Mildred Delgado

BUS211a

Group 1

Public vs. Private Transportation

Transportation is an interesting topic, especially when comparing public and private transport when gas prices fluctuate. The goal of this project was to understand if there was an effect on public and private transportation when the gas prices peaked and the orange line shut down in Boston. In order to do this, group one set up a hypothesis to test if the data frames were biased to meet its goal. They prepared the data, removed some missing values, separated public and private transportation, analyzed the number of visits, and added some graphics with trends to draw conclusions. Some of the considerations for this analysis were to count the number of visits to gas stations and public transportation stations under the POI category. Therefore, since the datasets presented information about the necessary POIs, this project was suitable for further consideration. To evaluate this process, it is necessary to review the reasoning behind the codes used.

At first, when joining the data frames places, and patterns, the team used the function merge without removing the repetitive columns to avoid making the environment noisy. They could have used the function left\_join and removed columns 2:5 and 11:14. Moreover, at the time to look for missing values under the top\_category column, using the keyword “MBTA” under the location name was an interesting approach; however, the downside is that there may be some locations that should be considered under stations but because they do not have “MBTA” listed, they were left out. It would have been fascinating if they run some tests using the naics\_code column and filter it by 447110 first, which is the naics code for gasoline stations. After finding the missing values, they looked for specific names under top\_category such as “Transit and Ground Passenger Transportation” and “Rail Transportation”, which they may have researched in Google to be very common in Boston; however, there is a possibility that other gas stations that do not match those listed were left out. Additionally, there could be double counting from people sharing a car and are considered under the private transport category or an overlap from other businesses. Even though these are some of the disadvantages, the datasets are still a good source to drive some analyses.

Besides, the team used the Chi-square test, which is a statistical way learned throughout the course, to see if the sample was biased. To get the observed values, they did some manual sum and obtained the public transportation observed (pts\_obs). It would have been preferable if they used two separate codes to get both results (gas\_obs and pt\_obs) to avoid any errors from inserting the numbers manually. Especially when the data gets bigger, it will be difficult to get track of the numbers and add them up to get the desired results. When running the expected counts for the actual number of gas stations and public transportation POIs, they may have searched the number online and applied the formula, which was very creative to do. Then, they obtained 0.69 as the result of their hypothesis test. This was a great way to see how to use Chi-square for the hypothesis testing.

Next, team one filled out the missing values found under the top category column. This helped to clean up more the data for further analysis because instead of removing the missing values or just leave them in blank, they entered the values that they got from the other non-empty rows. They also grouped the data by date\_range\_start, which help to get the total amounts of visits to gas stations per month. Most groups did this do avoid double counting. It was a good idea to create the new column “Cate” to distinguish private from public transport to pursue the main goal. Nevertheless, instead of using a number for the month, they could have used the complete or an abbreviated word of the month such as Jan, Feb, and Mar to make it easier for the reader to go through the monthly data.

Finally, the fact that they used the methods of normalization and standardization was very unique. They tried both approaches, which made the argument stronger. The team ended up using the method of normalization, which is helpful mostly when people do not know the distribution of the data or when the distribution is not a bell curve. In this case, the distribution does not follow a bell curve figure, so it was great that they chose the normalization method. At the time of presenting the results, it was nice that they used ggplot to show the trend in a friendlier and visualizer way. As I mentioned before, writing the months in words instead of numbers would have helped the reader to understand better the graphic. Furthermore, if they have named the months, it was not necessary to write (starting from January) under xlab.

In overall, transportation and the effects that the increase in gas prices can produce is a relevant topic that affect most users. These datasets were useful to analyze if there was a specific trend and draw some conclusions from it such as those stated by the team “there is a decrease in private transportation when gas prices rise and consumers seems sensitive to price” and “there is no definite conclusion, there is a slight rise in public transport and no rise in private transport”. As in most statistic approaches, there are usually some downsizes such as if the datasets are biased or are not representative. Some statistical tools such as the chi-square, normalization, and standardization methods can be helpful to test the datasets.