The Battle of the Neighbourhoods – Report

1. Introduction & Business Problem :

Problem Background:

The City of New York is the most populous city in the United States. It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business opportunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major centre for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, fashion, and the arts in the United States.

This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analysed carefully. The insights derived from analysis will give good understanding of the business environment which helps in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

Problem Description:

A restaurant is a business which prepares and serves food and drink to customers in return for money, either paid before the meal, after the meal, or with an open account. The City of New York is famous for its excellent cuisine. Its food culture includes an array of international cuisines influenced by the city's immigrant history.
1. Central and Eastern European immigrants, especially Jewish immigrants - bagels, cheesecake, hot dogs, knishes, and delicatessens
2. Italian immigrants - New York-style pizza and Italian cuisine
3. Jewish immigrants and Irish immigrants - pastrami and corned beef
4. Chinese and other Asian restaurants, sandwich joints, diners, and coffee houses are ubiquitous throughout the city
5. Mobile food vendors - Some 4,000 licensed by the city
6. Middle Eastern foods such as falafel and kebabs examples of modern New York street food
7. It is famous for not just Pizzerias, Cafe's but also for fine dining Michelin starred restaurants. The city is home to "nearly one thousand of the finest and most diverse haute cuisine restaurants in the world", according to Michelin.

So it is evident that to survive in such competitive market it is very important to strategically plan.

Various factors need to be studied in order to decide on the Location such as:

1. New York Population and Demographics
3. Are there any Farmers Markets, Wholesale markets nearby so that the ingredients can be purchased fresh to maintain quality and cost?
4. Are there any venues like Gyms, Entertainment zones, Parks nearby where floating population is high?
5. Who are the competitors in that location?
6. Cuisine served / Menu of the competitors
7. Segmentation of the Borough
8. Untapped markets and Saturated markets

Even though XYZ Company Ltd need to choose the correct location to start its first venture. If this is successful they can replicate the same in other locations. First move is very important thereby choice of location is very important.

Target Audience:

To recommend the correct location, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighbourhood of New York city will be best choice to start a restaurant. The Management also expects to understand the rationale of the recommendations made.
This would interest anyone who wants to start a new restaurant in New York City.

Success Criteria:

The success criteria of the project will be a good recommendation of Borough/Neighbourhood choice to XYZ Company Ltd based on Lack of such restaurants in that location and nearest suppliers of ingredients.

2. Data :

One city will be analysed in this project: **New York City**
We will be using the below datasets for analysing New York city

Data 1: Neighbourhood has a total of 5 boroughs and 306 neighbourhoods. In order to segment the neighbourhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood.

This dataset exists for free on the web. Link to the dataset is:
https://geo.nyu.edu/catalog/nyu_2451_34572

	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

Data 2: Second data which will be used is the DOHMH Farmers Markets and Food Boxes dataset.

In this we will be using the data of Farmers Markets.
https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets-and-Food-Boxes/8vwk-6iz2
Website-https://www.grownyc.org/greenmarketco/foodbox

GrowNYC's Fresh Food Box Program is a food access initiative that enables under-served communities to purchase fresh, healthy, and primarily regionally grown produce well below traditional retail prices. A farmers' market is often defined as a public site used by two or more local or regional producers for the direct sale of farm products to consumers. In addition to fresh fruits and vegetables, markets may sell dairy products, fish, meat, baked goods, and other minimally processed foods.

Data 3: For the below analysis we will get data from wikipedia as given below :

1. New York Population
2. New York City Demographics
3. Cuisine of New York city

https://en.wikipedia.org/wiki/New_York_City
https://en.wikipedia.org/wiki/Economy_of_New_York_City
https://en.wikipedia.org/wiki/Portal:New_York_City
https://en.wikipedia.org/wiki/Cuisine_of_New_York_City

Data 4: New York city geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighbourhood. We will use the Foursquare API to explore neighbourhoods in New York City. The below is image of the Foursquare API data.

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

3. Methodology :

Business Understanding :

Our main goal is to get optimum location for new restaurant business in New York City for XYZ Company.

Analytic Approach :

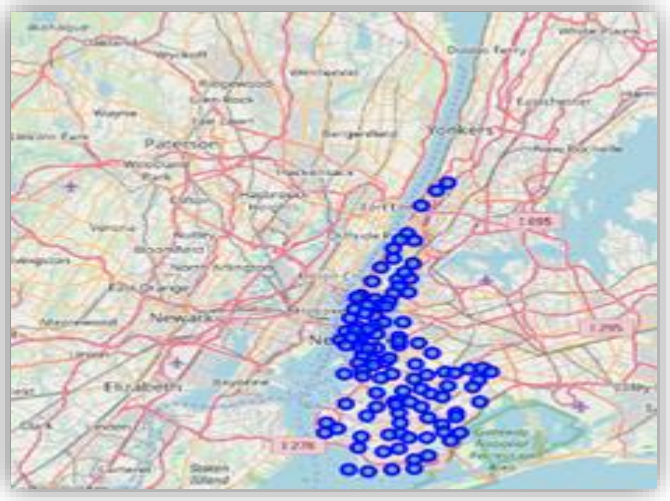
New York city neighbourhood has a total of 5 boroughs and 306 neighbourhoods. In this project we are clustering Manhattan and Brooklyn. This is done because of the following Exploratory data analysis.

Exploratory Data Analysis:

Data 1- New York city Geographical Coordinates Data

1. In this we load the data and explore data from newyork_data.json file.
2. Transform the data of nested python dictionaries into a pandas dataframe.
3. This dataframe contains the geographical coordinates of New York city neighbourhoods.
4. This data will used to get Venues data from Fouresquare.
5. We used geopy and folium libraries to create a map of New York city with neighbourhoods super imposed on top.

New York neighbourhood visualization



Data 2- Second data which is used is the DOHMH Farmers Markets and Food Boxes dataset.

In this we will be using the data of Farmers Markets data. There are totally 144 Farmers Markets in New York city. Highest number are in Manhattan and Brooklyn. And lowest in Queens, Bronx and Staten Island.

Data 3- To analyse New York city Population, Demographics and Cuisine, scrapped the data from Wikipedia pages given above in the data section. We used BeautifulSoup python library. Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping

New York Population: Insights from the data:

1. Manhattan (New York County) is the geographically smallest and most densely populated borough.
2. Manhattan's (New York County's) population density of 72,033 people per square mile (27,812/km²) in 2015 makes it the highest of any county in the United States and higher than the density of any individual American city.
3. Brooklyn (Kings County), on the western tip of Long Island, is the city's most populous borough.
4. Queens (Queens County), on Long Island north and east of Brooklyn, is geographically the largest borough.

	Borough	County	Estimate_2017	square_miles	square_km	persons_sq_mi	persons_sq_km
0	The Bronx	Bronx	1,471,160	19,570	42.10	109.04	34,653
1	Brooklyn	Kings	2,648,771	23,900	70.82	183.42	37,137
2	Manhattan	New York	1,664,727	378,250	22.83	59.13	72,033
3	Queens	Queens	2,358,582	31,310	108.53	281.09	21,460
4	Staten Island	Richmond	479,458	23,460	58.37	151.18	8,112
5		City of New York	8,622,698	806.863	302.64	783.83	28,188
6		State of New York	19,849,399	1,547.116	47,214	122,284	416.4

Data 4- Foursquare: We used the Foursquare API data to explore neighbourhoods in New York city:

NewYork city geographical coordinates data is utilized as input for the Foursquare API, that has been leveraged to provision venues information for each neighbourhood.

Brooklyn and Manhattan:

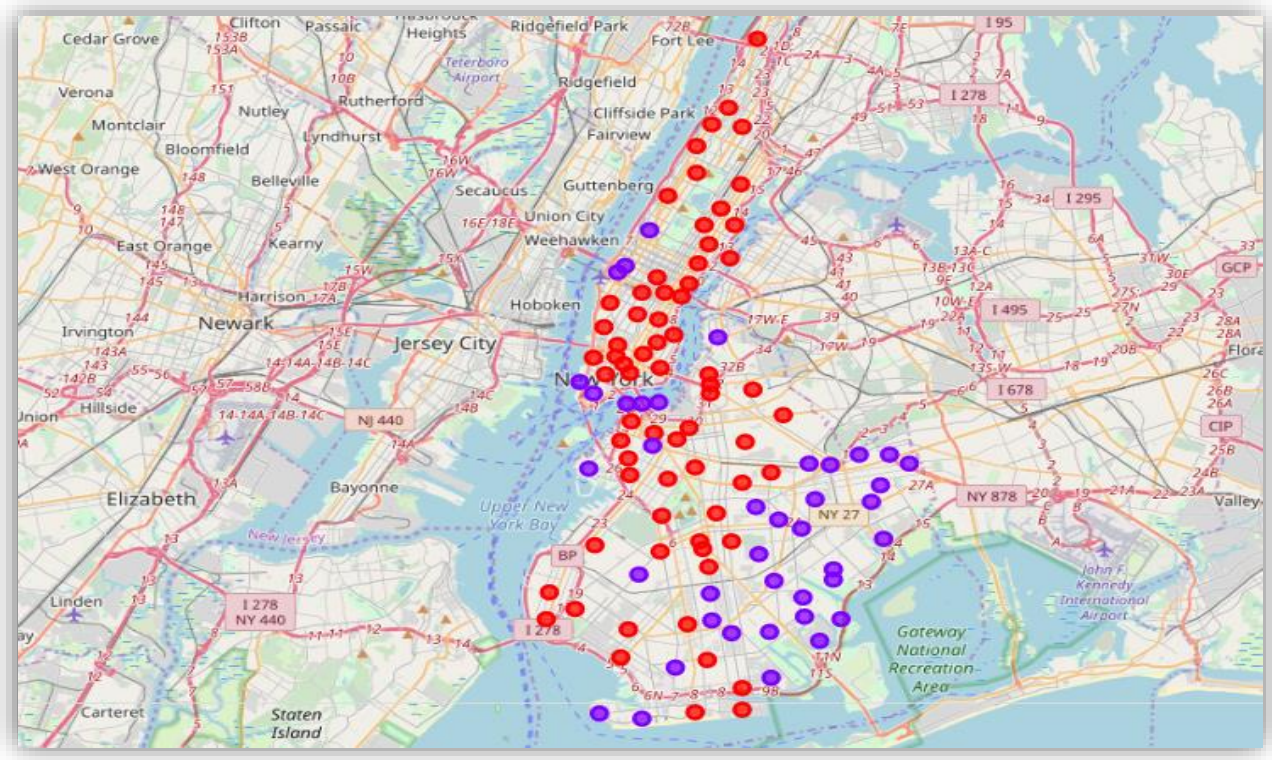
Brooklyn and Manhattan Venues:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Brooklyn and Manhattan Visualization :
Using the geographical coordinates of each neighbourhood foursquare API calls are made to get top 200 venues in a radius of 1000 meters. The venues data is as given below:

Brooklyn and Manhattan Venues Visualization:

Generated the below Brooklyn and Manhattan Venues Visualization. The "BM_venues" dataframe has 9708 venues and 397 unique venue types.

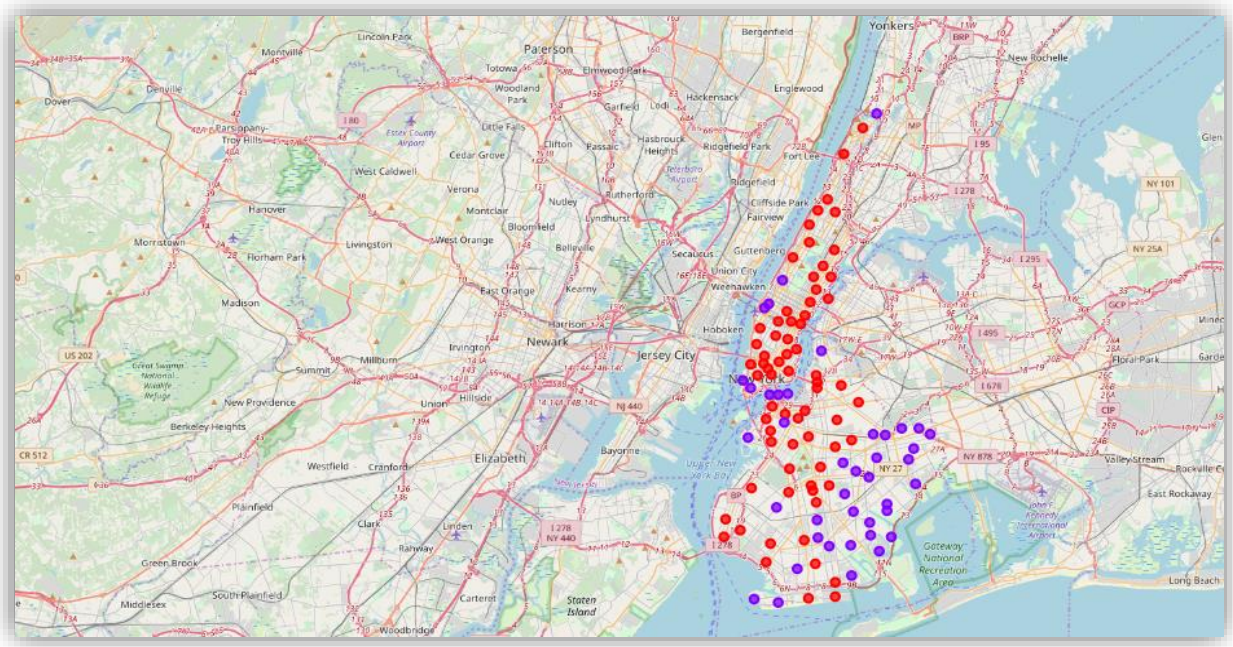


4. RESULT:

From this venues data we filtered and used only the restaurant data for Brooklyn & Manhattan clustering and Bronx, Queens and Staten Island clustering. As we focussed only on restaurants business.

Neighbourhood K-Means clustering based on mean occurrence of venue category:
To cluster the neighbourhoods into two clusters we used the K-Means clustering Algorithm. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

Brooklyn & Manhattan:
In the below Map Visualization, we can see the different types of clusters created by using K-Means for Brooklyn & Manhattan.



Cluster0: The Total and Total Sum of cluster0 has smallest value. It shows that the market is not saturated.

Cluster1: The Total and Total Sum of cluster1 has highest value. It shows that the markets are saturated. Number of restaurants are very high.

There are no untapped neighbourhoods in Brooklyn and Manhattan.

5. DISCUSSION:

1. There is scope to increase Farmers markets in Bronx, Queens and Staten Island.
2. There is scope to explore cuisines of various countries in Bronx, Queens and Staten Island.
3. In Manhattan and Brooklyn restaurants of cuisines of many countries are available. So if risk can be taken with great menu on board. It also shows people love eating cuisines of various countries.

6. CONCLUSION:

This analysis is performed on limited data.
This may be right or may be wrong. But if good amount of data is available there is scope to come up with better results. If there are lot of restaurants probably there is lot of demand.
Brooklyn and Manhattan has high concentration of restaurant business. Very competitive market. Bronx, Queens and Staten Island also has good number of restaurants but not as many as required. So this can be explored.
As per the neighbourhood or restaurant type mentioned like Indian Restaurant analysis can be checked.
A venue with lowest risk and competition can be identified.