

The Battle of Neighborhoods

Opening a Pizza Shop in Toronto

Capstone Project – Final Report

Week 2

Mustafa Ayub

January 31st, 2021

Contents	Page Number
<hr/>	
1. Introduction	2
1.1 <i>Background:</i>	2
1.2 <i>Business Problem:</i>	2
1.3 <i>Target Audience:</i>	2
2. Data Requirement	3
2.1 <i>List of postal codes of Toronto from Wikipedia:</i>	3
2.2 <i>Geographical coordinates of Toronto:</i>	3
2.3 <i>Foursquare API:</i>	4
3. Methodology	5
3.1 <i>Data pre-processing and exploration:</i>	5
3.2 <i>Predictive modeling and clustering:</i>	8
4. Results and Discussion	10
4.1 <i>Visualizing Clusters:</i>	10
4.2 <i>Cluster Analysis:</i>	
4.2.1 <i>Cluster 0 (red)</i>	11
4.2.2 <i>Cluster 1 (dark blue)</i>	11
4.2.3 <i>Cluster 2 (light blue)</i>	12
5. Conclusion	13
6. References	14

1 Introduction

1.1 Background:

Toronto, a beautiful city, is well known for being a diverse and multicultural city. Almost half its population was born outside Canada. Multiculturalism has proven to be the backbone of Toronto and a key strength. One of the key factors bringing people together in such a multicultural city is variety of food. While a mix of multicultural and ethnic foods remain a bonding factor among people, there are some other food items that are in high demand by everyone; regardless of their ethnicity or background. Let's take an example - Pizza!!!

According to a report by the Technomic, Pizza remains one of the most popular and cravable foods in the foodservice landscape. It is simple food, yet comes with sharable portions, can be enjoyed in the venue, or takeout to the parks / home / birthday & office parties, etc.

1.2 Business Problem:

An investor who recently moved to Canada from Europe is willing to invest in a food business that is liked by majority of Canadians; regardless of their ethnic background. After doing some research, he has decided that Pizza shop would be an ideal option, but would like to know *where is the best location to open his Pizza shop in Toronto*.

Though this capstone project, I will leverage my data science knowledge gained throughout the course along with key data (Wikipedia, Geospatial coordinates and Foursquare API) to identify where would be an ideal location for a pizza shop.

1.3 Target Audience:

- The investor who is willing to invest in Toronto to open up the pizza shop
- Other Data analysts and Data scientists who may leverage the analysis for other similar projects
- Anyone interested in finding more about the Toronto neighborhoods and what it has to offer

2 Data Requirement

In order to complete this analysis, we would need access to the data from reliable sources. Having gone through week3 assignment and being introduced to the following reliable sources of the data, I will be using them throughout the Capstone project.

2.1 List of Postal code of Canada from Wikipedia:

The data is in a tabular format and we can use different techniques (i.e., BeautifulSoup) to retrieve the information from the Wikipedia table. This data will help provide the list of the following Pieces of information:

[“https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M”](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

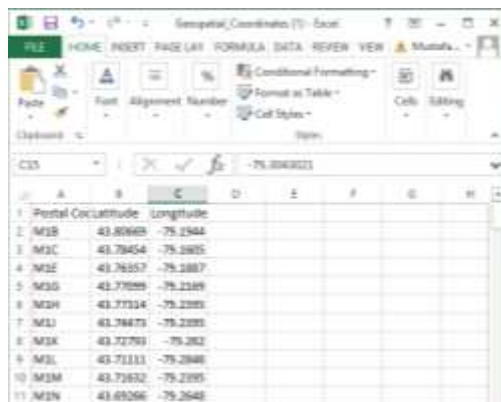
Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned
M9A	Eatonville	Islington Avenue, Humber Valley Village
M1B	Scarborough	Midtown, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson

Figure 1: list of postal code

2.2 Geospatial coordinates of Toronto:

This data (.csv file) will allow us to obtain the geospatial coordinates of each borough that will help us during our data analysis along with the data obtained from Wikipedia and Foursquare API. This file will allow us to pull the following info:

[“https://cocl.us/Geospatial_data”](https://cocl.us/Geospatial_data)



	Postal Code	Latitude	Longitude
1	M1B	43.80609	-79.2944
2	M1C	43.79454	-79.2895
3	M1D	43.76557	-79.2887
4	M1E	43.77089	-79.2389
5	M1F	43.77314	-79.2393
6	M1G	43.78873	-79.2393
7	M1H	43.72793	-79.282
8	M1J	43.71113	-79.2846
9	M1K	43.73632	-79.2393
10	M1L	43.69266	-79.2648
11	M1N	43.69266	-79.2648

Figure 2: Geospatial coordinates

2.3 Foursquare API:

Foursquare is a social location service that allows users to explore the world around them. The Foursquare API allows application developers to interact with the Foursquare platform and obtain necessary info; i.e., pull a list of venues along with its features, etc. For this project, I will be interested more into the following data fields (just to name a few) in order to be able to navigate through Toronto, build clusters, etc.:

<https://developer.foursquare.com/docs>

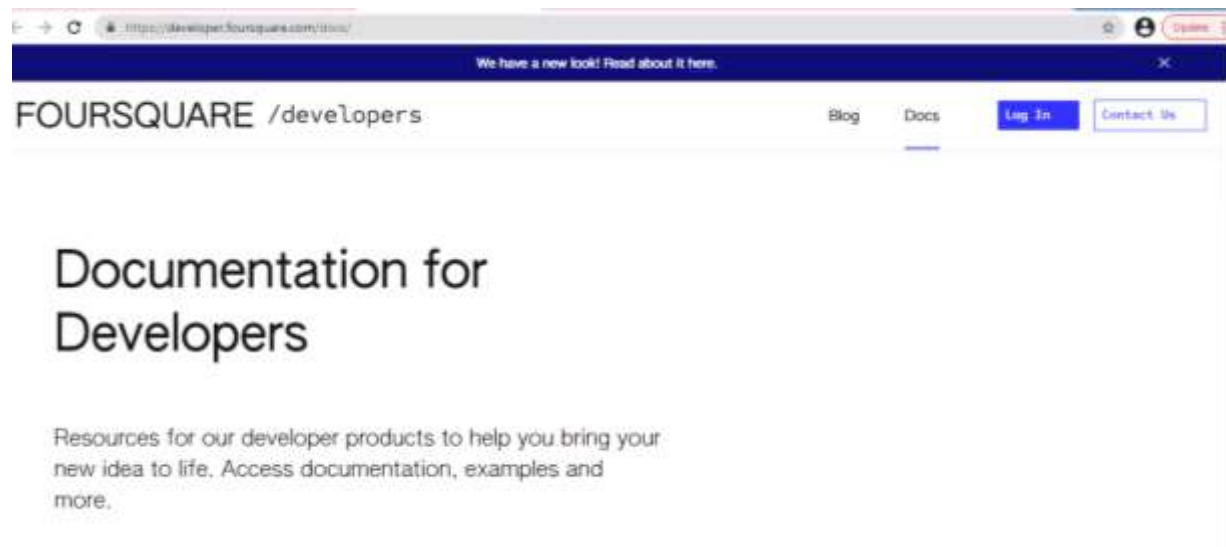


Figure 3: Foursquare API developer site

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Careful & Reliable Painting	43.752622	-79.331957	Construction & Landscaping
2	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop
3	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
4	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena

Figure 4: pre-processed data sample Foursquare

3 Methodology

3.1 Data pre-processing and exploration:

Once all the sources of data were identified, the next step was pre-processing the data. The goal of the exercise was to format the data in such manner that can be leveraged throughout analysis. The first data table we wanted to import into Python was the list of postal codes from Wikipedia website.

While there are many different approaches, *BeautifulSoup* methodology was followed to scrape the list of postal code from Wikipedia page and import into Python.

	Postal Code	Borough	Neighborhood
0	M1A\n	Not assigned\n	Not assigned\n
1	M2A\n	Not assigned\n	Not assigned\n
2	M3A\n	North York\n	Parkwoods\n
3	M4A\n	North York\n	Victoria Village\n
4	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n

Figure 5: raw scraped postal code data from wikipedia

Looking at the data, there are few items that jumped immediately to address:

- Remove the unnecessary suffix “\n” from the columns
- Exclude any boroughs that are unassigned
- Make sure all neighborhoods are assigned to a borough

Following pre-processing of the postal code data, the following desired output was achieved:

	Postal Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 6: pre-processed postal code data

The next data source required for the analysis was to obtain the latitude and longitude of each neighborhood. This was particularly important as it was used in creating the Folium map and cluster analysis later in the analysis.

Similarly while there are numerous ways to import this data into Python, the methodology used here was simply importing it from the Geospatial coordinates file (.csv) provided during the course. Once the data was imported into Python, it was merged with the formatted postal code data to create the following data frame:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 7: Merged data frame from postal code and geospatial data

This final data frame had all the data columns that was required, along with data from Foursquare API, to complete the analysis. This data could help pinpoint any neighborhood, along with the postal code, borough, latitude and longitude.

The data frame contained 10 boroughs and 103 neighborhoods. While this does not seem a lot, it would still be quite challenging for someone to decide where they would like to open a Pizza shop. Using Folium map, a visual representation of the different neighborhoods was generated to show how the neighborhoods were spread out across Toronto.



Figure 8: Visual representation of Toronto neighborhoods

While the Folium map shows how the neighborhoods were spread out in Toronto, we would want to overlay the representation of all venues available in each of those neighborhoods.

The Foursquare API allows application developers to interact with the Foursquare platform and obtain necessary info; i.e., pull a list of venues along with its features, etc. Leveraging Foursquare API, the lists of all available venues in Toronto were imported into Python. After grouping the data frame by neighborhoods, the following data frame was created; containing 271 unique venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Figure 9: Foursquare API pre-processed data

Furthermore, a plot was created to see the top 10 venues and identify how pizza shop ranks against other venues.

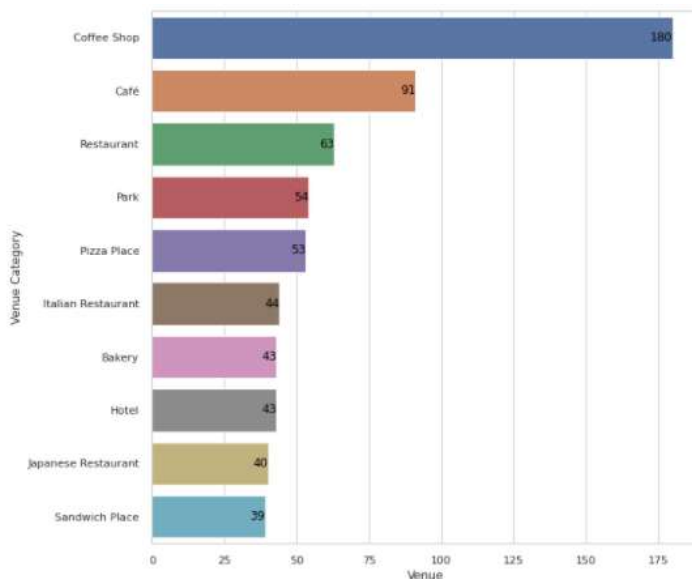


Figure 10: Top 10 venues in Toronto

The plot above indicates that Pizza shop is top 5th venue in the database; with a total of 53 pizza shops spread out across Toronto. This is amazing as it confirms the investor's initial research that the demand for Pizza shop is high in Toronto!

3.2 Predictive Modeling and Clustering

Looking at the data obtained from Foursquare API, the first step would be to prepare the data for machine learning using a two-step approach:

- Used one-hot-encoding to convert the categorical data into numerical data
- Grouped the data by neighborhood to obtain the average frequency of the venue occurrence per each of the given neighborhoods in the dataset

Neighborhood	Type	Amusement Store	Art Store	Art Food Court	Art Gift	Art Lounge	Art Salon	Art Terrace	Art Restaurant	Art Shop	Art Studio	Art Gallery	Art Museum	Art & Crafts Store	Bar Restaurant	Bar & Sports	Auto Garage	Auto Workshop	BBQ Joint	Bakery Store	Book Shop	Bakery	Bar	Bar	Beach Club	Beach Club
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Alderwood Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Bathurst Water Village Highgate Dundasville North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Bedford Park Lawrence Silver Star	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 11: Foursquare API pre-processed data post one-hot-encoding exercise

Furthermore a summarized view of relationship between neighborhoods and pizza place was created as follows:

	Neighborhood	Pizza Place
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.333333
2	Bathurst Manor, Wilson Heights, Downsview North	0.045455
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.043478

Figure 12: Neighborhood to Pizza place relationship

Above data suggests that there are some areas that have no pizza places while some has a fair number of pizza places; i.e., one out of three venues in Alderwood, Long Branch is a pizza shop!

The next activity was grouping similar neighborhoods containing similar number of pizza stores together and understanding their underlying pattern. *KMean* clustering, an unsupervised machine learning method, was used to complete this task. First activity was to determine the optimum number of centroids (*k*) for the data that would represent the center of each cluster. While there are many different techniques of determine the optimal number of centroids or *k*, “KElbow” technique was leveraged for this exercise:

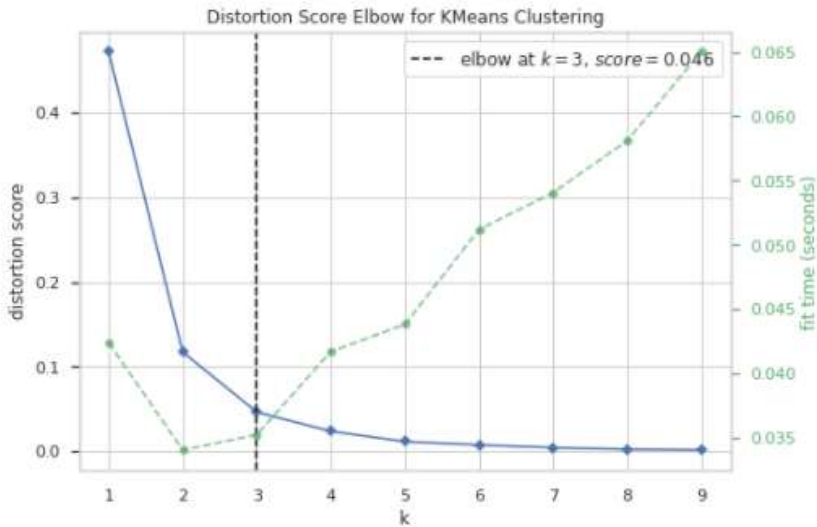


Figure 13: KElbow output; indicating $k = 3$

The figure above indicates that the neighborhoods can be divided into three different clusters. This is evident from the dotted line intersecting with the dark blue line at $k = 3$.

Now that the number of k was defined, the data frame was divided into 3 clusters using *KMean* clustering method:

```
: kclusters = 3

to_grouped_clusters = pizza_df.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(to_grouped_clusters)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

53]: array([0, 2, 1, 0, 1, 0, 0, 0, 1, 0], dtype=int32)

Figure 14: KMean clustering screen shot

4 Results & Discussion

4.1 Visualizing clusters:

Following the *KMean* clustering exercise, below is the revised table highlighting the relationship between neighborhoods and pizza places; only difference is that a cluster has been applied for each respective neighborhood:

	Neighborhood	Pizza Place	Cluster Labels
0	Agincourt	0.000000	0
1	Alderwood, Long Branch	0.333333	2
2	Bathurst Manor, Wilson Heights, Downsview North	0.045455	1
3	Bayview Village	0.000000	0
4	Bedford Park, Lawrence Manor East	0.043478	1

Figure 15: data frame with added cluster labels

For example, looking at the above two neighborhoods belonging to cluster 1 have very similar frequencies; i.e., 0.045 and 0.043.

Now that each neighborhood is assigned a cluster label, a visual representation using the Folium map with color coded location identifiers was created as follows:

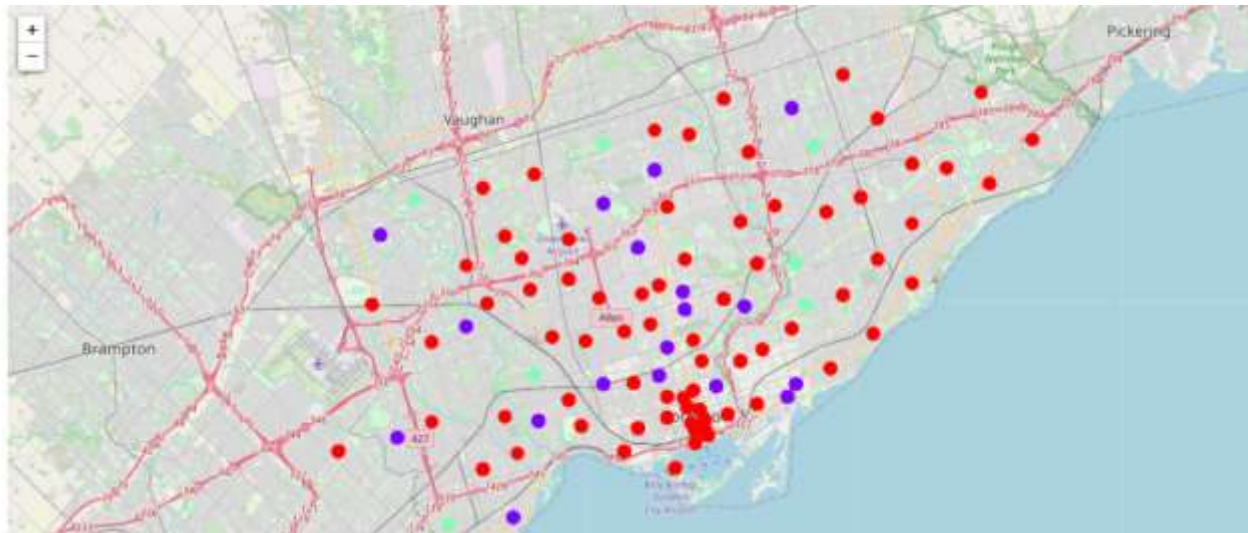


Figure 16: Visual representation of clustered neighborhoods

The map displays how the neighborhoods are spread out along with their cluster labels as follows:

- *Cluster 0: red*
- *Cluster 1: dark blue*
- *Cluster 2: light blue*

4.2 Cluster Analysis:

Let's take a look at what the clusters imply:

4.2.1 Cluster 0 (Red):

Let's examine Cluster 0 - red

```
df_clusters.to_merged.loc[df_merged['cluster_label'] == 0, df_merged.columns[[0] + list(range(1, df_merged.shape[1]))]]
df_clusters.head()
```

	Neighborhood	Pizza Place	Cluster Label	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	0.0	0	43.79422	-79.28222	Penelope's Breakfast & Lunch	43.792370	-79.282220	Breakfast Spot
55	Parkdale-Rosedale	0.0	0	43.54595	-79.45525	Coor House	43.550550	-79.455550	Restaurant
56	Parkdale-Rosedale	0.0	0	43.54595	-79.45525	Donat Restaurant & Wine Bar	43.546225	-79.455225	Italian Restaurant
59	Parkdale-Rosedale	0.0	0	43.54595	-79.45525	Mar Sae	43.546795	-79.455310	Eastern European Restaurant
59	Parkdale-Rosedale	0.0	0	43.54595	-79.45525	Revue Cinema	43.551112	-79.455951	Movie Theater

Figure 17: cluster 0 output

Cluster 0 is mainly centered around downtown Toronto and area. While this may be a great cluster due to the higher number of office buildings, it will also be very challenging due to high rent as well as existing well established other 18 Pizza shops in the area. So cluster 0 is out of question!

4.2.2 Cluster 1 (Dark blue):

Now let's look at Cluster 1 - dark blue

```
df_clusters.to_merged.loc[df_merged['cluster_label'] == 1, df_merged.columns[[0] + list(range(1, df_merged.shape[1]))]]
df_clusters.head()
```

	Neighborhood	Pizza Place	Cluster Label	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
18	Danville	0.117947	1	43.754324	-79.355750	Provenance Pizzeria Series	43.755230	-79.355945	Pizza Place
19	Danville	0.117947	1	43.754324	-79.355750	Pizzeria	43.757251	-79.355520	Pizza Place
19	Danville	0.117947	1	43.754324	-79.355750	Shoppers Drug Mart	43.757559	-79.355550	Pharmacy
8	Business Reply Mail Processing Centre, South C.	0.035524	1	43.852744	-79.321555	East End Garden Centre & Hardware	43.854554	-79.324471	Garden Center
19	Danville	0.117947	1	43.754324	-79.355750	Bellet Tru-Cuisine	43.755555	-79.355220	Thai Restaurant

Figure 18: cluster 1 output

Cluster 1 seems to have the most number of Pizza shops (25). Since the goal of the investor is to open up an authentic pizza shop (i.e., not franchised), this will also be a tough competition to win easily. While he could overcome the competition in the long run, it may be hard to realize any benefits immediate. Cluster 1 also seems to be out of question!

4.2.3 Cluster 2 (Light blue):

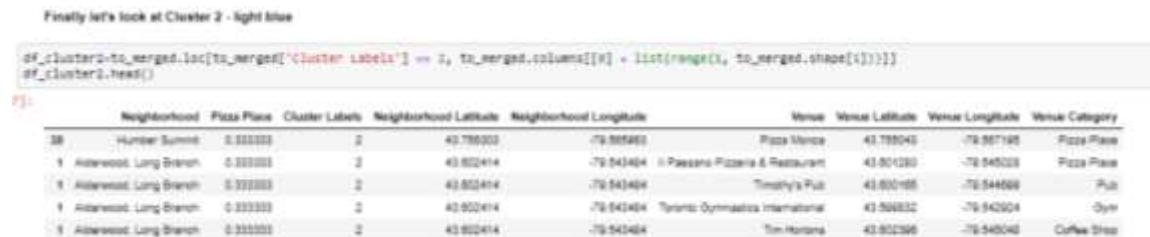


Figure 19: Cluster 2 output

Cluster 2 has the least number of pizza shops (10) with density in only few neighborhoods. Is that potentially due to the fact that people in cluster 2 are not interested in pizza? This is hard to tell as no survey was conducted as part of this project to see what people like in those neighborhoods. However, this may be an area where the investor could enter the market with an authentic pizza shop that could cater to the population in the neighborhood.

Now that the option has been narrowed down to cluster 2, let's look at the number of venues per neighborhood in cluster 2 and identify an area where we have a higher presence of various venues but lower concentration of pizza shops.

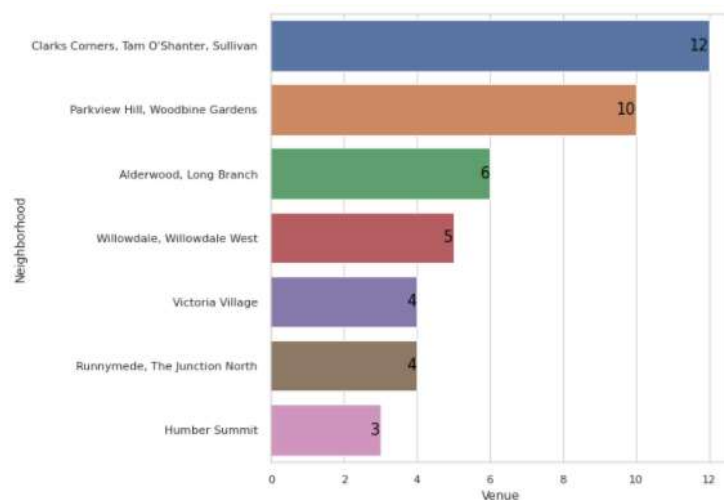


Figure 20: Cluster 2 neighborhoods with number of venues

Looking at the plot above, we see that “Clarks Corners, Tam O’shanter, Sullivan” have a higher concentration of venues. This is a great area since it seems people in these areas are interested in restaurants, cafes, etc. The investor does not need to worry about bringing traffic into the area and he just needs to market his presence, deliver authentic pizza and maintain his quality service.

Going a step even further, the data suggests that pizza shops only represent 16% of the venues in this neighborhood.

Neighborhood	Pizza Place
Clarks Corners, Tam O'Shanter, Sullivan	0.166667

Figure 21: Pizza shop concentration in this neighborhood

Based on this analysis, it is recommended that the investor could open his authentic pizza shop in Clarks Corners, Tam O'Shanter, Sullivan. This is an optimal area; lots of venues suggesting interest for dining out but less number of existing pizza shops.

5 Conclusions

While the need for big data is increasing on a daily basis, similarly the need for data storage and analysis grows at the same pace. Having acknowledge that, data science is the technique of discovering hidden patterns in the data to draw a conclusion.

In this project, we were able to demonstrate how data science can be applied to various problems; in this case finding the best location to open a pizza shop.

We used data from various sources (Wikipedia, Foursquare API), prepared data for analysis, conducted necessary analysis, applied machine learning techniques to eventually identify the best location for the pizza shop.

6 References

- Neighborhood by postal code
 - Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geospatial Coordinates (Latitude & Longitude)
 - CSV file: https://cocl.us/Geospatial_data
- Available Venues per Neighborhood
 - Foursquare API: <https://developer.foursquare.com/docs>
- Toronto Populations
 - Population Stat : <https://populationstat.com/canada/toronto>
- Statistics about Pizza
 - Technomics: <https://www.technomic.com/reports/consumer/consumer-trend-reports/pizza>