# Applied Data Science Capstone

## Capstone Project - The Battle of Neighborhoods

Created by Matej Kroc

November 14, 2019

# Contents

# 1 Introduction

## 1.1 Problem Description

In this project, which is part of the Data Science Capstone course on Coursera, we attempt to give protentional stakeholders a better description of the Toronto, Canada neighborhoods. This is done for the purpose of recognizing which area/neighborhood is the best for opening a new restaurant. Finding appropriate place for a new restaurant is crucial in nowadays business. Finding gap on the market in certain area or building your restaurant in an area that is similar to other location where certain type of a restaurant is popular can be the breaking point for successful business.

## 1.2 Approach

We collected the data about neighborhood restaurants and analyzed them. Location information was obtained via the Foursquare API. By using data mining technique k-mean clustering we categorize our dataset into different clusters based on similarities in restaurants available in the areas. It the end, we analyzed different types of restaurants and what cluster would be the best option for the investor to build the new restaurant in.

# 2 Data description

Data for this project were obtained from different sources, cleaned and add to a single data frame. Python and Jupyter Notebook were used during the analysis.

We needed to obtain names and locations of neighborhoods in Toronto, Canada. Names of different boroughs and neighborhoods together with their postcodes were extracted from Wikipedia page containing information about Canadian postcodes. Only postal codes where the first letter is M were extracted since postal codes beginning with M are located within the city

of Toronto. We omitted postal codes that have no borough and neighborhood assigned. Some neighborhoods are relatively small; therefore, some postal codes are the same for more than one neighborhood. We merged the data for these areas to obtain data with a unique postal code as an identifier for each neighborhood. Next, using the geocoder library available for Python we obtained location coordinates for each neighborhood based on the postal code. The final dataset contains 103 different postal codes. The first 5 entries in the dataset can be seen in Table 1. The map of Toronto with the marked neighborhood's location can be seen in Figure 1. An interactive map is also available at GitHub repository (GitHub, nbviewer(renders maps)) .

| | Postcode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| 1 | M4A | North York | Victoria Village | 43.725882 | -79.315572 |
| 2 | M5A | Downtown Toronto | Harbourfront, Regent Park | 43.654260 | -79.360636 |
| 3 | M6A | North York | Lawrence Heights, Lawrence Manor | 43.718518 | -79.464763 |
| 4 | M7A | Queen's Park | Queen's Park | 43.662301 | -79.389494 |

*Table 1: Dataset with neighborhood locations.*


*Figure 1: The map of Toronto with marked neighborhoods.*

Then, using Foursquare API and endpoint explore with specified ID code category, we obtained venues that are categorized as food venues. We used a limit of 100 venues and a radius 500 meters. Venues were then grouped in a data frame and relative frequencies computed. Top 5 venues for 3 neighborhoods together with their relative frequencies are displayed in Figure 2.

```
----Adelaide, King, Richmond----
                 venue  freq
0           Restaurant  0.08
1                Café  0.08
2       Sandwich Place  0.06
3     Asian Restaurant  0.06
4  American Restaurant  0.05


----Agincourt----
                      venue  freq
0  Latin American Restaurant  0.25
1             Sandwich Place  0.25
2         Chinese Restaurant  0.25
3             Breakfast Spot  0.25
4          Afghan Restaurant  0.00


----Agincourt North, L'Amoreaux East, Milliken, Steeles East----
                      venue  freq
0           Asian Restaurant   1.0
1          Afghan Restaurant   0.0
2  Middle Eastern Restaurant   0.0
3                Pizza Place   0.0
4         Persian Restaurant   0.0
```

*Figure 2: Sample of venues and their relative frequencies.*

# 3 Methodology

For our analysis we used unsupervised machine learning technique called $k$-means clustering. This approach allowed us to cluster the neighborhoods based on their similar restaurants.

## 3.1 $K$-means Clustering

$K$-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Important part of the clustering process was to choose the optimal number of clusters $k$. We applied the algorithm to our data with possible number of clusters $k$ ranging from 1 to 10. Then computed the sum square error (SSE) and plot the results (Figure 3).

$$SSE = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left|\left|x_i - c_k\right|\right|^2$$

where $w_{ik} = 1$ for data point $x_i$ if it belongs to the cluster $k$. Otherwise, $w_{ik} = 0$. $c_k$ is the centroid of $x_i$'s cluster.

It can be seen in Figure 3 that with increasing number of clusters the SSE is decreasing. But that is expected since the SSE is approaching $0$ for $k$ approaching the number of neighborhoods. We used the "elbow" method to decide for the optimal value of $k$. Based on the scatter plot we decided for $k = 3$ (3 clusters).

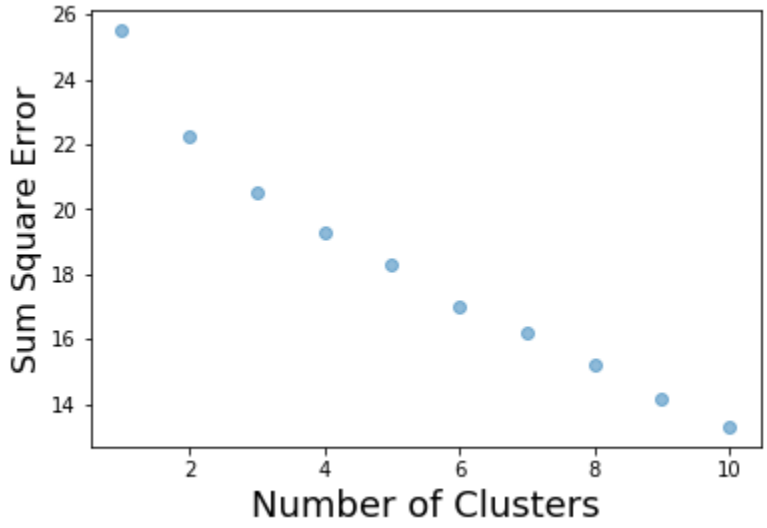


*Figure 3: Number of clusters vs. SSE*

# 4 Results

## 4.1 Master data frame

Using $k$-means algorithm on average frequencies of occurrence of venues in the neighborhoods we assigned cluster labels to each neighborhood. Clusters are labeled $0, 1, and\ 2$, respectively.

In can be seen distribution of clusters in our dataset. We can observe that most of the data were assigned to the cluster 1, precisely 83% of locations. Other two clusters 0 and 2 represent 10.5% and 6.5%, respectively.

| Cluster Labels | Postcode |
|---|---|
| 0 | 10 |
| 1 | 76 |
| 2 | 6 |

*Table 2: Distribution of cluster labels*

We created a table consisting of the 10 most common venues in each neighborhood and merged this table with our neighborhood locations and cluster labels. First 5 rows of the merged table can be seen below (Table 3).

| | Postcode | Borough | Neighborhood | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | M3A | North York | Parkwoods | 43.753259 | -79.329656 | 2 | BBQ Joint | Fast Food Restaurant | Wings Joint | Filipino Restaurant | Diner | Doner Restaurant | Donut Shop |
| 1 | M4A | North York | Victoria Village | 43.725882 | -79.315572 | 1 | Deli / Bodega | Portuguese Restaurant | French Restaurant | Pizza Place | Wings Joint | Ethiopian Restaurant | Dim Sum Restaurant |
| 2 | M5A | Downtown Toronto | Harbourfront, Regent Park | 43.654260 | -79.360636 | 1 | Café | Restaurant | Mexican Restaurant | Bakery | Breakfast Spot | Chinese Restaurant | Seafood Restaurant |
| 3 | M6A | North York | Lawrence Heights, Lawrence Manor | 43.718518 | -79.464763 | 1 | Hot Dog Joint | BBQ Joint | Bakery | Vietnamese Restaurant | Argentinian Restaurant | Fish & Chips Shop | Donut Shop |
| 4 | M7A | Queen's Park | Queen's Park | 43.662301 | -79.389494 | 1 | Sushi Restaurant | Diner | Japanese Restaurant | Wings Joint | Sandwich Place | Middle Eastern Restaurant | Creperie |

*Table 3: Merged table.*

Map of the Toronto's neighborhood with their cluster assignments can be seen in Figure 4.

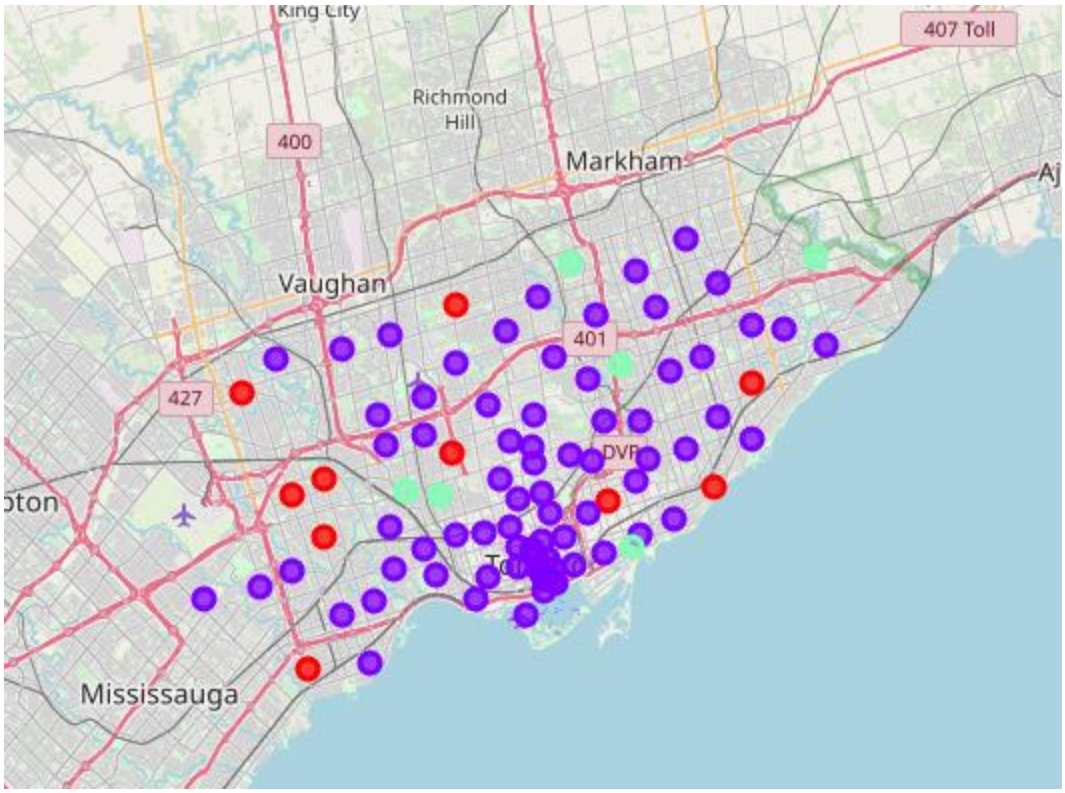


*Figure 4: Neighborhood clustering.*

## 4.2 Clusters

In this section we examined the separate clusters and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we then assigned a name to each cluster.

### 4.2.1 Cluster 0

This cluster consists of 10 neighborhoods. Based on the closer look on the most common venues we can see that the most common venue is Pizza Place, followed by Wings Joint and Fast Food Restaurant. We named this cluster: „Pizza and Fast Food venues".

### 4.2.2 Cluster 1

Cluster 1 is the biggest cluster in our set. It has 76 entries and this cluster is centered in downtown Toronto. However, there are also neighborhoods on suburbs that belong to cluster 1. This cluster

most common venues are mostly Café, Pizza Places and various restaurants, most commonly Sushi Restaurants. Based on the structure and the location of these places we assumed that these are more expensive restaurants or cafes for social and family events. We could name this cluster: „Premium location and high-end places“.

### 4.2.3 Cluster 2

This cluster consist only of 6 neighborhoods. These places scattered across the city but mostly away of the downtown area. Based on the common venues we assumed that this cluster has majority of ethnic restaurants like Mediterranean, Indian, or Philippine's restaurants with fast food chains being right behind. Cluster name in this case could be „Ethnic restaurants".

## 5 Discussion

As intended this study was developed to help cluster the city of Toronto based on its food venues to help protentional investors with finding a good place for their restaurant. The study was conducted on 103 Toronto neighborhoods based on their postal codes which determined their location in the city.

Different clustering algorithms could be used for this study and could potentially yield different results due to the differences between the neighborhoods. We used k-means clustering algorithm with specified number of clusters $k = 3$. We decided for three clusters based on the elbow method. However, it could be seen in Figure 3 that the elbow point isn't obvious. But using more clusters seemed unnecessary since when 4 or more clusters were used some clusters only consisted of one location.

Neighborhoods were then assigned to clusters and visualized on map of Toronto. We created a table of 10 most common venues for each neighborhood and describe clusters characteristics based on the most common venues in each cluster.

# 6 Conclusion

As a result, we were able to cluster Toronto's neighborhoods and derive some characteristics of their food venues. However, since cluster 1 consisted of majority of locations, further analysis of these neighborhoods would be beneficial for protentional investors in these areas.