CAPSTONE PROJECT

Neighbourhood selection for business expansion

Nikhil Sahu
[Email address]

Table of Contents

Business Idea
Neighbourhood selection for a new office in a new city to expand the business.

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.

Expansion to a new region may not be as easy as it seems. There may be various challenges that a company may face while selecting an appropriate neighbourhood. Issues such as cost of the neighbourhood, distance from the business centre of the city, transportation, etc. may become an especially important factor in deciding the fate of the new business.

One way to mitigate this issue is to open the new office in a neighbourhood which is similar to the neighbourhood of our existing office. 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
To identify the best neighbourhood to open a new office we will need the following data:

1. Neighbourhood Data for New York City.
   https://www.baruch.cuny.edu/nycdata/population-geography/neighborhoods.htm

2. 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 clustered based on these features/attributes. Based on how similar the neighbourhoods are to that of the current one, we can identify the new neighbourhoods which can be our potential location for setting up a new office in NYC.

Methodology
Data Preparation
For Foursquare
The neighbourhood data from the link was extracted using the Beautiful soup and then added to dataframe. Latitude and longitude for each of the neighbourhood was extracted and added to the dataframe using Geolocation from Geopy library. We removed the neighbourhoods for which the latitude/latitude was not able available. The final processed dataset is as follows:

	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

The dataset consists of 315 neighbourhoods of New York.

We then added the data for our office in San Francisco to this dataframe and the final dataset now contains 316 rows.

312	Staten Island	Ward Hill	40.632918	-74.082918
313	Staten Island	Westerleigh	40.621215	-74.131809
314	Staten Island	Willowbrook	40.60316	-74.138476
315	Staten Island	Woodrow	40.543439	-74.197644
316	Financial District, SF	Financial District	37.794163	-122.399263

For Modelling

Before modelling we need to provide basis on which the model can separate neighbourhoods from each other. This can be based purely on the location, populations etc.

In our case, we need to define 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). The requests returned a total 10636 results a sample of which has been indicated below.

	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

One of the venue categories was “Neighborhood” which would have created issues during the modelling phase, hence in all the venues with venues category defined as “Neighbourhood” has been removed, thus the final unique venues count is 439.

For modelling, the venues category is now converted to categorical variables by creating dummy variables. This dataset is now consolidated neighbourhood wise to get a mean score for each of the venue categories.

Now the dataset looks as the following:

	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	

After dropping Borough and Neighbourhood column, the dataset is ready for modelling.

The top 10 venue category (basis frequency) for each neighbourhood has been indicated below:

	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

Modelling

One way to identify a similar neighbourhood in New York city is to cluster neighbourhoods in New York and based on features identify which cluster is similar to the current neighbourhood.

In order to do so, we have already added the San Francisco neighbourhood in the New York neighbourhoods dataset, so if cluster, our current neighbourhood will form cluster with the neigbourhoods in New York.

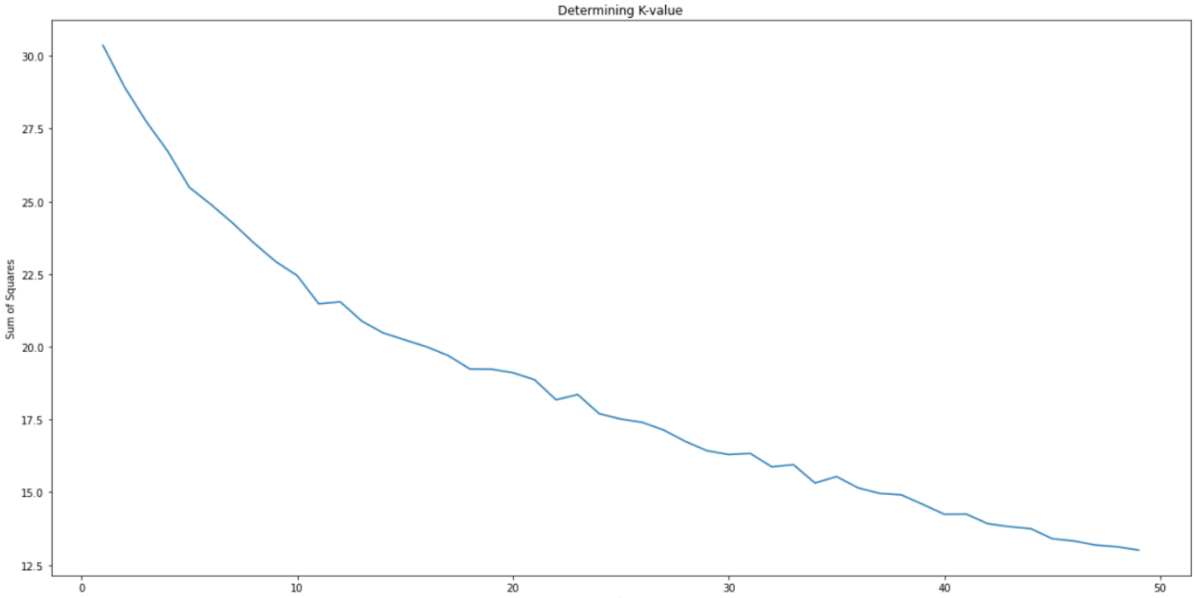
In order to cluster, we have used K-means clustering as it is relatively simple to implement, can handle large datasets, guarantees convergence and don't require us to determine the starting centroids of clusters. However, in our modelling, we need to define the K or how many clusters are to be formed post which the model separates the neighbourhoods into those many clusters using Euclidean distance/similarity amongst the neighbourhoods.

In order to determine, k value, the model was with k-values starting from 1 to 50 and WCSS (within clusters squared sum) is used as a determinant for identify the optimal K value. The WCSS is plotted against the k value and wherever a kink is formed, that is taken as the k value.

The WCSS for the 50 runs are indicated below:

```
[30.364083499853304, 28.931798333796863, 27.75231308829518, 26.726744971794094, 25.484556259485668, 24.89657642972803, 24.26292924708035, 23.5671033770635, 22.9335438747115
5, 22.44873164669437, 21.480422088473745, 21.54730028075086, 20.88115453859288, 20.47409786712654, 20.237149913341852, 20.000177551173145, 19.700349568583995, 19.2386245824
14733, 19.231948040113704, 19.108944141074318, 18.868866629138864, 18.178032886482573, 18.36085718394965, 17.700229201689023, 17.516415613888444, 17.402276241067074, 17.134
540354146985, 16.743163498759355, 16.426871902704324, 16.29628425134416, 16.335172865278142, 15.872679229180273, 15.947543958719317, 15.31646618726706, 15.539046432135004,
15.149368511351636, 14.96101933009104, 14.91097384282348, 14.582025915063507, 14.24124022112134, 14.25176029118952, 13.921388361413932, 13.816747418871069, 13.748171591214
486, 13.402435524438738, 13.324471626291345, 13.182979332746914, 13.126669548634807, 13.011137393386925]
```

The plot of WCSS vs K is as below:

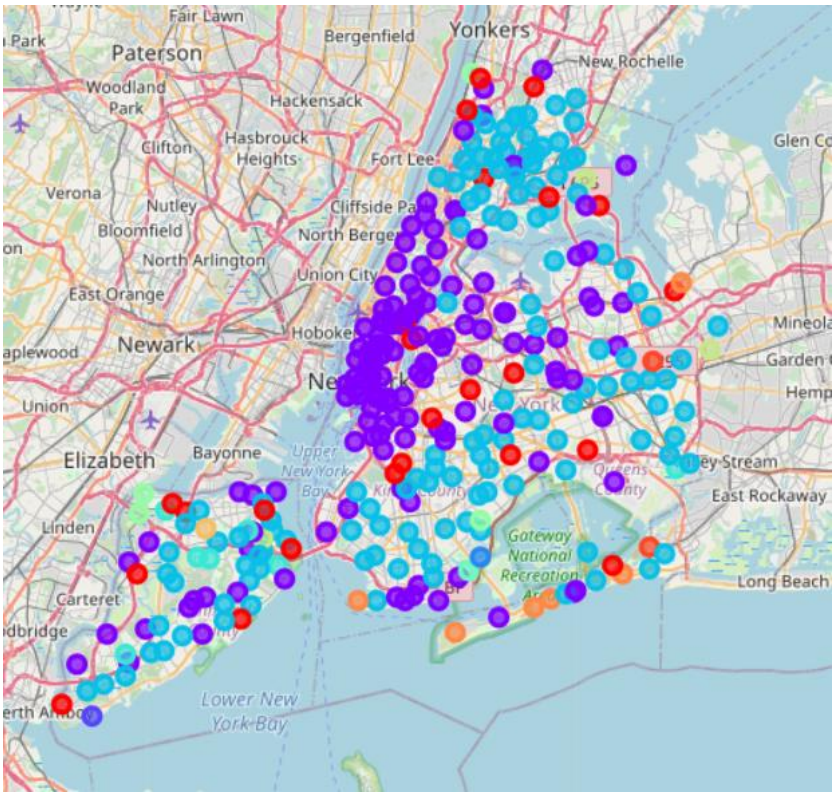


From the above plot, we can observe that there is kink formation at K =12 which is taken as the final number of clusters that can be formed.

After inserting the Cluster label into the dataset, the dataset is as follows:

	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	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	Brooklyn	Bath Beach	40.60185	-74.000501	4	Bank	Pizza Place	Chinese Restaurant	Supplement Shop	Japanese Restaurant	Italian Restaurant	Hookah Bar	Middle Eastern Restaurant	Tea Room	Cantonese Restaurant
1	Brooklyn	Bay Ridge	40.633993	-74.014584	1	Chinese Restaurant	Dessert Shop	Seafood Restaurant	Noodle House	Bar	Malay Restaurant	Hotpot Restaurant	Tennis Court	Park	Vietnamese Restaurant
2	Brooklyn	Bedford Stuyvesant	40.683436	-73.941249	1	Coffee Shop	Café	Pizza Place	Bar	Deli / Bodega	Playground	Cosmetics Shop	Southern / Soul Food Restaurant	Fried Chicken Joint	Cocktail Bar
3	Brooklyn	Bensonhurst	40.604977	-73.993406	4	Chinese Restaurant	Bank	Bakery	Pizza Place	Mobile Phone Shop	Cantonese Restaurant	Japanese Restaurant	Supplement Shop	Fast Food Restaurant	Shoe Store
4	Brooklyn	Bergen Beach	40.620382	-73.906803	4	Japanese Restaurant	Liquor Store	Donut Shop	Plaza	Pizza Place	Chinese Restaurant	Peruvian Restaurant	Food	Bus Station	Supermarket

Visualizing the clusters on the map:



Our office is found to be clustered in cluster label 1 along with 135 other neighbourhoods. On a preliminary level, it is possible to recommend that all these 135 neighbourhoods can be a potential location for opening a neighbourhood.

However, a further analysis is done to narrow down potential locations. The method adopted for this purpose, is to consider Cluster 1 and then recluster them into very small cluster sizes so that the number of neighbourhoods within our new cluster is very small. By using high number of clusters (k) we are indirectly shortlisting the neighbourhoods with minimum distance or in order words, neighbourhoods which are the most similar to our target neighbourhood.

With k=75, we are able to shortlist 5 neighbourhoods from New York City which are the most similar to our neighbourhood. The following table lists those 5 New York neighbourhoods.

	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

Results

While clustering our office was found to be in cluster 1 along with another 135 neighbourhoods. Neighbourhoods which are most similar to our neighbourhood was identified using k-means clustering with very high values of k (as a proxy of Euclidean Distance). Based on the following are the key results:

- Potential 5 neighbourhoods where the new office can be opened in New York are:
    - 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.

Discussions

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 any 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.
- The techniques applied in this model is pertaining to only New York city, but the same can be applied to any number of cities in the world.

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.

Conclusion

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.

The k-means clustering method, though easy to implement, has a few limitations. For example, identifying the optimal number of clusters is not an exact science. We will have to put our domain knowledge to make better sense of the output that we get from the model. Also, we need to understand the importance of the input variables as it may lead to overfitting of the model and thus can give inaccurate results. To identify any overlapping variables domain knowledge is required and an in-depth understanding of the variables is also required. Further, data analytics can help us understand any underlying connection between the input variables.