Applied Data Science Project – Coursestone Project

Battle of the Neighborhoods – Introduction

Problem Background:

New York City is the most well-known city in the United states and there is a large amount of data available about the city online. It has a diverse financial and ethnical culture. It provides lot of business opportunities and an environment that encourages business.

This of course means that it is also a highly competitive market with many players. Due to cost the entry barrier is also quite high which means that making the right strategical decision when entering is important. This can be supported by using data science methods as taught in Coursestone Project.

Problem Description:

A restaurant has approached us to help them choose the correct location for their new restaurant in the New York city. Due to multicultural environment, the New York city has cuisine from all over the world (e.g. European, middle eastern, Asian, Chinese and many other cuisines). The way this food is also sold differs, as there are street vendors, food trucks, restaurants and fast food places all over the city.

This makes the problem more complex as the competition in this market is high and the customer already has a lot of choices all throughout the city. This means that our client needs to consider all factors they possibly can when choosing the right location.

Key stakeholders:

The key stakeholders is the restaurant that has hired us to perform this research. As they are the management and they are paying us to do this, they expect to have the results presented in a well-structured manner so that the results can be used to make management (Strategical) decisions.

Success Criteria:

The success criteria for this project is defined as being able to recommend the key stakeholders the best neighborhood in Newyork where there can set-up their restaurant based on two main factors:

- Minimal competition
- Availability of fresh produce

Data

The scope of this Project is the New York city.

Initially we gather the data on the New York city neighborhoods using information found online : (Link and Link2)

Using these two dataset, we were able to gather data on Boroughs (5 total), Neighborhoods (306 total) and Latitude/Longitude coordinates we will need for Foursquare processing.

	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

To identify the needed sources for fresh produce we used a single dataset available online on farmer market locations in NY city (Link)

	FacilityName	Service Category	Service_Type	Address	Address 2	Borough	ZipCode	Latitude	Longitude	AdditionalInfo	StartDate	EndDate	Monday	Tuesday	Wednesday	Thursday
0	Inwood Park Greenmarket	Farmers Markets and Food Boxes	Farmers Markets	Isham St bet Seaman & Cooper	NaN	Manhattan	10034	40.869009	-73.920320	Open year-round	NaN	NaN	NaN	NaN	NaN	NaN
1	82nd Street Greenmarket	Farmers Markets and Food Boxes	Farmers Markets	82nd St bet 1st & York Aves	NaN	Manhattan	10028	40.773448	-73.948954	Open year-round	NaN	NaN	NaN	NaN	NaN	NaN
3	125th Street Farmers Market	Farmers Markets and Food Boxes	Farmers Markets	125th St & Adam Clayton Powell Jr Blvd	NaN	Manhattan	10027	40.808981	-73.948327	Market open dates: 6/13/2017 to 11/21/2017	06/13/2017	11/21/2017	NaN	10am-7pm	NaN	NaN
4	170 Farm Stand	Farmers Markets and Food Boxes	Farmers Markets	170th St & Townsend Ave	NaN	Bronx	10452	40.840095	-73.916827	Market open dates: 7/5/2017 to 11/22/2017	07/05/2017	11/22/2017	NaN	NaN	2:30pm-6:30pm	NaN
5	175th Street Greenmarket	Farmers Markets and Food Boxes	Farmers Markets	175th St bet Wadsworth Ave & Broadway	NaN	Manhattan	10033	40.845956	-73.937813	Market open dates: 6/29/2017 to 11/30/2017	06/29/2017	11/30/2017	NaN	NaN	NaN	8am-5pm

To have a better insight into the population, demographics and cuisine in NY we used the following datasets from Wikipedia and scraped the web.

1. New York Population and Demographics on wikipedia (Link, Link1)
2. Cuisine of New York city on wikipedia(Link)

Using Foursquare and the new York city neighborhood coordinates we were able to identify Venues information and category. This allowed us to observe our competition.

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
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	Loeser's Delicatessen	40.879242	-73.905471	Sandwich Place

Methodology :

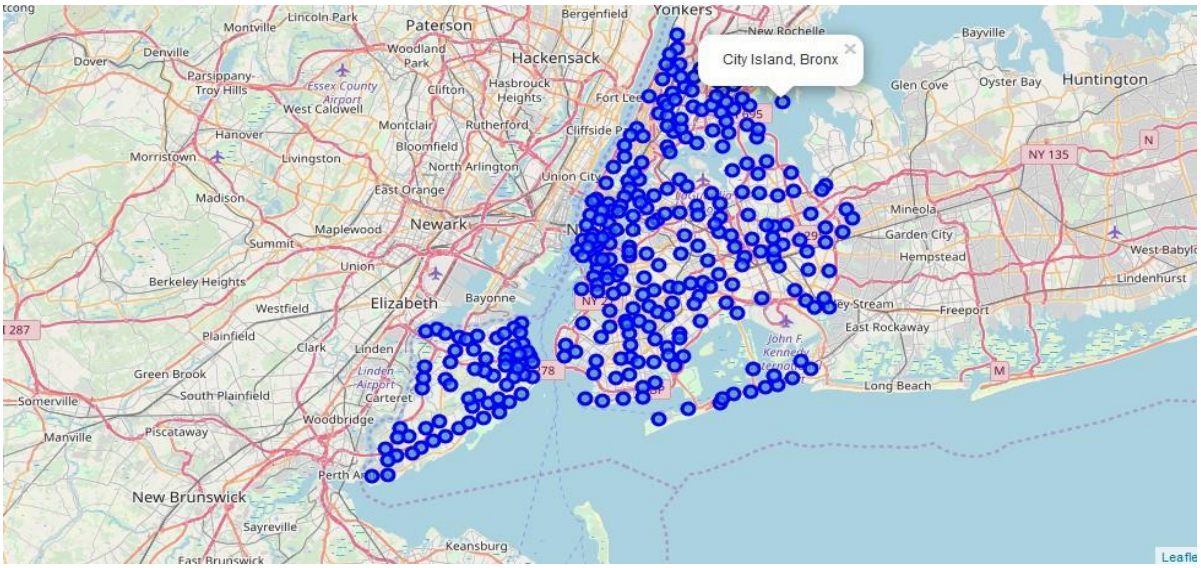
Business Understanding :

Our data mining efforts is to find the best location for the restaurant to have the highest probability of success.

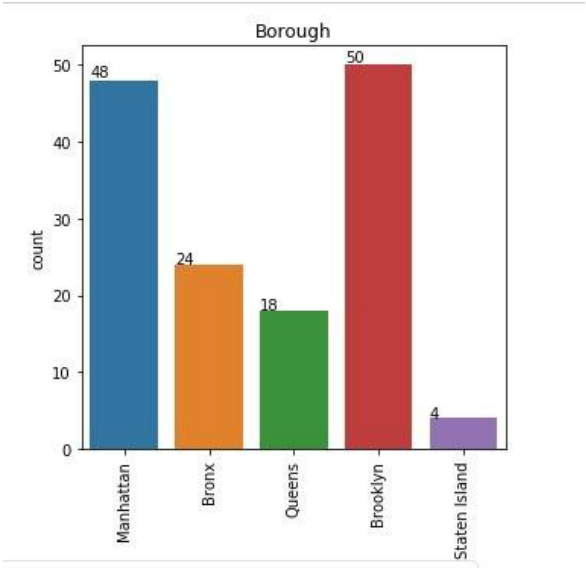
Data analysis/Understanding:

New York city has a total of 5 boroughs and 306 neighborhoods.

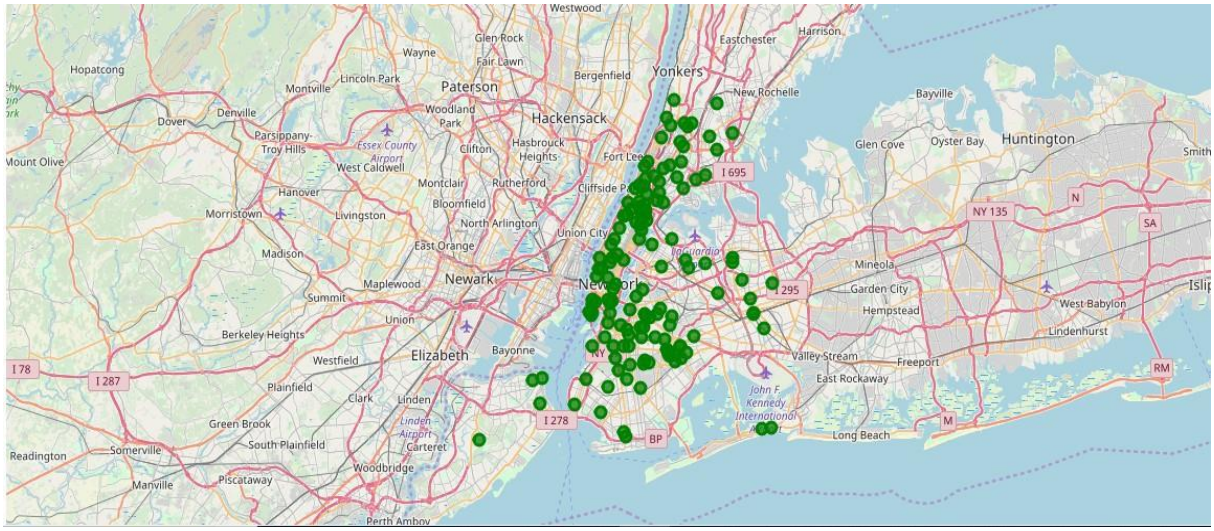
In the below screenshot all neighborhoods around NY can be seen.



There are in total 144 Farmers Markets in NY city. Highest number are in Manhattan and Brooklyn and the lowest in Queens, Bronx and Staten Island. In the below screenshot we can see the frequency of farmer markets in each Borough.



In the below screenshot we can see all the locations of Farmer Markets across NY city.



By scraping the data from Wikipedia pages, we were able to analyze the data of NY city population, demographics and cuisine. The following observations were made:

- Manhattan is the most densely populated borough with 72 thousand people per square mile (According to data from 2015).
- Brooklyn is the most populated borough
- Queens is the largest borough by landmass

	Borough	County	Estimate_2017	square_miles	square_km	persons_sq_mi	persons_sq_km
0	Manhattan	New York	1,664,727	22.83	59.13	72,033	27,826
1	The Bronx	Bronx	1,471,160	42.10	109.04	34,653	13,231
2	Brooklyn	Kings	2,648,771	70.82	183.42	37,137	14,649
3	Queens	Queens	2,358,582	108.53	281.09	21,460	8,354
4	Staten Island	Richmond	479,458	58.37	151.18	8,112	3,132
5		City of New York	8,622,698	302.64	783.83	28,188	10,947
6		State of New York	19,849,399	47,214	122,284	416.4	159

In the below screenshot is a summary of the racial distribution of NY city.

	Racialcomposition	2010	1990	1970	1940
0	White	44.0%	52.3%	76.6%	93.6%
1	—Non-Hispanic	33.3%	43.2%	62.9%	92.0%
2	Black or African American	25.5%	28.7%	21.1%	6.1%
3	Hispanic or Latino (of any race)	28.6%	24.4%	16.2%	1.6%
4	Asian	12.7%	7.0%	1.2%	–

As it can be seen, the diversity of the different racial profiles has been steadily increased over the decades.

Based on the scraped data from Wikipedia the following cuisines are the most preferred in each borough:

New York city : Most Preferred Food in New York City –Italian, Purto Rican, Mexican, Jewish, Indian, Pakistani & Dominican.

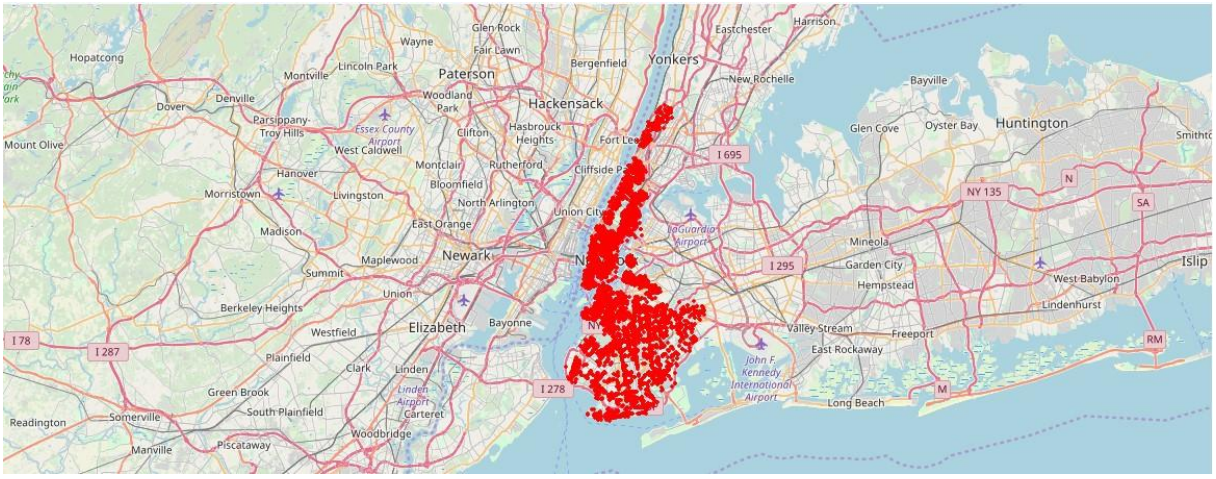
- **Brooklyn -**Most Preferred Food in Brooklyn is –Italian, Purto Rican & Mexican.
- **Manhattan -** Most Preferred Food in Manhattan is – Italian, American, Puerto Rican and Indian.
- **Queens -** Most Preferred Food in Queens is – Indian, Irish, Pakistani and Mexican.
- **The Bronx -** Most Preferred Food in The Bronx is – Italian, Puerto Rican, Albanian and Dominican.
- **Staten island –** Italian, Mexican, Indian, Russian, Sri Lankan , Arab, Pakistani and Polish.

Using Foursquare and New york city neighborhood coordinates we gathered data on venues around New York city.

Below table contains Brooklyn and Manhattan venues:

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
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	Loeser's Delicatessen	40.879242	-73.905471	Sandwich Place

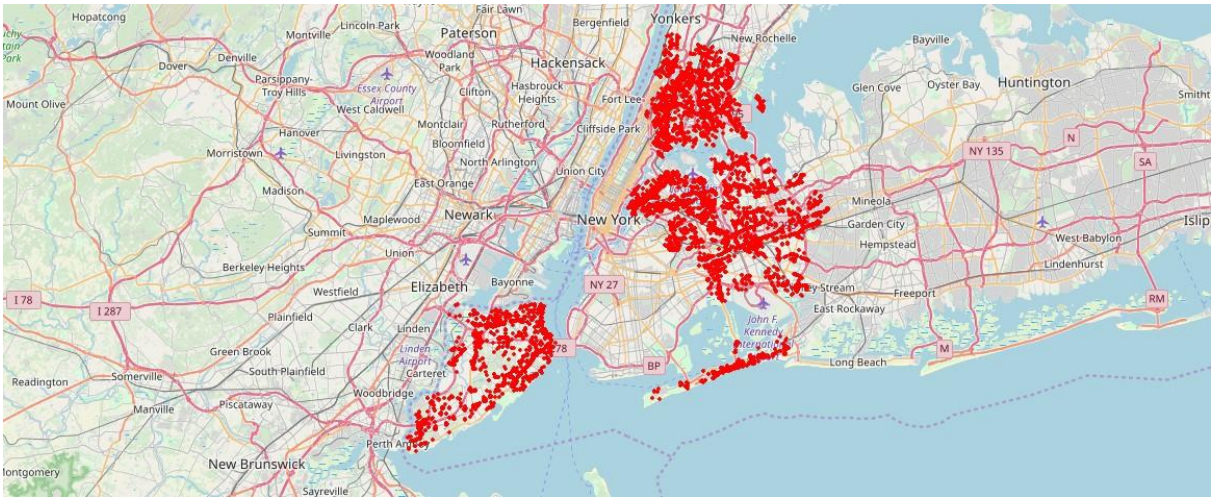
In Screenshot below we can see Brooklyn and Manhattan Venues map. There are in total 9708 venues and 397 unique venue types.



Below table contains Bronx, Queens and Staten Island venues

	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Ali's Roti Shop	40.894036	-73.856935	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy

In Screenshot below we can see Bronx, Queens and Staten Island Venues Map. Bronx, Queens and Staten Island has in total 10805 venues and 387 unique venue types.



RESULTS :

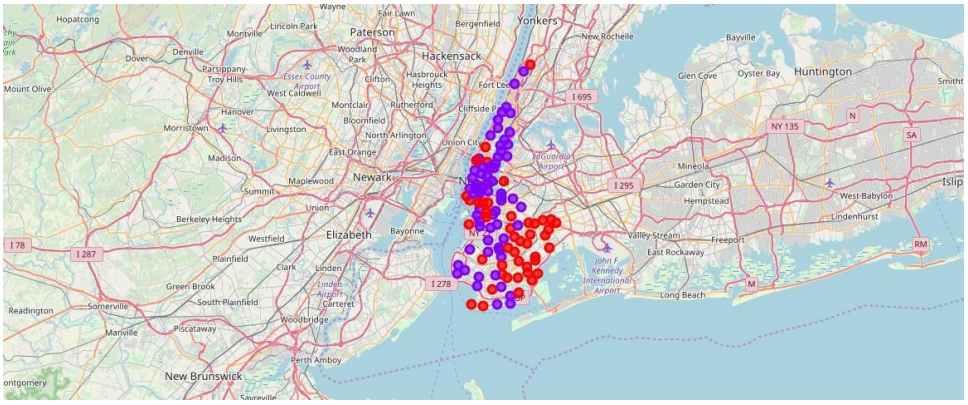
From this venues data we filtered and used only the restaurant data for Brooklyn & Manhattan clustering and Bronx, Queens and Staten Island clustering. As we focussed only on restaurant business.

Neighborhood K-Means clustering based on mean occurrence of venue category :

To cluster the neighborhoods 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.



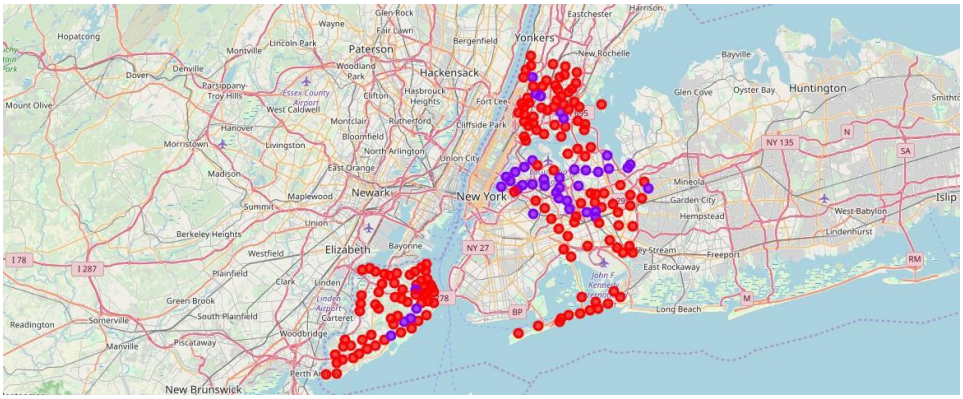
Cluster0: The Total and Total Sum of cluster0 has smallest value. It shows that the market is not saturated.

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 neighborhoods in Brooklyn and Manhattan.

Bronx, Queens and Staten Island:

In the below Map Visualization, we can see the different types of clusters created by using K-Means for Bronx, Queens and Staten Island.



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.

DISCUSSION:

- The market in Brooklyn and Manhattan is highly saturated when compared to Bronx, Queens or Staten island.
- There is scope to explore cuisines of various countries in Bronx, Queens and Staten Island.
- In Manhattan and Brooklyn restaurants of cuisines of many countries are available, so it shows that there is demand for these cuisines.
- Manhattan and Brooklyn is most saturated due to being the most populous and dens areas of New York. Restaurants in this area are in closer proximity to potential customers than in other areas.

CONCLUSION:

The analysis performed in this data science project is very limited in scope and barely scratches the surface of the complexity of entering such a competitive market as restaurant business in New York.
Based on the results, Bronx, Queens and Staten Island are the least saturated markets. The advice would be to gather further data and explore possibility of opening a restaurant in this area as there might be market share still available.
To grab some of the market share from Manhattan or Brooklyn, the advice is to stay near those areas as to be able to also tap into their market when there is a higher demand than supply.

Further analysis is needed to identify which restaurant type and cuisine category would lead to the lowest risk and competition. Perhaps with more analysis and deeper understanding of the situation it would be possible to enter the market in Manhattan or Brooklyn with more robust research.