**Summary and Pre-processing of Dataset(s)**

***Trips Data***

Trips data, from the trips.csv file, described bike trips with the following information:

* id
* duration (seconds)
* start\_date (m/d/y h/m)
* start\_station\_name
* start\_station\_id
* end\_date (m/d/y h/m)
* end\_station\_name
* end\_station\_id
* bike\_id
* subscription\_type
* zip\_code

Exploratory Data Analysis (EDA)

EDA was performed to briefly overview the trips data.

None of the columns included NA or infinity values, however zip\_code contained fifty “0” values.

Frequencies of categorical variables, subscription\_type, start\_station\_name and end\_station\_name, were determined and plotted (Fig 1-3).

Numerical data was plotted (Fig 4). In Figure 4, we can see that id, start\_station\_id, end\_station\_id, duration, and bike\_ids are plotted as numerical. However, only duration of bike trip make sense as a numerical value.

Numerical data was also summarized by its metrics in Table 1. Here we can see ids of trips, bikes, start and end station, and duration as numerical value. However, only duration of bike trip make sense as a numerical value.

**Figure 1. Frequency of Subscription Types for Bike Riders**

A diagram of a customer

Description automatically generated with medium confidence

**Figure 2. Frequency of Bike Trips by Start Station**

**A rainbow colored line with white text

Description automatically generated**

**Figure 3. Frequency of Bike Trips by End Station**

**A rainbow colored line with white text

Description automatically generated**

**Figure 4. Count of ids, duration of bike trips, start\_station\_ids, end\_station\_ids and bike\_ids**

**A graph of different colored bars

Description automatically generated**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 1. Metrics of Trips Numerical Data | | | | | |
|  | **Mean** | **Std\_dev** | **Variation\_coef** | **Skewness** | **IQR** |
| Id | 362366.29122 | 129430.63831 | 0.3571818 | 0.02422033 | 223218 |
| Duration | 1131.96740 | 30816.16160 | 27.2235416 | 539.65122964 | 405 |
| Start\_station\_id | 57.79163 | 17.30435 | 0.2994267 | -1.42240299 | 20 |
| End\_station\_id | 57.78113 | 17.39386 | 0.3010301 | -1.4-566006 | 20 |
| Bike\_id | 427.59535 | 152.63705 | 0.3569661 | -0.50121144 | 211 |

Cancelled Trips

Cancelled trips were categorized as any trips which started and ended at the same station, and were less than 3 minutes (or 180s) in duration.

In the trips data, 1082 trips were categorized as cancelled, removed from trips data, and their ids saved to removed\_trips.csv, alongside the outliers.

Outliers

For the trips data, 152464967 outliers were removed from the duration variable, and their ids saved to removed\_trips.csv.

Outliers were points of data that were 1.5x IQR below Quantile 1, and 1.5x IQR above Quantile 3 (Table 2).

We can see changes in the max, min, mean and medium after outliers are removed (Table 3). Specifically, we see reduction in the max value, and the mean value closer to the median value, compared to the duration data with outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2. Quantiles and Interquartile Range of Duration | | | |
|  | **Q1** | **Q3** | **IQR** |
| Duration (s) | 354 | 732 | 378 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3. Max, Min and Measures of Central Tendency, before and after removing outliers from Duration | | | | |
|  | **Max** | **Min** | **Mean** | **Median** |
| With outliers | 17270400 | 180 | 961.7805 | 514 |
| Without outliers | 1298 | 180 | 534.0878 | 492 |

***Weather Data***

Weather data, from the weather.csv file, described the weather with the following information:

* 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
* events
* zip\_code
* city

Exploratory Data Analysis (EDA)

EDA was performed to briefly overview the weather data.

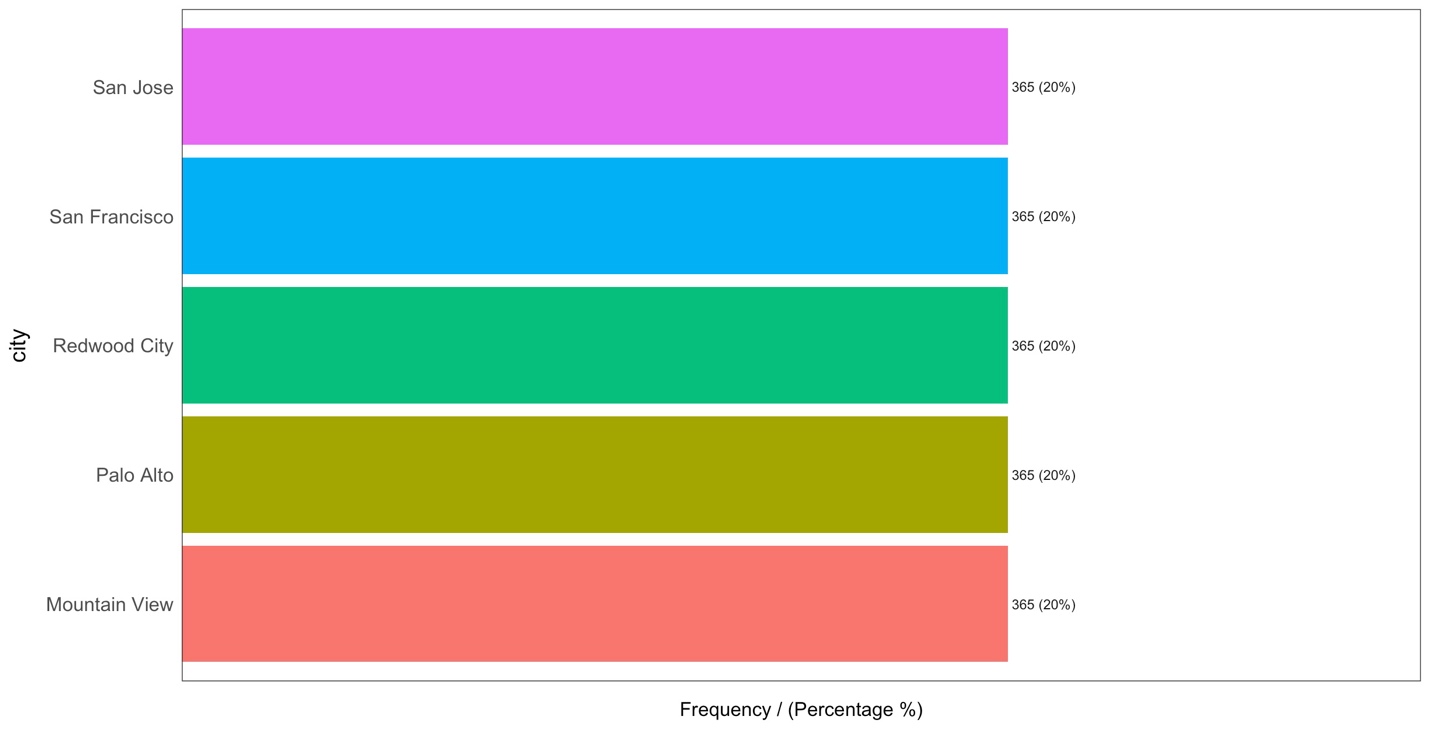
All variables are characters or integers. We see “0” values in cloud\_cover, precipitation\_inches, mean\_wind\_speed\_mph and min\_visibility\_miles. We also see NA values in mean\_visibility\_miles, max\_visibility\_miles, min\_visibility\_miles, max\_gust\_speed\_mph.

Frequencies of categorical variables, citiy, events and precipitation\_inches, were determined and plotted (Fig 5-7). Here it can be noted, the precipitation\_inches variable is better as a numerical value than categorical.

Numerical data was plotted (Fig 8). In Figure 8, we can see that 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, zip\_code are numerical values.

Numerical data was also summarized by its metrics in Table 4.

**Figure 5. Frequencies of Cities**



**Figure 6. Frequency of Weather Events**

A screenshot of a computer

Description automatically generated

**Figure 7. Frequency of Precipitation Amount**

A screenshot of a computer

Description automatically generated

**Figure 8. Count of Numerical Data in Weather**

A graph of different colored bars

Description automatically generated with medium confidence

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 4. Metrics of Weather Numerical Data | | | | | |
|  | **Mean** | **Std\_dev** | **Variation\_coef** | **Skewness** | **IQR** |
| max\_temperature\_f | 71.0263 | 8.264167 | 0.1163536 | 0.2495052 | 12 |
| mean\_temperature\_f | 62.03397 | 6.750564 | 0.1088204 | - 0.1466389 | 11 |
| min\_temperature\_f | 52.82849 | 6.67368 | 0.1263273 | - 0.4771881 | 10 |
| max\_visibility\_miles | 10.86013 | 2.620587 | 0.2413035 | 2.881776 | 0 |
| mean\_visibility\_miles | 9.970815 | 1.622407 | 0.1627156 | 1.782374 | 0 |
| min\_visibility\_miles | 8.105727 | 3.037298 | 0.3747101 | -1.032912 | 3 |
| max\_wind\_Speed\_mph | 16.43836 | 7.317416 | 0.4451428 | 7.474993 | 7 |
| mean\_wind\_speed\_mph | 6.106301 | 3.046728 | 0.4989482 | 0.4568894 | 4 |
| max\_gust\_speed\_mph | 22.68559 | 9.091113 | 0.400744 | 4.925805 | 7 |
| cloud\_cover | 3.000548 | 2.303815 | 0.7677982 | 0.3367737 | 4 |
| zip\_code | 94325 | 404.6709 | 0.004290176 | 1.323103 | 238 |

Data cleaning

The weather dataset included many variables with NA values which were removed and saved to removed\_weather.csv.

***Station Data***

Station data, from the station.csv file, described the bike stations with the following information:

* id
* name
* lat
* long
* dock\_count
* city
* installation\_date

EDA was not performed on station data as it is not used for downstream analyses.

**Findings**

***Rush hours on weekdays***

The rush hours on weekdays are 9 am, 10 am and 6 pm.

Volume of trips per hour, per weekday, are stored in highest\_volumehours.csv and outlined in Figure 9. Weekdays are coded as 1: Monday, 2: Tuesday, 3: Wednesday, 4: Thursday, 5: Friday.

**Figure 9. Volume of trips on weekdays per hour**



***10 most frequent start and ending during weekday rush hours***

The 10 most frequent start stations during rush hours are outlined in Table 5, and the 10 most frequent end stations during rush hours are outlined in Table 6.

|  |  |
| --- | --- |
| Table 5. 10 Most Frequent Start Stations (Weekday Rush hours) | |
| **Station Name** | **Number of Trips** |
| San Francisco Caltrain (Townsend at 4th) | 48320 |
| San Francisco Caltrain 2 (330 Townsend) | 3321 |
| Market at Sansome | 2291 |
| 2nd at Townsend | 2277 |
| Temporary Transbay Terminal (Howard at Beale) | 2066 |
| Steuart at Market | 2005 |
| Market at 10th | 1837 |
| Townsend at 7th | 1777 |
| 2nd at South Park | 1664 |
| Harry Bridges Plaza (Ferry Building) | 1640 |

|  |  |
| --- | --- |
| Table 6. 10 Most Frequent End Stations (Weekday Rush hours) | |
| **Station Name** | **Number of Trips** |
| San Francisco Caltrain (Townsend at 4th) | 4511 |
| San Francisco Caltrain 2 (330 Townsend) | 2694 |
| Townsend at 7th | 2598 |
| 2nd at Townsend | 2558 |
| Market at Sansome | 2533 |
| Temporary Transbay Terminal (Howard at Beale) | 2444 |
| Steuart at Market | 2091 |
| 2nd at South Park | 2026 |
| Market at 10th | 1875 |
| Powell Street BART | 1672 |

***10 most frequent start and ending during weekends***

The 10 most frequent start stations during weekends are outlined in Table 7, and the 10 most frequent end stations during weekends are outlined in Table 8.

|  |  |
| --- | --- |
| Table 7. 10 Most Frequent Start Stations (Weekends) | |
| **Station Name** | **Number of Trips** |
| San Francisco Caltrain (Townsend at 4th) | 4515 |
| San Francisco Caltrain 2 (330 Townsend) | 2849 |
| Harry Bridges Plaza (Ferry Building) | 2818 |
| Embarcadero at Sansome | 2426 |
| 2nd at Townsend | 2411 |
| Market at Sansome | 2218 |
| Temporary Transbay Terminal (Howard at Beale) | 2053 |
| Townsend at 7th | 2037 |
| Steuart at Market | 1969 |
| Market at 4th | 1961 |

|  |  |
| --- | --- |
| Table 8. 10 Most Frequent End Stations (Weekends) | |
| **Station Name** | **Number of Trips** |
| San Francisco Caltrain (Townsend at 4th) | 5953 |
| Harry Bridges Plaza (Ferry Building) | 3326 |
| San Francisco Caltrain 2 (330 Townsend) | 2675 |
| Market at Sansome | 2669 |
| 2nd at Townsend | 2607 |
| Townsend at 7th | 2325 |
| Embarcadero at Sansome | 2204 |
| Steuart at Market | 2163 |
| Powell Street BART | 2116 |
| Market at 4th | 2062 |

***Average Utilization of Bikes***

To calculate average utilization of bikes, total time of bikes used was divided by total time in month, then converted to percents (Table 9).

|  |  |
| --- | --- |
| Table 9. Average Bike Utilization, per Month | |
| **Month** | Utilization % |
| January | 415.2562 |
| February | 319.9517 |
| March | 419.2785 |
| April | 450.9296 |
| May | 488.3874 |
| June | 516.2299 |
| July | 536.5481 |
| August | 530.7793 |
| September | 554.1146 |
| October | 606.0571 |
| November | 448.2694 |
| December | 342.6274 |

***Correlation***

A correlation matrix was created between the duration of bike trips and various weather variables. The three highest correlated weather conditions with duration of bike trip are latitude and longitude with a positive correlation, and cloud\_cover with a negative correlation.

Table 10 outlines the correlations of weather variables with duration of bike trip, and the three highest correlations are indicated by asterisks.

|  |  |
| --- | --- |
| Table 10. Correlation between Bike Duration and Weather Conditions | |
| **Weather Condition** | Correlation |
| lat | 0.11947414 \*\* |
| long | 0.027042954 \*\* |
| dock\_count | 0.021897865 |
| max\_temperature\_f | 0.014856356 |
| mean\_temperature\_f | 0.011198260 |
| min\_temperature\_f | 0.009301808 |
| max\_visibility\_miles | 0.009223888 |
| mean\_visibility\_miles | 0.007359057 |
| min\_visibility\_miles | 0.006386815 |
| max\_wind\_Speed\_mph | 0.006336481 |
| mean\_wind\_speed\_mph | 0.005965717 |
| max\_gust\_speed\_mph | 0.004295895 |
| cloud\_cover | -0.111483796 \*\* |