BTC1855 Midterm Project

Angela Bakaj

“Bay Area Bike Rental Operation Research”

**Objective:** The data science team is interested in creating a predictive model that will predict the number of bikes leaving each station and the number of bikes being returned to each station in the next three days. The insights from this report will help the maintenance teams to better plan their bike/dock maintenance operations.

**Contents:**

1. Exploratory Data Analysis (EDA) on all the datasets in question
2. Data Cleaning Summary
3. “Rush Hours” Analysis
4. Weekend Analysis
5. Average Bike Utilization
6. Weather Condition – Rental Activity Correlations

**1) Exploratory Data Analysis (EDA)**

The following EDA was performed on each dataset (including the “station\_data” for good measure) to summarize the important characteristics, assess missingness, and identify outliers. The associated figures were all created using the “funModelling” package in R and are located in the end of this report under the **appendix**. The “describe()” function in the “Hmisc” package in R was particularly helpful in summarizing the features of the data to inform cleaning.

1. **Station Data**

|  |  |
| --- | --- |
| ***Features*** | 70 observations (meaning 70 stations recorded) across 7 different variables including the station ID, the name, the latitude and longitude, the dock count, the city, and the installation date. |
| ***describe( ) function insights*** | No missing data across the entire frame |

1. **Trip Data**

|  |  |
| --- | --- |
| ***Features*** | 326339 observations (meaning this many individual recorded trips) across 13 different variables including the trip ID, the duration in seconds, the start and end date, the starting and ending station names and id, the bike ID, the subscription type, the zip code, and the start hour and day of the week. |
| ***funModelling plots*** | Figure **1** & **2** respectively represent the frequency distributions of starting and ending stations for the trips. The three most frequent stations are consistent between the figures.  Figure **3** depicts the frequency of both subscription types for the trips; majority (85.1%) were subscribers and only 14.9% were customers.  Figure **4** depicts the numeric variables including ID, duration, start and end station IDs, and bike ID. The plots of the starting and ending station IDs are quite similar, indicating that the majority of the trips have ended at the same station in which they had started. The bike ID plot also offers meaningful insight as this plot depicts the frequency of use for each individual bike. |
| ***describe( ) function insights*** | The lowest durations were 60 seconds and the highest was 17270400 seconds. Missing and invalid data entries were numerous (14251 of them) in the zip-codes. |

1. **Weather Data**

|  |  |
| --- | --- |
| ***Features*** | 1825 observations (meaning an observation per day for a year for 5 cities) across 15 variables including the date, the maximum, mean, and min temperature, the maximum, mean, and min visibility in miles, the maximum, mean, and min wind speed in miles per hour, the precipitation in inches, the cloud cover, events, zip code, and city. |
| ***funModelling plots*** | Figure **5** depicts the frequency of events reported in the dataset; however, the vast majority (80.7%) were missing entries.  Figure **6** depicts the frequencies of precipitation recorded, with the vast majority (84.6%) being 0 inches in precipitation.  Figure **7** depicts the numeric values in the weather dataset including maximum, minimum and mean values for each temperature, visibility in miles, wind speed in miles per hour, as well as cloud coverage and zip code. |
| ***describe( ) function insights*** | 9 missing data points in visibility variables, 451 missing data points in max gust speed variable, and a numerous amount (1473) of missing entries for the events column. |

**2) Data Cleaning Summary**

1. **Station Data**

As is, the station data was relatively clean. There were no missing or invalid values. In anticipation of downstream analysis, the date entries in “installation\_date” were changed to POSIX format.

1. **Trip Data**

The trip data required several preprocessing steps as “nil” values, long zip codes, and letter-containing zip codes were each found in the “zip\_codes” column. To limit such zip codes to valid American zip codes, according to Google, valid American zip codes are between the highest value of 99950 in Ketchikan, AK and the lowest value of 00501 in Holtsville, NY.

The “duration” column has some invalid values and outliers to handle. Trips under 3 minutes as those are considered cancelled trips, so all values less than 180 seconds were set to NAs. There also existed some very, very large durations that span over days. To limit these, I chose to take the time required to travel the distance between two cities with the farthest distance, being San Francisco and San Jose and 5 hours in duration by bike, and multiply that time by 2, being 10 hours or 36,000 seconds. This was done in efforts to replicate the longest simple trip that may be taken in a single day. The 99th percentile for the duration in seconds is 22,000; meaning, then by limiting the duration to 36,000 or less, I have not severely reduced my dataset unreasonably.

Again, in anticipation of downstream analysis, the date entries in both “start\_date” and “end\_date” columns were changed to POSIX format.

1. **Weather Data**

Evidently, there was an extreme amount of missing data in the “events” column, which all have been assigned to NA.

In anticipation of potential downstream analysis, the columns, “city”, “events” and “cloud cover” were each converted to factor form.

A presence of many “T” values in precipitation were evident. Online search shows that this may represent “trace” values of precipitation which mean less than 0.01. To account in some way for these trace values, instead of setting them to NA, I chose to assign them to values of 0.005 as this is less than 0.01 which is the standard trace amount. I chose 0.005 as it is the middle value between the lowest possible trace value (0.001) and highest possible trace value (0.009).

Again, in anticipation of downstream analysis, the date entries in the “date” columns were changed to POSIX format.

**3) “Rush Hours” Analysis**

**Histogram:**

**A graph of a number of bars

Description automatically generated**

**Findings:** From the histogram we see that the “rush hours” were determined to be from 7-9am and 4-6pm.

**Top 10 Most Frequent Starting and Ending Stations:**

|  |  |
| --- | --- |
| **Weekday Rush Hours** | |
| ***Starting Stations*** | ***Ending Stations*** |
| San Francisco Caltrain (Townsend at 4th) | San Francisco Caltrain (Townsend at 4th) |
| San Francisco Caltrain 2 (330 Townsend) | San Francisco Caltrain 2 (330 Townsend) |
| Temporary Transbay Terminal (Howard at Beale) | Market at Sansome |
| Harry Bridges Plaza (Ferry Building) | 2nd at Townsend |
| 2nd at Townsend | Temporary Transbay Terminal (Howard at Beale) |
| Steuart at Market | Harry Bridges Plaza (Ferry Building) |
| Market at Sansome | Townsend at 7th |
| Townsend at 7th | Steuart at Market |
| Market at 10th | Embarcadero at Sansome |
| Embarcadero at Sansome | 2nd at Southpark |

**Findings:** From the table we see the top 10 most frequent starting and ending stations during the weekday rush hours determined from the histogram. The stations were nearly completely consistent between the starting and ending lists, with the only differences being the presence of Market at 10th in starting stations and 2nd at Southpark in ending stations. The order between lists was generally similar as well.

**3) Weekend Analysis**

**Histogram:**

**A graph of a number of bars

Description automatically generated**

**Findings:** From the histogramwe see that the bike rental activity begins to increase steadily around 8am, reaching its peak hours around 11am to 3pm, followed by a steady decrease.

**Top 10 Most Frequent Starting and Ending Stations:**

|  |  |
| --- | --- |
| **Weekend (Saturday & Sunday)** | |
| ***Starting Stations*** | ***Ending Stations*** |
| Harry Bridges Plaza (Ferry Building) | Embarcadero at Sansome |
| Embarcadero at Sansome | Harry Bridges Plaza (Ferry Building) |
| Market at 4th | Market at 4th |
| Embarcadero at Bryant | Powell Street BART |
| 2nd at Townsend | San Francisco Caltrain (Townsend at 4th) |
| Powell Street BART | 2nd at Townsend |
| San Francisco Caltrain (Townsend at 4th) | Embarcadero at Bryant |
| Grant Avenue at Columbus Avenue | Steuart at Market |
| Market at Sansome | Market at Sansome |
| Powell at Post (Union Square) | Grant Avenue at Columbus Avenue |

**Findings:** From the table we see the top 10 most frequent starting and ending stations during the weekend. The stations were nearly completely consistent between the starting and ending lists, with the only differences being the presence of Powell at Post in starting stations and Steuart at Market in ending stations. The order between lists was generally similar as well.

**4) Average Bike Utilization**

**Bar-plot:**

**A graph of blue bars

Description automatically generated**

**Findings:** The total time used for each month was determined by summing all the trip durations per month. The average utilization was then determined by the total time used by the total time available per each month. Generally, utilization is greater in the summer months in comparison to the winter months. This may be explained by less optimal weather conditions (ex. rain, snow, colder temperatures, poor visibility) for biking which are characteristic to the winter months.

**5) Weather Conditions – Rental Activity Correlations**

**Correlation Plots by City:**

**A screenshot of a computer

Description automatically generated**A screenshot of a computer

Description automatically generated

**Palo Alto**

**Mountain View**

A screenshot of a computer

Description automatically generated**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**San Francisco**

**San Jose**

**Redwood City**

**Findings:** To determine the potential correlations between weather conditions and bike rental activity, the “weather\_data” and “trip\_data” were joined together. Each corresponding city and date where then summarized by the total number of trips on each specific date and the duration of each trip. The above correlation plots were then constructed to assess the impact of the relevant weather conditions on the number of trips and the duration of these trips, in other words, the bike rental activity in general. It was found that better weather conditions resulted in a positive correlation with the number of trips and the duration of each trip. It was also found that there was a negative correlation between the number of trips and duration of these trips with the precipitation and cloud cover. This is expected as rain or signs of rain coming are not considered optimal conditions for biking.

**Appendix:**

**A graph of different colored lines

Description automatically generated with medium confidence**

**Figure 1:** Frequency plot depicting the frequency of start stations as a percentage on the Y-axis and the names of the stations themselves on the X-axis.

**A graph showing different colored numbers

Description automatically generated with medium confidence**

**Figure 2:** Frequency plot depicting the frequency of end stations as a percentage on the Y-axis and the names of the stations themselves on the X-axis.

**A graph of a number of people

Description automatically generated with medium confidence**

**Figure 3:** Frequency plot showing the frequency of each subscription type (customer and subscriber) as a percentage on the X-axis and the subscription types themselves on the Y-axis.

**A graph of different colored bars

Description automatically generated with medium confidence**

**Figure 4:** Histograms depicting the numeric variables of the trips dataset including the ID, the duration in seconds, the start and end stations, and the bike ID.

**A graph with different colored squares

Description automatically generated**

**Figure 5:** Frequency plot of the frequency of the events as a percentage on the X-axis and the different events themselves on the Y-axis.

**A screen shot of a graph

Description automatically generated**

**Figure 6:** Frequency plot of the frequency of precipitation recorded in inches as a percentage on the X-axis and the different amounts of precipitation in inches themselves on the Y-axis.

**A graph of different colored bars

Description automatically generated with medium confidence**

**Figure 7:** Histograms depicting the numeric variables of the weather dataset including maximum, mean, and minimum variables for each temperature, visibility in miles, max wind speed in miles per hour, as well as maximum gust speed, cloud coverage, and zip codes.