Team 7

Lian Liu | Guangji Tang | George Nekwaya | Qinghao Yang

• What is your specific area of interest?

Overview:

Our team plans to explore the Safegraph data of consumer patterns of Restaurants and Other Eating Places in Boston, Massachusetts. More specifically, we decided to focus on Restaurants and Other Eating Places which offer Chinese Food as a whole or part of their serving.

Problem Questions/Objectives

Using the STAR Framework:

SITUATION

We plan to use our findings from the data to make informed data driven decisions with regards to 1) whether or not it is viable to open up a Chinese restaurant, 2) based on the consumer pattern data, find out if there are Chinese restaurant clusters and if the clusters influence consumer behavior .Then establish about three ideal locations in Boston that we can consider to open up the restaurant, 3) based on the demand data, we would consider the selection of the restaurant surface area. Do we select a size similar to average restaurant sizes or one which is smaller or larger to be able to serve our target market effectively. 4) Lastly, we wish to make use of the data from existing Restaurant and Other Eating Places to give us some operational insights, a) opening and closing hours, b) peak hours data for adequate staffing for weekly, monthly and yearly operations, c) how to advertising and attract more consumers based on the consumer pattern.

TASK

Aspects of the consumer behavior of visitors to the restaurants. Firstly, we want to get a general view of the dataset and examine it to interpret and understand the patterns and behaviors of the consumers to Restaurants and Other Eating Places in Boston.

Load the following packages from the library: readr, stringr, lubridate, dplyr, tidyverse, ggplot, RSQLite and DBI. Import the Safegraph data your\_data\_aug\_31\_2022\_0013am.csv and your\_data\_sep\_23\_2022\_0456pm.csv into R Studio. We will call the summary function and examine the 28,228 rows and 22 columns and the type of data they contain. Identify missing data/inconsistent data using the table function with the is.na function across all the columns, table function lets us see how many null entries are observed in contrast to the total number of rows. We see which information is redundant or irrelevant by making use of the unique function. Filter, sort, the dataset and have it in a usable format/table. In detail, in order to have a general view of the dataset, we first look through all the columns and pay attention to the contents in the columns, whether it is a column with numeric data or a column with string. We also will dig into it, find out some facts like if one column contains numeric data, which kind of numeric data is in it exactly, how we may use it or if it is irrelevant to our final goal. This would give us an overall idea of what to expect and what we can use in this dataset. Then we will find missing values of each column in the dataset. After this, we need to determine what we need to do with these missing values based on how they will affect our estimation and the content in this column.

We would try to answer the questions about the basic information of the consumers of Chinese restaurants, like which Chinese restaurants are more popular, when is the most popular period of visits, how many times are consumers returning to the same restaurants and how far the consumers travel for food.

Applying data manipulation techniques, we can use queries with the SQL function or just use the R language. To answer this question, we need to filter the data to get all the Chinese restaurants at first. We would join Places and Patterns into one dataset and use the NAICS code and names of the POIs to select all the Chinese restaurants since the NAICS code contains information about the POIs’ business category. Then to validate our results, we would use top\_category, sub\_category and category\_tags to check if all the Chinese restaurants are selected. After finding all the Chinese restaurants, we would explore the consumer patterns each month since the data is collected monthly. We could use columns of visit\_counts and visit\_counts\_by\_day to compare the popularity between restaurants and popularity of the same restaurants on different days. Columns like popularity\_by \_hour and popularity\_by\_day could be used to find when is the most popular period of the restaurant. We could also use visitor\_home\_cbgs, visitor\_home\_aggregation, and distance\_from\_home to analyze how far consumers are willing to travel to visit the targeted restaurants. All the steps above could give us a detailed view of consumer patterns of the Chinese restaurants in the Boston area, which can be used to decide the location to build a new Chinese restaurant.

Our second task is to explore the changes in consumer behavior from January to August 2022. We want to analyze the differences between the consumer patterns of different periods.

(How?) Based on the above mentioned step, we can now aggregate the findings according to our objectives to derive insights like daily, weekly, monthly or yearly trends in consumer behavior. Then we will make figures to visualize the information. We could compare the popular hours and the popular days of each restaurant and compute the differences between months to find if there are significant changes. Then based on the changes of consumer behavior, we could estimate the trends of consumers’ preferences of Chinese restaurants, which could be used as evidence of our recommendation of the opening of the next Chinese restaurant. We may give suggestions of operating a Chinese restaurant as well since we will know the peak hour and day, and the sources of customers.

• Why is it more important than other areas?

Eating out is one of the most common and popular activities, which makes it valuable to analyze the behavior of those who went to the restaurants and other eating places. By performing analysis on their cell phone tracking data, we can gain insights into people’s consumption habits when eating out.

Besides, the number of observations marked as restaurants is the highest. There are 12478 observations in the category of “Restaurants and Other Eating Places”, while the second largest category “Museums, Historical Sites, and Similar Institutions” has only 4152 observations. Such big data size allows us to perform an in-depth analysis.

• Why is the cell phone tracking data fit for your interest?

The cell phone tracking data provides information on the POI’s industry and location. It also provides tracks of consumer behavior, including their destinations and time of visits, which allows us to find the shared characteristics of the consumers and the connections between different POIs. For example, we can find out when people go to Chinese restaurants most (holidays?) and if people are far from home to eat.

ACTION

When the TASK is approved, we dive into ACTION where we put theory into practice and CODE!

RESULT

Now that we have run our codes/queries and made sense of the datasets. How are we going to share and present our findings?

Data visualization

After we calculate the popularity (the customer volume) of each Chinese restaurant, we would show them on a map. The map will have every Chinese restaurant shown as a circle with the relative popularity as its size. The map will show the audience the popularity of Chinese restaurants in different regions in the Boston area. We will use Tableau and ggplot2 in R to plot the map.

We will also use the line chart to display the trends of consumer preferences. We calculate the differences of consumer behavior from January to August 2022 above and we will show the popularity of the top 5 popular restaurants and bottom 5 restaurants in a line graph. We will use ggplot2 in R to plot the graph. We can also show a chart with all the restaurants’ popularity on it with color as different areas in Boston.The popularity of restaurants on different days within a week, popularity by hours could be shown in the same way.

Consumer behavior in the restaurants could be shown in a bar plot. We could use the bar to show customers’ dwell time inside the restaurants. The height of the bar is the total count of visits of the restaurant each month and the bar will be divided into several parts to show the proportion of different dwell time. We could show the top 5 popular restaurants in this way so that we could learn from the most popular restaurants when opening a new restaurant.

Feedback from Yeabin,

Hi Group 7

This is an absolutely interesting idea. Let’s get started now. One small problem is how to find Chinese stores in the SafeGraph. We can make this.

Let’s do the following:

(The Steps here support the explanation in the above write up)

1. How do you classify the restaurants in SafeGraph data?

a. Use NAICS code

- naics\_code: 722511, sub\_category: Full-Service Restaurants; Total entries: 1,537

- naics\_code: 722513, sub\_category: Limited-Service Restaurants; Total Entries: 410

Filter by naics\_code 722511-722513 (722512 does not exist or is null)

b. Once you figured this out, we can try a different search algorithm for Chinese restaurants

Total entries: 28,228, Top\_category: 2,427, sub\_category: 1,947, category\_tags: Chinese Food : 136 entries (full & limited)

2. How many of them (restaurants in general)?

Total entries: naics\_code: (722511 - 722513) = 28,228 (Total Restaurants in Boston)

Full- Service Restaurants : 1,537 for naics\_code: 722511; (of which sub\_category: Chinese Food = 133)

Limited-Service Restaurants: 410 for naics\_code: 722513; (of which sub\_category: Chinese Food = 3)

a. Figure out the largest and smallest

Try to see how many restaurants there are in the Boston area first. Talk with your teammates. And then we can set the time to talk about it. Let's speed up the process.

How to validate the results.

Filter by, “Food”, “Chinese”, “Chinese Food”, “Dragon”,”Hot Pot”, “Shanghai”, “Hong Kong”, “Gourmet”, “Speciality” ,”Szechuan”,”Chengdu” ,”Cha Siu”,“Dim Sum”etc. to explore the results for consideration.

Other Chinese restaurants under different names : Some tags could generalize by putting Chinese food under Asian Food.

In order to check if there are restaurants not tagged correctly, we ran a filter of tags with “Asian” and without “Chinese”. In the 335 filtered observations, we didn’t find any restaurant with a recognizable Chinese name or other Chinese element.

Catering has some tags which fall under the NAICS code we are interested in, but not all of them.

The Green Dragon Tavern, Chinese Dragon, Dragon Bowl, Blue Dragon are other names, but only a few have corresponding tags for Chinese Food.

Living Root Dragon Boat (NAICS CODE: 713110) - This is not relevant to our investigation.

Sample Presentation Guide:

Team and Project Introduction

In our investigation using the Safegraph datasets, we aim to explore the points of interest and the patterns of data concerned with Chinese Restaurants in Boston, MA.

Our aim is to use the findings of the investigation to make informed decisions on where to open up a new Chinese restaurant and how best to advertise it using print media at designated locations where consumers of Chinese food would normally visit.

We aggregated the visit counts and limited the results to Chinese restaurants and used ggplot to get a visual outline of the findings. We were able to notice a High concentration of around 2,959 visits to Chinese restaurants within the ChinaTown area. Medium concentrations of 2,000 visits were observed scattered in other locations around Boston. Low concentrations of 1,000 visits were also observed in this manner.

We wanted to see which restaurants and locations were popular and profitable.

We made the following assumptions:

1. For the Chinese restaurants with a monthly visit equal to or below 30, we excluded them from our investigation as they could be either not profitable, not popular or no longer operational.
2. The Average dwelling time differs according to the nature of the Chinese restaurant, as some restaurants allow for Pick ups on orders, some do deliveries, this is very common nowadays with service providers like Uber Eats, Doordash etc. This contributes to a lower average dwell time, which might give a false impression of the restaurant being less popular than it actually is.
3. A really high number of visits could be misleading as some of the restaurants could be placed in a mall or complex and traffic and/or dwell time into the nearby shops could be counted as visits to the Chinese restaurants we are investigating. - Using the boxplot we approached this by considering the restaurants in the ¾ percentile to be popular by visit counts.
4. The distance from home to the restaurant might not directly be related to the popularity of the restaurant as consumers might have cars and they dont mind the drive and would also visit other places to maximize their commute. The transportation system could also be affordable to allow ease of long distance commutes for non drivers.

Conclusion:

We have decided to open a new restaurant (for example) outside ChinaTown, towards the North/East/West/South of Boston, because …..

The restaurant will be a Limited Service Restaurant/Full Service Restaurant ….

We have calculated an advertisement index and it shows that most consumers of Chinese Foods, would visit Cosco etc. after/before and we will use this an indicator for selecting and prioritizing the ads of our new restaurant

Popularity based on months and visits:

April Peak- increase staff working hours : Spring Break we assume that results in more students being free to go into Boston to visit various POIs.

June/July - less hours to workers : Summer Break results in lower visits as we assume that more people take advantage of this to travel outside of Boston.

*Dataset not up to date or safegraph data has some errors, as this downward trend was observed across all restaurants at the same periods. So we could not make assumptions based on Chinese restaurants*

Distance from home is not a reliable measure toward gauging the popularity of the Chinese Restaurants.

The average distance to Chinese restaurants reflects the proportion of travelers who visited the restaurants. As the figure demonstrates, there is no clear difference between Chinatown group and non-Chinatown group. And for both groups, there is a valley in February and a peak in July.

Our hypothesis is:

As visitors from other states as well as incoming students to Boston Universities or Summer Schools or Internships across the US can inflate the average distance traveled.

As a validation of our hypothesis, we plot the average distance to all POIs and figure out there is also a valley in February and a peak in July, which proves that more travelers are coming to Boston in July.

So that we suggest restaurants can do some promotion which targets travelers in July.

Advertising strategies based on data analysis.

1. To advertise at the related brands with same month visits as Chinese restaurants.
   1. We used the related\_brand column to find the brands that Chinese restaurants customers went to the most.
   2. Restaurant owners can negotiate with brands to do advertising. But they also need to consider the cost since the cost of advertising is not included in our analysis.
2. To advertise in front of Competitor Chinese Restaurants.
   1. For example, owners can use our map figure to find popular restaurants near their location to advertise.