IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

A SHOPPING MALL FOR SANTO DOMINGO ESTE, DOMINICAN REPUBLIC

28.JAN.2021

José Luis D’Andrade
Distrito Nacional, Dominican Republic

Introduction

Shopping malls have been increasingly important in modern society. Our visits are not limited to buying things any more. They have become the children's playground. Adults spent the day walking through elegant alleys equipped with benches, flowers and even palms as a means of exercising. Today, most shopping malls have many restaurants, bars, cafes or even hairdressers, beauty salons, gyms, cinemas and other entertainment attractions which enables us to fulfil a lot of different needs in the area of a single building. Social life is gradually transferring from the areas of old towns and main streets to shopping centres.

The **Distrito Nacional** is a subdivision of the **Dominican Republic** enclosing the capital Santo Domingo. Before 2001, the Distrito Nacional was a large area included in what is now known as Santo Domingo Province. The Law 163-01 created the province and separated the Distrito Nacional from other municipalities. **Santo Domingo Este** was created.

Santo Domingo Este is across the Ozama River which divides the east and west sections of metropolitan Santo Domingo. It is more residential and less commercially developed, but it has experienced growth since its creation, with new malls, department stores, racetrack, water parks, aquarium and many other attractions.

Business Problem

The Distrito National has a high density of housing and businesses. Transportation is a growing issue. In the last two decades many shopping centers have experienced a decline in attracting visiting public and a drop in their commercial activities. The current economic climate and culture of "new is better than old" has left many commercial centers built in the 80's, 90's and early 2000, vacant and disused. It may be time to look elsewhere when thinking about new commercial plazas.

Santo Domingo Este, has a booming economy which is rapidly attracting the interest of many not just as a living destination, but for investing purposes as well. The rapid growth poses a problem when trying to decide where to open a business.

Following a data science methodology and utilizing machine learning techniques, this work aims to provide a guide to answer investors, developers, construction industry in general business questions. Geospatial analysis can help us to select the best location for opening a new shopping mall in the city of Santo Domingo Este.

The objective of this work is to analyze and select the best location to open a new shopping mall in the city of Santo Domingo Este. Following a data science methodology and utilizing machine learning techniques, this work aims to provide a guide to answer the business question:

Considering the issues in the Distrito Nacional, and the sustained growth of the east side of metropolitan Santo Domingo, where in the city of Santo Domingo Este would be the best location to build a new shopping mall?

Description of the data

I. List of neighbourhoods in Santo Domingo Este

The scope of this project is constrained to the city of Santo Domingo Este, the second most important municipality of the province of Santo Domingo.

II. Latitude and longitude coordinates of neighbourhood

Geocoding is the process of transforming a description of a location, such as an address, or a name of a place, to a location on the earth's surface. The resulting locations are output as geographic features with attributes, which can be used for mapping or spatial analysis. We will geocode neighbourhoods and process venues in the surrounding latitude and longitude.

III. Venue data

Data of businesses in the vicinity of the geocoded neighbourhoods. We will use this data for cluster analysis.

Data Sources, APIs and Python Libraries

Government. I have worked with government data in the past and I am familiar with this government page https://www.one.gob.do/ . It is Dominican Republic’s head department in charge of statistics. There are several **databases**..

Foursquare API. After obtaining geolocation data, we will use **Foursquare API** and, if necessary, **Google Places API**, to get the **venue data** for those neighbourhoods. Foursquare has one of the largest databases of places and is used by over 125,000 developers.

Foursquare API will provide many categories of the venue data. Our main interest is the **Shopping Mall category**. This is a project that will make use of many data science skills from web scraping, working with API (Foursquare, Google), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Python Libraries. We will get geographical coordinates using Python Geocoder package which will give us coordinates of the neighbourhoods. Other libraries to be used:

Pandas: For creating and manipulating vectors and matrices.

Folium: For data visualization, to visualize the neighborhood cluster distribution.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.

Map of Dominican Republic

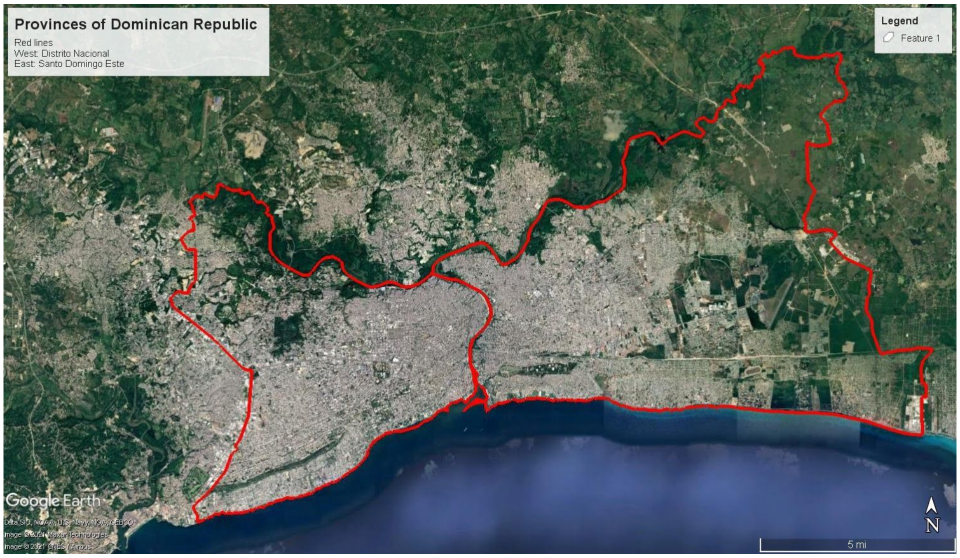
The map shows the island of La Hispaniola, one of the few shared by more than one country in the world. The left portion is Haiti. On the right we show Dominican Republic provinces.

Provinces of Dominican Republic
Red lines
West: Distrito Nacional
East: Santo Domingo Este
Legend
Feature 1
Google Earth
100 mi
N

Geographic Location of Interest

The following image shows the nation's Capital, on the left, namely Distrito National and on the right Santo Domingo Este.

The left side is approximately 91 square kilometers. It has the majority of shopping malls, banks and



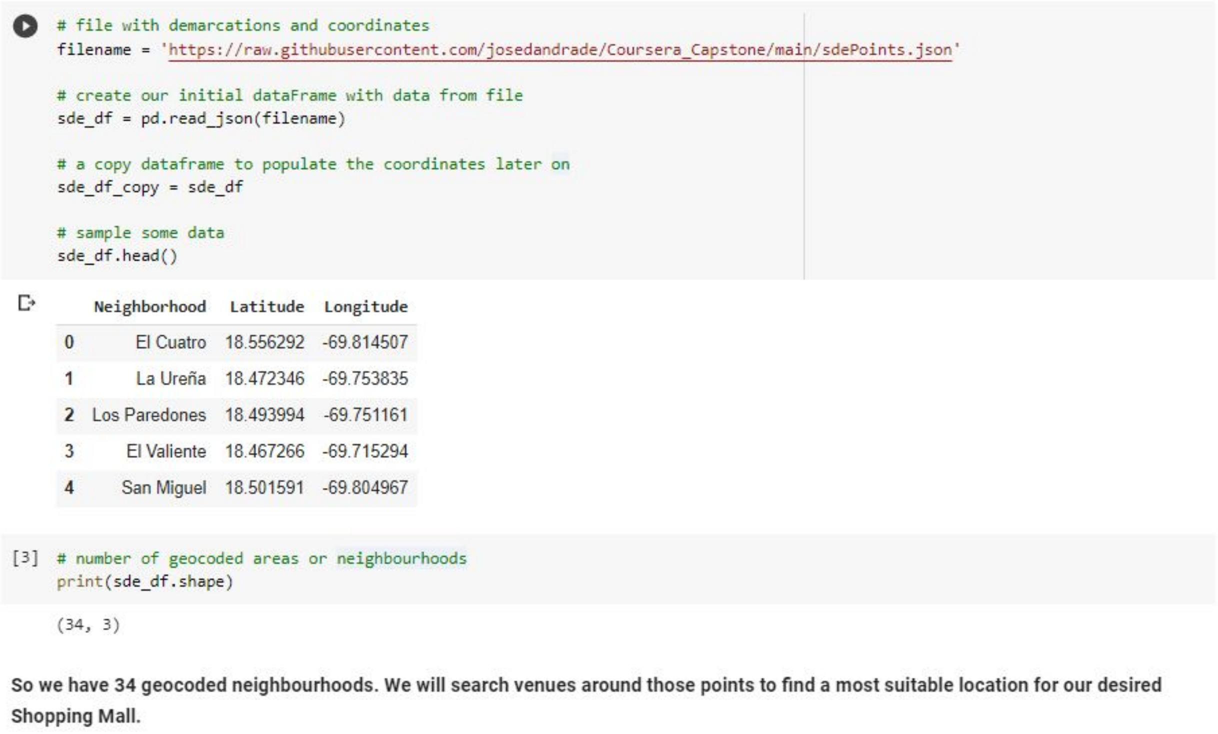
restaurants in the Dominican Republic. And on the right, with 106 square kilometers, rapid growth and only a short distance away to the main airport, tourism destinations, and many opportunities there lies Santo Domingo Este, our focus of interest in this study.

Methodology

Geospatial analysis can help us to select the best location for opening a new shopping mall in the city of Santo Domingo Este. We will follow a data science methodology and utilize machine learning techniques to create a model.

We followed a process of selecting neighbourhoods for geocoding. This is the process of transforming a description of a location, such as an address, or a name of a place, to a location on the earth's surface. The resulting locations are output as geographic features with attributes, which can be used for mapping or spatial analysis. The features are expressed in terms of latitude and longitude, or coordinates, of neighbourhoods.

Initially, we had 308 housing projects. According to the city municipality they are located **in 34 demarcations**. We will refer to them **as neighbourhoods**.



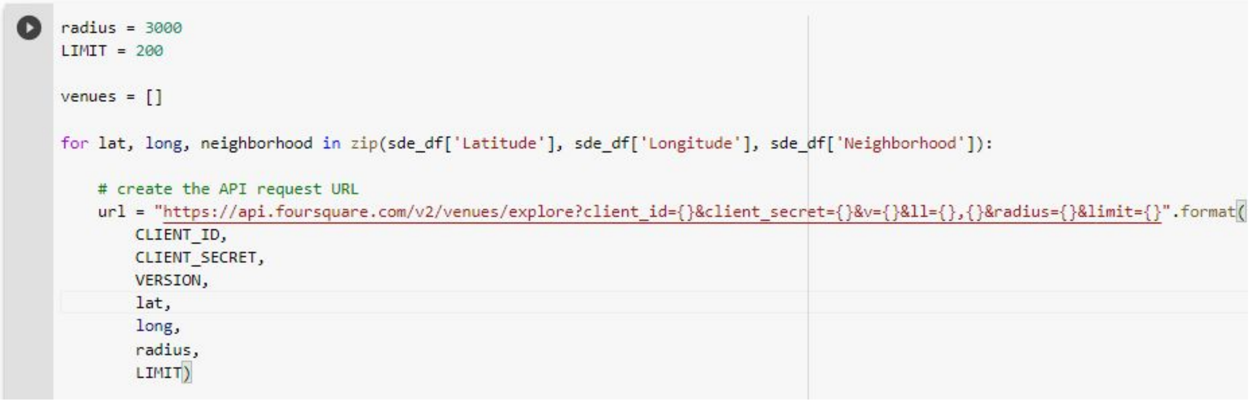
```
# file with demarcations and coordinates
filename = 'https://raw.githubusercontent.com/josedandrade/Coursera_Capstone/main/sdePoints.json'

# create our initial dataFrame with data from file
sde_df = pd.read_json(filename)

# a copy dataframe to populate the coordinates later on
sde_df_copy = sde_df

# sample some data
sde_df.head()
```

	Neighborhood	Latitude	Longitude
0	El Cuatro	18.556292	-69.814507
1	La Ureña	18.472346	-69.753835
2	Los Paredones	18.493994	-69.751161
3	El Valiente	18.467266	-69.715294
4	San Miguel	18.501591	-69.804967

```
[3] # number of geocoded areas or neighbourhoods
    print(sde_df.shape)

    (34, 3)
```

So we have 34 geocoded neighbourhoods. We will search venues around those points to find a most suitable location for our desired Shopping Mall.

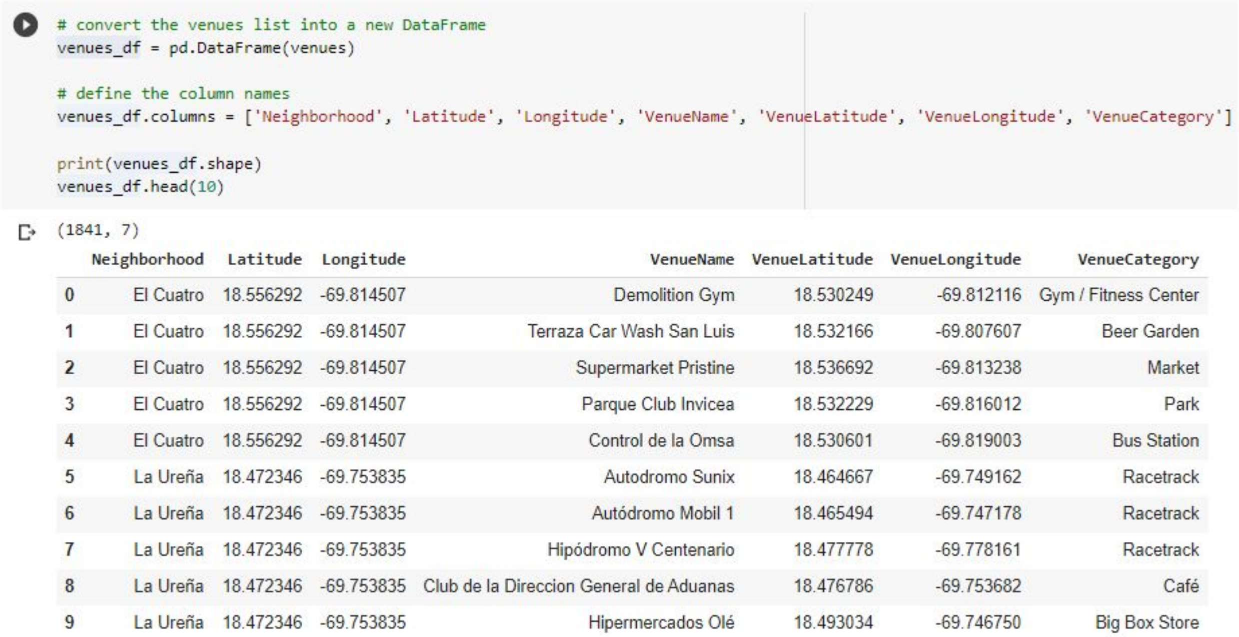
For every neighbourhood, within a radius of 2 km, we searched for a maximum of 200 venues. We used FOURSQUARE services to get access to global Points Of Interest data and rich content, such as Shopping Mall, Restaurant, etc. The resulting venue categories are new attributes that describe neighbourhoods.



```
radius = 3000
LIMIT = 200

venues = []

for lat, long, neighborhood in zip(sde_df['Latitude'], sde_df['Longitude'], sde_df['Neighborhood']):

    # create the API request URL
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)
```

We were able to locate **1841 venues**.

Sampling our data



```
# convert the venues list into a new DataFrame
venues_df = pd.DataFrame(venues)

# define the column names
venues_df.columns = ['Neighborhood', 'Latitude', 'Longitude', 'VenueName', 'VenueLatitude', 'VenueLongitude', 'VenueCategory']

print(venues_df.shape)
venues_df.head(10)
```

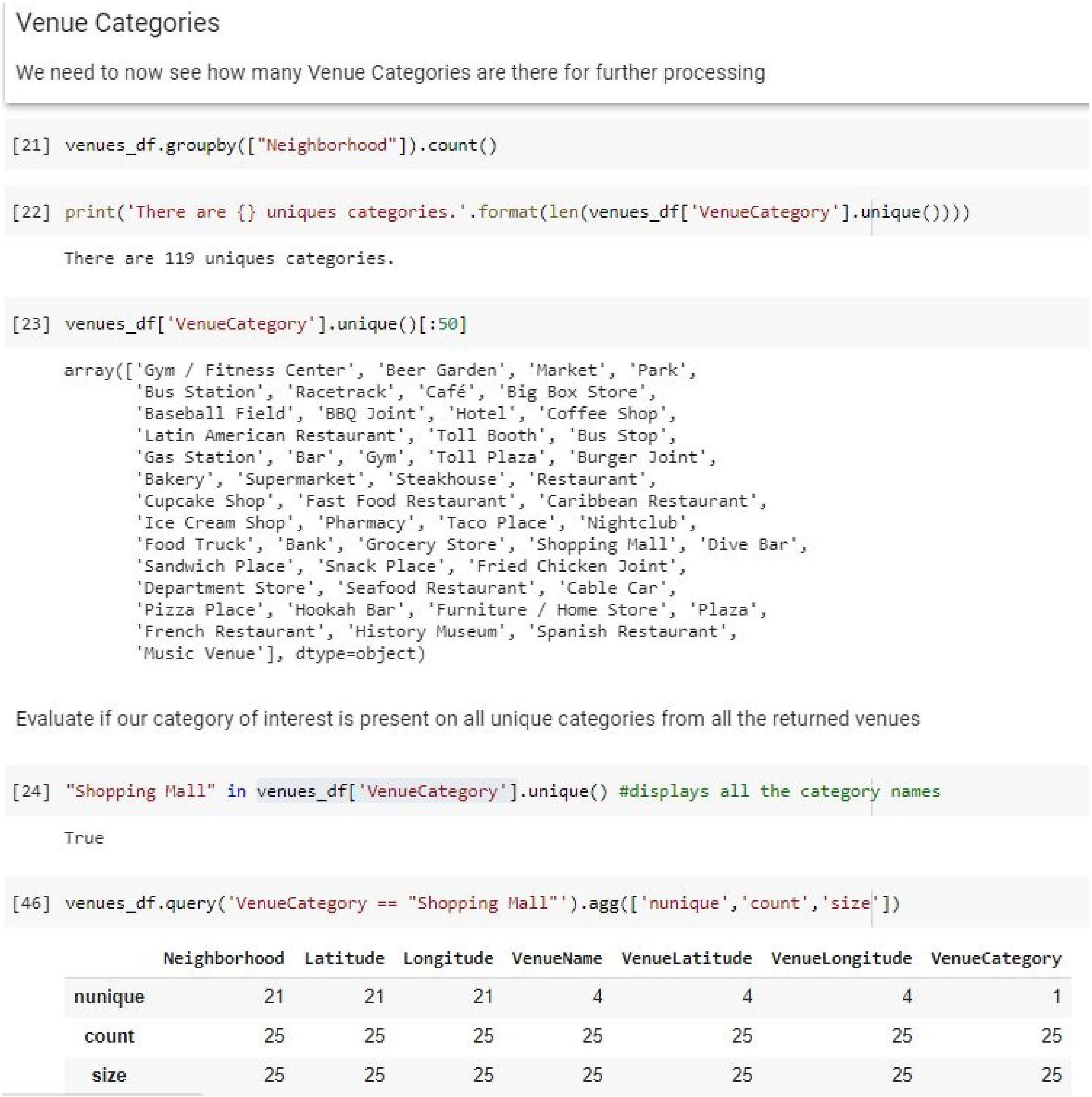
```
(1841, 7)
```

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	El Cuatro	18.556292	-69.814507	Demolition Gym	18.530249	-69.812116	Gym / Fitness Center
1	El Cuatro	18.556292	-69.814507	Terraza Car Wash San Luis	18.532166	-69.807607	Beer Garden
2	El Cuatro	18.556292	-69.814507	Supermarket Pristine	18.536692	-69.813238	Market
3	El Cuatro	18.556292	-69.814507	Parque Club Invicea	18.532229	-69.816012	Park
4	El Cuatro	18.556292	-69.814507	Control de la Omsa	18.530601	-69.819003	Bus Station
5	La Ureña	18.472346	-69.753835	Autodromo Sunix	18.464667	-69.749162	Racetrack
6	La Ureña	18.472346	-69.753835	Autódromo Mobil 1	18.465494	-69.747178	Racetrack
7	La Ureña	18.472346	-69.753835	Hipódromo V Centenario	18.477778	-69.778161	Racetrack
8	La Ureña	18.472346	-69.753835	Club de la Direccion General de Aduanas	18.476786	-69.753682	Café
9	La Ureña	18.472346	-69.753835	Hipermercados Olé	18.493034	-69.746750	Big Box Store

We have all the venues in our location of interest, **Santo Domingo Este**, and it's demarcations, the 34 neighbourhoods. We have collected all venues within a radius of 3 km of every neighbourhood center.

Every observed neighbourhood is now being described by all of its venues, coordinates and a category of venue. This is useful to query if our desired category is present.

We are also interested in knowing all the categories that came up. We found **119 categories**.



Venue Categories

We need to now see how many Venue Categories are there for further processing

```
[21] venues_df.groupby(["Neighborhood"]).count()
```

```
[22] print('There are {} uniques categories.'.format(len(venues_df['VenueCategory'].unique())))

     There are 119 uniques categories.
```

```
[23] venues_df['VenueCategory'].unique()[:50]

     array(['Gym / Fitness Center', 'Beer Garden', 'Market', 'Park',
            'Bus Station', 'Racetrack', 'Café', 'Big Box Store',
            'Baseball Field', 'BBQ Joint', 'Hotel', 'Coffee Shop',
            'Latin American Restaurant', 'Toll Booth', 'Bus Stop',
            'Gas Station', 'Bar', 'Gym', 'Toll Plaza', 'Burger Joint',
            'Bakery', 'Supermarket', 'Steakhouse', 'Restaurant',
            'Cupcake Shop', 'Fast Food Restaurant', 'Caribbean Restaurant',
            'Ice Cream Shop', 'Pharmacy', 'Taco Place', 'Nightclub',
            'Food Truck', 'Bank', 'Grocery Store', 'Shopping Mall', 'Dive Bar',
            'Sandwich Place', 'Snack Place', 'Fried Chicken Joint',
            'Department Store', 'Seafood Restaurant', 'Cable Car',
            'Pizza Place', 'Hookah Bar', 'Furniture / Home Store', 'Plaza',
            'French Restaurant', 'History Museum', 'Spanish Restaurant',
            'Music Venue'], dtype=object)
```

Evaluate if our category of interest is present on all unique categories from all the returned venues

```
[24] "Shopping Mall" in venues_df['VenueCategory'].unique() #displays all the category names

     True
```

```
[46] venues_df.query('VenueCategory == "Shopping Mall"').agg(['nunique','count','size'])
```

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
nunique	21	21	21	4	4	4	1
count	25	25	25	25	25	25	25
size	25	25	25	25	25	25	25

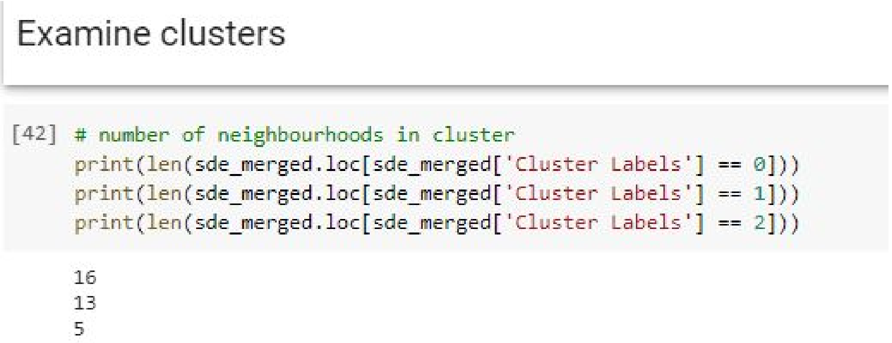
We also know there are Shopping Malls in the area. Because of the radius of search some neighborhood might show as having a shopping mall. There were **21 neighborhoods** and **25 venues** categorized as such.

This concludes the data gathering aspect of our study. We are going to use this data for analysis and to produce the report on optimal locations for a new Shopping Mall.

Our choice of machine learning algorithm is K-means clustering. With K-means we can group data based on the similarity. It's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

Results

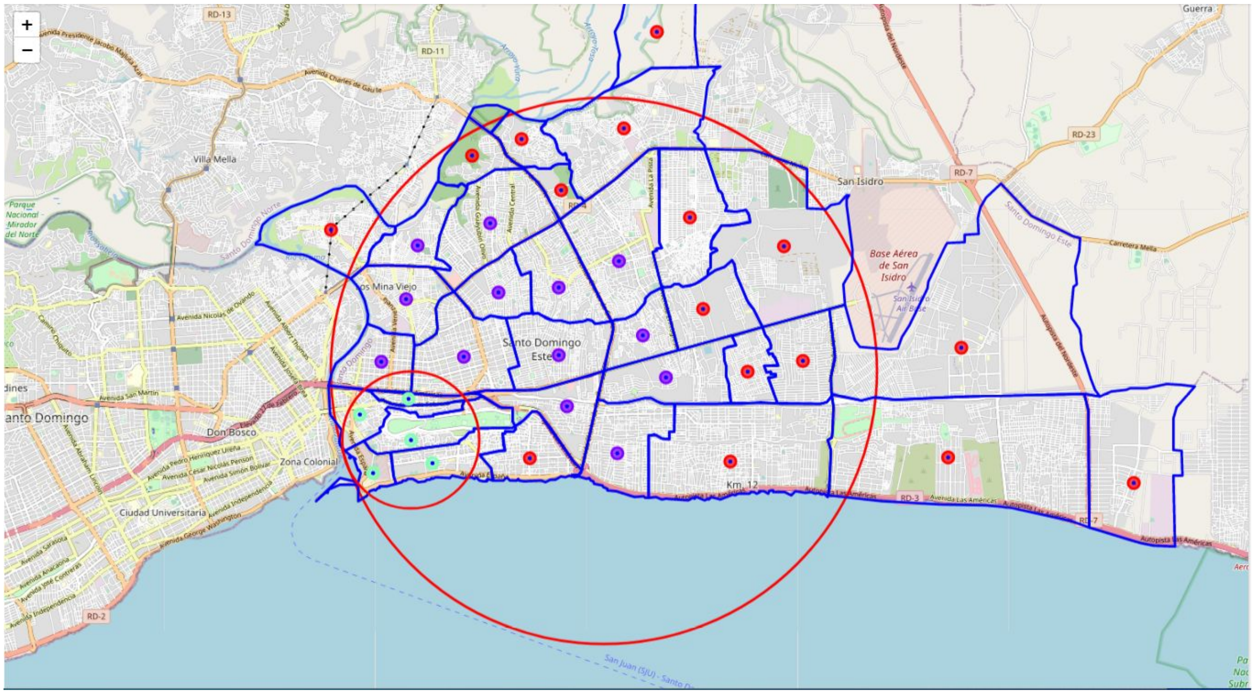
With every observed neighbourhood and its attributes we were able to find similar neighbourhoods. The model created **3 clusters** of locations that meet some basic requirements established in discussion with stakeholders.



Examine clusters

```
[42] # number of neighbourhoods in cluster
     print(len(sde_merged.loc[sde_merged['Cluster Labels'] == 0]))
     print(len(sde_merged.loc[sde_merged['Cluster Labels'] == 1]))
     print(len(sde_merged.loc[sde_merged['Cluster Labels'] == 2]))
```

```
16
13
5
```



A good number of shopping locations are in the central area of Santo Domingo Este, with the highest number in cluster 1 and almost the same moderate number in cluster 0.

It seems there is opportunity and high potential areas to open new shopping malls as there is no competition from existing malls in Cluster 2 (small circle).

Shopping malls in cluster 0 and 1 are likely suffering from intense competition due to high concentration of shopping locations.

Conclusion

This project recommends property developers to open new shopping malls in neighbourhoods in cluster 2 where there is no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 0 with moderate competition.

One other observation is the surrounding areas in cluster 2. A Museum, Aquatic Park, Aquarium and Racetrack are among the attractions in the area that are a walking distance away or a few minutes driving. And, better yet, there is unbuilt land.