Project Report

The Battle of Neighborhoods

Author: Samuel Bisaso

Table of Contents

Introduction:

Background

New York is said to be the most populous city in the United States, with a population of 8,398,748 as of 2018, distributed over a land area of about 302.6 square miles. And the New York metropolitan Areas has a population of about 18,593,22 People according to the world Atlas website, making New York (NY) also the most densely populated major city in the United States, Located at the southern tip of the state of New York, which is also the centre of the New York metropolitan area, also said to be the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities.

New York is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. It has exerted a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. It is Situated on one of the world's largest natural harbours, New York City consists of five boroughs, each of which is a separate county of the State of New York. And its these five boroughs Brooklyn, Queens, Manhattan, The Bronx, and Staten Island that were consolidated into a single city in 1898.

Brief history about New York: During the precolonial era, New York City was inhabited by Algonquian Native Americans, including the Lenape. Their homeland, known as Lenapehoking, included Staten Island, Manhattan, the Bronx, the western portion of Long Island (including the areas that would later become the boroughs of Brooklyn and Queens), and the Lower Hudson Valley. It is said that New York became the most populous urbanized area in the world in the early 1920s, overtaking London. The metropolitan area surpassed the 10 million mark in the early 1930s, becoming the first megacity in human history. Since then people have travelled through and inhabited the New York area. The currently receives a lot of tourists and thousands of immigrants look at it as their number one destination in United states.

Problem Description

Lets say you live one part of New York and would like to relocate or trying to find a new job in an area which has a neighbourhood similar or closely the same as your current neighbourhood, mainly providing you with your favourite amenities and other types of venues that exist in the neighbourhood, such as gourmet fast food joints, pharmacies, parks, graduate schools. Searching on the internet may not be enough to give you a clear suggestion whether these locations are similar or not. In a situation where you have received a very nice job for which you have to relocate if at all you have accepted the job offer, then finding information about these areas might be difficult. There is a need to have a best alternative which has already done the analysis based to the venues and locations of the neighbourhoods to provide you with a more precise and more reliable suggestions.

Say you live on the west side of the city of Toronto in Canada. You love your neighbourhood, mainly because of all the great amenities and other types of venues that exist in the neighbourhood, such as gourmet fast food joints, pharmacies, parks, graduate schools and so on. Now say you receive a job offer from a great company on the other side of the city with great career prospects. However, given the far distance from your current place you unfortunately must move if you decide to accept the offer.

Wouldn't it be great if you are able to determine neighbourhoods on the other side of the city that are the same as your current neighbourhood, and if not perhaps similar neighbourhoods that are at least closer to your new job?

Objective

The aim of this report is to explore, study and analyse the neighbourhoods of New York city and group them into similar clusters and, to analyse those clusters to gather meaningful information. This information could be used in finding out best neighbourhoods that are the same, similar or even better than your current neighbourhood for your own preferences.

Target Audience

This report provides useful information to people wishing to relocate to New York city who are in search for new neighbourhoods proved to be similar to their current neighbourhood or even more better depending on their likes and preferences

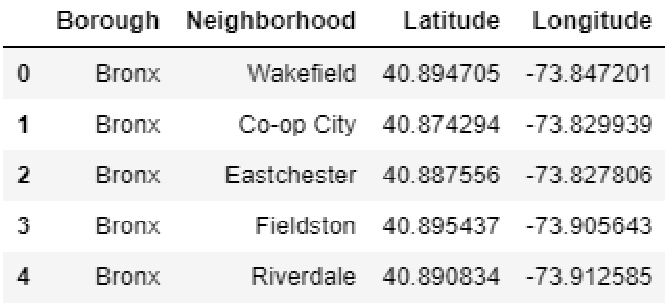
Data Description

To consider the objective of this study, the New York Neighbourhood has a total of 5 boroughs and 306 neighbourhoods. And in order to segment the neighbourhoods and explore them, essentially need a dataset that also contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood. This dataset exists for free on the web. And the link to the dataset is here: https://geo.nyu.edu/catalog/nyu_2451_34572 . The information obtained was then transformed into a panda’s data frame for further analysis.

Methodology:

Download Data and gathering the data into a Pandas data frame

To start with our analysis, data was downloaded, coded and transformed into a pandas data frame, here it was essential to transform the data of nested python dictionaries into pandas data frame by looping through the data, was able to fill the data frame one row at a time. The resulting data frame was examined to make sure the dataset contains all the 5 boroughs and 306 neighbourhoods.

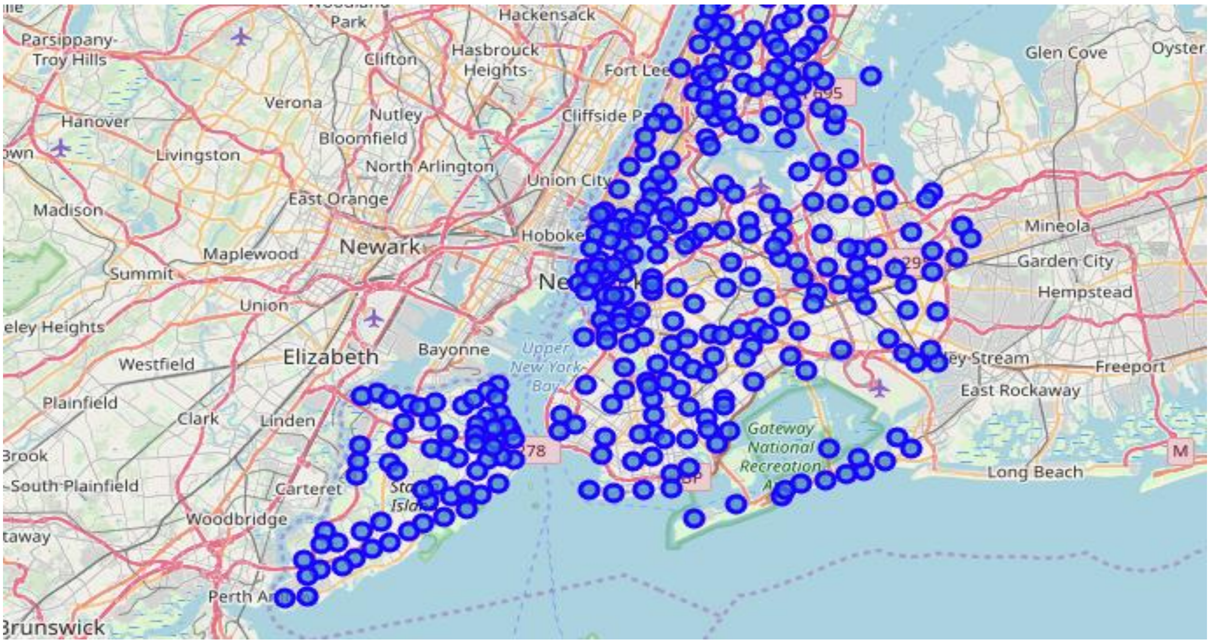


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Generating a map of New York with plotted Neighbourhood data on it

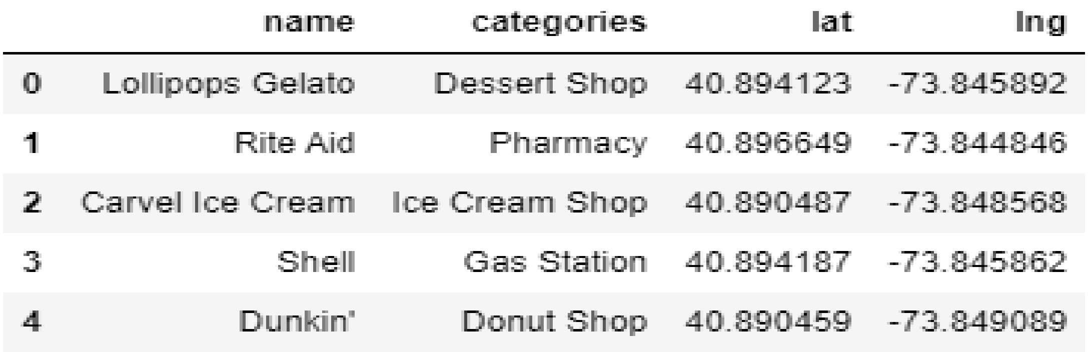
Used geopy library to get the latitude and longitude values of New York City in preparation to plot the Neighbourhood on it.

We then use the python **folium** library to visualize geographic details of New York city and its boroughs. Created a map of New York with boroughs superimposed on top using the latitude and longitude values to get the visual as below:



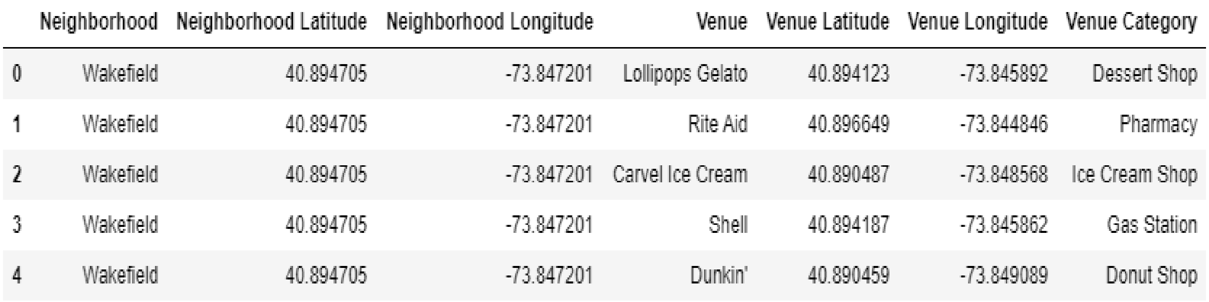
Utilizing Foursquare API to explore the Neighbourhoods

To next, started utilizing the Foursquare API to explore the neighbourhoods and then create segments out of these. The LIMIT parameter was set to **100**, which would limit the number of venues returned by the Foursquare API and the radius of 500 meter. Below is a visual of the list of Nearby Venues for the first neighborhood.



	name	categories	lat	lng
0	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Rite Aid	Pharmacy	40.896649	-73.844846
2	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568
3	Shell	Gas Station	40.894187	-73.845862
4	Dunkin'	Donut Shop	40.890459	-73.849089

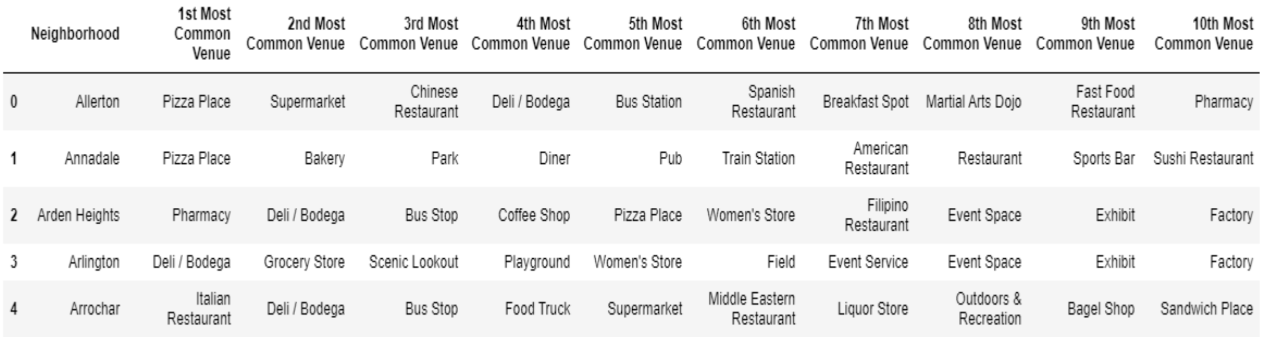
A new function is the created that will repeat the process above for all the neighbourhoods in New York. With this function, it give us a list of all venues present in New York city. Here is a table showing the first five values of this data frame.



	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	Shell	40.894187	-73.845862	Gas Station
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Analyzing each of the neighborhoods

By using a One Hot Encoding, the neighborhood is used to group data, and find out the top ten venues present in each neighborhood.



	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	Allerton	Pizza Place	Supermarket	Chinese Restaurant	Deli / Bodega	Bus Station	Spanish Restaurant	Breakfast Spot	Martial Arts Dojo	Fast Food Restaurant	Pharmacy
1	Annadale	Pizza Place	Bakery	Park	Diner	Pub	Train Station	American Restaurant	Restaurant	Sports Bar	Sushi Restaurant
2	Arden Heights	Pharmacy	Deli / Bodega	Bus Stop	Coffee Shop	Pizza Place	Women's Store	Filipino Restaurant	Event Space	Exhibit	Factory
3	Arlington	Deli / Bodega	Grocery Store	Scenic Lookout	Playground	Women's Store	Field	Event Service	Event Space	Exhibit	Factory
4	Arrochar	Italian Restaurant	Deli / Bodega	Bus Stop	Food Truck	Supermarket	Middle Eastern Restaurant	Liquor Store	Outdoors & Recreation	Bagel Shop	Sandwich Place

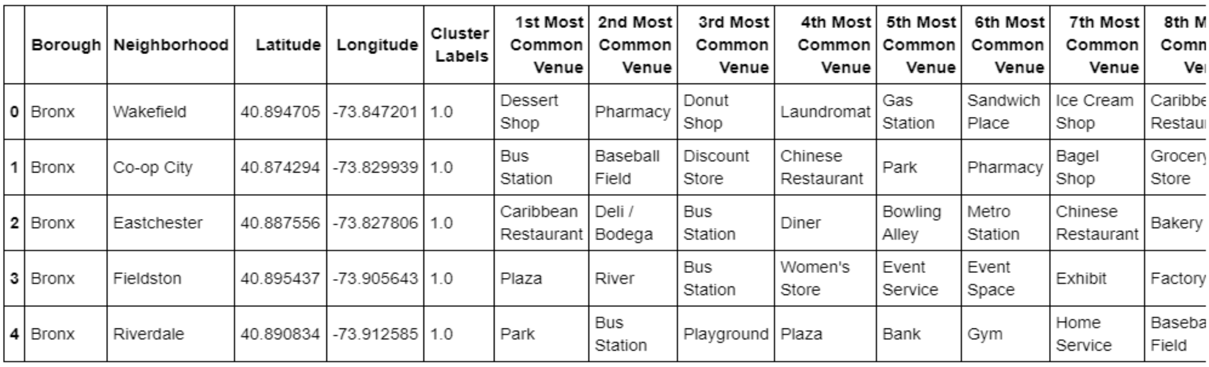
Then found out that there are some common venue categories in the neighbourhoods. Which prompted to then use the unsupervised learning K-means algorithm to cluster the neighbourhoods. clearly noting that K-Means algorithm is one of the most common method for clustering in unsupervised learning.

We use a K-**cluster** value of 11 to split the neighbourhoods into 11 different clusters based on the similarity they have concerning the venues they contain.

Results:

Adding the Cluster Labels to the Venue Data

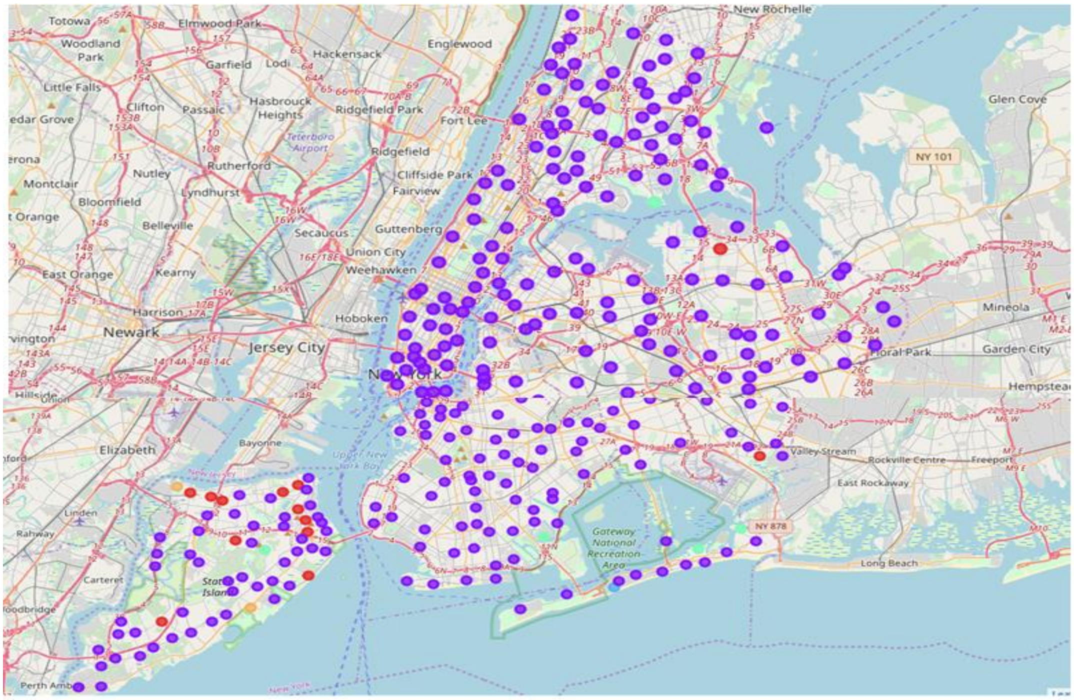
The table below portrays the clustered data and also showing the top 10 most common venues



	Borough	Neighborhood	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 M Comn Ve
0	Bronx	Wakefield	40.894705	-73.847201	1.0	Dessert Shop	Pharmacy	Donut Shop	Laundromat	Gas Station	Sandwich Place	Ice Cream Shop	Caribbe Restau
1	Bronx	Co-op City	40.874294	-73.829939	1.0	Bus Station	Baseball Field	Discount Store	Chinese Restaurant	Park	Pharmacy	Bagel Shop	Grocery Store
2	Bronx	Eastchester	40.887556	-73.827806	1.0	Caribbean Restaurant	Deli / Bodega	Bus Station	Diner	Bowling Alley	Metro Station	Chinese Restaurant	Bakery
3	Bronx	Fieldston	40.895437	-73.905643	1.0	Plaza	River	Bus Station	Women's Store	Event Service	Event Space	Exhibit	Factory
4	Bronx	Riverdale	40.890834	-73.912585	1.0	Park	Bus Station	Playground	Plaza	Bank	Gym	Home Service	Baseba Field

Visualisation of the resulting Clusters

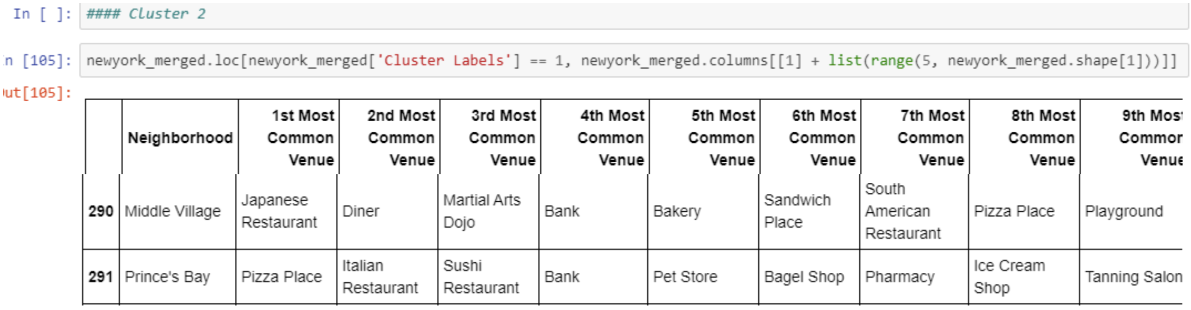
Matplotlib and folium packages were used to visualize the clusters on a map of New York.



Discussion:

The analysis was carried out with an intent to find out similar neighbourhoods for a person relocating within the city of New York.

As we analyse the results section, we can analyse the clusters and see similar neighbourhoods in different parts of the city. For example, if we compare the different neighbourhoods clustered in cluster 2.



```
In [ ]: #### Cluster 2

n [105]: newyork_merged.loc[newyork_merged['Cluster Labels'] == 1, newyork_merged.columns[[1] + list(range(5, newyork_merged.shape[1]))]]

ut[105]:
```

	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 Mos Commor Venue
290	Middle Village	Japanese Restaurant	Diner	Martial Arts Dojo	Bank	Bakery	Sandwich Place	South American Restaurant	Pizza Place	Playground
291	Prince's Bay	Pizza Place	Italian Restaurant	Sushi Restaurant	Bank	Pet Store	Bagel Shop	Pharmacy	Ice Cream Shop	Tanning Salon

As seen in the table above, if someone wished to move from a suburb in Queens to Staten island. If a person’s current location were in the Neighbourhood of Middle village in Queens, which has venues like Banks, bakeries and restaurants nearby, the person, would like to relocate to a neighbourhood like Prince Bay in Staten island which also has venues like Banks and Restaurants. This could just be one of the numerous examples of how our data analysis can help people relocate from one part of the city to another which like their current localities.

The five resultant Clusters could be named as follows

- I: Neighborhoods that have around Bus lines, financial and legal services and Italian Resturants.
- II: Neighborhoods that have around pharmacies, restourants and banks.
- III: Neighborhoods that have around the beach, women's store and fish and chips shops
- IV: Neighborhood that have around Parks, restaurants and event spaces.
- V: Neighborhoods that have around Bars and event service

Conclusion:

Nevertheless, location data can be used to solve a lot of societal problems by providing better solution that can help the people make informed decisions, be it starting a new business, visiting a new place by finding interesting venues or relocate within a city. Like seen in the example provided in this report, data was used to cluster neighbourhoods in New York based on the most common venues in those neighbourhoods. In the same manner, data can also be used to solve other problems, which most people face in bigger cities.

References:

Data source: https://geo.nyu.edu/catalog/nyu_2451_34572

https://www.worldatlas.com/citypops.htm

Foursquare API: https://developer.foursquare.com/docs/api/

Link to the Notebook

https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/04358e92-c19c-496c-86cb-5768ec7448d1/view?access_token=bb85c16151043f893f2da5654a8d85bda020a2a880ac57d0d7e64fea3b400965