IBM Data Science Capstone Project Report

Problem Description:

We want to create a map that recommends similar neighbourhoods to people in New York city based on where they currently live in case they need to move.

New York City is one of the most unaffordable cities on earth. Owning a home in the city is nothing but a far beyond dream for the majority of people in the city. The average homeownership rate of the city is just 33% (as of 2018). Because of this most people rent, which makes it more likely for them to have to move at some point in their lives. I would like to make this process easier by implementing a recomender system which tells the person moving which neighbourhoods are similar to the one where he lives based on data about venues nearby.

Data:

- For this project we are going to need data about the neighbourhoods in NY, which can be found here:
    - https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

- We are also going to be using data from the FoursQuare API in order to gather venue data for each one of the neighbourhoods.

- In addition to this we will use google's Map API to geocode our neighbourhoods into Latitude and Longitude coordinates for the use of the FourSquare API.

Methodology:

- First, I webscraped the wikipedia page to get a table of neighbourhoods in New York city.
- I realized that there were too many neighbourhoods in New York City, which made my analysis hard. This is because if the neighbourhoods are too close together then when i call the foursquare API for venues then there will be an overlap between neighbourhoods and then the clustering algorithm that i was going to use wouldn't give as meaningfull results.
    - Therefore i decided to use the New York Community boards as a proxy for neighbourhoods since there is less of them but they should still be just as representative of the underlying neighbourhoods without the venue overlap.

- I Webscraped all of the neighbourhoods into a dataframe:

	Neighbourhoods
0	Bronx CB 1
1	Bronx CB 2
2	Bronx CB 3
3	Bronx CB 4
4	Bronx CB 5

  -
- Then for each one of those neighbourhoods i GeoCoded its latitude and longitude using Googles Map API and appended it to the dataframe:

	Neighbourhoods	Latitude	Longitude
0	Bronx CB 1	40.819666	-73.913235
1	Bronx CB 2	40.821118	-73.892096
2	Bronx CB 3	40.833152	-73.896798
3	Bronx CB 4	40.843377	-73.910068
4	Bronx CB 5	40.856771	-73.910510

  -
- After this i used the Foursquare API to fetch the venues for each of these places, i appended them all onto a big list and then filtered it for duplicates in order to get a list of all possible venues:

```
1 print("There are "+str(len(set(get_unique_list(total_list))))+" unique venue categories in NY")
```

-
```
There are 288 unique venue categories in NY
```

- Using this information, i went neighbourhood by neighbourhood adding the ammount of venues of each category that can be found in that specific neighbourhood. For example Bronx CB 1 has 3 bike shops and 2 cafes, while Manhattan CB 5 has 3 banks, one sushi restaurant and 3 pediatricians.
  - I appended this information to my dataframe to get the following result:

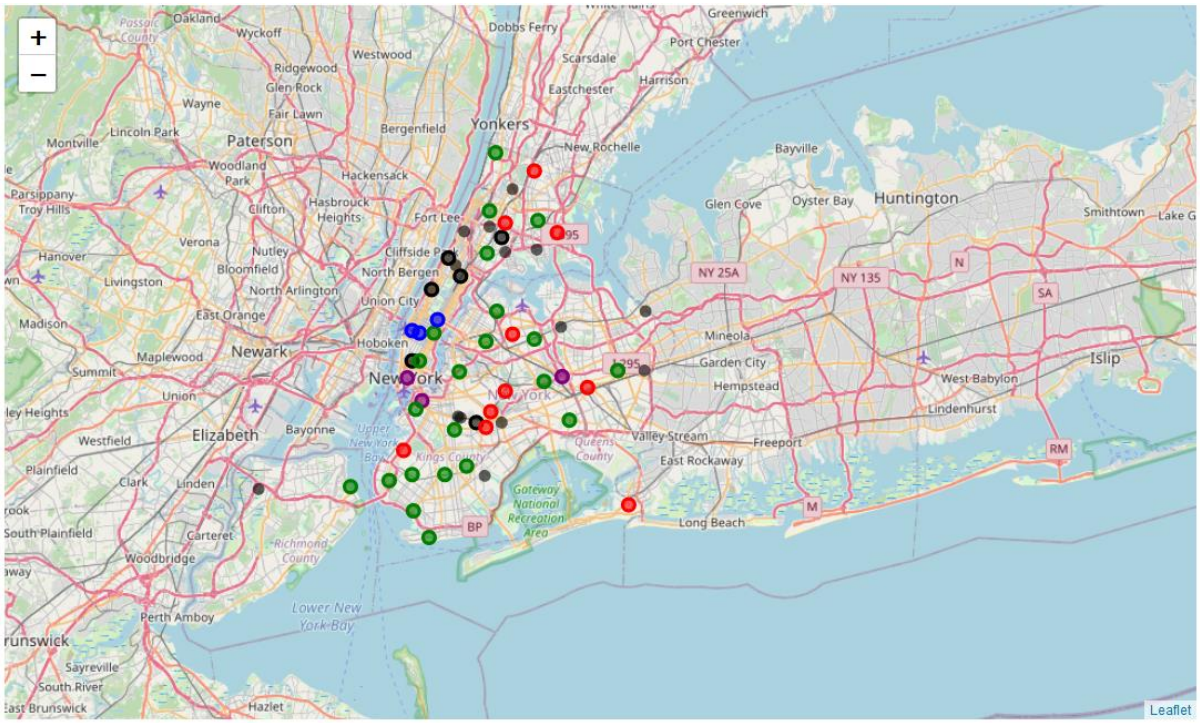
	Neighbourhoods	Latitude	Longitude	Sushi Restaurant	Eye Doctor	Japanese Curry Restaurant	Church	Theme Park Ride / Attraction	Veterinarian	Business Center	...	Chocolate Shop	Climbing Gym	Clothing Store	Insura O
0	Bronx CB 1	40.819666	-73.913235	0	0	0	0	0	0	0	...	0	0	0	
1	Bronx CB 2	40.821118	-73.892096	0	0	0	0	0	0	0	...	0	0	0	
2	Bronx CB 3	40.833152	-73.896798	0	0	0	2	0	0	0	...	0	0	0	
3	Bronx CB 4	40.843377	-73.910068	0	1	0	0	0	0	0	...	0	0	0	
4	Bronx CB 5	40.856771	-73.910510	0	0	0	0	0	0	0	...	0	0	0	
5	Bronx CB 6	40.845830	-73.892824	0	0	0	1	0	0	0	...	0	0	0	
6	Bronx CB 7	40.874687	-73.885022	0	1	0	0	0	0	0	...	0	0	0	
7	Bronx CB 8	40.906457	-73.903927	0	0	0	0	0	0	0	...	0	0	0	
8	Bronx CB 9	40.823446	-73.857068	0	0	0	0	0	0	0	...	0	0	2	
9	Bronx CB 10	40.837449	-73.834185	0	0	0	0	0	0	0	...	0	0	0	
10	Bronx CB 11	40.847732	-73.856198	0	0	0	0	0	0	0	...	0	0	0	
11	Bronx CB 12	40.890900	-73.859152	0	0	0	0	0	0	0	...	0	0	0	
12	Brooklyn CB 1	40.718159	-73.945254	0	0	0	0	0	0	0	...	0	0	0	
13	Brooklyn CB 2	40.693526	-73.987648	0	0	0	0	0	0	0	...	0	0	0	

-
- After doing this i deployed a K means clustering algorithm to cluster neighbourhoods based on their similarity in venues.

- After deploying the model i appended each neighbourhood label at the end of the dataframe and then used that information to plot the map using the Folium Library:

Results:

Here is the final map output of my program:



Discussion:

After having lived in New York myself for about 5 years i can confidently say that the algorithm does a surprisingly good job at finding similar neighbourhoods.

Conclusion:

I believe this has been a usefull and meaningfull problem which i believed allowed me to used the tools which we learned for in the course. Im pretty happy with the results. I believe the final result is of value and useful for the intended purpose.