

# COURSERA IBM DATA SCIENCE CAPSTONE PROJECT

---

## New City Same Hood

---

*Authors:*

William NGANA  
195801950

Toronto, ON

Date Started: Feb,2020

# **1 Introduction**

## **1.1 Background**

Toronto is Canada's largest city, I has many different neighbourhoods with many communities. As a prospective business owner you when establishing your business you need to choose an area that would be optimal for you. The area that a business is built in can make or break it. Build in the wrong place a business can fail, but built in the right place the business will thrive and grow. Having a proper understanding of the community and the business that are already there is of great importance to prospective business owners.

## **1.2 Problem**

Some data that can be looked at to figure out the best place to build you new business would be, the restaurants that already exist in the area, the demographics of the area, for example building an Italian restaurant in Lil' Italy would probably be a good idea. As well since the rent price in Toronto is so high we can look at price range. If you are looking to start a high end restaurant you wouldn't want to build in an with a concentration of fast food. That will be the aim of the project to locate the best areas in the City of Toronto to build a business by trying to groups the different post codes in Toronto and seeing which areas are similar.

## **1.3 Interest**

This problem would be of interests to anyone looking to start a restaurant or getting into the restaurant business in Toronto and trying to identify a place to put it.

## 2 Data Acquisition and Cleaning

### 2.1 Sources of Data

The main source of data that I'm going to use in this project comes from here this is a table of all the boroughs and neighbourhoods in Toronto. I will also be making calls to ForSquare api and from there be able to extract relevant venue data for each borough. The api call will then return the a list of all the top 50 venues in the area and the type of venue it is.

### 2.2 Data Cleaning and Feature Selection

The data that was scrapped from wiki and the data that was retrieved from the api call have been combined to form one table. This was done in several steps that I will outline below

The data frame that was the result of scraping the wiki page had various miss in data points, so I am needed to clean up the data a bit. I will do this by dropping the rows in which the value of the borough is "Not assinged". Then I will combine all the rows with the same postcode value and combine the neighbourhoods separated by a comma. Using data from online I then added the Longitude and Latitude for each postcode.

The next step was to take the data frame I had and for each Post Code determine the top 50 venues that were around it. This was done with the `getNearbyVenues` method which returned a dataframe with the post code the latitude and longitude of the neighbourhood the venue and the longitude and latitude of the venue as well as the venue category. For this particular project I didn't require all of the columns so I'm going to create a new dataframe where the Neighborhood and venue Latitude/Longitude is dropped. For this project what we are look for most is the kinds of venues in each area. So our features will be the Postal Code of the area and the number of each kind of venue of each possible categories. To do this we will first need to extract the unique categories from the dataframe then sum up how many of each are in a particular postal code.

[ PostalCode,Venue category 1,Venue category 2,...] and so on will be the columns of the data frame. The final step was to determine how many of those occure in each neighbourhood. To do this we first need to group by the neighbourhood then count the number of occurrences of a particular type in a neighborhood.

## 3 Analysis

### 3.1 Clustering

The next step in this project was the analysis of all the data. To do so I used the k-means clustering technique. That is because the data that I have is unlabelled and I would like to organize them into groups of neighbourhoods that are similar. For the clustering I set the number of clusters to 5 and then fit the data to the model. The resulting cluster labels were inserted into the original dataframe with the latitude and longitude. A map was then created with the neighbourhoods with the same clustering label shown in the same color.

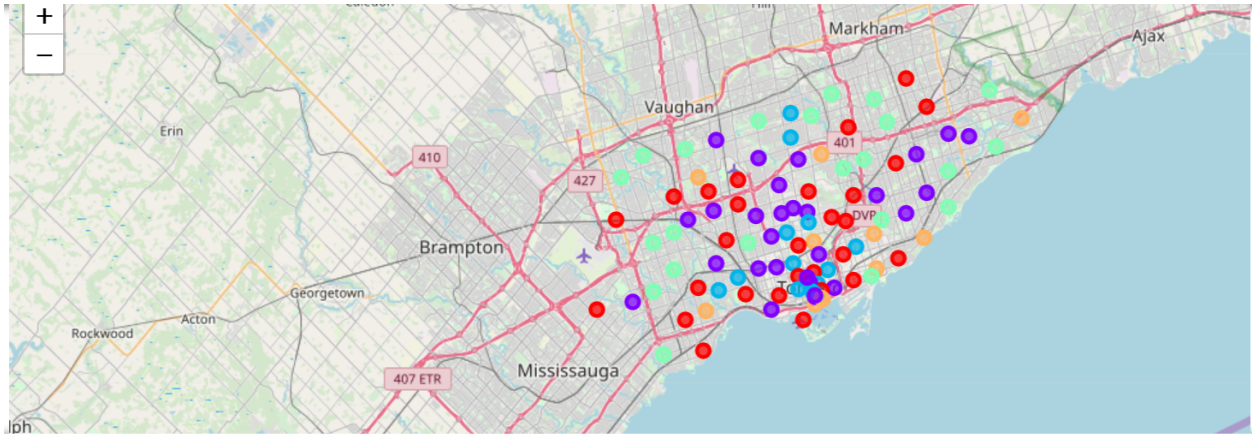


Figure 1: An image of Toronto and the different clustering labels

### 3.2 Cluster Analysis

From this the next step was to look at each cluster and try and determine what makes each of the different clusters unique. To do this I used a pie chart to try and figure out the split of different venues in different clusters.

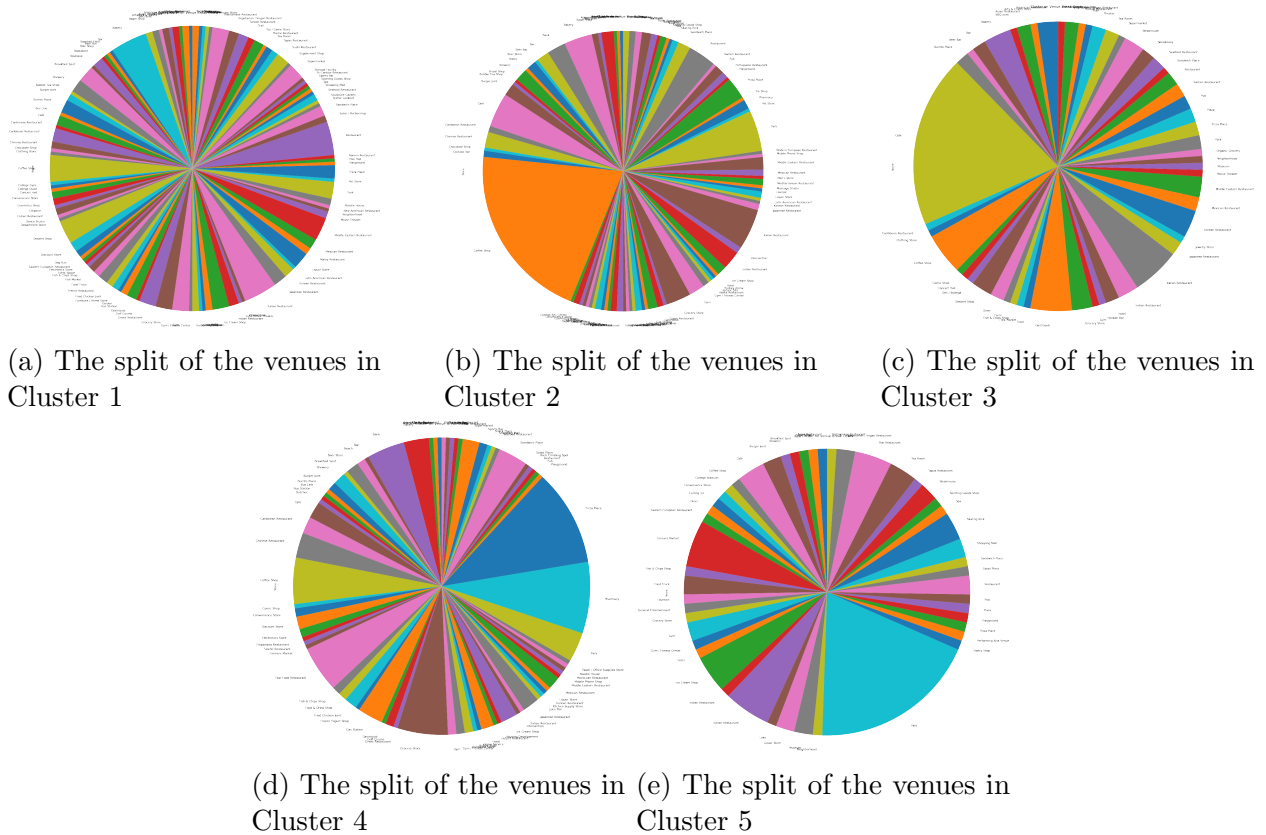


Figure 2: A look at all of the clusters

## 4 Discussion

As we can see the split between all of the clusters differs greatly. When looking at cluster 1 in particular you see that there isn't one major venue that dominates the cluster. It is comprised of 28 different postal codes and 112 different venues, of which the restaurant, coffee shop, desert shop, bakery and Italian restaurant. Going back to the original problem posed at the start of this project. What kind of business would succeed in this particular cluster based on the results I'd say that for the most part this cluster has no specific winner.

Looking at Cluster 2, this cluster contains 28 postal codes and 89 different venue categories. Of those 89 the dominate venue category is coffee shop. This would indicate that in these particular postal codes the most popular venues by far are coffee shops so it would be the best place to start a coffee shop.

Looking at Cluster 3, this cluster contains 13 postal codes and 54 different venue categories. Of those 54 the dominate venue category is cafe. This would indicate that in these particular postal codes the most popular venues by far are cafes so it would be the best place to start a cafe.

Looking at Cluster 4, this cluster contains 22 postal codes and 77 different venue categories. Of those 77 the dominate venue category is a split between pizza places and pharmacies. This would indicate that in these postal codes the businesses that do well are those two.

Looking at Cluster 5, this cluster contains 11 postal codes and 49 different venue categories. Of those 49 the dominate venue category is park. This would indicate that in these particular postal codes if you were to start a business the best thing would be to do something park related.

## 5 Conclusion

### 5.1 Next Steps

This is a very elementary look at the different postcodes in the city of Toronto. The next step into this analysis would be to look at a more in depth break down of the venues. Grouping each of them in to specific larger categories. For example all the restaurants get grouped together and from there we can find out in which postcodes restaurants do best in in general and from there we can the have a breakdown and look what kind or restaurant does better where. That or we can take a closer look at the current clusters that we have and look at economics of the areas. The different rent prices in the postcodes, and a demographic breakdown. All this would give us more information that will allow for a better grouping and for us to be better equipped to cluster the different postcodes in Toronto.