The Battle of Neighborhoods

Introduction

Manhattan is the most densely populated of New York City's 5 boroughs. It's mostly made up of Manhattan Island, bounded by the Hudson, East and Harlem rivers. Among the world's major commercial, financial and cultural centres, it's the heart of "the Big Apple."

As it is highly developed city so cost of doing business is also of the highest. Restaurants are one of the most frequently started small businesses, yet have one of the highest failure rates. Survivors need a powerful strategic advantage: a sound business plan and feasibility study prior to opening. A business plan precisely defines your business, identifies your goals, and serves as your firm's resume.

The basic components include a current and pro forma balance sheet, an income statement, and a cash flow analysis. It helps to allocate resources properly, handle unforeseen complications, and make good business decisions. Because it provides specific and organized information about your company and how you will repay borrowed money, a good business plan is a crucial part of any loan application. Additionally, it informs personnel, suppliers, and others about your operations and goals.

Of course, as with any business decision, opening a new dessert restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the dessert restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the Manhattan borough, New York to open a dessert restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Manhattan, New York, if a property developer is looking to open a new dessert restaurant, where would you recommend that they open it? This project is timely as the city is currently suffering from oversupply of restaurants.

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Manhattan, New York. This defines the scope of this project which is confined to Manhattan, the most densely populated of New York City's 5 boroughs. Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

- Venue data, particularly data related to dessert restaurants. We will use this data to perform clustering on the neighbourhoods.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Battery Park City	47	47	47	47	47	47
Carnegie Hill	48	48	48	48	48	48
Central Harlem	13	13	13	13	13	13
Chelsea	50	50	50	50	50	50
Chinatown	50	50	50	50	50	50
Civic Center	49	49	49	49	49	49
Clinton	48	48	48	48	48	48
East Harlem	14	14	14	14	14	14
East Village	50	50	50	50	50	50
Financial District	47	47	47	47	47	47
Flatiron	50	50	50	50	50	50

Link to the data set is:
(https://geo.nyu.edu/catalog/nyu_2451_34572) contains a list of neighbourhoods in the 5 boroughs , with a total of 306 neighbourhoods.

- Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the dessert restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from working with API (Foursquare), data cleaning, data wrangling, to machine learning (K means clustering) and map visualization (Folium).

Methodology

New York city has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. Fortunately, the link to the dataset is available (https://geo.nyu.edu/catalog/nyu_2451_34572). We will do web scraping and clean the dataset to extract the list of neighbourhood's data. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in New York city. Then, we slice the original DataFrame and create a new DataFrame of the Manhattan data. We need to get the geographical coordinates of Manhattan borough in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the Manhattan. Next, we will use Foursquare API to get the dessert restaurant venues that are within a radius of 1000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Dessert Restaurants" data, we will filter the "Dessert Restaurant" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based

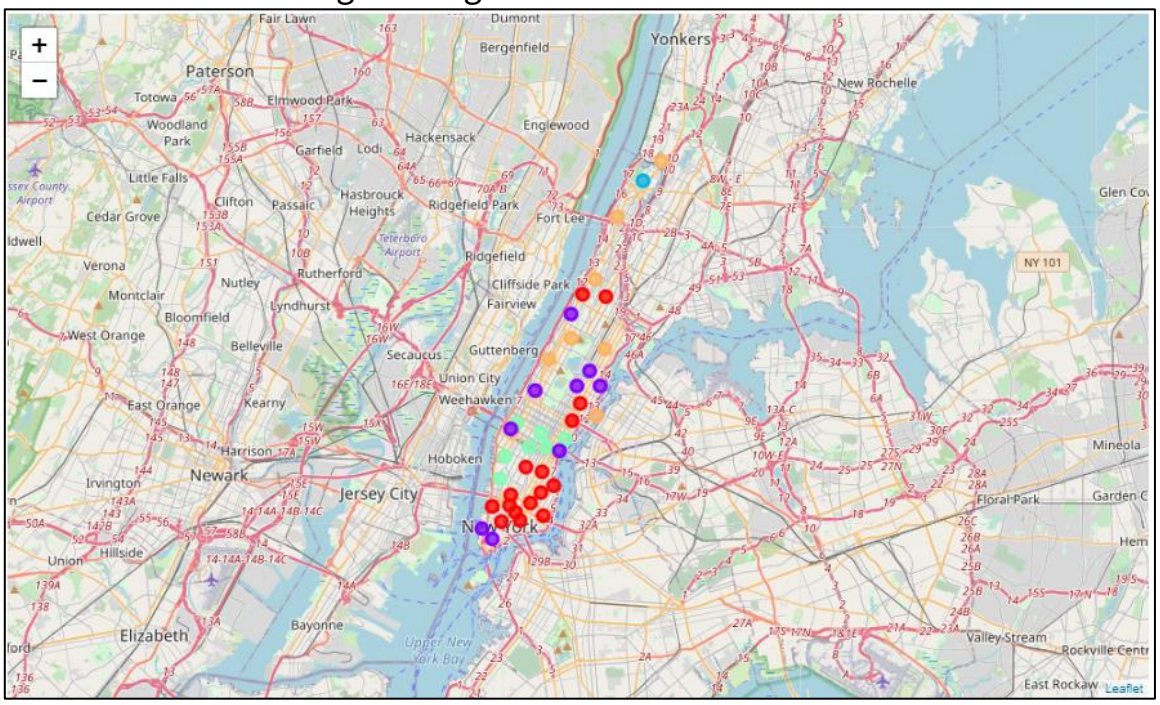
on their frequency of occurrence for “Dessert restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of dessert restaurant while which neighbourhoods have fewer number of dessert restaurants. Based on the occurrence of dessert restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new dessert restaurant.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence for “Dessert Restaurants”:

- Cluster 0: Neighbourhoods with high concentration of dessert restaurants.
- Cluster 1: Neighbourhoods with moderate number of dessert restaurants.
- Cluster 2: Neighbourhoods with very low number to no existence of dessert restaurants.
- Cluster 3: Neighbourhoods with moderate number of dessert restaurants.
- Cluster 4: Neighbourhoods with moderate number of dessert restaurants.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in blue colour, cluster 3 in mint green colour and cluster 4 in light orange colour.



Discussions

As observations noted from the map in the Results section, most of the dessert restaurants are concentrated in the Noho neighborhood, with the highest number in cluster 0 and moderate number in cluster 1, cluster 3 and cluster 4. On the other hand, cluster 2 has very low number to no dessert restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new dessert restaurant as there is very little to no competition from existing dessert restaurants. Meanwhile, dessert restaurant in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of dessert restaurants. From another perspective, the results also show that the oversupply of dessert restaurants mostly happened in the Noho neighborhood, with the Inwood area still have very few dessert restaurants. Therefore, this project recommends property developers to capitalize on these findings to open new dessert restaurants in neighbourhoods in cluster 2 with little to no competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations the best locations to open a new dessert restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new dessert restaurant.