Bakery store in Santo Domingo

Junior Peña

July 6, 2020

1. Introduction

1.1 Background

In the Dominican Republic people don't tend to follow certain rules when opening a new business. The first and principal motivator is "I need to open my business wherever I want because I know the product, I have is going to attract people". This thinking works for some for a few months, but it does not for the majority. Usually, we don't think the location as a relevant factor when taking this important step, but is one of the first things we need to figure it out. It's true that the brands attract people, but this is when you have a product well established and known. In most cases you need to go where people will see you regardless if they are looking for the thing you sell. Therefore, it is necessary to be sure where exactly is the best location in order to maximize the customer flow and the business profits.

1.2 Problem

Josh is a recent college graduate who is looking for places in Santo Domingo, Dominican Republic to open a new bakery store. He already has the employees and resources necessary to open the business, but he needs to determine where exactly is the best place (neighborhood) to open it.

In Santo Domingo, there are many bakery stores in every neighborhood, so it is necessary to choose the best place in order to maximize customer flow and profits. He knows the competition is rough, but also there are many bakery stores who open and stay in business successfully. Nevertheless, others don't have the same fate and close in just months. Josh needs to be presented a model to determine where is the best place where he can find success.

1.3 Interest

This is interesting to anyone who is looking for a clear understanding of the importance of choosing the location for your next business. Also, to the people who want to learn how to implement a model like this and obtain the best results.

2. Data acquisition and cleaning

2.1 Data sources

In this scenario the data we are going to use is:

A csv file with a list of all the neighborhoods in Santo Domingo. This file contains three columns: the postal code for each neighborhood, borough and the neighborhood name. This is going to be helpful because it's the base file which we are going to use to create the dataset and the model.

A csv file with the coordinates of each neighborhood. This file contains three columns: the postal code for each neighborhood, the latitude of each neighborhood and the longitude of each neighborhood. This is necessary because it contains the exact location of each neighborhood which is essential to determine where is the best place to open the bakery.

The Foursquare location data to create the model based on the coordinates. This will help us to obtain the venues in each location.

2.2 Data cleaning

There were multiple problems with the data files which we needed to fix before starting to implement the model. First, the coordinates data was arranged in multiple pages, so we merged all the 546 rows in one single sheet. This was performed manually because the web page does not have a way to download a document with this information. Second, as the data is in Spanish, there were some special characters used not recognized in the English language which had to be manually removed. Third, we used the list of all neighborhoods and group them by Postal code and Borough in the notebook. This was done because there were some neighborhoods with the same postal code, so we joined them together. After this, we exported the csv file and added the coordinates of each neighborhoods with a VLOOKUP function in Excel.

All the columns were checked to see if missing values were present. As we are dealing with a dataset with few columns, none of them had missing values. Finally, after all the changes and grouping, we had a dataset consisting in 4 Boroughs and 207 neighborhoods.

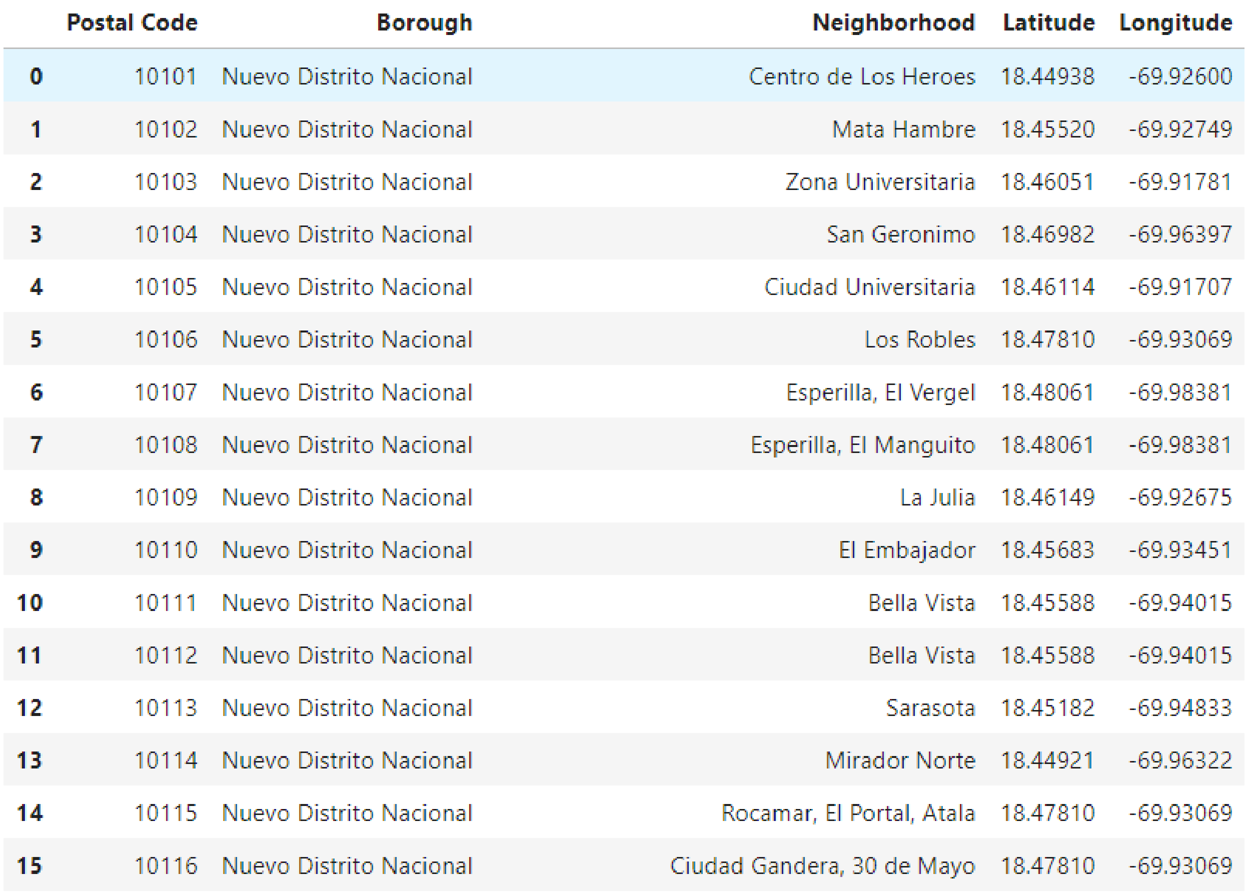
2.3 Feature selection

Kept features	Dropped features	Reason for dropping
Postal Code	N/A	Data without conflict or null values
Borough	N/A	Data without conflict or null values
Neighborhood	N/A	Data without conflict or null values
Latitude	N/A	Data without conflict or null values
Longitude	N/A	Data without conflict or null values

3. Exploratory Data Analysis

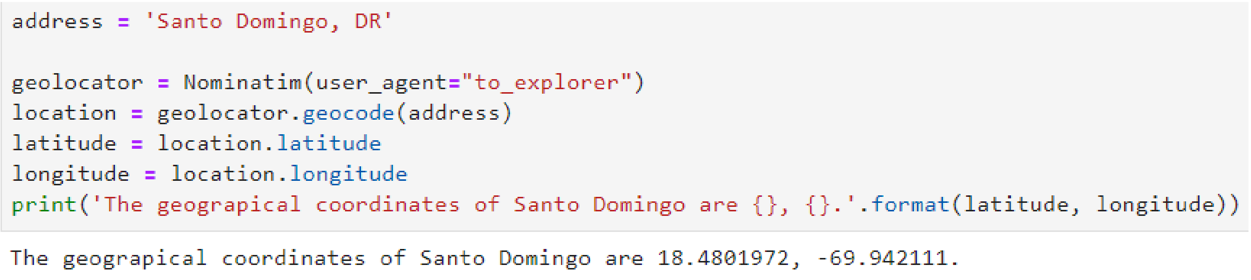
3.1 Cluster the neighborhoods in Santo Domingo

After grouping the variable specified in the data cleaning part, we had a dataset with five columns as shown below. Keep in mind that the neighborhoods are merged by postal code and Borough.



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	10101	Nuevo Distrito Nacional	Centro de Los Heroes	18.44938	-69.92600
1	10102	Nuevo Distrito Nacional	Mata Hambre	18.45520	-69.92749
2	10103	Nuevo Distrito Nacional	Zona Universitaria	18.46051	-69.91781
3	10104	Nuevo Distrito Nacional	San Geronimo	18.46982	-69.96397
4	10105	Nuevo Distrito Nacional	Ciudad Universitaria	18.46114	-69.91707
5	10106	Nuevo Distrito Nacional	Los Robles	18.47810	-69.93069
6	10107	Nuevo Distrito Nacional	Esperilla, El Vergel	18.48061	-69.98381
7	10108	Nuevo Distrito Nacional	Esperilla, El Manguito	18.48061	-69.98381
8	10109	Nuevo Distrito Nacional	La Julia	18.46149	-69.92675
9	10110	Nuevo Distrito Nacional	El Embajador	18.45683	-69.93451
10	10111	Nuevo Distrito Nacional	Bella Vista	18.45588	-69.94015
11	10112	Nuevo Distrito Nacional	Bella Vista	18.45588	-69.94015
12	10113	Nuevo Distrito Nacional	Sarasota	18.45182	-69.94833
13	10114	Nuevo Distrito Nacional	Mirador Norte	18.44921	-69.96322
14	10115	Nuevo Distrito Nacional	Rocamar, El Portal, Atala	18.47810	-69.93069
15	10116	Nuevo Distrito Nacional	Ciudad Gandera, 30 de Mayo	18.47810	-69.93069

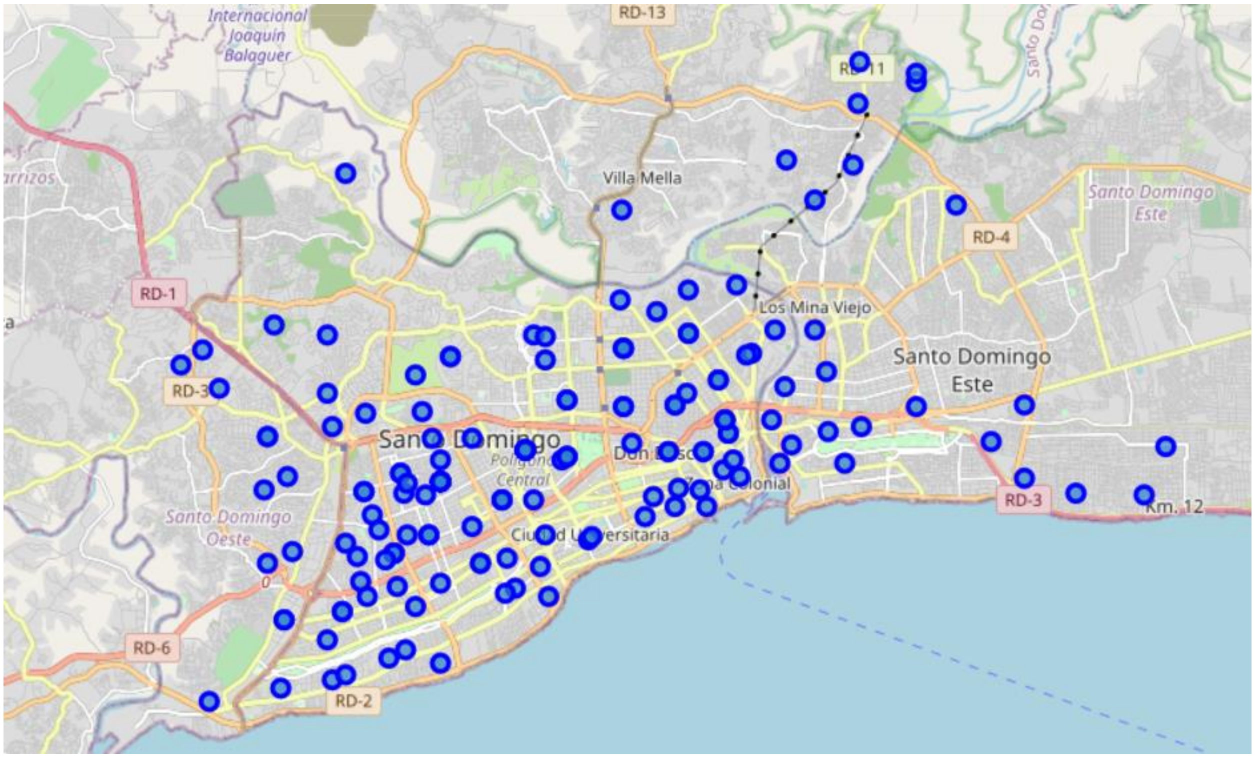
Using the geopy library we get the latitude and longitude of Santo Domingo to be able to show the map of the whole city and mark each neighborhood inside it. In order to define an instance of the geocoder, we need to define a user_agent. We will name our agent to_explorer, as shown below.



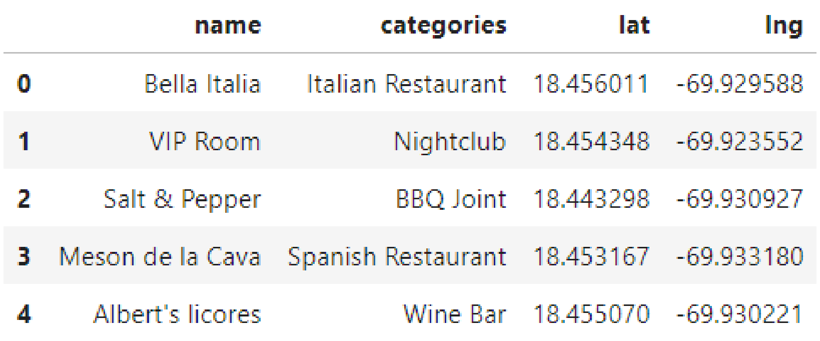
```
address = 'Santo Domingo, DR'

geolocator = Nominatim(user_agent="to_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinates of Santo Domingo are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinates of Santo Domingo are 18.4801972, -69.942111.
```

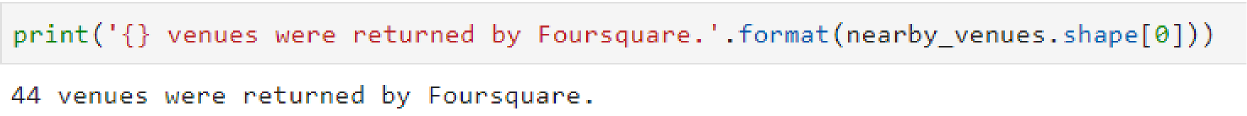
After this, we can plot the map of the city with its neighborhoods.



In order to obtain the venues for each location, we need to connect to the Foursquare API using our unique Client ID and Client Secret. Once the connection is established, we can start visualizing the neighborhoods in our dataset. We can see that the first one is 'Centro de Los Heroes' with the coordinates of 18.449379999999998, -69.926. To see the top 50 venues that are in this neighborhood we define the variables of 'Limit' and 'Radius' for the number of venues and the radius (in meters) respectively. The results are shown in a json format, so we first define a function to obtain the categories and name of the venues and then we structure the information into a panda's data frame.



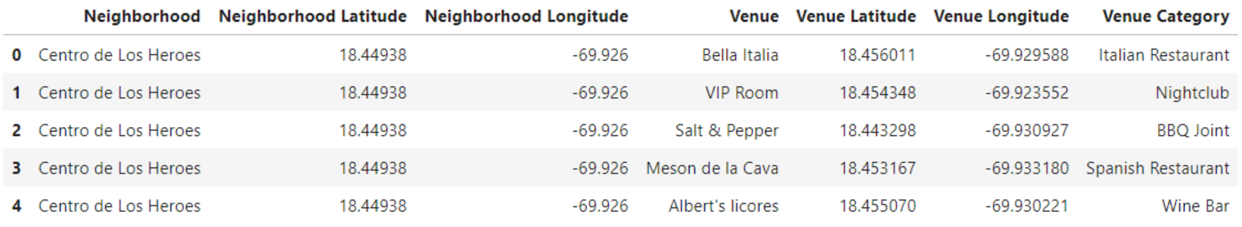
	name	categories	lat	lng
0	Bella Italia	Italian Restaurant	18.456011	-69.929588
1	VIP Room	Nightclub	18.454348	-69.923552
2	Salt & Pepper	BBQ Joint	18.443298	-69.930927
3	Meson de la Cava	Spanish Restaurant	18.453167	-69.933180
4	Albert's licores	Wine Bar	18.455070	-69.930221



```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

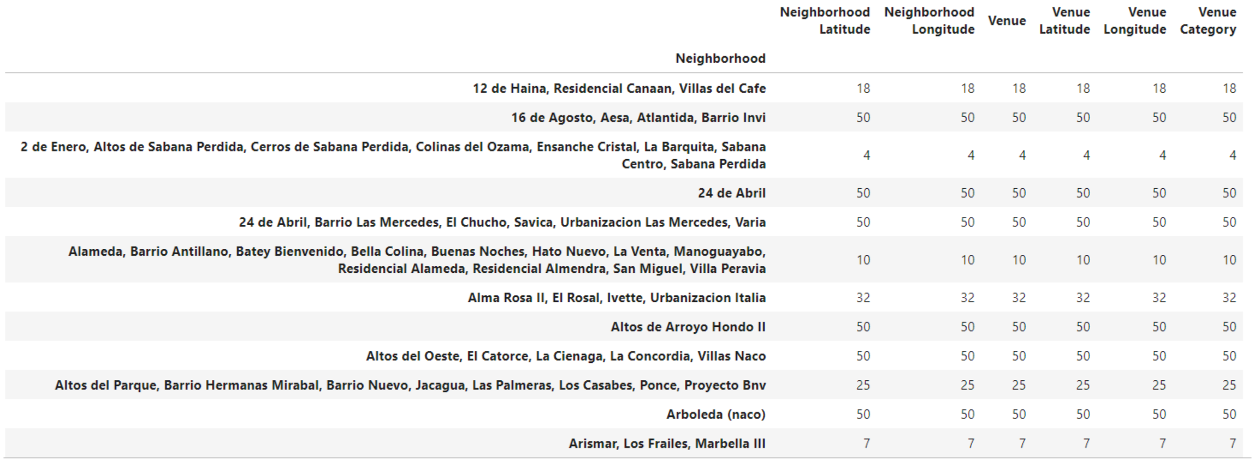
```
44 venues were returned by Foursquare.
```

Next we create a function to repeat the same process to all the neighborhoods in Santo Domingo and merge the information in our original data frame.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro de Los Heroes	18.44938	-69.926	Bella Italia	18.456011	-69.929588	Italian Restaurant
1	Centro de Los Heroes	18.44938	-69.926	VIP Room	18.454348	-69.923552	Nightclub
2	Centro de Los Heroes	18.44938	-69.926	Salt & Pepper	18.443298	-69.930927	BBQ Joint
3	Centro de Los Heroes	18.44938	-69.926	Meson de la Cava	18.453167	-69.933180	Spanish Restaurant
4	Centro de Los Heroes	18.44938	-69.926	Albert's licores	18.455070	-69.930221	Wine Bar

We can group the information by neighborhood to visualize how many venues are in each one of them.

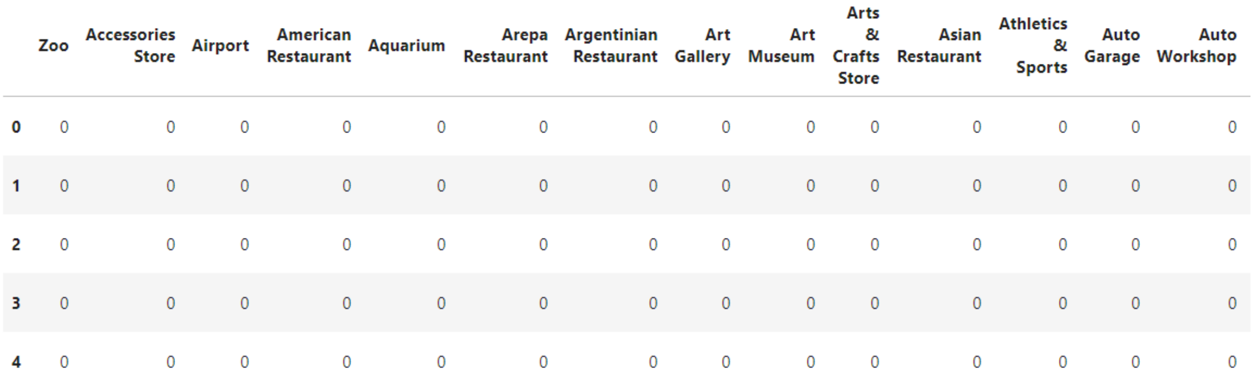


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
12 de Haina, Residencial Canaan, Villas del Cafe	18	18	18	18	18	18
16 de Agosto, Aesa, Atlantida, Barrio Invi	50	50	50	50	50	50
2 de Enero, Altos de Sabana Perdida, Cerros de Sabana Perdida, Colinas del Ozama, Ensanche Cristal, La Barquita, Sabana Centro, Sabana Perdida	4	4	4	4	4	4
24 de Abril	50	50	50	50	50	50
24 de Abril, Barrio Las Mercedes, El Chucho, Savica, Urbanizacion Las Mercedes, Varia	50	50	50	50	50	50
Alameda, Barrio Antillano, Batey Bienvenido, Bella Colina, Buenas Noches, Hato Nuevo, La Venta, Manoguayabo, Residencial Alameda, Residencial Almendra, San Miguel, Villa Peravia	10	10	10	10	10	10
Alma Rosa II, El Rosal, Ivette, Urbanizacion Italia	32	32	32	32	32	32
Altos de Arroyo Hondo II	50	50	50	50	50	50
Altos del Oeste, El Catorce, La Cienaga, La Concordia, Villas Naco	50	50	50	50	50	50
Altos del Parque, Barrio Hermanas Mirabal, Barrio Nuevo, Jacagua, Las Palmeras, Los Casabes, Ponce, Proyecto Bnv	25	25	25	25	25	25
Arboleda (naco)	50	50	50	50	50	50
Arismar, Los Frailes, Marbella III	7	7	7	7	7	7

Using the unique argument, we can see that there are 232 different categories of venues.

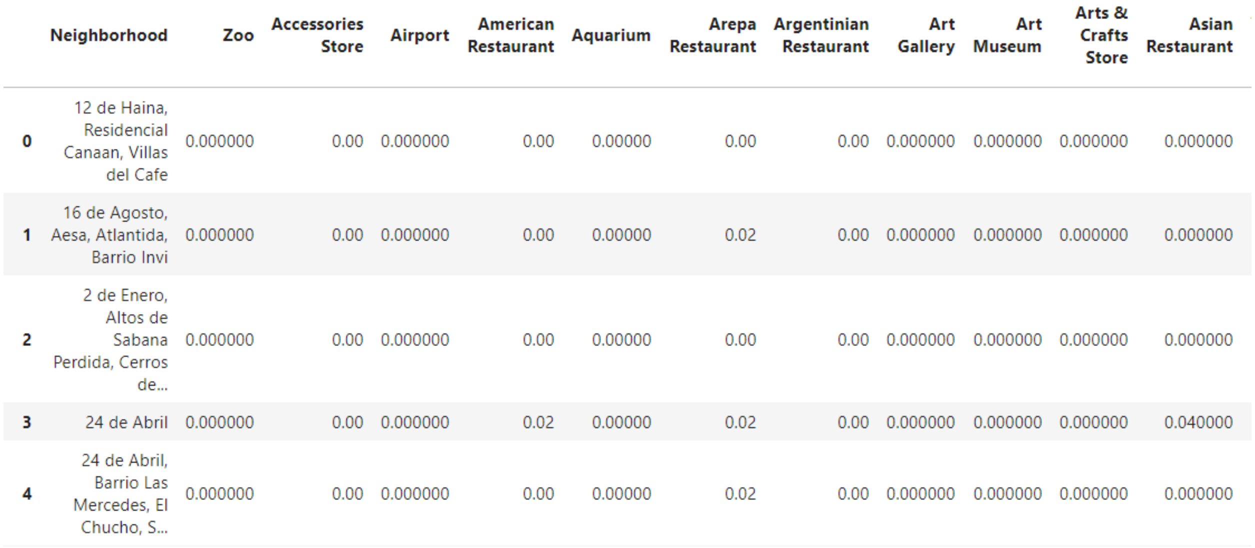
3.2 Analyze each Neighborhood

Using one hot encoding, we can create a new data frame to have each venue category as a column with a resulting shape of 6440 rows and 232 columns.



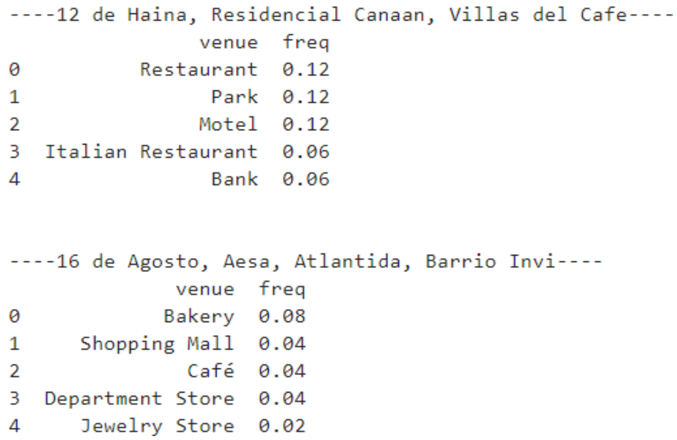
	Zoo	Accessories Store	Airport	American Restaurant	Aquarium	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Now, let’s group the rows by neighborhood and by taking the mean of the frequency of occurrence of each category.



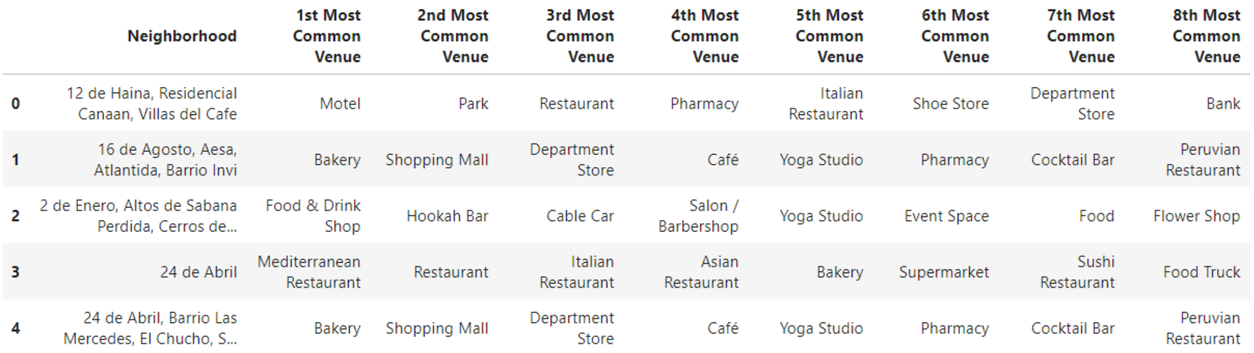
	Neighborhood	Zoo	Accessories Store	Airport	American Restaurant	Aquarium	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	12 de Haina, Residencial Canaan, Villas del Cafe	0.000000	0.00	0.000000	0.00	0.00000	0.00	0.00	0.000000	0.000000	0.000000	0.000000
1	16 de Agosto, Aesa, Atlantida, Barrio Invi	0.000000	0.00	0.000000	0.00	0.00000	0.02	0.00	0.000000	0.000000	0.000000	0.000000
2	2 de Enero, Altos de Sabana Perdida, Cerros de...	0.000000	0.00	0.000000	0.00	0.00000	0.00	0.00	0.000000	0.000000	0.000000	0.000000
3	24 de Abril	0.000000	0.00	0.000000	0.02	0.00000	0.02	0.00	0.000000	0.000000	0.000000	0.040000
4	24 de Abril, Barrio Las Mercedes, El Chucho, S...	0.000000	0.00	0.000000	0.00	0.00000	0.02	0.00	0.000000	0.000000	0.000000	0.000000

After this, we are dealing with a dataset of 160 rows and 232 columns. We can create a 'For Loop' to print each neighborhood along with the top 5 most common venues.



```
----12 de Haina, Residencial Canaan, Villas del Cafe----
                venue  freq
0          Restaurant  0.12
1                Park  0.12
2               Motel  0.12
3  Italian Restaurant  0.06
4                Bank  0.06


----16 de Agosto, Aesa, Atlantida, Barrio Invi----
              venue  freq
0            Bakery  0.08
1     Shopping Mall  0.04
2              Café  0.04
3  Department Store  0.04
4     Jewelry Store  0.02
```

Next, we define a new function and a loop to save these new values in a data frame with the ten top venues per neighborhood.

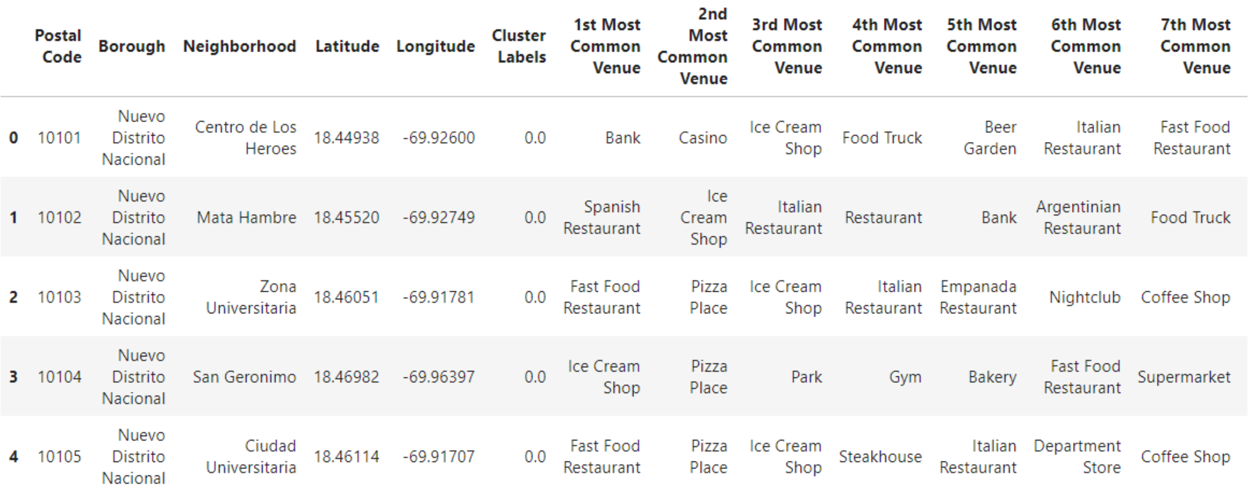


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	12 de Haina, Residencial Canaan, Villas del Cafe	Motel	Park	Restaurant	Pharmacy	Italian Restaurant	Shoe Store	Department Store	Bank
1	16 de Agosto, Aesa, Atlantida, Barrio Invi	Bakery	Shopping Mall	Department Store	Café	Yoga Studio	Pharmacy	Cocktail Bar	Peruvian Restaurant
2	2 de Enero, Altos de Sabana Perdida, Cerros de...	Food & Drink Shop	Hookah Bar	Cable Car	Salon / Barbershop	Yoga Studio	Event Space	Food	Flower Shop
3	24 de Abril	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck
4	24 de Abril, Barrio Las Mercedes, El Chucho, S...	Bakery	Shopping Mall	Department Store	Café	Yoga Studio	Pharmacy	Cocktail Bar	Peruvian Restaurant

4. Classification modeling

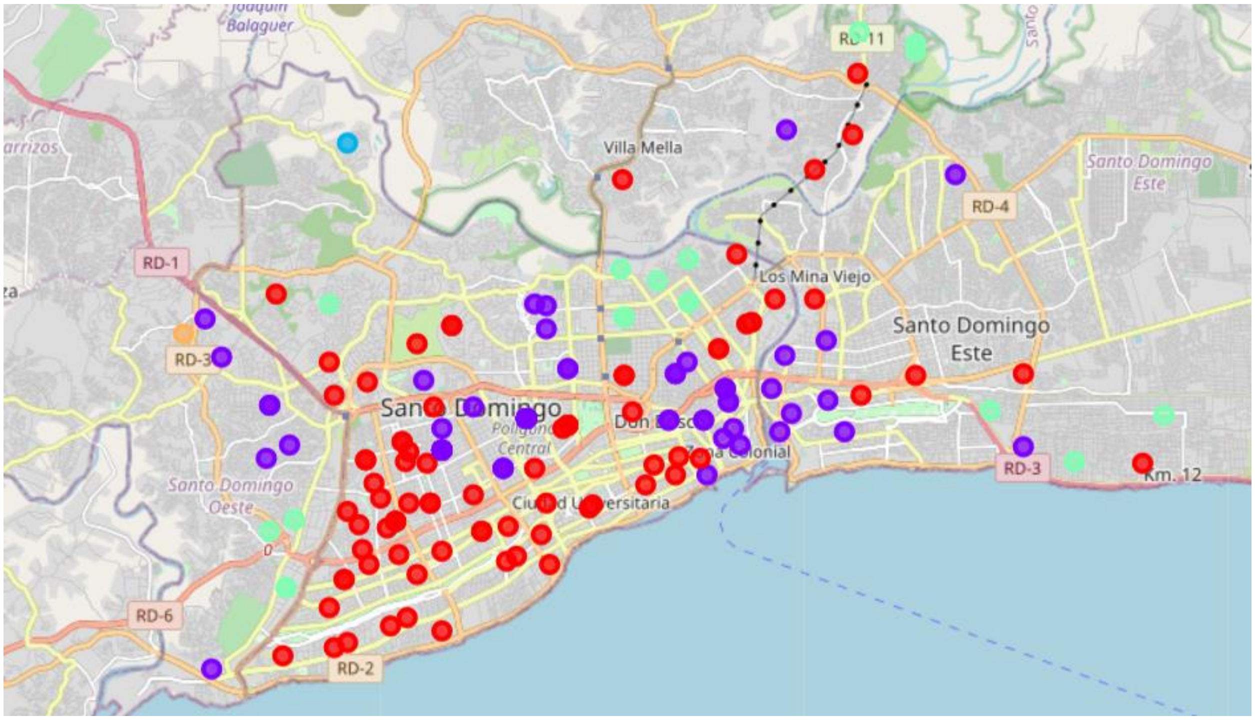
4.1 Cluster Neighborhoods

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. We are going to run K-means to divide the neighborhoods into 5 clusters and create a new data frame that includes the cluster as well as the top 10 venues for each neighborhood.



	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	10101	Nuevo Distrito Nacional	Centro de Los Heroes	18.44938	-69.92600	0.0	Bank	Casino	Ice Cream Shop	Food Truck	Beer Garden	Italian Restaurant	Fast Food Restaurant
1	10102	Nuevo Distrito Nacional	Mata Hambre	18.45520	-69.92749	0.0	Spanish Restaurant	Ice Cream Shop	Italian Restaurant	Restaurant	Bank	Argentinian Restaurant	Food Truck
2	10103	Nuevo Distrito Nacional	Zona Universitaria	18.46051	-69.91781	0.0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Italian Restaurant	Empanada Restaurant	Nightclub	Coffee Shop
3	10104	Nuevo Distrito Nacional	San Geronimo	18.46982	-69.96397	0.0	Ice Cream Shop	Pizza Place	Park	Gym	Bakery	Fast Food Restaurant	Supermarket
4	10105	Nuevo Distrito Nacional	Ciudad Universitaria	18.46114	-69.91707	0.0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Steakhouse	Italian Restaurant	Department Store	Coffee Shop

Also, we are going to visualize the resulting clusters in the following map.

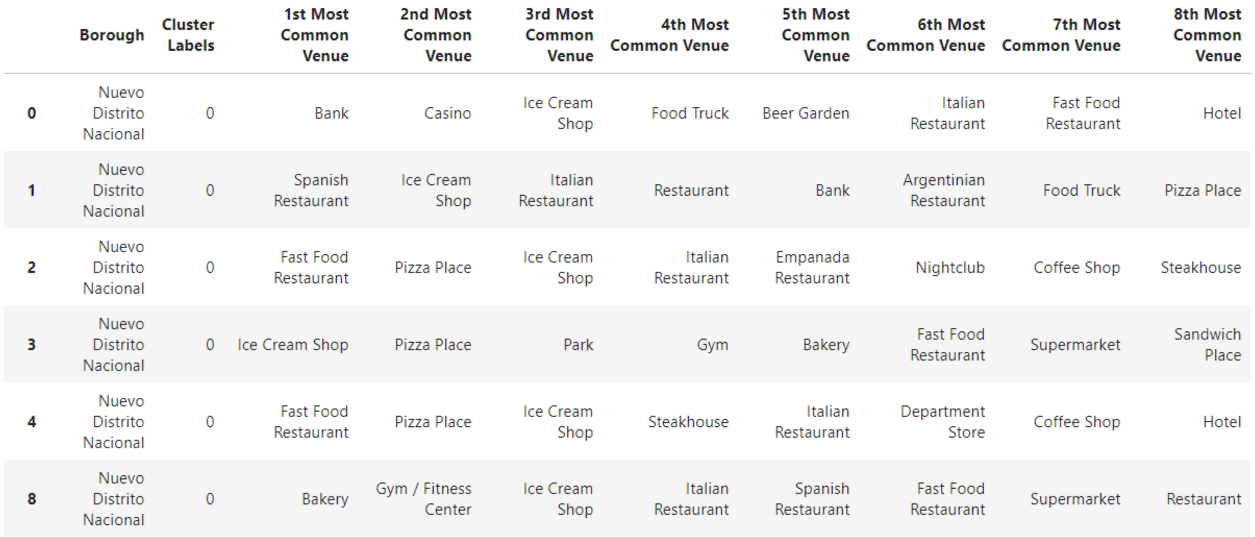


4.2 Examine Cluster

Now that we have the five clusters created, we can examine them to see which venues are present in each Cluster.

4.2.1 Cluster 1

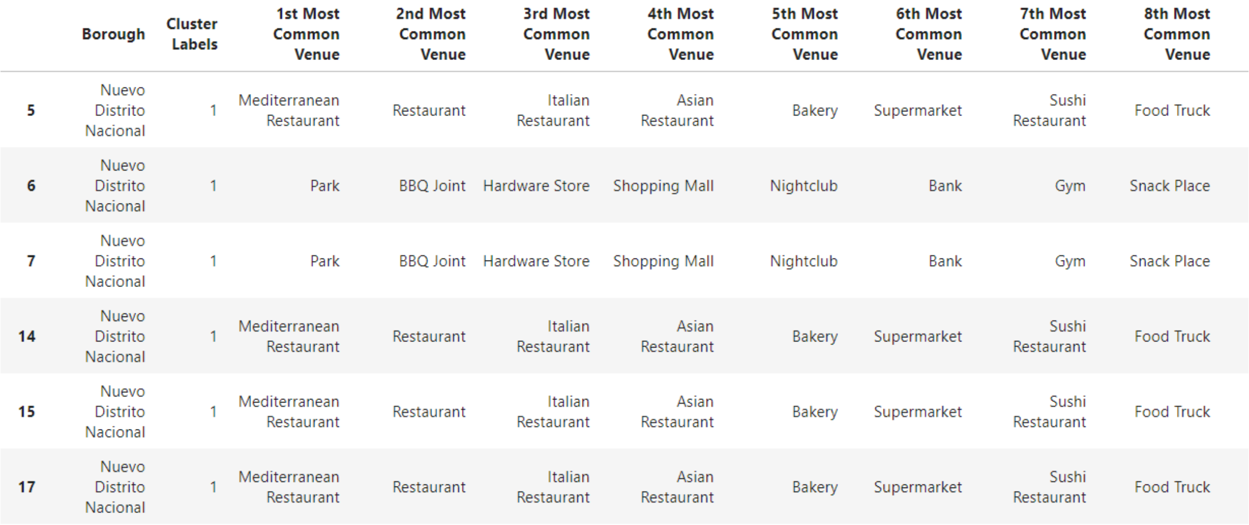
The venues more characteristic of the first cluster are: Restaurants, Pizza Places, Ice Cream Shops and Supermarkets.



	Borough	Cluster Labels	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
0	Nuevo Distrito Nacional	0	Bank	Casino	Ice Cream Shop	Food Truck	Beer Garden	Italian Restaurant	Fast Food Restaurant	Hotel
1	Nuevo Distrito Nacional	0	Spanish Restaurant	Ice Cream Shop	Italian Restaurant	Restaurant	Bank	Argentinian Restaurant	Food Truck	Pizza Place
2	Nuevo Distrito Nacional	0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Italian Restaurant	Empanada Restaurant	Nightclub	Coffee Shop	Steakhouse
3	Nuevo Distrito Nacional	0	Ice Cream Shop	Pizza Place	Park	Gym	Bakery	Fast Food Restaurant	Supermarket	Sandwich Place
4	Nuevo Distrito Nacional	0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Steakhouse	Italian Restaurant	Department Store	Coffee Shop	Hotel
8	Nuevo Distrito Nacional	0	Bakery	Gym / Fitness Center	Ice Cream Shop	Italian Restaurant	Spanish Restaurant	Fast Food Restaurant	Supermarket	Restaurant

4.2.2 Cluster 2

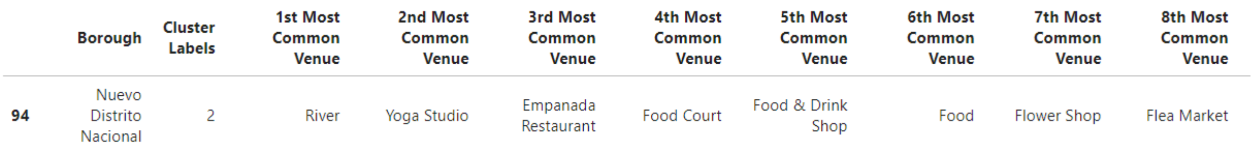
The venues more characteristic of the second cluster are: Mediterranean Restaurants, Bakeries, Parks, Nightclubs, Gyms and Lounges.



	Borough	Cluster Labels	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
5	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck
6	Nuevo Distrito Nacional	1	Park	BBQ Joint	Hardware Store	Shopping Mall	Nightclub	Bank	Gym	Snack Place
7	Nuevo Distrito Nacional	1	Park	BBQ Joint	Hardware Store	Shopping Mall	Nightclub	Bank	Gym	Snack Place
14	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck
15	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck
17	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck

4.2.3 Cluster 3

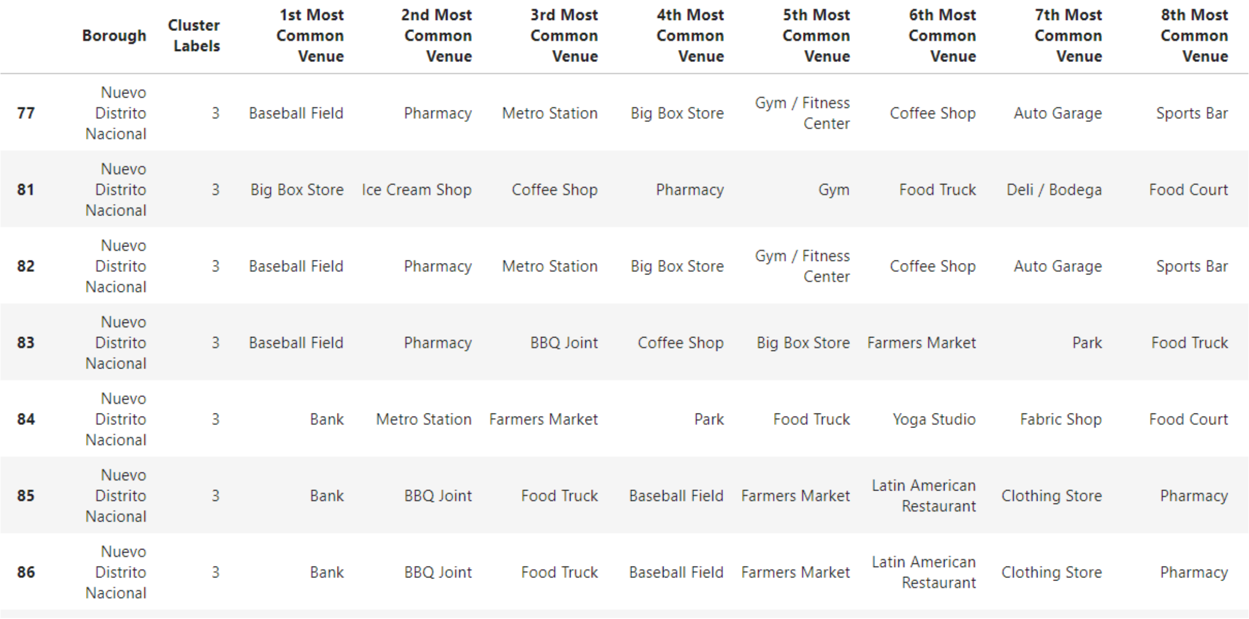
There in only one neighborhood in the third cluster with the venues of River, Yoga Studio, Empanada Restaurant, Food Court and so on.



	Borough	Cluster Labels	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
94	Nuevo Distrito Nacional	2	River	Yoga Studio	Empanada Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market

4.2.4 Cluster 4

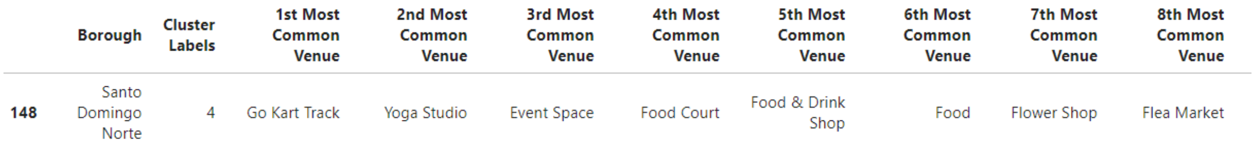
The venues more characteristic of the fourth cluster are: Baseball fields, Banks and Pharmacies.



	Borough	Cluster Labels	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
77	Nuevo Distrito Nacional	3	Baseball Field	Pharmacy	Metro Station	Big Box Store	Gym / Fitness Center	Coffee Shop	Auto Garage	Sports Bar
81	Nuevo Distrito Nacional	3	Big Box Store	Ice Cream Shop	Coffee Shop	Pharmacy	Gym	Food Truck	Deli / Bodega	Food Court
82	Nuevo Distrito Nacional	3	Baseball Field	Pharmacy	Metro Station	Big Box Store	Gym / Fitness Center	Coffee Shop	Auto Garage	Sports Bar
83	Nuevo Distrito Nacional	3	Baseball Field	Pharmacy	BBQ Joint	Coffee Shop	Big Box Store	Farmers Market	Park	Food Truck
84	Nuevo Distrito Nacional	3	Bank	Metro Station	Farmers Market	Park	Food Truck	Yoga Studio	Fabric Shop	Food Court
85	Nuevo Distrito Nacional	3	Bank	BBQ Joint	Food Truck	Baseball Field	Farmers Market	Latin American Restaurant	Clothing Store	Pharmacy
86	Nuevo Distrito Nacional	3	Bank	BBQ Joint	Food Truck	Baseball Field	Farmers Market	Latin American Restaurant	Clothing Store	Pharmacy

4.2.5 Cluster 5

There in only one neighborhood in the fifth cluster with the venues of Go Cart Track, Yoga Studio, Event Space and so on.



	Borough	Cluster Labels	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
148	Santo Domingo Norte	4	Go Kart Track	Yoga Studio	Event Space	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market

5. Results

The following were the results obtained in each of the five clusters created:

- First Cluster. In the column of first most common venue we see a tendency of business like different types of restaurants (Spanish, Fast Food, Vegetarian and Caribbean), Ice Cream Shops, Pizza places, Metro Stations, Food Trucks, BBQ and Burger Joints. So basically, in this cluster we can see businesses in the food market.
- Second Cluster. The businesses most common in this clusters are the ones destined to gather people or to offer a special kind of food like Mediterranean and Chinese restaurants, Bakeries, Bars, Café, Nightclubs, Sports Clubs, Lounges
- Third Cluster. This one is contained just by one neighborhood with the most common venue as River.
- Fourth Cluster. Health, money transactions and big gathering places are the most common venues in this clusters, like Baseball Fields, Banks and Pharmacies.
- Fifth Cluster. This one is contained just by one neighborhood with the most common venue as Go Kart Track.

6. Discussion

Josh is trying to determine the best place to open a bakery in Santo Domingo and by the K-means clustering, we can determine that the second cluster contains the most common bakery venues. Interestingly, these bakeries are not in the Borough of Nuevo Distrito Nacional (where the most people live and the most developed borough) but in Santo Domingo Oeste (one of the least developed Boroughs in Santo Domingo. These means that the people who live in the neighborhoods of these borough tend to buy much more in bakeries than in the neighborhoods of different boroughs. If Josh wants to guarantee the success of his business, then the neighborhoods in Santo Domingo Oeste are the ones with the highest possibilities of achieving that.

7. Conclusion

In this study, I analyzed the most common venues of all the neighborhoods in the city of Santo Domingo, Dominican Republic to determined where is the best place to open a Bakery Store. After adjusting the data, I used the K-means clustering method to divide the city into five different groups and determined which one had bakeries as the most common venues. This model could help anybody to analyze the data of any city or country and determine specific characteristics to be able to answer multiple questions.