Comparing North America's Biggest Cities

Roberto Cabrer

1 Introduction

Data Science techniques are becoming ubiquitous because of their potential to extract 'hidden' patterns in data. As we've seen in the courses in this specialization, diverse computational abilities are needed in order to extract valuable information from data.

As now it's our turn to propose an idea to use the acquired knowledge, for my Capstone project I'll use Foursquare location data from the largest city in each of the top 3 most populous countries in North America, to see if there are similarities among them. The cities to be compared are Mexico City in Mexico, New York in the United States and Toronto in Canada.

With the location (longitude and latitude) of the neighborhoods in this cities, we will extract venue information using Foursquare, that will be used as features in the clustering technique. The kind of information to be extracted is that of the presence of different business included in the Foursquare City Guide (`https://foursquare.com/city-guide`). Once the neighborhoods are clustered through the unsupervised machine learning technique, we will try to find similarities and dissimilarities between cities.

From the clustering of neighborhoods and comparison of these cities, huge amount of information can be extracted, we will focus on the expansion of a business or a brand from one of the cities to the other. Obtained insights can help stakeholders decide wether it is feasible to expand a known successful business into another city, trying to reduce risks by knowing if the cities are similar or not.

As the data is inherently related to its location, the results from this experiment can also give valuable information about the business offerings by neighborhood that can be used, for example, to decide where to locate a physical store in the case this is needed.

2 Data

The data I'll be needing is the location of neighborhoods in each of the cities, that is latitudes and longitudes, from where we will extract the venue information using Foursquare.

For this project I will use datasets from Mexico City, New York and Toronto. Datasets from Toronto and New York have been used previously in other labs,

Figure 1: Mexico City's data and the query used to extract the Postal Code data for Mexico City. The website belongs to the Mexican Postal Service.

besides these two datasets I will include data from Mexico City obtained from an HTML source. As these datasets come from different sources they will need slightly different cleansing procedures.

2.1 Mexico City data

The dataset for Mexico City is extracted from the Mexican Postal Service website (found at the url `https://www.correosdemexico.gob.mx/SSLServicios/ConsultaCP/Descarga.aspx`). This dataset contains Postal Codes that will help us retrieve location data for each neighborhood in Mexico City using the geocoder package.

The dataset comes as a HTML file with Borough, Postal Code and Neighborhood information as shown in Figure 1. Other information is included in the file, nonetheless it does not add value so it won't be used.

{"type":"FeatureCollection","totalFeatures":306,"features":[{"type":"Feature","id":"nyu_2451_34572.1","geometry":
{"type":"Point","coordinates":[-73.84720052054902,40.89470517661]},"geometry_name":"geom","properties":
{"name":"Wakefield","stacked":1,"annoline1":"Wakefield","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.84720052054902,40.89470517661,-73.84720052054902,40.89470517661]}},
{"type":"Feature","id":"nyu_2451_34572.2","geometry":{"type":"Point","coordinates":
[-73.82993910812398,40.87429419303012]},"geometry_name":"geom","properties":{"name":"Co-op City","stacked":2,"annoline1":"Co-
op","annoline2":"City","annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox":
[-73.82993910812398,40.87429419303012,-73.82993910812398,40.87429419303012]}},
{"type":"Feature","id":"nyu_2451_34572.3","geometry":{"type":"Point","coordinates":
[-73.82780644716412,40.887555677350775]},"geometry_name":"geom","properties":
{"name":"Eastchester","stacked":1,"annoline1":"Eastchester","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.82780644716412,40.887555677350775,-73.82780644716412,40.887555677350775]}},
{"type":"Feature","id":"nyu_2451_34572.4","geometry":{"type":"Point","coordinates":
[-73.90564259591682,40.89543742690383]},"geometry_name":"geom","properties":
{"name":"Fieldston","stacked":1,"annoline1":"Fieldston","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.90564259591682,40.89543742690383,-73.90564259591682,40.89543742690383]}},
{"type":"Feature","id":"nyu_2451_34572.5","geometry":{"type":"Point","coordinates":
[-73.9125854610857,40.890834493891305]},"geometry_name":"geom","properties":
{"name":"Riverdale","stacked":1,"annoline1":"Riverdale","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.9125854610857,40.890834493891305,-73.9125854610857,40.890834493891305]}},
{"type":"Feature","id":"nyu_2451_34572.6","geometry":{"type":"Point","coordinates":
[-73.90281798724604,40.88168737120521]},"geometry_name":"geom","properties":
{"name":"Kingsbridge","stacked":1,"annoline1":"Kingsbridge","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.90281798724604,40.88168737120521,-73.90281798724604,40.88168737120521]}},
{"type":"Feature","id":"nyu_2451_34572.7","geometry":{"type":"Point","coordinates":
[-73.91065965862981,40.87655077879964]},"geometry_name":"geom","properties":{"name":"Marble
Hill","stacked":2,"annoline1":"Marble","annoline2":"Hill","annoline3":null,"annoangle":0E-11,"borough":"Manhattan","bbox":
[-73.91065965862981,40.87655077879964,-73.91065965862981,40.87655077879964]}},
{"type":"Feature","id":"nyu_2451_34572.8","geometry":{"type":"Point","coordinates":
[-73.86731496814176,40.898272612213805]},"geometry_name":"geom","properties":
{"name":"Woodlawn","stacked":1,"annoline1":"Woodlawn","annoline2":null,"annoline3":null,"annoangle":0E-
11,"borough":"Bronx","bbox":[-73.86731496814176,40.898272612213805,-73.86731496814176,40.898272612213805]}},
{"type":"Feature","id":"nyu_2451_34572.9","geometry":{"type":"Point","coordinates":
[-73.8793907395681,40.87722415599446]},"geometry_name":"geom","properties":
{"name":"Norwood","stacked":1,"annoline1":"Norwood","annoline2":null,"annoline3":null,"annoangle":0E-11,"borough":"Bronx","bbox":
[-73.8793907395681,40.87722415599446,-73.8793907395681,40.87722415599446]}},
{"type":"Feature","id":"nyu_2451_34572.10","geometry":{"type":"Point","coordinates":

Figure 2: For New York City if the URL is accessed, the json file will pop-up presenting this structure.

2.2 New York City data

The data of New York City is stored in the IBM cloud and is requested from the URL `https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json`. The data comes in the form of a json file (Figure 2). In this file the neighborhood information with latitudes and longitudes can be found to create the desired dataframes.

2.3 City of Toronto data

Toronto's data is extracted from an article in wikipedia. The table contains Postal Codes, Borough name, and the Neighborhoods present in each Borough (Figure 3). This data can be found at `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`. and is extracted using Beautiful Soup, a package useful to parse data contained in HTML format to Python dictionaries that later can be used to create Pandas DataFrames.

As with the Mexico City data, the information extracted does not contain latitudes and longitudes for the neighborhoods, these have to be extracted through the postal codes using the geocoder package.

The final desired dataframes will contain columns 'Borough', 'Neighborhood', 'Latitude' and 'Longitude' as this is the information needed to do requests to the Foursquare API.

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M (except M0R and M7R) are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, a handful of individual special-purpose codes in the M0R FSA are assigned to Gateway Commercial Returns, 4567 Dixie Rd, Mississauga as a merchandise returns label for freepost returns to high-volume vendors such as Amazon and the Shopping Channel.[3]

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 Queen's Park (Ontario Provincial Government)	M8A Not assigned	M9A Etobicoke (Islington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B Not assigned	M3B North York (Don Mills) North	M4B East York (Parkview Hill / Woodbine Gardens)	M5B Downtown Toronto (Garden District, Ryerson)	M6B North York (Glencairn)	M7B Not assigned	M8B Not assigned	M9B Etobicoke (West Deane Park / Princess Gardens / Martin Grove / Islington / Cloverdale)
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	M2C Not assigned	M3C North York (Don Mills) South (Flemingdon Park)	M4C East York (Woodbine Heights)	M5C Downtown Toronto (St. James Town)	M6C York (Humewood-Cedarvale)	M7C Not assigned	M8C Not assigned	M9C Etobicoke (Eringate / Bloordale Gardens / Old Burnhamthorpe / Markland Wood)
M1E Scarborough (Guildwood / Morningside / West Hill)	M2E Not assigned	M3E Not assigned	M4E East Toronto (The Beaches)	M5E Downtown Toronto (Berczy Park)	M6E York (Caledonia-Fairbanks)	M7E Not assigned	M8E Not assigned	M9E Not assigned

Figure 3: Table present at the mentioned wikipedia URL. From here postal code information is extracted.

3 Methodology

The above presented data needs to be parsed into dataframes for its analysis. As the data from each city comes from different sources, each dataset needs a slightly different cleansing procedure.

3.1 Data cleansing

I decided to use data from previous labs in addition to the Mexico City postal codes that I found.

New York's data is the easiest to parse since the function to extract the information from the json file has been given earlier in the labs.

Mexico City and Toronto's data, obtained from the HTML file and the wikipedia webpage respectively, come in HTML format and needs to be parsed to dataframes, either by directly reading it with pandas (if the only information contained is the table) or by using Beautiful Soup to find the relevant tags to create dictionaries with the information that need to be in the dataframe. In both cases the data that we found is Borough, Neighborhood and Postal Code but not Latitude and Longitude as we need. We need to use the geocoder API to get the Latitude and Longitude information from each of the Postal Codes.

Additionally, for Mexico City's data, we discarded Boroughs at the periphery of the city to reduce dataframe size.

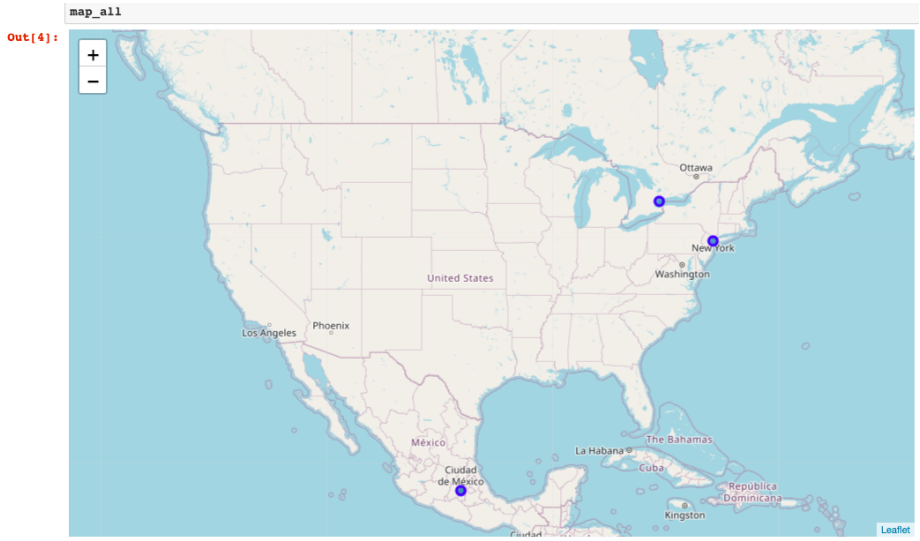

Figure 4: Map of North America with markers presenting the location of the cities to be compared.

3.2 Data visualization

Once the latitudes and longitudes for each neighborhood are obtained, we will have 3 dataframes, one for each of the cities, that contain Borough, Neighborhood, Latitude and Longitude information. With this we can start a simple exploratory analysis through visualization.

Firstly, to draw some context on the analyzed data, I decided to create a map of North America with the cities superimposed as markers (Figure 4). It can be seen that New York and Toronto are close to each other with respect to the distance between any of this cities and Mexico City.

Visualizing Mexico City (Figure 5), we can see that it is a landlocked city with neighborhoods expanding from an apparent city center.

New York (Figure 6) is a coastal city facing the Atlantic ocean conformed of what seems to be interconnected pieces of land.

Toronto (Figure 7) is a city surrounded by lakes with neighborhoods distributed in a mesh-like manner.

All the presented maps have been created using Folium, a library to create maps and the markers needed through latitude and longitude information. Once we obtain the results of the clustering technique, this kind of maps will be of great help to obtain insights and draw conclusions.

3.3 Foursquare Venue Information

Once the latitudes and longitudes for each neighborhood are obtained, we request venue information from Foursquare using their API. We can request sev-

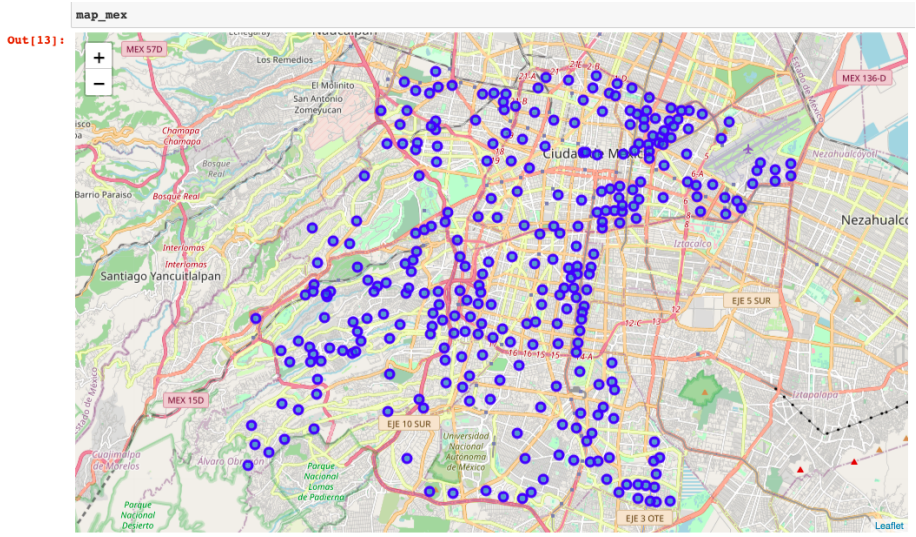

Figure 5: Mexico City's map with its neighborhoods superimposed, using location data.

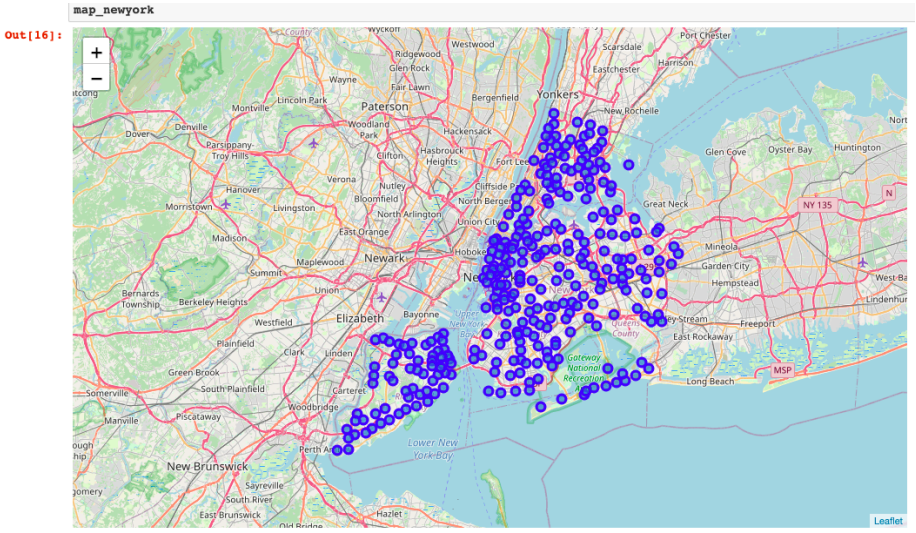

Figure 6: Map of New York's neighborhoods obtained from the IBM cloud URL.

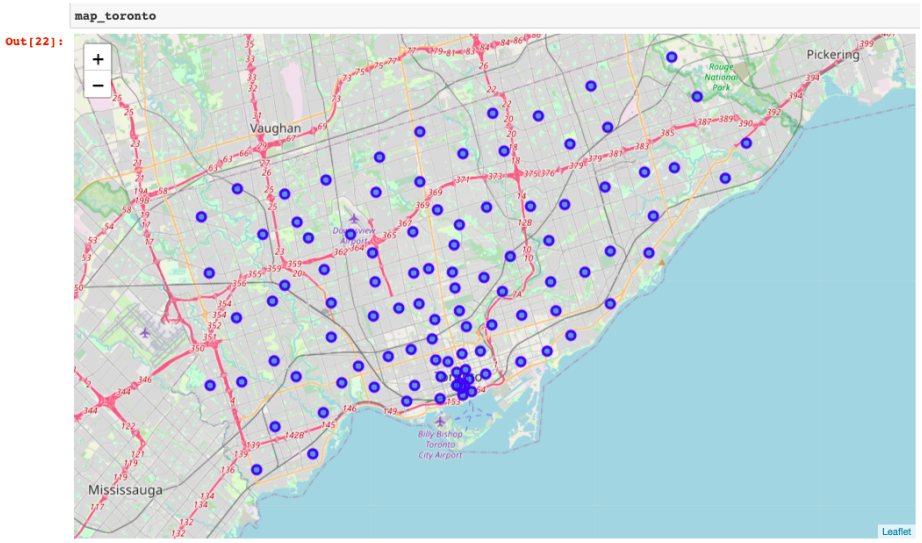

Figure 7: City of Toronto and the neighborhoods used in the dataframe, as markers.

eral types of information from the API, in this case we will use venue information, that, given a location (latitude-longitude pair) the API will return the names and types of venues for a certain desired radius.

It is important to remember that to make requests to the Foursquare API a Client ID and a Client Secret are needed, these are personal codes that can be obtained by creating a developer account in Foursquare.

After the requests are made, we will obtain a dataframe with a great quantity of information, since Foursquare can have information of hundreds of venues in the selected radius for each neighborhood. To ease the manipulation of this data, we encode it using one-hot encoding that will result in a sparse matrix of 0s and 1s, where a 1 represents that the neighborhood (stored as rows) contains that Venue Category (feature column) and 0 indicates that there is none of that class offerings.

Once having the encoding we can calculate the mean occurrence of each type of Venue Category, that are the values to be used as features in the clustering technique.

3.4 K-means clustering

The k-means clustering technique is an unsupervised machine learning algorithm that, given unlabeled data, has the ability to find some correspondence trough the concept of distance.

This technique starts by randomly assigning the desired number of cluster centers to the data, then calculates the distance from each center to each sample

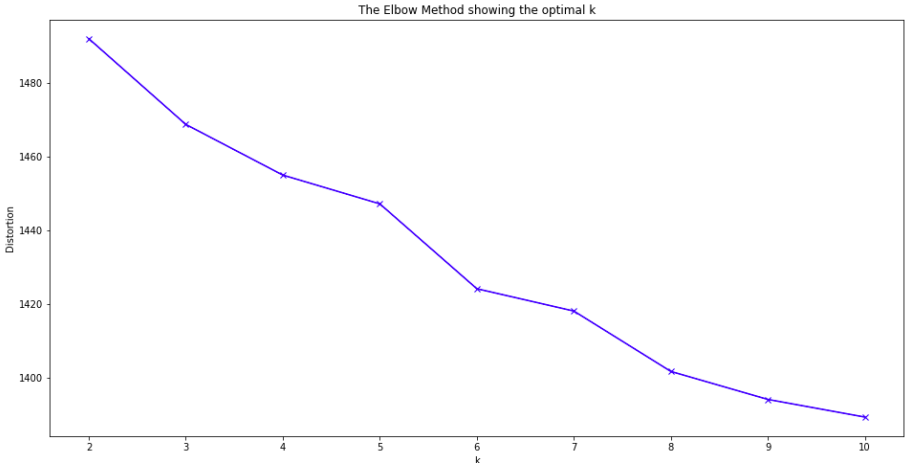

Figure 8: Elbow method for the selection of the number of clusters. Note that for this experiment there is no clear elbow.

in the dataset, the closest samples to each of the cluster centers are labeled using the index for that cluster center, we calculate the mean of the samples belonging to a cluster and make this the new cluster center. From where the process repeats iteratively until the cluster centers have minimal variation, reducing the variance among samples in the same cluster.

It is important to note that the neighborhoods in each cluster are similar to each other in terms of the features included in the dataset.

We will then compare the occurrence of the same clusters in each of the cities to see if they are similar or not.

As described, this technique has one setback, the number of cluster to be used is uncertain. We need to find parameters that help us decide how many clusters are needed.

3.4.1 The elbow method

One way of creating a clustering model with a 'good' number of clusters (that reduces the error but at the same time lets the model be general) is using the so called "elbow method".

The elbow method consist in creating different models with an ascending number of clusters in a given range, and ploting the error of the model (Figure 8), we then select the model where there is a change in the slope of the error function, graphically this looks like an elbow. This will ensure that the model reduces the error but at the same time has the most generalization capabilities.

Adding too much clusters will make the error smaller, but will make generalization of the model harder. So the elbow method tries to overcome this by using a model that fits well to the data and at the same time is not overfitted to it.

The elbow method gives an easy form of calculating the number of clusters to be used, nonetheless, sometimes it is hard to select the number of clusters since the error function can present more than one inflection point or "joint" or even none.

3.5 Cluster center closeness analysis

Another possible setback of k-means is that we do not know what there is in each of the clusters, what features are included to group neighborhoods in such a way. Cluster don't provide this information for themselves, it needs to be extracted through context.

In order to shed some light on the content of each cluster I decided to analyze the top venues in the neighborhoods whose features are "closest" (under a euclidean metric) to the cluster center they have been assigned to. This metric can be related to how "good" is the fit of a certain Neighborhood to its assigned cluster. Another reason to do this analysis is because we have a lot of samples with a lot of features that are difficult to map in a low dimensional space that lets us visually select the most representative samples of each cluster.

4 Results

In this experiment I decided to join the three cities dataframes to apply the clustering method to a single dataset. For this I concatenated datasets maintaining the columns and requested the Foursquare venue information, then with all the venues from the 3 cities I encoded the dataset to pass it through the clustering algorithm.

As seen earlier in Figure 8 for my experiment the elbow method was not very conclusive, as the error function presented various changes in slope. I decided to use 4 clusters since a slight change in slope is present in the function. Also a small number of clusters will let us maintain the model's simplicity and its ability to be generalizable.

Table 1 presents the results from the experiment:

Table 1: Occurrence by cluster in the joined dataset

Cluster Label	Quantity
0	299
2	235
3	126
1	45

Since we did the analysis using a joined dataset, we need to extract the information for each of the cities. The share in percentage of cluster by city, that is, how many of each cluster there are in relation to the total number of clusters by city, is shown in Table 2.

Table 2: Cluster share by city

Cluster	Mexico City [%]	New York [%]	Toronto [%]
0	7.213115	69.666667	68.316832
1	0.655738	12.333333	5.940594
2	76.065574	0.333333	1.980198
3	16.065574	17.666667	23.762376

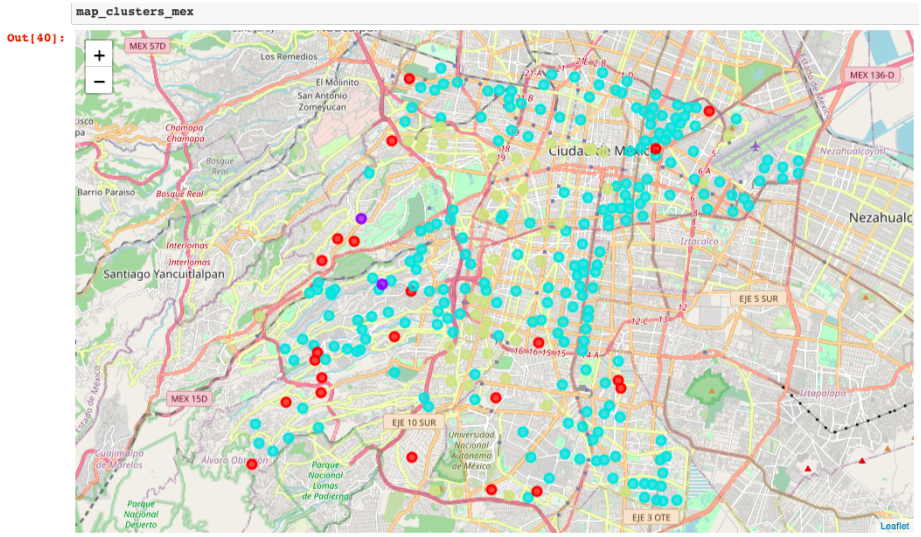

Figure 9: Mexico City's clusters obtained from the Foursquare Venue information data.

Analyzing the results by city we can see that Mexico City (Figure 9) has the biggest share of neighborhoods in cluster 2 (blue markers), some share of cluster 3 labels (yellow markers), slight presence of cluster 0 (red markers), and almost no presence of cluster 1 (purple markers).

In New York (Figure 10) we can mainly find neighborhoods labeled under cluster 0, a similar presence between cluster 3 and cluster 1 labels (although cluster 1 labels are more disperse) and a really small presence of cluster 2.

In Toronto (Figure 11) we can mainly find neighborhoods labeled under cluster 0, the next biggest share is that of neighborhoods in cluster 3, slight presence of cluster 1 and a pair of neighborhoods under cluster 2.

4.1 Results from the cluster center closeness analysis

Trying to obtain context of what features there are in each of the clusters I did the mentioned closeness analysis between cluster centers and neighborhoods. from where I saw that,

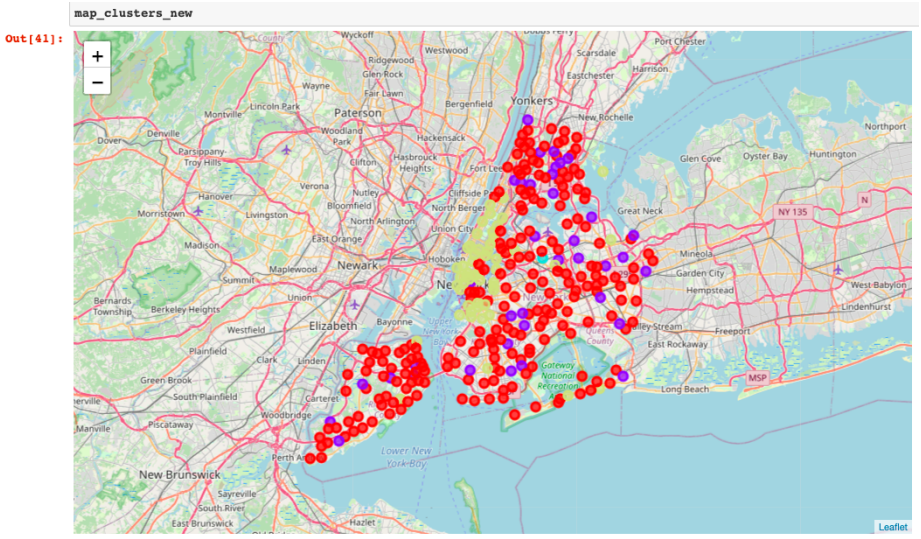

Figure 10: Obtained clusters for the City of New York.

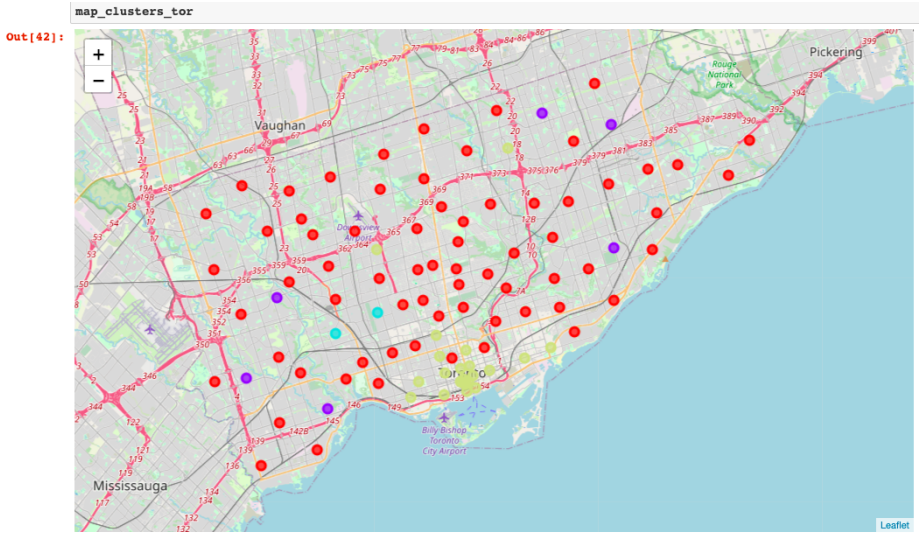

Figure 11: Toronto's clustering results.

- Clusters 0 groups neighborhoods with diverse food offerings and a diversity of businesses. This cluster seems to group residential zones.
- Cluster 1 has mainly asian food offerings and some other restaurants as well as a variety of shops and markets. This cluster may group commercial zones.
- Cluster 2 groups places where there is mainly Mexican food offerings and some businesses. This cluster seems to group mexican-like neighborhoods.
- Cluster 3 has high presence of coffee shops, bars and diverse fast-food like food offerings, as well as some stores. This cluster seems to group leisure districts.

This results are open to interpretation since this is not the "strongest" method mathematically speaking.

5 Discussion

The k-means result can be biased towards the Mexico City data since there were more datapoints in that dataframe. It is recommended to use a similar number of datapoints to avoid bias.

The elbow method gives an easy form of calculating the number of clusters to be used, nonetheless, sometimes it is hard to select the number of clusters since the error function can present more than one inflection point or none. This makes the method prone to subjectiveness.

6 Conclusions

About the methods:

- The obtained results contain helpful information about how the cities are conformed and offer an easy form of comparing them.
- The data and the methods used offer not only numeric advantages but also the capability to interpret this data physically, conforming a visual aid on where we can find different offerings.

About the cities:

- Cluster 0 seems to group residential zones in the northern cities.
- Judging by the food offerings present in cluster 1, this cluster might represent a much more diverse/multicultural community. This cluster is highly present in Toronto and New York

- Cluster 2 represents mainly neighborhoods in Mexico City as it contains many Mexican cuisine venues. This might be why there is a small presence of this cluster in the northern cities.
- Cluster 3 is a cluster that we can see in all of the cities, all of them seem to have places to spend some leisure time. This might be the point where we can find most similarity among all the analyzed cities.
- Toronto and New York might be so similar because of their geographical closeness, if we compare the distance from Mexico City to any of these other cities. Also, Mexico City is a landlocked city that might have other kind of structure and dynamics that make it different from the northern cities.