



# **The battle of the Neighborhoods**

Visit or invest in Punta Cana  
Dominican Republic

**Capston Project**  
IBM Data Science Professional Certificate

Erickson Figueroa  
IT Data Analyst  
December 2020

# 1. Introduction

## 1.1 Background

The Dominican Republic is a democratic nation with a population of over 10 million people. Its territory is stretching 48,442 square kilometers (18,704 square miles). That's just over twice the size of New Jersey, or nearly the size of Costa Rica. Its coveted shores face the Caribbean Sea in the south and the Atlantic Ocean in the north. As of 2017, 6,187,542 overnight visitors and more than 1,100,000 cruise passengers make their way to the Dominican Republic every year.

**Punta Cana** is a resort town within the Veron municipal district, in the municipality of Higüey, in La Altagracia Province, the easternmost province of the Dominican Republic. The area has beaches and balnearios which face both the Caribbean Sea and the Atlantic Ocean. The town has many hotels and resorts, restaurants, shopping centers, cabins, nightclubs, water parks, fast food points, cultural shopping stores etc.

Based on national statistics is the most tourist destination in the country.

**It is more than 50% of those arriving each year.**

## 1.2 Problem

Knowing the above information about Punta Cana, we need to answer the next questions:

- Which are the most visited venues?
- Which neighborhood has the most venues?

Answering these questions tourists and investors will know what place is better to visit or invest and start a new business based on the most popular venues.

We will do a statistical analysis to understand the data and found analytical insights, the Machine Learning's K-MEAN algorithm to cluster the venues by neighborhoods and visualize them on a map.

## 1.3 Interest

- Whose want to visit or invest in Veron Punta Cana.
- The Dominican government to take important decisions about tourism.
- Maybe to help local businesses in the area to do marketing plans and know their competitors.

# 2. Data acquisition and cleaning

## 2.1 Data sources

The principal data source was from the National Bureau of Statistics. It's about census data that contains information like the province, municipality, municipal district, section, neighborhood, latitude, longitude, population, etc. Its format is in an excel XLS file. The last update of this data source was in 2010 because the census occurs every ten years.

This dataset goes to be used to get the venues from the Foursquare API.

link here:

<https://www.one.gob.do/censos>

The second data source it's about Tourist flow in the year 2019 by Nationalities. This dataset goes to be used to show some plots like the top arriving countries, total tourist by month, by location e.g., North America, South America, Asia, Europe, etc. It would help the investors to know which are the nationalities that come every month and give them ideas for future marketing plans or maybe help the visitors to know which are the cultures around those tourist destinations.

Link here:

<https://www.bancentral.gov.do/a/d/2537-sector-turismo>

## Dataset description

*population\_dr\_2010.xls*

Column	Description
RID	Id of the rows
Reg	Region code of the province
Prov	Province code
Mun	Municipality code
Md	Municipality district code
Zon	Zone code
Secc	Section code
Neig	Neighborhood code
Neighborhood	Neighborhood name
Longitude	Longitude
Latitude	Latitude
Province	Province name
Municipality	Municipality name
Municipal District	Municipal district name
Section	Section name
Zone	Zone name
Zone_Total	Total of the population int the zone
Men	Population of the men in the neighborhood
Women	Population to the women in the neighborhood
Total_M_W	Total population both men and women in the neighborhood

*tourist\_flow\_2019.xls*

Column	Description
LOCATION	North America, South America etc.
<b>COUNTRY</b>	Country name
MONTHS	January, February, March etc.
Mun	Municipality code

### 2.1 Data cleaning

The first steps of this process were manual to delete some columns, translate all the columns from Spanish to English I did it all using Excel without programing any code line. The next processes like filter data to take only the neighborhoods from the district of Veron Punta Cana, ordering, cleaning, merge with the Foursquare API, etc. I will use python's library Pandas.