COURSERA CAPSTONE – THE BATTLE OF NEIGHBORHOODS

Recommending a neighborhood for a new café shop in Toronto, Canada

Problem Description

- This project is a hypothesis for an entrepreneur who would like to open a new business (a Café shop) in the capital city of Canada, Toronto, and would like to search for a suitable neighborhood for such kind of business.
- The entrepreneur would like to avoid strong competition as he/she is new to the business by targeting a neighborhood with no or little number of Café shops!

Data Acquisition and Cleaning

1. Scrapping the following Wikipedia page to get needed information about different neighborhoods of Toronto:
   - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

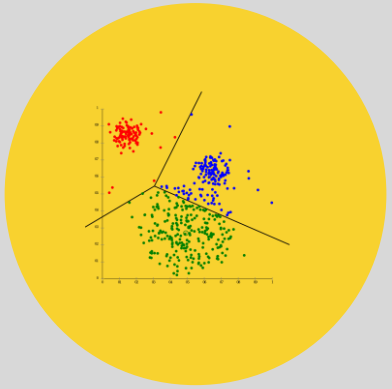
Data Acquisition and Cleaning

- Using Foursquare API to get more info about venues of each neighborhood in Toronto such as name, location and categories

Methodology

DATA ACQUISITION AND PREPARATION

EXPLORATORY DATA ANALYSIS



CLUSTERING NEIGHBORHOODS

Data Acquisition and Preparation

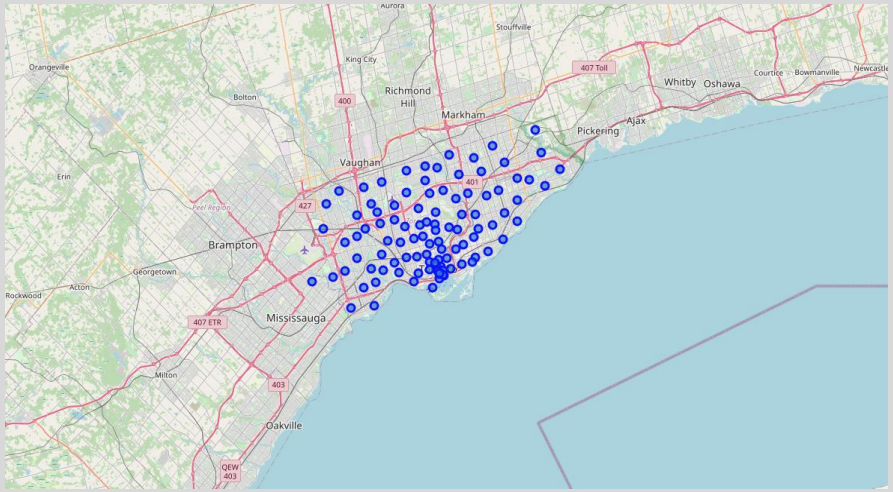
- Scrapping Wikipedia webpage to get needed information about Toronto neighborhoods
- Getting neighborhoods' coordinates

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

- Using Foursquare API to get location data

Exploratory Data Analysis

- Visualizing Toronto and its neighborhoods
- Creating a data-frame with pandas one hot encoding for the venue categories
- Calculating the mean of the frequency for every category



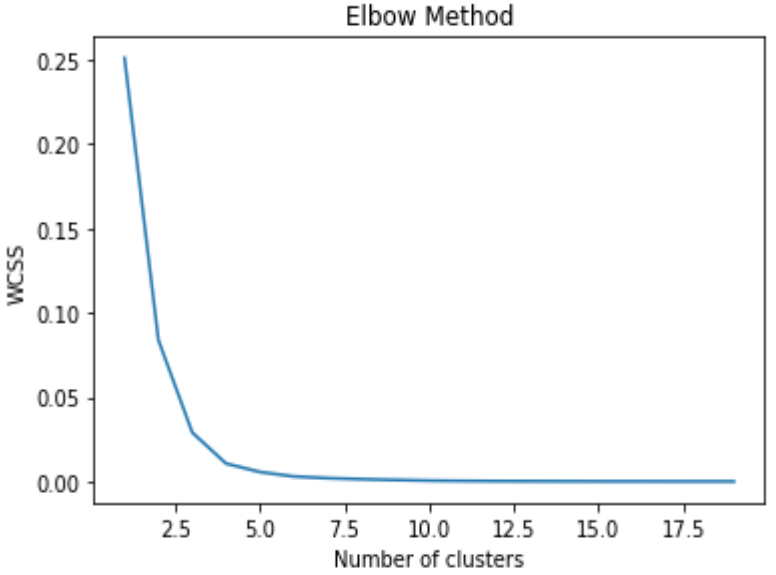
Clustering Neighborhoods

- DBScan Clustering

```
from sklearn.cluster import DBSCAN
import sklearn.utils
from sklearn.preprocessing import StandardScaler
sklearn.utils.check_random_state(1000)
Clus_dataSet = mid_df_clustering_cp.copy()
Clus_dataSet = np.nan_to_num(Clus_dataSet)
Clus_dataSet = StandardScaler().fit_transform(Clus_dataSet)

# Compute DBSCAN
db = DBSCAN(eps=0.3, min_samples=10).fit(Clus_dataSet)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_
mid_df_clustering_cp["Cluster Labels"]=labels

realClusterNum=len(set(labels)) - (1 if -1 in labels else 0)
clusterNum = len(set(labels))
print('Number of Clusters: ', clusterNum)

# A sample of clusters
mid_df_clustering_cp.head(20)
```

```
Number of Clusters:  3
```

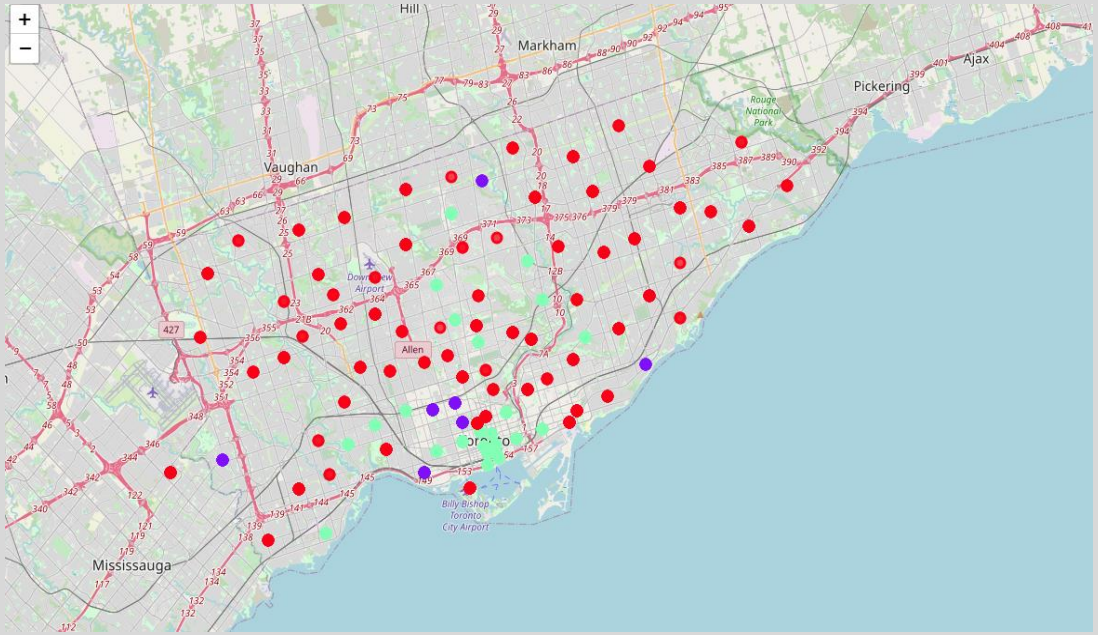
- KMean clustering



Results

The clustering output of both used clustering algorithms shows that we can cluster Toronto neighborhoods into 3 distinct clusters in terms of the number of available café shops in every neighborhood. The clusters are the following:

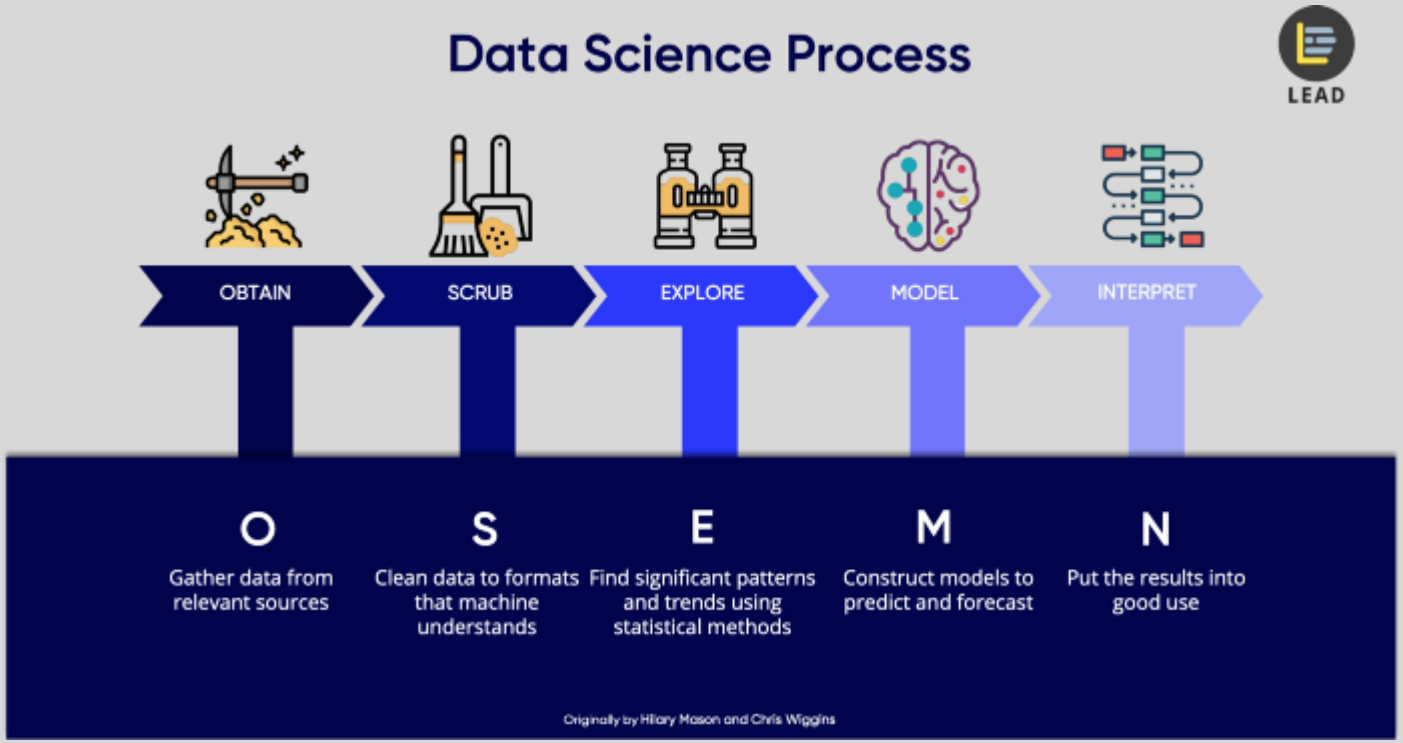
1. Cluster 0: Neighborhoods with little or no café shops
2. Cluster 1: Neighborhoods with high number of café shops
3. Cluster 2: Neighborhoods with some café shops



Discussion

- In this project we were able to use location data to visualize and analyze the neighborhoods of Toronto city, and segment and cluster these neighborhoods according to the existence and popularity of café shops.
- One main limitation of this project could be that we considered only a single factor for segmenting neighborhoods to find the most suitable ones for opening a new café shop which is the existence and frequency of café shops! However, to achieve a solid result, many other factors could be put in consideration when looking for the optimal location such as population density, average income of residents, taxes and rent costs for every neighborhood. Future research and projects could take into considerations all of these factors and possible other factors to reach better results.

Conclusion


Data Science Process
LEAD
OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
O
Gather data from relevant sources
S
Clean data to formats that machine understands
E
Find significant patterns and trends using statistical methods
M
Construct models to predict and forecast
N
Put the results into good use
Originally by Hilary Mason and Chris Wiggins