# **I) Specific area of interest and its importance**

# This paper will analyze the trends of public and private transportation in the Boston area from January 1st, 2022 to September 21st, 2022 and look at major changes during the same period. Specifically, the analysis will be focused on how private and public transportation was impacted by the increase of gas prices. In the United States, gas prices peaked on June 13th, 2022 at [$5.10 per gallon](https://ycharts.com/indicators/us_gas_price) and look at individual preferences of transportation before and after this time. We would also like to explore the impact that the shutdown of the orange line on [August 19th, 2022](https://www.boston.gov/departments/mayors-office/orange-line-shutdown-boston) had on public and private transportation in Boston. Lastly, we would like to explore the impact of the start of the new school year for Boston Public schools on September 8th, 2022. We want to analyze the influence of all these major events on transportation preference. The trends in public and private transportation are important as sustainability has become a key topic in Boston and around the world. Boston has set a [carbon neutral date in 2050](https://www.boston.gov/environment-and-energy/reducing-emissions), which is heavily reliant on personal transport reduction, and the implementation of more public transportation as well as better city planning to accommodate other modes of transportation.

**II) How does the cell phone tracking data fit our interests?**

The Safegraph Data, both places and patterns when joined, has top\_catergory, sub\_category, naics\_ codes, which will all be helpful in determining the usage of both personal cars and public transportation. There is an acceptable range of answers to be searched for each entry. Visitor counts can then be analyzed. The top\_category and sub\_category columns can be searched for keywords such as “transit” and “parking lots.” The NAICS codes we wish to use are searchable on the [census website](https://www.census.gov/naics/), from there we can search for NAICS codes applicable to public and private transport areas.

Certain columns such as city, region, or postal code, among many more, will not be useful in our analysis, as we know our data has been taken from Boston.. Certain columns such as device\_type may prove useful if we choose to look into the demographics of individuals using public versus private transportation, however, this will come after collecting columns that fit observations we are interested in. We will need to remove duplicates, and empty entries, in both rows and columns.

From this, we can look at the proportion of different people using public transportation versus private transportation and be able to infer the consequences of each distribution, whatever it may be.

# **III) Main challenge in answering the question**

One challenge our project has is determining which questions to focus on answering and therefore, which dependent and independent variables to include. The places and patterns dataset, merged together include many possible directions to take regarding the usage of both personal cars and public transportation. While analyzing the direction we want this project to take, we made the decision to focus on analyzing the relative change between the popularity of private and public transportation from January 2022 through September of the same year.

Another challenge we face is the number of rows that will be plausible to analyze. By deciding on a number of questions, there will be specific criteria that each observation will have to meet, which concludes in a small fraction of the total merged dataset, since the POIs we will analyze only concern public transportation and parking services.

The final noteworthy challenge our project will face is our results and more specifically the applicability of our results. The tests and analysis that we will do on our dataset will be limited to the observations that meet our criteria of acceptable data entries, as discussed prior– entries missing data and duplicates will not be included. While our methods can be thorough, we may not be able to draw any clear conclusions about the “whole picture”, because of the insufficient sample size.

# **IV) Data manipulation to answer the question**

We need to do basic data cleaning to our data set so that our analysis is based on sound and valid observations. This will also make sure the analysis yields reliable and meaningful insights.

The following methods will be used:

To begin, we are planning to remove duplicate and/or irrelevant columns. This can be done by looking at the data set and deciding if we keep all the columns. Given the nature of joining two existing datasets, the merge data set contains a few duplicate columns such as City, State, Zip Code, Brand, etc. Likewise, we can also see that there are several columns in the dataset that may not be useful enough for our analysis, for example, the Parent Placekey and phone number columns. Removing duplicate and irrelevant columns can make our data set thinner, smaller, and overall, easier to manipulate. This will create a more meaningful data analysis and a more robust performance overall.

Next, we can identify potential structural errors. It may not be the case for our SafeGraph dataset, but it’s vital to check whether we have structural errors here. Common errors include multiple expressions for empty values, inconsistent data types within columns, or misplaced data in columns. By screening the data we have, we can make sure if the data set has clearly incorrect column types. We can make sure that columns consisting of strings, numeric values, and date times are correct.

Our third step is to find outliers and determine whether to exclude them. We have a number of methods to determine if we have outliers in some of the most important numeric columns. Given the relatively simple nature of the numeric columns we have, it makes sense to use sample variance methods to see how the data is distributed. We still have to see whether it’s suitable to exclude the outliers. It is very unlikely we have extreme values for Visit Counts, Distance from Home, etc.

The last step we are planning to do is handling missing values in observations: even though we are excluding many irrelevant columns from our analysis, it’s inevitable that we still have missing values in the columns that we are interested in. We are planning to look at some of the important columns to investigate whether there are missing values and whether we should eliminate the observations or fill them in with other statistical numbers.

I really liked Part I. The questions are not only well-defined, and they address significant matters in real life. Analyzing the public transportation and parking services would provide some hints to your questions, although the link is relatively weak. We might infer the transportation choices from the traffic on these POIs, but we need a more convincing story to work on this project. We need to add/update them.

Here is what I want you to do:

1. Find all the POIs relevant to the public transportation and parking services
   1. Provide me the specific ways in words
      1. The list of keywords in top\_category and sub\_category
      2. The list of NAICS code
      3. Are there POIs could not be filtered out using the methods in (i) and (ii)
         * If so, how could you find them?
   2. Provide me the result
      1. Summary statistics of POIs
      2. Find the number of raw visitors for the corresponding POIs
         * Summary statistics of visitors
         * Add time (month) dimension if necessary

Send me the result by 11th. If you want to talk with me, please use:

<https://calendly.com/ymoon-econ/30min_moon>

Motivation: Good

Answer Strategy: Fair

Writing quality: Poor (please review Part II and III)