Capstone Project - The Battle of Neighborhoods

IBM Data Science Professional Certificate

Ebrima B Sawaneh

1. Introduction/Business Problem

The objective of this project is to help the management of Banjul Pizza Ltd. The firm is considering the possibility of opening Pizza restaurants in Toronto Canada. The initial research shows that Toronto is ethnically diverse with people from different background who loves pizza. The management want to find out the best possible place to open their first pizza restaurant. As one of the largest cities in Canada, with immigrant cultures, Toronto is a business-minded city which already have many pizza restaurants. Therefore, the aim is to open the first restaurant in a neighborhood with lesser pizza restaurants thereby a possible lesser competition too.

As a member of the project team and data science professional, I have been tasked to offer solution to the problem of opening the first restaurant in the right place with less potential competition in the city of Toronto.

2. Data Sources

This project will make use of data from sources to offer a data driven solution. Therefore, we will focus on data collection that are relevant to Toronto. For instance, we will need to know all the borough in Toronto with related neighborhoods and post code. We all need to geo location data such as longitude and latitude of each place too. The source below will offer data needed for the analysis:

- Wikipedia - There is a special page on the Wikipedia that has data about borough, neighborhood and postal code of Toronto.
- Toronto Geospace information – Initial review shows that Wikipedia data does not contain information about longitude and latitude. Therefore, we will use downloaded CVS file that already has the geo data for all neighborhoods in Toronto.
- Foursquare location data– Now we data about the surrounding venues near each location. We will use the Foursqure API tool to collect venues that are nearby each geo location.

These three sources will supply us with data that will be transform and explore using relevant data science methodologies.

3. Methodology

3.1 Data collection

The data of borough, neighborhood and postal code were first collected from the Wikipedia page and converted into table using BeautifulSoup package. I then put the table into a Pandas data frame. The initial data shows that a number of boroughs have not been assigned a name. I therefore drop such names from the dataf rame. Additionally, some neighborhoods have no name and I replace those NaN with the name of the borough. The next key data source was the geolocation of postal codes (https://cocl.us/Geospatial_data) in Toronto into another table. Then I combined the borough table and geolocation into one data frame.

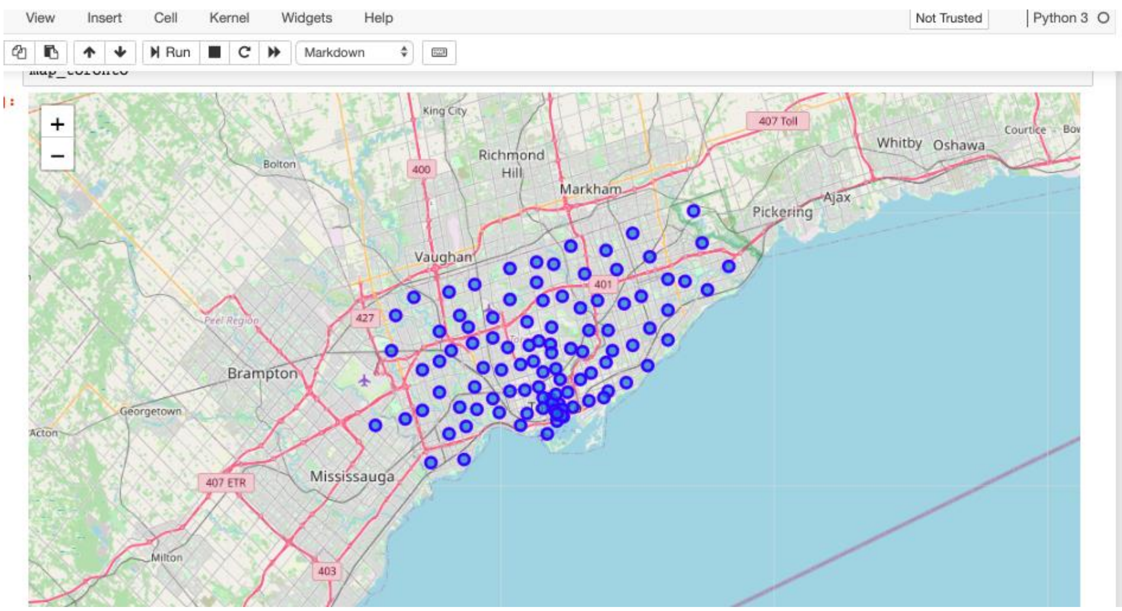
In [68]:
```
# Now merging both DataFrames
torontodata = pd.merge(df2,geo,on='Postal Code')
torontodata.head()
```

Out[68]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

3.2 Data exploration

Some basic values were calculated including the shape of the table, the number of unique boroughs etc. I then used the Folium to display the neighborhood on the map of Toronto.



I also used the Foursquare API from my free account to access venues within 1000 from radius. The report limits to only 100 venue. The collected data from Json file are converted into Panda data frame. Then I filter pizza venues only and merge the results with the first data of borough with geolocation.

In [112]:
```python
# Let's sort the results by Cluster Labels
print(toronto_Pizza_Clustering_merged2.shape)
toronto_Pizza_Clustering_merged2.sort_values(["Cluster Labels"], inplace=True)
toronto_Pizza_Clustering_merged2
```

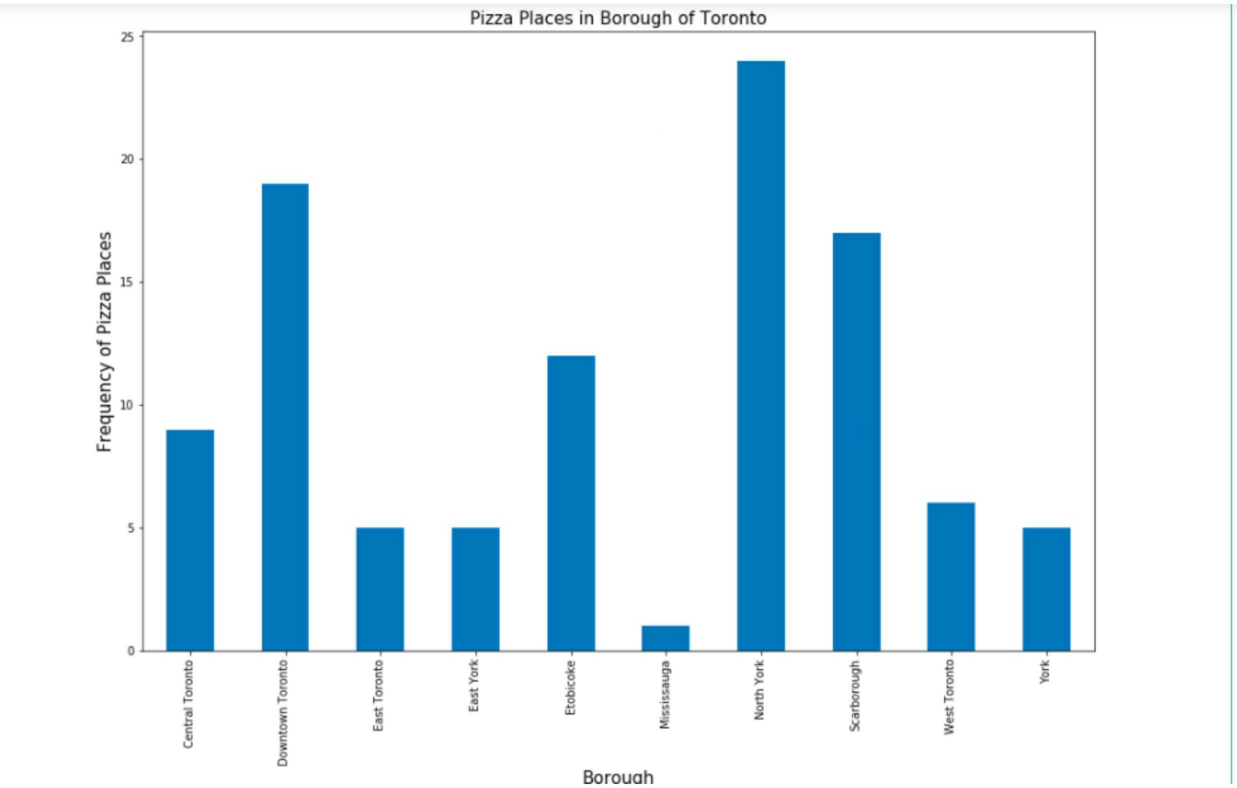
```
(103, 7)
```

Out[112]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Pizza Place	Cluster Labels
51	M6L	North York	North Park, Maple Leaf Park, Upwood Park	43.713756	-79.490074	0.058824	0
40	M3K	North York	Downsview	43.737473	-79.464763	0.053571	0
43	M3N	North York	Downsview	43.761631	-79.520999	0.053571	0
38	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029	0.050000	0
44	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	0.043478	0
35	M4J	East York	East Toronto, Broadview North (Old East York)	43.685347	-79.338106	0.048193	0
32	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	0.060606	0
52	M9L	North York	Humber Summit	43.756303	-79.565963	0.045455	0
53	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476	0.045455	0
55	M4M	East Toronto	Studio District	43.659526	-79.340923	0.050000	0
27	M2H	North York	Hillcrest Village	43.803762	-79.363452	0.052632	0

The number of venues in neighborhood were converted into dummy variables and grouped by taking the mean of frequency of occurrence.

A filter of pizza venues was put into another data frame and I used Matplotlib to display pizza venue per borough.



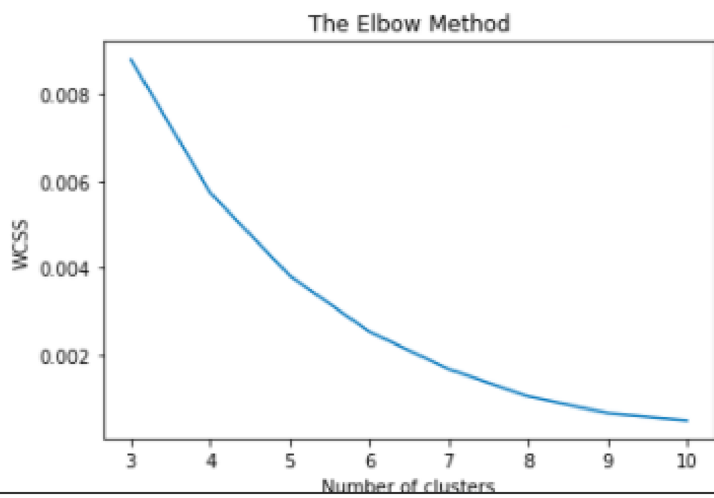
4. Clustering

Considering the need to explore venue density within a location, K-mean clustering was the most suitable method to achieve the result. Therefore, an elbow method was used to assess the suitable number of clusters, which gives me 6.

The K-mean clustering was initiated, and result cluster was merge with other table from the exploratory stage.

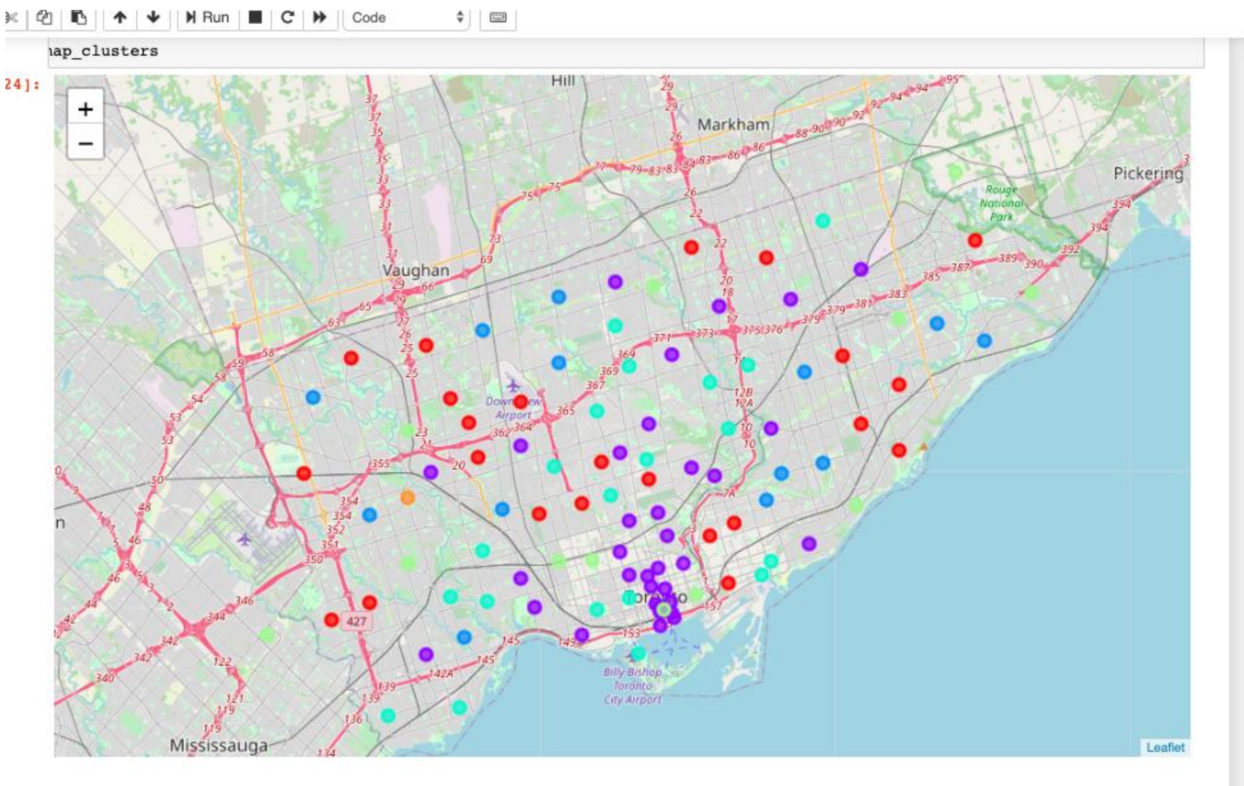
[105]:

```
# Using the elbow method to find the optimal number of clusters
# import k-means from clustering stage
from sklearn.cluster import KMeans

# Matplotlib and associated plotting modules
from matplotlib import pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors

toronto_Pizza_Clustering = toronto_Pizza.drop('Neighborhood', 1)


wcss = []

for i in range(3, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter= 50)
    kmeans.fit(toronto_Pizza_Clustering)
    wcss.append(kmeans.inertia_)
plt.plot(range(3, 11), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



5. Results

The folium map was used to display neighborhood by cluster.



The of the analysis from different tables were summarized by weighting of the number of venues in each cluster.

Let's group the results by summing the Cluster Labels

```
In [120]: toronto_Pizza_Clustering_merged2.groupby(['Cluster Labels'])['Pizza Place'].agg('sum')

Out[120]: Cluster Labels
          0    1.206009
          1    0.467172
          2    1.020216
          3    0.656773
          4    0.000000
          5    0.121212
          Name: Pizza Place, dtype: float64
```

6. Discussion & Conclusion

The management of firm wants to avoid heavy competition against existing Pizza. Therefore, they want to open their first Pizza restaurant in locations where there as less pizza restaurants. We can see from the K-Mean clustering results that Cluster 4 is the cluster which has less Pizza venues while cluster 3 has more pizza venues. I will then conclude that cluster 4 should be the firms first Pizza location. However, the firm should do further research about the boroughs within the cluster as there are other variables that may impact sales such as traffic, population, crime rate etc.