Capstone Project Report

Introduction : Business Problem

A company that is based in New York City wants to create a new office in Toronto. They want to put their office in the neighborhood having similar venues to the one they are based in so that the coworkers that will work there can be in similar conditions to the one based in New York. My job as a Data Scientist is to use data analysis methods to find that neighborhood.

Data

Data sources

In order to retrieve Toronto's neighborhoods I scraped the list of Canada's postal codes from Wikipedia and read a CSV file containing the coordinates of Toronto's neighborhoods. With these coordinates I accessed the Foursquare API in order to find the venues of each neighborhood and put them in a Dataframe.

Having the coordinates of the New York office, I accessed the Foursquare API to find the venues around it and put them in another dataframe.

Data preprocessing

I append the 2 dataframes into one having one row per venue with the features being the venue, the venue category,the neighborhood, and the coordinates of the neighborhood and the venue.

Then I used the one hot encoding technique with the resulting dataframe so that the dataset can be fit into a machine learning model : I got rid of the ‘Venue Category’ feature and replaced it with one feature for each category having a 0 or 1 value.

The final dataframe contains one row per neighborhood with the features being the proportion of each venue category in that neighborhood,

Methodology

My objective is to cluster the neighborhoods of Toronto into different groups and then try to predict the cluster the New York office would belong to in order to find the neighborhoods that would be the most similar to the office’s surroundings. I decided to use the K-means method.

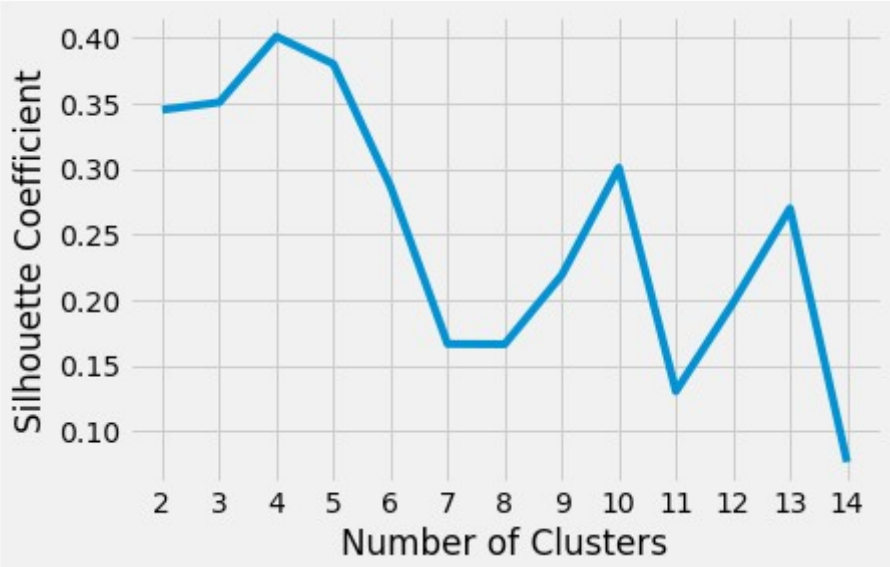
First, I try to find the best number of clusters by using the silhouette method : I fit a new Kmeans model with the dataset and a different number of clusters and I choose the number of clusters which gives the best silhouette score.

With the number of clusters defined I fit a Kmeans model with the Toronto's rows so that each neighborhood has an assigned cluster label.

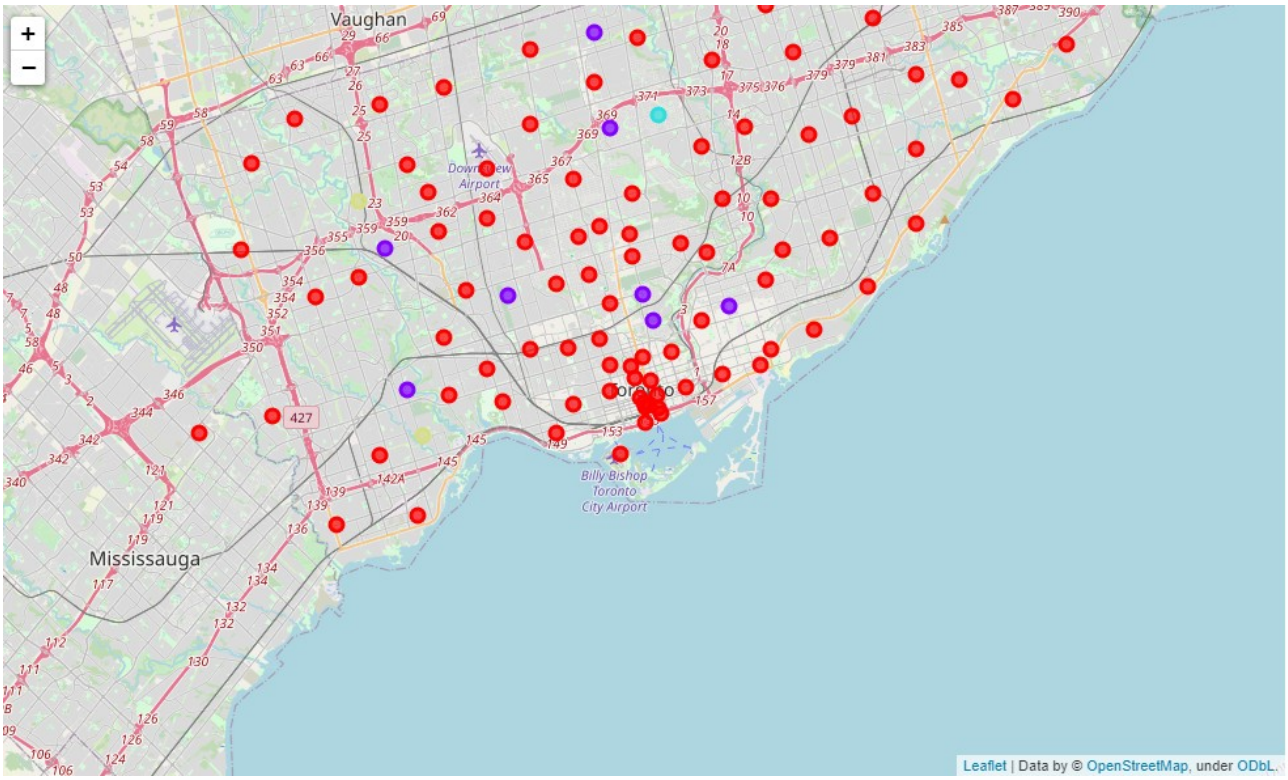
Then I use the trained model to predict the cluster that the New York office would belong to.

Results

According to the results from the silhouette method, the best number of clusters is 4.

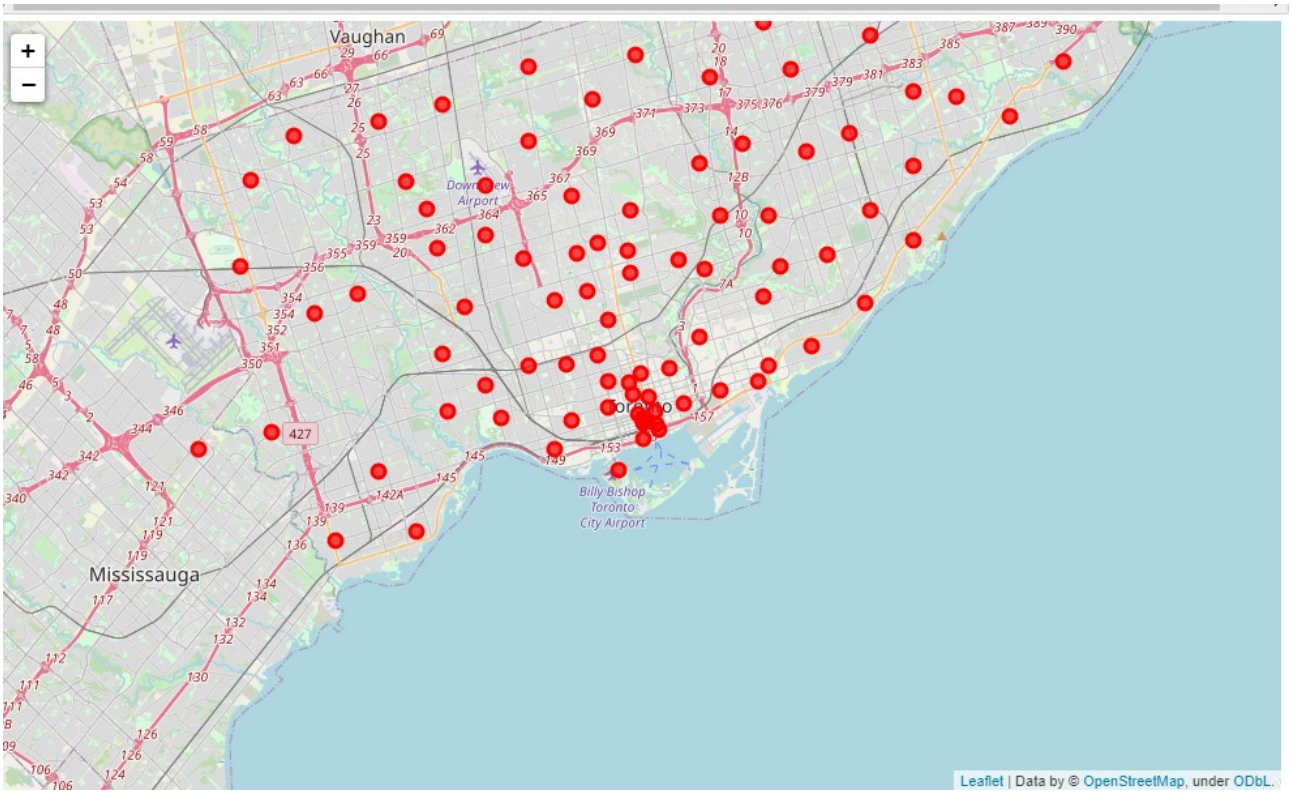


The Kmeans machine learning shows that the majority of the neighborhoods belong to a single cluster of similar venues, and the 3 other clusters are much smaller



The different clusters of similar Toronto’s neighborhoods

Finally we try to predict which neighborhoods the NY's office surrounding would be the most similar to by predicting the cluster it would belong. It seems the NY office would fit best in the first cluster.



The first cluster of Toronto's neighborhoods

Discussion

The results seem to show that the New York's office surroundings is similar to most neighborhoods of Toronto. A recommendation for the company would be to put their new Toronto's office in any of those 89 neighborhoods. A problem with the Kmeans clustering method is that most neighborhoods are put in the same cluster (89/97) therefore it doesn't allow us to discriminate much between the neighborhoods. A solution to this would be to use a different clustering technique like DBSCAN or to find different data sources.

Conclusion

Our purpose was to find a location for a new office in Toronto in a neighborhood that is like the surroundings of the New York office the company is based in. Therefore we did get data that allowed us to get the venues and the venue categories of Toronto's neighborhoods and New York that we then used as feature for Kmeans clustering. We used the silhouette method to determine the ideal number of clusters then we used a Kmeans model to fit the toronto's neighborhoods and assign

clusters to them. Then we used the model to predict the cluster of neighborhoods that are the most similar to the New York office’s surroundings.

The neighborhoods of this cluster would be a good choice to put the new office inbut the clustering did not discriminate enough between neighborhoods and we should rather use a different clustering method to answer the question or find another data source.