CAPSTONE PROJECT

NEIGHBORHOOD SELECTION FOR BUSINESS EXPANSION

PROBLEM STATEMENT

Over a period, every successful business may reach saturation in its current region/market. Hence, in order to ensure the continuous growth of the company, it becomes prudent to expand to other regions to gain more businesses.

Currently, we have our office in San Francisco (Address: *353 Sacramento St, San Francisco, CA 94111, United States*) and a new office is required to be set up in New York City.

The target audience of this report would be anyone who wants to identify suitable neighbourhoods in a new city or in the same city to open a new office. This can be extended to opening of multiple offices in multiple cities across the world.

DATA

We have utilised the following data to decide on a neighbourhood for our new office:

- **Neighbourhood Data for New York City:**

  https://www.baruch.cuny.edu/nycdata/population-geography/neighborhoods.htm

- **Foursquare Data:**

  Using foursquare data, we can get the attributes/features of the neighbourhoods both for our current office neighbourhood in San Francisco and of all the neighbourhoods in New York City. These neighbourhoods will be then compared and based on how similar the neighbourhoods are to that of the current one, we can identify the new potential neighbourhoods in NYC.

METHODOLOGY

Data Preparation:

We extracted the NYC neighbourhoods from the given link using Beautiful Soup library, added Latitude and Latitude to our dataset and also added the our current office address to get to the below DataFrame:

	Borough	Neighbourhood	Latitude	Longitude
0	Brooklyn	Bath Beach	40.60185	-74.000501
1	Brooklyn	Bay Ridge	40.633993	-74.014584
2	Brooklyn	Bedford Stuyvesant	40.683436	-73.941249
3	Brooklyn	Bensonhurst	40.604977	-73.993406
4	Brooklyn	Bergen Beach	40.620382	-73.906803
5	Brooklyn	Boerum Hill	40.685626	-73.984171
6	Brooklyn	Borough Park	40.633993	-73.996806
7	Brooklyn	Brighton Beach	40.579644	-73.961111
8	Brooklyn	Broadway Junction	40.679192	-73.903354
9	Brooklyn	Brooklyn Heights	40.696085	-73.995028
10	Brooklyn	Brownsville	40.667236	-73.906798

Data Modelling:

- Before modelling we need to provide basis on which the model can separate neighbourhoods from each other. We defined the characteristics of neighbourhoods in terms of nearby locations/venues. Using Foursquare API, we can get the nearby venues as per our requirement. For our modelling purpose, we took all venues within the radius of 500m from our neighbourhood (with a limit of 100).

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Brooklyn	Bath Beach	40.60185	-74.000501	King's Kitchen	40.603844	-73.996960	Cantonese Restaurant
1	Brooklyn	Bath Beach	40.60185	-74.000501	Lutzina Bar&Lounge	40.600807	-74.000578	Hookah Bar
2	Brooklyn	Bath Beach	40.60185	-74.000501	Lenny's Pizza	40.604908	-73.998713	Pizza Place
3	Brooklyn	Bath Beach	40.60185	-74.000501	Planet Fitness	40.604567	-73.997861	Gym / Fitness Center
4	Brooklyn	Bath Beach	40.60185	-74.000501	Grotta Azzurra	40.603611	-73.995381	Pizza Place

- The unique 439 venues category from Foursquare API were converted to categorical variables, and then consolidated neighbourhood wise to get a mean score for each of the venue categories.

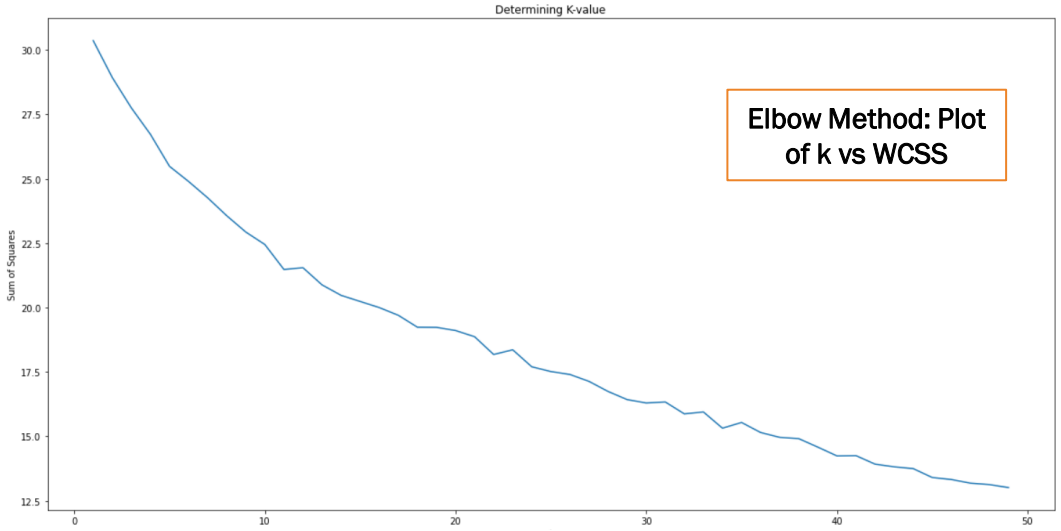
	Borough	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Food Court	Airport Lounge	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Ar R
0	Bronx	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
1	Bronx	Bathgate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
2	Bronx	Baychester	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
3	Bronx	Bedford Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
4	Bronx	Belmont	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.017857	0.0	0.0	0.0	0.0	0.0	0.0	

METHODOLOGY

- We used the top 10 venues (basis frequency) for our further analysis.

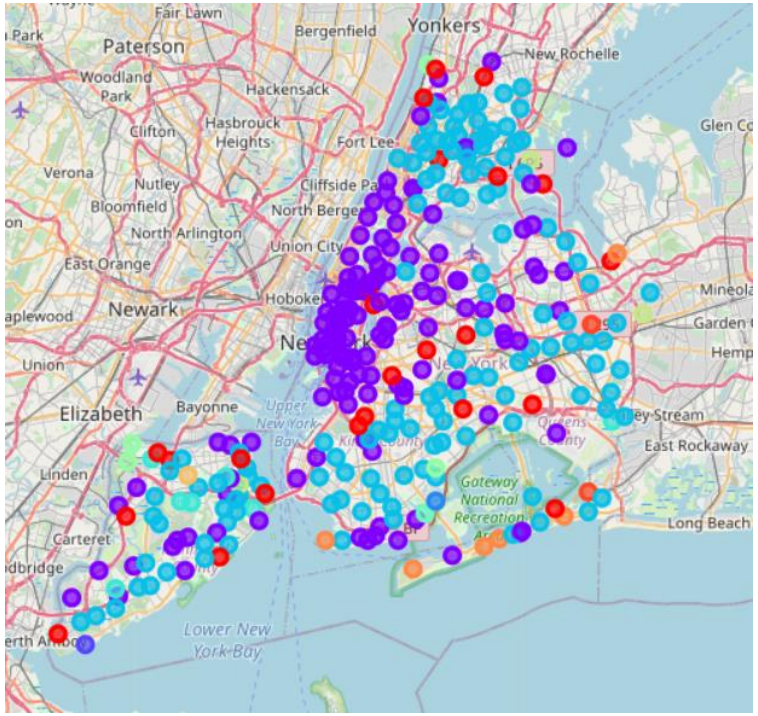
	Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	Allerton	Bakery	Sandwich Place	Discount Store	Donut Shop	Pharmacy	Fast Food Restaurant	Pizza Place	Trail	Bank	Bar
1	Bronx	Bathgate	Restaurant	Bus Station	Donut Shop	Supermarket	Spanish Restaurant	Pharmacy	Pizza Place	Lounge	School	Sandwich Place
2	Bronx	Baychester	Pharmacy	Print Shop	Deli / Bodega	Donut Shop	Pizza Place	Historic Site	Playground	Chinese Restaurant	Grocery Store	Construction & Landscaping
3	Bronx	Bedford Park	Chinese Restaurant	Diner	Mexican Restaurant	Pizza Place	Deli / Bodega	Sandwich Place	Fried Chicken Joint	Baseball Field	Train Station	Grocery Store
4	Bronx	Belmont	Italian Restaurant	Pizza Place	Bakery	Deli / Bodega	Dessert Shop	Fish Market	Food & Drink Shop	Donut Shop	Pharmacy	Chinese Restaurant

- To create clusters, we have used K-means clustering algorithm. In order to define K, we applied the Elbow Method optimizing the WCSS value.



Elbow Method: Plot of k vs WCSS

- Initially, we got 135 neighbourhoods that matched with our current office location, but to narrow down our options, we applied K means clustering to those 135 neighbourhoods and with a k value of 75 ww were able to shortlist 5 neighbourhoods best suited for our requirements.



Visualizing the clusters on NYC Map

RESULTS

After the second round of clustering with 135 neighbourhoods, we got potential 5 neighbourhoods where the new office can be opened in New York as follows:

- East Williamsburg, Brooklyn
- Financial District, Manhattan
- Greenwich Village, Manhattan
- Murray Hill, Manhattan
- Wall Street, Manhattan

Out of these locations, **East Williamsburg, Brooklyn is the most like our neighbourhood (identified by using k=100)**. It is observed that the most common venue is Coffee Shop and restaurants which seems to be plausible as there usually it has been observed that there are many food establishments near offices as they provide good business.

	Borough	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
82	Brooklyn	East Williamsburg	Coffee Shop	Bakery	Gym	Pizza Place	Shoe Store	Italian Restaurant	Ice Cream Shop	Bar	Bookstore	French Restaurant
130	Financial District, SF	Financial District	Coffee Shop	Gym	Japanese Restaurant	Men's Store	Food Truck	Sushi Restaurant	New American Restaurant	Park	Restaurant	Cocktail Bar
140	Manhattan	Financial District	Coffee Shop	American Restaurant	Italian Restaurant	Gym	Steakhouse	Café	Falafel Restaurant	Hotel	Pizza Place	Sandwich Place
143	Manhattan	Greenwich Village	Coffee Shop	American Restaurant	Italian Restaurant	Wine Bar	Sandwich Place	Yoga Studio	Cosmetics Shop	Bar	Food Truck	Steakhouse
158	Manhattan	Murray Hill	Hotel	Japanese Restaurant	Coffee Shop	Sandwich Place	American Restaurant	Bar	Gym / Fitness Center	Gym	Cocktail Bar	Bakery
172	Manhattan	Wall Street	Coffee Shop	American Restaurant	Gym	Cocktail Bar	Salad Place	Mexican Restaurant	Steakhouse	Italian Restaurant	Café	Bar

DISCUSSIONS AND CONCLUSIONS

Even though the model has provided us with a list of potential neighbourhoods, it still can improve. Some of the scenarios which can be incorporated are as follows:

- If a particular feature is required to emphasized or removed, then the same can be achieved by tweaking the values of feature set of our current neighbourhood. The tweaked neighbourhood can be clustered with New York neighbourhood to identify a desired neighbourhood.
- If any feature is not important then that feature can be dropped from the dataset.
- One of issues encountered while modelling was that it was very difficult to find the optimal k-value for cluttering. The plot between the WCSS and k didn't clearly show any elbow, so I have projected and backtracked the graph to an approximate k-value of 12.

Finally, using k-means clustering method and foursquare API for extracting venues, I was able to identify potential neighbourhoods for setting up a new office in New York city. This model can be easily extended to any number of cities and can also include specific requirements for neighbourhood. This model is not specific to offices itself, it may be applied to any office, restaurants, chains, workshops, factories etc. The possibilities are infinite, we just to be careful in understanding the limitations of the model and not totally rely on it. A user's domain knowledge and data science can go together and complement each other, thus producing significant results.