

# Comparing North America's Biggest Cities

Roberto Cabrera

IBM Applied Data Science Capstone

2021

# Table of Contents

Introduction

Data

Methodology

Results

Discussion

Conclusions

# Introduction

- ▶ Data Science techniques are becoming ubiquitous because of their potential to extract 'hidden' patterns in data.
- ▶ 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.

# Introduction

- ▶ With the location (longitude and latitude) of the neighborhoods in these cities, we will extract venue information using the Foursquare API.
- ▶ 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.

# Introduction

- ▶ 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 whether 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.
- ▶ The results from this experiment can give valuable information about the business offerings by neighborhood that can be used, for example, to decide where to locate a physical store.

## 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.
- ▶ I will use datasets from Mexico City, New York and Toronto.
- ▶ Datasets from Toronto and New York have been used previously in other labs,
- ▶ besides these two datasets I will include data from Mexico City obtained from an HTML source.

# Mexico City data

Extracted from the Mexican Postal Service (found at <https://www.correosdemexico.gob.mx/SSLServicios/ConsultaCP/Descarga.aspx>). This dataset contains Postal Codes that will help us retrieve location data using geocoder. The dataset comes as a HTML file with Borough, Postal Code and Neighborhood information.

SERVICIO POSTAL MEXICANO

### Consulta de Códigos Postales

Estado: Ciudad de México  
Municipio: T o d o s  
Asentamiento:  
Código Postal:

Ayuda Buscar

Código Postal	Asentamiento	Tipo de Asentamiento	Municipio	Estado	Ciudad	Clave de Oficina
07890	Nueva Tenochtitlán	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07000
07899	La Malinche	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07000
07900	Cuchilla del Tesoro	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07910	San Juan de Aragón VII Sección	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07918	San Juan de Aragón VI Sección	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07919	Ex Ejido San Juan de Aragón Sector 32	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07920	El Olivo	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07920	San Juan de Aragón	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07930	Indeco	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07939	Héroes de Chapultepec	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001
07940	Ex Ejido San Juan de Aragón Sector 33	Colonia	Gustavo A. Madero	Ciudad de México	Ciudad de México	07001

Figure: Mexico City's data and query used to extract the Postal Code data for Mexico City.

# New York City data

Stored in the IBM cloud at the URL [https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDriverSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork\\_data.json](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDriverSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json). The json file contains neighborhood information with latitudes and longitudes.

```
{"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": "Kawfield", "stacked": 1, "annoline1": "Kawfield", "annoline2": null, "annoline3": null, "annoangle": 0E-11, "borough": "Bronx", "bbox": [-73.84720052054902, 40.89470517661]}, {"type": "Feature", "id": "nyu_2451_34572.2", "geometry": {"type": "Point", "coordinates": [-73.82993910812398, 40.87429419303012]}, "geometry_name": "Co-op", "properties": {"name": "Co-op City", "stacked": 2, "annoline1": "Co-op", "annoline2": "City", "annoline3": null, "annoangle": 0E-11, "borough": "Bronx", "bbox": [-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]}, {"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]}, {"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]}, {"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]}, {"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 Hill", "annoline2": null, "annoline3": null, "annoangle": 0E-11, "borough": "Manhattan", "bbox": [-73.91065965862981, 40.87655077879964]}, {"type": "Feature", "id": "nyu_2451_34572.8", "geometry": {"type": "Point", "coordinates": [-73.86731496814176, 40.89827261213805]}, "geometry_name": "geom", "properties": {"name": "Woodlawn", "stacked": 1, "annoline1": "Woodlawn", "annoline2": null, "annoline3": null, "annoangle": 0E-11, "borough": "Bronx", "bbox": [-73.86731496814176, 40.89827261213805]}, {"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]}, {"type": "Feature", "id": "nyu_2451_34572.10", "geometry": {"type": "Point", "coordinates": [-73.8793907395681, 40.87722415599446]}}]
```

Figure: New York City json file structure.

# City of Toronto data

Extracted from an article in wikipedia [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and is extracted using Beautiful Soup. The table contains Postal Codes, Borough name, and the Neighborhoods present in each Borough.

## 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<sup>[1]</sup> via its applications for such smartphones as the iPhone and BlackBerry,<sup>[2]</sup> 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](#).<sup>[3]</sup>

M1A Not assigned	M2A Not assigned	M3A North York (Parkwood)	M4A North York (Victoria Village)	MSA Downtown Toronto (Regent Park / Hartforth)	MBA North York (Lawrence Manor / Lawrence Heights)	M7A Queen's Park (Ontario Provincial Government)	M8A Etobicoke (Bington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B Not assigned	M3B North York (Don Mills)	M4B East York (Parkview Hill / Woodbine Gardens)	M5B Downtown Toronto (Garden District, Ryerson)	M6B North York (Glencoe)	M7B Not assigned	M8B Etobicoke (West Donor Park / Princess Gardens / Martin Grove / Islington / Clarendale)
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	M2C Not assigned	M3C North York (Don Mills)	M4C East York (Woodbine Heights)	M5C Downtown Toronto (St. James Town)	M6C York (Humewood-Cedaville)	M7C Not assigned	M8C Etobicoke (Englefield / Bloordale Gardens / Old Bantamthorpe / Markland Wood)
M1E Scarborough (Glenwood / Morningside / West Hill)	M2E Not assigned	M3E Not assigned	M4E East Toronto (The Beaches)	M5E Downtown Toronto (Benzby Park)	M6E York (Caledonia-Fairbanks)	M7E Not assigned	M8E Not assigned

Figure: Table with information from Toronto.

## Data

- ▶ The data needs to be parsed into dataframes for its analysis.
- ▶ Mexico City and Toronto's data 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.
- ▶ 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.

# Visualizing Neighborhoods in Cities

It can be seen that New York and Toronto are close to each other with respect to the distance between any of these cities and Mexico City.

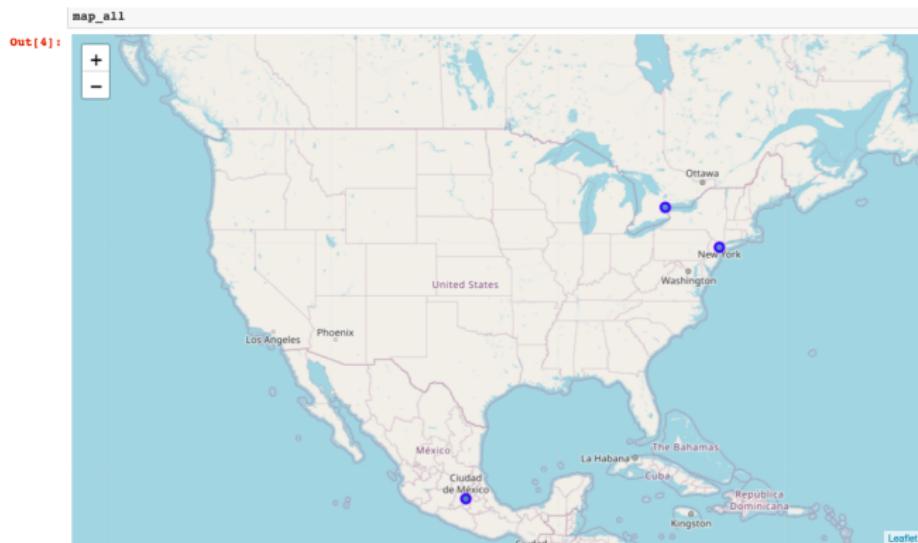
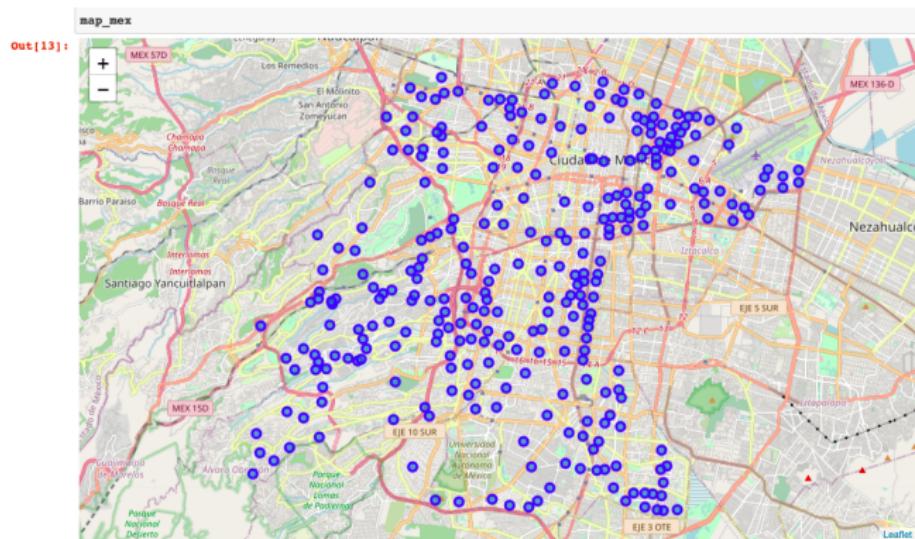


Figure: Map of North America with markers presenting the location of the cities to be compared.

# Mexico City

Landlocked city with neighborhoods expanding from an apparent city center. We discarded Boroughs at the periphery of the city to reduce dataframe size.



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

# New York

Coastal city facing the Atlantic ocean conformed of what seems to be interconnected pieces of land.

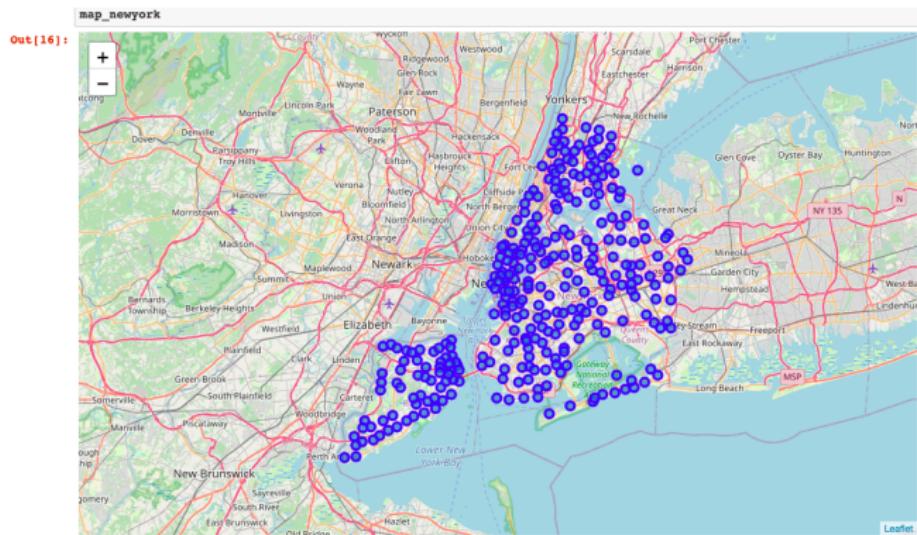


Figure: Map of New York's neighborhoods obtained from the IBM cloud URL.

# Toronto's Neighborhoods

City surrounded by lakes with neighborhoods distributed in a mesh-like manner.

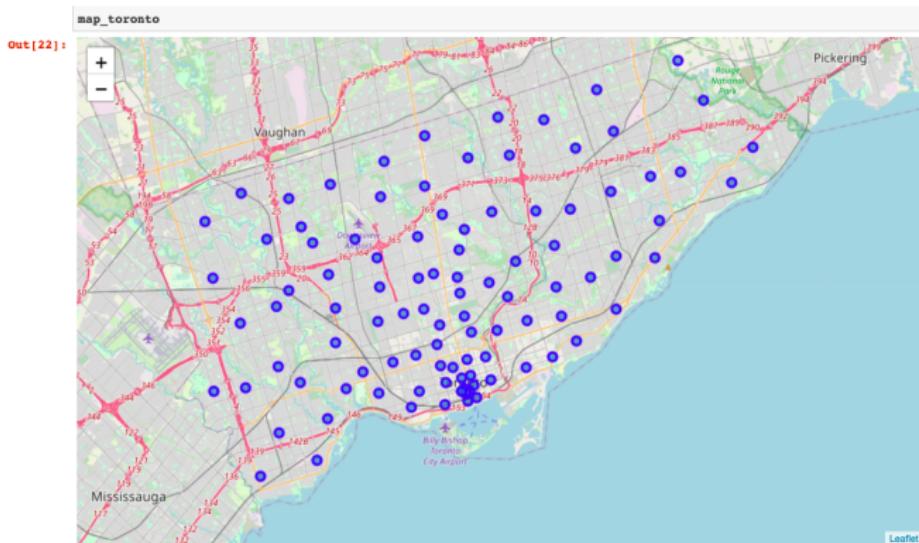


Figure: City of Toronto and the neighborhoods used in the dataframe, as markers.

## 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 several 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.

## Foursquare Venue Information

- ▶ 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.

# K-means clustering

Unsupervised machine learning algorithm that, given unlabeled data, has the ability to find some correspondence through the concept of distance.

## K-means clustering

Steps:

- ▶ Randomly assigning the desired number of cluster centers to the data,
- ▶ Calculates the distance from each center to each sample in the dataset,
- ▶ The closest samples to each of the cluster centers are labeled using the index for that cluster center,
- ▶ The mean of the samples belonging to a cluster and make this the new cluster center.

The process repeats iteratively until the cluster centers have minimal variation, reducing the variance among samples in the same cluster.

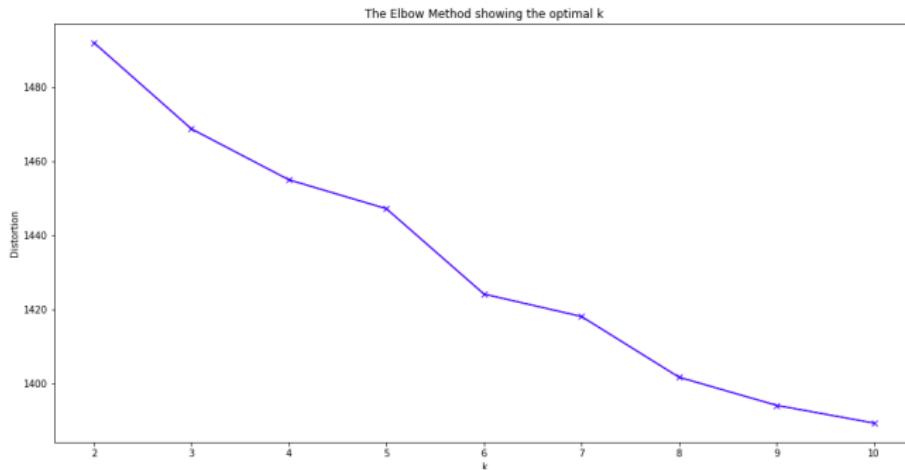
## K-means clustering

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.

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.

## The elbow method

The method consist in creating different models with an ascending number of clusters in a given range, then ploting the error of each model, then selecting the model where there is a change in the slope of the error function,



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

## The elbow method

- ▶ 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.

## Results

- ▶ Joined the three cities dataframes to apply the clustering method to a single dataset.
- ▶ 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.
- ▶ The elbow method was not very conclusive, I used 4 clusters as a slight change in slope is present in the function around  $k = 4$ .

# Results

The results from the clustering experiment were:

**Table:** Occurrence by cluster in the joined dataset

Cluster Label	Quantity
0	299
2	235
3	126
1	45

# Results

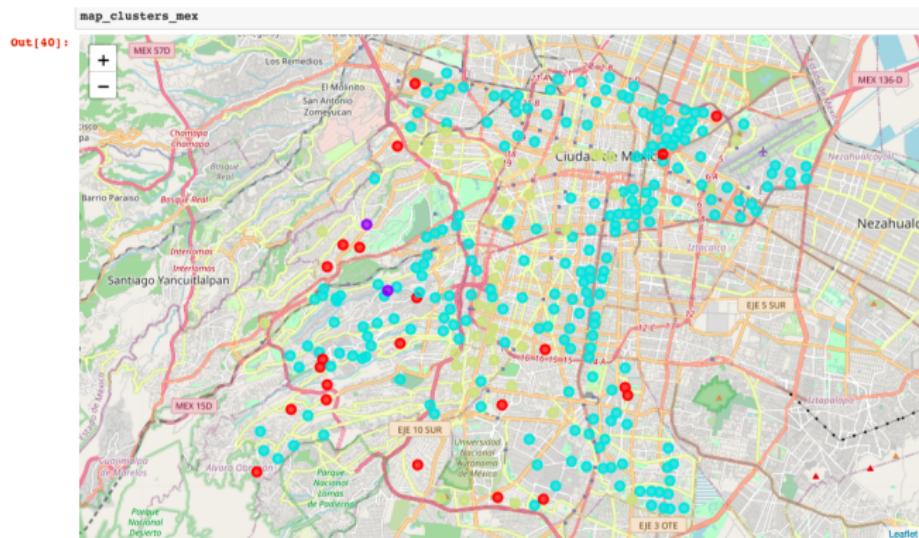
The share in percentage of cluster by city, or how many of each cluster there are in relation to the total number of clusters by city, is:

Table: 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

## Results

Mexico City 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).



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

## Results

In New York 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.

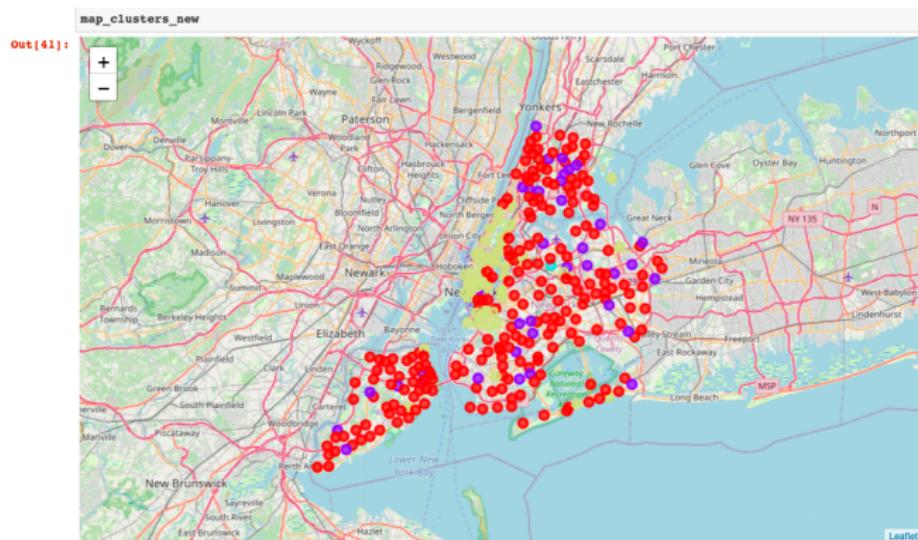


Figure: Obtained clusters for the City of New York.

## Results

In Toronto 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.

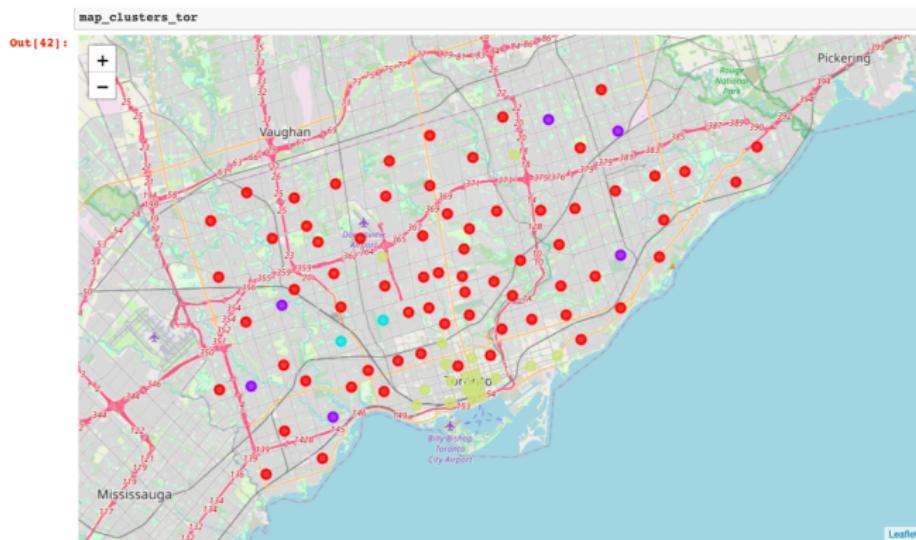


Figure: Toronto's clustering results.

## 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,

- ▶ 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.

## 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.

# Conclusions

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.

## Conclusions

- ▶ 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.

# 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 can we find different offerings.