The battle of the Neighborhoods

Visit or invest in Punta Cana Dominican Republic

Capstone 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 venus?
- 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.
COUNTRY	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.

3. Methodology

- Collect the data from the National Bureau of Statistics.
- Organizing the dataset's columns and select the features that only be needed.
- Reading the .xls file and using Pandas library to filter the data only from Punta Cana District.
- Using the Geopy library to get the latitude and longitude values of the Punta Cana district.
- Visualize each point into a map.
- Using the Foursquare API to get the venues based on the latitude and longitude of my dataset.
- Find out how many unique categories can be curated from all the returned venues.
- Analyze Each Neighborhood and add the one-hot encoding technique for venue category.
- Calculating the mean of the frequency of occurrence for each category to know what's the top 10 venues for each neighborhood.
- Using a K-Mean cluster algorithm to do the clusters and add the label cluster in the dataset.
- Visualize the resulting clusters in a new map.
- Examine each Cluster and determine the discriminating venue categories that distinguish each cluster.
- The last step was some basic exploratory data analysis and plots like Top 5 venue category by frequency of occurrence, Top 5 neighborhoods by venues total, Top 5 countries with the most tourism flow, etc.

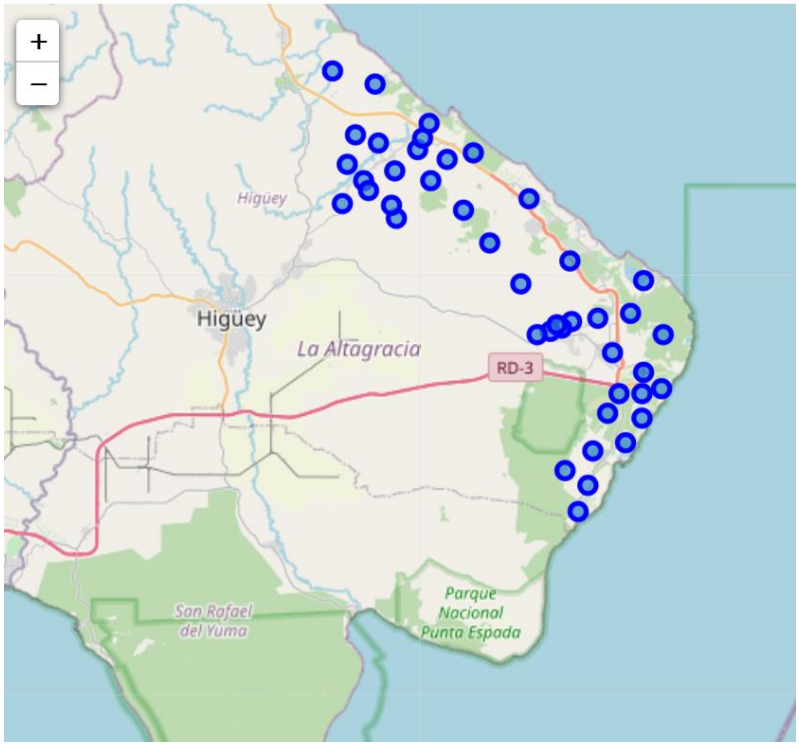
An example of datasets:

Population dataset

	RID	Reg	Prov	Mun	Md	Zon	Secc	Neig	Neighborhood	Longitude	Latitude	Province	Municipality	Municipal District
0	0	7	7	1	1	2	4	1	LA JAGUA	-71.721277	18.932071	ELIAS PIÑA	COMENDADOR	COMENDADOR
1	1	7	7	1	1	2	8	3	GUAZUMAL	-71.664995	18.884885	ELIAS PIÑA	COMENDADOR	COMENDADOR
2	2	7	7	1	3	2	4	6	LA TINAJA O TINAJITA	-71.705507	18.846651	ELIAS PIÑA	COMENDADOR	GUAYABO (D. M.).
3	3	7	7	1	1	2	8	5	LA CUNA	-71.695479	18.902046	ELIAS PIÑA	COMENDADOR	COMENDADOR
4	4	7	7	3	1	2	2	8	PALMA CANA	-71.633865	18.855488	ELIAS PIÑA	EL LLANO	EL LLANO

Tourism flow

	LOCATION	COUNTRY	Unnamed: 2	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
0	North America	Canada	NaN	178634	210177	261513	203185	210175	254169	203194	139243	61633	78672	103236
1	North America	United States	NaN	126503	124662	131579	95233	46211	36792	45493	40416	20630	27438	68442
2	North America	Mexico	NaN	25523	29392	23687	23744	13742	11558	17632	20199	7727	13121	13960
3	Central America And The Caribbean	Aruba	NaN	25062	19931	20233	18013	10502	10241	12449	13993	15256	19362	24774
4	Central America And The Caribbean	Caicos and Turks, Islands	NaN	19280	16209	15970	16153	19747	16127	16180	14160	12100	13614	16107



The result of the visualization for each Latitude and longitude Punta Cana district.

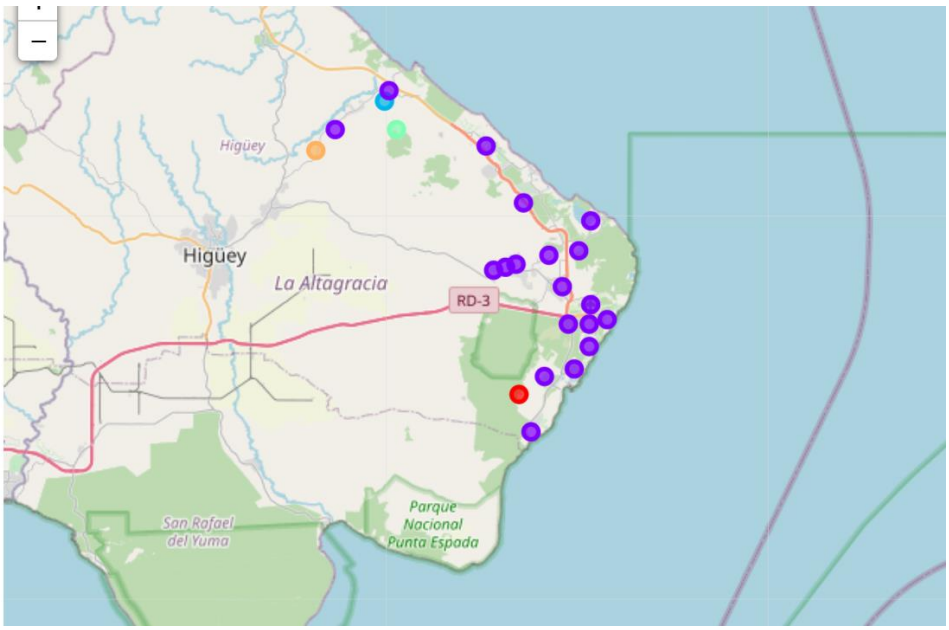
Using the Foursquare API, we returned the venues for each neighborhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue
BELLO AMANECER	2	2	2
BORRACHON	9	9	9
BÁVARO	8	8	8
CABEZA DE TORO	17	17	17
CABO SAN RAFAEL	7	7	7
CAOBA	4	4	4
CARACOLITO	16	16	16
CAÑADA HONDA	1	1	1
EL SALADO	1	1	1
EL SAMARITANO	4	4	4
EL SUERO	1	1	1
GUIRI GUIRI	3	3	3
LA CURVA O LA CULATA	14	14	14
LA JARDA DE HOYO CLARO	5	5	5
LA JINA	1	1	1

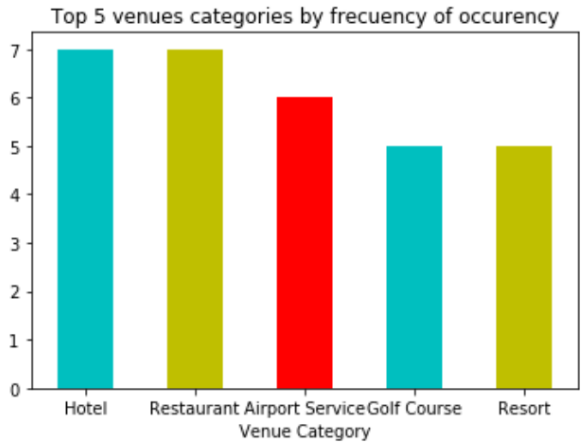
Getting the mean of the frequency of occurrence for each category to know what's the top 10 venues for each neighborhood.

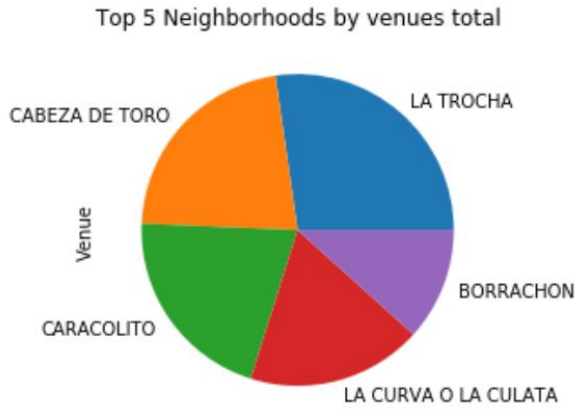
Neighborhood	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Asian Restaurant	Auto Garage	BBQ Joint	Bagel Shop	Bakery
BELLO AMANECER	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000	0.0	0.000
BORRACHON	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000	0.0	0.000
BÁVARO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.125	0.125	0.0	0.125
CABEZA DE TORO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000	0.0	0.000
CABO SAN RAFAEL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000	0.0	0.000

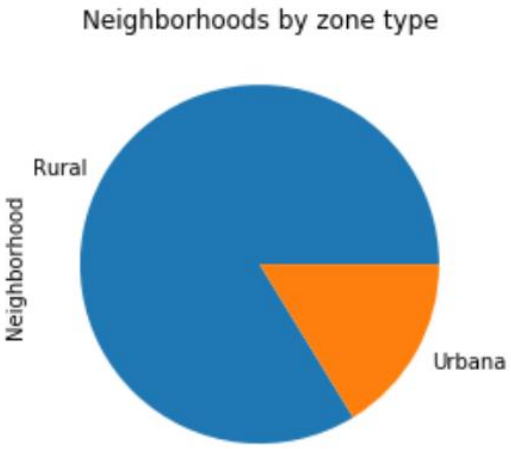
When I did the cluster in the code section, then put it into a new map for each group or cluster label

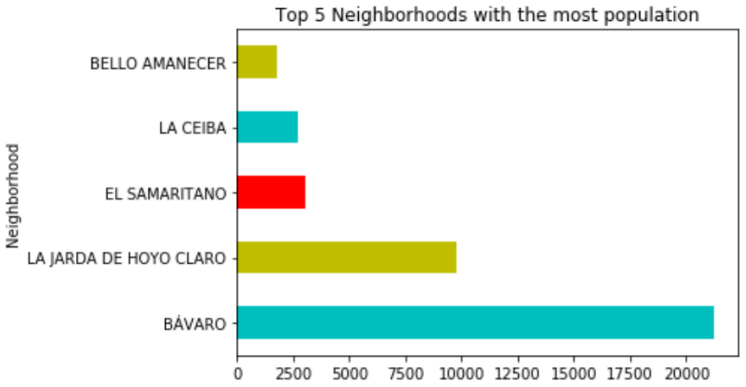


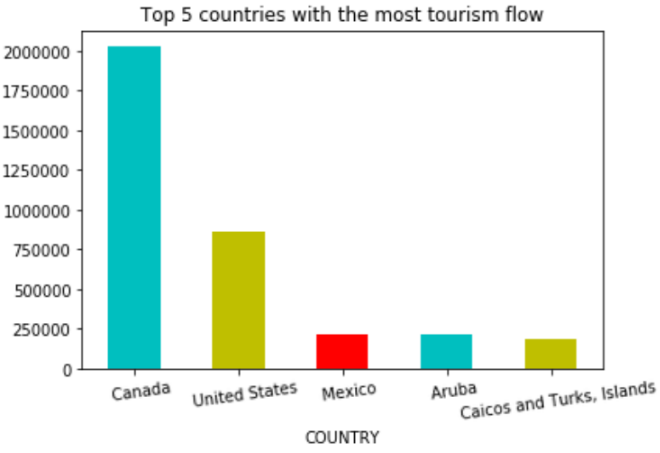
Basic exploratory data analysis and plots

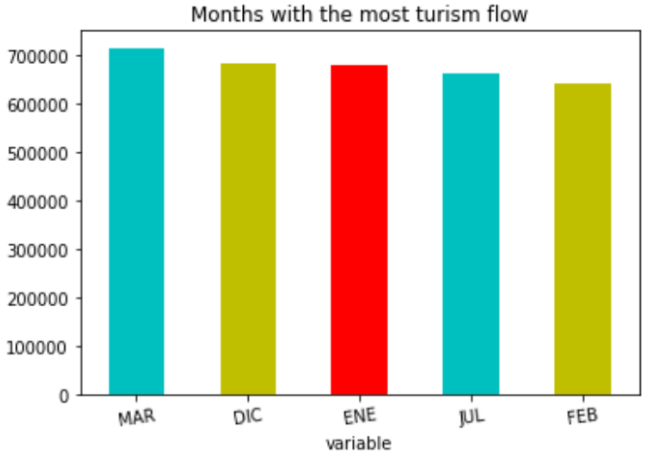






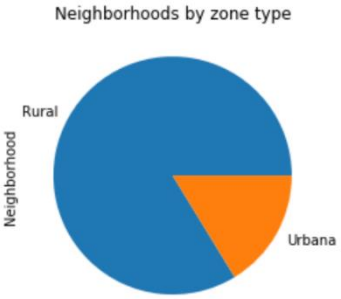






4. Results

After executing the K-Means clustering algorithm 5 clusters were created the most of them are in cluster number one and the rural zone.



Example of the most common places

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
BELLO AMANECER	Pharmacy	Hotel	Sports Bar	Department Store	Caribbean Restaurant	Casino	Cave	Clothing Store	Coffee Shop
BORRACHON	Nature Preserve	Japanese Restaurant	Casino	Fishing Spot	Hotel	Hotel Bar	Resort	Restaurant	Comedy Club
BÁVARO	Grocery Store	Auto Garage	Restaurant	Baseball Field	Rock Club	BBQ Joint	Bakery	Supermarket	Fast Food Restaurant
CABEZA DE TORO	Bar	Resort	Beach	Japanese Restaurant	Café	Lounge	Buffet	Restaurant	Brazilian Restaurant
CABO SAN RAFAEL	Golf Course	Resort	Hotel	Steakhouse	Mediterranean Restaurant	Beach	Caribbean Restaurant	Cosmetics Shop	Cave

5. Discussion

Most of the places returned by cluster #1 were in the rural zone because there are many resorts, hotels, beaches, tourist complexes, etc.
If you are a tourist, you will be many interested in this area, but if you are an investor, you will need to evaluate the next:

- The population of the area
- Property prices
- The business type that operates there and their number by categories
- Don't forget to check the urban area. This part depends on the business that you want to start.

6. Conclusion

Based on the above result, most visit places were in the rural zone of Punta Cana. If you are a tourist, you will be interested in these places, but if you are an investor, you will need to evaluate some criteria before starting a new business in those areas.

References and sources

https://www.one.gob.do/
https://www.bancentral.gov.do/a/d/2537-sector-turismo
https://www.godominicanrepublic.com/punta-cana/
https://en.wikipedia.org/wiki/Punta_Cana