Coursera Capstone – The battle of Neighborhoods

Opening a New Café Shop in Toronto, Canada

Fatima Al Alobaidi

8/18/20

Applied Data Science Capstone

Problem Background

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.

This 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!

Problem Description

The objective of this capstone project is to use both location-based API like Foursquare API, Data Science methods and Machine Learning Algorithms to help a new entrepreneur in finding an appropriate neighborhood with no or little Café shops to open a new Café shop. This project aims to answer the following question: If a new entrepreneur would like to open a new Café shop in Toronto and avoid strong competition with other nearby Café shops, in which neighborhood he/she should consider opening it?

Target Audience

This project is targeted for new entrepreneur who wish to start a new business in Toronto by opening a new Café shop in one of its neighborhoods!

Success Criteria

The success criteria of this project is to recommend some neighborhoods or venues in Toronto for opening a Café shop where there is no or little Café shops in the same area to avoid business conflict and competition!

Data Description

This project requires exploring, segmenting and clustering neighborhoods of Toronto city.

Two main data sources were used in this project:

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
   This Wikipedia page provides an html table with the list of postal codes along with boroughs and neighborhoods of Toronto city.
2. Foursquare API to get more info about venues of each neighborhood in Toronto such as name, location and categories

Methodology

Data Preparation

Scrapping Wikipedia webpage to get needed information about Toronto neighborhoods

We started data preparation by scrapping the following Wikipedia webpage using *BeautifulSoup* Python library:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Then, a get request was made to fetch the raw HTML contents. After that, we parsed the html content and read the target table from the scrapped Wikipedia page, which contains lists of postal codes and corresponding neighborhoods and boroughs of Toronto, and convert it to pandas dataframe.

The created dataframe looks like the following:

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Getting neighborhoods' coordinates

Geopy library was used to retrieve the latitude and longitude of every neighborhood in our dataframe resulting in the following dataframe:

	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

At this stage, we make use of *Foursquare API* to get needed information about the top 100 venues for every neighborhood in Toronto within a radius of 500 meters.

Exploratory Data Analysis

Visualizing Toronto and its neighborhoods

Using Folium library, we create a map of Toronto and added marker for every neighborhood in Toronto to better visualize the city of Toronto and its various neighborhoods.

Creating a data-frame with pandas one hot encoding for the venue categories

Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(toronto_venues['Venue Category'].unique())))
```

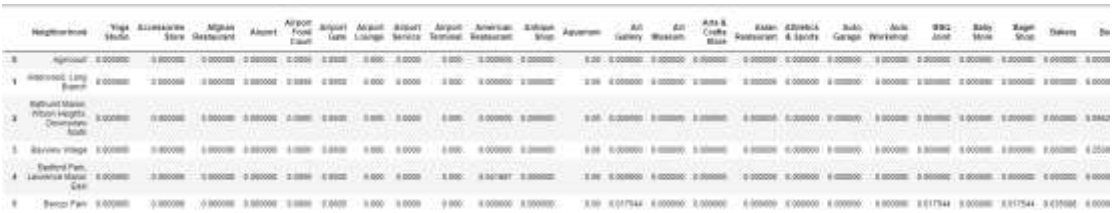
```
There are 273 uniques categories.
```

Analyzing Each Neighborhood

```
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]

toronto_onehot.head()
```

calculating the mean of the frequency for every category

At this stage, we group our dataframe on neighborhood column and calculate the mean of the frequency of occurrence of each venue category.



Machine Learning Approach

Both DBscan and KMean clustering where used and compared to cluster the venues and neighborhoods of Toronto based on mean frequency of number of café shops.

DBScan Clustering

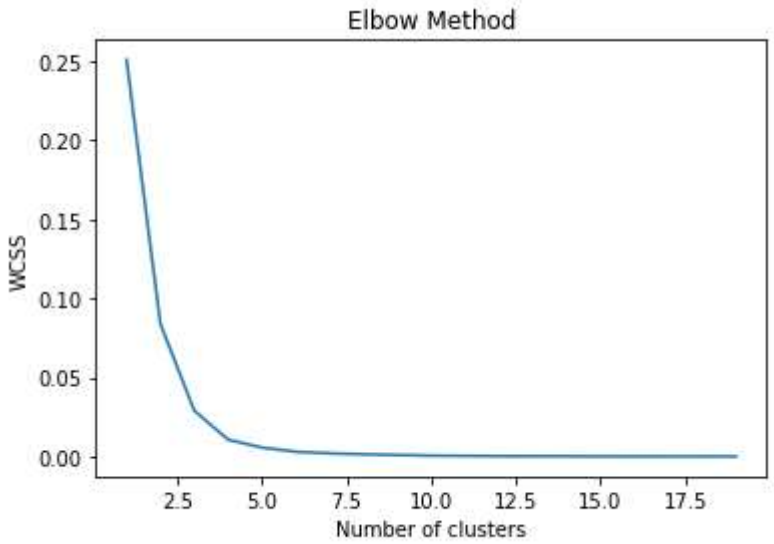
The following code was used to apply cluster Toronto neighborhoods using DBScan which resulted in creating 3 different clusters.

```
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

Elbow method were used to find the optimal k, which was found to be 3 as illustrated by the following figure.



Both clustering algorithms gave the same number of clusters (3) and resulted in almost the same clusters which give more confidence on the found results and clusters!

The 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

The clusters are visualized in the above map with cluster 0 in green color, cluster 1 in purple color, and cluster 2 in red color.

Observations and Recommendations

Café shops in Toronto are concentrated in the second cluster which contains the following neighborhoods: Bayview Village, Birch Cliff, Cliffside West, Brockton, Parkdale Village, Exhibition Place, Christie, Eringate, Bloordale Gardens, Old Burnhamthorpe, The Annex, North Midtown, Yorkville and University of Toronto, Harbord. The third cluster contains some Café shops, but they are much less than the second cluster. However, venues and neighborhoods in the first cluster rarely have Café shops, and most of them do not have a single one!

Putting these results in mind in addition to the target condition of the entrepreneur who would like to avoid any competition by targeting neighborhoods with little or no Café shops, we can clearly say that the entrepreneur should avoid neighborhoods and venues of the second cluster and look for a good place in either cluster 1 or 3 (preferably cluster 1) to avoid any possible competition from nearby 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

In this project we have applied the Data Science lifecycle by identifying and understanding business problem, gathering required data, cleaning and wrangling data, analyzing data, performing appropriate machine learning algorithm by using and comparing both DBScan and KMean clustering algorithms, and finally analyzing and visualizing identified clusters.