Neighborhood Clustering and Segmentation: Toronto, Canada (City of Neighborhoods):

1 Introduction

1.1 Business Problem:

The battle of Neighborhoods is a data science project, aimed at grouping similar neighborhoods into the same clusters with the end result being useful information which a user can leverage on to take important business decisions. This project will be targeted on which neighborhoods look promising to set up a new African restaurant.

1.2 Background Description:

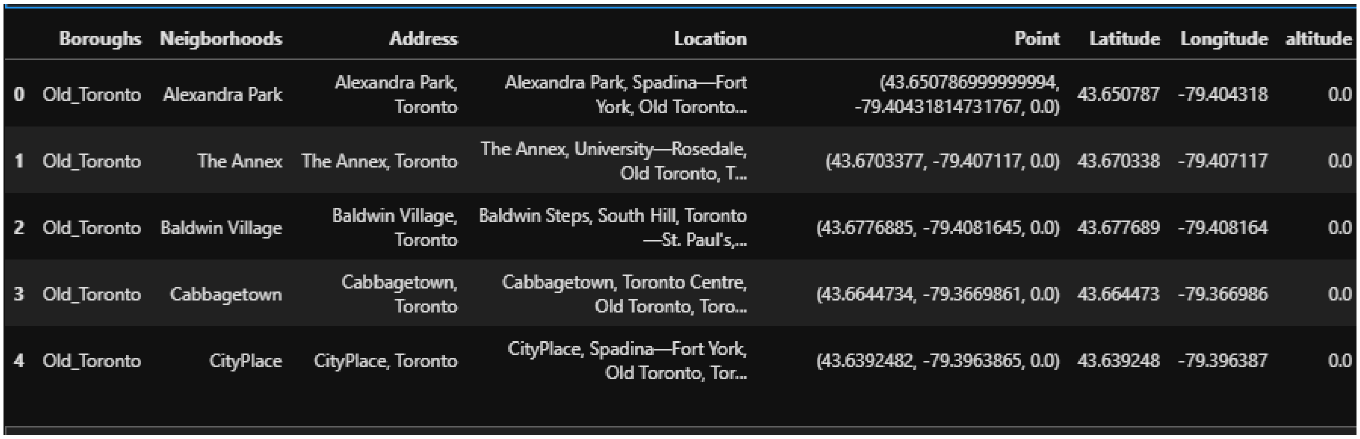
This projects focuses on Neighborhoods in Toronto, Canada. Toronto is the capital city of the Canadian province of Ontario and financial capital of Canada. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America [1]. So it makes sense that the commercial hub will be where most business will be aiming at due to the large purchasing power in the area. The diversity resulting from the movement of a lot of immigrants from several parts of the world for work and settlement is an added advantage because a variety of businesses can thrive there and will meet the needs of different people.

The city of Toronto is made up of 6 Boroughs. There are over 140 neighborhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighborhoods within city limits [2]. Some of the boroughs are further subdivided into sub-boroughs such as the borough called "old" Toronto which is further sub dived into four boroughs. The other boroughs are East York, Etobicoke, North York, Scaborough and York.

2 Data Description

The data used for this project comprised of the Neighborhoods name, position data (Latitude and Longitude) and the most popular venues in each neighborhood. The dataset was not readily available and so I had to:

- Use the **Beautiful Soup** [3] library to scrap the names of the Neighborhood and their respective boroughs from this Wikipedia page [2]. While the corresponding latitude and longitude data for each neighborhood was obtained using the **Nominatim** module from the **geopy** library. A head slice of the data is shown below.



	Boroughs	Neigborhoods	Address	Location	Point	Latitude	Longitude	altitude
0	Old_Toronto	Alexandra Park	Alexandra Park, Toronto	Alexandra Park, Spadina—Fort York, Old Toronto...	(43.650786999999994, -79.40431814731767, 0.0)	43.650787	-79.404318	0.0
1	Old_Toronto	The Annex	The Annex, Toronto	The Annex, University—Rosedale, Old Toronto, T...	(43.6703377, -79.407117, 0.0)	43.670338	-79.407117	0.0
2	Old_Toronto	Baldwin Village	Baldwin Village, Toronto	Baldwin Steps, South Hill, Toronto —St. Paul's,...	(43.6776885, -79.4081645, 0.0)	43.677689	-79.408164	0.0
3	Old_Toronto	Cabbagetown	Cabbagetown, Toronto	Cabbagetown, Toronto Centre, Old Toronto, Toro...	(43.6644734, -79.3669861, 0.0)	43.664473	-79.366986	0.0
4	Old_Toronto	CityPlace	CityPlace, Toronto	CityPlace, Spadina—Fort York, Old Toronto, Tor...	(43.6392482, -79.3963865, 0.0)	43.639248	-79.396387	0.0

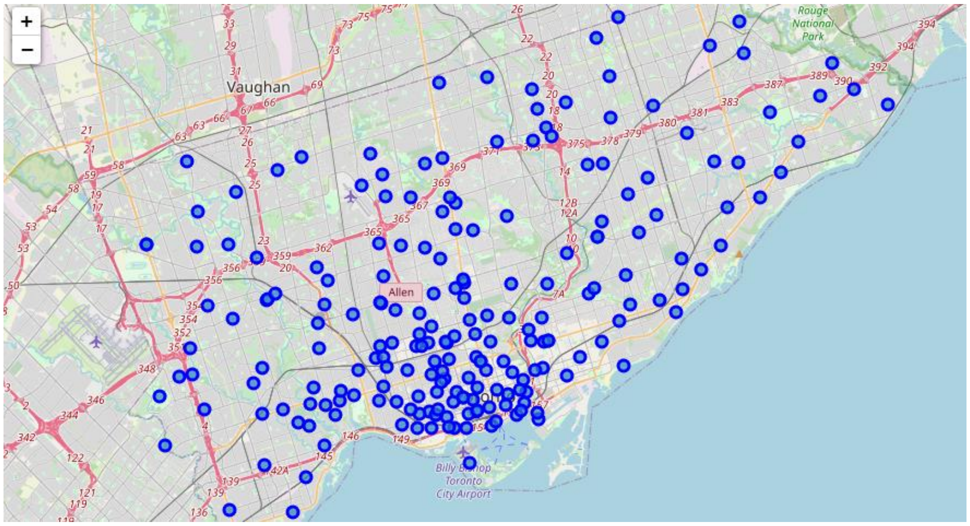
- The most popular venues for each neighborhood was obtained using the **Foursquare** [4] API. The datasets were then eventually merged.

The data was grouped based on the basis of Neighborhood and the mean of the frequency of occurrence of certain venues were taken, this data was then fed into the K-Means algorithm which then clustered similar Neighborhoods together.

3 Methodology

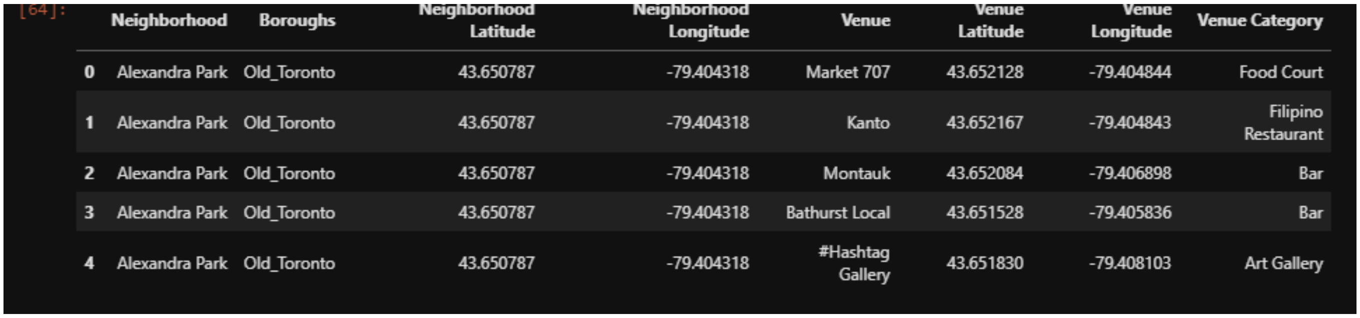
The base data that formed the foundation of this project as stated earlier is a result of the combination of results obtained from scraping with BeautifulSoup and the Nominatim module for location data in the goopy library. The final result returned 213 Neighborhoods of which some positional data were not obtained from the Nominatim Library and had to be manually imputed using a customized function. Also the neighborhood **Briar Hill-Belgravia** [5] which is now included with the neighborhood **Fairbank** had to be dropped since they have the same latitude and longitude value.

Making use of Python’s Folium library, I created a map of Toronto on which I superimposed the neighborhoods points using the Longitude and Latitude data as can be seen in the image below.



Map of Toronto with Neighborhoods location.

In line with the goal of this project which is to ascertain the most viable neighborhood or neighborhoods in which a certain new business can be established, the Foursquare API was used to get the most popular venues in each neighborhood setting the limit and radius of search around each neighborhood at which the API call should operate to 100 and 500 meters respectively. A total of five thousand six hundred and twenty three (5623) venues was returned from the API call, this result returned by the API is not always constant and changes based on the time which the call was made as a list of popular venues in a neighborhood can either increase or decrease due to different factors. Also data was not returned for two neighborhoods Port Union (Centennial Scarborough) and West Deane Park and thus they were dropped from the neighborhood. A head slice of the dataset is shown below:

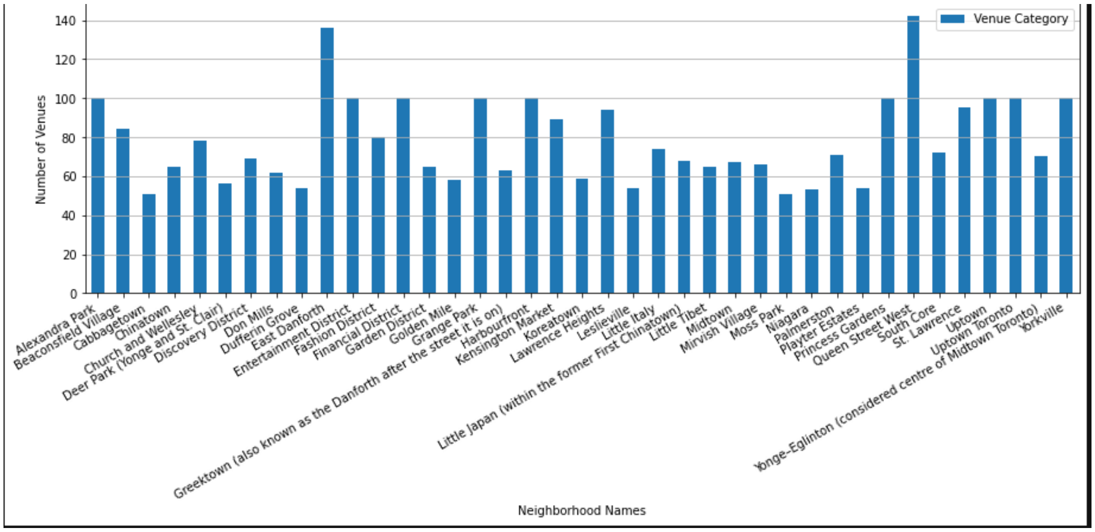


[64]:

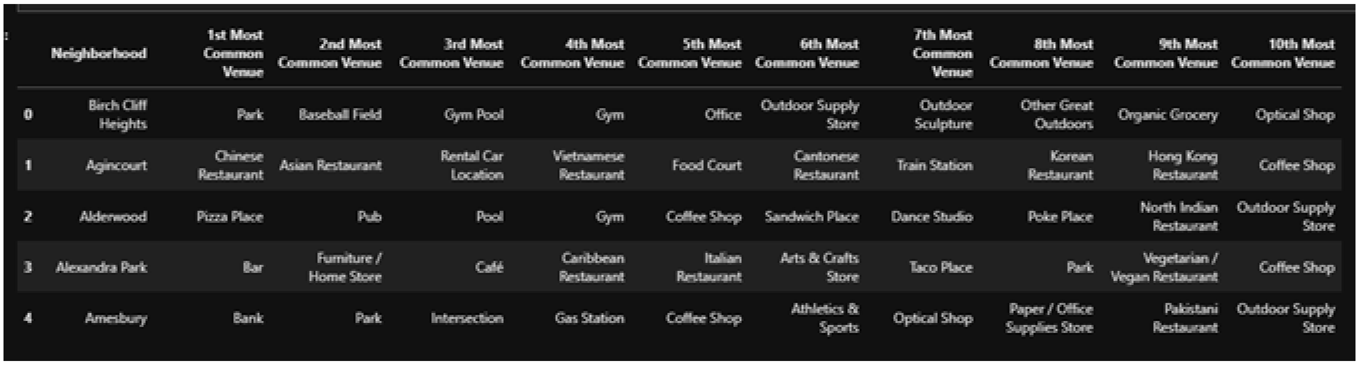
	Neighborhood	Boroughs	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alexandra Park	Old_Toronto	43.650787	-79.404318	Market 707	43.652128	-79.404844	Food Court
1	Alexandra Park	Old_Toronto	43.650787	-79.404318	Kanto	43.652167	-79.404843	Filipino Restaurant
2	Alexandra Park	Old_Toronto	43.650787	-79.404318	Montauk	43.652084	-79.406898	Bar
3	Alexandra Park	Old_Toronto	43.650787	-79.404318	Bathurst Local	43.651528	-79.405836	Bar
4	Alexandra Park	Old_Toronto	43.650787	-79.404318	#Hashtag Gallery	43.651830	-79.408103	Art Gallery

Grouping by the neighborhoods name to get the number of venues returned for each neighborhood resulted in Queen Street West having the highest amount of venues returned numbering up to one hundred and forty two (142) and the neighborhoods Bayview Woods-Steels, Bedford Park, The Bridle Path, West Rouge all having the lowest number of venue returned with values of one (1) each.

The graph below shows neighborhoods with more than 50 venues returned from the API call.



The dataset containing the neighborhoods and most popular venues in each contains a total of three hundred and fifty four (354) unique venue categories. Using one-hot encoding [6] followed by calculating the mean of the frequency of occurrence of a particular venue category in a neighborhood the top ten (10) most popular venues for each neighborhood was generated.

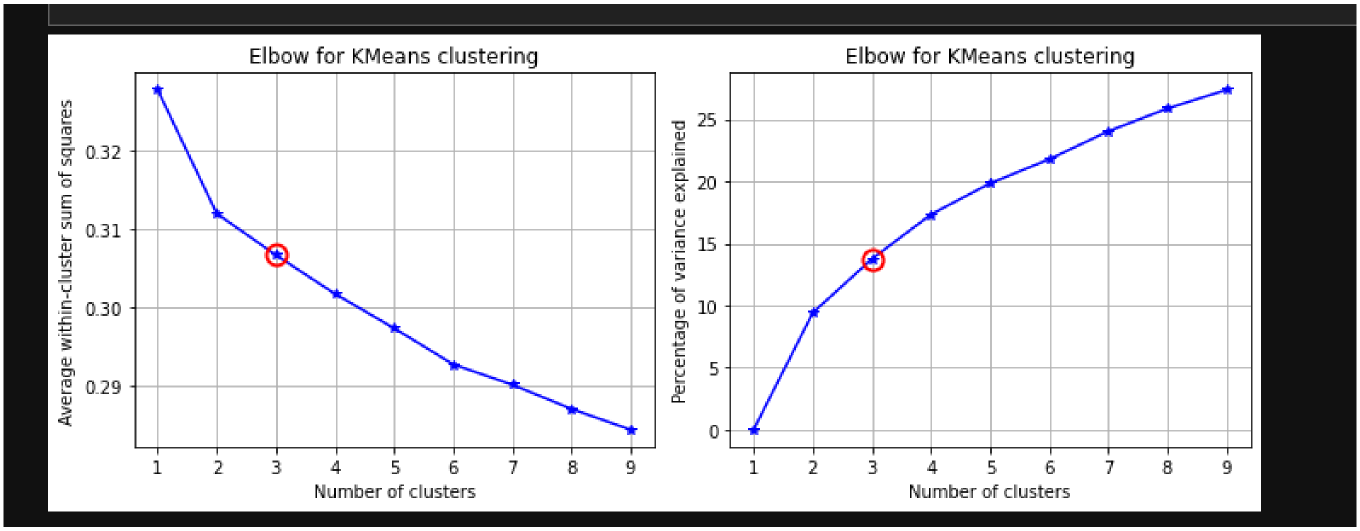


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Birch Cliff Heights	Park	Baseball Field	Gym Pool	Gym	Office	Outdoor Supply Store	Outdoor Sculpture	Other Great Outdoors	Organic Grocery	Optical Shop
1	Agincourt	Chinese Restaurant	Asian Restaurant	Rental Car Location	Vietnamese Restaurant	Food Court	Cantonese Restaurant	Train Station	Korean Restaurant	Hong Kong Restaurant	Coffee Shop
2	Alderwood	Pizza Place	Pub	Pool	Gym	Coffee Shop	Sandwich Place	Dance Studio	Poke Place	North Indian Restaurant	Outdoor Supply Store
3	Alexandra Park	Bar	Furniture / Home Store	Café	Caribbean Restaurant	Italian Restaurant	Arts & Crafts Store	Taco Place	Park	Vegetarian / Vegan Restaurant	Coffee Shop
4	Amesbury	Bank	Park	Intersection	Gas Station	Coffee Shop	Athletics & Sports	Optical Shop	Paper / Office Supplies Store	Pakistani Restaurant	Outdoor Supply Store

A total of 5623 venues returned with only 354 unique categories across all neighborhoods shows that same venue categories occur in more than one neighborhood and leveraging on this the K-Means algorithm which is a common unsupervised machine learning algorithm is used to group into clusters neighborhoods that are very similar based on the type of venue categories that occur frequently.

The elbow method was used to determine the optimal number of clusters to use which will have an impact on the Sum of Squared Error. The Sum of Squared Error (SSE) gives us information on how well a certain dataset is clustered, with a lower SSE value pointing to a high intra-cluster similarity and a higher SSE value pointing to low intra-cluster similarity. With an increase in clusters, the SSE tends to decrease toward 0. The SSE is zero if it is equal to the total number of data points in the dataset, as at this stage each data point becomes its own cluster, and no error exists between the cluster and its center. So the goal with the elbow method is to choose a small value of k that has a low SSE.

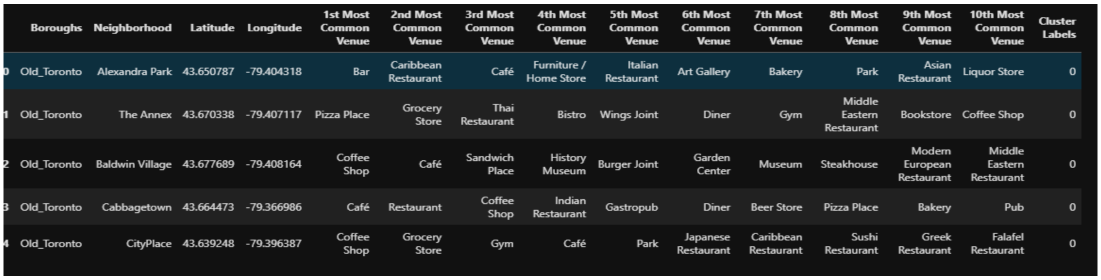
From the result as can be seen in the graph generated the optimal number of cluster to be used is three (3).



4 Results

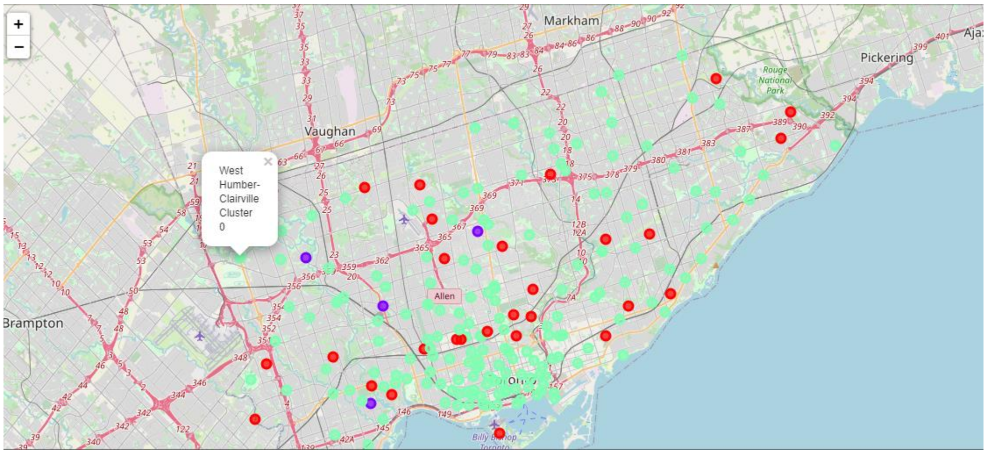
Applying the result of the K-Means algorithm on the dataset and merging the generated cluster labels to the dataset. Here is the merged table with cluster labels for each borough

Making use of the Folium library a clustered Neighborhood map of Toronto was generate



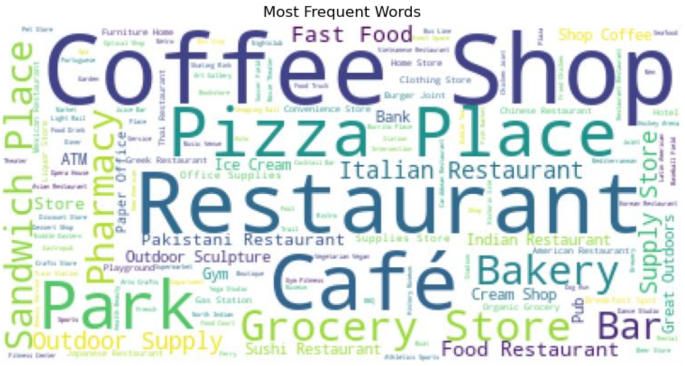
	Boroughs	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Labels
0	Old_Toronto	Alexandra Park	43.650787	-79.404318	Bar	Caribbean Restaurant	Café	Furniture / Home Store	Italian Restaurant	Art Gallery	Bakery	Park	Asian Restaurant	Liquor Store	0
1	Old_Toronto	The Annex	43.670338	-79.407117	Pizza Place	Grocery Store	Thai Restaurant	Bistro	Wings Joint	Diner	Gym	Middle Eastern Restaurant	Bookstore	Coffee Shop	0
2	Old_Toronto	Baldwin Village	43.677689	-79.408164	Coffee Shop	Café	Sandwich Place	History Museum	Burger Joint	Garden Center	Museum	Steakhouse	Modern European Restaurant	Middle Eastern Restaurant	0
3	Old_Toronto	Cabbagetown	43.664473	-79.366986	Café	Restaurant	Coffee Shop	Indian Restaurant	Gastropub	Diner	Beer Store	Pizza Place	Bakery	Pub	0
4	Old_Toronto	CityPlace	43.639248	-79.396387	Coffee Shop	Grocery Store	Gym	Café	Park	Japanese Restaurant	Caribbean Restaurant	Sushi Restaurant	Greek Restaurant	Falafel Restaurant	0

Making use of the Folium library a clustered Neighborhood map of Toronto was generate

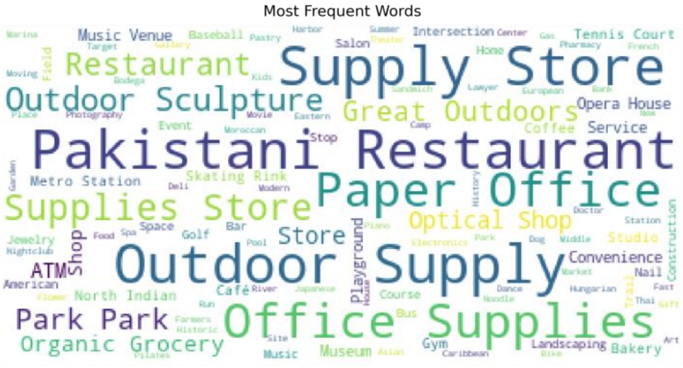


Renaming the clusters to the name of the most popular venues across all ten most common venues was done using the results obtained from creating a **Word Cloud.**

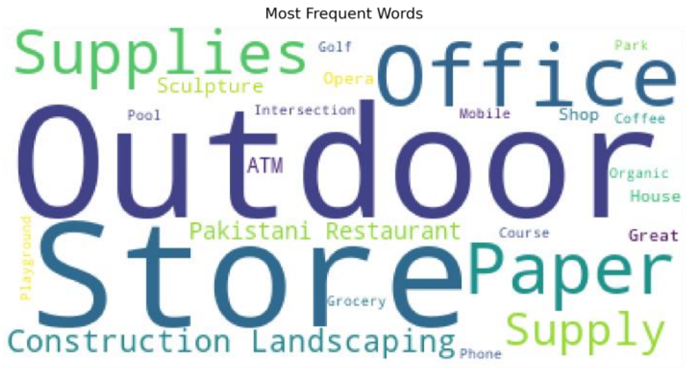
For Cluster 0: Coffee Shop



For cluster 1: Pakistani Restaurant



For cluster 2: Outdoor Supply Store and Outdoor Sculpture



5 Discussion and Conclusion

Based on the results from above and paying particular attention to the clusters generated. Cluster 0 also called the **Coffee Shop** cluster contains the highest amount of neighborhoods with similarities and this can be attributed to the fact that majority of the neighborhoods in this cluster are from the 'old' Toronto borough that has the highest amount of neighborhoods hence the heavy similarity. Taking into consideration the business objective of this project which is to find a neighborhood that will increase the feasibility of establishing and growing a new African restaurant, Taking a looking at cluster 0 giving the high population and number of restaurants that are already well established a high purchasing power is expected in such area which is needed for optimal business growth ,however taking into consideration the heavy competition that will most likely exist amongst similar businesses might provide a bit of challenge for a newly established business plus the fact that some customers might already have favorite venue which they might not be willing to change. Taking into consideration cluster 1 also called the **Pakistani Restaurant** as derived from the word cloud generated from this cluster and exploring the common type of restaurants present in this cluster suggests that the populace here might be Asian and might not be a good market base for African dishes.

Considering cluster 2 this cluster might not provide the necessary factors needed for a new African restaurant to make optimal business profit based on two reasons derived from the results above. Firstly the most common venues have little to do with restaurants except for the occurrence of Pakistani restaurants in some neighborhoods which leads to the second reason that thus there will be little purchasing power or consumer interest for an African restaurant in any of the neighborhoods found in this cluster.

So in conclusion neighborhoods in cluster 0 relative to other clusters provide the necessary factors such as large population and high purchasing power which can be leveraged on to produce optimum profit for the new African restaurant.

REFERENCES

1 https://en.wikipedia.org/wiki/Toronto

2 https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto

3 https://www.crummy.com/software/BeautifulSoup/bs4/doc/

4 https://developer.foursquare.com/

5 https://en.wikipedia.org/wiki/Fairbank,_Toronto/

6 https://www.educative.io/edpresso/one-hot-encoding-in-python/