San Francisco Bike Usage Report

August 6, 2024

Kimberly Corneil

BTC1855H

Dr. Sebnem Kuzulugil

**Introduction to the Data: Exploratory Data Analysis**

For the analysis of the Bay Area Bike Share's operation in the year 2014, three files were provided, referred to from here on out as “trip”, “station”, and “weather”. All figures in the Exploratory Data Analysis (Figures 1-12) were generated using functions in the package “funModelling” in R.

Exploratory Data Analysis of Trip

Trip is a data frame containing information about 326339 bike trips across 11 categories of information. These categories include the trip’s ID, the duration of the trip, the date the trip started on, the date the trip ended on, the name (and ID in another column) of the station that the trip started on, the name (and ID in another column) of the station that the trip ended on, the ID of the bike used, the subscription type the user had, and the zip code. The data was provided as either a character or an integer.

Character Variables

Figures 1 and 2 depict the frequency distribution of stations from which the bike trips started and ended on, respectively. The top three starting and ending stations are the same, however variation occurs afterwards. There were no observed missing values for either variable.

Figure 3 shows that the majority (85.11%) of bike users are subscribers, while the rest are customers and there are no missing values.

The other character variables include start date, end date, and zip code. There are no missing values for the start or end dates. Zip code contains missing data (1480 missing entries and 50 0’s). Start and end date are currently characters but will need to be coded as dates and separated by their respective times for further analysis.

A graph showing a number of cities

Description automatically generated

**Figure 1:** Frequency plot showing the percentage frequency of the various start station names on the X-axis and the names of the starting stations on the Y-axis.

A graph showing a number of cities

Description automatically generated

**Figure 2:** Frequency plot showing the percentage frequency of the various end station names on the X-axis and the names of the ending stations on the Y-axis.

A screenshot of a graph

Description automatically generated

**Figure 3:** Frequency plot showing the percentage frequency of the two subscription types available to ride the bikes with the subscription type on the Y-axis and percentage frequency on the X-axis.

*Numeric Variables*

Figure 4 shows histograms generated for all numeric variables in “trip”. ID is a trip identifier, and each observation takes on a unique value, and as such viewing the distribution of ID is not an important metric for understanding the data. Duration describes the length of each bike trip (assumed to be in seconds, although no unit of measurement was given). The histogram for duration appears to be greatly right skewed. This distribution suggests that most of the values in duration are smaller values or shorter trips, while there may be outliers or large values that cause the observed distribution. Duration ranges from 60 seconds to 17270400 seconds. Start station ID, end station ID, and bike ID are numeric identifiers used to identify where trips start and end or the bike that was used. Start and end station ID, however, have similar distributions, which suggests that many trips may have begun and ended at the same station. As well, the most frequent starting and ending stations, as seen in Figures 1 and 2, respectively, likely have numeric identifiers around 60 to slightly above 75. Finally, the most frequently used bikes, as shown in bike ID, provide insight as to which bikes are used more or less often, which may be valuable to determine which bikes require maintenance if they are often avoided due to potential physical or visual abnormalities.

A graph of different colored bars

Description automatically generated

**Figure 4:** Histograms of the distribution of the five integer variables in “trip”. All graphs show the count on the Y-axis and the distribution values on the X-axis. Moving from the top left, the variables depicted are ID, duration, ID of the start station, ID of the end station, and bike ID.

Exploratory Data Analysis of Weather

Weather is a data frame containing information about 1825 observations across 15 categories of information. There are four character variables: date, precipitation (in inches), events, and city. There are eleven integer variables: the maximum, mean, and minimum values for temperature, the maximum, mean, and minimum values for visibility (reported in miles), the maximum and mean wind speed (reported in miles per hour (mph)), the maximum gust speed, as well as cloud cover and zip code.

Character Variables

Figure 5 shows an even distribution of observed weather events among the five cities included in the data set. The 365 events observed align with 365 days in a non-leap year, with separate weather events per day per city. Figure 6 shows the frequency of reported events, with most reports having no events reported (80.71% of the values are empty) and other observed events including variations of rain, fog, and thunderstorms. This variable may present issues in an analysis as it fails to quantify observations and contains missing data. Figure 7 shows the frequencies of reported precipitation amounts, with most observations reporting 0 inches of precipitation (84.55%). Precipitation has mainly numeric reporters except for T, representing 4% of observations. This may present issues for downstream analysis of an undefined reporting value while the rest of the values appear to be in inches.

The other character variable is date, which does not contain any missing observations and will need to be transformed into a proper date identifier for downstream analysis.

A colorful rectangular bars with text

Description automatically generated with medium confidence

**Figure 5:** Frequency plot showing the percentage frequency of the cities observed for weather events on the X-axis and the names of the cities on the Y-axis.

A graph with different colored squares

Description automatically generated

**Figure 6:** Frequency plot showing the percentage frequency of the event observed on the X-axis and the name description of the event on the Y-axis.

A screenshot of a graph

Description automatically generated

**Figure 7:** Frequency plot showing the percentage frequency of the amount of precipitation observed on the X-axis and the name or value of precipitation on the Y-axis.

*Numeric Variables*

Figure 8A-E shows the distribution of each numeric variable in “weather” across each city, as it is presumed that each city may experience a different distribution of weather events. The distribution for the maximum, mean, and minimum temperatures appears slightly different for all cities, with the histograms for mean temperature being relatively normal, the maximum temperature being slightly right skewed and the minimum temperature being slightly left skewed. All temperature values range from 32-102 °F.

Maximum visibility in Mountain View and San Jose appears to be uniformly distributed, while in Palo Alto it appears trimodal with little distribution otherwise. In Redwood City and San Jose, the maximum visibility is greatly left skewed, which may indicate the presence of excessive low-values. The mean visibility in Palo Alto takes on a loosely normal distribution, while in all other cities it is greatly left skewed. Palo Alto also has a minimum visibility that is right skewed, while for all the other cities it is again greatly left skewed. Notably, Palo Alto has a different histogram bin distribution than the other cities, with the largest value on the X-axis being 20 for the maximum, mean, and minimum visibility plots, while all other cities do not exceed 10. Reported visibility values range from 0-20 miles.

The maximum wind speed for Mountain View, Palo Alto, and Redwood City is extremely right skewed, with the largest displayed bin values ranging from 90 to 120 mph, which may indicate outliers of large values. San Francisco and San Jose have displayed bin values with a maximum of 30 or 40 mph and as such their distribution appears relatively normal and slightly right skewed for San Francisco. The mean wind speeds for all cities have right skewed distribution, with the severity of skewed distribution varying. Maximum gust speed was also binned differently across cities, with Mountain View and San Francisco’s largest displayed bin values being 125 and 100 mph, respectively, while all other cities have a largest bin value of 40 mph. All the cities display a right skewed distribution for maximum gust speed, with that of Mountain View and San Francisco being the most pronounced. Reported wind and gust speeds range from 0 to122 mph.

Cloud cover in all cities appears to have a non-symmetric bimodal distribution, that is right skewed for all cities except San Francisco. Cloud cover is reported from 0-8. Finally, although zip code appears as a numeric variable, across all cities it has uniform distribution and is an identifier that is numeric not a measurement.

When all cities are included in the analysis, in Figure 8F, maximum temperature is right skewed, while minimum temperature is left skewed and mean temperature is relatively normally distributed. Maximum visibility is right skewed with three main bins and most values falling in one bin. Mean and minimum visibility also have most values falling in one bin and are slightly right skewed. This may indicate the presence of outliers in visibility. Further all visibility measurements have 9 NAs and thus information is missing for certain measurements. Maximum wind speed, mean wind speed, and maximum gust speed are all right skewed as well, which may also indicate the presence of outliers. Maximum gust speed contains 451 NAs which indicates a significant amount of missing information. Cloud cover has a non-symmetric bimodal distribution, that is right skewed. Finally, although zip code is a numerically coded identifier, it has three main bins within which most zip codes fall.

A screenshot of a graph

Description automatically generated A graph of different colored bars

Description automatically generated with medium confidence

B

D

C

A

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph of different colored squares

Description automatically generatedA graph of different colored bars

Description automatically generated with medium confidence

F

E

**Figure 8:** Histograms of the distribution of numeric variables in “weather” across 11 variables. All graphs show the count on the Y-axis and the distribution values on the X-axis. Moving from the top left, the variables depicted are maximum temperature (°F), mean temperature (°F), minimum temperature (°F), maximum visibility (miles), mean visibility (miles), minimum visibility (miles), maximum wind speed (mph), mean wind speed (mph), maximum gust speed (mph), cloud cover, and zip code. The cities represented are **A)** Mountain View, **B)** Palo Alto, **C)** Redwood City, **D)** San Francisco, **E)** San Jose, and **F)** all cities.

Exploratory Data Analysis of Station

Station is a data frame containing information about 70 observations across 7 categories. There are three character variables: name, city, and installation date. There are two integer variables: ID and dock count, and two numeric variables: latitude and longitude. The data provides information on the 70 possible starting stations, which are also included in “trip”, although in trip there are 74 possible starting stations. This discrepancy may be due to stations added later on or excluded in the station analysis. Although the reason for trip containing an additional four stations is unknown, it will not affect the downstream analysis of determining the top 10 starting and ending stations.

Character Variables

Figure 9 shows an even distribution amongst all stations included in the data set, which is as expected as they all appear once for each separate observation. Figure 10 shows the distribution of which city each station is in, with the majority of stations in San Francisco, followed by San Jose, Redwood City, Mountain View, and Palo Alto. Finally, Figure 11 shows how many stations were installed on each day, with three stations being installed in 2014, which may skew results from the analysis of the most frequent starting and ending stations as all stations were not analyzed for the same time period. These stations are Broadwood St at Battery St (ID number 68), Mezes Park (ID number 69), and Ryland Park (ID number 70).

A chart of different colored lines

Description automatically generated with medium confidence

**Figure 9:** Frequency plot showing the percentage frequency of the stations observed on the X-axis and the names of the stations on the Y-axis.

A graph with different colored squares

Description automatically generated

**Figure 10:** Frequency plot showing the percentage frequency of the cities each stations observed is in on the X-axis and the names of the cities on the Y-axis.

A graph with different colored bars

Description automatically generated

**Figure 11:** Frequency plot showing the percentage frequency of date of installation of a station on the X-axis and the names of the cities on the Y-axis.

*Numeric Variables*

Figure 12 shows the distribution of numeric and integer variables in station. Station ID is an identifier, and since all stations are represented equally in the data frame (Figure 9), each identifier appears only once as well. Latitude and longitude are used to determine the location of stations and are also numerically coded in the “station” data frame but are location identifiers. Finally, dock count is divided into four main sections, three separated bins and a fourth section with a slightly continuous distribution.

A screenshot of a graph

Description automatically generated

**Figure 12:** Histograms of the distribution of the four numeric and integer variables in “station”. All graphs show the count on the Y-axis and the distribution values on the X-axis. Moving from the top left, the variables depicted are ID, latitude, longitude, and dock count.

**Removal of Suspected Cancelled Trips**

As decided by the data analysis team, it is suspected that trips starting and ending at the same station with a duration of less than three minutes are cancelled trip. The units of duration have not been specified, however, in the context of the data it can be assumed that duration has been provided in seconds. The total number of suspected (and removed) cancelled trips is 1082. The IDs can be found in the file labelled “cancelled\_trips2.csv” in the main branch of the included GitHub repository named “Midterm”.

**Removal of Outliers**

Trip

Outlier determination is not required for all of the variables in the trip data set, as for some it is not applicable. ID and bike ID were not deemed applicable for outlier analysis as they are both identification of either a trip or a specific bike used. Start station name (and ID), as well as end station name (and ID), were also not used for outlier identification as there is a specified number of stations within each city that a trip can start from and as such outliers do not apply.

Trip duration, however, can be affected by outliers. Since the bottom part of the data, in cancelled trips, has been cleaned by removing suspected cancelled trips, for the purpose of outlier determination, I am only analyzing outliers exceeding the duration of a reasonable bike ride. After removing cancelled trips, duration ranges from 61 seconds to 17270400 (about 200 days) seconds. Further support for my decision to not remove trips with an excessively short duration is that a bike ride between two stations may only take a minute (60s), as can be seen in Figure 13A. An excessively long bike duration is defined as over 7 hours. My reasoning for the 7-hour cut-off is that the distance between the two furthest cities is around 58 miles and would take about 5-6 hours to bike, as estimated in Figure 13B. The 7-hour cut-off is such that it includes a reasonable full-day bike if customers would like to bike from San Francisco and San Jose, including an extra hour for breaks. It is assumed that if customers would also like to bike the way back, that they would dock the bike at a station to complete tasks or a rest in the respective city without being charged for having the bike when it is not in use. 7 hours is also above the 99th percentile (above ~2.6 hours), so cases where customers do wish to complete the longest bike ride on this map are rare.

Figure 4 shows right-skewed distribution for duration, as mentioned in the Exploratory Data Analysis. This indicates outliers of an excessive duration. A logarithmic histogram can show a more accurate representation of the data without the effect of the outliers, as can be seen in Figure 14A. The removal of trips over 7 hours removed 1271 observations, and while a non-logarithmic histogram still appears largely left-skewed with most of the observations falling in one bin (Figure 14B), a log10 scaled histogram appears more normally distributed and only slightly left-skewed. The IDs of the removed cases can be found in the file labelled “trips\_outliers4.csv” in the main branch of the included GitHub repository named “Midterm”.

A screenshot of a map

Description automatically generated A map of a city

Description automatically generated

B

A

**Figure 13:** Screenshot of Google Maps predicted duration to bike from **A)** the station Market at 10th to South Van Ness at Market and **B)** San Francisco to San Jose.

A graph of a graph of a bike ride

Description automatically generatedA graph of a bicycle ride

Description automatically generated

C

B

A

A graph of a bicycle ride

Description automatically generated

**Figure 14:** Histograms of the distribution of bike ride duration in trip. All graphs show the frequency on the Y-axis and the distribution values on the X-axis. **A)** Frequency of the duration in seconds on a log10 scale prior to outlier removal. **B)** Frequency of the duration in seconds after outlier removal without the use of a log10 scale. **C)** Frequency of the duration in seconds on a log10 scale after to outlier removal.

Station

Outlier determination is not required for any of the variables in station. Name, city, ID, and installation date are all clearly defined. The reported longitude and latitude all have reasonable coordinates, as does the number of docks.

Weather

Outlier determination is not required for all variables in weather. Date, for example, is between a set range of days in 2014 and is not subject to outliers. Similarly, zip code is also not subject to outliers as it is also an identifier.Events is also a character variable that is not well defined but has a fixed number of levels.The maximum, mean, and minimum temperatures collectively range from 32°F to 102°F, both of which are reasonable temperatures, with relatively non-skewed distributions for the three temperature measurements amongst the five recorded cities. The maximum, mean, and minimum visibility scores collectively range from 0 to 20 miles, and is not clearly defined for how the measurement was taken. A human eye can theoretically see longer than 20 miles, however at ground level curvature of the Earth limits this visibility to about 3 miles.1 Since the measurement was not stated to be visibility by a human eye, 20 miles is a reasonable visibility distance and therefore these larger values should be included in the data. Cloud cover is reported on a scale of 1-8 which is a defined range that was not exceeded and therefore outliers do not need to be removed.

Maximum wind and maximum gust speed in miles per hour (mph) have some unreasonably high values, as speeds of over 75 mph are defined as hurricane force winds, and California has not got a hurricane since 1939.2,3 As such, any measures from the maximum wind and gust speeds that exceeded 75 mph were excluded by changing the values to NA. Entire rows were not removed as other weather data for that day may still be applicable. Mean wind speed had a maximum of 19 mph, which is reasonable and classified by as a light breeze.2 After removal of the outlying wind speeds, histograms of maximum wind speed and gust speed distributions are shown in Figure 15.

Although precipitation is presented as a character, it will need to be converted into a numeric variable later in the analysis of weather condition correlation with bike trips. The maximum value obtained by precipitation was 3.36 inches, and although this is excessive it is not unreasonable. Heavy rain is defined as as 0.3 or inches per hour, and rain could continue for 10-11hours to reach 3.36 inches.4

**A graph of a wind speed

Description automatically generated**

A

A graph of a speed test

Description automatically generated

B

**Figure 15:** Histograms of the distribution of **A)** Maximum wind speeds after outlier removal, and **B)** Maximum gust speeds after outlier removal. All graphs show the frequency on the Y-axis and the distribution values on the X-axis

**Rush Hours**

To determine the highest volume hours on weekdays, a histogram (Figure 16) was created to represent the frequency of bike rentals that began on the half hour, where a half-hour is defined as a start time is rounded down to the nearest half hour (ie if a bike trip started at 1:15 the half hour it started on is 1:00). Weekdays were counted as any trip that began on a Monday-Friday.

A weekday rush hour was defined as any time with a frequency of over 10000, making the weekday rush hours between 7:30-9:30 and 16:00-18:30. These defined times were then used to determine the 10 most frequent starting and ending stations during weekday rush hours, which can be seen in Table 1.

To determine the highest volume hours on weekends, a histogram (Figure 17) was created using the same method and definition for half hour increments as with weekdays.

A rush hour was defined as any time with a frequency of over 1700, making the weekday rush hours from 11:30 - 16:30. These defined times were then used to determine the 10 most frequent starting and ending stations during weekday rush hours, which can be seen in Table 2.

A graph showing the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of the number of

Description automatically generated

**Figure 16:** Histogram showing the frequencies of start times (on the Y-axis) and start times by the half-hour (on the X-axis) during weekdays.

**Table 1:** 10 most frequently used start and end stations during weekday rush hours.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Start Station Name** | **Start Station ID** | **Number of Departing Trips During Weekday Rush Hours** | **End Station Name** | **End Station ID** | **Number of Arriving Trips During Weekday Rush Hours** |
| San Francisco Caltrain   (Townsend at 4th) | 70 | 401 | San Francisco Caltrain   (Townsend at 4th) | 70 | 1500 |
| Townsend at 7th | 65 | 383 | San Francisco Caltrain 2 (330 Townsend) | 69 | 672 |
| 2nd at Townsend | 61 | 373 | Harry Bridges Plaza (Ferry Building) | 50 | 507 |
| Market at Sansome | 77 | 349 | Steuart at Market | 74 | 397 |
| 2nd at South Park | 64 | 343 | Temporary Transbay Terminal (Howard at Beale) | 55 | 360 |
| Embarcadero at Sansome | 60 | 338 | Market at Sansome | 77 | 328 |
| Temporary Transbay Terminal (Howard at Beale) | 55 | 304 | Powell Street BART | 39 | 264 |
| Embarcadero at Folsom | 51 | 300 | 2nd at Townsend | 61 | 245 |
| San Francisco Caltrain 2 (330 Townsend) | 69 | 294 | Embarcadero at Sansome | 60 | 244 |
| Market at 10th | 67 | 285 | Townsend at 7th | 65 | 231 |

A graph showing the number of frequency

Description automatically generated

**Figure 17:** Histogram showing the frequencies of start times (on the Y-axis) and start times by the half-hour (on the X-axis) during weekends.

**Table 2:** 10 most frequently used start and end stations during weekend rush hours.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Start Station Name** | **Start Station ID** | **Number of Departing Trips During Weekday Rush Hours** | **End Station Name** | **End Station ID** | **Number of Arriving Trips During Weekday Rush Hours** |
| Embarcadero at Sansome | 60 | 172 | Embarcadero at Sansome | 60 | 180 |
| Harry Bridges Plaza (Ferry Building) | 50 | 159 | Harry Bridges Plaza (Ferry Building) | 50 | 160 |
| 2nd at Townsend | 61 | 75 | Powell Street BART | 39 | 88 |
| Market at 4th | 76 | 73 | Market at 4th | 76 | 80 |
| Powell Street BART | 39 | 72 | 2nd at Townsend | 61 | 68 |
| Embarcadero at Bryant | 54 | 63 | Embarcadero at Vallejo | 48 | 63 |
| San Francisco Caltrain   (Townsend at 4th) | 70 | 61 | Townsend at 7th | 65 | 62 |
| Grant Avenue at Columbus Avenue | 73 | 57 | San Francisco Caltrain   (Townsend at 4th) | 70 | 59 |
| Market at 10th | 67 | 56 | Steuart at Market | 74 | 59 |
| Embarcadero at Vallejo | 48 | 52 | Embarcadero at Bryant | 54 | 56 |

**Utilization**

The amount the bikes were used by month was determined by summing the total duration of bike usage during the month and dividing by the number of days in the month to get the results displayed in Table 3. The results of the total hours of bike usage per month per day can also be seen in Figure 18. The average duration per day is a better representation than the total number of hours of bike usage in a month as some months have a different number of days which may lead to an overly or underly represented amount of use.

**Table 3:** Total and average daily use of bikes per month.

|  |  |  |
| --- | --- | --- |
| **Month** | **Total Duration (hours)** | **Average Duration (Hours) Per Day** |
| January | 5216 | 168 |
| February | 4032 | 144 |
| March | 5755 | 186 |
| April | 6067 | 202 |
| May | 6825 | 220 |
| June | 6830 | 228 |
| July | 7467 | 241 |
| August | 7722 | 249 |
| September | 7020 | 234 |
| October | 7296 | 235 |
| November | 5313 | 177 |
| December | 4082 | 132 |

A graph of a number of bikes

Description automatically generated with medium confidence

**Figure 18:** Bar graph showing the average daily utilization of bikes in hours during each month. The months are represented on the X-axis and the average amount of hour of bike usage per day is on the Y-axis.

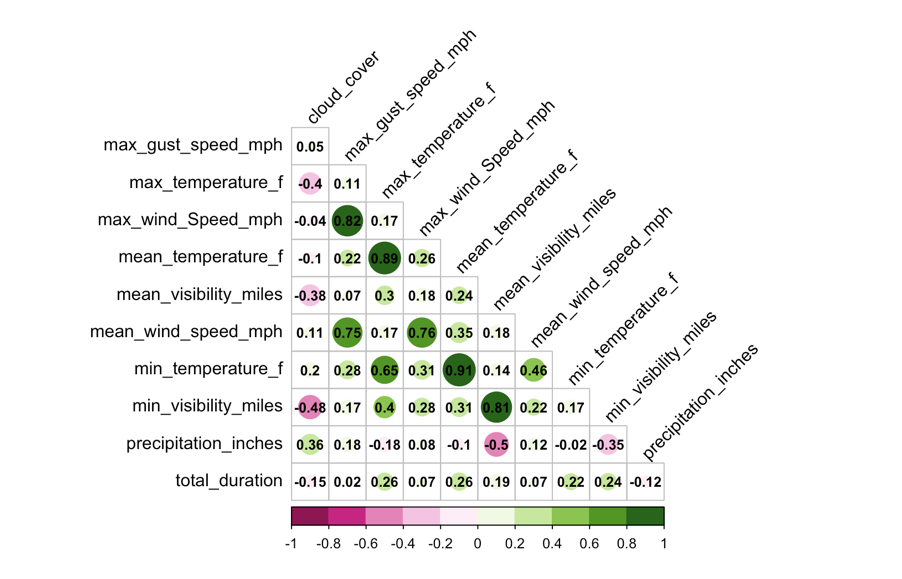
**Correlation**

To analyze whether there is a correlation between weather patterns and the bike usage (measured in total duration for the day) a correlation analysis using the corrplot package in R was performed. A new data set, called “start\_weather” was created, which consists of 1825 observations and 14 variables. There is one factor variable, city, which consists of the five cities dealt with in the data sets. There is one date variable, which are all possible dates in 2014. The other 12 variables are numeric, consisting of the 10 weather descriptors from the “weather” dataset (maximum, mean, and minimum temperature, maximum, mean, and minimum visibility, maximum, and mean wind speed, maximum gust speed, and cloud cover). Precipitation was also converted to a numeric variable, with all values of T converted to 0.011. This was done under the assumption that T meant true, indicating that precipitation had occurred. Since the value of precipitation, however, was not explicitly described, 0.011 is 0.001 inches above the lowest precipitation value of 0.01. The last numeric variable is the total duration that was biked in the given day for each city. The assumption was made that weather correlations affected the beginning of bike ride trips more than the end, as when a person is deciding their form or transportation and whether to bike the weather at the beginning of the trip will more likely affect that choice. As such, total trip duration per day and the weather of cities was only analyzed for the city in which the trip began. The events variable from “weather” was excluded from the analysis as it contains only qualitative information that can likely be described by numeric variables (ie precipitation can quantitively describe rain or thunderstorms, and cloud cover may be sufficient in describing fog).

Figure 19 shows correlation plots for each of the cities analyzed across the specified variables. Mountain View, Redwood City, and San Jose all have maximum visibility omitted in the analysis, as an error occurred when running the correlation that there was no standard deviation in this variable. For Mountain View and San Jose this can be seen in Figure 8A and E, respectively. For Redwood City, however, Figure 8C does not appear to have no standard deviation, and a direct standard deviation analysis showed that maximum visibility had a standard deviation of ~3.12, however for the purpose of this analysis it was omitted so that the corrplot package could be used properly. Notable findings from the correlation analysis on the total duration biked have been highlighted by city and are summarized in Table 4.

**Table 5:** Summary of correlation between starting city weather patterns and total daily bike usage.

|  |  |
| --- | --- |
| **City** | **Notable Correlations with Total Bike Usage** |
| Mountain View | * Weak positive correlation between temperature (maximum, mean, and minimum) and total daily biking duration. * Weak positive correlation between minimum visibility and total daily biking duration. |
| Palo Alto | * Weak positive correlation between maximum temperature and total daily biking duration. * Weak positive correlation between visibility (maximum, mean, and minimum) and total daily biking duration. |
| Redwood City | * No significant correlations between any weather condition and total daily biking duration |
| San Francisco | * Weak negative correlation between cloud cover and total daily biking duration. * Moderate positive correlation between temperature (maximum, mean, and minimum) and total daily biking duration. * Moderate positive correlation between visibility (mean and minimum) and total daily biking duration. * Moderate negative correlation between precipitation and total daily biking duration. |
| San Jose | * Weak negative correlation between cloud cover and total daily biking duration. * Weak to moderate positive correlation between temperature (maximum, mean, and minimum) and total daily biking duration. * Weak to moderate positive correlation between visibility (mean and minimum) and total daily biking duration. * Weak negative correlation between precipitation and total daily biking duration. |

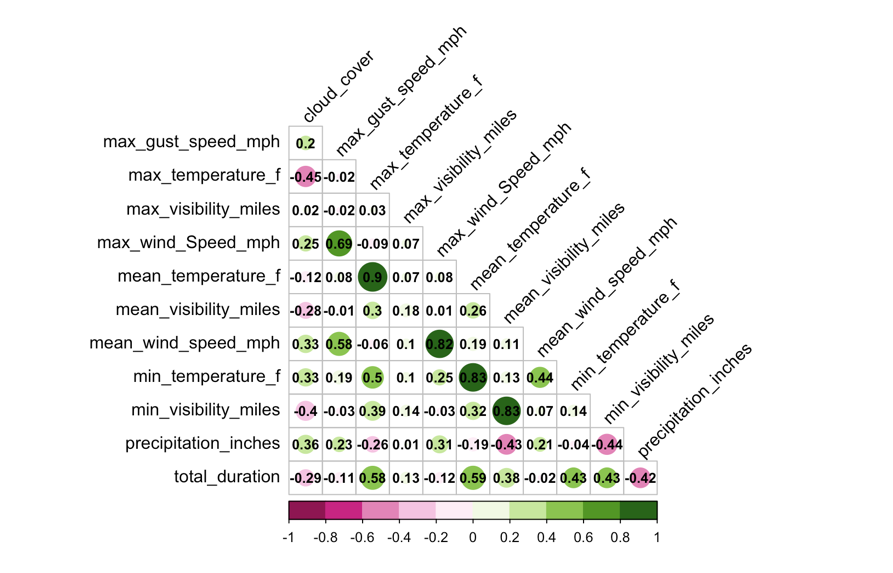
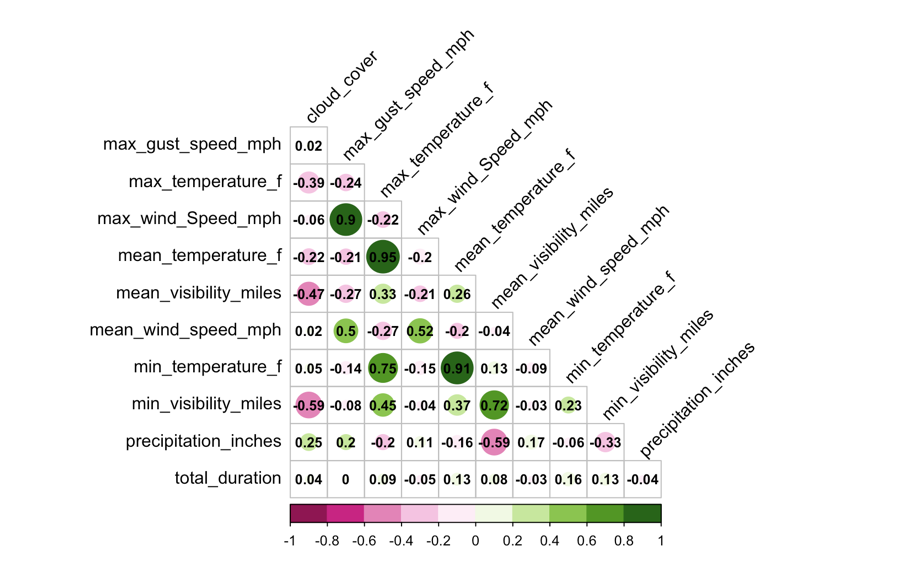
A graph showing the speed of a speedometer

Description automatically generated with medium confidence

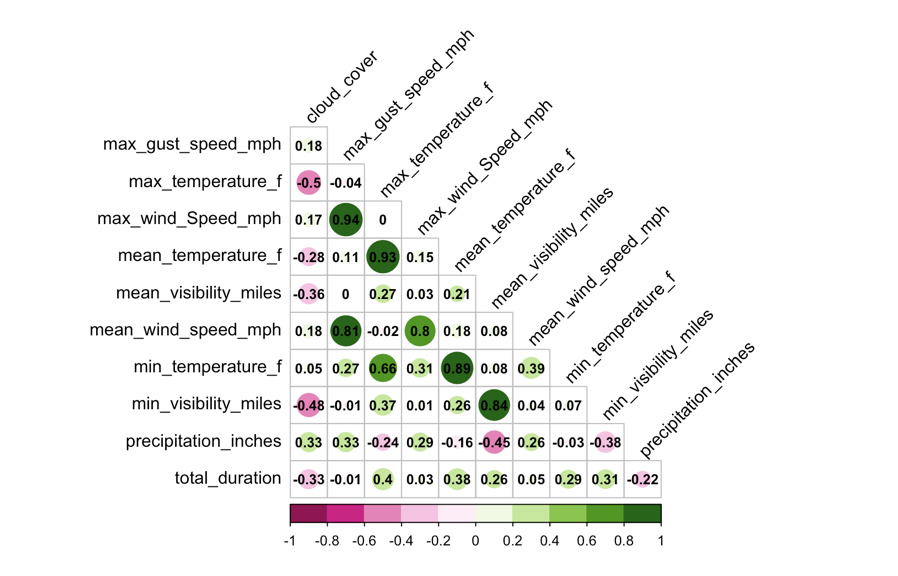
D

B

A



C



E

**Figure 19:** Correlation plots of weather patterns and bike duration where positive correlation is represented in green and negative correlation is in pink. Strong correlations have a larger circle and a value closer to 1 or -1. Total duration represents the total duration of bike rides for a given day in 2014 in the represented. Correlation plots are across the cities **A)** Mountain View, **B)** Palo Alto, **C)** Redwood City, **D)** San Francisco, and **E)** San Jose.

Work Cited

1. How far can the human eye see. (2023, June). *Warby Parker*. Retrieved from <https://www.warbyparker.com/learn/how-far-can-the-human-eye-see>
2. *Estimating wind speed.* (n.d.). National Weather Service. Retrieved from <https://www.weather.gov/pqr/wind>
3. Singh, S. (2023, August). Has California ever had a hurricane? One expert says tropical storm threat from Hilary is "nearly unprecedented". *CBS News*. Retrieved from [https://www.cbsnews.com/news/hurricane-hilary-forecast-california-tropical- storm-nearly-unprecedented/](https://www.cbsnews.com/news/hurricane-hilary-forecast-california-tropical-%09storm-nearly-unprecedented/)
4. Skilling, T. (2018, September). Do meteorologists have specific definitions for drizzle, light rain, steady rain, heavy rain, downpour? *Chicago’s Very Own WGN9*. Retrieved from [https://wgntv.com/weather/weather-blog/do-meteorologists-have- specific-definitions-for-drizzle-light-rain-steady-rain-heavy-rain-downpour/](https://wgntv.com/weather/weather-blog/do-meteorologists-have-%09specific-definitions-for-drizzle-light-rain-steady-rain-heavy-rain-downpour/)