Final Assignment – Capstone Project

- 04th December 2020 -

I. Introduction

I.1 Background

John lives and works in 522 E 189th St Belmont, Bronx, New York. He has been offered a new job opportunity to work in Staten Island, NY which is far from his home in his Belmont neighbourhood.

John is willing to accept the job offer if he manages to find a neighbourhood in Staten Island similar to his Belmont neighbourhood where he can move in. Joh have a preference for neighbourhoods with many bakery shops. We will use our data science powers to generate a few most promising neighbourhoods based on this criterion. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by John.

I.2 Objective

We can help john take a decision by looking for similar neighbourhoods in Staten Island where he can live near his new job. A recommender system could be built based on a clustering approach combined with Foursquare location data from Staten Island neighbourhoods and Belmont neighbourhood.

The above business problem could be resumed to the following question: what Staten Island neighbourhoods could be similar to John's neighbourhood: Belmont?

II. DATA PREPARATION

II.1 Data sources

We will need data from Staten Island, NY and Belmont, Bronx, NY. The data must include:

- Neighborhood name, latitude and longitude of each neighborhood in Staten Island.
- Neighborhood name, latitude and longitude of Belmont.

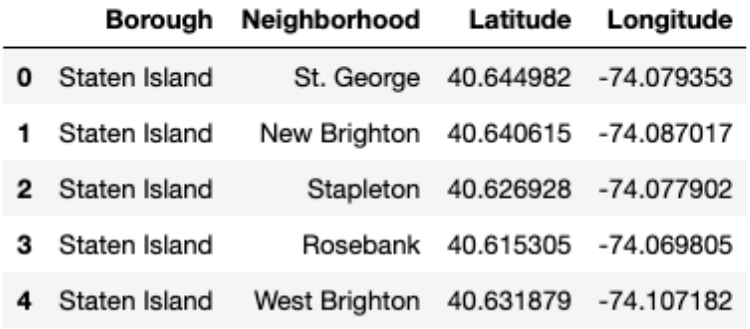
Then, we analyze each neighborhood by calling foursquare venues and their categories in order to rank each neighborhood from Staten Island. The same analysis must be done on John's neighborhood while normalizing the ranking of the neighborhoods from Staten island and John's neighborhood. Data is fetched from the following link: '*https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN SkillsNetwork/labs/newyork_data.json*'

The data contains 5 boroughs and 306 neighborhoods from New York. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood for visualization purposes.

We add then John's address which is: 522 E 189th St Belmont The Bronx, NY.

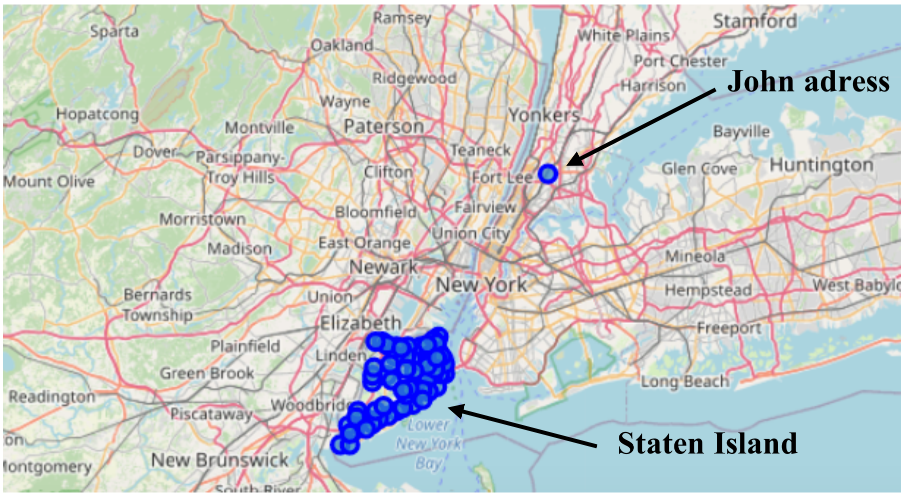
II.2. Data cleaning

We are interested in data from Staten Island, so a slice for the neighborhoods in Staten island is what we will keep for our analysis.



	Borough	Neighborhood	Latitude	Longitude
0	Staten Island	St. George	40.644982	-74.079353
1	Staten Island	New Brighton	40.640615	-74.087017
2	Staten Island	Stapleton	40.626928	-74.077902
3	Staten Island	Rosebank	40.615305	-74.069805
4	Staten Island	West Brighton	40.631879	-74.107182

The Staten Island data contains 63 neighborhoods where John can potentially compare to his Belmont neighborhood. Our aim is to cluster this data using Foursquare location data.



John adress

Staten Island

II.3. Feature Selection

Now that we have our locations data, let's use Foursquare API to get informations on each neighborhood.

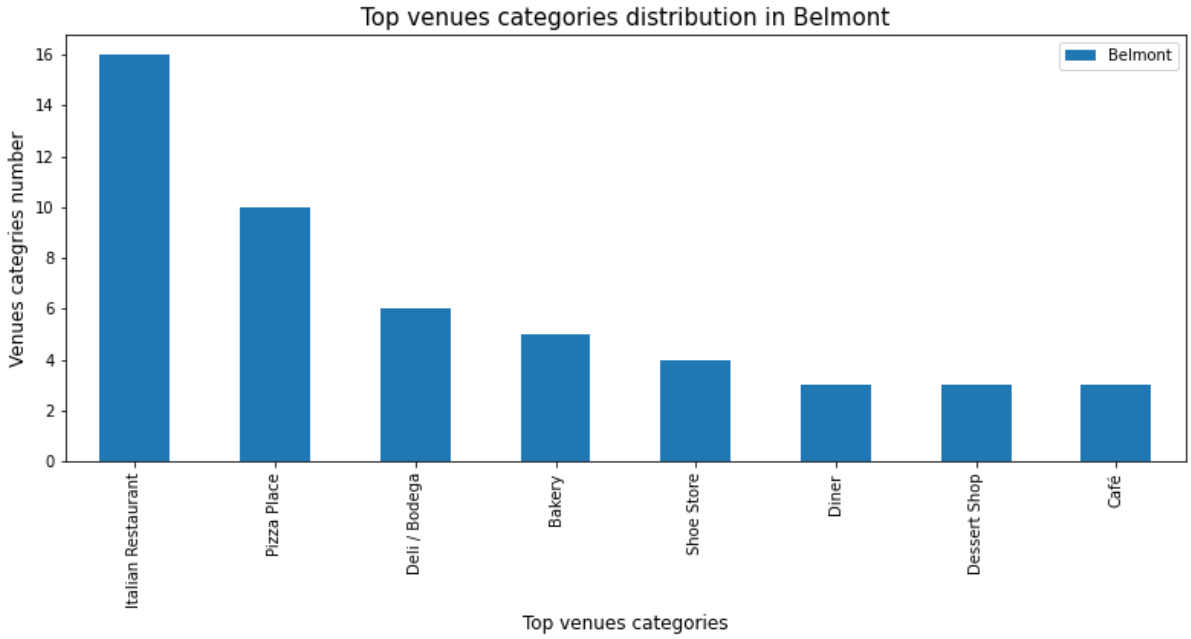
We are interested in venues call in order to categorize each of our neighborhood. Venue category are given by regular call to Foursquare API which return the venues categories (Italian restaurant, museum, fitness, spa, pizza place, outlet mall,…).

We choose venues in a radius of 800 m for each neighborhood. There are 212 venue categories that we base our analysis upon.

II. DATA ANALYSIS

II.1. Distribution of venues categories

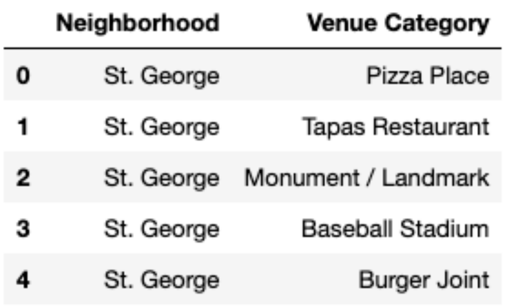
John's address in Belmont contains mostly the following abundant venues categories:



John current neighborhood in Belmont contains many Italian restaurants and a descent number of food and beverages proximity such as bakery, deli/bodega and pizza place. A neighborhood in Staten Island having a similar distribution of the venue categories would be a good fit for our objective.

II.2. Scaling the features

Location data extracted from Foursquare API is organized by all the venues in a neighborhood and the relative category, here, as an example for St. George neighborhood:

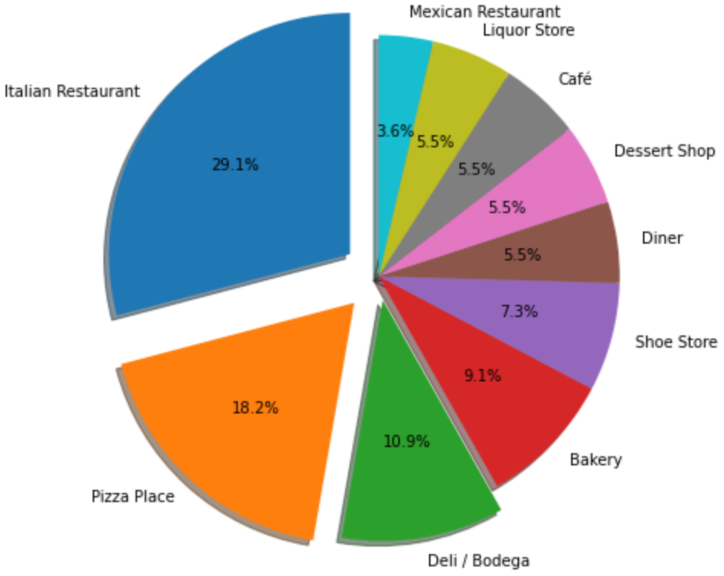


	Neighborhood	Venue Category
0	St. George	Pizza Place
1	St. George	Tapas Restaurant
2	St. George	Monument / Landmark
3	St. George	Baseball Stadium
4	St. George	Burger Joint

Venues categories are represented by categorical data. The clustering algorithm isn't directly applicable to categorical variables because Euclidean distance function isn't really meaningful for discrete variables. Hence, one- hot encoding approach is used to convert these categorical data to numerical data.

We divide each venue categories by the sum of all venues categories in order to have numerical values ranging between 0 and 1 which will allow the algorithm to compare all the neighborhoods.

The previous top venues categories in Belmont can be scaled to:



Top Venues categories distribution in Belmont

We are looking for neighborhoods in Staten Island with a significant distribution of the following venues categories: Italian restaurant, Pizza Place, Deli/Bodega, Bakery.

III. CLUSTERING THE DATA

II.2. K-means Algorithm

Clustering is a set of techniques used to partition data into groups, or clusters. Clusters are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters.

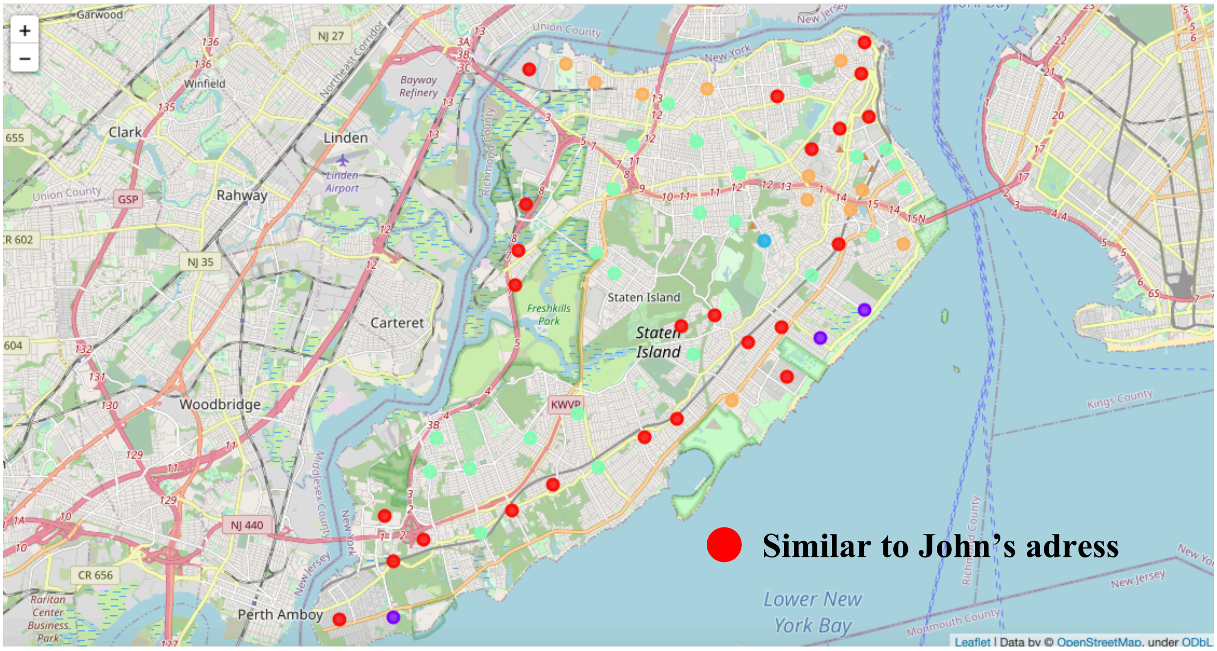
K-means algorithm is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. K-means will partition the neighbourhoods into n groups based on the distribution of venues categories. The customers in each cluster are similar to each other in terms of the features included in the dataset.

II.2. K-means parameters

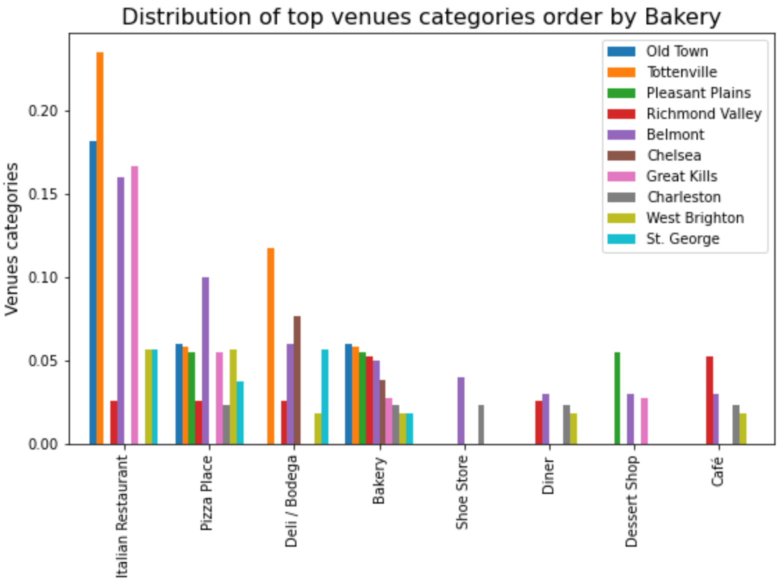
- Venues categories: We choose the 10 top venues categories from our data in order to cluster the neighborhoods. Base on the above pie chart, Belmont venues categories distribution are less representative starting the 10th most common venue.
- Clusters number: we choose to cluster the neighborhoods in five categories as an arbitrary value.

IV. Results and discussion

After clustering the results and plotting the neighbourhoods by latitude and longitude, we obtain the following results:



The results show that there are 24 neighbourhoods in Staten Island similar to John's address in Belmont. Exploring the most common venue distribution of the 10 top neighbourhoods ordered by bakery is shown on the right:



We can see that the four most common venues categories in Belmont are well represented in the clustered neighbourhoods in Staten Island.

There are four neighbourhoods having approximately the same distribution of bakeries such as Belmont: Old Town, Tottenville, Pleasant Plains and Richmond Valley.

Tottenville and Great Kills have a significant distribution of the most common venues in Belmont but the less common venues are not well represented while West Brighton is well distributed between many most common venues of Belmont.

We can advise John to look for a new place in Tottenville if he is very interested in finding the four most common venues categories of Belmont (Italian restaurant, Pizza place, Deli/Bodega and bakery). Otherwise, if John is more interested in the overall most common venues categories, we could advise him West Brighton or Richmond valley.

IV. Conclusion

The purpose of this project was to identify Staten Island neighbourhoods similar to John's Bronx address in Belmont. By identifying venues categories from foursquare API, we have grouped each neighbourhood in Staten Island and Belmont by the top venues categories.

Clustering of these neighbourhoods was performed in order to identify best Staten Island neighbourhoods having the 10 most common venues categories such as John' address in Belmont which are italian restaurant, pizza place, deli/bodega and bakry.

Final decision on optimal neighbourhoods gives John the choice between 24 neighbourhoods having the similar common venues categories distribution such as his current address.

We have made recommendations of 5 neighbourhoods from the 24 neighbourhoods found.

John knows now that there are similar neighbourhoods in Staten Island and he could accept the new job offer. A further analysis could be improving the decision by narrowing the final list of 24 neighbourhoods by adding features such as proximity to his new work, proximity to airport or rent price.