**Bay Area Bike Rental Operation Research: Analysis of Bicycle Trip and Weather Patterns**

Olivia Paunkoff and Papiha Joharapurkar

University of Toronto Mississauga

MSC2011H: Special Topics in Biomedical Communications

Dr. Sebnem Kuzulugil

July 13, 2022

**Introduction**

This project features the thorough exploration of the San Francisco Bay Area Bike Share dataset. This dataset provides details on quick, easy and affordable bike trips around the San Francisco Bay Area. Our goal was to explore several questions regarding to bike-patterns and weather-patterns, however we mainly focused our efforts on developing a predictive model that would be able to predict the number of bikes leaving each station and being returned to each station.

**Exploratory Data Analysis for Trip Data**

*First approach to the data*

In the first approach to the “trip.csv” data set, the “glimpse” function was used to obtain an overview of the data. The data set includes a total of 326,339 observation with 11 variables including: id, duration, start\_date, start\_station\_name, start\_station\_id, end\_date, end\_station\_name, end\_station\_id, bike\_id, subscription\_type and zip\_code.

The “status” function was then used which indicated that “zip\_code” was the only variable with missing values. The zip code variable had 1493 (0.457% of total) missing values and 50 values equal to zero or 0.0153% of total. Since these values were less than 20% missing, they were not problematic and therefore were not removed. Moreover, the data types for each variable were obtained. All variables were set to the correct data type except for start and end dates which were coded as characters. Start and end dates will need to be changed to a date format at a later time. In addition, there were 70 unique start and end station ids, however, 74 unique start and end station names. This finding prompted a deeper analysis of the start and end station names to search for duplicates or misspelled station names. Using the function “unique” two misspelled station names were identified at first glance: "Washington at Kearny" and “Post at Kearny" with the correct spelling being “Kearney”. To fix this issue, the incorrectly spelled stations names were recoded with the correct spelling. Upon a closer examination of the data set, it was observed that “San Jose Government Center" and "Santa Clara County Civic Center" are both coded under station ID 25. In addition, "Broadway at Main” and "Stanford in Redwood City" were both coded under station ID 80. As "Broadway at Main” and “San Jose Government Center" stations are not included in the “station.csv” data set, they were recoded to "Stanford in Redwood City" and "Santa Clara County Civic Center" by their respective corresponding station IDs as it was assumed the names were entered incorrectly by error. The “status” function was then used again to confirm there were now 70 unique stations IDs and 70 unique station (starting and ending) names.

*Analyzing categorical variables*

Analysis of categorical variables using the “freq” function from the “funModelling” R package examined the frequency in which starting and ending stations were seen in trip indicated the top 10 most frequent start stations in data to be: San Francisco Caltrain (Townsend at 4th), Harry Bridges Plaza (Ferry Building), San Francisco Caltrain 2 (330 Townsend), 2nd at Townsend, Embarcadero at Sansome, Temporary Transbay Terminal (Howard at Beale), Market at Sansome, Steuart at Market, Townsend at 7th, and Market at 4th **(Table 1)** The top 10 ending stations were similar: San Francisco Caltrain (Townsend at 4th), Harry Bridges Plaza (Ferry Building), San Francisco Caltrain 2 (330 Townsend), Market at Sansome, 2nd at Townsend, Embarcadero at Sansome, Townsend at 7th, Steuart at Market, Temporary Transbay Terminal (Howard at Beale), and Powell Street BART **(Table 2)**.

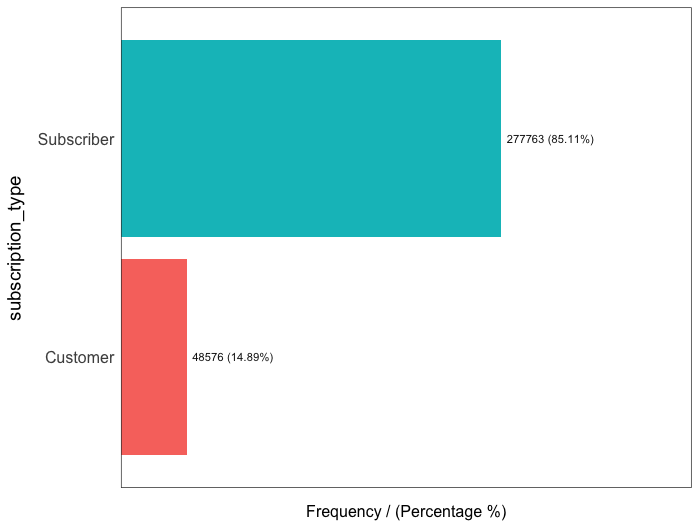
|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Start Station** | **Frequency of Rides** | **%** |
| 1 | San Francisco Caltrain (Townsend at 4th) | 25144 | 7.70 |
| 2 | Harry Bridges Plaza (Ferry Building) | 15536 | 4.76 |
| 3 | San Francisco Caltrain 2 (330 Townsend) | 15132 | 4.64 |
| 4 | 2nd at Townsend | 12935 | 3.96 |
| 5 | Embarcadero at Sansome | 12879 | 3.96 |
| 6 | Temporary Transbay Terminal (Howard at Beale) | 12748 | 3.91 |
| 7 | Market at Sansome | 12518 | 3.84 |
| 8 | Steuart at Market | 11579 | 3.55 |
| 9 | Townsend at 7th | 11002 | 3.37 |
| 10 | Market at 10th | 9980 | 3.00 |

**Table 1:** *Ten most frequent starting stations overall*. Number of total rides left from each station with the percent of total ride left from each station.

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **End Station** | **Frequency of Rides** | **%** |
| 1 | San Francisco Caltrain (Townsend at 4th) | 33213 | 10.18 |
| 2 | Harry Bridges Plaza (Ferry Building) | 15692 | 4.81 |
| 3 | San Francisco Caltrain 2 (330 Townsend) | 15333 | 4.70 |
| 4 | Market at Sansome | 14816 | 4.54 |
| 5 | 2nd at Townsend | 14064 | 4.31 |
| 6 | Embarcadero at Sansome | 13449 | 4.12 |
| 7 | Townsend at 7th | 12225 | 3.75 |
| 8 | Steuart at Market | 11713 | 3.59 |
| 9 | Temporary Transbay Terminal (Howard at Beale) | 11464 | 3.51 |
| 10 | Powell Street BART | 10232 | 3.14 |

**Table 2:** *Ten most frequent ending stations overall*. Number of total rides arrived at each station with the percent of total ride arrived at each station.

In addition, analysis of the subscription type found that 85.11% of rides are by “Subscribers” and 14.89% of rides are by “Customers”, those without a subscription **(Figure 1)**.

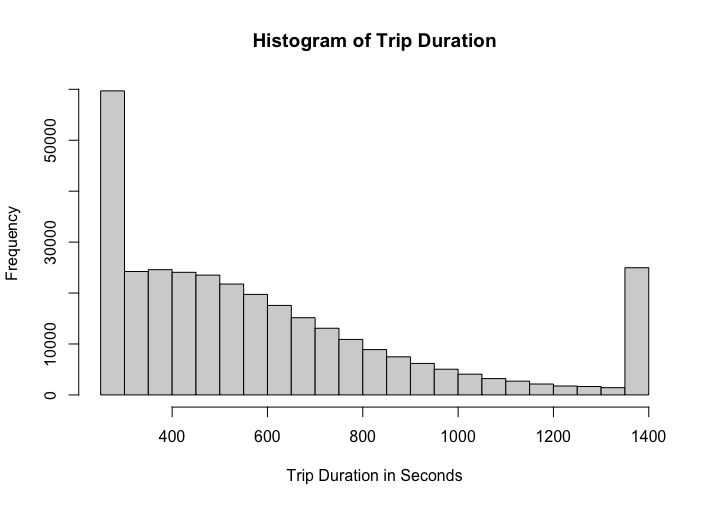


**Figure 1:** *Subscription types.*

*Analyzing numerical variables and addressing outliers*

Analysis of numerical variables using the “describe” function from the “Hmisc” R package was done to obtain information about the trip durations. The mean value of trip duration was of trips in the dataset was 1131.967 sec, with the five highest durations being 644 771, 715 339, 716 480, 720 454, and 1 727 040 seconds long and the five lowest durations all being approximately 60 seconds or less. Here, trips that were less than 2 minutes (120 seconds long) were removed from the data set using the filter function from the “tidyverse” package; about 2,499 observations were removed from the dataset after completing this step.

Outliers were identified using the equation 1.5\*IQR + Q3 for identifying the upper limits and 1.5\*IQR - Q1 for identifying the lower limits. There were approximately 65,195 outliers if the upper and lower limits were to be removed. To prevent loss of a massive number of data points the outliers were identified and reassigned to be the upper and lower limits as opposed to removing them entirely. A histogram of duration trip frequency was created to examine the distribution and to check for any extreme skewedness **(Figure 2)**. Using the “describe” function from the “ Hmisc” package a final check of the data was done to ensure outliers had been addressed. After modifying outliers to be either upper or lower limits, the new mean of trip duration was 598.9 seconds. The modified trip data set was named “trip\_clean2” and saved as a RDS file for later use.



**Figure 2:** *Histogram of trip durations in seconds.* Note frequencies at each end of the range are large due to the reassignment of outliers to upper and lower limits.

**Exploratory Data Analysis for Weather Data**

*First approach to the data*

To initiate our exploration of the weather.csv file, the “glimpse” function from the dplyr package was used to obtain an overview of all the variables from the weather dataset. In this primary step, the following details were noted:

The date column was formatted in character-form instead of as date form, which necessitated a modification to the latter form. In addition, the max\_temperature\_f, mean\_temperature\_f and min\_temperature\_f variables, along with the max\_visibility\_miles, mean\_visibility\_miles, min\_visibility\_miles variables, and the max\_wind\_Speed\_mph, mean\_wind\_speed\_mph, max\_gust\_speed\_mph variables were provided in integer form. These would need to be coded to numeric form for future analyses, such as correlation plots. Moreover, the precipitation (inches) was incorrectly coded as characters instead of integers, due to the fact that there were a few values coded as “T”. This value encodes “trace” for trace-amounts of precipitation, and would need to be changed to 0. The cloud-cover variable seems to be coded as an integer but could be coded as a factor as there are different categories of cloud-cover. In addition, the events variable seem to be properly coded as characters, however the output also indicated that there are some variables coded as “ “ which would need to be changed to NA. The zipcode variable is also correctly coded as an integer, but should be coded as a factor. Based on this initial examination, the variables were recoded and reformatted before performing any of the future steps of the EDA.

*Analyzing categorical variables*

The next approach to the exploratory analysis involved performing an assessment of the categorical data through the “freq” function. There were only a few categorical variables assessed, these encompassed the cloud\_cover, events, zip\_code and city variables and the most relevant plots are shown here. The output revealed that there were 9 levels of the cloud\_cover variable, ranging in frequency with 0-cloud cover being the most frequent and 8-cloud\_cover category as being the least frequent as shown in Figure 3A. In addition, the events variable indicated that 80.71% (representing 1473 data values) of the data was composed of NA-values, with the next most frequent events composing of rain and fog. Lastly, there were 5 levels associated with the cities, related to the 5 levels of zipcodes, which are evenly distributed in frequency, as shown by Figure 3B.

Timeline

Description automatically generatedChart, bar chart

Description automatically generated

**Figure 3***: Assessment of Categorical Variables.* A) Frequency of Weather Events (Left) B) Frequency of Dates for Each City (Right)

*Analyzing numerical variables*

After assessing the categorical variables, the variables remaining were numerical in form, and were thus explored through the plot\_num function. For this function, barplots were provided for all the numerical variables, which helped us develop initial insights into each of variables as can be seen in Figure 4. The minimum-visibility (miles), maximum-gust-speed (mph) variables, along with max\_gust\_speed\_mph and precipitation\_inches variables show relatively skewed and unbalanced distributions which can be attributed to the presence of outliers that would need to be removed, supporting previous analyses. The remaining numerical variables seem to have a relatively normal distribution, without any discerning outliers.

Chart

Description automatically generated

**Figure 4**: *Assessment of Numerical Variables*. The minimum-visibility (miles), maximum-gust-speed (mph), max\_gust\_speed\_mph and precipitation\_inches variables display skewed-distributions.

After plotting the numerical variables, the profiling-num function was performed to obtain an understanding of each variable’s distribution and range. This function demonstrated that there is a high standard deviation for the max\_temperature\_f, mean\_temperature\_f, min\_temperature\_f, max\_wind\_Speed\_mph and max\_gust\_speed\_mph variables. In addition, the variation coefficient demonstrated that the precipitation\_inches variable is associated with more variance as opposed to the remaining variables that are centered around the mean, inferring less accuracy associated with this variable if it were to be chosen as a possible predictor. The skewness metric informs on the measure of asymmetry, and this function demonstrated that the variables of max\_visibility\_miles, mean\_visibility\_miles, max\_wind\_Speed\_mph, max\_gust\_speed\_mph and precipitation\_inches variables have greater skewness-values and therefore greater likelihood of observing outliers in these variables. The kurtosis variable describes the distribution for the tails, where higher number indicates presence of outliers, and this was demonstrated by the higher values shown for the max\_visibility\_miles, mean\_visibility\_miles, max\_wind\_Speed\_mph, max\_gust\_speed\_mph and precipitation\_inches variables. There was also a high IQR range for the max\_temperature\_f, mean\_temperature\_f, and min\_temperature\_f variables obtained.

The describe function was used to obtain an overall overview of both the numerical and categorical variables. The date-column has 144 unique values but 720 dates available, with the remaining 1105 values being missing values and representing 60.5% of the date-data. These values range from 2020-01-10 to 2020-12-12. Since this date-data is being used for combining with the trip-data based on date, values with missing dates will be removed. The max\_temperature\_f column has 46 unique values, ranging from 50 to 102. For the mean\_temperature\_f variable, there are 37 distinct values ranging from 44 to 84, and similarly, there are 37 distinct values for the min\_temperature\_f variable, ranging from 32 to 69. The max\_visibility\_miles has 9 missing values, and 9 distinct values ranging from 5 to 20. This variable was also shown to have many outliers, which will need to be removed. The related column of mean\_visibility\_miles has 9 missing values, and 17 distinct values ranging from 4 to 20, the higher values representing outlier-values will also will be removed. The min\_visibility\_miles variable has 9 missing values and 17 distinct values ranging from 4 to 20, but this column does not have any outliers as shown by previous analyses. The max\_wind\_Speed\_mph variable has no missing values, but 34 distinct values ranging from 4 to 122, and this variable has been noted to contain outliers, which will be removed. The mean\_wind\_speed\_mph variable has no missing values, but 20 distinct values ranging from 0 to 19, but no presence of outliers. The max\_gust\_speed\_mph variable has 451 missing values, and 43 distinct-events, ranging from 6 to 114, and previous analyses indicate this variable has outliers which will be removed. For the precipitation inches variable, there are 74 unique values ranging from 0 to 3.36, but this variable has been noted to contain outlier values, which will be removed. For the cloud-cover variable, there are no missing values and 9 distinct events ranging from 0 to 8, with no discernable outliers according to previous analyses. For the event variable, there are 5 distinct-events however with 1473 missing values. For the zip-code variable, there are 0 missing values but 5 distinct events. This variable is also similar to the output from the city-variable, which retrieved 0 missing values but 5 distinct events, with both of these variables retrieving equivalent proportions for the frequency of each of the values. For the variables with missing values which are the date, max\_visibility\_miles, mean\_visibility\_miles, min\_visibility\_miles, max\_gust\_speed\_mph and events variables, the missing-values were not recoded to 0 or other values such as mean/minimum/maximum as these entries instead represent events where values were simply not observed and recorded, and it would be incorrect to impute values in place of them.

*Addressing outliers*

The previous exploratory analyses were helpful in uncovering the variables that necessitated outlier-removal in order to ensure high accuracy of future results. Firstly, we initially removed all the available data were date-information was not available for reasons explained previously. The method of removing outliers was consistent for many of the same variables from the trip-dataset, with some exceptions, where outliers were identified using the equation 1.5\*IQR + Q3 for identifying the upper limits and 1.5\*IQR - Q1 for identifying the lower limits. To start, the IQR quantile and IQR functions were applied on the variable of max\_visibility\_miles, but since identical values for both the upper limit and lower limit were retrieved, therefore instead, another method of outlier recoding was used. This method that was used was the percentile-method, where data values were recoded to the lower and upper limits if observation values were beyond the 2.5% or 97.5% percentile respectively. This alternative approach allowed for the successful recoding of 2 lower-limit outliers and 0 for upper limit of max\_visibility\_miles. Similarly for the mean\_visibility\_miles variable, identical values for the upper limit and lower limit were retrieved, and the percentile method of outlier removal was performed. This allowed for the recoding of 9 lower-limit values and 12 higher-limit values. The determination of disparate values for the upper and lower limit for the variable of max\_wind\_Speed\_mph was also successful, allowing for those outlier-values to be recoded from the dataset based on the IQR method. There were 2 lower-limit outliers recoded and 23 higher-limit outliers recoded. The IQR method of outlier removal was also successful for the max\_gust\_speed\_mph variable allowing for the recoding of 0 lower-limit but 18 higher limit outliers. Finally, the method of outlier removal based on percentiles was used for the variable of precipitation\_inches, allowing for 0 lower-limit but 18 higher-limit outliers to be recoded.

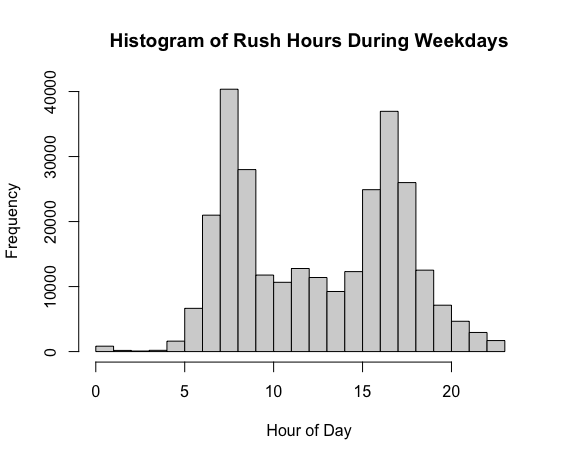
**Outliers for Station**

Alongside the weather and data-files, the station file was also examined. With the structure function, the station file was found to provide many insights. To note, the id-variables were correctly coded as numeric and station name and city variables were correctly coded in character form. Additionally, the latitude and longitude variables were also correctly coded as numeric alongside the dock-count variable. However, the installation-dates were coded as characters. In order to identify any outliers in association to the main trip-file, the number of unique rows was determined, which was retrieved to be 70. After this, the number of unique stations was determined within the trip-file was determined by using the end-station-name, although this could have been replaced with start-station-name. This resulted in the total output of 74 unique values. This value’s discrepancy with the value from the station file suggested that there were 4 additional values that had been improperly coded in the trip-data file. In order to determine which station-names had been improperly coded in the trip-data, an anti-join function was applied with the trip-data and the station-file data to determine which station-names did not match with the station-data. This function found that the San Jose Government Center, Broadway at Main, Washington at Kearny and Post at Kearny values had not been provided in the station-data and must have been therefore, been improperly coded in the trip-file data.

**Rush-Hours and 10-Most Frequent Starting/Ending Stations**

*Establishing the Highest Volume Hours on Weekdays*

The modified “trip\_clean2.rds” was imported and will be the main analysis on the trip data set moving forward. In order to establish highest volume hours on weekdays, the start and end trip dates first had to be recognizable as dates. Using the “as.POSIXct” function the dates were converted from characters to dates. A new data frame called “rush\_hours” was created, and using the “tidyverse” package, weekends, which were assigned value of 1 (for Sunday) and 7 (for Saturday) using the “wday” function, were filtered out from the starting dates in the data frame. A new column of hours, corresponding to the time in the starting dates column was created and data was grouped by starting hour. Through closer investigation by the “table” function and by using a histogram to model the data, the hours of 8:00 - 8:59am, 5:00 - 5:59pm, and 9:00 - 9:59am were the most popular times to start trips on weekdays (**Figure 5)**. Starting times were chosen to represent rush hours as this would be the times that customers were picking up bikes and would therefore be the times when the shop is most busy with customers.



***Figure 5:*** *Histogram representing the most frequent times to start trips.* Hours of 8:00 - 8:59 am, 5:00 - 5:59pm, and 9:00 - 9:59 am are the most frequent times for riders to start their trip on weekdays.

Rush hours of 8:00 - 8:59 am, 5:00 - 5:59pm, and 9:00 - 9:59 am on weekdays are expected values as they closely align with the workday patterns of the average 9:00am to 5:00pm worker. In addition, the least common times to start a trip are between the hours of 1:00am - 4:00 am, which are hours when most people tend to be asleep.

*Determining the 10 Most Frequent Starting and Ending Stations on Weekdays*

In order to determine the 10 most frequent starting stations on weekdays, the same steps were followed as to establish the rush hours for ending stations. The data frame was then grouped by starting station name and then only trips leaving at hours between 8:00 - 8:59 am, 9:00 - 9:59 am and 5:00 - 5:59 pm were included in the analysis. The starting stations and their corresponding frequencies were then put in table format and ordered from most to least frequent in terms of number of trips left during the previously specified hours. The top 10 starting stations at rush hours were: San Francisco Caltrain (Townsend at 4th), San Francisco Caltrain 2 (330 Townsend), Temporary Transbay Terminal (Howard at Beale), Harry Bridges Plaza (Ferry Building), Steuart at Market, 2nd at Townsend, Market at Sansome, Townsend at 7th, Beale at Market, and Market at 10th **(Table 3)**.

|  |  |
| --- | --- |
| **Start Station** | **Frequency of Rides** |
| San Francisco Caltrain (Townsend at 4th) | 10977 |
| San Francisco Caltrain 2 (330 Townsend) | 6966 |
| Temporary Transbay Terminal (Howard at Beale) | 5816 |
| Harry Bridges Plaza (Ferry Building) | 5512 |
| Steuart at Market | 4647 |
| 2nd at Townsend | 4321 |
| Market at Sansome | 3958 |
| Townsend at 7th | 3236 |
| Beale at Market | 3093 |
| Market at 10th | 3056 |

**Table 3:** *Ten most frequent starting stations on weekdays*. Frequency of rides represents the numbers of trips left at each starting station during rush hours on weekdays.

To determine the 10 most frequent ending stations during rush hours on weekdays the same process to determine rush hours for starting stations was followed. This analysis returned the same rush hours at the ending stations as the starting stations being: 8:00 - 8:59 am, 9:00 - 9:59 am and 5:00 - 5:59 pm. The new data frame for frequencies of ending stations during weekdays was then filtered to only include rush hours and then made into a table, using the “table” function and ordered from most to least frequent. In order, the top 10 most frequent ending stations during rush hours on weekdays were: San Francisco Caltrain (Townsend at 4th), San Francisco Caltrain 2 (330 Townsend), Market at Sansome, Harry Bridges Plaza (Ferry Building), 2nd at Townsend, Townsend at 7th, Temporary Transbay Terminal (Howard at Beale), Steuart at Market, Embarcadero at Sansome, and Powell Street BART **(Table 4)**.

|  |  |
| --- | --- |
| **End Station** | **Frequency of Rides** |
| San Francisco Caltrain (Townsend at 4th) | 31506 |
| San Francisco Caltrain 2 (330 Townsend) | 14295 |
| Market at Sansome | 13631 |
| Harry Bridges Plaza (Ferry Building) | 12439 |
| 2nd at Townsend | 12365 |
| Townsend at 7th | 11139 |
| Temporary Transbay Terminal (Howard at Beale) | 10666 |
| Steuart at Market | 10418 |
| Embarcadero at Sansome | 10046 |
| Powell Street BART | 8512 |

**Table 4:** *Ten most frequent ending stations on weekdays*. Frequency of rides represents the numbers of trips arriving at each ending station during rush hours on weekdays.

Apart from a few discrepancies, the most frequent starting and ending stations on weekdays appeared to mostly align with one another, which is to be expected. Specifically, the ending stations that were not included in the top 10 starting stations were Powell Street BART and Embarcadero at Sansome. Conversely, the two starting stations that were not included in the top 10 ending stations were Beale at Market and Market at 10th **(Table 3-4)**.

*Determining the 10 Most Frequent Starting and Ending Stations on Weekends*

The process for determining the most frequent stations on weekends was largely like the previous analysis except data was filtered to remove weekdays instead of weekends. A new data frame called “rush\_hours\_wkend” was created, and using the “tidyverse” package, weekdays, which are assigned values of 1 (Monday) to 6 (Friday) were filtered out. The 10 most frequent starting stations on weekends were: Harry Bridges Plaza (Ferry Building), Embarcadero at Sansome, Market at 4th, Embarcadero at Bryant, 2nd at Townsend, Powell Street BART, San Francisco Caltrain (Townsend at 4th), Grant Avenue at Columbus Avenue, Market at Sansome, and Powell at Post (Union Square) **(Table 5)**.

|  |  |
| --- | --- |
| **Start Station** | **Frequency of Rides** |
| Harry Bridges Plaza (Ferry Building) | 3169 |
| Embarcadero at Sansome | 3128 |
| Market at 4th | 1661 |
| Embarcadero at Bryant | 1591 |
| 2nd at Townsend | 1546 |
| Powell Street BART | 1488 |
| San Francisco Caltrain (Townsend at 4th) | 1365 |
| Grant Avenue at Columbus Avenue | 1299 |
| Market at Sansome | 1091 |
| Powell at Post (Union Square) | 1091 |

**Table 5:** *Ten most frequent starting stations on weekend*. Frequency of rides represents the numbers of trips beginning at each starting station during weekends.

The analysis of the most frequent end stations during weekends was done in the same way as the previous analysis, however, with the replacement of starting station to ending stations data from “trip\_clean2” data set. Moreover, the results indicated a large portion of most frequent stations to be very similar to the starting station analysis, however, with slightly different ordering: Harry Bridges Plaza (Ferry Building), Embarcadero at Sansome, Market at 4th, Embarcadero at Bryant, 2nd at Townsend, Powell Street BART, San Francisco Caltrain (Townsend at 4th), Grant Avenue at Columbus Avenue, Market at Sansome, and Powell at Post (Union Square) **(Table 6)**.

|  |  |
| --- | --- |
| **End Station** | **Frequency of Rides** |
| Embarcadero at Sansome | 3377 |
| Harry Bridges Plaza (Ferry Building) | 3182 |
| Market at 4th | 1876 |
| Powell Street BART | 1680 |
| San Francisco Caltrain (Townsend at 4th) | 1655 |
| 2nd at Townsend | 1586 |
| Embarcadero at Bryant | 1361 |
| Steuart at Market | 1215 |
| Market at Sansome | 1118 |
| Grant Avenue at Columbus Avenue | 1107 |

**Table 6:** *Ten most frequent ending stations on weekend*. Frequency of rides represents the numbers of trips finishing at each end station during on weekends.

Overall, the most frequent starting/ending stations on weekends and weekdays appeared to closely align with one another. Interestingly, there appears to be a slight discrepancy in stations that are more popular on weekends versus weekdays. For example, some of the top 10 starting stations for weekends included Grant Avenue at Columbus Avenue, Powell at Post (Union Square), Embarcadero at Bryant, Embarcadero at Sansome, and Market at 4th all of which were not included in the top 10 stations starting list on weekdays. For ending stations, top 10 that were not in the ending station analysis for weekends included: Market at 4th, Embarcadero at Bryant, and Grant Avenue at Columbus Avenue. It is worth noting that San Francisco Caltrain (Townsend at 4th) and 2nd at Townsend stations-maintained popularity throughout each analysis.

**Average Utilization of Bikes**

In order to calculate the average utilization of bikes overall per month, a new data frame called “bike\_utl” from the “trip\_clean2” data frame was created. Data was first grouped by month and then an additional data frame was created and using the “aggregate” function to sum the total trip durations corresponding to each month. An additional column for the number of days in each month was added as well and a column representing the total seconds in each month which multiplies the days in the month by the average seconds in a day (86400 seconds). The total bike utilization per month was calculated by adding a column that performs a division of total seconds used per month by the total seconds in the given month. Utilization results for each month can be seen in Table 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Month** | **Total duration of rides in seconds** | **Days per month** | **Seconds per month** | **Average utilization per month** |
| January | 14 126 404 | 31 | 2678400 | 5.274195 |
| February | 10 840 342 | 28 | 2419200 | 4.480961 |
| March | 14 884 029 | 31 | 2678400 | 5.557060 |
| April | 15 603 402 | 30 | 2592000 | 6.019831 |
| May | 17 384 746 | 31 | 2678400 | 6.490720 |
| June | 17 933 709 | 30 | 2592000 | 6.918869 |
| July | 19 085 815 | 31 | 2678400 | 7.125827 |
| August | 19 051 250 | 31 | 2678400 | 7.112922 |
| September | 18 706 073 | 30 | 2592000 | 7.216849 |
| October | 20 160 155 | 31 | 2678400 | 7.526940 |
| November | 14 722 858 | 30 | 2592000 | 5.680115 |
| December | 11 440 882 | 31 | 2678400 | 4.271536 |

**Table 7:** *Total bike utilization per month.* Average utilization was calculated by dividing total duration of trips per month in seconds divided by total time in month in seconds.

**Weather and Bike Rental Patterns**

To determine the impact of weather patterns on bike-rental patterns, the cleaned and filtered data from the previous steps were retrieved. A dataframe with both the trip-data and station-data was then made, using the left\_join function. This was performed in order to obtain information regarding the city of each station, which were necessary details since the weather-dataset contained weather patterns based on each city and zipcode, but not according to station-names. After creating this merged dataframe, the variables of id (id.x corresponding to the ID from the trip-dataset), bike-ids, station and end-station ids, subscription-type and zipcodes which were originally from the trip-dataframe were removed and the variables of id (id.y corresponding to the ID from the station-data), installation-date, start-station name, end-station-name, dock-count, long and lat columns were also removed. These variables were removed as they were irrelevant to the analysis of bike-rental frequency and weather-patterns. In order to then obtain the bike-rental frequency, the data was grouped by start-date and city to obtain the total number of counts of bikes used per day per city, as opposed to variables such as duration. It was important to stratify the data by city as well as date, as the weather-data was divided by city, not only date. Therefore, the next step involved joining this newly created dataframe containing the start-date, city, and number of trips per day information with the weather-dataframe. The unique-identifiers in this case were the variables of start-date/date and city. In order to perform correlation analysis with just the numerical variables, all character or factor variables such as start\_date, events, and city character-variables were moved. Then, correlation analysis was performed for each of the cities. The events-variable was removed as it was a factor/character-value, and therefore was not included in the quantitative-correlation analysis. For the following correlation plots, freq refers to the number of trips completed per day.

For San Francisco as can be seen in Figure 6, a positive but weak correlation can be observed for the variables of max-temperature-f and mean-temperature, and a negative moderate correlation for precipitation inches, as well as a weaker negative correlation for maximum and minimum wing speed, and cloud-cover.

Chart, bubble chart

Description automatically generated

**Figure 7**: *Correlation Plot for San Francisco*.

For the city of Mountain View as can be seen in Figure 8, a positive but weak correlation can be seen for the variables of max-temperature-f and mean-temperature and weak negative correlations for the variables of max/mean wind speed, and gust speed. The variables of cloud-cover and precipitation inches demonstrate a marginally better, borderline moderate negative correlation.

Chart, bubble chart

Description automatically generated

**Figure 8**: *Correlation Plot for Mountain View*.

The city of Palo Alto similarly showed a weak positive correlation for max and mean temperature, as well as for minimum visibility in miles, and a weak negative correlation for minimum visibility and precipitation inches, as well as for events-types and gust-speed. However, a moderate negative correlation for maximum and average wind speed, as well as cloud cover. Interestingly, there were not any variables that were insufficient for correlation analysis, seen by the lack of ‘?’ compared to the other plots for the other cities as seen in Figure 9.

Chart

Description automatically generated

**Figure 9**: *Correlation Plot for Palo Alto*

As seen in Figure 10, which displays the correlation plot for Redwood City, there is only weak positive correlation shown for all variables, with the exception of average wind speed, which shows a borderline-moderate correlation.

Chart, scatter chart, bubble chart

Description automatically generated

**Figure 10**: *Correlation Plot for Redwood City*

Chart, bubble chart

Description automatically generated

**Figure 11:** *Correlation Plot for San Jose*

For the town of San Jose as shown in Figure 11, there are weak positive correlations shown for the mean and minimum visibility variables, with borderline moderate correlation for the max/mean/min temperature variables. Moreover, a weak negative correlation for max/mean wind speeds and gust-speeds with cloud cover, and for precipitation inches is also seen.

To provide a summary of the weather patterns for all these plots, all the cities demonstrate a weak-positive correlation associated with the max/min/mean temperature variables where increased frequency of bike rides are associated with increased temperature. In addition, this correlation was also shown to be true for the max/minimum/mean visibility (miles) variables, where increased visibility is associated with increased frequency of bike rides. On the other hand, weak-negative correlations are demonstrated with the speed variables, where increasing speed is associated with lower frequency of bike rides. In addition to this, weak-negative correlations were exhibited for precipitation, where increased precipitation is associated with lower frequency of bike rides. Similar to this variable, greater-cloud-cover values being associated with lower frequency of bike rides. These trends were however not noticed in Redwood City, where increased frequency of bike rides were associated not only with greater temperatures but also, negative weather events such as greater wind-speed and greater precipitation or cloud-cover.

**Conclusion**

The analysis concluded that the common rush hours during weekdays were 8:00 - 8:59 am, 9:00 - 9:59 am and 5:00 - 5:59pm. These results were consistent for both starting and ending stations and likely align with workday patterns of the average worker in the Bay Area. Overall, the bike starting and ending stations on weekdays aligned well. There were discrepancies in the most popular stations for weekends in comparison to weekdays. Specifically, the Market at 4th, Embarcadero at Bryant, and Grant Avenue at Columbus Avenue stations seemed to be busier on the weekends in comparison to weekdays. Interestingly, San Francisco Caltrain (Townsend at 4th) and 2nd at Townsend stations-maintained popularity throughout each analysis. In addition, an increase in utilization of bikes is seen in the months of July – October. Conversely, a decrease in utilization was seen in December, January, and February. When it comes to the expected weather patterns, weather events thought to be associated with the months of July – October include increased temperatures and greater visibility, along with lower wind-speeds and lower precipitation with the converse expected to present in December, January, and February. The analysis of weather patterns also supports the previous results, where greater frequency of bike rides were shown to be associated with positive or increasing values for temperatures and greater visibility as opposed to lower frequency with greater wind-speeds and precipitation values. However, due to the limitations of missing data which necessitated missing-value imputation for the weather and trip-information, further analysis to determine the impact of weather patterns on bicycle utilization could be helpful to inform maintenance decisions.