Capstone Project
Neighbourhood Segmentation and Clustering using Foursquare API.

João Vitor Padilha Ramos

Setember 05, 2020

1. Introduction

1.1 Background
There are people all over the world who want to undertake. Whether because they want to remedy an existing pain in the market or simply because they make money where no one saw an opportunity. Based on that, I made an analysis where I show what would be the best region to put a Coffee Shop in the city of Toronto. And why Coffee Shop? Because Canada is among the 10 largest coffee-consuming countries in the world. And why Toronto? Because it is the most populous city in Canada. That said, let's understand better about the problem.

1.2 Problem
The objective of this project is to analyze and select the best locations in the city of Toronto, for the opening of a new Coffe Shop. This project is mainly focused on geospatial analysis of the city of Toronto to understand what would be the best place to open a new store. Using data science methodology and machine learning techniques such as clustering, this project aims to provide solutions to answer the business question: In the city of Toronto, if an entrepreneur is looking to open a new Coffee Shop, where would you recommend that he open?

1.3 Interest
This project is aimed at all people who are thinking of undertaking or who want to learn about a Fourquare API tool, which can assist us in analysis for different types of problems.

2. Data acquisition and cleaning

2.1 Data sources
I did a web scraping of a bit of data taken from a wikipedia site, where it contains the following information:
- Postal Code
- Borough
- Neighborhood

```
[6]: df_clean.head()
```

[6]:

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Follow the link below the source of the data used:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2.2 Data cleaning

I removed all lines with Borough = Not assigned from our database.

2.3 Adding coordinates

I added the geographic coordinates (Latitude and Longitude) to my database to facilitate our analysis and use of the K-Means algorithm later on.

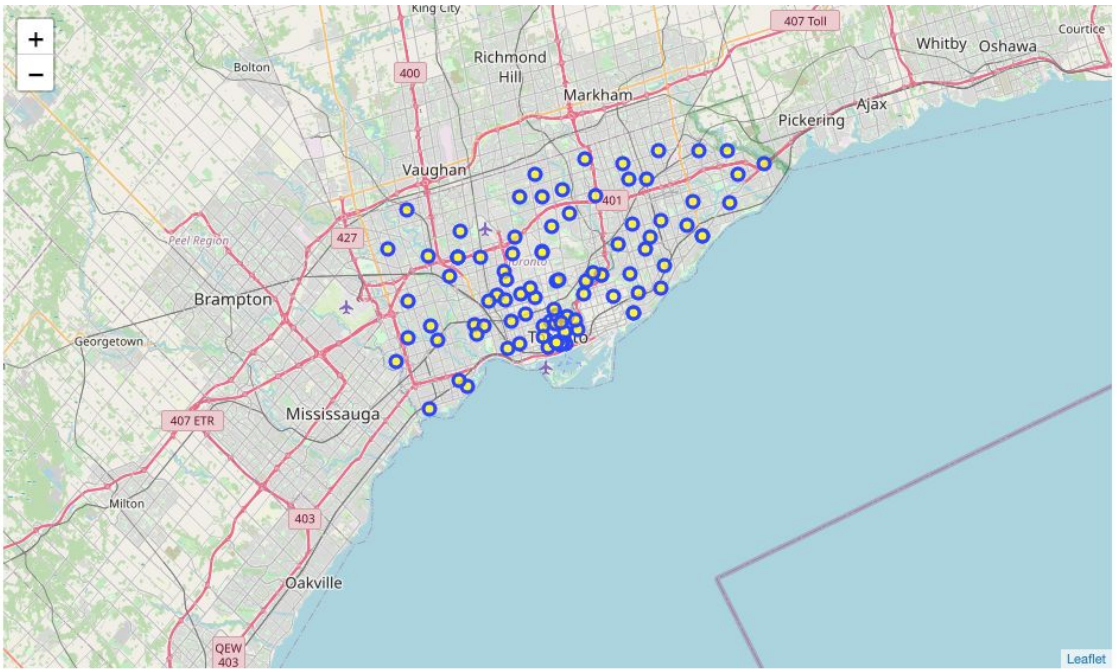
```
[11]: # check the neighborhoods and the coordinates
print(df_clean.shape)
df_clean.head()
```

```
(103, 5)
```

[11]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.686588	-79.409996
1	M4A	North York	Victoria Village	43.731540	-79.314280
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.659743	-79.361561
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.723570	-79.437110
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.666622	-79.393264

[14]:

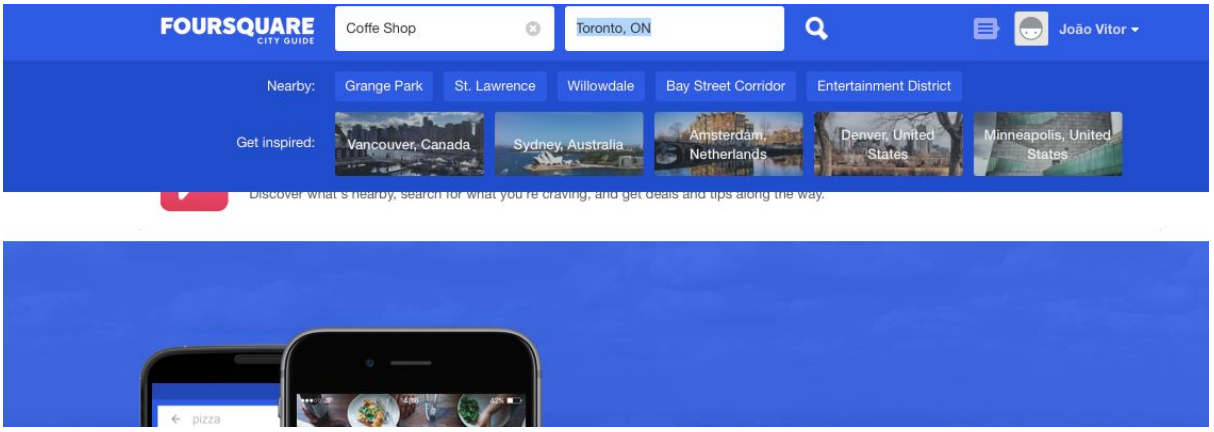


3. Exploratory Data Analysis

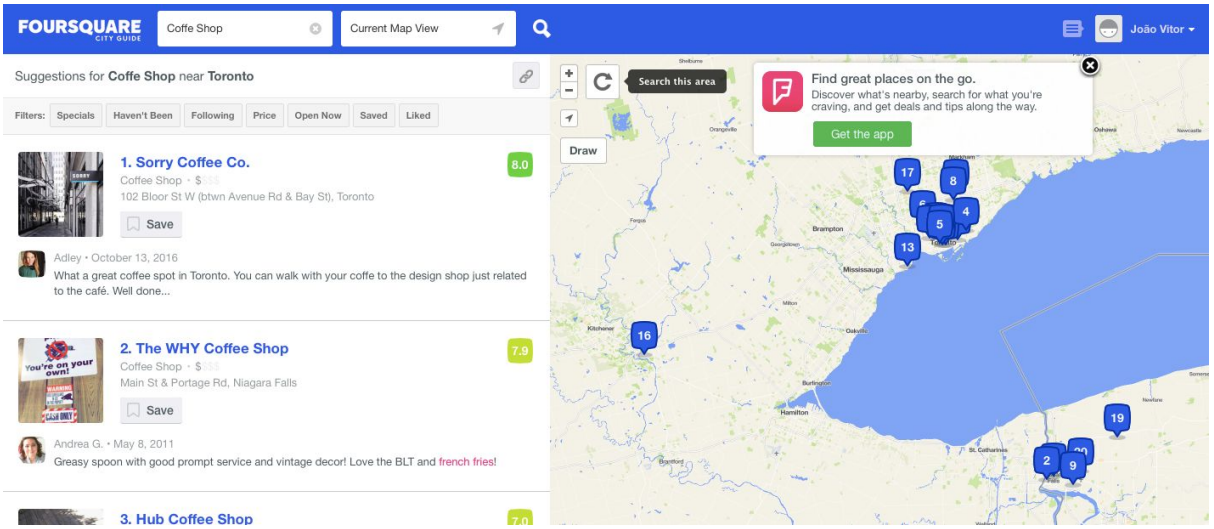
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103 entries, 0 to 102
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Postal Code    103 non-null    object 
 1   Borough        103 non-null    object 
 2   Neighbourhood  103 non-null    object 
 3   Latitude       103 non-null    float64
 4   Longitude      103 non-null    float64
dtypes: float64(2), object(3)
memory usage: 4.1+ KB
```

4. Methodology

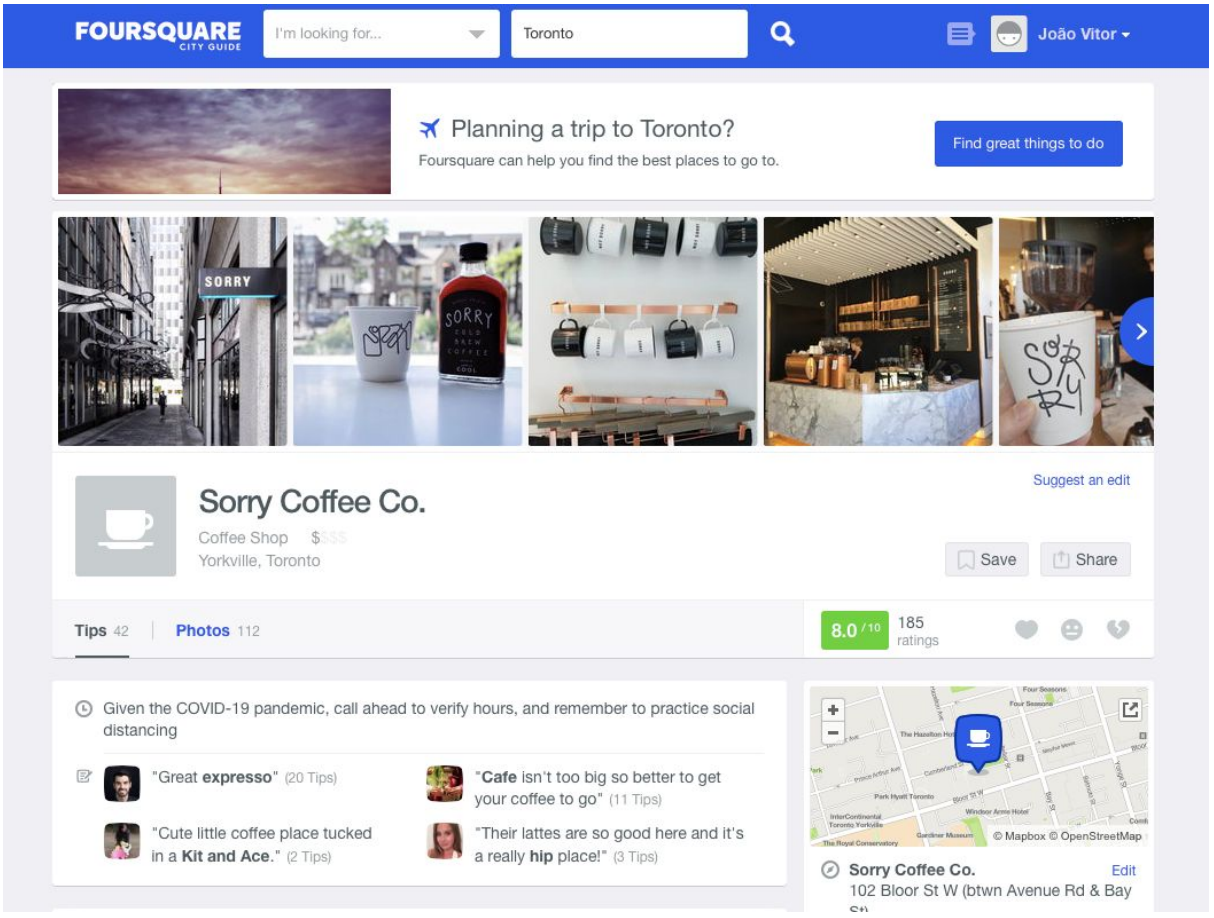
The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there is really nothing to download from your server.



On the left, you see all the Coffee Shops and their names, category and address for each location in Toronto. On the right, you see the map of the places on the left.



If you click on any Coffee Shop, you will be redirected to this page, where you will see all the information in the Foursquare dataset about the chosen element. This includes name, full address, opening hours, tips and images that users have posted about the establishment. In the same way, you can explore the other places in Toronto or anywhere in the world.



To explore Foursquare click on the link below:
https://foursquare.com/download

5. Model

We use the clustering k-means which is a Clustering method that aims to partition n observations among k groups, where each observation belongs to the group closest to the average. This results in a division of the data space in a Voronoi Diagram.

6. Results

There are 37 Coffee Shops in cluster 0 and this cluster contains all the neighborhoods that have the lowest amount of this type of establishment. Cluster 1 contains 48 Coffee Shops, the cluster that contains the largest number of neighborhoods, but with an intermediate quantity of Coffee Shops. Cluster 2 contains 18 stores, the cluster with the fewest neighborhoods, but with the highest concentration of Coffee Shops.

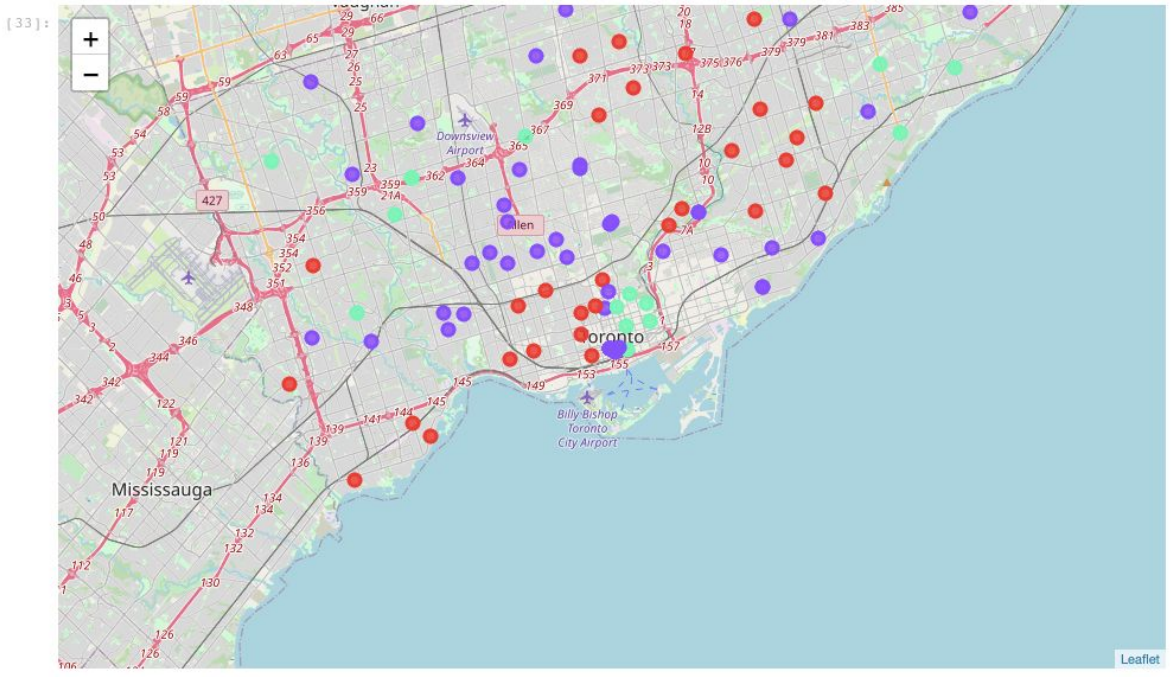
The results of the K-means cluster show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

• Cluster 0: Neighborhoods with a much smaller number of stores.

• Cluster 1: Neighborhoods with a moderate concentration of stores.

• Cluster 2: Neighborhoods with a high concentration of stores.

We see the results of the grouping on the map with cluster 0 in red, cluster 1 in purple and cluster 2 in green.



7. Conclusions

A good number of Coffee Shops are concentrated in the downtown area of Toronto. Cluster 0 has a very low number of stores. This represents a great opportunity and areas of high potential for the opening of new stores, since there is little or no competition. Meanwhile, Cluster 2 Coffee Shops are probably suffering intense competition due to the excess supply and the high concentration of stores in the same segment. Analyzing only these points presented, we conclude that, for those interested in opening a Coffee Shop in Toronto, we strongly suggest the opening of new stores in the regions related to cluster 0 neighborhoods, where there is little or no competition.

We believe that competition between cluster 1 neighborhoods is moderate and that there is still room for new stores, but there is a greater risk of competition. Finally, entrepreneurs should avoid cluster 2 neighborhoods, which already have a high concentration of Coffe Shops and suffer intense competition. (However we can emphasize that regardless of the competition being high, if the service provided is excellent, the quality of the product is good and you have a differential, there is no competition that can resist a good deal).

We can apply the rationale of this project to several other areas or problems and what was presented is just an example of what we can do with this tool. In this project, we consider only one factor, that is, the frequency of occurrence of shopping centers, there are other factors such as population and income of residents that can influence the decision to locate a new Coffe Shop.