Battle of Neighborhoods

Full Report

1. Introduction/Business Problem

For this specific problem or situation, we would like to group the different neighborhood in the city by segmenting the different venues in a neighborhood by its venue category, and consequently group neighborhoods together that incorporate similar kind of neighborhoods. By grouping all the similar kinds of neighborhood, this will help us to know which places to consider when a person wants to move out to another area.

By identifying identical neighborhood in different cities this will help people feel more like home and make better decision on which neighborhood to choose base on data driven analysis.

2. Data

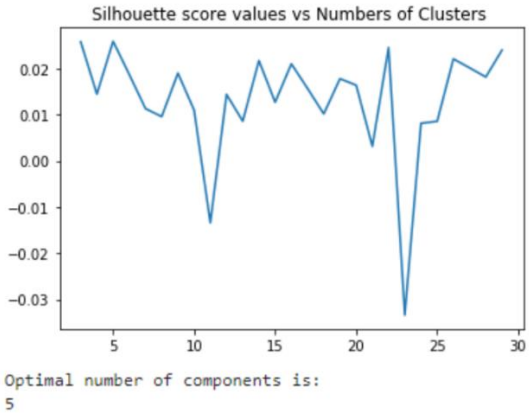
For this Project we will use two data bases (NY and Toronto) that contains information such as neighborhoods, coordinates and postcodes. Before doing the analysis, it was important to clean the data and put it in a proper format to continue. Many of this job was done in the previous projects of this course. After finishing with data wrangling, we proceeded to come up with different venues that the different venues have to offer in both cities. To retrieve venue information, we used the Foursquare API where we found common features for different shops, restaurants and many others; where we decided to create one table with both information to perform a cluster analysis later.

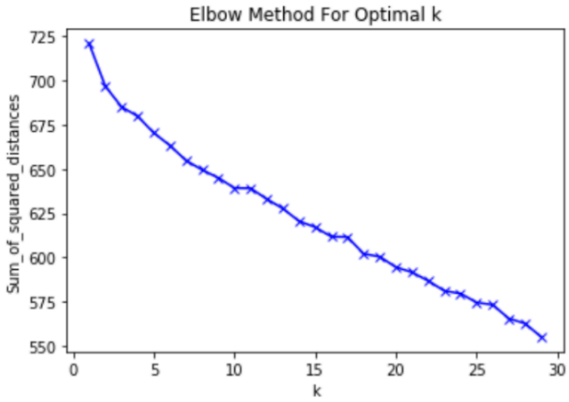
3. Methodology

The purpose of this project consists in grouping similar neighborhood in NYC and Toronto. The first step in our methodology was to determine the optimal cluster number. For this case, a k-means algorithm was used which consists on a simple unsupervised machine learning algorithm that groups a dataset into a user-specific number of clusters. Our goal is to choose a small value of k that has a low SSE. Additionally, a random

initialization was implemented to overcome issues, where many iterations on different random initializations were performed in order to find the best set of convergence in this case.

4. Results

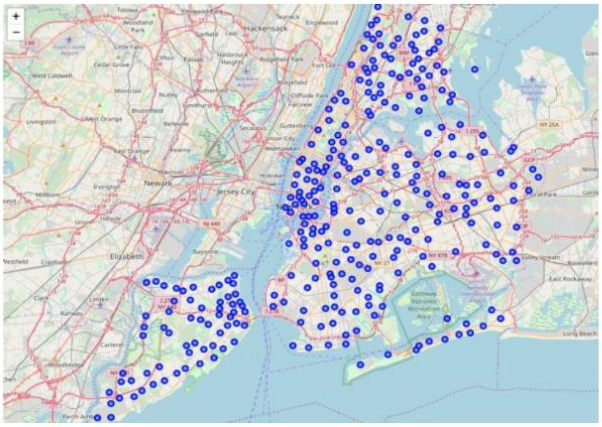


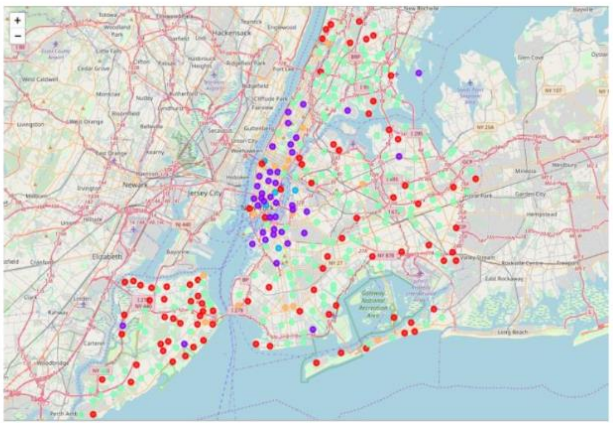


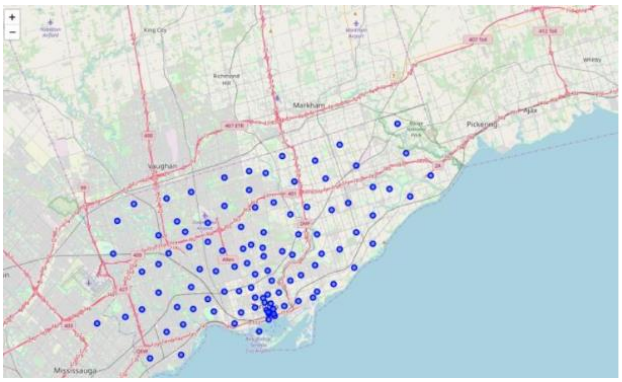
To find the Optimal Number of Cluster we use the Silhouette score and it confirms a total of five cluster where a peak can be seen. Using a different approach (Elbow Method), we can´t see a distinctive elbow, but we can still conclude that 15 cluster would be a reasonable choice for K.

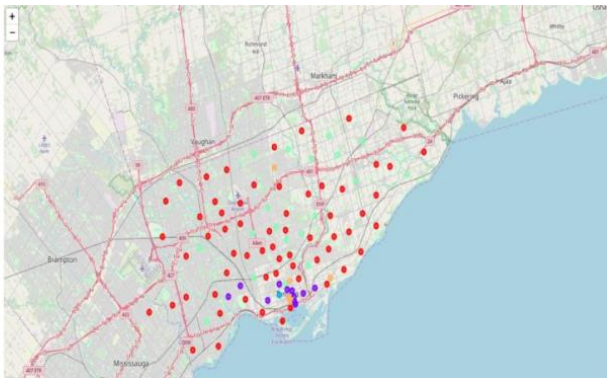
Exploratory analysis – Clusters

Different clusters for the city of 1. NY and 2. Toronto.









5. Discussion

Some of the limitations found in this project where that there was a huge feature space but very few numbers of samples. It would be beneficial to add more data to improve our analysis.

We identified some outlier in our data. For more detailed analysis we would need to filter the outliers.

In this case we only used k mean algorithm, but probably other algorithms ca be applied to compare the results and choose the best approach to use for this specific problem.

In the case we had location data on a deeper level, for instance at neighborhood level may result in better grouping of similar data points which eventually may result is better clustering. The study here is being needed by visualizing the data and clustering information on the map of the City of New York and Toronto.

6. Conclusion

Nowadays, many people need to move from one place to another for different reasons such as work, school, family issues, etc. By implementing a neighborhood recommendation based on location data is something to be considered very useful for companies developing new building for example. It can also improve a lot the organization of a city and not many resources need to be used for this project because the information in most of the cases can be found free online. But we will need a great team of data scientist to develop this project in further detail 😊