*Exploratory Data Analysis (EDA)*

An EDA was conducted for each of the 3 datasets (station, trip, and weather) separately. This allowed for the understanding and general overview of the data characteristics. Through the use of the “funModeling” package, key variables were identified for downstream analysis to achieve the project objectives. Furthermore, certain variables were identified as irrelevant or redundant, and missing data patterns were shown as well. This was useful in preparation for the data cleaning stage.

*EDA – station*

* Initially 70 observations (rows) of 7 variables (columns)
* installation\_date variable was noted to be in character form
* 5 unique cities in the dataset
* city identified as a key categorical variable to analyze
* Other than dock\_count, none of the integer variables appeared to be relevant for downstream analysis
* No missing values in the dataset

A graph with different colored squares

Description automatically generated

**Figure 1.** Frequency of bike rental stations in Bay Area cities.

* San Francisco has the most bike rental stations (more than double the amount in San Jose)

*EDA – trip*

* Initially 326 339 observations (rows) of 11 variables (columns)
* Both date variables (start\_date and end\_date) were noted to be in character form
* 1493 missing values present in the column for zip\_code
* ID numbers are all unique
* subscription\_type is clearly shown for categorical variables
* duration is a relevant numerical variable to be analyzed
* 74 distinct station names

*A chart of different colored numbers

Description automatically generated with medium confidence*

**Figure 2.** Frequency of ending stations for bike rentals in the Bay Area.

*A chart of a number of stations

Description automatically generated with medium confidence*

**Figure 3.** Frequency of starting stations for bike rentals in the Bay Area.

* Most frequently used station for both starting and ending is San Francisco Caltrain (Townsend at 4th)
* Top 3 most frequent are the same for starting and ending stations

*A screenshot of a graph

Description automatically generated*

**Figure 4.** Comparing the customer base for bike rentals that use the service through a subscription to users that pay-per-ride.

* With more users opting to use the subscription service, this may indicate that the bike rental service can improve upon the ability to rent a bike without a subscription; this could mean lowering the pay-per-use cost, and would allow for the bikes to reach a wider customer base

*EDA – weather*

* Initially 1825 observations (rows) of 15 variables (columns)
* date variable noted to be in character form
* precipitation\_inches is in character form; has some integer values, but values under 0.01 are classified as trace (denoted by T)
* Zeroes found in min\_visibility\_miles, mean\_wind\_speed\_mph, precipitation\_inches, and cloud\_cover
* NA values in all the visibility variables and max\_gust\_speed\_mph
* Analysis is relevant for all the weather measurement values, as they are numeric (temperature, visibility, wind speed, cloud cover)
* 1473 missing values in events, 451 missing values in max\_gust\_speed\_mph, 9 missing values in each of the visibility variables

A graph of different colored bars

Description automatically generated with medium confidence**Figure 5.** Analysis for key weather measurements in the Bay Area in the year 2014. Note that zip\_code is included in this plot as part of the EDA, but the analysis is not relevant.

*Data Cleaning*

In this stage, duplicate copies of each dataset were created to store the cleaned data. Blank values were replaced with NA, and variable names/syntax within each dataset was checked for consistency. Cancelled trips, defined as trips with a duration less than 3 minutes starting and ending at the same station, were removed from the dataset. Outlier thresholds were determined, and outlier values were removed from the dataset.

*Data Cleaning – station*

* installation\_date converted from character to POSIX format
* No duplicate rows, NA values, or blank values in the dataset

*Data Cleaning – trip*

* Blank values in zip\_code column replaced with NA (rows were not removed)
* No duplicate rows in the dataset
* start\_date and end\_date converted to POSIX format from character type
* Cancelled trips (trips with a duration less than 3 minutes starting and ending at the same station) were identified and removed from the dataset
* 1082 rows were removed from the dataset as cancelled trips
* A csv file containing the trip IDs for the cancelled trips is saved in the GitHub repository for this project, under the file name “cancelled\_trips.csv”
* From the EDA, the p\_01 value is 128.0 and the p\_99 value is 13311.62 for the duration column (in seconds); trips shorter than 128 seconds (approx. 2 mins) or longer than 13311 seconds (approx 3.7 hours) could still be realistic bike rental timeframes
* Unrealistic trip length can be considered any trip under 1 minute (60 seconds) in length, or over 6 hours in length (21600 seconds); values outside of this range (below 60 and above 21600) will be treated as outliers, and these rows were identified and removed from the dataset
* 1622 rows were removed from the dataset as outliers
* A csv file containing the trip IDs for outlier trips is saved in the GitHub repository for this project, under the file name "outlier\_trips.csv"

*Data Cleaning – weather*

* Changed column name from max\_wind\_Speed\_mph to max\_wind\_speed\_mph to maintain consistency with other variable names
* Converted date to POSIX format from character type
* Replaced blank values with NA in the events column of the dataset
* No duplicate rows in the dataset
* Trace precipitation, denoted by T, prevents the precipitation\_inches column from being numeric; the midpoint value of 0 and the lowest non-zero value of 0.01, which is 0.05, was imputed in the data for trace precipitation

*Rush Hours*

* The midpoint of each trip was used to determine rush hours; this was done to ensure that the entire time spent on the bike was considered for each ride, and also preventing the data from being skewed towards starting or ending times
* Separate data frames were created for trips on weekdays and weekends

A graph of blue bars

Description automatically generated

**Figure 6.** Frequency of trips for bike rentals in the Bay Area on weekdays.

* Viewing the histogram, the hours of weekdays with the highest trip volume (rush hours) were determined to be 7:00-10:00 AM and 4:00 to 7:00 PM
* This makes sense, as these hours are typically when people commute to and from work or school on weekdays
* To determine the most frequent stations during rush hours, a separate data frame was created that only includes trips with a midpoint within rush hour

**Table 1.** 10 most frequent starting stations for bike rentals during rush hours (7:00-10:00 AM and 4:00 to 7:00 PM) in the Bay Area.

|  |  |
| --- | --- |
| **Station Name** | **Frequency** |
| San Francisco Caltrain (Townsend at 4th) | 17772 |
| San Francisco Caltrain 2 (330 Townsend) | 10121 |
| Temporary Transbay Terminal (Howard at Beale) | 9514 |
| Harry Bridges Plaza (Ferry Building) | 7878 |
| 2nd at Townsend | 7275 |
| Steuart at Market | 7217 |
| Market at Sansome | 6530 |
| Townsend at 7th | 6384 |
| Market at 10th | 5465 |
| Embarcadero at Sansome | 5324 |

**Table 2.** 10 most frequent ending stations for bike rentals during rush hours (7:00-10:00 AM and 4:00 to 7:00 PM) in the Bay Area.

|  |  |
| --- | --- |
| **Station Name** | **Frequency** |
| San Francisco Caltrain (Townsend at 4th) | 24037 |
| San Francisco Caltrain 2 (330 Townsend) | 10489 |
| Market at Sansome | 8274 |
| 2nd at Townsend | 8122 |
| Temporary Transbay Terminal (Howard at Beale) | 7712 |
| Harry Bridges Plaza (Ferry Building) | 7656 |
| Townsend at 7th | 7178 |
| Steuart at Market | 7112 |
| Embarcadero at Sansome | 5202 |
| 2nd at South Park | 4737 |

**Table 3.** 10 most frequent starting stations for bike rentals during weekends in the Bay Area.

|  |  |
| --- | --- |
| **Station Name** | **Frequency** |
| Harry Bridges Plaza (Ferry Building) | 3137 |
| Embarcadero at Sansome | 3094 |
| Market at 4th | 1627 |
| Embarcadero at Bryant | 1589 |
| 2nd at Townsend | 1529 |
| Powell Street BART | 1469 |
| San Francisco Caltrain (Townsend at 4th) | 1353 |
| Grant Avenue at Columbus Avenue | 1287 |
| Market at Sansome | 1085 |
| Market at 10th | 1072 |

**Table 4.** 10 most frequent ending stations for bike rentals during weekends in the Bay Area.

|  |  |
| --- | --- |
| **Station Name** | **Frequency** |
| Embarcadero at Sansome | 3344 |
| Harry Bridges Plaza (Ferry Building) | 3150 |
| Market at 4th | 1838 |
| Powell Street BART | 1649 |
| San Francisco Caltrain (Townsend at 4th) | 1632 |
| 2nd at Townsend | 1584 |
| Embarcadero at Bryant | 1376 |
| Steuart at Market | 1196 |
| Grant Avenue at Columbus Avenue | 1086 |
| Market at Sansome | 1086 |

1

San Francisco Caltrain (Townsend at 4th)

17772

2

*Bike Utilization*

* Rather than calculating the average utilization for each individual bike determined by bike\_id, an average value was calculated across all bikes used within a single month
* This determines the months in which bikes are used more often, so an average value calculated across all bikes is sufficient
* Average for individual bikes is not necessary, and applicability would be limited as variations between different bikes are not known (all bikes are looked at as the same in the dataset) i.e. if there were different types of bikes mapped out in the dataset, then it would make more sense to calculate the average for individual bikes
* There are 687 unique bike IDs in the dataset, but all bikes may not be used throughout the month
* For the unique bikes that are used throughout the month: they are assumed to be available for use 24/7 for the entire duration of the month
* Total time was calculated as multiplying the total amount of time in the month by the number of unique bikes for that month (theoretically, 100% utilization would mean that all unique bikes for the month were used 24/7 for the entire duration of the month)
* Total duration was calculated as the sum of duration of all bike rides throughout the month
* Average utilization was calculated with ((total duration) / (total time))

**Table 5.** Average utilization of rental bikes in the Bay Area for the year 2014.

|  |  |
| --- | --- |
| **Month** | **Average Bike Utilization (%)** |
| January | 1.159 |
| February | 0.995 |
| March | 1.230 |
| April | 1.325 |
| May | 1.470 |
| June | 1.542 |
| July | 1.621 |
| August | 1.653 |
| September | 1.521 |
| October | 1.544 |
| November | 1.162 |
| December | 0.932 |

* Average utilization is highest in the summer months (peak in August), with a marked increase from February to March and a decrease in use from October to November
* This is expected, with more users willing to rent bikes as the weather gets warmer
* December has the lowest average utilization
* Average bike utilization throughout the year is 0.01346, or 1.346%
* Utilization is consistently low throughout the year, with each bike spending majority of time docked every month; this may be because the rental service needs to keep as many bikes as possible always docked at each station, so that users do not run into a situation where they cannot find a bike at a station

*Weather Condition Analysis*

* Created a new dataset that combines trip and weather data together by first combining trip and station data together
* Trip and station data were joined together through the start\_station\_name and name columns, as it allows for trip data to be combined with the weather data through city (city column present in both weather and station datasets)
* Times were removed from the dates (midpoint used to ensure that each trip is only associated with a single date
* Variables that were non-numeric or irrelevant to the analysis were removed
* All numeric variables were used to create a correlation matrix using pairwise complete observations; instead of excluding an entire row with missing data, this method allows for only pairs of values with missing data to be excluded, maximizing data utilization

A diagram of a weather forecast

Description automatically generated

**Figure 7.** Correlation matrix comparing bike ride duration to key weather measurements in the Bay Area for the year 2014. Note that the comparisons for key weather measurements between each other are included; however, the analysis is not relevant.

**Table 6.** Correlation coefficients between bike ride duration and key weather measurements.

|  |  |
| --- | --- |
|  | Correlation coefficient with Duration |
| Maximum Temperature | 0.01240 |
| Mean Temperature | 0.01007 |
| Minimum Temperature | 0.00489 |
| Max Visibility | 0.06176 |
| Mean Visibility | 0.03678 |
| Minimum Visibility | 0.03117 |
| Maximum Wind Speed | -0.0080 |
| Mean Wind Speed | 0.00525 |
| Maximum Gust Speed | -0.0094 |
| Precipitation | -0.0212 |
| Cloud Cover | -0.0366 |

* When comparing trip duration to key weather measurements, the highest correlation coefficient calculated is with Max Visibility, followed by Mean Visibility and Minimum Visibility; this would typically indicate that an increase in these measurements would result in an increased number of trips, but the correlation is still very weak (coefficient less than 0.1)
* Negative correlation coefficient observed with Cloud Cover, Precipitation, Maximum Gust Speed, and Maximum Wind Speed
* This would indicate that these measurements are associated with a decreased number of trips, but the correlation is weak with these variables as well (greater than -0.1).
* Overall, the weather measurements are very weakly correlated with trip duration over the course of a single day; this may be due to limitations of this analysis
* In the future, other trip metrics must be analyzed with their correlation to weather measurements, such as the number of trips in a single day or the average utilization throughout the day
* A clearer correlation may be shown if weather measurements were to be analyzed only during higher volume trip times (e.g. during rush hours)
* Furthermore, future analyses should not be grouped with all the cities together; a separate analysis should be done for each city, as unique weather patterns may exist in different areas