Capstone Project
The Battle of Neighborhoods

Voyo Tenev
March 24, 2020

1. Introduction

This project is about helping several stakeholders decide where it is the most optimal place to open a highly sophisticated gym and fitness center in Toronto. Given that Toronto is the most populous city in Canada and one of the highest ranked cities in the world on health and high quality of living, opening a gym there is an easy decision for an investment group. With that said we first need to explore the current gym market in the city and decide where it will be the smartest location to open a gym to maximize the profits.

The investors are interested in neighborhoods that meet the following criteria:

- The neighborhood should have average or above average population
- Since it is a 24 hours gym a higher percentage of people younger than 45 is preferred
- Since it is a highly sophisticated gym, the membership will be higher so the household incomes in that area should be average or above average.

With the data gathered by the explorations for these criteria, the goal is to find and recommend the optimal areas to the investors to open their gym. They will use this data to find places to buy or rent for their business.

Also, this information can be shared with other investors that are looking to open a new gym or a recreational center.

2. Data

The data for the needs of this project will come from the following sources:

- City of Toronto Neighborhoods: https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050
- City of Toronto Neighborhoods demographics data: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
- Foursquare API to collect information on other gyms and competitors in Toronto

After cleaning and preparing the data this is how our final data looked

	index	Borough	Neighborhood	Latitude	Longitude	Population	Population Density	Average Income	Number of Gyms
0	2	North York	Parkwoods	43.753259	-79.329656	26533.0	5349.0	34811.0	0.0
1	3	North York	Victoria Village	43.725882	-79.315572	17047.0	3612.0	29657.0	0.0
3	5	North York	Lawrence Heights	43.718518	-79.464763	3769.0	1178.0	29867.0	0.0
4	6	North York	Lawrence Manor	43.718518	-79.464763	13750.0	6425.0	36361.0	0.0
7	10	Scarborough	Rouge	43.806686	-79.194353	22724.0	791.0	29230.0	0.0
8	11	Scarborough	Malvern	43.806686	-79.194353	44324.0	5003.0	25677.0	0.0
13	17	Downtown Toronto	Garden District	43.657162	-79.378937	8240.0	15846.0	37614.0	2.0
18	24	Etobicoke	Princess Gardens	43.650943	-79.554724	9288.0	2249.0	80607.0	0.0
19	25	Etobicoke	West Deane Park	43.650943	-79.554724	4395.0	2063.0	41582.0	0.0
20	26	Scarborough	Highland Creek	43.784535	-79.160497	12853.0	2505.0	33640.0	0.0
21	27	Scarborough	Rouge Hill	43.784535	-79.160497	11167.0	2878.0	32858.0	0.0
22	28	Scarborough	Port Union	43.784535	-79.160497	12450.0	2310.0	48117.0	0.0
23	30	North York	Flemingdon Park	43.725900	-79.340923	21287.0	8760.0	23471.0	2.0
26	33	Downtown Toronto	St. James Town	43.651494	-79.375418	14666.0	63765.0	22341.0	1.0
29	38	Etobicoke	Eringate	43.643515	-79.577201	8008.0	3282.0	34789.0	0.0
30	39	Etobicoke	Markland Wood	43.643515	-79.577201	10240.0	3507.0	51695.0	0.0
32	41	Scarborough	Guildwood	43.763573	-79.188711	12820.0	2688.0	40806.0	0.0
33	42	Scarborough	Morningside	43.763573	-79.188711	11472.0	4112.0	27139.0	0.0
34	43	Scarborough	West Hill	43.763573	-79.188711	25632.0	2676.0	27936.0	0.0
35	46	East Toronto	The Beaches	43.676357	-79.293031	20416.0	5719.0	67536.0	0.0

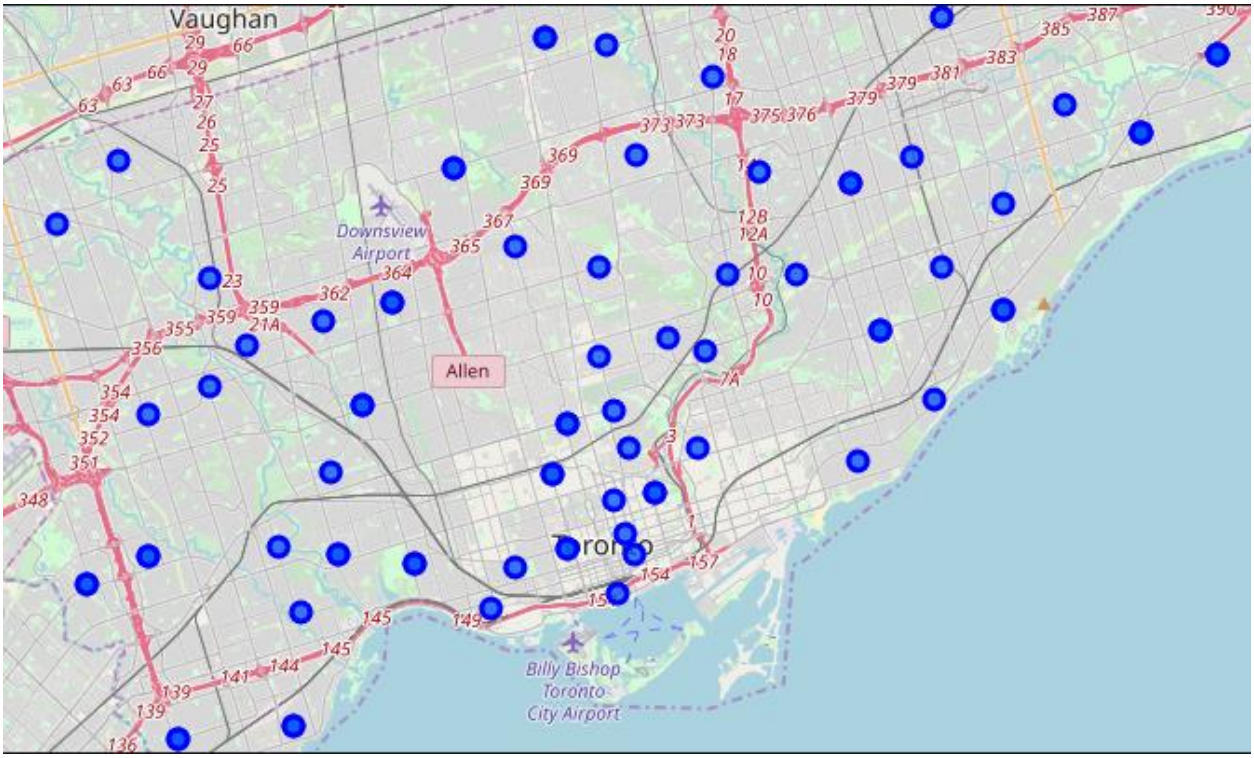
1. Part of the final dataframe

3. Methodology and Analysis

After cleaning and preparing the data we will firstly apply some basic visually exploratory techniques to the data. Then replace some of the data with normalized data. And in the end perform cluster analysis of the data to find the best cluster of neighborhoods to open our gym.

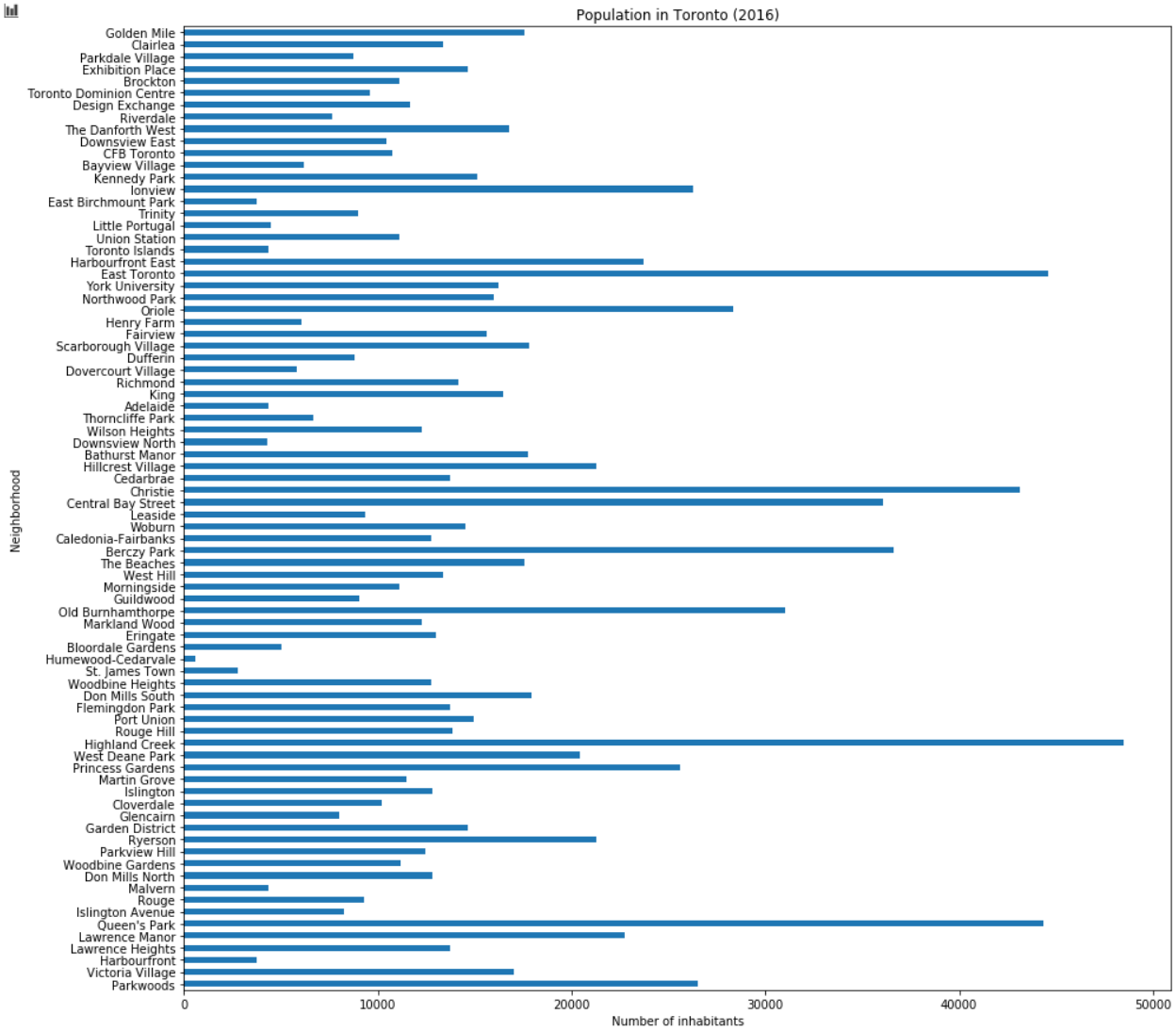
3.1 Exploratory Data Analysis

First, we create a map with circle markers for each neighborhood, in order to visualize the location of the neighborhoods.

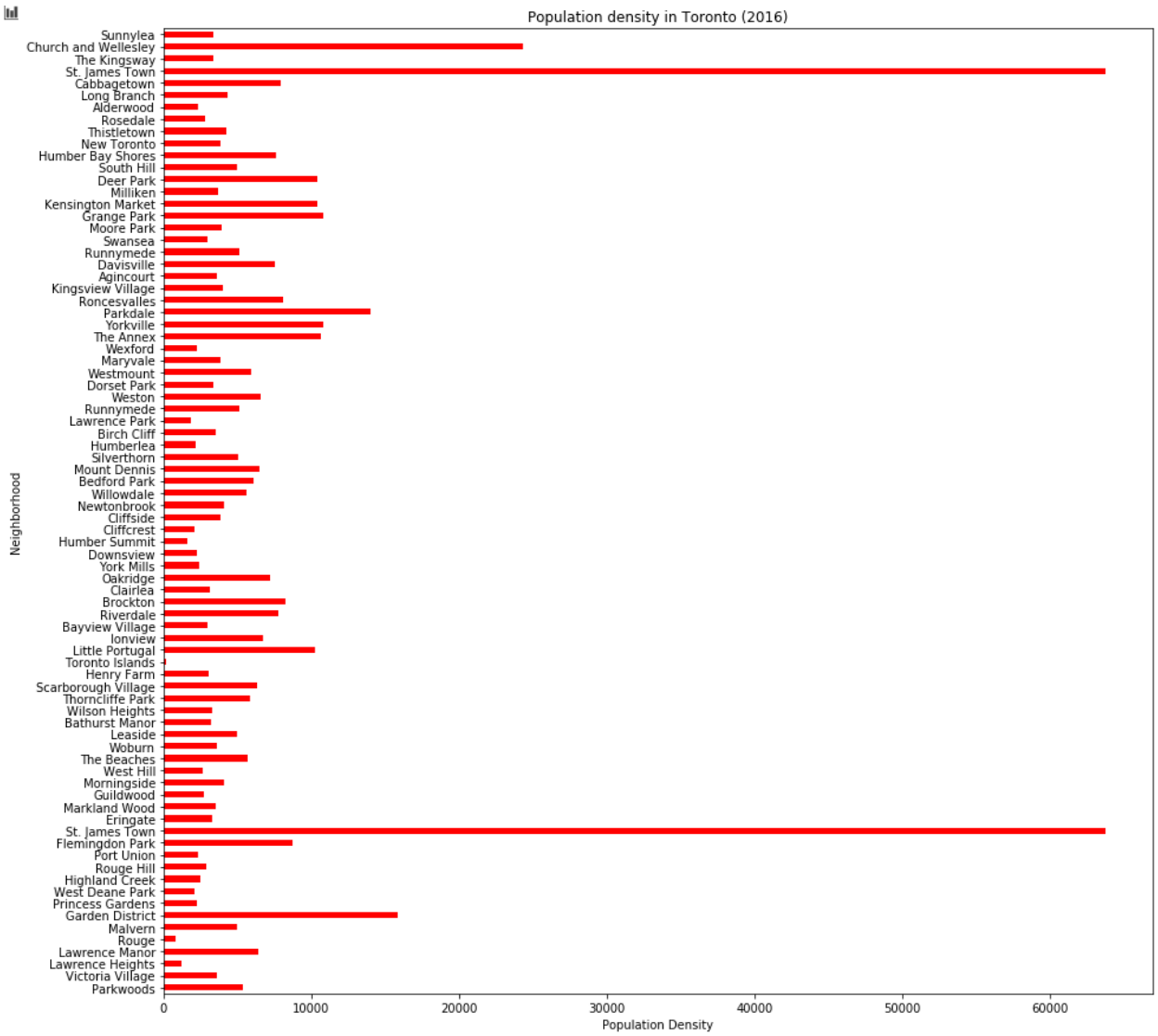


2. Map of locations of each neighborhood

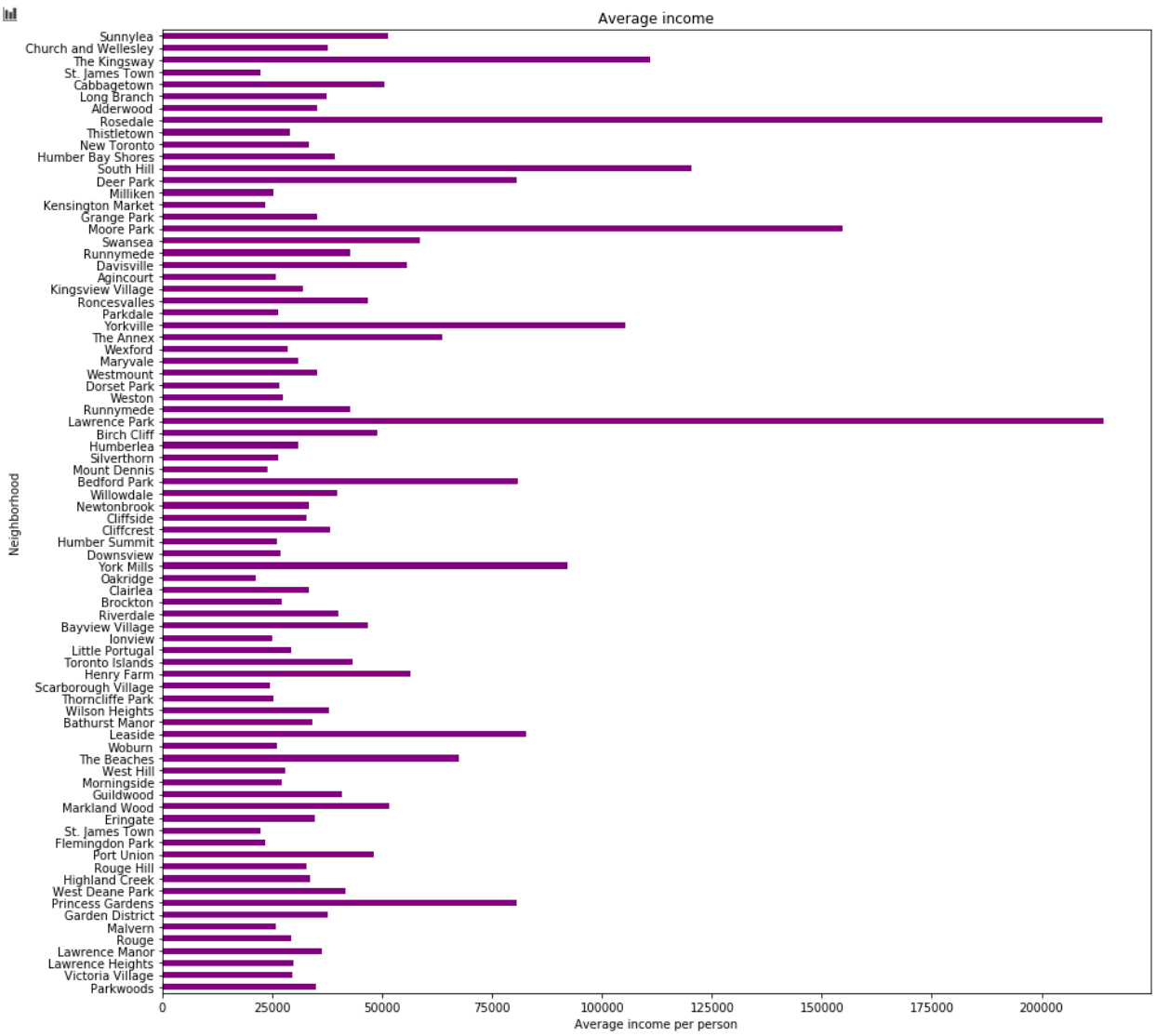
Second, we visualize the other important features that influence our decisions.



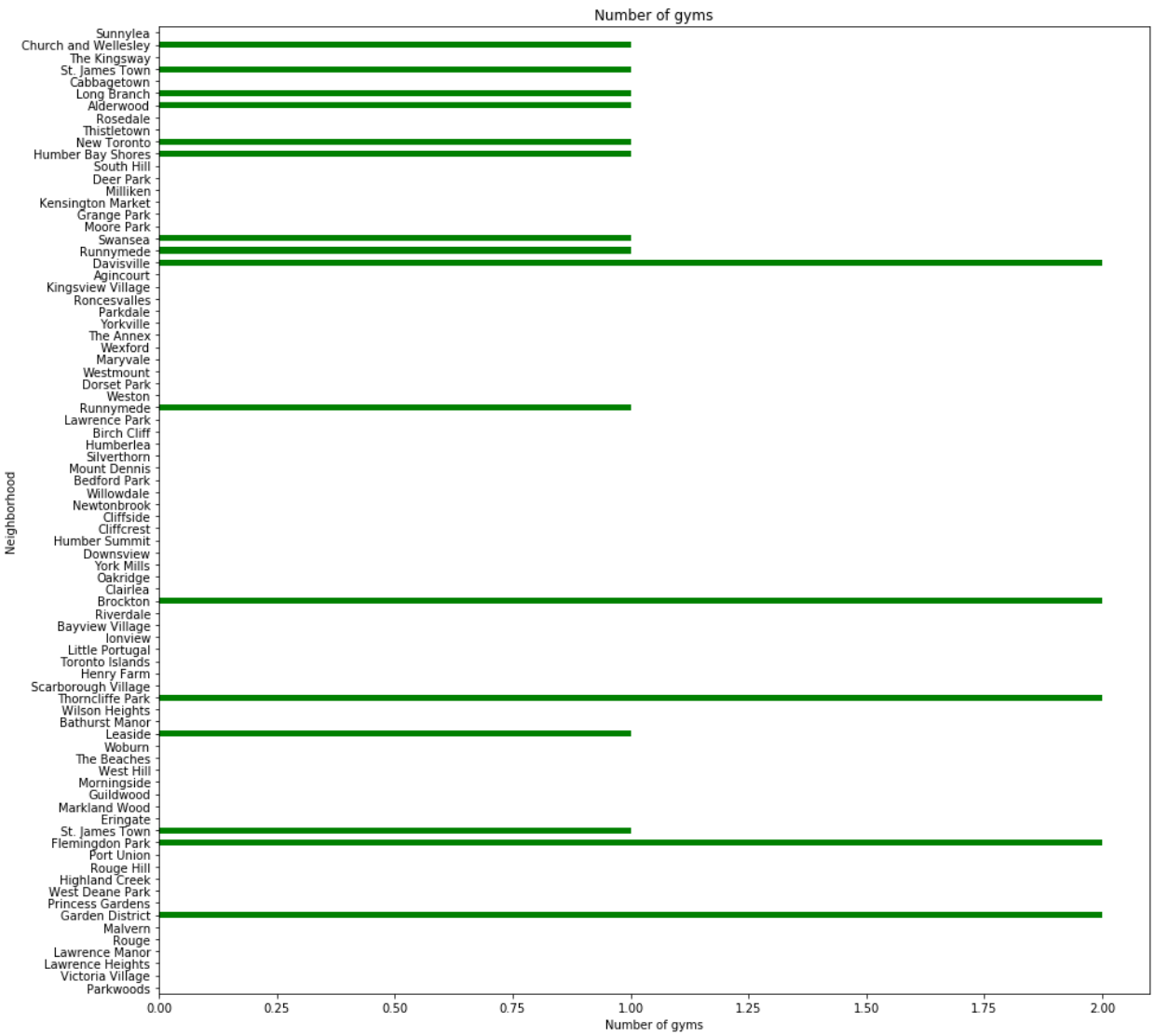
3. Number of inhabitants in each neighborhood graph



4. Population density per neighborhood graph



5. Average income per neighborhood graph



6. Number of gyms in each neighborhood graph

3.2 Cluster Analysis

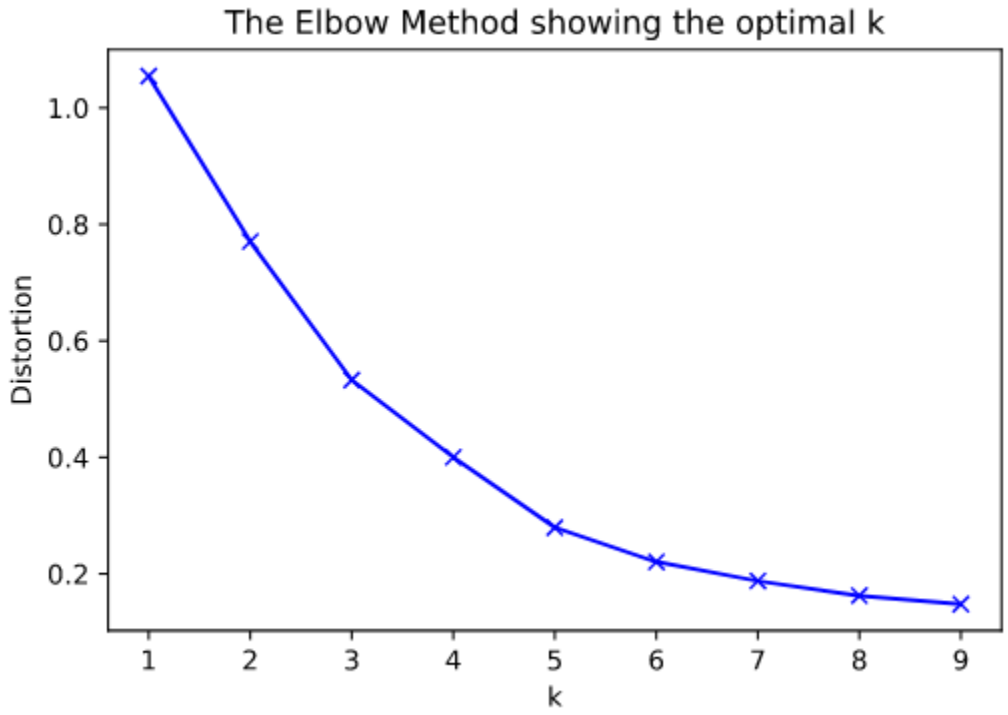
In order to identify clusters of similar neighborhoods the unsupervised learning method the K-means algorithm was applied.

Preceding that the unnecessary columns Borough, Neighborhood, Latitude, Longitude and Population Density were removed, and the columns Number of Gyms and Population were merged into one column representing the number of gyms per 1000 people in each neighborhood.

	Average Income	Number of gyms per 1000 people
0	34811.0	0.000000
1	29657.0	0.000000
3	29867.0	0.000000
4	36361.0	0.000000
7	29230.0	0.000000
8	25677.0	0.000000
13	37614.0	0.242718
18	80607.0	0.000000
19	41582.0	0.000000
20	33640.0	0.000000

7. The resulting dataframe

To identify the optimal number of clusters the Elbow method was used

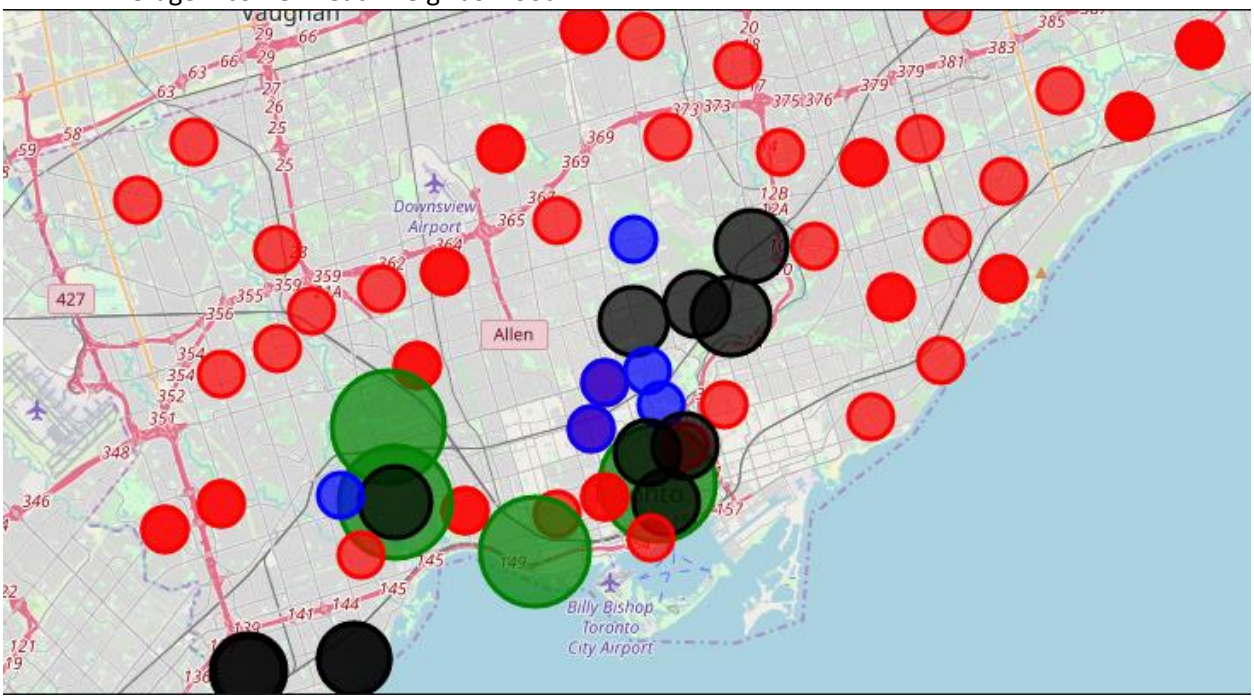


8. The resulting Elbow Method graph

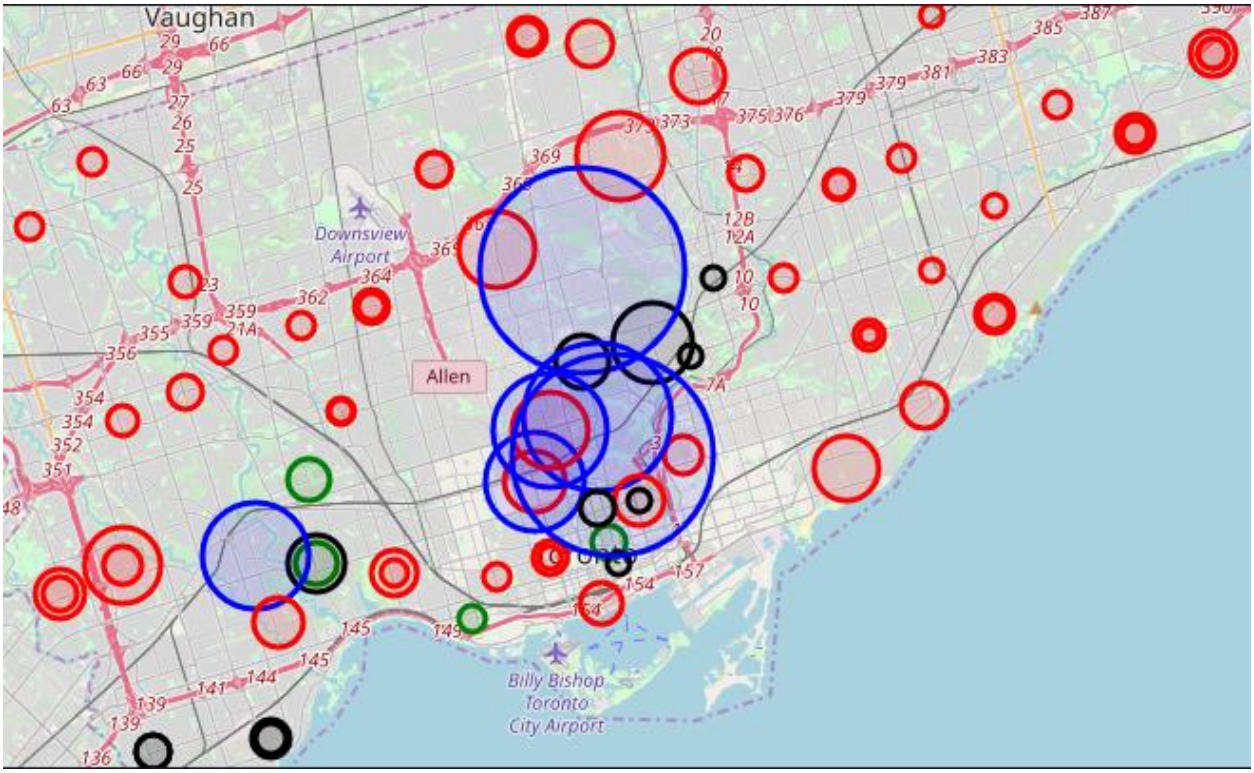
Four number of clusters were chosen based on the results from the Elbow method.

After that two maps were created each representing:

- Number of gyms in each neighborhood
- Average income in each neighborhood

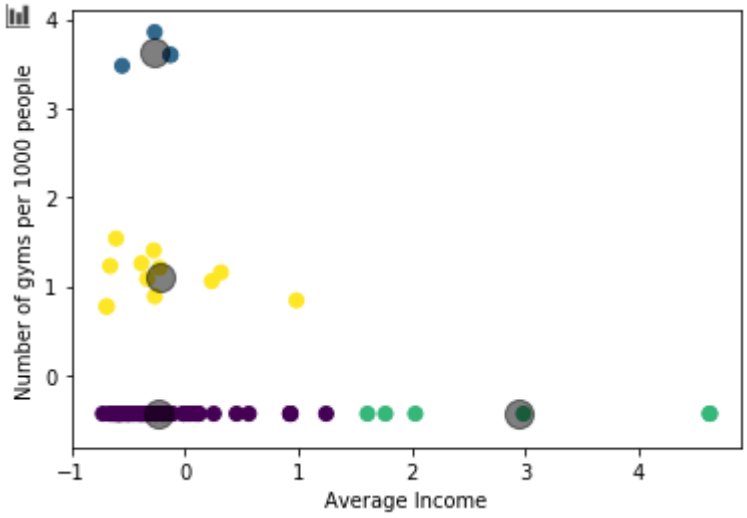


9. Number of gyms per 1000 people in each neighborhood



10. Average Income in each neighborhood

After that a scatter plot was created to represent the clusters.



11. Scatter plot representing the clusters of neighborhoods

From the scatter plot above we can see there is one cluster (the cluster in green) with the highest average income while having 0 gyms.

4. Results

From the four clusters that were defined one came out as an outsider. From the above data we can see that Cluster 2 has the 2nd lowest competition while having the highest average income.

From this data we can only conclude that the neighborhoods in the 2nd Cluster are the most ideal, however in real life more data such as rent price or price for property price for the sports center may play an even bigger role.

5. Conclusion

In conclusion, this simple data analysis was performed in order to find the the most optimal neighborhoods in Toronto, Canada for opening a new highly sophisticated gym and fitness center. In the analysis multiple statistical tools and methods were used in order to explore and visualise the data. Clustering was used to identify the most optimal neighborhoods in the city.

In the end, it is obvious that the neighborhoods in the 2nd cluster: Lawrence Park, Yorkville, Moore Park, South Hill, Rosedale and The Kingsway would be ideal for opening a highly sophisticated and expensive gym and fitness center.