Coursera Capstone Project

IBM Data Science Specialization

Opening an Indian Cuisine Restaurant in New York City, USA

By: Tushar Singhal

August 2020

1. Introduction

1.1 Background

While opening a restaurant can be a very lucrative business, a lack of demand causes many restaurants to close within the first year of opening. There are many different factors that can account for a restaurant's success such as location, competition and quality of the food. This is an important question that every business owner must face when choosing whether to open a restaurant or not, as well as location of the business. To demonstrate the process of picking a location for a client opening a business, the project will focus on answering were to open the restaurant. If there are too many Indian Restaurants in the local vicinity, the profitability of the restaurant will be severely decreased. Additionally, starting a restaurant in a location with higher income would increase the profitability of the business over starting in a poorer area.

1.2 Business Problem

The following question: "If the client wanted to open an Indian Restaurant in New York City, what areas are the best options to open the restaurant?" For an Indian Restaurant, the location and competition are both determined by where the restaurant is opened.

1.3 Target audience of this project

This project is mainly useful to an anyone who wishes to open an Indian Cuisine Restaurant in New York City. The insight from the project will be helpful for determining the best possible location of the restaurant. It will help it understanding whether there is a lot or little competition in a given neighbourhood of the city. Accordingly, the person who is interested in opening it can take the necessary decision with the help of the gathered intel.

2. Data

To solve the problem, the following data is needed:

- List of neighbourhoods in New York City. This defines the scope of this project which is confined to the city of New York.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

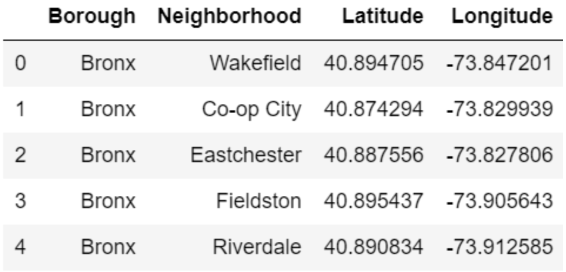
- Venue data.

Sources of data and methods to extract them,

The data for the neighbourhoods of New York is obtained from previous week of the course. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data in order to help us to solve the business problem put forward. This is a project will make use of many data science skills, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

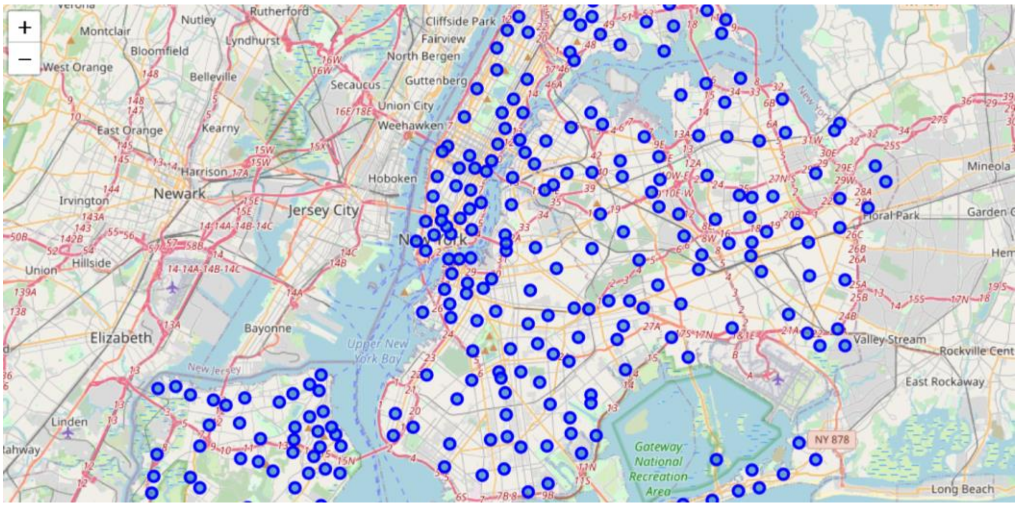
3. Methodology

First the json file containing the New York neighbourhood data set was loaded. It contains the boroughs, neighbourhoods and their respective coordinates. Then the data is loaded into a pandas data frame named neighbourhoods, only the required data is stored in it.



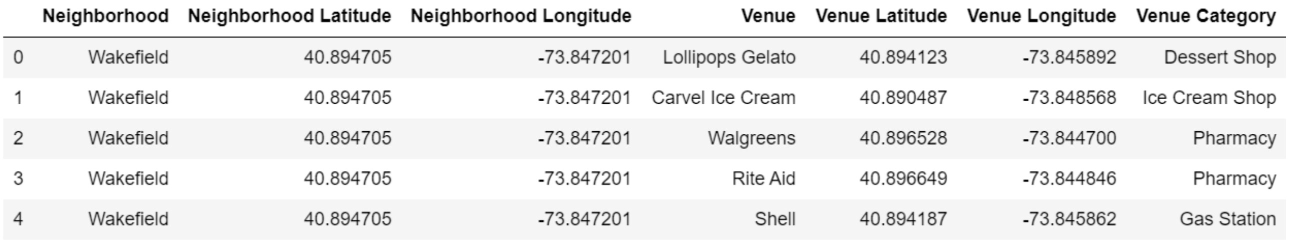
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

In total there are 306 unique neighbourhoods and using the folium library this is visualized as shown below



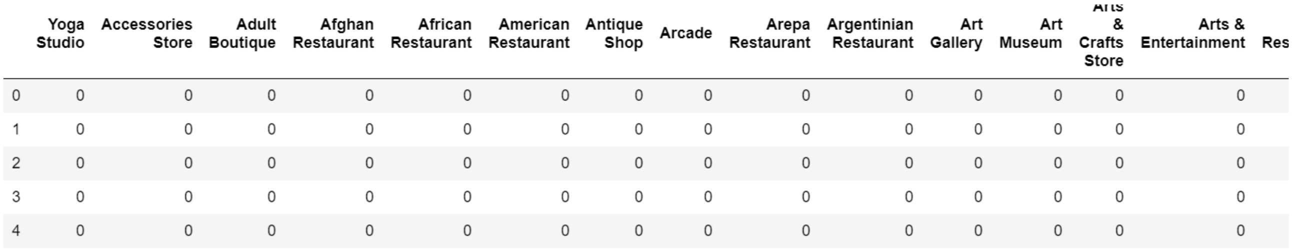
3

Using the Foursquare API up to 100 venues in a radius of 500m in each neighbourhood is obtained. This is done with the help of REST API. The data is stored in nyc_venues data frame.



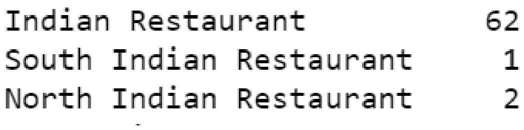
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

In total there are 429 unique venue categories. The data frame is then onehot encoded according to venues.



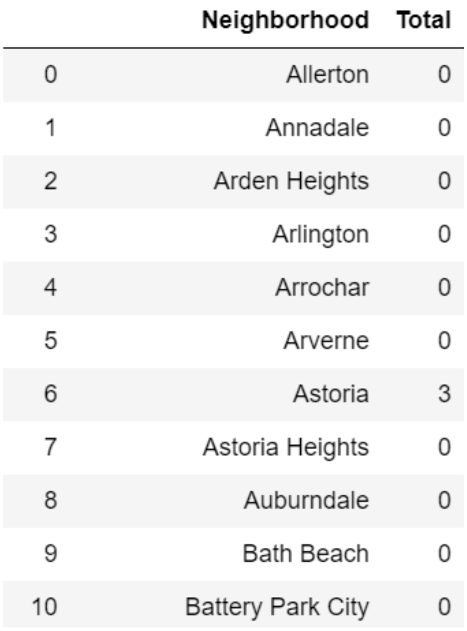
	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Res
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

There are three different kinds of Indian cuisine restaurants namely, Indian, South Indian and North Indian. In total there are 65 Indian restaurants combined.



```
Indian Restaurant          62
South Indian Restaurant     1
North Indian Restaurant     2
```

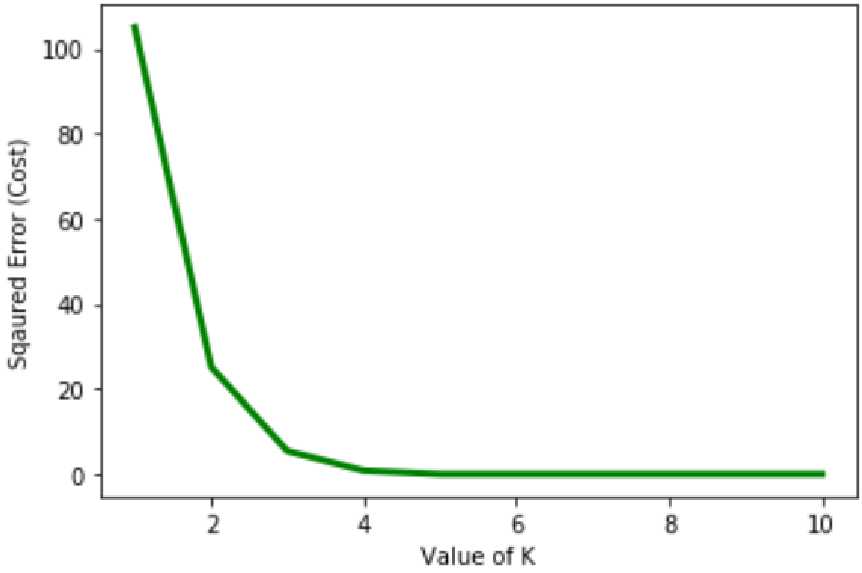
Then the data frame is grouped by neighbourhood and summed. A new data frame consisting of only the Indian cuisine restaurants is created and the three different kinds of Indian Restaurants are summed to a single total.



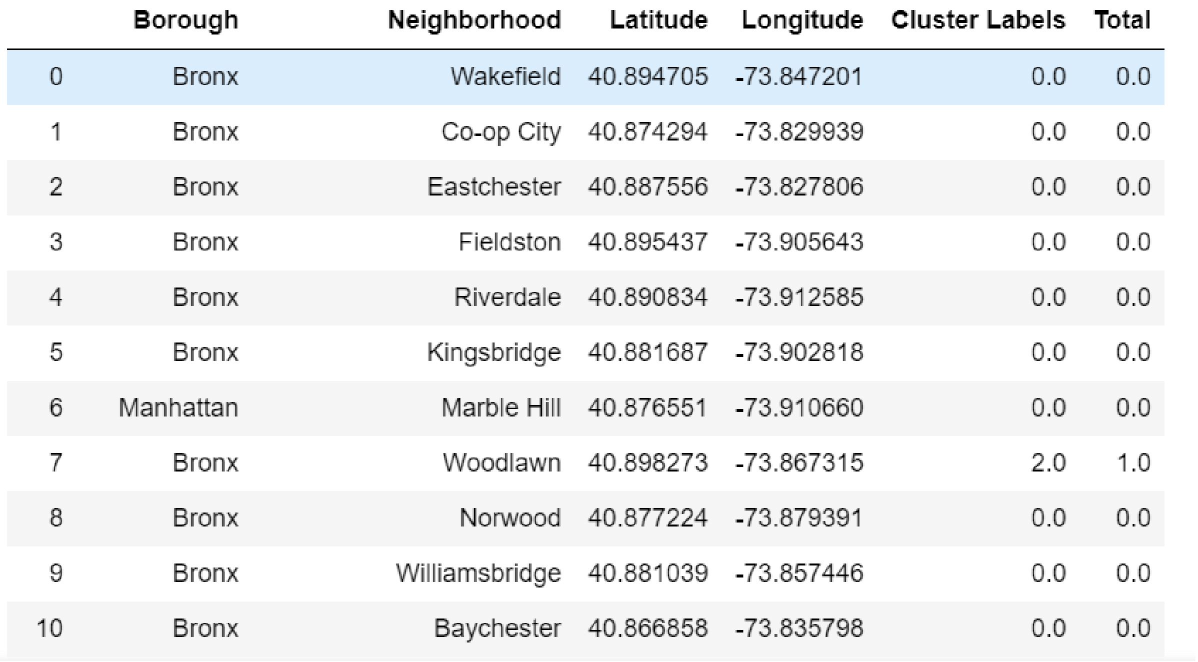
	Neighborhood	Total
0	Allerton	0
1	Annadale	0
2	Arden Heights	0
3	Arlington	0
4	Arrochar	0
5	Arverne	0
6	Astoria	3
7	Astoria Heights	0
8	Auburndale	0
9	Bath Beach	0
10	Battery Park City	0

Now that the data is prepared for analysis, we can use k-means clustering to identify the areas which have either a smaller number of restaurants or where they are clustered to find the optimal location for a new one. k-means clustering is used here as this is unsupervised form of learning.

To determine the value of k a graph of cost Squared error vs k is plotted using matplot.pyplot library.



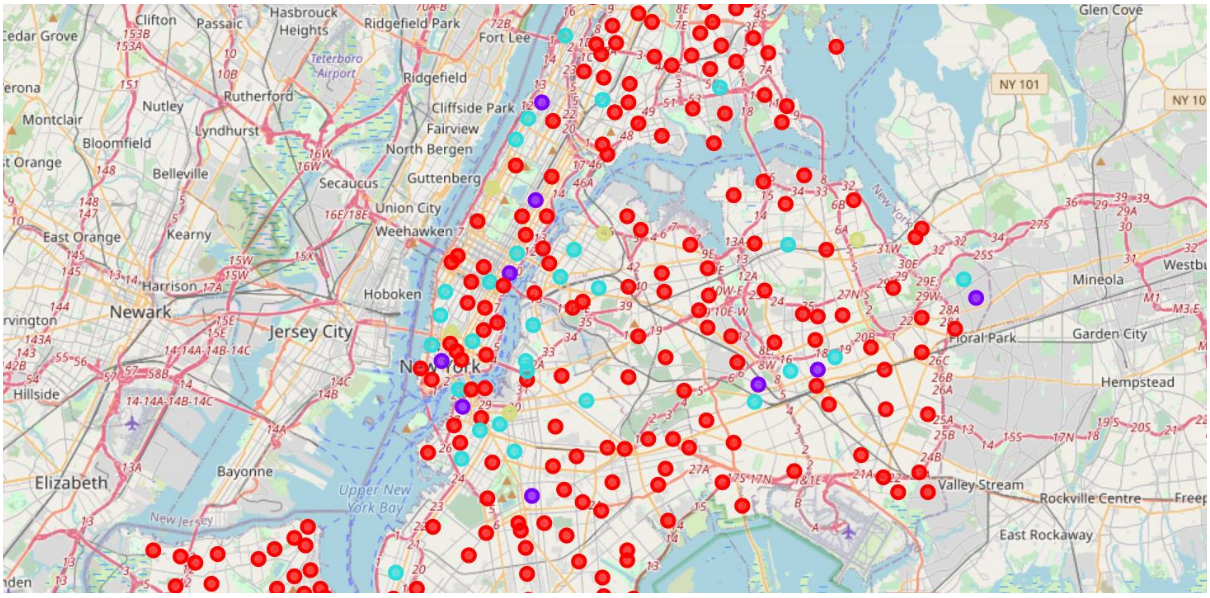
The elbow point is at 4 we can choose k as 4. After choosing k clustering is performed on the data to obtain the cluster labels. The obtained labels is merged with neighbourhood name and location data.



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total
0	Bronx	Wakefield	40.894705	-73.847201	0.0	0.0
1	Bronx	Co-op City	40.874294	-73.829939	0.0	0.0
2	Bronx	Eastchester	40.887556	-73.827806	0.0	0.0
3	Bronx	Fieldston	40.895437	-73.905643	0.0	0.0
4	Bronx	Riverdale	40.890834	-73.912585	0.0	0.0
5	Bronx	Kingsbridge	40.881687	-73.902818	0.0	0.0
6	Manhattan	Marble Hill	40.876551	-73.910660	0.0	0.0
7	Bronx	Woodlawn	40.898273	-73.867315	2.0	1.0
8	Bronx	Norwood	40.877224	-73.879391	0.0	0.0
9	Bronx	Williamsbridge	40.881039	-73.857446	0.0	0.0
10	Bronx	Baychester	40.866858	-73.835798	0.0	0.0

The type of the values of Cluster Labels is changed to integer from float to prepare for displaying using folium.

4. Result



Cluster Number	Colour	Number of Restaurants
0	Red	0
1	Purple	2
2	Teal	1
3	Gold	3 or more

In cluster number 3 i.e. Gold coloured 3 or 4 Indian cuisine restaurants are there.

5. Discussions

From the results of the clustering it can be inferred that the best neighbourhoods would be those which have either 0 or 1 Indian restaurant at the most. These neighbourhoods are represented by Red and Teal colour respectively. On the other hand red coloured clusters are predominantly in southern part of New York these neighbourhoods could be suburbs, so these could be avoided. The best areas are either the northern or Southern Manhattan were there are 0 or 1 restaurants. Even Brooklyn or Queens have good places to setup. At any cost Purple and Gold coloured clusters should be avoided as they won’t lead to higher profits due to competition already existing there.

6. Conclusion

Opening a restaurant is a complex task that can lead to a large monetary loss if not done properly. Thus, extensive research about the area would greatly increase the likelihood of the restaurant succeeding. From the project above, I demonstrated the workflow necessary for a client to determine what area the restaurant should open. This project can further be improved upon by taking demography and income of the households, this will give an idea of the spending power in that neighbourhood.