IBM Applied Data Science Capstone project

This document reports my project for the course IBM Applied Data Science Capstone. The project is required to make use of the Foursquare API. It is based on a hypothetical—but realistic—business problem.

Introduction

A coffee chain has multiple locations in New York City and are looking to open their first location in Toronto, Canada. However, not all location in NYC are equally productive. The task is to find the best neighborhood for the new Toronto location. Their location in Williamsburg (Bedford and S 1st, Brooklyn) is their most successful location and sales are much higher here than in their location in the Upper East Side (Madison and E 80th, Manhattan; fictional information). They think that the local business environment plays a role. An approach to solve problem is to find a Toronto neighborhood where the local business environment resembles Williamsburg more so than the Upper East side.

Data

Foursquare API allows us to locate businesses in an area and access user provided scores. This can provide data sets that allow us to cluster together neighborhoods based on the available businesses in an area. Such data sets allow us to build a clustering model that can group together neighborhood from both NYC and Toronto neighborhoods, based on their similarity. The challenge will be to identify a model that captures the difference between Upper East Side and Williamsburg. Such model could then be used to cluster Toronto neighborhoods as well. The data sets will be calls to the Foursquare API for venues in the relevant area that return venue name/id, location, categories, as well as user rating information.

Methodology

I solve the problem by establishing a KMeans clustering model based on the NYC neighborhoods and then cluster the Toronto neighborhoods using the same cluster centers.

The picture below shows the NYC neighborhood used in the model in yellow and the two coffee shop locations in navy blue. It includes neighborhoods from Manhattan, Bronx, and Brooklyn.

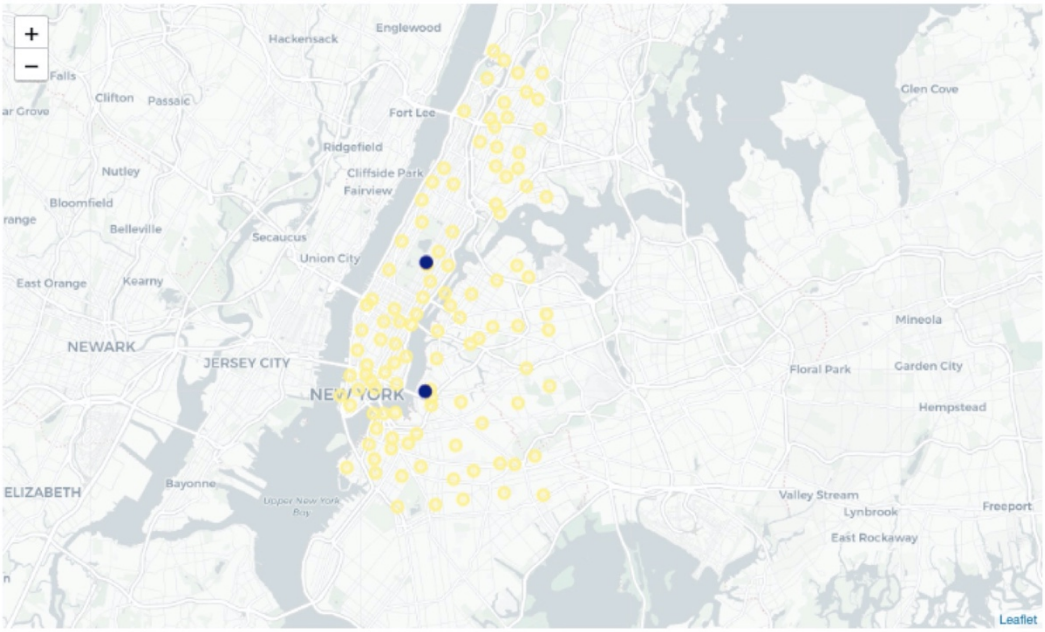

Figure 1 Map of neighborhoods in the model (yellow) with the two coffee shop locations (navy blue) in Williamsburg and Upper East Side.

Using these neighborhoods, I call the Foursquare API and request information about up to 100 nearby businesses in a radius of 1000 m of the latitude and longitude. Models with k = 1–20 is computed and the sum of squared distance to cluster centers is plotted for each model. This shows that, as expected, the distance gets smaller with more clusters. I decide to use 12 clusters in the final model as this is the closest to an "elbow point".

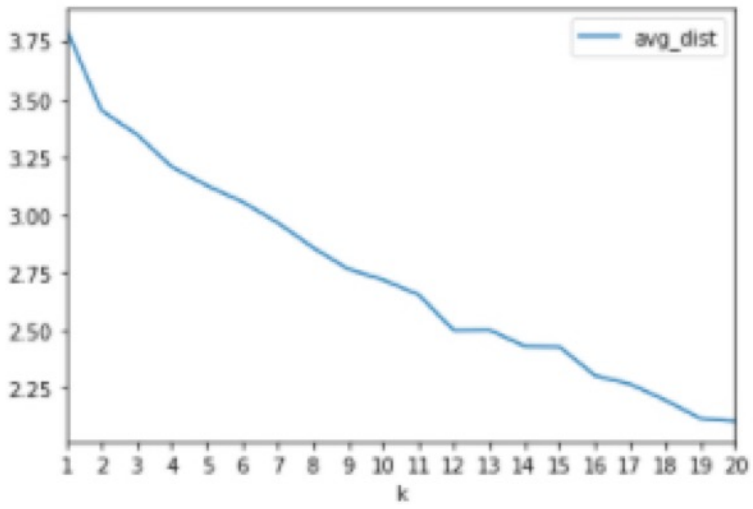

Figure 2 Plot of sum of squared distance to cluster centers versus number of clusters (k). Not average distance.

We can visualize the clustering on a map. Here we notice that the coffee shop locations are right next to cluster 11 (Upper East Side) and cluster 4 (Williamsburg).

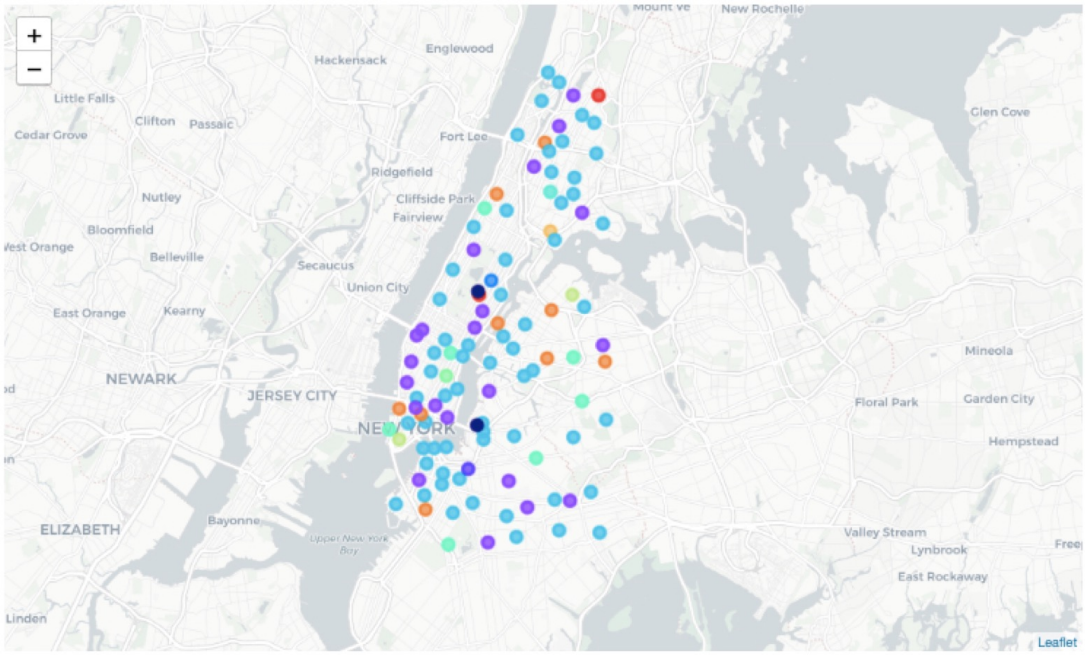

Figure 3 Color coded map of NYC neighborhood clusters. Coffee shop locations are still in navy blue.

I can import the data for Toronto neighborhoods (postal codes) and make a KMeans models with then NYC cluster centers predefined.

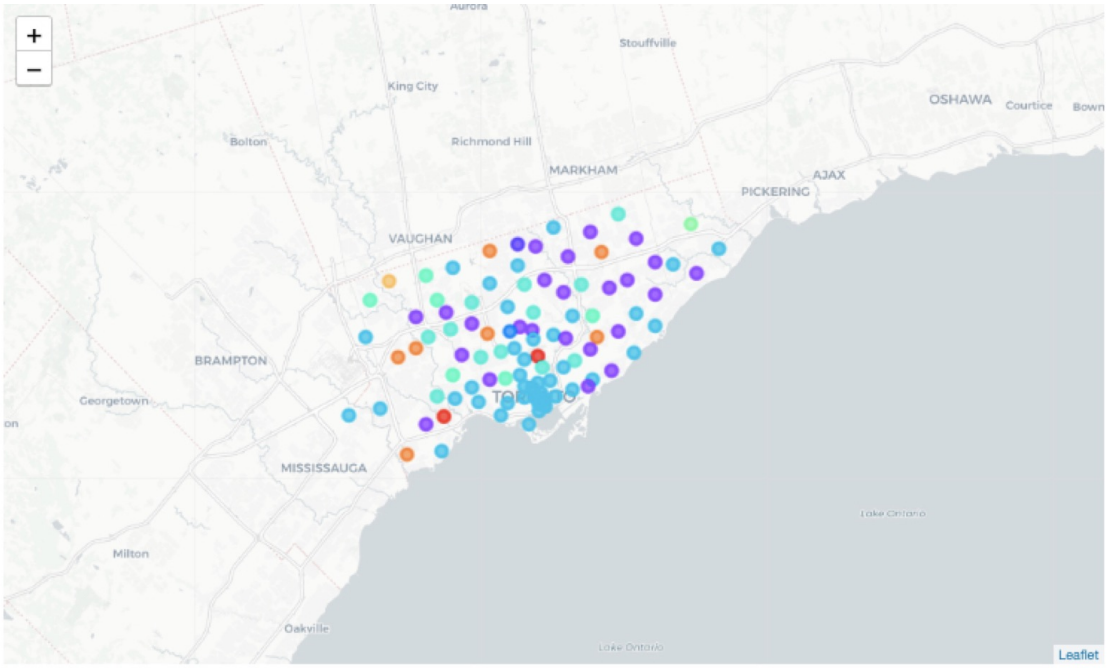

Figure 4 Color coded map of Toronto postal code clusters.

Results

The clustering for Toronto clearly shows that there are many locations that have the same composition of businesses nearby as Williamsburg, Brooklyn. In fact, 45 postal codes assign to the NYC cluster 4. Most of them are in the central part of Toronto. In contrast only one postal code assign to the NYC cluster 11 like the Upper East Side.

Discussion

First of all, one would assume that the coffee shop chain hopes that nearby business composition will be indicative of sales at the location. It is, not surprisingly, probably much more complex since so many Toronto neighborhoods resemble their successful location in Williamsburg. Instead, the success at the Williamsburg location could be associated with many other (confounder) attributes e.g. what the population that shops in area looks like in terms of age and socioeconomic status or perhaps even just how many people shop or live in the area to begin with.

One potential problem with the analysis is that categories are perhaps too specific e.g. the success of a local business might have little to with whether or not the nearby popular restaurant serves African cuisine and more to do with the target audience of said restaurant. Because the categories are so specific, cluster differentiation could be driven by relatively few but very unique business categories in an area. The user score of nearby businesses or the number of visitors (as is provided by Google) would be interesting to implement into the model.

The analysis does not incorporate any information branding of nearby business. Regardless of the user score or categories, a driving factor for the Williamsburg location could be that it is located next to businesses with a particular brand profile e.g. an expensive coffee shop will prefer to be next to the Apple Store compared to McDonalds. This kind of information is very difficult to approach with data science and is likely easier to incorporate at a later stage (after data science) on case-by-case basis.

Conclusion

The analysis shows that there are many areas in Toronto that have a similar nearby business profile to that of Williamsburg, Brooklyn. In fact, there are many more (45) of those areas than areas that resemble the Upper East Side. However, the analysis is also incomplete as it does not incorporate a profile of shoppers in the areas, as well as branding and popularity information on the nearby business types.