Toronto’s Neighborhoods Clustering based on Attractiveness of Living and Venue Types

Mirena Trifonova

07 May 2020

Abstract

World biggest capitals can be really intimidating to explore, be it for business, education, tourism, etc. Their diversity and scale makes any generalization challenging. And, yet clustering as unsupervised machine learning algorithm for grouping unlabeled data comes in handy. The current analysis makes use of K-means clustering for grouping Toronto’s neighborhoods on the basis of a) their attractiveness of living and b) venue types. Data are obtained from https://torontolife.com/neighbourhood-rankings/ and from https://foursquare.com/. The first source is especially useful with its list of Toronto Neighborhoods, their latitude and longitude and ranking based on 10 criteria: housing, safety, transit, shopping, health, entertainment, community, diversity, education and employment. For each criteria the neighborhoods get ranking score form 0 to 100. The clustering based on these 10 criteria reveals that the optimal split of Toronto’s neighborhoods is in 3 clusters (groups). The first group is the largest in terms of numbers of neighborhoods being grouped here and the worst rated. It encompasses almost half of Toronto’s 140 neighborhoods and is, namely, the most diverse and least secure one. The second cluster, which is the smallest (based on number of neighborhoods) has best community, housing conditions and is safest. The cluster that marks the highest average rating is the third. It scores best in 6 out of the 10 ranking criteria, namely – employment, education, entertainment, health, shopping, transit. The analysis concludes with another clustering based on neighborhoods venue categories obtained via Foursquare API.

Keywords: Toronto’s neighborhoods, K-means, clustering, neighborhoods ranking, attractiveness of living, venue categories.

Introduction

The projects targets people who want to weight the relative advantages of living in a particular area of Toronto. These could be future students looking for the most dynamic areas, or areas with better education. These could be people planning to move there or to send their children to learn, who might be interested in levels of education, safety, employment, etc. in the neighborhood, or future investors in Toronto.

Toronto is among the 5 biggest North American cities, divided into 140 neighborhoods and as such offers wide variety of opportunities, so wide that some might need a more summarized picture of it.

Data

The project looks at Toronto neighborhoods in order to cluster them into groups based on criteria related to attractiveness of living and venue types. It takes advantage of two sources of information:

1. https://torontolife.com/neighbourhood-rankings/#
2. https://foursquare.com/

The first source is especially useful with its list of Toronto Neighborhoods, their latitude and longitude and ranking based on 10 criteria: housing, safety, transit, shopping, health, entertainment, community, diversity, education and employment. For each criteria the neighborhoods get a ranking form 0 to 100. Interestingly enough, the names of the neighborhoods differ from other sources of information to another. This is due to the fact that different data sources might group together some neighborhoods, or split others into different areas, or even use old-fashioned names, which hinders comparability. Therefore, it is better when all data: neighborhoods and their latitude and longitude, come from the same source of information, as is here the case.

The project uses K-means clustering. Hence, another point worth consideration is that K-means clustering needs some distance metric, for example Euclidean distance, which means that K-means algorithm is not directly applicable for categorical variables because Euclidean distance function is not really meaningful for discrete variables. All relevant data from the website of Toronto Life is continuous, moreover, the ranking for each category lies between 0 and 100, which implies that we do not need to rescale or normalize the data.

Data have been scraped using Selenium. The respective script could be found in torontolife_scraper.ipynb file on this repo: Link to Repo. Selenium is convenient when using dynamic web elements, which contents change as JavaScript executes, as is the case with changing ratings.

The second source of information, Foursquare website, could be used for exploring venues and clustering neighborhoods based on their entertainment profile. Foursquare API will be used to explore each neighborhood and return maximum 100 venues within radius of 500 meters of its longitude and latitude. Changing the maximum number of venues and radius from the neighborhood location, within which the venues are located, would affect the results. For example, when exploring Toronto neighborhoods venues within a radius of 1000 meters and a limit of maximum 200 venues. There were more than 6000 venues but 1/3 were overlapping among neighborhoods. So, the radius were restricted to 500 meters and the maximum number of venues to 100 and thus, there are roughly 2500 number of unique venue locations out of around 2700 total venues.

Again, as mentioned earlier, K-means clustering uses some distance metric, for example Euclidean distance, which means that it is not directly applicable for categorical variables. Hence, venue categories will be converted to dummy variables with value of 1 if such venue is present in the neighborhood and 0 otherwise.

Methodology

K-means clustering is unsupervised learning algorithm that groups data based on similarities between data points. It is a type of partitioned-based clustering that divides the data points into 'k' non-overlapping clusters in which each observation belongs to the cluster with the nearest centroid. The main objective of K-Means clustering is to minimize the squared distance of between the observation and the centroid of the cluster to which it belongs and, at the same time, to maximize the distance from other cluster centroids.

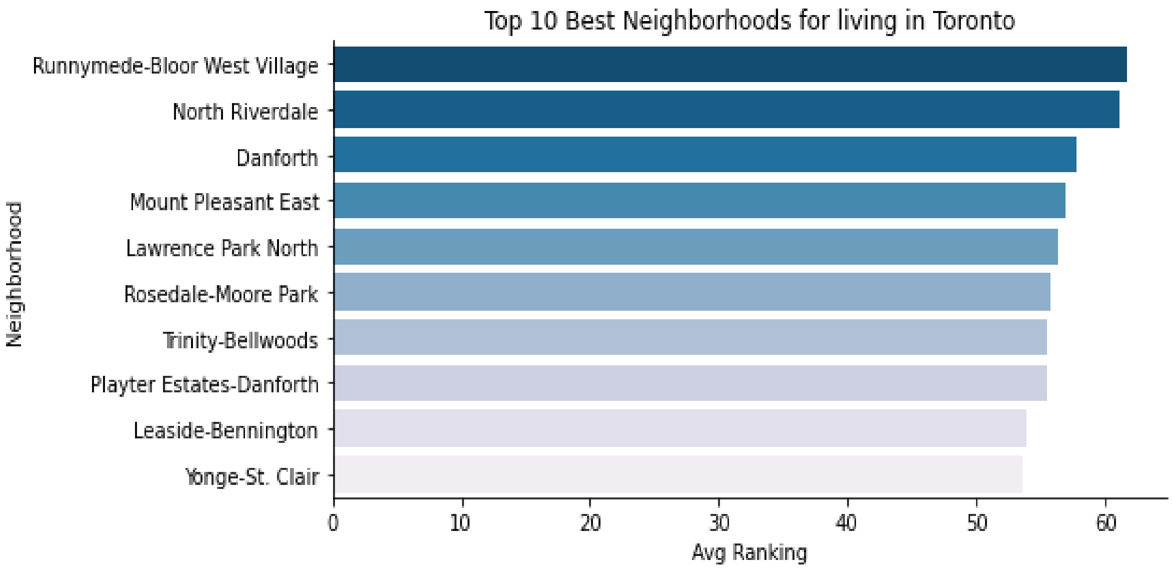
When implementing K-means clustering the number of clusters must be prespecified. One of the approaches for determining the optimal number of 'k' is by using the Elbow Method.

Here, Euclidean distance is employed as a measure of the dissimilarity between the data points and more specifically between each observation and its centroid. The Elbow Method is applied for determining the optimal number of clusters, 'k', and Python and its scikit-learn machine learning library are used.

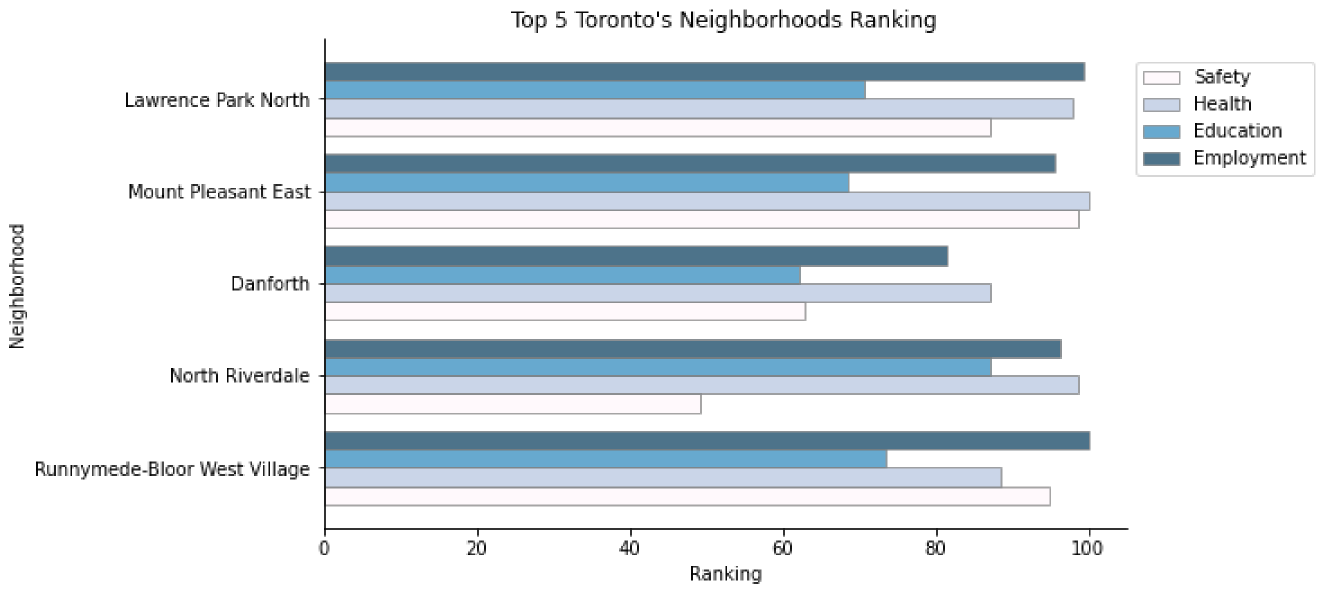
K-means algorithm relies on the correct choice of number of clusters, k. Therefore, the clustering is run across different values of k. However, increasing the number of clusters will always reduce the distance of centroids to data points. It is namely the squared distance between each point and its centroid that we are trying to minimize, i. e. to minimize the within cluster sum of squared errors. This means that increasing k will always decrease the error. Hence, K is determined as the elbow point on the elbow graph, where the rate of decrease in the error sharply shifts.

Exploratory Data Analysis and Empirical Findings

Toronto Life website provides data on Toronto's 140 Neighborhoods ranking based on 10 criteria: housing, safety, transit, shopping, health, entertainment, community, diversity, education and employment. These 10 criteria are used when grouping Toronto neighborhoods based on their attractiveness of living. Different approaches could be used to get an overall idea of the average rating/ranking of each neighborhood. One, would be to place equal importance on each of the 10 criteria and calculate a weighted average rating per neighborhood. Another approach, might weight higher some of these 10 criteria while other would receive a lower weight. The next graph shows the average ranking of the top 10 Toronto neighborhoods by placing equal importance on each of the 10 criteria: housing, safety, transit, shopping, health, entertainment, community, diversity, education and employment:

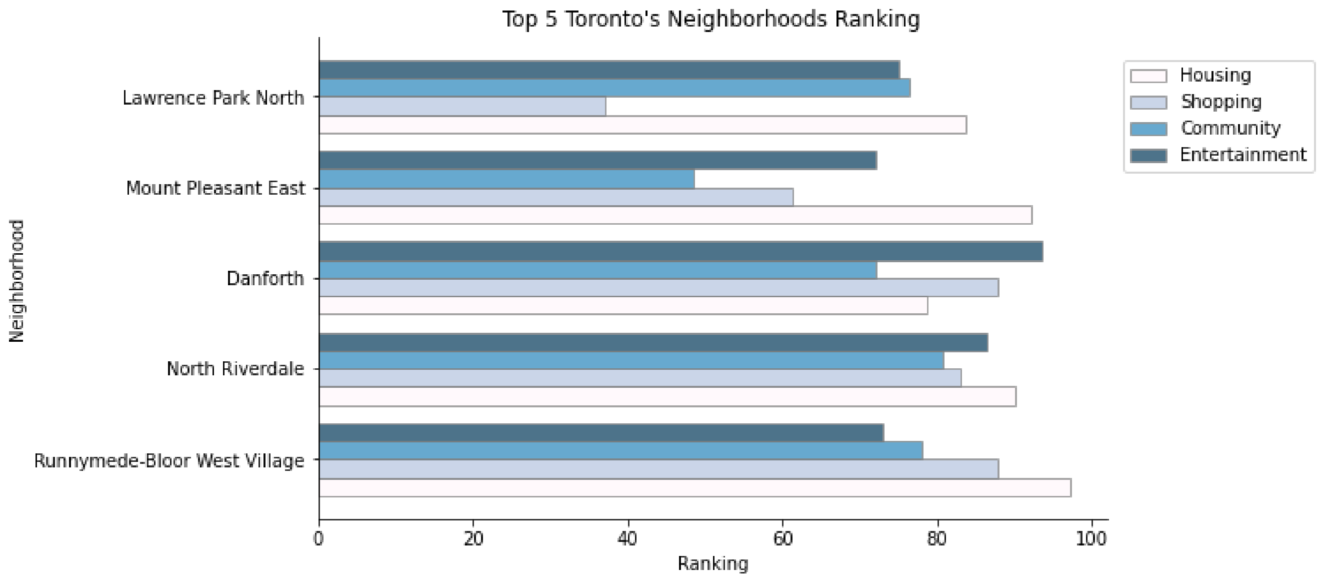


If we plot only safety, education, employment and health for the top five Toronto's neighborhoods based on average ranking, we get the following:



As evident from the graph, the top five Toronto neighborhoods seem to have high ranking and close scores based on employment and health, but differ in safety and education. It would be interesting to see *whether neighborhoods with namely high employment and health rankings are clustered together*.

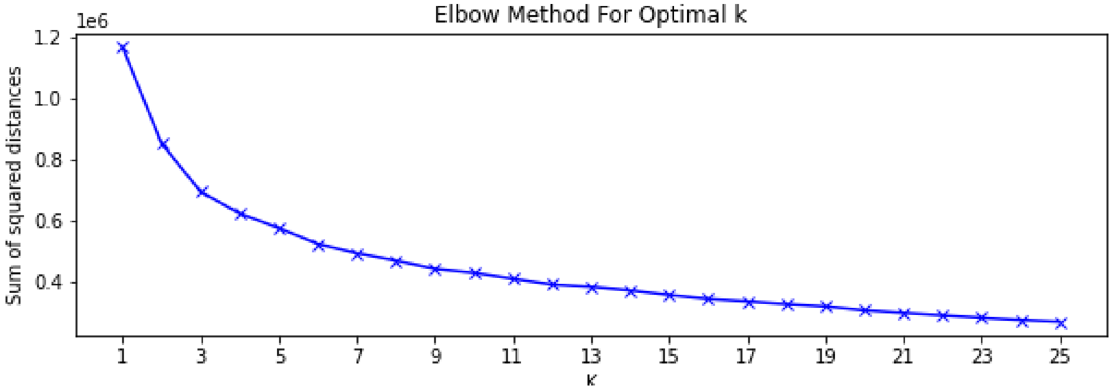
As next, we can see that among *housing, entertainment*, shopping and community the first two *seem to have high relevance for the ranking of top 5 Toronto neighborhoods* based on average rating.



In order *to see whether the relative importance of these criteria is the same for the grouping of the neighborhoods as it is for the average rating*, the analysis proceed with K-means clustering and determining the optimal number of clusters for the grouping of the neighborhoods based on their attractiveness of living.

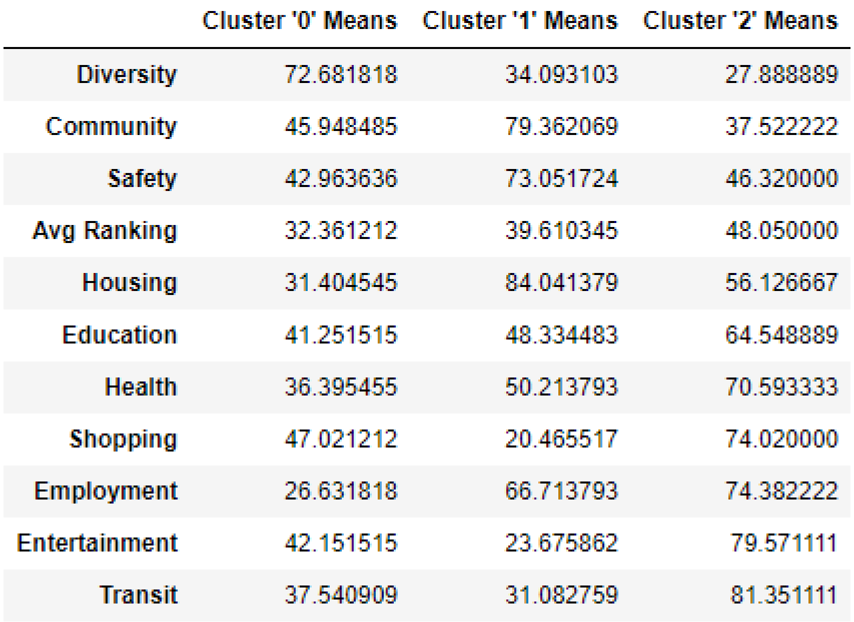
Clustering Based on Attractiveness of Living

K-means algorithm relies on the correct choice of number of clusters, k. Therefore, the clustering is run across different values of 'k'. However, increasing the number of clusters will always reduce the distance of centroids to data points. It is namely the squared distance between each point and its centroid that we are trying to minimize, i. e. to minimize the within cluster sum of squared errors. This means that increasing k will always decrease the error. Hence, K is determined as the elbow point on the elbow graph, where the rate of decrease in the error sharply shifts.

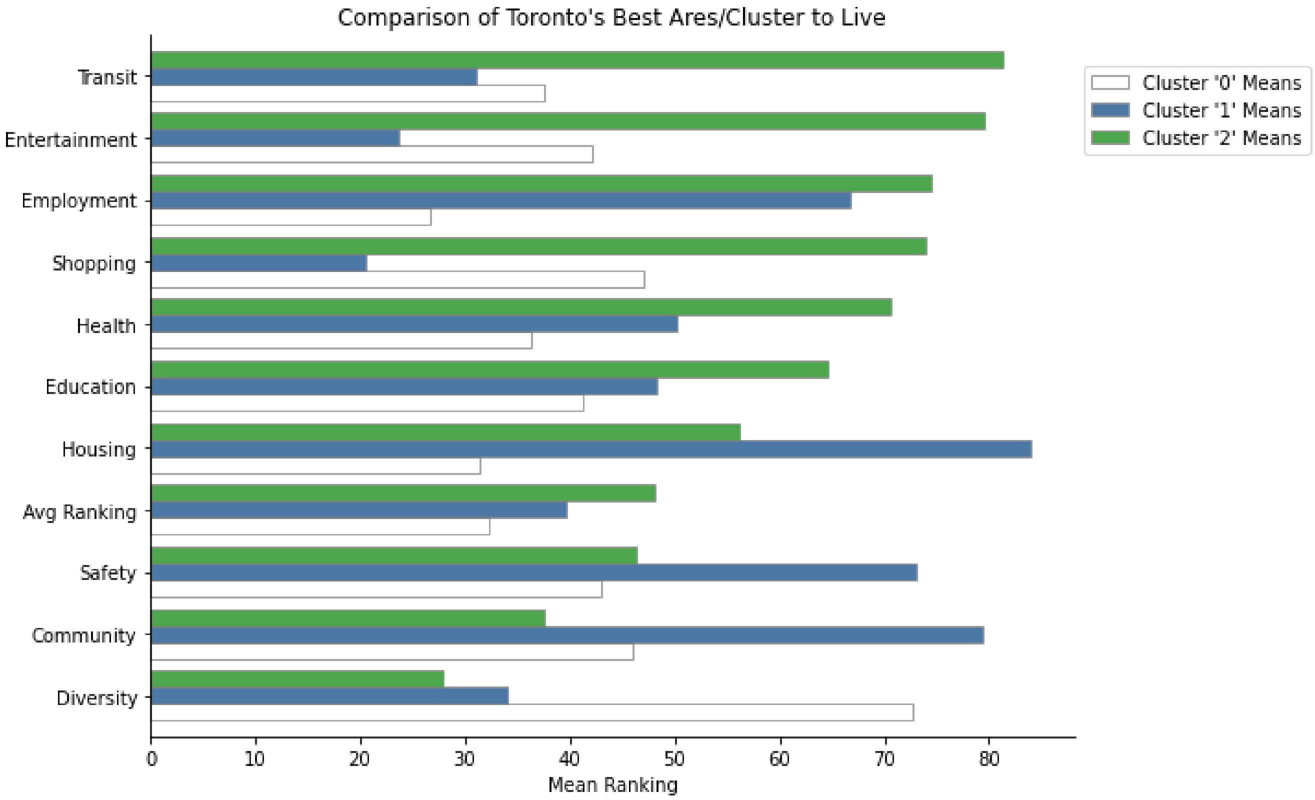


The graph indicates that the point where the rate of decrease in the error shifts in at K equals 3. After that the shift in the error slows.

Next, the K-means clustering is run with optimal K of 3 and in effect the results are as follows:



	Cluster '0' Means	Cluster '1' Means	Cluster '2' Means
Diversity	72.681818	34.093103	27.888889
Community	45.948485	79.362069	37.522222
Safety	42.963636	73.051724	46.320000
Avg Ranking	32.361212	39.610345	48.050000
Housing	31.404545	84.041379	56.126667
Education	41.251515	48.334483	64.548889
Health	36.395455	50.213793	70.593333
Shopping	47.021212	20.465517	74.020000
Employment	26.631818	66.713793	74.382222
Entertainment	42.151515	23.675862	79.571111
Transit	37.540909	31.082759	81.351111



The third cluster, shown *in green*, contains 46 neighborhoods and is the *second largest cluster* in terms of number of neighborhoods. It has *highest average ranking* and highest employment, education, entertainment, health, shopping and transit, meaning that the second largest cluster (based on number of neighborhoods) is *the leader in 6 of 10 ranking criteria*.

The second cluster, shown *in blue*, has only 27 neighborhoods and is *the smallest* of the three (based on number of neighborhoods). It has *the last but one average ranking* and *highest average scores for community, safety, housing*.

The first cluster, shown *in white*, is *the largest one* (based on number of neighborhoods). It includes 67, *almost half of the 140 neighborhoods* and yet it is *the worst graded*. It is *the leader in terms of diversity*. Diversity is seen as the percentage of visible minorities, people whose mother tongues are not French or English, and first- and second-generation immigrants. However, it yields *the worst average results for education, employment, health, safety and housing*.

The fact that almost half of the neighborhoods fall there means that Toronto is a really diverse city, but its diversity is not what makes it attractive for living. Obviously, as the highest average ranking implies, what makes an area attractive for living is employment, education, entertainment, health, shopping and transit.

So, all in all, it could be concluded that Toronto neighborhoods are clustered into three groups. The first and most numerous is most diverse, where people from different countries live together, but it has worst conditions for education, work, healthcare, worst housing and safety.

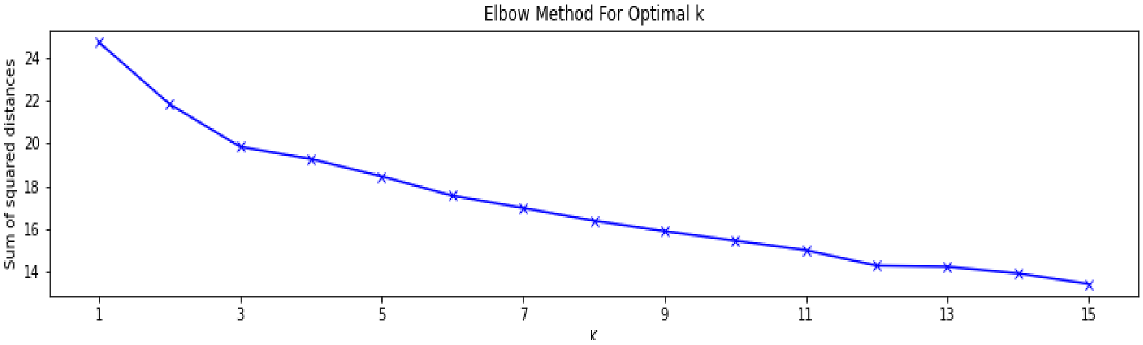
On the other hand, the best ranked neighborhoods provide best conditions for education, work, healthcare, entertainment, shopping, traveling.

And there is the smallest cluster (based on number of neighborhoods), with highest safety, housing and community scores.

Obviously, it would be interesting to see the average age of the population in the clusters, as the smallest (based on number of neighborhoods) is the safest one, with best community and housing performances. And there is the average-sized cluster with best opportunities for education, work, entertainment, healthcare, traveling.

Clustering Based on Venue Types

The second part of the analysis makes use of Foursquare API for exploring venues and clustering neighborhoods based on their venue types. For each neighborhood we get maximum 100 venues within radius of 500 meters of its the longitude and latitude. Again, similarly to the clustering based on attractiveness of living, the optimal number of clusters, K is determined via the Elbow Method. The squared distance between each point and its centroid for different K-s is shown next.



Since the squared distance from the observations to their closest centorids is what we try to minimize, K is determined as the elbow point on the elbow graph, where the rate of decrease in the error sharply shifts. Here, it is K equals 3.

The neighborhoods are clustered based on venue category, meaning that the categories will serve as features. K-means algorithm uses some distance metric, for example Euclidean distance, which means that it isn't directly applicable for categorical variables. Hence, venue categories will be converted to dummy variables with value of 1 if such venue is present in the neighborhood and 0 otherwise. So, we apply 'one hot encoding' to the venue categories. After that, the rows are grouped by neighborhood and by taking the mean of the frequency of occurrence of each venue category and the resulting dataframe is used for the clustering.

The results for the three clusters reveal that:

- The first cluster contains 110 neighborhoods.

The most common venues that account for 51.18% of the 1st most common venues in the neighborhoods in the cluster is/are Coffee Shops, Cafés, Italian Restaurants.
- The second cluster contains 2 neighborhoods.

The most common venues that account for 100.00% of the 1st most common venues in the neighborhoods in the cluster is/are Pools.
- The third cluster contains 28 neighborhoods.

The most common venues that account for 51.59% of the 1st most common venues in the neighborhoods in the cluster is/are Parks.

Conclusion

All in all, Toronto neighborhoods clustering based on their *attractiveness of living*, determined as the average ranking in each of these 10 categories: housing, safety, transit, shopping, health, entertainment, community, diversity, education and employment seems *perfectly plausible*. The grouping of neighborhoods in three groups – one, that is smallest in terms of number of neighborhoods, with the last but one average rating, best safety and community rankings; other, that has the best average rating, which is average-sized cluster with best opportunities for education, work, entertainment, healthcare, traveling and third, that is the most numerous with highest diversity but worst conditions for education, employment, health, safety and housing raises at least one ‘hot’ question. Obviously, it would be interesting to examine the average age of the population in the clusters in further analysis.

References

1. Foursquare 2020, viewed on 07 May 2020, <https://foursquare.com/>.
2. Toronto Life 2020, *The Ultimate Neighbourhood Rankings*, viewed 07 May 2020, <https://torontolife.com/neighbourhood-rankings/ >.