IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

A SHOPPING MALL FOR SANTO DOMINGO ESTE, DOMINICAN REPUBLIC

17.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.

Business Problem

The province of Santo Domingo has two competing municipalities. There is the National District, and there is Santo Domingo Este, a booming economy which is rapidly attracting the interest of many in the National District not just for investing purposes, but as a living destination.

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 similarities with the National District, 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

This is required in order to plot the map and also to get venues data.

III. Venue data

Data of shopping malls. We will use this data to perform clustering on the neighbourhoods.

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**. One of such databases contains **demographic information** about every province, municipality down to **neighbourhoods** in the country.

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.

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 dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

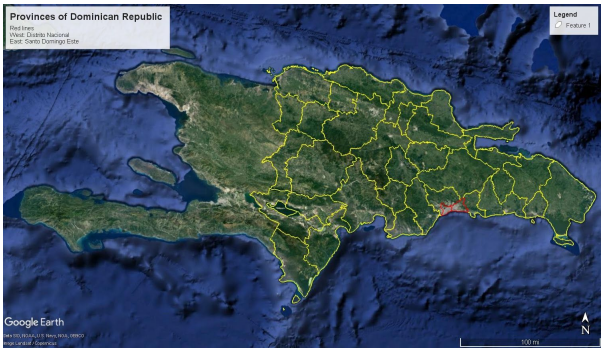
Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.

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).

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.



Geographic Area of Interest

The following image shows the nation's Capital, on the left, namely National District ("Distrito National" in spanish) and on the right Santo Domingo Este.

The left side is our National District municipality. With approximately 91 square kilometers it has the majority of shopping malls, banks and restaurants in the Dominican Republic. But 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.

Cluster Analysis

We will perform a cluster analysis. Our task will be grouping sets of neighbourhoods, which will be plotted on the above map. Those of the same group (called a cluster) being more similar to each other than to those in other groups will give us an idea of where there are opportunities to develop a shopping mall and where there are too many already. This is a main task of exploratory data mining, and a common technique for statistical data analysis.

To properly perform the cluster analysis we should encode neighbourhoods with geolocation information.

jupyter finalassignment DN Y GRAN SANTO DOMINGO Last Checkpoint: L

File Edit View Insert Cell Kernel Widgets Help

Run Markdown

In [19]:
```
df_SD = df_SD.rename(columns={'Location': 'Neighborhood'})
```

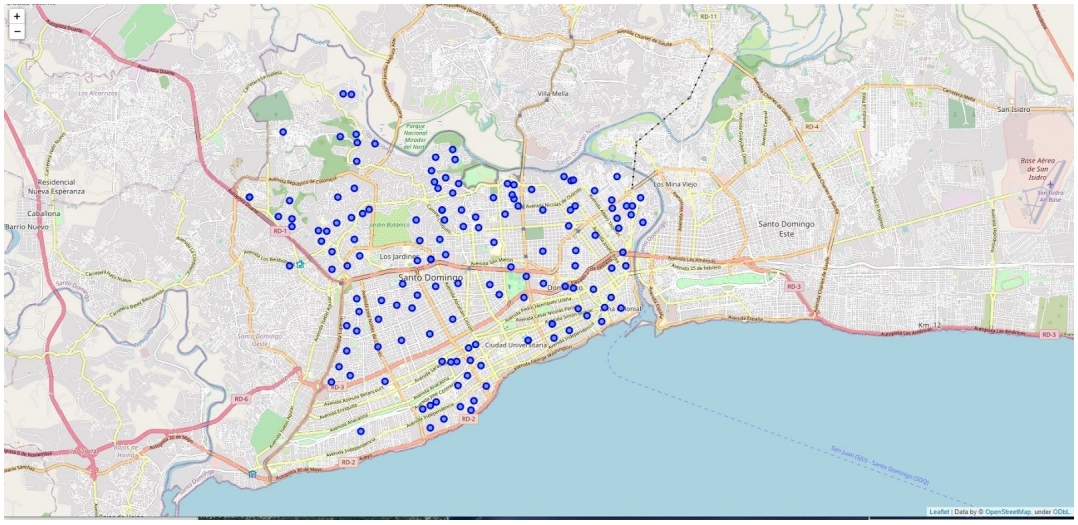
In [20]:
```
# check the neighborhoods and the coordinates
print(df_SD.shape)
df_SD.head()
```

```
(397, 3)
```

Out[20]:

	Neighborhood	Latitude	Longitude
0	24 DE ABRIL	18.474302	-69.849325
1	30 DE MAYO	18.517191	-69.882794
2	AEROPUERTO INTERNACIONAL	18.575710	-69.981480
3	AGUACATE ADENTRO	18.599690	-69.833560
4	AHORCA LOS PERROS	18.638070	-69.606580

In [21]:
```
# get the coordinates of Distrito Nacional
address = 'Distrito Nacional, DOM'
```



After finding neighbourhoods, we connect to Foursquare API to gather information about venues inside each and every neighborhood. The data retrieved within a specified distance (Venue, Venue Latitude, Venue Longitude, Venue Category) will help us derive clusters.