Report on 2014 San Francisco, Bay Area Bike Share Trips

BTC1855, Midterm

Prof. Sebnem Kuzulugil

Zachery Chan, 1005469012

Github Repo: https://github.com/SirAcia/BTC1855-Midterm.git

Aug 6, 2024

Introduction

The SF Bay Area Bike Share (SFBABS) focuses on providing affordable, accessible, and easy bike travel in and around the San Francisco Bay Area. This report examines the annual data on bike trips as part of the SF Bay Area Bike Share.

The overarching focus of this report is to provide insight into the available 2014 data for training of a predictive model to forecast bike usage, specifically the number of bikes entering and leaving a given station, within the upcoming 72 hours. This predictive model can decrease supply costs to SF Bay Area Bike Share HQ by providing insight for maintenance teams to better plan station, dock, and bike maintenance.

This report examines the SF Bay Area Bike Share rental data between January 1, 2014 and Dec 31, 2014. Additionally, daily weather data was collected for the various cities in which SFBABS operates as well as the station data for each of the bike stations in the rental network. The goal of this report is to conduct pre-processing of the data (including correcting structural & data issues, identifying cancelled trips, and identifying outliers) as well as insight to guide development of the previously mentioned predictive model including identifying peak hours/rush hours, identifying most used stations during peak hours, average utilisation, and identifying strong correlations of bike usage with weather.

Description & Pre-Processing of Weather, Station, and Trip Datasets

The data for this report is comprised of 3 sources: weather, station, and trip data. These datasets aggregate daily weather recordings across the 5 cities where SFBABS currently operates (San Jose, San Francisco, Redwood City, Palo Alto, and Mountain View), descriptive data for each station in SFBABS's bike rental network, and data on each trip conducted in 2014, respectively.

The largest data frame is the trip data, comprised of 326,339 observations across 11 variables:

- ID,
- Trip Duration (seconds),
- Trip Start Date,
- Start Station Name,
- Start Station ID,
- Trip End Date,
- End Station Name,
- End Station ID,
- Bike ID,
- Subscription Type, and
- Zip code.

The weather data consists of 1825 observations across 15 variables including:

- 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 (inches),
- Cloud Cover,
- Weather Events,
- Zip Code, and
- City.

Lastly, the station dataset consists of 70 observations across 7 variables including:

- Start Station ID,
- Station Name,
- Latitude,
- Longitude,
- Dock Count,
- City, and
- Installation Date.

Pre-processing of each dataset involved identifying structural issues with the raw data. Notably, the raw data is written as character strings which were subsequently converted to their correct class type. These conversions included: converting dates to POSIX across all 3 data sets, factoring "Station Name" & "ID" across datasets, factoring "Subscription Type" in Trip, factoring weather events in Weather and factoring the "City" variable in both the weather and station data. Additionally, data issues were identified including a typo in "Weather Events" (a "Rain" event was recorded as "rain") and blank strings (i.e. "") were identified in the trip and weather data sets. These blank strings were converted to NAs.

Basic exploratory analysis revealed a significant positive skew in trip duration data, as trip durations theoretically range from 0 to infinity. To address this skew, outliers were identified using an upper limit set at three times the standard deviation and a lower limit set at twice the 25th percentile. This mixed-method approach was chosen to effectively manage the positive skew. Defining a lower limit using standard deviation was impractical (since $mean - \sigma < 0$, resulting in a negative lower limit). Thus, applying the standard deviation for the upper limit accounted for variance and a small positive mean, while basing the lower limit on the IQR considered the overall data spread, effectively managing the large cluster of data close to 0.

Conceptually, the upper and lower limits correspond to approximately 180 seconds and 26 hours. In the author's opinion, these limits are reasonable, as a bike rental lasting just over a day seems lengthy, while a ride of less than 3 minutes indicates an extremely short trip between stations. Extreme outliers in the data included trips lasting 4000+ and 200+ hours and trips between

stations lasting approx. 1 minute. Both ends of the spectrum appear unfeasible and improbable. Overall, 8147 outliers were removed, the ids of these outliers are stored in "BTC_1855_Midterm_Outliers.csv".

Another aspect of pre-processing was the removal of “cancelled trips”. A “cancelled trip” is defined as bike trip lasting less than 3 minutes as well as starting and ending at the same station. In total, 1082 trips met these criteria and were removed. The ids of cancelled trips are stored in “BTC_1855_Midterm_Cancelled_Trips.csv".

Exploratory Analysis of Cleaned Data

Exploratory analysis was conducted after pre-processing. After removal of the outliers and cancelled trips, the number of observations for trips and weather were 317110 and 1825 respectively.

Examining the cleaned trip data, the mean trip time is approx. 17 minutes with an approx. IQR of [6, 12.5] minutes in length. Shown in figure 1, the majority of trips lie below 30-minute mark with few trips stretching longer than 30 minutes. Examining the spread of trips across the year, illustrated in figure 2, bike rentals are relatively consistent throughout the spring and summer with slight increases in June - September, with sharper decreases in the winter months of December and January.

Figure 1: Exploratory Descriptive Analysis (EDA) Plots for Cleaned Trip Data

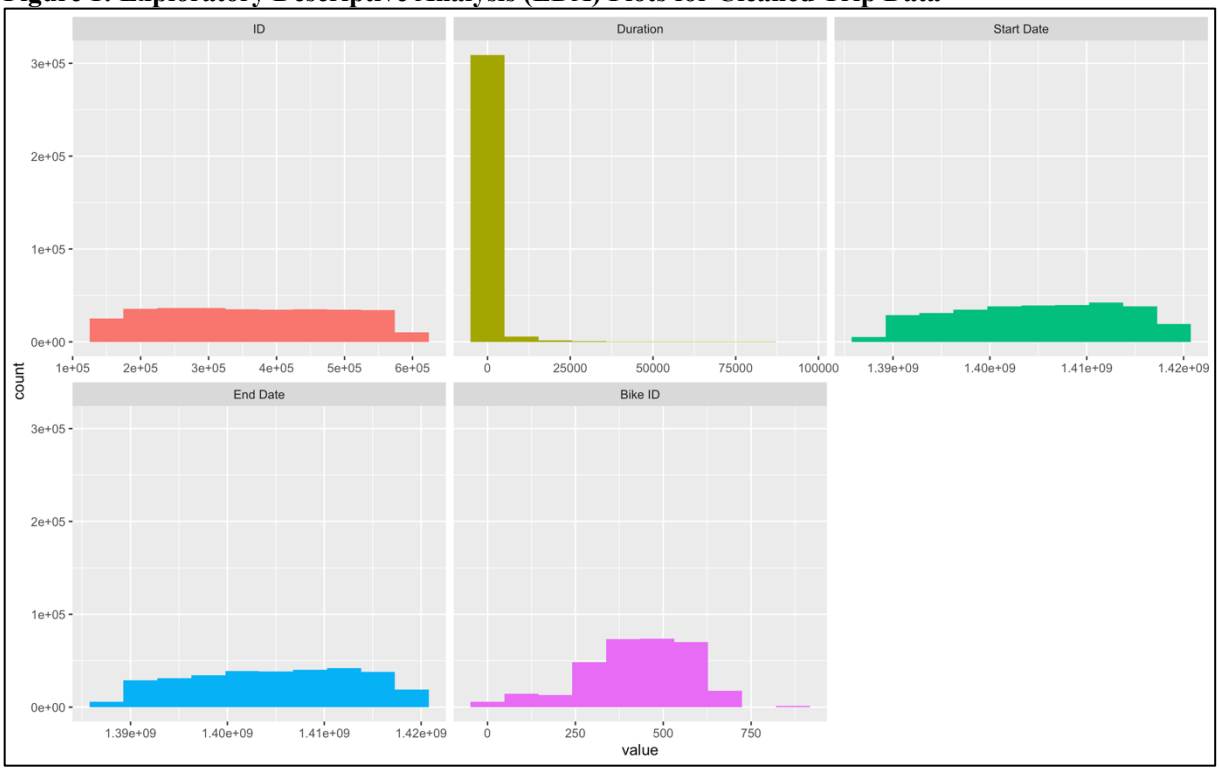

Figure 1: *EDA Histograms from left to right: ID frequency, trip duration frequency, trip start date frequency, trip end date frequency, and bike ID. Notably, the histograms for start and end dates of trips suggest that trips are relatively consistent throughout the year with the exception of fewer trips in the winter months of December and January. These plots were created using functions from the HMisc and funModeling packages in R.*

Figure 2: Histogram of Trip Duration in 2014

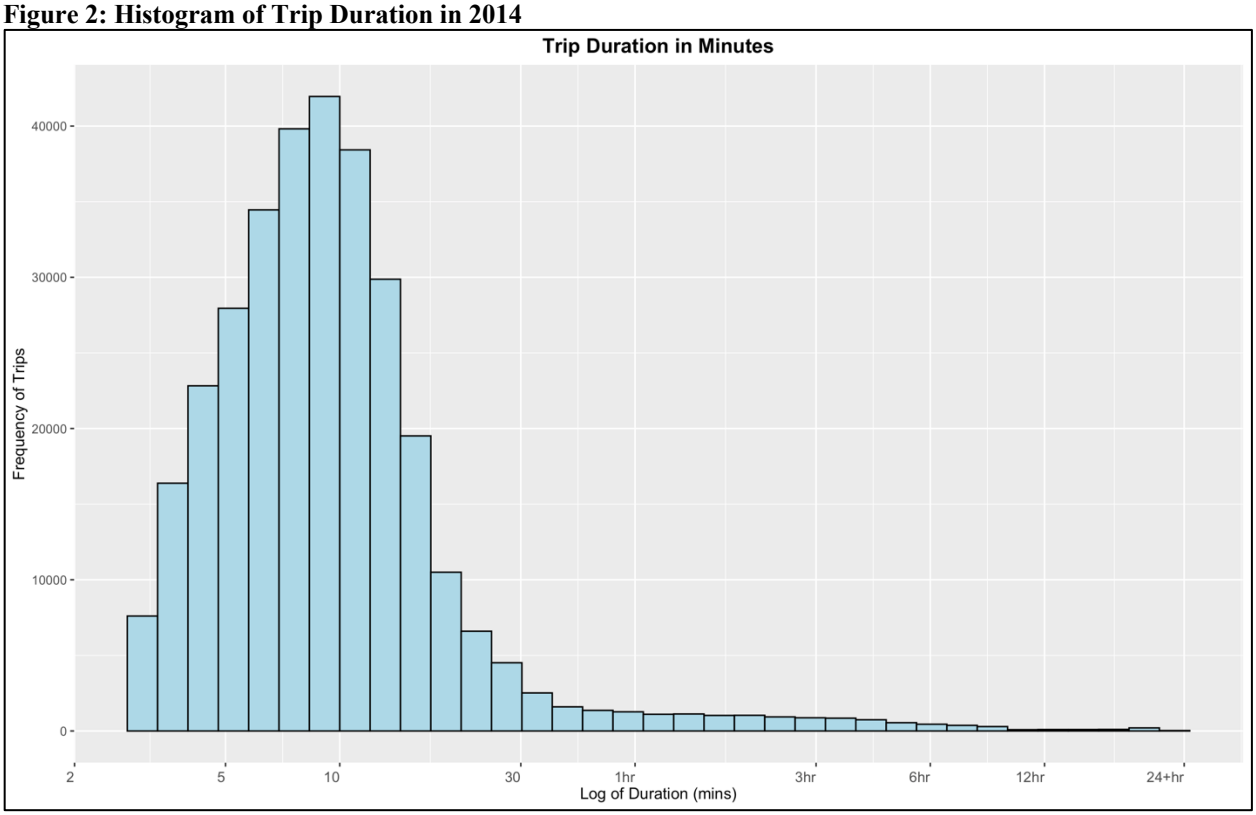

Figure 2: *Histogram of trip duration for Trip data set. This histogram displays the frequency of trips in 2014 by duration. This histogram displays log_10 transformed trip duration data for easier visualization. From this histogram, it is evident the majority of trips fall below a 30-minute duration threshold. This plot was created using functions in the ggplot2 package in R.*

Examining the most used stations across all of 2014, shown in figures 3 & 4, there is an overlap between the three most used start and end stations. San Francisco Caltrain (Townsend at 4th, ID:70), Harry Bridges Plaza (Ferry Building, ID:50), and San Francisco Caltrain (330 Townsend, ID:69) were the most used start and end stations in all of 2014.

Figure 3: Exploratory Descriptive Analysis (EDA) Plots for Cleaned Trip Data

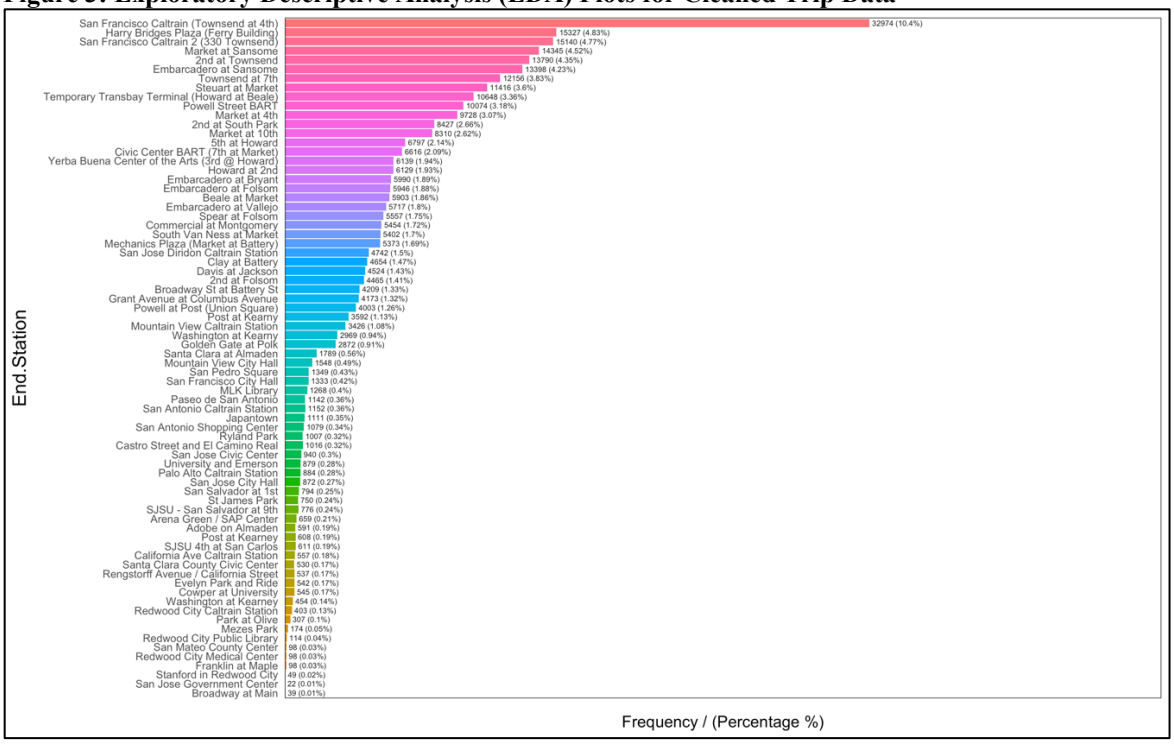

Figure 3: *Bar graph of SFBABS start station usage in 2014. This plot was created using functions in the ggplot2 package*

Figure 4: Exploratory Descriptive Analysis (EDA) Plots for Cleaned Trip Data

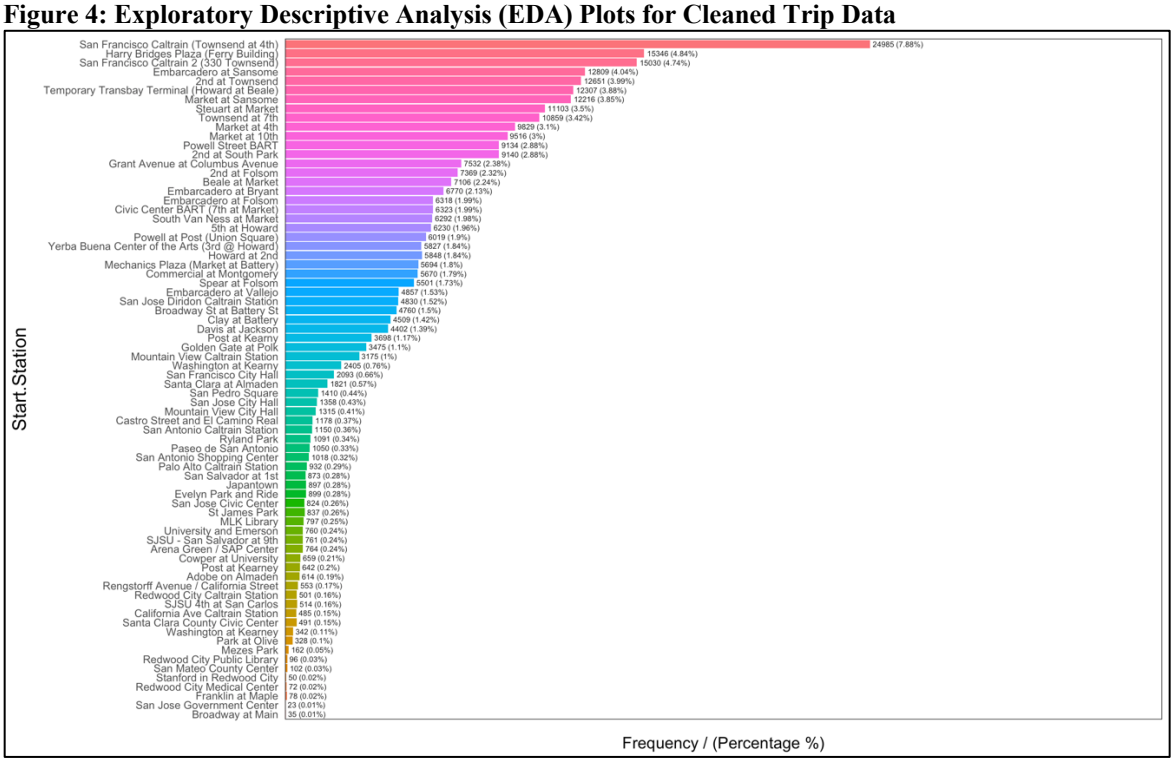

Figure 4: *Bar graph of SFBABS end station usage in 2014. Comparing figure 3 with figure 4, the topmost used stations are exactly the same: San Francisco Caltrain (Townsend at 4th, ID:70), Harry Bridges Plaza (Ferry Building, ID:50), and San Francisco Caltrain (330 Townsend, ID:69) This plot was created using functions in the ggplot2 package.*

Pulling insight from analysis desxcribed later in this report, figure 5 depicts the frequency of trips in each city, with over 80% of trips occuring in the city of San Francisco. Additionally, shown in figure 6, the majportiy (approx. 85%) of bike renters in 2014 are subscribers to SFBABS.

Figure 5: Trip Frequency by City in 2014

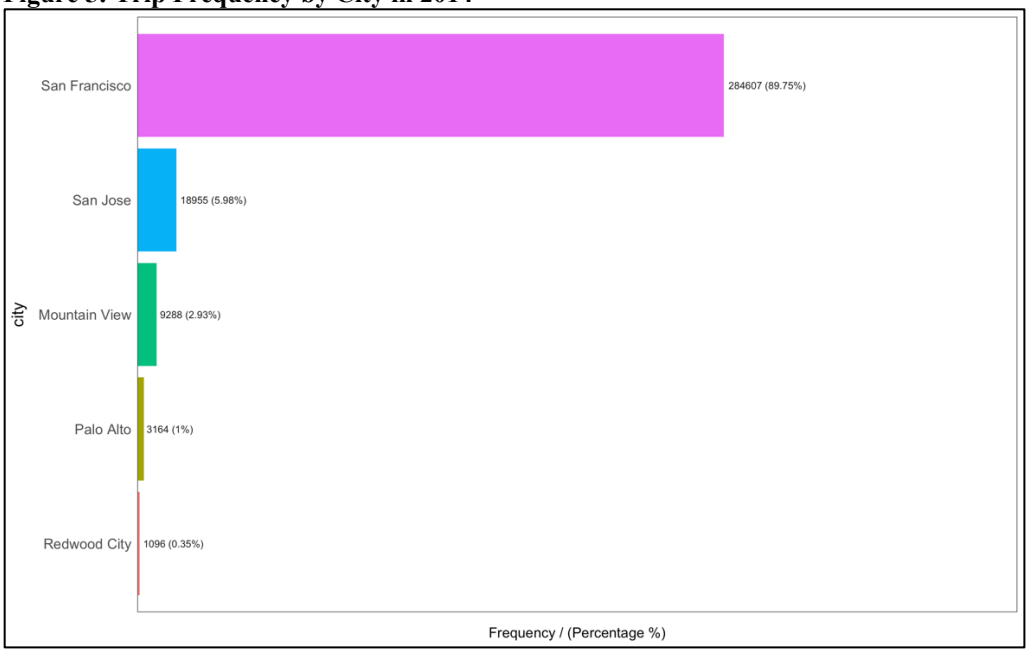

Figure 5: *Bar graph of trips per city in 2014. The majority of trips occurred in San Francisco as compared to the other 4 cities SFBABS operates in (San Jose, Mountain view, Palo Alto, and Redwood City). This plot was created using functions from the HMisc and funModeling packages in R.*

Figure 6: Proportion of subscribers as customers in 2014

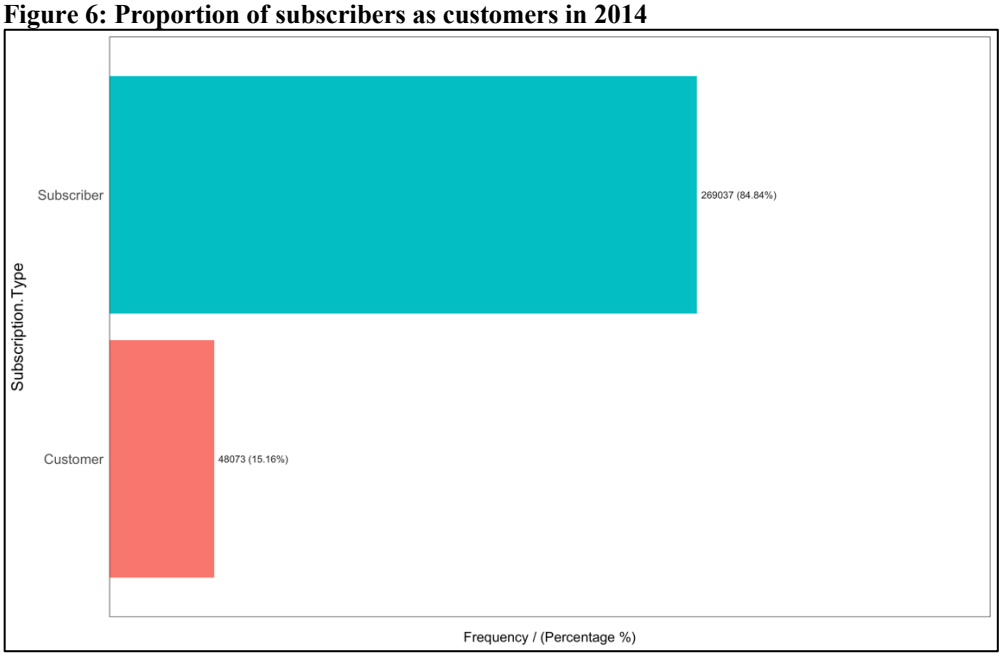

Figure 6: *Bar graph of 2014 trips by subscription type. This plot was created using functions from the HMisc and funModeling packages in R.*

Examining the weather data shown in Figure 7, the average daily temperature and visibility across all days of the year are approximately 62°F (16.68°C) and 9.9 miles, respectively. The temperature remained relatively consistent within a range of 45-65°F (7.22-18.33°C). Interestingly, there was very little variation in the average visibility, with an interquartile range (IQR) of [10, 10]. Figure 8 illustrates the weather events recorded in the five cities, which included rain, thunderstorms, fog, and a mix of events, with rain being the most commonly observed. Approximately 80% of the year saw no major weather events, correlating with the abundance of NAs in the recorded precipitation data.

Figure 7: Exploratory Descriptive Analysis (EDA) Plots for Cleaned Weather Data

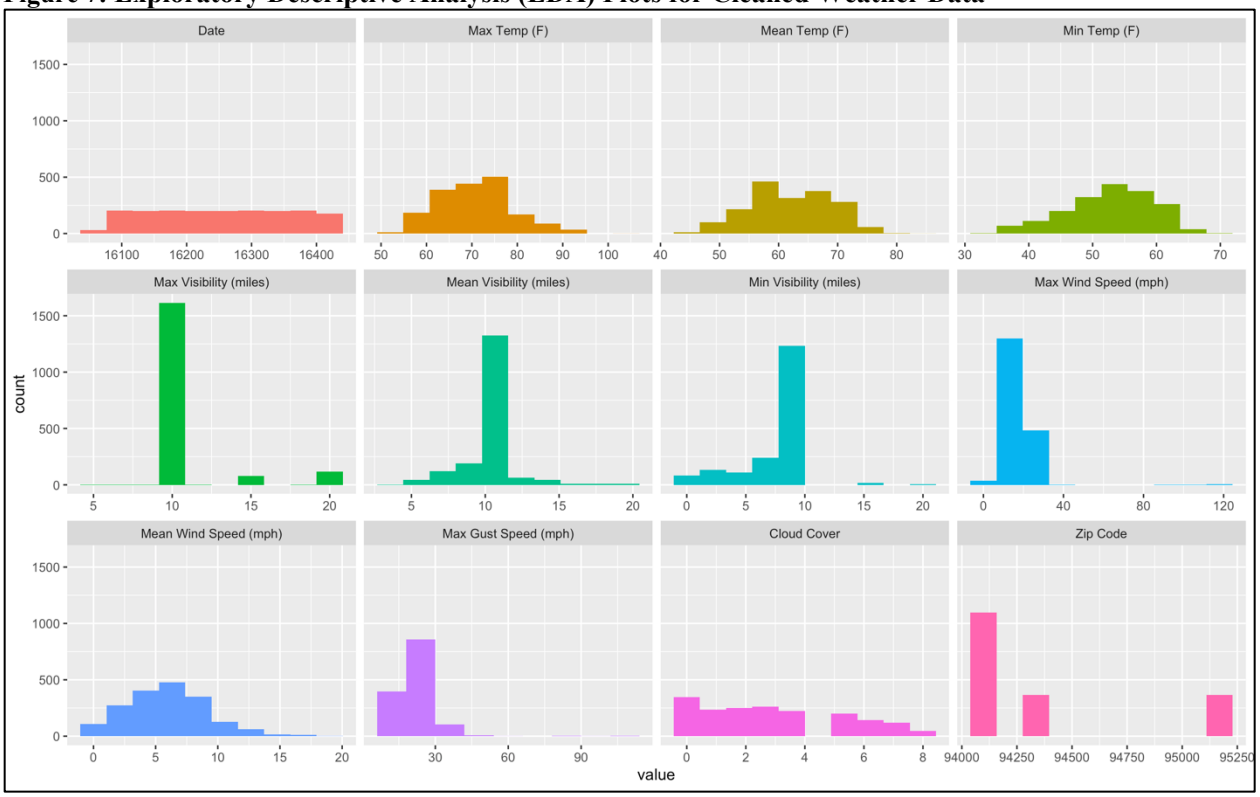

Figure 7: *EDA frequency histograms for daily weather data from left to right: Date, Max Temperature (°F), Mean Temperature (°F), Min Temperature (°F), Max Visibility (miles), Mean Visibility (miles), Min Visibility (miles), Max Wind Speed (mph), Mean Wind Speed (mph), Max Gust Speed (mph), Cloud Cover, and Zip Code. Examining the histograms for temperature suggest a mild temperature range of 45-65°F (7.22-18.33°C). Similarly, examining the histograms for visibility indicate little variation in daily visibility indicated by the similar clustered frequencies around 10 miles. These plots were created using functions from the HMisc and funModeling packages in R.*

Figure 8: Graph of Weather Events Across Cities in 2014

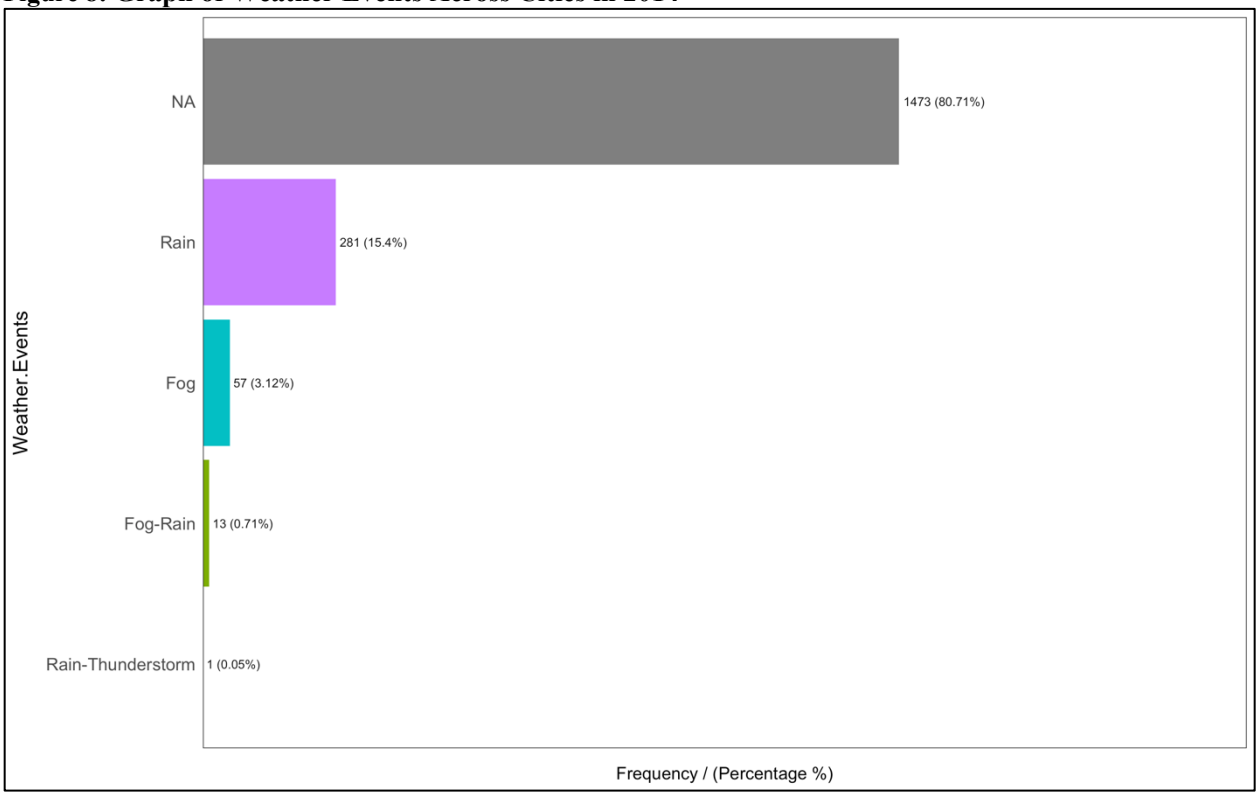

Figure 8: *Bar Graph of Frequency of Weather Events Across San Francisco, San Jose, Mountain View, Palo Alto, and Redwood City. Overall in 2014, the greater San Francisco area appears to have mild weather with few adverse weather events, the most common being rain. This plot were created using functions from the HMisc and funModeling packages in R.*

Key Findings

Overall, the data analysed in this report suggest that rush hours for bike rentals in the San Francisco Bay Area are approx. 7:00am – 9:00am in the morning and 4:00pm – 6:00pm in the evening. Figure 10 compares the frequency of trips per hour for each day of the week. As seen in figure 9, this trend holds true across all weekdays but does not hold for the weekends. Instead, for weekends the busiest hours are around 12:00pm – 3:00pm in the afternoon.

Figure 9: Plots of Bike Rentals Throughout the Day in 2014

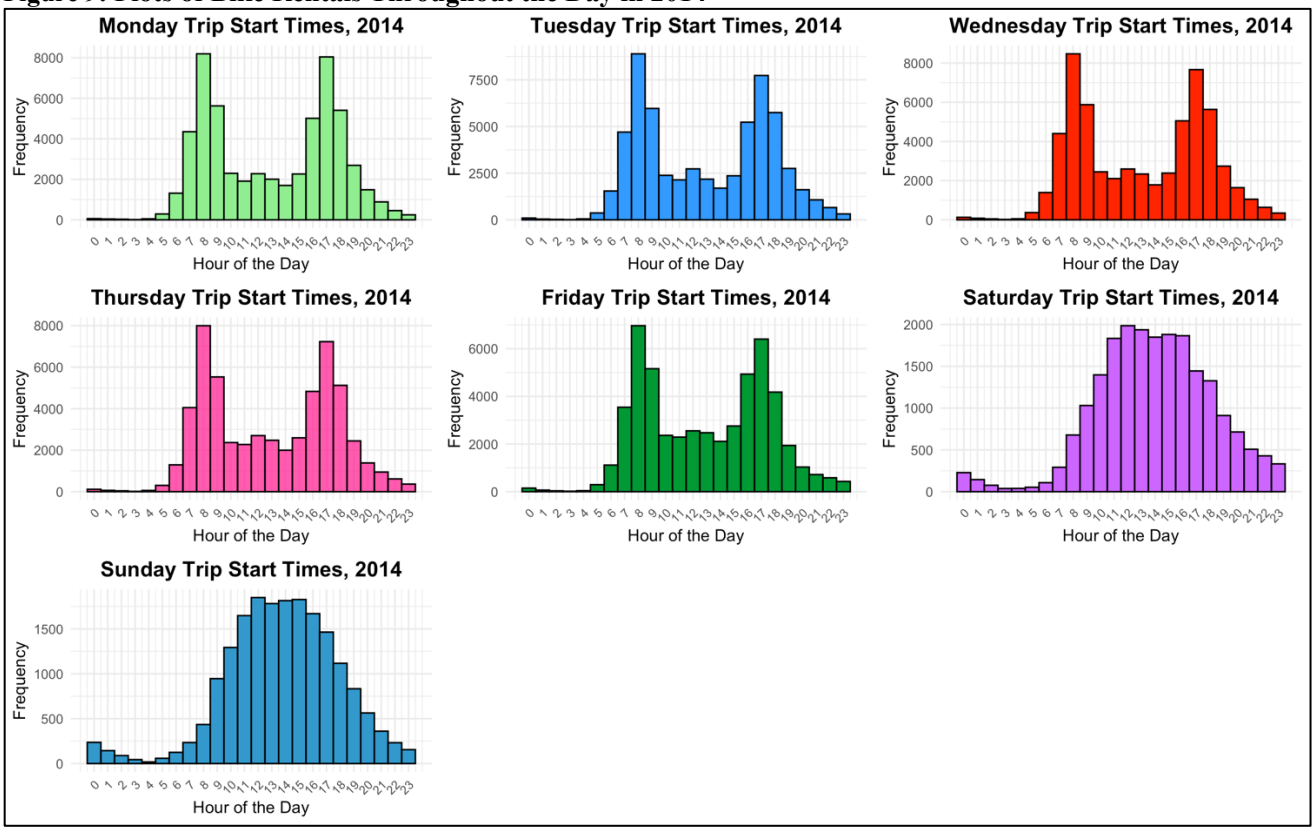

Figure 9: *Plots of Trip Frequency per Hour for Each Day of the Week in 2014. From left to right: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday Trip Start Times. This figure visualizes that rush hours are only on weekdays (M-F) between the hours of 7-9am and 4-6pm. These plots were made in R using functions from the ggplot2 package.*

The 10 most used stations during rush hour are listed in table 1, with corresponding graphs in appendix A. Unsurprisingly, there is significant overlap between the most frequented start and end stations, indicating that San Francisco Caltrain (Townsend at 4th) and San Francisco Caltrain 2 (330 Townsend) are likely some of the busiest stations in the SFBABS network.

Table 1: Ranking of Most Frequented Stations During Rush Hours in 2014

Rank	Start Station?	Station	ID	Frequency		Ending Station?	Station	ID	Frequency
1	Start	San Francisco Caltrain (Townsend at 4th)	70	19313		End	San Francisco Caltrain (Townsend at 4th)	70	26579
2	Start	San Francisco Caltrain 2 (330 Townsend)	69	10703		End	San Francisco Caltrain 2 (330 Townsend)	69	11380

3	Start	Temporary Transbay Terminal (Howard at Beale)	55	10399		End	Market at Sansome	77	9234
4	Start	Harry Bridges Plaza (Ferry Building)	50	8809		End	2nd at Townsend	61	9097
5	Start	2nd at Townsend	61	8289		End	Harry Bridges Plaza (Ferry Building)	50	8721
6	Start	Steuart at Market	74	7922		End	Temporary Transbay Terminal (Howard at Beale)	55	8622
7	Start	Market at Sansome	77	7246		End	Townsend at 7th	65	7786
8	Start	Townsend at 7th	65	7009		End	Steuart at Market	74	7670
9	Start	Embarcadero at Sansome	60	6322		End	Embarcadero at Sansome	60	5918
10	Start	Market at 10th	67	5944		End	San Francisco Caltrain (Townsend at 4th)	70	26579

Table 1: *Table of Start and End Stations During Rush Hours. This table lists the 10 most frequented stations during rush hours (7-9am & 4-6pm) in 2014. From this table, the data suggests that San Francisco Caltrain (Townsend at 4th) and San Francisco Caltrain 2 (330 Townsend) are the most used stations in 2014*

Similarly, the 10 most used stations during the weekends are also shown in table 2 with corresponding graphs in appendix B. Comparing table 1 and 2, there are significant differences in the most frequented stations suggesting different consumer trends and demands between rush hour and non-rush hour times. Moreover, there is a significant decrease in the overall volume of trips to each station with there being significantly less trips during the weekends.

Table 2: Ranking of Most Frequented Stations During Weekends in 2014

Rank	Start Station?	Station	ID	Frequency		Ending Station?	Station	ID	Frequency
1	Start	Harry Bridges Plaza (Ferry Building)	50	3164		End	Embarcadero at Sansome	60	3368

2	Start	Embarcadero at Sansome	60	3116		End	Harry Bridges Plaza (Ferry Building)	50	3174
3	Start	Market at 4th	76	1661		End	Market at 4th	76	1877
4	Start	Embarcadero at Bryant	54	1603		End	Powell Street BART	39	1677
5	Start	2nd at Townsend	61	1546		End	San Francisco Caltrain (Townsend at 4th)	70	1660
6	Start	Powell Street BART	39	1487		End	2nd at Townsend	61	1592
7	Start	San Francisco Caltrain (Townsend at 4th)	70	1361		End	Embarcadero at Bryant	54	1384
8	Start	Grant Avenue at Columbus Avenue	73	1299		End	Steuart at Market	74	1223
9	Start	Market at Sansome	77	1095		End	Market at Sansome	77	1110
10	Start	Powell at Post (Union Square)	71	1090		End	Grant Avenue at Columbus Avenue	73	1098

Table 2: *Table of Start and End Stations During Weekends. This table lists the 10 most frequented stations during the weekends in 2014. From this table, the data suggests that there are significant differences in the most used stations between peak and non-peak hours.*

The average utilisation rate was calculated by:

$$Average\ utilisation\ rate\ per\ month = \frac{Total\ bike\ time\ per\ month}{Total\ time\ available\ oper\ month}$$
$$= \frac{\sum(Trip\ duration\ (secs))}{(Days\ in\ month) \cdot (24) \cdot (60) \cdot (60) \cdot (Total\ number\ of\ bikes)}$$

Effectively, this report calculates average utilisation as the percentage of the total bike time per month during which a bike was being ridden.

Shown in figure 10, the average utilisation rate per month was calculated with the highest utilisation of 1.83% occurring in August of 2014. The summer months of June – September 2014 had the average highest utilisation, with rates above 1.7%. This correlates with previous

exploratory analysis suggesting the highest number of trips per month occurred in the summer months.

Figure 10: Average Monthly Utilisation Rate in 2014

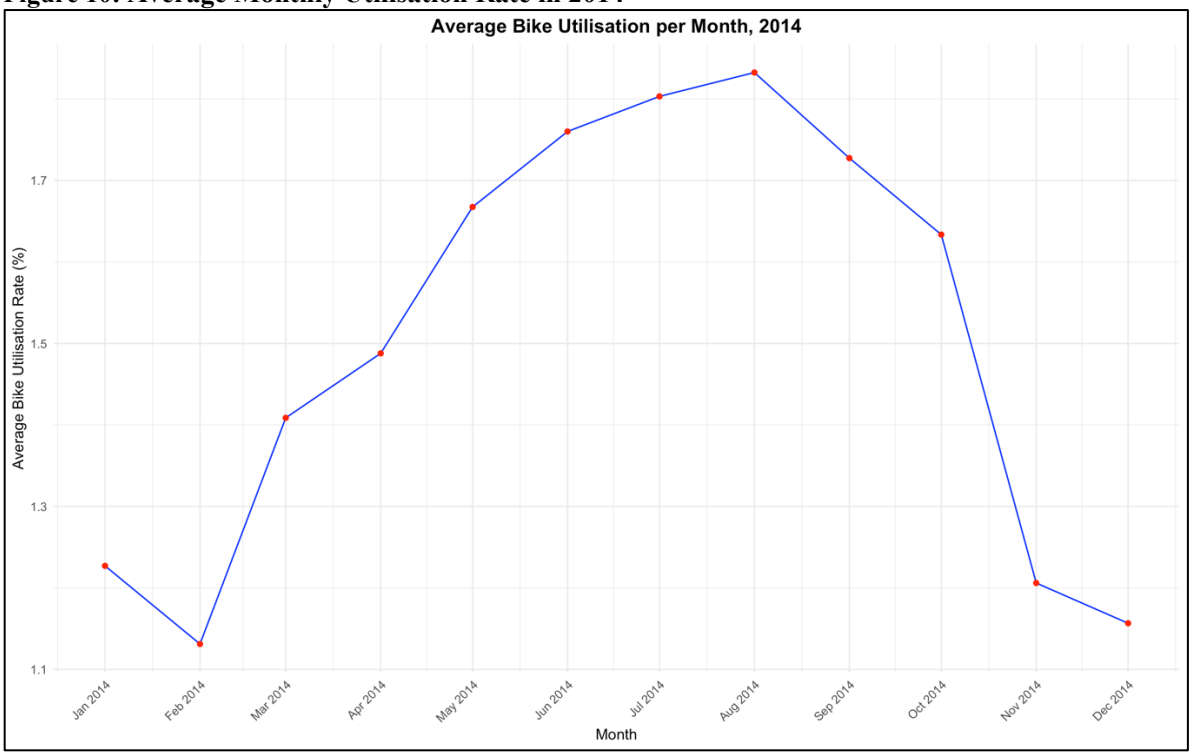

Figure 10: *Plot of Average Utilisation Rate per Month in 2014. This plot graphs the average utilisation time (total trip duration/total bike time available) per month. From this analysis, it suggests that August has the highest utilisation rate of approx. 1.83% with other summer months (June to September) having higher average utilisation compared to the rest of the year. This plot was made in R using functions from the ggplot2 package.*

Examining the relationship between trip and weather data, this report focuses on the correlation between a trip's duration and various weather variables. Most variables within the trip dataset are categorical and/or factored, making it difficult to accurately calculate a correlation coefficient. For example, even if city or station ID are not considered factors, there is no logical way to order the levels and therefore no accurate way to calculate a correlation calculation. The correlation matrix between trip duration, start date, and various weather variables is shown in Figure 11, with the highest correlations listed in Table 3. Importantly, these correlations were computed using pairwise observations due to the high number of missingness/NAs in certain variables such as Weather Events. Utilising pairwise observations for when calculating correlation coefficients ensures the maximum number of observations will be computed, maintaining integrity of the data analysis despite the high degree of missingness clustered in some variables.

Table 3: Correlations between Trip and Weather Variables

Weather Variable	Direction w/ Duration	Duration Correlation Coefficient
Max Temp F	+	0.00505
Mean Temp F	+	0.00217
Min Temp F	-	-0.00139

Max Visibilty (miles)	+	**0.0556**
Mean Visibility (miles)	+	**0.0278**
Min Visibility (miles)	+	**0.0176**
Max Wind Speed (mph)	-	-0.00557
Mean Wind Speed (mph)	+	0.000791
Max Gust Speed (mph)	-	-0.00941
Precipitation (Inches)	-	-0.00298
Cloud Cover	-	-0.0268
Weather Events	-	-0.0154

Table 3: *Table of Correlation Coefficients Between Trip and Weather Variables. Overall, there are no variables with a strong correlation with trip duration. The weather variables that have the strongest correlation to trip duration are Max, Mean, and Min Visibility. Importantly, these correlations were computed using pairwise observations due to the high number of missingness/NAs in certain variables such as Weather Events.*

Figure 11: Correlation Matrix for Trip and Weather Data

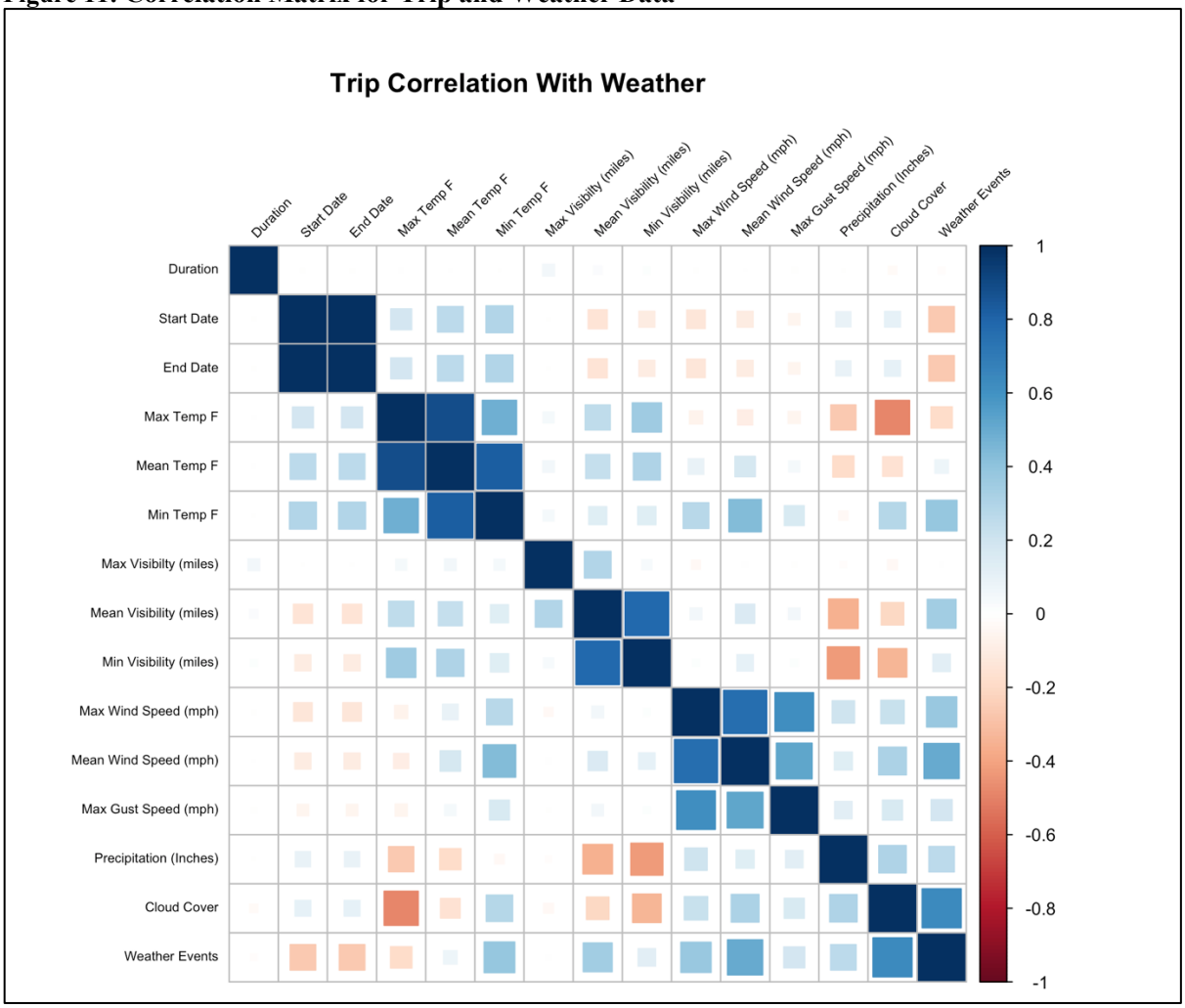

Figure 11: *Correlation Matrix for Trip and Weather Data. Overall there is no significant correlation between any weather variable and trip duration. The highest and lowest correlations with trip duration are Max Visibility (0.0556) and Weather Events (-0.0154). Importantly, Weather Events was coded in terms of severity with Fog = 1, Fog-Rain = 2, Rain = 3, and Rain-Thunderstorms = 4 which explain the negative correlation of trip duration with more severe weather events. This plot was made in R using the corrplot() function from corrplot() package.*

Discussion

Examining the differences in bike rentals between rush hours and weekends can provide valuable insights into consumer trends and demand. Stations saw significantly more usage during rush hours, logging over 20,000 trips during rush hours as compared to approx. 3000 on weekends. The large difference in volume and significant differences in most frequented stations suggest different consumer usage trends between weekday peak hours and weekends. This likely suggests that SFBABS is currently being used a commuting method for consumers in addition to providing more leisurely and/or sight-seeing biking activities.

The significant differences in overall usage on weekends compared to weekdays suggest that larger windows for maintenance could be scheduled during these times. For instance, maintenance for busy weekend stations could be scheduled during the weekdays and vice versa. Additionally, scheduling maintenance during January or December might be advantageous due to lower utilization rates during these months as the weather in these winter months, ranging between 45-65°F (7.22-18.33°C), is unlikely to hinder maintenance activities.

Our analysis revealed no strong correlations between any weather variables and trip duration, likely suggesting a much more complex relationship between daily weather and a bike trip. Rather this relationship is likely influenced by multiple factors beyond just weather. The highest correlations were with maximum visibility (miles), mean visibility (miles), and minimum visibility (miles), with correlation coefficients of 0.0556, 0.0278, and 0.0176, respectively. These correlations were computed using pairwise observations, ensuring the maximum number of observations were included despite the high degree of missingness in variables such as “Weather Events” and “Precipitation”. Given the mountainous geography and renown scenic views of the San Francisco area, the strong correlation of trip duration with visibility variables may be related to consumers being more inclined to bike as a method of sightseeing.

Despite low correlation coefficients, the overall directions of the correlation coefficients aligned with expectations. Negative correlation coefficients were observed for maximum wind speed, maximum gust speed, precipitation, and weather events, indicating that these factors negatively impact trip duration. Conversely, positive correlations were found for mean temperature and maximum temperature, reflecting that biking is more popular during warmer weather. Weather events were factored with respect to severity (Fog = 1, Fog-Rain = 2, Rain = 3, and Rain-Thunderstorms = 4), so the negative correlation is expected, as more severe weather conditions likely affect outdoor activities such as biking. Importantly, the high degree of missingness likely significantly hindered the accuracy of the calculated correlation coefficients for weather events and precipitation as previous EDA indicated approximately greater than 80% of the 2014 logged no weather events (i.e. sunny skies). Despite accounting for the lack of data through conducting pairwise observations for correlation, the correlation coefficient will likely significantly with more data from additional years.

While this report’s data has been cleaned for downstream analysis, it is important to note that trip duration was recorded in seconds and also stored in POSIX format. While POSIX format is advantageous for data integrity and calculation, it can be difficult to interpret in downstream

analysis. Therefore, trip duration as a measure of seconds was retained for easier comprehension of analysis and subsequent work.

Appendices Table of Contents

Appendix A: Trip Stations During Rush Hours in 2014

A1. Introduction

This appendix contains the graphs showing the frequency of start and end stations during rush hours for SFBABS stations in 2014.

A2. Start Stations During Rush Hour

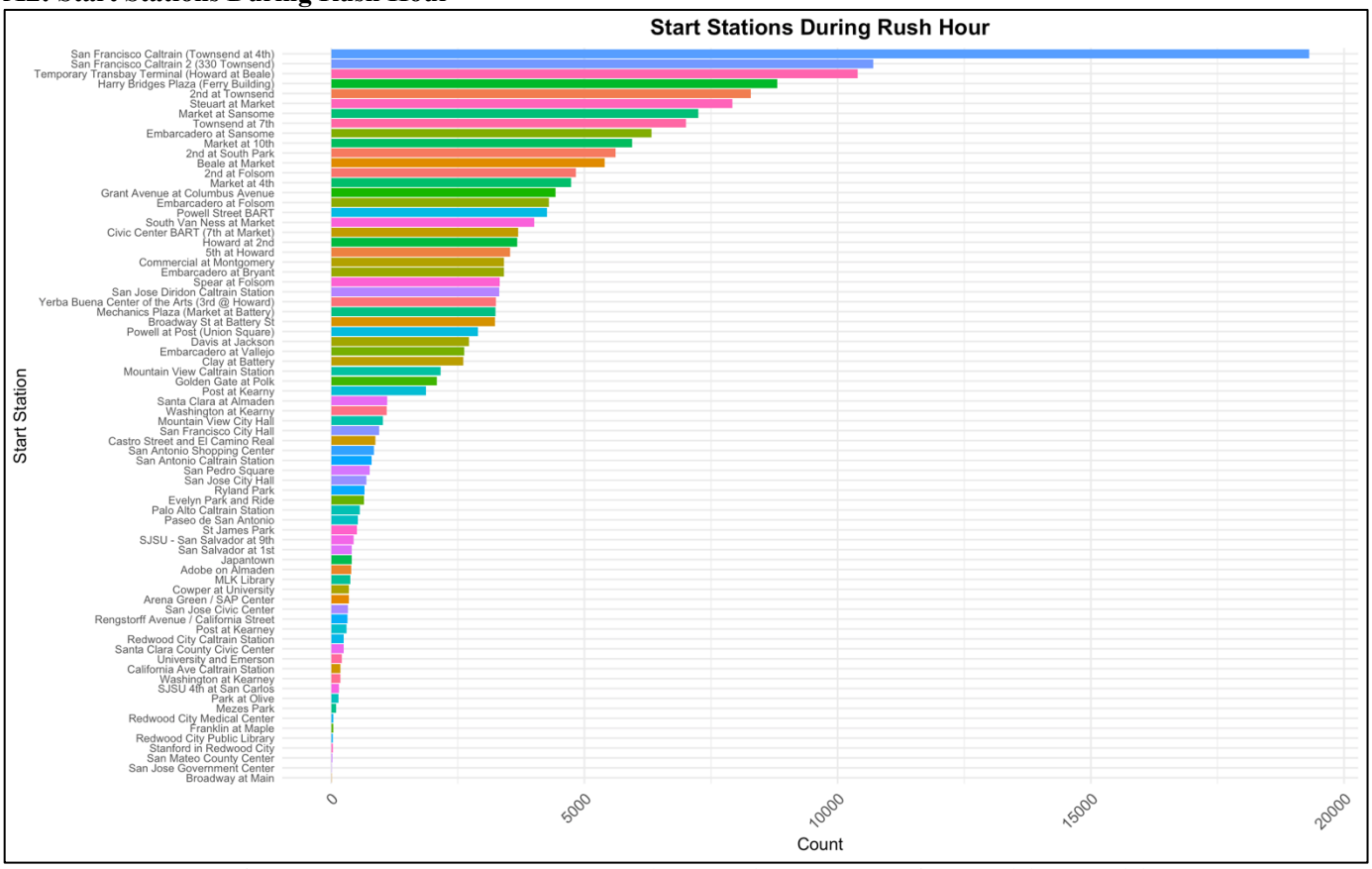

Figure A2: *Histogram of Start Station Frequency During 2014 Rush Hours. The top 3 stations frequented during rush hours see significantly more use than other stations, with San Francisco Caltrain (Townsend at 4th), San Francisco Caltrain 2 (330 Townsend, and Temporary Transbay Terminal (Horward at Beale) logging over 10,000 trips during rush hours over the course of 2014. This plot was made in R using functions from the ggplot2 package.*

A3. End Stations During Rush Hour

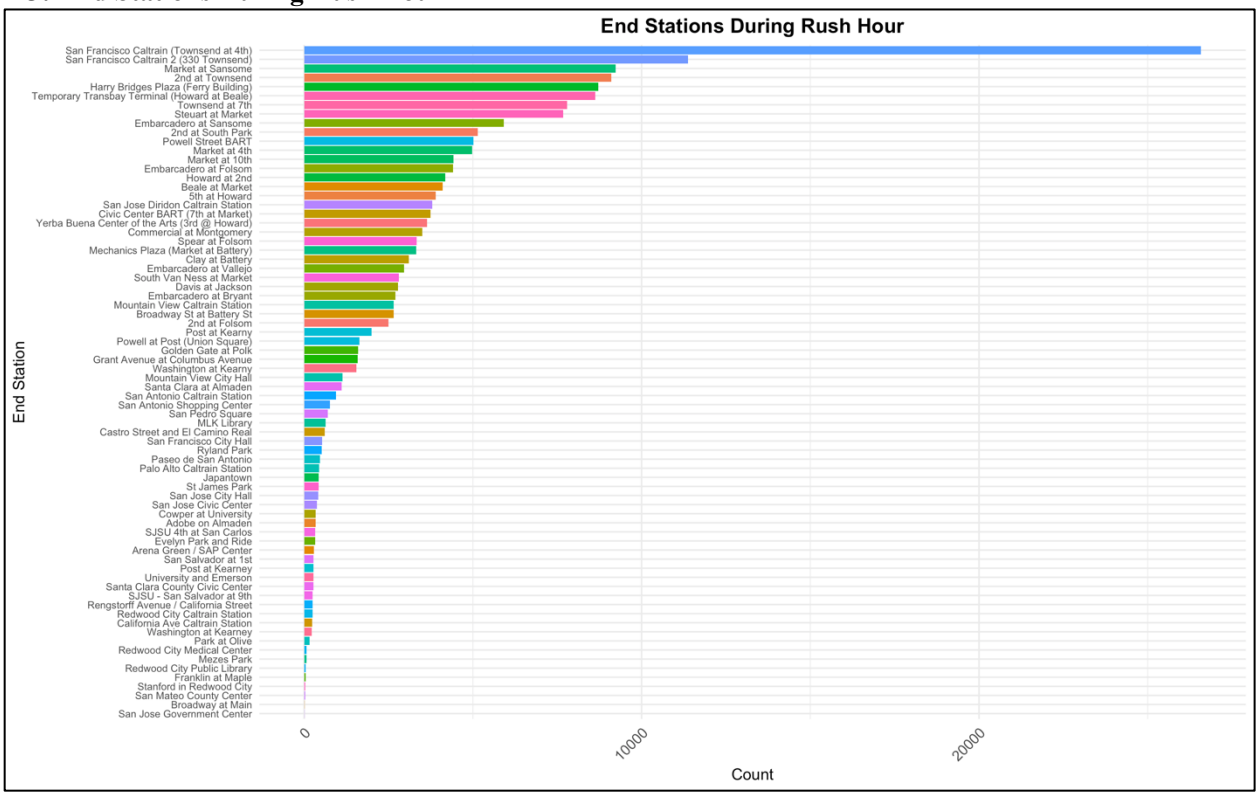

Figure A3: *Histogram of End Station Frequency During 2014 Rush Hours. Similar to figure A2, there is significant overlap between the most used start and end stations during rush hours in 2014. This plot was made in R using functions from the ggplot2 package.*

Appendix B: Station Frequency During Weekends in 2014

B1. Introduction

This appendix contains the graphs showing the frequency of start and end stations during weekends for SFBABS stations in 2014.

B2. Start Stations During Weekends

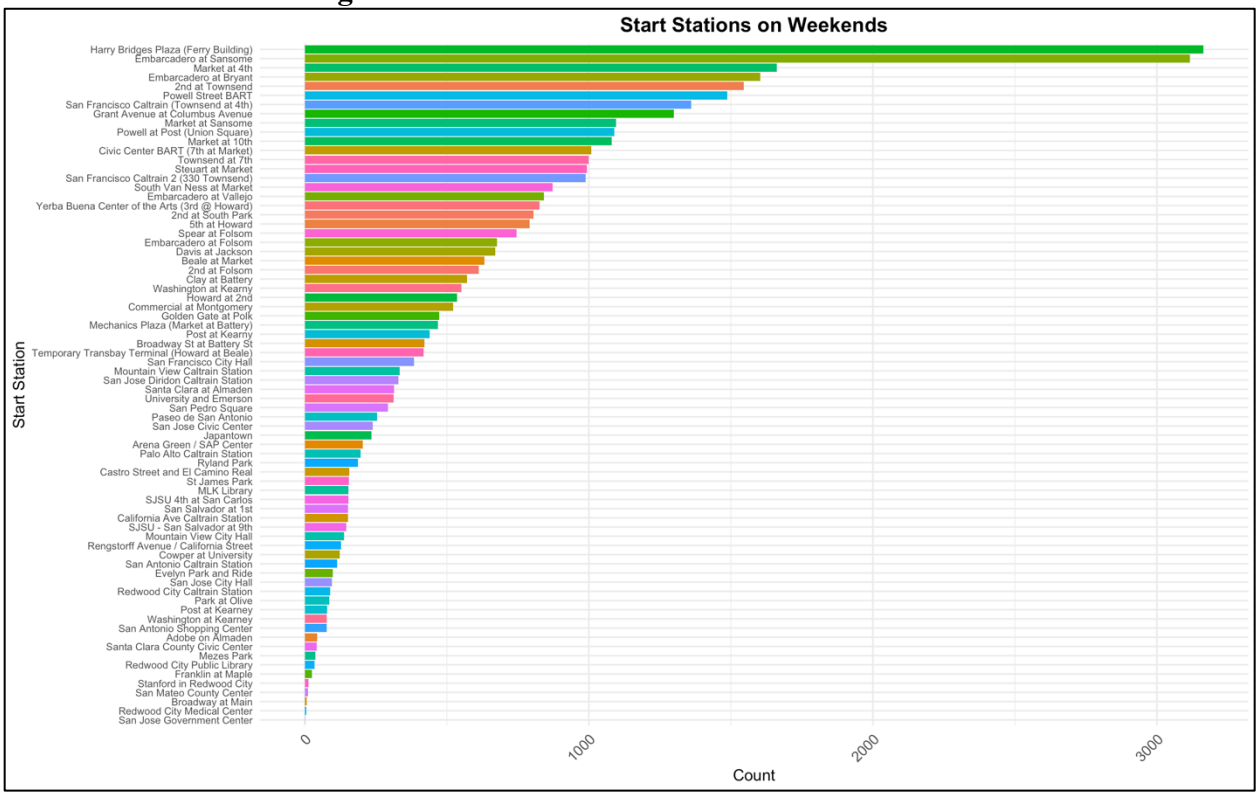

Figure B2: *Histogram of Start Station Frequency During Weekends. Comparing to figures A1 and A2, there is significant differences in the number of trips and most used stations between rush hours and the weekends. The top 3 most frequented stations during the weekends are Harry Bridges Plaza (Ferry Building), Embarcadero at Sansome, and Market at 4th. This plot was made in R using functions from the ggplot2 package.*

B3. End Stations During Weekends

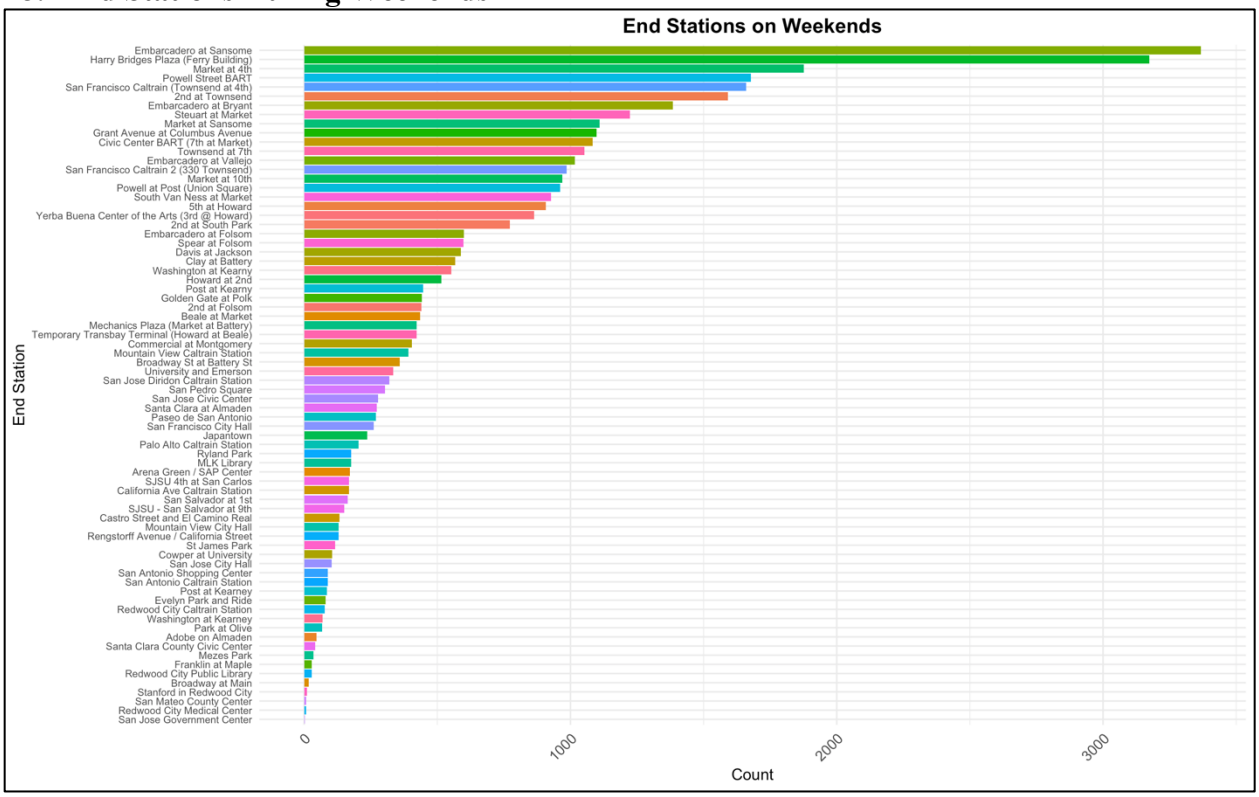

Figure B3: *Histogram of Start Station Frequency During Weekends. Comparing to figures A1 and A2, there is significant differences in the number of trips and most used stations between rush hours and the weekends. The top 3 most frequented end stations during the weekends are the same as the top 3 most used start stations: Harry Bridges Plaza (Ferry Building), Embarcadero at Sansome, and Market at 4th. As there are distinct differences in bike rental between rush hours and weekends, this may suggest different consumer trends between peak hours on the weekends vs weekdays. This plot was made in R using functions from the ggplot2 package.*