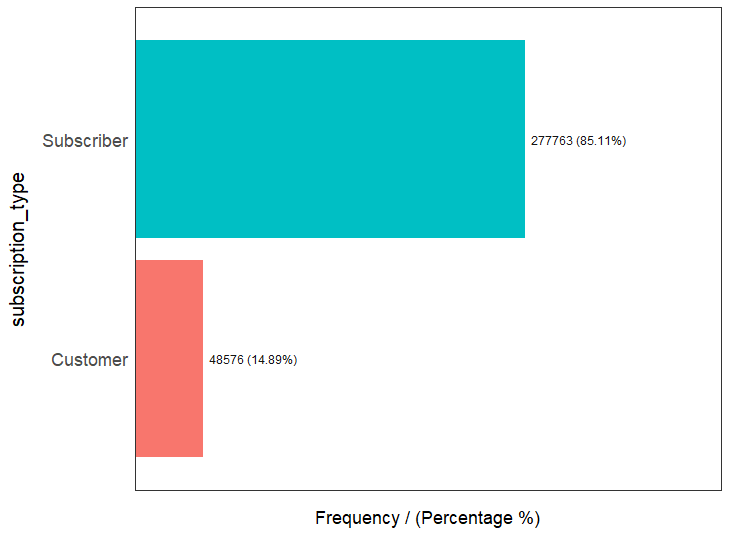
Coding in R Midterm - Bay Area Bike Rental Operation Research

Ahmed Mokhtar

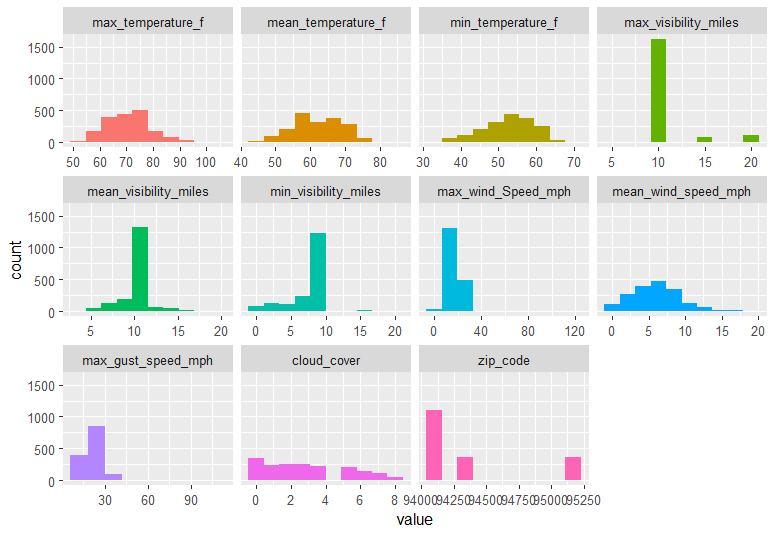
**Introduction and Exploratory Data Analysis**

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

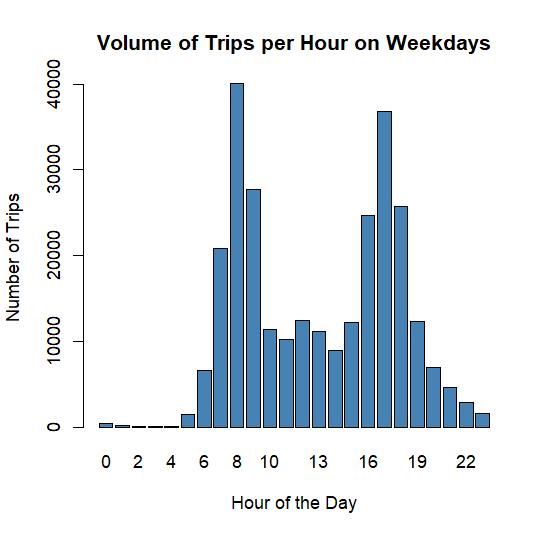
A detailed EDA was performed on all three datasets. These EDAs were used by the team of data analysts to guide the process of data cleaning. Generally speaking, all three datasets were fairly clean. Figure 2 illustrates relevant data distributions for weather factors in the “weather” dataset. It can 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.

**Rush Hour Analysis**

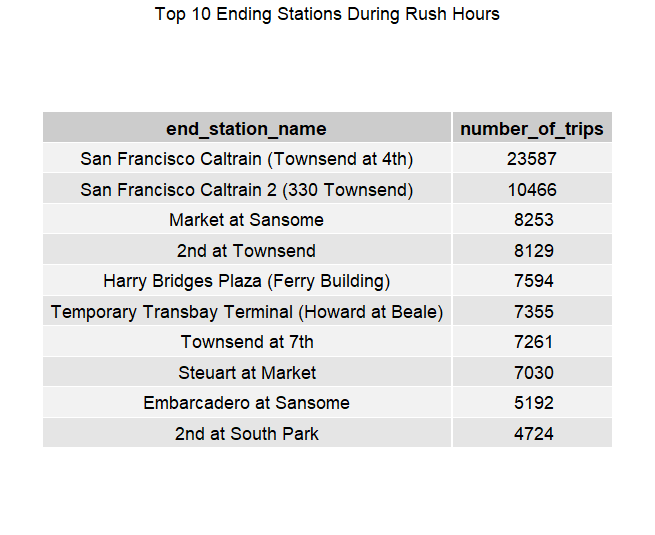
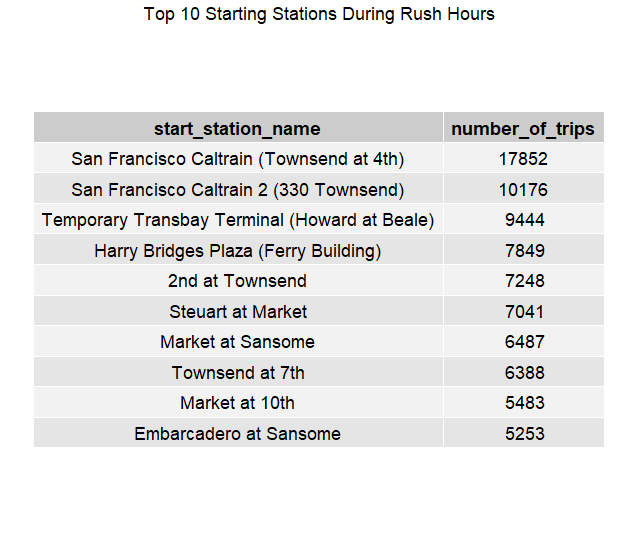
The next step in our analysis was to establish a breakdown of the number of trips per hour on weekdays. The idea behind this is that this would provide insight into when bikes are highest in demand, and allow us to potentially understand the population demographic that uses the bike rental services most frequently.



**Figure 3**: Breakdown of the total number of trips by hour on weekdays.

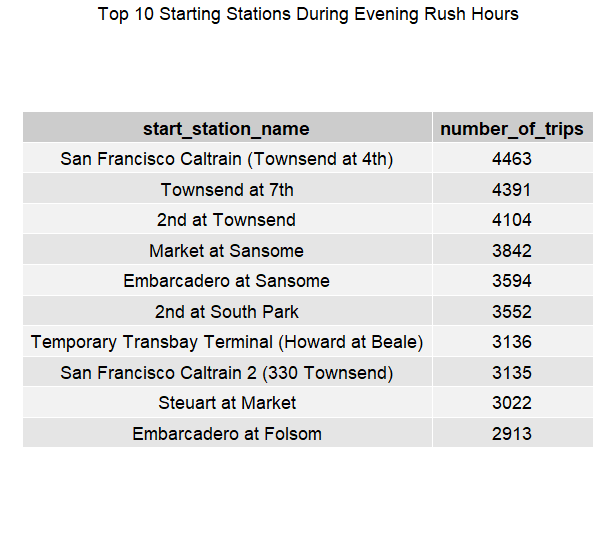
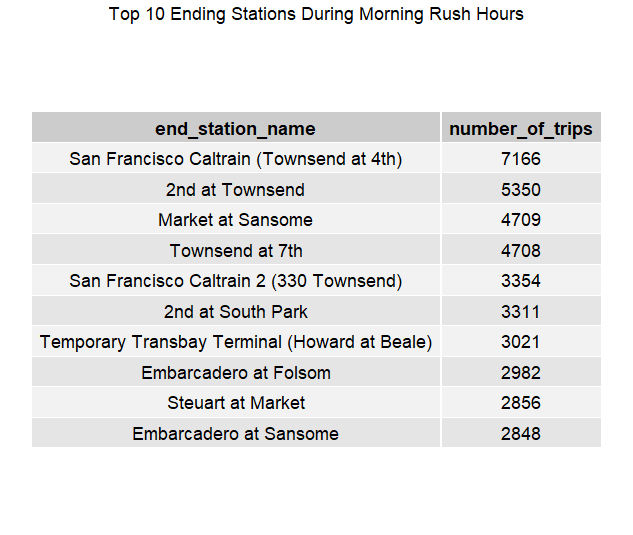
Figure 3 clearly shows some of the trends that might be expected. Bike rentals at midnight and into sunrise hours are very close to 0. This can be attributed that most people are asleep at this time. In contrast, it can be observed that bike rentals peak twice a day. The first peak phase is between the hours of 7-9 am. This can be attributed to the fact that a large segment of the population is commuting to their day jobs. A similar trend can be observed in the hours of 4-6 pm, which is the time when most people finish their workday. These findings are consistent with the fact that the majority of bike users in the Bay Area are subscribers, who rely on the use of bikes to commute to and from work on a daily basis. Another very interesting finding is that at 12pm, there is a small increase in bike rentals. This can be attributed to noon being the traditional lunch time that most people have.

**Station Frequencies**



**Figure 4**: Illustration of the most common starting and ending stations during rush hours.

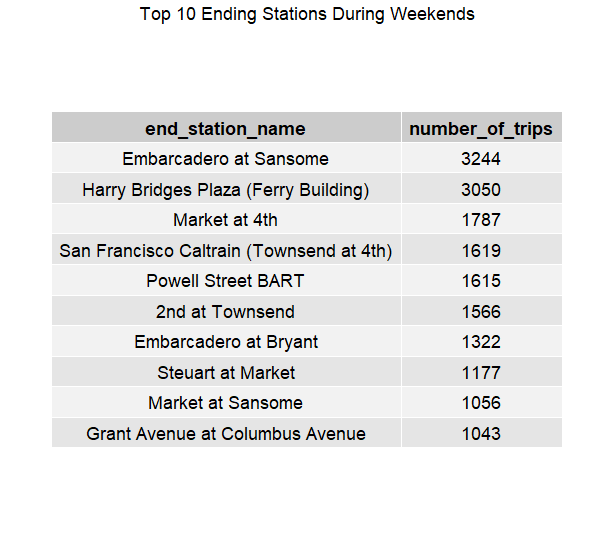
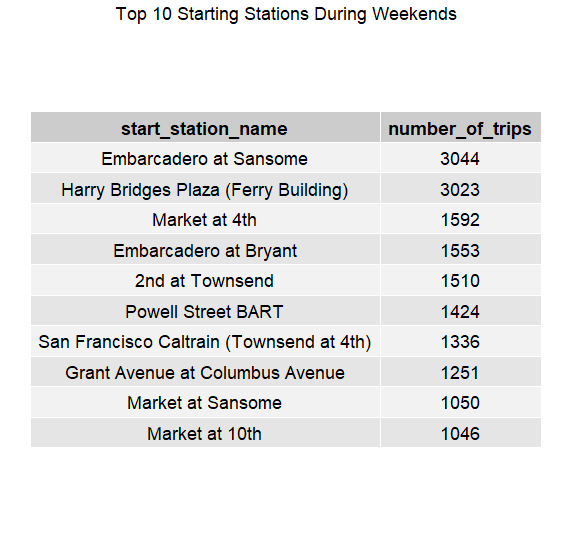
Now that we have an understanding of the two rush hour peaks and some of the potential causalities behind why they occur, we can start analyzing the traffic per station. Interestingly, the two top stations in terms of starting and ending during the rush hours are the same in both cases. However, in order to develop a better understanding of the meaning behind this, I decided to examine this data based on the rush hour that they occur. I collected the top ten starting and ending stations during morning and evening rush hours separately, and the results were interesting.



**Figure 5:** Illustration of the most common ending stations during morning rush hours, and most common starting stations during evening rush hours.

Following the theory that workers commute to work in the morning rush, and commute back home in the evening hours, it can be assumed that the morning end station and the evening start stations are the same, since they arrive and depart from the same place. Figure 5 shows data that tests that hypothesis. It is indeed observed that all 10 morning ending stations are the same as the 10 most commonly used evening starting stations. It can be further observed that most of these stations maintain about the same level of traffic during those times.

Now that we have an understanding behind the traffic during weekdays, and have broken it down by rush period, it is also necessary to understand the station frequencies during weekends.

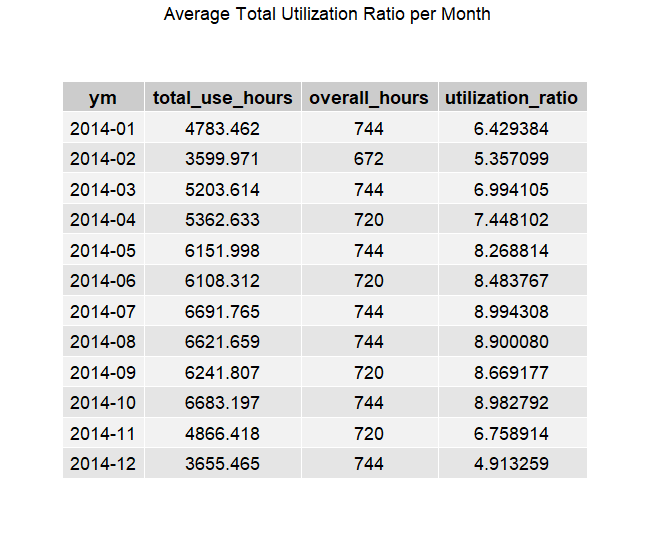


**Figure 6**: Illustration of the most commonly used starting and ending stations on weekends.

Traffic flow on weekends can be more ambiguous than weekdays. However, given that most people have their typical workdays on weekdays, weekends are usually spent relaxing. This can be observed by the significant reduction in the number of rides during weekends compared to just rush hours on weekdays. However, one interesting observation from Figure 6 is that the top stations on weekends are completely different than those on weekdays, and are also the same for starting and ending. This can be potentially attributed to the fact that these stations are closer to more recreational destinations in the Bay Area. As such, workers and children who are free on weekends exhibit vastly different traffic behaviours. The second most common station for both starting and ending trips on weekends is Harry Bridges Plaza, which is a common tourist and recreational attraction.

**Utilization Analysis**

Now that we have an understanding of the typical bike use during the week. It is necessary to understand the varying bike utilization during the year. In order to do that, the utilization ratio was calculated. The utilization ratio is calculated by dividing the time that bikes were being used, by the total time available. To perform this calculation, I opted to pool the data into individual months of the year. This was followed by calculating the ratios as a whole, and not per bike. Since this research aims to review bike operations in the Bay Area on a systemic level, I figured that gathering the total utilization ratio is much more beneficial and meaningful to that purpose, that calculating indvidual bike ratios.

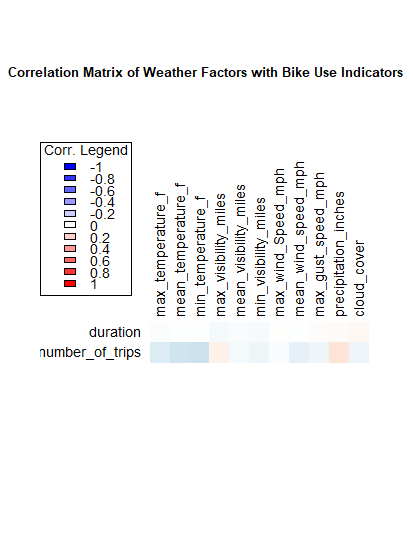


**Figure 7**: Illustration of the average total utilization ratio per month of the year.

Figure 7 illustrates that there is significant variance in bike utilization throughout the year. It appears that bikes are more frequently used in the months of May-October, with a peak in July. Whereas the ratio plummets in the months of December-February. These findings make sense considering that people are more likely to resort to other forms of public transportation during the winter months considering the weather difference. It is also noticeable that bike use in December specifically is much lower than the rest, which can be attributed to Christmas time, when people spend more time at home.

**Weather Analysis**

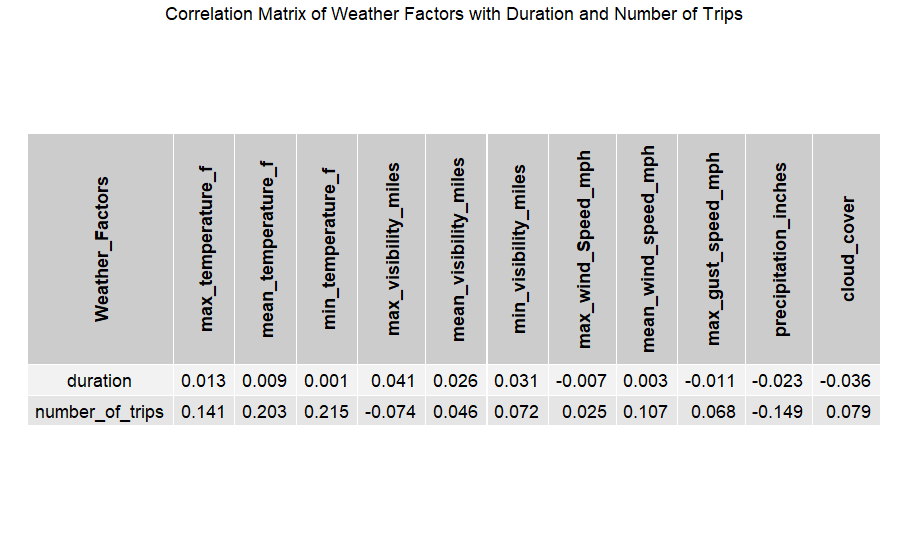
Now that we have established that weather could be a potential influencer to lower utilization ratios in the winter months, it is necessary to investigate that hypothesis. The weather database contained pertinent daily weather data corresponding to each city in the dataset. The first step that was performed was joining all of the data sets into one, in order to streamline the work process for the data science team. This can be found in the R script. After joining all of the datasets together, with the relevant city and weather information corresponding to each ride, the correlation analysis was performed. Rather than comparing correlations between weather factors together, I thought it would be much more insightful to see how weather factors influence indicators of bike use. The two indicators selected were duration and number of trips per day. Since the number of trips per day metric was not included in the dataset, it was calculated with R, and attached next to each trip.



**Figure 8:** Correlation plot of the various weather factors, and their influence on duration and number of trips per day.

As shown in Figure 8, there does not appear to be any stong correlations between the various weather factors in the dataset and the bike use indicators. The strongest correlation present appears to be mean temperature and the number of trips, however this remains a weak correlation. It can be further observed that duration does not seem to have any correlation at all with weather factors. This can be attributed to the fact that the distance that workers or subscribers cover on each trip is the same, assuming that they only use the bikes to commute to work and back. As such, varying weather conditions do not influence the distance that they have to cover. However, although it is weak, there is more correlation observed between number of trips and the weather factors. This can be attributed to the fact that regular customers, who are not subscribers, are more likely to use public transport, and avoid using bikes even as a method of recreation. As such, the number of trips does go down because there is less randomness induced by non-subscribers. Overall, these findings do no reflect the expected trend that can be observed in the utilization table. It may be attributed to the limitations of the weather metrics of the dataset. In order to perform a more meaningful investigation, I suggest consulting with field expert about what weather metrics to use in order to build the predictive model. The appendix section of this report contained a figure with the exact correlation values.

**Appendix**



**Figure 9:** Exact correlation values for the weather correlation analysis.

**References**

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

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