Name: Shipeng Ji

Assigned Group: 1

Group 1 conducted a study on the effects of gas prices and public transportation disruptions on transportation trends in Boston. They used cell phone data to track visits to gas stations and public transportation points of interest. After validating the data was not significantly biased according to a Chi-square test, they then went on to analyze the data and test their hypothesis. Group 1 found that the increase in gas prices had a negative impact on the number of visits to gas stations, indicating that people were sensitive to the price increases. They also found that the shutdown of Orange Line during August had a negative impact on the number of visits to public transportation points of interest, likely due to the inconvenience it caused for people who relied on that line.

While this study has some strengths, such as using a Chi-square test to validate the data and clearly articulating their research questions and hypotheses, it is important to note that the results may not be representative of long-term trends in Boston. Also, there may be other factors at play that could affect transportation trends in the city, such as changes in population or employment levels.

The findings of this study could be useful for policymakers and transportation authorities in Boston as they work to improve and maintain the city's transportation systems. Further research would be needed to fully understand the complex factors that influence transportation in Boston and other cities.

Group 1 used the tidyverse and caret packages for data manipulation and visualization, respectively. The code first loads two datasets using read\_csv, and then merges them using the merge function. The resulting dataset is explored using the glimpse function. The code then performs a Chi-square test on the data to determine if the sample is biased in terms of the number of points of interest (POIs) for gas stations and public transportation stations. To do this, the code filters the rows with missing values in the top\_category column, and then groups and counts the POIs in each category. The resulting data is then used to calculate the observed and expected number of POIs in each category, and the p-value is calculated using the pchisq function. As they found that the p value was not small enough to reject the null hypothesis, indicating that the sample is not biased, they moved forward to clean and manipulate the sample data. Specifically, they created a new column to divide the top category values into private and public transportation categories, translated the date field into the corresponding month, and scaled the data using normalization. They then used ggplot to create a graph showing the trend of private and public transportation usage over the period of their data set. The results showed that private transportation usage was consistently higher than public transportation usage and that the gap between the two categories decreased when guest prices increased, but the increase in public transportation usage was not substantial. In conclusion, they found that gas price increases did lead to a decrease in private transportation, but it was not necessarily replaced with increased use of public transportation. However, they were unable to reach a definite conclusion about the effect of the orange line shut down on transportation patterns.

Overall, Group 1’s code is clearly written and well-documented. It includes comments explaining the purpose of each step and uses appropriate functions and techniques for data manipulation, filtering, and statistical testing. It also follows a logical workflow, starting with data loading and preparation, followed by a chi-square test to check for bias in the data, and ending with an analysis of the data to test the research hypotheses. However, it seems they were focusing too much on the data preparation section, while paid less attention to generate finding from the data. Therefore, one potential improvement to the code might be to include more visualizations or summary statistics to help readers better understand the characteristics of the data and the relationships. It would also be helpful to conduct regression analysis to further examine their hypothesis.

As extensions to Group 1’s research question, more outputs will be generated if researchers could:

1. Gather data from multiple years to see if there are any seasonal or annual patterns in transportation usage.
2. Examine the effect of other factors on transportation patterns, such as weather, events, or promotions.
3. Expand the analysis to include other modes of transportation, such as biking or walking, and comparing their usage to private and public transportation.
4. Analyze the effect of different types of public transportation, such as bus, subway, or rail, on transportation patterns.
5. Examine the effect of different price points or ranges on transportation usage, to see if there is a threshold at which people are more likely to switch to public transportation.
6. Compare the result of Boston with those of other cities or regions to see if the patterns of transportation usage are similar or different.