Coursera Capstone

IBM Applied Data Science Capstone

Opening A New Restaurant In New York, US

By: Arnav Singh

June 2020

Introduction

A restaurant is a place where people can have a meal by paying for it. For many people it is a great way to relax and enjoy themselves during weekends, holidays, or any special occasion. They can visit a restaurant of their choice and order whatever they want to have. People can also order and get delivered their order at their home or office if the restaurant allows the delivery services. The quality of food, the ambience of the restaurant attracts customers to itself. But, one property on which success of a restaurant heavily depends on is its location. Location of a restaurant plays a vital role as it decides how accessible the restaurant is to its customer and how often an unknown potential customer can come across the restaurant. Every restaurant can deliver food to a limited area which are located nearby to it, but if there is no office or high population areas with possible customers nearby then the delivery services offered will be of no advantage. Or if it is located near some industrial waste or if there is no proper tidiness in its neighbourhood, then also people will not like to come and eat at that place. Hence, the location of a restaurant should be nearby some residential areas, offices, parks, shopping-malls etc. and its surroundings should be clean and tidy as well so that people get attracted towards it. Selecting the correct location for opening a restaurant requires serious consideration and is more complicated than it seems to. Selection of location can determine whether a restaurant will be successful or not.

Business Problem

The objective of this project is to analyse and provide a group of favourable locations to open a new restaurant in New York, US. Using data science methodologies and machine learning algorithms, this project aims to provide solution to the problem: *If an investor wants to open a new restaurant in New York, US where would you recommend that they open it?*

Target Audience

This project is useful for those investors who are looking open or invest in a new restaurant in New York, US. Every year, 55 million tourists pour into New York attracted by its vast options for entertainment, shopping and dining. With 25,000 eating places, restaurants are a vital part of the economy. But in a city

that is more competitive than ever, this project will benefit investors to understand the recent trends in the food industry and assist them to find the correct place to open their new restaurant to maximize their profits.

Data

For finding the solution, following data is required:

- List of neighbourhoods in New York. This defines the scope of this project which is confined to New York.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and retrieve the venue data.
- Venue data, particularly related to restaurants. We will use this data to perform clustering on the neighbourhoods.

Source of data and methodologies

The Spatial Data Repository NYU (https://geo.nyu.edu/catalog/nyu_2451_34572) contains a list of neighbourhood in New York, with a total of 306 neighbourhoods. We will be using Pandas and Requests data frame of Python to extract the data. Then we will get geographical coordinates using Python Geocoder package. After that we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database with 105+ million places and is used by over 125,000 developers. The API will provide many categories of the venue data, we are particularly interested in the restaurant category in order to help us to solve the business problem. After cleaning and wrangling data generated by the API we will use machine learning algorithm K-means clustering and visualize the results using Folium package of Python.

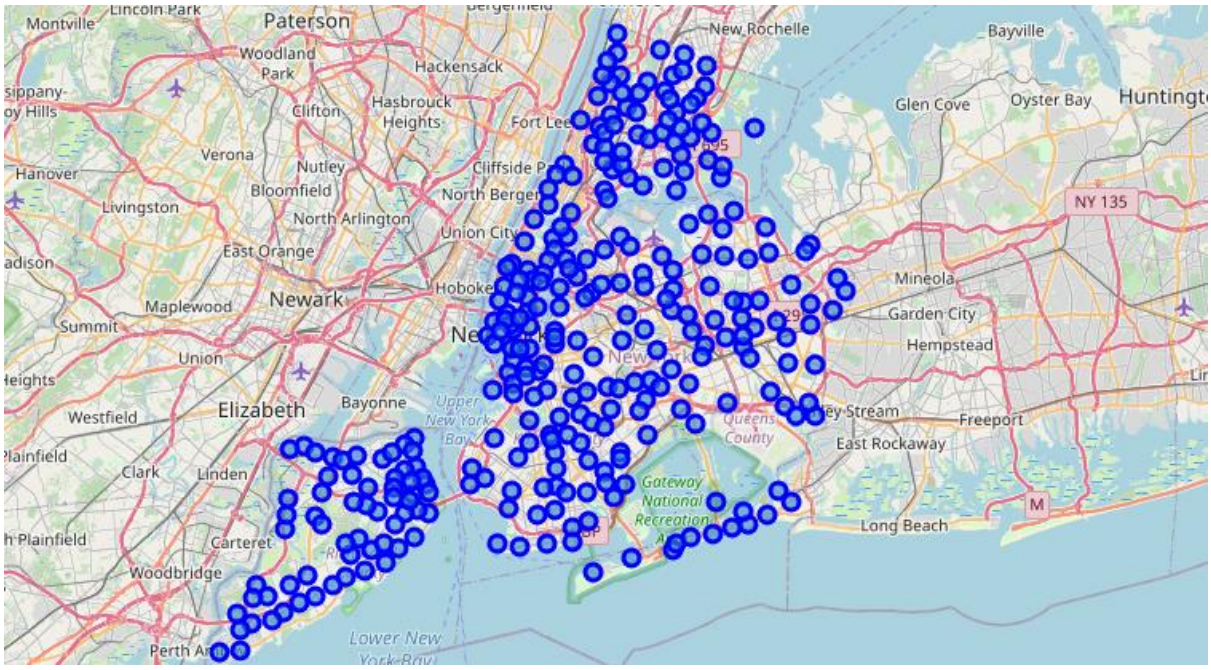
Methodology

We will start our project by retrieving the neighbourhoods' data for New York from spatial data repository NYU (https://geo.nyu.edu/catalog/nyu_2451_34572) . Then we will do web scraping using Python's requests and json packages to extract the required data from the response from the NYU repository. However, we will be needing geographical coordinates of the neighbourhoods for our analysis. We will use

geocoder package for getting coordinates. The finalised data with coordinates will look like the following:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

After gathering the data, we will populate the data into pandas dataframe and the visualize the neighbourhoods in a map using the folium package. This allows us to perform sanity check to make sure that the geographical data returned by geocoder is relevant. The map developed will represented as:



Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register for a Foursquare developer account in order to obtain the Foursquare ID and secret key. We make API calls to Foursquare by passing the geographical coordinates of the neighbourhoods. It will return venue data in JSON format. Then we extract the venue data from the response and load it into the dataframe at corresponding rows and columns. The resulting dataframe will look like:

	Borough	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Bronx	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

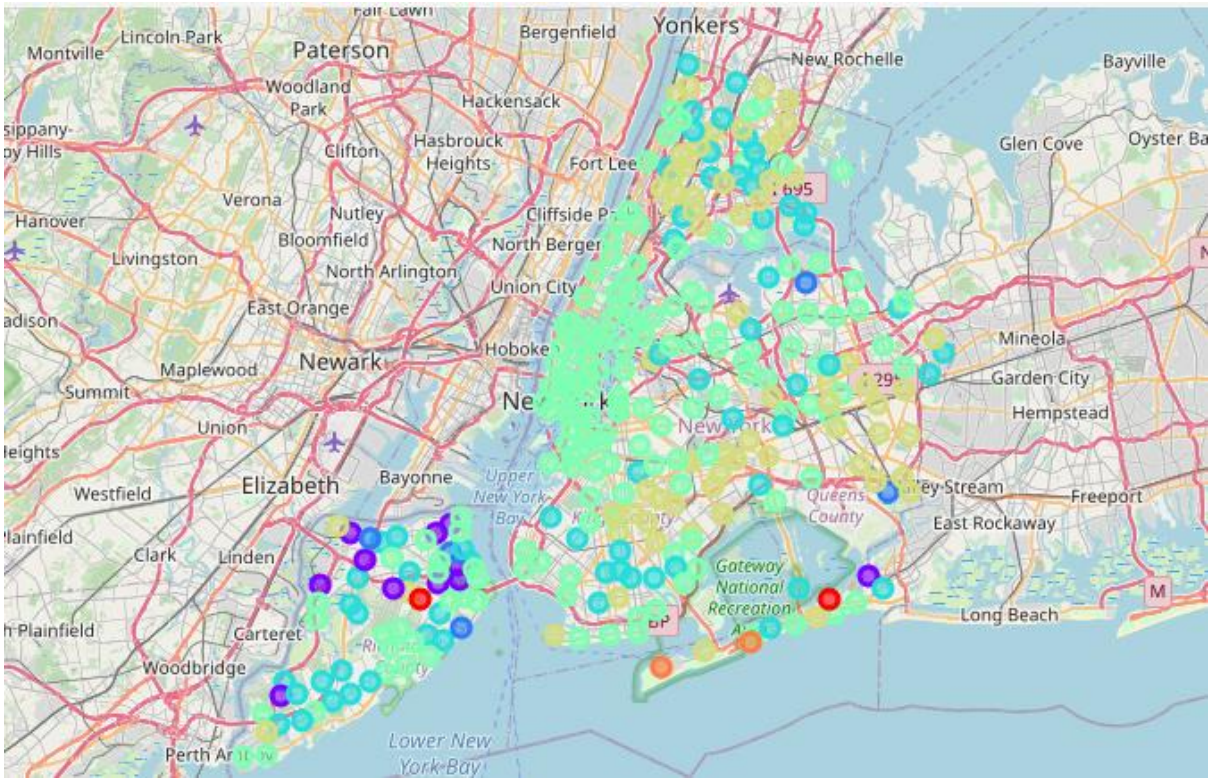
Then we convert the dataframe into a dummy data table with category section being the key. We take category as the key for generation of dummy table to get knowledge about the neighbourhoods and their elements. To group similar neighbourhoods together we perform k-means clustering. K-mean clustering algorithm identifies k number of centroids and the allocates every data point to the nearest cluster while regularly updating centroid according to new data points added to a cluster. It is very simple and highly suited machine learning algorithm that suited this project. We will cluster neighbourhoods into 7 clusters. The result will allow us to specify which neighbourhoods are similar to each other. The result of k-mean clustering will be represented as :

	Borough	Neighborhood	Latitude	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Allerton	40.865788	-73.859319	3	Pizza Place	Deli / Bodega	Supermarket	Intersection	Fast Food Restaurant	Grocery Store	Gas Station	Breakfast Spot	Bus Station	Spanish Restaurant
1	Bronx	Baychester	40.866858	-73.835798	5	Donut Shop	Mexican Restaurant	Arcade	Pet Store	Bank	Pizza Place	Electronics Store	Discount Store	Fast Food Restaurant	Sandwich Place
2	Bronx	Bedford Park	40.870185	-73.885512	5	Chinese Restaurant	Diner	Mexican Restaurant	Supermarket	Pizza Place	Sandwich Place	Pharmacy	Fried Chicken Joint	Deli / Bodega	Smoke Shop
3	Bronx	Belmont	40.857277	-73.888452	3	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Dessert Shop	Bank	Donut Shop	Fried Chicken Joint	Coffee Shop	Sandwich Place
4	Bronx	Bronxdale	40.852723	-73.861726	4	Chinese Restaurant	Supermarket	Park	Eastern European Restaurant	Pizza Place	Mexican Restaurant	Breakfast Spot	Bank	Spanish Restaurant	Gym

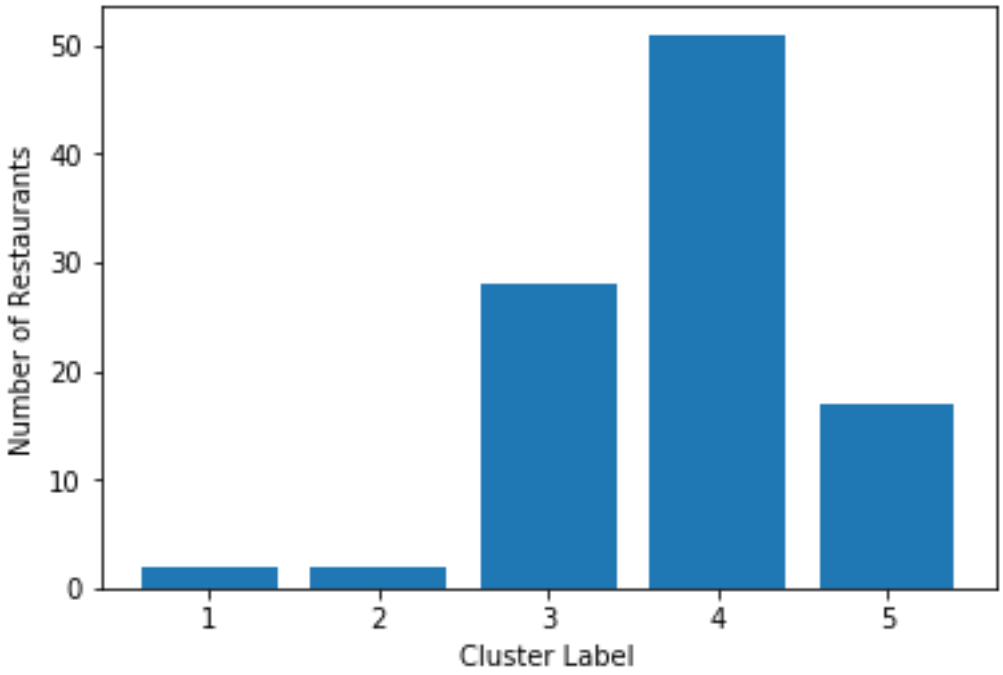
Using this dataframe we can analyse which neighbourhoods are more favourable for opening a new restaurant. We can also represent different clusters on map using folium:

Colours for representing different clusters -

- - Cluster 1
- - Cluster 2
- - Cluster 3
- - Cluster 4
- - Cluster 5
- - Cluster 6
- - Cluster 7



Now we will analyse different clusters with help of a bar chart to find how many restaurants they cover -



This will help us to identify which neighbourhood is suitable for opening a new restaurant.

Results

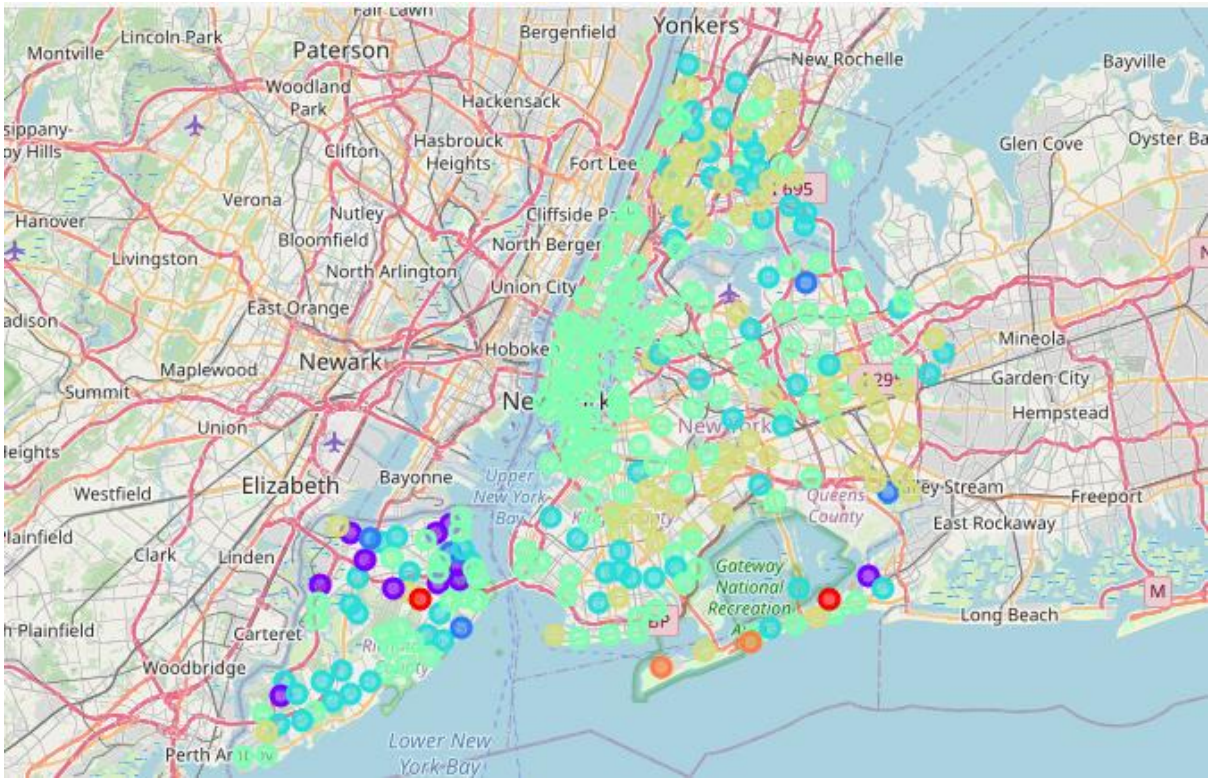
The results from the k-means clustering and quantitative analysis of different clusters show that we can categorize neighbourhoods into 7 clusters having different number of restaurants as following:

1. Cluster 1- Neighbourhoods with low number of restaurants and high number of unfavourable factors
2. Cluster 2- Neighbourhoods with low number of restaurants and moderate number of unfavourable factors
3. Cluster 3- Neighbourhoods with moderate number restaurants and high number of favourable factors
4. Cluster 4- Neighbourhoods with high number of restaurants and high number of favourable factors
5. Cluster 5- Neighbourhoods with moderate number restaurants and moderate number of favourable factors
6. Cluster 6- Neighbourhoods with very low number of restaurants and high number of unfavourable factors
7. Cluster 7- Neighbourhoods with no restaurants nearby and high number of unfavourable factors

Clusters represented on map:

Colours for representing different clusters -

- - Cluster 1
- - Cluster 2
- - Cluster 3
- - Cluster 4
- - Cluster 5
- - Cluster 6
- - Cluster 7



Discussion

As observations noted from the map in the results section, highest number of restaurants are concentrated in the area covered by cluster 4. Likewise, cluster 3 and cluster 4 also have moderate number of restaurants. Cluster 1 and cluster 2 have low number of restaurants. Cluster 7 is not favourable for opening a new restaurant as there are many unfavourable elements present in it. On the other hand, cluster 6 has very low number of restaurants and it also contains some favourable places like beaches and parks but because of presence of high number unfavourable factors it very unlikely to find a place at correct market value for investing money to open a new restaurant there. Cluster 4 provides a very high potential as it covers the main visiting spots of the city but there are already a high number of restaurants present there, opening a new one will be highly competitive in that area. Cluster 1 and cluster 2 has less competition for a restaurant but they cover a large portion of city having farms, factories and construction sites. Lastly cluster 3 and cluster 5 remain our only choices with large number of visiting spots to attract potential customers and moderate amount of restaurant present. Presence of other

restaurants will provide a little competition but it won't be that much lethal as it can be in regions of cluster 4.

Limitations and Suggestions for Future Research

In this project, we only consider two factors i.e. similarity between neighbourhoods and frequency of restaurants among them. There are many other factors such as population and income of residents that could influence the location for opening a new restaurant, frequency of offices and other work places nearby and cost of real state in the neighbourhood. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred location to open a new restaurant. In addition, this project uses free Sandbox Tier Account of Foursquare API that come with limitations on number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain better results.

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 7 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 3 and cluster 4 are most favourable locations to open a new restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding highly competitive and unfavourable areas in their decision to open a new restaurant.