**Report of Findings**

*Introduction*

As part of this course project, I will be analyzing the Hubway bike sharing system in Boston and the near suburbs. The program managers have requested insights into the system usage, particularly to better prepare for inquiries from city managers at budget time. My analysis will focus on answering the following questions: Which stations are the most and least used? Which stations are considered "home-based" (i.e., ride ends and begins there) and which are "commuter" (i.e., ride begins there and ends elsewhere)? For commuter stations, are the sending stations characterized by morning (net outflow in the morning), evening (net outflow in the evening), or both (balanced)? Is there a problem with bikes being taken but not returned? Were there any stations at capacity, to recommend for expansion? Which age group or gender used ride shares more? Any particular week days or hours that have more bike borrowing? Is trip duration dependent on the age of the customer, or is it dependent on distance between stations? Do registered or casual users have more bike trips? Is there a correlation between age and duration of the trip? Are there any particular days where capacity is exceeded? Is there a seasonality to bike rentals? Are there outliers?

To start, I will assemble the necessary data and formulate a strategy for completing the project. It is important to ensure I understand the questions being asked, and I will seek clarification if necessary. As part of my analysis, I will address the program managers' questions, as well as the additional questions posed. I will identify any stations that are at or near capacity and recommend them for expansion, if necessary. To understand the demographics of bike share users, I will analyze the age and gender of the customers who use the system. I will explore the usage patterns of the system by day of the week and time of day. To understand how trip duration is affected, I will investigate whether it is dependent on the age of the customer or the distance between the stations. To understand the usage patterns of registered versus casual users, I will compare the number of bike trips made by each group. I will explore whether there is a correlation between age and duration of the trip. I will investigate whether there are any particular days where capacity is exceeded and identify any potential reasons. To understand whether there is a seasonality to bike rentals, I will explore whether bike usage varies throughout the year. I will identify any outliers in the data that may be affecting my analysis. By addressing these questions, I hope to provide valuable insights into the usage patterns of the Hubway bike sharing system and help program managers better understand the system's usage and potential areas for improvement.

*Results*

1. *Which are the most used stations?*

22 1924.0

38 1592.0

36 1571.0

42 1209.0

16 1028.0

33 979.0

60 978.0

52 952.0

53 931.0

1. 900.0

*2.* *Which are the least used stations?*

93 3.0

92 4.0

96 13.0

95 16.0

87 19.0

77 20.0

94 20.0

97 21.0

85 22.0

83 23.0

*3. Which are “home-based” stations (i.e. ride ends and begins there)?*

There are 1374 stations which are home based.

1. *Which are “commuter” stations (ride begins there and ends elsewhere)?*

There are 18624 are commuter stations.

1. *For the commuter stations, are the sending stations morning (net outflow in the morning), or evening (net outflow in the evening), or both (balanced)?*

PM 14636

AM 3988

1. *Is there an issue with bikes taken and not returned?*

No, there were 2 rows with no start station and no end station but there were no rows with a start station and no end station meaning the bikes were not returned. This means all bikes taken were returned in this data set.

By using the status variable, we can see that closed is 19998.

*Are there any other questions you think you should be prepared to answer?*

1. *Which age group or gender used ride shares more?*

Male used ride shares more being 3/4 of users.

1. *What percent were registered vs casual users?*

65 percent were by registered users, indicating casual users are a smaller percent.

1. *Which months had the largest rentals?*

July-September are the months with the largest percentage of bike rentals and this is their summer season.

1. *Which day is the busiest and slowest?*

Friday is the busiest. Saturday is the slowest day.

1. *Which are the Outliers?*

I first found the z-scores for all of the observations using the standardization formula:

[(value-mean)/standard deviation]. Then I choose 2 as the cut off for the z scores to identify any outliers. There are 97 outliers in the duration column out of 19998. There are 60 observations with a z score greater than 3.

1. *Did any stations exceed Capacity?*

I used pd.concat to concatenate the two data frames. A new column/data frame was created called 'difference of capacity' which yielded all negative results. This was then set to greater than 0 in a new created column/data frame called exceeds\_df. This data frame was empty since all of the values were negative. Hence, none of the stations exceeded capacity.