**Battle of the Neighborhoods**

Finding the best Neighborhood in Toronto using Data Science

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

Table of Contents

1. Introduction
2. Target Audience
3. Data Overview
4. Methodology
5. Discussion
6. Conclusion

**1. Identifying the Business Problem**

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 neighborhoods 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.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the ‘best’ neighborhood in Toronto to open a restaurant. Pizza and Pasta are one of the most bought dishes in Toronto originating from Italy. Toronto is the fourth largest home to Italians with a population of over 500k, there are numerous opportunities to open a new Italian restaurant. Through this project, we will find the most suitable location for an entrepreneur to open a new Italian restaurant in Toronto, Canada.

2. Target Audience

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

3. Data Overview

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Italian restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open an Italian restaurant.

3.1 Data acquisition:

Source 1: Toronto Neighborhoods via Wikipedia



Figure 1:Wikipedia Page showing List of Neighborhoods in Toronto with respective Postal Codes

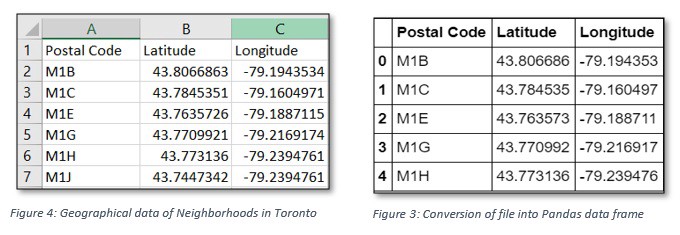
1. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

The Wikipedia site shown above provided almost all the information about the neighborhoods. It included the postal code, borough and the name of the neighborhoods present in Toronto. Since the data is not in a format that is suitable for analysis, scraping of the data was done from this site (shown in *figure2*).



Figure 2: Data that was scraped from Wikipedia site and put into Pandas data frame

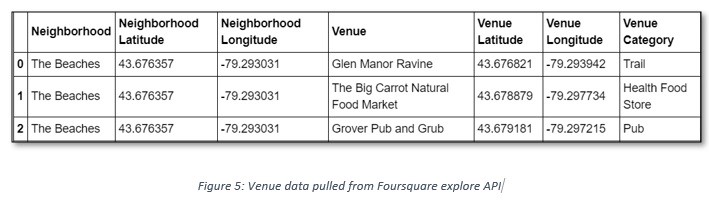
Source 2: Geographical Location data using Geocoder Package



2. <https://cocl.us/Geospatial_data>

The second source of data provided us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so we had to attach it to a Pandas data frame(shown in figure 3).

Source 3: Venue Data using Foursquare



We performed a bit of data cleansing. It is seen through figure 5 (above) that the neighborhoods are grouped by the name of the neighborhood, so data clustering is made easier later.

4. Methodology

4.1 — Data Cleansing

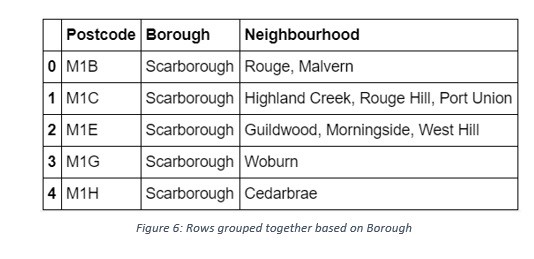
After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Only the cells that have an assigned borough will be processed. Borough’s that were not assigned get ignored.

2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in *Figure2*row 4.

3. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on the borough as shown below.



Using the Latitude and Longitude collected from the Geocoder package, we merged the two tables together based on Postal Code.

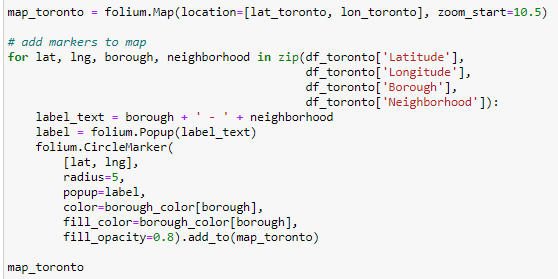


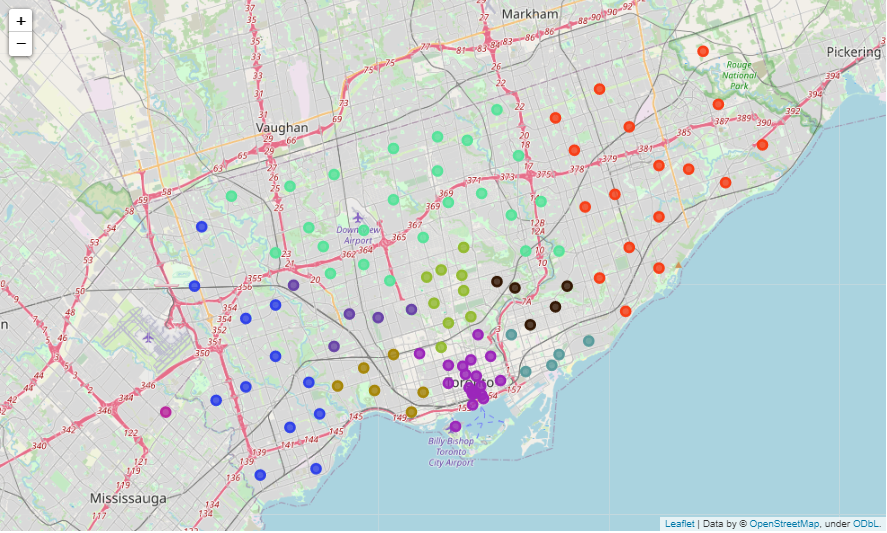
After, the venue data pulled from the Foursquare API was merged with the table above providing us with the local venue within a 500-meter radius shown below.



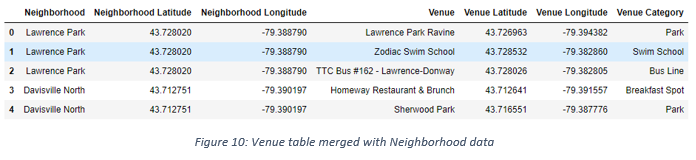
4.2 — Data Exploration

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in.



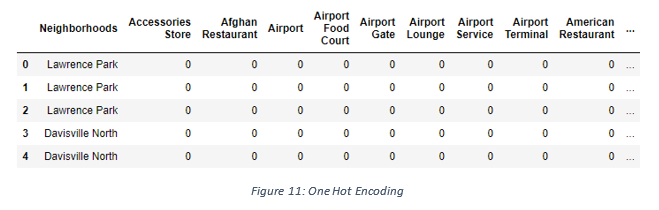
This snippet of code provided us with the map below:

Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Italian Restaurants all over Toronto. There was a total of 45 Italian Restaurants in Toronto. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

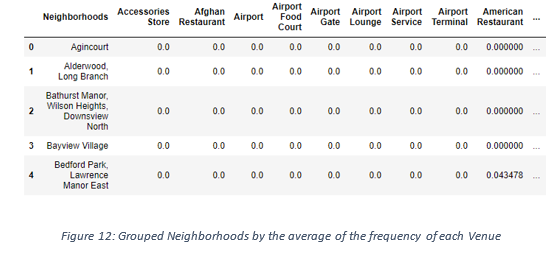


4.3 — Machine Learning

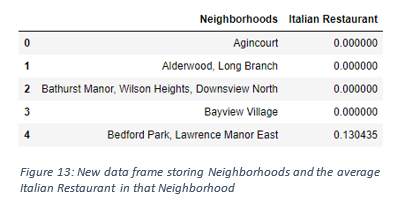
Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were in each neighborhood.



Then we grouped those rows by Neighborhood and by taking the **average** of the frequency of occurrence of each Venue Category.

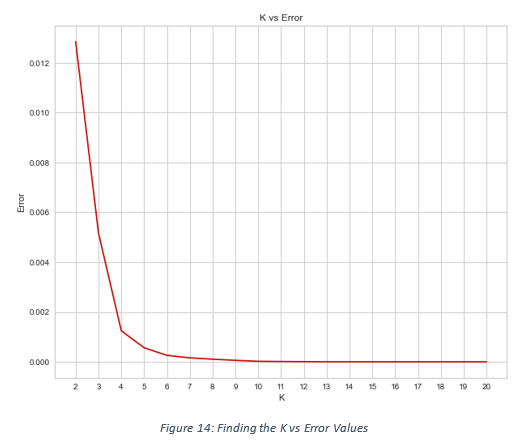


After, we created a new data frame that only stored the Neighborhood names as well as the mean frequency of Italian Restaurants in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

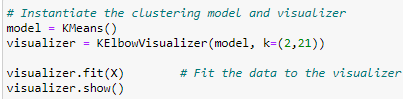


K-Means Clustering

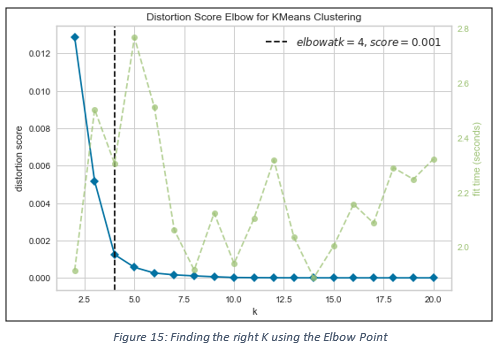
To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Italian Restaurants in that Neighborhood. To do this we used **K-Means**clustering. To get our optimum K value that was neither overfitting nor underfitting the model, we used the **Elbow Point** Technique. In this technique, we ran a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has the sharpest turn. In our case, we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.



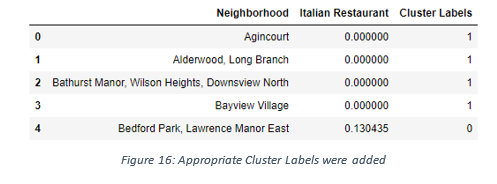
Then we used a model that accurately pointed out the optimum K value. We imported ‘*KElbowVisualizer*’ from the *Yellowbrick package.*Then we fit our K-Means model above to the Elbow visualizer.



This gave the model below:

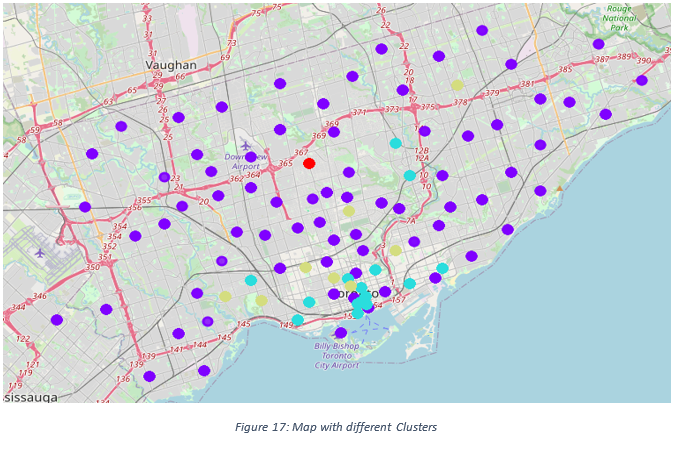


We just integrated a model that would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighborhoods that had a similar mean frequency of Italian Restaurants were divided into 4 clusters. Each of these clusters was labelled from 0 to 3 as the indexing of labels begins with 0 instead of 1.



After, we merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a new Italian Restaurant in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.

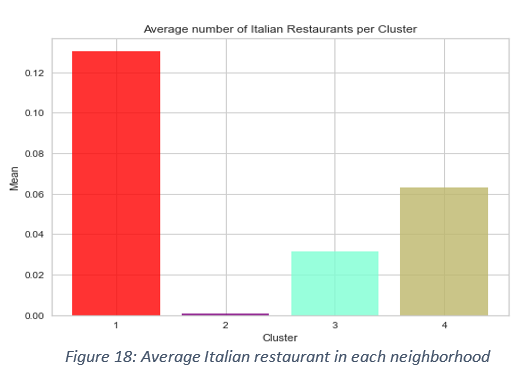
* Cluster 1 — Red
* Cluster 2 — Purple
* Cluster 3 — Turquoise
* Cluster 4 — Dark Khaki

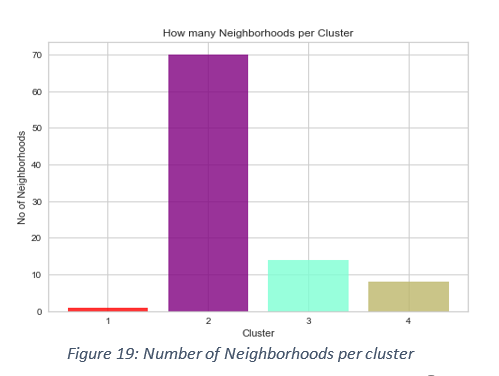


The map above shows the different clusters that had a similar mean frequency of Italian restaurants.

4.4 — Data Analysis

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one let's check the total amount of neighborhoods in each cluster and the average Italian Restaurants in that cluster. From the bar graph that was made using Matplotlib (figure 18), we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the least neighborhoods (1) while cluster 2 has the most (70). Cluster 3 has 14 neighborhoods and cluster 4 has only 8. Then we compared the average Italian Restaurants per cluster.

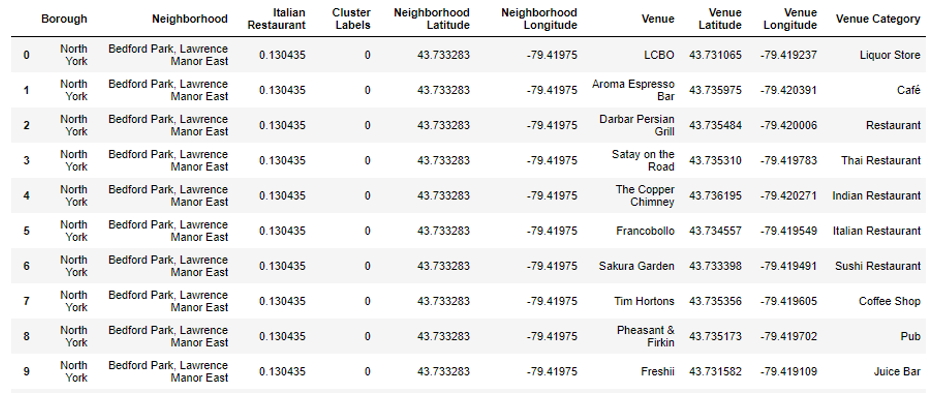




Cluster Analysis

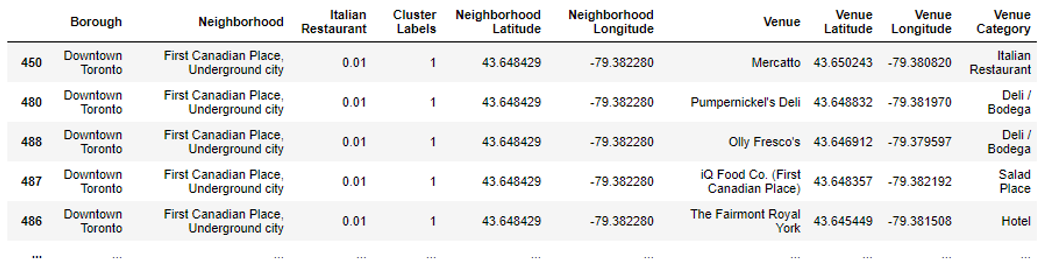
This information is crucial as we can see that even though there is only 1 neighborhood in Cluster 1, it has the highest number of Italian Restaurants (0.1304) while Cluster 2 has the most neighborhoods but has the least average of Italian Restaurants (0.0009). The average of the average Italian Restaurant made up the data for Figure 18. Also, from the map, we can see that neighborhoods in Cluster 2 are the most sparsely populated. Now let’s analyze the Clusters individually (Note: these are just snippets of the data).

*Cluster 1(Red):*



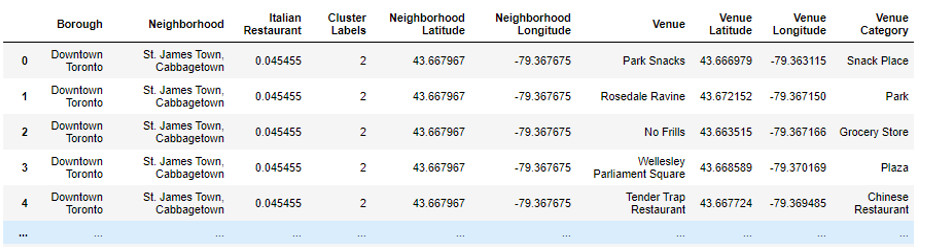
Cluster 1 was in the North York area. Bedford and Lawrence Manor East were the two Neighborhoods that were in that cluster. Cluster 1 had 19 unique Venue locations and out of those only 3 were Italian Restaurants. Cluster 1 had the highest average of Italian Restaurants equating to 0.130435. The reason why the average of Italian Restaurants is the highest is that all these Restaurants are in two neighborhoods, Bedford and Lawrence Manor East.

*Cluster 2 (Blue):*



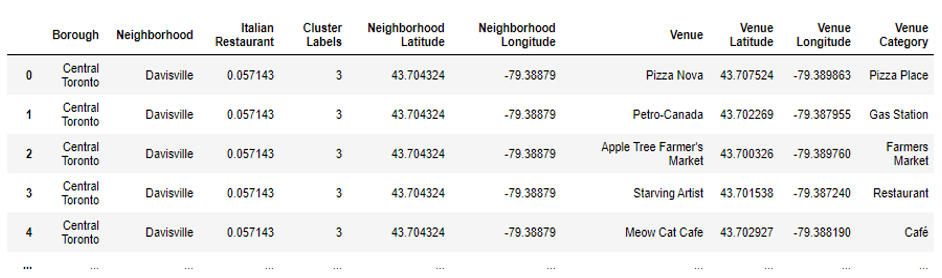
There was a total of 70 neighborhoods, 229 different venues and only 1 Italian Restaurant. Therefore, the average amount of Italian Restaurants that were near the venues in Cluster 2 is the lowest being 0.01. In the map, we can see that nodes of Cluster 3 were dispersed all throughout Toronto making it one of the most sparsely populated clusters.

*Cluster 3 (Turquoise):*



Cluster 3 had the second to lowest average of Italian Restaurants. Cluster 3 was mainly located in the Downtown area but also had some neighborhoods in West Toronto, East Toronto and in North York. Neighborhoods such as Ryerson, Toronto Dominion Center, Don Mills, Garden District, Queen’s Park and many more were included in this cluster. There was a total of 176 unique venues and out of those 27 were Italian Restaurants.

*Cluster 4 (Dark Khaki):*



Cluster 4 venues were in the Downtown, West, East and Central Toronto areas as well as Scarborough. Neighborhoods such as Central Bay Street, University of Toronto, Central Bay Street and Riverdale were some of the neighborhoods that made up this cluster. There was a total of 91 unique Venues in Cluster 4 with 16 Italian Restaurants. This made up the second-highest average of Italian Restaurants in that cluster which was approximately 0.063.

Therefore, the ordering of the average Italian Restaurant in each cluster goes as follows:

1. Cluster 1 (≈0.1304)

2. Cluster 4 (≈0.0632)

3. Cluster 3 (≈0.0317)

4. Cluster 2 (≈0.0009)

5. Discussion:

Most of the Italian Restaurants are in cluster 1 represented by the red clusters. The Neighborhoods located in the North York area that have the highest average of Italian Restaurants are Bedford Park and Lawrence Manor East. Even though there is a huge number of Neighborhoods in cluster 2, there is little to no Italian Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second last average of Italian Restaurants. Looking at the nearby venues, the optimum place to put a new Italian Restaurant in Downtown Toronto as there are many Neighborhoods in the area but little to no Italian Restaurants, therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Adelaide and King, Fairview, etc. which is in Cluster 2. Having 70 neighborhoods in the area with no Italian Restaurants gives a good opportunity for opening a new restaurant. Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the Italian population across neighborhoods as this can play a huge factor while choosing which place to open a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

6. Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was like how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the Beautifulsoup Web scraping Library. We also visualized utilizing different plots present in seaborn and Matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.