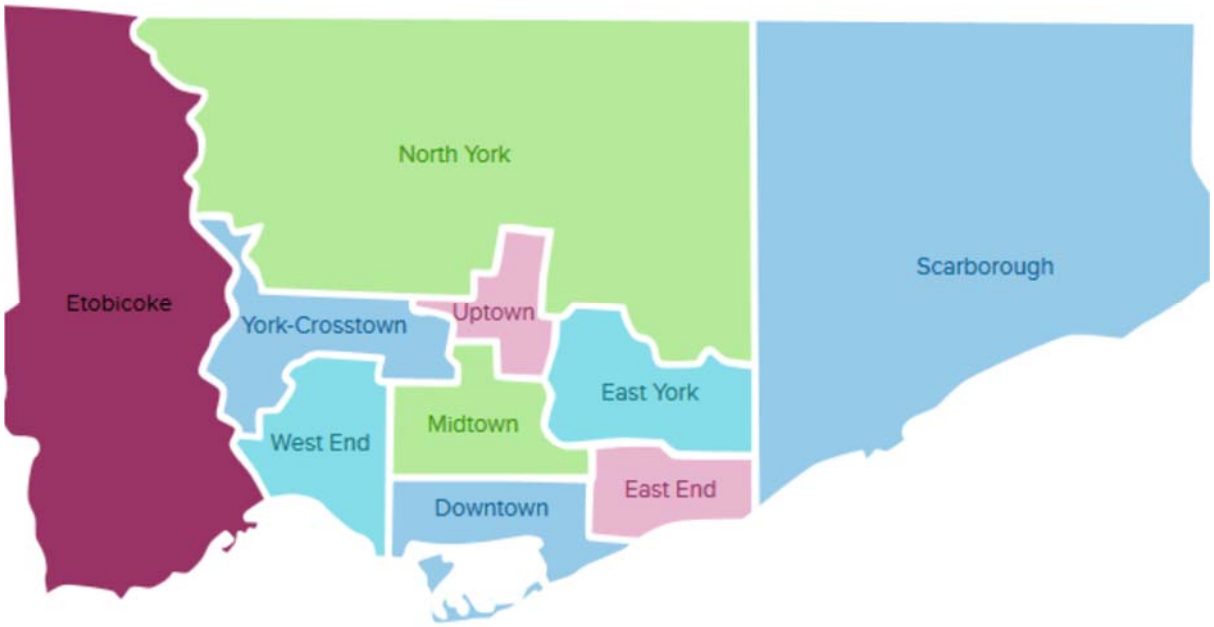
Battle of Neighborhoods

Battle between Toronto neighborhoods



By,

Prashant Shingate, Ph.D.

Introduction

Toronto is the capital city of the province of Ontario. It's a dynamic metropolitan city with vast number of diverse groups. It is one of the densely populated city in Canada with 5.9 million population. Hence, it is a premium location for business.

In Toronto the food industry is one of major industry. The location and type of food channel are key parameters, which strongly affect the business turnover. The population density decreases as one move away from the city center hence the distance of the business venue from city center will affect the number of customers. However, the number of other food channels in other word number of competitors is also high in the proximity of city center.

Besides location, the type of food channel determines cost for the business, profitability, scalability and other challenges. Hence, it is also important to know the type of food channels already present at the location before finalizing business plan.

The aim of my project is to identify the location for a food channel, which is reasonably close to the city center and have less number of competitors. In addition, the type of competitor food channels in the location viz. restaurant, bar, fast-food channel, etc. will definitely help stakeholders to take well-informed decision.

Target Audience:

All entrepreuners willing to enter in a food industry in Toronto area.Specially different food-chain owners or franchise will be best target audience for this study.

Data Sets:

Following datasets and information will be required for this project:

1) For this project, I will be using postal code information on neighborhoods within different in Toronto. This dataset is available at following location:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

   i) Table containing postal code information:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights,Lawrence Manor
5	M7A	Downtown Toronto	Queen's Park

2) The coordinates for each postal code using csv file obtained from following URL:
https://cocl.us/Geospatial_data

   i) Table containing coordinates for postal codes:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

3) Foursquare API to extract information on Venue location and Venue type within Toronto city.
   Table containing 'Venue' information extracted from Foursquare API:

	name	categories	lat	lng
0	Downtown Toronto	Neighborhood	43.653232	-79.385296
1	Japango	Sushi Restaurant	43.655268	-79.385165
2	Rolltation	Japanese Restaurant	43.654918	-79.387424
3	Sansotei Ramen 三草亭	Ramen Restaurant	43.655157	-79.386501
4	Tsujiri	Tea Room	43.655374	-79.385354

Methodology:

Following steps need to be followed for successful execution of this project:

1) Download and preprocess data using different libraries like beatifulsoup4, requests, numpy and pandas.

2) Save data in tabular format as a dataframe. Both the input table looks as follow after processing:

   ii) Table containing postal code information:

   iii) Table containing coordinates for postal codes:

   iv) Table containing 'Venue' information extracted from Foursquare API:

3) Then calculate the distance of each venue from Toronto city center (43.6532, -79.3832) using distance function from 'geopy' python library to calculate distance of each venue from City center.

4) Then create a master table by merging all the data-table. The master table will contain several important features.

For Example:

   i) Borough name

ii) Neighborhood name
iii) Neighborhood coordinates
iv) Venue name
v) Venue Coordinates
vi) Venue Category
vii) Distance from city center, *etc.*

5) Estimation of Pearson's correlation coefficient and regression analysis between the distance from city center and number of venues will be calculated using 'scipy'

6) Then using k-mean clustering method, cluster all the venues based on their location and check for most common type of food channels for each cluster.

7) Visualize all these cluster on the map using Foursquare API and folium library

8) Recommend the stakeholder a location proximal to city center with least number of competition, and warn them about most common type of food channels/competitors in the area.

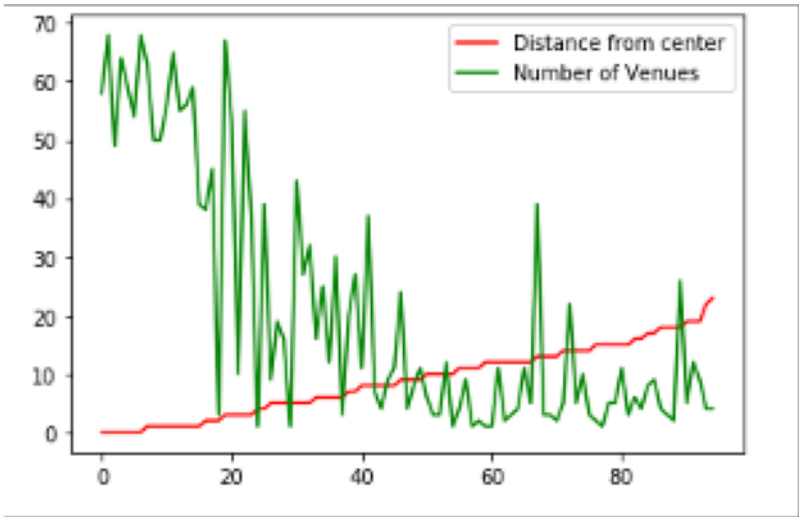
Results

Merged table:

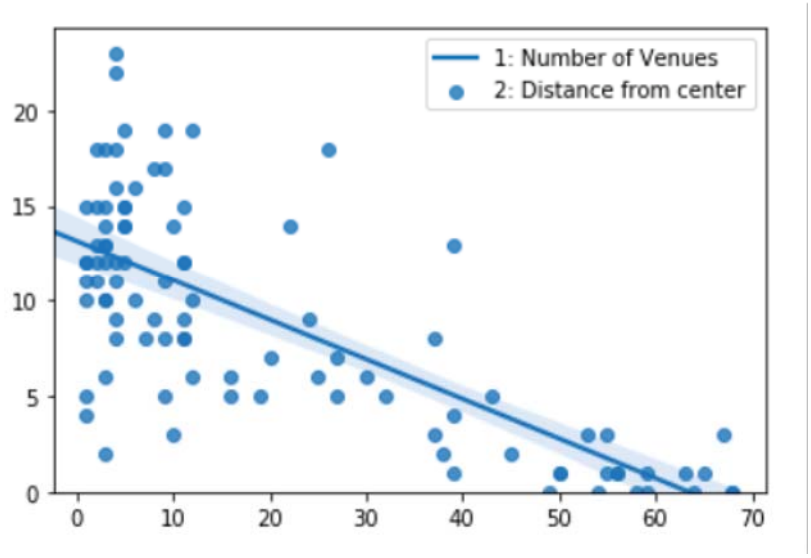
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	MainCategory	Venue Category	Borough	Postal Code	Distance
0	Victoria Village	43.725882	-79.315572	The Frig	43.727051	-79.317418	Restaurant	French Restaurant	North York	M4A	9.7438088475991 km
1	Victoria Village	43.725882	-79.315572	Pizza Nova	43.725824	-79.312860	Other food	Pizza Place	North York	M4A	9.7438088475991 km
2	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Cafe/Bakery	Coffee Shop	North York	M4A	9.7438088475991 km
3	Victoria Village	43.725882	-79.315572	JJ Bean	43.720229	-79.313561	Cafe/Bakery	Café	North York	M4A	9.7438088475991 km
4	Harbourfront	43.654260	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant	Restaurant	Downtown Toronto	M5A	1.8240779832081802 km

Relationship between distance from city center and number of venues:

From following the graph it is evident that as distance from city center increases increases the number of venues decreases. **The Pearson's correlation coefficient between them is -0.74 with 9 × 10E-18.**

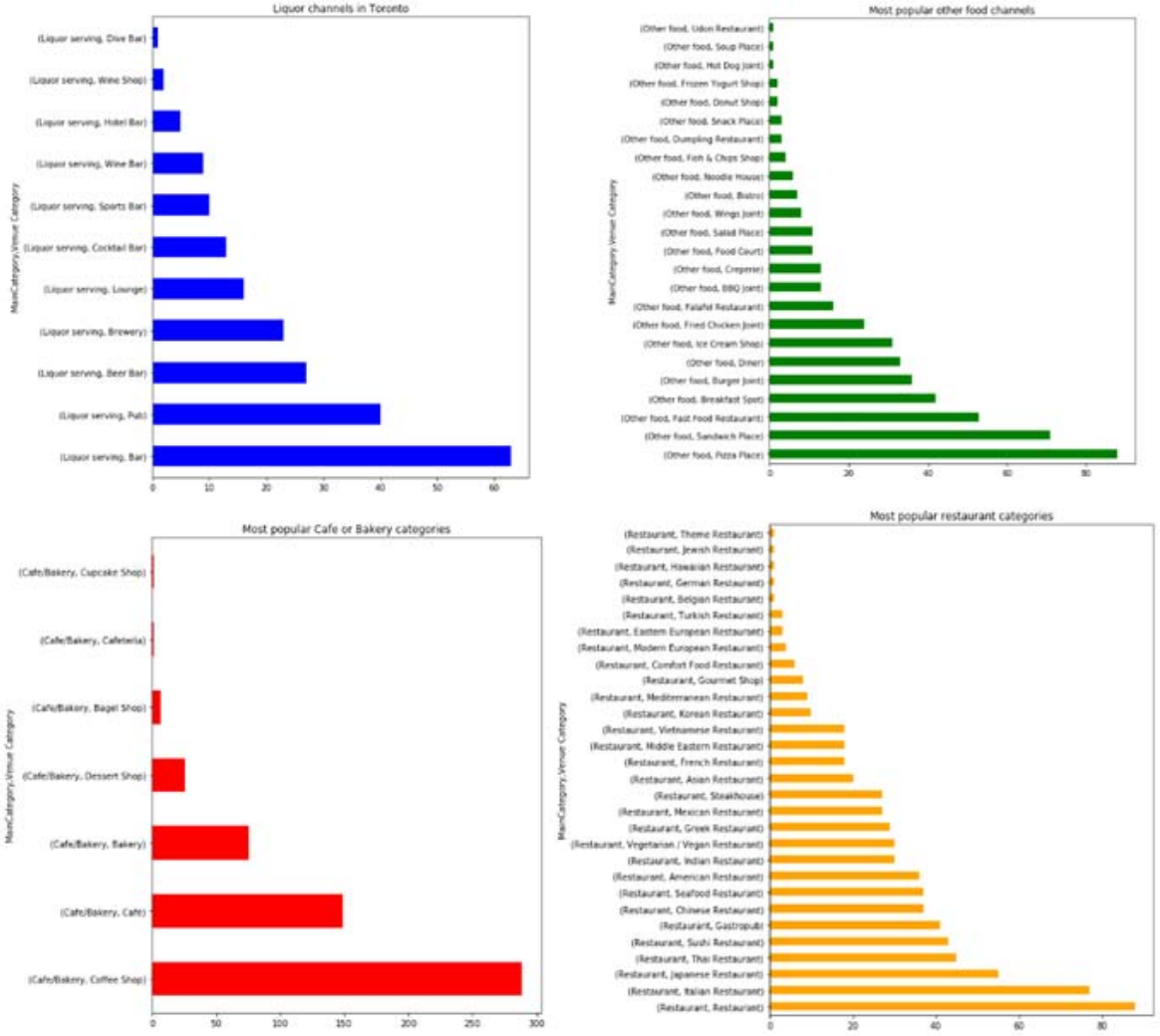


The regression plot between number of venues and the distance from city center is shown in following graph.



Distribution of venues in Toronto:

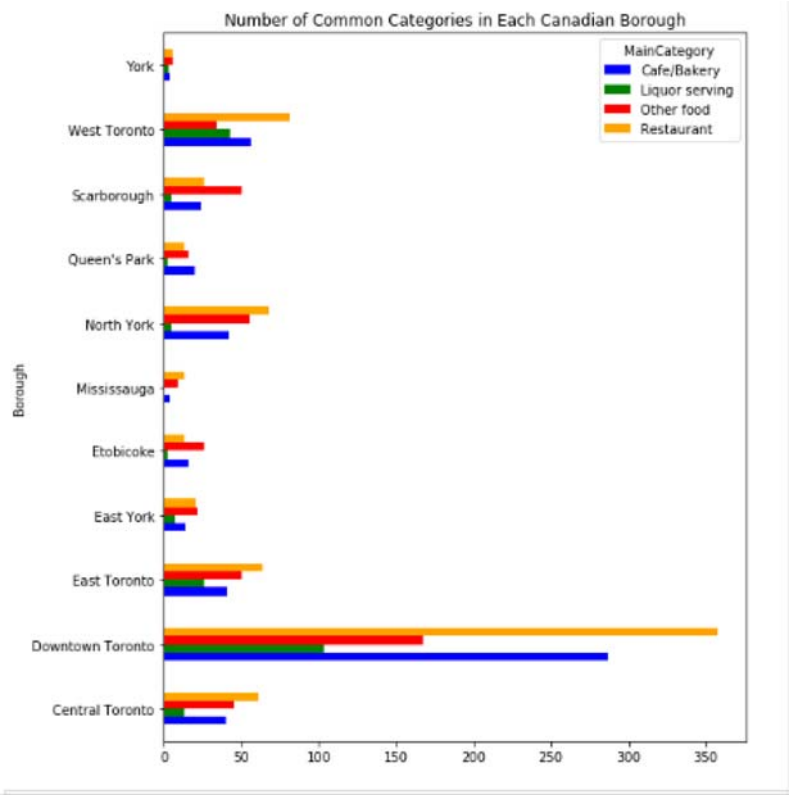
The distribution of liquor serving channels, restaurants, cafes/bakery and other food channels are shown in blue, orange, red and green color for each borough.



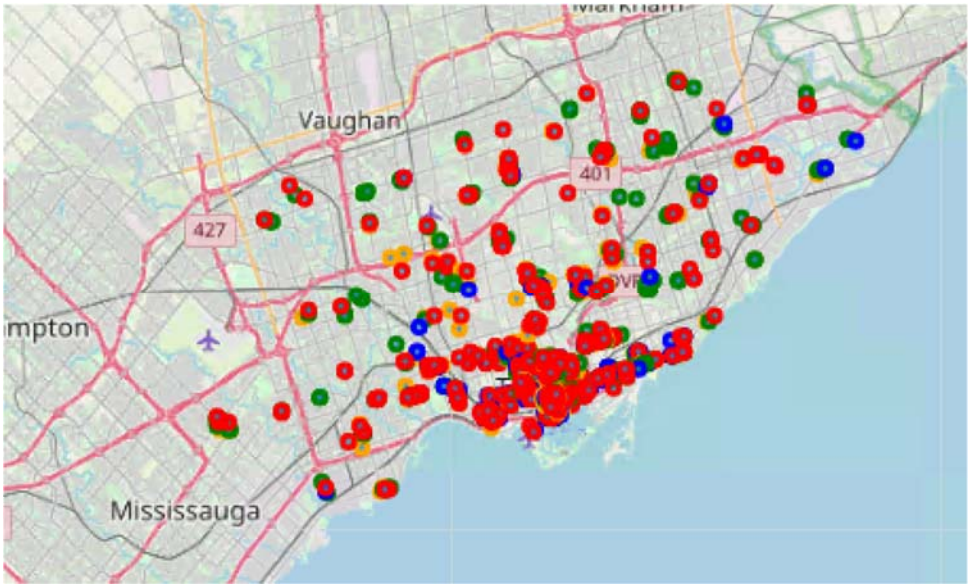
It is evident that few types of venues have larger frequency compared to other.

Number of venues in each borough:

It is evident from following figure that the ‘Downtown Toronto’ has largest number of venues compared to any other borough. Also the highest number of venue is ‘restaurant’ compared to other venues however it is now always true. Hence we need to perform a k-mer clustering to understand the different types of venues based on types of venues present within them.



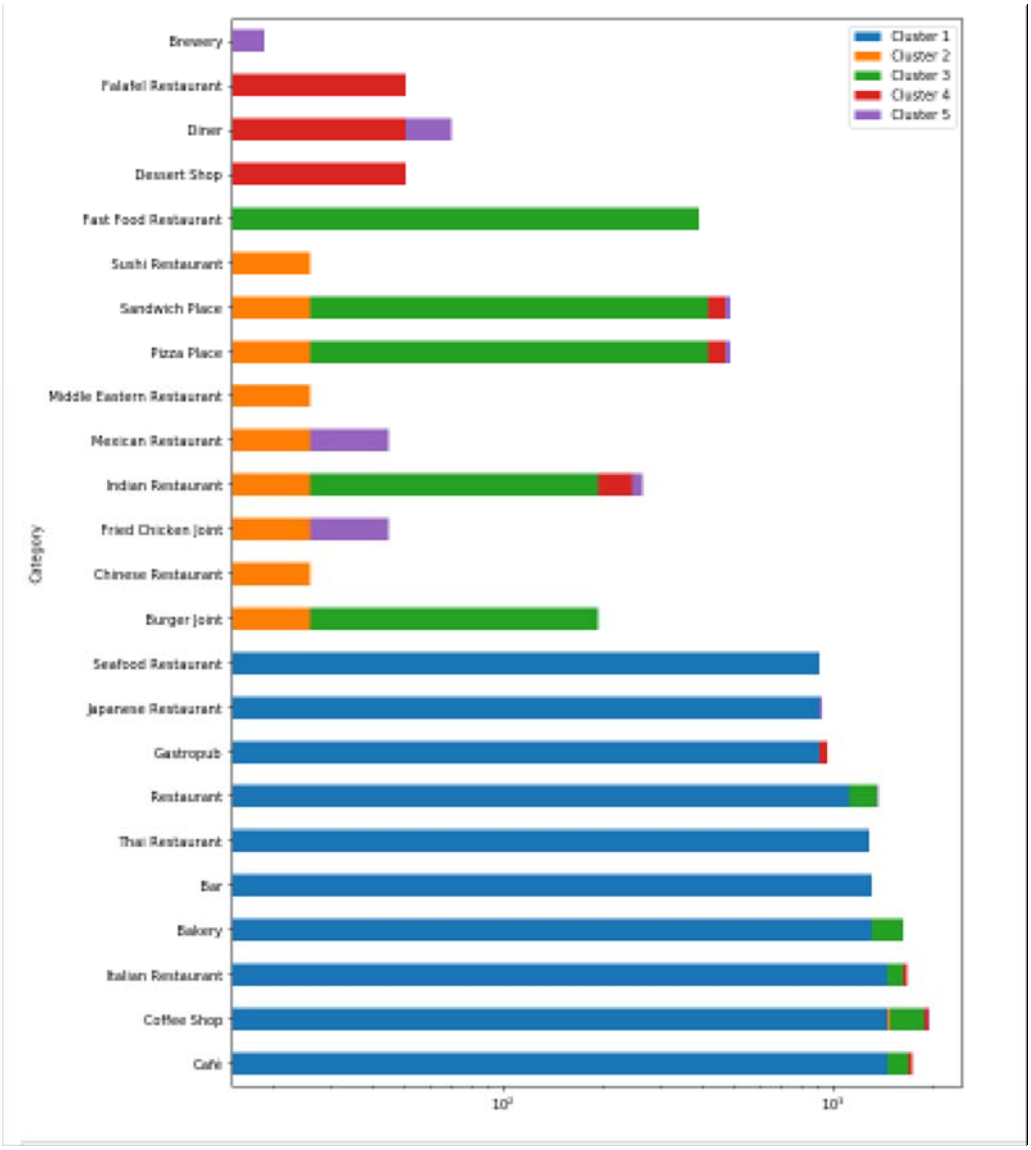
The distribution of different venues on Toronto map



K-mer clustering

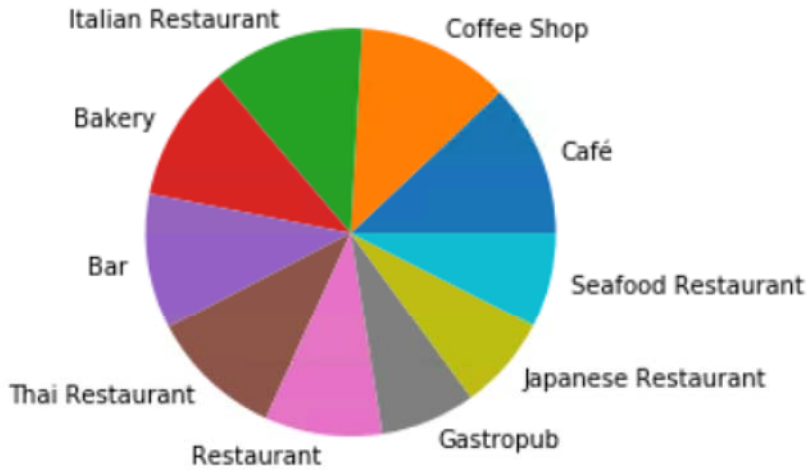
With k-mer clustering method, all the neighborhoods are successfully clustered into 5 clusters.

The popularity of different venues within these clusters are represented in following figures. It is clear from the figure that different venue categories are popular in different neighbourhood clusters and it is not a homogenous clusters.



The most popular venues in each clusters are shown in following pie charts:

Cluster 1

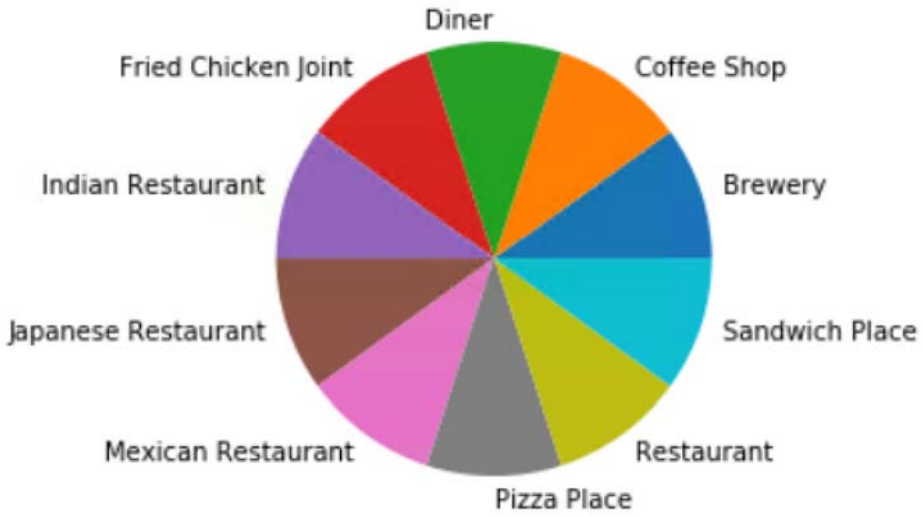


Cluster 2

Cluster 3

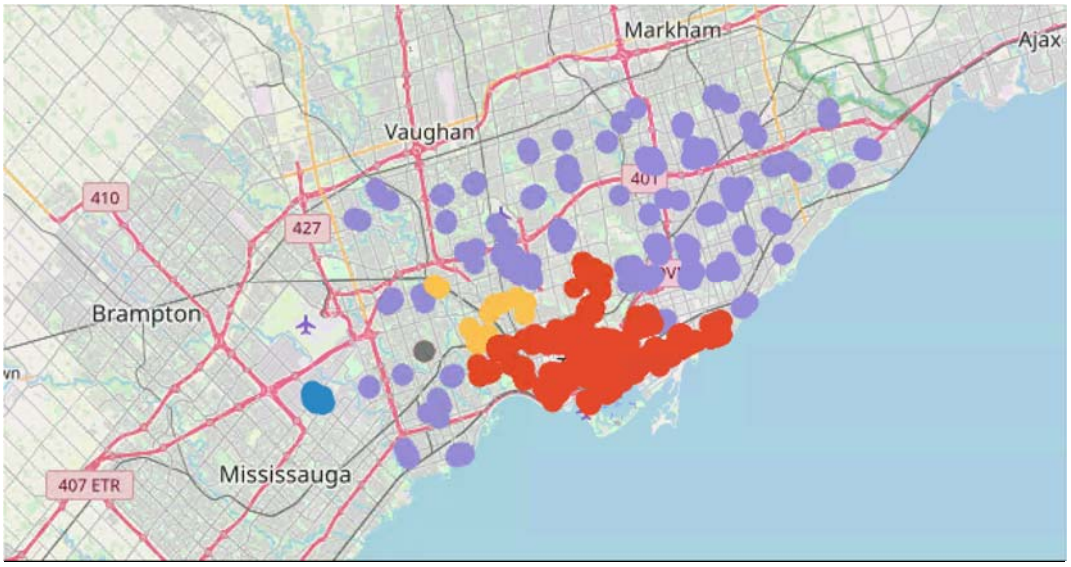
Cluster 4

Cluster 5



Discussion and Recommendations

Based on the k-mer clustering and analysis of distance from city center it is evident that the cluster number '2' (shown in orange color in following map) is fairly close to city center however it has comparatively fewer number of venues and I will recommend the stake holders to consider neighborhoods within this cluster.



Distribution of different clusters in Toronto

Regarding type venue:

From following Pie chart it is clear that there plenty of restaurants and moderate number of café/bakery in neighbourhoods within cluster 2. However there are few almost no liquor serving channels like bars and pubs hence I strongly recommend them to consider 'liquor serving channels' in these neighbourhoods.

Conclusions

The strong and negative correlation between number of venues and distance from city center is confirmed.

It is clear that different venue categories are popular within different neighborhood clusters and it is not a homogenous clusters.

The restaurants are most common type of venues In Toronto.

The bars/pubs in neighborhoods within cluster 2 seems the best choice for the stakeholders.