BAKERY STORE IN SANTO DOMINGO

Junior Peña

INTRODUCTION

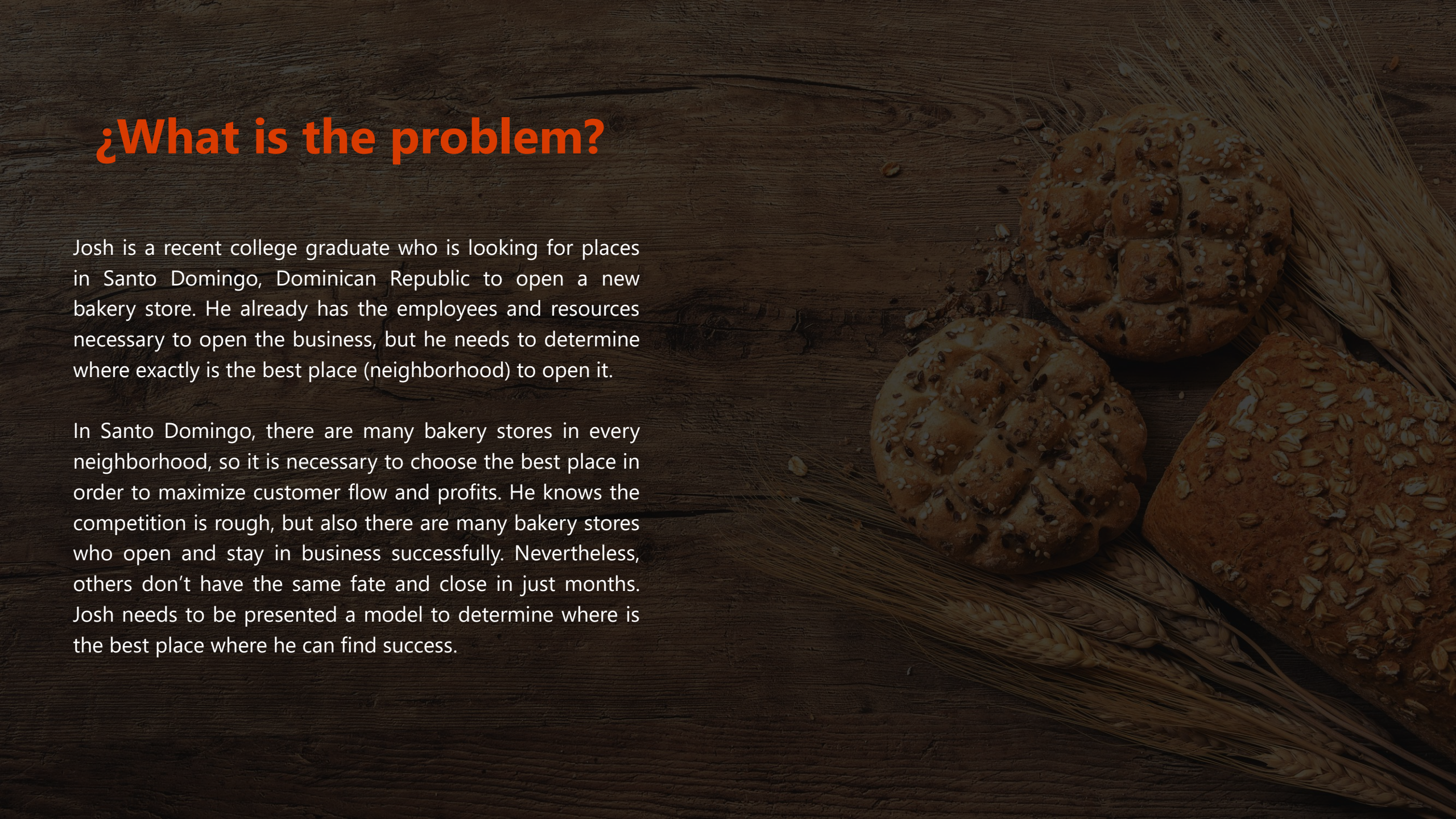
Background

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.

¿What is the 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.



DATA ACQUISITION AND CLEANING

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.

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.

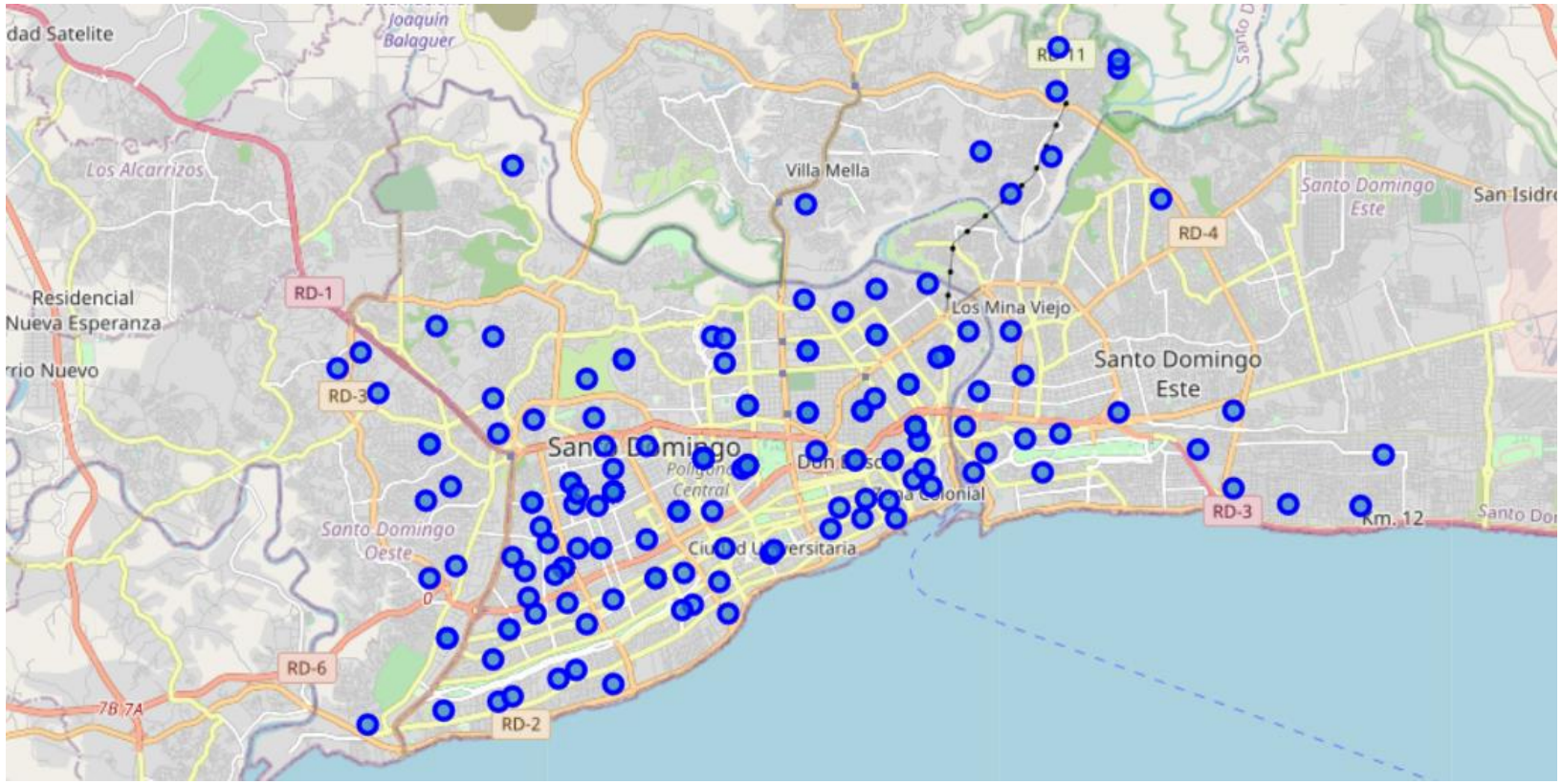
Feature Selection

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

EXPLORATORY DATA ANALYSIS

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. 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.



Cluster the neighborhoods in Santo Domingo

We define a function to obtain the categories and name of the venues and then we structure the information into a panda's data frame. We can group the information by neighborhood to visualize how many venues are in each one of them.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
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

Analyze each Neighborhood

Using one hot encoding, we can create a new data frame to have each venue category as a column and then 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

Analyze each Neighborhood

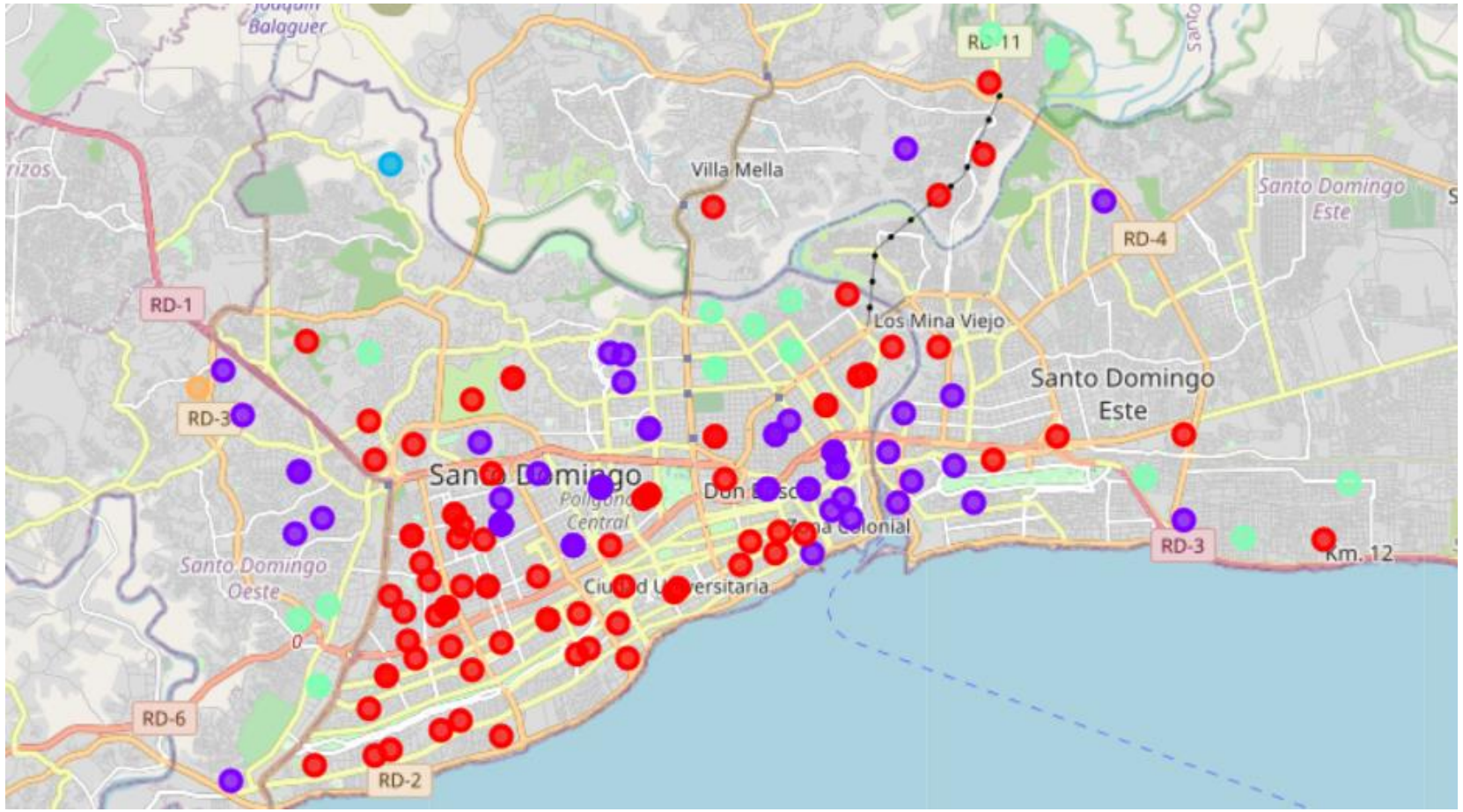
We can create a 'For Loop' to print each neighborhood along with the top 5 most common venues. Next, we define a new function and a loop to save these new values in a data frame with the ten top venues per neighborhood.

	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	9th Most Common Venue	10th Most Common Venue
0	12 de Haina, Residencial Canaan, Villas del Cafe	Motel	Park	Restaurant	Pharmacy	Italian Restaurant	Shoe Store	Department Store	Bank	Toll Booth	Pier
1	16 de Agosto, Aesa, Atlantida, Barrio Invi	Bakery	Shopping Mall	Department Store	Café	Yoga Studio	Pharmacy	Cocktail Bar	Peruvian Restaurant	Pet Store	Pizza Place
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	Flea Market	Film Studio
3	24 de Abril	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop
4	24 de Abril, Barrio Las Mercedes, El Chucho, S...	Bakery	Shopping Mall	Department Store	Café	Yoga Studio	Pharmacy	Cocktail Bar	Peruvian Restaurant	Pet Store	Pizza Place

CLASSIFICATION MODELING

Cluster Neighborhoods

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.



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	9th Most Common Venue	10th Most Common Venue
0	Nuevo Distrito Nacional	0	Bank	Casino	Ice Cream Shop	Food Truck	Beer Garden	Italian Restaurant	Fast Food Restaurant	Hotel	Sandwich Place	Beer Store
1	Nuevo Distrito Nacional	0	Spanish Restaurant	Ice Cream Shop	Italian Restaurant	Restaurant	Bank	Argentinian Restaurant	Food Truck	Pizza Place	Empanada Restaurant	Latin American Restaurant
2	Nuevo Distrito Nacional	0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Italian Restaurant	Empanada Restaurant	Nightclub	Coffee Shop	Steakhouse	Food Court	Mediterranean Restaurant
3	Nuevo Distrito Nacional	0	Ice Cream Shop	Pizza Place	Park	Gym	Bakery	Fast Food Restaurant	Supermarket	Sandwich Place	Italian Restaurant	Empanada Restaurant
4	Nuevo Distrito Nacional	0	Fast Food Restaurant	Pizza Place	Ice Cream Shop	Steakhouse	Italian Restaurant	Department Store	Coffee Shop	Hotel	Indian Restaurant	Food Court
8	Nuevo Distrito Nacional	0	Bakery	Gym / Fitness Center	Ice Cream Shop	Italian Restaurant	Spanish Restaurant	Fast Food Restaurant	Supermarket	Restaurant	Furniture / Home Store	Gift Shop
9	Nuevo Distrito Nacional	0	Pizza Place	Gym	Spanish Restaurant	Ice Cream Shop	Argentinian Restaurant	Food Truck	Italian Restaurant	Gym / Fitness Center	Coffee Shop	American Restaurant

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	9th Most Common Venue	10th Most Common Venue
5	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop
6	Nuevo Distrito Nacional	1	Park	BBQ Joint	Hardware Store	Shopping Mall	Nightclub	Bank	Gym	Snack Place	Diner	Dim Sum Restaurant
7	Nuevo Distrito Nacional	1	Park	BBQ Joint	Hardware Store	Shopping Mall	Nightclub	Bank	Gym	Snack Place	Diner	Dim Sum Restaurant
14	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop
15	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop
17	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop
24	Nuevo Distrito Nacional	1	Mediterranean Restaurant	Restaurant	Italian Restaurant	Asian Restaurant	Bakery	Supermarket	Sushi Restaurant	Food Truck	Pizza Place	Wine Shop

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	9th Most Common Venue	10th Most Common Venue
94	Nuevo Distrito Nacional	2	River	Yoga Studio	Empanada Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Film Studio	Fast Food Restaurant

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	9th Most Common Venue	10th 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	Fast Food Restaurant	Falafel Restaurant
81	Nuevo Distrito Nacional	3	Big Box Store	Ice Cream Shop	Coffee Shop	Pharmacy	Gym	Food Truck	Deli / Bodega	Food Court	Cupcake Shop	Food & Drink Shop
82	Nuevo Distrito Nacional	3	Baseball Field	Pharmacy	Metro Station	Big Box Store	Gym / Fitness Center	Coffee Shop	Auto Garage	Sports Bar	Fast Food Restaurant	Falafel Restaurant
83	Nuevo Distrito Nacional	3	Baseball Field	Pharmacy	BBQ Joint	Coffee Shop	Big Box Store	Farmers Market	Park	Food Truck	Dim Sum Restaurant	Diner
84	Nuevo Distrito Nacional	3	Bank	Metro Station	Farmers Market	Park	Food Truck	Yoga Studio	Fabric Shop	Food Court	Food & Drink Shop	Food
85	Nuevo Distrito Nacional	3	Bank	BBQ Joint	Food Truck	Baseball Field	Farmers Market	Latin American Restaurant	Clothing Store	Pharmacy	Flea Market	Film Studio
86	Nuevo Distrito Nacional	3	Bank	BBQ Joint	Food Truck	Baseball Field	Farmers Market	Latin American Restaurant	Clothing Store	Pharmacy	Flea Market	Film Studio

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	9th Most Common Venue	10th 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	Film Studio	Fast Food Restaurant

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.

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.

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.

THANK YOU!

Junior Peña

junior03b@gmail.com