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

**Mingwei Cui: 1007039018**

**Exploratory Data Analysis**

**Trip data**

The trip data stored in “trip.csv” contains 326, 339 observations with 11 columns. Those columns are Id, Duration, Start\_date , Start\_station\_name , Start\_station\_id, End\_date . End\_station\_id, End\_station\_name , Bike\_id, Subscription \_type and zip\_code

With status() in R we found that all variables are properly stored with no NAs and empty cells except for the column zip\_code. We would not be focusing on zip\_code for the sake of this analysis, therefore would not be working on missing values. To begin with our EDA we would like to look at the variable trip duration. Trip durations were found to have mean and median of 1132 and 551. However plotting duration with a boxplot in figure 1. suggest that there exist some extreme values. Those extreme values would be dealt with later when we discuss our strategy with outliers.

A white rectangle with black lines

Description automatically generated

Figure 1. Boxplot with duration of bike rides in Bay Area.

Next, we move on and inspect the start and stop stations. Caltrain station is most popular start and end station followed by Harry Bridges Plaza and Caltrain 2 station. The top 10 station are mostly shared between staring and sharing stations as shown in figure 2.

|  |  |
| --- | --- |
| A |  |
| B |  |

Figure 2. A Top 10 most frequent start stations by bike trips. B, Top 10 most frequent end stations by bike trips.

Lastly, we would like to look at subscription levels of users in this data. Most bikers in the dataset are subscribers with only around 15% of customers as shown in figure 3.

A graph with a number of squares and numbers

Description automatically generated with medium confidence

Figure 3. Level of subscription of users in Bay area bike rental.

**Weather Data**

The weather data stored in “weather.csv” contains 1825 observations, with 15 columns. Figure 4 shows all numeric variables displayed in a histogram.

A screenshot of a graph

Description automatically generated

Figure 4. Overview of weather data contains histogram of for all numeric variables in the dataset.

**Cancelled Rides and Outliers**

We remove all the invalid observations such as outliers and cancelled rides. Cancelled rides are observation that last shorter than 180 in duration (3 minutes) and start/end as the same station. We identified 1082 of those observations and removed them from the dataset.

To identify outliers, we defined our boundaries as 1.5\*IQR from the 1st and 3rd quantile. Any value beyond this would be considered as invalid. The trip\_id of outliers and cancelled rides are too long or too short therefore could not be included in this report. Removed trips have ID stored in “cancelled.id” and “out.id” in the R script. After removal of these observations let us display the boxplot again. Compared to figure1 it is obvious that some extreme values have been removed to give a more sensible boxplot shown in in figure 5. Cancelled rides are not meaningful as they are not considered as an actual ride. Outliers, specifically those that have unusually long durations (i.e. 20hrs + ) , are also not considered. Those were likely due to glitches in the system, or unreturned bikes and are not representative of actual rental profile in the Bay Area.

A diagram of a bike ride

Description automatically generated

Figure 5. Boxplot of Ride Durations with cancelled rides and outliers removed.

**Rush-hours and Weekends**After removing cancelled trips and outliers, we intend to identify rush hours. According to figure 6, I have identified rush hours as 6-8 in the morning and 15-17 in the afternoon from Monday to Friday. During weekdays, rush hours of bike rental matches working hours. This suggest that bike rides during weekdays are considered a method of commute.

A graph of a number of different colored bars

Description automatically generated with medium confidence

**Figure 6. Histogram of start times and end times of bike rides overlaid during weekdays**. Start and end time both peaks between 6-8 in the morning and 15-17 in the afternoon. Start time is displayed in purple-blue bars while end time is displayed in pink bars. Overlapping bars are shown in purple.

|  |  |  |  |
| --- | --- | --- | --- |
| Top 10 Most Frequent Starting Station | | Top 10 Most Frequent ending Station | |
| Weekends | Rush-hours | Weekends | Rush hours |
| #1 Embarcadero at Sansome: 2145 | 1# San Francisco Caltrain (Townsend at 4th): 14072 | #1 Harry Bridges Plaza (Ferry Building): 2344 | #1 San Francisco Caltrain (Townsend at 4th) 12843 |
| #2 Harry Bridges Plaza (Ferry Building): 1924 | #2 Temporary Transbay Terminal (Howard at Beale): 7907 | #2 Embarcadero at Sansome: 1664 | #2 Temporary Transbay Terminal (Howard at Beale) 7126 |
| #3 Market at 4th: 1266 | #3 San Francisco Caltrain 2 (330 Townsend) : 7185 | #3 Market at 4th: 1507 | #3 San Francisco Caltrain 2 (330 Townsend) 6493 |
| #4 2nd at Townsend:1232 | #4 Harry Bridges Plaza (Ferry Building) : 6554 | #4 Powell Street BART: 1378 | #4 Harry Bridges Plaza (Ferry Building) 5642 |
| #5 Embarcadero at Bryant :1232 | #5 2nd at Townsend : 5807 | #5 San Francisco Caltrain (Townsend at 4th): 1355 | #5 2nd at Townsend 5381 |
| #6 Powell Street BART :1147 | #6 Steuart at Market :5624 | #6 2nd at Townsend: 1269 | #6 Steuart at Market 5063 |
| #7 San Francisco Caltrain (Townsend at 4th):1080 | #7 Townsend at 7th : 5058 | #7 Embarcadero at Bryant: 1125 | #7 Townsend at 7th 4748 |
| #8 Grant Avenue at Columbus Avenue :1028 | #8 Market at Sansome : 4872 | #8 Steuart at Market: 976 | #8 Market at Sansome 4504 |
| #9 Market at 10th :877 | #9 Embarcadero at Sansome : 4385 | #9 Townsend at 7th : 922 | #9 Embarcadero at Sansome 4145 |
| #10 San Francisco Caltrain 2 (330 Townsend) :871 | #10 Market at 10th : 3870 | #10 Market at Sansome : 914 | #10 2nd at South Park 3570 |

**Table1. Top 10 busiest starting and ending station during rush hours and weekends.**

Based on the established rush hours we were able to identify the top 10 busiest start and end station during rush hours and weekends. Those stations are listed in table 1. Rush hour starting and ending stations are very much the same sets of stations.

**Utilization and Correlation**

**A graph of a bike

Description automatically generated with medium confidence**

**Figure 7. Bar graph of Utilization ratio of bikes in Bay Area by Month.** Utilization ratio is given by the “total time of bike used”/ “total time in months”. Given this table it could be seen that utilization is highest during late summer and early fall.

To calculate utilization, observations (rides) are first grouped by months to calculate bike utilization by month. Given this figure 7 it could be seen that utilization is highest during late summer and early fall. Utilization significantly dips in the month of January, February, and December, potentially due to the winter season. This change in utilization ratio by month could therefore be explain by the change in weather patterns.

A screenshot of a computer

Description automatically generated

**Figure 8. Correlation plot of rental pattern and weather data**. Level of correlation is given by the intensity of the colours. Blue indicates positive correlation while red indicates negative correlation.

Based on figure 8, we are able to investigate the correlation between trip patterns and weater data. For this purpose of this analysis, we would not discuss correlations between weather variables. “#trips” and “Total.duration” were two columns derived from trip.data and they strongly correlate with each other. This correlation is intuitive as increase in number of trips would likely increase the total duration of riders on bikes. Both variables have a weak negative association with maximum and mean visibility which is odd. As we would think higher visibility is indicative of good weather, we should expect a positive associative between visibility and level of bike rental. Precipitation and Events have a very weak negative association with total.duration and #trips which is also unusual. I would expect a strong negative association between them as the weather events it what people are most concerned about when it comes to outdoor activities. Other than the ones already mentioned, I did not find any other meaningful correlations which is not what I expected.