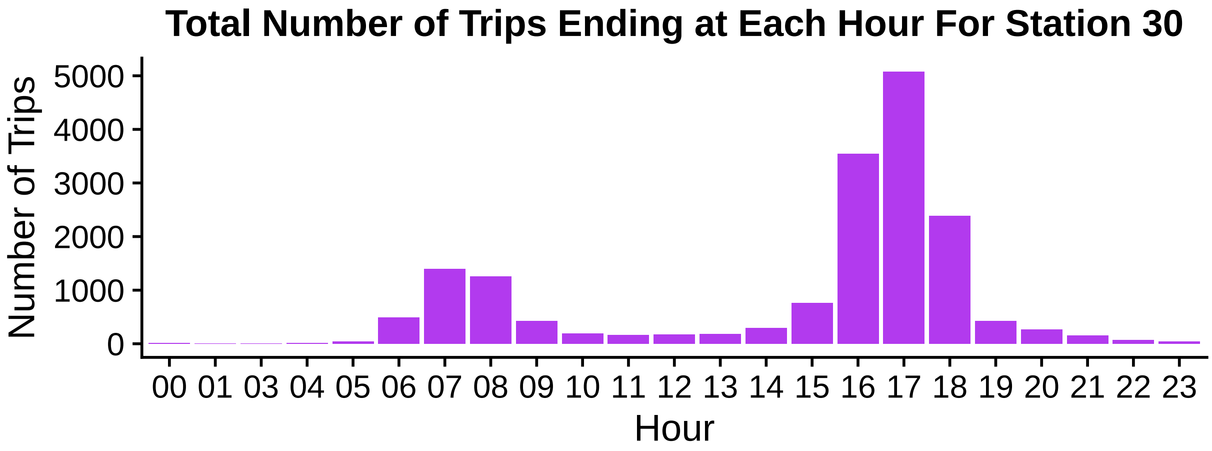
Station 30 is located on the corner of Townsend St and 4th St in San Francisco. This is right near the San Francisco Caltrain Station, which is the north end of the Caltrain commuter rail and also a major transit hub for the area. The proximity of this bike station to a major transit station could explain its popularity.

Here is the location of station 30 on a map:



Now I examine how the number of trips fluctuate throughout the day, on average. The following plots show the number of trips starting and ending at each hour.



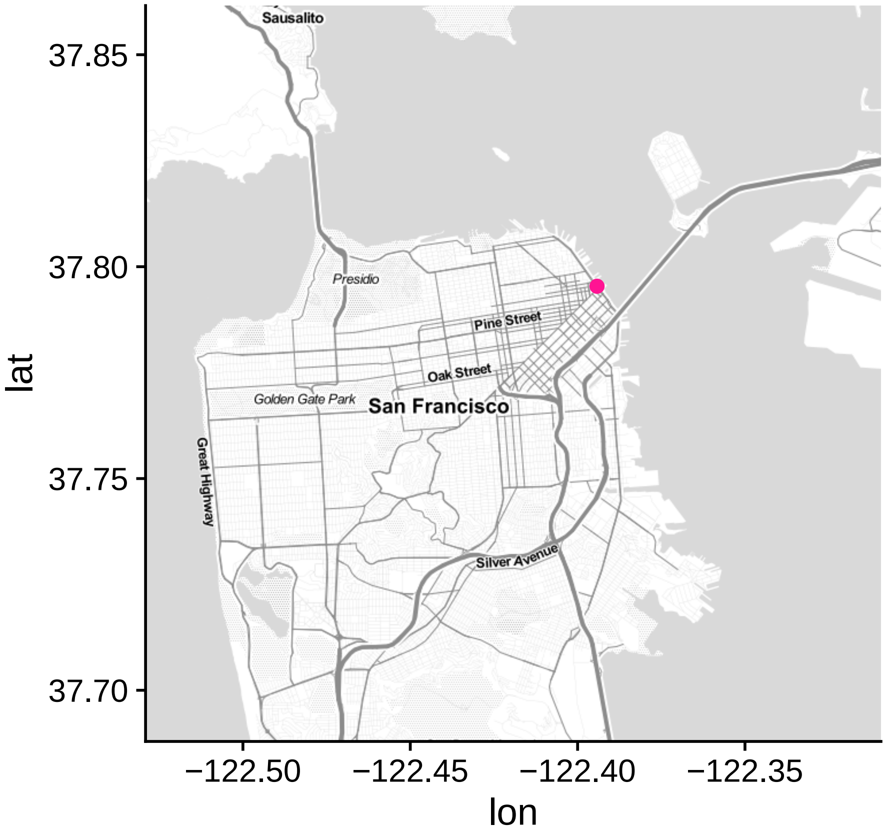


There is a clear trend, with two peaks for both the start and end hour. Furthermore, the greater peak of the two is opposite when looking at start hour versus end hour. It is clear that the most trips start around 8 o'clock in the morning and end around around 5 o'clock in the evening.

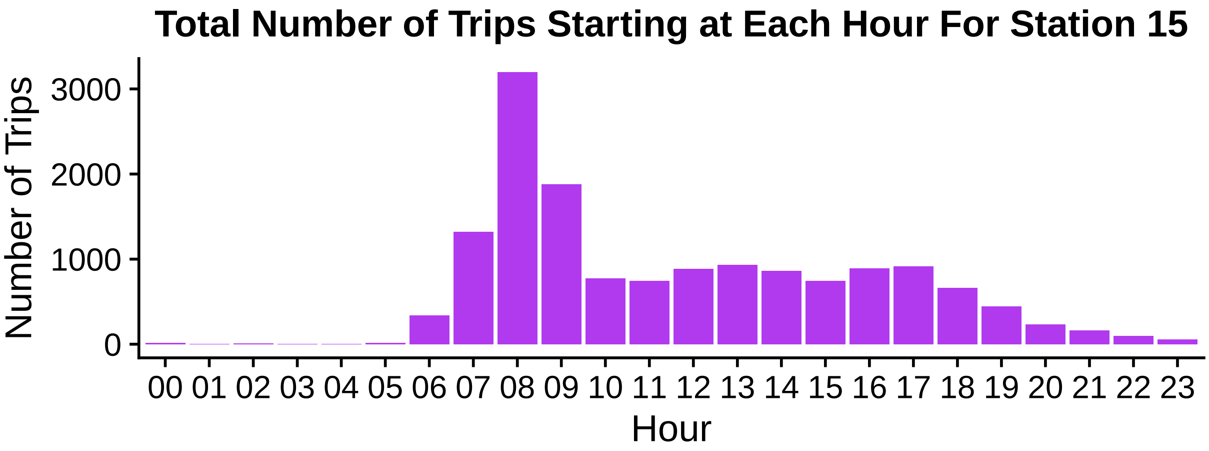
Station 15

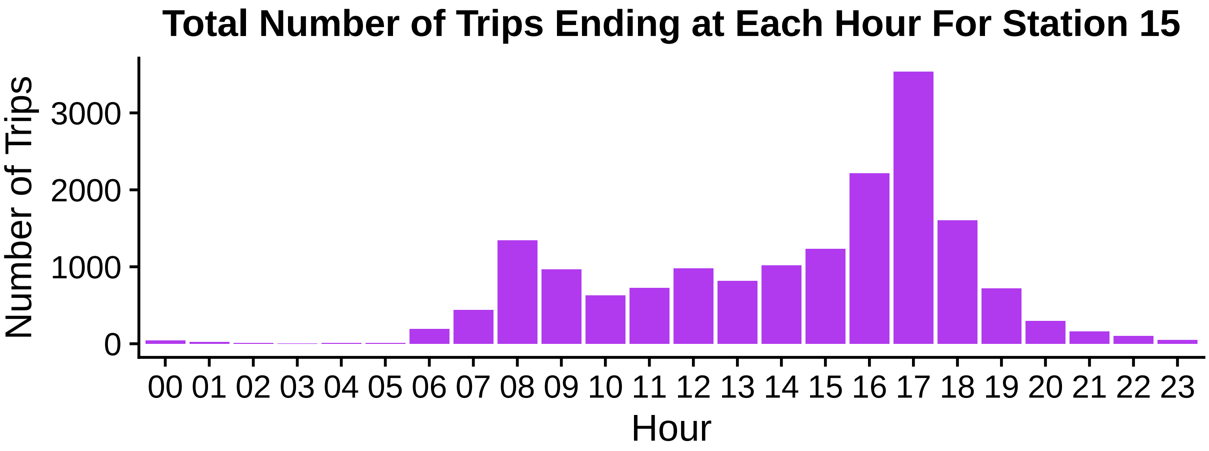
Station 15 is located at the Harry Bridges Plaza right next to the Ferry Building and the Port of San Fracisco. As with station 30, the proximity of sttion 15 to a major transportation hub, now a ferry instead of a train, explains its popularity.

Here is the location of station 15 on a map:



Now I examine the trips fluctuate throughout the day, on average. The following plots show the number of trips starting and ending at each hour.



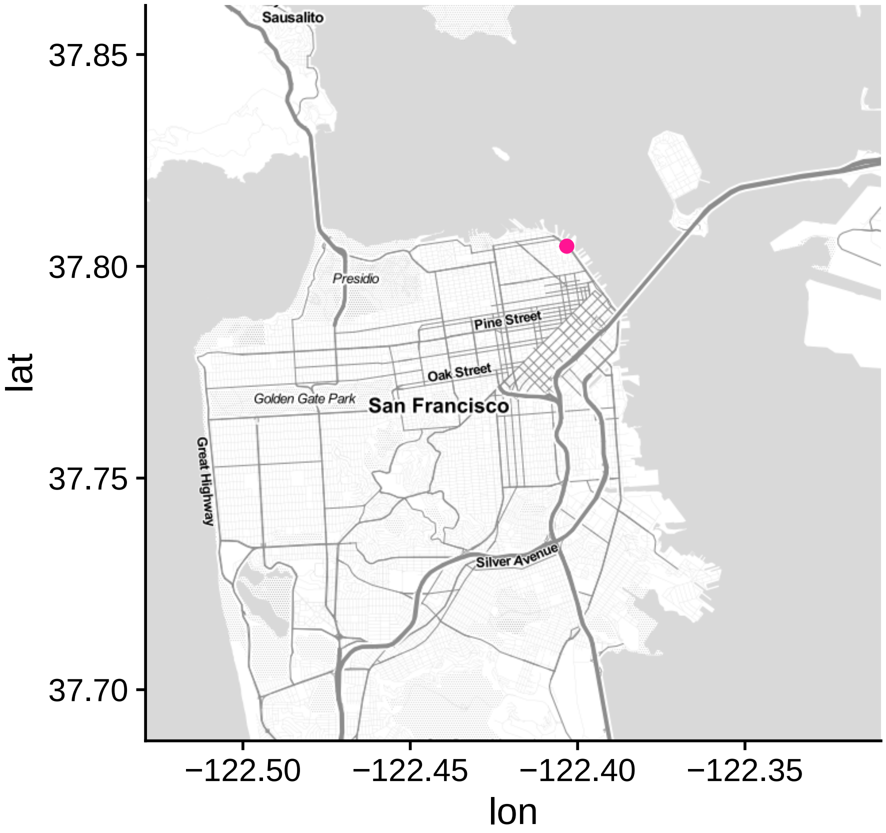


There is a clear peak for both the starting hour and ending hour, but unlike station 30, the second peak is less defined. However, the reverse symmetry among the two plots, as with station 30, is clear in the above plots.

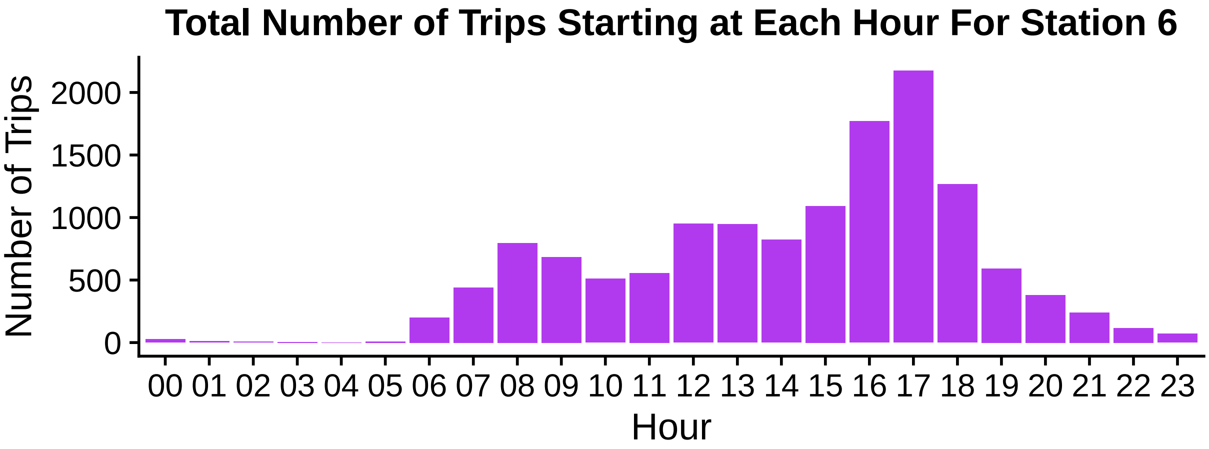
Station 6

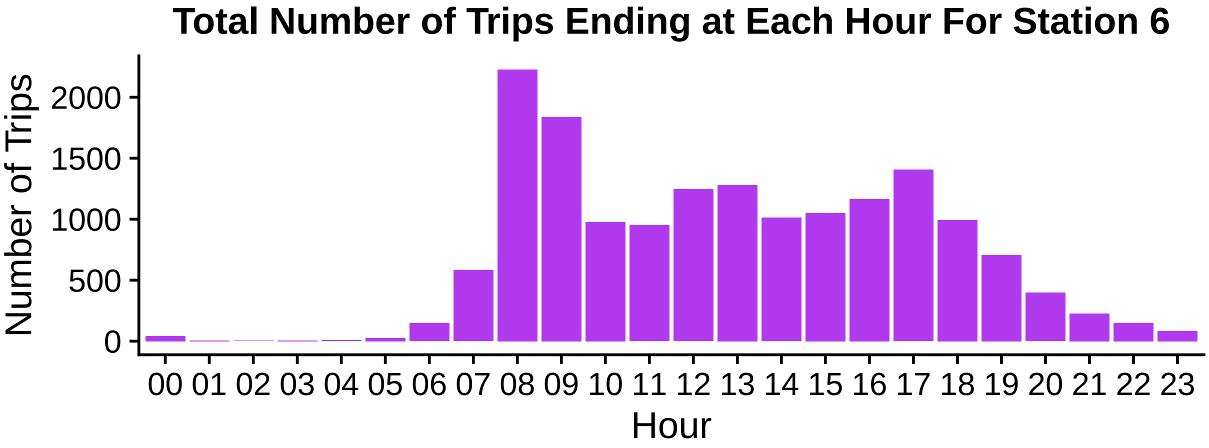
Station 6 is located on The Embarcadero at Sansome St, right across the street from the James R. Herman Cruise Terminal. This tourist destination is sure to bring in a large number of people, which explains the large number of trips that begin and end at station 6.

Here is the location of station 6 on a map:



Now I examine how the trips fluctuate throughout the day, on average. The following plots show the number of trips starting and ending at each hour.

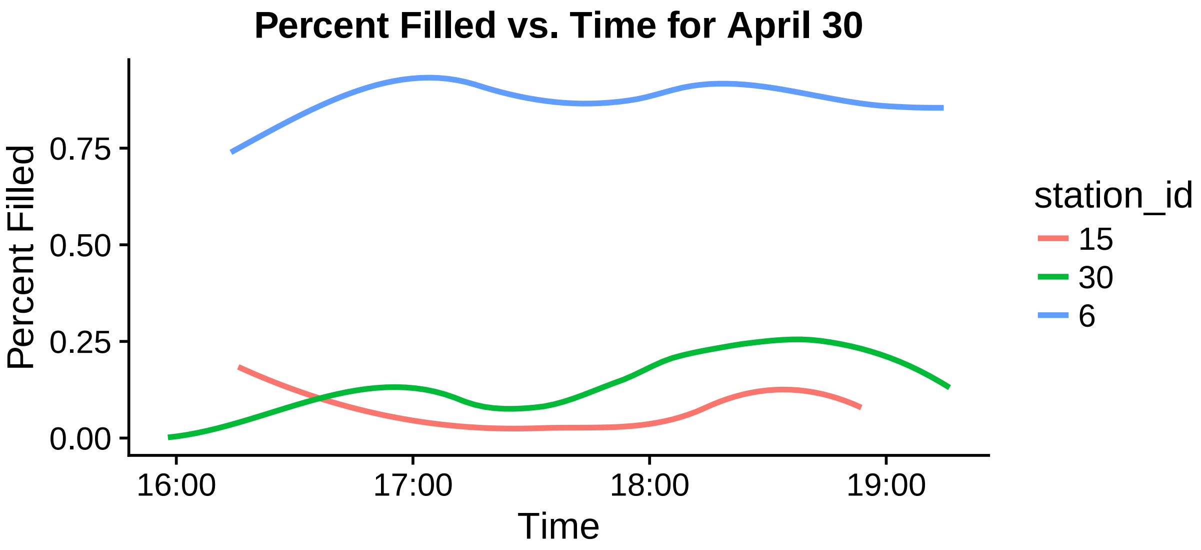




Real Time Station Status

Knowing that historically, stations 6, 15, and 30 are the more popular with regards to number of trips, I will now look at real time data for these specific stations. I saved data at various time points across a single day and combined all the data into a single data set, real_time_data. This work is done in real_time.R. The data can easily be updated and imported through real_time.R. As long as the data set is saved to a file called real_time_data.csv with the same specifications and columns given in real_time.R, the analysis here will work.

The data I retrieved was from April 30, from 4pm to 7pm.



One Sample t-test

data: Percent Filled for Station 15

t = 5.0964
df = 10
p-value = 0.0004665

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.04173866 0.10658670
mean of x: 0.07416268

One Sample t-test

data: Percent Filled for Station 6

t = 54.572
df = 10
p-value = 1.034e-13

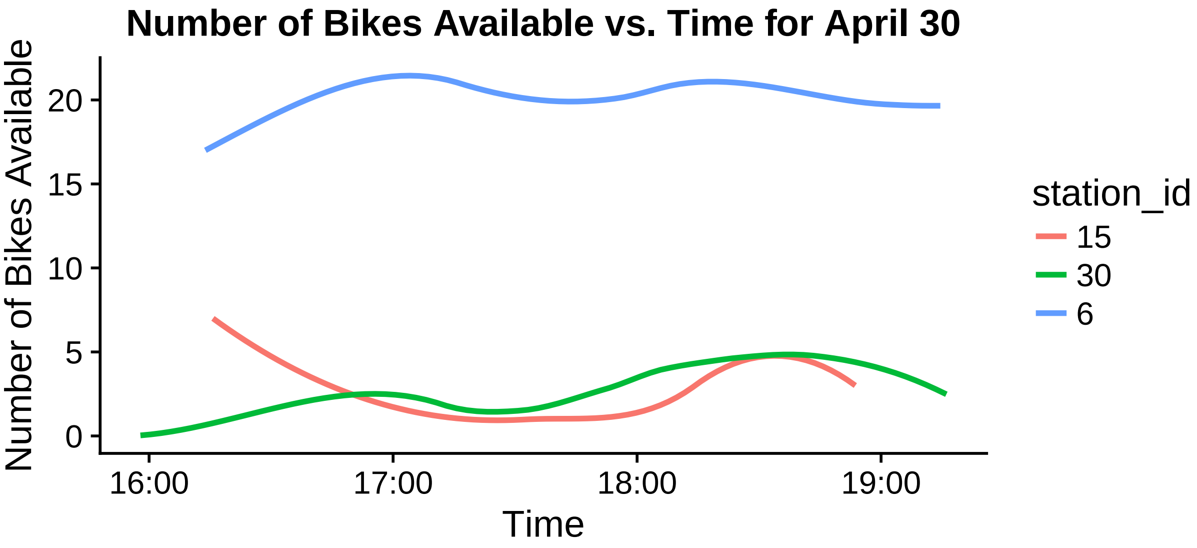
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.8378527 0.9091829
mean of x: 0.8735178

One Sample t-test

data: Percent Filled for Station 30

t = 5.4294
df = 10
p-value = 0.0002891

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.08745515 0.20919556
mean of x: 0.1483254



One Sample t-test

data: Number of Bikes for Station 15

t = 5.0964
df = 10
p-value = 0.0004665

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 1.586069 4.050295
mean of x: 2.818182

One Sample t-test

data: Number of Bikes for Station 6

t = 54.572
df = 10
p-value = 1.034e-13

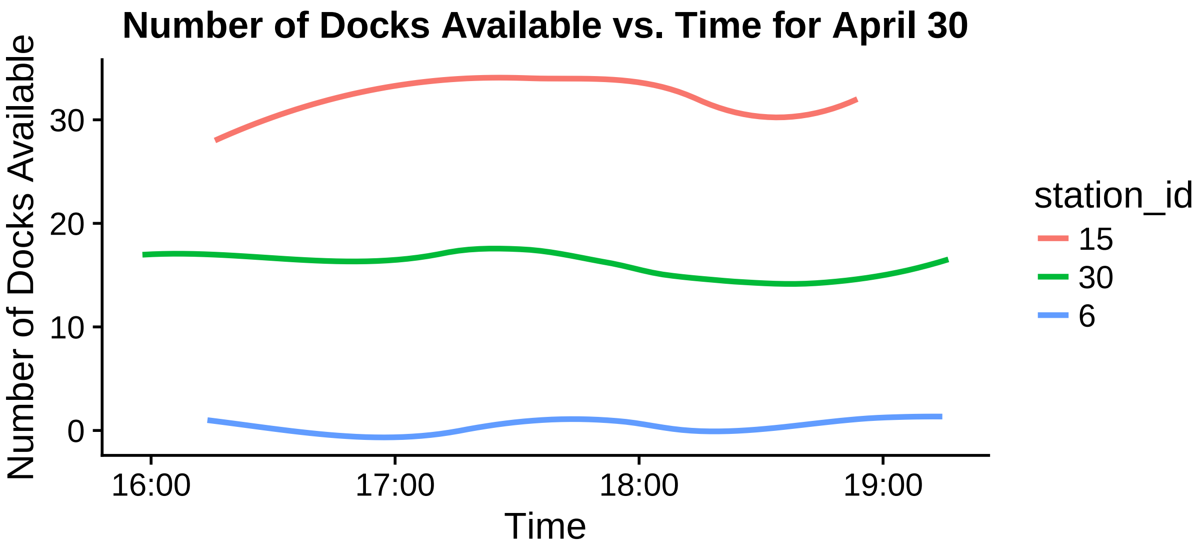
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 19.27061 20.91121
mean of x: 20.09091

One Sample t-test

data: Number of Bikes for Station 30

t = 5.4294
df = 10
p-value = 0.0002891

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 1.661648 3.974716
mean of x: 2.818182



One Sample t-test

data: Number of Docks for Station 15

t = 58.197
df = 10
p-value = 5.448e-14

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 30.94971 33.41393
mean of x: 32.18182

One Sample t-test

data: Number of Docks for Station 6

t = 3.1305
df = 10
p-value = 0.01068

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.183430 1.089297
mean of x: 0.6363636

One Sample t-test

data: Number of Docks for Station 30

t = 35.777
df = 10
p-value = 6.911e-12

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 15.00355 16.99645
mean of x: 16