Setting shop in the neighborhoods of Toronto

Introduction

Toronto is the provincial capital of Ontario and the most populous metropolitan area in Canada with a population close to 3 million (as of July 2018). It is the most multicultural diverse city on the planet with over 180 languages and dialects being spoken. It's estimated that over half of Toronto's residents were born outside Canada. Toronto is also a global hub of commerce, technology, entertainment, culture and is constantly ranked as one of the world's most livable and competitive cities.

Business Problem

Mark already owns a courier business, delivering documents and large parcels to homes in the boroughs of Toronto. Recently he has observed a rise in people preferring to buy organic food and fresh farm produce because more and more people are adopting conscious health lifestyle. He is now thinking of expanding his income stream by offering a service to people or restaurants that would like organic and fresh farm produce delivered directly to their door step. In so doing, Mark eliminates the time consuming necessity of people going to the market every day. He has talked to some farmers and they have agreed to sell their products at a discount price, provided that he is able to move a large volume of goods to market quickly. His potential clients are office workers, restaurants or individuals with a taste for organic food. Having saved some money, from his other business he is now looking for a borough/neighborhood where he can build/rent a warehouse. The criteria for this borough/neighborhood should be:

1. The population of people must be large enough so that he has a ready supply of customers.
2. Average income of people must be high enough to be able to buy this type of food.
3. Rent or buying the warehouse should be in an ideal location that is not costly.
4. A reasonable number of restaurants that can potentially be his customers, e.g., more family themed restaurants instead of fast food outlets.

This analysis can also be applicable to:

- An individual looking for an apartment to rent or buy in Toronto,
- A restaurateur looking to open a new restaurant in any neighborhood,
- Or this can serve as a generalized information source on the neighborhoods of Toronto.

Data Section

Source of Data

The data I used in this project was gathered from a variety of online sources using web scraping libraries such as Beautiful Soup. The **first data set** for the project involved web scrapping a Wikipedia page on the demographics of the neighborhoods of Toronto.

The link to the page is: *https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods*. From this page we can extract relevant census data applicable to this study such, population for each neighborhood, population density per square kilometer for each neighborhood, average income per neighborhood etc. From the first data set we can obtain answers to points 1 and 2 as laid out in the Business section problem.

Two answer points 3 and 4 we need to explore the neighborhoods in Toronto. This involves clustering and segmenting the neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a data set that contains the Toronto boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood. The **second data set** was created by combing two data sources.

1. A data source that contains a list of the boroughs and the neighborhoods. This kind of information can be found by web scrapping the Wikipedia page: *https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'*.

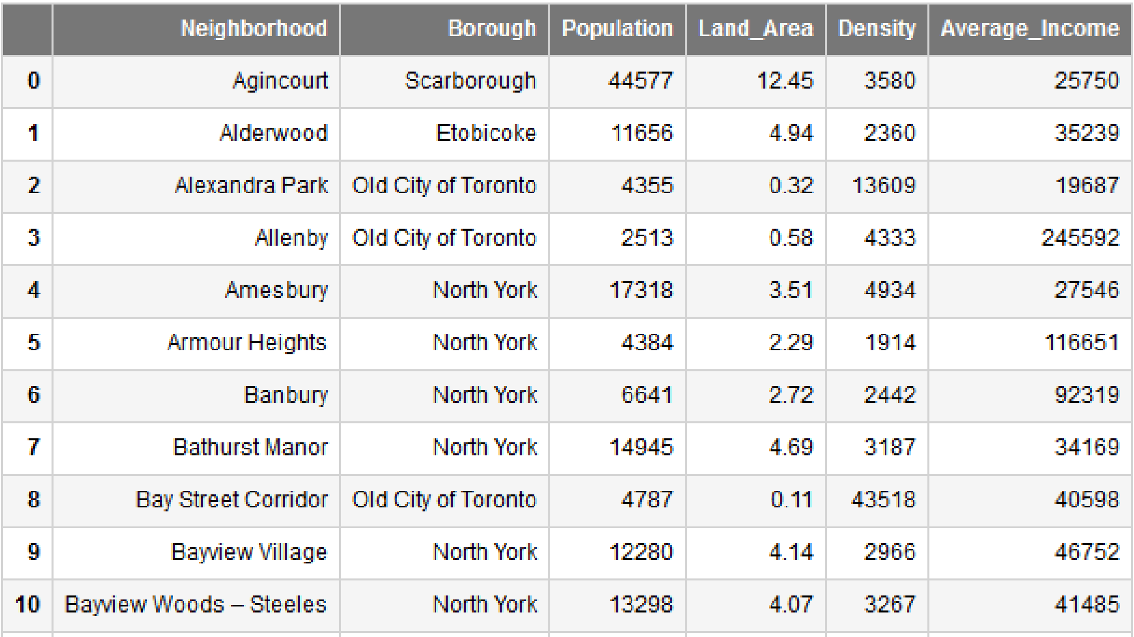
2. A geospatial data source that will gives the exact locations of the neighborhoods within Toronto using latitude and longitude values. To obtain the geographical coordinates the Geocoder python package can be used. In my case I obtained a ready made *csv* file that has the geographical coordinates of each postal code from:*http://cocl.us/Geospatial_data*.

    - With this data set and using the Foursquare API I can obtain data about different venues in different neighborhoods of each borough. The Foursquare Places API provides location based experiences with diverse information about venues. The Folium library is used for map visualizations in all cases.

Methodology

Exploratory Data Analysis

The Wikipedia page of the demographics contains the table of the boroughs and neighborhoods of Toronto, area, population and average income etc. I have used Beautifulsoup4 and pandas library to create the initial data-frame. For a clean and understandable data frame, I have maintained the names of the boroughs as they appear in the Wikipedia page. The reader may notice that the names of the boroughs scrapped from the two Wikipedia pages are different. The final table used for data exploration is shown below:



	Neighborhood	Borough	Population	Land_Area	Density	Average_Income
0	Agincourt	Scarborough	44577	12.45	3580	25750
1	Alderwood	Etobicoke	11656	4.94	2360	35239
2	Alexandra Park	Old City of Toronto	4355	0.32	13609	19687
3	Allenby	Old City of Toronto	2513	0.58	4333	245592
4	Amesbury	North York	17318	3.51	4934	27546
5	Armour Heights	North York	4384	2.29	1914	116651
6	Banbury	North York	6641	2.72	2442	92319
7	Bathurst Manor	North York	14945	4.69	3187	34169
8	Bay Street Corridor	Old City of Toronto	4787	0.11	43518	40598
9	Bayview Village	North York	12280	4.14	2966	46752
10	Bayview Woods – Steeles	North York	13298	4.07	3267	41485

Figure 0.0.1: *A snap shot of the final table of demographics after webscrapping and cleaning.*

To get a glimpse of the structure of the neighborhoods the following charts were produced:

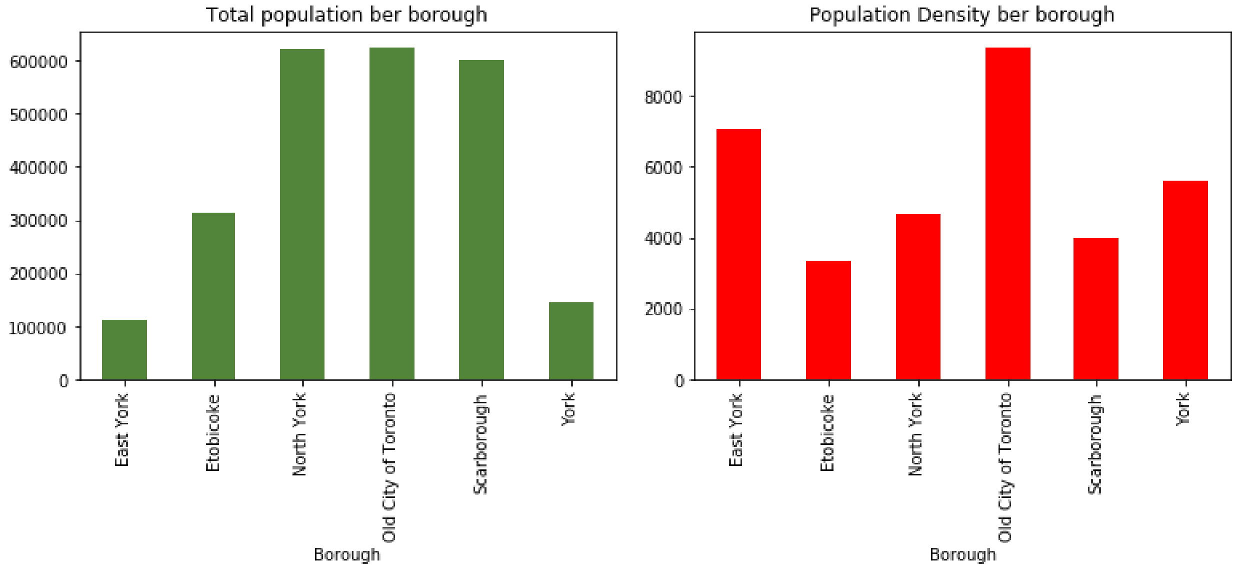


Figure 0.0.2: *Bar charts depicting total population in numbers and population density (per km^2) in each borough.*

From the above bar charts we can see that the Old City of Toronto has the largest population of all boroughs at 624 900, with North York a close second with 621 000. Even though the Old City of Toronto and North York may have roughly the same population, the population densities are very different. North York has half the density of the Old City of Toronto.

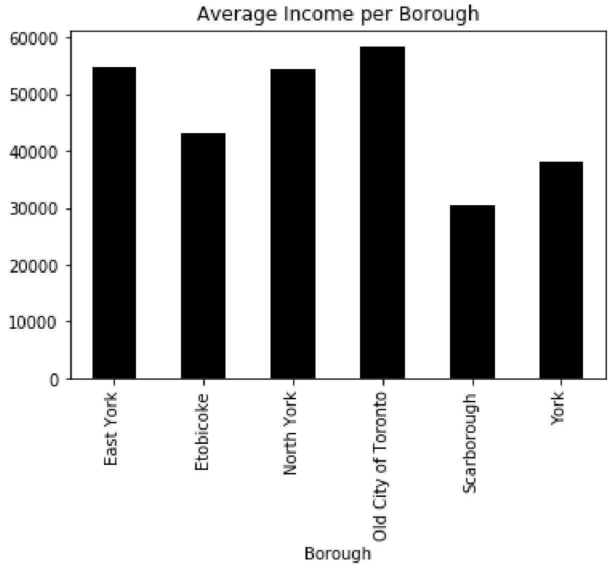


Figure 0.0.3: *A bar graph of the average income per borough.*

In terms of average income, Old city of Toronto has highest average income of the boroughs, with $58 400. Other boroughs are not far behind, with East York and North

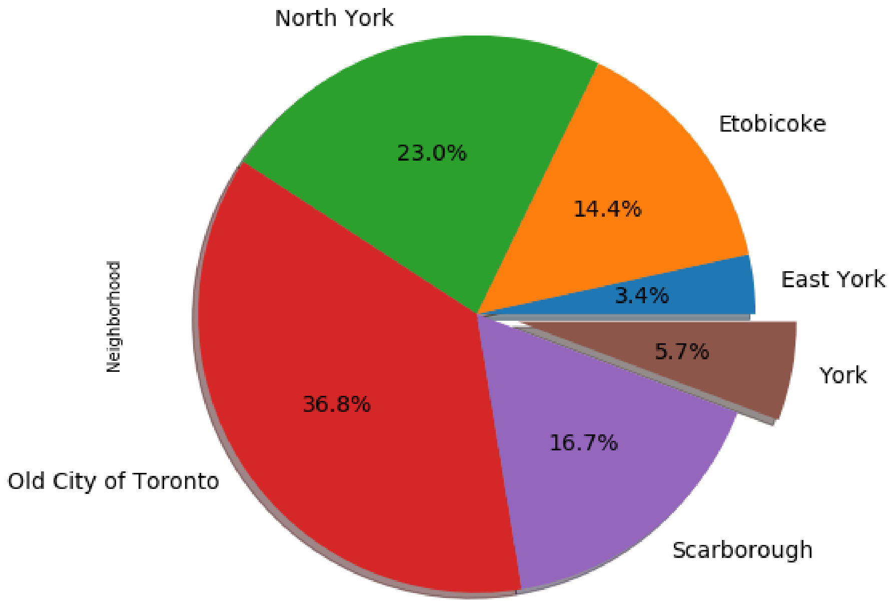


Figure 0.0.4: *Pie chart showing percentage of neighborhoods per borough of the total.*

York having an average income of $54 600 and $54 400 respectively.

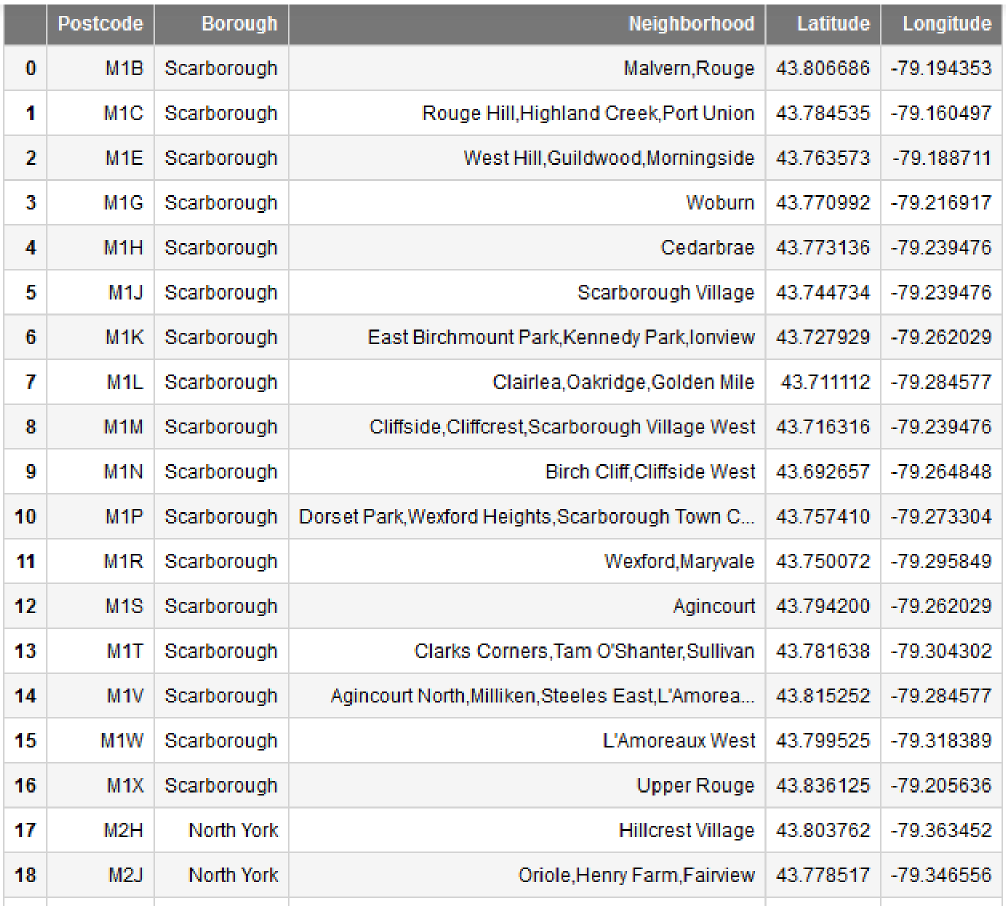
The pie chart answers the question of *'Which borough has the highest number of neighborhoods?'*.

The Old City of Toronto has the most number of neighborhoods, as seen from the pie chart above. Old City of Toronto has 37% of the total neighborhoods in Toronto, which equates to 64 out of 174. To view the numbers that were used to generate the charts, the reader is encouraged to see the Jupyter notebook.

Clustering by Venues per neighborhood using Foursquare Data

K-means clustering was used to answer points 3 and 4 in the Business problem section on page 2. Scikitlearn's KMeans clustering was used to determine similar neighborhoods based on a restaurant as venue category. The point here is to find a ware house location that is in a borough that has the most family oriented restaurants and has the largest number of neighborhoods. But also the location to be within a 5km radius to homes, offices etc. Added to being in a high income neighborhood would result in more business opportunity because of the higher disposable income of the residents.

From the preliminary data exploration on page 4, it is clear that the Old City of Toronto borough is the clear choice for Mark to look for a ware house for his delivery business. It has the highest population density and average income. Therefore in our clustering we will focus on the Old City of Toronto neighborhoods.



	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern,Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill,Highland Creek,Port Union	43.784535	-79.160497
2	M1E	Scarborough	West Hill,Guildwood,Morningside	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park,Kennedy Park,Ionview	43.727929	-79.262029
7	M1L	Scarborough	Clairlea,Oakridge,Golden Mile	43.711112	-79.284577
8	M1M	Scarborough	Cliffside,Cliffcrest,Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff,Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park,Wexford Heights,Scarborough Town C...	43.757410	-79.273304
11	M1R	Scarborough	Wexford,Maryvale	43.750072	-79.295849
12	M1S	Scarborough	Agincourt	43.794200	-79.262029
13	M1T	Scarborough	Clarks Corners,Tam O'Shanter,Sullivan	43.781638	-79.304302
14	M1V	Scarborough	Agincourt North,Milliken,Steeles East,L'Amorea...	43.815252	-79.284577
15	M1W	Scarborough	L'Amoreaux West	43.799525	-79.318389
16	M1X	Scarborough	Upper Rouge	43.836125	-79.205636
17	M2H	North York	Hillcrest Village	43.803762	-79.363452
18	M2J	North York	Oriole,Henry Farm,Fairview	43.778517	-79.346556

Figure 0.0.5: *Final data frame to be used for segmenting and clustering using the K-Means algorithm.*

Figure 0.0.5 shows the data frame that was used in clustering. The process of web scraping and cleaning of the data is detailed in the Jupyter notebok accompanying this report.

Below is a visualization of the neighborhoods in question superimposed on a Toronto map, created using the Folium library.

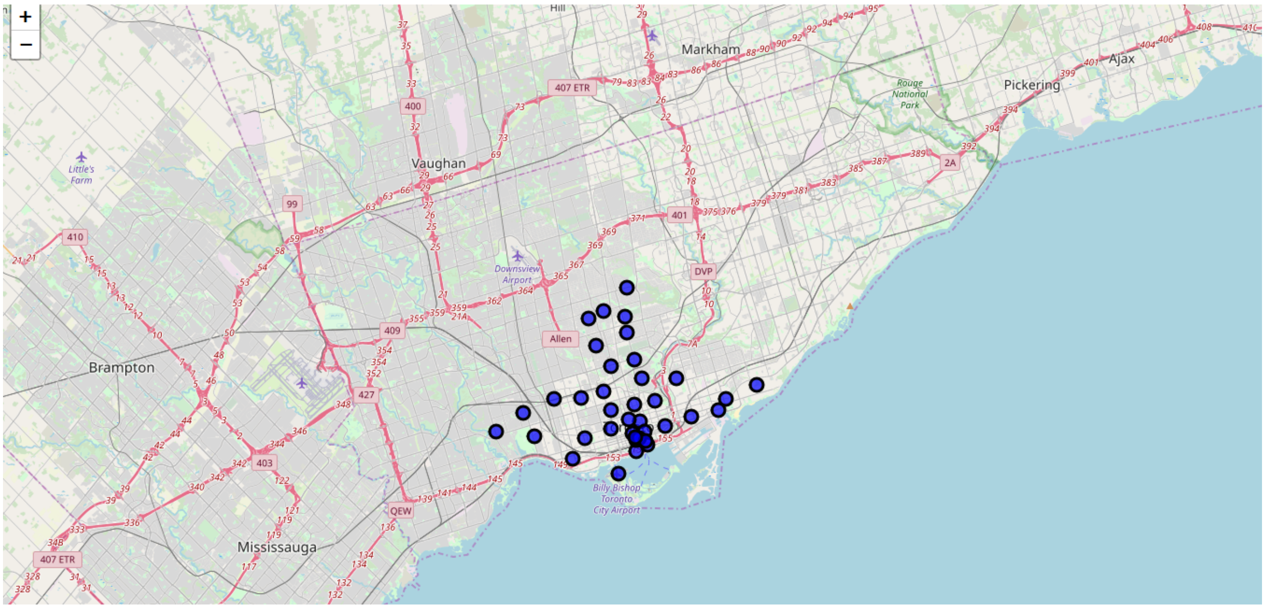


Figure 0.0.6: *Superimpose and visualize the 'Toronto' neighborhoods in the city of Toronto map.*

Results and Discussion

K-Means clustering algorithm was run on five different clusters based on the venue category. The clusters are shown below:

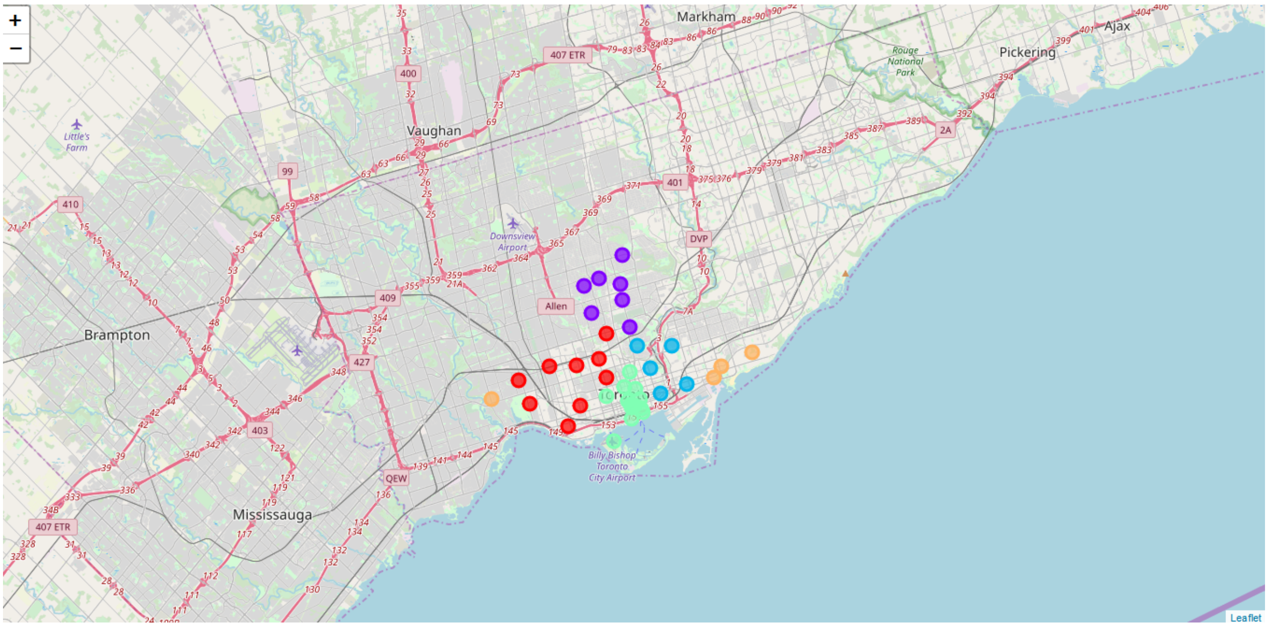
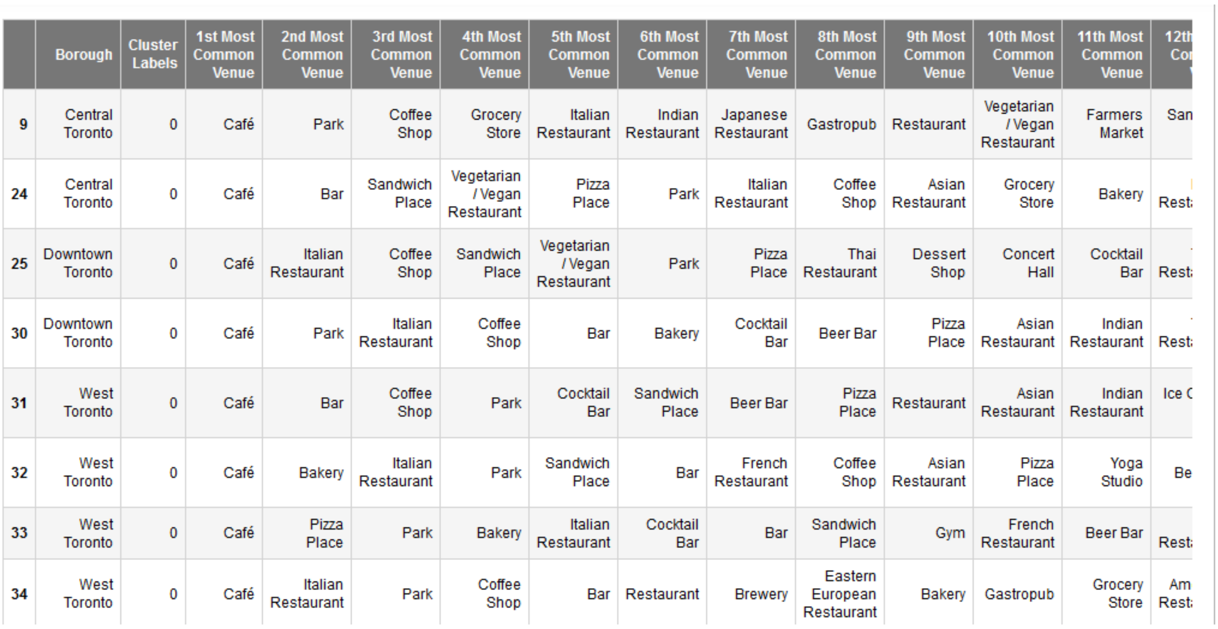


Figure 0.0.7: *The five clusters generated from the K-Means algorithm.*

Each cluster shows a list of neighborhoods with their respective top venue categories. All the clusters are almost similar in size and are not that far dispersed from each other.



	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Co
9	Central Toronto	0	Café	Park	Coffee Shop	Grocery Store	Italian Restaurant	Indian Restaurant	Japanese Restaurant	Gastropub	Restaurant	Vegetarian / Vegan Restaurant	Farmers Market	San
24	Central Toronto	0	Café	Bar	Sandwich Place	Vegetarian / Vegan Restaurant	Pizza Place	Park	Italian Restaurant	Coffee Shop	Asian Restaurant	Grocery Store	Bakery	Rest
25	Downtown Toronto	0	Café	Italian Restaurant	Coffee Shop	Sandwich Place	Vegetarian / Vegan Restaurant	Park	Pizza Place	Thai Restaurant	Dessert Shop	Concert Hall	Cocktail Bar	Rest
30	Downtown Toronto	0	Café	Park	Italian Restaurant	Coffee Shop	Bar	Bakery	Cocktail Bar	Beer Bar	Pizza Place	Asian Restaurant	Indian Restaurant	Rest
31	West Toronto	0	Café	Bar	Coffee Shop	Park	Cocktail Bar	Sandwich Place	Beer Bar	Pizza Place	Restaurant	Asian Restaurant	Indian Restaurant	Ice C
32	West Toronto	0	Café	Bakery	Italian Restaurant	Park	Sandwich Place	Bar	French Restaurant	Coffee Shop	Asian Restaurant	Pizza Place	Yoga Studio	Be
33	West Toronto	0	Café	Pizza Place	Park	Bakery	Italian Restaurant	Cocktail Bar	Bar	Sandwich Place	Gym	French Restaurant	Beer Bar	Rest
34	West Toronto	0	Café	Italian Restaurant	Park	Coffee Shop	Bar	Restaurant	Brewery	Eastern European Restaurant	Bakery	Gastropub	Grocery Store	Am Rest

Figure 0.0.8: *A snapshot of part of cluster 1.*

The results from the exploratory data analysis and clustering can be listed as:

1. Cafe's are the most popular type of eatery in Toronto neighborhoods.

2. Cluster 1 and Cluster 2, mainly found in the boroughs of Downtown Toronto, West Toronto and Central Toronto has a lot of restaurants that I would call family themed as compared to fast food outlets (see Figure 0.0.8).

3. Fast food outlets are not many in the neighborhoods analyzed.

What we find from the clustering is that there is definitely a lost of restaurants in the 'Toronto' that could serve as Mark's customers. Since the clusters are not far dispersed from each other, the best thing for Mark would be to find a central location that is at best equidistant, i.e., a ware house location that is in the middle of the five clusters on the Toronto map. This would make it fast and easy to transport goods to all the neighborhoods. Ideally a centrally located warehouse should shorten shipping times since there's an increased chance that customers are geographically close to the warehouse.

Conclusion

From this project we can see how machine learning can be used in a real life data science project to gain insight to a business problem. In this case K-Means clustering was used to segment and cluster the neighborhoods of Toronto to find the best location for a ware house . Using open source Python libraries and Foursquare API's I was able to leverage web data to answer pressing questions to a given problem. I would however note that, I assumed the rental prices for buildings are the same for the 'Toronto' labeled boroughs. This is obviously not the case. Also the rental price will depend on the size of the building per square meter. This project could be expanded to include the effect of rental price on the choice of location. This improved project would could serve as a recommendation study for a professional realtor.