Ranking Neighborhoods based on Pizza Restaurants

IBM Data Science Capstone Project

Written by Muhammad Saleh Anwar

Introduction:

The goal of this project is to cluster neighborhoods in Toronto based on the Pizza restaurants located in the neighborhood. If you are a pizza lover in Toronto or any city in the world, you would like to know which neighborhoods in your city have the best options available for Pizza restaurants.

This project will answer the question: Can we rank the neighborhoods in Toronto based on the Pizza restaurant options available? In order to answer this question, two main factors are going to be analyzed. One is the number of Pizza restaurant options available in a neighborhood. The more options are available in a neighborhood, the more choices its residents have. Therefore, the number of options available should affect the ranking of the neighborhood. The other factor is the popularity of the Pizza restaurant options in the neighborhood. There are different metrics available that can be utilized for gauging popularity of any restaurant. Most people rely on ratings provided by platforms such as Foursquare, Yelp and Google to decide on a particular restaurant. Customers can log on to these platforms and submit reviews, ratings or likes for a particular restaurant. For this project, the number of likes associated with each Pizza restaurant is extracted using the Foursquare API. These two factors are used to cluster the neighborhoods in the Toronto based on Pizza restaurants.

Data:

In order to cluster the neighborhoods in Toronto, the list of all neighborhoods in Toronto is required. This data is extracted from the Wikipedia page titled 'List of Postal Codes of Canada' (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=942655364). This page only contains the postal codes starting with 'M' which is the convention used for the neighborhoods in Toronto. Along with the postal codes, it also contains the name of Borough and the neighborhood name.
In addition to the list of all neighborhoods in Toronto, the geospatial data (latitude and longitude) for the neighborhoods in Toronto is also required. This is needed to perform

Foursquare API queries as well as to plot the neighborhood on a map using Folium for visualization purposes. This dataset was provided during this course and it can be downloaded from http://cocl.us/Geospatial_data. This dataset lists all the postal codes in Toronto with their latitude and longitude.

List of all restaurants in a particular neighborhood was required for this project. This was extracted through a Foursquare API call. This API call provided a list of all venues for the neighborhoods in Toronto. Using pattern matching, the list of restaurants and pizza restaurants were extracted from the venue's dataset. The number of likes submitted for each pizza restaurant was also obtained through a Foursquare API call.

In summary, the following data was collected and used for the project:

- Postal Codes in Toronto (Wikipedia)
- Geospatial data for Toronto neighborhoods (provided in previous assignment)
- List of all venues for the neighborhoods in Toronto (Foursquare API call)
- Number of likes submitted for each Pizza restaurant in Toronto (Foursquare API call)

Methodology:

1) Data Extraction & Cleaning:

The first step is to clean the data. The list of Postal Codes in Toronto is extracted from Wikipedia. The extracted dataset contains one row entry for every Borough and Neighborhood. Below are the first 5 rows in the dataset:

	Postcode	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	Harbourfront

From the rows displayed above, it can be seen that some postal codes have not been assigned to Boroughs or Neighborhoods. Those postal codes are going to be excluded from the dataset. Additionally, any postal codes with multiple neighborhoods will be combined together in one

row will all the neighborhoods listed (separated by commas). The resulting dataset is shown below:

	Postcode	Borough	Neighbourhood
1	M1B	Scarborough	Rouge, Malvern
2	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
3	M1E	Scarborough	Guildwood, Morningside, West Hill
4	M1G	Scarborough	Woburn
5	M1H	Scarborough	Cedarbrae
6	M1J	Scarborough	Scarborough Village
7	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
8	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
9	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
10	M1N	Scarborough	Birch Cliff, Cliffside West
11	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...
12	M1R	Scarborough	Maryvale, Wexford
13	M1S	Scarborough	Agincourt
14	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter
15	M1V	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...
16	M1W	Scarborough	L'Amoreaux West
17	M1X	Scarborough	Upper Rouge
25	M2H	North York	Hillcrest Village
26	M2J	North York	Fairview, Henry Farm, Oriole

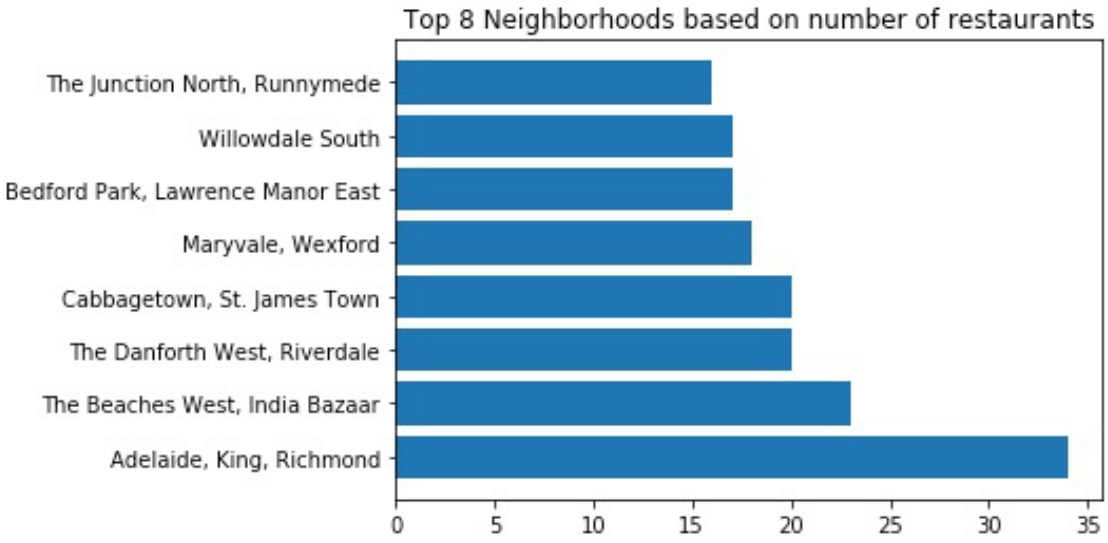
The geospatial data contains one row entry for each postal code and contains the longitude and latitude information. This data set is merged with the neighborhood data and the resulting dataset is show below:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

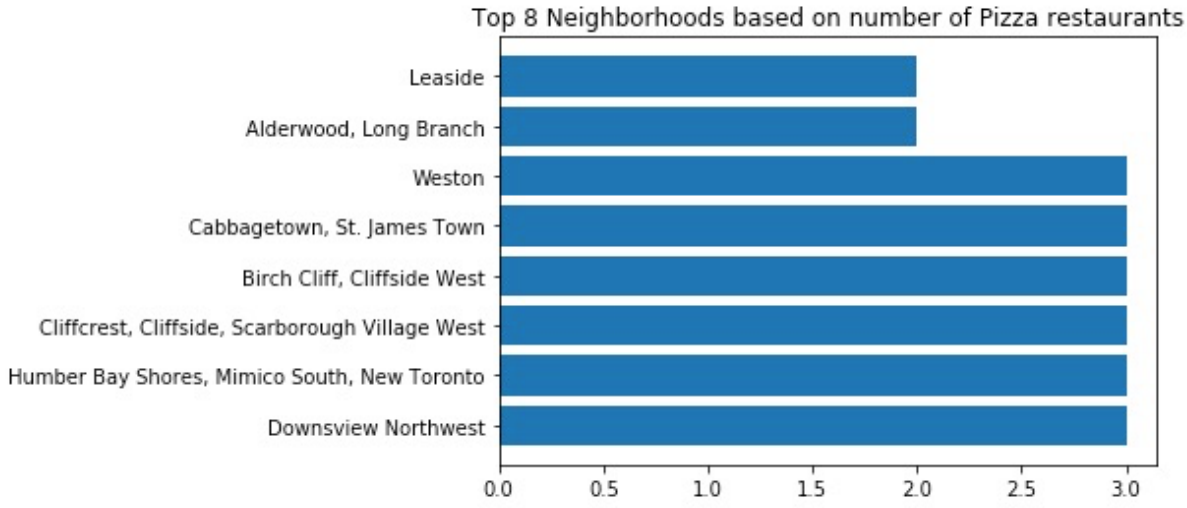
List of all the venues located in a particular neighborhood is extracted by performing an API call through Foursquare API. This is done for each Postal Code using the latitude and longitude information. The postal code of the venue is also extracted from the Foursquare output which is later used to assign the venues to each postal code in our Postal Code dataset.

The output from Foursquare API call contains a list of all the venues in the neighborhoods of Toronto. For the purposes of our project, the data related to restaurants and more specifically Pizza restaurant is of interest. A text search criterion is created based on the Venue Categories outputted by the API call. This criterion is applied to extract a list of just restaurants from the overall venues list.

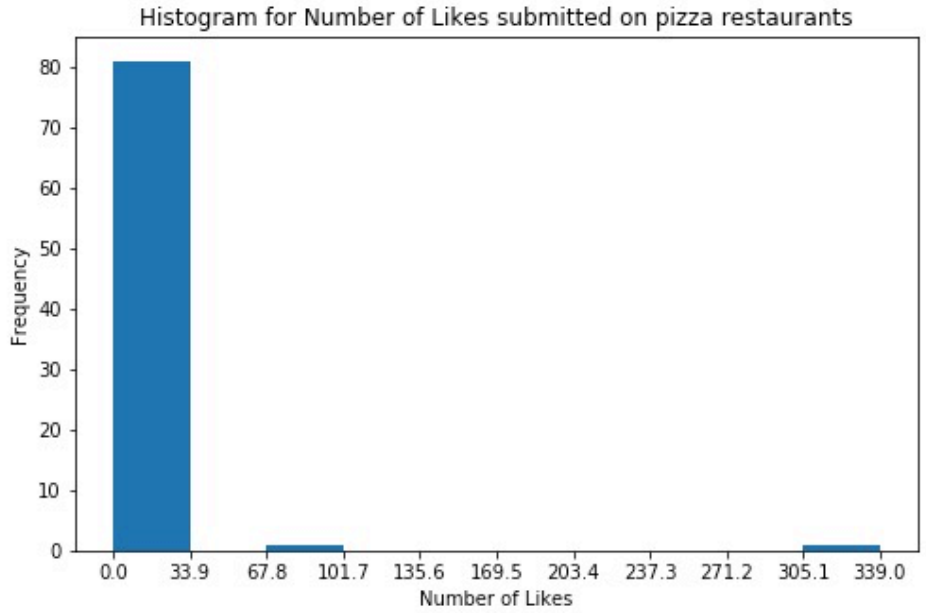
The top 8 neighborhoods based on the number of restaurants is shown in the bar chart below:



Another text search criterion is applied to extract a list of just Pizza restaurants from the venues dataset. The top 8 neighborhoods based on the number of Pizza restaurant is shown in the bar chart below:

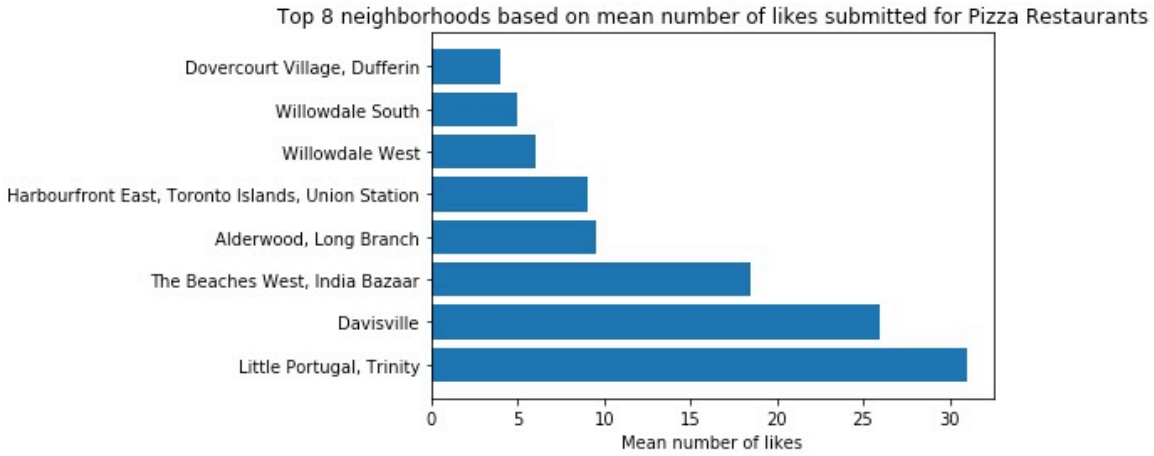


The number of likes submitted for the Pizza restaurants is extracted by making a Foursquare API call. The number of likes for each restaurant are then analyzed through a histogram as seen below:



From the histogram, it can be seen that any values over 33 can be considered as outliers. If those outliers are included in the dataset, the results will significantly skewed in favor of the neighborhoods with those high number of likes. In order to present a more balanced clustering of neighborhoods, the number of likes for reach restaurant is capped at 33.

The number of likes is then grouped by neighborhood to calculate the mean number of likes for each neighborhood. The top 8 neighborhoods based on the mean number of likes are displayed in a bar chart below:



2) Data Clustering:

The feature set for Clustering is obtained by merging the following three datasets:

- Postal code dataset (obtained from Wikipedia)
- Number of Restaurants and Pizza Restaurants for each postal code/neighborhood (Foursquare API)
- Mean number of likes for each postal code/neighborhood (Foursquare API)

The first five rows of the feature set is shown below:

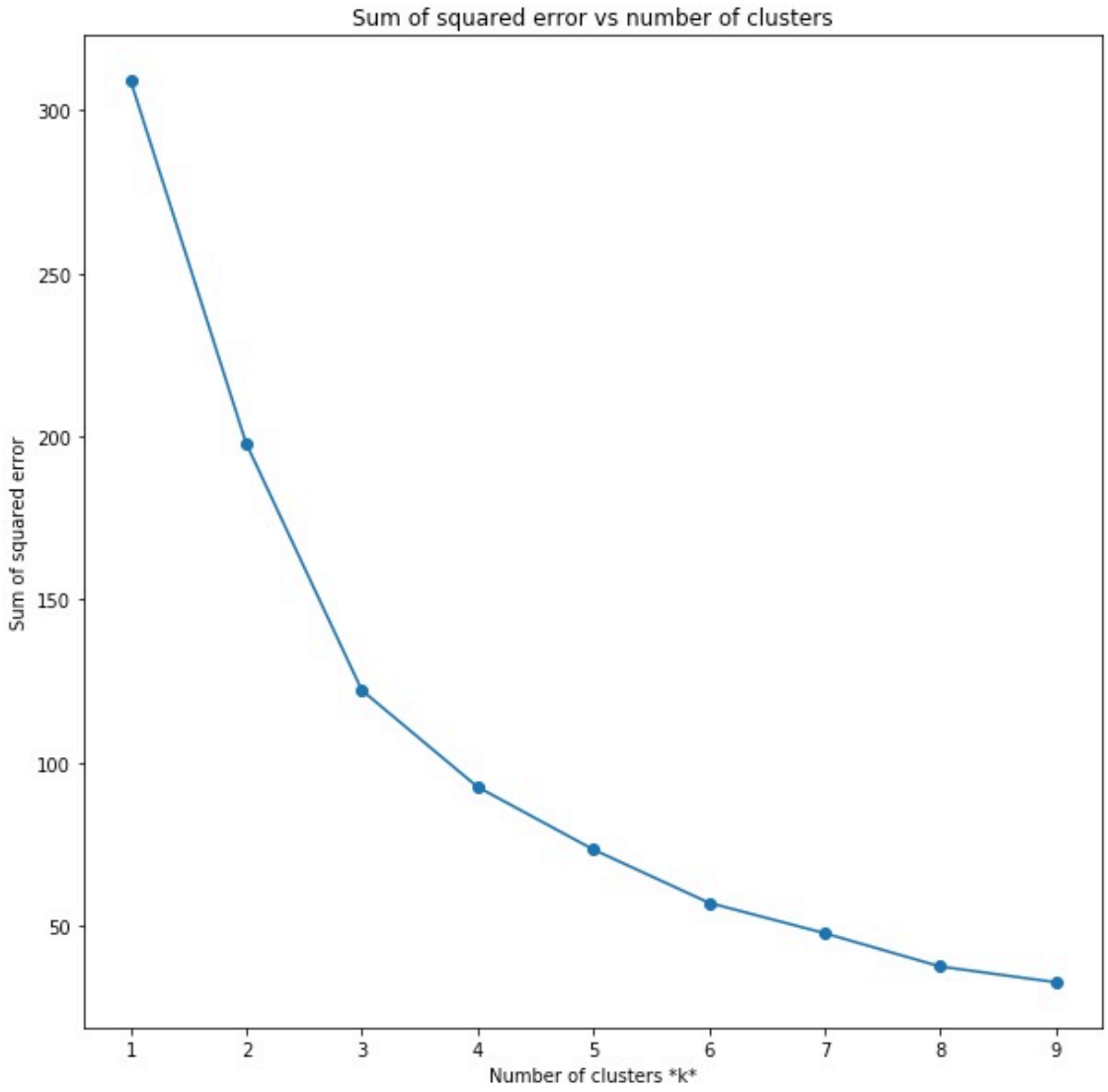
	PostalCode	Borough	Neighbourhood	Restaurant_Count	Pizza_Count	Likes
0	M1B	Scarborough	Rouge, Malvern	12.0	1.0	0.0
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	6.0	1.0	0.0
2	M1E	Scarborough	Guildwood, Morningside, West Hill	9.0	NaN	NaN
3	M1G	Scarborough	Woburn	5.0	2.0	1.0
4	M1H	Scarborough	Cedarbrae	14.0	1.0	1.0

The missing values are resolved by replacing them with 0. The values for Restaurant Count, Pizza Count and Likes are standardized by removing the mean and scaling it to unit variance which are then fed into the clustering algorithm.

K-means algorithm is an iterative algorithm that tries to partition the dataset into K distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (the mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

One of the drawbacks of K-means algorithm is that the number of clusters is pre-defined. In order to find the optimal number of clusters, the elbow method is used. This method gives an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. The number of clusters k is picked at the spot where SSE starts to flatten out and forming an elbow.

Below is the plot of sum of squared error for the different values of k:



It is evident from the plot that the sum of squared error decreases rapidly until k=3. After that the rate of decrease slows down. Hence, the value of k=4 is selected for K-Means algorithm.

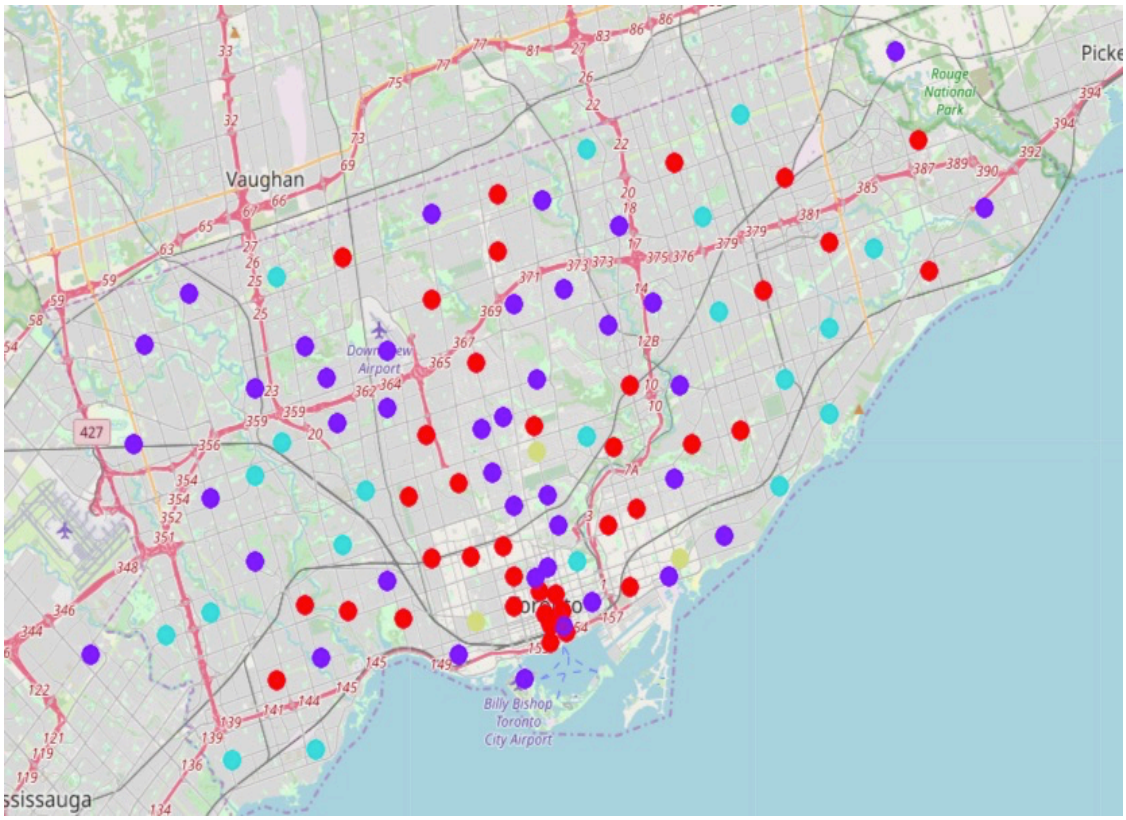
Results:

Clustering was performed on the neighborhood dataset though the use of K-Means algorithm with k=4. The neighborhoods in Toronto were divided into four (4) distinct groups based on the number of restaurants, number of pizza restaurants and the mean number of likes for the Pizza restaurants.

The table below shows the different clusters along with the mean values of the features:

Labels	Restaurant_Count	Pizza_Count	Likes
0	11.000000	0.525000	0.950000
1	2.750000	0.275000	0.375000
2	10.400000	2.300000	1.733333
3	17.333333	1.666667	25.166667

Below visualization shows the different neighborhoods in Toronto. The color of label represents what cluster they belong to.

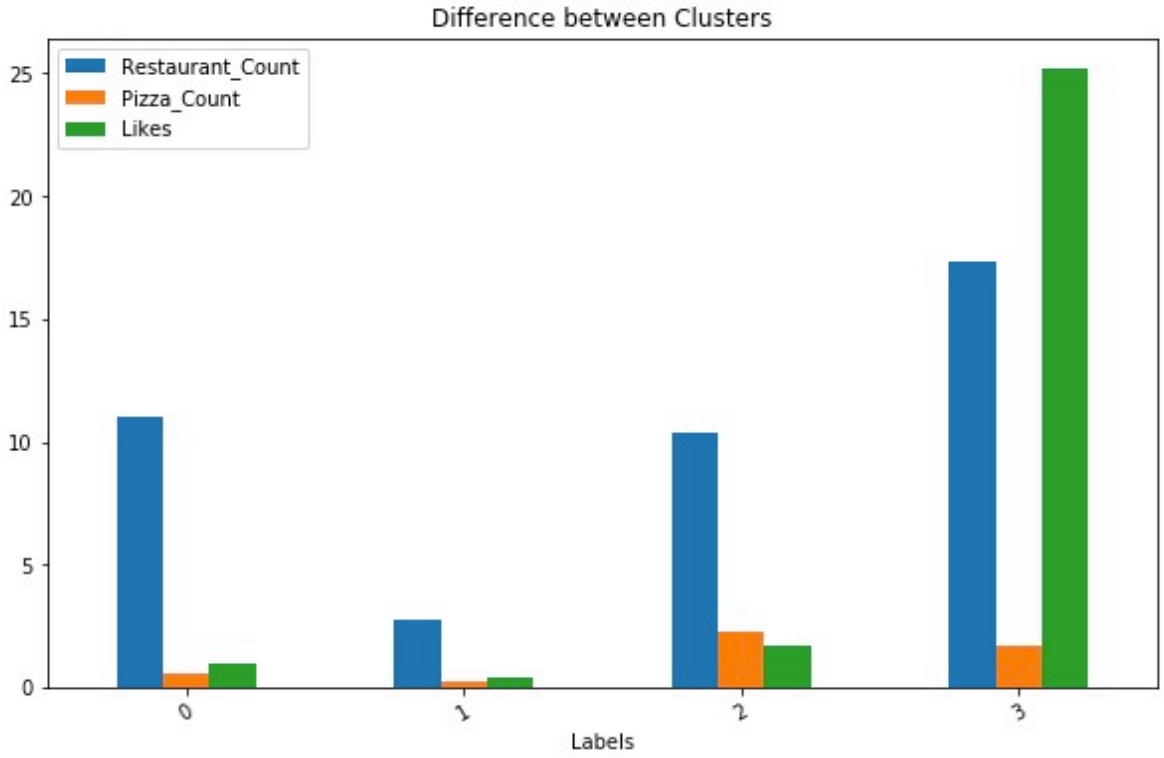


Color	Cluster
	0
	1
	2
	3

Discussion:

Through clustering, the neighborhoods in Toronto were split into four distinct clusters. These four clusters are different from each other based on the number of restaurants, number of Pizza restaurants and mean number of Likes submitted for their Pizza restaurants.

The difference between clusters is shown the bar chart below:



Cluster 0 and Cluster 1 have low mean number of Pizza restaurants and low mean number of Likes. The difference between these two clusters is the mean overall number of restaurants.

Cluster 2 have a high number of Pizza restaurants, but it has a relatively low mean number of Likes compared to Cluster 3.

Cluster 3 has a really high mean number of Likes which shows that it has the most popular Pizza restaurant options out of all the neighborhoods in Toronto.

The neighborhoods included in Cluster 3 are shown below. It can be seen that the neighborhoods in this cluster include Pizza restaurants with very high mean number of Likes.

	PostalCode	Borough	Neighbourhood	Restaurant_Count	Pizza_Count	Likes	Labels	Latitude	Longitude
42	M4L	East Toronto	The Beaches West, India Bazaar	23.0	2.0	18.5	3	43.668999	-79.315572
47	M4S	Central Toronto	Davisville	13.0	1.0	26.0	3	43.704324	-79.388790
77	M6J	West Toronto	Little Portugal, Trinity	16.0	2.0	31.0	3	43.647927	-79.419750

These neighborhoods are visualized on the folium map below:

Conclusion:

Through the application of K-means clustering algorithm, the neighborhoods in Toronto were clustered based on the Pizza restaurants located in those neighborhoods. We were able to rank the neighborhoods based on the Pizza restaurants and we concluded the neighborhoods with best features are Davisville, Little Portugal / Trinity and The Beaches West/India Bazaar.

In this project, the features used were number of Restaurants, number of Pizza restaurants and number of Likes submitted for the Pizza restaurants. All of this data was extracted from Foursquare. This adds bias to our model as only what source was used to collect the data used in our model. It is possible that Foursquare is not the most popular platform for majority of Toronto residents.

In order to improve this project and reduce bias, other platforms such as Yelp and Google can also be used to extract this data. Additionally, other features can be used to perform clustering. These features could include comments from users, ratings etc. This may require use of algorithms such as sentiment analysis, but it could serve as a good indicator for what the customer feels about a restaurant.

Through this project, we applied various data science tools to answer a real-world question. We applied these tools to a real-world dataset. We complete various steps of data extraction, cleaning, modelling, visualization and evaluation. However, data science is an iterative process. Therefore, the findings from this project are to be evaluated and any feedback is to be implemented into the methodology to generate improvements.