CAPSTONE PROJECT

INTRODUCTION

The problem I pretend to solve is very particular but can be extremely important for me. Since I´m moving to Manhattan I need to decide where exactly I should move, considering the next features:

1. I´m mexican, so I want to live somewhere with a lot of Mexican restarants.
2. I enjoy working out, so I must be able to find gyms nearby.
3. I like walking in Parks.
4. I really love Italian Food.
5. I´m interested on doing Yoga.

DATA ACQUISITION AND CLEANING

To success in my research I will need to extract data regarding Manhattan from the Neighborhoods and Boroughs in New York.

This data can be obtained from this website: https://cocl.us/new_york_dataset, and for the analysis we´ll require information coming from Foursquare to look for relevant information according to our goals.

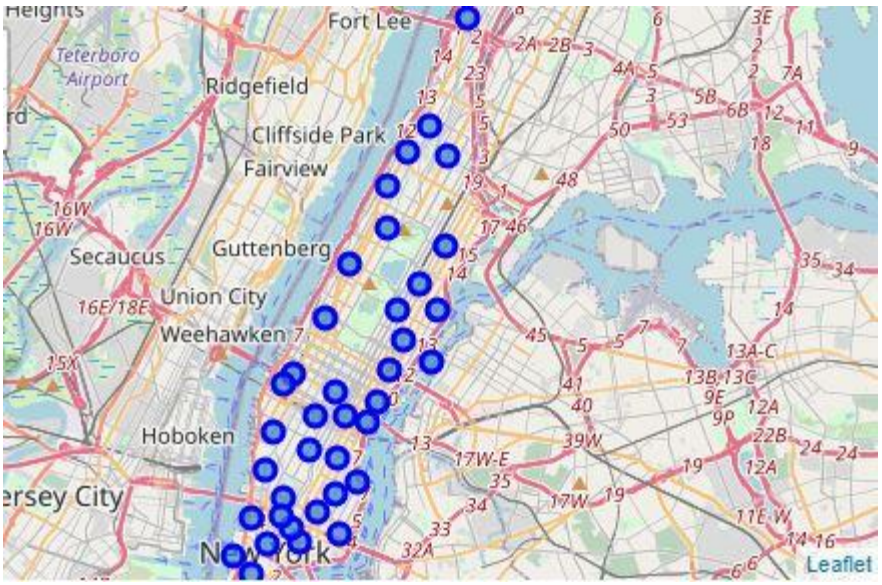
The link gives us a lot of information related to New York City but in order to have all in order our main interests will be the columns Borough, Neighbourhood, Latitude and Longitude. To get our data frame as we want I had to clean the original file it was necessary to have knowledge of how to manage data frames using the package pandas.

This is finally the cleaned data frame:

	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

EXPLORATORY DATA ANALYSIS

Thanks to a package it was possible to know the latitude and longitude of the place I was about to analyse, given this values I was able to map the city of Manhattan with all the Neighbourhoods it has.



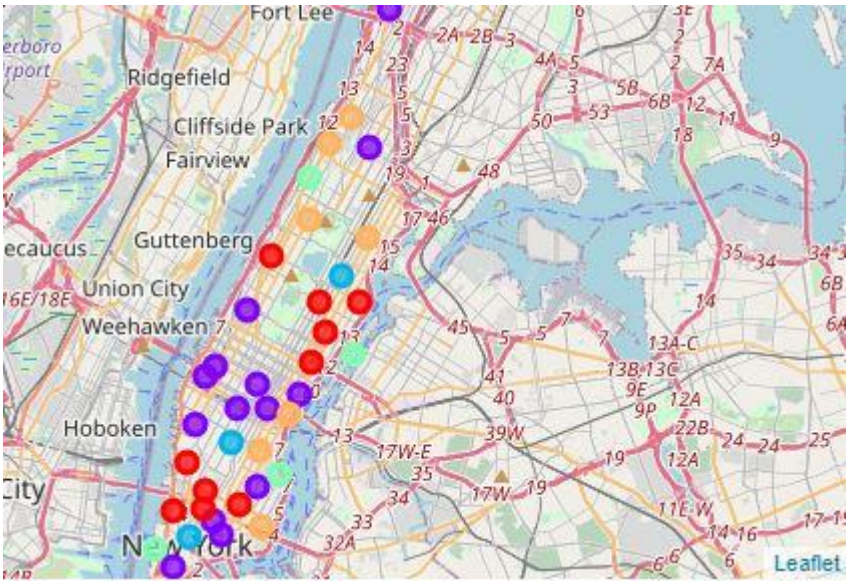
After I had the location of each neighbourhood I can proceed to analyse each zone using the Foursquare API, from where I can get more information. The important data from this new dataset is the category of the venues, because with it we can be able to see the distribution of category for each neighbourhood.

Our main interest was only 5 columns of our data frame which were Neighbourhood, Mexican Restaurant, Gym, Park, Italian Restaurant and Yoga Studio.

	Neighborhood	Mexican Restaurant	Gym	Park	Italian Restaurant	Yoga Studio
0	Battery Park City	0.016667	0.066667	0.10000	0.033333	0.000000
1	Carnegie Hill	0.011628	0.034884	0.00000	0.034884	0.034884
2	Central Harlem	0.000000	0.023810	0.02381	0.000000	0.000000
3	Chelsea	0.000000	0.010000	0.02000	0.030000	0.000000
4	Chinatown	0.020000	0.000000	0.00000	0.000000	0.010000

MACHINE LEARNING TECHNIQUES

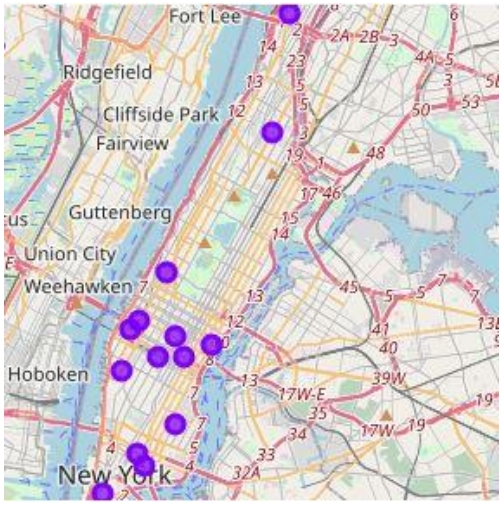
Clustering the Neighbours was our next step, using a machine learning technique called K-means we managed to differentiate each neighbourhood according to its distribution in 5 possible Clusters.



Finally, the las step is to know in which cluster we are according to our preferences. In order to succeed we needed to use another machine learning technique to classify me called K-Nearest Neighbours, it is used to determine in which group you are depending on your values.

RESULTS

The algorithm determines that I would be part of cluster 1, which belongs all the following neighbourhoods.



CONCLUSION

Now we know all the possible places where people who like Mexican food, parks, yoga, Italian food and working out can live. Although we found a lot of places in Manhattan where I can be happy I will always choose being a home, and home is where my family is.