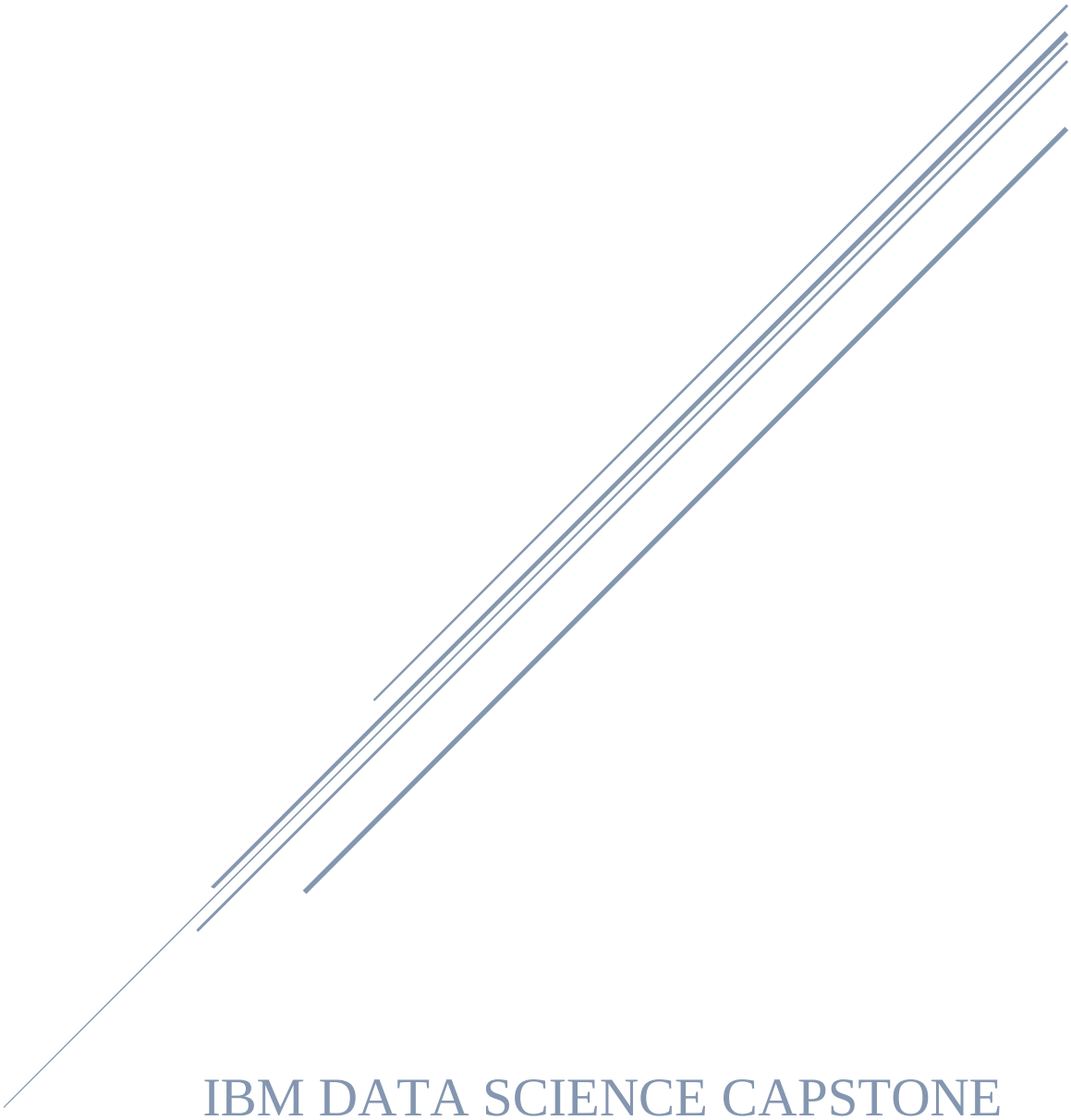
CLUSTERING TORONTO'S NEIGHBORHOODS BY SUSTAINABILITY INDICATORS USING K-MEANS CLUSTERING

AZIM SHAMSHIEV



IBM DATA SCIENCE CAPSTONE

1. Introduction

The goal of this project is to evaluate the sustainability performance of 140 neighborhoods in the City of Toronto. To this end, the k-means clustering algorithm, a popular (unsupervised) machine learning technique, will be drawn upon using a range of variables pertaining to the three domains of sustainability – social, economic and environmental. The clustering technique will help partition the neighborhoods into groups (clusters) and identify their distinct characteristics.

The need for such analysis arises from the fact that the evaluation of social, economic and environmental indicators is often performed in isolation from one another. However, analyzing sustainability requires a comprehensive look at the state of social, economic and environmental well-being of a unit. In addition, grouping neighborhoods by a single indicator is relatively simple, while doing the same with multiple indicators is a complex task and requires more advanced methods. Therefore, this project attempts to comprehensively examine relevant indicators using the K-means model and propose sustainability profiles by which neighborhoods can be grouped.

Sustainability-based clustering can be used in setting a benchmark for tracking the progress of neighborhoods over a period of time. It can also enable a status quo comparison between neighborhoods at a given time. In more practical terms, it can be a valuable tool in solving real-world business problems such as locating optimal neighborhoods for housing development projects. For example, green buildings in sustainable neighborhoods can be made a key element of a company's brand and business model.

2. Data collection

Sustainability can be measured using different criteria. A specific choice of indicators depends on the scope of a study and the availability of data. I used the following three indicators to examine the three dimensions of sustainability in Toronto's neighborhoods:

- Green spaces (environmental)
- Healthy food index (social)
- The number of businesses (economic)

Foursquare API and the City of Toronto's Open Data Portal were the main sources of data for the project. Initial data on the amount of green spaces was acquired from Foursquare. More complete data on green spaces and the other two indicators were accessed through the City of Toronto.

3. Methodology

A number of data science procedures and tools were utilized in the project to make sure the analysis is methodologically robust and accurate in terms of its results.

3.1. Normalization

Since the neighborhoods are not of the same size, the amount of green spaces and businesses were normalized by each neighborhood's population. Thus, green spaces are presented in square kilometers per 1000 people and businesses as the number of businesses per 1000 people.

3.2. Metric selection for green spaces

I used Python's Folium library to visualize the locations of neighborhoods. As per the requirement of the project, I also used Foursquare API to extract venues for each neighborhood. I set the limit for venues at 100 and the radius of 1000 meters from the center of neighborhoods. I further narrowed my search to parks and initially used them as proxies for green space. A limitation of using Foursquare is that it only gives the number of parks in a location and disregards their actual areas. Therefore, where possible, an area-based metric must be applied since it gives a much more accurate picture of green spaces. Fortunately, the city of Toronto provides data on the total area of green spaces by neighborhoods. Moreover, the data includes not only parklands but also utility corridors and utility areas such as soccer fields. Since such data was publicly available, I substituted the number of parks with the total area of green spaces as my metric for the environmental dimension of sustainability.

3.3. K-means clustering

For clustering, I used the K-means model which is widely used in machine learning because it is simple and yet powerful. It is a distance-based algorithm aiming to partition the entire data into groups or clusters. It assigns each instance to the closest centroid (cluster center). Centroids can

be initialized randomly but eventually become the mean of all instances in their clusters. Therefore, clustering is a repetitive process which continues until either a pre-defined number of iterations is reached or centroids (as the means of their clusters) stops to change. The goal is that instances must be as close as possible to their own centroids and as far as possible from the centroids of other clusters.

3.4. The elbow method

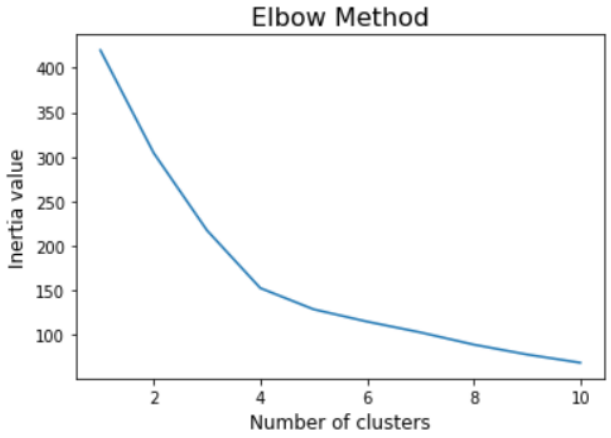
In K-means, *k* or the number of clusters is input manually, the procedure known as hyper-parameter tuning. The elbow method is a technique that helps to identify the optimal number of clusters. The method compares inertia (intra-cluster sum of square distances) with the number of clusters. The goal is to find the region ("elbow of the curve") where inertia drops considerably and the curve begins to bend. The choice of a specific cluster number in the region is a matter of intuition. It should also be kept in mind that beyond the region identified through the method inertia can be decreased only marginally but at a high computational cost.

3.5. Discretization and outlier treatment

To analyze the clusters, variables without outliers were divided into three equal intervals. The intervals were labelled as "low", "medium", and "high". The same was implemented for variables containing outliers, however, outliers were defined as a separate group, i.e. "very high".

4. Results

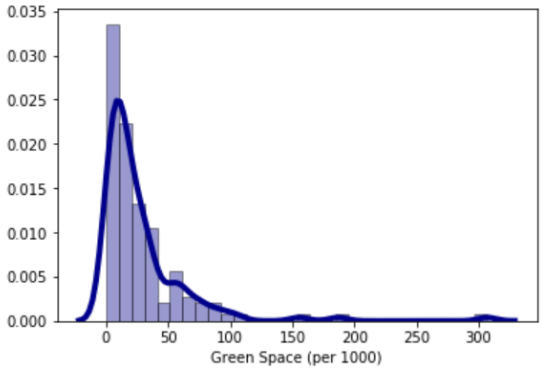
The implementation of the elbow method suggests the optimal number of clusters lies between 4 and 6. From this range, the number of clusters was set at 5.

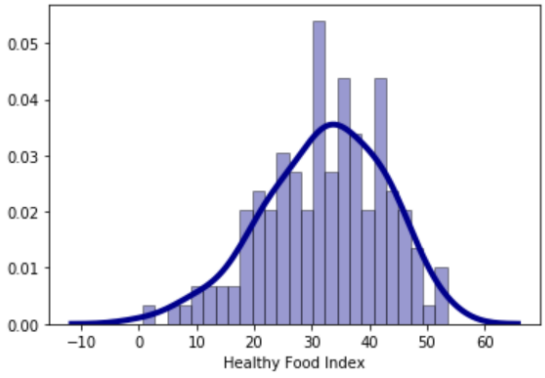


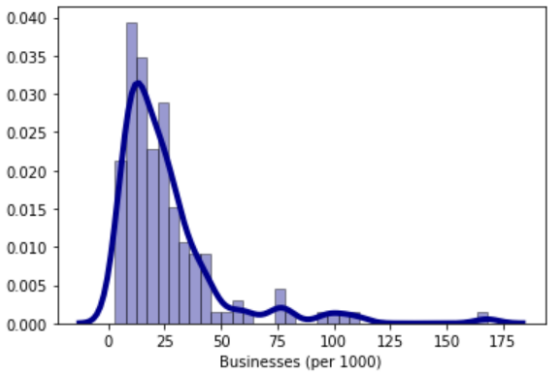
Below is the sorted break-down of neighborhoods by 5 clusters after running the model. The clusters are unevenly distributed with the majority of neighborhoods located in clusters 0, 2, and 3.

Clusters	Number of Neighborhoods
0	55
3	37
2	33
4	12
1	3

For a meaningful comparison of clusters, variables had to be divided into ranges. The graph below illustrates bins and density curves for each variable. They were used to detect outliers and discretize the variables.



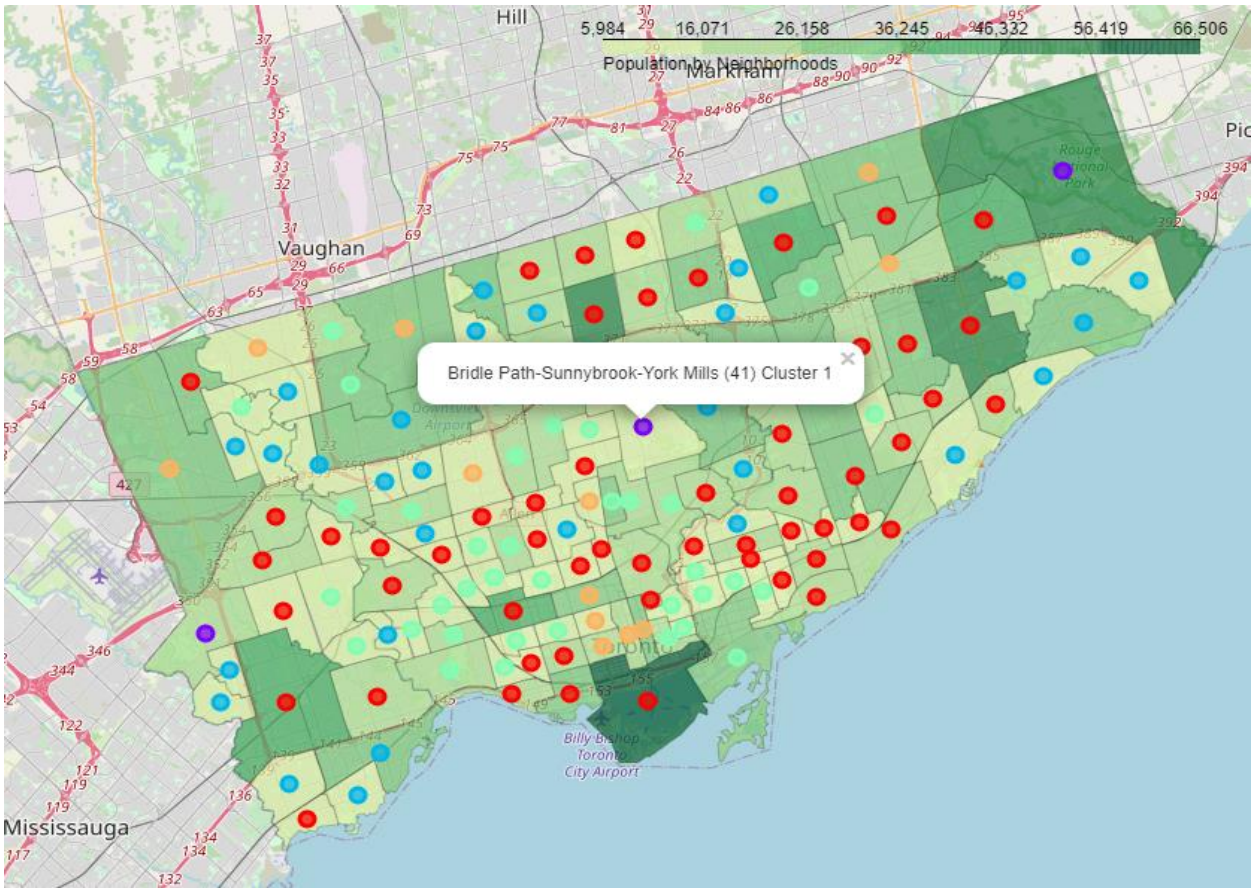




Healthy food index shows a normal-like distribution, whereas both green space and businesses are right skewed with some extreme outliers. The limits for outliers were identified using the 3xIQR rule (Q1-3xIQR for lower limit and Q3+3xIQR for upper limit). Below is the summary of ranges for each variable.

	Ranges	Green spaces	Healthy Food Index	Businesses
0	Low	0-36	0-19	0-29
1	Medium	37-71	20-36	30-54
2	High	72-107	37-54	55-79
3	Very high (outliers)	107-307	na	80-168

The following map shows the cluster labels superimposed on the choropleth map of the city's population.



Cluster 0 – Red | Cluster 1 – Violet | Cluster 2 – Blue | Cluster 3 – Green | Cluster 4 -Orange

5. Discussion

First, it should be noted that in most cases indicators within the clusters do not fall within a single range. The table below illustrates that only healthy food index in cluster 3 and businesses in cluster 1 have a single range, high and low, respectively. The rest are represented by two or three ranges. This implies that in choosing neighborhoods one should be look into clusters as well as specific ranges within the clusters. To aid with this task, the number of neighborhoods in each range were provided in brackets.

	Clusters	Green Space	Healthy Food Index	Businesses
0	Cluster 0	Low (47) - Med (8)	Med (46) - High (9)	Low (46) - Med (49)
1	Cluster 1	Very High (3)	Low (1) - Med (2)	Low (3)
2	Cluster 2	Low (19) - Med (9) - High (15)	Low (14) - Med (19)	Low (28) - Med (5)
3	Cluster 3	Low (33) - Med (1) - High (3)	High (37)	Low (25) - Med (12)
4	Cluster 4	Low (10) - Med (2)	Med (8) - High (4)	High (7) - Very High (5)

Second, although among the 5 clusters of neighborhoods none show high scores on all three variables, some clusters, with some trade-offs, appear to be more suitable options than others. Given such situation, the main criterion was to avoid low ranges and capture the most amount of high or very high ranges with some medium ranges as trade-offs. With this criterion, South Riverdale in cluster 3 appears to be the best option for our business problem (housing development) since it has high scores on green space and healthy food, and a medium score on businesses. However, with the exception of this particular neighborhood, the cluster does not contain any other neighborhoods with at least two medium indicators despite the fact that the entire cluster shows high score for healthy food.

Neighborhood	Cluster Labels	Green Space (per 1000)	Healthy Food Index	Businesses (per 1000)
South Riverdale (70)	3	High	High	Med

Cluster 4 seems to be another option. More specifically, it has 2 neighborhoods (Humber Summit & West Humber Clairville) which fall within the medium range for green space and medium to very high ranges on the other two variables.

Neighborhood	Cluster Labels	Green Space (per 1000)	Healthy Food Index	Businesses (per 1000)
Humber Summit (21)	4	Med	Med	Very High
West Humber-Clairville (1)	4	Med	Med	High

6. Conclusion

As sustainability becomes ever more important with a plethora of environmental, social and economic challenges, a holistic assessment of our well-being, both locally and globally, becomes

an imperative. This stresses the need to benchmark and monitor the performance of our neighborhoods on sustainability indicators.

As noted at the outset, clustering with multiple variables can be a challenging task. Therefore, while our aspirational goal is to look for cluster(s) with the best values for all indicators, our practical goal, as in most real-life scenarios, is to find the right balance and trade-off between imperfect alternatives and competing priorities. Ultimate trade-offs will often depend on available options, resources, and priorities of stakeholders. In this regard, K-means clustering can be a good starting point to group and analyze large numbers of local units such as neighborhoods with multiple variables at hand.