**APPLIED DATA SCIENCE CAPSTONE**

**Introduction**

In order to succeed in any business, one has to be one of the best in that particular field offering solutions that one’s competitors cannot. XYZ housing agents thrive on this motto. We provide our clients with the best locations to rent/buy a house based on their most important necessities. Using data science, we analyze the city of New York to come up with suggestions of locations based on various criteria given to us by our clients. Let’s say you want a park within 500 m of your location. Done! But you also want to have a school along with the park in that radius. Done! You just have to name your criteria and leave the searching to us. With data science, we quickly come up with the best solutions for you.

**Problem for this study**

A client has moved from Los Angeles to New York City. The client is a single person who is a fitness freak. He is looking for locations which have a gym/fitness center/yoga studio for his fitness activities, a supermarket/farmer’s market/department store and a pharmacy within a walking distance of 750 m. He also does not want the place to be clustered with too many restaurants and bars and is looking for a quiet place. The client also wishes to be connected to the center of the city.

**Methodology/Approach and data used.**

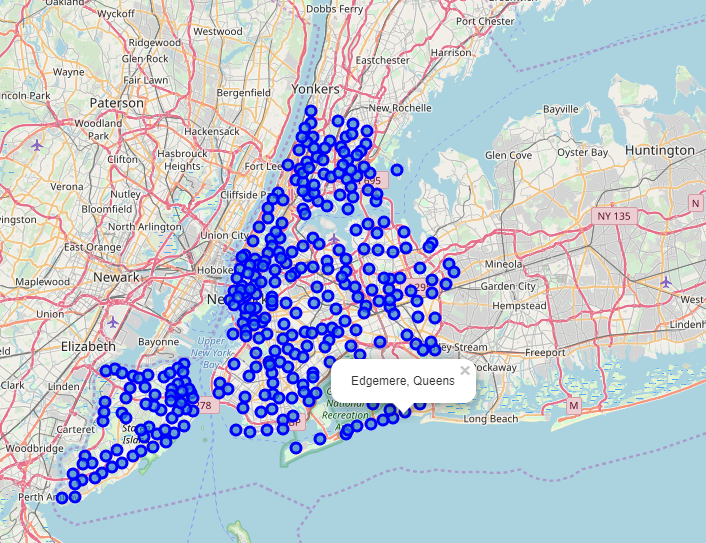
The following link is used to get the latitude and longitude data of NYC:

<https://geo.nyu.edu/catalog/nyu_2451_34572>

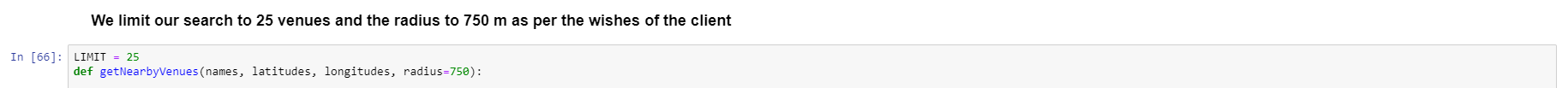
Preliminary analysis was done to identify the various boroughs of NYC.



It was observed there are 5 boroughs. On visualization using a folium map, we can see that Queens and Staten Island are not directly connected/nearby the center of the city and hence we exclude these 2 boroughs from further analysis.



We then use the Foursquare API which allows us to explore and search for venues in various places across the world. We use the search, explore and venues (end points) options for performing the analysis. We limit the search radius to 750 m as per the clients request.



We then use the k-means clustering analysis in order to cluster the neighborhoods based on the most common venues found in each neighborhood. We limit this to 10 and divide each borough into 3 clusters. This gives us an idea of the most common venues in each neighborhood which will give us an idea of which neighborhood has a gym, park and supermarket in their vicinity.

For example, using the above data set in the link, the following pic after a k-means clustering analysis of the borough of Manhattan in New York City shows that yoga studio is one the 5 most common venues in the neighborhood of Marble Hill. But none of the above criteria are satisfied for the client. Indeed it can be seen that this neighborhood is filled with restaurants and coffee shops suggesting a busy and bustling neighborhood and not a calm one as the client wishes.



**Results**

Based on the clustering analysis performed, we identified 3 neighborhoods in the borough of Bronx and 1 each in the borough of Manhattan and Brooklyn.

On examination of the clusters, it was found that Manhattan and Brooklyn are more popular for their restaurants and tourist spots and they may not be the most ideal locations for a peaceful stay.

The identified neighborhoods are as follows:

Bronx – Westchester Square, Mott Haven and Concourse Village

Manhattan – Stuyvesant Town

Brooklyn – New Lots

Of all these places, Concourse Village is the best recommended as it has the least number of commercial establishments among its popular venues and is more popular for venues relating to the daily activities such as pharmacies, diners gyms, fitness centers which is what the client wishes for.

**Discussions**

Of course, the above study contains some limitations which are highlighted as follows:

1. It does not consider the average rental values of properties.
2. It does not consider the safety of the neighborhoods.
3. The socio-economic data of the client is neglected.

But this study is gives a preliminary idea of how to filter neighborhoods based on a client’s wishes. Increasing the number of criteria can be more beneficial but the general process of using data science for filtering purposes remains the same.

**Conclusions**

In this study, we analyzed the various neighborhoods of NYC to identify suitable locations to rent a house based on certain pre-specified criteria. The foursquare app along with the k-means clustering analysis was used to perform this analysis. After analysis, we could narrow down the search to 5 neighborhoods and suggested the best one among these 5 neighborhoods.