**Bay Area Bike Rental Operation Data Analysis Report**

MSC2011H Midterm

Imran Rhemtulla & Archana Jeyagaran

July 13, 2022

**Exploratory Data Analysis of the Weather Dataset**

Pre-processing steps were applied before data analyses began. “T” values representing trace amounts of precipitation within the precipitation\_inches variable were converted to 0 and the variable was converted to numerical. “rain” was converted to “Rain” in the events variable to properly group the variable. Within the events variable, there were 1473 observations that were null. These observations were converted to “No Event”, indicating that no weather event took place. The variable zip\_code was converted to a categorical variable from numerical as there were only 5 zip codes, one for each of the cities described in the dataset.

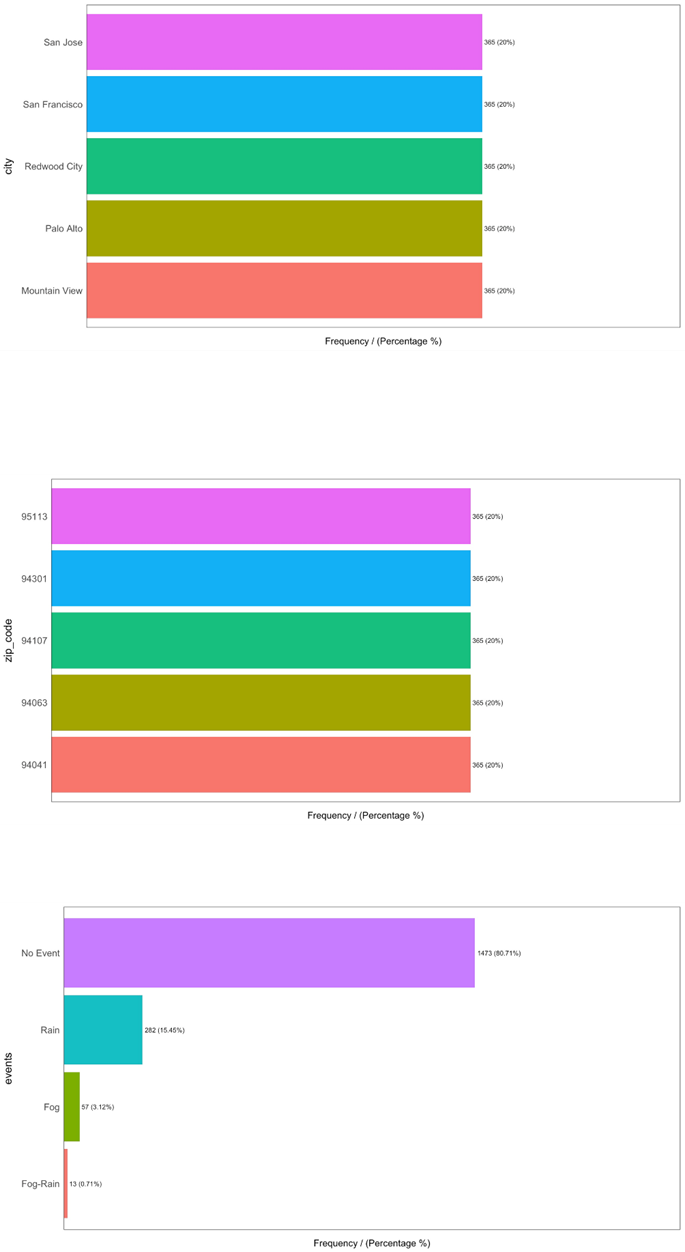
Within the weather dataset there are a total of 1825 observations and 15 variables. 5 (0.27%) observations for each date were present in the data set and ranged from 1/1/2014 to 6/15/2014. In the city variable there were 365 (20%) observations from Mountain View, 365 (20%) observations from Palo Alto, 365 (20%) observations from Redwood City, 365 (20%) observations from San Francisco, and 365 (20%) observations from San Jose (Figure 1). Similarly, the zip codes within the zip\_codes variable were general and represent each of the regions. 365 (20%) observations were present for 94041 (Mountain View), 365 (20%) observations for 94301 (Palo Alto), 365 (20%) observations for 94063 (Redwood City), 365 (20%) observations for 94107 (San Francisco), and 365 (20%) observations for 95113 (San Jose) (Figure 1) Of these dates, 1473 (80.71%) were recorded as undergoing no weather event across all cities, 282 (15.45%) were noted to experience rain, 57 (3.12%) had fog documented across all cities, and 13 (0.71%) had both fog and rain.

         Within the numerical variables, 1616 (88.55%) of observations for precipitation\_inches were zero. 451 observations for maximum gust speed were missing (24.71%), thus maximum gust speed had less than 80% of all values as non-NA values.  The mean of maximum temperature was 71.03 with a standard deviation (SD) of 8.26. The mean of mean temperature was 62.04 with an SD of 6.75. The mean of minimum temperature was 52.83 with an SD of 6.67.  The mean of max visibility was 10.86 with an SD of 2.62.  The mean of mean visibility was 9.97 with an SD of 1.62. The mean of minimum visibility was 8.11 with an SD of 3.04.  The mean of maximum wind speed 16.43 was with an SD of 7.32.  The mean of mean wind speed was 6.11 with an SD of 3.05. The mean of maximum wind speed was 22.69 with an SD of 9.09. Precipitation had an extremely high variation of coefficient of 5.97 and broad 98-percentile range of 9-45.81, indicating the presence of outliers. The mean precipitation was 0.03 with an SD of 0.18 as well. The variable describing cloud cover had a mean of 3, with an SD of 2.30. Representative histograms can be found in Figure 2.

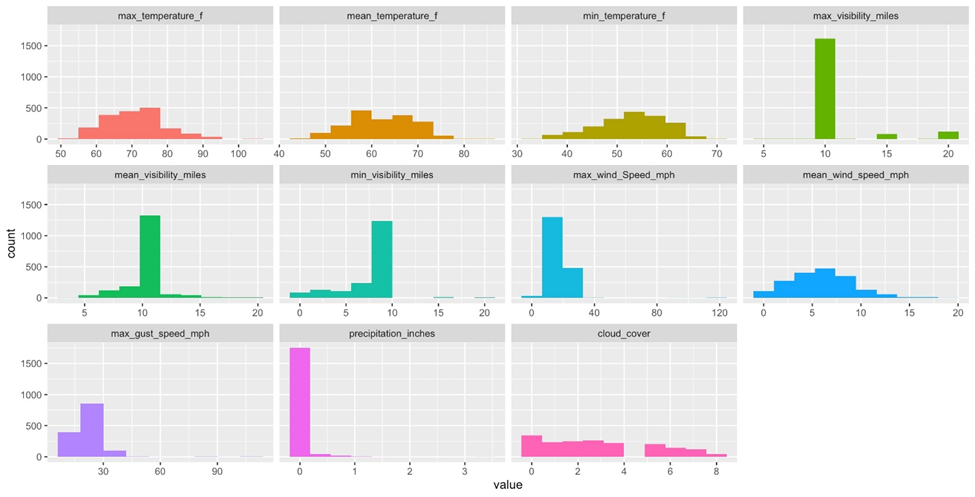
**Exploratory Data Analysis of the Trip Dataset**

The trip dataset contained 326,339 trip observations of 11 variables from 01/01/2014 to 09/09/2014. Within the 11 variables there was one numerical variable, duration, and the rest of the variables were categorical; id, start station id, end station id, bike id, start date, start station name, end date, end station name, subscription type, and zip code. The zip code variable contains a total of 50 observations that were recorded as 0’s and can be interpreted as NAs; all other data was complete.

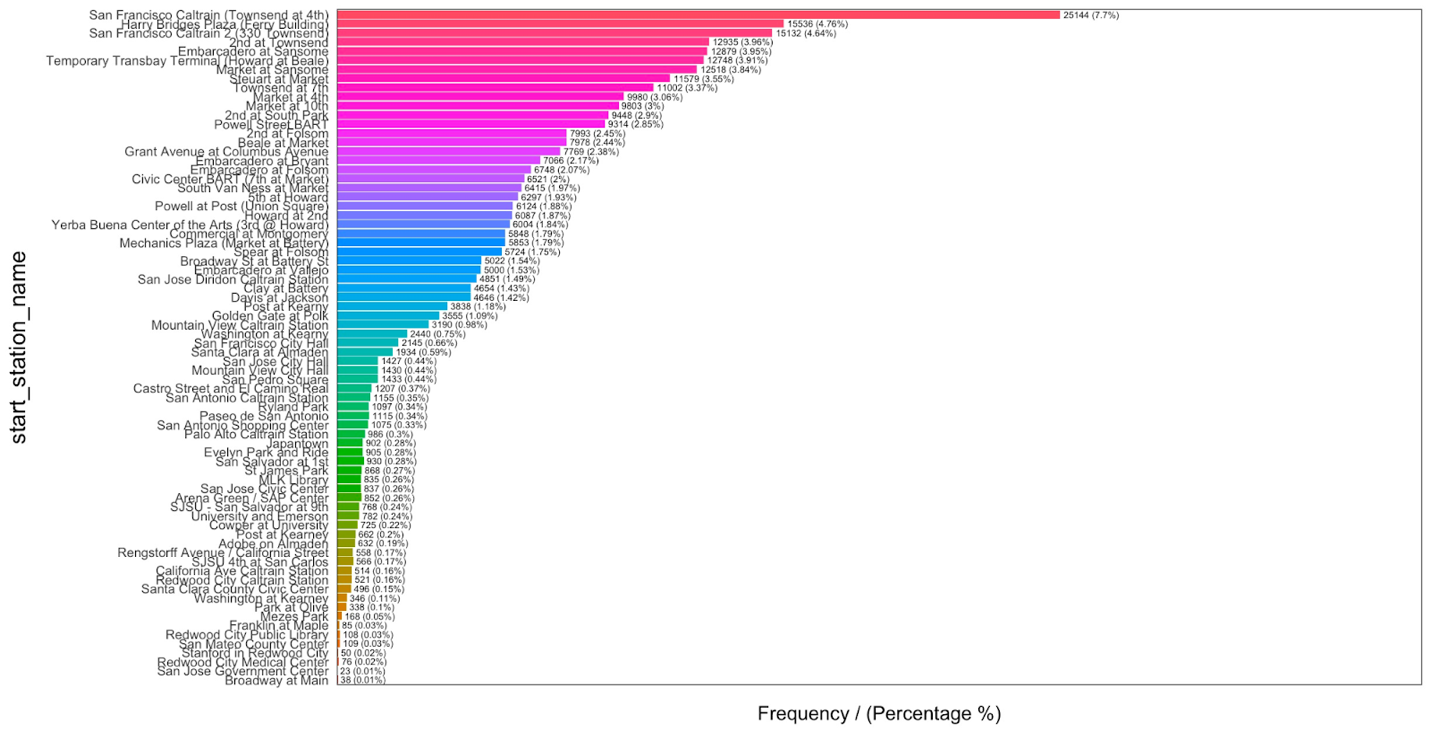
There was an average trip duration in seconds was ~1132 seconds. The duration data is also skewed to the right, has a high variance (27), and high standard deviation (30816). Out of the 74 different stations, San Francisco Caltrain (Townsend at 4th), had the highest frequency with 25144 (7.7%) start station observations (Figure 3) and 33213 (10.18%) end station observations (Figure 4). Figure 5 displays the frequency of trips that were made by subscribers and customers. It was found that ~85% (277763 observations) of the trips from this data set were from subscribers and the other ~15% (48576 observations) of the trips were from the customers.



**Figure 1.** Barplots of categorical variables city, zip\_code, and events within the weather dataset.



**Figure 2.** Representative histograms of numerical variables max\_temperature\_f, mean\_temperature\_f, min\_temperature\_f, max\_visibility\_miles, mean\_visibility\_miles, min\_visibility\_miles, max\_wind\_Speed\_mph, mean\_wind\_speed\_mph, max\_gust\_speed\_mph, max\_gust\_speed\_mph, precipitation\_inches, cloud\_cover.



**Figure 3:** Bar plot of categorical variable start\_station\_name with the trip dataset.

Histogram

Description automatically generated with medium confidence

**Figure 4:** Bar plot of categorical variable end\_station\_name with the trip dataset.

A picture containing rectangle

Description automatically generated

**Figure 5:** Bar plot of categorical variable subscription\_type with the trip dataset.

**Cancelled Trips**

Within the trip dataset cancelled trips were identified to be trips with duration less than 2 minutes. It was found that there was a total of 2499 cancelled trips within this dataset, these observations are stored in the trip\_data\_cancelled dataframe and removed from the original dataset.

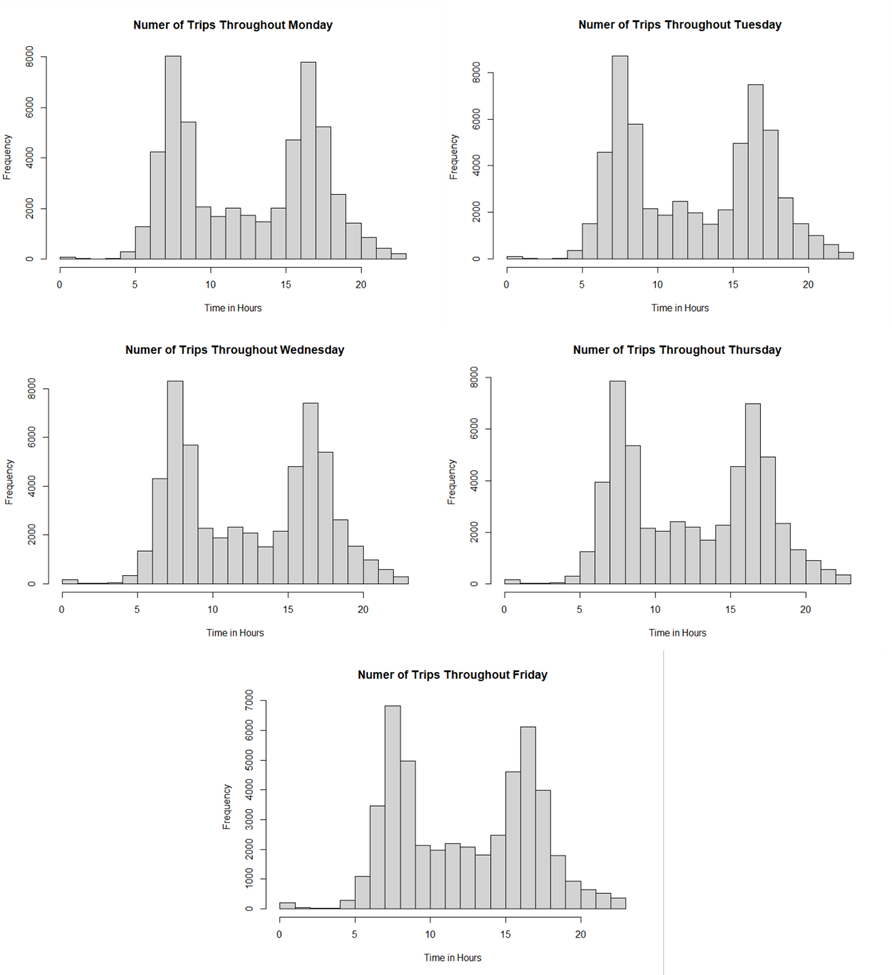
**Outliers**

The outliers for the trip and weather dataset were evaluated by constructing box plots for each numerical variable and using the boxplot.stats( ) function in R to identify the outliers. The function essentially looks for values of the datapoints which lie beyond the extremes of the whiskers of the boxplot and removes them per the 1.5IQR rule. There were a total of 24927 outliers in the trip dataset this data is stored in the trip\_outliers dataframe, and all outliers were extracted from the duration variable. The cleaned trip dataset is stored in the trip\_no\_outliers dataset.

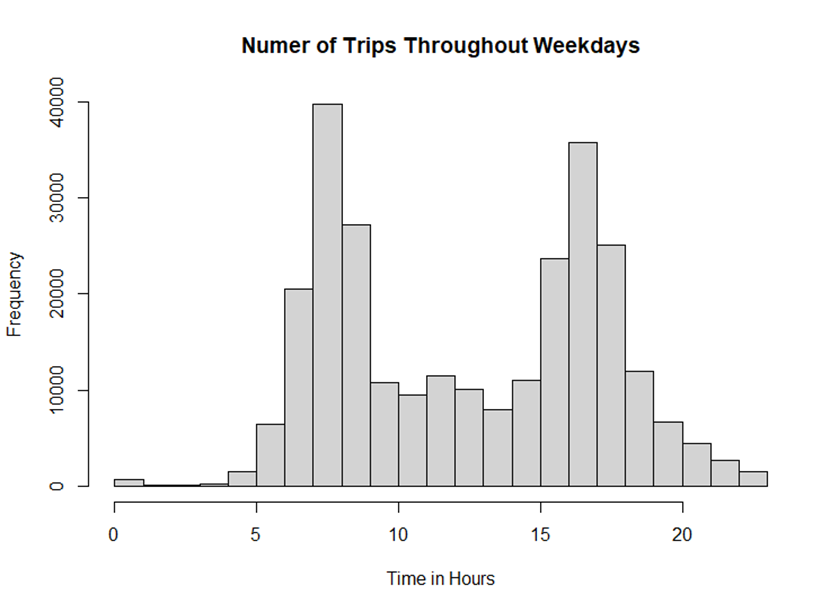
There were a total of 80 outliers in the weather dataset and were stored in the weather\_no\_outliers dataset. Outliers were evaluated for the following variables within weather dataset: mean temperature, minimum temperature, maximum temperature, mean visibility, minimum visibility, maximum visibility, mean windspeed, maximum wind speed, maximum gust speed, and precipitation. Mean temperature did not contain any outliers. Minimum visibility contained outliers, however the distribution was onesided and thus the outliers were removed. The boxplot for precipitation also showed the data too be heavily and the outliers were ignored for this variable as well. Both mean and minimum visibility had outliers with an IQR of 0 so the outliers were ignored for these variables. The minimum and maximum temperature variables each had one outlier that were stored in the mintemp\_outrows and maxtemp\_outrows dataframes respectively. Mean wind had 19 outliers that were stored in mean\_wind\_outrows, maximum wind had 14 outliers that are stored in max\_wind\_outrows, and maximum gust had 44 outliers that were stored in max\_gust\_outrows. The cleaned weather dataset was stored in the weather\_data dataset.

**Rush Hours on Weekdays**

Histograms indicated that the busiest hours of each weekday individually were 8 AM and 5 PM based on number of trips beginning during these hours (Figure 6). A combined histogram including trips from all weekdays corroborated this result (Figure 7). A total of 39,744 (14.77%) trips began between 8:00 and 9:00 AM. 35,738 (13.27%) trips began between 5:00 and 6:00 PM on weekdays.



**Figure 6.** Histograms of the number of starting trips per hour for individual weekdays.

**Figure 7.** Histogram of the number of starting trips per hour for all weekdays combined.

**Busiest Stations During Rush Hours and Weekends**

During the morning rush hour (8 AM) on weekdays, the ten busiest starting stations in descending order were San Francisco Caltrain (Townsend at 4th) (ID: 70), Harry Bridges Plaza (Ferry Building) (ID: 50), San Francisco Caltrain 2 (330 Townsend) (ID: 69), Temporary Transbay Terminal (Howard at Beale) (ID: 55), Steuart at Market (ID: 74), Grant Avenue at Columbus Avenue (ID: 73), 2nd at Townsend (ID: 61), Embarcadero at Bryant (ID: 54), Civic Center BART (7th at Market) (ID: 72), and South Van Ness at Market (ID: 66) (Figure 5A).  Trips starting at these stations during this time represented 8.03% of all trips in the dataset.

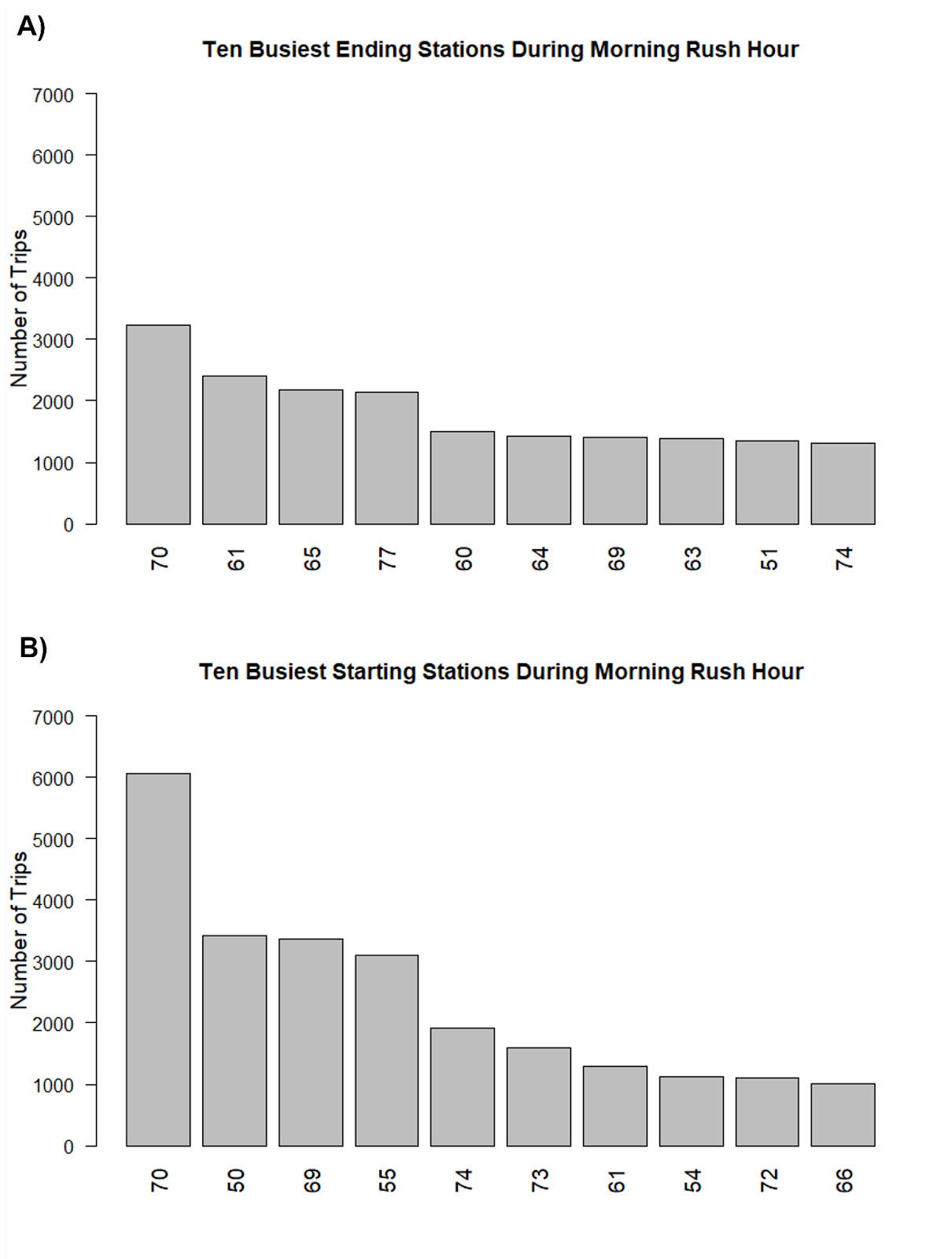
During the morning rush hour (8 AM) on weekdays, the ten busiest ending stations in descending order were San Francisco Caltrain (Townsend at 4th) (ID: 70), 2nd at Townsend (ID: 61), Townsend at 7th (ID: 65), Market at Sansome (ID: 77), Embarcadero at Sansome (ID: 60), 2nd at South Park (ID: 64),  San Francisco Caltrain 2 (330 Townsend) (ID: 69), Howard at 2nd (ID: 63), Embarcadero at Folsom (ID: 51), and Steuart at Market (ID: 74) (Figure 8B). Trips ending at these stations during this time represented 6.13% of all trips in the dataset.

During the evening rush hour (5 PM) on weekdays, the ten busiest starting stations in descending order were Townsend at 7th (ID: 65),  San Francisco Caltrain (Townsend at 4th) (ID: 70), 2nd at Townsend (ID: 61), Market at Sansome (ID: 77), 2nd at South Park (ID: 64),  Embarcadero at Sansome (ID: 60),  Steuart at Market (ID: 74), San Francisco Caltrain 2 (330 Townsend) (ID: 69), Embarcadero at Folsom (ID: 51), and Commercial at Montgomery (ID: 45) (Figure 9A). Trips starting during this time at these stations represented 5.00% of all trips in the dataset.

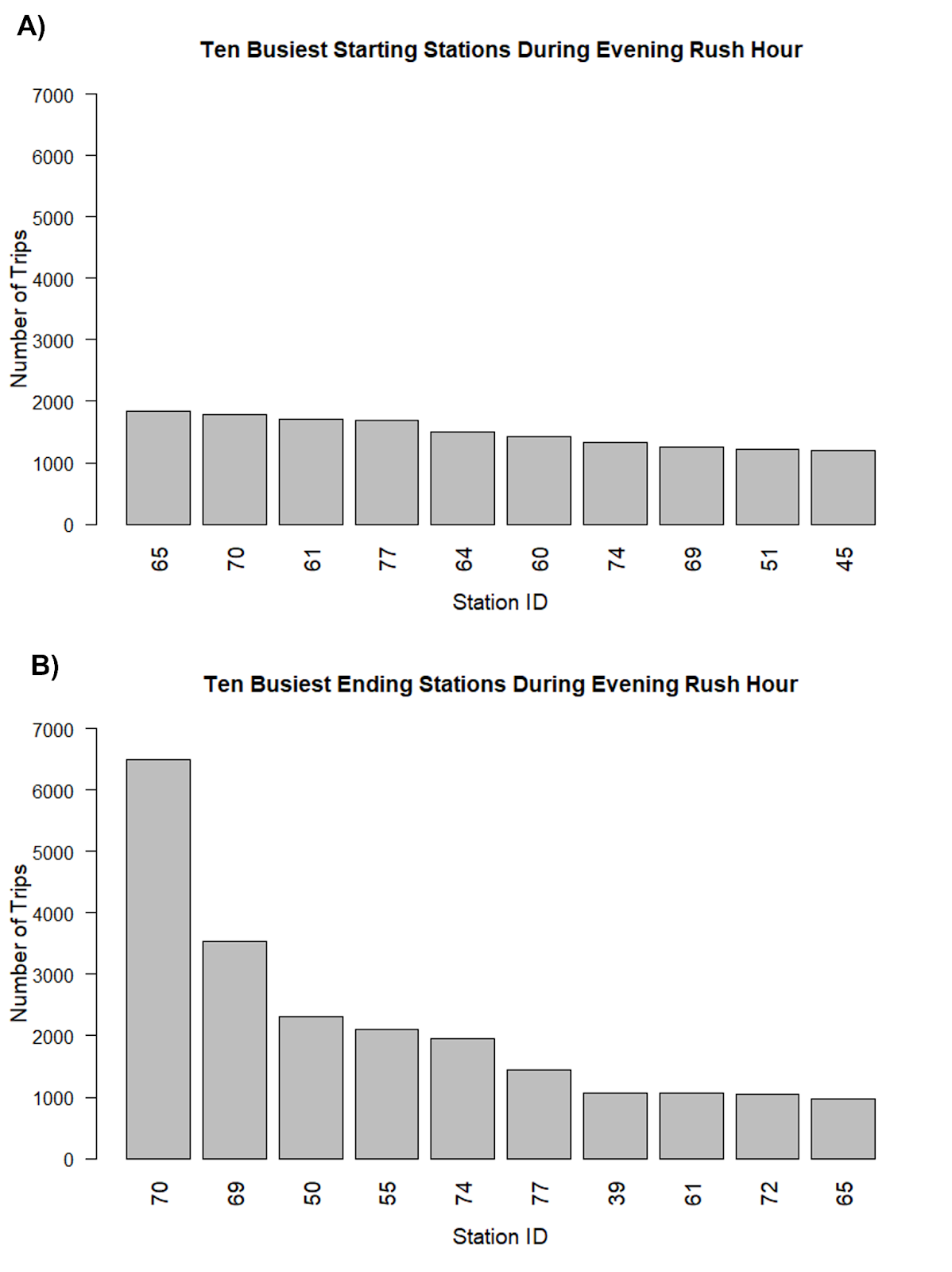
During the evening rush hour (5 PM) on weekdays, the ten busiest ending stations in descending order were San Francisco Caltrain (Townsend at 4th) (ID: 70), San Francisco Clatrain 2 (330 Townsend) (ID: 69),  Harry Bridges Plaza (Ferry Building) (ID: 50),  Temporary Transbay Terminal (Howard at Beale) (ID: 55), Steuart at Market (ID: 74), Market at Sansome (ID: 77),  Powell Street BART (ID: 39), 2nd at Townsend (ID: 61), Civic Center BART (7th at Market) (ID: 72), and Townsend at 7th (ID: 65) (Figure 9B). Trips ending during this period at these stations represented 7.37% of all trips in the dataset.

The ten busiest starting stations during the weekends in descending order were Embarcadero at Sansome (ID: 60), Harry Bridges Plaza (Ferry Building) (ID: 50), Market at 4th (ID: 76), 2nd at Townsend (ID: 61), Embercadero at BART ID (ID: 54), Powell Street BART (ID: 39), San Francisco Caltrain (Townsend at 4th) (ID: 70), Grant Avenue at Columbus Avenue (ID: 73), San Francisco Caltrain 2 (330 Townsend) (ID: 69), and Townsend at 7th (ID: 65) (Figure 10A). Trips starting on the weekend at these stations represented 4.28% of all trips in the dataset.

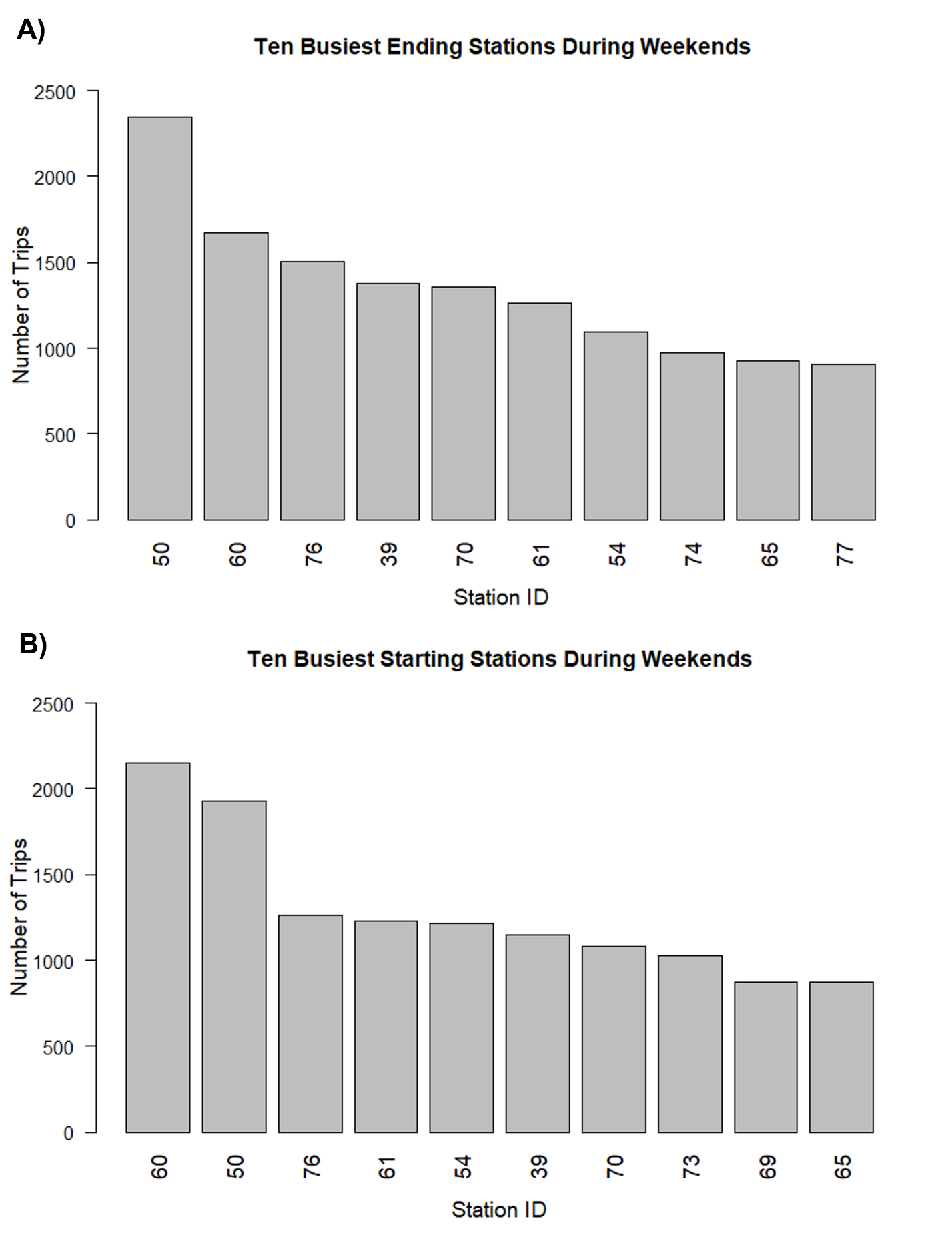
The ten busiest ending stations during the weekends in descending order were  Harry Bridges Plaza (Ferry Building) (ID: 50), Embarcadero at Sansome (ID: 60), Market at 4th (ID: 76), Powell Street BART (ID: 39), San Francisco Caltrain (Townsend at 4th) (ID: 70), 2nd at Townsend (ID: 61),  Embercadero at BART ID (ID: 54),  Steuart at Market (ID: 74), Townsend at 7th (ID: 65),  Market at Sansome (ID: 77) (Figure 10B). Trips starting on the weekend at these stations represented 4.49% of all trips in the dataset.

****

**Figure 8.** Barplots of ten busiest starting (A) and ending (B) stations during the morning rush hour.

****

**Figure 9.** Barplots of ten busiest starting (A) and ending (B) stations during the evening rush hour.

****

**Figure 10.** Barplots of ten busiest starting (A) and ending (B) stations during the weekends

**Average Utilization of Bikes for Each Month**

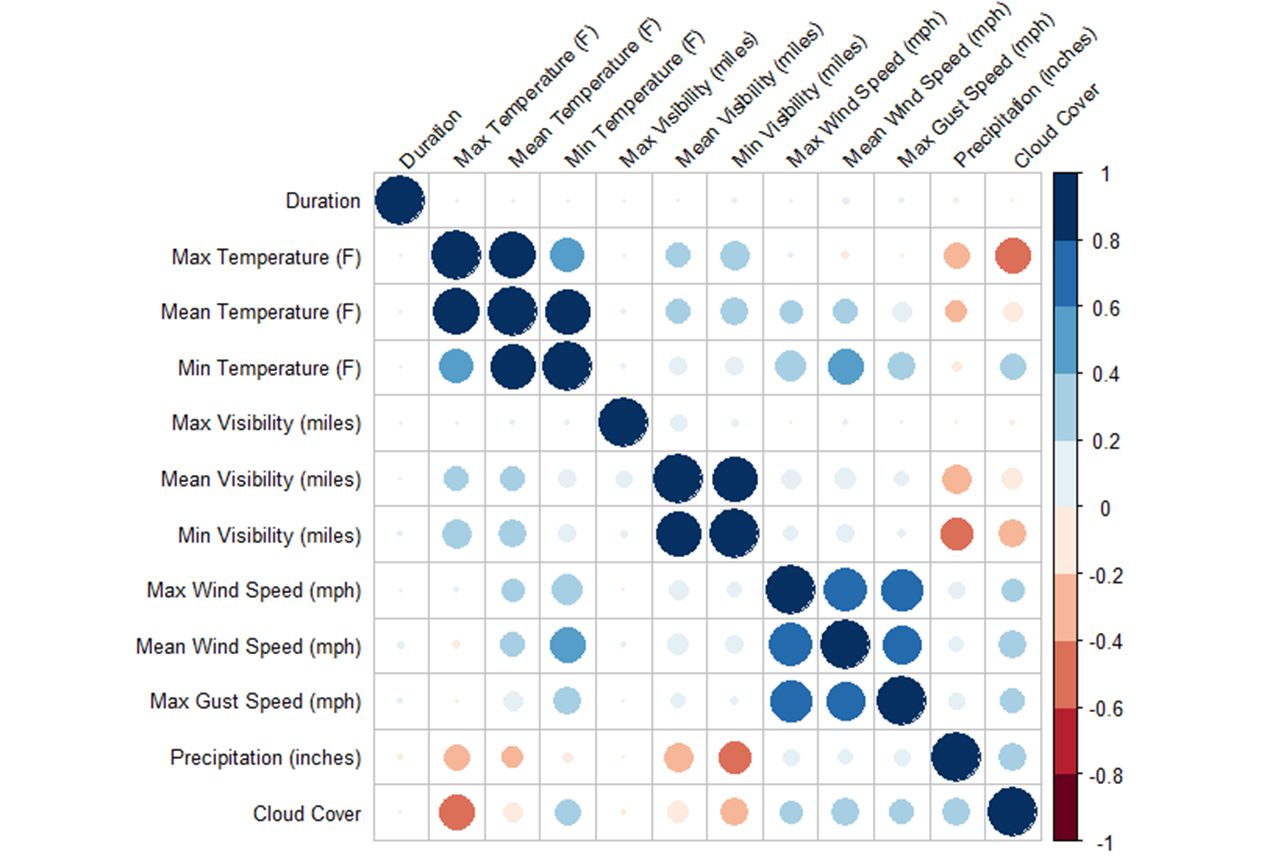
The average utilization of bikes for each month was calculated in respect to the start date, average duration per month, and the total time in each month (Table 1). The total duration for each month was summed and the total duration for each specific month was determined to perform the analysis.

|  |  |
| --- | --- |
| Month | Average Utilization (%) |
| January | 437 |
| February | 371 |
| March | 441 |
| April | 487 |
| May | 516 |
| June | 561 |
| July | 569 |
| August | 562 |
| September | 598 |
| October | 631 |
| November | 483 |
| December | 356 |

**Table 1**: Calculated monthly average utilization in percentage from the trip dataset for the year of 2014.

**Correlation between Different Weather Metrics and Trip Duration**

The weather metrics most closely correlated to trip duration were mean wind speed (*r* = 0.0279), maximum gust speed (*r* = 0.0164), and precipitation (*r* = -0.0156) (Figure 11). There was no discernible relationship between weather events and the number of trips as the number of trips per each weather event resembles the proportion of days for each weather seen in the events variable for each city. Maximum temperature and cloud cover had the strongest negative correlation between weather metrics that measure different characteristics of weather (temperature, visibility, wind speed, precipitation, cloudiness) (*r* = -0.4936) (Figure 8). Mean wind speed and minimum temperature had the strongest correlation between weather metrics that measure different characteristics of weather (*r* = 0.5250) (Figure 8).

****

**Figure 11**. Matrix of correlations between different weather metrics captured across different dates and cities within the weather dataset and duration of trips for the same date and location from the trips dataset.