Capstone Project Report

The Battle of Neighbourhoods

1 Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada with a population of 2,731,571 in 2016. Toronto is a center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

1.1 Problem Statement

The city generates lots of data which can be utilized to answer questions and help the residents of Toronto. The data is scattered, comes from different sources and is difficult to analyze. The objective of this project will be to gather data for Toronto city neighbourhoods from different sources and present it in a structured way so that following useful questions could be answered.

- Determine what types of venues exist around different neighbourhoods of Toronto and in what quantity.
- Compare neighbourhoods of Toronto city based on different parameters of interest such as demographics and venues.
- Segment and cluster neighbourhoods
- Help the audience in useful decision making

1.2 Audience:

The targeted audience can be the entrepreneurs, businessmen and investors such as a restaurant owner who wants to open a new restaurant in Toronto. The data would also be helpful for a person who is relocating and wants to compare available options for the residence.

2 Data

The data will give information different aspects of neighbourhoods. The aspects will be demographics parameters such as population, area, income etc. of the people in different neighbourhoods. The other aspects will cover the points of interest and the venues that exist around different neighbourhoods.

The data will be collected from following sources.

- Toronto Open Data is a good source for collecting demographics data. This is a government website and a useful data source.
  https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/
- Data regarding venues will be collected using Foursquare.
  https://foursquare.com/v/city-of-toronto/4c50d7d7250dd13a12fa377c
- Some useful data can be collected from Wikipedia pages as well.
  https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto
  https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

Once the data from different sources is aggregated in a structured form, it will be analyzed to answer the questions relevant to a particular audience and their interest. Unsupervised machine learning algorithms such as k-means clustering will be used to find similarity, differences and patterns that exist among different neighbourhoods.

3 Methodology

In this section, all the tasks were that performed for achieving the desired objectives are discussed.

3.1 Data Preparation & Exploratory Data Analysis

The data was imported from the sources previously mentioned in the data section. The raw data was pre-processed before the data analysis stage.

- Data Pre-Processing for List of Neighbourhoods
  - removed neighbourhoods assigned no borough
  - grouped neighbourhoods having same postal code
  - accommodated boroughs with missing neighbourhoods
  - Imported location data and merged it with neighbourhoods dataframe

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

- Data Pre-Processing for demographics of Toronto neighbourhoods
  - Imported data from Wikipedia
  - matched neighbourhoods with previous data
  - dropped rows whose match was not found and filled missing values

	Neighbourhood	Population	Land area (km2)	Density (people/km2)	Average Income
0	Agincourt	44577	12.45	3580	25750
1	Agincourt North, L'Amoreaux East, Milliken, St...	96830	18.86	5177	26092
2	Albion Gardens, Beaumond Heights, Humbergate, ...	16790	3.97	4229	28955
3	Alderwood, Long Branch	21281	7.16	3348	36263
4	Bathurst Manor, Downsview North, Wilson Heights	65290	24.96	2924	32966

- In the next step, nearby venues were obtained for each neighborhood using foursquare. The procedure used in week 3 assignment was followed for this purpose.

	PostalCode	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M1B	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	M1C	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	M1C	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
3	M1E	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place
4	M1E	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

- Exploring neighbourhoods and preparing Venues Data
    - Venues were grouped by neighbourhoods.
    - Venue count for each neighborhood was calculated
    - Most popular venues were determined for each neighborhood
    - One hot encoding was done for venue category
    - The venues data was merged into data frame.

	Neighbourhood	Venue Count	Top 3 Venues	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	Acces
0	Adelaide, King, Richmond	100	Coffee Shop, Café, Bar	Coffee Shop	Café	Bar	Steakhouse	Thai Restaurant	Hotel	Sushi Restaurant	Bakery	American Restaurant	Asian Restaurant	
1	Agincourt	4	Sandwich Place, Lounge, Breakfast Spot	Sandwich Place	Lounge	Breakfast Spot	Clothing Store	Drugstore	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	
2	Agincourt North, L'Amoreaux East, Milliken, St...	2	Playground, Park, Drugstore	Playground	Park	Drugstore	Diner	Discount Store	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	
3	Albion Gardens, Beaumond Heights, Humbergate, ...	11	Grocery Store, Pharmacy, Coffee Shop	Grocery Store	Pharmacy	Coffee Shop	Sandwich Place	Discount Store	Fried Chicken Joint	Beer Store	Fast Food Restaurant	Pizza Place	Japanese Restaurant	
4	Alderwood, Long Branch	9	Pizza Place, Pharmacy, Gym	Pizza Place	Pharmacy	Gym	Sandwich Place	Coffee Shop	Athletics & Sports	Skating Rink	Pub	Dog Run	Dim Sum Restaurant	

Now we have enough data (i-e demographics and venues information) for each neighborhood. Each neighborhood has a unique name and its corresponding features can be accessed using this unique name.

3.2 Segmenting and Clustering Neighbourhoods

3.2.1 Clustering with-respect-to Venues

- Used k-means for clustering neighbourhoods by venues using one-hot encoding data
- Estimated Optimal Number of Clusters. The optimal number was in this case was 3.
- train final k-means model using 3 clusters
- analyzed the cluster centers and came up with a description for each cluster.
- Visualize the clusters by venue count
- Visualize the clustered neighborhoods on the map

3.2.2 Clustering with-respect-to Demographics

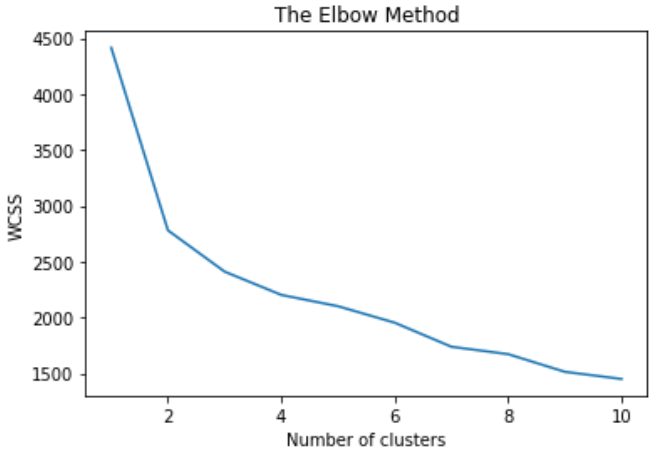
- Used k-means for clustering neighbourhoods by demographics
- Estimated Optimal Number of Clusters. The optimal number was in this case was 5.
- train final k-means model using 5 clusters

- analyzed the clusters and came up with a description for each cluster.
- Visualize the clusters by venue count
- Visualize the clustered neighborhoods on the map

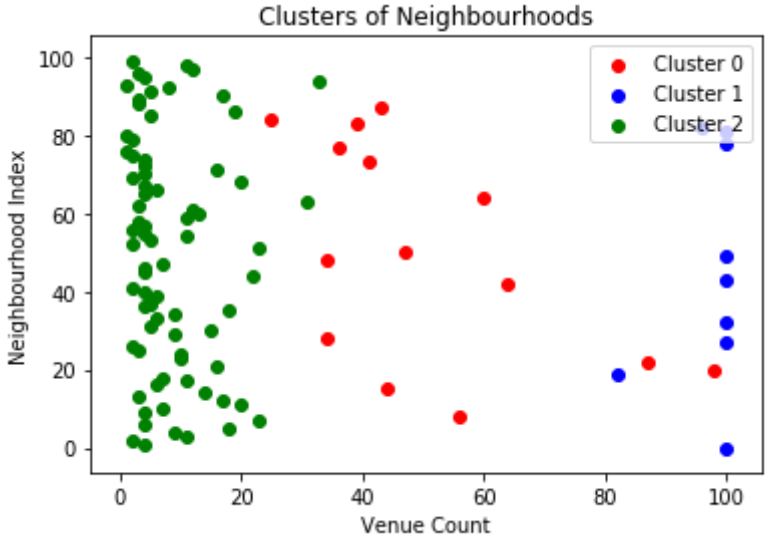
4 Results

4.1 Results for Clustering by Venues

- Optimal number of clusters was 3.



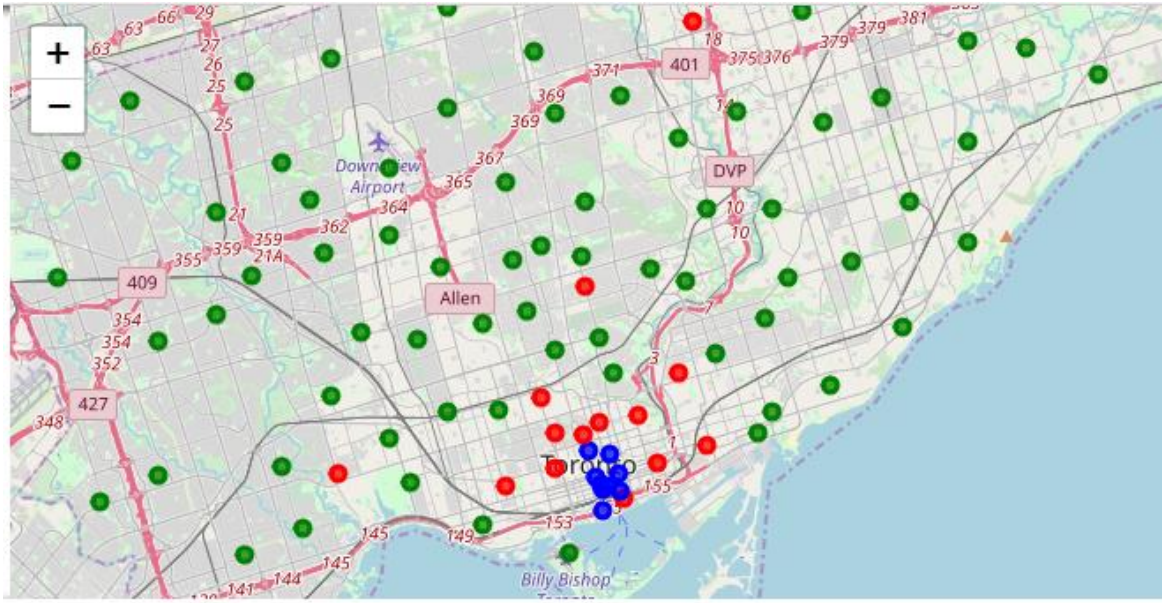
- On visualizing the clusters by venue count, it seemed like clusters are differentiated by venue counts.



- Cluster centers explain the intuitive description for each cluster.

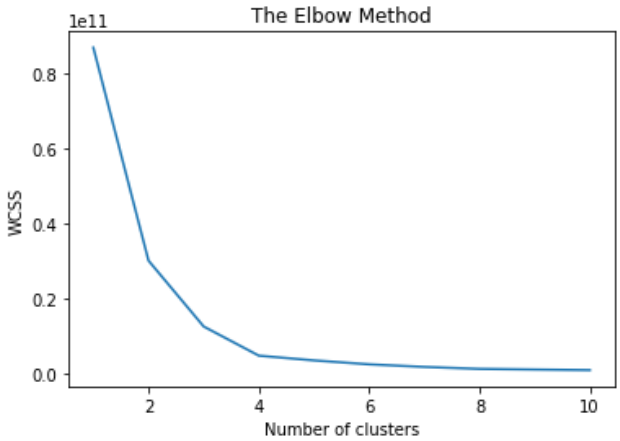
	Cluster Label	Neighbourhood Count	Venue Count Mean	Description
0	0	14	50.571429	Mid Venue Count
1	1	9	97.555556	High Venue Count
2	2	77	8.389610	Low Venue Count

- Visualize the clustered neighborhoods on the map

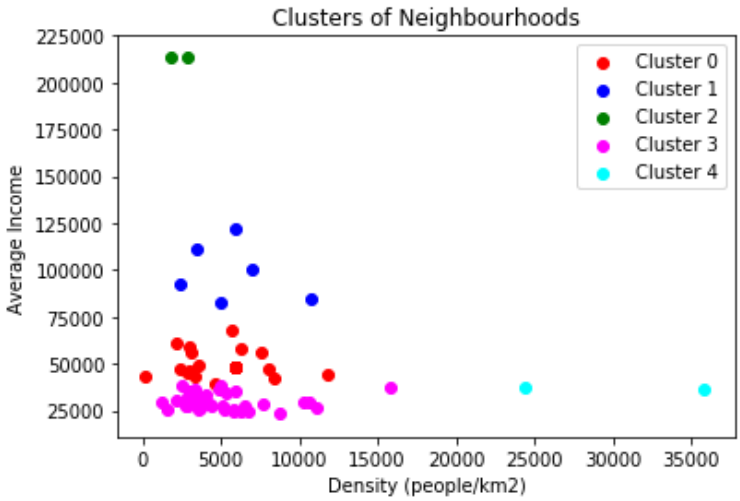


4.2 Results for Clustering by Demographics

- Optimal number of clusters was 5.



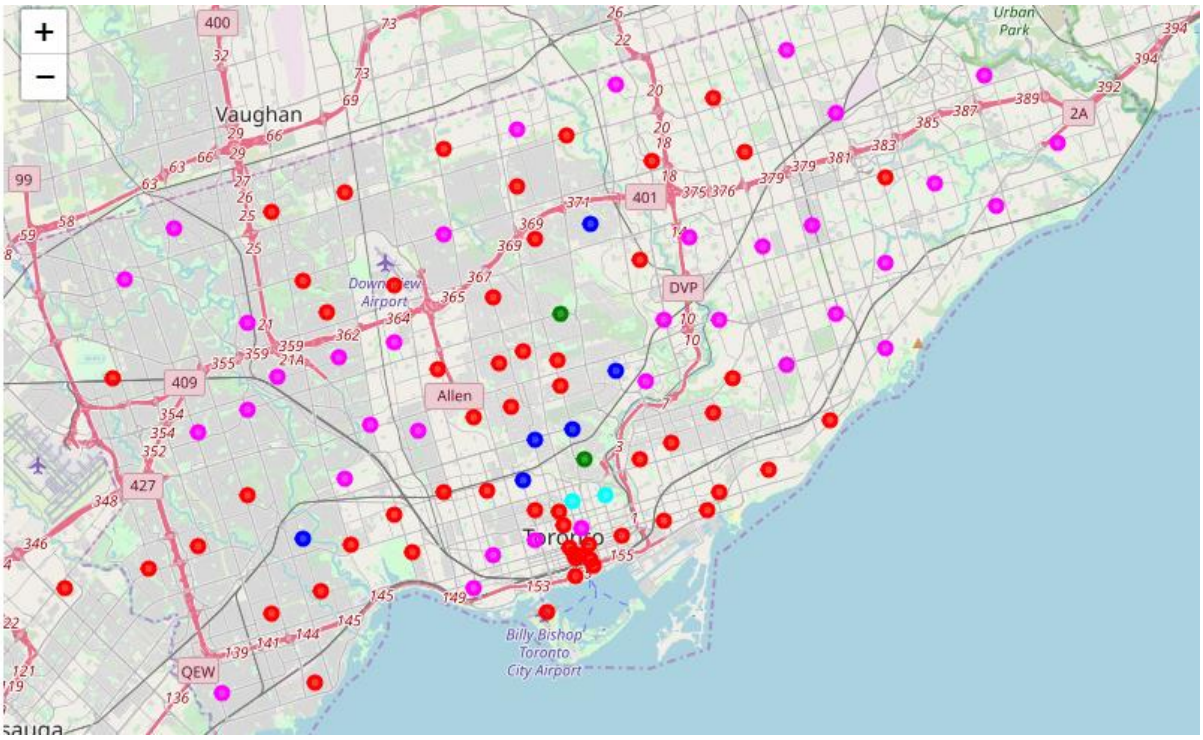
- Scatter plot was used for visualizing the clusters.



- Cluster centers gave an intuitive explanation for each cluster.

	Cluster Label	Density (people/km2)	Average Income	Description
0	0	5566.310345	48996.275862	Mid Dense Mid Income
1	1	5717.666667	98826.666667	Mid Dense High Income
2	2	2324.500000	214025.500000	Low Dense Very High Income
3	3	5234.571429	30101.542857	Mid Dense Low Income
4	4	30106.000000	37011.000000	High Dense Low Income

- Visualize the clustered neighborhoods on the map



5 Discussion

The data from different sources helped us in understanding the neighbourhoods of Toronto. Using machine learning clustering algorithms, we were able to find optimal number of clusters and segment the neighbourhoods based on demographics and the venues respectively. This shows that there exists similarities and differences among different neighbourhoods. So we achieved our objectives of comparing and clustering the neighbourhoods.

From the results section, it can be seen that neighbourhoods can be divided into three clusters based on venues; 1) High Venue Count, 2) Mid Venue Count, and 3) Low Venue Count. Moreover, we also obtained top 3 venues for each neighborhood. This information will be extremely useful for someone who is relocating and wants to compare different neighbourhoods for residence. High or mid venue count would be preference in this case.

Furthermore, he can narrow down his selection using the info about top 3 venues based on his taste and interest.

Similarly, we also observed neighbourhoods can be divided 5 clusters based on the demographics; 1) Mid Dense Low Income, 2) Mid Dense Mid Income, 3) Mid Dense High Income, 4) Low Dense Very High Income, 5) High Dense Low Income.

This information is extremely useful for entrepreneurs, businessmen, investors and marketing groups. For example, a restaurant owner can look at the demographics and already present venues to decide a location for his new restaurant. Marketing groups can use this information to design and plan marketing campaigns for products. Some of the products will be relevant to one cluster of people and other products will be relevant to other clusters.

It was also observed that city center is more crowded and venue count is higher. As we go farther from the city center, the population density and the venue count decrease as well. This was an expected result which validates the authenticity of our results.

6 Conclusion

Using the data science methodology and the relevant tools, we were able to make sense of the open source data and answered some important questions. We were able to analyze neighbourhoods and discover groups of neighborhoods with similar characteristics. The information is extremely useful for the targeted audience. We can refine this approach by considering other data sources as well and will be able to answer more relevant questions.