Coursera Data Science Capstone Project: Battle of the Neighborhoods

Introduction: In this project I will be helping a business investor decide which part of Toronto to start his Coffee Shop outlet. In this project I will consider three major factors that will help determine the best neighborhood to start a coffee shop in Toronto. These factors are:

- Viability: This refers to how feasible opening a coffee shop in any of the neighborhoods is. This viability index will be created based on the existing coffee shops in that region. This is based on the assumption that if there is a reasonable amount of coffee shops in a region it implies that the business is relatively profitable.
- Environmental Factors: Coffee shops do not stand in isolation. The presence of various establishments such as banks, sports centers, parks etc. within a region help to drive more traffic in such neighborhoods which can improve the conditions for the proposed coffee shop.
- Population Density: People are constantly looking for a 'third place' to spend to spend time that's neither work or home, so the more the people in a given location the more likely it is for the coffee shop to attract traffic.

In this project I will also try to analyze the various parts of Toronto city to see similarities between neighborhoods and discuss the upsides to moving into any of these Neighborhoods.

Who should be interested in this project:

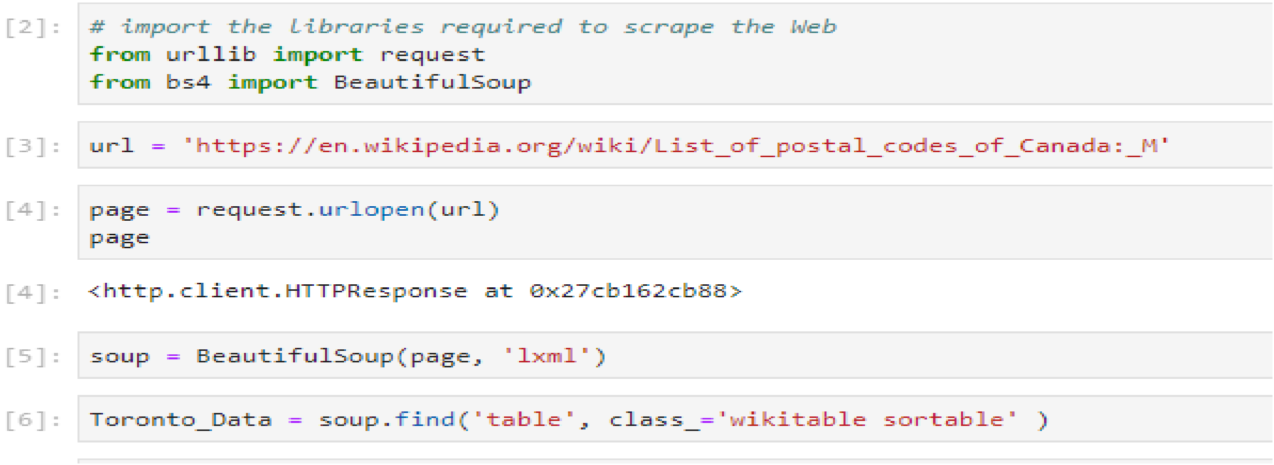
- Newbie Coffee shop investors who want to have a rough idea of how coffee Shops have been thriving in Neighborhoods in Toronto and which Neighborhood to startup in.
- People interested in the city of Toronto who need a broad overview of the nature of the neighborhoods in Toronto

Data Sources: In this project four major data sources were used.

- Neighborhood Data: This data contains a list of all the Neighborhoods in Toronto alongside their corresponding Boroughs and postal codes. This data was scraped from this Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) using the BeautifulSoup python library.
- Location Data: This data contains the Longitude and Latitude of all the neighborhoods in Toronto. This Data was generated using the Geocoder python package.
- Venue Data: This data contains a list of the common venues within each of the neighborhoods. This data was obtained using the Foursquare API.
- Population Density Data: This data contains information about the population densities of certain neighborhoods in the city of Toronto. This data was obtained from this Wikipedia page (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods).

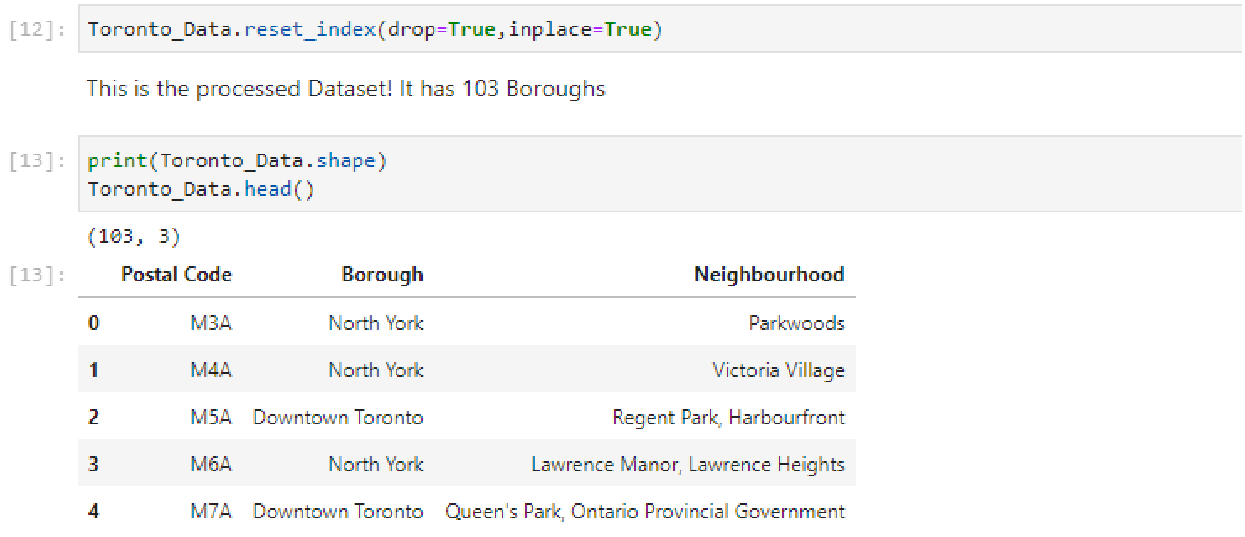
Methodology:

The first step in this project involved obtaining the Neighborhood Data from the this Wikipedia page(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). This data contains a list of all the neighborhood in Toronto alongside their corresponding Postal Codes. This data was scraped using the BeauifulSoup python package, converted to a pandas dataframe and then cleaned so as to make it possible to process and analyze.



```
[2]: # import the libraries required to scrape the Web
     from urllib import request
     from bs4 import BeautifulSoup

[3]: url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

[4]: page = request.urlopen(url)
     page

[4]: <http.client.HTTPResponse at 0x27cb162cb88>

[5]: soup = BeautifulSoup(page, 'lxml')

[6]: Toronto_Data = soup.find('table', class_='wikitable sortable' )
```

Figure 1 Obtaining the Neighborhood Data Using BeautifulSoup



```
[12]: Toronto_Data.reset_index(drop=True,inplace=True)
```

This is the processed Dataset! It has 103 Boroughs

```
[13]: print(Toronto_Data.shape)
      Toronto_Data.head()
```

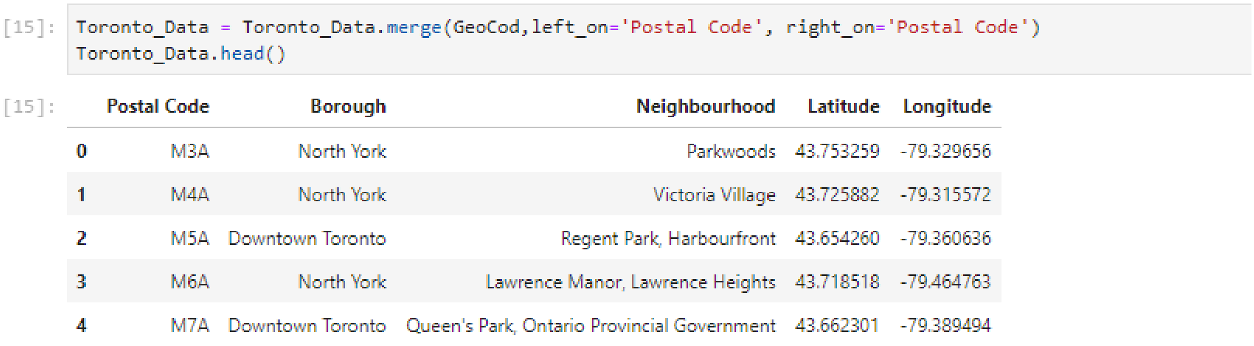
```
(103, 3)
```

[13]:

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2 The processed data in the Pandas Dataframe

The Location Data (Longitude and Latitude) from the Geocoder package was joined to the Dataframe so as to make it possible to make calls to the FourSquare API.



```
[15]: Toronto_Data = Toronto_Data.merge(GeoCod,left_on='Postal Code', right_on='Postal Code')
      Toronto_Data.head()
```

[15]:

	Postal Code	Borough	Neighbourhood	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

Figure 3 DataFrame with the Longitude and Latitude data added

A visualization of these Neighborhoods was generated using the Folium Package.

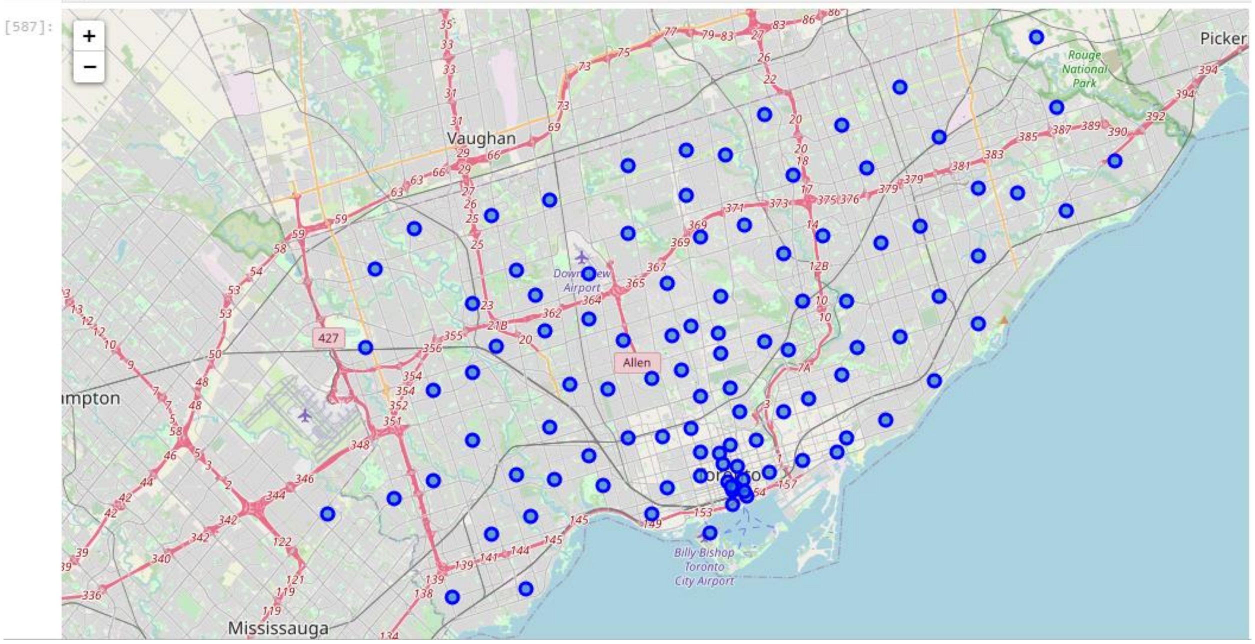
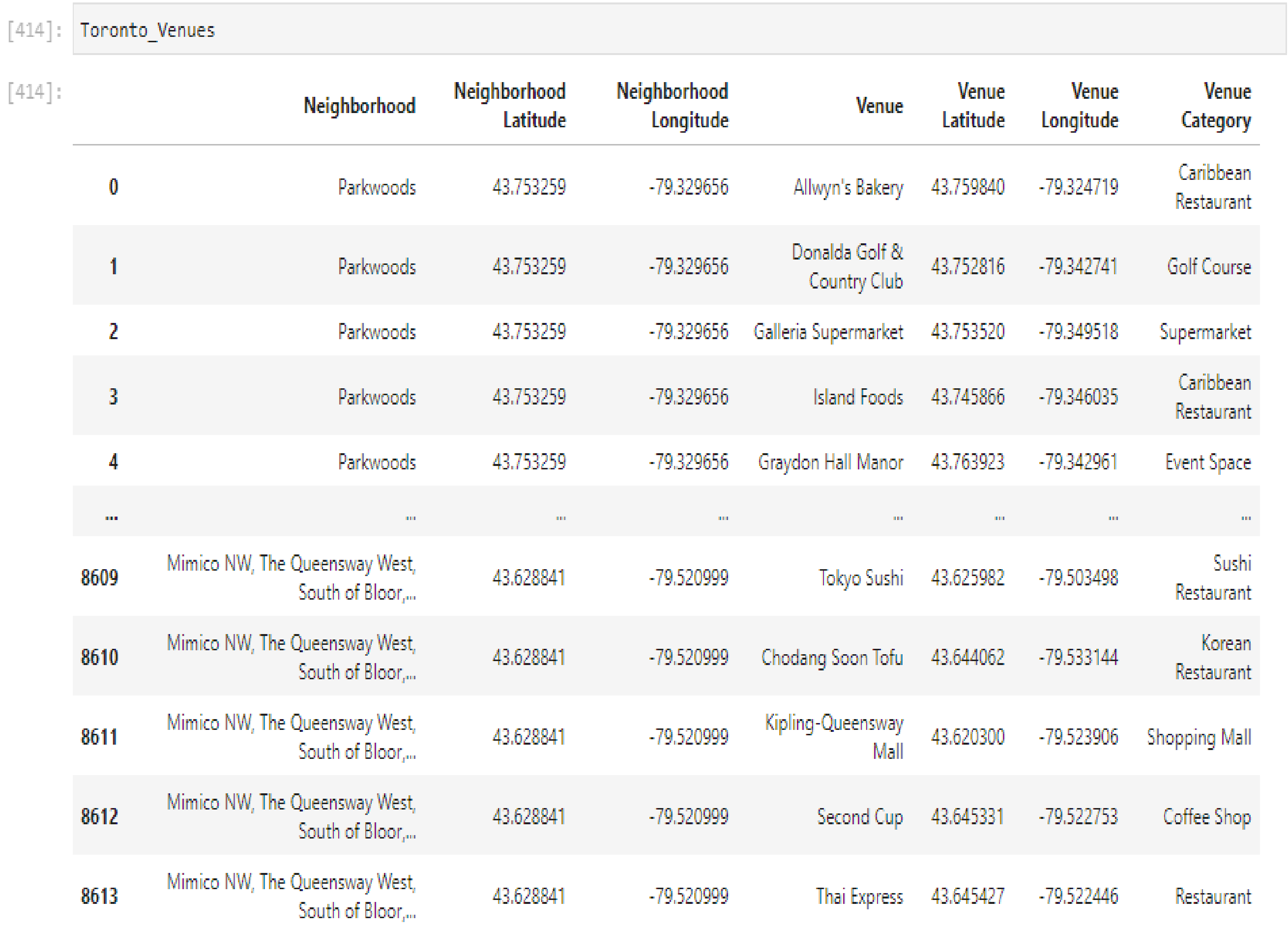


Figure 4 neighborhoods in Toronto

The Venue Data was obtained from the Foursquare API. This data contains the top 200 venues within a 2000-meter radius of each neighborhood. It returned 8615 venues for all the neighborhoods and 319 unique venues within these neighborhoods. Of these 8615 venues returned, the most common venue category was the coffee shop category.

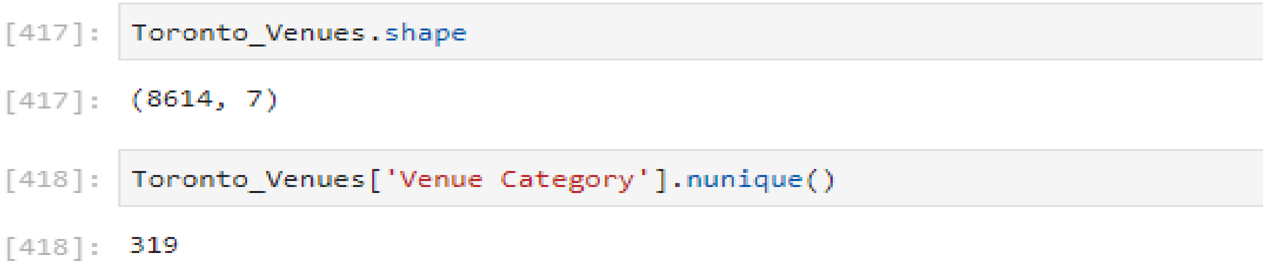


```
[414]: Toronto_Venues
```

[414]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant
1	Parkwoods	43.753259	-79.329656	Donalda Golf & Country Club	43.752816	-79.342741	Golf Course
2	Parkwoods	43.753259	-79.329656	Galleria Supermarket	43.753520	-79.349518	Supermarket
3	Parkwoods	43.753259	-79.329656	Island Foods	43.745866	-79.346035	Caribbean Restaurant
4	Parkwoods	43.753259	-79.329656	Graydon Hall Manor	43.763923	-79.342961	Event Space
...	...	...	...	...	...	...	...
8609	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Tokyo Sushi	43.625982	-79.503498	Sushi Restaurant
8610	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Chodang Soon Tofu	43.644062	-79.533144	Korean Restaurant
8611	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kipling-Queensway Mall	43.620300	-79.523906	Shopping Mall
8612	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Second Cup	43.645331	-79.522753	Coffee Shop
8613	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Thai Express	43.645427	-79.522446	Restaurant

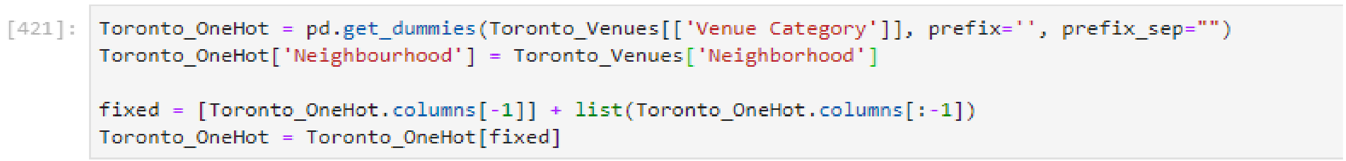
Figure 5 List of the top 200 Venues in all the neighborhoods in Toronto



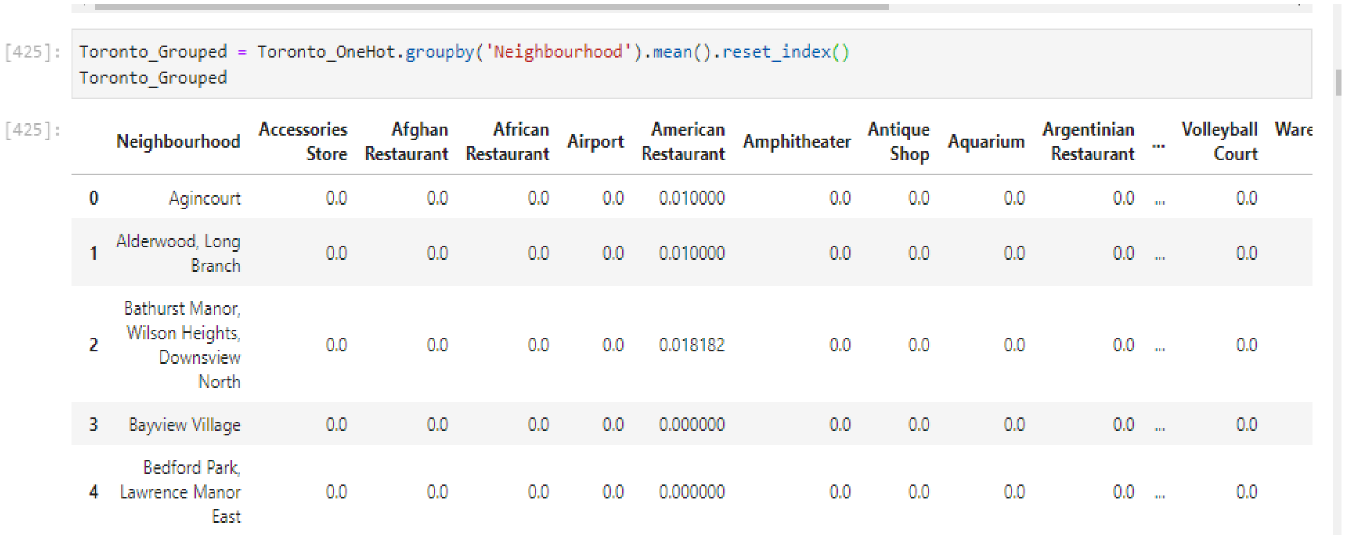
```
[417]: Toronto_Venues.shape

[417]: (8614, 7)

[418]: Toronto_Venues['Venue Category'].nunique()

[418]: 319
```

Figure 6 319 Unique Venue categories

Using the method of one hot encoding these distinct venues were converted to a binary series and grouped by neighborhood. This was done to make it easy for machine learning to work with the data.



```
[421]: Toronto_OneHot = pd.get_dummies(Toronto_Venues[['Venue Category']], prefix='', prefix_sep="")
       Toronto_OneHot['Neighbourhood'] = Toronto_Venues['Neighborhood']

       fixed = [Toronto_OneHot.columns[-1]] + list(Toronto_OneHot.columns[:-1])
       Toronto_OneHot = Toronto_OneHot[fixed]
```



```
[425]: Toronto_Grouped = Toronto_OneHot.groupby('Neighbourhood').mean().reset_index()
       Toronto_Grouped
```

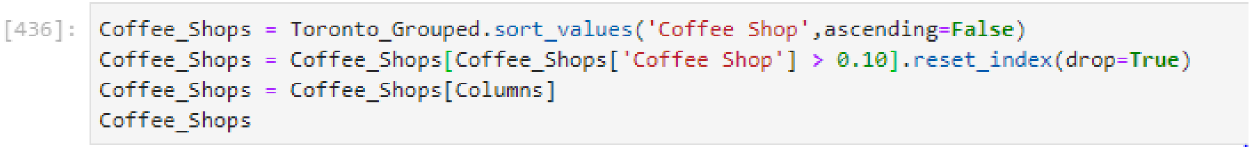
[425]:

	Neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Amphitheater	Antique Shop	Aquarium	Argentinian Restaurant	...	Volleyball Court	Ware
0	Agincourt	0.0	0.0	0.0	0.0	0.010000	0.0	0.0	0.0	0.0	...	0.0	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.010000	0.0	0.0	0.0	0.0	...	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.018182	0.0	0.0	0.0	0.0	...	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	

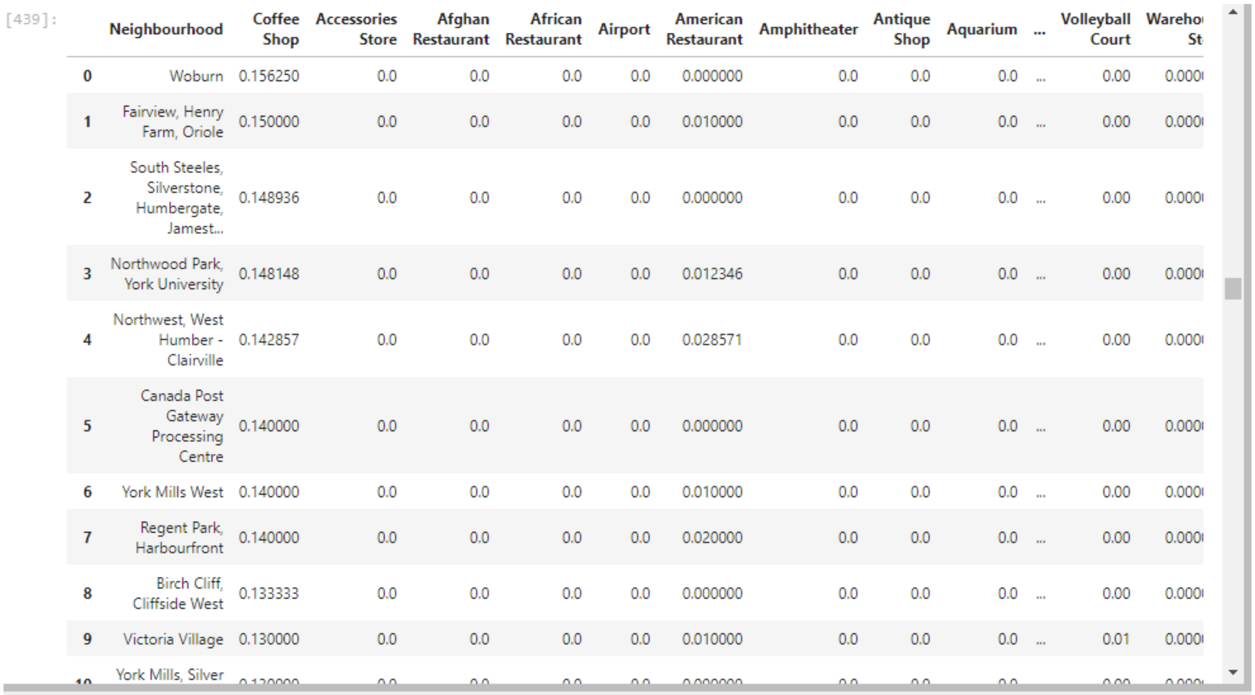
Figure 7 Encoded Data grouped by Neighborhood

At this point all the data required for this project have been obtained, cleaned and processed.

In this phase only neighborhoods that have at least 10% of their generated venues as coffee shops are considered. This narrowed the total number of neighborhoods to be considered down from 99 to 10.



```
[436]: Coffee_Shops = Toronto_Grouped.sort_values('Coffee Shop',ascending=False)
       Coffee_Shops = Coffee_Shops[Coffee_Shops['Coffee Shop'] > 0.10].reset_index(drop=True)
       Coffee_Shops = Coffee_Shops[Columns]
       Coffee_Shops
```



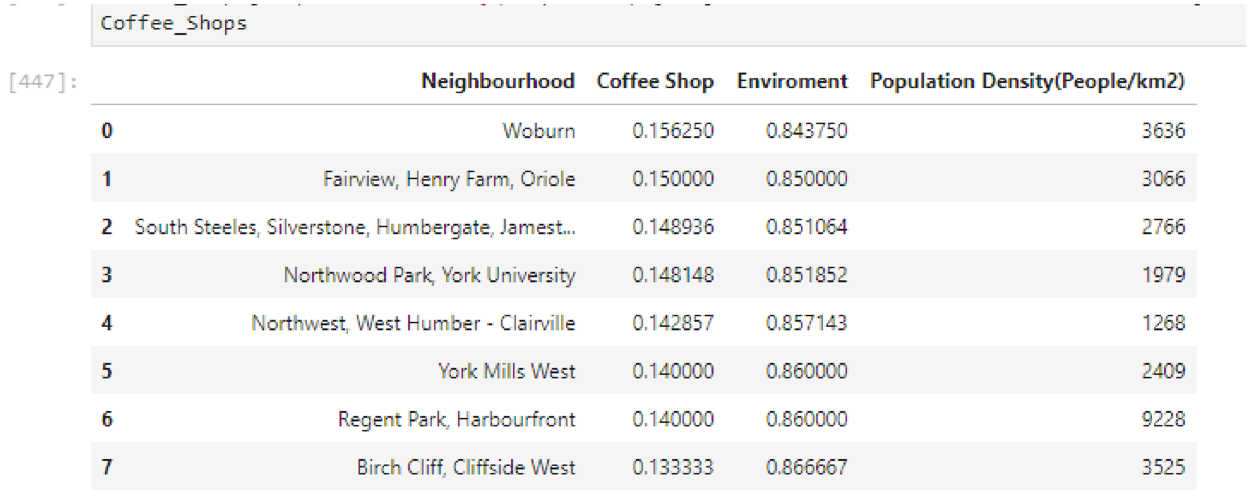
[439]:

	Neighbourhood	Coffee Shop	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Amphitheater	Antique Shop	Aquarium	...	Volleyball Court	Wareho St
0	Woburn	0.156250	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.00	0.000
1	Fairview, Henry Farm, Oriole	0.150000	0.0	0.0	0.0	0.0	0.010000	0.0	0.0	0.0	...	0.00	0.000
2	South Steeles, Silverstone, Humbergate, Jamest...	0.148936	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.00	0.000
3	Northwood Park, York University	0.148148	0.0	0.0	0.0	0.0	0.012346	0.0	0.0	0.0	...	0.00	0.000
4	Northwest, West Humber - Clairville	0.142857	0.0	0.0	0.0	0.0	0.028571	0.0	0.0	0.0	...	0.00	0.000
5	Canada Post Gateway Processing Centre	0.140000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.00	0.000
6	York Mills West	0.140000	0.0	0.0	0.0	0.0	0.010000	0.0	0.0	0.0	...	0.00	0.000
7	Regent Park, Harbourfront	0.140000	0.0	0.0	0.0	0.0	0.020000	0.0	0.0	0.0	...	0.00	0.000
8	Birch Cliff, Cliffside West	0.133333	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.00	0.000
9	Victoria Village	0.130000	0.0	0.0	0.0	0.0	0.010000	0.0	0.0	0.0	...	0.01	0.000
10	York Mills, Silver	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	...	[illegible]	[illegible]

Figure 8 Neighborhoods which have at least 10% of their trending Venues as Coffee shops

The Environment data for each of these shortlisted neighborhoods was obtained by comparing the nature of the sites they have around them. Also the population density of each of these neighborhoods was obtained. These values were then combined and used to form a comparison index.

Note that two of these 10 neighborhoods were dropped as their population density data were not readily accessible, hence only 8 neighborhoods are in the shortlist.

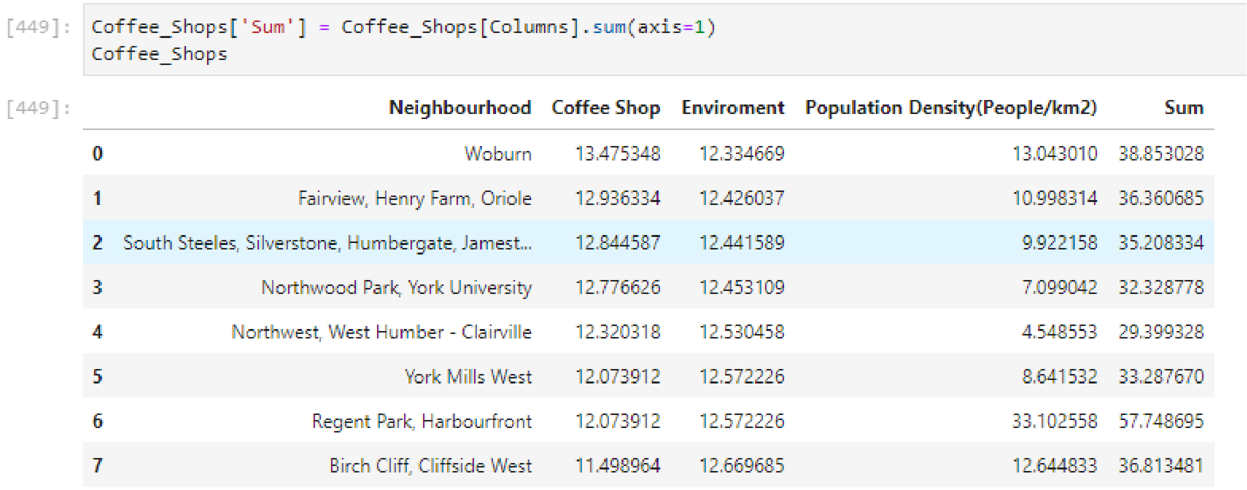


```
Coffee_Shops
```

[447]:

	Neighbourhood	Coffee Shop	Enviroment	Population Density(People/km2)
0	Woburn	0.156250	0.843750	3636
1	Fairview, Henry Farm, Oriole	0.150000	0.850000	3066
2	South Steeles, Silverstone, Humbergate, Jamest...	0.148936	0.851064	2766
3	Northwood Park, York University	0.148148	0.851852	1979
4	Northwest, West Humber - Clairville	0.142857	0.857143	1268
5	York Mills West	0.140000	0.860000	2409
6	Regent Park, Harbourfront	0.140000	0.860000	9228
7	Birch Cliff, Cliffside West	0.133333	0.866667	3525

Figure 9 Shortlisted neighborhoods



[449]:
```
Coffee_Shops['Sum'] = Coffee_Shops[Columns].sum(axis=1)
Coffee_Shops
```

[449]:

	Neighbourhood	Coffee Shop	Enviroment	Population Density(People/km2)	Sum
0	Woburn	13.475348	12.334669	13.043010	38.853028
1	Fairview, Henry Farm, Oriole	12.936334	12.426037	10.998314	36.360685
2	South Steeles, Silverstone, Humbergate, Jamest...	12.844587	12.441589	9.922158	35.208334
3	Northwood Park, York University	12.776626	12.453109	7.099042	32.328778
4	Northwest, West Humber - Clairville	12.320318	12.530458	4.548553	29.399328
5	York Mills West	12.073912	12.572226	8.641532	33.287670
6	Regent Park, Harbourfront	12.073912	12.572226	33.102558	57.748695
7	Birch Cliff, Cliffside West	11.498964	12.669685	12.644833	36.813481

Figure 10 Comparison Index for the shortlisted neighborhoods

Finally, the KMeans clustering algorithm was run on the whole dataset to try and group the various neighborhoods based on their similarity to each other. The top 10 venues for each neighborhood was extracted so as to make the cluster comparison easier and then the clusters were then created. The best run of the algorithm produced a total of three clusters.

[586]:

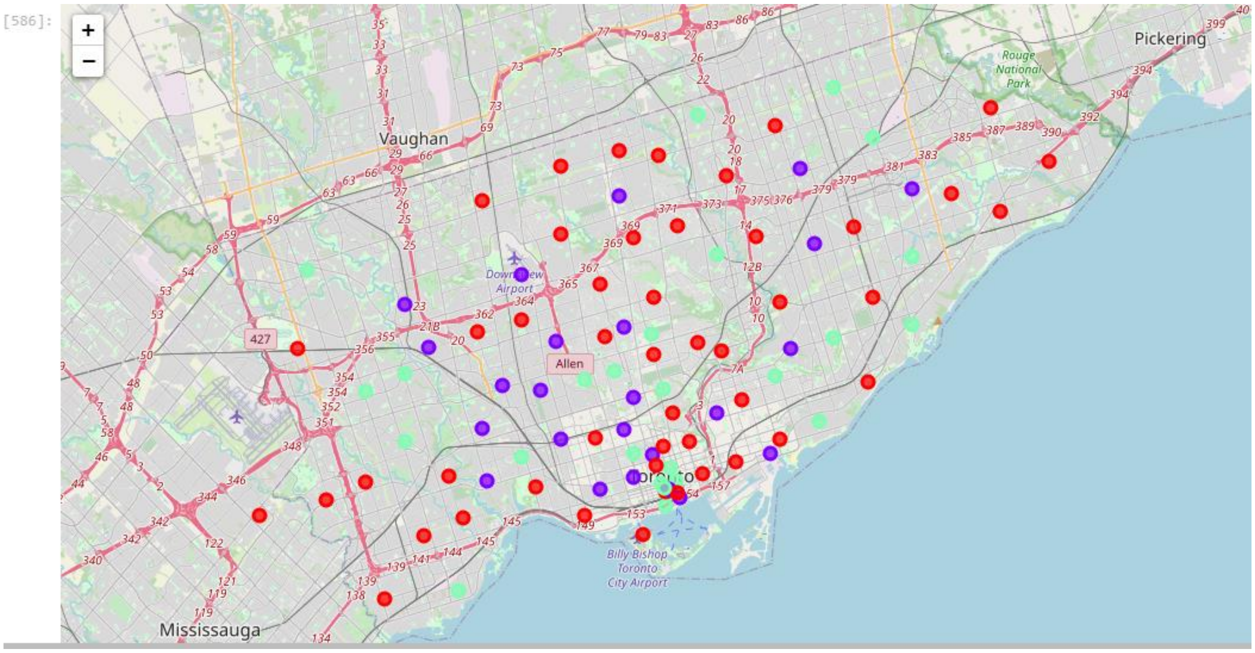
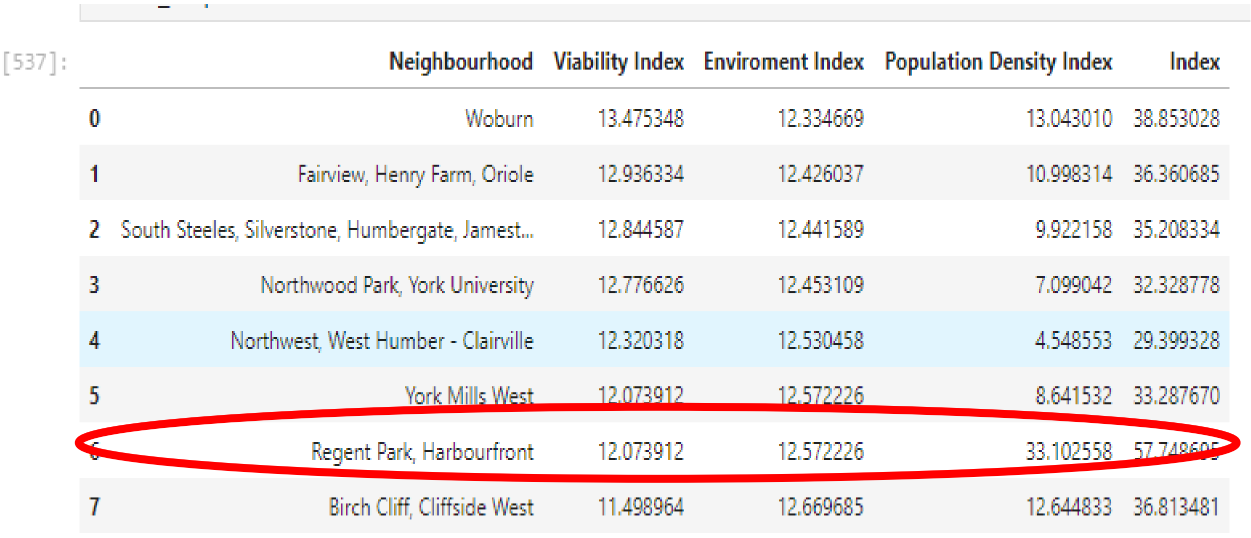


Figure 11 Image showing the three clusters based on the nature of their Venues

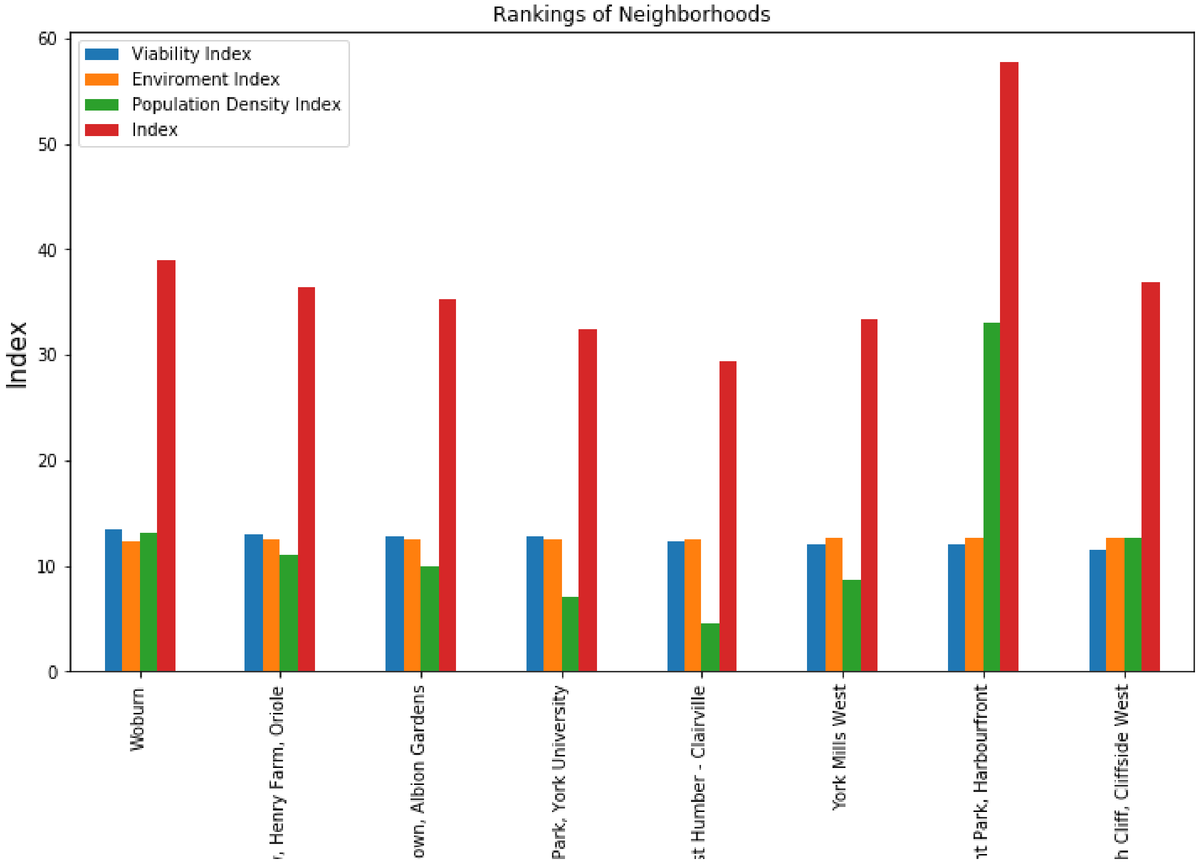
Results and Discussion

For the first part of the project the best neighborhoods for the coffee shop have been narrowed down to 8 neighborhoods. The best neighborhood can then be selected based on the index created.



[537]:

	Neighbourhood	Viability Index	Enviroment Index	Population Density Index	Index
0	Woburn	13.475348	12.334669	13.043010	38.853028
1	Fairview, Henry Farm, Oriole	12.936334	12.426037	10.998314	36.360685
2	South Steeles, Silverstone, Humbergate, Jamest...	12.844587	12.441589	9.922158	35.208334
3	Northwood Park, York University	12.776626	12.453109	7.099042	32.328778
4	Northwest, West Humber - Clairville	12.320318	12.530458	4.548553	29.399328
5	York Mills West	12.073912	12.572226	8.641532	33.287670
6	Regent Park, Harbourfront	12.073912	12.572226	33.102558	57.748695
7	Birch Cliff, Cliffside West	11.498964	12.669685	12.644833	36.813481



Based on the Index Harbourfront, Regent Park is the best neighborhood to start a coffee shop. It is important to note that this location has a high population density and the current number of coffee shops in this neighborhood are above average. Based on the assumptions used, this implies that the coffee shops in this neighborhoods seem to be thriving and the competition is not excessive, thereby making this neighborhood the ideal location to start a coffee shop.

Clustering Results

The clustering algorithm generated three clusters.

Cluster 0: This cluster consists majorly of a variety of restaurants, such as Japanese restaurants, Vietnamese restaurants, Middle Eastern restaurants, Sushi restaurants and Korean restaurants. This cluster has a total of 48 neighborhoods. It seems like a good place if you want to experience various cultures in Toronto.

Cluster 1: This cluster consists majorly of parks, banks, clothing stores furniture stores and hotels. This cluster has a total of 24 neighborhoods.

Cluster 2: This cluster consists majorly of Shopping malls, Liquor store, banks, pharmacies, hotels and gastropubs. This cluster has a total of 25 neighborhoods. The neighborhoods in this cluster seem to be a good location for job seekers.

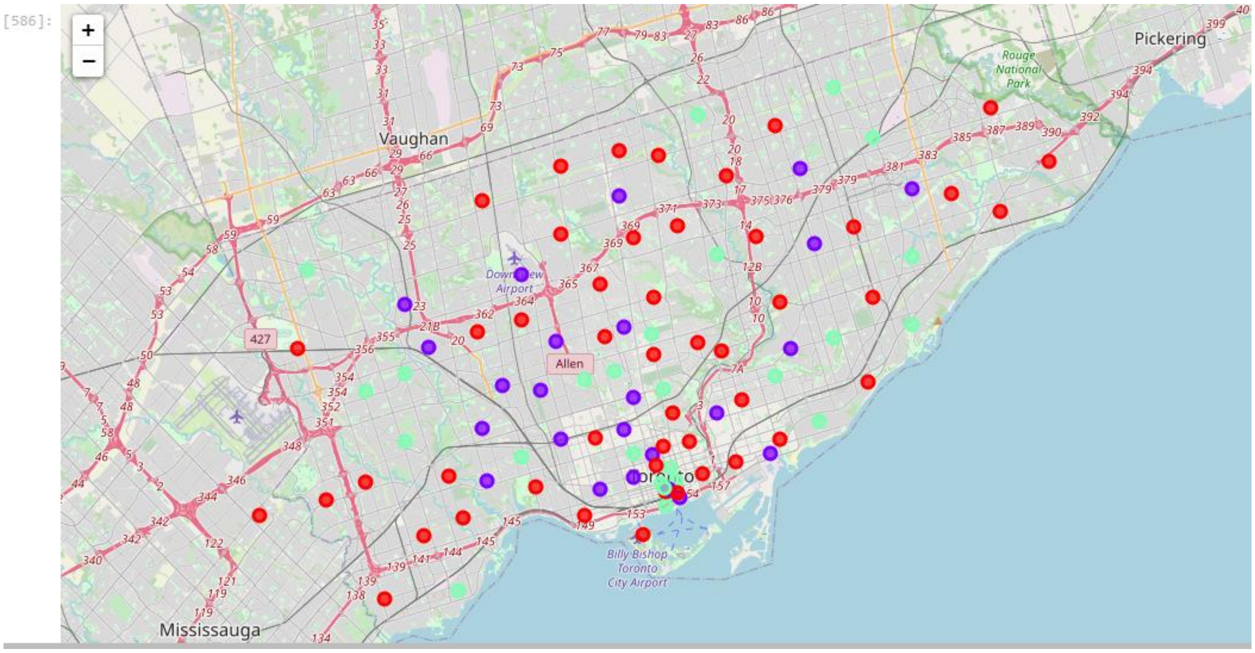


Figure 12 Image showing the three clusters based on the nature of their Venues

From the Diagram above,

The Red circles represent cluster 0

The Green circles represent cluster 1

The Purple circles represent cluster 2

Conclusion

In this project I was able to help a hypothetical investor decide which neighborhood in Toronto to launch a coffee shop based on the viability, environment and population density. Also, I did some light analysis on the structure of the neighborhoods in Toronto.

This Data science capstone project has helped me explore various aspects of data science by covering topics like web scraping, data cleaning, statistics and machine learning. Through this project I have a better understanding of data science and its applications in day to day life.