IBM

DATA SCIENCE PROFESSIONAL CERTIFICATE

Karol Kozlowski

Recommending attractive locations for restaurants using data analysis methods

Warsaw, 2020

1. Introduction

1.1 Background

Establishing a restaurant involves setting many criteria and having a vision of what you want to deliver to your potential customers. I mean here such aspects as the type of cuisine served, the definition of the type of customers for which it should be a place, decor, location, etc..

Recommendation systems are now very popular in many aspects such as e-commerce services, movie rating (Netflix), social networking and a lot of services helping users finding preferred items, in this work I will focus on (LBSN) location-based social networks.

1.2 Business Problem

This analysis is for those who know what kind of cuisine they want to serve in their restaurant. I will focus on identifying good/not good locations for this restaurant. The choice must be based on an analysis of many factors such as: the population of the district, the popularity of the area among people, the proximity of popular public transport, the proximity of other urban infrastructure and the most important is the number of units in the area offering the same type of cuisine.

2. Data acquisition and cleaning

2.1 Data sources.

I used data sources such as:

- List of Toronto postal codes from WIKIPEDIA.
- Geospatial Coordinates CSV file for Toronto postal codes from http://cocl.us/Geospatial_data
- Foursquare API - to explore venues in Toronto neighbourhoods

2.2 Data cleaning

At the beginning, I read the HTML text using BeautifulSoup library from the wikitable with postal codes of Toronto.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0, however, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, suggesting t

Postal Code	Borough	Neighborhood
M1A	Not assigned	
M2A	Not assigned	
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Malvern, Rouge

Next, I dropped rows with blank “Borough” values, filled “Neighborhood” positions when “Borough” for that neighborhood was known. After that I grouped each Postal Code with all corresponding neighborhoods.

Out[14]:

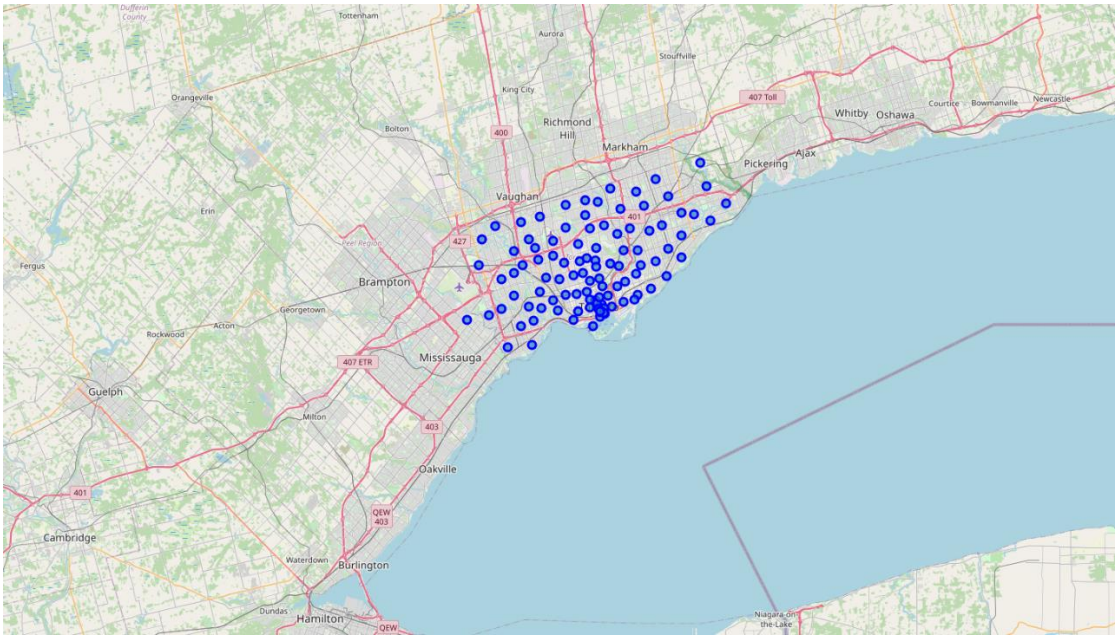
	Postal Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

2.3 Insert Geo-spatial Coordinates

I imported a csv file with geolocation data into the script. Then I combined the postal codes from the table with the corresponding geolocation data. As a result of these actions I got dataframes with new Latitude and Longitude columns added.

	Postal Code	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Then I could plot a map of the neighborhoods using Folium Library:



2.4 Foursquare API utilizing

In order to get information about interesting places in the district I used Foursquare. Using their API I downloaded the top 100 places within 500 meters from the point inserted by the geolocation data for the district.
The downloaded data was in JSON format, so it had to be extracted and transformed to be placed in DataFrame.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Id	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	4e8d9dcdd5fbbbb6b3003c7b	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	4e6696b6d16433b9ffff47c3	KFC	43.754387	-79.333021	Fast Fooc Restaurar
2	Parkwoods	43.753259	-79.329656	4cb11e2075ebb60cd1c4caad	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	4c633acb86b6be9a61268e34	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	4bbe904a85fbb713420d7167	Tim Hortons	43.725517	-79.313103	Coffee Shop

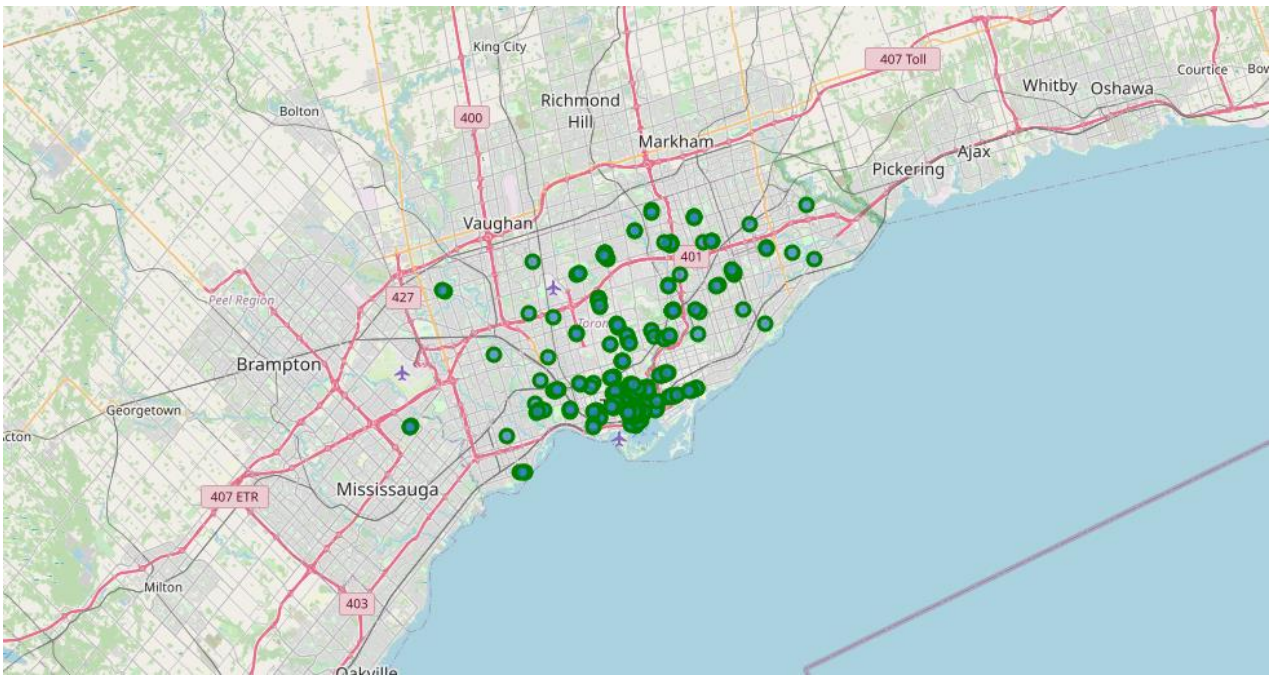
Then I will perform top venue exploration for all neighborhoods and merge it to one data frame, that will include neighborhoods and their top rated venues, and because our main goal is to recommend a neighborhood for a new restaurant, we need to filter our data, so we will have only venues categories that contains the word 'Restaurant'

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Id	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	4e6696b6d16433b9ffff47c3	KFC	43.754387	-79.333021	Fast Food Restauran
1	Victoria Village	43.725882	-79.315572	4f3ecce6e4b0587016b6f30d	Portugril	43.725819	-79.312785	Portugues Restauran
2	Victoria Village	43.725882	-79.315572	4d689350b6f46dcb77ee15b2	The Frig	43.727051	-79.317418	French Restauran
3	Regent Park, Harbourfront	43.654260	-79.360636	5612b1cc498e3dd742af0dc8	Impact Kitchen	43.656369	-79.356980	Restauran
4	Regent Park, Harbourfront	43.654260	-79.360636	53a22c92498ec91fda7ce133	Cluny Bistro & Boulangerie	43.650565	-79.357843	French Restauran

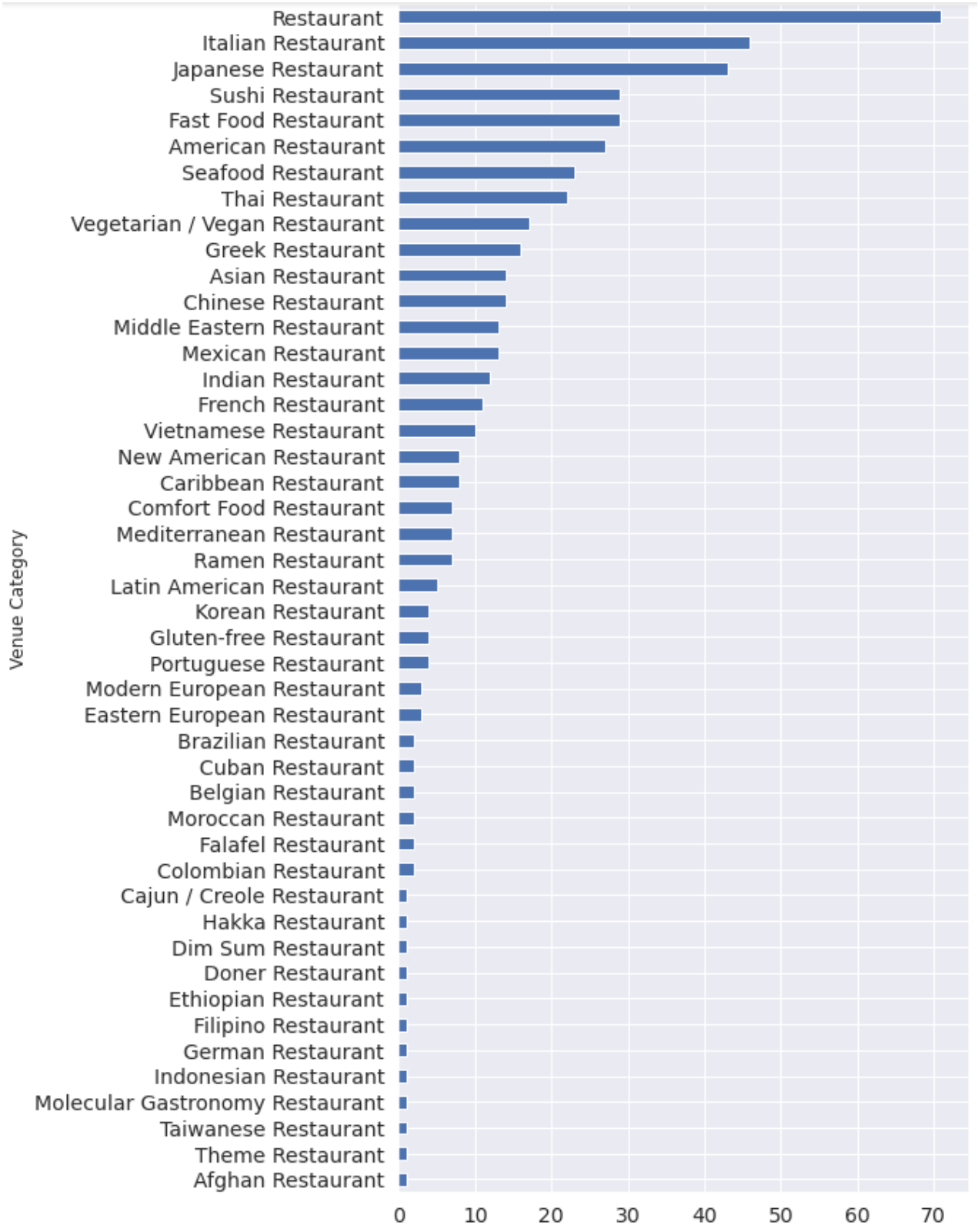
3. Methodology section (EDA - Exploratory Data Analysis)

From the dataset, we filtered out the neighborhood which have venue contains the word 'Restaurant' . After filtering out the the data, our dataset shape is (494, 8) that means that the data contains 494 rows (Restaurant venues) and 8 relevant columns.

Here is a map of all the top rated restaurant in Toronto:



Here is the distribution of the restaurant cuisines:



As you can see, the 3 most numerous restaurant types are general, Italian and Japanese.

4. Results section

Then for each one of the neighborhood we can extract the top frequent restaurant categories:

```
----Agincourt----
                                    cuisine  freq
0  Venue Category_Latin American Restaurant   1.0
1           Venue Category_Afghan Restaurant   0.0
2        Venue Category_Moroccan Restaurant   0.0
3         Venue Category_Italian Restaurant   0.0
4        Venue Category_Japanese Restaurant   0.0


----Bathurst Manor, Wilson Heights, Downsview North----
                                         cuisine  freq
0                 Venue Category_Sushi Restaurant  0.33
1                       Venue Category_Restaurant  0.33
2        Venue Category_Middle Eastern Restaurant  0.33
3                Venue Category_Afghan Restaurant  0.00
4  Venue Category_Molecular Gastronomy Restaurant  0.00
```

As we can see for "Bathurst Manor, Wilson Heights, Downsview North" three types of restaurants have the same frequency. These are: Sushi Restaurant, General Restaurant and Middle Eastern Restaurant.

	Neighborhood	1st Most Common cuisine	2nd Most Common cuisine	3rd Most Common cuisine	4th Most Common cuisine	5th Most Commo cuisine
0	Agincourt	Venue Category_Latin American Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Dim Sum Restaurant	Venue Category_German Restaurant	Venue Category_French Restaurant
1	Bathurst Manor, Wilson Heights, Downsview North	Venue Category_Sushi Restaurant	Venue Category_Restaurant	Venue Category_Middle Eastern Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Cuban Restaurant
2	Bayview Village	Venue Category_Japanese Restaurant	Venue Category_Chinese Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Dim Sum Restaurant	Venue Category_German Restaurant
3	Bedford Park, Lawrence Manor East	Venue Category_Restaurant	Venue Category_Italian Restaurant	Venue Category_Comfort Food Restaurant	Venue Category_American Restaurant	Venue Category_Thai Restaurant
4	Berczy Park	Venue Category_Seafood Restaurant	Venue Category_Restaurant	Venue Category_Greek Restaurant	Venue Category_Thai Restaurant	Venue Category_Vegetari / Vegan Restauran

5. Clustering Modeling

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)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

I will use popularity recommendation filtering approach in order to make recommendations.

Popularity Recommendation : Given the number of play counts (occurrences) for items songs, movies, whatever, just sort by descending ordering and recommend everyone what's popular.This actually makes a good shelf for any content platform, such as Netflix, Spotify or Amazon.

I performed a qualitative comparison of the existing techniques used in the Data I got from Foursquare, followed by a clustering on the type of the services and the location features the Foursquare utilizes to perform the recommendations.

First of all I performed one hot encoding on the data to get a binary represented data set for each neighborhood venue(Restaurant)

	Neighborhood	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Venue Categor / Creole Restaur
0	Parkwoods	0	0	0	0	0	0
1	Victoria Village	0	0	0	0	0	0
2	Victoria Village	0	0	0	0	0	0
3	Regent Park, Harbourfront	0	0	0	0	0	0
4	Regent Park, Harbourfront	0	0	0	0	0	0

Then I grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category I got :

	Neighborhood	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Venue Categor / Creole Restaur
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0
1	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0
2	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0
3	Bedford Park, Lawrence Manor East	0.0	0.1	0.0	0.0	0.0	0.0
4	Berczy Park	0.0	0.0	0.0	0.0	0.0	0.0

A dataframe representing each neighborhood and the mean frequency of occurrence for each restaurant cuisine in it.

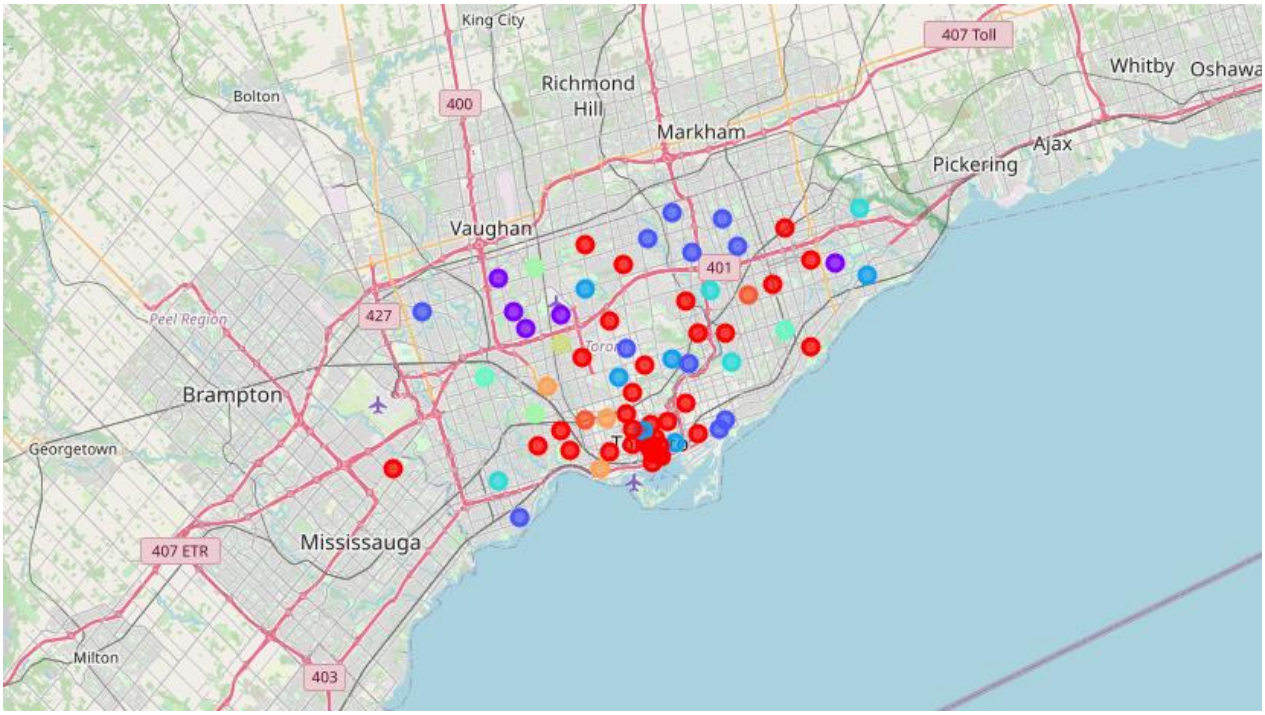
In my solution I chose K = 10, so I want to split neighborhoods to 10 clusters.

6. Discussion section

For simplicity we can see clusters size:

```
Cluster Labels
0    35
1     5
2    11
3     6
4     4
5     2
6     2
7     1
8     3
9     2
```

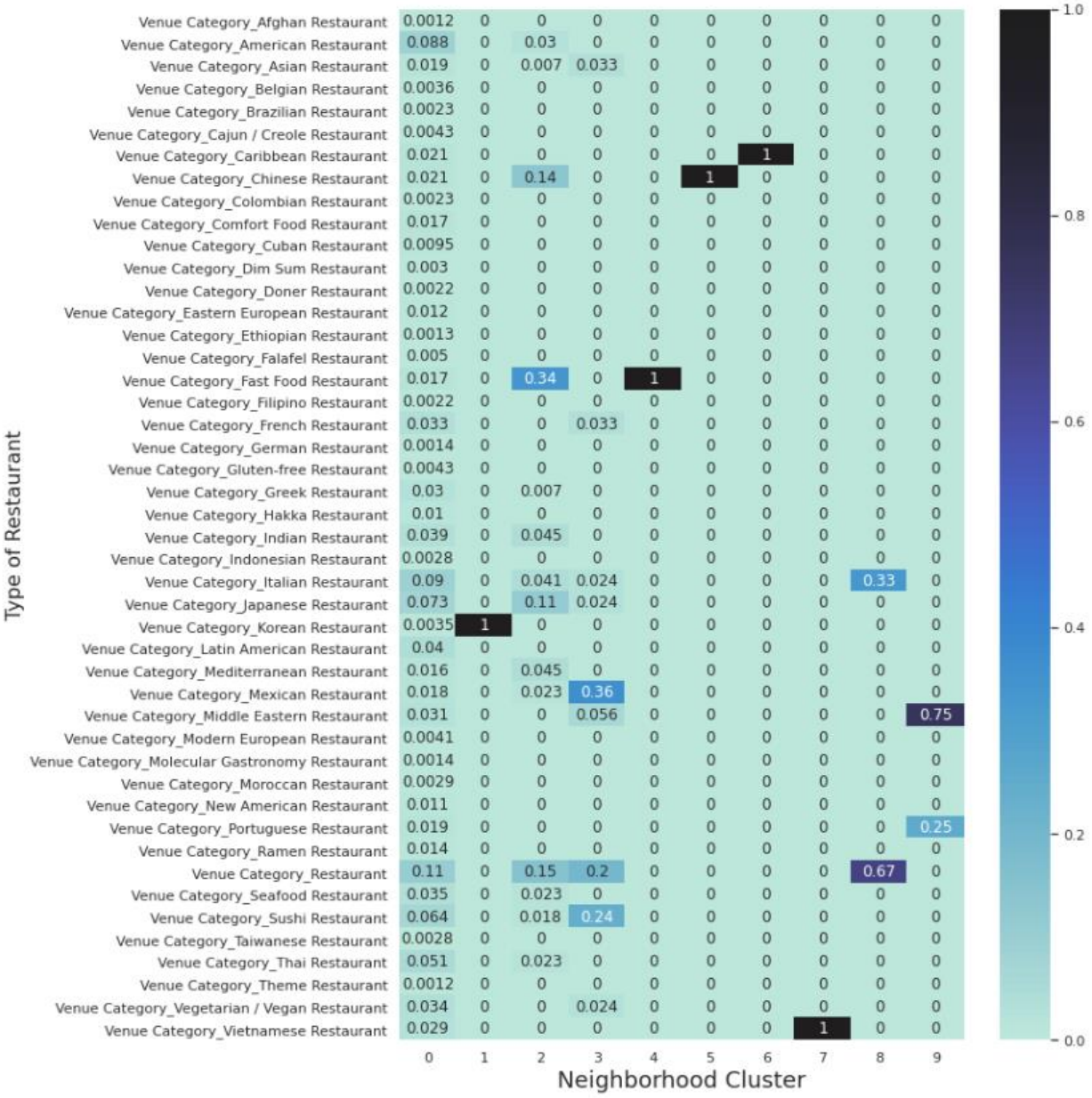
Easily we can see that cluster 0 is the biggest one, that means that all neighborhoods in cluster 0 are similar with top rated restaurants cuisines, Now, let’s plot a map of neighborhood and colored clusters.



in this stage, I merged cluster labels to the dataset, so the new dataset will be :

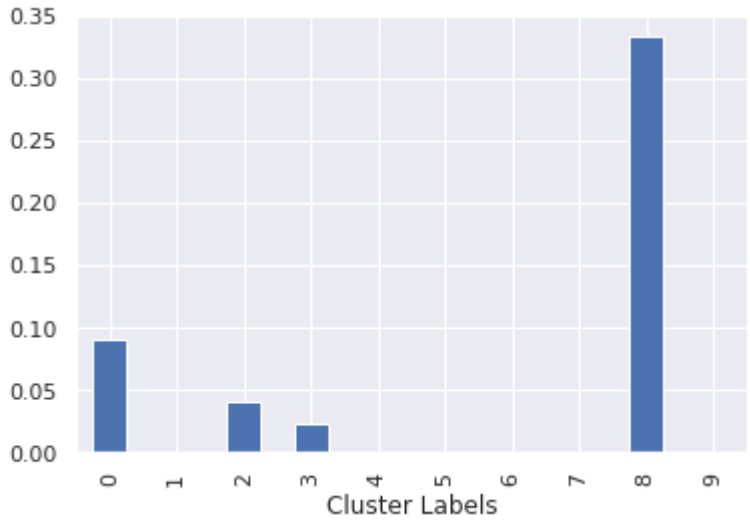
Cluster Labels	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Venue Category_Cajun / Creole Restaurant
0	0.001212	0.08777	0.018684	0.003608	0.002273	0.004329
1	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.02972	0.006993	0.000000	0.000000	0.000000
3	0.000000	0.00000	0.033333	0.000000	0.000000	0.000000
4	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
5	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
6	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
7	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
9	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000

Then I grouped the dataset by cluster to get the mean of the frequency occurrence of occurrence of each cluster, we then I examined each cluster by a heatmap.


Type of Restaurant
Venue Category_Afghan Restaurant 0.0012 0 0 0 0 0 0 0 0 0
Venue Category_American Restaurant 0.088 0 0.03 0 0 0 0 0 0 0
Venue Category_Asian Restaurant 0.019 0 0.007 0.033 0 0 0 0 0 0
Venue Category_Belgian Restaurant 0.0036 0 0 0 0 0 0 0 0 0
Venue Category_Brazilian Restaurant 0.0023 0 0 0 0 0 0 0 0 0
Venue Category_Cajun / Creole Restaurant 0.0043 0 0 0 0 0 0 0 0 0
Venue Category_Caribbean Restaurant 0.021 0 0 0 0 0 1 0 0 0
Venue Category_Chinese Restaurant 0.021 0 0.14 0 0 1 0 0 0 0
Venue Category_Colombian Restaurant 0.0023 0 0 0 0 0 0 0 0 0
Venue Category_Comfort Food Restaurant 0.017 0 0 0 0 0 0 0 0 0
Venue Category_Cuban Restaurant 0.0095 0 0 0 0 0 0 0 0 0
Venue Category_Dim Sum Restaurant 0.003 0 0 0 0 0 0 0 0 0
Venue Category_Doner Restaurant 0.0022 0 0 0 0 0 0 0 0 0
Venue Category_Eastern European Restaurant 0.012 0 0 0 0 0 0 0 0 0
Venue Category_Ethiopian Restaurant 0.0013 0 0 0 0 0 0 0 0 0
Venue Category_Falafel Restaurant 0.005 0 0 0 0 0 0 0 0 0
Venue Category_Fast Food Restaurant 0.017 0 0.34 0 1 0 0 0 0 0
Venue Category_Filipino Restaurant 0.0022 0 0 0 0 0 0 0 0 0
Venue Category_French Restaurant 0.033 0 0 0.033 0 0 0 0 0 0
Venue Category_German Restaurant 0.0014 0 0 0 0 0 0 0 0 0
Venue Category_Gluten-free Restaurant 0.0043 0 0 0 0 0 0 0 0 0
Venue Category_Greek Restaurant 0.03 0 0.007 0 0 0 0 0 0 0
Venue Category_Hakka Restaurant 0.01 0 0 0 0 0 0 0 0 0
Venue Category_Indian Restaurant 0.039 0 0.045 0 0 0 0 0 0 0
Venue Category_Indonesian Restaurant 0.0028 0 0 0 0 0 0 0 0 0
Venue Category_Italian Restaurant 0.09 0 0.041 0.024 0 0 0 0 0.33 0
Venue Category_Japanese Restaurant 0.073 0 0.11 0.024 0 0 0 0 0 0
Venue Category_Korean Restaurant 0.0035 1 0 0 0 0 0 0 0 0
Venue Category_Latin American Restaurant 0.04 0 0 0 0 0 0 0 0 0
Venue Category_Mediterranean Restaurant 0.016 0 0.045 0 0 0 0 0 0 0
Venue Category_Mexican Restaurant 0.018 0 0.023 0.36 0 0 0 0 0 0
Venue Category_Middle Eastern Restaurant 0.031 0 0 0.056 0 0 0 0 0 0.75
Venue Category_Modern European Restaurant 0.0041 0 0 0 0 0 0 0 0 0
Venue Category_Molecular Gastronomy Restaurant 0.0014 0 0 0 0 0 0 0 0 0
Venue Category_Moroccan Restaurant 0.0029 0 0 0 0 0 0 0 0 0
Venue Category_New American Restaurant 0.011 0 0 0 0 0 0 0 0 0
Venue Category_Portuguese Restaurant 0.019 0 0 0 0 0 0 0 0 0.25
Venue Category_Ramen Restaurant 0.014 0 0 0 0 0 0 0 0 0
Venue Category_Restaurant 0.11 0 0.15 0.2 0 0 0 0 0.67 0
Venue Category_Seafood Restaurant 0.035 0 0.023 0 0 0 0 0 0 0
Venue Category_Sushi Restaurant 0.064 0 0.018 0.24 0 0 0 0 0 0
Venue Category_Taiwanese Restaurant 0.0028 0 0 0 0 0 0 0 0 0
Venue Category_Thai Restaurant 0.051 0 0.023 0 0 0 0 0 0 0
Venue Category_Theme Restaurant 0.0012 0 0 0 0 0 0 0 0 0
Venue Category_Vegetarian / Vegan Restaurant 0.034 0 0 0.024 0 0 0 0 0 0
Venue Category_Vietnamese Restaurant 0.029 0 0 0 0 0 0 1 0 0
0 1 2 3 4 5 6 7 8 9
Neighborhood Cluster
1.0
0.8
0.6
0.4
0.2
0.0

7. Conclusion section

Conclusion can be taken from the previous heatmap, Dark color means that the restaurant cuisine is very common in the corresponding neighborhood cluster, for example if someone is willing to open a new Italian Restaurant, we can recommend him where not to open his new restaurant.



The occurrence of restaurants serving Italian cuisine is frequent in cluster no. 8. Therefore, it is unreasonable to open restaurants with such cuisine in a district qualified for this cluster.