# Coursera Capstone Project - The Battle of Neighborhoods

As a part of the Data Science Capstone Project, I have chosen to compare the cities of Lima-Peru (where I live) and Toronto-Canada (which was used as a case example during the course), and explore the neighborhoods of these two cities, using the Foursquare location data.

A GitHub link to the Notebook with code and graphics can be found at the end of this report.

## Introduction

The objective of this project is to compare the neighborhoods of Lima and Toronto cities and determine how similar or dissimilar they are.

**Toronto** is the provincial capital of Ontario and the most populous city in Canada. It is a center of business, finance, arts, and culture; and is recognized as one of the most multicultural and cosmopolitan cities in the world.

As Toronto is a known global city, let's take an overview of Lima city.

**Lima** is the capital and the largest city and the most populous metropolitan area of Peru, with a population of 8,574,974 and a population density of 3,208.8 inhabitants per square kilometer.

From Wikipedia ( https://en.wikipedia.org/wiki/Lima ):

> *Economy*: Lima is the country's industrial and financial center and one of Latin America's most important financial centers, home to many national companies and hotels.
>
> *Food*: Lima is known as the *Gastronomical Capital of the Americas*. In the 21st century, its restaurants became recognized internationally. Since 2011, several Lima restaurants have been recognized as among The World's 50 Best Restaurants. Lima is the Peruvian city with the greatest variety and where different dishes representing South American cuisine can be found.
>
> *Subdivisions*: Lima is made up of thirty densely populated **districts**, each headed by a local mayor and the Mayor of Lima.
>
> The city's historic center is located in the Cercado de Lima district, locally known as simply *Lima*, or as "El Centro" ("*Downtown*") and it is home to most of the vestiges the colonial past, the Presidential Palace, the Metropolitan Municipality and, Chinatown and dozens of hotels, some operating and some defunct, that cater to the national and international elite.
>
> The upscale *San Isidro* District is the city's financial center. It is home to politicians and celebrities. San Isidro has parks, including Parque El Olivar, which is home to olive trees

imported from Spain during the seventeenth century. The Lima Golf Club, a prominent golf club, is located within the district.

Another upscale district is *Miraflores*, which has luxury hotels, shops and restaurants. Miraflores has parks and green areas, more than most other districts. Larcomar, a shopping mall and entertainment center built on cliffs overlooking the Pacific Ocean, featuring bars, dance clubs, movie theaters, cafes, shops, boutiques and galleries, is also located in this district. Nightlife, shopping and entertainment center around Parque Kennedy, a park in the heart of Miraflores.

*La Molina, San Borja, Pueblo Libre, Santiago de Surco* -home to the American Embassy and the exclusive Club Polo Lima-, and *Jesús María* – home to one of the largest parks in Lima, El Campo De Marte – are the other five wealthy districts.

*Barranco*, which borders Miraflores by the Pacific Ocean, is the city's bohemian district, home or once home of writers and intellectuals including Mario Vargas Llosa, Chabuca Granda and Alfredo Bryce Echenique. This district has restaurants, music venues called "peñas" featuring the traditional folk music of coastal Peru and Victorian-style chalets. Along with Miraflores it serves as the home to the foreign nightlife scene.

The previous information from Lima will be relevant when performing the analysis in further detail.

Some of the target audience that may be interested in this project are:

- companies willing to expand their businesses in locations with similar characteristics;
- travel companies that wish to segment boroughs in different cities around the world based on their similarities;
- other data scientist learners who want to expand their knowledge on how to apply a data science methodology to solve different problems.

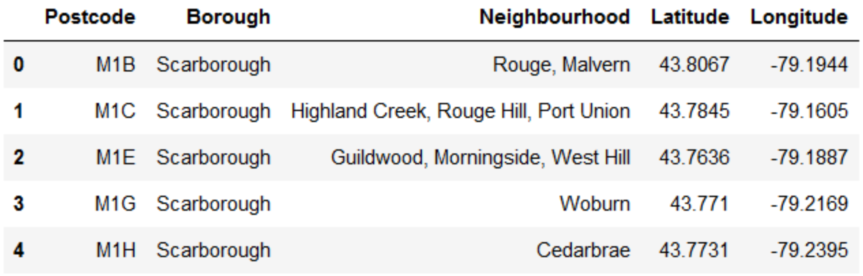
## Information needed and source of data

To be able to compare the two cities described previously, it is necessary to collect data of boroughs, neighborhoods and their coordinates:

- for the city of Toronto, I will use the list of postal codes where the first letter is M, to get the boroughs and neighborhoods. The data source comes from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  And for the coordinates of each location in Canada, the data will be extracted from http://cocl.us/Geospatial_data
- the city of Lima is divided into districts (which are equivalent to the boroughs in Toronto), and each district is broken down in what is called 'urbanizations' (which are equivalent to the neighborhoods in Toronto), although the coordinates for the urbanizations are not easily available. For Lima, coordinates will be for the districts, and the data source will come from https://www.antipodas.net/coordenadaspais/peru/lima.php#coordenadas

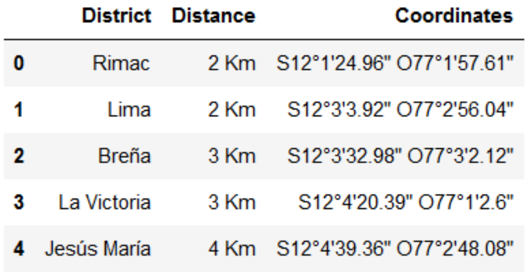
- to scrape information from web pages, the Beautiful Soup library will be used;
- to get the most common venues given the coordinates of each location, the Foursquare API will be used.

After web scraping the URL with Toronto postal codes using the Beautiful Soup library, and joining it with the Geospatial_data, the following sample dataframe was obtained, with coordinates for each neighborhood:



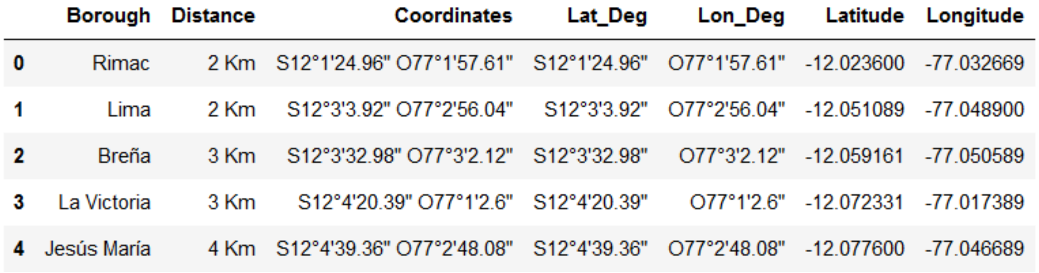
|  | Postcode | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Rouge, Malvern | 43.8067 | -79.1944 |
| 1 | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | 43.7845 | -79.1605 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.7636 | -79.1887 |
| 3 | M1G | Scarborough | Woburn | 43.771 | -79.2169 |
| 4 | M1H | Scarborough | Cedarbrae | 43.7731 | -79.2395 |

And after web scraping the URL with Lima coordinates using the Beautiful Soup library, the following sample dataframe was obtained, with coordinates in degrees for each district:



|  | District | Distance | Coordinates |
|---|---|---|---|
| 0 | Rimac | 2 Km | S12°1'24.96" O77°1'57.61" |
| 1 | Lima | 2 Km | S12°3'3.92" O77°2'56.04" |
| 2 | Breña | 3 Km | S12°3'32.98" O77°3'2.12" |
| 3 | La Victoria | 3 Km | S12°4'20.39" O77°1'2.6" |
| 4 | Jesús María | 4 Km | S12°4'39.36" O77°2'48.08" |

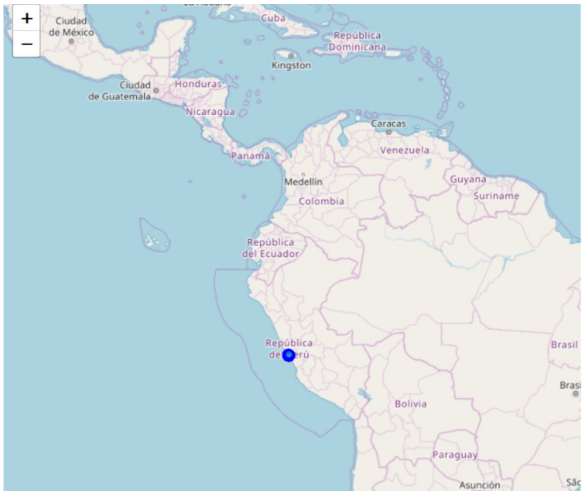
As the Lima coordinates were in degrees in a single column, it was necessary to split them, and then use a function to transform it to decimals. Also, the District column was renamed to Borough:

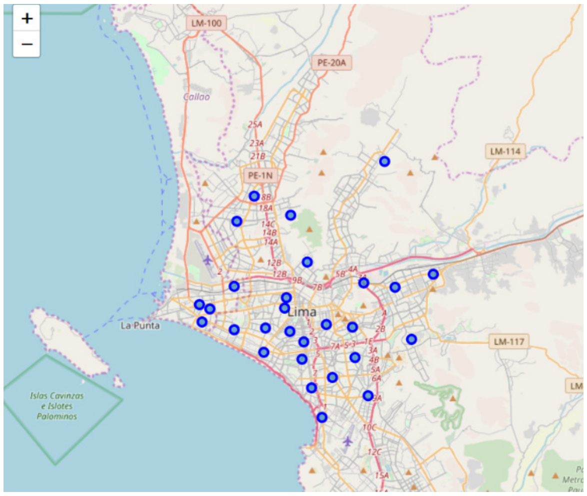


|  | Borough | Distance | Coordinates | Lat_Deg | Lon_Deg | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| 0 | Rimac | 2 Km | S12°1'24.96" O77°1'57.61" | S12°1'24.96" | O77°1'57.61" | -12.023600 | -77.032669 |
| 1 | Lima | 2 Km | S12°3'3.92" O77°2'56.04" | S12°3'3.92" | O77°2'56.04" | -12.051089 | -77.048900 |
| 2 | Breña | 3 Km | S12°3'32.98" O77°3'2.12" | S12°3'32.98" | O77°3'2.12" | -12.059161 | -77.050589 |
| 3 | La Victoria | 3 Km | S12°4'20.39" O77°1'2.6" | S12°4'20.39" | O77°1'2.6" | -12.072331 | -77.017389 |
| 4 | Jesús María | 4 Km | S12°4'39.36" O77°2'48.08" | S12°4'39.36" | O77°2'48.08" | -12.077600 | -77.046689 |

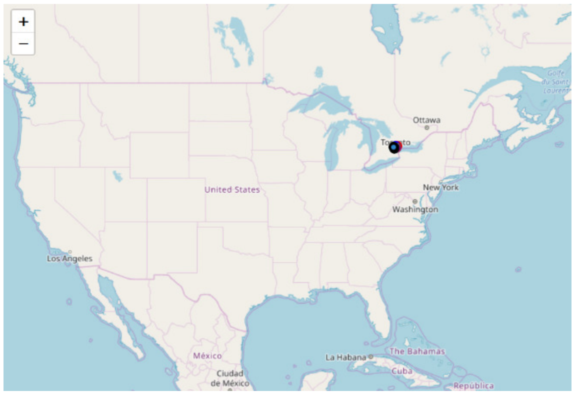
At this point, I used Folium to visualize each of the locations from Lima and Toronto in the map. As I am familiar with the boroughs of Lima but not from Toronto, and to make it easier for me to understand how neighborhoods were distributed in Toronto, I assigned different colors to each borough.

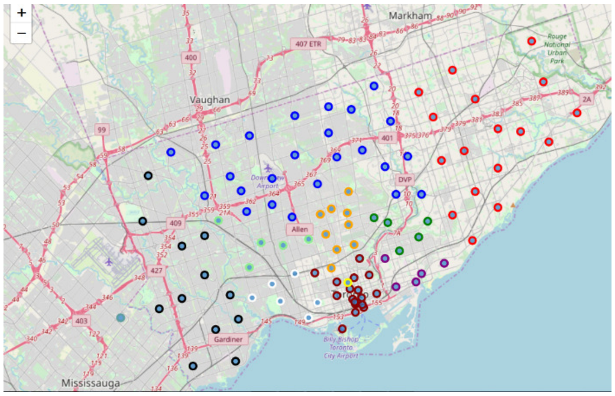
Map at the left side shows the city within the country for reference.
Map at the right shows the boroughs, and in addition for Toronto, the neighborhoods.





For Lima, and to simplify the analysis, I chose 11 of the 28 boroughs represented in the map, which are the most relevant districts in terms of tourism, business, and finance centers; those districts are located mainly at the south and east from the Downtown Lima district.



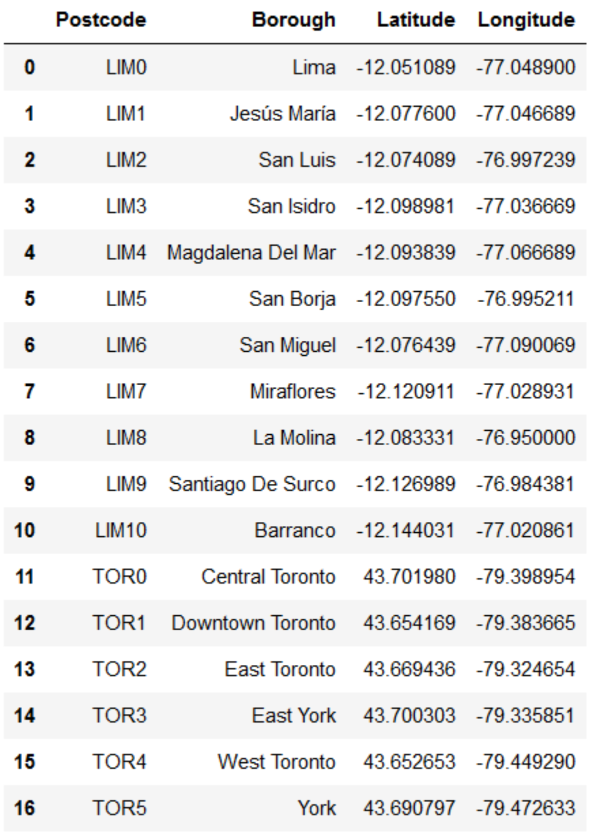


For Toronto, and to simplify the analysis too, I chose boroughs closer to the Downtown Toronto, which based on the map distribution would be West Toronto, East Toronto, Central Toronto, York and East York.

As it would not be equivalent to compare boroughs from Lima with neighborhoods from Toronto, for the purpose of this project I decided to compare boroughs from each city.

For that purpose, I used the Geocoder library to obtain the coordinates of the selected boroughs; however, I got strange coordinates for a couple of Boroughs in Toronto, so I decided to use an average of coordinates instead.

With the selection of the boroughs from Lima and Toronto, and merging both dataframes, the following dataset was obtained. A dummy postcode was generated for Lima boroughs (starting with LIM) and for Toronto boroughs (starting with TOR).



| | Postcode | Borough | Latitude | Longitude |
|---|---|---|---|---|
| **0** | LIM0 | Lima | -12.051089 | -77.048900 |
| **1** | LIM1 | Jesús María | -12.077600 | -77.046689 |
| **2** | LIM2 | San Luis | -12.074089 | -76.997239 |
| **3** | LIM3 | San Isidro | -12.098981 | -77.036669 |
| **4** | LIM4 | Magdalena Del Mar | -12.093839 | -77.066689 |
| **5** | LIM5 | San Borja | -12.097550 | -76.995211 |
| **6** | LIM6 | San Miguel | -12.076439 | -77.090069 |
| **7** | LIM7 | Miraflores | -12.120911 | -77.028931 |
| **8** | LIM8 | La Molina | -12.083331 | -76.950000 |
| **9** | LIM9 | Santiago De Surco | -12.126989 | -76.984381 |
| **10** | LIM10 | Barranco | -12.144031 | -77.020861 |
| **11** | TOR0 | Central Toronto | 43.701980 | -79.398954 |
| **12** | TOR1 | Downtown Toronto | 43.654169 | -79.383665 |
| **13** | TOR2 | East Toronto | 43.669436 | -79.324654 |
| **14** | TOR3 | East York | 43.700303 | -79.335851 |
| **15** | TOR4 | West Toronto | 43.652653 | -79.449290 |
| **16** | TOR5 | York | 43.690797 | -79.472633 |

## Methodology

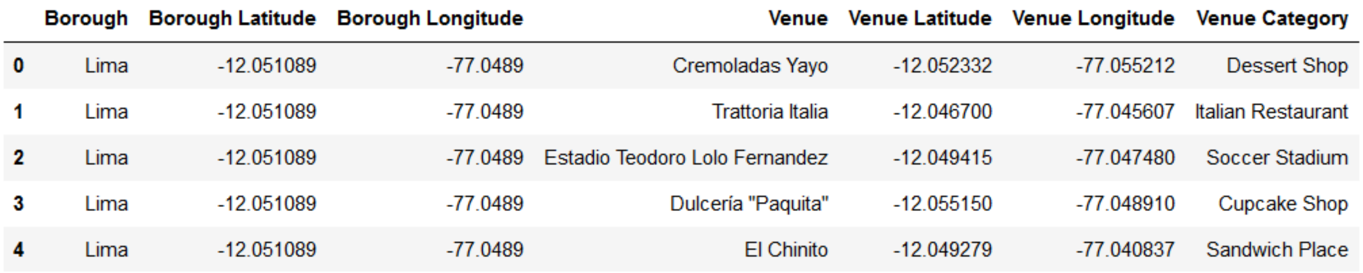
In the previous sections, I defined the dataset that will be used to compare the boroughs from Lima and Toronto.

### Foursquare API

In this step, and with the help of the Foursquare API which provides up-to-date location data, I retrieved information of venues around the borough locations. Among other parameters, the number of venues to retrieve by location was set to 100, and the radius to search venues within this distance (in meters) from the specified location was also set.

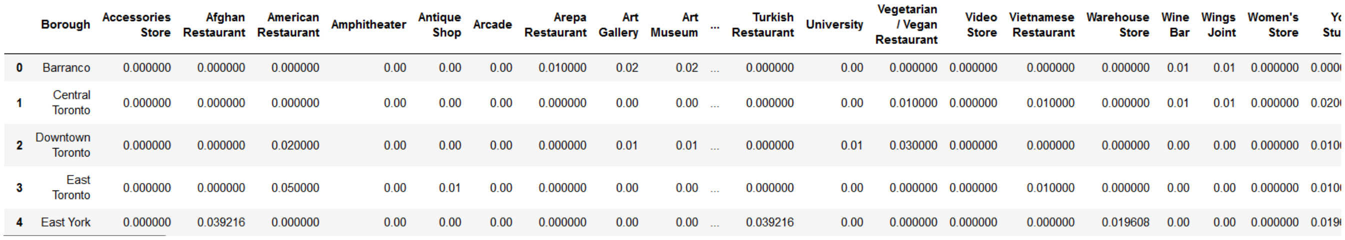
To estimate the radius, I built a function to calculate the minimum distance between the boroughs in each city, and then get an average of these minimum distances. The average for Lima was 3,0 km and the average for Toronto was 4,4 km. For the purpose of this project, I decided to choose the smallest average distance, that is 3.0 km, therefore a radius of 1.5 km (1500 meters) from each location.

After extracting the relevant information from each venue from the response provided by the Foursquare API, the data was put in a dataframe as shown below:



| | Borough | Borough Latitude | Borough Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Lima | -12.051089 | -77.0489 | Cremoladas Yayo | -12.052332 | -77.055212 | Dessert Shop |
| **1** | Lima | -12.051089 | -77.0489 | Trattoria Italia | -12.046700 | -77.045607 | Italian Restaurant |
| **2** | Lima | -12.051089 | -77.0489 | Estadio Teodoro Lolo Fernandez | -12.049415 | -77.047480 | Soccer Stadium |
| **3** | Lima | -12.051089 | -77.0489 | Dulcería "Paquita" | -12.055150 | -77.048910 | Cupcake Shop |
| **4** | Lima | -12.051089 | -77.0489 | El Chinito | -12.049279 | -77.040837 | Sandwich Place |

The quantity of unique venue categories was 227 considering all boroughs from Lima and Toronto. To analyze each borough further, the data was hot-encoded for each venue category and then grouped by each borough; and finally, the mean of the one-hot encoded venue categories was obtained, which resulted in the following dataset:



| | Borough | Accessories Store | Afghan Restaurant | American Restaurant | Amphitheater | Antique Shop | Arcade | Arepa Restaurant | Art Gallery | Art Museum | ... | Turkish Restaurant | University | Vegetarian / Vegan Restaurant | Video Store | Vietnamese Restaurant | Warehouse Store | Wine Bar | Wings Joint | Women's Store | Yc Stu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Barranco | 0.000000 | 0.000000 | 0.000000 | 0.00 | 0.00 | 0.00 | 0.010000 | 0.02 | 0.02 | ... | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.01 | 0.01 | 0.000000 | 0.000 |
| **1** | Central Toronto | 0.000000 | 0.000000 | 0.000000 | 0.00 | 0.00 | 0.00 | 0.000000 | 0.00 | 0.00 | ... | 0.000000 | 0.00 | 0.010000 | 0.000000 | 0.010000 | 0.000000 | 0.01 | 0.01 | 0.000000 | 0.020 |
| **2** | Downtown Toronto | 0.000000 | 0.000000 | 0.020000 | 0.00 | 0.00 | 0.00 | 0.000000 | 0.01 | 0.01 | ... | 0.000000 | 0.01 | 0.030000 | 0.000000 | 0.000000 | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.010 |
| **3** | East Toronto | 0.000000 | 0.000000 | 0.050000 | 0.00 | 0.01 | 0.00 | 0.000000 | 0.00 | 0.00 | ... | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.010000 | 0.000000 | 0.00 | 0.00 | 0.000000 | 0.010 |
| **4** | East York | 0.000000 | 0.039216 | 0.000000 | 0.00 | 0.00 | 0.00 | 0.000000 | 0.00 | 0.00 | ... | 0.039216 | 0.00 | 0.000000 | 0.000000 | 0.000000 | 0.019608 | 0.00 | 0.00 | 0.000000 | 0.019 |

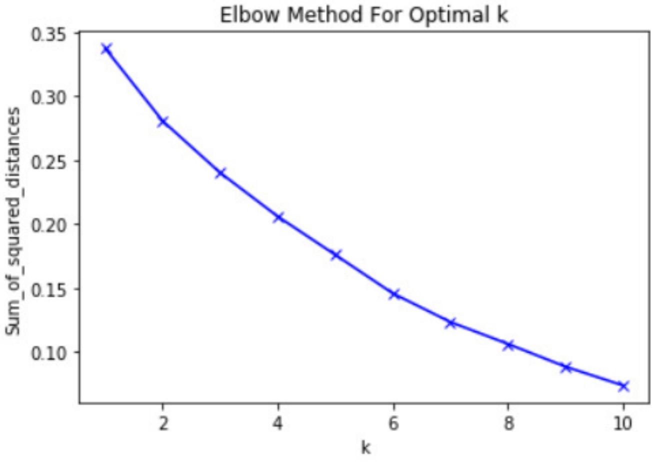
## Clustering

At this point we can make use of an algorithm to compare similar boroughs in Lima and Toronto.

The K-means clustering is one of the simplest and popular unsupervised machine learning algorithms, with the objective to group similar data points together in clusters and discover underlying patterns.

The K-means algorithm was selected to cluster the boroughs.

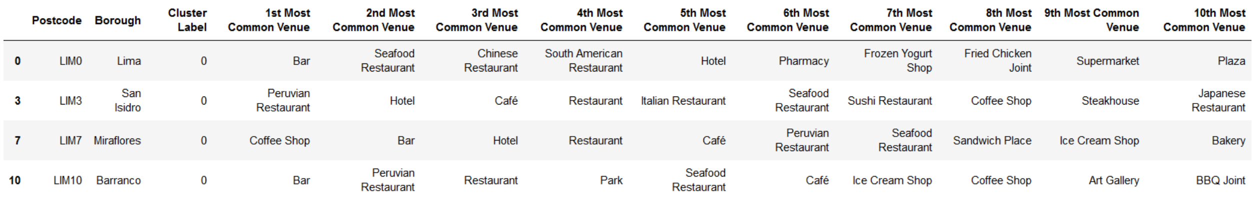
The number of clusters must be defined in order to use the K-means algorithm. In order to do that, I will use the Elbow method to determine the optimal number of clusters to be used in the algorithm. The chart below shows the results of the Elbow method for a range of 1 to 10 clusters (k).



The ‘elbow’ in this method cannot always be unambiguously identified. We can visualize an initial ‘elbow’ for k=2, although not a definitive number can be set. For the purpose of this project and being only 17 the total number of boroughs selected from Lima and Toronto, I decided to choose k=3 clusters, instead of k=2 which would have not given too much room for analysis.

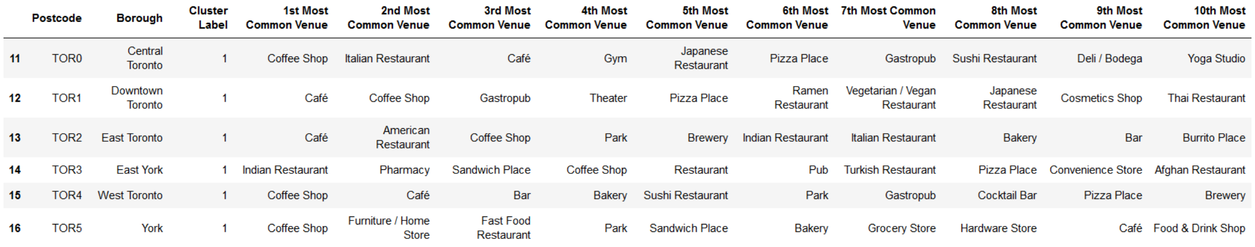
Let’s see how the boroughs were clustered using the K-means algorithm. The charts also show the most common venues in each borough:

**Cluster 0:**



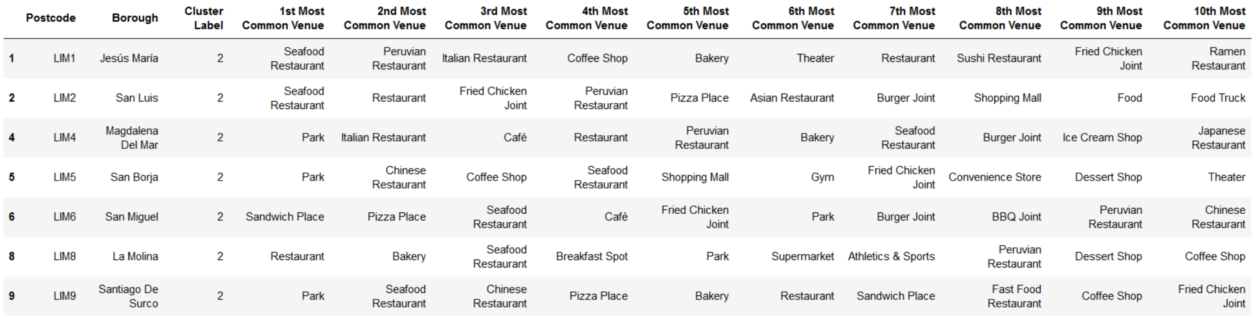
| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LIM0 | Lima | 0 | Bar | Seafood Restaurant | Chinese Restaurant | South American Restaurant | Hotel | Pharmacy | Frozen Yogurt Shop | Fried Chicken Joint | Supermarket | Plaza |
| 3 | LIM3 | San Isidro | 0 | Peruvian Restaurant | Hotel | Café | Restaurant | Italian Restaurant | Seafood Restaurant | Sushi Restaurant | Coffee Shop | Steakhouse | Japanese Restaurant |
| 7 | LIM7 | Miraflores | 0 | Coffee Shop | Bar | Hotel | Restaurant | Café | Peruvian Restaurant | Seafood Restaurant | Sandwich Place | Ice Cream Shop | Bakery |
| 10 | LIM10 | Barranco | 0 | Bar | Peruvian Restaurant | Restaurant | Park | Seafood Restaurant | Café | Ice Cream Shop | Coffee Shop | Art Gallery | BBQ Joint |

**Cluster 1:**



| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | TOR0 | Central Toronto | 1 | Coffee Shop | Italian Restaurant | Café | Gym | Japanese Restaurant | Pizza Place | Gastropub | Sushi Restaurant | Deli / Bodega | Yoga Studio |
| 12 | TOR1 | Downtown Toronto | 1 | Café | Coffee Shop | Gastropub | Theater | Pizza Place | Ramen Restaurant | Vegetarian / Vegan Restaurant | Japanese Restaurant | Cosmetics Shop | Thai Restaurant |
| 13 | TOR2 | East Toronto | 1 | Café | American Restaurant | Coffee Shop | Park | Brewery | Indian Restaurant | Italian Restaurant | Bakery | Bar | Burrito Place |
| 14 | TOR3 | East York | 1 | Indian Restaurant | Pharmacy | Sandwich Place | Coffee Shop | Restaurant | Pub | Turkish Restaurant | Pizza Place | Convenience Store | Afghan Restaurant |
| 15 | TOR4 | West Toronto | 1 | Coffee Shop | Café | Bar | Bakery | Sushi Restaurant | Park | Gastropub | Cocktail Bar | Pizza Place | Brewery |
| 16 | TOR5 | York | 1 | Coffee Shop | Furniture / Home Store | Fast Food Restaurant | Park | Sandwich Place | Bakery | Grocery Store | Hardware Store | Café | Food & Drink Shop |

**Cluster 2:**

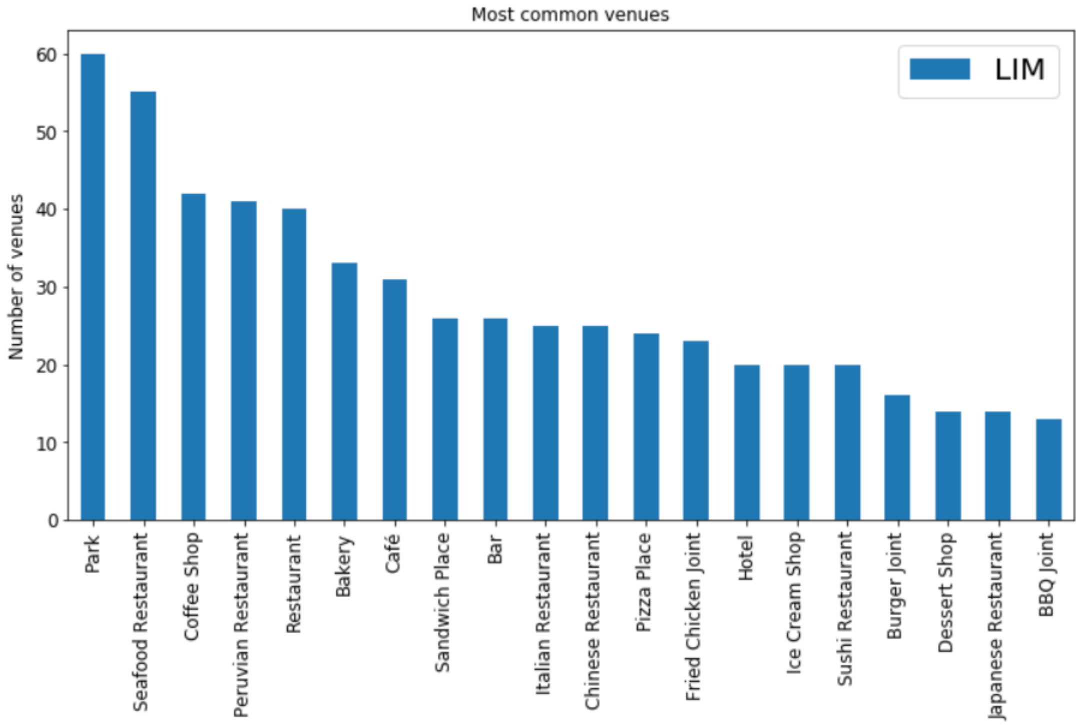


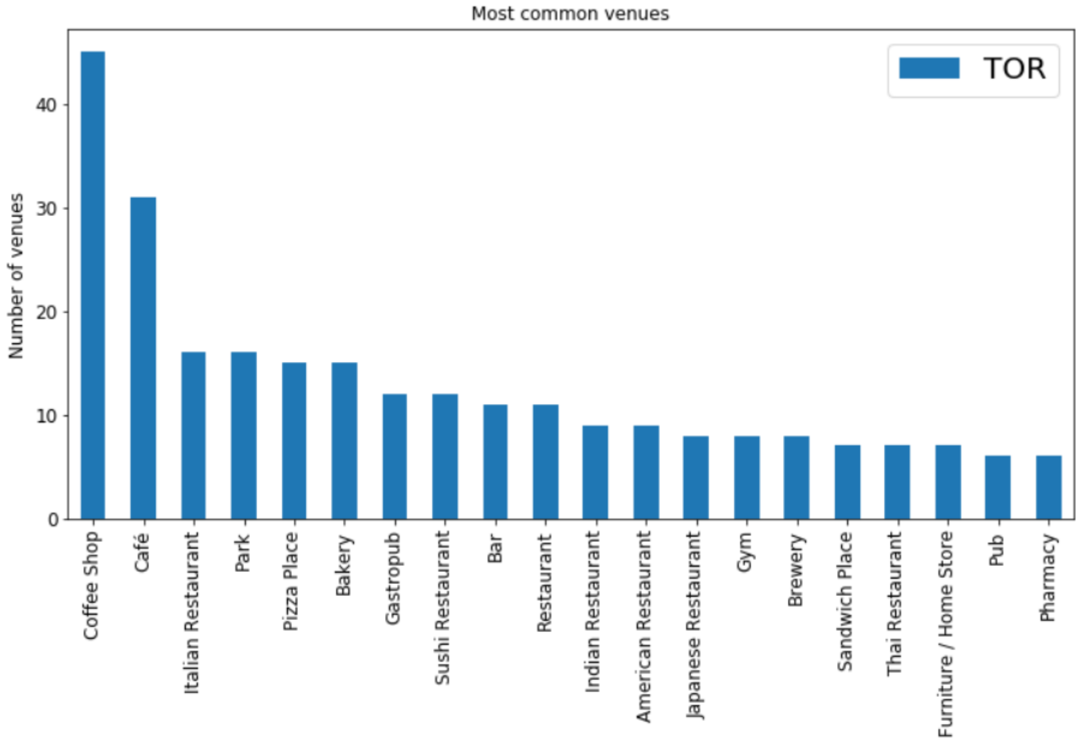
| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LIM1 | Jesús María | 2 | Seafood Restaurant | Peruvian Restaurant | Italian Restaurant | Coffee Shop | Bakery | Theater | Restaurant | Sushi Restaurant | Fried Chicken Joint | Ramen Restaurant |
| 2 | LIM2 | San Luis | 2 | Seafood Restaurant | Restaurant | Fried Chicken Joint | Peruvian Restaurant | Pizza Place | Asian Restaurant | Burger Joint | Shopping Mall | Food | Food Truck |
| 4 | LIM4 | Magdalena Del Mar | 2 | Park | Italian Restaurant | Café | Restaurant | Peruvian Restaurant | Bakery | Seafood Restaurant | Burger Joint | Ice Cream Shop | Japanese Restaurant |
| 5 | LIM5 | San Borja | 2 | Park | Chinese Restaurant | Coffee Shop | Seafood Restaurant | Shopping Mall | Gym | Fried Chicken Joint | Convenience Store | Dessert Shop | Theater |
| 6 | LIM6 | San Miguel | 2 | Sandwich Place | Pizza Place | Seafood Restaurant | Café | Fried Chicken Joint | Park | Burger Joint | BBQ Joint | Peruvian Restaurant | Chinese Restaurant |
| 8 | LIM8 | La Molina | 2 | Restaurant | Bakery | Seafood Restaurant | Breakfast Spot | Park | Supermarket | Athletics & Sports | Peruvian Restaurant | Dessert Shop | Coffee Shop |
| 9 | LIM9 | Santiago De Surco | 2 | Park | Seafood Restaurant | Chinese Restaurant | Pizza Place | Bakery | Restaurant | Sandwich Place | Fast Food Restaurant | Coffee Shop | Fried Chicken Joint |

One of the patterns that can be seen from the previous charts is that boroughs from the same city are grouped together in the same cluster: boroughs from Lima are in Cluster 0 and 2, while boroughs from Toronto are in Cluster 1.

Other pattern is that the most common venues in Cluster 0 and in Cluster 2 also, are Restaurants in Lima, and in Cluster 1 are Café and Coffee Shops in Toronto.

Let’s bar chart the top venue categories to confirm this pattern.





From the previous charts, we can see that the proportion of all kinds of Restaurants is very high in Lima, as well as the Café and Coffee Shops in Toronto, which seems to be predominating when the algorithm performs the clustering.

In Lima, Restaurants represents 30% and Café and Coffee Shops represents 8% of all venues.

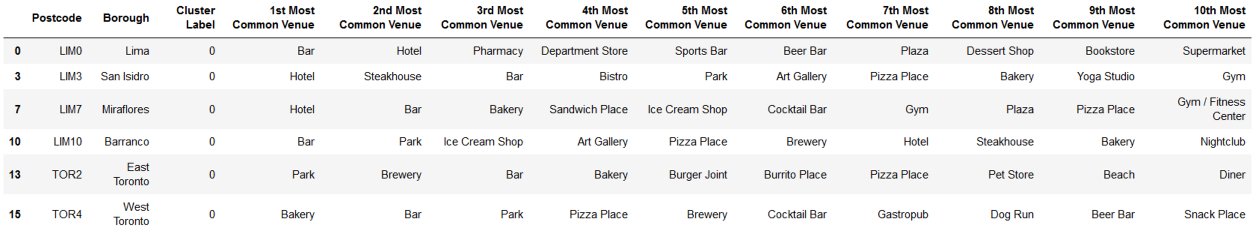
In Toronto, Restaurants represents 26% and Café and Coffee Shops represents 16% of all venues.

Doing a little research, you can notice that Canada ranks at the top in coffee consumption across the globe. The same can be said about the Restaurants in Lima: it is not for nothing that Lima is known as the *Gastronomical Capital of the Americas* as we saw in the introduction section.

What if we want to compare Lima and Toronto but without these two venue categories? We can do the same exercise, and cluster the boroughs excluding all Restaurants and Café / Coffee Shops venues.

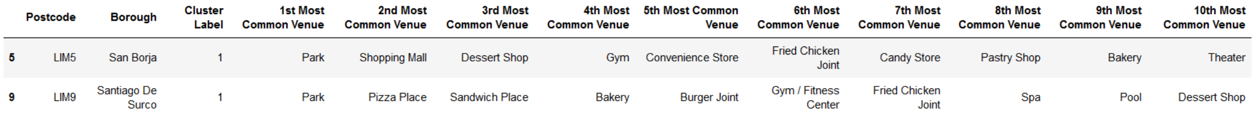
After removing all Restaurants and Café / Coffee Shops venue categories from the dataset, and applying again the K-means algorithm, the new clustering is shown below:

**Cluster 0:**



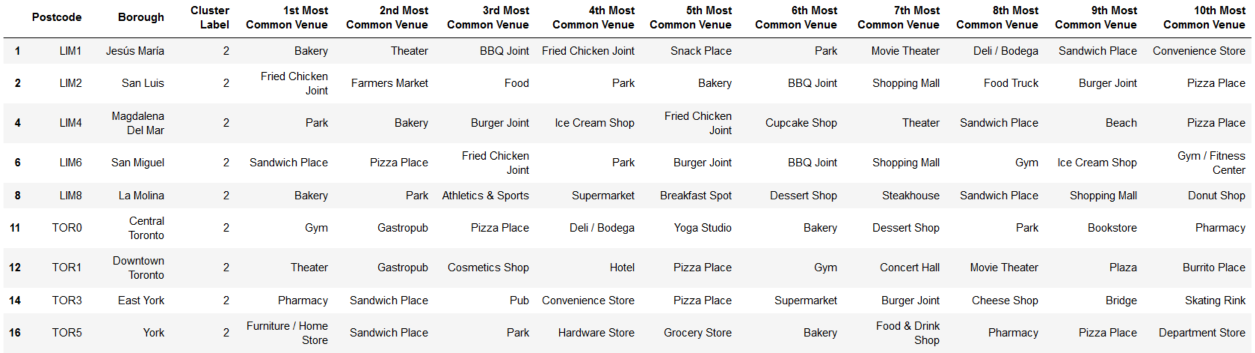
| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LIM0 | Lima | 0 | Bar | Hotel | Pharmacy | Department Store | Sports Bar | Beer Bar | Plaza | Dessert Shop | Bookstore | Supermarket |
| 3 | LIM3 | San Isidro | 0 | Hotel | Steakhouse | Bar | Bistro | Park | Art Gallery | Pizza Place | Bakery | Yoga Studio | Gym |
| 7 | LIM7 | Miraflores | 0 | Hotel | Bar | Bakery | Sandwich Place | Ice Cream Shop | Cocktail Bar | Gym | Plaza | Pizza Place | Gym / Fitness Center |
| 10 | LIM10 | Barranco | 0 | Bar | Park | Ice Cream Shop | Art Gallery | Pizza Place | Brewery | Hotel | Steakhouse | Bakery | Nightclub |
| 13 | TOR2 | East Toronto | 0 | Park | Brewery | Bar | Bakery | Burger Joint | Burrito Place | Pizza Place | Pet Store | Beach | Diner |
| 15 | TOR4 | West Toronto | 0 | Bakery | Bar | Park | Pizza Place | Brewery | Cocktail Bar | Gastropub | Dog Run | Beer Bar | Snack Place |

**Cluster 1:**



| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | LIM5 | San Borja | 1 | Park | Shopping Mall | Dessert Shop | Gym | Convenience Store | Fried Chicken Joint | Candy Store | Pastry Shop | Bakery | Theater |
| 9 | LIM9 | Santiago De Surco | 1 | Park | Pizza Place | Sandwich Place | Bakery | Burger Joint | Gym / Fitness Center | Fried Chicken Joint | Spa | Pool | Dessert Shop |

**Cluster 2:**



| | Postcode | Borough | Cluster Label | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LIM1 | Jesús María | 2 | Bakery | Theater | BBQ Joint | Fried Chicken Joint | Snack Place | Park | Movie Theater | Deli / Bodega | Sandwich Place | Convenience Store |
| 2 | LIM2 | San Luis | 2 | Fried Chicken Joint | Farmers Market | Food | Park | Bakery | BBQ Joint | Shopping Mall | Food Truck | Burger Joint | Pizza Place |
| 4 | LIM4 | Magdalena Del Mar | 2 | Park | Bakery | Burger Joint | Ice Cream Shop | Fried Chicken Joint | Cupcake Shop | Theater | Sandwich Place | Beach | Pizza Place |
| 6 | LIM6 | San Miguel | 2 | Sandwich Place | Pizza Place | Fried Chicken Joint | Park | Burger Joint | BBQ Joint | Shopping Mall | Gym | Ice Cream Shop | Gym / Fitness Center |
| 8 | LIM8 | La Molina | 2 | Bakery | Park | Athletics & Sports | Supermarket | Breakfast Spot | Dessert Shop | Steakhouse | Sandwich Place | Shopping Mall | Donut Shop |
| 11 | TOR0 | Central Toronto | 2 | Gym | Gastropub | Pizza Place | Deli / Bodega | Yoga Studio | Bakery | Dessert Shop | Park | Bookstore | Pharmacy |
| 12 | TOR1 | Downtown Toronto | 2 | Theater | Gastropub | Cosmetics Shop | Hotel | Pizza Place | Gym | Concert Hall | Movie Theater | Plaza | Burrito Place |
| 14 | TOR3 | East York | 2 | Pharmacy | Sandwich Place | Pub | Convenience Store | Pizza Place | Supermarket | Burger Joint | Cheese Shop | Bridge | Skating Rink |
| 16 | TOR5 | York | 2 | Furniture / Home Store | Sandwich Place | Park | Hardware Store | Grocery Store | Bakery | Food & Drink Shop | Pharmacy | Pizza Place | Department Store |

In this new clustering, in Clusters 0 and 2, boroughs from Lima get now mixed with boroughs from Toronto.

Was the clustering from this second exercise expected, i.e., without Restaurants and Café / Coffee Shops?

Let's see the analysis in the Results section.

## Results

For this second exercise, I was expecting boroughs that are known as tourism, business and finance centers grouped together, i.e.: Central, East, West, and Downtown Toronto, with Lima, Miraflores, San Isidro and Barranco.

But data from Foursquare seems to be more associated to tourism venues rather than business or finance locations.

I was also surprised that the number of Hotel venues was very low when analyzing all the venues pulled from Foursquare. Maybe the limit of venues set to 100 was not big enough to include a more representative number of venues from each category in each borough.

Looking to the previous charts with the clustering results, we observe that:

- Cluster 0: Lima, Miraflores, San Isidro and Barranco were grouped together with East and West Toronto, close to my expectation. The most common venues in this cluster are Bars, Hotels, Bakeries, Breweries, Parks.
- Cluster 1: the main driver seems to be the Parks venue category.
- Cluster 2: might be grouping several venues where there is no a predominant one.

## Observations

In the clustering of the first exercise, we saw that the main driver for grouping the boroughs were the Restaurant venues and the Café / Coffee Shop venues. The results applied by the K-means algorithm were consistent with the number of venues in each of these categories that was pulled from Foursquare.

These first clustering, however, did not allow us to determine which boroughs were similar when comparing Lima and Toronto, because boroughs from Lima were clustered together, with the Restaurants venues being the influential factor; while boroughs from Toronto were clustered together, with the Café / Coffee Shop venues being the influential factor.

Excluding the Restaurants and Café / Coffee Shops venue categories from the analysis, would allow me to compare Lima and Toronto in other aspects. This does not diminish the importance of the excluded two categories, as we already saw that they are a very important aspect of Lima and Toronto cities.

When excluding the Restaurants venue categories, it became obvious that many other venue categories may be grouped together to have a better view of categories affecting the clustering results, e.g., venues like Fried Chicken Joint and Burger Joint can be grouped as Fast Food. This of course will have the K-means algorithm to provide different results, but explanation might be easier.

Regarding the K-means algorithm used for clustering, it needed some parameters to be set before running the model. Two of them are the random_state and the number of clusters:

- for the first parameter, random_state, I set it to a fixed value to avoid having different results while working in this project;
- about the second parameter, the Elbow method used to determine the optimal number of clusters did not gave a strong suggestion. Still, considering that the dataset consisted of only 17 boroughs, the number of clusters would probably be between 3 and 5 clusters if I had to give a rough estimate.

To be considered also in this analysis was the number of venues and radius or distance to be searched from the borough locations when using the Foursquare API. To simplify the analysis, the number of venues was limited to 100 for each borough; and while Toronto distances between boroughs were greater than those in Lima, the radius to search for venues was set to 1500 meters for all locations, which was the average for Lima. These two parameters might be adjusted to get more precise results if needed.

## Conclusion

The methodology followed in this report is an example of how real-life data science projects can be.

The algorithms, APIs, and tools used are just among other approaches that can be used for similar analysis.

Some of the considerations, simplifications and limitations of this project have been described in the previous sections.

Finally, I hope that the objective of comparing the boroughs from Lima and Toronto cities leveraging the Foursquare location data has been accomplished in this report.

The full Capstone Project code can be found in the following link:

https://github.com/JoseLPachecoB/github-example/blob/master/Capstone_Lima%20and%20Toronto%20cities.ipynb