**Bay Area Bike Rental Operation Research**

Breanna Vasko, Ning Xu

University of Toronto Mississauga

[MSC2011H](https://q.utoronto.ca/courses/262687): Special Topics in Biomedical Communications

Dr. S.K.

July 13th, 2022

Data from the Bay Area Bike Operations was collected over 12 months in the year 2014. Information about the bike trips, weather, and bike stations were collected and organized into three separate data frames. All figures throughout this report can be referred to in the Appendix section.

**Preliminary Data**

**Exploratory Data Analysis**

An exploratory data analysis of both the trips data set and weather data set were completed. The most frequently used starting and ending station is the San Francisco Caltrain station. The following most frequently used stations can be viewed in Figure 1 and 2. Descriptive statistics were viewed for numeric variables within the trips data set. Descriptions of the duration of customers and subscribers trips can be found in Table 1.

Table 1. Description of statistics for the duration of trips from the trips data set in seconds.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trips Data Set | Mean | IQR | Standard Deviation | Sample size(n) |
| Duration of trip: | 1131.97 | 405 | 30816.16 | 326339 |

The amount of precipitation for each event can be viewed in Figure 3. No precipitation occurred for 1542 trips out of the entire data set. Descriptive statistics of all numerical variables in the weather data set can be found in Table 2.

Table 2. Descriptive Statistics of Numeric Values from the Weather Data Set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Mean** | **Standard Deviation** | **IQR** |
| max\_temperature\_f | 71.03 | 8.26 | 12 |
| mean\_temperature\_f | 62.03 | 6.75 | 11 |
| min\_temperature\_f | 52.83 | 6.67 | 10 |
| max\_visibility\_miles | 10.86 | 2.62 | 0 |
| mean\_visibility\_miles | 9.97 | 1.62 | 0 |
| min\_visibility\_miles | 8.12 | 3.04 | 3 |
| max\_wind\_Speed\_mph | 16.44 | 7.32 | 7 |
| mean\_wind\_speed\_mph | 6.10 | 3.05 | 4 |
| max\_gust\_speed\_mph | 22.69 | 9.09 | 7 |
| cloud\_cover | 3.00 | 2.30 | 4 |

**Outliers**

Any trips with a duration of less than 2 minutes (120 seconds) were considered as a canceled trip and were removed from the trips data set. This removed 2499 observations from the trips data set. An extremely large value of 17270400 was observed in the durations column in the trips data set which was significantly larger than the rest of the values in the column leading to the removal of this specific value. To remove any outliers from the weather data set, any values within any numeric columns that were not within 99.9% of the data range were removed. This led to the removal of 522 observations from the weather data set.

**Findings**

After cleaning the data frames, rush hours during the weekdays were found where the trip volume was the highest at the starting and ending stations. The peak hours can be observed in Table 1.0. Weekdays with peak hours for highest trip volume.

|  |  |  |
| --- | --- | --- |
| Day of the Week | Start Station Peak Hours | End Station Peak Hours |
| Monday | 8-9AM | 5-6PM |
| Tuesday | 8-9AM | 8-9AM |
| Wednesday | 8-9AM | 8-9AM |
| Thursday | 8-9AM | 8-9AM |
| Friday | 8-9AM | 8-9AM |

Next, the 10 most frequent starting and ending stations during these rush hours can be found in Table 1.1. The stations were found by first keeping the trip data where the starting and ending days were weekdays, and then finding the starting and ending station names for the corresponding rush hours demonstrated in Table 1.0. Then, the starting and ending station names are put into individual tables and sorted according to their frequencies in a descending order. Lastly, the first 10 station names in the ordered table along with their frequencies were found and used to compute Table 1.1. The starting and ending stations with the highest frequencies during the rush hours on weekdays were San Francisco Caltrain (Townsend at 4th) and San Francisco Caltrain (Townsend at 4th), respectively.

Table 1.1. Ten most frequent starting and ending stations during the rush hours on weekdays.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank\* | Starting Station | Starting Station Frequency | Ending Station | Ending Station Frequency |
| 1 | San Francisco Caltrain (Townsend at 4th) | 6143 | San Francisco Caltrain (Townsend at 4th) | 4292 |
| 2 | Harry Bridges Plaza (Ferry Building) | 3476 | Market at Sansome | 2070 |
| 3 | San Francisco Caltrain 2 (330 Townsend) | 3398 | San Francisco Caltrain 2 (330 Townsend) | 1798 |
| 4 | Temporary Transbay Terminal (Howard at Beale) | 3135 | 2nd at Townsend | 1796 |
| 5 | Steuart at Market | 1934 | Townsend at 7th | 1784 |
| 6 | Grant Avenue at Columbus Avenue | 1620 | Steuart at Market | 1506 |
| 7 | 2nd at Townsend | 1297 | Harry Bridges Plaza (Ferry Building) | 1463 |
| 8 | Embarcadero at Bryant | 1128 | Temporary Transbay Terminal (Howard at Beale) | 1425 |
| 9 | Civic Center BART (7th at Market) | 1123 | Embarcadero at Sansome | 1375 |
| 10 | South Van Ness at Market | 1015 | Embarcadero at Folsom | 1207 |

\*(Most frequent = 1; least frequent = 10)

The 10 most frequent starting and ending stations during the weekend can be found in Table 1.2. These stations were found by a similar method as the one for Table 1.1, but more sophisticated and less efficient. This is done by first keeping the trip data where the starting and ending days were weekends. Then, a copy of the station data with only the station name column was made. Two more columns were added to the copy, with each counting the number of times each starting and ending stations were used, respectively. Another copy of the station names was also made for looping.

A nested for loop was used to iterate through the copy of station names and the starting or ending station names. The corresponding frequency in the frequency column of the station data copy incremented by 1 when the starting or ending station name from the trip data matched with the current station name. Then, the copy of station data was first ordered in a descending order according to the frequency of the starting stations. The first 10 stations were the 10 most frequent starting stations during the weekend. Then, the copy of station data was ordered in a descending order according to the frequency of the ending stations. The first 10 stations were the 10 most frequent ending stations during the weekend. These station names along with their frequencies can be found in Table 1.2. The starting and ending stations with the highest frequencies during the weekends were Harry Bridges Plaza (Ferry Building) and Embarcadero at Sansome, respectively.

Table 1.2. Ten most frequent starting and ending stations during the weekends.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank\* | Starting Station | Starting Station Frequency | Ending Station | Ending Station Frequency |
| 1 | Harry Bridges Plaza (Ferry Building) | 3169 | Embarcadero at Sansome | 3376 |
| 2 | Embarcadero at Sansome | 3128 | Harry Bridges Plaza (Ferry Building) | 3175 |
| 3 | Market at 4th | 1661 | Market at 4th | 1871 |
| 4 | Embarcadero at Bryant | 1591 | Powell Street BART | 1675 |
| 5 | 2nd at Townsend | 1546 | San Francisco Caltrain (Townsend at 4th) | 1653 |
| 6 | Powell Street BART | 1488 | 2nd at Townsend | 1584 |
| 7 | San Francisco Caltrain (Townsend at 4th) | 1365 | Embarcadero at Bryant | 1354 |
| 8 | Grant Avenue at Columbus Avenue | 1299 | Steuart at Market | 1213 |
| 9 | Powell at Post (Union Square) | 1091 | Grant Avenue at Columbus Avenue | 1098 |
| 10 | Market at Sansome | 1091 | Market at Sansome" | 1098 |

\*(Most frequent = 1; least frequent = 10)

The average utilization of bikes for each month can be found in Table 1.3. It should be noted that there were two assumptions made. One of them was that the average utilization of bikes for each month was interpreted as the total time(duration) used/total time (total number of times bikes were used) in month, so it means the average duration in seconds per bike rental in a month. The second one was that when the starting and ending months were not equal, the number of times the bikes were used was one in both the starting and ending months.

First, the total utilization (duration) of bikes and the number of times bikes were used for each month, where the starting and ending months were equal, were found. The dataset analysis was separated according to whether the starting and ending months were equal or not, because if the starting and ending months were not equal, the duration needed to be divided into the corresponding months.

Then, for the part of the dataset where the starting and ending months did not equal, the total time in seconds from the starting date (time) to the end of the staring month was found. This number was stored into a new column. Then, the number was subtracted from duration, and the result was put into another new column. The duration only needed to be divided into two components because the maximum difference between starting and ending months was one. Two subsets of the dataset were created. Each of them contained only the utilization and the number of times bikes were used for adding to the starting or ending months, respectively.

The two subsets were merged with the table containing only the equal starting and ending months data, and then the utilization and number of times bikes were used in each month were summed accordingly. The total utilization in each month was divided by the corresponding number of times bikes were used to obtain the average utilization of bikes for each month. The results can be found in Table 1.3. December had the highest average utilization of bikes.

Table 1.3. Average utilization of bikes for each month.

|  |  |
| --- | --- |
| Month | Average Utilization of Bikes (sec/bike) |
| 1 | 1016 |
| 2 | 1044 |
| 3 | 1159 |
| 4 | 1117 |
| 5 | 1134 |
| 6 | 1150 |
| 7 | 1128 |
| .g. 8 | 1153 |
| 9 | 1049 |
| 10 | 970 |
| 11 | 911 |
| 12 | 1226 |

To investigate on whether the weather conditions have an impact on the bike rental patterns, weather variables such as temperature, wind speed, gust speed, cloud cover, visibility were used to create a correlation matrix with the duration of bike rental trips, which can be seen in Figure 4. The highest correlations should be in dark blue or dark red colours. The matrix was created from the new dataset combining trip data with the weather data using the “city” and “date” columns. The correlation coefficients between duration and the weather variables can be seen in Table 1.4. The correlation coefficients were all close to zero in Table 1.4, and they were all transparent in Figure 4, which indicates that there was no correlation between bike rental duration and the weather variables.

Table 1.4. Correlation coefficients showing relationship between bike rental duration and weather variables.

|  |  |
| --- | --- |
| Weather Variables | Correlation Coefficient With Duration |
| Max Temperature (F) | 0.0011 |
| Mean Temperature (F) | -0.0011 |
| Min Temperature (F) | -0.0028 |
| Max Visibility (miles) | 0.0059 |
| Mean Visibility (miles) | 0.0055 |
| Min Visibility (miles) | 0.0065 |
| Max Wind Speed (mph) | -0.0054 |
| Mean Wind Speed (mph) | -0.0044 |
| Max Gust Speed (mph) | -0.0084 |
| Precipitation (inches) | 0.0011 |
| Cloud Cover | -0.0152 |

**Appendix**

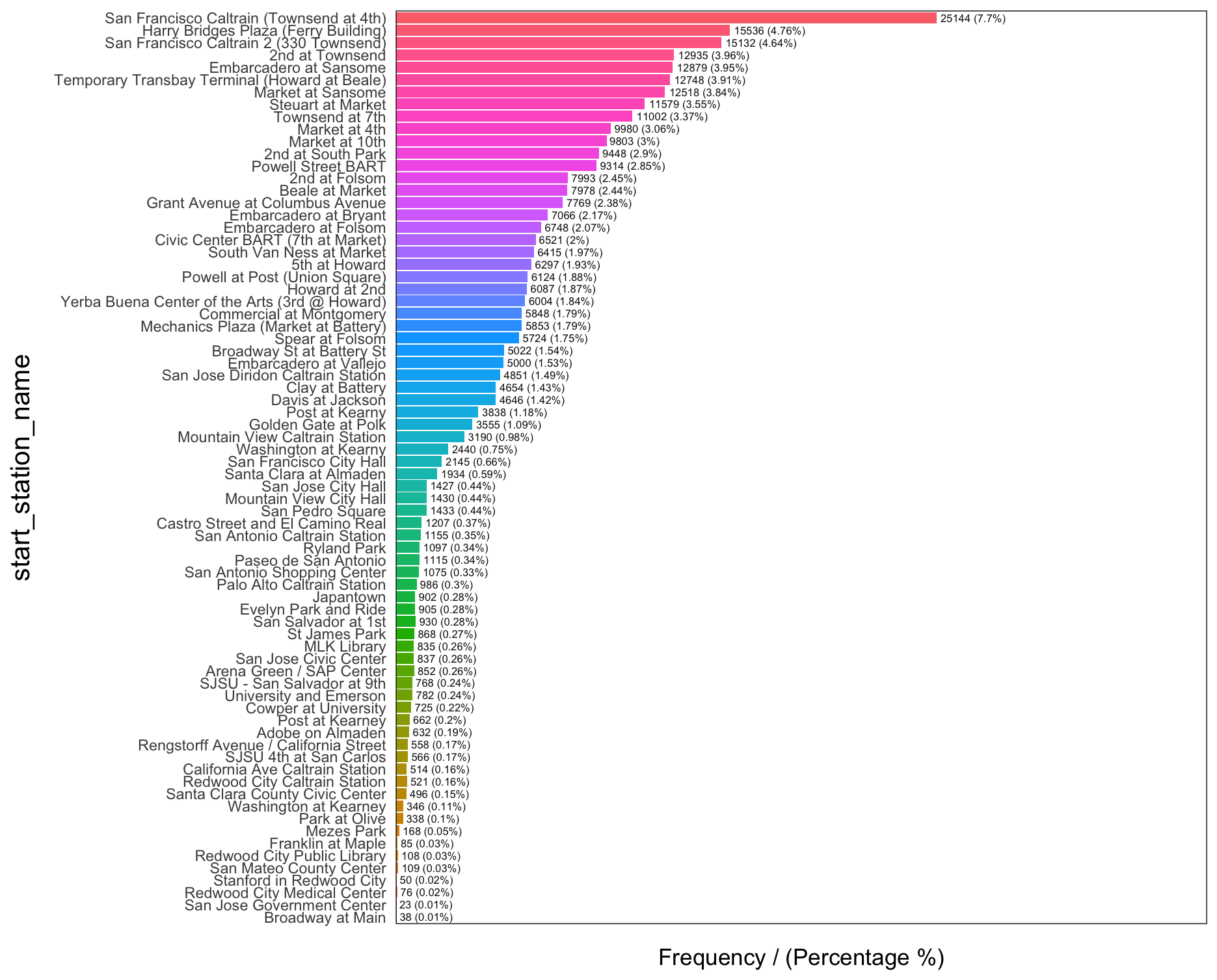


Figure 1. Frequency of Starting Stations Used

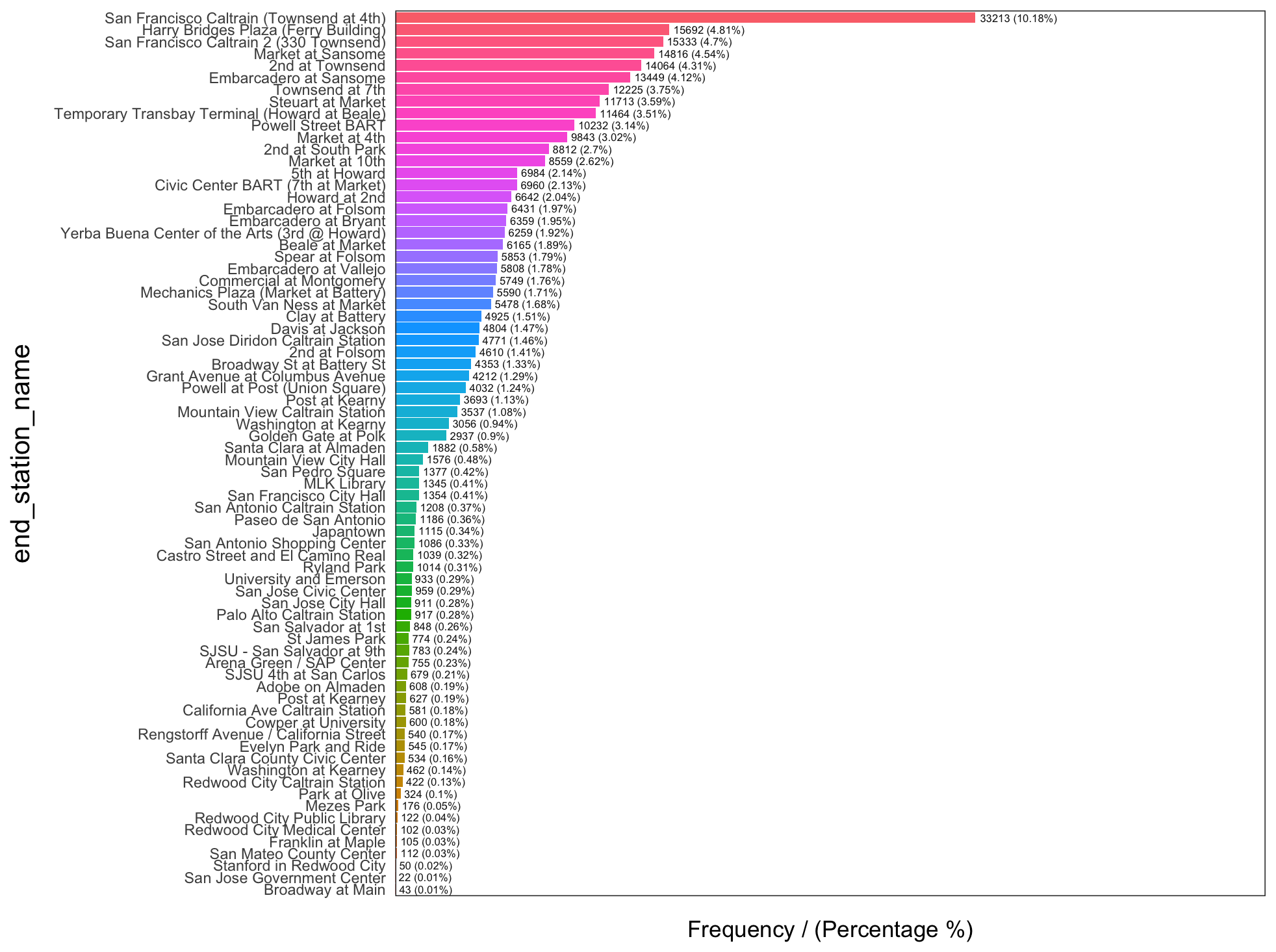


Figure 2. Frequency of Ending Stations Used

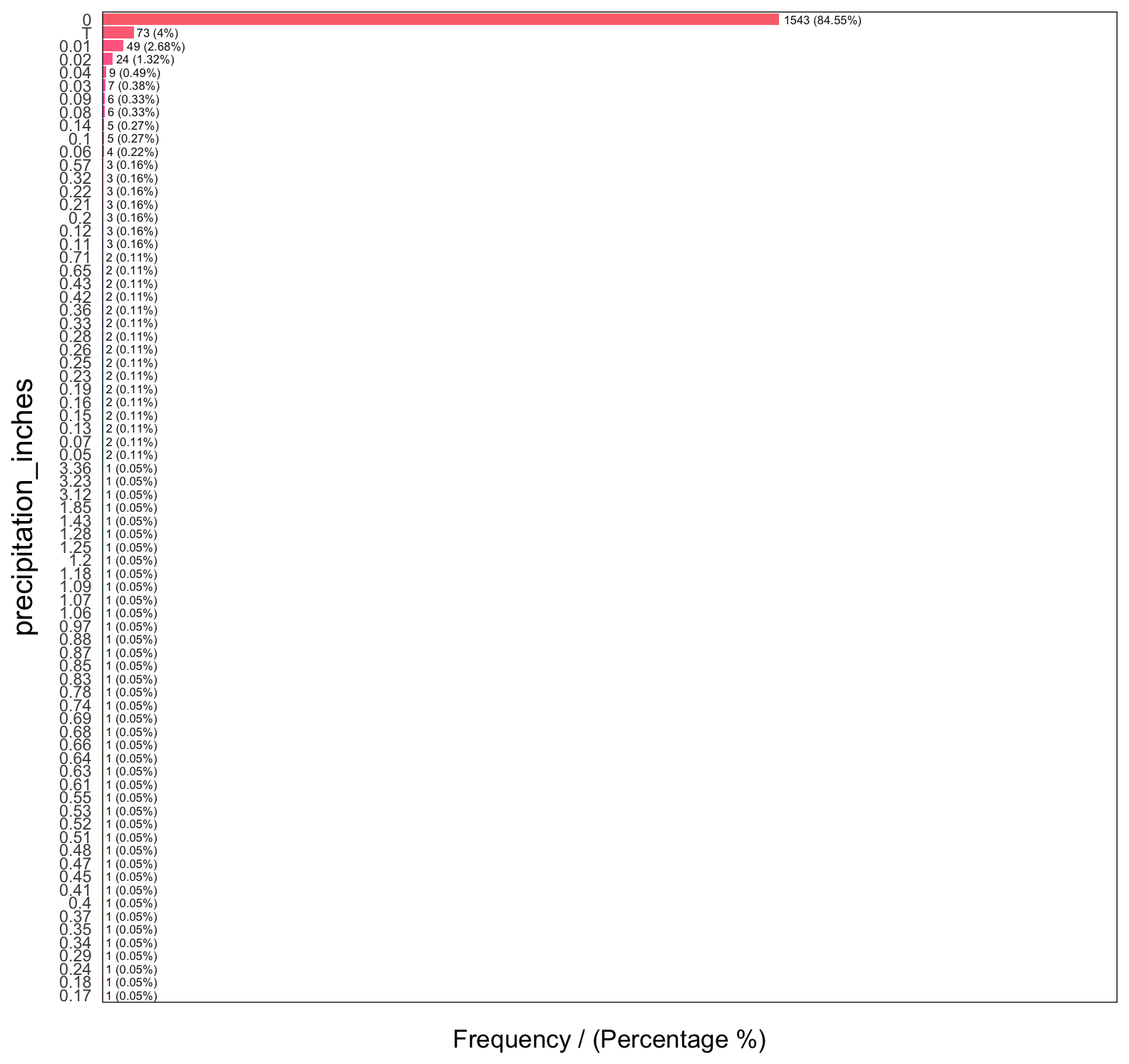
****

Figure 3. Trip frequencies and their amount of Precipitation

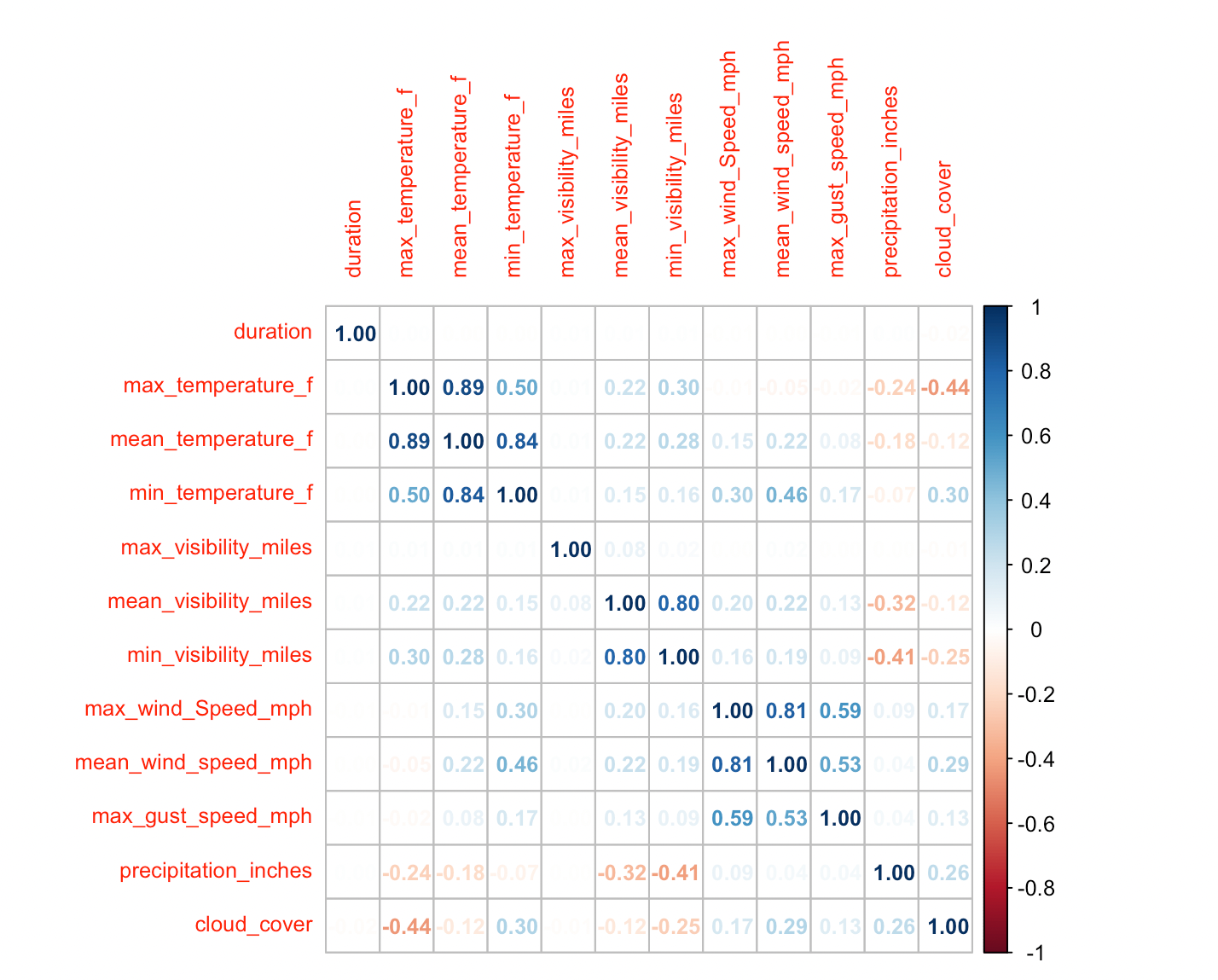


Figure 4. Correlation matrix plot showing relationship between bike rental duration and weather variables.

​​