**Comparing Toronto with the most important South American cities**

**Which Toronto’s neighborhoods are more like Sao Paulo (Brazil), Buenos Aires (Argentina), Asuncion (Paraguay) and Montevideo (Uruguay)?**

**Lucas Mielke**

**November 20, 2019**

**Summary**

[**1. Introduction 3**](#_Toc25177570)

[**2. Data 3**](#_Toc25177571)

[**Methodology 4**](#_Toc25177572)

[**Methodology Part 1a - Creating Toronto Data - Scrape Wikipedia List of postal codes 4**](#_Toc25177573)

[**Methodology Part 1b - Creating Toronto Data - Scrape Geospatial Data 5**](#_Toc25177574)

[**Methodology Part 2 - Adding the 4 South American cities to the Analysis 5**](#_Toc25177575)

[**Methodology Part 3 - Combining all the data 5**](#_Toc25177576)

[**Methodology Part 4 - Creating a function to connect to Foursquare API and using it 6**](#_Toc25177577)

[**Methodology Part 5 - Normalizing and creating cluster using Kmeans 8**](#_Toc25177578)

[**Conclusions 11**](#_Toc25177579)

# Introduction

Toronto is a very cosmopolitan city. In this city, there are several immigrants who have moved to it, either temporarily for a trip, taking a medium-term course, graduating or even permanent changes due to work or marriage.

Among the immigrants of this city, there are a considerable number of people from South America.

Migration is not always easy. South American immigrants face language, cultural, climate barriers. In this context, living in a slightly more similar place with your place of origin may be helpful.

In this context, this project aims to compare 4 major South American cities with Toronto neighborhoods and find those that are most like these cities.

The cities analyzed will be:

* Sao Paulo (Brazil),
* Buenos Aires (Argentina),
* Asuncion (Paraguay) and
* Montevideo (Uruguay).

# Data

The analysis of this project will include the following data:

* Foursquare API, which will provide the most common establishments from all analyzed locations (Toronto, Canada + 4 South American Cities described at introduction): <https://developer.foursquare.com/>
* List of Neighborhoods and Postcodes in Toronto, provided by wikipedia, that will serve as the basis of Toronto's analysis: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
* Toronto geospatial data, Where will we have the coordinates of each neighborhood of Toronto: <https://cocl.us/Geospatial_data>.
* Geopy Geocoders, which will be used to convert the location of the 4 south American cities into geographic coordinates: <https://pypi.org/project/geopy/>

For this project, we will create a database that will contain the neighborhoods of Toronto and the 4 South American cities analyzed, with their respective commercial establishments and geographic information. This base will be used in a machine learning algorithm to create clusters (kmeans - <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>) and as a result we will have combinations of South American cities with the neighborhoods of Toronto.

As a first step, let's capitalize Wikipedia's neighborhood data and neighborhood geospatial data, do some cleaning and treatment, and combine these two bases into one called "toronto\_data".

Second step, let's create 4 databases with 4 South American cities (Sao Paulo, Buenos Aires, Asuncion, and Montevideo) with latitude and longitude using Geopy. The bases will be "saopaulo\_data", "buenosaires\_data", "asuncion\_data" and "montivideu\_data".

In the third step we will combine the above 5 bases into one. This base, which will contain a list of neighborhoods and geographic information (latitude and longitude) will be the basis for the Foursquare API.

In the fourth step we will create a function of capturing establishments based on geographic information and we will apply them to all locations.

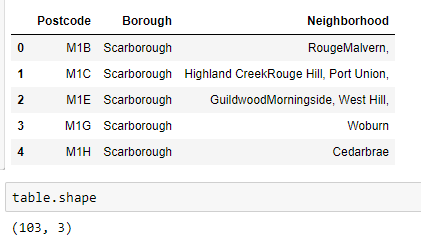
In the fifth step the data with the establishments of each locality will be normalized and used to create cluster using Kmeans.

In the sixty step we will analyses the clusters and were the South American cities are classified.

# Methodology

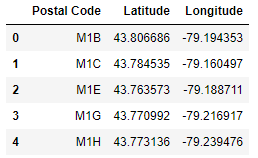
## **Methodology Part 1a - Creating Toronto Data - Scrape Wikipedia List of postal codes**

We scrape a table in Wikipedia using pandas, following the instructions at <https://stackoverflow.com/questions/55234512/how-to-scrap-wikipedia-tables-with-python> and we removed rows that were “not assigned”. Then we grouped the neighborhoods of duplicated postal codes. The final row had 103 rows and 3 columns.



## **Methodology Part 1b - Creating Toronto Data - Scrape Geospatial Data**

Then we scraped a table from Geospatial data (<https://cocl.us/Geospatial_data>). This table have the geographic information (latitude and longitude) by postal code.



This Table was merged with the first table (in part 1a)

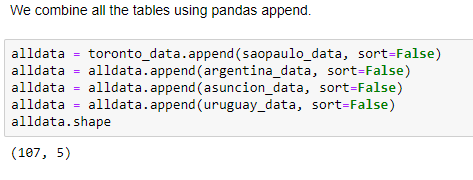
## **Methodology Part 2 - Adding the 4 South American cities to the Analysis**

Using Geopy we got the Latitude and longitude of the 4 analyzed cities of south America. Example of Sao Paulo is shown below. The “Borough” was named “South America” to help identifying these cities.



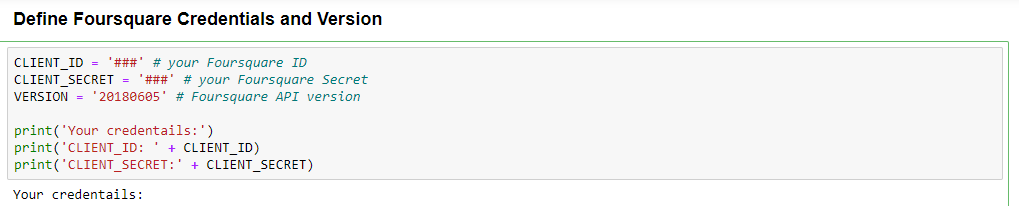
## **Methodology Part 3 - Combining all the data**

All the data (Toronto’s neighborhoods with Latitude and Longitude, Sao Paulo’s, Buenos Aires’, Asuncion’s and Montevideo’s latitude and longitude) was combined in one dataframe having 107 rows and 5 columns.



## **Methodology Part 4 - Creating a function to connect to Foursquare API and using it**

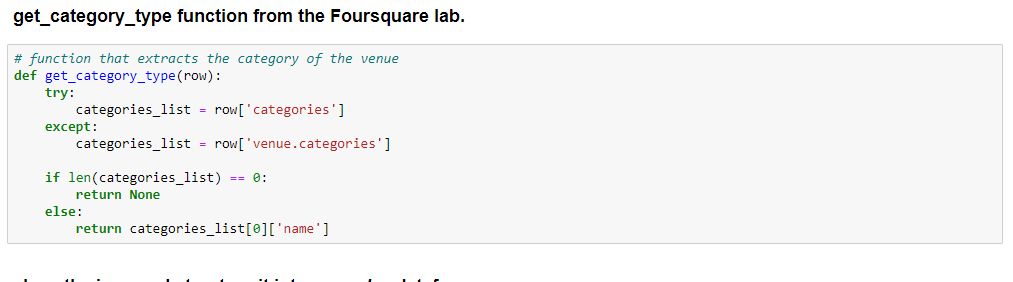
We used Foursquare API to get venues of all locations (107 rows at “alldata” dataframe - above). We add our Client ID and Password from Foursquare developer tool.



We created an API as below.



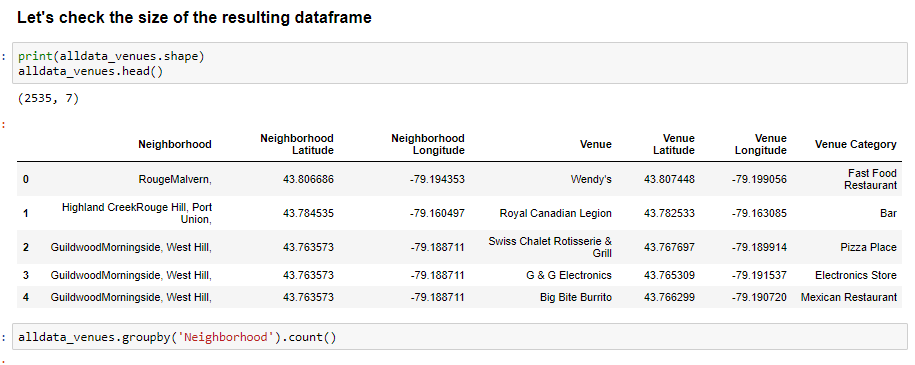
Then we created a function to get category type and venues.



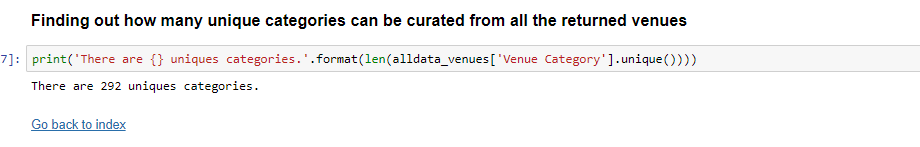
Then we get all the locations.



Then we put all the data in one dataframe as below.

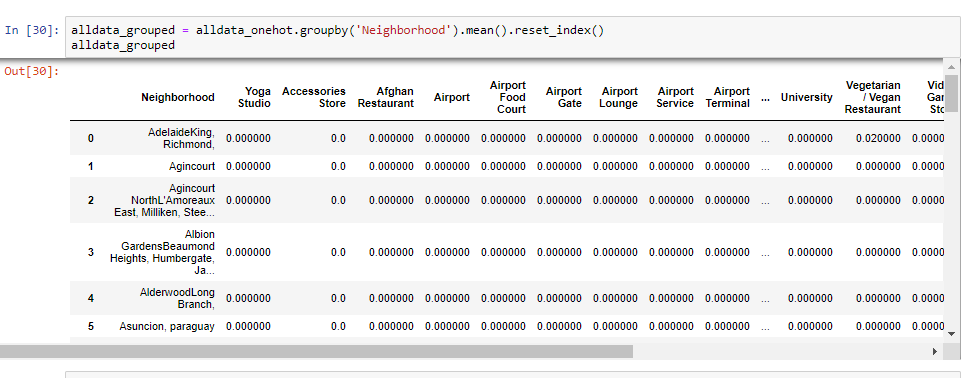


The final dataframe has 292 unique categories of venues.

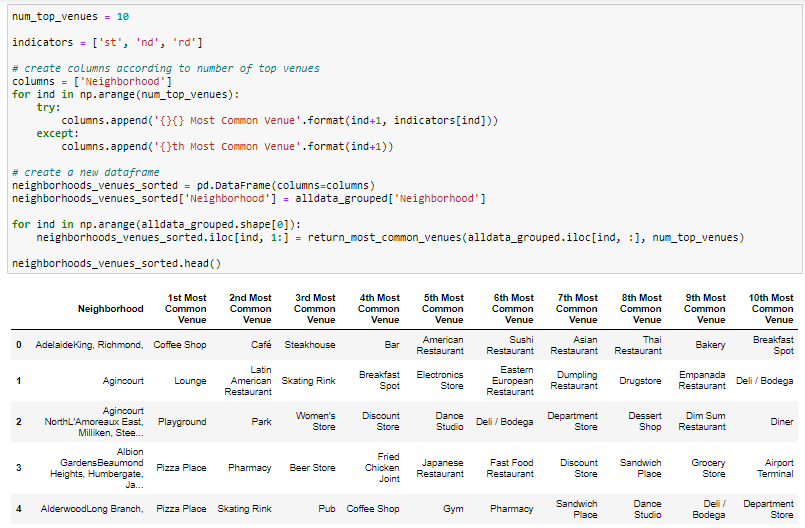


## **Methodology Part 5 - Normalizing and creating cluster using Kmeans**

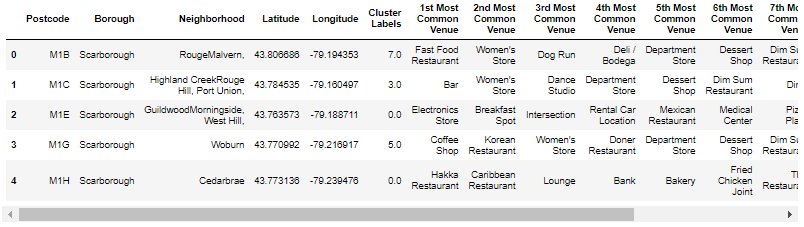
All the data was normalized to be used in Kmeans.



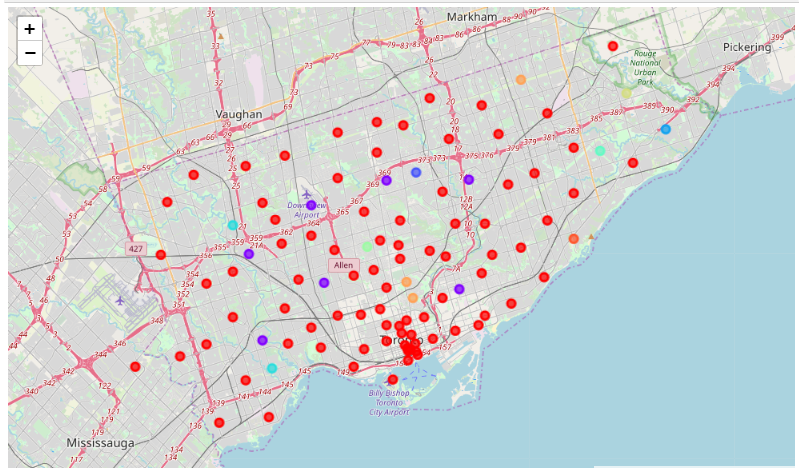
And we created a dataframe that list the top 10 venues from each location.



We ran the Kmeans model, classing the locations in 10 clusters and merged the clusters with the dataframe above, creating a list that classify each neighborhood in one location and lists the 10 most common venues.



The clusters in Toronto was as below

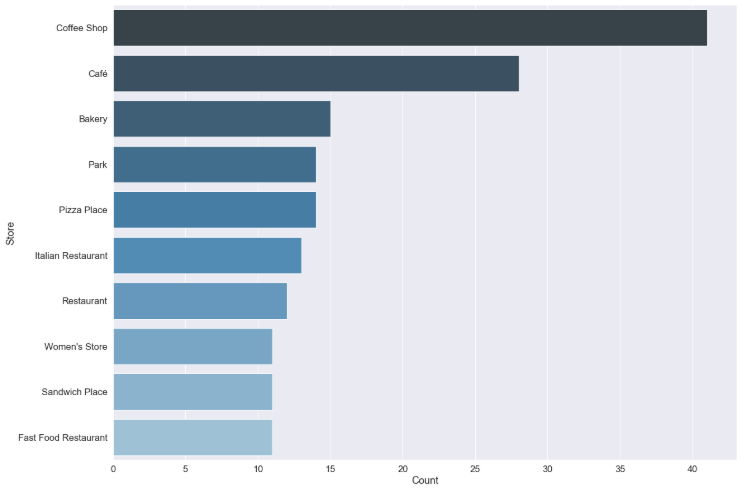


Most of the location was in cluster 1 (red)…

As the south American cities…



The four South American cities are in cluster 1, the biggest one. In this cluster are neighborhoods like GuildwoodMorningside, East Birchmount ParkIonview, Kennedy Park, West Hill, Clarks CornersSullivan, Tam O'Shanter. Let’s Plot the top 10 venues of the cluster 1, the cluster that contains the 4 south American cities. We can see below that the most common venue is Coffee Shop by far, followed Cafés, Bakery, Parks...



# Conclusions

In this work we take public data from neighborhoods of Toronto, Canada (list of neighborhoods and geographic information) as well as the location of 4 major Latin American cities (Sao Paulo - Brazil, Asuncion - Paraguay, Buenos Aires - Argentina and Montevideo - Uruguay).

We combined this information into one database, which contained locations and geographic information. This combined base contained 107 locations to be analyzed.

A connection has been made to the Foursquare API and a function has been created to fetch Venues for each of the locations. Running this function in the combined database found 292 unique venues.

The combined database (with locations and venues) contains 2535 rows per 292 columns.

The database containing the neighborhood list and each venue type was normalized and used with Machine Learning. We ran a Kmeans algorithm that clustered all neighborhoods into 10 clusters based on the venues of every 1. Finally, we created a dataframe listing the 10 most common Venus by neighborhood and added the created clusters.

Cluster distribution was highly concentrated in cluster 1, which contained 89 neighborhoods, including the 4 South American cities. Cluster 1 is quite abundant in coffee shops and cafes and generally contains plenty of restaurants.