CAPSTONE PROJECT

The Battle of Neighborhoods

Introduction
I chose Queens, NYC to be the focal area of my project. As a former resident of Queens, I'm aware of the 'hip' areas that are becoming very pricey all the way around. Bars are very controversial to some but definitely a staple in not only NYC but the world. They can be seen as bringing people together or for business owners, an easy way to make money. If you build it, they will come. There are certain "it" neighborhoods in Queens that are considered "hip" and loaded with bars and restaurants. It would not be wise to go in and open a bar in an area already saturated.

Business Problem
The purpose of this capstone project is to find the best place in Queens, NYC to open a bar. This project uses data science techniques and Machine Learning Clustering techniques to solve the business problem.
Business Problem : Where is the best place in Queens, NYC to open a bar?

Target Audience
Someone who wants to open a bar in Queens, NYC.

Data Description

I will be using following data to solve the business problem:
1) Dataset which contains the data of New York neighborhoods.
2) Co-ordinates of different neighborhoods in Queens, New York.
3) Data related to bars in Queens..

Data Extraction
1) The New York dataset which contains neighborhoods in New York is available free on the web. It also contains latitude and longitude of the locations in New York.
2)To get the co-ordinates of neighborhoods in Staten Island, I will use Geocoder Package.
3)To get the data about bars in QueensI will use Foursqare API.

Methodology
First, I need to get the data related to the neighborhoods in New York. This dataset provided by NYU contains all the information related to neighborhoods in New York. I only need boroughs, neighborhoods, latitude and longitude of the neighborhoods in the dataset. Create the dataframe to store this data. The dataframe has 5 boroughs and 306 neighborhoods. After importing all NYC neighborhoods, I then got only information about neighborhoods in Queens.
After getting the data of neighborhoods in Queens, I used Foursquare API to get the nearby venues within 500 meter area. For this, I created an account on Foursquare API to get Client ID and Client Secret. I need this information to access locations on Foursquare API. After taking the

neighborhoods, I grouped them by neighborhoods and taking the mean on the frequency occurrences of each venue category. This is the preprocessed data for Clustering.
I used K-Means Clustering Method to group different venues in group. K-Means Clustering is the simplest and most widely used unsupervised machine learning algorithm for clustering. I created 8 clusters based on frequency for 'Bar'. I chose to do 8 because there were quite a lot of results for bar in Queens due to ‘up and coming areas’. Without making more clusters, it would have been hard to identify and narrow down which areas.

Result

I could not get the map for the clusters to populate even though all the other maps in the notebook populated.

The 8 clusters are :
1) Cluster 1 - Had the most variety of neighborhoods. Lots of little neighborhoods here that could have potential to be good to open a bar.
2) Cluster 2 - Only had 2 neighborhoods with about 10 bars each.
3) Cluster 3 - Had a 3 neighborhood variety. Most of them had a decent amount of bars, definitely not the most. .
4) Cluster 4 - Had a decent variety of neighborhoods, most with a small concentration of bars per neighborhood
5) Cluster 5 - The 3 neighborhoods all have a decent concentration of bars (Long Island City, Astoria, Sunnyside Gardens)
6) Cluster 6 - About 4 neighborhoods including 2 highly populated with bars (Kew Gardens, College Point)
7) Cluster 7 - Single neighborhood(Jackson Heights) high concentration of bars.
8) Cluster 8 - Single neighborhood (Bayside) with high concentration of bars

Observations

It looks like the majority of bars are in cluster 3 and cluster 4 Long Island City, Sunnyside Gardens, Astoria, Ridgewood, and Little Neck. Cluster 0 has some neighborhoods that do not have a large concentration of bars and therefore would be a great choice to open a bar: Rochdale, Cambria Heights, Floral Park, Douglaston.

Conclusion

In this project, I have determined business problem, performed data preprocessing, applied Machine Learning K-Means Clustering method to solve the business problem.