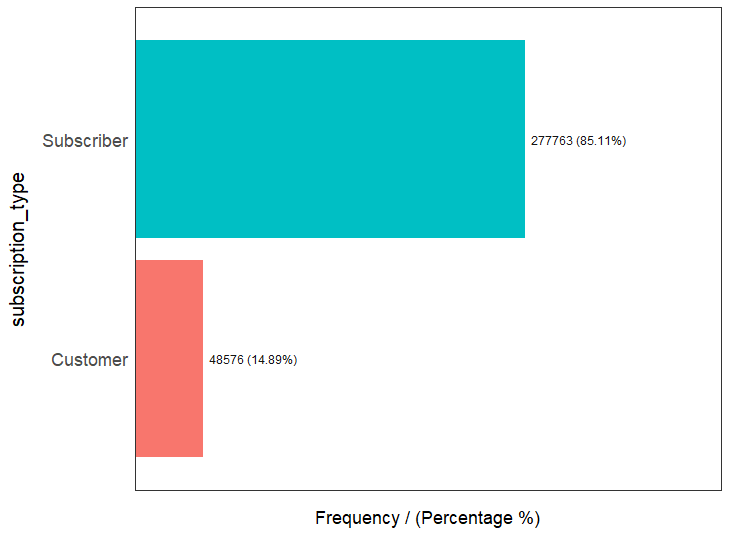
Coding in R Midterm - Bay Area Bike Rental Operation Research

Ahmed Mokhtar

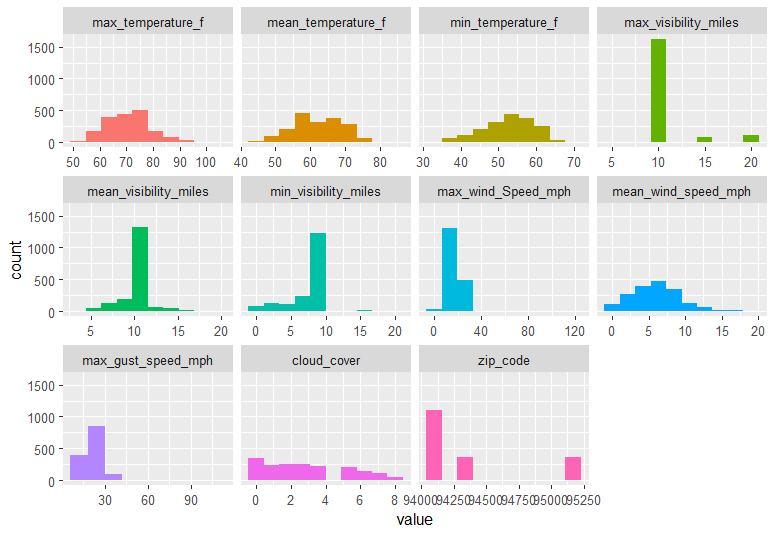
**Introduction and Exploratory Data Analysis**

This report is prepared as part of the Bay Area Bike Rental Operation Research at SF Bay Area Bike Operations HQ. The datasets used in the making of this report spans from January 1, 2014, to December 31, 2014. This encompasses data on the flow of bikes, including when and where they were picked up an returned, station information, and weather information for the specified time period. The three datasets are called “trips”, “station” and “weather” respectively.



**Figure 1:** Distribution of Bike User Status

To start, we will begin with previewing the trips dataset. Figure 1 provides a breakdown of the status of users. 85% of the trips recorded in 2014 were by subscribers, whereas 15% were by customers. Subscribers are defined as individuals with annual or 30-day memberships, whereas customers are individuals with 24-hour or 3-day memberships. (Parry, 2016) The high proportion of subscribers compared to customers illustrates that bike rentals are a commodity that is essential to the lives of hundreds of thousands of people in Bay Area. Additionally, it also reveals that the majority of the customer base are returning, which will be essential to interpreting data such as the utilization ratio and the station frequency analyses.



**Figure 2:** Further EDA for the “weather” Dataset

A detailed EDA was performed on all three datasets. These EDAs were used by the team of data analysts to guide the process of data cleaning. Generally speaking, all three datasets were fairly clean. Figure 2 illustrates relevant data distributions for weather factors in the “weather” dataset. It can be observed that most of the categories exhibit a distribution that is quiet expected. For example, the mean temperature histogram shows a fairly normal distribution, which is to be expected given the fluctuation of temperatures throughout the year.

**Data Cleaning**

Although the data was fairly well structured and organized, there was some cleaning that was performed. The first item of cleaning that was performed was removing trip data for trips that were less than 3 minutes, and were both started and returned at the same station. These trips are void since they were cancelled. As such, I opted to remove these rides from the data to avoid skewing the duration times. The trip ids for all cancelled trips can be found the file “cancelled\_trips\_ids.csv” in the R\_Midterm\_Project repo. Next, in order to perform effective analyses on the duration times, which a key indicator that will be used later in the correlation tests, I opted to remove outliers from the duration column. I classified outliers as any data point that falls outside a range that encompasses 98% of the dataset. This ensures that extremely long or extremely short rides that are not actually representative of most rides are removed from the dataset. The trip ids for the excluded outliers can be found in the “outlier\_ids.csv” file in the R\_Midterm\_Project repo. Lastly, the last major item of cleaning that I performed was replacing “T” denoting trace amounts from the precipitation column with 0.005. The dataset collected precipitation levels in inches. However, on days with levels of precipitation that were less than 0.01, and were considered trace, I marked them with a value of 0.005. (Parry, 2016) The assumption I am making is that 0.005 represents the mean of all trace amounts of precipitation that are below 0.01, and were not recorded.

**References**

Parry (2016). Codebook from Data Source. *Kaggle*.

<https://www.kaggle.com/datasets/benhamner/sf-bay-area-bike-share/discussion/23165>