IBM Capstone Project week 5

Introduction

This is a Capstone initiative for IBM Data Science Professional Certificate. As part of this, I am going to recommend an area in Manhattan borough of New York City which would be best suited for opening a restaurant, considering some important factors based on available data. Manhattan appears to be the most densely populated of the five boroughs of New York City so before starting a new venture, like a restaurant, it is important to thoroughly study the different areas in this borough, the existence of already established food joints in those areas and a number of other factors. I would carry out the necessary work required to assist someone in finding the most suitable area.

Business Problem

The aim of this capstone project is to locate the most appropriate area for an entrepreneur to open a successfully running and revenue generating restaurant in Manhattan, New York City. By making use of different data science techniques, machine learning algorithms, Clustering, this project aims to provide a response to the following business question:

In the Manhattan borough of New York City, if someone is looking to open a restaurant, where would you recommend that they open it?

Target Audience

The entrepreneurs who would like to find a location to open a restaurant in Manhattan, New York.

Data

For finding an area in the most densely populated area, we would require the following data:

- List of neighbourhoods in Manhattan, New York, USA
- Latitude and Longitude of these neighbourhoods
- Top ten venues for these neighbourhoods to see the existence of different features that are helpful for opening an eatery, like a restaurant.

Data Extraction Strategies

- Using extracted Json Data for New York neighbourhoods
- Extracting longitude and latitude information of these neighbourhoods using the GeoPy Geocoder package
- Using Foursquare API to get venue details for these neighbourhoods

Methodology

In order to get the list of neighbourhoods in Manhattan, New York City, I used the data available in https://cocl.us/new_york_dataset This data is freely available here as well: https://geo.nyu.edu/catalog/nyu_2451_34572

I then transformed the data into a pandas dataframe and used Geopy library to get the latitude and longitude values of New York City. One of the boroughs that are densely populated and also a centre hub of business appears to be Manhattan. This is the borough in which we'll be looking for an area.

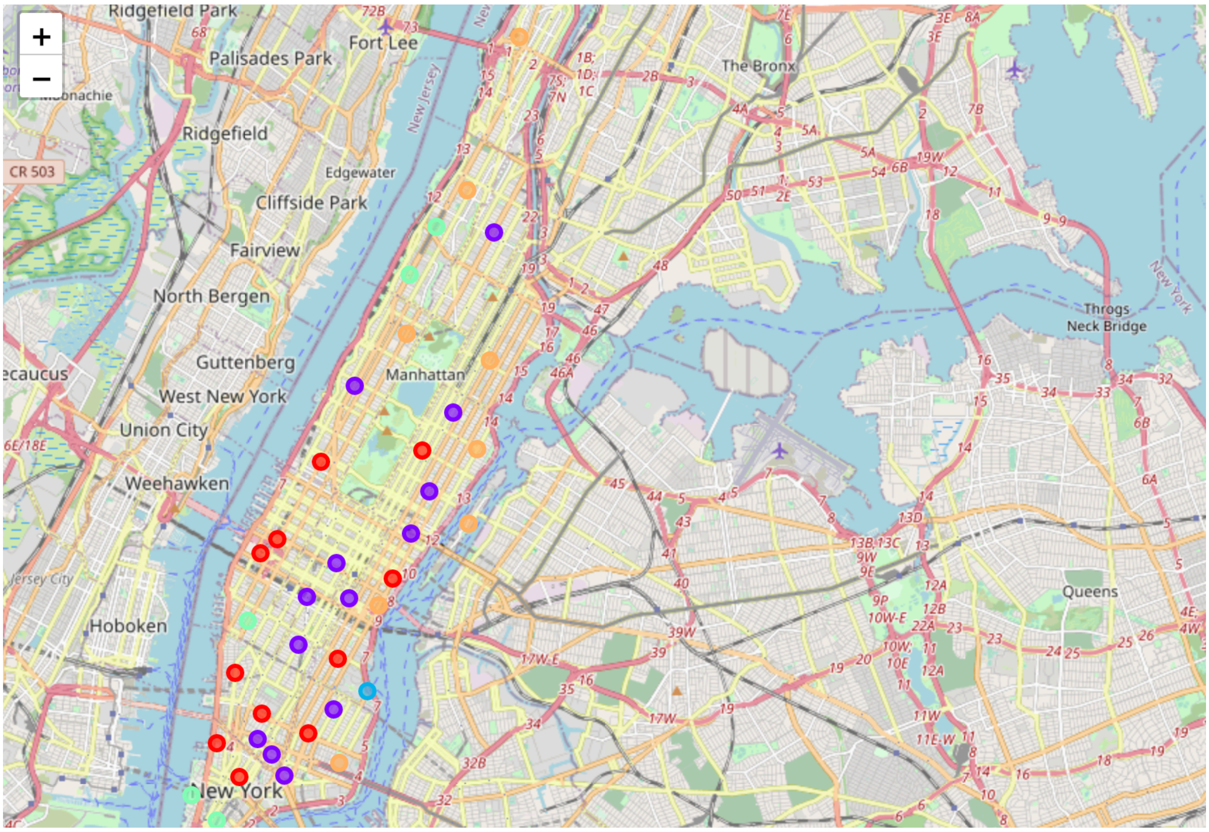
To get venue details from FourSquare, we need coordinates of Manhattan area. For this, I used the GeoCode from GeoLocator package. To visualise the neighbourhoods, I created a folium map. Next, I use Foursquare API to pull a list of top 10 venues in the radius of 500 metres. To use the API, I have a developer account in Foursquare. Using this, I get the names, categories, latitude and longitude of the nearby venues. It becomes easier at this stage to find out how many unique categories can be curated from all the returned venues.

The next step is to analyse each neighbourhood by grouping rows by neighbourhood and by taking the mean of the frequency of occurrence of each category. This information would be required for doing the clustering at a later stage.

Once I have clarity on the top venues in a neighbourhood, I then proceed towards clustering of neighbourhoods. **K-means clustering** is one of the simplest and popular unsupervised machine learning algorithms. **K-means** looks for a fixed number (**k**) of **clusters** in a dataset." A **cluster** refers to a collection of data points aggregated together because of certain similarities. I ran k-means to cluster the neighbourhood into 5 clusters based on the top 10 venues to determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, I can come up with results on which area would be most suitable for the purposes of opening a restaurant.

Results:

Clusters:



The results from k-means clustering show that we can categorise Manhattan neighbourhoods into 5 clusters based on how many discriminating venue categories are there in each cluster.

Cluster 1-3: A lot of restaurants already in existence.
Cluster 4: No other restaurant but other venues like Park, Playground and recreational activity grounds in vicinity.
Cluster 5: Hardly any restaurants but a seafood restaurant in vicinity

Discussions

- Most of the restaurants are available in Cluster 1 – 3 so opening another one in these areas could be very competitive and may not attract much business
- Clusters 4-5 look suitable for consideration of opening a restaurant

Recommendations

Most of the restaurants are in Clusters 1-3.

Looking at the nearby venues, Stuyvesant Town would be quite suitable for opening a new restaurant.

Conclusion

- The neighbourhoods in cluster 4 offer the most preferrable location to open a new restaurant.
- Findings and observations from this project will help the relevant entrepreneurs to zero in on high potential areas while avoiding the areas in other clusters that already have restaurants as top 10 venues. This should provide useful information to stakeholders who are looking for a suitable area for opening a restaurant in a suitable location in Manhattan.