Capstone Project

THE BATTLE OF NEIGHBOURHOODS

April 3rd, 2021

Table of Contents

1. Introduction

This project is the final assignment of the IBM Professional Data science Certificate. This certificate consists of a series of 10 courses in which data science skills, including Data Science Methodology, data mining and analysis with python has been studied and applied. The main objectives for this project are the following:

- Leverage location data provided by Foursquare and the Data of Amsterdam website
- Applying data science skills in machine learning and data visualization

The results will be presented in this report as well as in a presentation and a Jupyter notebook published on Github.

1.1 Business Problem:

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571, it is the most populous city in Canada and the fourth most populous city in North America. The way we dine out has been transformed in the last decade with independent cafes and restaurants opening on every corner and food stalls booming throughout multiple cities. We want to support our stakeholders in finding the optimal location within the city where to open a new restaurant. It is not so easy to decide on a venue for such a business: Beyond providing great food, wine, and service, your job is to find a place where food businesses start as an idea and quickly grow into a brick and mortar venue. We may presume that the majority of the existing restaurants are concentrated in the city centre and most touristic venues, to benefit from the influx of short-term visitors. It may then be a good option to look for venues with fewer existing restaurants, so not to be hindered by excessive competition, so as to attract a clientele of national residents that are accustomed to varieties of cuisines and may build up a base of regular and affectionate clients. We will then be working with the 39 Toronto neighbourhoods and use the “Foursquare API” to build a venue data analysis and subsequently to cluster the various locations in order to identify the most promising places for their business.

1.2 Audience and Stakeholders

Because of the business problem and the approach in using machine learnings this project could be potentially interesting to the following groups

- Entrepreneurs who want to start a food business in Toronto
- Brokers to advise people on buying properties based on their business plan

The insights found in this project can be of great help for people wanting to start their own restaurant in Toronto. Furthermore, brokers helping their clients in finding available venue and advise them on the neighbourhoods.

2. Data

To find an answer to the business problem we will make use of the following two data sources:

- Geographical data provided by the Canada Data Website
- Location data provided by Foursquare

In this chapter these sources are further introduced and explained.

2.1 Location Data neighbourhoods

The Canada data website (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains postal code and list of the Canadian boroughs. The data is extracted using the BeautifulSoup library and stored in a dataframe.

2.2 Venue Data: The foursquare database

The second source which is used in this project is the Foursquare database. Foursquare (https://foursquare.com) is a company which build and maintains a massive dataset of accurate location data. This data is freely available using the RESTful API. After registration on their website on can do a limited number of queries with different search parameters so called “endpoints” and collect a dataset containing all the interesting venues a location has to offer. Figure 3 shows what a RESTfull API url would looks like:

```
# create the API request URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.
```

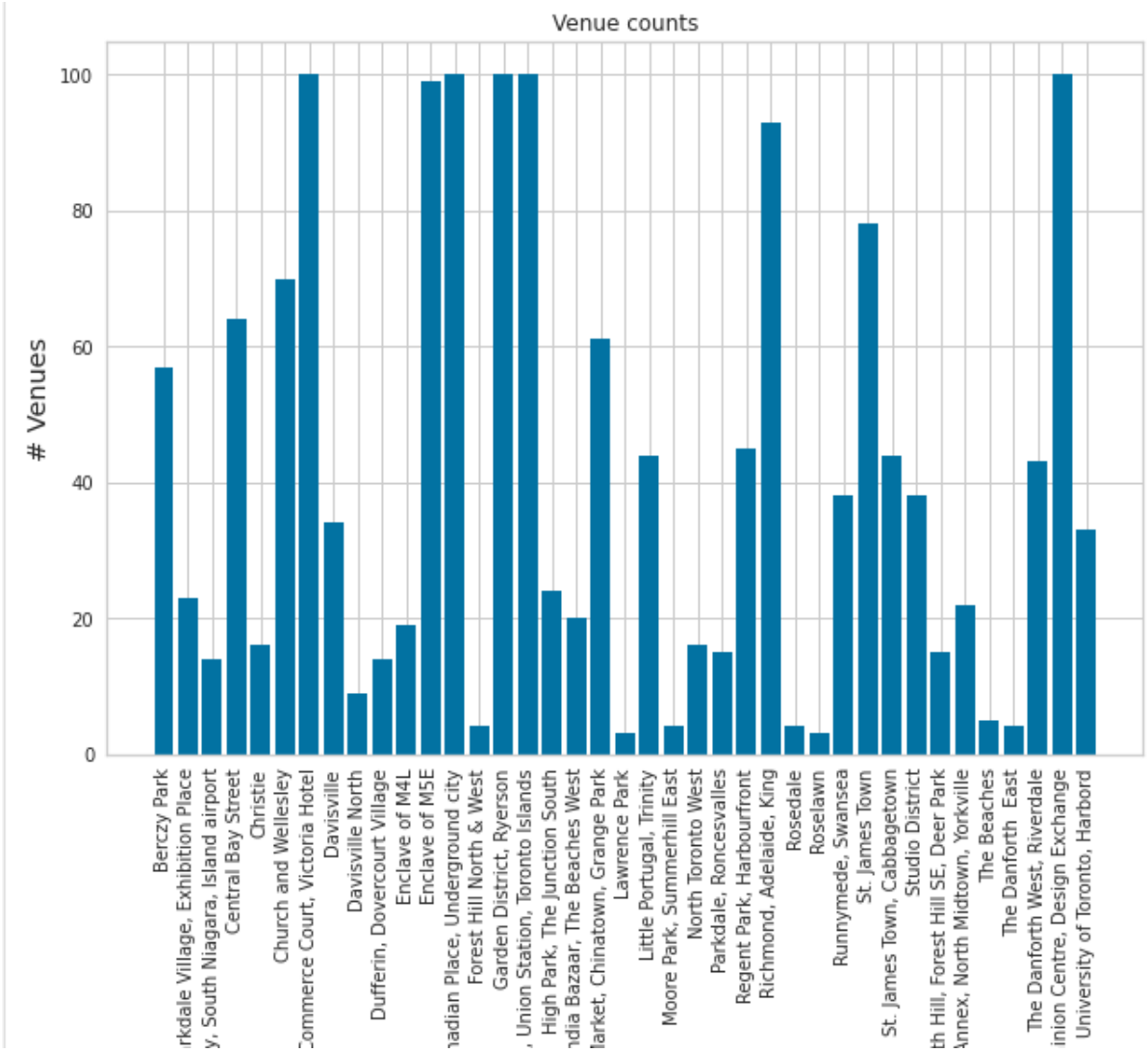
As show in the above figure, one of the limitations of this API is the search method “radius” around a location of interest. Since neighbourhoods in cities like Toronto tends to have irregular forms and sizes there is a need for a more specific definition of the search area. If the radius has been defined to small, interesting venues could be missed in the neighbourhood. On the other side, choosing a to large radius will result in duplicate venues and venues assigned to the wrong neighbourhood. Therefore, we have chosen a radius of 500.

3. Methodology

This chapter will describe the methods used for analysis, statistical testing of the machine learning model.

3.1 Data Mining

The below graph shows the line graph between the most popular venues and neighbourhood. This graph explains us that the neighbourhood containing maximum popular venues may be a good place to open a restaurant.



After the data has been acquired from the foursquare database using a radius of 500. The data has been cleaned from duplicates, pre-processed and put into a test data frame.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Cheese Shop	Pharmacy
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Convenience Store	Italian Restaurant
2	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Harbor / Marina	Airport Food Court	Airport Gate
3	Central Bay Street	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Burger Joint
4	Christie	Grocery Store	Café	Park	Coffee Shop	Nightclub

4. Machine learning: k-means clustering

The machine learning technique applied in this project is K-Means Clustering. This chapter contains a brief explanation what K-Means Clustering is and why is has been applied. Furthermore, the specific application on the dataset will be discussed as well as parameter choice and the validation of the model.

4.1 What is k-means and why is the method of choice?

K-Means Clustering is a machine learning technique for calculating the similarity and dissimilarity of points in a given dataset. It is a form of unsupervised machine learning being capable to handle unlabelled data which is the case of our feature set. By calculating the distances between data points clusters are formed by two optimizations.

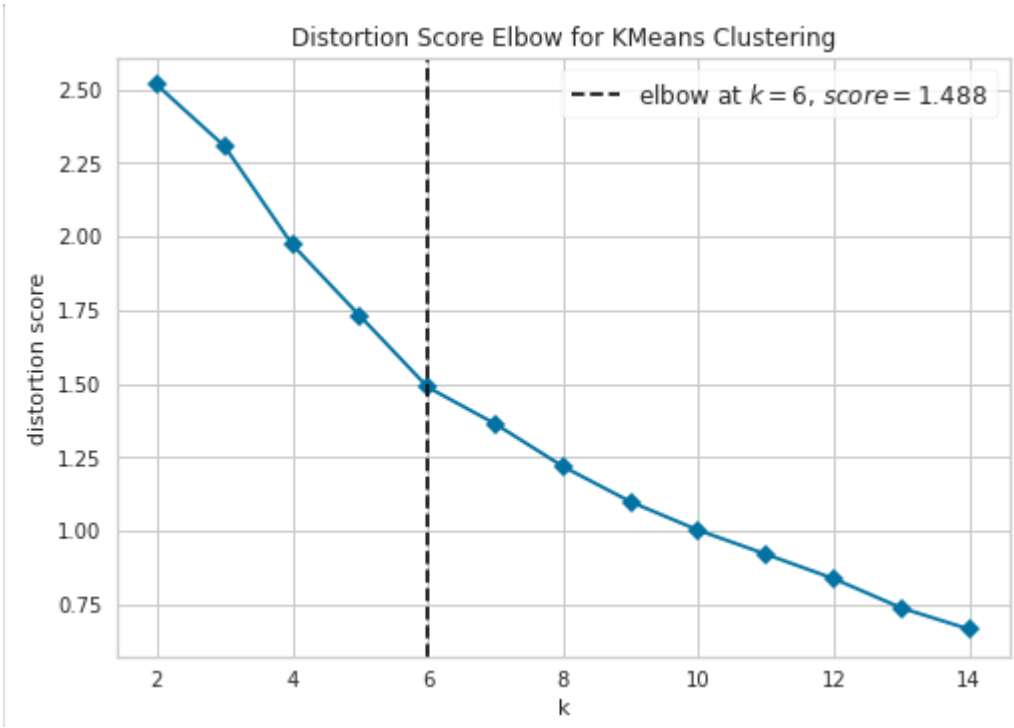
Create clusters where distance of points within the clusters are minimized and the distance between de clusters is maximized. Since we are trying to separate the neighbourhood based on unlabelled features and finding and those similar neighbourhoods which will have the ideal feature values for starting a restaurant.

The most critical part of the K-Means clustering is determining the best number of k. K naturally will be within the following to extreme cases When K is chosen 1 than there will be only one cluster containing all datapoints.

Cluster distance is maximized (no other cluster available) but there is no separation of data points. When K is chosen very large (for example 99 as the number of neighbourhoods), every cluster might contain 1 datapoint. This is also not useful since in this case there is no clustering at all. To find an optimal value of K an iterative approach has been chosen. The clustering algorithm has been applied on the data set using different values of k and 'quality' of the clustering examined using the commonly used Elbow method. in the next chapter the results of these examinations will be shown and discussed.

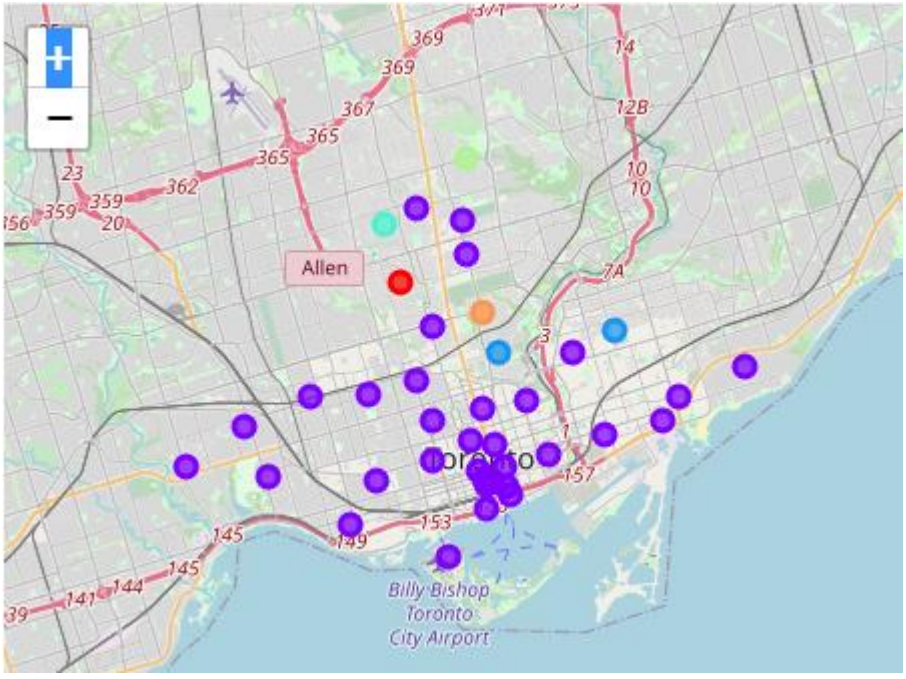
4.1.1 Elbow Method

In the elbow method multiple runs of the K-means algorithm are executed over a range of values for K. In our case the range for k is 1-10, which is commonly used. For each run the inertia score is calculated on the datapoints after clustering. Inertial is the sum of the distances of data points to their closest cluster centre. An increasing number of K will result in a decrease in Inertia. In the below figure, the calculated inertia is plot against the number of clusters. There is a slight elbow change of steepness visible at k=6, which is the commonly advised optimal point for K.



5. Results

This chapter covers the results achieved in this project. In below figure, the neighbourhoods are plotted into a folium map. The colours of the marker points are corresponding with the calculated labels of the clustering



Clusters are assigned by following colors:

- Red: Cluster 0, containing 1 neighbourhood

- Purple: Cluster 1, containing 33 neighbourhoods
- Dark blue: Cluster 2, containing 2 neighbourhoods
- Light blue: Cluster 3, containing 1 neighbourhood
- Light Purple: Cluster 4, containing 1 neighbourhood
- Orange: Cluster 5, containing 1 neighbourhood

Now we have visualized the clusters in a map, the cluster data has to be reviewed in more detail to address the characteristics and form conclusions about the possibilities to start a restaurant.

Cluster 0

These are neighbourhoods mainly located outside of the touristic city center. and therefore, less interesting for starting a restaurant for tourist and day visitors. Could be a good fit for starting a restaurant for locals

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
21	Central Toronto	2	0	Mexican Restaurant	Trail	Jewelry Store	Sushi Restaurant	Yoga Studio

Cluster 1

The most popular neighbourhoods of the city. Plenty entertainment venues and fairly saturated with restaurants! Hard areas to start new business and therefore not advisable.

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Downtown Toronto	1	1	Coffee Shop	Pub	Park	Bakery	Breakfast Spot
1	Downtown Toronto	1	1	Clothing Store	Coffee Shop	Italian Restaurant	Middle Eastern Restaurant	Cosmetics Shop
2	Downtown Toronto	1	1	Café	Coffee Shop	Cosmetics Shop	Cocktail Bar	Gastropub
3	East Toronto	4	1	Pub	Health Food Store	Pizza Place	Trail	Yoga Studio
4	Downtown Toronto	1	1	Coffee Shop	Cocktail Bar	Bakery	Cheese Shop	Pharmacy
5	Downtown Toronto	1	1	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Burger Joint
6	Downtown Toronto	1	1	Grocery Store	Café	Park	Coffee Shop	Nightclub

Cluster 2

These are neighbourhoods mainly surrounded by parks and therefore, could be a good fit for starting a restaurant for locals

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	East York/East Toronto	East York/East Toronto	2	Park	Convenience Store	Metro Station	Yoga Studio	Museum
33	Downtown Toronto	1	2	Park	Playground	Trail	Yoga Studio	Movie Theater

Cluster 3

Neighbourhoods outside of the city centre and low reachability by public transport. Low restaurant counts but unpopular area for going out, therefore not advisable

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
19	Central Toronto	2	3	Garden	Fast Food Restaurant	Pool	Yoga Studio	Mediterranean Restaurant

Cluster 4

Neighbourhoods outside of the city centre and low reachability by public transport. Low restaurant counts but unpopular area for going out, therefore not advisable.

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
18	Central Toronto	2	4	Park	Bus Line	Swim School	Yoga Studio	Movie Theater

Cluster 5

Neighbourhoods outside of the city centre and low reachability by public transport. Low restaurant counts but unpopular area for going out, therefore not advisable

	Borough	Label	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
29	Central Toronto	2	5	Restaurant	Lawyer	Trail	Tennis Court	Movie Theater

6. Conclusion

During the final stage of the IBM Data Science certificate course machine learning has proven to be an effective tool for discovering insights from large amounts of data. The Foursquare database is an interesting data source but has its limitations when it comes to data precision. This precision can be enhanced by using additional data cleaning and especially using Geopandas object for assigning venues to the correct neighbourhoods.

Opening a restaurant is a complex job and finding a promising area to start can be of great help. Using K-means clustering we have been able to group the neighbourhoods based on their common features.

The following neighbourhoods have been highlighted: -

- Central Toronto (cluster 0)
- East Toronto & DownTown Toronto (Cluster 2).

These neighbourhoods are potentially interesting starting an restaurant because of good transportability, relatively low restaurant competition and enough close to popular venues which could promote enough traffic of potential customers.