**BTC1855 Midterm Report**

1. [Introduction](#intro)
2. [Description and summary of dataset](#summary)
3. [Preprocessing](#data)
4. [Findings](#findings)

**Introduction**

The SF Bay Area Bike Operations dataset includes information about bike stations, individual trips, and weather conditions. The objective of this analysis is to clean and preprocess the data, perform exploratory data analysis, and extract insights that will help the data science team create a predictive model for bike usage patterns.

**Description and summary of dataset**

An exploratory data analysis (EDA) was conducted on the three datasets, following Data Science Heroes’ code template. Focus will be placed on variables that will be used for further data analysis and interpretation.

**Station Data**

This dataset consists of 70 observations of bike stations in the Bay Area, each with a unique ID, name, latitude, longitude, dock count, city, and installation date. The stations are located across five cities, with the majority in San Francisco (50%) and the rest distributed among San Jose, Mountain View, Redwood City, and Palo Alto (Figure 1). The installation dates span from August 2013 to early 2014, with most stations installed in August 2013 (Figure 2).

**A graph with different colored squares

Description automatically generated**

**Figure 1.** Bar graph displaying distribution of stations in five cities in the Bay Area

A graph of a number of colored bars

Description automatically generated with medium confidence

**Figure 2**. Bar graph displaying distribution of installation dates of bike stations spanning 2013-2014

**Trip Data**

The dataset consists of 326,339 observations and 13 variables related to bike trips. Some of the information included is trip duration (Table 3), bike IDs, start and end dates, and the type of user (subscriber vs. customer) (Figure 3). An important thing to note is the large skew in trip duration, where there seems to be one trip that is substantially longer in duration than others. Further analysis may need to look into this specific trip and whether it should be removed.

**Table 3.** Summary statistics for trip data’s duration variable

|  |  |
| --- | --- |
| **Statistic** | **Duration (seconds)** |
| **Minimum** | 60 |
| **Maximum** | 17270400 |
| **Mean** | 1132 |
| **Standard Deviation** | 30816.2 |

A graph with blue and red squares

Description automatically generated

**Figure 3.** Bar graph displaying distribution of subscription type of user

**Weather EDA**

The dataset consists of 1,825 observations with 15 variables, capturing weather measures and conditions. It includes weather conditions recorded per day in 2014 across five cities in the Bay Area: San Francisco, Mountain View, Palo Alto, Redwood City, and San Jose. The following are summary statistics for some of the major variables such as temperature, visibility, wind, precipitation, and events (Table 4-8; Figure 4-5). It is important to note the large number of NAs in events (Figure 4) and the high number of zeroes in precipitation\_inches (Figure 5).

**Table 4.** Summary statistics of temperature variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Max Temperature (°F)** | **Mean Temperature (°F)** | **Min Temperature (°F)** |
| Minimum | 50 | 44 | 32 |
| Maximum | 102 | 84 | 69 |
| Mean | 71 | 62 | 52.8 |
| Standard Deviation | 8.3 | 6.8 | 6.7 |

**Table 5.** Summary statistics of visibility variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Max Visibility (miles)** | **Mean Visibility (miles)** | **Min Visibility (miles)** |
| Minimum | 5 | 4 | 0 |
| Maximum | 20 | 20 | 20 |
| Mean | 10.9 | 10 | 8.1 |
| Standard Deviation | 2.6 | 1.6 | 3 |

**Table 6.** Summary statistics of wind variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Max Wind Speed (mph)** | **Mean Wind Speed (mph)** | **Max Gust Speed (mph)** |
| Minimum | 4 | 1 | 0 |
| Maximum | 122 | 34 | 43 |
| Mean | 16.4 | 6.1 | 22.7 |
| Standard Deviation | 7.3 | 3 | 9.1 |

**Table 7.** Summary statistics of precipitation and cloud cover

|  |  |  |
| --- | --- | --- |
| **Statistic** | **Precipitation (inches)** | **Cloud Cover** |
| Minimum | 0 | 0 |
| Maximum | 0.73 | 8 |
| Mean | 0.03 | 3 |
| Standard Deviation | 0.18 | 2.3 |

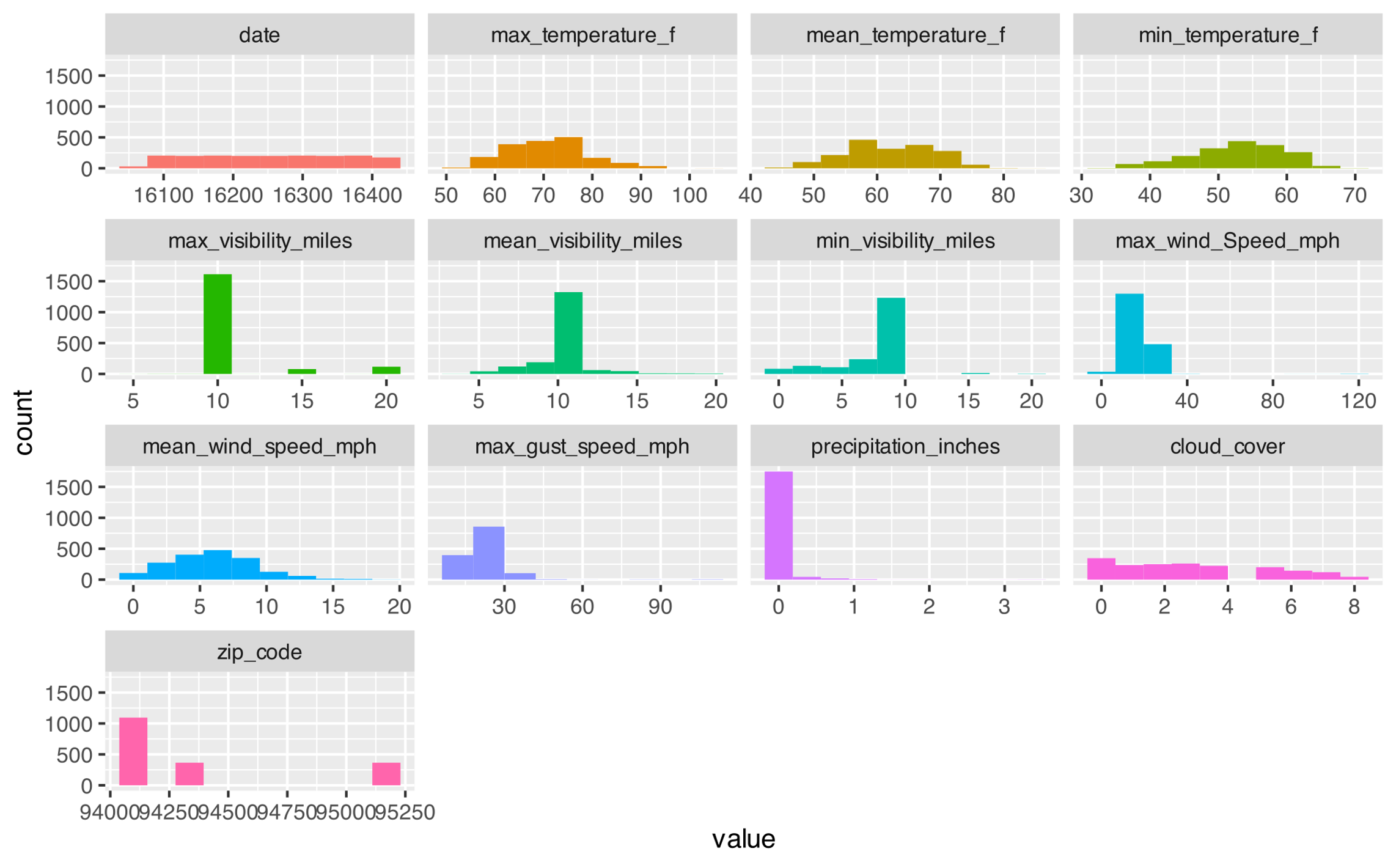
**Table 8.** Frequency of weather events

|  |  |  |
| --- | --- | --- |
| **Event** | **Frequency** | **Percentage (%)** |
| None | 1,472 | 80.7 |
| Rain | 279 | 15.3 |
| Fog | 60 | 3.3 |
| Fog-Rain | 14 | 0.8 |

A graph with different colored squares

Description automatically generated

**Figure 4.** Frequency of weather events including NAs

****

**Figure 5.** Overview of thefrequency of each weather variable

**Data Preprocessing**

**Station Data Cleaning**

Preprocessing station data involved reviewing the dataset's summary to understand its structure and contents. After checking to see if the data was tidy, with consistent variables, appropriate column names, and whether there were NA and blank entries, a check was done for duplicates. This step was important to maintain data integrity and ensure that each record was unique. The original dataset is considered tidy and had no missing values. Since installation\_date will not be used for downstream analysis using the lubridate package, it was not converted to the Date class.

**Trip Data Cleaning**

For the trip data, a similar check was done for tidiness and duplications. To enhance the clarity and usability of the dataset, the start\_date and end\_date columns, initially combined with time information, were separated into distinct start\_date and start\_time as well as end\_date and end\_time columns. Knowing we had downstream analysis on dates and times, it seemed beneficial to separate it into columns.

The start\_date and end\_date columns were converted into the Date class, which standardized the date formats and allowed for analyses to be done using the lubridate package. Additionally, any blank entries in the dataset were replaced with NA to handle missing values appropriately. This step ensured that the dataset was ready for subsequent analysis and modeling.

**Weather Data Cleaning**

The weather dataset underwent a similar review process. One specific adjustment made was converting entries of "T" in the precipitation\_inches column to 0.01. This adjustment represented trace amounts of precipitation and ensured all entries in the precipitation\_inches column were consistent numeric values.

The date column in weather data was also converted to the Date class to ensure that date operations and downstream joining of datasets could be performed correctly. Blank entries in the dataset were also replaced with NA.

**Cancelled Trips**

A cancelled trip was determined by selecting entries with the same start and end station name as well as a duration of less than 3 minutes. There was a total of 1082 cancelled trips.

See “cancelled\_trips.csv” in the repository for a list of cancelled trips that were removed from the dataset.

**Removal of Outliers**

**Station Data**

For the station dataset, outlier detection was not conducted. This was due to the absence of measurable data that could be categorized as outliers. The dataset mostly contains categorical or nominal information about the stations, such as names, locations, and coordinates, lacking quantitative measurements that could include outliers.

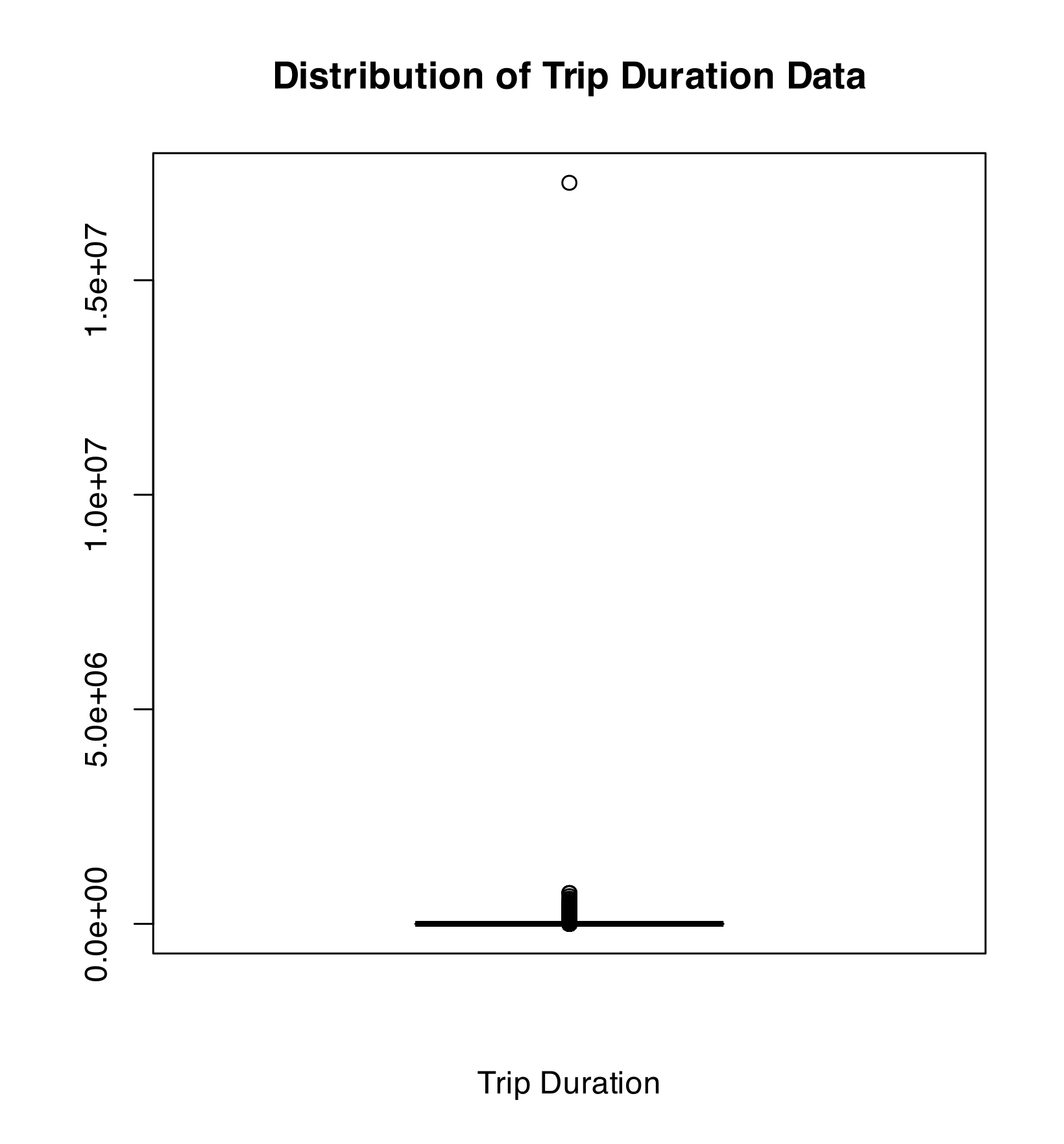
**Trip Data**

In order to remove outliers, a function was created based on the Interquartile Range (IQR). This function calculates the first quartile (Q1), the third quartile (Q3), and IQR for the defined column, followed by the upper and lower bounds for identifying outliers. Any values outside of these bounds were considered outliers. The multiplier 1.5 was chosen as it is the standard threshold used in statistical analysis.

* **Lower Bound**: Q1 - (1.5 \* IQR)
* **Upper Bound**: Q3 + (1.5 \* IQR)

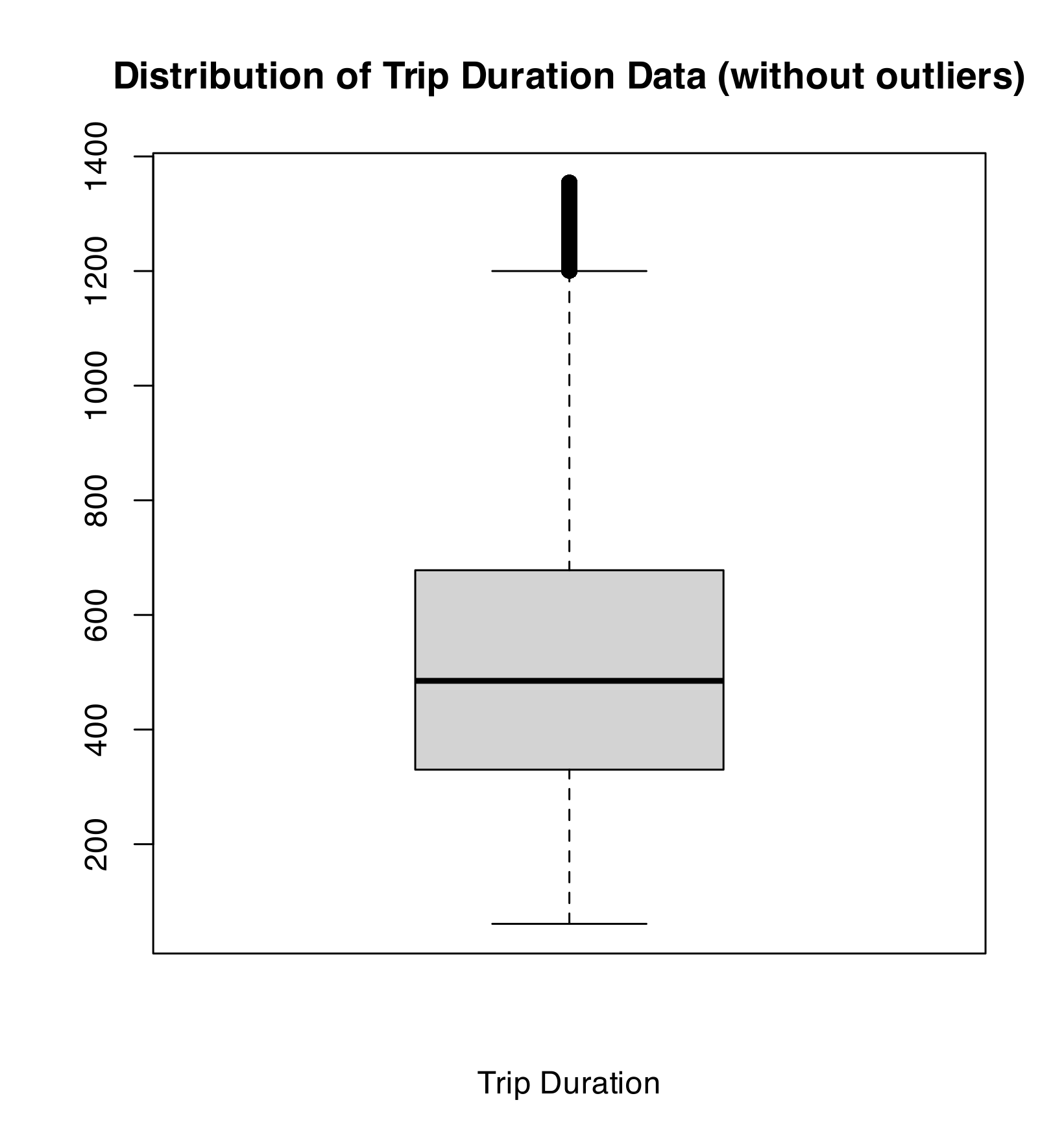
The focus was placed on the duration column of trip data since much of the downstream analysis involving numeric values is done on the duration:

As we can see from the boxplot in Figure 6, there is a prominent outlier in the trip duration data. Upon further investigation, this outlier was identified as a customer, not a subscriber. While it might make some sense for a subscriber to have a longer trip due to their regular access and potential to keep the bike for extended periods, the fact that this outlier is a customer suggests that the exceptionally long duration may be due to unusual circumstances. For instance, it is more plausible that the bike was not returned properly, rather than being used for an extended period as might be the case with a subscription. It was also noted that most of the outliers were customers rather than subscribers (Figure 8).

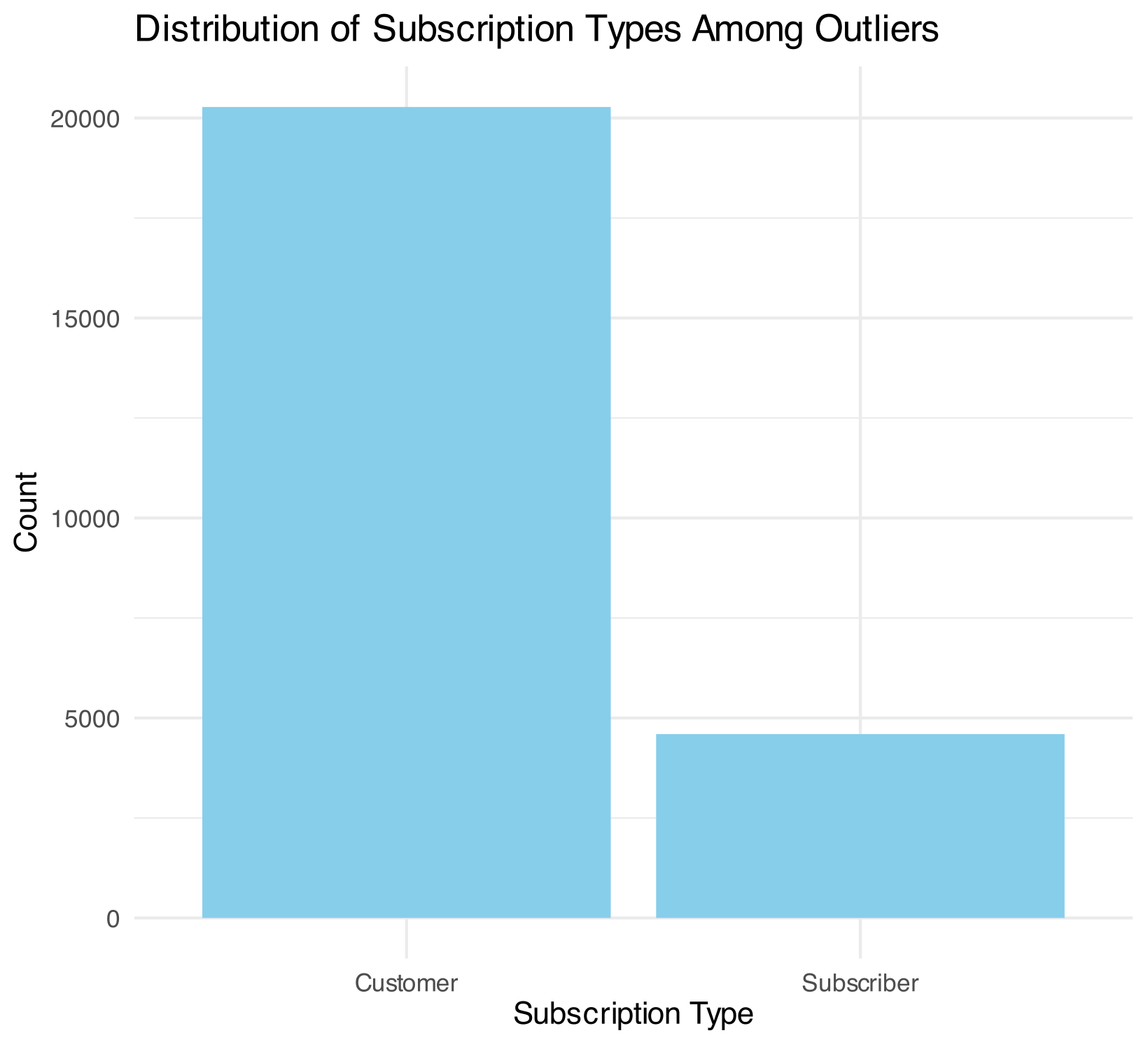


**Figure 6.** Boxplot of trip duration data before removal of outliers

This finding supports the removal of this outlier from our analysis, as it is unlikely to reflect typical trip behavior. Removing such outliers helps ensure that our analysis is based on data that accurately represents typical trip durations and usage patterns (Figure 7).



**Figure 7.** Boxplot of trip duration data after removal of outliers



**Figure 8.** Distribution of subscription types (customer or subscriber) among outliers

A total of 24,874 outlier observations were removed from the dataset. See “trip\_outliers\_ids.csv” in the repository for a list of outlier trips that were removed from the dataset.

**Weather Data**

Similarly, the outlier detection function was utilized to identify outliers across multiple numeric weather variables. During the exploratory data analysis (EDA), it was noted that the majority of the values for precipitation\_inches were zeros. This high frequency of zero values indicates that there was very little variation in the precipitation data. Given this low variation, conducting outlier detection on this variable was deemed unnecessary. As a result, precipitation\_inches was excluded from outlier detection processes and will not be included in the downstream analysis.

A list of outliers was then compiled across the different variables and used to filter out and remove the corresponding outliers from the dataset.

A total of 600 outlier observations were removed from the dataset. See “weather\_outliers.csv” in the repository for a list of outlier indices that were removed from the dataset.

**Findings**

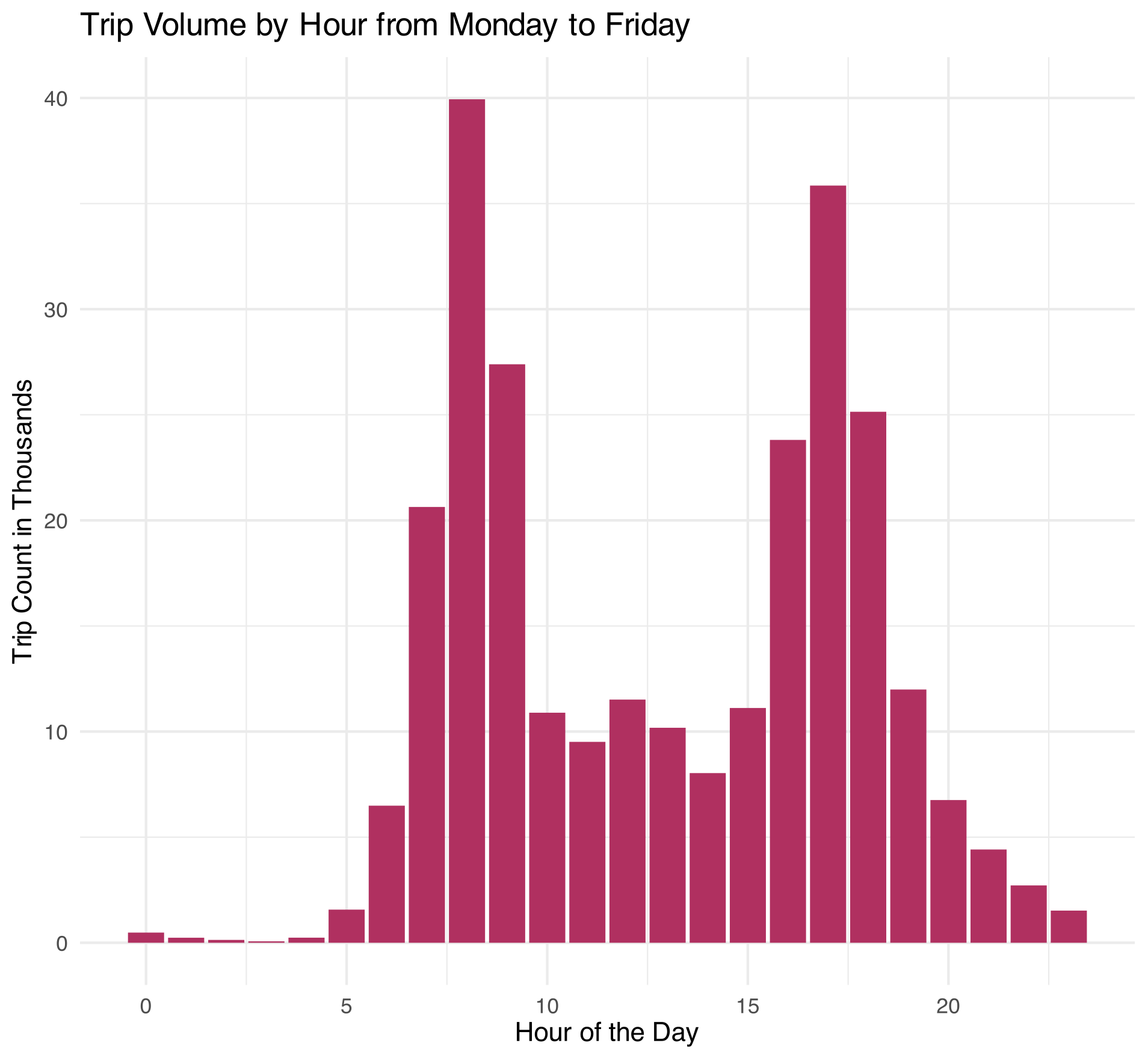
The following analyses show the relationship between bike usage and daily commuting patterns. Understanding these peak hours is crucial for optimizing bike availability and station management. We can use these insights to develop the predictive model that forecasts bike demand accounting for time of day.

*Hours of weekdays where the trip volume is highest*

First, we focused on identifying peak usage hours across all weekdays. By extracting the day of the week and the hour from the start times of the trips, we were able to group the data and calculate the number of trips occurring in each hour for each weekday. This analysis revealed that the highest volume of bike trips occurred during morning and evening rush hours, specifically around 8 AM (Hour 8) and 5 PM (Hour 17) (Table 9; Figure 3).

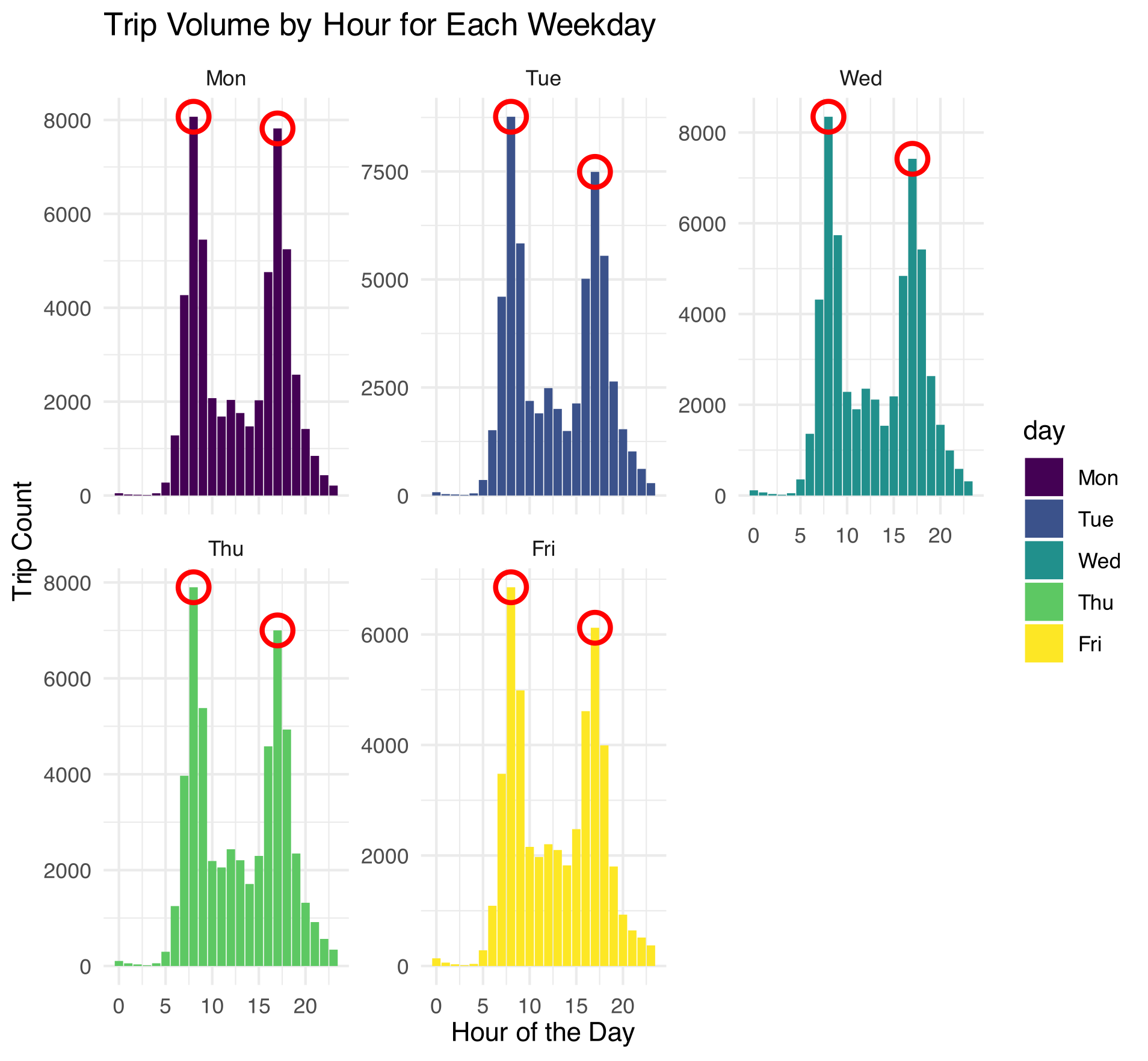
**Table 9.** Top 10 rush hours by trip count across weekdays

|  |  |
| --- | --- |
| **Hour** | **Trip Count** |
| 8 | 39938 |
| 17 | 35853 |
| 9 | 27391 |
| 18 | 25142 |
| 16 | 23809 |
| 7 | 20635 |
| 19 | 11991 |
| 12 | 11514 |
| 15 | 11116 |
| 10 | 10893 |

****

**Figure 9.** Bar graph of trip counts per hour across weekdays with peak hours at 8 AM (Hour 8) and 5 PM (Hour 17)

We then identified the top peak hour for each weekday, which consistently showed peak activity at 8 AM (Hour 8) and 5 PM (Hour 17) for each day (Figure 4).



**Figure 10.** Bar graphs displaying hours with highest trip counts for each weekday

Next, we analyzed the ten most frequent starting and ending stations at 8 AM (Hour 8) and 5 PM (Hour 17) across all weekdays (Table 10-13; Figure 11-12).

*Ten most frequent starting stations and ending stations during the rush hours established*

**Table 10.** Top 10 most frequent starting stations at 8 AM rush hour on weekdays

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| San Francisco Caltrain (Townsend at 4th) | 6070 |
| Harry Bridges Plaza (Ferry Building) | 3426 |
| San Francisco Caltrain 2 (330 Townsend) | 3363 |
| Temporary Transbay Terminal (Howard at Beale) | 3111 |
| Steuart at Market | 1992 |
| Grant Avenue at Columbus Avenue | 1592 |
| 2nd at Townsend | 1296 |
| Embarcadero at Bryant | 1123 |
| Civic Center BART (7th at Market) | 1107 |
| Beale at Market | 1015 |

**Table 11.** Top 10 most frequent starting stations at 5 PM rush hour on weekdays

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| Townsend at 7th | 1843 |
| San Francisco Caltrain (Townsend at 4th) | 1777 |
| 2nd at Townsend | 1699 |
| Market at Sansome | 1687 |
| 2nd at South Park | 1524 |
| Embarcadero at Sansome | 1418 |
| Steuart at Market | 1338 |
| San Francisco Caltrain 2 (330 Townsend) | 1261 |
| Embarcadero at Folsom | 1219 |
| Commercial at Montgomery | 1207 |

**Table 12.** Top 10 most frequent ending stations at 8 AM rush hour on weekdays

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| San Francisco Caltrain (Townsend at 4th) | 3230 |
| 2nd at Townsend | 2408 |
| Townsend at 7th | 2189 |
| Market at Sansome | 2149 |
| Embarcadero at Sansome | 1492 |
| Embarcadero at Folsom | 1436 |
| 2nd at South Park | 1434 |
| San Francisco Caltrain 2 (330 Townsend) | 1397 |
| Howard at 2nd | 1392 |
| Temporary Transbay Terminal (Howard at Beale) | 1315 |

**Table 13.** Top 10 most frequent ending stations at 5 PM rush hour on weekdays

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| San Francisco Caltrain (Townsend at 4th) | 6499 |
| San Francisco Caltrain 2 (330 Townsend) | 3541 |
| Harry Bridges Plaza (Ferry Building) | 2323 |
| Temporary Transbay Terminal (Howard at Beale) | 2125 |
| Steuart at Market | 1963 |
| Market at Sansome | 1451 |
| 2nd at Townsend | 1097 |
| Powell Street BART | 1073 |
| Civic Center BART (7th at Market) | 1047 |
| Townsend at 7th | 980 |

A graph of a number of stations

Description automatically generated with medium confidence

**Figure 11.** Bar graph displaying top 10 starting stations during 8 AM (Hour 8) and 5 PM (Hour 17) rush hours on weekdays

A graph of a number of stations

Description automatically generated with medium confidence

**Figure 12.** Bar graph displaying top 10 ending stations during 8 AM (Hour 8) and 5 PM (Hour 17) rush hours on weekdays

Finally, we found the ten most frequent starting and ending stations during the weekends (Table 14-15, Figure 13-14).

*Determine the 10 most frequent starting stations and ending stations during the weekends.*

**Table 14.** Top 10 most frequent starting stations over the weekend

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| Embarcadero at Sansome | 2145 |
| Harry Bridges Plaza (Ferry Building) | 1924 |
| Market at 4th | 1266 |
| 2nd at Townsend | 1232 |
| Embarcadero at Bryant | 1232 |
| Powell Street BART | 1147 |
| San Francisco Caltrain (Townsend at 4th) | 1080 |
| Grant Avenue at Columbus Avenue | 1028 |
| Market at 10th | 877 |
| San Francisco Caltrain 2 (330 Townsend) | 871 |

**Table 15.** Top 10 most frequent ending stations over the weekend

|  |  |
| --- | --- |
| **Station Name** | **Trip Count** |
| Harry Bridges Plaza (Ferry Building) | 2344 |
| Embarcadero at Sansome | 1664 |
| Market at 4th | 1507 |
| Powell Street BART | 1378 |
| San Francisco Caltrain (Townsend at 4th) | 1355 |
| 2nd at Townsend | 1269 |
| Embarcadero at Bryant | 1125 |
| Steuart at Market | 976 |
| Townsend at 7th | 922 |
| Market at Sansome | 914 |

A graph of different colored bars

Description automatically generated

**Figure 13.** Bar graph displaying top 10 starting stations over the weekend

A graph of different colored bars

Description automatically generated

**Figure 14.** Bar graph displaying top 10 ending stations over the weekend

*Average utilization of bikes for each month*

To analyze bike usage patterns across different months, the total time bikes were used each month was calculated by adding up the trip durations. Additionally, we identified the number of unique bikes used each month based on bike IDs. By considering the number of bikes, we can account for the total available bike time. This approach allows us to better understand usage relative to bike availability.

To evaluate average bike utilization, the total available bike time was multiplied with the number of bikes by the number of days in the month and the total minutes per day. The average utilization ratio for each bike per month was then calculated by dividing the total time used by the total time available (Table 16).

**Table 16.** Average utilization of bikes for each month (total time used/total time in month).

|  |  |
| --- | --- |
| **Month** | **Utilization Ratio (Total Time Used/Total Time Available)** |
| Jan | 0.00734 |
| Feb | 0.00626 |
| Mar | 0.00710 |
| Apr | 0.00775 |
| May | 0.00837 |
| Jun | 0.00918 |
| Jul | 0.00921 |
| Aug | 0.00909 |
| Sep | 0.00942 |
| Oct | 0.01004 |
| Nov | 0.00771 |
| Dec | 0.00617 |

*Investigating weather variables*

To investigate the weather variables and the ones that would be most useful to our analysis of bike rental patterns, the daily trip frequency was calculated. This frequency data was then joined with the main trip dataset to provide a comprehensive view of trip counts alongside other variables. We integrated city information from the station data to link each station to its city and joined this with weather data using both zip code and city. This created a final dataset incorporating trip details, station information, and weather variables (Figure 15).

A diagram of a station

Description automatically generatedA black background with different colored rectangles

Description automatically generated

**Figure 15.** Joining station data, trip data, and weather data into one dataset using matching variables

For the correlation analysis, numeric variables related to trip counts and weather conditions were chosen. After cleaning the data by removing rows with missing values, the correlation matrix was created to assess the strength of relationships between these variables (Figure 16). The correlation analysis revealed no strong or even moderate correlations between trip data (including trip count) and weather variables. As mentioned previously, the variable precipitation\_inches was excluded from the analysis due to its low variation and the high number of zeroes.

A screenshot of a computer

Description automatically generated

**Figure 16.** Plot of correlation matrix between numerical trip data variables and weather variables