**Bay Area Bike Rental Analysis Report**

**Description and Summary of dataset**

Three datasets were used and analyzed in this report. The station.csv, trip.csv, and weather.csv datasets. Briefly, they contain the following information:

* Station.csv:
  + Contains data that represents a station where users can pick up or return bikes
* Trip.csv:
  + Contains data about individual bike trips in the US from stations listed in station.csv
* Weather.csv:
  + Contains data about the weather on all 365 days in 2014 for 5 cities in the US (San Francisco, Redwood City, Palo Alto, Mountain View, and San Jose)

**Station dataset**

The station dataset contains a variety of stations in the Bay Area that include its station ID, name, geographic location (longitude and latitude), number of total docks in the station, its landmark city location, and its installation date.

A brief inspection of the dataset revealed that it has no missing values, and each observation are unique. However, a single structural issue is detected in the installation\_date variable, which are listed as character. Using the lubridate package in R, they are easily converted into a date format.

**Trip dataset**

The trip dataset has data on 326,339 bike trips in the Bay Area that includes their unique trip ID, duration of their trip, starting and ending date and time, starting and ending station name and ID, the bike ID used for the trip, user’s subscription type, and user’s zip code.

Brief inspection of the dataset show that each observation is unique, but like the station dataset entries, the starting and ending trip dates are of character structure. However, as the dates also include the starting time, they are converted first into a POSIXct structure.

We also observe that there are erroneous entries for the zip codes such that some of the zip codes have 1 to 2 digits, characters, and mix of numbers and characters. Since all the three datasets focuses on the Bay Area in the US, a valid zip code should only include numbers ranging from 501 to 99950. Thus, zip code entries apart from 501 to 99950 are converted into NA.

Using the FunModeling package in R, we perform an exploratory data analysis on this dataset and found that the San Francisco Caltrain (Townsend at 4th), Harry Bridges Plaza (Ferry Building), and San Francisco Caltrain 2 (330 Townsend) stations are among the top three starting and ending stations for trips in 2014. This information is illustrated in Figure 1 below. Moreover 85% of the trips are done by subscribers.

A long shot of a chart

Description automatically generated with medium confidenceA chart of a number of cities

Description automatically generated with medium confidence

**Figure 1. Percent frequency of start and end station names for bike trips in the Bay Area in 2014.**

Furthermore, part of our EDA is also ploting the histogram for the counts of each numerical variables in the data. With 8 numerical variables, 8 histogram are generated and depicted in Figure 2 below. Inpecting them further, two of the histogram sticks out with due to the possibility of having outliers, the trip duration and zip code.

Considering the majority of US zip codes in the Bay Area are above 90000, the resulting histogram is not absurd. The small frequency count of 0 zip codes are those we have previously set as NA in our analysis.

The duration histogram indicate that a few of the trips might have a very long duration. This may indicate some possibility of outliers in the dataset where they may have been recorded erroneously, which we will deal with shortly.

A screenshot of a graph

Description automatically generated

**Figure 2. Histogram of numerical variables from the trip.csv dataset.** This includes the trip ID, trip duration, starting date, starting station ID, ending date, ending station ID, bike ID, and zip codes for 326,339 trips.

The remaining 6 histograms do not indicate any potential issues with outliers. Additionally, while they are numeric, these histogram might not be useful in describing them as they are simply an identifier or a date.

**Weather dataset**

The weather dataset contains data on weather conditions, including maximum, mean, and minimum temperature in Farenheit, maximum, mean, and minimum visibility in miles, maximum and mean wind speed in mph, maximum gust speed in mph, precipitation inches, weather events on that day, and the city and zip code of the recording.

These observations span across the entire calendar year in 2014 for 5 cities, San Francisco, Redwood City, Palo Alto, Mountain View, and San Jose. Collectively, they result in the 1825 observations in the dataset (365 x 5).

While there appear to be missing values in some of the weather observations for some cities (i.e., missing wind speed entry for a specific date in SF), we’ve decided not to remove them to ensure that each cities have a complete weather recording for the entire calendar year (keeping the 365 entries for each city), despite some criteria may not be recorded.

The dates are also converted into a date structure like the previous two datasets. Moreover, we also noted that the precipitations are not consistent across all observation where some are recorded as T (trace), and other in numerical forms. It is believed that trace number of precipitations are between 0.00 and 0.01. Here, we have arbitrarily picked 0.001 as a replacement for the T recordings, ensuring that all the observations are consistent (are numeric).

Performing the exploratory data analysis again confirmed that there is an equal amount of weather observations for the 5 cities, with 365 observations each as depicted in Figure 3 below.

A colorful rectangular bars with text

Description automatically generated with medium confidence

**Figure 3. Frequency of weather observations across 5 cities in Bay Area.** This figure illustrates the frequency of daily weather observations across five cities in the Bay Area: San Francisco, Redwood City, Palo Alto, Mountain View, and San Jose. Each city has exactly 365 observations, representing daily data collection for one year, ensuring consistent and comprehensive weather monitoring across the region.

Looking at the histogram plots for each numerical variables in our dataset below also shows most of the weather observations are precise with few outliers. We can argue that some outliers might be present in the maximum gust speed and precipitation inches observations, with recordings reaching 90 mph gust and 3 inches of precipitations. However, these observations are still considered valid and may be indicative of severe weather conditions, bearing in mind that they are annual weather recordings. Moreover, in the interest of keeping all the observations, we’ve decided not to remove these extreme values.

Briefly, we can see that there is a relatively normal distribution in the temperatures and average wind speed, which agrees with the dynamic nature of these variables within a calendar year. While not instantly visible, the maximum and average visibility also appeared to assume a normal distribution. However, the minimum visibility, maximum wind speed, maximum gust speed, and precipitation inches appeared to be negatively skewed. These findings are therefore realistic as weather conditions are often on the lower end (mostly calm and stable throughout the year), especially around the Bay Area in the US.

A screenshot of a graph

Description automatically generated

**Figure 4. Histogram of numerical variables from the weather.csv dataset.** This includes the date, maximum, mean, and minimum temperature in Fahrenheit, maximum, mean, and minimum visibility in miles, maximum and mean wind speed in mph, maximum gust speed in mph, precipitation in inches, cloud cover, and zip codes.

**Cancelled Trips**

Among the 326,339 recorded bike trips, some of the trips are cancelled. Cancelled trips are those trips where the duration is less than 3 minutes with the same starting and ending stations.

Fortunately, only a very small portion of the trips were cancelled, which is 0.33%. This amounts to 1,082 trips out of a total of 326,339. The trip IDs are recorded in the cancelled\_trip\_ids.csv file. This proportion is visualized in Figure 5 below.

A blue rectangle with black text

Description automatically generated

**Figure 5. Frequency of cancelled bike trips.** Cancelled trips are defined as those where the start and end station are the same, and the duration is less than 3 minutes. Most trips (99.67%) were not canceled, while only 0.33% of trips were canceled.

**Outliers**

As previously described, we are only concerned in removing the outliers from the trip duration variable in the trip dataset. Figure 6 below is a boxplot of the current observed trip durations, highlighting the presence of outliers with exceptionally long duration.

A graph of a trip duration

Description automatically generated

**Figure 6. Boxplot of bike trip duration.** The boxplot describes the distribution of bike trip durations (in seconds) for all non-canceled trips. Circular points indicate the outliers, which are 1.5 times the interquartile range (IQR). The median and IQR range are not visible due to the extreme outliers.

Here, we establish the outliers being any trip duration greater or less than 1.5 times the interquartile range (IQR) beyond the 3rd and 1st quartiles (Q3 + 1.5 IQR and Q1 – 1.5 IQR), respectively. The IQR is the difference between the third quartile (75%) and the first quartile (25%), which equates to 405 seconds in our dataset.

With this criterion, we’ve eliminated 24,873 observations, which is roughly 7.6% of the uncancelled trips. This results in a remaining of 300,384 observations in the final, cleaned trip observations. This cleaned data is described in Figure 7 below. The outlier ids are recorded in the outlier\_ids.csv file accompanying this report.

Looking at the updated boxplot on Figure 7 below, we can have a better depiction of the box and also the mean, first (25th percentile of the data), and third (75th percentile of the data) quartiles, and whiskers, which extends from the box to the smallest and largest values within 1.5 times the IQR from the first and third quartiles, respectively.

Alternatively, we can also plot the histogram depicting the frequency for different trip durations, as depicted in Figure 8 below. This histogram clearly shows a slight negative skew with most trips lasting between 400 – 500 seconds, which is around 6-9 minutes. This finding conforms with the previous findings that the top starting and ending trip stations are the San Francisco Caltrain (Townsend at 4th) and Harry Bridges Plaza (Ferry Building) stations. Using Google maps, we can see that it indeed takes on average 9 minutes to travel via bike between these two stations.

A graph with lines and numbers

Description automatically generated

**Figure 7. Boxplot of bike trip duration after removing outliers.** The boxplot describes the distribution of bike trip durations (in seconds) for all non-canceled trips, after removing outliers. The interquartile range (IQR) is 405 seconds, and outliers are defined as any duration greater than 1.5 times the IQR above the third quartile or below the first quartile. The plot shows the median trip duration as a horizontal line within the box, with the whiskers extending to the most extreme values within 1.5 times the IQR. Circles beyond the whiskers indicate the outliers from this cleaned data.

A graph of a trip duration

Description automatically generated

**Figure 8. Histogram of bike trip duration after removing outliers.** The histogram describes the frequency of bike trip durations (in seconds) for all non-canceled trips, after removing outliers. The distribution is slightly negatively skewed, with most trips lasting between 400 and 500 seconds.

**Rush Hour**

Based on the cleaned trip dataset, we can determine the rush hour where most of the bike trips are happening during the weekday. Figure 9 below shows the trip counts by daily hour during the weekdays.

Figure 9 shows a bimodal distribution of trip counts in the 24-hour day period during the weekdays with one peak at around 8-9 AM and another peak around 5-6 PM. Therefore, the rush hours are 8 AM, 9 AM, 5 PM, and 6 PM. This finding agrees with the fact that traffic rush hours are usually between 8-9 AM when people are getting to work and again between 5-6 PM when people are getting off from work.

We can also see a very low number of trip counts between 12 AM to 5 AM where most people are resting and sleeping. The trip numbers during the day (between 10 AM to 3 PM) are steady at around 10,000 trips which is also reflective of the fact that these are great times where many people would be biking around the city as the sun is still out.

A graph of blue bars

Description automatically generated

**Figure 9. Barplot of bike trip counts during different hours of the day.** The plot appears bimodal with two peaks, at 8 Am and 5 PM.

**Weekday Rush Hour Stations**

Further analysis revealed that the top 2 starting and ending station are the San Francisco Caltrain (Townsend at 4th) and San Francisco Caltrain 2 (330 Townsend) stations. As expected, both bike stations are located at the San Francisco train stations, specifically the California commuter rail line serving the SF peninsula and Santa Clara Valley. The high volume of trips starting and ending at this station could indicate that most bike commuters take the Caltrain to head to work and return after work.

However, there are slight variation between the remaining 8 starting and ending stations. Figure 10 illustrates the trip counts for the top 10 starting stations during the weekday rush hour, and Figure 11 illustrates the trip counts for the top 10 ending stations.

Other popular starting stations include Temporary Transbay Terminal (Howard at Beale), Harry Bridges Plaza (Ferry Building), and Steuart at Market. Other popular ending stations include Market at Sansome, 2nd at Townsend, and Temporary Transbay Terminal (Howard at Beale).

A graph of blue bars

Description automatically generated

**Figure 10. Barplot of the top 10 starting bike stations during weekday rush hours.** The highest number of trips start at the San Francisco Caltrain (Townsend at 4th) station, followed closely by the San Francisco Caltrain 2 (330 Townsend) station.

A graph of a number of stations

Description automatically generated

**Figure 11. Barplot of the top 10 ending bike stations during weekday rush hours.** The highest number of trips start at the San Francisco Caltrain (Townsend at 4th) station, followed closely by the San Francisco Caltrain 2 (330 Townsend) station.

**Weekend Popular Stations**

Weekend trip analysis shows a striking difference in the highest starting and ending bike stations compared to the weekday trips. During the weekend, the top 3 starting stations are Embarcadero at Sansome, Harry Bridges Plaza (Ferry Building), and Market at 4th. The top 3 ending stations are similar to the starting stations, but with the Harry Bridges Plaza being the one with the highest ending trip counts. The top 10 starting and ending stations are also illustrated in Figure 12 and 13 below.

As expected, these top 3 stations are located in tourist and leisure areas. The Embarcadero at Sansome station is located near the SF pier while Harry Bridges Plaza is a famous public hang out square in the city. Similarly, Market at 4th station is located near one of the city’s markets.

Other popular weekend trip starting stations include 2nd at Townsend, Embarcadero at Bryant, and Powell Street BART stations. On the other hand, popular weekend ending trip stations include Powell Street BART, San Francisco Caltrain (Townsend at 4th), and 2nd at Townsend stations.

A graph of different colored bars

Description automatically generated with medium confidence

**Figure 12. Barplot of the top 10 ending bike stations during weekends.** The highest number of trips start at the Embarcadero at Sansome station, followed closely by the Harry Bridges Plaza (Ferry Building) station.

A graph of a number of stations

Description automatically generated with medium confidence

**Figure 13. Barplot of the top 10 ending bike stations during weekends.** The highest number of trips start at the Harry Bridges Plaza (Ferry Building) station, followed closely by the Embarcadero at Sansome station.

**Average Monthly Bike Utilization**

The average monthly bike utilization is calculated by dividing the total time used (total duration of the trips in that month) by the total time available in that month. The resulting percent utilization is listed in Table 1 below.

Summing up the entire monthly trip duration showed that October has the longest total trip duration in 2014 among the 5 cities with 16,910,897 seconds or roughly 4697 hours spanning over 31,864 trips. This duration is almost twice the total trip duration in the month with the shortest duration, February, with only 8,981,880 seconds or roughly 2495 hours over 17,690 trips. One possible reason is that February is considered one of the colder months in the year.

It should be noted that all these trips throughout 2014 are not performed using a single bicycle. Analyzing the unique number of bike IDs showed that there are 687 bikes in total. Thus, the average monthly bike utilization for a single bike should consider the total number of bikes available. The formula used to calculate average monthly utilization is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Month** | **Total duration (seconds)** | **Total trips** | **Total time in month (seconds)** | **Average utilization** | **Percent utilization** |
| **January** | 11,718,971 | 22,682 | 1,840,060,800 | 0.00637 | 0.637% |
| **February** | 8,981,880 | 17,690 | 1,661,990,400 | 0.00540 | 0.540% |
| **March** | 11,835,638 | 22,362 | 1,840,060,800 | 0.00643 | 0.643% |
| **April** | 12,650,823 | 24,056 | 1,780,704,000 | 0.00710 | 0.710% |
| **May** | 13,835,604 | 25,903 | 1,840,060,800 | 0.00752 | 0.752% |
| **June** | 14,556,587 | 25,703 | 1,780,704,000 | 0.00817 | 0.817% |
| **July** | 15,252,430 | 28,464 | 1,840,060,800 | 0.00829 | 0.829% |
| **August** | 15,063,555 | 28,283 | 1,840,060,800 | 0.00819 | 0.819% |
| **September** | 15,506,966 | 29,352 | 1,780,704,000 | 0.00871 | 0.871% |
| **October** | 16,910,897 | 31,863 | 1,840,060,800 | 0.00919 | 0.919% |
| **November** | 12,526,366 | 23,940 | 1,780,704,000 | 0.00703 | 0.703% |
| **December** | 9,541,043 | 18,285 | 1,840,060,800 | 0.00519 | 0.519% |

**Table 1. Monthly bike utilization statistics for 2014 in the Bay Area.**

Looking at the percent utilization above, the average utilization of these bikes is relatively low, with all values being below 1%. As expected, October, with the longest trip duration, has the highest percent utilization for each bike. Interestingly, while December’s trip duration is longer than February’s, it has the least percent utilization, which is due to having more days in that month.

This low average utilization is explained by the high number of bikes available (687 unique bikes). It is likely that most of these bikes are being docked throughout the day, as described previously in Figure 9, where there is a very low volume of trips happening during the nighttime. This suggests that the availability of bikes often exceeds the demand, especially during non-peak hours. Another reason could be that certain bikes are more frequently used while others remain idle.

**Weather Condition Impact on Rental**

Combining the trip and weather datasets, we performed a correlation matrix analysis to investigate whether weather conditions impact rental patterns, specifically the duration of bike trips. The resulting correlation matrix is shown in Figure 14 below.

A graph of weather data

Description automatically generated with medium confidence

**Figure 14. Correlation matrix of trip duration and weather conditions.** The figure shows the correlation matrix between bike trip durations and weather conditions. Blue boxes indicate positive correlation while red boxes indicate negative correlation. The darker the colour, the darker the corresponding correlation is. It appears that the correlations between trip duration and weather conditions are very weak, indicating that weather have minimal impact on the duration of bike trips.

As expected, the strongest correlation appears diagonally, which is each variable correlated to itself. We also observe strong correlations among the three temperature statistics, the three visibility statistics, and between the wind and gust measurements.

However, the duration of bike trips shows very weak correlations with weather conditions, indicated by the near-zero correlation values represented by almost white squares. Nevertheless, a more in-depth look into the specific correlation values reveals a minimal positive correlation with temperature, visibility, wind, and gust conditions, while there is a slight negative correlation with precipitation and cloud cover. These duration-specific correlation values are detailed in Table 2 below.

In summary, as expected, during warmer temperatures and better visibility conditions, people are slightly more likely to go on bike trips. Conversely, during periods of higher cloud cover, which may indicate a likelihood of rain, or during rainfall, people are less likely to go on bike trips. Interestingly, there is a minimal positive correlation between bike trip duration and wind/gust conditions, suggesting that people might engage in slightly longer bike trips during windy conditions. However, it is essential to note that these correlation values are very small, indicating that weather conditions do not significantly affect bike trip patterns in terms of duration.

|  |  |
| --- | --- |
| **Weather Conditions** | **Correlation with Duration** |
| Min Temperature (oF) | 0.012 |
| Mean Temperature (oF) | 0.015 |
| Max Temperature (oF) | 0.013 |
| Min Visibility (miles) | 0.016 |
| Mean Visibility (miles) | 0.007 |
| Max Visibility (miles) | 0.008 |
| Max Wind Speed (mph) | 0.002 |
| Mean Wind Speed (mph) | 0.022 |
| Max Gust Speed (mph) | 0.003 |
| Precipitation (inches) | -0.016 |
| Cloud Cover | -0.005 |

**Table 2. Correlation values between trip duration and various weather conditions.**