Group 2

Chenyi Wei, Li Zhou, Jenny Wang, Sicheng Yun

After the group discussion, we decided to insist on our previous topic, which focuses on Boston's fast-food industry. More specifically, our research question is: from the time’s perspective, does fast food restaurants' pattern of in-store traffic keep the same as what people generally perceived before the quarantine policy(which we can use April 18 2022 as a segment line, after April 18 2022, people are no longer forced to wear mask) published by the department of Public Health of Mass.gov). Under the mask policy, people may consume more fast food because its convenience, no need to eat in. To achieve this goal, we need to first differentiate fast-food restaurants from various industries shown in the safe graph data set, and then manipulate the dataset so that the statistics appear in an analysis-friendly way.

To achieve this target, we need to first filter fast-food restaurants out of the entire industry's information in Boston. In this case, we discussed several ways to differentiate fast-food restaurants from various industries.

One way we think about this is to sort the data by using the information included in the naics\_code column. As the naics\_code column includes specific codes that refer to the industry’s type, observing such a code is a convenient way to filter fast-food restaurants from other industries. We may choose to join the place and pattern dataset using the ‘street\_address’ key. The advantage of such a way is convenience compared to other methods we will discuss in the following paragraphs since it does not require outside resources except the searching of the code that represents fast-food restaurants. However, when searching for such code, we find that we can only find the group which presents Food Services and Drinking Places, it is hard to say which code will specifically represent the fast-food industry since we don’t know the exact scope of each subcode.

The other method we figured out is to filter the fast-food restaurants by hand. As our targeted data is in a comparably small range, which is the Boston area, we may first use some outside resources like yelp which lists all the fast-food restaurants in the Boston area, and then filter these restaurants by hand. The positive side of such a manual manipulation method is the correctness of the filtered data since human beings may identify the different names that refer to the same restaurant, while the negative side is the complexity of the filtering process. In this case, when observing the outside resources, we find a dataset in Kaggle that lists the fast-food restaurants' information inside the US, which led to the creation of the following method.

The method is the combination of both the outside resources and the Safegraph dataset. In this situation, if we can find an outside resource that contains the list of fast-food restaurants in the Boston area, we may use the inner join function to filter the fast-food restaurants in the original dataset. After the online research process, we found that the following link represents the information that fits such a requirement.

<https://www.kaggle.com/datasets/datafiniti/fast-food-restaurants>

By observing the variables presented in this dataset, we found that although the file includes fast-food restaurants in a country range, detailed information like the city and state names are also included, allowing us to filter the fast-food restaurants in the Boston area. After the filtering step for the outside resource, we may do some data cleaning to both data sets to keep the name column consistent before merging the two datasets. The advantage of such a way is that the programming-based process makes the whole filtering process easier to finish. However, there are also some disadvantages, which will be discussed in the following paragraphs.

One disadvantage of such a filtering way is the potential time bias. For instance, there may exist a time gap between when the two resources were collected. As the fast-food industry is an industry that experiences comparably fast replacement and the one that was seriously affected by the pandemic, the existence of the time gap may cause problems with the accuracy of the final data. In our case, since the list of US fast-food restaurants was created in 2018, the four-year time gap results in the fact that the new dataset may fail to record fast-food restaurants in Boston that enter the market after 2018. And if the missing restaurants present a significant effect on the variation of the number of visitors on a different date, there may exist a time-related sampling error in our analysis. Good point.

Besides, using such a method also leads to a concern about data consistency. Since we choose the inner joint as the way to combine the two datasets, the consistency of the baseline itself becomes a concerning issue. Ideally, all the targeted restaurants have the same name which makes the merging process easier to conduct. However, as R is a capital-sensitive code and the merging function only works with the data that have the same value, a slight difference, like a variation between the capital and smaller letter and a different name referring to the same restaurant may lead to missing data problems. In this case, we may need to slightly modify the two data sets to improve consistency. For example, we may change all the characters in the name columns to small letters and may do some research to find out all the potential names that refer to the same restaurant.

After the filtering process, the next step is to select columns and manipulate the sorted dataset to make it analysis-friendly. Considering our focused topic, we are interested in the number of visits to fast food restaurants. We firstly select the relative columns(location\_name\_data, data\_range\_start, data\_range\_end, row\_visit\_counts, visits\_by\_day).

The columns "date\_range\_start" and "date\_range\_end" represent information about the time of the month, and the time span is exactly one month (e.g. "date\_range range\_start" is 2022-07-01 04:00:00, "date\_range\_end" is 2022-08-01 04:00:00, which represents the data of July). Therefore, we can replace these two columns with one column named "month" when processing. We can use the mutate function in the tidyverse package.

In order to avoid version conflicts mentioned in Thursday’s conference, we decided to choose the date column instead of the one that included weekdays since the latter one is in Jason's file and cannot be easily manipulated inside R. In this case, we may consider separating daily in-store traffic by using the parse function and set comma as the determinant factor. Finally, we have the number of visits per day as a separate category to get the number of visits per month corresponding to each day (31 columns in total).

Good. The questions are not only well-defined, and they address significant matters in real life. The key is how to filter the fast-food restaurants. I want to see more strategies.

Here is what I want you to do:

1. Find all the POIs relevant to the fast-food restaurants
   1. Provide me the specific ways in words
      1. … we may change all the characters in the name columns to **small letters** and **may do some research** to find out all the potential names that refer to the same restaurant.
      2. How do you approach? Show me the VERY SPECIFIC approach
   2. Provide me the result
      1. Summary statistics of POIs for each strategy
      2. Find the number of raw visitors for the corresponding strategy
         * 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: Fair