The Battle of Neighborhoods

Md Nazrul Islam

Coursera IBM Data Science Capstone Project

Motivation

- Suppose, a person Mr. X has been living in Toronto for 8 sweet years of his/her life.
- Now Mr. X has to leave and relocate to Manhattan NY due to his job relocation.
- Mr. X has been used to a particular lifestyle for a longtime in Toronto. He may like to go to Chinese restaurants for lunch, maybe he loves to visit some parks in the weekends.
- Now, Mr. X needs to choose a neighborhood in Manhattan which has all the amenities he is used to in his neighborhood in Toronto at a close proximity.

Objective

- Applying k-mean clustering algorithm to cluster the neighborhood based on their similarities in different amenities and venues.

- For defining success we will try to figure out the optimal cluster size by doing some exploratory data analysis on different clusters and trying to observe their similarities.

- Based on the clusters Mr. X can be offered some options which he can choose for his future residence.

Workflow

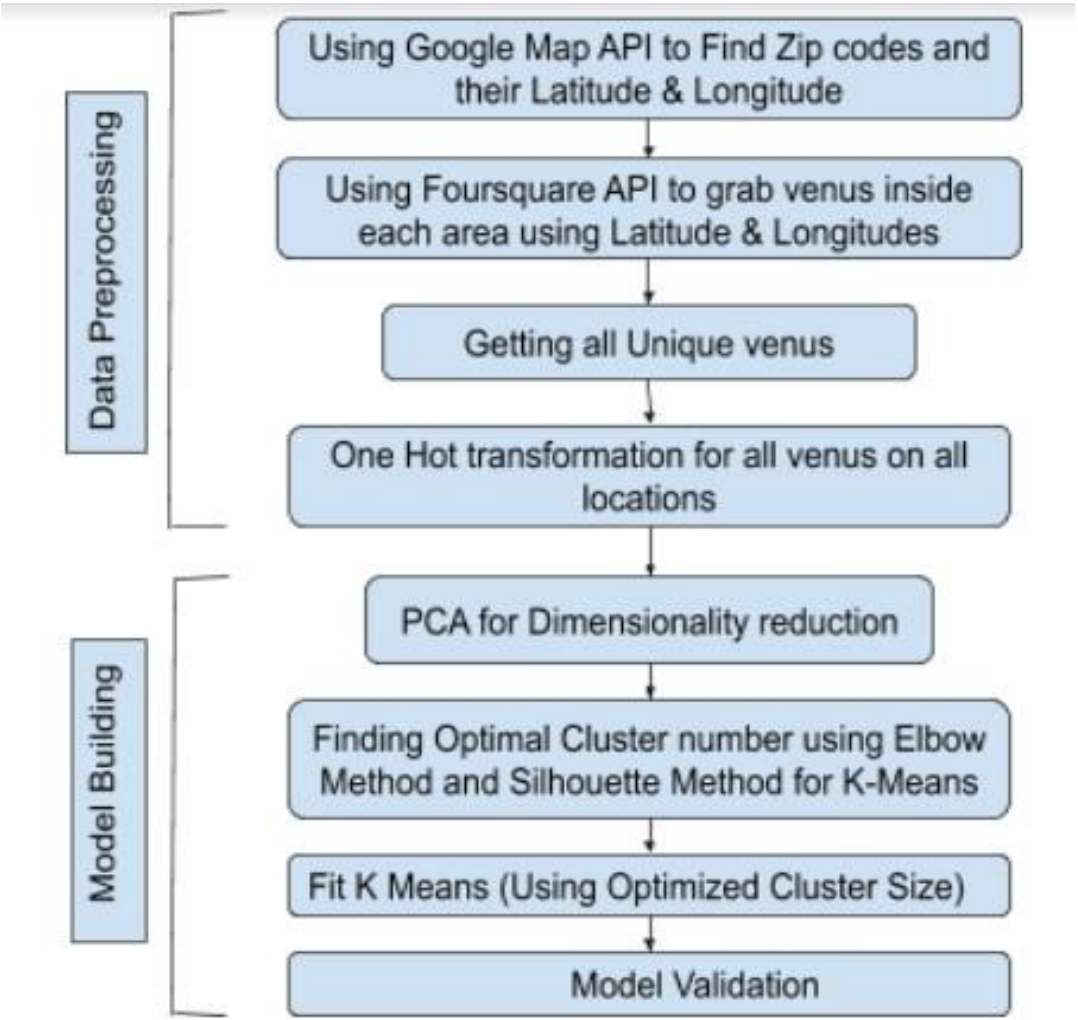


Fig 1. Neighborhood segmentation work flow chart

Selecting Principal Components

- Applying PCA Components analysis will reduce the number of variables to a smaller set but Confining all the attributes of neighborhoods

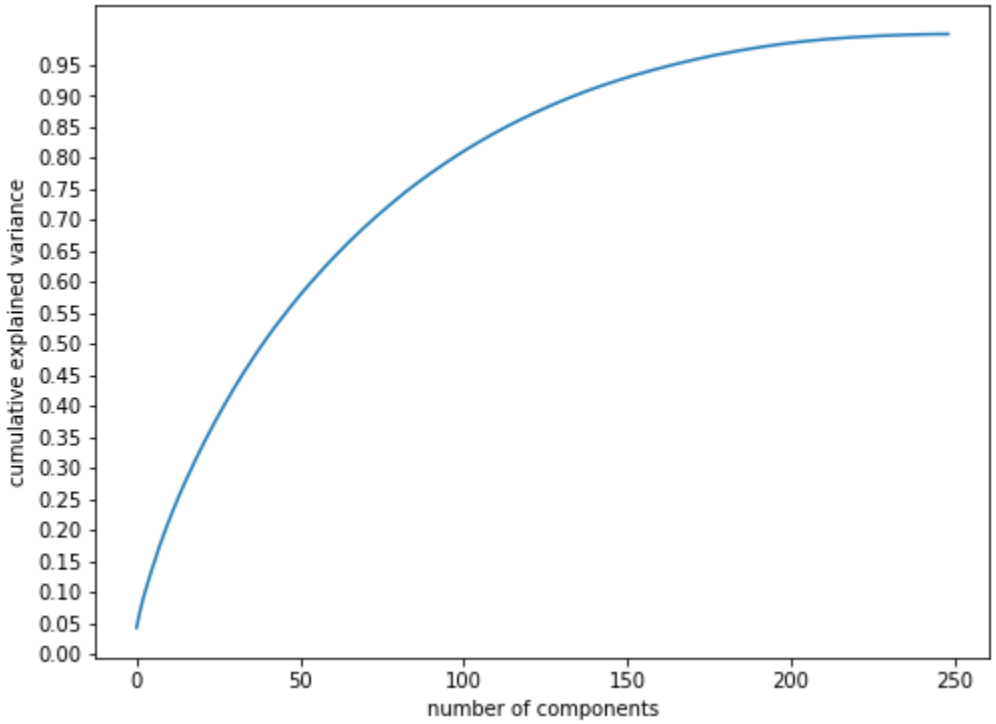


Figure 2. Selecting Principal Components

Elbow Method

- Through Elbow method determine the optimal number of clusters

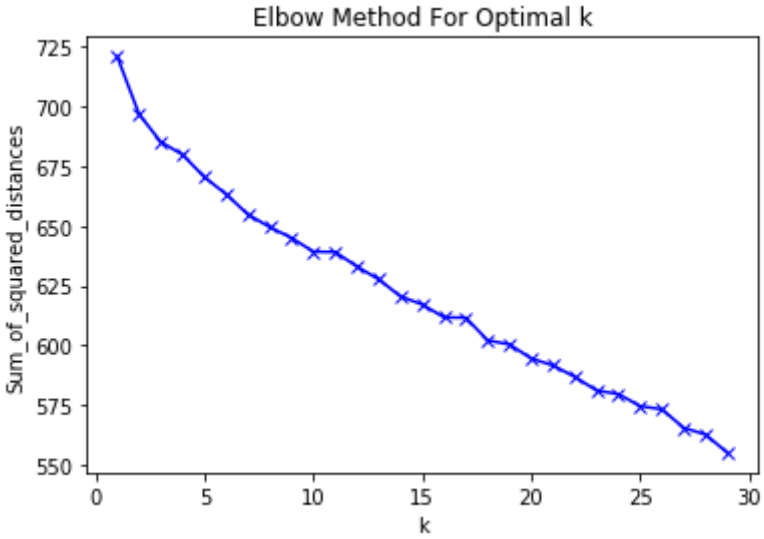


Figure 3. Elbow found at k =5 (K is number of Clusters)

Silhouette Score

- Through Silhouette score validate optimal number of clusters

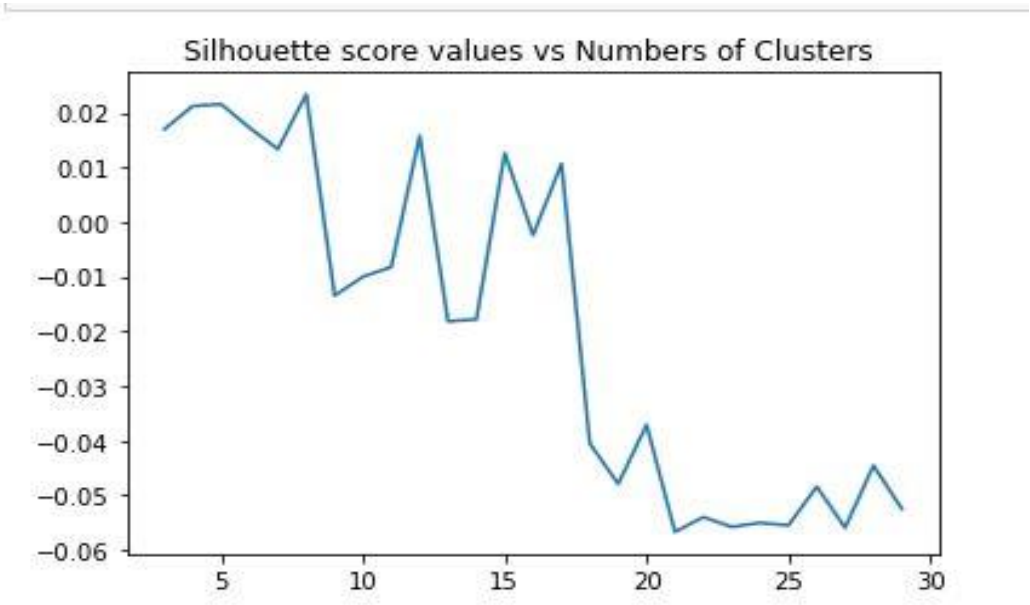
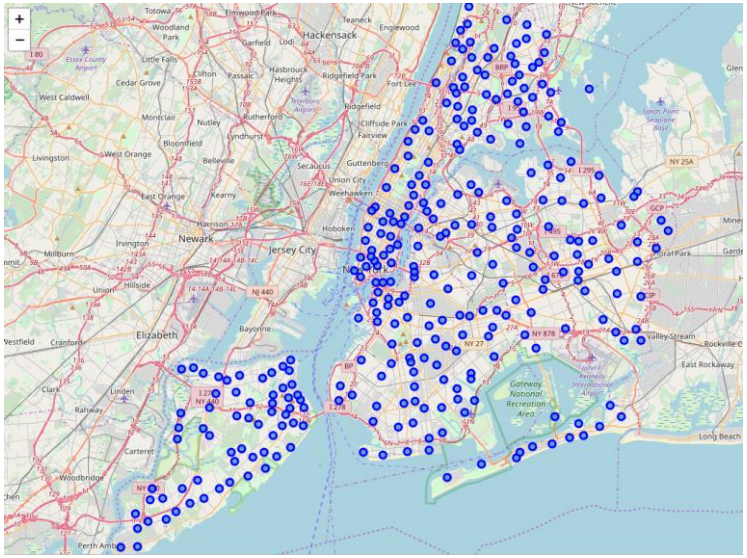
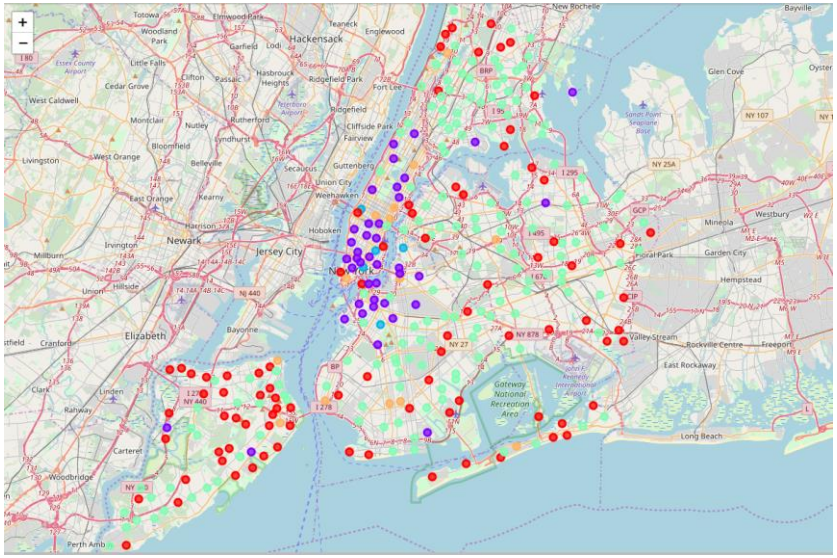


Figure 4. Silhouette Score confirms optimal Cluster Number

Cluster Visualization



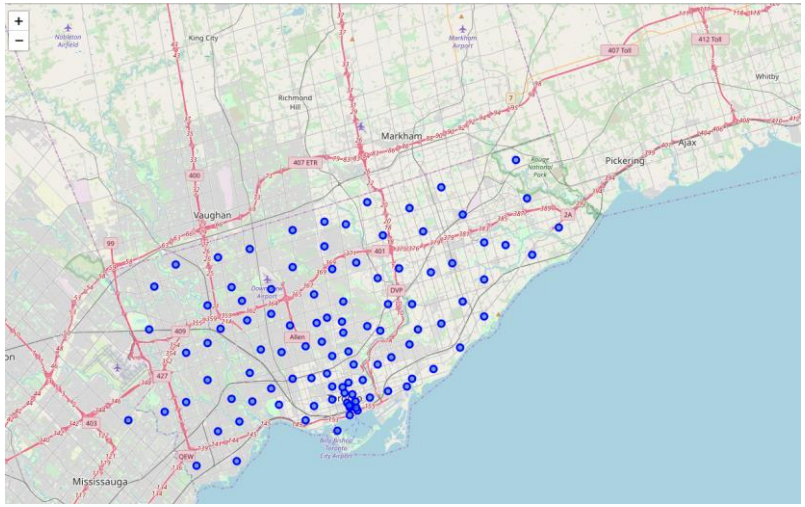
Before Clustering

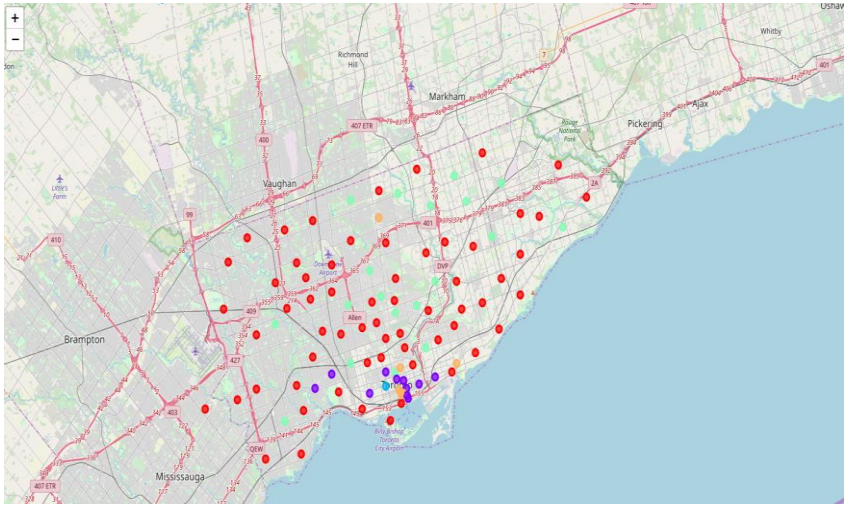


After Clustering

Figure 5. Manhattan, NY neighborhoods with the assigned cluster label

Cluster Visualization





Before Clustering

After Clustering

Figure 6. Toronto neighborhoods with the assigned cluster label

Comparison of neighborhoods

- First column is the cluster ID, second column is the number of neighborhoods of Manhattan NY that matches the attributes for cluster 1, third column is for the Toronto neighborhoods

Cluster ID	Manhattan, NY no. of Neighborhoods	Toronto No. of Neighborhoods
1	16	0
2	108	36
3	54	8
4	0	2
5	127	4

- A person living in a neighborhood in cluster <i> can move to any neighborhood of Manhattan NY which has the same attributes for that cluster. Here i is in {1,2,3,4,5}.
- Ref: https://github.com/nazrulcse2k/data_science_nazrul/blob/master/NY%20vs%20Toronto.ipynb

Observations

- If Mr. X lives in any neighborhood of Toronto bearing the attributes of either Cluster 2 or 4 then he will have a lots of choices in Manhattan, NY to choose as new neighborhood.

- If Mr. X lives in a neighborhood of Toronto under cluster 3 then he will still have choices from Manhattan but options are not at large as others.

- If Mr. X lives in a neighborhood of Toronto under cluster 4 then he will not have much choice to make, he will have to choose some neighborhood in Manhattan based on some other options.

- Ref: https://github.com/nazrulcse2k/data_science_nazrul/blob/master/NY%20vs%20Toronto.ipynb

Conclusion

- The project work was only done on the zip codes of New York and Toronto, which includes 401 zip codes each having 150 features even after dimensionality reduction with PCA. If features are increased as well as cluster numbers then Mr. X may have more options to choose for each cluster.
- Note that we have a huge feature space but limited number of samples. We can collect data from entire United States and Canada, which will make our dataset well balanced and model can be applied for relocation problem between US and Canada.
- From figure 5 and 6, one can spot certain outlier in our data. Those outliers can be more fine-tuned applying adaptive machine learning rules for more robust clustering.

THANK YOU