Kevin Xiao, reflection paper on Group 5 project

1. Introduction

The industry Group 5 is interested in is the Snack and Nonalcoholic Beverage Bars within the restaurant industry. They focus on this sub-category by filtering the “Restaurants and other eating places” top category, in order to look at a specific group of stores/POIs for their research problem: how the weak economy affects large coffee house chains in Boston. The group defines the large coffee house brands as brands which have more than ten stores in Boston. In their presentation slides, the group also list the question they plan to answer is that how consumer behaviors changed during and post the pandemic.

It makes sense for the group to use SafeGraph data for their analysis, as SafeGraph data describes the traffic patterns over a period of time in this calendar year with visit counts to certain point of interests. It would be beneficial for the group to explore the visits to focused locations of coffee house stores over the time, in which way the group will be able to pick up trends and answer their research question. During the presentation the group also elaborated that the audiences of their analysis would be potential coffee shop owners and those who will consider entering the business, thus, it’s expected all these questions are addressed later in the project.

1. Representative Data

Their data cleaning process, or data wrangling as the group calls it, is relatively brief and simple. The group joins the two tables using “playcekey” as joined key and does a left join to include all the rows from the Patterns data. Then the team proceeds to filter the data by NAICS code to only looking at the Snack and Nonalcoholic Beverage Bars POIs, and finally filtering by individual brand names to include the three brands they are going to do analysis on. They will be using a relatively niche set of data containing only three brands which have 152 locations in total. This marks the end of the data processing of the group.

It's interesting to lean about the steps the group takes to test if the data set is biased. At first the group tries to test if the store counts among the brands are biased. Instead of using any statistical testing methods, the group provides the actual store location counts in Boston as the population statistics; then compares the observed number of stores in the sample directly to the number of stores they acquired. It’s not mentioned in the presentation nor in the R script by the group about how and where they obtain the actual store location counts from.

The second part of the data bias test the group performs is about location distribution. Although there are graphs provided and outliers noted and eliminated, the audience may find the purpose of this location distribution bias test confusing. It’s not explained in the presentation the statistical method behind performing the testing based on the graphs. The scatter plot shown resembles the locations of coffee stores within the boundary of City of Boston, and it’s not clear to the author why this will entail unbiased from the data sample.

Lastly, the group has a third part of data sample bias test to be the raw visit count distribution. In this step the group uses the bootstrap strategy to eliminate the possible bias in the sample by bootstrapping 50,000 times to result in a medium number. By repeating sampling, the group can get the numbers of raw visit counts without the influence of a heavily left-screwed data distribution.

1. Code Review

One of the potential challenges for the group is the limitation in the time frame the data set includes. In order to analyze the pattern in a special period, longer time frame or comparable year-to-year data might be necessary. The group did not address this limitation in the presentation. The other challenges the group may face is the relatively large amount of missing data. The group did a great job elaborating on this topic in the presentation. Facing with the large number of rows containing null values, the group stated and acknowledged the limitation this brings to the project.

In general, the code script from the group is easy to understand. However, the structure of the script has room to improve, as well as the comments and explanations for the steps taken and commands used. The reader of the scripts may find some lines of codes exists without notes and comments. It’s without doubt the scripts is clearly written and well formatted, though.