North York Neighborhoods Final Project

1. Introduction

A Retail company is opening a new store in North York Borough. In order to do that the company is studying different neighborhood to open a new Supermarket store identifying similar localities based on the neighborhood where they already have a Store with good Sales records. In order to do that the Data Science area will use K Means clustering to identify clusters with similar (or dissimilar) localities. The Stakeholders are the Business Development Area, and they have to find the right place to build the New Store. This analysis can be extended to recommending localities to new and upcoming Supermarket. The analysis can also be extended to analyze the average spending for each locality's in terms of probable average spending, where would you be recommended to open it, and so on.

2. Data Description

The data we use has the following columns. Every Supermarket (or store) contains the following variables:

• Supermarket id: Unique id of every Supermarket across various cities of the world

• Supermarket Name: Name of the Supermarket

• Country Code: Country in which Supermarket is located

• City: City in which Supermarket is located

• Address: Address of the Supermarket

• Neighbor

• Locality: Location in the city

• Locality Verbose: Detailed description of the locality

• Longitude: Longitude coordinate of the Supermarket's location

• Latitude: Latitude coordinate of the Supermarket's location

The collected Spatial data has been stored in the Comma Separated Value. Each Supermarket in the dataset is uniquely identified by its Supermarket Id. We will use Web Scraping for the neighborhood and the Foursquare API to find different Supermarket types of each neighborhood. For this study we will use the next venue category:

• Supermarket

• Market

• Grocery Store

• Floating Market

• Farmers Market

• Fish Market

• Food & Drink Shop

• Beer Store

• Dairy Store

• Butcher

• Liquor Store

3. Methodology

In this section we will collect and prepare all the data, then we will analyse the data. Finally, we will use unsupervised machine learning technique (K means) to find similarity between neighborhood.

3.2 Web Scraping

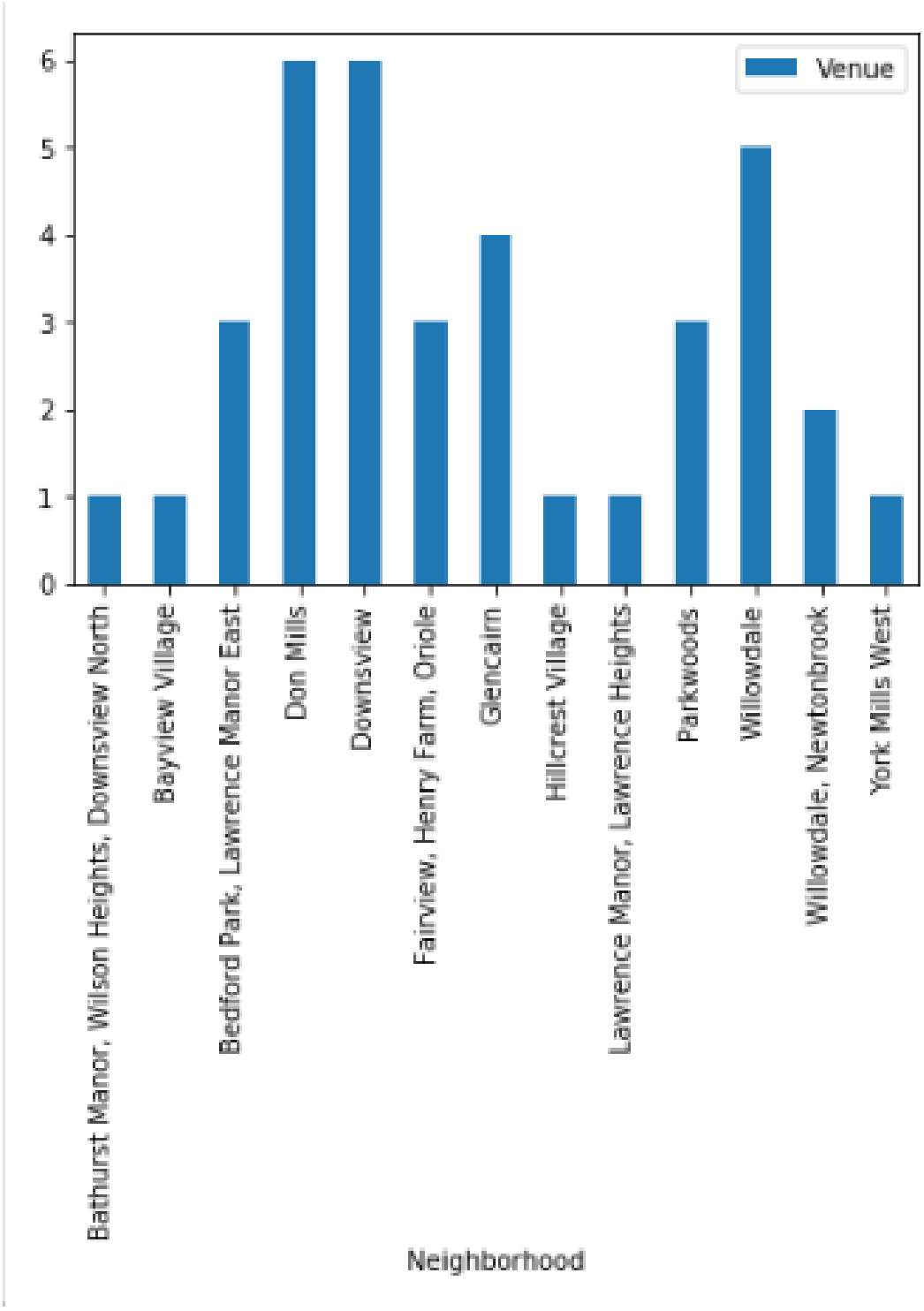
We will use the Beautifulsoup librarie to web scrap the Wikipedia table with the different neighbords. After doing this request we will filter the neighbors with NaN values and not assigned values. Then we will load and merge a .csv file with the Latitud and Longitude of each neighborhood. Finally, we will filter the “North York” Borough

3.3 Forsquare API

After building the dataframe we will use the Forsquare API to obtain the different venue categories per neighbors. After that we will filter all the interesting venue types for this analysis like:

- Supermarket
- Market
- Grocery Store
- Floating Market
- Farmers Market
- Fish Market
- Food & Drink Shop
- Beer Store
- Dairy Store
- Butcher
- Liquor Store

In the next figure we will see the amount of venue per neightbord where , for example, Don Mills and Downsview are the neightbord with more Stores.



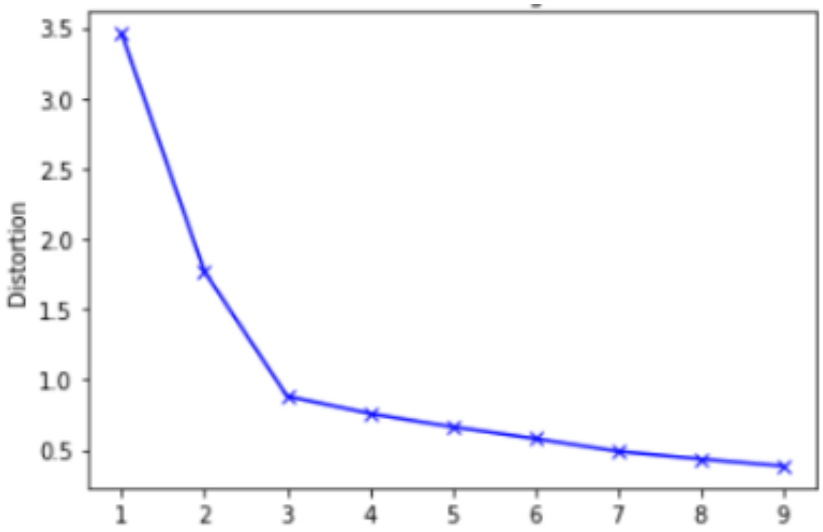
Bar Plot

Other interesting analysis is evaluate which is the most common Venue of each neighbord. Here we can see how,on one hand, in Don Mills the most common venue is Supermarket, on the other hand in Downsview the most common venue is Grocery Store.

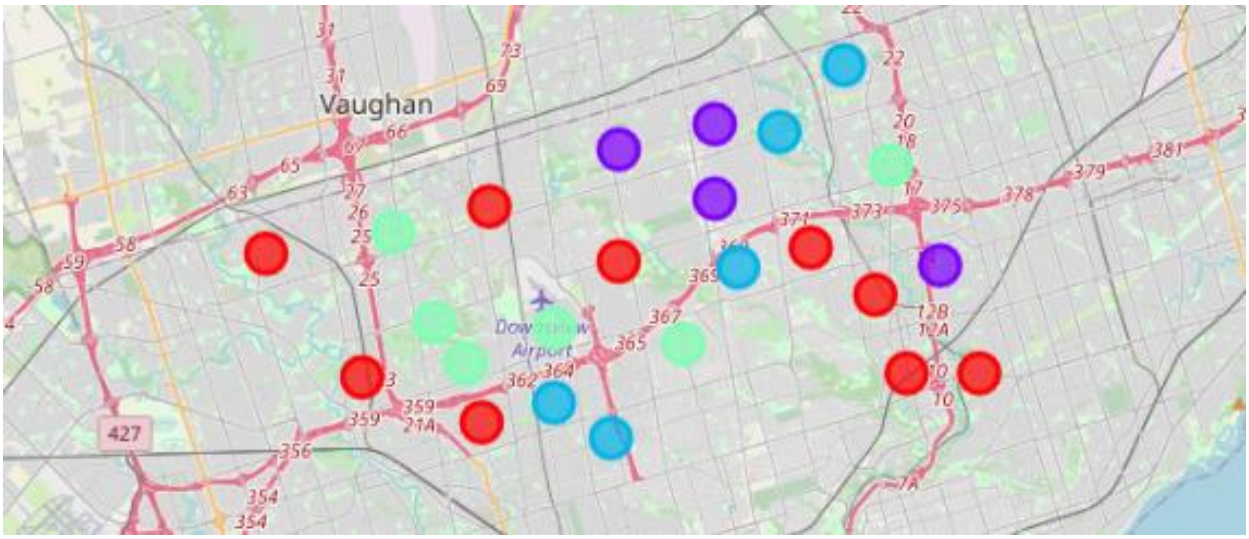
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Bathurst Manor, Wilson Heights, Downsview North	Supermarket	Liquor Store	Grocery Store
1	Bayview Village	Grocery Store	Supermarket	Liquor Store
2	Bedford Park, Lawrence Manor East	Liquor Store	Grocery Store	Butcher
3	Don Mills	Supermarket	Beer Store	Liquor Store
4	Downsview	Grocery Store	Liquor Store	Beer Store
5	Fairview, Henry Farm, Oriole	Liquor Store	Grocery Store	Beer Store
6	Glencairn	Grocery Store	Supermarket	Liquor Store
7	Hillcrest Village	Grocery Store	Supermarket	Liquor Store
8	Lawrence Manor, Lawrence Heights	Grocery Store	Supermarket	Liquor Store
9	Parkwoods	Supermarket	Grocery Store	Food & Drink Shop
10	Willowdale	Grocery Store	Supermarket	Beer Store
11	Willowdale, Newtonbrook	Supermarket	Grocery Store	Liquor Store
12	York Mills West	Grocery Store	Supermarket	Liquor Store

4. Results

After preparing the data and normilizing the feutures we use the Kmeans technique in order to find similarity between each neighboard. We will iterate diferent values of K to find the optimal value.



K-means Plot



Map Plot

5. Discussion

As we can see in the Results chapter the average distance between data points and the centroid (Error) is decreasing,however, we should use 3 different Cluster according to the “Elbow Method”.

6. Conclusion

In this project we use Web scraping technique to obtain table for internet, we consume an API to obtain and build features to train an unsupervised machine learning method as K–means and find similarity between neighborhood using only information about the Supermarket store. We also optimize the number of cluster that we should use, with all this information we can evaluate which cluster is more similar and which location is optimal to open new Store.