Clustering Neighborhoods for supporting tourists decisions

Thiago F. C. de Oliveira

July, 2020

Abstract

This is a report of the Applied Data Science Capstone, the last course of the IBM Data Science Professional Certificate. I used data from Foursquare API and transitland to cluster Toronto's neighborhoods in order to help tourists to decide which neighborhood to stay during a vacation, based on venues price, mobilitiby and proximity of the tourist attractions.

1 Introduction

1.1 Backgroud

We all know that planning a trip can be a quite difficult task. You have to look for plane tickets, passports, schedule and for a place to stay. We also know that frequently things can get out of control, even though you did the best travel itinerary. And when you loose control, it's possible that you loose a lot of money and time too. And all that was supposed to be about relaxing, becomes a nightmare.

You may have found the cheapest bedroom in Airbnb, or the most confortable hotel. But before you check in, you realize that is far away from everything. Or that all the restaurantes around are too expensive, or even that there isn't a bus stop around and you'll spent a lot of money in cabs.

But what if you can get all the information about the neighborhood you will be staying? How easy is to get the subway, or how many options would you have to eat if you decide to take a walk? Wouldn't it be great to have all neighborhoods of your destination clustered according to its characteriscs?

The objective of this work is to provide a cluster analysis of Toronto's neighborhoods in order to support tourists decision in choosing a place to stay in a vacation travel.

1.2 Problem

Gathering data from Foursquare API, containing venues in the neighborhood, venues category and also venues price. I'll also get the location of all subway and bus stations and tourist attractions. With all this information, I'll apply k-Means clustering and label neighborhoods according to price, venues variability, mobility and proximity to tourist attractions.

1.3 Interest

This work can give insight to any traveler going to Toronto and can be also be interesting to Travel Agencies, improving costumer satisfaction in choosing a place to stay in a vacation. With the right data, this work can be applied to any city.

2 Data

Three different sources will be required for this project. With all of them gathered together, it will be possible to assemble a model to use in the k-Means clustering algorithm.

First, the list with all postal codes, boroughs and neighborhoodsd of Toronto. This list is avaiable in wikipedia, and the table with all this information will be scraped with BeautifulSoup library.

For the use of Foursquare API, it's also requiered latitude and longitude data. This can be done with GeoPy, but during the course a list with latitude and longitude was provided, so I'll use it.

After this, data will be acquiered by the Foursquare API. I'll get all venues related with tourism activities: restaurants, malls, tourist attractions, beachs and parks. I'll also use premium calls in the Foursquare API to retrieve venues prices.

Foursquare Categories can often be too specific. To better cluster the neighborhoods, I'll use a hierarchy of categories, provided by Foursquare in here. So, each category will be mapped in one of the ten hierarchies: Arts Entertainment, College University, Event, Food, Nightlife Spot, Outdoors Recreation, Professional Other Places, Residence, Shop Service and Travel Transport.

Finally, I'll use transit.land database to get all subways and bus station, in order to measure the mobility of each neighborhood.

3 Exploratory Analysis

With all the neighborhood and postal codes data, scraped in the wikipedia page, and the geographic data, provided by Coursera, it's possible to check all the nighborhoods analyzed in this work in the Figure 1

After all the information required is obtained, we can begin to ask some questions to the data. The first one is: how is the variability of categories? Figure 2 answer this question:

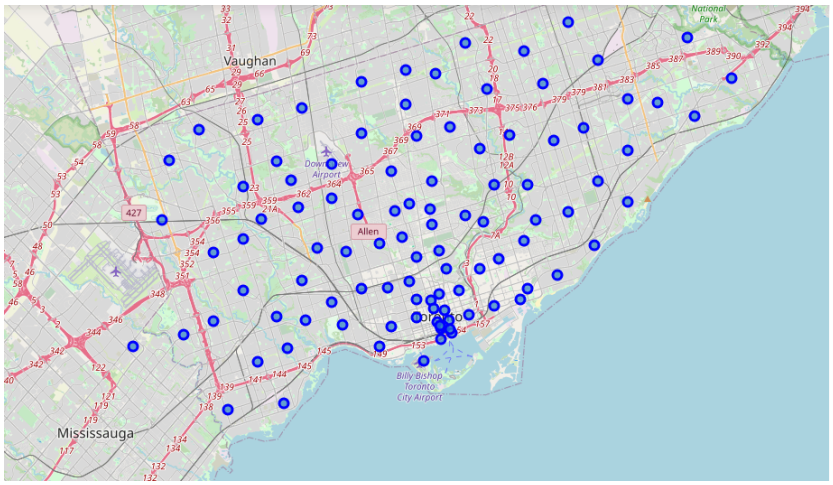

Figure 1: Toronto's neighborhood studied in this work.

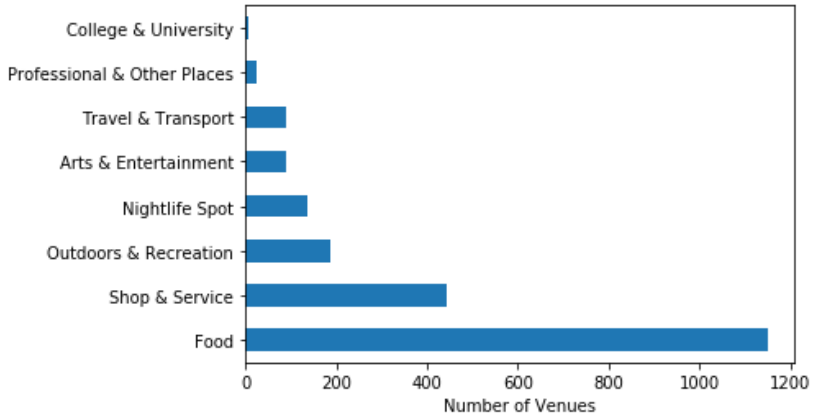

Figure 2: Number of venues of each category.

We can see that most of Foursquare venues in Toronto are food related. So, eating will be no problem while staying there. The second most frequent venue is Shop Service, almost one third of the Food category venues. The is also a relevant number of Outdoors Recreation Venues, so, if you are a tourist that likes Nature and Camping, Toronto can be a good call.

Well, after checking the venues categories distribution, we can now begin to understand the prices in this city. Food prices are retrieved from a premium call in the Foursquare API, and, unfortunatelly, it's possible to make only 500 calls per day. Due this situation, I just got the prices to food related venues.

The price is given in a scale from 0 to 4, where 0 representes a cheap restaurant, and 4 a very expensive one. The histogram with the prices mean in each neighborhood is show in the Figure 3:

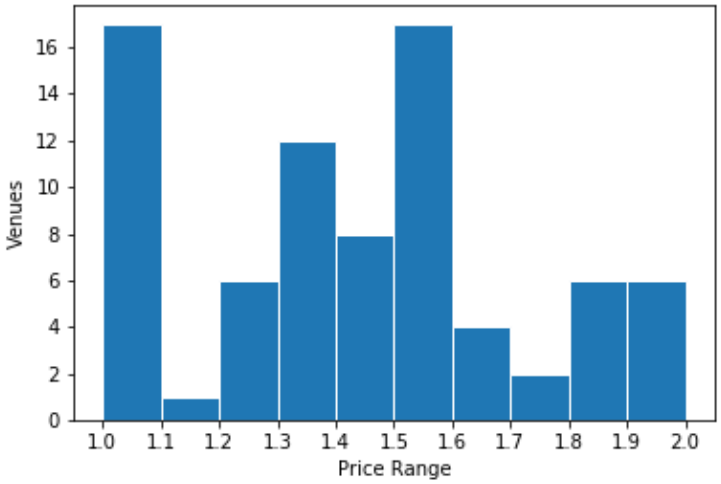

Figure 3: Price distribution in Toronto

The first conclusion we can get is that the price in Toronto is definitely not so high. There are no neighborhood with mean price bigger than 2, and the most commum prices are between 1-1.1 and 1.5-1.6. So, if you don't have much money to travel, Toronto is a very good option when it comes to food prices.

We can, finally, try to understand the mobility in Toronto. The data retrieved from transit.land got all the bus and subway station within a radius of 1 km within each neighborhood geolocation. As expected, there are more data related to bus station. Figure 4 shows the top 10 neighborhoods when it comes to bus station:

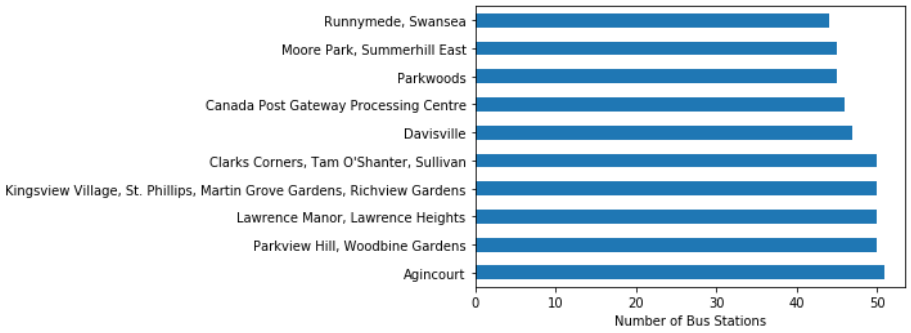

Figure 4: Top 10 neighborhoods in number of bus stations.

Figure 5 shows the top 5 neighborhoods in number of subway stations:

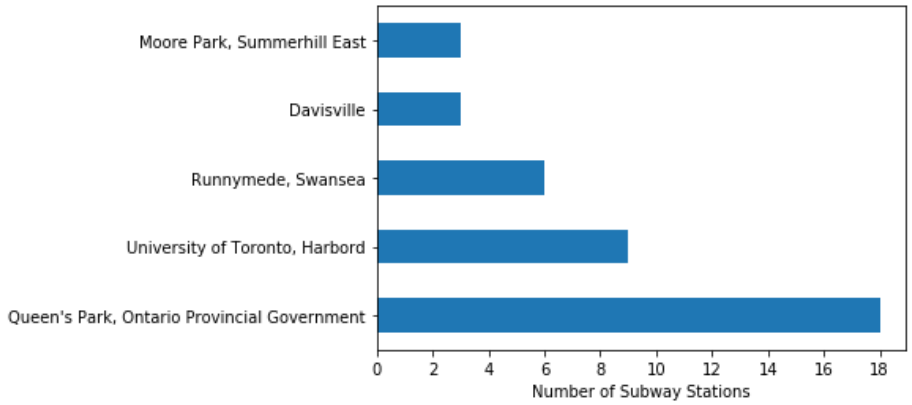

Figure 5: Top 5 neighborhoods in number of subway stations.

So, if mobility is really important and you don't want to spend money with cabs, you must check this neighborhoods.

4 Clustering Toronto's Neighborhoods

Before clustering using k-Means clustering, I normalized all the data. The figure 6 shows the final DataFrame. The features selected were the frequency of each category in each neighborhood, the normalized number of of bus and subway stations and mean price range in each neighborhood.

	Borough	Neighborhood	#_Bus	#_Subway	price	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
0	North York	Parkwoods	0.882353	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.500000	0.000000
1	North York	Victoria Village	0.823529	0.0	1.000000	0.250000	0.000000	0.750000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Downtown Toronto	Regent Park, Harbourfront	0.431373	0.0	0.700000	0.111111	0.000000	0.444444	0.044444	0.111111	0.044444	0.222222	0.022222
3	North York	Lawrence Manor, Lawrence Heights	0.980392	0.0	0.750000	0.000000	0.000000	0.117647	0.000000	0.058824	0.058824	0.764706	0.000000
4	Downtown Toronto	Queen's Park, Ontario Provincial Government	0.529412	1.0	0.600000	0.060606	0.030303	0.606061	0.060606	0.090909	0.030303	0.121212	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...
93	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
94	Downtown Toronto	Church and Wellesley	0.000000	0.0	0.756098	0.041096	0.000000	0.561644	0.109589	0.095890	0.013699	0.136986	0.041096
95	East Toronto	Business reply mail Processing Centre, South C...	0.000000	0.0	0.750000	0.000000	0.000000	0.235294	0.058824	0.235294	0.000000	0.352941	0.117647

Figure 6: DataFrame used in the model.

After the feature selection, the next step was the choice of the right k. The method chosen for this task was the elbow method. It consists of plotting the within cluster mean squared distance for a range of k. Then, the right k is the one where the elbow of the curve is.

Figure 7 shows clearly that the right k is k=3.

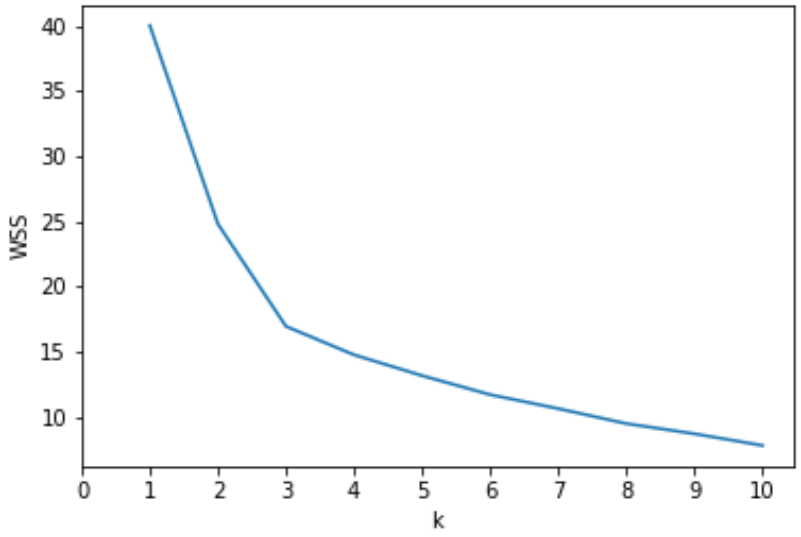

Figure 7: The elbow method, used to choose the right k.

5 Results and Discussion

With k=3, the result of the clustering is shown in the Figure 9:

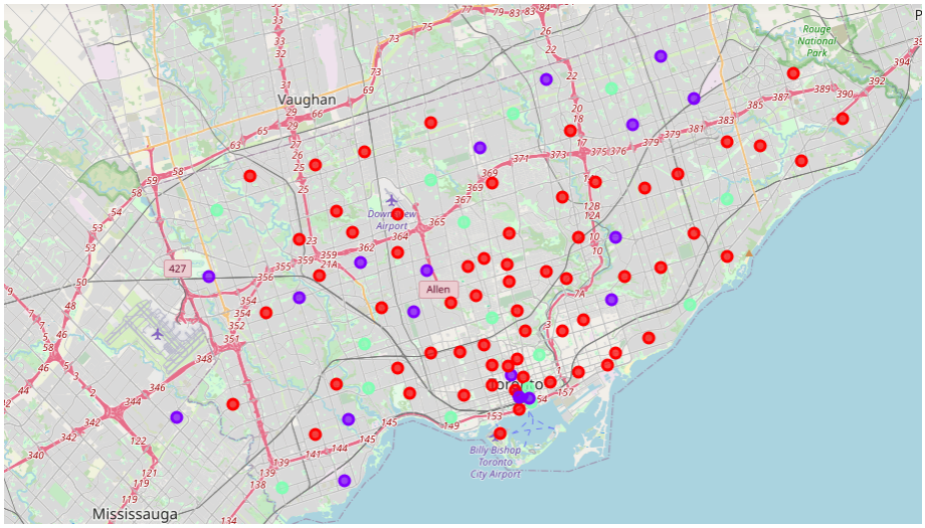

Figure 8: Toronto's neighborhoods clustered.

The three clusters labels are: Red, Purple and Blue. The first comment that can be done, looking at the map, is the high number of Red neighborhoods, spread all over the city.

Table 1 shows the number of each cluster's occurrence:

Table 1: Clusters occurrence in Toronto.

Cluster	Occurrence
Red	63
Purple	20
Blue	15

The majority of neighborhoods are Red. The number of Purple and Blue are almost the same. Another way to get deeper insights about each cluster characteriscs is by grouping all the results of each one and studying the mean or mode of each one. We can see this in Figure **??**:

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	#_Bus	#_Subway	price
Red	Food	Shop & Service	Outdoors & Recreation	0.148459	0.030864	0.560112
Purple	Food	Shop & Service	Travel & Transport	0.225490	0.000000	0.604491
Blue	Food	Shop & Service	['Nightlife Spot' 'Travel & Transport']	0.057516	0.022222	0.582284

Figure 9: Clusters characteristcs.

About venues, we can see that 1st and 2nd most commom venues are Food and Shop Services. The difference comes in the 3rd most commum venues. For Red Cluster, it's Outdoors and Recreation. As for Purple Cluster, it's Travel Transport. Finally, for Blue Cluster, we have a tie: Travel Transport and Nightlife Spot.

Regarding the mobility, this numbers represent the mean for each cluster of the normalized number of bus and subway stations. For the Purple Cluster, we have the highest number of bus stations, but no subway stations at all. The Red Cluster is the second in bus stations options, and the first in number of subway stations. As for the Blue Cluster, there are few options of both subway and bus stations.

Finally, we can analyze the price. The most expensive cluster is the Purple one. It's good to keep in mind that this Cluster don't have subway stations, so it can become even more expensive. The second when it comes to price is the Blue Cluster. And, the cheapest one is the Red Cluster.

6 Conclusion

In this study, I analyzed the neighborhoods in Toronto, and clustered them according to variability of venues in each neighborhood, mobility and venues prices. This classification provide insight to tourists. When choosing a place to stay on Airbnb, or book a hotel, it is possible to know a priori the characteristics of the neighborhood, and take this into account when planning your trip.

If you are a young tourist who likes to go out at night, perhaps the Blue Cluster is a good option. However, if you don't have that much money, taking advantage of Red Cluster's low prices and mobility options will help keep you within budget.