New Venue implementation Analysis

Toronto - Canada

Executive Summary

This data science project is designed to support the decision of choosing the best neighborhood in Toronto to establish a new restaurant.

Table of Contents:

Introduction

This project is designed to support the decision of choosing the best neighborhoods in Toronto to establish a new Restaurant.

Background:

- Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016, The city's population grew by 4.3% (111,779 residents) between 2006 and 2011, and 4.5% (116,511) between 2011 and 2016.

- Toronto is an international centre for business and finance. Generally considered the financial capital of Canada, Toronto has a high concentration of banks and brokerage firms on Bay Street, in the Financial District. The Toronto Stock Exchange is the world's seventh-largest stock exchange by market capitalization. The city is an important centre for the media, publishing, telecommunication, information technology and film production industries,

- Toronto is one of Canada's leading tourism destinations. In 2017, the Toronto-area received 43.7 million tourists, of which 10.4 million were domestic visitors and 2.97 million were from the United States, spending a total of $8.84 billion. Toronto has an array of tourist attractions, and a rich cultural life.

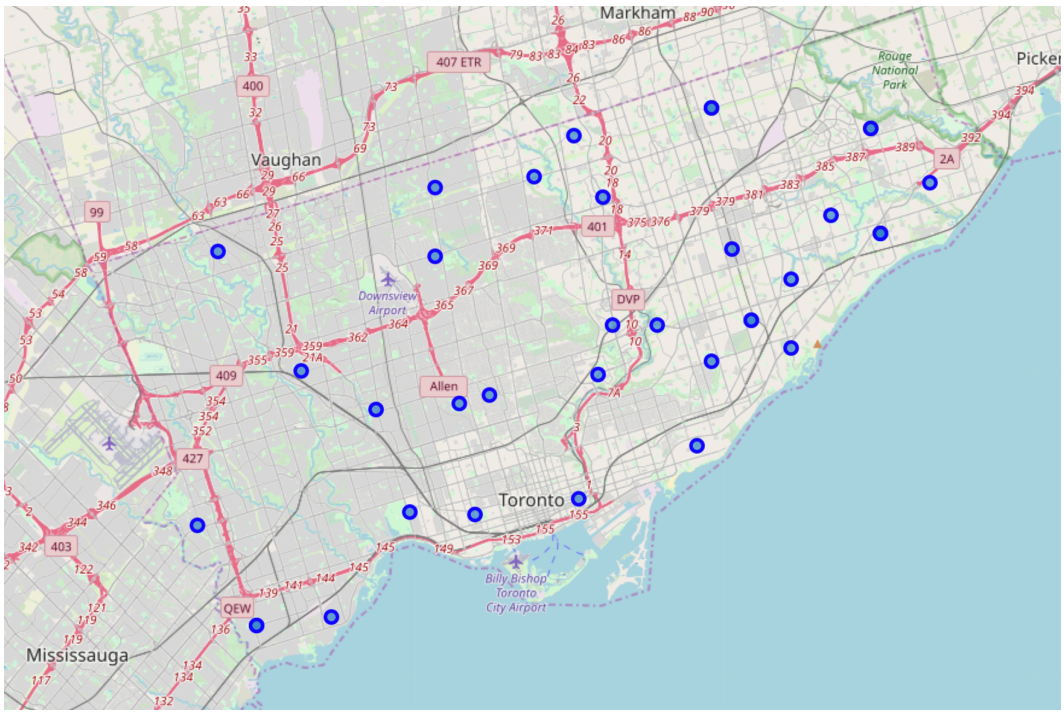
The Problem Question:

In this fast growing city, we can see that the demand on services like restaurants is increasing, and the problem that face Everyone who consider opening a new business in Toronto is:
What is the best neighborhood in Toronto to establish my new business?

Data Collection

For the present problem we will use data collected from 3 main sources:

- **Wikipedia**: This will be used to gather all the neighborhoods in Toronto.



CITY OF TORONTO

- **Neighborhood Profiles** :provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighborhood. The data is based on tabulations of 2016 Census of Population data from Statistics Canada https://www.toronto.ca
- **Foursquare API**: Will be used to get the venues surrounding each specific neighborhood in order to cluster them and have a better understanding of how involved a neighborhood is in the healthy foods sector. Each source and gathering methods is shown in the code cells below
- **Geopy**: This library will be used to get the coordinates (latitudes and longitudes) for Toronto as well as for each of its neighborhoods;

Methodology

For this specific case problem, the sequence of the Data Science methodology was deliberately changed, the machine-learning algorithm was executed before the exploratory data analysis since it represented a new source of extra information for the analysis.

Machine Learning Algorithm

For the present analysis, the chosen algorithm was an unsupervised learning model for data clustering. The reason for the present choice is due to the fact that the analysis aims to identify similar neighborhoods that could be potential site for the new venue, based on Income and the presence of F&B venues on those neighborhoods. Therefore, the K-means algorithm was the choice to identify such clusters.

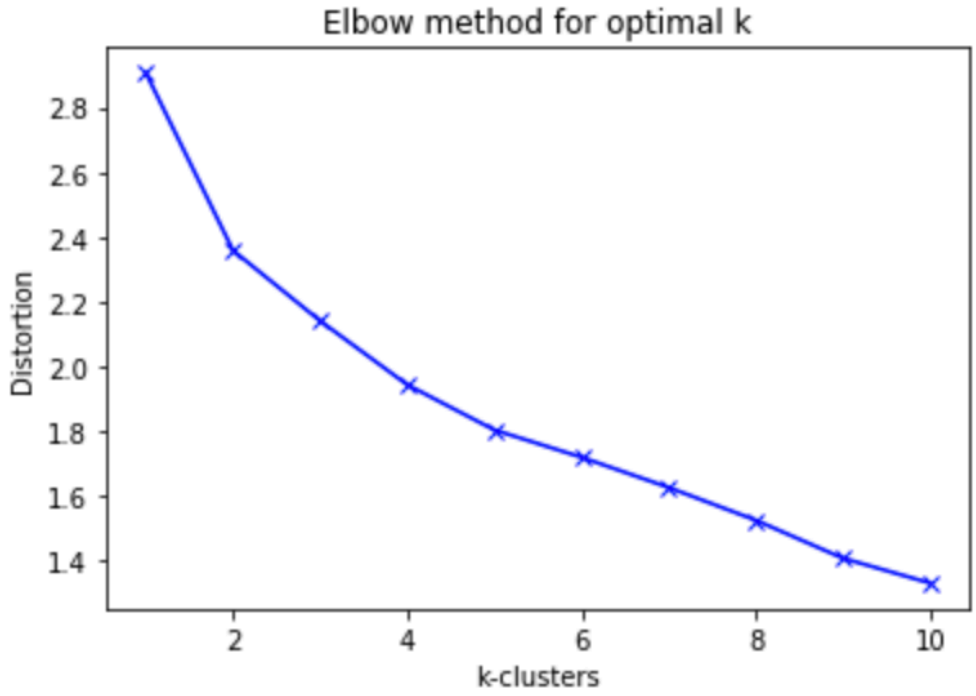
The K-Means algorithms, as stated, is a model of unsupervised learning where it tries to minimize the Euclidean distance between each observation and their cluster's centroid, while maximizes the distances between them and observations from other clusters, all that based on the number of clusters that the user specified.

That presents a challenge for the current analysis on how many clusters should be used to gather the maximum information without overfitting the model. For that matter, we performed the elbow method. This method consists on running the model many times with a different number of clusters (as known as k) and storing the accumulated distortion – measured by the Euclidean distance between the points and their cluster's centroid – so we check when the information gain starts to be minimal, that would define the optimal k.
Finally, the model was created using the following attributes:

1) Income,
2) Number of restaurants,
3) Number of markets,

4) Number of bar and
5) Number of cafes.

A standard scale of said attributes were performed before the model execution to ensure no bias taken towards a specific attribute.
When performing the *elbow* method to find the optimal *k* number of clusters, no clear point



THE ELBOW METHOD, DISTORTION LOSS BY NUMBER OF CLUSTER IN THE MODE

could be made from the curve – Figure 2. That happened because the data is too sparse (too many zeros in the dataset), however we can notice that the distortion loss after 8 clusters started to be a little flatter than before, therefore we continued the analysis with 8 clusters in total.

Cluster Label	Number of Neighborhoods
0	**9**
1	**7**
2	**5**
3	**5**
4	**4**

Cluster Label	Number of Neighborhoods
5	**3**
6	**2**
7	**1**

COUNT OF NEIGHBORHOODS IN EACH CLUSTER

Exploratory Data Analysis

Now based on the clusters output from the machine learning algorithm an analysis on the clusters' characteristics could be conducted, first an outlook on each cluster's means for the selected attributes were executed.

Cluster Labels	Value	Restaurants	Markets	Bar	Cafe
5	14869.500000	26.500000	2.000000	2.000000	6.500000
2	6071.142857	36.571429	4.142857	0.714286	5.714286
3	5931.200000	20.800000	4.000000	3.200000	14.800000
7	5796.000000	28.750000	1.250000	11.000000	12.250000
0	5134.400000	22.000000	1.800000	4.400000	9.600000
4	4208.666667	36.333333	1.000000	0.333333	4.666667
1	2865.666667	27.888889	2.333333	0.888889	7.000000

Clusters' Attribute mean

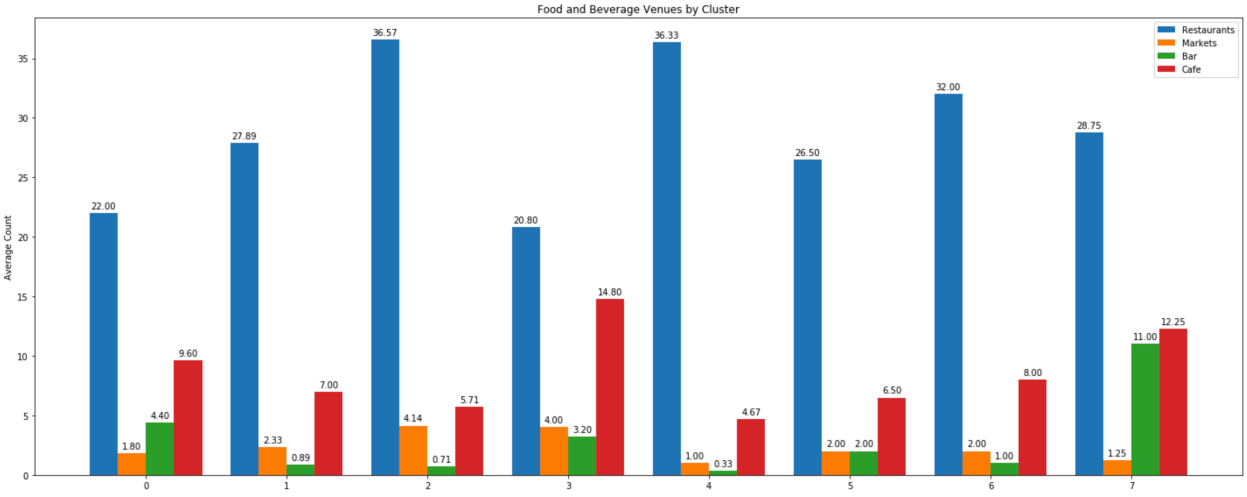
In terms of F&B, we have two clusters that contains a high average – 5 and 4 – and some clusters with a reasonably low average – 1, 0 and 7. Now visualizing the venues attributes on a bar chart we have as shown in Figure 3.

Some observations can arise from the chart in Figure 3:

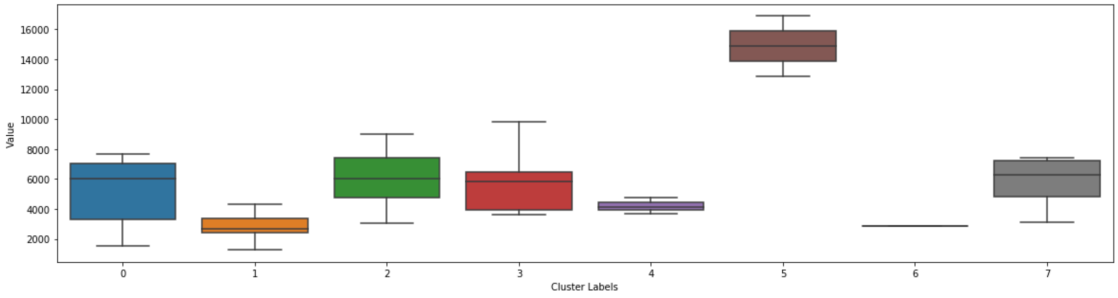
- Clusters 2, 4 has the large number of restaurants and the lowest number of bars.

- Cluster 7 has the highest number of bars.

- although cluster 5 is the highest income cluster but it has low number of F&B venues.



CLUSTERS' VENUE ATTRIBUTES



BOXPLOT DISTRIBUTIONS OF INCOME FOR EACH CLUSTER

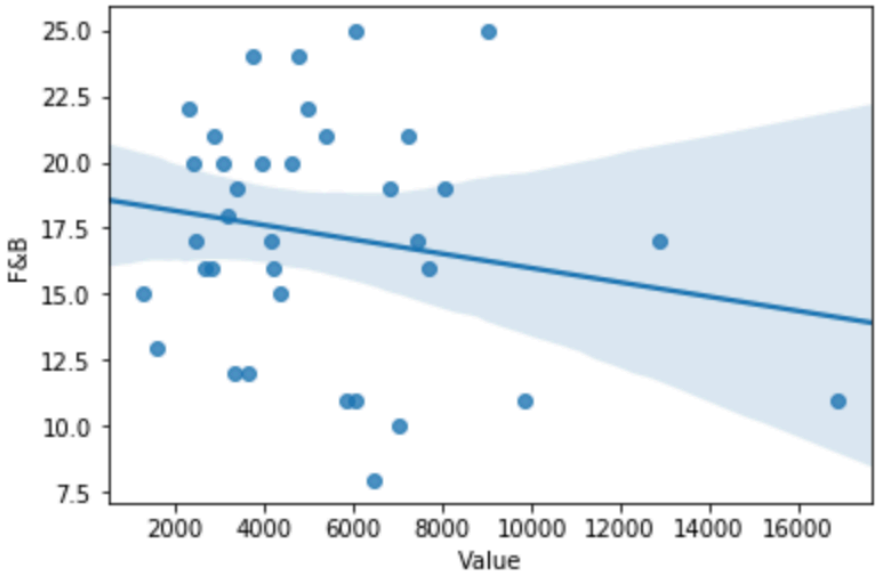
In order to drawn better conclusions, we should analyze the observations from the Figure 3 with Figure 4 where we can see HDI distributions for each cluster.

- The data seems to show a correlation between the number of healthy venues and the HDI score of the neighborhoods. For instance, the clusters with the highest counts of healthy venues – 2, 4 and 5 – are also the clusters with the highest HDI scores, Table 8 shows correlation matrix and Figure 5 shows a scatterplot of HDI and Healthy Venues

count, based on the observation of those two visual there is a conclusion that they are, in fact, correlated

	Value	F&B
Value	1.000000	0.038436
F&B	0.038436	1.000000

CORRELATION MATRIX OF INCOME AND NUMBER OF VENUES



SCATTER PLOT OF NEIGHBORHOODS INCOME AND NUMBER OF F&B VENUES

That observation proves a concept common on region, which is that healthy food and lifestyle is usually more expensive and, therefore, only people with a higher monetary power can afford. Another concept behind that 0.6 correlation is the fact that with a higher education and development comes a higher sense on the need to be healthy, therefore, it is safe to say that the higher the HDI of a given neighborhood the higher is the likelihood of the population embracing a new Health Foods Market. So, the potential neighborhoods should have a high HDI.

Based on that two clusters stand out from the rest with potential neighborhoods for the new venue, they the 2 and 7 clusters, the reasons are shown below and the neighborhood spread across the map on Figure 6.

- Cluster 2 concentrates a lot of high HDI neighborhoods (ranging around 0.95), has a lot of gyms which might means that the population on those neighborhoods are very healthy conscious, however the amount of Healthy Markets there seems very low scoring only 0.4 on average, which gives a potential growth possibility of the health sector;

- Cluster 7, although it's got a wider HDI distribution – we might be able to check some of the neighborhoods in the 3rd or 4rth quartile to stay in the high HDI range – this cluster contains neighborhoods with a high amount of gyms and no Healthy venues, Markets or even Restaurants, making them new territory for the health sector.

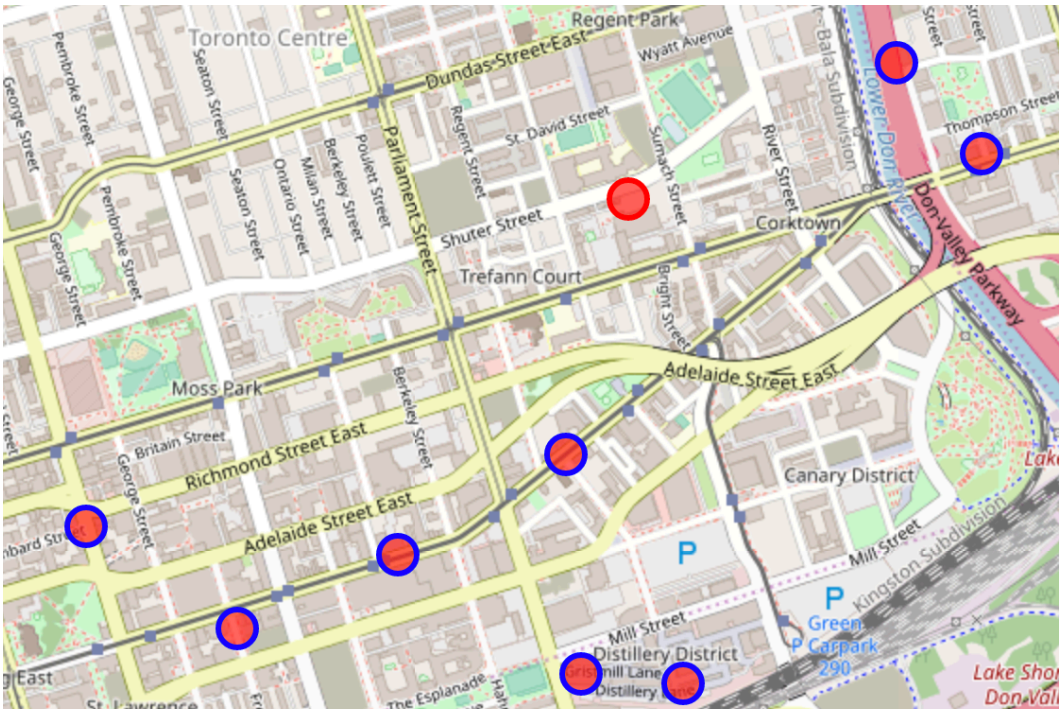
Result

Finally, based on the described methodology and analysis process for the present case, the top neighborhoods for the new venue lays in the clusters 2 and 7, utilizing the engineered feature, the HH Index, to gather the top 5 neighborhoods we end up with the list shown in Table

	Neighbourhood	Cluster Labels	F&B	value index
3	The Beaches	2	25	14787.073171
0	Malvern	2	22	12079.024390
1	Flemingdon Park	2	25	9923.736264
9	Forest Hill North	2	19	8854.945055
10	Long Branch	2	20	5664.197531

TOP 5 VALUE INDEX NEIGHBORHOODS IN CLUSTER 2

And as a recommendation spot, by taking the centroid between the gyms in the best HH Index scoring neighborhood – The Beaches – the optimal spot would be as illustrated on Figure 7. That assumption takes as premise the fact that the population usually prefer to go to the gym that is closer to their homes, therefore a region with many gyms would mean a great number of health conscious residents.



OPTIMAL SPOT BASED ON BEST VALUE INDEX NEIGHBORHOOD'S CAFES

Discussion

Based on the observed data and conducted analysis' insights we might infer that – due to the correlation between the Human Development index and the amount of healthy venues in the neighborhood – for an optimal spot to establish a new Health Foods Market, we might look for a high HDI. In addition, due to the beginning stage that the health sector is in Brazil, especially in Sao Paulo, we can look for the lack of those kind of venues in the selected neighborhoods to assess how saturated the market is for the given area.

Finally, the present analysis can recommend that the entrepreneurs interested in establishing a new Health Food venue can focus their attentions to neighborhoods with high HDI and low number of already established health markets, such as the top five recommended by the model
1) The Beaches,
2) Malvern,
3) Flemingdon Park,
4) Forest Hill North, and
5) Long Branch

Conclusion

In summary, the presented analysis could achieve the desired information, which was to come up with the best spots for establishing a new Health Food Market. The recommendation was to focus on high HDI and low already established Health Market count, especially the top five listed in the previous sections.