Clustering Toronto neighborhoods for people with children

Agnė Pučilauskaitė

2021 08 07

1. Introduction

1.1 Background

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016 it is the most populous city in Canada and the fourth most populous city in North America. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Even life in Toronto sounds exciting but for adults with children it could be really challenging. Traffic, long distances between places is not comfortable for all people but for people with childrens it creates more difficulties.Therefore for people with children is really important if neighborhood will be comfortable for life with kids, they prefer that in their chosen neighborhood would be more schools, parks and playgrounds, easy access to hospitals and also that grocery stores would be near home.

1.2 Problem

Data that might contribute to determining which Toronto neighborhood is more suitable for life with childrens includes schools amount, parks and playground amount, hospitals amount and grocery store amount around neighborhoods. This project aims to cluster Toronto neighborhoods in clusters which could help people with children to choose neighborhoods for living.

1.3 Interest

People with childrens who is thinking to start life in Toronto or people with childrens which want to change Toronto neighborhood in similar Toronto neighborhood or to change Toronto neighborhood to different Toronto neighborhood to improve they life.

2. Data acquisition and cleaning

2.1 Data sources

Data of the Toronto neighborhood was used from Foursquare API. To get data from Foursquare several parameters should be described.

1. Latitude and Longitude. First of all, to get Toronto neighborhoods Wikipedia page [https://en.wikipedia.org/ wiki/List_of_postal_codes_of_Canada:_](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_) was scraped. Then Latitude and Longitude were added from 'GeoSpatial Dataset.csv' which was provided in IBM Applied Data Science Capstone course.
2. Date of data or version. In this study it was used 2021 - 07 - 21.
3. Search query. It was performed 5 Foursquare API calls, where search parameters were: 1 - school, 2 - park, 3 - playground, 4 - hospital, 5 - grocery stores.
4. Radius. A radius limitation was added: for schools, hospitals and grocery stores it was searched 2 km around the neighborhood latitude and longitude and for parks and playgrounds - 3 km.

2.2 Data cleaning and Feature selection

Data from Foursquare were combined in tables. In tables were a lot of not relevant information because Foursquare returns all venues in which name or category it found a search query word. For example: in received Park data it was included categories of areas like bus stops, parking, etc. Therefore data cleaning was performed:

- For schools data it was left just data which category is equal to 'School', 'Elementary School', 'High School', 'Daycare', 'Middle School'. Because clustering will use just one feature of schools, schools data categories were renamed to 'schools'.
- For Parks and Playground data it was left which category is equal to 'Park' or 'Playground'. Because clustering will use just one feature of parks/playgrounds, Parks and Playground data categories were renamed to 'parks/playgrounds'.
- For Hospital data it was left data which category is equal to 'Hospital', 'Medical Center', 'Doctor's Office', 'Emergency Room'. Because clustering will use just one feature of hospitals, Hospitals data categories were renamed to 'hospitals'.
- For Grocery store data it was left data which category is equal to 'Convenience Store', 'Miscellaneous Shop', 'Grocery Store', 'Supermarket', 'Food & Drink Shop', 'Market'. Because clustering will use just one feature of grocery stores, Grocery store data different categories were renamed to 'grocery store'.

After cleaning all different category data was combined in one table. Because some neighborhoods do not have parks/playgrounds or hospitals near them in the table was found several NaN values, these NaN values were filled in with 0.

3.0 Exploratory Data Analysis

To understand more about data exploratory data analysis was conducted.

In Figure 1 is shown the distribution of total schools per neighborhood. Data of schools is not normally distributed. Most neighborhoods have between 13 and 19 schools. Also in Figure 5 we could see that the minimum schools amount per neighborhood is 2 and maximum 23, average amount schools in Toronto per neighborhood is 13.3.

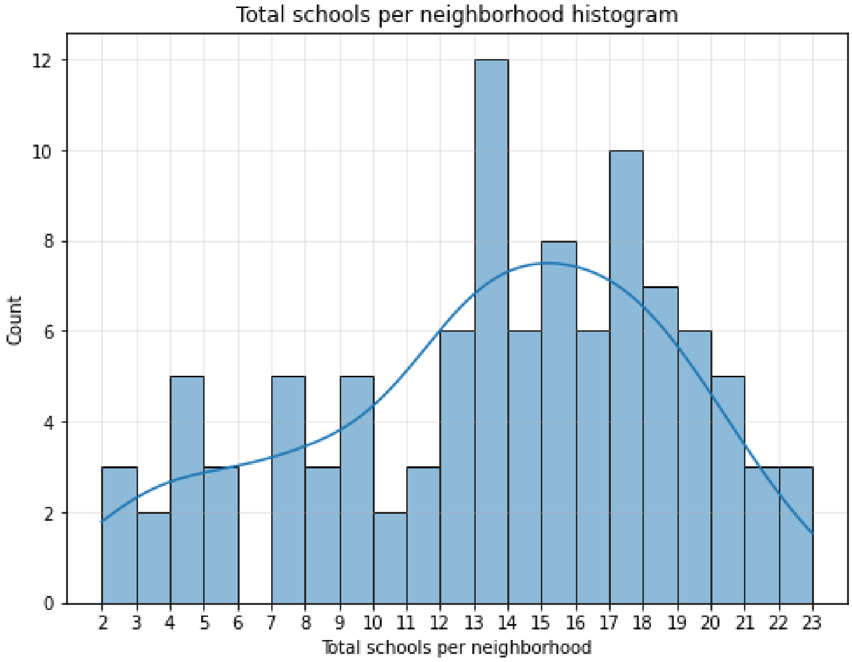

Figure 1 . Histogram of total schools amount per neighborhoods

In Figure 2 is shown the distribution of total parks/ playgrounds per neighborhood. Data of parks and playgrounds is not normally distributed, we could see two peaks: first peak - neighborhoods which do not have or have small amounts of parks and playgrounds around them, second peak - parks and playground between 16 and 20. Average amount of parks/playgrounds in Toronto per neighborhood is 9.9 (Figure 5).

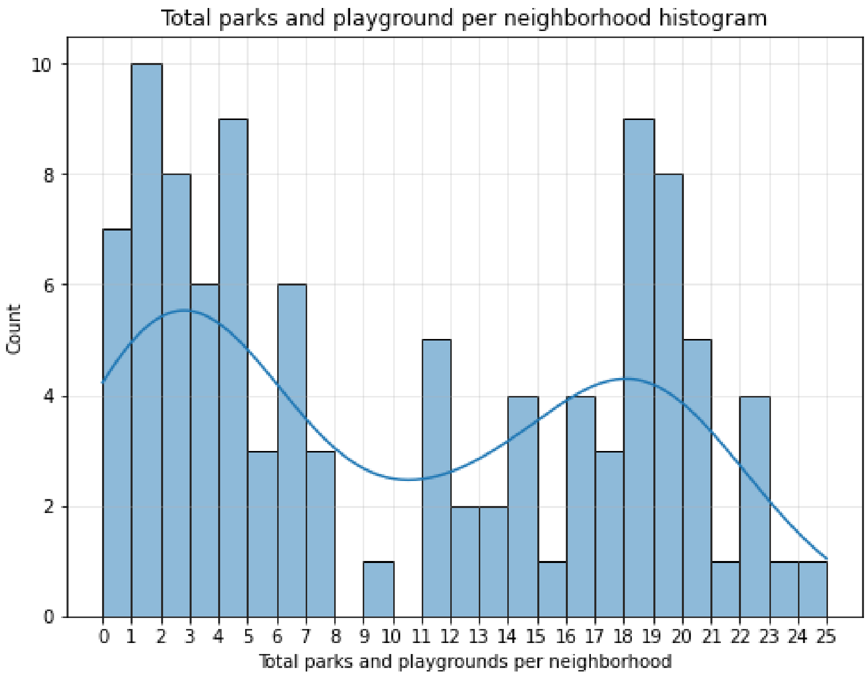

Figure 2 . Histogram of total parks and playgrounds amount per neighborhoods

In Figure 3 is shown the distribution of total hospitals per neighborhood. Data of hospitals is not normally distributed - most neighborhoods do not have or have just 1 hospital around them. In Figure 5 we could see that in data has outliers, these neighborhoods have a high amount of hospitals in their area. Average amount of hospitals in Toronto per neighborhood is 5.7 (Figure 5).

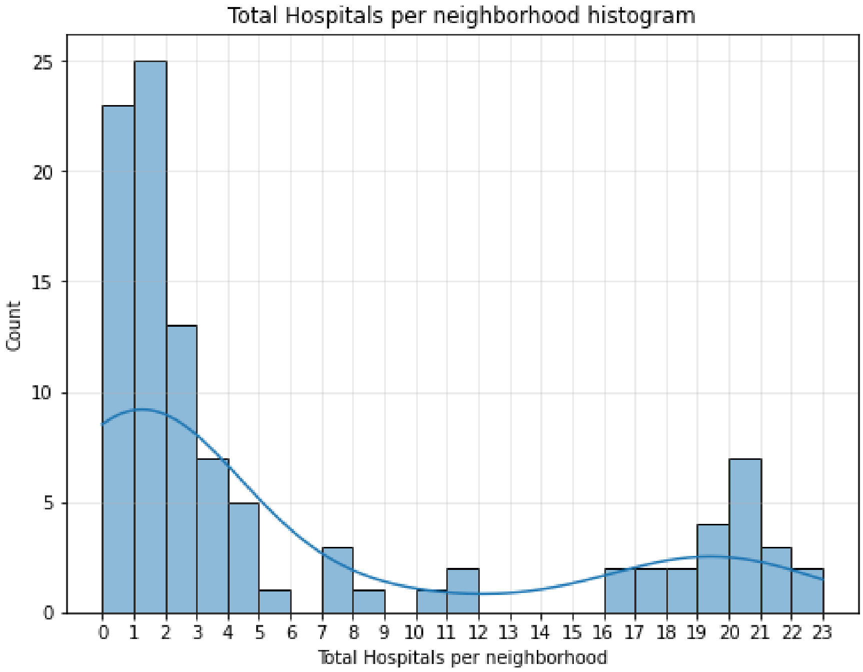

Figure 3 . Histogram of total hospitals amount per neighborhoods

In Figure 4 is shown the distribution of total Groceries stores per neighborhood. Most neighborhoods have between 4 and 8 grocery stores around their area. According to Figure 5 boxplot, data of groceries stores has outliers: these neighborhoods have higher amounts of grocery stores around them. Average amount of hospitals in Toronto per neighborhood is 5.6 (Figure 5).

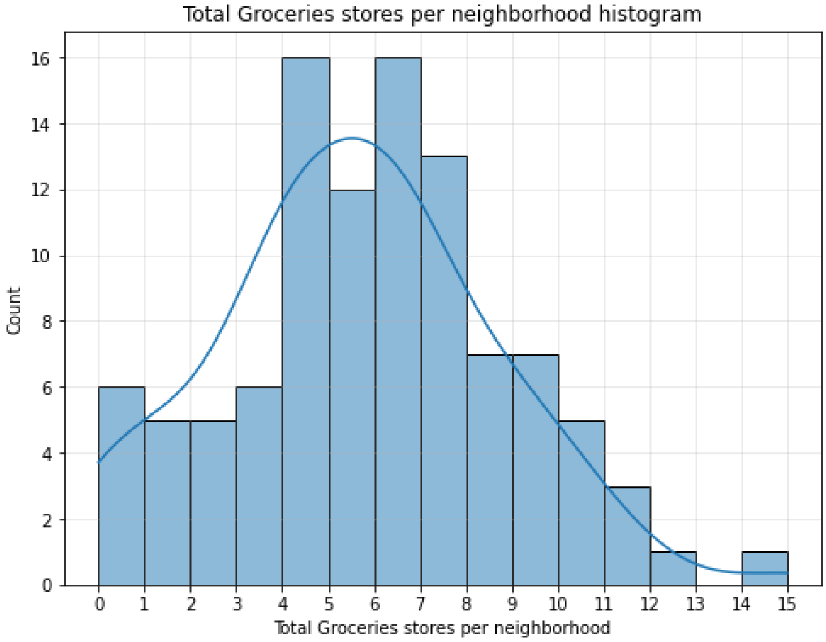

Figure 4 . Histogram of total groceries stores amount per neighborhoods

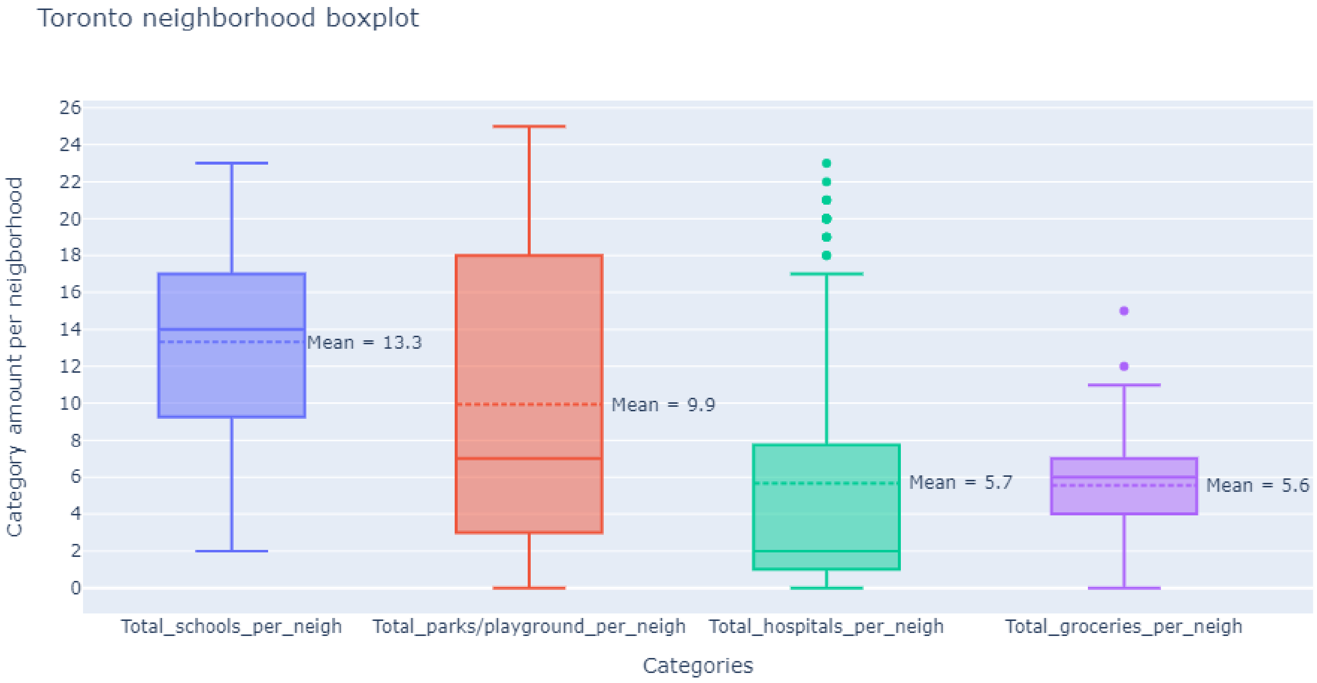

Figure 5. Toronto neighborhood boxplot

4. Predictive model

In these studies it was used k-means clustering method to group Toronto neighborhoods according to their similarities.

First task was to find the best cluster amount to cluster Toronto data. It was calculated the Inertia value of k-means clustering when it was used from 1 to 10 clusters. In Figure 6 is shown how the Inertia value decreases as cluster amount increases. The higher Inertia value changes are from cluster amount 1 to 3. From cluster amount 4 Inertia value starts to decrease slower. Even higher amount of clusters have lower Inertia values, It is not practical to have many clusters. It will be more difficult for people to interpret results and to see differences between clusters. It was decided to analyse results more when Toronto data is clustered in 4, 5 and 6 clusters.

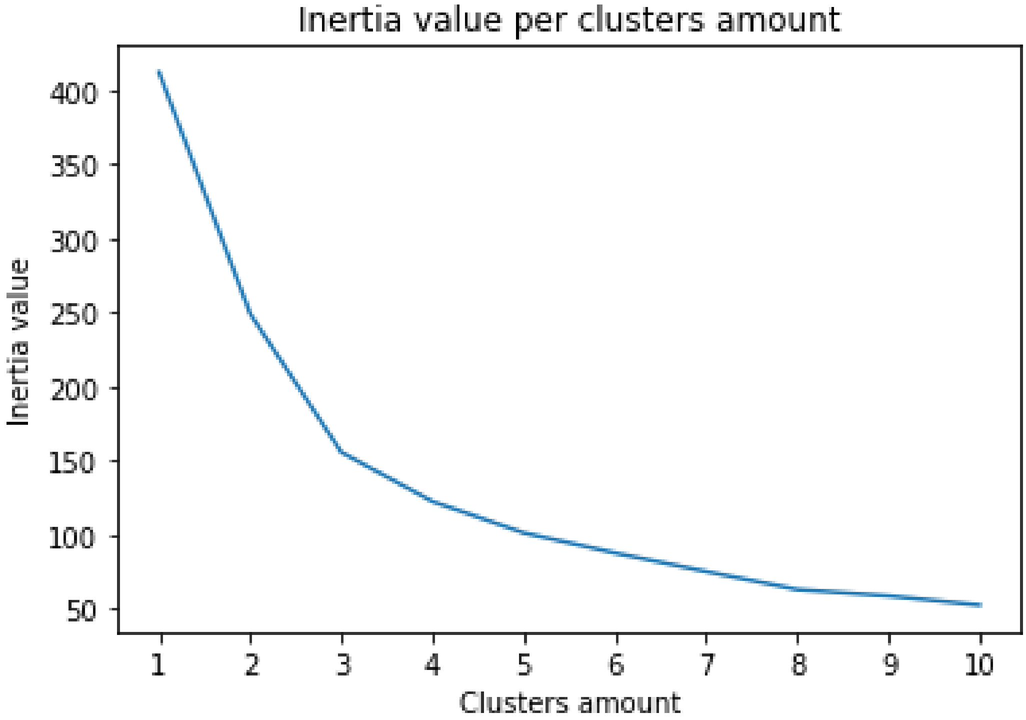

Figure 6. Inertia value per cluster amount

To interpret cluster was used this approach:

- If the difference between cluster feature average and all Toronto data feature average is not lower or higher than 10 %, it means that in the cluster are neighborhoods which features are similar to Toronto data features average and feature in cluster is described as AVERAGE.
- If the difference between cluster feature average and all Toronto data feature average is lower than 10 %, it means that in clusters are neighborhoods which have lower than

Toronto data features average values and features in clusters are described as LOWER.

- If the difference between cluster feature average and all Toronto data feature average is higher than 10 %, it means that in clusters are neighborhoods which have higher than Toronto data features average values and features in clusters are described as HIGHER.

In table 1 it shows how Toronto neighborhoods are grouped if they are clustered in 4 clusters. A model with 4 clusters has high feature variation per cluster, for example, in cluster label 0 is included Toronto neighborhoods which have 2 schools and Toronto neighborhoods which have 20 schools.

Table 1. Toronto Neighborhood data clustered in 4 cluster results

Cluster label	**Schools amount / Average amount per cluster**	**Parks/Playground amount / Average amount per cluster**	**Hospital amount / Average amount per cluster**	**Grocery stores amount / Average amount per cluster**
0	Lower / 11.8 ± 4.6	Lower / 2.5 ± 2.2	Lower / 2.1 ± 3.1	Lower / 2.4 ± 2.8
1	Lower / 7.8 ± 4.5	Higher / 17.6 ± 2.4	Higher / 19.4 ± 1.8	Higher / 6.3 ± 1.7
2	Higher / 16.2 ± 3.2	Higher / 18.4 ± 3.6	Lower / 2.5 ± 1.6	Higher / 8.9 ± 2.4
3	Higher / 16.7 ± 3.5	Lower / 5.3 ± 3.9	Lower / 1.4 ± 2.2	Average / 5.7 ± 1.6

In table 2 it shows how Toronto neighborhoods are grouped if they are clustered in 5 clusters. To compare a model with 5 clusters with a model with 4 clusters, clusters have a lower difference between minimum and maximum features values. Therefore we could say that a model with 5 clusters is more accurate than a model with 4 clusters.

Table 2. Toronto Neighborhood data clustered in 5 cluster results

Cluster label	**Schools amount / Average amount per cluster**	**Parks/Playground amount / Average amount per cluster**	**Hospital amount / Average amount per cluster**	**Grocery stores amount / Average amount per cluster**
0	Higher / 17.4 ± 3.1	Lower / 5.3 ± 3.8	Lower / 1.6 ± 2.4	Lower / 2.1 ± 1.5

1	Lower / 7.8 ± 4.5	Higher / 17.6 ± 2.4	Higher / 19.4 ± 1.8	Higher / 6.3 ± 1.7
2	Higher / 16.4 ± 3.2	Higher / 18.6 ± 3.5	Lower / 2.5 ± 1.7	Higher / 9.1 ± 2.4
3	Lower / 8.0 ± 3.1	Lower / 1.4 ± 1.5	Lower / 2.5 ± 3.8	Lower / 3.4 ± 1.7
4	Higher / 15.0 ± 2.9	Lower / 4.6 ± 3.8	Lower / 1.4 ± 2.2	Higher / 6.3 ± 1.2

In table 3 it shows how Toronto neighborhoods are grouped if they are clustered in 6 clusters. In a model with 6 clusters data started to be clustered more per 3 different groups - Higher, Lover, Average feature amount per neighborhood. This pattern was not seen in a model with 5 clusters. Model with 6 clusters is more accurate than models with 4 and 5 clusters, because variation per cluster between features is lower.

Table 3. Toronto Neighborhood data clustered in 6 cluster results

Cluster label	**Schools amount / Average per cluster**	**Parks/Playground amount / Average per cluster**	**Hospital amount / Average per cluster**	**Grocery stores amount / Average per cluster**
0	Higher / 17.7 ± 3.0	Lower / 7.1 ± 4.0	Lower / 1.1 ± 1.1	Higher / 5.8 ± 1.6
1	Average / 13.1 ± 2.1	Higher / 16.0 ± 3.4	Higher / 18.4 ± 1.8	Average / 6.1 ± 2.2
2	Lower / 4.7 ± 1.7	Higher / 18.4 ± 0.7	Higher / 19.9 ± 1.6	Average / 6.4 ± 1.4
3	Lower / 9.8 ± 3.5	Lower / 1.6 ± 1.4	Lower / 2.2 ± 3.6	Lower / 4.5 ± 1.8
4	Higher / 16.4 ± 3.2	Higher / 18.6 ± 3.5	Lower / 2.5 ± 1.7	Higher / 9.1 ± 2.4
5	Higher / 15.6 ± 2.8	Lower / 3.5 ± 2.2	Lower / 2.0 ± 2.7	Lower / 1.2 ± 1.1

Even a model with 6 clusters is more accurate than models with 4 and 5 clusters, it is more clear to distinguish neighborhoods if we could to interpret cluster data with LOWER or HIGHER descriptions .Therefore a model with 5 clusters is chosen to cluster Toronto data.

In Figure 7 is shown how clusters from Toronto neighborhoods model with 5 clusters spread through Toronto. We could see some patterns, in city center are neighborhoods which have label 1, in this area neighborhoods have higher amounts of parks/playgrounds, hospitals and grocery stores but lower schools amount, going from city center schools amounts starts to increase and neighborhoods with labels 0, 2 and 4 have higher amount of schools around them. Neighborhoods with label 3 will be mostly not attractive to families with childrens, in these neighborhoods there are lower amount of schools, parks/playgrounds, hospitals, grocery shops, these areas mostly are settled farther from city center than other clusters labels neighborhoods .

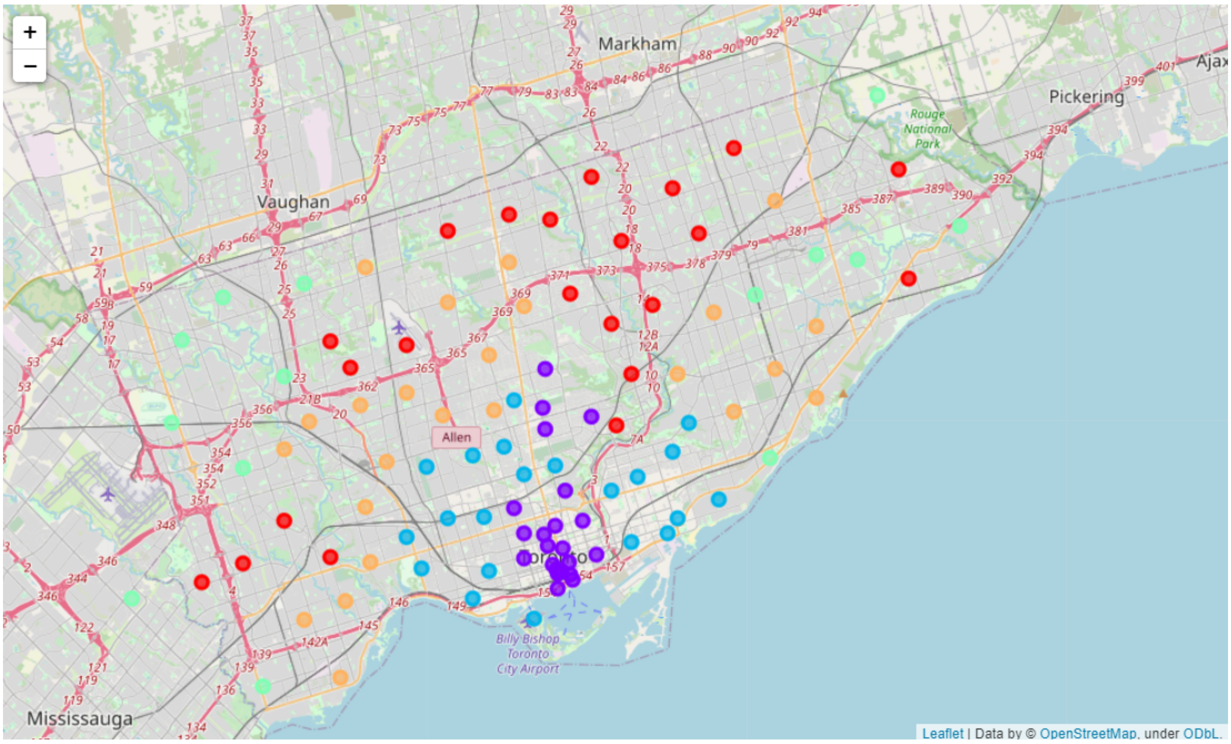
+
−
Markham
Vaughan
Pickering
Ajax
Rouge National Park
Allen
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 7. Toronto neighborhoods are clustered in 5 clusters: red - cluster label 0, purple - cluster label 1, blue - cluster label 2, green - cluster label 3, orange - cluster label 4.

5. Conclusion

In this study a k-mean clustering model was built for Toronto neighborhoods data. Toronto neighborhoods data was clustered in 5 clusters according to schools, parks/playgrounds, hospitals, grocery amounts per neighborhood. This study could be helpful to people with childrens who are thinking of starting life in Toronto or want to change their neighborhood to a similar neighborhood or more suitable for life with childrens.

6. Discussion

In this study Toronto neighborhoods were clustered according to 4 features: schools, park/playground, hospitals and grocery stores amount. To improve the model it could be added more features which are important for people with childrens.

Another improvement possible to get information which neighborhoods are still in the developing process and they feature characteristics will drastically change in time.