Hasan Abdo

BTC 1855 - Coding in R

Midterm Assignment

Bay Area Bike Rental Operation Research

Tuesday August 6th, 2024

Table of Contents

Exploratory Data Analysis (EDA)

- The data comes in 3 different data sets. "Station" data, "Trip" data, and "Weather" data. For the purposes of clarity, the EDA will be divided by data set, where I will describe my findings/interpretations of each data set separately.

Station Data

- Data set contains 7 variables, each of which have 70 observations.

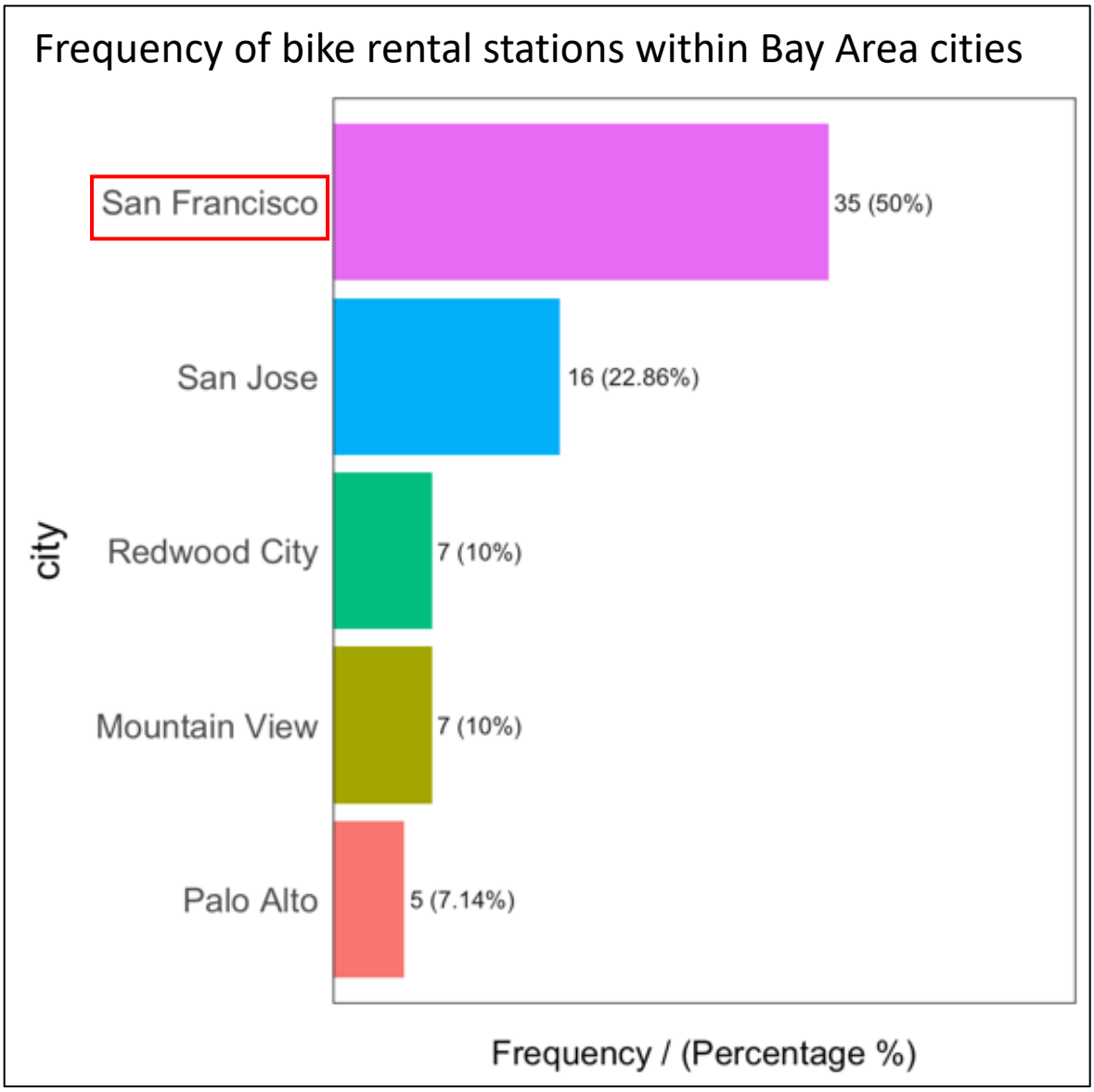

Figure 1: Frequency plot showing the frequency of stations within the 5 cities within the Bay Area.

- Data set shows that there are 5 distinct cities within the Bay Area. I constructed a plot to show which cities have the most stations. San Francisco is shown to have the highest number of stations, with 35, followed by San Jose with 16.
- Most frequent "installation date" was 8/23/2013.
- None of the variables in this dataset have any missing values.
- There don't seem to be any outliers to be removed in this data set.

Trip Data

- Data set contains 11 variables, each of which have around 326,339 observations.

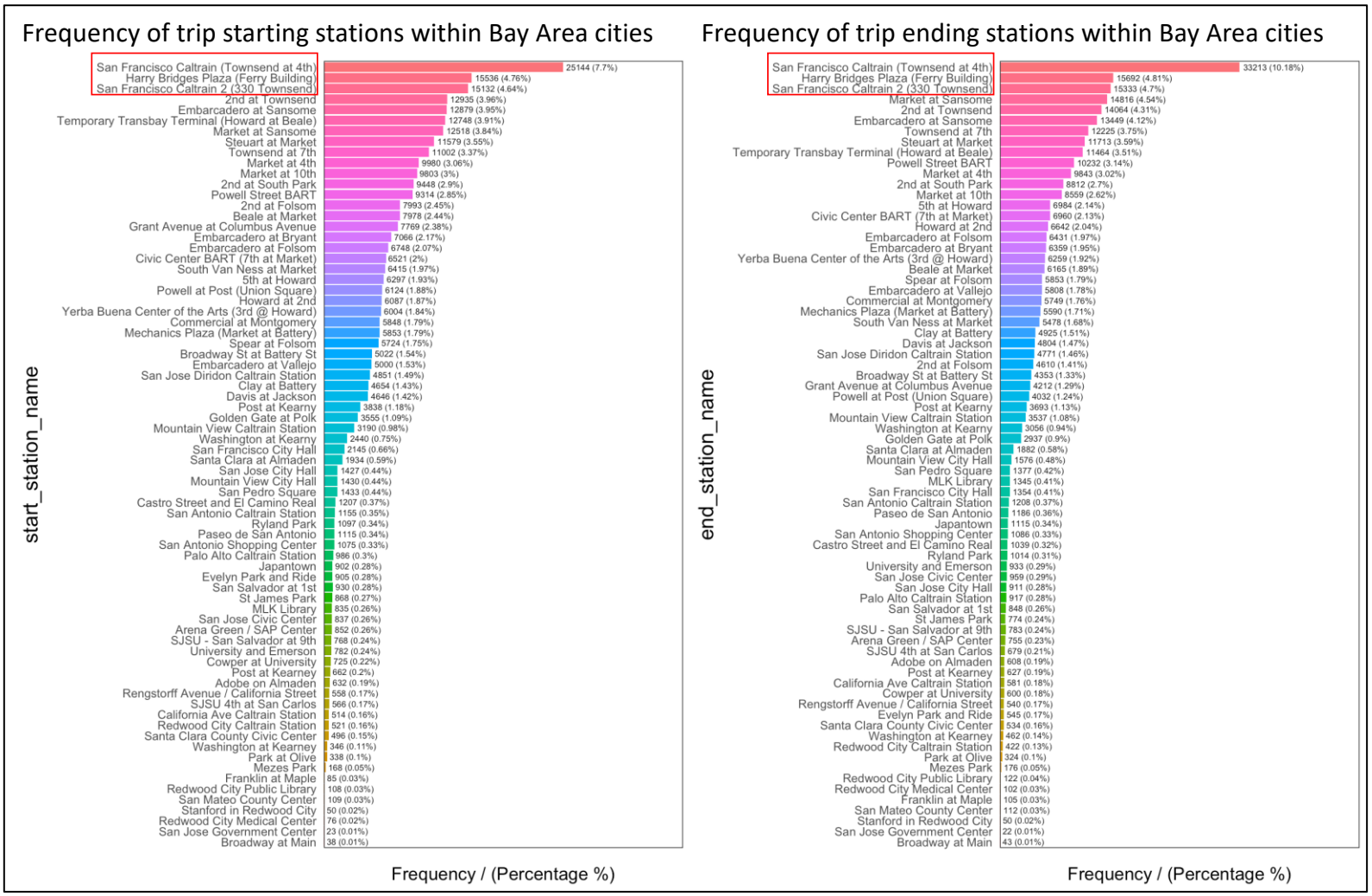

Figure 2: Frequency plot showing the frequency of trip start and end stations within the Bay Area.

- The most common start and end stations for trips were San Francisco Caltrain (Townsend at 4th), Harry Bridges Plaza (Ferry building), San Francisco Caltrain 2 (330 Townsend).

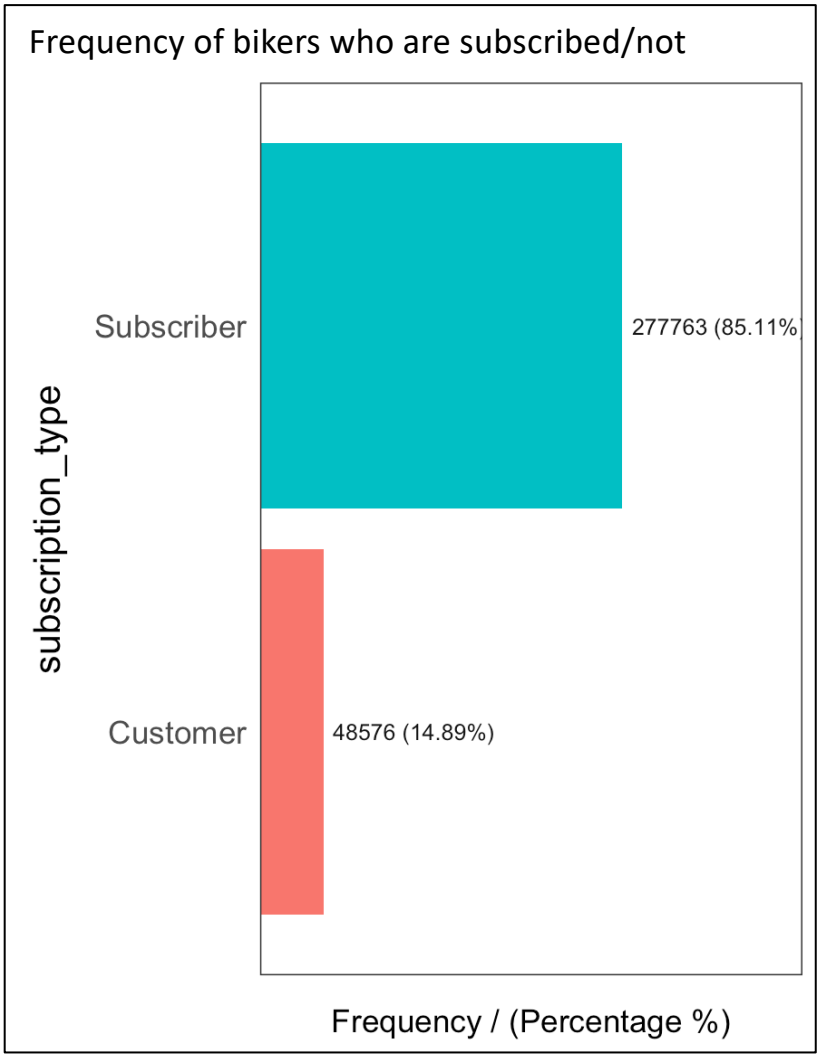

Figure 3: Frequency plot showing the frequency of bikers who are subscribed and those that use the service without subscription (customers).

- Most of the bikers taking trips were subscribed according to the data, around 85% of the trips were taken by subscribers.
- Shortest recorded trip time was a minute, with the longest being 199 days.
- There are 1493 missing entries in the zip code variable, with other observations that jump out as inappropriate as 0,1, “nil”, etc.

Weather Data

- This data set has 15 variables with around 1825 observations each.
- A couple of variables have missing values, those need to be investigated.

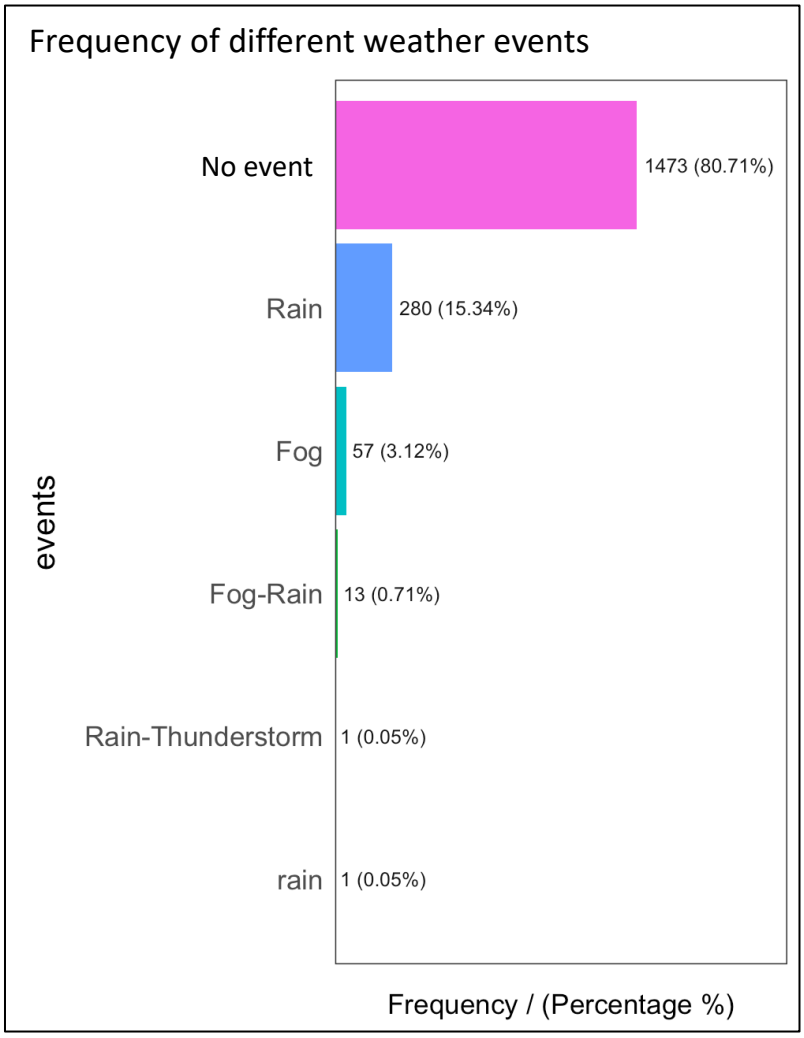

Figure 3: Frequency plot showing the weather events within the 5 Bay Area cities.

- May need to switch the "" in the events variable to NAs or "No event" indicating no weather event on those days.

Data Cleaning:

Station Data

- Did not do any cleaning to this data set. The data set was already clean, no missing variables, or any notable outliers.

Trip Data

- Main thing removed from this data set were inappropriate zip codes. These were not removed from the data set, but these zip codes were converted to NAs. This is because there are other valid observations for this data set (other variables), that would be removed with the zip code values, resulting in loss of data. These values include:

Table 1: Table showing the removed Zip Codes that were converted to NA values within the trip dataset.

Zip Codes converted to NAs (Invalid zip codes)
" "
v6z2x
nil
M4S1P
99999
9990540
0
1
100
1000
10000
100004

- It was indicated that "cancelled trips" would be those where the trip starts and ends at the same station, along with the duration being less than 3 minutes. There are 1082 of those trips within the dataset which were subsequently removed. A list of those IDs can be found here.
- I also removed the data point where the duration of the bike trip lasted for 199 days. The next closest data point was around 8 days. The ID of this specific trip was: 568474. It was removed from the data set.

Weather Data

- Removed completely blank values within the event column within the data set and replaced it with "No Event" indicating no weather event for that day. There were also 2 different rain events, one indicated as "rain" and the other as "Rain". I combined those two values as "Rain" to avoid any confusion.
- Within the precipitation column, there were "trace" values that are supposed to be values for when precipitation is less than 0.01. I manually imputed these values to 0.009, which is less than 0.01.

Rush Hour analysis

To ensure the accuracy of our rush hour analysis, I chose to exclude trips that exceed 5 hours. This decision is based on the assumption that most individuals use their bikes for commuting between their homes and workplaces. The longest possible commute within the Bay Area, from San Francisco to San Jose, typically takes around 5 hours by bike, according to Google Maps. While this duration might be shorter with electric bikes, excluding based on longer trips helps maintain a more conservative dataset, preventing any potential loss of data for later analysis. There were 2094 of those observations, and they were excluded from the rush hour analysis.

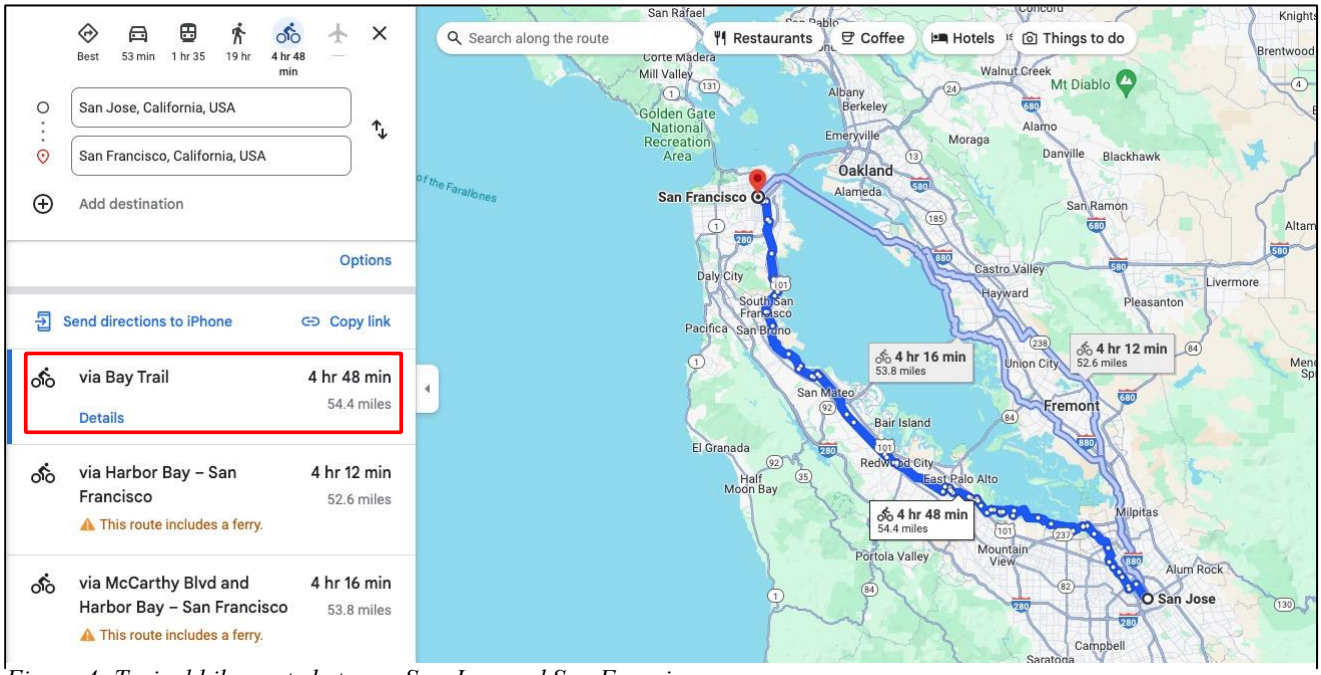

Figure 4: Typical bike route between San Jose and San Francisco.

- Rush hours were extracted from a “mid-point” time value. This value was included in order to best relate start and end dates/times to have 1 date/time stamp that can be used to extract the hours from.

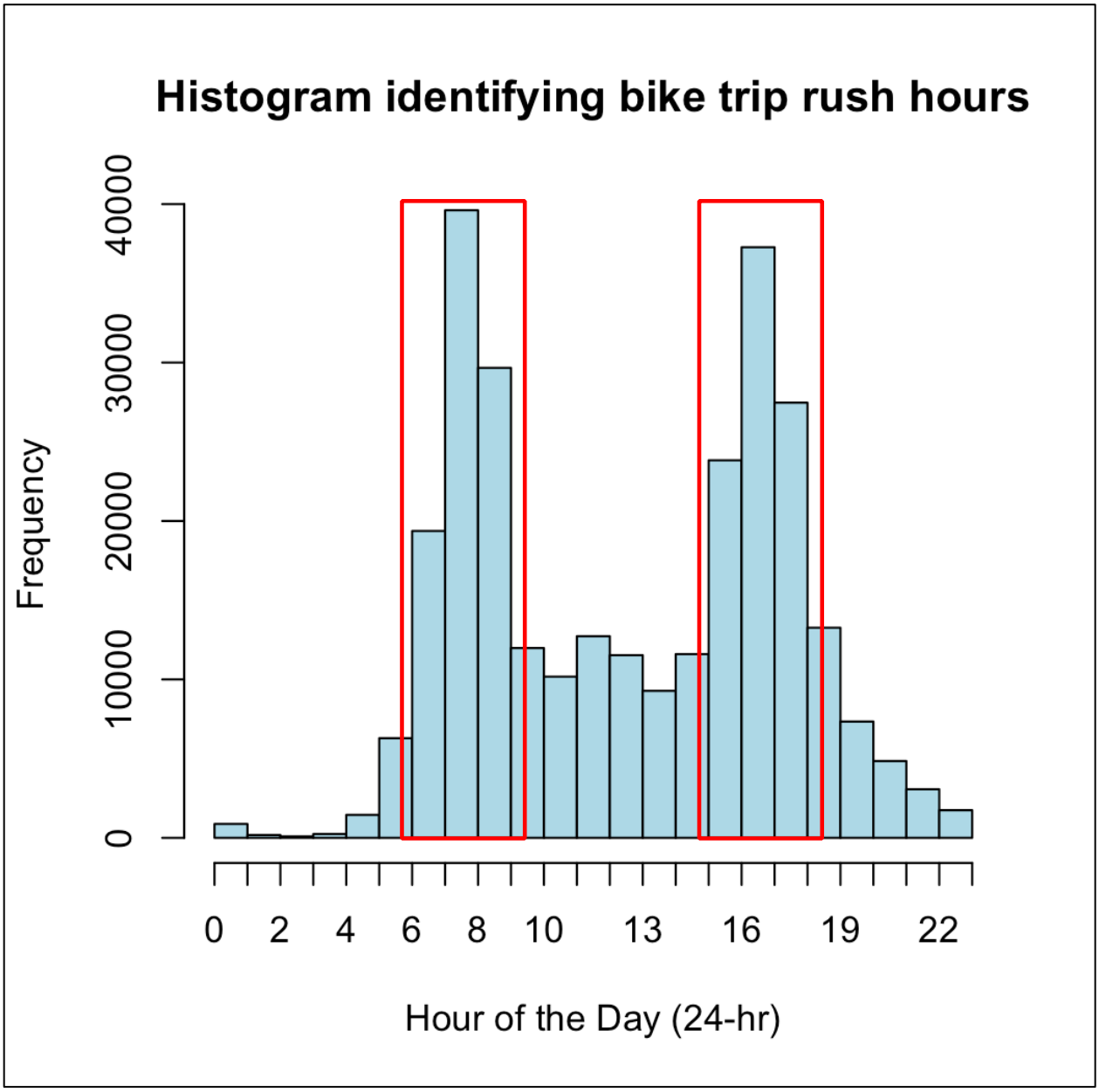

Figure 5: Histogram identifying weekday bike trip rush hours. These hours are roughly identified within the red boxes.

- Looking at the weekday hours histogram, it was determined that the rush hours for bike trips are from 7-10 am, and 4-7pm. Assuming bikers are taking a separate trip to and from work, this data aligns with typical workdays, meaning the highest peaks are when people are commuting to work and back from work.
- The 10 most common start and end stations during those rush hours can be seen in the next tables.

Table 2: Table showing the top 10 most frequently used start stations (in descending order, 1 = most) during established rush hours.

Station order	Station name
1	San Francisco Caltrain (Townsend at 4th)
2	San Francisco Caltrain 2 (330 Townsend)
3	Temporary Transbay Terminal (Howard at Beale)
4	Harry Bridges Plaza (Ferry Building)
5	2nd at Townsend
6	Steuart at Market
7	Market at Sansome
8	Townsend at 7th
9	Market at 10th
10	Embarcadero at Sansome

Table 3: Table showing the top 10 most frequently used end stations (in descending order, 1 = most) during established rush hours.

Station order	Station name
1	San Francisco Caltrain (Townsend at 4th)
2	San Francisco Caltrain 2 (330 Townsend)
3	Market at Sansome
4	2nd at Townsend
5	Temporary Transbay Terminal (Howard at Beale)
6	Harry Bridges Plaza (Ferry Building)
7	Townsend at 7th
8	Steuart at Market
9	Embarcadero at Sansome
10	2nd at South Park

Weekend stations

- It is important to note that not the same data was used for the weekend analysis. Since data for the rush hours excluded those trips that were longer than 5 hours, assuming people went to work on those days, the same assumption doesn't apply, since people don't work on weekends. Therefore, the removed data was also used in addition to the rush hour data for a completer and more reserved look at the data (preventing any potential loss).

Table 4: Table showing the top 10 most frequently used start stations (in descending order, 1 = most) during the weekends.

Station order	Station name
1	Harry Bridges Plaza (Ferry Building)
2	Embarcadero at Sansome
3	Market at 4th
4	Embarcadero at Bryant
5	2nd at Townsend
6	Powell Street BART
7	San Francisco Caltrain (Townsend at 4th)
8	Grant Avenue at Columbus Avenue
9	Market at Sansome
10	Powell at Post (Union Square)

Table 5: Table showing the top 10 most frequently used end stations (in descending order, 1 = most) during the weekends.

Station order	Station name
1	Embarcadero at Sansome
2	Harry Bridges Plaza (Ferry Building)
3	Market at 4th
4	Powell Street BART
5	San Francisco Caltrain (Townsend at 4th)
6	2nd at Townsend
7	Embarcadero at Bryant
8	Steuart at Market
9	Market at Sansome
10	Grant Avenue at Columbus Avenue

Utilization

- Average monthly utilization was calculated using the formula: duration of utilization per month/ time in a month, where time in a month is approximately 2628000 seconds.
- The trend observed here is that utilization peaks during the summer months, with the highest being during August and July. Utilization begins to fall around the winter month. Cooler weather likely discourages people to bike and use other methods of transport instead.
- Table including average monthly utilization values follows. Utilization here doesn't have units; however it could be treated as a monthly rate, which can be used to compare monthly utilization relative to other months throughout the year.

Table 6: Average monthly bike utilization rates for monthly comparison

Month	Average Utilization
January	9.14
February	7.44
March	10.65
April	11.05
May	12.09
June	12.93
July	13.31
August	13.47
September	12.43
October	12.36
November	8.43
December	9.08

Weather Correlation

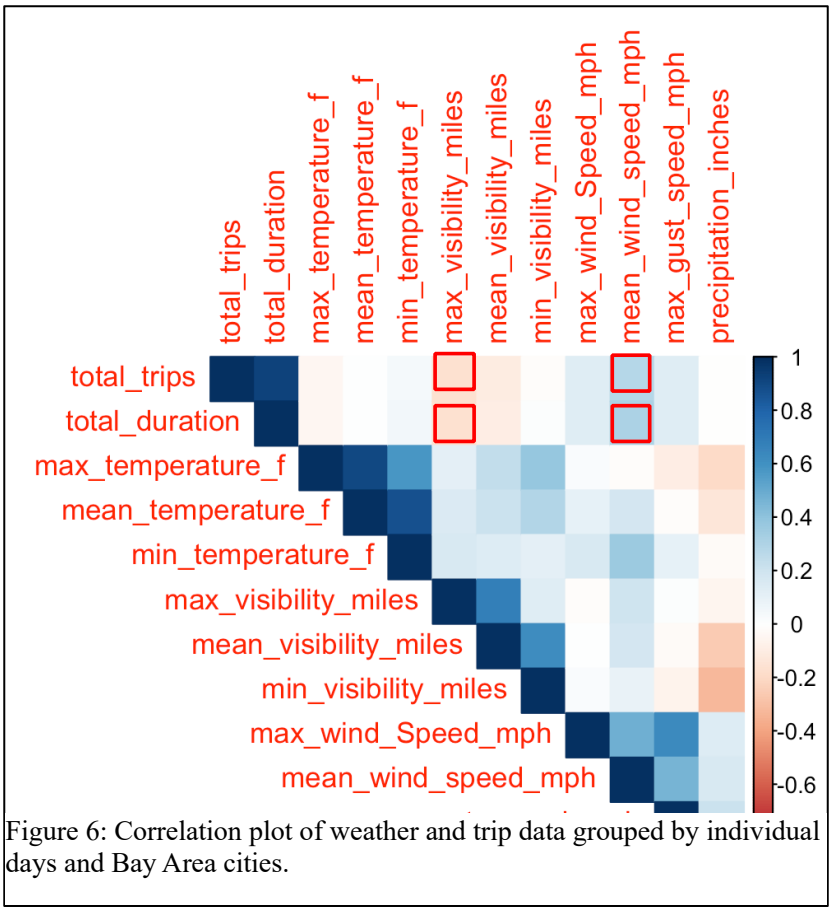

Figure 6: Correlation plot of weather and trip data grouped by individual days and Bay Area cities.

- The correlation between weather and trip data did not show any meaningful correlation. Of note however, the amount of max visibility did appear to have a negative correlation on the number of trips taken per day, as well as the overall duration of trips taken on an individual day. The correlation was -0.16 and -0.16 for total trip and total trip duration when correlated with max visibility respectively.
- Another of note correlation was that of the total number of daily trips and total duration of daily trips with the mean wind speed, which was a slight positive correlation. The correlation between total number of daily trips and mean wind speed was 0.28 while the correlation between total daily trip duration and mean wind speed was 0.32.