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

Trinley Palmo

BTC1855H

August 6th, 2024

**Brief description and summary of the dataset**

The dataset utilized for this report has been transformed from the second year of Bay Area Bike Share’s operation and downloaded from Kaggle (<https://www.kaggle.com/datasets/benhamner/sf-bay-area-bike-share>). The dataset contained three files: *stations.csv*; *trips.csv*; and *weather.csv*.

*station.csv*

This dataset contained data on each bike station within the Bay Area where users can go to rent or return bikes. It has 70 rows/observations and 7 columns. The 7 columns are as follows:

1. *id*, which contained ID numbers that identified each unique record.
2. *name*, which contained names of the stations.
3. *lat*, which contained the latitude measurements of each station.
4. *long*, which contained the longitude measurements of each station.
5. *dock\_count*, which contained the number of docks at each station.
6. *city*, which contained the names of the cities in which the station was located.
7. *installation\_date*, which contained the dates of when the station was installed.

It contained data on 70 stations. The average latitude and longitude of the station’s location are 37.59 and -122.2 degrees, respectively. The average number of docks at a station is about 17-18. Each station is found in one of five cities: San Jose, Redwood City, Mountain View, Palo Alto, and San Francisco. The earliest a station was installed was 2013, August 15th, and the latest was 2014, April 9th.

*trips.csv*

This dataset contained data on each individual Bay Area Bike Share trip. It has 326339 rows and 11 columns. It contains data on 326339 trips. The 11 columns of measurements are as follows:

1. *id*, which contained ID numbers that identified each unique record.
2. *duration*, which contained the duration of the trip in seconds.
3. *start\_date*, which contained the date (month/day/year format) and time (hour:minute format) of when the trip started.
4. *start\_station\_name*, which contained the name of the station where the trip started. In other words, it is the station where the bike was rented from.
5. *start\_station\_id*, which contained the ID number of the station where the trip started.
6. *end\_date*, which contained the date (month/day/year format) and time (hour:minute format) of when the trip ended.
7. *end\_station\_name*, which contained the name of the station where the trip ended. In other words, it is the station where the bike was returned to.
8. *end\_station\_id*, which contained the ID number of the station where the trip ended.
9. *bike\_id*, which contained the ID number of each bike that was rented for the trip.
10. *subscription\_type*, which contained the subscription type of the user.
11. *zip\_code*, which contained the home zip code of the user.

The plots below show the distribution of numerical variables in the dataset containing trip data through histograms. For most variables, there is an even spread of their values. Based on the distribution of start\_station\_id and end\_station\_id, it seems that the stations where people begin their trips the most are similar to those that people end their trips at. There are also certain bikes that are used the most often among all bikes. Duration (of the trips) seems to have extreme values and outliers, based on the spread and the x axis labels. Outliers will have to be identified and removed from this column.**A screenshot of a graph

Description automatically generated**

**Figure 1. Distribution of numeric variables from the `trips` dataset.** Various histograms presented for `id`, `duration`, `start date`, ‘start station id`, `end date`, `end station id`, and `bike id`.

For the cities, the trips dataset covers an equal number of records among the 5 cities. However, note that this does not imply that all records will have data available for all variables.

A graph with different colored rectangles

Description automatically generated

**Figure 2. Distribution of the categorical variable, `city`, from the `trips` dataset.** The bar-chart shows the frequency and percentage proportion of trip records across five cities in the Bay Area: San Jose, San Francisco, Redwood City, Palo Alto, and Mountain View. Each city has an equal number of records available.

The two plots below provide an overview of all stations within the dataset and their distribution. Based on these plots, the top starting and ending station is San Francisco Caltrain (Townsend at 4th), while the bottom starting and ending station is Broadway at Main.

A rainbow colored line with white background

Description automatically generated

**Figure 3. Distribution of the categorical variable, `start station name`, from the `trips` dataset.** The bar-chart shows the frequency and percentage proportion of trips starting from various stations. The station names are listed on the y-axis. San Francisco Caltrain (Townsend at 4th) has the highest frequency among all start stations.

A rainbow colored line with white text

Description automatically generated

**Figure 4. Distribution of the categorical variable, `end station name`, from the `trips` dataset.** The bar-chart shows the frequency and percentage proportion of trips ending at various stations. The station names are listed on the y-axis. San Francisco Caltrain (Townsend at 4th) has the highest frequency among all end stations.

*weather.csv*

This dataset contained daily weather information for each service region. It contains 1825 rows and 15 columns. The 15 columns are as follows:

1. *date*, which contained the date on which the measurements were taken on.
2. *max/mean/min\_temperature\_f* (3 columns), which contained the maximum, mean, and minimum temperatures in Fahrenheit for each day.
3. *max/mean/min\_visibility\_miles*, which contained the maximum, mean, and minimum visibility in miles for each day.
4. *max/mean\_wind\_speed\_mph*, which contained the maximum and mean wind speed in miles per hour for each day.
5. *max\_gust\_speed\_mph*, which contained the maximum gust speed in miles per hour for each day.
6. *precipitation\_inches*, which contained the amount of precipitation in inches for each day.
7. *cloud\_cover*, which contained the level of cloud cover for each day, from 0-8 where 0 is clear skies.
8. *events*, which contained the special weather event that occurred that day, such as fog, rain, or fog-rain.
9. *zip\_code*, which contained the zip code of the region where the measurements were taken.
10. *city*, which contained the city where the measurements were taken.

A graph of different colored bars

Description automatically generated with medium confidence

**Figure 5.** **Distribution of numeric weather metric variables from the `weather` dataset.** Various histograms presented for `date`, `max temperature (f)`, `mean temperature (f)`, ‘min temperature (f)`, `max visibility (miles)`, `mean visibility (miles)`, `min visibility (miles)`, `max wind speed (mph)`, `mean wind speed (mph), `max gust speed (mph)`, `precipitation (in)`, and `cloud cover`.

The plots above show the distribution of numerical variables in the dataset containing weather data through histograms. For some variables (e.g., date, mean\_temperature\_f, etc.) there is an even or normal distribution of their values. For the visibility, max wind speed, and max gust speed measures, there seems to be an extremely skewed distribution of their measurements. From the spread of precipitation data, it seems like it doesn’t rain much within the Bay Area. Overall, there seems to be a presence of extreme values in several variables. However, the skewness and extreme values could be due to the natural variation of weather conditions.

The plot below shows the distribution of events within the weather dataset. There are 5 events, none (which was what NAs imputed to), rain, fog, fog-rain, and rain-thunderstorm. Most trips had no events, followed by rain at 15.4%. This corresponds to the minimum amount of precipitation noted above. Rain-thunderstorms were only found in one trip record. Overall, there is an uneven spread of events between the trips.

A graph with purple and blue squares

Description automatically generated

**Figure 6. Distribution of the categorical variable, `events`, from the `weather` dataset.** The bar-chart shows the frequency and percentage proportion of daily weather records with the various weather events. `None` was the most common, followed by `rain`, `fog`, `fog-rain`, then `rain-thunderstorm`.

**Pre-processing applied to the dataset**

Some generic pre-processing that were applied to all datasets were converting the class of certain variables. For instance, all dates and datetimes were in the `stations`, `weather`, and `trips` datasets were converted to Date classes. Numeric variables, like precipitation\_inches (in `weather`) and were also converted to numeric classes.

*weather*

Before converting to numeric class, the non-numeric values found in the `precipitation\_inches` column in `weather`, specifically “T”, were imputed to 0.001. “T” represented amount less than 0.01 inches. Unique values in the `events` column were also checked in `weather`, where two events called “Rain” and “rain”. To prevent this, I changed all the values to lowercase which resolved the issue. Empty strings were also found in `weather`, specifically under `events`. Those values were imputed to “none” to reflect that there was no special event that occurred that day. Rows with NAs were then removed from the dataset, resulting in 452 rows dropped. They were removed because we will be using this information later to conduct a correlation analysis for each city and each day between daily weather metrics and daily bike rental patterns. Observations with information missing will not be useful.

*trips*

The `zip\_code` column in `trips` tracks the home zip codes of the subscribers/users. This is not required for the analysis, so it was dropped.

**Requirements from team and findings**

*Cancelled Trips*

To find potentially cancelled trips, I looked for those that started and ended at the same station, and had a duration of less than 3 minutes. I found 1082 trips that satisfied these criteria. The trip IDs for them can be accessed by looking for the cancelled\_id variable in the R script or found in the `cancelled\_trip\_ids.csv` file. These were then removed from the dataset.

*Outliers*

To find the outliers, I first created a histogram with the duration data in the trips dataset. The histogram showed that there were extreme values that resulted in an inequal distribution of the duration data.

A graph of a number of hours

Description automatically generated

**Figure 7. Distribution of trip durations from original dataset.** The histogram shows the distribution of trip durations from the cleaned, original dataset. The data is highly skewed, with the majority clustering at lower values and a long tail towards the higher values.

To better understand the data and better visualize data spanning multiple orders of magnitude, I created a histogram with the log10 scale on the duration data. It compresses the range, allowing us to see potential patterns. This histogram showed that most of the data were under s.

A graph of a trip duration

Description automatically generated

**Figure 8. Distribution of log-transformed trip durations.** The histogram shows the distribution of trip durations after applying a log10 transformation, displaying underlying distribution patterns.

Determining and removing outliers based on the interquartile range (IQR) is widely accepted and often used in statistical analysis. As a result, I used this method to calculate the upper and lower limits, which were 1354.5 and -265.5, respectively. The upper limit worked well with the fact that most data were under 1000s and the lower limit allowed us to keep shorter trips. I then removed the outliers such that only the observations that had a duration within these established limits were kept. This resulted in 24873 removed trips. Checking with the histogram revealed a more even distribution.

A graph of a distribution of trip durations

Description automatically generated

**Figure 9. Distribution of trip duration after outlier removal.** The histogram shows the more symmetrical and clear distribution of trip durations after removing outliers using the interquartile range (IQR) method.

The trip IDs can be accessed by looking for the outliers\_trips\_id in the R script or found in the `outliers\_trip\_ids.csv` file.

*Highest Volume Hours on Weekdays*

To find the rush hours during the weekdays, I first filtered for all trips that occurred on a weekday. I then created an active trip hour tracker that tracks number of active trips during all 24 hours in a day. I then go through the filtered weekday trips to update the active trip hour tracker for each trip. The number of active trips in the tracker is increased by 1 each time for the corresponding hour. I then created a histogram to visualize the top 5 rush hours and then numerically found the top 5 rush hours by sorting the tracker in descending order.

A graph of blue bars

Description automatically generated

**Figure 10. Distribution of active trips per hour during weekdays.** Peak travel times observed at 8AM and 5PM, indicating high activity during morning and evening commute hours.

The top 5 rush hours were found, in order, to be 8AM, 5PM, 9AM, 6PM, 4PM. To simplify this into time ranges, the highest volume hours on weekdays would be from 8AM-9AM and from 4PM-6PM.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Hour (24hr)** | **Number of Active Trips** |
| 1 | 8 | 39938 |
| 2 | 17 | 35853 |
| 3 | 9 | 27391 |
| 4 | 18 | 25142 |
| 5 | 16 | 23809 |

*10 Most Frequent Starting Stations and Ending Stations During the Rush Hours on Weekdays*

To find these stations, I first created functions that will take in a dataframe that contains either trips that started or ended on a weekday, along with rush hours data. The functions will then filter for trips that started or ended during the rush hours, count the number of trips for each start and end station, then sort them in descending order to obtain the top 10 most frequent stations. Based on the rush hours established previously, the 10 most frequent starting and ending stations on weekdays were found to be:

*Start Stations*

|  |  |  |
| --- | --- | --- |
| **Rank** | **Start Station** | **Number of Active Trips** |
| 1 | San Francisco Caltrain (Townsend at 4th) | 13400 |
| 2 | San Francisco Caltrain 2 (330 Townsend) | 8720 |
| 3 | Temporary Transbay Terminal (Howard at Beale) | 7724 |
| 4 | 2nd at Townsend | 6616 |
| 5 | Harry Bridges Plaza (Ferry Building) | 6307 |
| 6 | Steuart at Market | 6275 |
| 7 | Market at Sansome | 5971 |
| 8 | Townsend at 7th | 5640 |
| 9 | 2nd at South Park | 4799 |
| 10 | Market at 10th | 4648 |

*End Stations*\*

|  |  |  |
| --- | --- | --- |
| **Rank** | **Start Station** | **Number of Active Trips** |
| 1 | San Francisco Caltrain (Townsend at 4th) | 21925 |
| 2 | San Francisco Caltrain 2 (330 Townsend) | 9724 |
| 3 | Market at Sansome | 7352 |
| 4 | Harry Bridges Plaza (Ferry Building) | 6943 |
| 5 | Temporary Transbay Terminal (Howard at Beale) | 6920 |
| 6 | 2nd at Townsend | 6791 |
| 7 | Steuart at Market | 6433 |
| 8 | Townsend at 7th | 6173 |
| 9 | 2nd at South Park | 4255 |
| 10 | Embarcadero at Sansome | 4151 |

\*Note: To ensure that the end stations accurately reflected the most frequent ending stations **during** the rush hours on weekdays, the dataset utilized was one that was filtered for trips that ended on a weekday.

*10 Most Frequent Starting Stations and Ending Stations During the Weekends*

The trips were first filtered into two datasets. One that started on a weekend and another that ended on a weekend. The two datasets were then used to determine the most frequent starting and ending stations accordingly. In each dataset, the number of trips for each start station and end station was counted and sorted in descending order to get the top 10 stations. They were found to be:

*Start Stations*

|  |  |  |
| --- | --- | --- |
| **Rank** | **Start Station** | **Number of Active Trips** |
| 1 | Embarcadero at Sansome | 2145 |
| 2 | Harry Bridges Plaza (Ferry Building) | 1924 |
| 3 | Market at 4th | 1266 |
| 4 | 2nd at Townsend | 1232 |
| 5 | Embarcadero at Bryant | 1232 |
| 6 | Powell Street BART | 1147 |
| 7 | San Francisco Caltrain (Townsend at 4th) | 1080 |
| 8 | Grant Avenue at Columbus Avenue | 1028 |
| 9 | Market at 10th | 877 |
| 10 | San Francisco Caltrain 2 (330 Townsend) | 871 |

*End Stations*

|  |  |  |
| --- | --- | --- |
| **Rank** | **Start Station** | **Number of Active Trips** |
| 1 | Harry Bridges Plaza (Ferry Building) | 2345 |
| 2 | Embarcadero at Sansome | 1664 |
| 3 | Market at 4th | 1508 |
| 4 | Powell Street BART | 1380 |
| 5 | San Francisco Caltrain (Townsend at 4th) | 1353 |
| 6 | 2nd at Townsend | 1269 |
| 7 | Embarcadero at Bryant | 1129 |
| 8 | Steuart at Market | 976 |
| 9 | Townsend at 7th | 921 |
| 10 | Market at Sansome | 914 |

*Average Utilization of Bikes for Each Month*

To calculate the average bike utilization per month for each bike, I used a for loop to complete actions for each unique bike (based on bike ID). For each bike, I obtained their specific set of data, grouped it by month, and calculated the total trip duration per month. I then calculated the monthly utilization rate by dividing the total trip duration by the total seconds in the month. This rate was also converted to a percentage to better understand the result. The monthly utilization rates for each bike for each month was saved to a singular dataframe containing the bike ID, the month (from 1-12), monthly utilization rate, and monthly utilization rate in percent. It was then sorted by bike ID. This data can be accessed through the .CSV file named `monthly\_utilization.csv`.

*Correlation*

To determine the impact of weather conditions on bike rental patterns, I needed to join the weather and trips data together. I decided to join them based on matching cities and days, since the weather data was based on those two measures. To do this, I first ensured that the trips data had sufficient city information. I did this by joining it with the `stations` dataset by the station names. I then filled the remaining missing city values that had available zip code information by matching it with zip codes from the weather data. Any remaining trip data with insufficient city information was removed as they did not have enough information to determine the region. I then calculated some bike rental patterns, specifically total rentals (n) and total duration (s), for each city and each date. Other daily weather metrics by date and city were also included. These data were then used to compute a correlation matrix, which was then visualized with a correlation heatmap. I then sorted the correlations by the absolute value of the correlation coefficients to determine the highest correlations.

A diagram of a weather forecast

Description automatically generated with medium confidence

**Figure 11. Correlation plot of bike rental patterns and weather metrics.** The correlation heatmap shows the correlation between various bike rental patterns and various weather metrics. The strength and direction of correlation coefficients were colour-coded. Darker intensity represented stronger correlations. Blue shades represented positive correlations, while red shades represented negative correlations.

The correlation coefficients can be accessed by looking for the `highest\_correlation` variable in the R script or found in the `trip\_weather\_correlation.csv` file. Please note that coefficients where the two variables are the same or where the two variables are both weather metrics are not included. The full correlation\_matrix can be accessed through the `correlation\_matrix` variable in the R script or found in the `full\_trip\_weather\_correlation.csv` file.

Looking at the correlation coefficients that were arranged in descending order of their absolute values, we see that total duration (s) and daily rentals (n) have the strongest correlation with the highest correlation coefficient of 0.9987. This strong positive correlation indicates that as the number of daily rentals increases, the daily total duration also increases. This is expected since more rentals should lead to longer total duration of trips during that day.

The next 6 highest correlations were mean temperature (F) and total duration, max temperature (F) with total duration, mean temperature with daily rentals, max temperature with daily rentals, min temperature with total duration, and min temperature with daily rentals. Their coefficients were found to be 0.2308, 0.2296, 0.2214, 0.2159, 0.1701, and 0.1677, respectively. The postive correlations suggest that as the average, maximum, and minimum temperatures increase, the daily total duration and the daily number of rentals increase as well. This may seem counterintuitive, since one would expect that the hotter weather sets up a more uncomfortable biking condition. However, the opposite can also be true, where the hotter weather could lead to individuals biking instead of walking, for example. Overall, this is a weak correlation, which suggests that the temperature does not significantly impact bike rental patterns.

There is also a weak negative correlation between events and total duration with a coefficient of -0.1558. An increase in event would mean worsened weather condition, as it was set up to be none, fog, rain, and rain-fog. This suggests that as the weather conditions worsen, there is a decrease in the total duration. This is expected since weather conditions like rain and rain-fog would make biking a difficult and dangerous activity. However, take note that it is also a weak correlation.

The above are the highest correlations that were found between daily bike rental patterns and daily weather metrics. Some of these seem counterintuitive, alongside others that are not explicitly stated in this report (e.g., weak positive correlation between mean wind speed (mph) with number of daily rentals and daily total duration). However, it’s important to consider that these correlations only looked at the relationship between a single weather metric and bike rental patterns. In reality, the impact of weather on riders’ choices are quite complex and multi-layered. Multiple weather metrics simultaneously create different comfortable and uncomfortable riding conditions. The weather metrics also only reflect the average or extremes of the measurements during the day. Riders may adjust their behaviour based on these conditions and time their rides to avoid the worst conditions. Overall, while there are certain notable correlations, majority were weak, requiring further research to determine the impact of weather conditions on bike rentals.