**SF Bay Area Bike Operations HQ – Data Analysis Report**

**Introduction:**

In this analysis, we consider 3 data sets:

* Station data, including 70 observations of 7 variables.
* Trip data, including 326339 observations of 11 variables.
* Weather data, including 1825 observations of 15 variables.

In our pre-processing (prior to EDA), we confirm that there are no duplicate observations across the data sets. We also change any blank spaces in the data sets to NAs for consistency of indicating missing values.

**EDA for Trip Data:**

Using our EDA commands for trip data, we identified that:

* ID variables are integers when they should be characters. Changed.
* Date variables are characters when they should be date type. Changed.
* There are 74 station start names but only 70 unique start station IDs. This means that different start stations have the same ID which should not be possible. To address this, we:
  + Correct spelling of start stations from “Kearny” to “Kearney”. This reduces the number of different start stations from 74 to 72.
  + Remove trip observations with either “San Jose Government Center” or “Standford in Redwood City” as the start station, since there is no associated station ID in the station data for these start stations. This reduces the number of different start stations from 72 to 70, which is consistent with the number of unique start station IDs.
    - Removed trips are recorded in the “trip\_removed” data frame, and IDs can be viewed using “trip\_removed$id”.
  + The exact same issue is present for end stations. The same steps were followed above for end stations to ensure the number of different end stations are consistent with the number of unique end station IDs.
* Zip code variable has zero values, which are not possible. Changed these values to NAs.
  + Interestingly, all NAs in the trip data set are in the zip code variable.

By examining the frequency of values for the categorical variables in trip data (note that this is prior to removal of cancelled trips and outliers), we plot (“quick and dirty” EDA, so plots not professionally formatted) and observe the following:

A close-up of a chart

Description automatically generated

* Top 3 starting stations are San Francisco Caltrain (Townsend at 4th) (ID# 70, with 25144 observations, representing 7.71% of all starting stations), Harry Bridges Plaza (Ferry Building) (ID# 50, with 15536 observations, representing 4.76% of all starting stations) and San Francisco Caltrain 2 (330 Townsend) (ID #69, with 15132 observations, representing 4.64% of all starting stations.

A collage of a rainbow colored line

Description automatically generated with medium confidence

* Top 3 ending stations are the same as the starting stations, being San Francisco Caltrain (Townsend at 4th) (ID# 70, with 33213 observations, representing 10.18% of all ending stations), Harry Bridges Plaza (Ferry Building) (ID# 50, with 15692 observations, representing 4.81% of all ending stations) and San Francisco Caltrain 2 (330 Townsend) (ID #69, with 15333 observations, representing 4.70% of all ending stations.

A blue and red squares with black text

Description automatically generated

* Most bike trip purchases come from subscribers (277726 observations, representing 85.12%) while the minority come from customers (48552 observations, representing 14.88%).

By examining the frequency distribution for the sole numerical variable in trip data (note that this is prior to removal of cancelled trips and outliers), we plot (“quick and dirty” EDA, so plots not professionally formatted) and observe the following:

A line graph with numbers and text

Description automatically generated

* The distribution of trip duration is extremely right skewed, meaning we have outliers at the right end. It may be helpful to remove outliers and then view the shape of the histogram data, potentially applying a transformation.
* The mean is 1131.41 seconds. This variable must have unit seconds, as identified by matching multiple observations’ duration values with their start and end date values. The standard deviation is 30821.07 seconds, which is extremely high, leading to a coefficient of variation of 27.241, also extremely high. The 95th percentile value is 1863 seconds while the 99th percentile value is 13304.12 seconds, indicating presence of outliers that have skewed our mean. Other parameters generated indicate non-normality and skewness of the data. Therefore, removal of outliers is necessary.

**EDA for Weather Data:**

Using our EDA commands for trip data, we identified that:

* Date variable is of character type when it should be date type. Changed.
* Zip code variable is of integer type when it should be character type. Changed.
* Changed variable name "max\_wind\_Speed\_mph" to "max\_wind\_speed\_mph" for consistency with other variables.
* Precipitation variable is of character type due to “T” values. These values were changed to “0.005” to align with meteorologist account (as per <https://wgntv.com/weather/what-are-traces-of-precipitation/>), then the entire variable was converted to numeric type.
* Cloud cover is of numeric type when it should be character type (collection of values were part of a scale, with no inherent units). Changed.
* It’s interesting to note that precipitation variable has quite a lot of zero values (1543/1825), while the events variable has quite a lot of missing values (1473/1825).

By examining the frequency of values for the categorical variables in weather data (note that this is prior to removal of cancelled trips and outliers), we plot (“quick and dirty” EDA, so plots not professionally formatted) and observe the following:

A bar graph with different colored bars

Description automatically generated

* Top 3 values for cloud cover are 0 (347 observations, representing 19.01%), 3 (260 observations, representing 14.25%) and 2 (250 observations, representing 13.7%).

A graph of weather forecast

Description automatically generated with medium confidence

* Ignoring NAs, (i.e., the vast majority), the top 3 values for events are rain (280 observations, representing 15.34%), fog (57 observations, representing 3.12%) and fog-rain (13 observations, representing 0.71%).

By examining the frequency distribution for the numerous numerical variables in weather data (note that this is prior to removal of cancelled trips and outliers), we plot (“quick and dirty” EDA, so plots not professionally formatted) and observe the following variables’ means and standard deviations:

|  |  |  |
| --- | --- | --- |
| Variable Name | Mean | Standard Deviation |
| max\_temperature\_f | 71.02630137 | 8.2641670 |
| mean\_temperature\_f | 62.03397260 | 6.7505642 |
| min\_temperature\_f | 52.82849315 | 6.6736799 |
| max\_visibility\_miles | 10.86013216 | 2.6205874 |
| mean\_visibility\_miles | 9.97081498 | 1.6224074 |
| min\_visibility\_miles | 8.10572687 | 3.0372980 |
| max\_wind\_speed\_mph | 16.43835616 | 7.3174163 |
| mean\_wind\_speed\_mph | 6.10630137 | 3.0467283 |
| max\_gust\_speed\_mph | 22.68558952 | 9.0911134 |
| precipitation\_inches | 0.03077534 | 0.1825588 |

We also observe:

A graph of different weather types

Description automatically generated with medium confidence

* Histograms for temperature have approximately normal distributions.

A graph of different weather types

Description automatically generated with medium confidence

* Histograms for visibility are approximately centrally distributed, with high frequency of central-tending values.

A graph of a weather

Description automatically generated with medium confidence

* Histograms for wind speed reveal right skewness, with possible outliers identified in maximum wind speed due to width of x-axis scale. For maximum wind speed, profiling reveals high coefficient of variation, skewness and kurtosis.

A graph of a graph

Description automatically generated

* Histogram for maximum gust speed reveals right skewness, with possible outliers identified due to width of x-axis scale. Profiling reveals high coefficient of variation, skewness and kurtosis.

A graph of a weather forecast

Description automatically generated

* Histogram for precipitation reveals right skewness, with possible outliers identified due to width of x-axis scale. Profiling reveals high coefficient of variation, skewness and kurtosis.

**Canceled Trips:**

Cancelled trips were identified based on the following criteria: any trip starting and ending at the same station, with duration less than 3 minutes. There were 1061 trips removed based on these criteria and removed trips were recorded in the “cancelled\_trips” data frame. IDs can be viewed using “cancelled\_trips$id”.

**Outliers:**

Outliers were identified for the sole numerical variable in trip data, being trip duration. The specific method of outlier removal was the interquartile method, with observations more than 1.5x the IQR below Q1 (i.e., -266 seconds, not possible so criterion ignored) or more than 1.5x the IQR above Q3 (i.e., 1354 seconds) being considered outliers. There were 24860 trips removed based on these criteria and outlier trips were recorded in the “trip\_outliers” data frame. IDs can be viewed using “trip\_outliers$id”.

**Weekday Rush Hours and Most Frequent Stations:**

Using the lubridate package, weekday rush hours were calculated based on trip start date and time, as this is when bikes are first considered “taken”. In the future, a more accurate measurement might include plotting not just the start or end time of individual trips, but rather each time block (e.g., 15 minutes) that individual bike were being used. The following histogram was constructed:

A graph of a bike trip

Description automatically generated

* Weekday rush hours were identified to be between 7:45 AM – 9:30 AM and between 4:45 PM – 6:15 PM.

The 10 most common starting stations during these weekday rush hours were identified to be:

1. San Francisco Caltrain (Townsend at 4th), with 10388 observations representing 15.85%.
2. San Francisco Caltrain 2 (330 Townsend), with 5867 representing 8.95%.
3. Temporary Transbay Terminal (Howard at Beale), with 4843 observations representing 7.39%.
4. Harry Bridges Plaza (Ferry Building), with 4328 observations, representing 6.60%.
5. Steuart at Market, with 3152 observations representing 4.81%.
6. Grant Avenue at Columbus Avenue, with 2638 observations representing 4.03%.
7. 2nd at Townsend, with 2319 observations representing 3.54%.
8. Beale at Market, with 1974 observations representing 3.01%.
9. Market at Sansome, with 1874 observations representing 2.86%.
10. Market at 10th, with 1799 observations representing 2.75%.

The 10 most common ending stations during these weekday rush hours were identified to be:

1. San Francisco Caltrain (Townsend at 4th), with 4839 observations representing 7.38%.
2. 2nd at Townsend, with 3902 observations representing 5.95%.
3. Market at Sansome, with 3699 observations representing 5.65%.
4. Townsend at 7th, with 3591 observations representing 5.48%
5. 2nd at South Park, with 2585 observations representing 3.94%.
6. Embarcadero at Folsom, with 2482 observations representing 3.79%.
7. Temporary Transbay Terminal (Howard at Beale), with 2362 observations representing 3.60%.
8. Steuart at Market, with 2313 observations representing 3.53%.
9. San Francisco Caltrain 2 (330 Townsend) with 2286 observations representing 3.49%.
10. Howard at 2nd, with 2135 observations representing 3.26%.

**Weekend Stations and Most Frequent Stations:**

The 10 most common starting stations during the weekends were identified to be:

1. Embarcadero at Sansome, with 2142 observations representing 7.21%.
2. Harry Bridges Plaza (Ferry Building) with 1923 observations representing 6.47%.
3. Market at 4th, with 1266 observations representing 4.26%.
4. 2nd at Townsend, with 1232 observations representing 4.14%.
5. Embarcadero at Bryant, with 1231 observations representing 4.14%.
6. Powell Street BART, with 1147 observations representing 3.86%.
7. San Francisco Caltrain (Townsend at 4th), with 1078 observations representing 3.63%.
8. Grant Avenue at Columbus Avenue, with 1028 observations representing 3.46%.
9. Market at 10th, with 877 observations representing 2.95%.
10. San Francisco Caltrain 2 (330 Townsend), with 871 observations representing 2.93%.

The 10 most common ending stations during the weekends were identified to be:

1. Harry Bridges Plaza (Ferry Building), with 2344 observations representing 7.89%.
2. Embarcadero at Sansome, with 1661 observations representing 5.59%.
3. Market at 4th, with 1507 observations representing 5.07%.
4. Powell Street BART, with 1378 observations representing 4.64%.
5. San Francisco Caltrain (Townsend at 4th), with 1354 observations representing 4.56%.
6. 2nd at Townsend, with 1269 observations representing 4.27%.
7. Embarcadero at Bryant, with 1125 observations representing 3.78%.
8. Steuart at Market, with 975 observations representing 3.28%.
9. Townsend at 7th, with 922 observations representing 3.10%.
10. Market at Sansome, with 914 observations representing 3.07%.

**Average Utilization of Bikes According to Month:**

The average utilization of all bikes according to month (total time used by all bikes / total time in month) is summarized in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Month | Total Time Used (s) | Total Time in Month (s) | Utilization Ratio |
| January | 11701808 | 2678400 | 4.368955 |
| February | 8977049 | 2419200 | 3.710751 |
| March | 11835070 | 2678400 | 4.418709 |
| April | 12650352 | 2592000 | 4.880537 |
| May | 13835880 | 2678400 | 5.165726 |
| June | 14556553 | 2592000 | 5.615954 |
| July | 15249968 | 2678400 | 5.693686 |
| August | 15060672 | 2678400 | 5.623011 |
| September | 15507053 | 2592000 | 5.982659 |
| October | 16910897 | 2678400 | 6.313806 |
| November | 12526558 | 2592000 | 4.832777 |
| December | 9541132 | 2678400 | 3.562251 |

* It appears that utilization is greatest in the summer and fall months, less so in the spring and lowest in the winter.

**Correlations for Bike Rental Patterns:**

To perform our correlation analysis, we must create one data frame to feed into the cor() and corrplot() functions. We can subset our data sets such that we keep only the variables of interest (must be numerical for above correlation functions) for correlation analysis and the key variables that will help link our data sets.

* From trip data, we keep the duration, start station ID and start date variables.
* From station data, we keep the station ID, city and dock count variables. We rename the ID variable to match the trip data’s ID variable for appropriate joining.
* From weather data, we keep all variables (for now) and rename the date variable to match the trip data’s start date variable for appropriate joining.

Once the data sets have been combined into one, we remove the non-numerical variables that originated from weather data. We also remove observations with NAs, as these observations will invalidate the correlation analysis. Indices for removed observations are recorded in “NA\_indices” variable. Finally we can run the cor() function on our combined data set and feed it into the corrplot() function for visualization, as seen below:

A screen shot of a data

Description automatically generated

* In terms of bike rental patterns, the variables of interest are trip duration and possibly dock count. Correlations with any other variables are quite low for both variables, with the highest correlation for trip duration being mean wind speed (0.0209) and the highest correlation for dock count being maximum visibility (-0.0232). In the future, it may be beneficial to examine their correlations with categorical variables using a different function in order to build the best predictive model.