Group 3

BUS 211A-3

November 4, 2022

Initial Report

American coffee culture has a long history. There are more than 180 million coffee drinkers in the United States, with an average of 4.5 kilograms of coffee per person per year; it is not only the world's largest coffee consumer but also a "third-wave coffee action" and "specialty coffee" pioneer.

In Boston, the demand for coffee is particularly huge because it is an academic and financial hub where a lot of students, faculty, and professionals go to coffee shops regularly to recharge themselves.

There are a lot of people consuming coffee to stay awake and be more productive at work and school. Additionally, many people go to coffee shops to have a quick meeting, meet friends, relax, and enjoy their free time over a book. Therefore, it is interesting to figure out if they would prefer franchises or independent coffee shops. At the same time, we also want to know what determines people's choice.

There are two databases, place, and pattern, which have detailed information about locations in the Boston area. By filtering out all the coffee shops in the Boston area, we can get a lot of information about coffee shops, including time, location, number of visitors, etc. Through the analysis of these data, we hope to be able to solve the problems raised above.

We need to consider the downsides of the strategies we are considering and find the optimal way to answer the question. For example, we could think about analyzing local coffee shops and franchises. Yet, we first need to know how we identify the stores and filter them. Furthermore, popular franchises will have way more visitors than independent stores. There will then be a gap between major coffee shops and local coffee stores. In this case, should we drop the stores that have very few visits or keep those columns? Moving forward, our task for this assignment is to establish the best way to select the coffee industry and the POI, then prepare the data.

Consequently, some potential challenge in pursuing this research is efficiently narrowing our dataset. We could experiment with adding and removing some keywords such as the word “coffee” and some franchise names like “Starbucks”. We could look for some industry insights, using random sampling to choose some coffee shops and look for common keywords to use in our search to clean up our dataset. Besides, another challenge is the missing values per column. We need to analyze if we should remove the columns with many empty cells such as brands, and columns with one single variable like city and regions. For columns with few empty cells, we need to consider if we should use the average, medium, or random sampling methods to fill out the missing values per column. Another way to approach this challenge is to choose related values as the base, for example, for the missing values in column distance\_from\_home, we could look for the same postal codes that contain this data and rewrite the same value or an approximation in the empty cells. We could use this approach to classify independent coffee shops and franchises. However, this option may not be time efficient as the size of the data is large. We could use this approach using the random sampling method to increase time efficiency.

Additionally, when combining the two datasets “places'' and “patterns”, we need to find out how to separate the columns that have multiple values such as visit\_by\_days, popularity\_by\_hour, and popularity\_by\_day. We could look for some functions that separate the data by comma and assign new columns to the results. However, since the data presents information for one month, the number of columns will be too large. We could create columns with the monthly averages. We also need to analyze if there is a relationship between location and brand popularity, which could be measured by the number of visits per location. Through using random sampling, we could see if there is a trend between these two variables. We would also like to pick other columns such as the average of popularity\_by\_day to see if there is a trend.

To answer this question of whether the customers prefer franchises or independent coffee shops and what affected their decisions, we need to manipulate the data first.

The general steps would be combining the two datasets, filtering out all the coffee shops via naics\_code 722513(which includes coffee shops) and category\_tags that contain “Coffee”; grouping the data by independent coffee shops and franchises; comparing the average monthly visits per category and reaching the conclusion based on the results.

There would be multiple approaches to grouping the data. The first would be looking up all the franchises operating in Boston, filtering them in the location\_name column, and grouping them as franchises.

Second, since most franchises have the same names across all locations, we could filter out all the rows with the same names in location\_name. This could be achieved by using unique () and/or distinct () in base R.

According to SafeGraph documents, if this POI is an instance of a larger brand that we have explicitly identified, this column will contain that brand name. We can filter the local coffee shops out by filtering all the rows that have blanks in the brands column.

To obtain the average monthly visits per category, we could divide the total number of shops per category by the total number of raw\_visit\_counts in each category.

By following these approaches, we expect to effectively answer our research questions.

Great. One concern is the purpose of the study is too descriptive, so I recommend you add some Covid situations in your question (optional). The questions are not only well-defined, and they address significant matters in real life.

I want to see more strategies on how to filter the coffee shops.

Here is what I want you to do:

1. Find all the POIs relevant to coffee shops
   1. Provide me the specific ways in words
      1. Need to have at least FOUR DIFFERENT strategies
         * Give me a FULL list of keywords if you want to filter based on the name of POIs
   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: Poor

Writing quality: Good