Selecting neighborhood for a new wine store in Toronto.

Miguel Tamarit

May 2020

1. Introduction

1.1 Background

Wine is a product often consumed at home and in restaurants.
Restaurants in particular have few stock so every day they must purchase products they use. The providers of wine for restaurants must be prepared to serve their customers in short time, so they preferably must have the store near their customers.
On other hand, the amount of wine people buy in a store is related to population in the neighborhood, and their income amount.

1.2 Problem

Data is widely available in this time, and it is the key to find the best places for any business to locate a new shop or office
This project will enable data to predict the best neighborhoods to open a new wine store.

1.3 Interest

A retail store wants to open a store in Toronto. This company sell wines to both people directly in the shop and to restaurants.
They want to open their store near their potential customers.
They think the best place to start business in the city will be a neighborhood with as many restaurants as possible, with more population (where presumably their products would be consumed more often), and with more average income where their more expensive (and more profitable) products will sell better.
They want me to find them several neighborhoods with this criterium, but finally will be they who select the final location.

2. Data acquisition and cleaning

2.1 Data sources

The population and income for every Toronto neighborhood can be found in wikipedia pages
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
Some web scrapping technics will be needed to get the data from this web page. The dataset is from several years ago, but this is not a quick changing data in percentage, so it will serve for purpose.
On the other hand, the restaurants in every neighborhood can be obtained from FourSquare using their API.

2.2 Data cleaning

The income and population for every Toronto neighborhood are columns in the referred wikipedia web page
Web scrapping is the technique to get this information from the page into a usable data frame

2.3 Feature selection

Data venues for Toronto neighborhood must be filtered to get only the restaurants and summed to get the amount of them in every neighborhood.
The income and population for every Toronto neighborhood are columns in the referred wikipedia web page. The only information to take is the next 3 columns: the name of the neighborhood, the average income, and population
Information tables from both sources must be combined into one table.

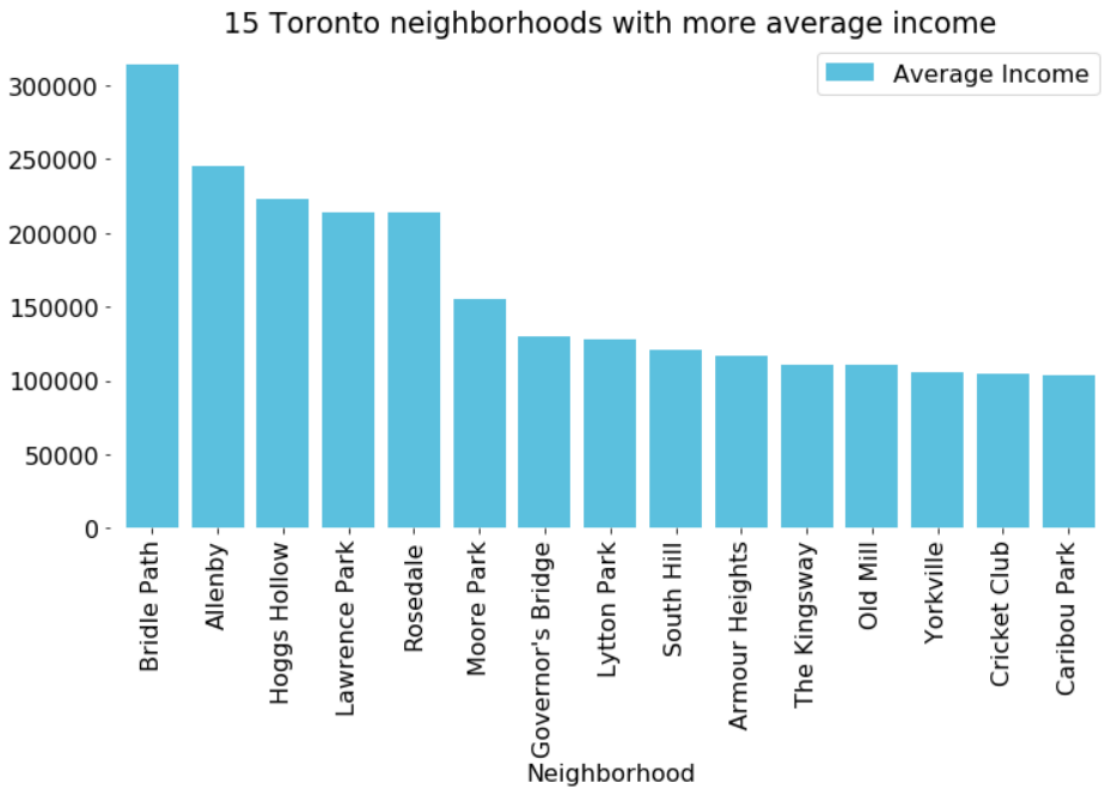
3. Methodology

3.1 Analyze Toronto neighborhoods by population

Get the population of every Toronto neighborhood and review the data. After taking the data from the Wikipedia, check some neighborhoods to be sure the data has been correctly obtained. Afterwards see the most populated neighborhoods

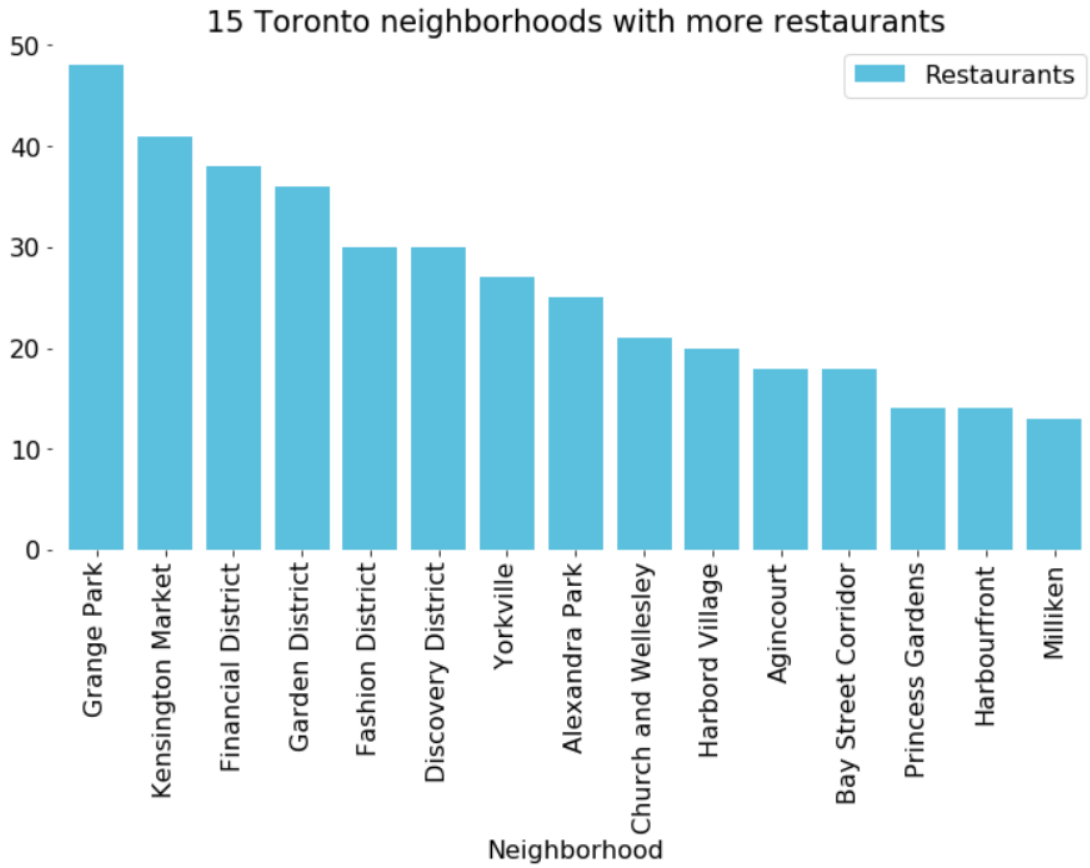
3.2 Analyze Toronto neighborhoods by average income.

Get the average income of every Toronto neighborhood and review the data.
After taking the data from the Wikipedia, check some neighborhoods to be sure the data has been correctly obtained. Afterwards see the neighborhoods with more average income



3.3 Analyze Toronto neighborhoods by number of restaurants

Get the number of restaurants of every Toronto neighborhood.
This information is taken from the Foursquare using its API.
After taking the data, check some neighborhoods using google or other similar tool to check the data is accurate. Afterwards see neighborhoods with more restaurants



3.4 Clustering

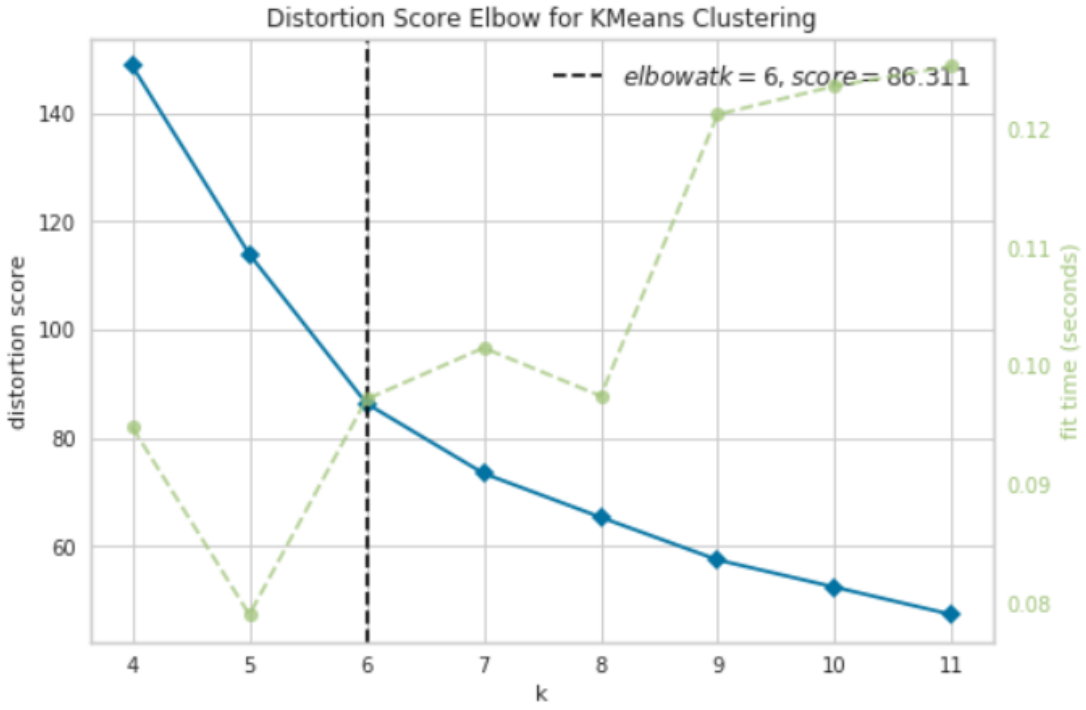
To group the similar neighborhoods depending on their population, average income and number of restaurants, I will use a clustering algorithm.
The selected clustering algorithm is k-means.
First, I must merge the information in a single dataset.

Merge the population, average income and number of restaurants in a single data frame, to be used in a k-means clustering

	Neighborhood	Population	Average Income	Restaurants
0	Agincourt	44577	25750	18
1	Alderwood	11656	35239	1
2	Alexandra Park	4355	19687	25
3	Allenby	2513	245592	3
4	Amesbury	17318	27546	0

Afterwards this information standardized with the “StandardScaler” to take every one of this three variables with the same weight por the clustering process, using it in a k-means clustering.

Once the information is standardized the best K for the k-means must be found. For this case 6 is the best K

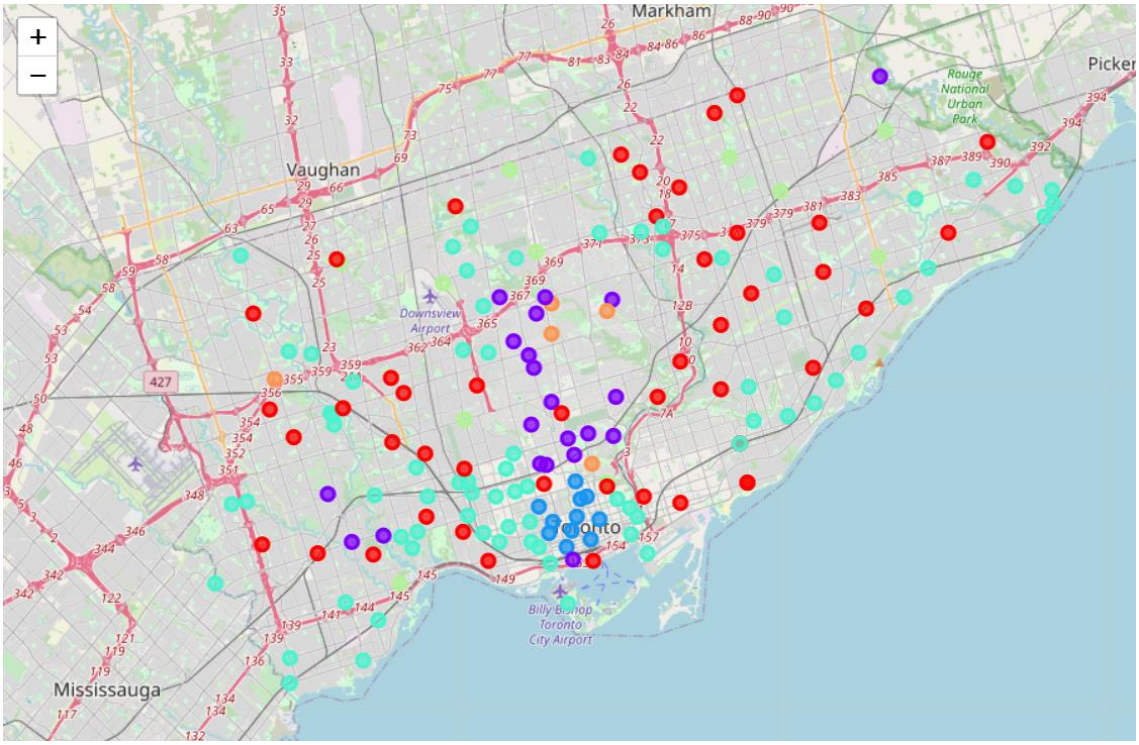


4. Results

The clustering k-means algorithm generates 6 groups of neighborhoods:

- cluster 0 is middle income, high population, no restaurants
- cluster 1 is high income, low population, no restaurants
- cluster 2 is middle or high income, low population, many restaurants
- cluster 3 is middle income, middle population, no restaurants
- cluster 4 is middle income, high population, few restaurants
- cluster 5 is high income, low population, few or no restaurants

They are showed in a folium map:



Cluster 0 is red
Cluster 1 is purple
Cluster 2 is blue
Cluster 3 is light blue
Cluster 4 is green
Cluster 5 is orange

5. Discussion

I used the K-means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 6, but a bigger value could be used in case the number of neighborhoods in a single cluster is too big.

The number of restaurants of every neighborhood has been obtained using a radius of 500 meters from the neighborhood center coordinates. But a different number could be used depending of the client's expertise.

The population of every neighborhood are data of 2012 year, this could be updated to improve the accuracy of the clustering, although given that the differences will be minimum, the results will change very few, if any.

6. Conclusions

In this study, I analyzed the best neighborhoods for opening a new wine store in Toronto based in the criteria for this customer, that is population, average income and number of restaurants. I retrieved the needed data and used it to cluster the neighborhoods to find the best places for the new store. This information can be used for new store openings in the same city.