**Identifying the best location to open a coffee shop in Toronto**

Vinicius Garcia

November 16, 2020

# Introduction

## Background

Location Data is data used to describe places and venues. This data contains many of the place’s attributes, such as geographical location, full address, working hours, category and so on, given its latitude and longitude values. It also identifies what types of venues are in a certain radius from that place, which means it can tell if there are hotels, restaurants or parks nearby. This Project aims to use Location Data to assist people deciding where is the best place to open a new coffee shop in the neighborhoods of Toronto.

## Problem

Finding the best location and opportunity is one of the key factors that can lead to success when it comes to coffee shop business. The neighborhood should have venues that people visit in the morning, so it is more likely for them to go to a coffee shop. However, if there is heavy competition from existing coffee shops, it will be more difficult for the business to turn out successful.

# Data acquisition and cleaning

## Data sources

For this project, a dataset containing all of Canada’s neighborhoods’ postal code and borough was acquired using a method of web scraping in a Wikipedia page (<https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=1012023397>) and only Toronto neighborhoods were selected. In order to get information about the venues around those neighborhoods, the Foursquare API was extremely important. It provides, given some latitude and longitude values, data such as what venues are there around that neighborhood, their exact locations and categories. As the API requires latitude and longitude values, it was necessary to merge the first dataset from Wikipedia with a dataset containing postal codes and their respective geospatial data.

## Data cleaning

From the Wikipedia’s dataframe, all the cells that don’t have an assigned borough were excluded. Also, all the “slashes” separations were changed to commas. In order to merge this dataframe with the geospatial one, it was necessary to check if the merge variable “Postal Code” was in the same format in both datasets and if it wasn’t, to standardize it. After the merging, the final dataframe was reduced to work only with Toronto neighborhoods.

## Interest

This project is destined to those who want to start a new coffee shop business in Toronto. Anyone that wants to learn about data science and location data can also be interested.

## Feature selection

# Methodology

## Foursquare access

In order to access the Foursquare API, a GET request was sent utilizing as parameters the API credentials and version, Toronto latitude and longitude values, the radius which the project will encompass and the limit of venues wanted. As a response to this request, a JSON was returned containing data about all venues in the location specified. However, the JSON format is not very suitable to work with when analysing data. Therefore, a dataframe was made out of it. The dataframe contained all the main data needed to the project, like the name of the venue, its geospatial data and to which category it belongs (restaurant, park, bakery etc.).

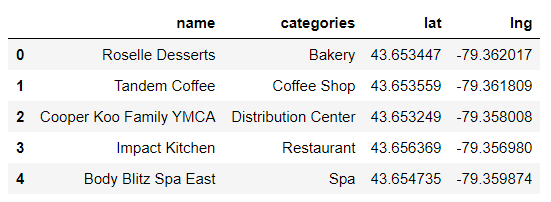


Figure 1. Sample of the dataframe generated from the JSON file.

## Exploring neighborhoods

Now that data about the venues was acquired, the next step was to merge it with the data we already had from the neighborhoods in Toronto. This produced a dataframe ready for the exploratory analysis.

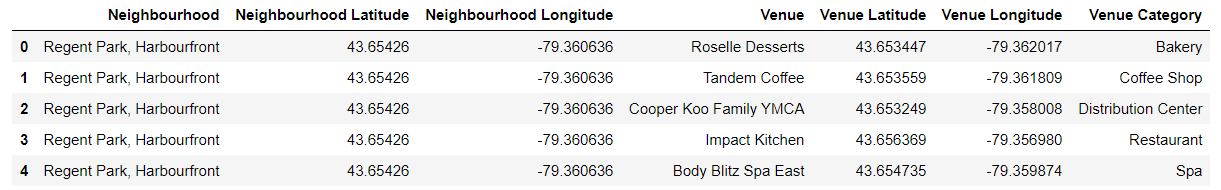


Figure 2. Sample of the dataframe with venues data including to which neighborhood it belongs.

After that, this dataframe was grouped by neighborhood to see how many venues registers were there to each neighborhood.

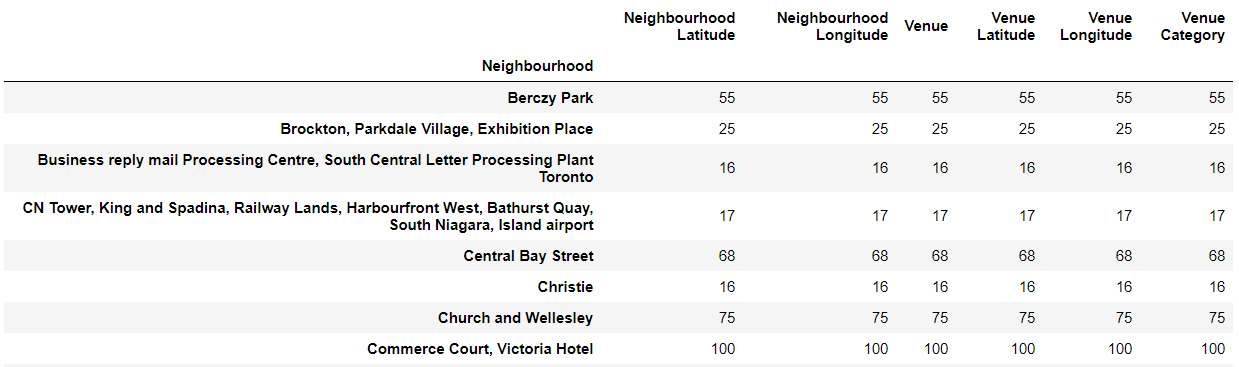


Figure 3. Sample of the grouped dataframe.

The next objective was to get the 5 most common venues to each neighborhood. In order to do that, a one-hot encoded dataframe was generated (transforming the categorical values in numerical ones) and it was grouped in a way to show the mean values of each venues inside each neighborhood register. That way, it was possible to see which venue had the highest mean value, that is, which are the most common venues to each neighborhood. That new info was used to generate a new dataframe in which each register was a neighborhood and the variables were the 5 most common venues regarding that neighborhood.

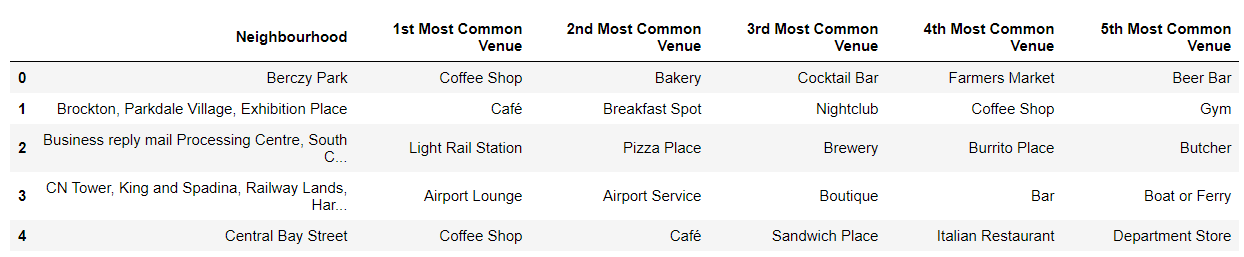


Figure 4. Sample of the dataframe containing the 5 most common venues in each neighborhood.

## Clustering the neighborhoods

In the interest of making it easier to choose a neighborhood to open a coffee shop, a machine learning technique called *clustering* was the choice for this project. Clustering is an unsupervised task that groups data in a natural way. The algorithm interprets the data and finds a way to group the similar points based on the features. One of the most used clustering algorithms is K-Means. It is relatively simple to implement and guarantees convergence, which makes it a very suitable choice for the problem we are trying to solve. With the purpose of initializing K-Means with an optimal K value, the Elbow Method was used. It consists in testing de algorithm with many different values for K and generating a graph that shows in which value the distortion stops changing significantly. After the number of clusters was determined, the algorithm assigned each neighborhood of the dataframe to a cluster. Now it was necessary to analyse each cluster and decide if there were adjustments to be made and which cluster is the best choice to open a coffee shop.

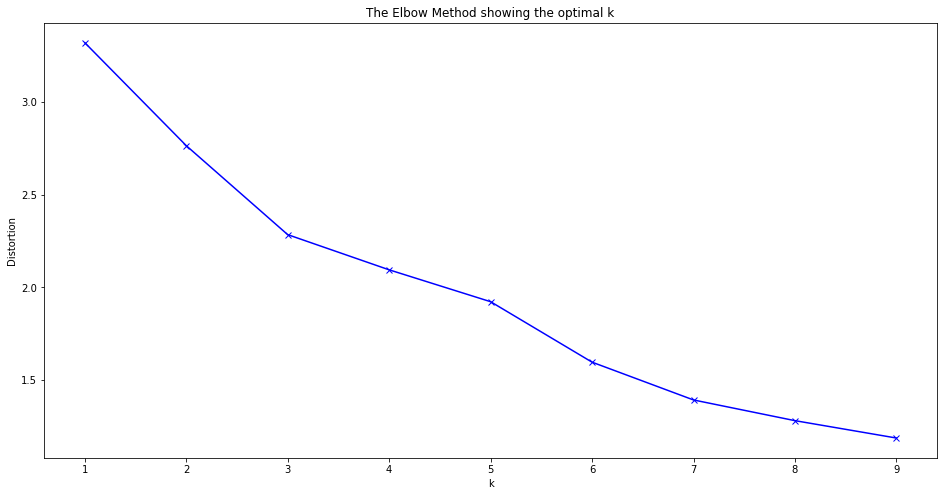
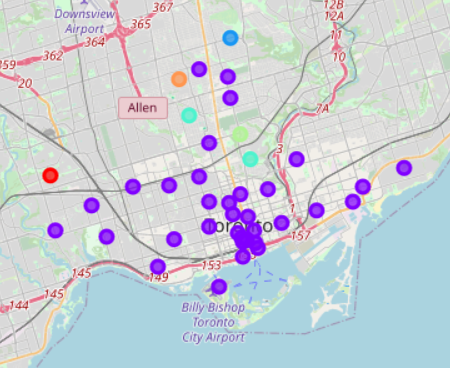


Figure 4. Elbow method.

# Results

The best number of clusters found was 6. The biggest cluster had 10 neighborhoods in it, while the second one had 2 and all the others had only 1 register. A folium map was generated to visualize the clusters.



# Discussion

As there were many clusters with just one neighborhood, it was needed to cluster some of those manually, analyzing the similarities among them. Based on the most common venues of each neighborhood and their functionalities and target audience, some smaller clusters were assigned to bigger ones. For example, in clusters 0, 3 and 4, there were trails and parks appearing more, as long as food venues. Cluster 2 (a bigger cluster) followed the same pattern so it involved the other clusters.

# Conclusion

After the clustering, it was clear that Final Cluster 2 had neighborhoods in which morning entertainment venues (parks, trails and playgrounds) are more common while restaurants are still present. Final Cluster 0, differs from the others but doesn't have anything very interesting to coffee shops' clients. As for Final Cluster 1, food venues like restaurants and coffee shops were more present. As it already has many coffee shops, it would be more difficult to deal with the competition. Final Cluster 2 doesn’t have many coffee shops and in addition, it is more likely to people who go to morning entertainment venues to be drawn to a coffee shop. In conclusion, cluster 0 was chosen as the cluster in which opening a coffee shop is more likely to succeed. The neighborhoods in that cluster are **Lawrence Park; Forest Hill North & West, Forest Hill Road Park; Moore Park, Summerhill East and Rosedale**.