Pashkov Daniel

22.10.2020



Capstone Project: Segmenting and Clustering Neighborhoods in Toronto City

[**1 Introduction: Description the Problem. 2**](#_Toc54233357)

[**2 Data cleaning and feature selection 3**](#_Toc54233358)

[**2.1 Data sourse 3**](#_Toc54233359)

[**2.2 Data cleaning 3**](#_Toc54233360)

[**3 Exploratory Analysis 6**](#_Toc54233361)

[**3.1 Explore Neighborhoods in Toronto 6**](#_Toc54233362)

[**3.2 Analyze Each Neighborhood 7**](#_Toc54233363)

[**4 Clustering Neighborhoods 8**](#_Toc54233364)

[**4.1 K-means Clustering Algorithm 8**](#_Toc54233365)

[**4.2 Examine Clusters 9**](#_Toc54233366)

[**5 Conclusion 12**](#_Toc54233367)

# Introduction: Description the Problem.

Today in modern city there are hundreds of venues and at each of its decade’s restaurants, shops, houses etc. For city government and businessman, who decide open new spa or park, or for citizens, who want to buy new home, very important know what is difference between city areas, which often unequal to districts borders.

So, needs some clusterisation, that dividing city on clusters based on most using places in that area. That is the main purpose of this project.

# Data cleaning and feature selection

## **2.1 Data sourse**

For this project neighborhoods data was scraped from [wikipedia page](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) using BeautifulSoup library (for more information [link here](https://beautiful-soup-4.readthedocs.io/en/latest/)).This data includes postal code for each borough.

For exploring neighborhoods in this project was use Foursquare API. Foursquare is a technology company that built a massive dataset of location data. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

## **2.2 Data cleaning**

Original data set have 217 neighborhoods. (Figure 1)

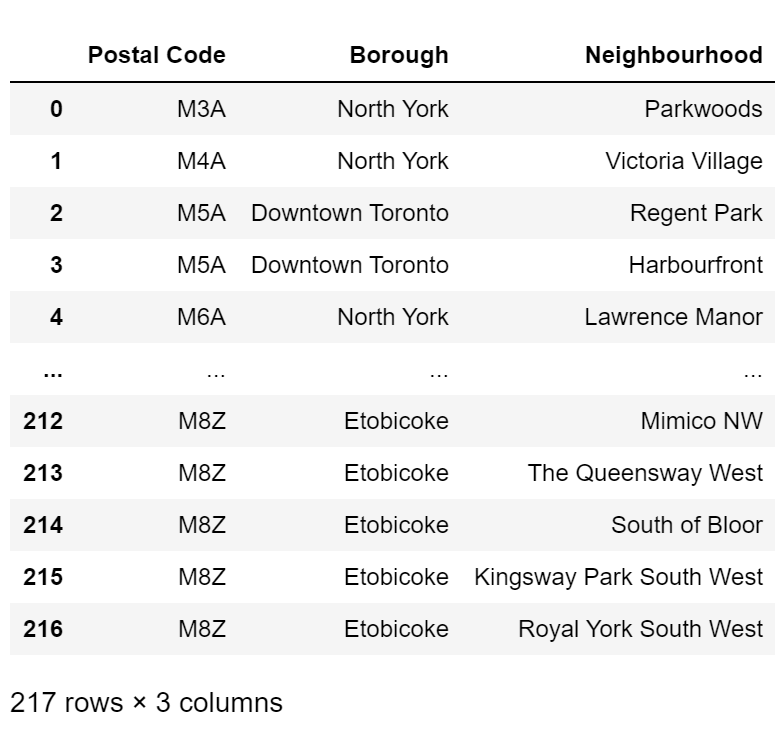


Figure 1. Original Dataset.

Further, get the coordinates for every neighborhood using Nominatim search engine and join this data with original dataset.

However, for some neighborhoods engine not found coordinates, so this row was dropped. After that, clean data includes 198 neighborhoods.(Figure 2, Figure 3)



Figure 2. Data with coordinates after cleaning.

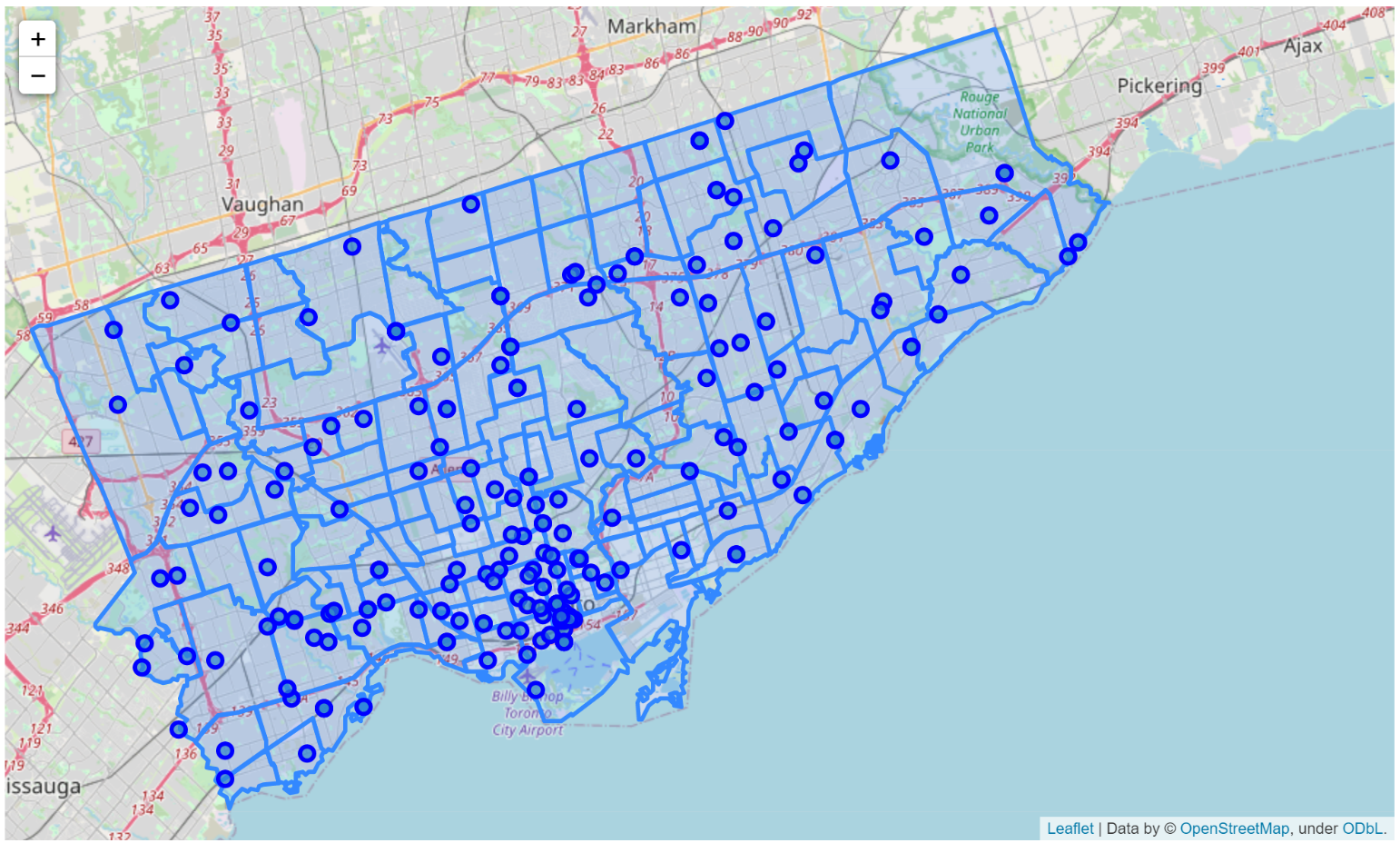


Figure 3. Neighborhoods in Toronto map. Straight line denotes districts borders.

According to map above (Figure 3), we see that some points located in a distance between others. So, for analyze neighborhoods needs calculate maximum distance between closest points, and this distance is about 3 kilometers.

# Exploratory Analysis

## **Explore Neighborhoods in Toronto**

For using Foursquare API was created account in [website for developer](https://foursquare.com/). This account

give credentials for connecting to the servers.

To get most common venues for each neighborhoods use explore request for venues with radius, which was calculated in previous chapter (half of 3 kilometers).

Example for explore request:

url = 'https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}

&ll={},{}&radius={}&limit={}'.format(CLIENT\_ID, CLIENT\_SECRET, VERSION, lat, lng, radius, LIMIT)

In example there are LIMIT parameters – is default REST API’s limit of value in response.

After, using response for this request (in JSON format), creates dataframe with next columns/parameters (Figure 4):

* + 1. 'Neighborhood',
    2. 'Neighborhood Latitude',
    3. 'Neighborhood Longitude',
    4. 'Venue',
    5. 'Venue Latitude',
    6. 'Venue Longitude',
    7. 'Venue Category'



Figure 4. Data from API response.

## **Analyze Each Neighborhood**

After getting venues, use onehot encoding venues category for each neighborhoods. In result – dataset there for each neighborhoods there are columns with 0/1 label for venues category. (Figure 5)

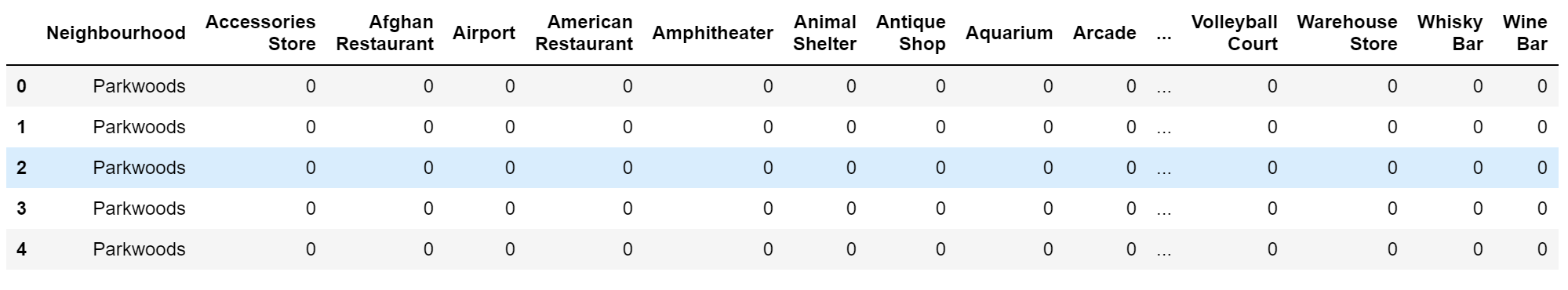


Figure 5. Data after onehot encoding.

Furter, first, group rows by neighborhood and by taking the mean of the frequency of occurrence of each category and, second, create the new dataframe with top 10 venues for each neighborhood. (Figure 6)



Figure 6. Dataset with feature columns.

# Clustering Neighborhoods

## **K-means Clustering Algorithm**

In this project for clustering using k-means algorithm. K-means can group data only unsupervised based on the similarity to each other. There are various types of clustering algorithms such as partitioning, hierarchical, or density based clustering.

K-means is a type of partitioning clustering. That is, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

For selecting the k-means parameters use the GridSearchCV library. After training, best models includes 7 clusters. In map below (Figure 7), each color of markers is different cluster.

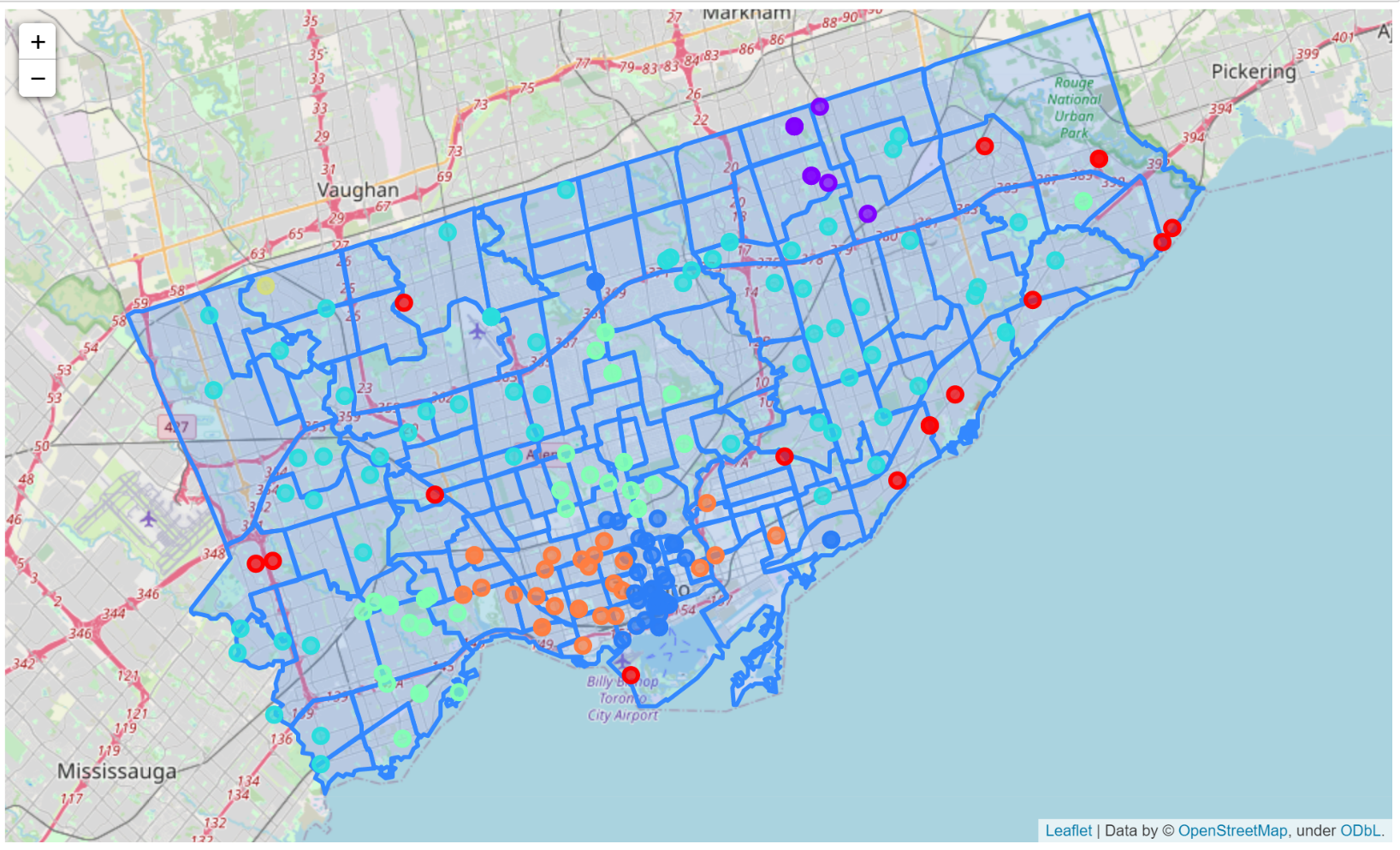


Figure 7. Toronto map with markered clusters.

## **Examine Clusters**

In this section, examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, possible assign a name to each cluster.

**Cluster 1**



According map and cluster description this cluster represents areas, which located on coast or near harbor and rivers. So, marked this cluster as 'Coasts Areas'.

**Cluster 2**



According the 1st common venues, obvious, it's 'Chinatown'.

**Cluster 3**



In this cluster the most common the coffee shops, cafes, pubs and points are located in center of Toronto, so this is 'City Center'.

**Cluster 4**



According number and location of this cluster, it's 'Low Cost Residential Areas'.

**Cluster 5**



This cluster include residential areas with parks, verious restaurants, spas. So it's 'Middle Cost Residential Areas'.

**Cluster 6**



This cluster are riddlest. One of the common venues is mexican restaurant and there are bar, store, so called it 'Latino-Americans Area'.

**Cluster 7**



The last cluster are closest to the city center, but it is residental area. So called cluster at 'Center Residential Areas'.

# Conclusion

Main purpose of this project is **segmenting and clustering neighborhoods in Toronto City**, based on a **real data**, which parsing on internet, using **BeautifulSoup** library.

Using **Nominatim search engine**, were receiving coordinates for each neighborhoods and after that, using **Forsquare API**, for each neighborhoods were find **top 10 most common venues** within a radius. Further, top venues transfer to ***k*-means algorithm** as a features, and using **GridSearchCV** library, was find **7 clusters** with the best accuracy. In 'Examine Clusters' sections each clusters was described and labeled.

In result there are clusters:

* **'Coasts Areas'**
* **'Chinatown'**
* **'Low Cost Residential Areas'**
* **'Middle Cost Residential Areas'**
* **'City Center'**
* **'Latino-Americans Area'**
* **'Center Residential Areas'**