**The Battle of Subway Stops**

Peilin Xin

July 23, 2020

**Introduction**

1. Background

Not every available space is right to open a restaurant. Choosing a location for a restaurant is just one of a few keys to profitability. There is a famous saying in real estate that also works for opening a restaurant: “You make your money when you buy.” Also, restaurant expert, Lorri Mealy, explains that one reason restaurants fail is a poor choice in location. She says a bad location is one of the biggest, if not THE biggest, reason restaurants fail.

1. Problem

We will use the power of data science to generate a few viable locations to open a restaurant. The decision will be based on the locations accessibility, target market, population base and competition.

1. Interests

There is so much to plan in advance to be prepared for opening a restaurant. Choosing a proper location can be determined for the future success of the restaurant. This project aims to help those who are interested in opening an American restaurant in Boston and confused about where to open it.

**Data acquisition**

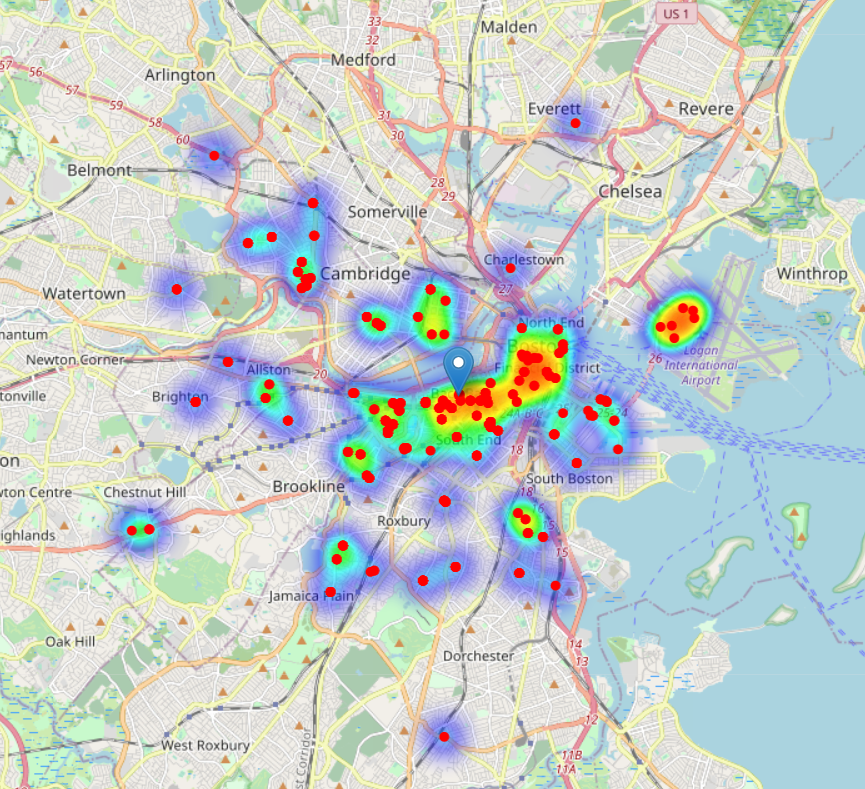
1. Data sources

The data sources come in two different parts. To learn the accessibility of a location, I need to have subway and bus stops information which I can download from the MBTA (Massachusetts Bay Transportation Authority)’s official website. Then I will use Foursquare API to extract information of the restaurants in my target market around the transportation stops to learn the density of the restaurants. Since Foursquare only supports 950 calls per day for a free account, I need several days to get all the information I need for this project.

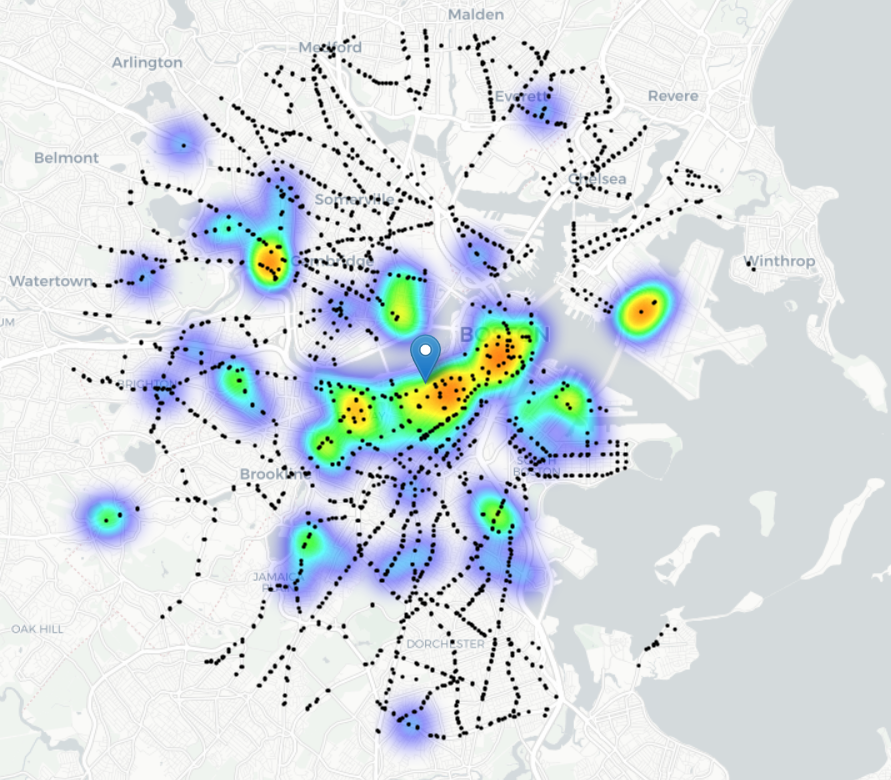
1. Data Cleaning

The most recent MBTA’s subway stops data has given the stops id, name, coordinates and a series of information and here are the data columns that I think could be useful for further analysis: stop\_name, stop\_lat, stop\_lon, zone\_id, location\_type, wheelchair\_boarding, municipality, on\_street, at\_street. After Looking at the data shape and number of null values in each column, I find out that there are a few columns missing a lot of values and there are no other data sources that can compensate. Filling the empties with calculated values are not acceptable in this project because even a slight discrepancy with fact would be misleading. So, if there are missing data in this location, then this location will not be considered in the research process.

**Methodology**

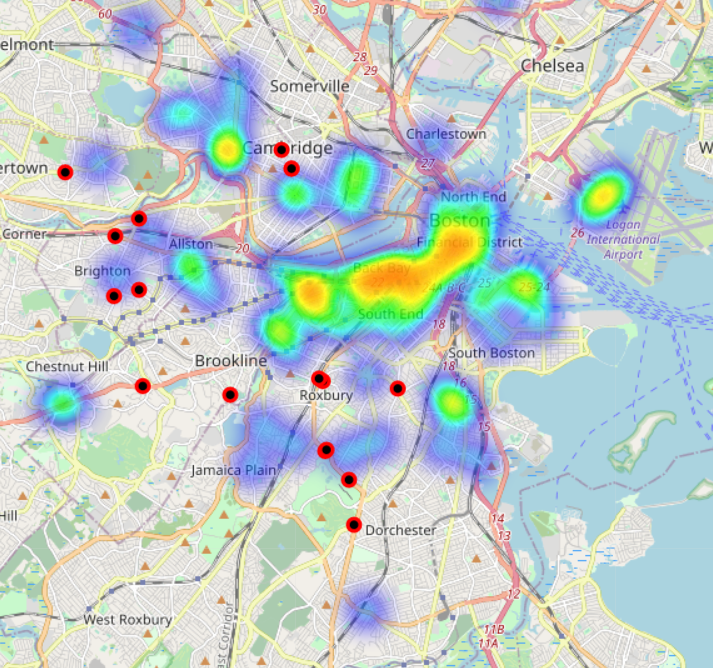
Since I have the exact coordinates of each commuter stop, I used the Foursquare API to search for the restaurants that were located within 350 meters of each stop and extract the information such as venue’s id, name, categories, coordinates, and distance from the stop. Since Foursquare only supports 950 regular calls for a free account per day and I have more than 2000 rows in research, it took me 3 days to gather all the information I need. After we got the restaurant data, I found out that only 1371 stops have restaurants within 350 meters from the stops. Also, the average number of restaurants in each stop is 2.53. The most crowded stop even has 61 restaurants in total. If we target the market to only American restaurants, we have 553 restaurants in the research range. Here is how the restaurants are distributed on the map.

We can see that the American restaurants are heavily distributed around the center of Boston. Besides that, the other restaurants are pretty much scattered around Boston. If we look at the comparison between the heatmap of the restaurants and the whole transportation system, we can see that there are still a lot of stops that have potential.



Though the transportation systems are well developed and cover most of the area in Boston, American restaurants are not covering where the transportation systems cover. That means there are more factors deciding whether to open a restaurant in certain locations such as the population density, the wealthiness of the population, and the major race ratio. For example, it’s usually not a good idea if you open a very traditional American restaurant in Korean town since the majority of the customers there are Asians. Also, the capacity of daily population movement in an area is a great indicator of whether to open a restaurant. However, it’s very difficult to get the accurate information of those I indicate, so I’m only going to focus on the restaurant density and accessibility of the location.

So, I found out the locations with no more than one restaurant nearby and the locations with no American restaurants within 340 meters. These conditions either satisfied the low restaurant density area or the accessibility of the locations. When I inner-joined 2 conditions, I got 17 locations that met both.



**Result**

After getting the potential locations that can be used to open a restaurant, I use Google API to geocode the actual addresses of these locations.

**Discussion**

Based on the result, we get 17 locations scattered around Boston. However, it’s just the very general look of the question. Only considering the accessibility and restaurant is not enough to solve the real world problem. We need more information such as the population density, the wealthiness of the population, the major race ratio, daily population movement, etc.

**Conclusion**

The purpose of this project is to identify viable locations for opening an American restaurant. After the effort I have done, we had a general view of how the problem can be solved. The location generated can help us learn the places that can be potentially viable for opening a restaurant. However, we need more information if we want to push the analysis further.