The Battle of Neighborhoods - Toronto

Applied Data Science Capstone by IBM on Coursera

- Neha Gupta

INTRODUCTION: BUSINESS PROBLEM

This project discusses the Neighborhoods in **Toronto** and what will be a suitable place to start a new business such as **Restaurants, Hotels**, etc.

Toronto is one of the most densely populated areas in **Canada**. Being the land of opportunity, it brings in a variety of people from different ethnic backgrounds to the core city of Canada, Toronto. Being the largest city in Canada with an estimated population of over 6 million, there is no doubt about the diversity of the population. Multiculturalism is seen through the various neighbourhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and many more. Downtown Toronto being the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.

INTRODUCTION: BUSINESS PROBLEM

The **Foursquare API** is used to access the venues in the neighborhoods. Since it returns less venues in the neighborhoods, we would be analyzing areas for which a countable number of venues are obtained. Then they are clustered based on their venues using Data Science Techniques. Here the **k-means** clustering algorithm is used to achieve the task. The optimal number of clusters can be obtained using silhouette score metrics. **Folium** visualization library can be used to visualize the clusters superimposed on the map of **Toronto** city.

This project is aimed towards **Entrepreneurs or Business** owners who want to start a new **Restaurant or Hotel** or grow their current business. The analysis will provide vital information that can be used by the target audience.

DATA REQUIREMENTS

The data required for the neighbourhoods in Toronto will be collected and prepared from Wikipedia using different techniques. The Geographical location of the neighbourhoods will be used from the Geocoder package. The Venue data will be collected from Foursquare which will be used to find a suitable place to start the business. Links used :

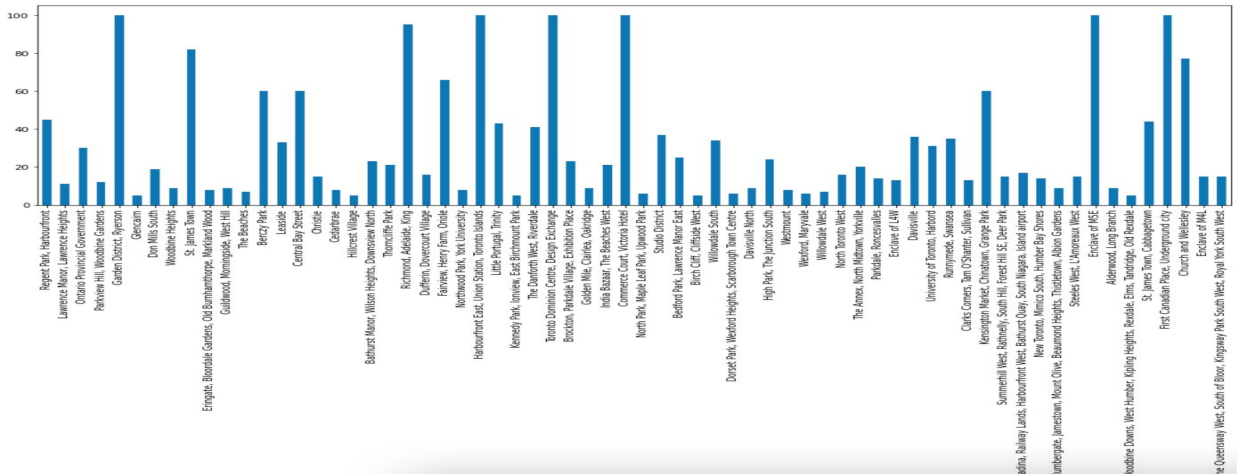
1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. https://cocl.us/Geospatial_data
3. https://foursquare.com/

METHODOLOGY

Now, we have the neighborhoods data of Toronto (**103 neighborhoods**). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **2130** venues have been obtained in the whole city and **272** unique categories. But as seen we have multiple neighborhoods with less than 5 venues returned. In order to create a good analysis let's consider only the neighborhoods with more than 5 venues. We can perform one hot encoding on the obtained data set and use it to find the 5 most common venue categories in each neighborhood. Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique has been used. To find the optimal number of clusters silhouette score metric technique is used. The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

ANALYSIS

Looking into the dataset we found that there were many neighborhoods with less than 5 venues which can be removed before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 5 or more than 5 venues were obtained.



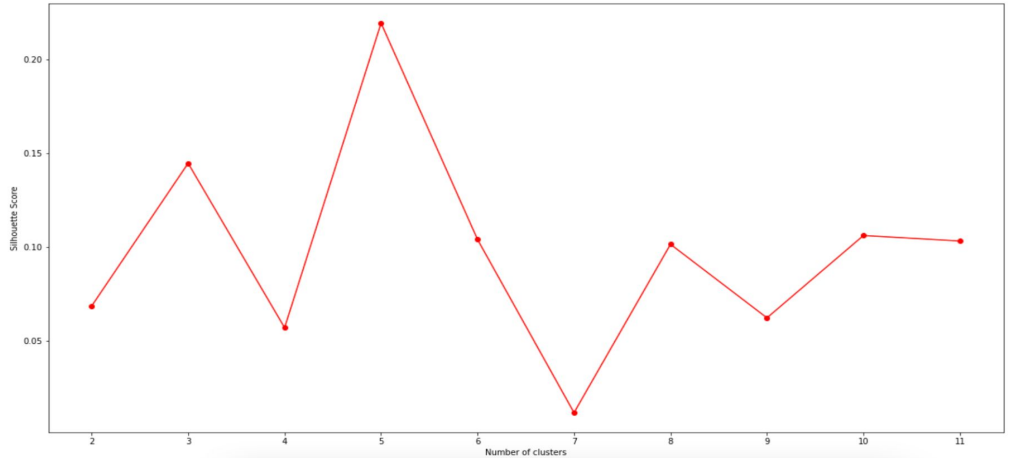
ANALYSIS

One hot encoding was performed on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. Then from the dataset the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence is extracted.

This dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the best value for K.

ANALYSIS

Here, the silhouette score is used to find the best value for K. A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot as shown below. From the plot we can see that a K value of 5 provides the best score. This K value is used for the K-Means Clustering Technique.



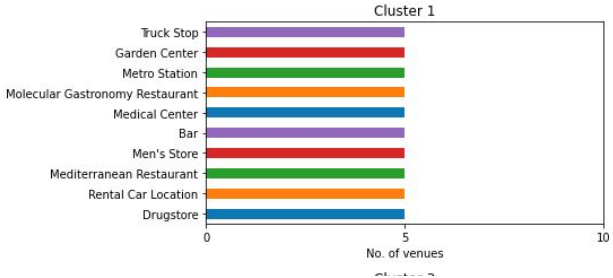
ANALYSIS

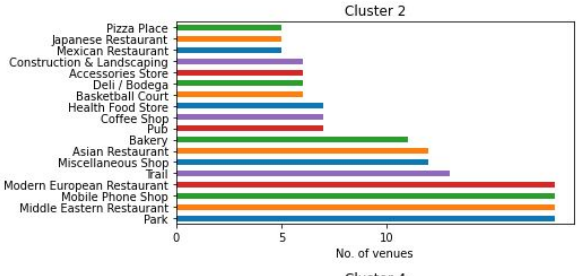
The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5 C
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa	3	Coffee Shop	Bakery	Park	Breakfast Spot	C
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant	3	Coffee Shop	Bakery	Park	Breakfast Spot	C

RESULTS AND DISCUSSION

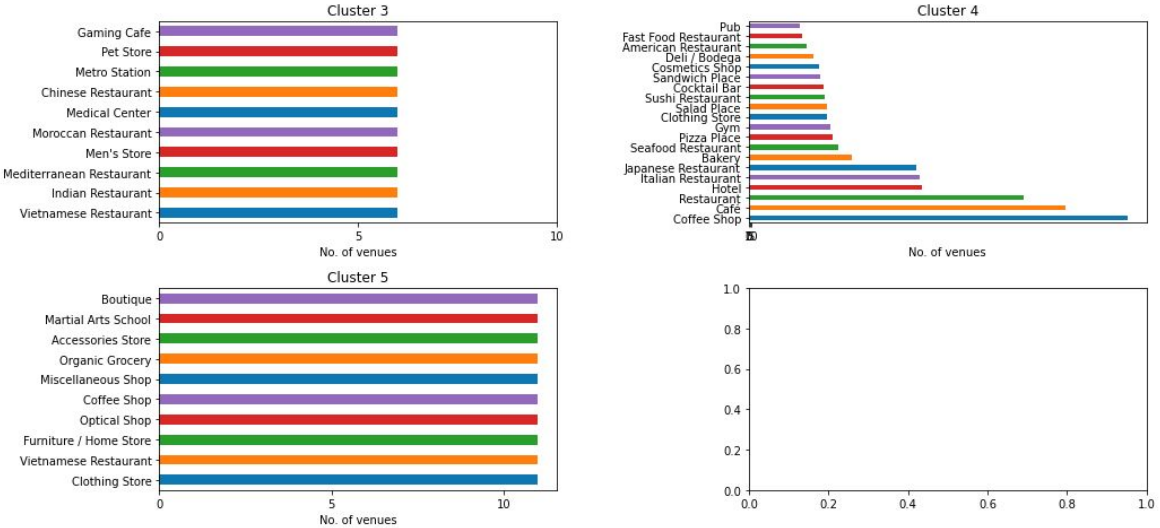
Using the clusters and the top venue categories let’s visualize the top 20 venue category in each Cluster for comparison





RESULTS AND DISCUSSION

This plot can be used to suggest valuable information to Business persons. Let's discuss a few examples considering they would like to start the following category of business.



RESULTS AND DISCUSSION

1. **Restaurant**

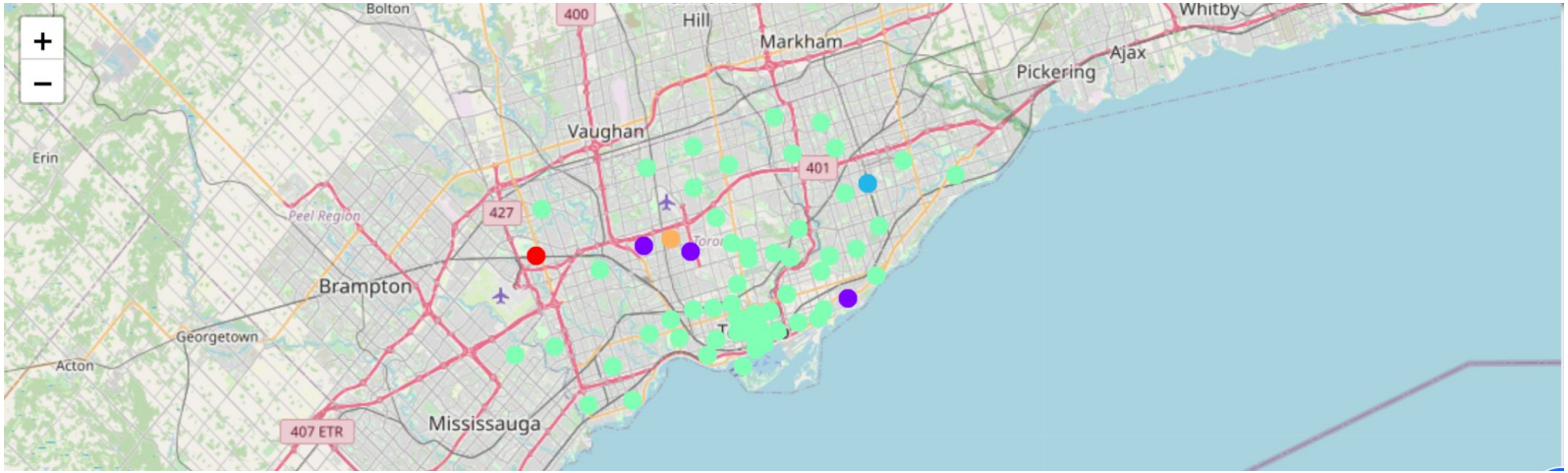
The neighborhood in cluster 2 has the greatest number of restaurants and also it has different kinds of restaurants such as American, Sushi, Seafood,etc. Hence opening one here is not the best choice. One can think of opening a specific type of restaurant in Cluster 1 or 3 since there are only a few specific types of places there. Cluster 5 is also a good choice for a Hotel or restaurants since there are less number of places there. Other factors such as places to be explored in the vicinity by the customers can also be considered by looking at the venues in the plot.

2. **Medical Centre**

The neighborhoods 1 and 3 have a notable number of medical stores whereas other clusters hardly have any. Hence the suitable cluster would be the Cluster 2 and Cluster 4 and 5. Cluster 5 has a Martial Arts School and many other shops which gives an advantage.

RESULTS AND DISCUSSION

Below figure shows a map of Toronto with the neighborhood clusters superimposed on top of it. This map can be used to suggest a vast location to start a new business based on the category.



DRAWBACKS

- A major limitation of this project was that the Foursquare API returned only few venues in each neighborhood.

- As a future improvement, better data sources can be used to obtain more venues in each neighborhood.

- This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model using KNN clustering algorithm

CONCLUSION

Purpose of this project was to analyze the neighborhoods of Toronto and create a clustering model to suggest personal places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 5 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 5 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster. A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only a few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.

Thank you!