The Battle of the Neighborhoods (Week 2)

Capstone Project

Applied Data Science Capstone by IBM / Coursera

Table of contents

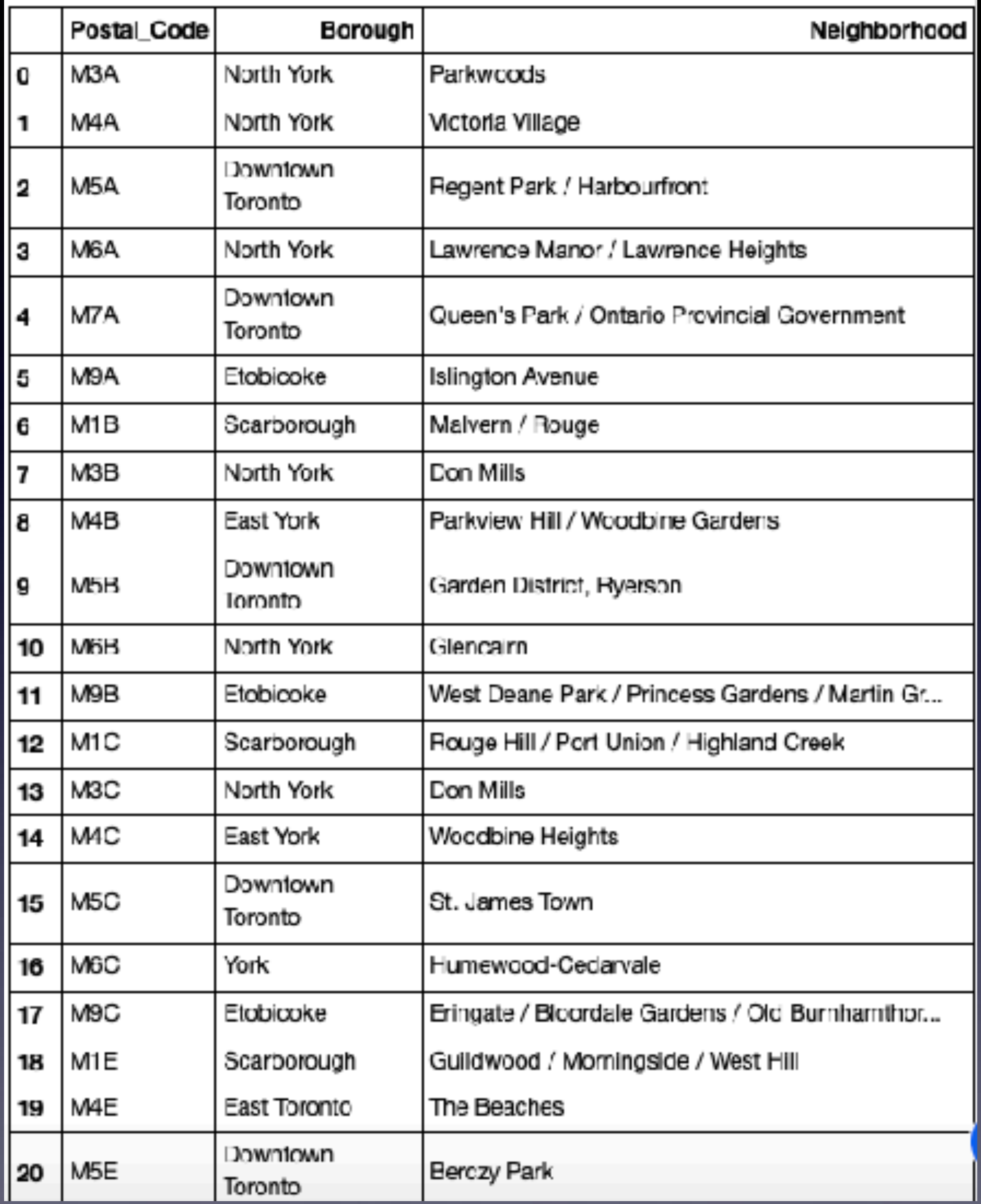
1. Introduction: Business Problem
2. Data
3. Methodology
4. Analysis
5. Results and Discussion
6. Conclusion

Introduction: Business Problem

- In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening a restaurant in Toronto, Canada.
- Here we will try finding if someone wants to open a new restaurant in the city which location is best suited for it keeping in mind the competitors and which income group of people will be attracted most to it based on the population of the neighborhood.
- Since there are lots of restaurants in Toronto, we will try to detect locations that are not already crowded with restaurants. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.
- We will use our data science powers to generate a few most promising neighborhoods based on this criteria.
- Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

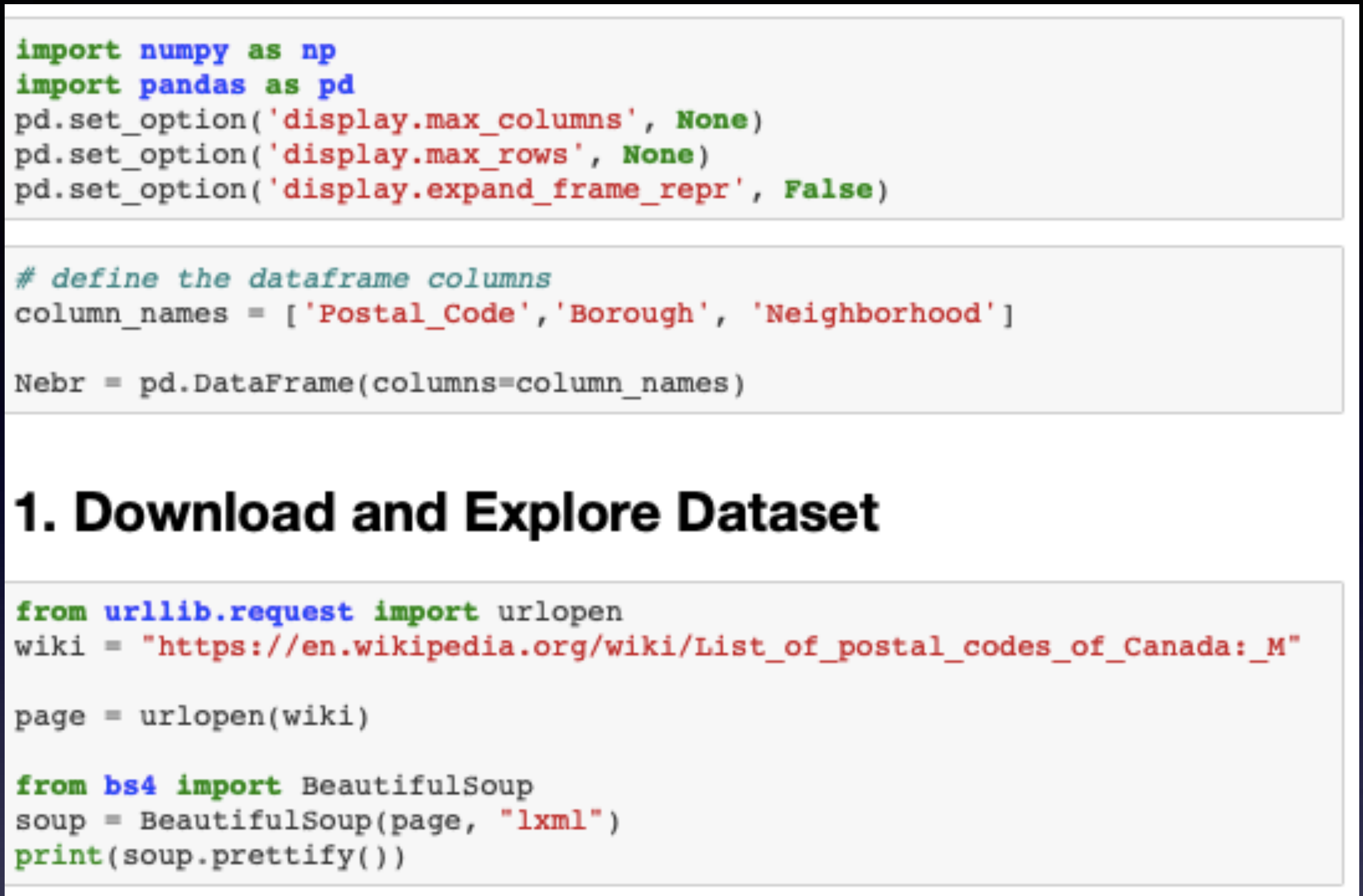
Data

- Based on definition of our problem, factors that will influence our decision are:
- All existing restaurants in the neighborhood (any type of restaurant)
- Age group of people with their income
- Distance of neighborhood from city center
- We decided to use regularly spaced grid of locations, revolved around city center, to define our neighborhoods.
- Following data sources will be needed to extract/generate the required information:
- Centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (Picture 1 & 2 next slide)
- The number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API(Picture 3 next slide)



	Postal_Code	Borough	Neighborhood
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
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern / Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill / Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson
10	M6B	North York	Glencairn
11	M9B	Etobicoke	West Deane Park / Princess Gardens / Martin Gr...
12	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
13	M3C	North York	Don Mills
14	M4C	East York	Woodbine Heights
15	M5C	Downtown Toronto	St. James Town
16	M6C	York	Humewood-Cedarvale
17	M9C	Etobicoke	Eringate / Bloordale Gardens / Old Burnhamthor...
18	M1E	Scarborough	Guildwood / Morningside / West Hill
19	M4E	East Toronto	The Beaches
20	M5E	Downtown Toronto	Berczy Park

Picture 2



```
import numpy as np
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
pd.set_option('display.expand_frame_repr', False)
```

```
# define the dataframe columns
column_names = ['Postal_Code','Borough', 'Neighborhood']

Nebr = pd.DataFrame(columns=column_names)
```

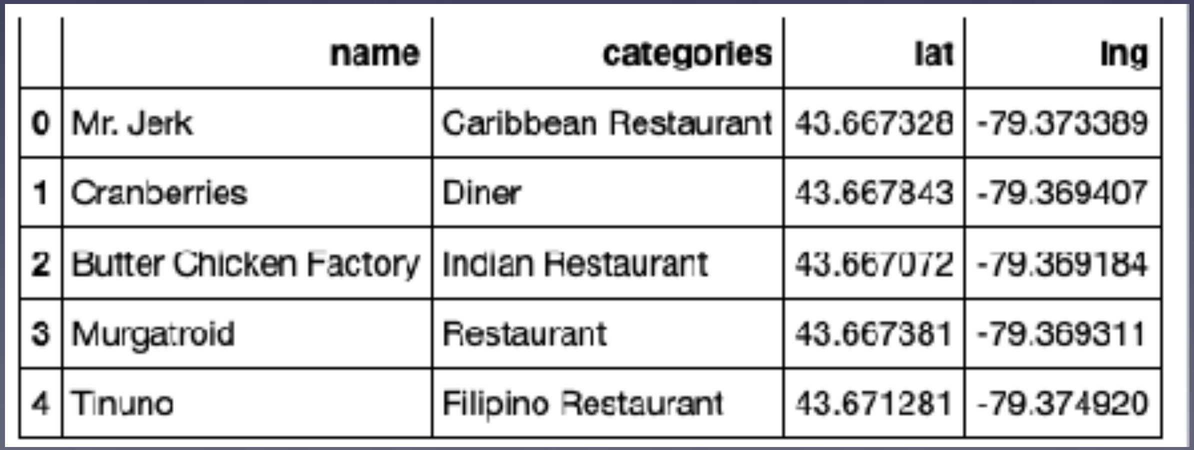
1. Download and Explore Dataset

```
from urllib.request import urlopen
wiki = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

page = urlopen(wiki)

from bs4 import BeautifulSoup
soup = BeautifulSoup(page, "lxml")
print(soup.prettify())
```

Picture 1



	name	categories	lat	lng
0	Mr. Jerk	Caribbean Restaurant	43.667328	-79.373389
1	Cranberries	Diner	43.667843	-79.369407
2	Butter Chicken Factory	Indian Restaurant	43.667072	-79.369184
3	Murgatroid	Restaurant	43.667381	-79.369311
4	Tinuno	Filipino Restaurant	43.671281	-79.374920

Picture 3

Methodology

- The main motto of this project is to find best location to open a new restaurant in Toronto, Canada based on competition in different locality and their population.

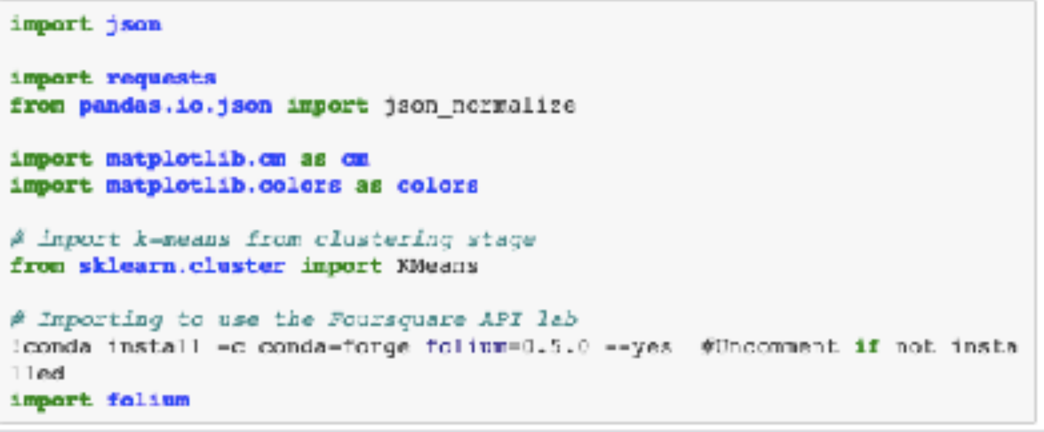
- So, to do this I have used 2 different data sets available as mentioned above. Those 2 data set contains Locality information of Toronto, different age group of people in the people, population.

- To solve the problem I am going to use “K-Means Clustering Algorithm ".

- K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).

- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

- The results of the K-means clustering algorithm are:
- The centroids of the K clusters, which can be used to label new data Labels for the training data (each data point is assigned to a single cluster)
- Also, I will be utilizing different maps in-order to give a clear vision to the target audience.
- Steps we took for the analysis:
- Collected required data: location and type (category) of every restaurant within our lat and lng. We have also the type of restaurants in particular locality.
- Explored the 'restaurant density' across different areas of Toronto - we will use K- mean to identify a few promising areas close to center with low number of restaurants and their type.
- Explored the most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with less restaurants in radius of 500 meters, We will present map of all such locations but also create clusters (using k-means clustering) of those locations to explore neighborhood.



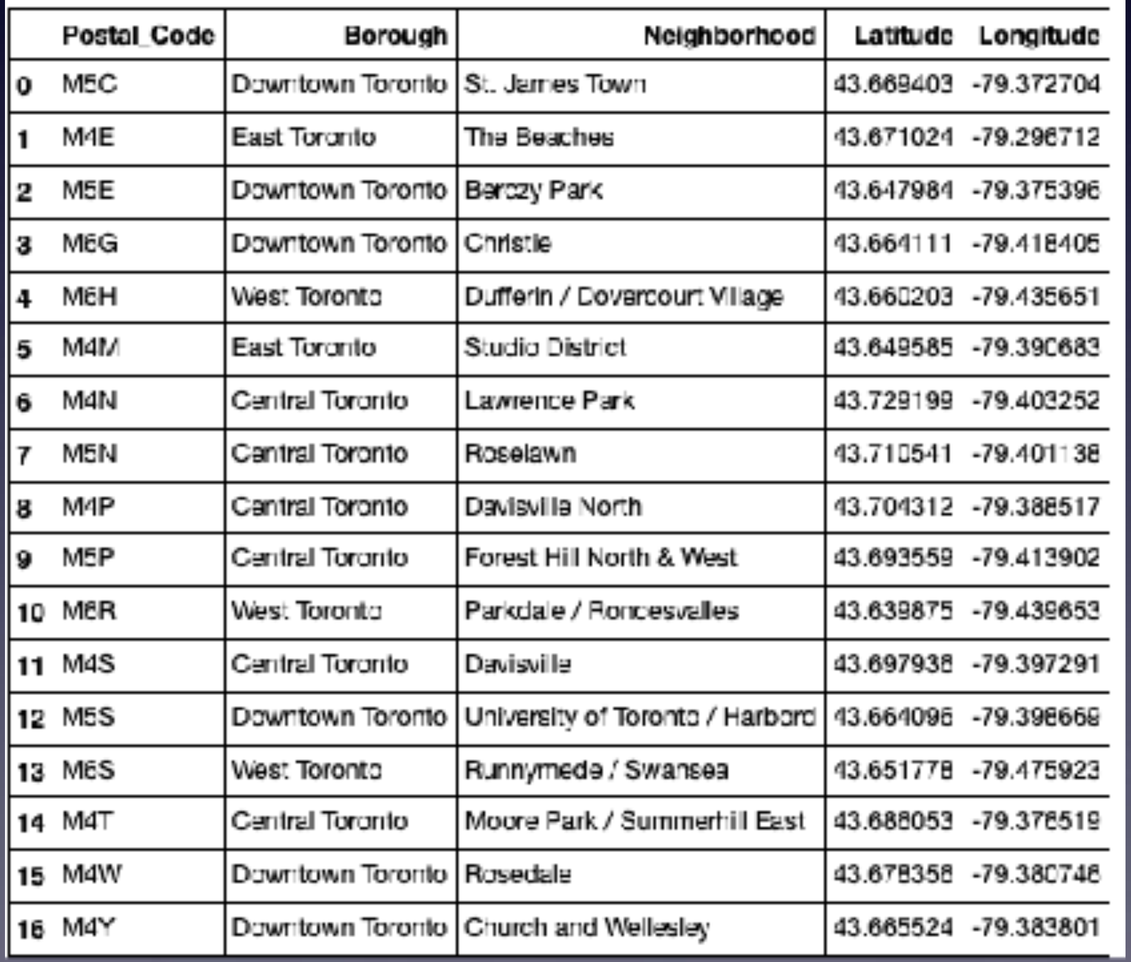
```
import json

import requests
from pandas.io.json import json_normalize

import matplotlib.cm as cm
import matplotlib.colors as colors

# import k-means from clustering stage
from sklearn.cluster import KMeans

# Importing to use the Foursquare API lab
!conda install -c conda-forge folium=0.5.0 --yes  #Uncomment if not insta
lled
import folium
```



	Postal_Code	Borough	Neighborhood	Latitude	Longitude
0	M5C	Downtown Toronto	St. James Town	43.669403	-79.372704
1	M4E	East Toronto	The Beaches	43.671024	-79.296712
2	M5E	Downtown Toronto	Berczy Park	43.647984	-79.375396
3	M6G	Downtown Toronto	Christie	43.664111	-79.418405
4	M6H	West Toronto	Dufferin / Dovercourt Village	43.660203	-79.435651
5	M4M	East Toronto	Studio District	43.649585	-79.390683
6	M4N	Central Toronto	Lawrence Park	43.729199	-79.403252
7	M5N	Central Toronto	Roselawn	43.710541	-79.401138
8	M4P	Central Toronto	Davisville North	43.704312	-79.388517
9	M5P	Central Toronto	Forest Hill North & West	43.693559	-79.413902
10	M6R	West Toronto	Parkdale / Roncesvalles	43.639875	-79.439653
11	M4S	Central Toronto	Davisville	43.697938	-79.397291
12	M5S	Downtown Toronto	University of Toronto / Harbord	43.664096	-79.398668
13	M6S	West Toronto	Runnymede / Swansea	43.651778	-79.475923
14	M4T	Central Toronto	Moore Park / Summerhill East	43.688053	-79.378519
15	M4W	Downtown Toronto	Rosedale	43.678358	-79.380748
16	M4Y	Downtown Toronto	Church and Wellesley	43.665524	-79.383801

Analysis

- Data Identification, capturing and cleaning.

- Combining different data source and sorting neighborhood based on Longitude and latitude

- Explore the Toronto's neighborhoods

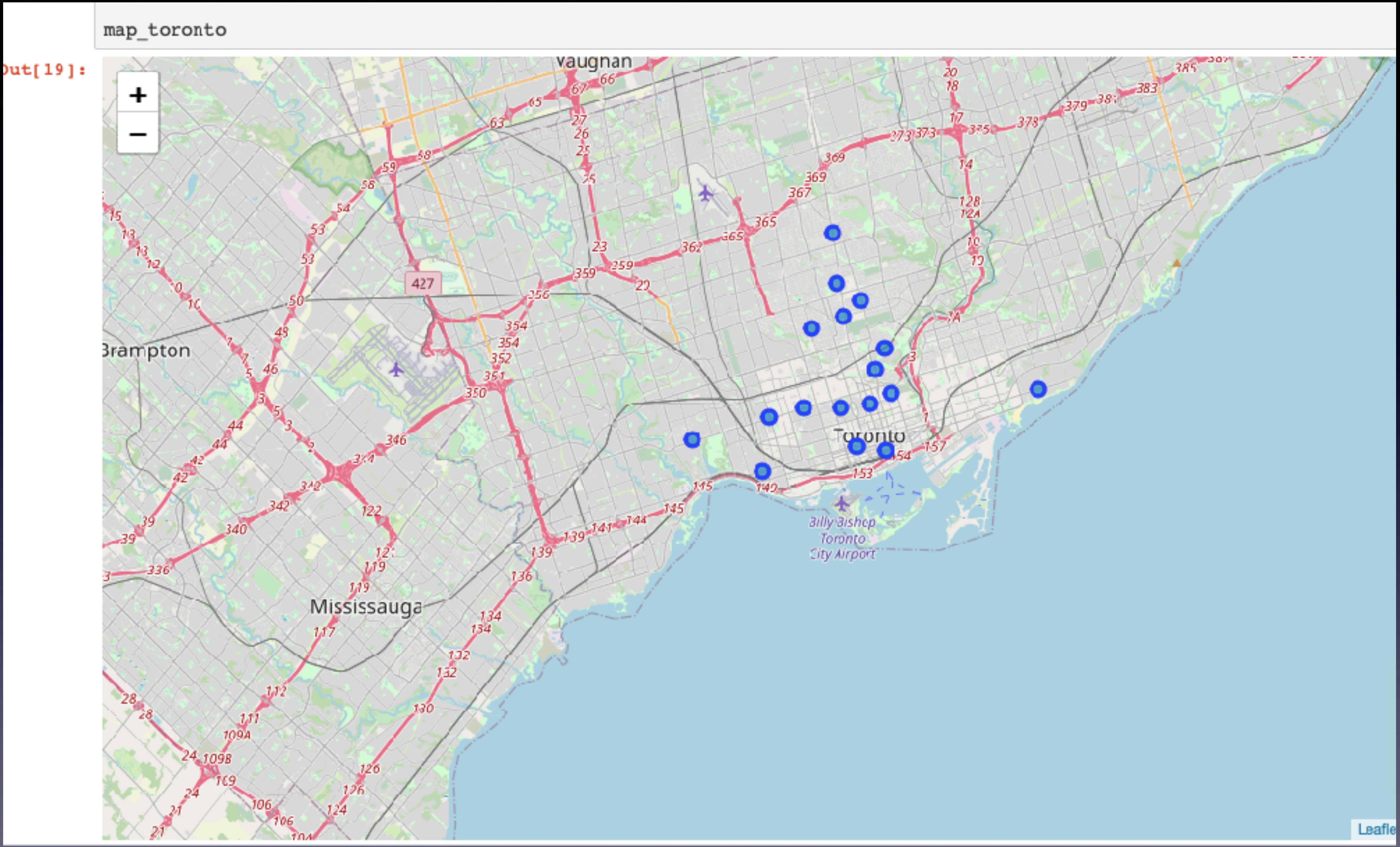
- Clustering

Data Identification, Capturing and Cleaning

- Search & Identify the relevant data source and capture it, here we are using wikipedia to get data about Toronto, Canada.
- Then we remove all the redundant value(data cleaning).
- Then we combine neighborhood similar Bronx. Now the data is clean and ready to use.

Combining different data source and sorting neighborhood based on Longitude and latitude

- Now, we will combine neighborhood dataset with postal address and dataset with Latitude & Longitude and save them it separate data frame.
- The resultant data frame with contain details about Postal code, Brough, Neighborhood, Latitude & Longitude.



```
map_toronto
```

```
Out[19]:
```

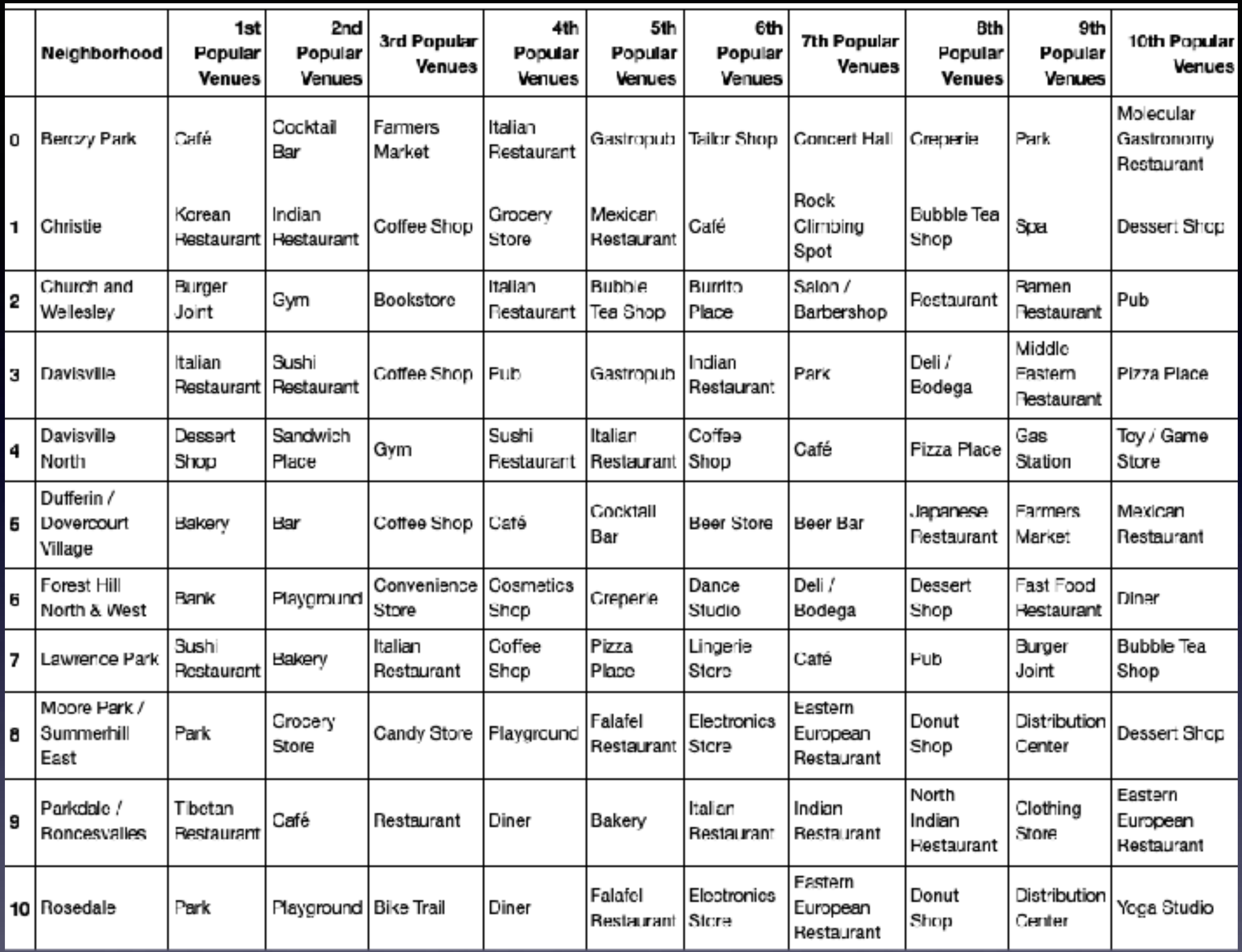
- Then visualize it using folium map

Explore the Toronto's neighborhoods

- Firstly, we explored all the neighborhoods in the city of Toronto, using the Latitude & Longitude data, using Foresquare API to get the Restaurant venues available in Toronto.

- Explore the unique categories in the neighborhood.Filter the Venues details for all possible ‘Restaurants’.

- Find each neighborhood along with the top most common venues.Identify the top 10 venues for each neighborhood.

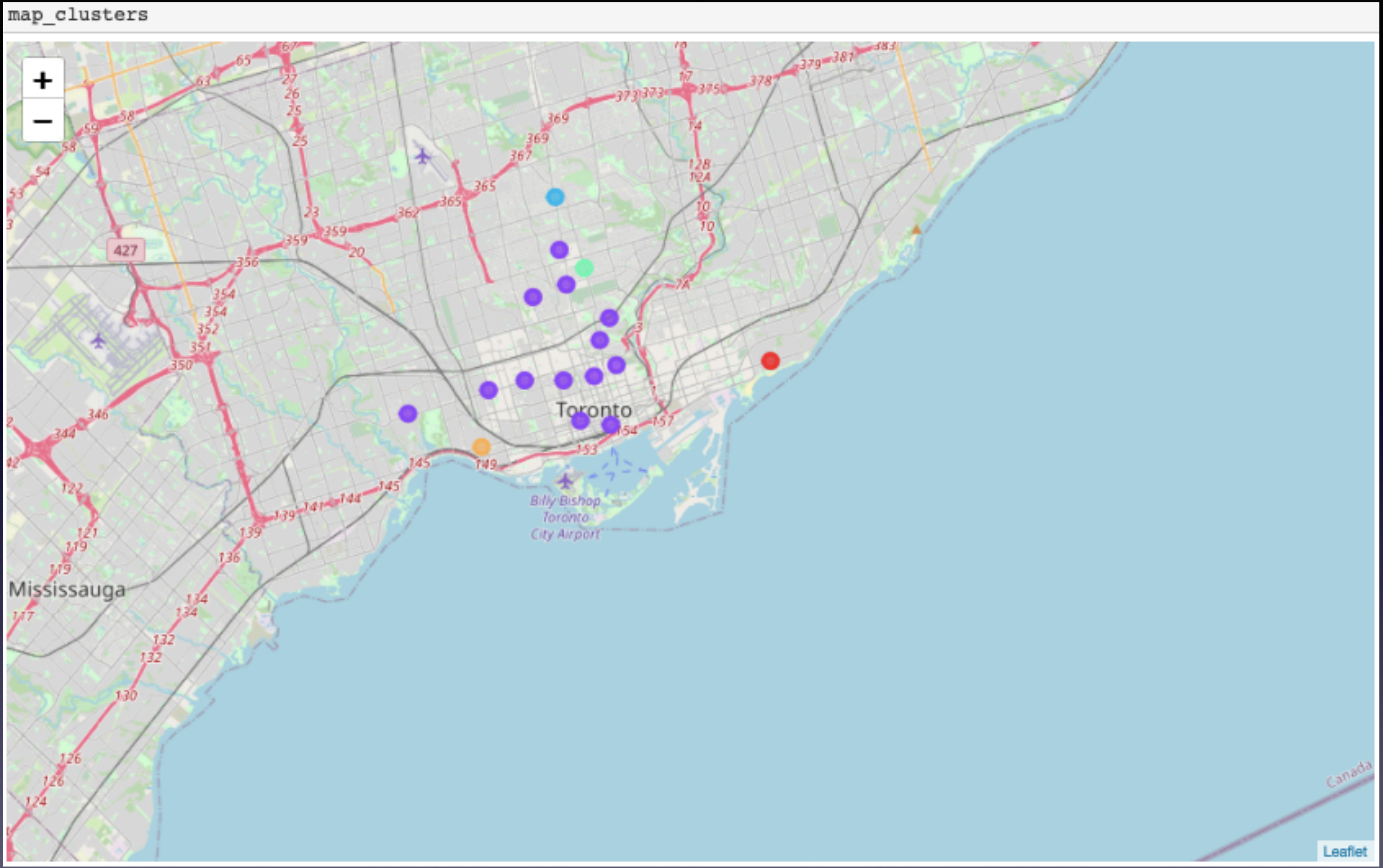


	Neighborhood	1st Popular Venues	2nd Popular Venues	3rd Popular Venues	4th Popular Venues	5th Popular Venues	6th Popular Venues	7th Popular Venues	8th Popular Venues	9th Popular Venues	10th Popular Venues
0	Berczy Park	Café	Cocktail Bar	Farmers Market	Italian Restaurant	Gastropub	Tailor Shop	Concert Hall	Creperie	Park	Molecular Gastronomy Restaurant
1	Christie	Korean Restaurant	Indian Restaurant	Coffee Shop	Grocery Store	Mexican Restaurant	Café	Rock Climbing Spot	Bubble Tea Shop	Spa	Dessert Shop
2	Church and Wellesley	Burger Joint	Gym	Bookstore	Italian Restaurant	Bubble Tea Shop	Burrito Place	Salon / Barbershop	Restaurant	Ramen Restaurant	Pub
3	Davisville	Italian Restaurant	Sushi Restaurant	Coffee Shop	Pub	Gastropub	Indian Restaurant	Park	Deli / Bodega	Middle Eastern Restaurant	Pizza Place
4	Davisville North	Dessert Shop	Sandwich Place	Gym	Sushi Restaurant	Italian Restaurant	Coffee Shop	Café	Pizza Place	Gas Station	Toy / Game Store
5	Dufferin / Dovercourt Village	Bakery	Bar	Coffee Shop	Café	Cocktail Bar	Beer Store	Beer Bar	Japanese Restaurant	Farmers Market	Mexican Restaurant
6	Forest Hill North & West	Bank	Playground	Convenience Store	Cosmetics Shop	Creperie	Dance Studio	Deli / Bodega	Dessert Shop	Fast Food Restaurant	Diner
7	Lawrence Park	Sushi Restaurant	Bakery	Italian Restaurant	Coffee Shop	Pizza Place	Lingerie Store	Café	Pub	Burger Joint	Bubble Tea Shop
8	Moore Park / Summerhill East	Park	Grocery Store	Candy Store	Playground	Falafel Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop	Distribution Center	Dessert Shop
9	Parkdale / Roncesvalles	Tibetan Restaurant	Café	Restaurant	Diner	Bakery	Italian Restaurant	Indian Restaurant	North Indian Restaurant	Clothing Store	Eastern European Restaurant
10	Rosedale	Park	Playground	Bike Trail	Diner	Falafel Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop	Distribution Center	Yoga Studio

Top 10 Vanues

Clustering

- With an assumption of 5 clusters, use K-Cluster algorithm to come up with 5 different clusters in Toronto with similar set of Venues.
- Explore each cluster and determine the discriminating venue categories that distinguish each cluster.
- Identify the clusters & Boroughs/Neighborhoods with Maximum number restaurants and there types.



```
map_clusters
```

Results and Discussion

- Our analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density fairly close to city center.We have 4 boroughs and 74 neighborhoods inside geographical coordinate of 43.653963, -79.387207.

- Based on our initial assumption of the cluster with maximum number of restaurants will have the best possibility to have a new restaurant due to the need in the area. Based on the resultant clusters it looks like Cluster 1 and Cluster 5 have higher number of restaurants than rest of the clusters.

- It is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area.

- Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusion

- Purpose of this project was to identify areas Toronto with low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant.

- By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants.

- Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Thanks