Starbucks in Quito, EC

Finding the Best Location for the First Starbucks in Quito

by Andrés Borja

Submitted as part of IBM's Data Science Specialization

Applied Data Science Capstone Project

a course by Coursera

Table of Contents

1. Introduction

1.1 Background

Latin America is the region from which Starbucks coffee buys the majority of its beans, and yet, it turns out it's among this is also the region where the world renowned coffee giant has the least amount of venues. Ecuador, in particular, does not have a Starbucks coffee shop, not even in the capital city, Quito. While it is true that up to a few years ago coffee culture was restricted to the privacy of Ecuadorian homes, in recent years this has rapidly changed. Even though Ecuador is not typically known for coffee production, it is slowly gaining a worldwide reputation of excellence. Coffee growing areas in the Andes, north of Quito and just south of Cuenca, have recently been recognized for the high quality of their product. As a consequence, Ecuadorians are drinking more and more coffee. Several roasters and small speciality coffee shops have started sprouting across Quito and some other cities, introducing locals and tourists alike to the wonders of great tasting coffee.

Two successful brands inspired by the Starbucks model have already opened in the country's biggest cities: Sweet & Coffee, a local company, and Juan Valdez, from Colombia. They have replicated the Starbucks model and their results have demonstrated that Ecuadorians do like drinking coffee outside their homes, surrounded by strangers, while working on their laptops. Of course, the target population is the country's booming middle class and the large number of upwardly mobile young professionals, a demographic heavily influenced by the U.S. and its culture. Such customers keep growing and were Starbucks to open a coffee shop in Quito, it would immediately be flooded by customers.

1.2 Business problem and audience

However, Quito is a big city; in order to guarantee its success, Starbucks should carefully and strategically plan where exactly to open its first store in Ecuador's capital. They should try to pick the best area or neighborhood based on the kind of venues that already exist there, which should be economically tied to coffee shops. By classifying the

neighborhoods of Quito in terms of its extant venues, Starbucks could make a better informed decision of where to open their first store in Ecuador's capital, thus improving the likelihood of their success. Fortunately, several data science and machine learning techniques can help Starbucks explore and cluster Quito's neighborhoods in order to better understand the city, its structure and its dynamics, so that they can accomplish their business goals.

2. Data

2.1 Data sources

In order to be able to classify Quito's neighborhoods on a map, we will need a database which includes all of the city's neighborhoods with their geographical coordinates. The city's municipality website has several georeferenced databases of the city's geopolitical and administrative divisions. We picked the databased named "Barrio - Sector", from their website.

The classification parameter will be the kinds of venues that exist in each neighborhood. This information will be obtained from Foursquare, a location data provider with information about all manner of venues and events within an area of interest. Such information includes venues names, categories, locations, menus, ratings and even pictures. We will be using using the Foursquare API to explore the venues surrounding the coordinates for each neighborhood, and then classifying the neighborhoods based on their assigned categories.

2.2 Data cleaning

The municipality's neighborhoods database includes all of the neighborhoods that make up the Metropolitan District of Quito, which is classified in terms of urban and rural boroughs. There are a total of 1268 neighborhoods in the MD of Quito. We decided to work with the urban area neighborhoods only, because rural areas are too remote and isolated, with very low population density (see Fig. 1). Almost no venues could be found

around the rural neighborhoods when exploring them through Foursquare, so it wouldn't make sense to include them in our analysis. The urban area is made up of 516 neighborhoods. Additionally, 28 of these urban neighborhoods did not have a name assigned in the database, so we decided to exclude them from the analysis as well. In the end we kept 488 neighborhoods from Quito's urban area.

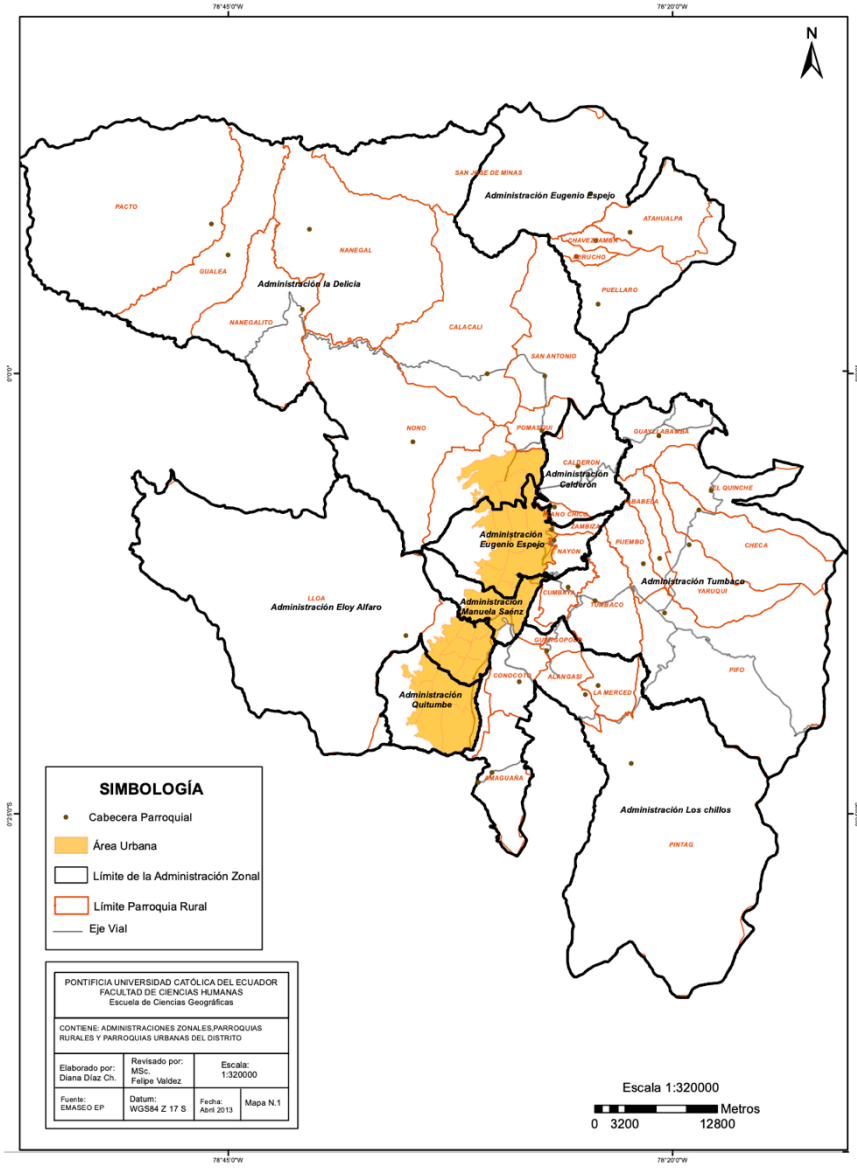


***Fig. 1**: Urban and rural buroughs in the Metropolitan District of Quito.*

After cleaning our data, we stored the urban area neighborhoods with their coordinates in a data frame with the following structure:

	ID	Neighborhood	Latitude	Longitude
0	01050014	NUEVA VIDA	-0.273928	-78.563715
1	01050026	VENCEREMOS	-0.272159	-78.564992
2	04070005	AVIACION CIVIL	-0.155464	-78.487854
3	01050017	S.MARTHA ALT CHIL	-0.275713	-78.561793
4	02010005	LA RAYA A	-0.255895	-78.543519

***Fig. 2**: Pandas data frame with geolocation information from the urban area neighborhoods of Quito.*

And thanks to their coordinates we were able to visualize the selected neighborhoods on a map created with the Folium library:

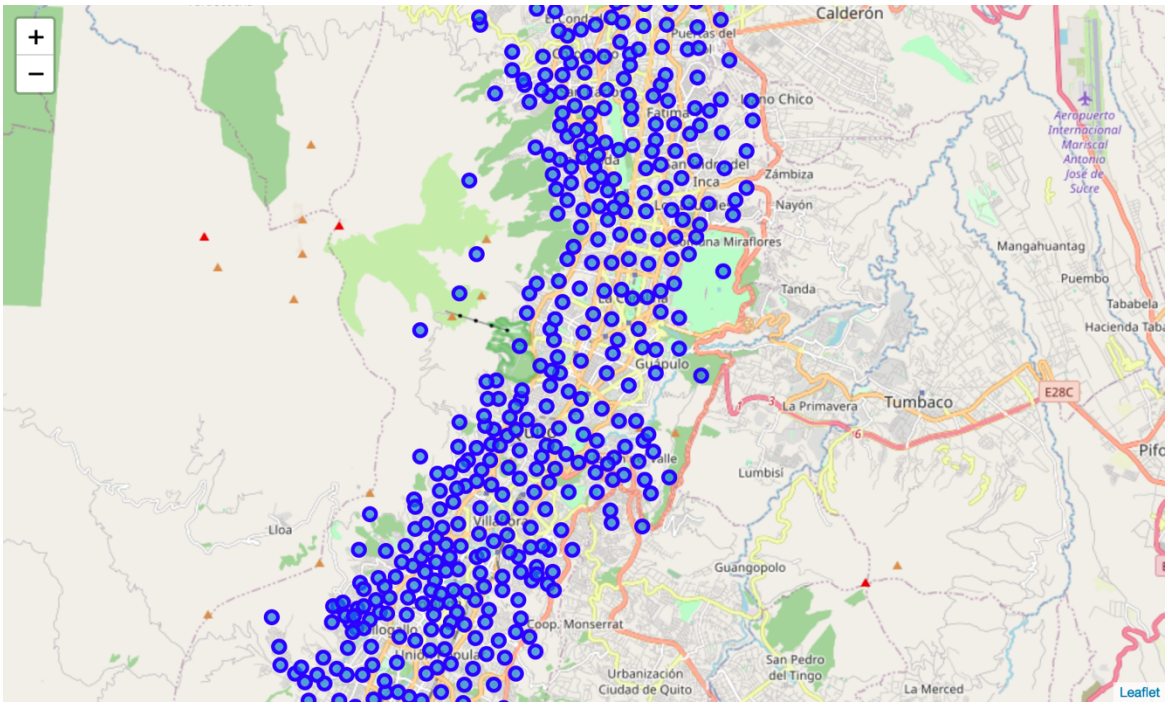


***Fig. 3:** Foilum map of Quito with urban neighborhoods superimposed.*

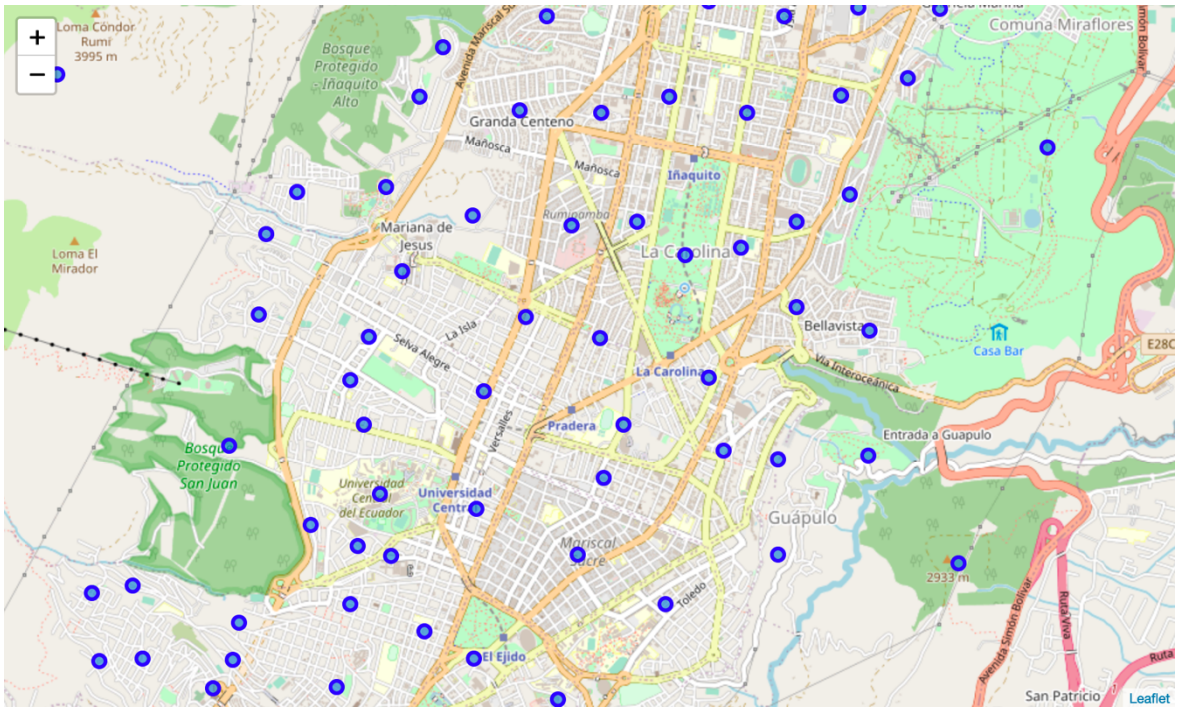


Fig. 4: *Close-up of the Folium map of Quito with urban neighborhoods superimposed.*

3. Method

3.1 Exploring the neighborhoods in Quito

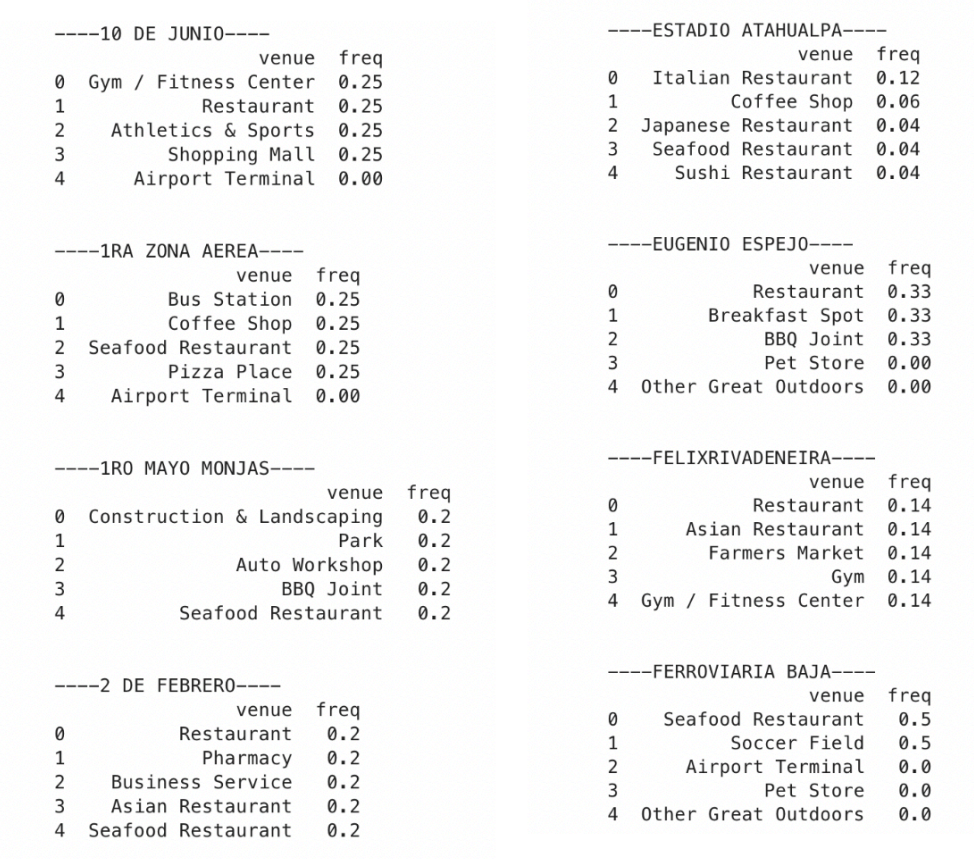
Once we obtained a map of the neighborhoods with their coordinates, the next step was exploring the venues information by extracting it from the Foursquare API. We called for the venues around each neighborhood's coordinates, establishing a radius of 500 meters around each neighborhood in order to obtain a certain number of venues (limit=50) in that radius. Specifically, we were interested in extracting the venues "categories" attributes. We created a function to repeat this process with all neighborhoods, generating a data frame that included all of the neighborhoods, its venues and the categories of said venues, all of this with georeferenced coordinates (longitude and latitude). The first few rows of this data frame can be seen in Fig. 5. By exploring this dataframe we found that there were 217 unique venue categories.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NUEVA VIDA	-0.273928	-78.563715	S29C Y OE12E	-0.272326	-78.563025	Bus Stop
1	NUEVA VIDA	-0.273928	-78.563715	Mariachi Fiesta Mexicana	-0.275757	-78.565411	Convention Center
2	VENCEREMOS	-0.272159	-78.564992	S29C Y OE12E	-0.272326	-78.563025	Bus Stop
3	VENCEREMOS	-0.272159	-78.564992	Mariachi Fiesta Mexicana	-0.275757	-78.565411	Convention Center
4	AVIACION CIVIL	-0.155464	-78.487854	Menestras del Primo	-0.155252	-78.488654	Restaurant
5	AVIACION CIVIL	-0.155464	-78.487854	Cachorros	-0.153951	-78.491237	Gym
6	AVIACION CIVIL	-0.155464	-78.487854	La Michoacana	-0.158246	-78.490567	Mexican Restaurant
7	AVIACION CIVIL	-0.155464	-78.487854	El Manglar De Las Conchas	-0.155171	-78.488566	Seafood Restaurant
8	AVIACION CIVIL	-0.155464	-78.487854	Los legitimos helados de paila de la Concepcion	-0.153525	-78.491681	Ice Cream Shop
9	AVIACION CIVIL	-0.155464	-78.487854	La tortilla	-0.153528	-78.490682	Arepa Restaurant

Fig. 5: Pandas *data frame with neighborhoods, venues, venue categories and coordinates for each.*

3.2 Analyzing each neighborhood's venue categories

The next step was finding out how many of each venue category existed in each neighborhood. For this, we used a process called one-hot encoding. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means clustering algorithm we will use later, all unique items under "Venue Category" are one-hot encoded, that is, transformed from categorical objects into integers. This way, we were able to explore the neighborhoods in terms of the frequency of occurrence of each venue category nearby. Figure 6 below shows a sample of some neighborhoods with the 5 most common venue categories near them and the relative frequency of each category in each neighborhood. With this information we were able to create a data frame containing the top 10 types of venues for each neighborhood in terms of their frequency.



```
----10 DE JUNIO----
                  venue  freq
0  Gym / Fitness Center  0.25
1            Restaurant  0.25
2    Athletics & Sports  0.25
3         Shopping Mall  0.25
4      Airport Terminal  0.00

----1RA ZONA AEREA----
                 venue  freq
0         Bus Station  0.25
1         Coffee Shop  0.25
2  Seafood Restaurant  0.25
3         Pizza Place  0.25
4    Airport Terminal  0.00

----1RO MAYO MONJAS----
                        venue  freq
0  Construction & Landscaping   0.2
1                        Park   0.2
2               Auto Workshop   0.2
3                   BBQ Joint   0.2
4          Seafood Restaurant   0.2

----2 DE FEBRERO----
                 venue  freq
0          Restaurant   0.2
1            Pharmacy   0.2
2    Business Service   0.2
3    Asian Restaurant   0.2
4  Seafood Restaurant   0.2

----ESTADIO ATAHUALPA----
                 venue  freq
0   Italian Restaurant  0.12
1         Coffee Shop  0.06
2  Japanese Restaurant  0.04
3   Seafood Restaurant  0.04
4     Sushi Restaurant  0.04

----EUGENIO ESPEJO----
                  venue  freq
0            Restaurant  0.33
1       Breakfast Spot  0.33
2             BBQ Joint  0.33
3             Pet Store  0.00
4  Other Great Outdoors  0.00

----FELIXRIVADENEIRA----
                  venue  freq
0            Restaurant  0.14
1      Asian Restaurant  0.14
2        Farmers Market  0.14
3                   Gym  0.14
4  Gym / Fitness Center  0.14

----FERROVIARIA BAJA----
                  venue  freq
0    Seafood Restaurant   0.5
1          Soccer Field   0.5
2      Airport Terminal   0.0
3             Pet Store   0.0
4  Other Great Outdoors   0.0
```

Fig. 6: *Frequency of venue categories within each neighborhood (sample).*

3.3 Clustering neighborhoods

K-means algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The venues data was then trained using a K-means clustering algorithm to get the desired clusters to base the analysis on. K-means was chosen as our dataset is pretty big and in such situations K-means will be computationally faster than other clustering algorithms. We tried different numbers of clusters when running the algorithm and in the end, we decided on 5 clusters as the ideal number, based on the spread of the data points and the proportionality of size of the resulting clusters with this number, unlike higher or lower numbers.

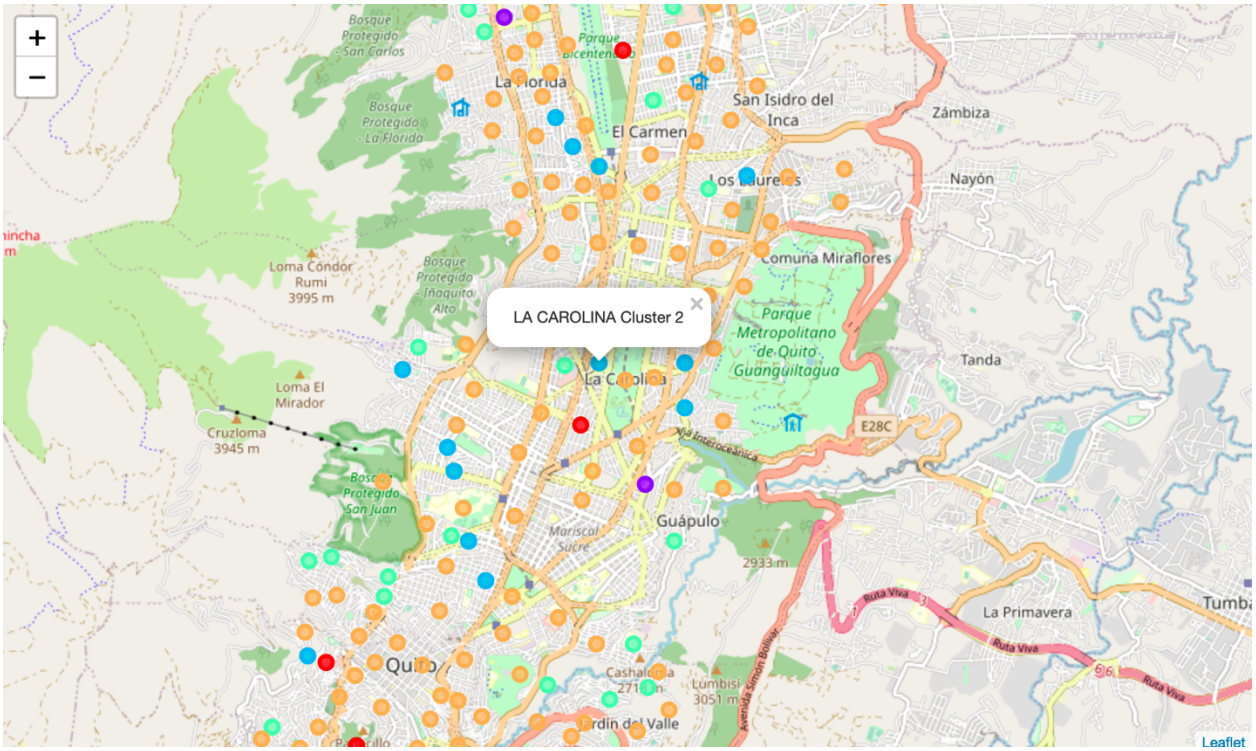


Fig. 7: *Map of Quito with clustered (k-means) neighborhoods superimposed.*

3.4 Examining resulting clusters

After training our data with the k-means clustering algorithm (k=5), we can examine the resulting clusters and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we observed the following:

- ***Cluster 1***: This cluster's most common venues seem to be mainly parks, zoos and fields, so it would not be a good match for Starbucks.

Cluster 1

```
uio_merged.loc[uio_merged['Cluster Labels'] == 0, uio_merged.columns[[2] +
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	C
17	STA. BARBARA BAJA	0	Park	Convenience Store	Nightclub	Plaza	Zoo	
27	MIRAFLORES ALTO	0	Park	Zoo	Empanada Restaurant	Food	Flower Shop	
109	SAN SALVADOR	0	Scenic Lookout	Park	Zoo	Donut Shop	Flea Market	
158	TOCTIUCO	0	Business Service	Park	Zoo	Empanada Restaurant	Flower Shop	
199	ARGELIA INTERMEDIA	0	Burger Joint	Park	Empanada Restaurant	Food	Flower Shop	
236	PAVON GRIJALVA	0	Park	Zoo	Empanada Restaurant	Food	Flower Shop	
276	CONSEJO PROVINCIAL	0	Park	Zoo	Empanada Restaurant	Food	Flower Shop	
298	CAMPO ALEGRE	0	Arts & Entertainment	Park	Zoo	Empanada Restaurant	Food	

- ***Cluster 2***: Neighborhoods in cluster 2 seem to be surrounded by bus stations, mostly, as well as miscellaneous categories not related with the food and beverages industry. Again, not the best match for our Starbucks shop.

Cluster 2

```
uio_merged.loc[uio_merged['Cluster Labels'] == 1, uio_merged.columns[[2] + li
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
22	CAUSAYLLACTA	1	Bus Station	Zoo	Empanada Restaurant	Food	Flower Shop
43	EL ROCIO	1	Bus Station	Fast Food Restaurant	Farmers Market	Soccer Stadium	Grocery Store
68	LA TOLA	1	Brewery	Park	Bus Station	Hostel	Food
71	AREA DE PROTECCION	1	Bus Station	Art Museum	Zoo	Empanada Restaurant	Food
106	ALVARO PEREZ INDEPENDIENTE	1	Convenience Store	Bus Station	Department Store	Zoo	Empanada Restaurant
133	LULUNCOTO	1	Bus Station	Breakfast Spot	Art Museum	Zoo	Entertainment Service
157	HUAYRALLACTA	1	Bus Station	Zoo	Empanada Restaurant	Food	Flower Shop
165	EL COMERCIO	1	Bus Station	Pharmacy	Auto Garage	Pet Store	Zoo
166	LOS LIBERTADORES	1	Martial Arts Dojo	Bus Station	Furniture / Home Store	Bar	Zoo
				Latin			

- ***Cluster 3***: Neighborhoods in cluster 3 seem to be surrounded by gyms, mostly, as well as zoos flea markets and flower shops. Again, not the best match for our Starbucks shop.

Cluster 3

```
uio_merged.loc[uio_merged['Cluster Labels'] == 2, uio_merged.columns[[2]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	COLINAS DEL SUR	2	Gym	Zoo	Flower Shop	Flea Market	Field
44	LA ESTANCIA	2	Gym	Zoo	Flower Shop	Flea Market	Field
143	YAGUACHI	2	Gym	Fried Chicken Joint	Food & Drink Shop	Flower Shop	Flea Market
177	LA LIBERTAD	2	Gym	Zoo	Flower Shop	Flea Market	Field
314	SANTA LUCICIA 2	2	Gym	Gift Shop	Food & Drink Shop	Flower Shop	Flea Market
320	S.FRANC HUARCAY	2	Gym	Zoo	Flower Shop	Flea Market	Field
359	BUENOS AIRES	2	Gym	Bus Line	Zoo	Flower Shop	Flea Market

- ***Cluster 4***: This is the first cluster in which its neighborhoods seem to have a high variety of restaurants, which are related to a coffee shop. This cluster is a good candidate for Starbucks to consider when opening their first coffee shop in Quito.

Cluster 4

```
uio_merged.loc[uio_merged['Cluster Labels'] == 3, uio_merged.columns[[2] +
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
11	PRESIDENCIA REPUBLICA	3	Seafood Restaurant	Fast Food Restaurant	Pizza Place	Soccer Field	BBQ Joint
35	MADRIGAL	3	Restaurant	Construction & Landscaping	Seafood Restaurant	Bed & Breakfast	Zoo
36	S.PEDRO MONJAS	3	Fast Food Restaurant	Restaurant	Clothing Store	Seafood Restaurant	Zoo
41	MIRAFLORES BAJO	3	Burger Joint	Fast Food Restaurant	Snack Place	Seafood Restaurant	Electronics Store
50	LOS ARRAYANES	3	Pizza Place	Burger Joint	Soccer Field	Seafood Restaurant	Electronics Store
61	VERTIENTES SUR	3	Seafood Restaurant	BBQ Joint	Zoo	Electronics Store	Food
73	PABLO ART SUAREZ	3	Construction & Landscaping	Health & Beauty Service	Seafood Restaurant	Grocery Store	Zoo
80	CHIMBACALLE	3	Print Shop	Diner	Chinese Restaurant	Seafood Restaurant	Zoo
83	MONGE DONOSO	3	Bus Station	Motel	Clothing Store	Seafood Restaurant	Zoo
98	LA VICTORIA	3	Seafood Restaurant	Fried Chicken Joint	Hotel	Farmers Market	History Museum

- ***Cluster 5***: This is the largest and least specific cluster we obtained. Its structure is hard to interpret, given that its most common venues don't seem to be related. It may or may not be a good fit for Starbucks' ideal neighborhood. However, on closer inspection, its neighborhoods seem to be all over the place, including areas far away from the city center and closer to the mountains and the city limits. Sadly, this geographical dispersion disqualifies cluster 5. Further analysis of this cluster should be conducted in order to better understand its underlying structure.

Cluster 5

```
uio_merged.loc[uio_merged['Cluster Labels'] == 4, uio_merged.columns[[2] + list(range(5
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	NUEVA VIDA	4	Convention Center	Bus Stop	Zoo	Electronics Store	Flower Shop
1	VENCEREMOS	4	Convention Center	Bus Stop	Zoo	Electronics Store	Flower Shop
2	AVIACION CIVIL	4	Restaurant	Pizza Place	Pharmacy	Seafood Restaurant	Sculpture Garden
3	S.MARTHA ALT CHIL	4	Bus Stop	Zoo	Empanada Restaurant	Food	Flower Shop
4	LA RAYA A	4	Restaurant	Cafeteria	Pizza Place	Latin American Restaurant	Garden
5	RUPERTO ALARCON	4	Electronics Store	Zoo	Food	Flower Shop	Flea Market
6	SOLANDA	4	Convenience Store	Plaza	Dessert Shop	Big Box Store	Park
7	EUGENIO ESPEJO	4	Breakfast Spot	Restaurant	BBQ Joint	Zoo	Empanada Restaurant
8	SANTIAGO 1	4	Supermarket	Pizza Place	Pharmacy	Shopping Mall	Zoo
9	S.ROSA CHIL 3ETP	4	Convention Center	Electronics Store	Food	Flower Shop	Flea Market

4. Results

Based on our analysis, the best cluster of neighborhoods Starbucks should consider when opening their first coffee shop in Quito, Ecuador, seems to be cluster 4. The neighborhoods in this cluster are surrounded by venues very much related to the food and beverages industry. These areas are popullated with all sorts of restaurants and cafes, making them the ideal place for a world renowned coffee shop to succeed. The neighborhoods in this cluster include:

> Presidencia República, Madrigal, San Pedro Monjas, Miraflores Bajo, Los Arrayanes, Vertientes Sur, Pablo Arturo Suárez, Chimbacalle, Monge Donoso, La Victoria, Barionuevo, La Chilena, Julio Matovelle, Recreo Clemencia, El Tejar, Agua Clara, La Kennedy, 6 de Diciembre, Ferroviaria Baja, Alma Lojana, 1ra Zona Aérea, Ejército Nacional, 1ro de Mayo, Unión y Progreso, Rumiñahui, Nueva Aurora II, Monjas Medio, Californ Bonanza, Jipijapa, 2 de Febrero, Santa Rita, Nazareth, La Ecuatoriana, San Isidro del Inca, and Ejército Nacional 2da etapa.

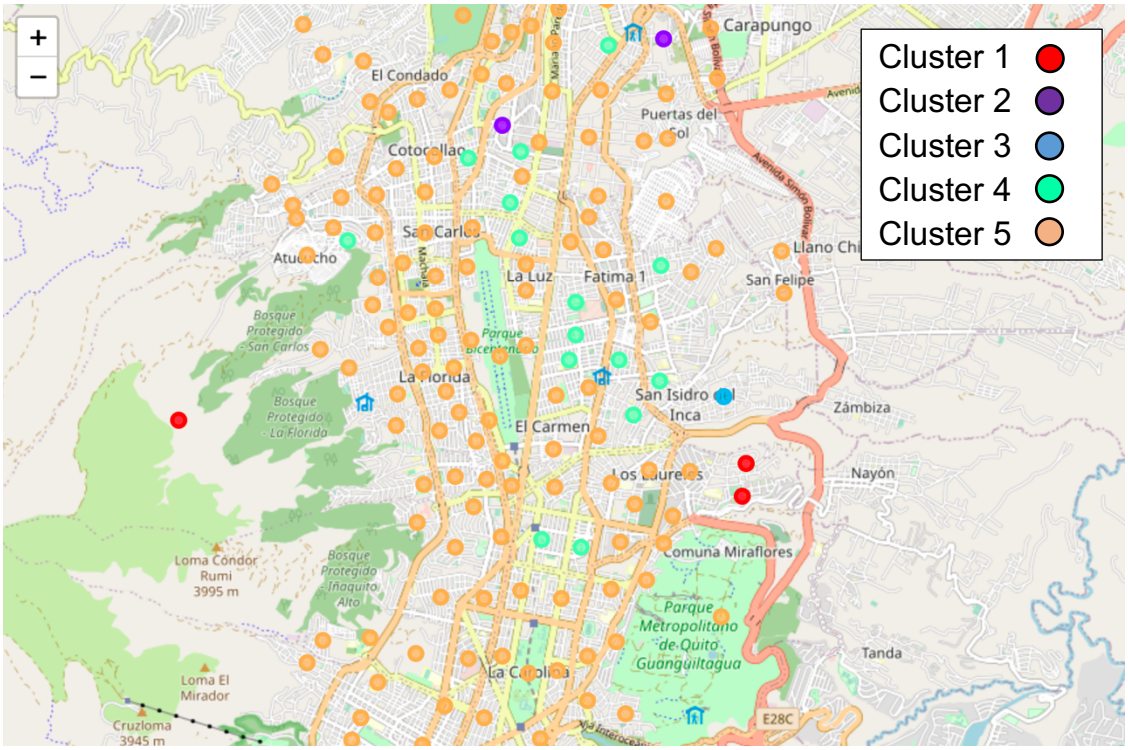


Figure 8: Map of Quito with clustered neighborhoods superimposed. The light green circles mark the location of neighborhoods that belong to cluster 4.

5. Discussion

In this data science project we set out to use machine learning algorithms in order to solve a problema for Starbucks, namely, to make a data-informed decission when selecting the possible location of their first coffee shop in Quito, Ecuador's capital. Using georeferenced data of the neighborhoods of Quito in combination with venues data from the Foursquare API, and training our data with a k-means clustering algorithm, we created 5 clusters of neighborhoods and based on its most common proximal venues, we determined that cluster 4 was the best fit.

However, one of the clusters (#5) did not seem to be grouped by any discernible criteria or clear pattern of its venue categories. We believe this is due to the fact that many of the neighborhoods did not have more than 1 or 2 venues in the established 500 meters radius. We would recommend repeating the analysis by only using neighborhoods with at least 5 venues in their surroundings. The number of neighborhoods may decrease dramatically, but the clusters could make more sense and help us understand the underlying structure of the data better.