**Student’s Name:** Djounia Saint-Fleurant

**Group Assigned:** Group 1

Group 1 has conducted their research about the use of public and private transportation in Boston. They have searched to determine whether the increase of gas price and the orange line shutdown had impacted people' decisions to use public or private vehicles. Their research question is interesting as long as they have used appropriate and resourceful datasets to process their analysis and draw their conclusions. The group wishes to analyze the sensitivity of the population to price changes and understand how the inconvenience caused by public transportation could push people to ride privately.

With two datasets (place and pattern) primarily based on POI (Point of Interest), it is difficult to determine the total number of visitors who travel using a private car or public transportation. However, the group came up with a great idea: using gas, bus, rail, and train stations/ routes as points of interests to determine the visit trends. Although we could tell it was the best method they could have found, it was full of limitations and raised many interrogations:

* What about people using electric vehicles?
* How will they consider alternative modes of transportation such as Uber and Lyft?
* How about public transportation specific to a group (school buses, for example)?
* What if people fill up their cars in neighborhood towns instead of in Boston?

Additionally, there are other limitations mentioned by the group, such as:

* the number of passengers in a car;
* the other modes of transportation (walking and biking);
* the time of the year, etc.

Those interrogations needed to be taken into account if the group wanted to ensure that the findings to their research questions were accurate. However, a quick glimpse at the Safegraph Datasets was enough to show they were insufficient to address those hypotheses efficiently.

**Main content 1: Representative Data**

As Group 1 is conducting its research in Boston, the sample it will choose to analyze should be representative of the population of Boston. In this case, there are two main characteristics in the picture: those who ride public transportation and those who use private vehicles. However, as mentioned above, the mode of transportation in Boston goes beyond buses, trains, rails, and gas-powered cars. In 2021, there were more than 18 million rideshares in Boston and approximately 21,500 electric vehicles in Massachusetts. Those data represent a significant amount that is not included in the chosen sample. By filtering and selecting a sample that excludes ride apps, taxis, scooters, and bikes, we can't say that it is representative of Boston's population.

The other consideration is that when the orange line shut down for a brief and specific period, the city didn't entirely stop public transportation on those routes. According to Mass Transit, the orange line allows 10,000 people to move around Boston. The city had to come up with strategies to facilitate those citizens who rely on public transportation to go to work. They introduced extensive bus-only areas, offered free 30-day passes to its Bluebike bike-share system, and made many other adjustments. If those POIs aren't included in the sample, they won't represent the population at the period they're trying to analyze.

Regardless, the data-cleaning methods used by the group were accurate. They merged the two datasets, looked for missing values, and assessed whether the dataset was biased. They found that they couldn't reject the null hypothesis and concluded that the dataset wasn't biased. Nonetheless, the fact that they had well-processed the data they were going to use didn't change much from the abovementioned issue. The Safegraph datasets (place and pattern) wouldn't help them answer their research questions efficiently.

**Main content 2: Code Review**

Looking closely at the codes, I realized that Group 1 had merged the two datasets and kept the total 55 variables while they could have used left\_join and only stowed the necessary columns for their analysis. Additionally, the chi-square test was performed over the entire dataset without prior selection of the observations they needed to answer their questions. I was expecting them first to filter and group\_by the variables and save it under a different name.

1) Messy Dataset

The first challenge they could have dealt with was a messy dataset. They have partially made a good selection of the part of the dataset they are planning to use and looked for missing values, although they haven’t specified whether they will keep or delete them.

2) Biased Dataset

For what has to do with the second challenge, they have conducted a thorough analysis to determine if the dataset was biased and concluded that it wasn't. The process followed is overall great, and the codes run well for this part.

**Conclusion**

The team did a great job with the overall analysis. However, the datasets don’t fit with the research questions and don’t allow to draw an accurate conclusion. If they had used a traffic or any transportation-related dataset, the results would have reflected the reality a little bit more.

**Bibliography**

Division, TNC. “2021 Rideshare Data Report.” *Mass.gov*, <https://www.mass.gov/info-details/2021-rideshare-data-report>

“Ma: What Boston Learned about Transit Planning from a Subway Shutdown ...” *Mass Transit*, <https://www.masstransitmag.com/management/news/21282898/ma-what-boston-learned-about-transit-planning-from-a-subway-shutdown>

Stacker, and Nexstar Media Wire. “See How Many Electric Vehicles Are Registered in Massachusetts.” *WWLP*, WWLP, 27 May 2022, <https://www.wwlp.com/news/local-news/see-how-many-electric-vehicles-are-registered-in-massachusetts/>