Project: Battle of Neighborhoods

Course: Applied Data Science Capstone

Best Restaurant Investment?

Parth Parikh

July 15, 2021

1. Introduction:

1.1 Background:

Restaurant investment has been quite a profitable business in most countries, even after Covid-19. This is especially true as several countries are having a successful vaccination program against the pandemic virus, which is leading to people visiting public places more than ever. This is good news for restaurant owners who were out of business last year during the pandemic. In countries like United States and Canada, the reopening has led to flooding of restaurants, thus opening a specialty restaurant seems a good investment currently. For this project, we have chosen the New York City from United States and City of Toronto from Canada to find out what restaurant business in them and where is most profitable.

1.2 Problem:

However, as easy it may sound, opening a restaurant business at this time is nothing like before. A lot of people have lost their financial stability either due to loss of a primary earner in family, or loss of job, financial losses in business, etc. and thus the general norms of what restaurant type is a good investment has changed significantly.

1.3 Interest:

This is where the use of data comes in. All the investors who are looking for a good restaurant investment are looking for ways to understand the market. In such circumstances, the use of data to find out which type of restaurants are majority of people preferring to go to and in what neighborhoods of the city, makes the investment decision easier and more accurate.

2. Data Acquisition and Processing:

2.1 Data Sources:

The neighborhoods and coordinate (latitude and longitude) data for the New York city can be found in this .json file that was provided by IBM previously in this course. The neighborhood data for Toronto was found in parts at two different places. The first part containing the postal codes, borough names, and respective neighborhood names was scraped from a Wikipedia page

found under this link. The second part of the data containing the latitude and longitude coordinates of each postal code was used from a previous project where IBM had provided the data here in this .csv format.

Next, the geopy library was used in order to find the coordinate data of a selected address/location when needed.

The Folium library was used to plot the neighborhoods onto a map. The restaurant related data such as their location, coordinates, trends, categories, etc. for each neighborhood in both New York City and Toronto were obtained via extensive use of Foursquare API.

2.2 Data Cleaning:

The raw json file for New York neighborhood data was converted into a pandas dataframe after which all the columns were dropped except 'Neighborhood Name', 'Latitude', and 'Longitude'. For the Toronto data, the Wikipedia scraped data was converted from html to a pandas dataframe. The second part from the csv file was also loaded as a pandas dataframe. Both sources were then merged into a single dataframe by postal codes, after which all columns were dropped except the ones same as New York data.

The Foursquare API provided the list of top-10 most visited restaurants for each neighborhood in both cities. This data was merged with the previous data of each city, thus giving a dataframe containing neighborhood name, coordinates of neighborhood, name of 10-most visited restaurants, coordinates of each restaurant in list, and the category of restaurant.

The data was then prepared for further analysis as needed.

3. Data Analysis:

3.1 Finding the optimum restaurant type:

In order to find the optimum restaurant type, the restaurant type that was consistently in top-10 list of most frequently visited restaurant type in each neighborhood has to be selected. For this, firstly, the data for each city was prepared by dropping the coordinate data and name of individual restaurants for each neighborhood. Using the method of "one hot encoding", the data was divided into a dataframe where each row represented a neighborhood and each column represented a unique restaurant type and the cells in between would denote whether the given category is in the top-10 of given neighborhood or not.

Next, the data was arranged as per the frequency of people visiting a given restaurant type of each neighborhood. The data would look something like this:

Neighborhood	1st Most Common Type of Restaurant	2nd Most Common Type of Restaurant	3rd Most Common Type of Restaurant	4th Most Common Type of Restaurant	5th Most Common Type of Restaurant	6th Most Common Type of Restaurant	7th Most Common Type of Restaurant	8th Most Common Type of Restaurant	9th Most Common Type of Restaurant	10th Most Common Type of Restaurant
Allerton	Pizza Place	Restaurant	Deli / Bodega	Chinese Restaurant	Fried Chicken Joint	Fast Food Restaurant	Breakfast Spot	Donut Shop	Food	Fish & Chips Shop
Annadale	American Restaurant	Pizza Place	Restaurant	Bakery	Diner	Sushi Restaurant	Food	Deli / Bodega	Eastern European Restaurant	Fish & Chips Shop
Arden Heights	Deli / Bodega	Pizza Place	Wings Joint	Food	Dosa Place	Dumpling Restaurant	Eastern European Restaurant	Empanada Restaurant	Ethiopian Restaurant	Falafel Restaurant
Arlington	American Restaurant	Deli / Bodega	Fast Food Restaurant	Caribbean Restaurant	Wings Joint	Food Court	Dumpling Restaurant	Eastern European Restaurant	Empanada Restaurant	Ethiopian Restaurant
Arrochar	Pizza Place	Italian Restaurant	Polish Restaurant	Mediterranean Restaurant	Restaurant	Deli / Bodega	Bagel Shop	Middle Eastern Restaurant	Dumpling Restaurant	Eastern European Restaurant

Then, the neighborhood column was dropped to prepare data for value counts; "series.value_count()" function was used to find how many neighborhoods was each type of restaurant was in top-10 of. Next, the data was sorted in descending order based on resultant value. The same process was repeated for both New York and Toronto datasets.

Since New York and Toronto has different number of total neighborhoods, the data for two cities could not directly be compared. Thus, in order to compare the data between two cities, the data was normalized by min-max technique.

The normalized values were added in the previously created dataframe that contained restaurant types and total neighborhoods where they were in top-10. Next, the new dataframe for both New York and Toronto were merged into a single new dataframe by restaurant types.

Next, normalized score of each restaurant type was added and the highest one was chosen as the optimum restaurant type to invest in since it was consistently in top-10 for maximum neighborhoods across both New York and Toronto.

3.2 Finding the optimum neighborhood:

Now that the optimum restaurant type was calculated, it was time to find the optimum neighborhood in each of the two cities to open a restaurant of optimum type.

Using the previously formed dataset from one hot encoding, the neighborhood/s would be sorted out where the optimum restaurant type is the most frequently visited. This neighborhood/s are our optimum one from both cities.

4. Results and Discussion:

4.1 Results:

From the thorough and exploratory data analysis of neighborhood and foursquare dataset for both New York City and Toronto, it was shown that Deli/Bodega was the restaurant type that was one of the most frequently visited in maximum neighborhoods of both New York City and Toronto. For such restaurant type, it was concluded from the statistical analysis that Arden Heights neighborhood in New York City and North Park, Maple Leaf Park, and Upwood Park neighborhoods in Toronto were the ideal ones to invest into.

4.2 Discussion:

A common question that can arise from reading above analysis is why so many statistics were performed when it would have been easier to just find out the most visited restaurant type over the whole city instead of going in each neighborhood. The answer to this is that when whole city is considered as one, the possibility of an outlier in each category affecting the results is very high. For example, there can be a famous Italian restaurant in New York City that is known for excellent food, however other Italian restaurants around the city may not be as good, but since the given, one is very good, Italian restaurant type is most frequently visited when looked at the New York statistics. But investing in an Italian restaurant by looking at this analysis would be wrong since the outlier has huge impact on the result.

However, when you divide the city in different parts, the possibility of any outlier affecting the overall result is minimum as a given restaurant type has to be frequently visited across maximum number of neighborhoods in ordered to be considered optimum for good investment rather than just one neighborhood.

Although the data analysis has created a strong foundation for our investment decision, there are several factors missing out from the analysis due to the scarcity of quality data on open-source platforms currently. These factors include competition, initial cost of setting up the restaurant, yearly maintenance cost, availability of labor, average cost per meal of surrounding restaurants in neighborhood to name a few. These factors are equally important in making a big investing decision as ours and when considered, they may change the result of best restaurant type and neighborhood to make the investment. Thus, further and more in-depth analysis pertaining to these factors would have to be conducted before coming out to a final, real-world, data-driven answer to our problem.

5. Conclusion:

As discussed in previous section, the analysis of different neighborhoods or New York and Toronto as well as different restaurant types has led us to an answer that Deli/Bodega type restaurants are good investment in Arden Heights neighborhood in New York City and North Park, Maple Leaf Park, and Upwood Park neighborhoods in Toronto. One thing to note however is that although the result above is data driven via exploratory data analysis, several other factors are to be considered before making this investment.