Capstone Project - The Battle of Neighborhoods

Applied Data Science Capstone by IBM, on Coursera (Last course to get the Data Science Professional Certificate)

Introduction: Business Problem

Aurora Coffee Shop is a Chinese company that was founded in the city of Shanghai in 1992, its particular way of preparing coffee and its own recipes make it unique worldwide. Due to the economic growth it has had in recent times, its shareholders decided to open new shops outside of China.

They consider that the United States (New York City) is a market in which they can succeed.

After several meetings, they have decided to focus their attention on **Brooklyn** since this borough is known for its cultural, social, and ethnic diversity, an independent art scene, distinct neighborhoods, and a distinctive architectural heritage. Another influencing factor is that since 2010, Brooklyn has evolved into a thriving hub of entrepreneurship and high technology startup firms, and of postmodern art and design.

As part of the company's **Data Science team**, I was tasked with recommending the two best areas for setting up their 2 new shops:

1. In the area where the 1st or 2nd most common venue is Coffee Shops
2. In another area where the data science team deems appropriate (borough analysis's results)

Data

Based on definition of our problem, following are the factors addressed:

- Number of existing Coffee Shops in Brooklyn
- Most interested venues

Following data sources will be needed to extract/generate the required information:

- NYC json file (https://cocl.us/new_york_dataset) containing features, names, coordinates, neighborhoods, boroughs and geometric properties of those boroughs within NYC
- Foursquare API to get the most common venues of each neighborhoods in Brooklyn

Data

1- Download all the dependencies (Their corresponding use has been documented within the comments of the code)

2- A) Download and explore the dataset (Using wget function to retrieve JSON File)

B) Transform the data into a pandas dataframe (Filling it with data from NYC json file)

3- A) Use Geopy library to get the latitude and longitude values of New York City

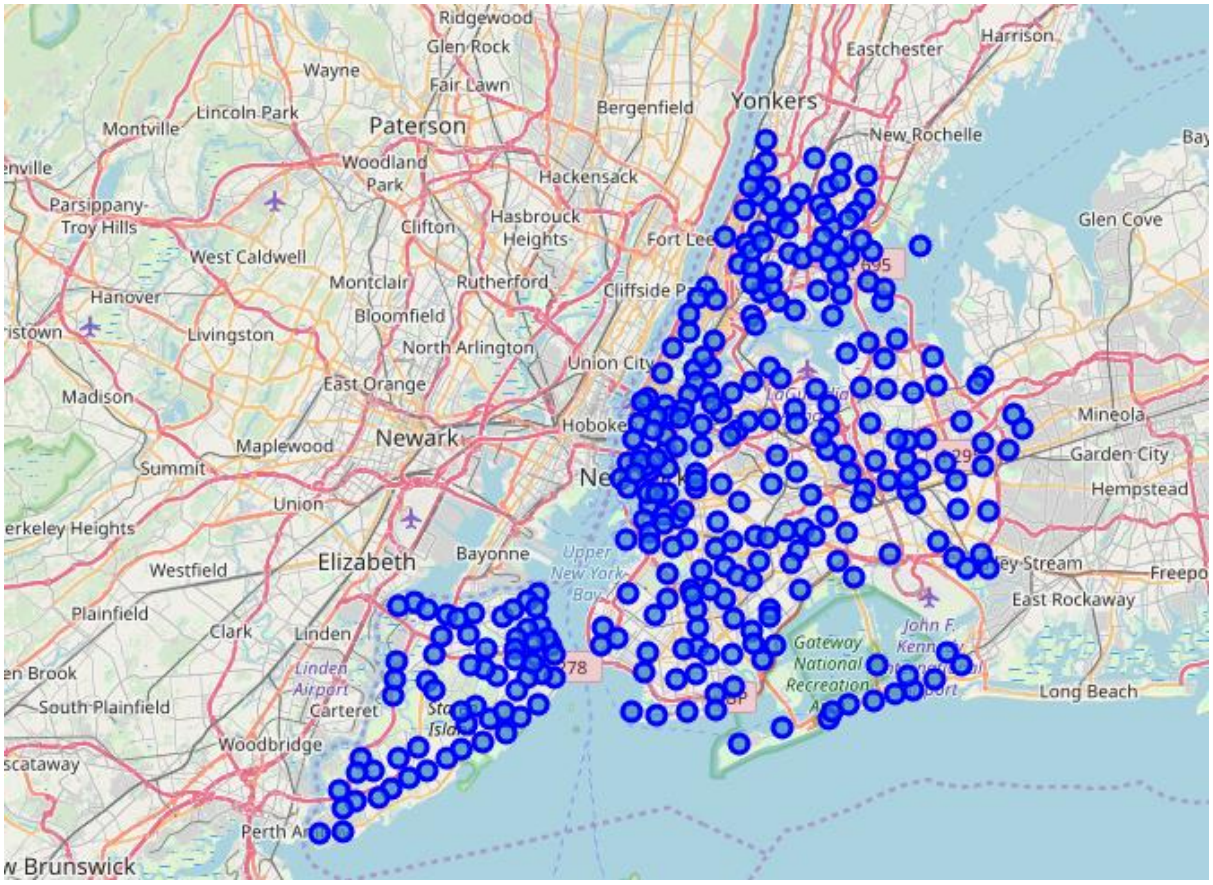
The geographical coordinates for NYC boroughs and neighbourhoods is extracted from the NYC json file.

We must do this step before the generation of a folium map.

The folium library allows for simple implementation of data visualisation via folium map.

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152
```

B) Create a map of New York with neighborhoods superimposed on top



C) The Data Science team wants to segment and cluster only the neighborhoods in Brooklyn.

We sliced the original dataframe and created a new dataframe of the Brooklyn data.

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

D) Create a map of Brooklyn with neighborhoods superimposed on top

4- Foursquare API

Foursquare is a technology company that built a massive dataset of a accurate location data.

Foursquare powers location data for Apple maps, Uber, Snapchat and many others.

Their API and location data are currently being used by over 100.000 developers.

URL: [Foursquare]

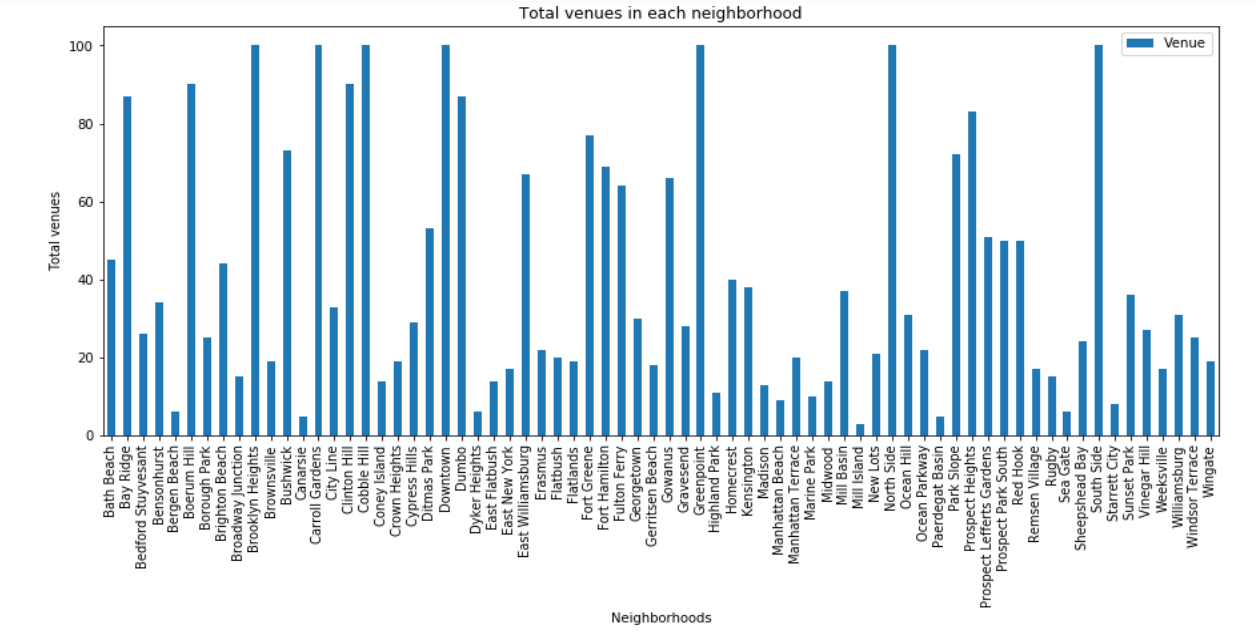
The calls return information that allow us to:

A) Explore all the neighborhoods in Brooklyn (Example: Bay Ridge)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	Ho' Brah Taco Joint	40.622960	-74.031371	Taco Place

B) Get the number of venues per each neighbourhood (Example)

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bath Beach	45	45	45	45	45	45
Bay Ridge	87	87	87	87	87	87
Bedford Stuyvesant	26	26	26	26	26	26
Bensonhurst	34	34	34	34	34	34
Bergen Beach	6	6	6	6	6	6
Boerum Hill	90	90	90	90	90	90
Borough Park	25	25	25	25	25	25
Brighton Beach	44	44	44	44	44	44
Broadway Junction	15	15	15	15	15	15



5- A) Analyze each neighborhood

k-means clustering algorithm only functions with numerical values:

The venue categories are not numerical values. This implies that venue categories need to be converted into numerical values.

Machine Learning applied: One-Hot Encoding, that quantifies categorical data.

B) Print each neighborhood along with the top 5 most common venues

The output is a dataframe containing neighbourhoods and their corresponding degrees of most common venues.

Examples:

```
----Dumbo----                    ----Park Slope----
            venue  freq                        venue  freq
0            Park  0.06        0  American Restaurant  0.06
1     Coffee Shop  0.06        1         Burger Joint  0.06
2  Scenic Lookout  0.05        2          Coffee Shop  0.06
3       Bookstore  0.05        3          Pizza Place  0.04
4            Café  0.05        4   Italian Restaurant  0.04
```

C) We put that into a pandas dataframe (function to sort the venues in descending order)

D) After that we print the Top 10 venues for each neighbourhood

6- A) Clustering neighborhoods (5 clusters)

K-means-Clustering:

Introduction to Clustering (Customer Segmentation): It is the practice of partitioning a customer base into groups of individuals that have similar characteristics. It can group data only unsupervised, based on the similarity of customers to each other.

What is a cluster?

A group of objects that are similar to other objects in the cluster and dissimilar to data points in other clusters.

Introduction to K- means:

It is a type of partitioning clustering. It divides the data into non-overlapping subsets (clusters) without any cluster-internal structure.

B) We should create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood

C) Visualize the clusters:

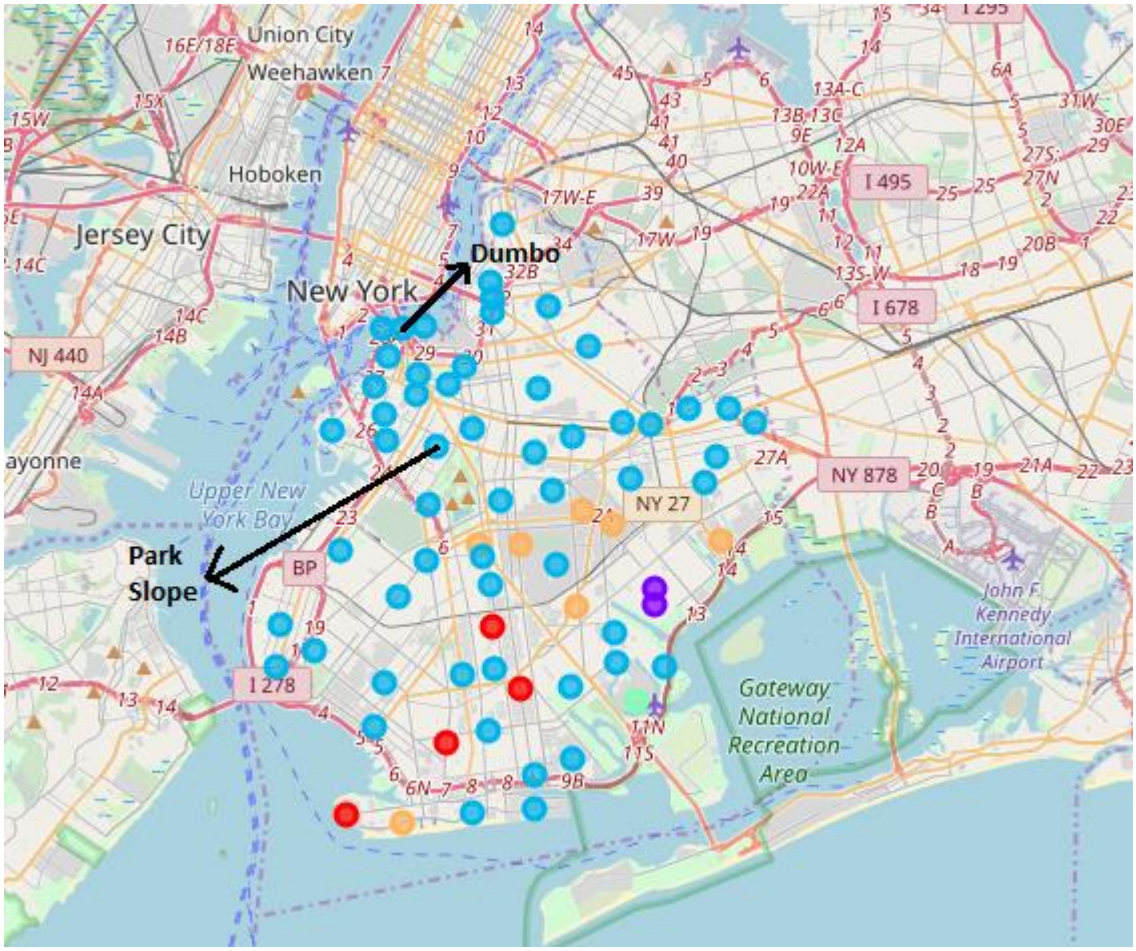
Cluster 1: Red

Cluster 2: Violet

Cluster 3: Blue

Cluster 4: Green

Cluster 5: Orange



7- Examine Clusters

This is an example of two of the neighborhoods of the cluster number 3 which contains in total 56 neighborhoods:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dumbo	Coffee Shop	Park	Bookstore	Scenic Lookout	Café	Ice Cream Shop	Bakery	Yoga Studio	Food Court	Dog Run
Park Slope	American Restaurant	Burger Joint	Coffee Shop	Italian Restaurant	Pizza Place	Falafel Restaurant	Caribbean Restaurant	Furniture / Home Store	Pet Store	Bakery

Results

After analyzing the dataset using different functions and techniques, some important data could be discovered in order to make the final decision about the two places to set up the new shops.

NCY is divided into 5 borough: Manhattan, Bronx, Queens, Staten Island and Brooklyn and 306 neighborhoods.

Brooklyn has 70 neighborhoods.

After analyzing Venues per each neighborhood, it turns out that there are 287 unique categories.

We believed it was very important to know what the neighborhoods with more venues were and the result was the following: Brooklyn Heights, Caroll Gardens, Cobble Hill, Downtown, Greenpoint, North Side and South Side.

Printing each neighborhood along with the top 5 most common venues: We focused our attention on neighborhoods that have coffee shops or similar points of interest as well as attractions where people are attracted to drink coffee (Take Away).

Taking into account the Top 10 venues for each neighborhood we decided to apply clustering (5 clusters) where we could clearly see that cluster 3 is where the largest number of neighborhoods are centred in which the 1st or 2nd common venue is a Coffee Shop or similar. For this reason, from there, we began to decide which of these neighborhoods would be the elect.

The decision will be made by combining the information obtained.

Discussion

Analyzing Cluster number 3 and since the shareholders of the company want to set up one of their new stores in a central location (surrounded by many competitors) we consider that **DUMBO** is a good option:

----Dumbo----

venue freq

0 Park 0.06 (people tend to drink coffee take away or after walking around the park prefer to spend time on a coffee shop)

1 Coffee Shop 0.06

2 Scenic Lookout 0.05 (Visitors from all over the world prefer this location while they are taking a stroll around Brooklyn)

3 Bookstore 0.05

4 Café 0.05

Visualizing the map of Brooklyn with clusters superimposed on top, we can see that DUMBO is very close to the two main bridges that connect the Borough with Manhattan (Brooklyn Bridge and Manhattan Bridge) so it is a key point for the entire borough.

We consider important that Dumbo is not among the neighborhoods with more venues since people usually prefer to avoid areas where a lot of shops and stores in the same place. For this same reason we believe it is necessary that the second place respect that principle.

PARK SLOPE is the neighborhood that recommends the data science team to set up the company's second shop.

First of all, the most important factor is that it is located in the same area as Prospect Park, a place visited by tourists throughout the year. At the same time we saw that it is very full of restaurants and that people tend to go to eat in those places, so a coffee after lunch is an excellent option for the customers. Also it could be seen that the third common venue is a Coffee Shop (Culture of coffee in the area).

----Park Slope----

	venue	freq
0	American Restaurant	0.06
1	Burger Joint	0.06
2	Coffee Shop	0.06
3	Pizza Place	0.04
4	Italian Restaurant	0.04

The Data Science team considers these two neighborhoods as the best to set up the first two coffees in the US, but it strongly recommends that after a trial period an analysis of the borough Manhattan begins to be aimed at the labour market in the area.

It is key for the success of this venture to design the shops differently since the target audience is completely different from that of Asia, so we already contacted the marketing and commercial department to analyze this point and start working on the topic.

Conclusion

The goal of this project was to determine the two best places to set up the two new shops in the US, specifically in Brooklyn. After analyzing the information received by applying the work methodology, the decision was made to choose two neighborhoods taking into account different factors detailed above.

As a member of the Data Science team, I considered necessary to gather more information from the two neighborhoods selected, so we had a meeting with members of human resources and the department of institutional relations, and the result was as follows:

DUMBO (Down Under the Manhattan Bridge Overpass) is the neighborhood in northwest Brooklyn with the best views of Manhattan. Home to Jane's Carousel and Brooklyn Bridge Park as well as classic cobblestone streets, and right near historic and undiscovered Vinegar Hill, DUMBO is a classic destination that is also home to many celebrities in the Clocktower apartment building. DUMBO is home to many wonderful restaurants, such as the original Grimaldi's pizzeria. One Girl Cookies is a must-visit down by the water before you go for a stroll through the park and catch breathtaking views of Manhattan. In the summer, movies are shown along the waterfront. Nearby neighborhoods: Brooklyn Heights, Cobble Hill, Downtown Brooklyn

Park Slope is located on the hill leading up to Prospect Park, the largest park in Brooklyn. (Prospect Park was created by Frederick Olmsted and Calvert Vaux, who also created Central Park in Manhattan) Park Slope has been listed as one of the most desirable neighborhoods to live in because of its quiet streets, good restaurants, good schools, and proximity to the Brooklyn Museum, Botanical Gardens, Prospect Park, access to public transit, and more. Park Slope is known for its historic brownstones and flickering gas-powered lamps in front of them, and the tree-lined streets offer available free parking, and young families and professionals wander the thin historic sidewalks.

The Data Science team of a company always must work together with all the departments of the organization having to handle variables from different areas that is sometimes better managed by another department.

Author: Nicolás Sacco

September 15, 2019