**IBM Applied Data Science Capstone – Coursera**



**Opening a new Coffee Shop in New York City, USA**

**Mahshid Gohari**

**08/19/2021**

Table of Contents

[Introduction 2](#_Toc80307408)

[Business Problem 3](#_Toc80307409)

[Target Audience of this project 3](#_Toc80307410)

[Sources of data and methods to extract them 4](#_Toc80307411)

[Methodology 4](#_Toc80307412)

[Results 5](#_Toc80307413)

[Discussion 6](#_Toc80307414)

[Future Research 7](#_Toc80307415)

[Conclusion 7](#_Toc80307416)

[References 8](#_Toc80307417)

# Introduction

New York, often called New York City to distinguish it from [New York State](https://en.wikipedia.org/wiki/New_York_(state)), or NYC for short, is the [most populous city](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) in the [United States](https://en.wikipedia.org/wiki/United_States). With a 2020 population of 8,804,190 distributed over 302.6 square miles (784 km2), New York City is also the [most densely populated](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population_density) major city in the United States. Located at the southern tip of the State of New York, the city is the center of the [New York metropolitan area](https://en.wikipedia.org/wiki/New_York_metropolitan_area), the largest [metropolitan area](https://en.wikipedia.org/wiki/Metropolitan_area) in the world by [urban area](https://en.wikipedia.org/wiki/Urban_area). With over 20 million people in its [metropolitan statistical area](https://en.wikipedia.org/wiki/Metropolitan_statistical_area) and approximately 23 million in its [combined statistical area](https://en.wikipedia.org/wiki/Combined_statistical_area), it is one of the world's most populous [megacities](https://en.wikipedia.org/wiki/Megacities). New York City has been described as the [cultural](https://en.wikipedia.org/wiki/Culture_of_New_York_City), financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports, and is the most photographed city in the world. Home to the [headquarters of the United Nations](https://en.wikipedia.org/wiki/Headquarters_of_the_United_Nations), New York is an important center for [international diplomacy](https://en.wikipedia.org/wiki/International_diplomacy), and has sometimes been called the [capital of the world](https://en.wikipedia.org/wiki/Caput_Mundi).

New York City is one of the most famous, romanticized and desired cities for living in the world due to its image portrayed in the pop culture. Some of the worlds' most popular content creators and influencers are located in the city for its stunning backgrounds and instagramable dining and coffee locations. Combining that with the overall growing trend for coffee and small business coffee shops makes it an attractive investment to open a new coffee shop specifically in NYC where this coffee shop is likely to receive a lot of attention on social media to drive revenue.

# Business Problem

The objective of this capstone project is to analyze and recommend the best neighborhoods in the city of New York, USA to open a new Coffee Shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Where would you recommend a new investor to open a new Coffee Shop in the city of New York?

# Target Audience of this project

This project is useful for any investors who are willing to open a new Coffee Shop in the city of New York. Based on the rankings provided by Trip Advisor in 2021, [Frisson Espresso](https://www.tripadvisor.com/Restaurant_Review-g60763-d7716799-Reviews-Frisson_Espresso-New_York_City_New_York.html) from New York ranked as the #1 Coffee Shop in USA. New York neighborhoods already have a number of Coffee Shop chains, specialty Coffee Shop chains and local Coffee Shop places

**Data**

Data required to approach the above business case should include New York City neighborhoods' and boroughs' geographical locations (latitude, longitude) and boroughs' population density.

New York City data containing the neighborhoods and boroughs latitudes and longitudes can be obtained from <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json>. While FourSquare API will be used to obtain data on existing coffee shops.

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

* List of neighborhoods in New York. This defines the scope of this project
* Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
* Venue data, particularly data related to Coffee Shop places which is required to perform clustering on the neighborhoods.

# Sources of data and methods to extract them

The<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json> contains a list of neighborhoods in and around New York, with a total of 306 neighborhoods and 5 boroughs. We will use the web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and ***beautifulsoup*** packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. 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 Coffee Shop places 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 web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

# Methodology

First step on this analysis to get a list of neighborhood s in the city of New York. Fortunately, the list is readily available in the IBM. Using the web scraping techniques supported by Python and beautifulsoup packages, all the neighborhood names are extracted. Before proceeding, the extracted list is compared to Wiki to confirm all neighborhoods are included in the result.

Next step is to remove duplicates and set the new value for neighborhoods that are in 2 or 3 boroughs we did this with concatenate borough and neighborhoods. Next is to find out the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, the Geocoder package is used to convert neighborhood address into geographical coordinates in the form of latitude and longitude. All gathered data is then populated to a pandas DataFrame and then visualized the neighborhoods through 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 city of New York.

Using Foursquare API, captured a list of top 100 venues that are within a radius of 5000 meters. To do so, I first registered a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then added made API calls to Foursquare passing in the geographical coordinates of the neighborhoods and the Foursquare credentials in a Python loop. Foursquare returned the venue data in JSON format and extracted the venue name, venue category, venue latitude and longitude. With the data, checked how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues.

In the next step, analyzed each neighborhood by grouping the rows by neighborhood 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 analyzing the “Coffee Shop” data, we will filter the “Coffee Shop Places” as venue category for the neighborhood s. Lastly, performed 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 neighborhoods into 3 clusters based on their frequency of occurrence for “Coffee Shop”. The results will allow us to identify which neighborhoods have higher concentration of Coffee Shop places while which neighborhoods have fewer number of Coffee Shop places. Based on the occurrence of Coffee Shop places in different neighborhoods, it will help us to answer the question as to which neighborhood s are most suitable to open new Coffee Shop.

# Results

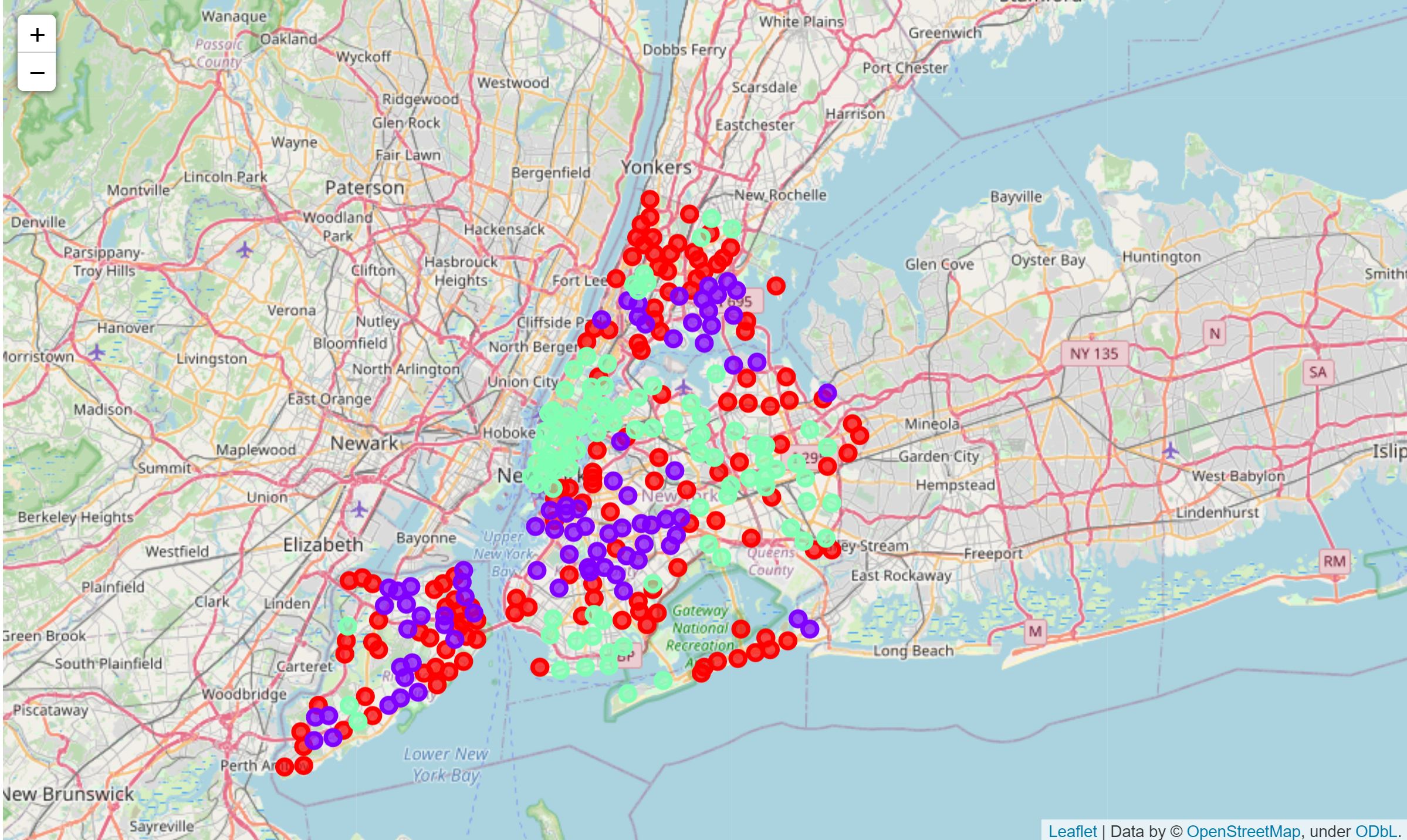
The results from the k-means clustering showing that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

#### Cluster 0: means that the neighborhoods that have 2 or 3 Coffee Shops, around 43%

#### Cluster 1: means that the neighborhoods that have 4 or 5 or 6 Coffee Shops, around 25%

#### Cluster 2: means that the neighborhoods that have 1 Coffee Shop, around 31%

The results of the clustering are visualized in the map below with cluster 0 in RED color, cluster 1 in PURPLE color, and cluster 2 in MINT GREEN color



# Discussion

Analysis showing that most of the Coffee Shops are concentrated in cluster 0 and some of the financial district area (around 43%). Cluster 1 have around 25% concentration and cluster 2 have around 31%. This is showing that new Coffee Shops can be opened in neighborhoods listed in cluster 1 as less competition compared to cluster 0 and 2. Meanwhile, Coffee Shops in cluster 0 are likely suffering from intense competition due to oversupply and high concentration. Coffee Shops with unique menus and special varieties can survive from the competition and can open on these neighborhoods.

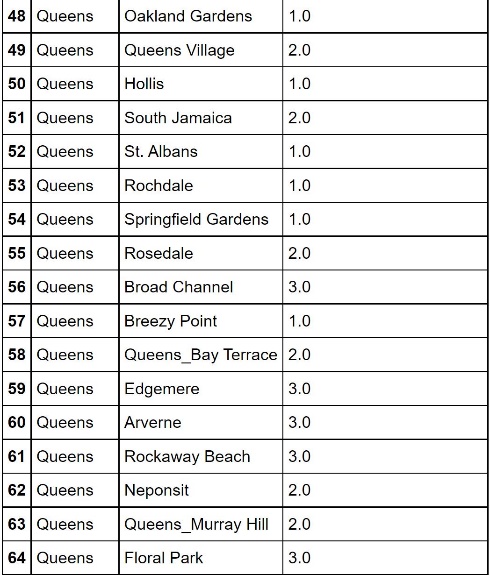
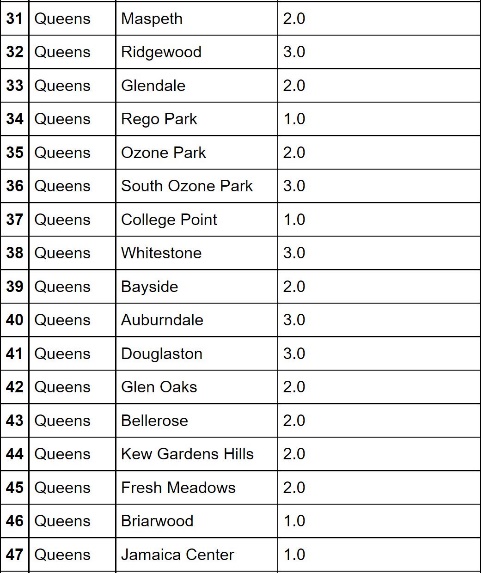
# Future Research

In this project, only considered one factor i.e., frequency of occurrence of Coffee Shops, there are other factors such as population, income of residents etc. can also influence the location decision of a new Coffee Shops. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Coffee Shops. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

# Conclusion

In this project, all steps in a data science analysis are applied like identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e., investors regarding the best locations to open a new Coffee Shops. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Coffee Shops. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Coffee Shops. So, we concentrate on cluster 1. we know that the Queens, Brooklyn and Manhattan are crowded and their density of population are high. So, we can choose these 2 boroughs but as you can see from above, Brooklyn has many coffee shops on 30 neighborhoods. So, we rather to choose Manhattan and Queens because Manhattan has 6 coffee shops on just 1 neighborhood (Hamilton Heights) and Queens on 7 neighborhoods. The other neighborhoods are on cluster 0 or 2. Thus, we choose other neighborhoods except those neighborhoods and eliminate 'Hamilton Heights', 'Middle Village', 'Beechhurst', 'Bayswater', 'Blissville', 'Little Neck', 'Malba', 'Far Rockaway'. Thus, the best neighborhoods to open new Coffee Shops are:







# References

* Wikipedia - <https://en.wikipedia.org/wiki/Boston>
* Wikipedia - <https://en.wikipedia.org/wiki/Greater_Boston>
* Beautiful Soup documentation - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
* Pandas documentation - <https://pandas.pydata.org/pandas-docs/stable/>
* Folium documentation - <https://python-visualization.github.io/folium/>
* Top 5 Coffee Shops in USA (Article: pizzatoday.com/news/pizza-headlines/guide-to-the-2018-national-best-pizzas-lists/)