IBM Advanced Data Science Capstone Final Project

Week 5 Report

Finding Best Neighbourhood to Start New Office in New York City

Project by: - Nanduri V P S Anirudh

INTRODUCTION

Background

The current situation in USA and rather the entire world is in such a way that even if someone says that the whole world is in a standstill, no one would disagree. However, one good thing about this time is that a lot of people have some free time on their hands where they are contemplating what can they do to better themselves. This also includes many people who will probably upskill themselves and would be looking for a better job when the situation around the world settles a little. It also includes quite a few businessmen, and new entrepreneurs who either would like to set up a new company or expand an existing company. Anyone looking to establish a decent sized organization or is thinking of expanding it would most likely contact real estate agents to figure out an ideal location for their new office. Real estate agents even if experienced in the field would still need to back up their claims for best suitable option for their clients with some real data. In this capstone project, the aim is to help such real estate agents to figure out the types of neighbourhoods and then recommend them neighbourhoods best suited for their client's needs.

Business Problem

The main business problem is to help real estate agents suggest their client's ideal or close to ideal locations/places in accordance to their requirements. Some of the questions which can be solved after going through the analysis are "suggest a good place to start my new office", "suggest me a place to move in near my office", etc.

Target Audience

The main audience for the analysis are the real estate agents in and around New York City, NY who cater to many clients. This analysis would help them understand behaviour of neighbourhoods in and around NYC with data to back them up which would significantly increase the trust their clients will have on them.

This analysis can also be used by people who are looking to shift their location. It would give them an idea of which neighbourhoods to target to look for a house.

DATA

Data Description

The data needed for this analysis would be information of neighbourhoods in New York City, NY with information of their borough and latitude and longitude. We would also be needing information of venues in the neighbourhood and which type of venues they are.

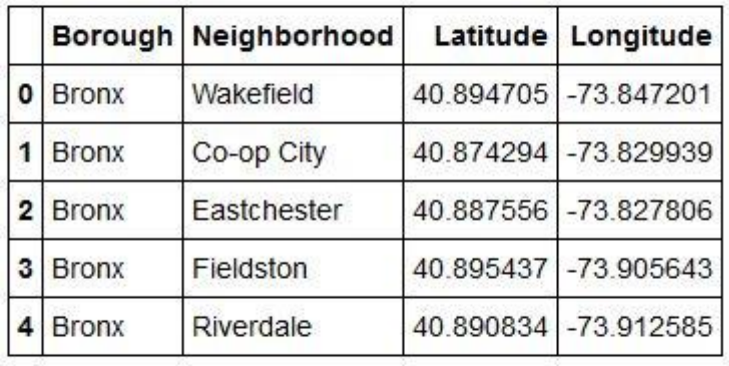
Data Sources

There are mainly two data required as stated above, one is data containing borough, neighbourhood, latitude, and longitude which is taken from json file available here and the information regarding venues will be extracted using Foursquare API while passing required inputs into API.

METHODOLOGY

Initial Data Analysis and Data Preparation

After reading the json file from here, the data is extracted and is converted into a pandas data frame so as to easily read and manipulate the data as needed. It is found that there are 5 boroughs and 306 neighbourhoods present in the data. The structure of the subset of the data frame can be found below in Table 1.



	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

Table 1. Neighbourhoods geo location along with its borough

Utilising the geographic co-ordinates available in the data and folium package in python, the neighbourhoods in New York City are visualized as shown in Image 1.

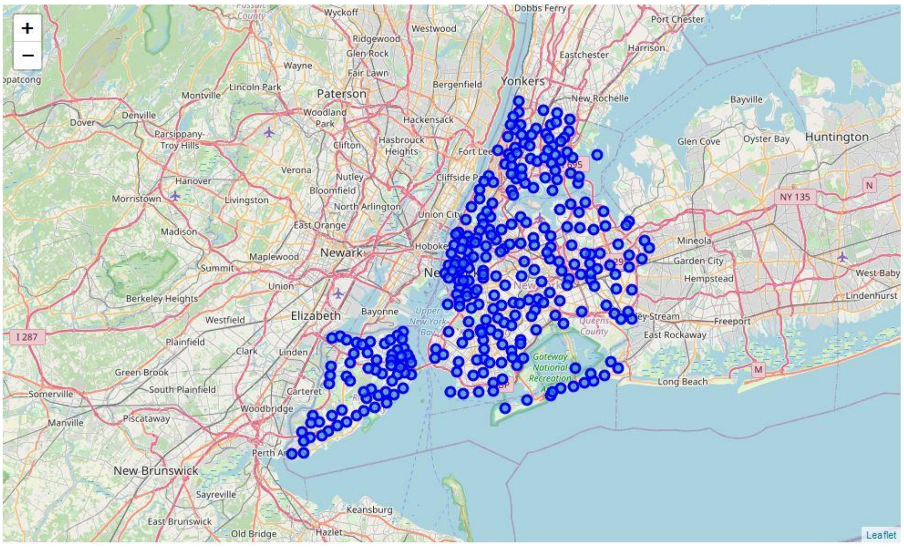
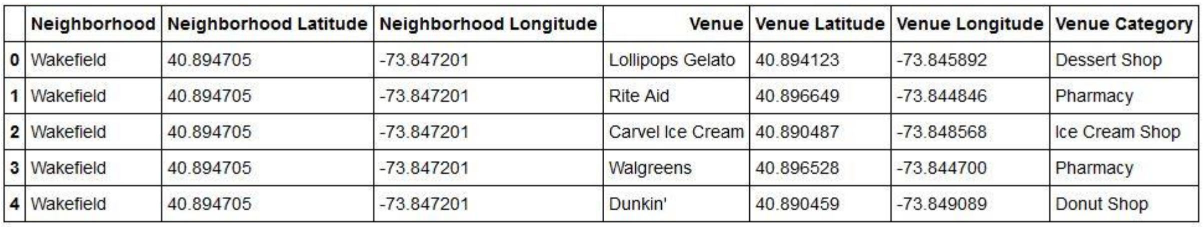


Image 1. Available locations of all neighbourhoods in New York City, NY

Now, using Foursquare API, information for a single neighbourhood is extracted in order to understand how the data is retrieved and how it can be worked upon. After understanding how to extract required information from the API call which consist of “Venue Name”, “Venue Category”, “Latitude”, and “Longitude” the same process is applied to all neighbourhoods under all 5 boroughs. Then, the information is extracted and is placed into a data frame to be worked on later. The subset of the data can be found below in Table 2.



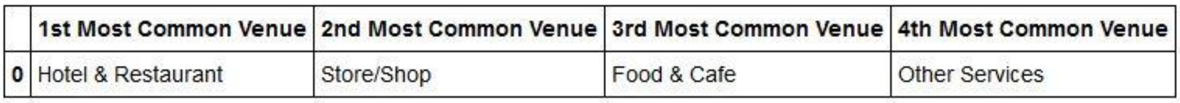
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Table 2. Table with all venues along with their category and geographic location

It might be possible that there might be a few neighbourhoods which do not return any information from API call. In the current case, it is found that information for 2 neighbourhoods were not fetched and they are Port Ivory and Howland Hook.

Feature Creation

From main data of Table 2, we can observe that there are many categories and when we try to work on these as is, we might not be able to get any useful information out of it. Hence, bucketing of “Venue Category” is done which will be called as “Venue Category Group” henceforth. The bucketing was done based on understanding of “Venue Category”. For example, “Venue Category” like Gym, Volleyball court, Indoor play area, comedy club are classified under “Fun, Fitness, & Relaxation”. All together from 421 “Venue Category” items, 14 “Venue Category Group” items were obtained. Here, using one-hot encoding method, information in “Venue Category Group” was taken from rows to columns and features were created. Information for each neighbourhood were grouped at these features. A simple analysis of most common values shows that in most of the neighbourhoods, Hotel & Restaurant are most frequent where as Store/Shop are the next frequent. This information of most common venue category can be observed in Table 3.

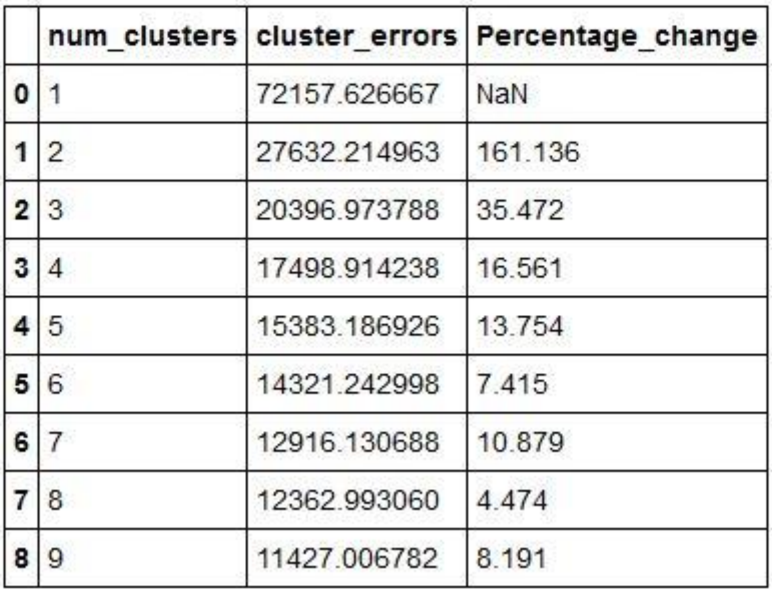


	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Hotel & Restaurant	Store/Shop	Food & Cafe	Other Services

Table 3. Most common venues in all neighbourhoods

Machine Learning Algorithm

The algorithm used in this project is K-Means algorithm which clusters the observations based on the distance between them. This particular algorithm was used to segment the observations as the intention here is to classify different neighbourhoods into number of clusters and understand general behaviour of each cluster. K-Means in particular was used mainly because of two reason which are its simplicity in implementation and its history of better results when compared to some more complicated algorithms. However, one input needed to run this algorithm is the number of clusters. For this, we can run the K-Means with multiple number of clusters and then extract the sum of squared distance measurement from each iteration which is given below in Image 2. The percentage change in sum of squared distance is also calculated to aid in finalizing the number of clusters which is given below as Table 4.



	num_clusters	cluster_errors	Percentage_change
0	1	72157.626667	NaN
1	2	27632.214963	161.136
2	3	20396.973788	35.472
3	4	17498.914238	16.561
4	5	15383.186926	13.754
5	6	14321.242998	7.415
6	7	12916.130688	10.879
7	8	12362.993060	4.474
8	9	11427.006782	8.191

Table 4. Number of clusters along with their sum of squared distance (error) and error change percentage

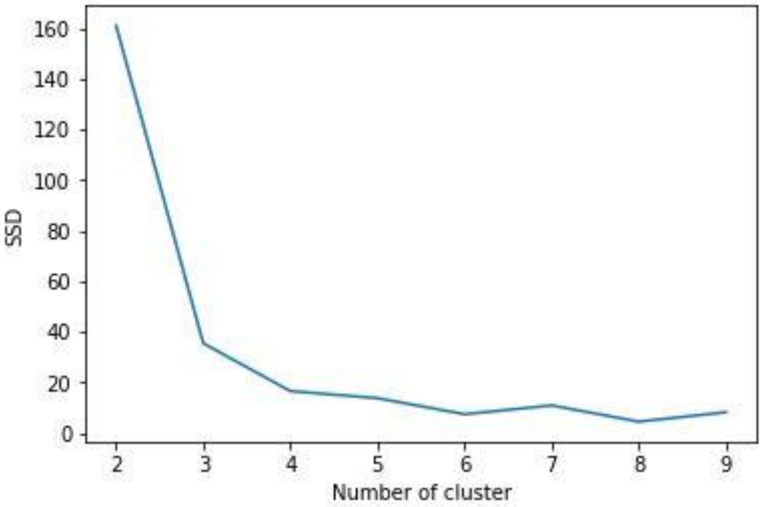


Image 2. Number of cluster vs their errors

From both the above table and image, it is determined that the optimal number of clusters for the analysis is 4 because the error percentage change from clusters 3 to 4 is significant while the error percentage change from clusters 4 to 5 is insignificant. It also confirms with the elbow graph as it has a bend at 4 cluster point.

RESULTS

The result of the clustering shows that all neighbourhoods are classified as one of 4 clusters. This is given in Image 3 which visually depicts the neighbourhoods and their assigned clusters.

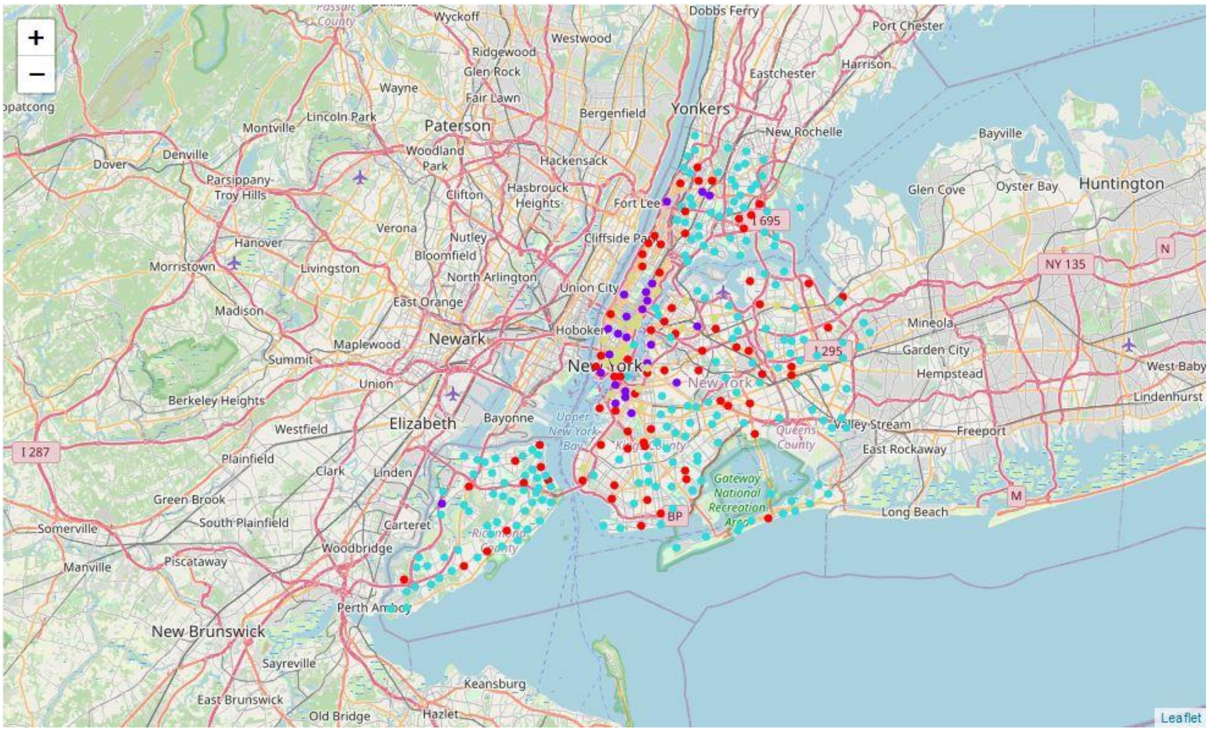
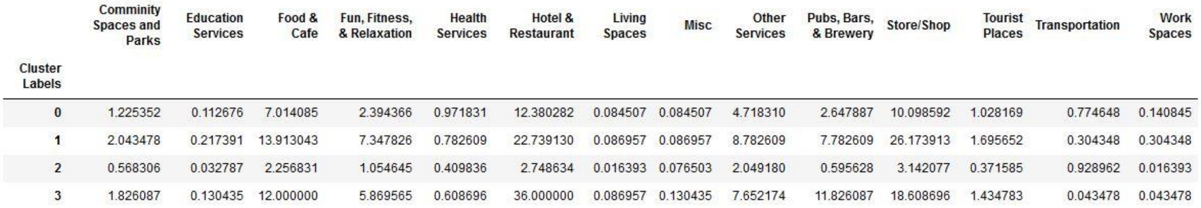


Image 3. New York City with its clustered neighbourhoods

DISCUSSIONS

The intended output of this project which is to understand the behaviour of each cluster can be derived from Table 5 below.



Cluster Labels	Comminity Spaces and Parks	Education Services	Food & Cafe	Fun, Fitness, & Relaxation	Health Services	Hotel & Restaurant	Living Spaces	Misc	Other Services	Pubs, Bars, & Brewery	Store/Shop	Tourist Places	Transportation	Work Spaces
0	1.225352	0.112676	7.014085	2.394366	0.971831	12.380282	0.084507	0.084507	4.718310	2.647887	10.098592	1.028169	0.774648	0.140845
1	2.043478	0.217391	13.913043	7.347826	0.782609	22.739130	0.086957	0.086957	8.782609	7.782609	26.173913	1.695652	0.304348	0.304348
2	0.568306	0.032787	2.256831	1.054645	0.409836	2.748634	0.016393	0.076503	2.049180	0.595628	3.142077	0.371585	0.928962	0.016393
3	1.826087	0.130435	12.000000	5.869565	0.608696	36.000000	0.086957	0.130435	7.652174	11.826087	18.608696	1.434783	0.043478	0.043478

Table 5. Average behaviour of each cluster

As we can see in the above table, neighbourhoods in cluster 0 have good health services, transportation, and work spaces near them. Could be considered as a decent place to live in for people who do not care that much about fun and relaxation.

Neighbourhoods in cluster 1 have most services being offered as well as most educational places when compared to other clusters while having a good number of work spaces, hotels, and parks. This would definitely be a good candidate for the customers who can spend a little for their accommodation as the neighbourhoods in this cluster are tend to be expensive considering that all factors are available here. This place is also good for people who want to expand business as this place clearly has a good footfall.

Neighbourhoods in cluster 2 seem to be behind all other neighbourhoods except for the fact that it has a good transportation facility. It is probably a place which can be developed later and might be a good place for future investment or real estate buys focused on generating rents.

Neighbourhoods in cluster 3 have the most hotels, shops, and pubs. This place is probably more ideal for setting up shops rather than living

FUTURE SCOPE

Future scope for the project can be to include the real estate pricing data along with the already existing feature set to create new clusters to see which neighbourhoods are highly priced with all facilities available and how do neighbourhoods which are relatively cheap compare to highly priced ones. Another way this can be achieved is by doing a regression model which will quantify the effect of each feature onto the real estate price. This will greatly help real estate agents in finding best neighbourhood based on the client’s choice and budget.

CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. real estate agents regarding the best neighbourhoods to reside in or start a new office space. To answer the business question that was raised in the introduction section, the answer proposed by this project is: neighbourhoods in cluster 1 are best candidates for both setting up a new office as well as living. Only probable down side in cluster 1 would be possible high rates as most of services and requirements are available. The next best place to set up office is in neighbourhoods of cluster 3. The second-best place to live is in neighbourhoods of cluster 0. It is highly likely that the prices here aren’t that high as was the case in neighbourhoods of cluster 1.