Exploring Puerto Rico

Steven Leonhart

A. Introduction

A.1 Background

Puerto Rico is a beautiful island that is located in the Caribbean region of the Americas. It is a great place for a tropical getaway. For the final part of the ***IBM Data Science Professional Certificate*** program, the Capstone Project, I was tasked with finding a problem that I can solve using real world data from Foursquare API and other places on the internet. While the assignment description involves exploring neighborhoods in a city or cities, I decided to do the island of Puerto Rico, taking Puerto Rico as my "city", and its municipalities as my "neighborhoods". I chose to do Puerto Rico because after traveling to the island for my brother's wedding, I enjoyed seeing the beaches, mountains, historic sites, and many other features the island had to offer. I also enjoyed getting to experience a culture that was very different from the one I'm a part of in my hometown of Chicago.

A.2 Problem

With this project, I hoped to learn more about what the island has to offer for both tourists and investors. A tourist who is looking to visit the island would ask

What parts of the island do I want to visit based on what I am interested in seeing the most?

An investor looking to invest in lodging and places to stay will have a very different perspective and might ask

1. **Where on the island would be a good place to open a new hotel?**
2. **Based on where I live on the island, would starting a Bed and Breakfast or an AirBnB on my property be a good investment for me?**

I was hoping to conduct data analysis that will allow me to answer each of the above questions.

A.3 Data

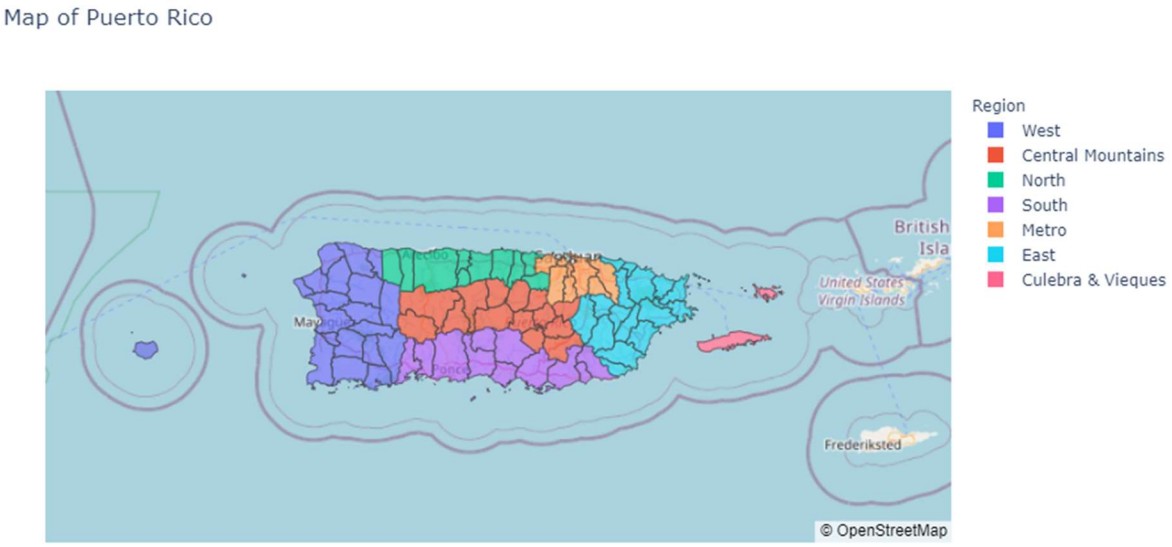
I used the following data to conduct my analysis for this project:

- A **GeoJSON** file titled **"First-level Administrative Divisions, Puerto Rico, 2015"** from the Stanford EarthWorks Library. [1] The .json file contains the boundaries of each municipality of Puerto Rico. I used it to create choropleth maps that were useful for visualizing the data in my analysis.
- **Google Maps** [2]. As the coordinates of the municipalities are not available in a nice table, I had to get them myself. Thus, I used Google Maps to look at Puerto Rico geographically in order to pick a coordinate point for each municipality that would likely give the best results when it comes to retrieving venues from Foursquare's API.
- **Foursquare API** [3]. I used the Foursquare API to get venues for each of Puerto Rico's municipalities.

B. Methodology

For this project, I used many of Python's built-in libraries in order to conduct my analysis. These include NumPy, Pandas, GeoPy, Requests, Scikit-learn, Plotly, and GeoPandas. First, I had to fetch the latitude and longitude coordinates for each municipality manually using Google Maps and then write them all into an Excel file that I saved as a CSV. The CSV file then had to be read into a Pandas DataFrame. I also had to import the GeoJSON file into GeoPandas GeoDataFrame, clean it up to slice off unneeded data to save memory and make it more useable, and then finally join it with the Pandas DataFrame in order to produce a master GeoDataFrame that I would use to retrieve the other data and produce my visualizations.

For the first part of this, I wanted to create a choropleth map that displayed each of Puerto Rico's 78 municipalities and 7 regions [4][5]. I used GeoPy to get the geographical coordinates of Puerto Rico.



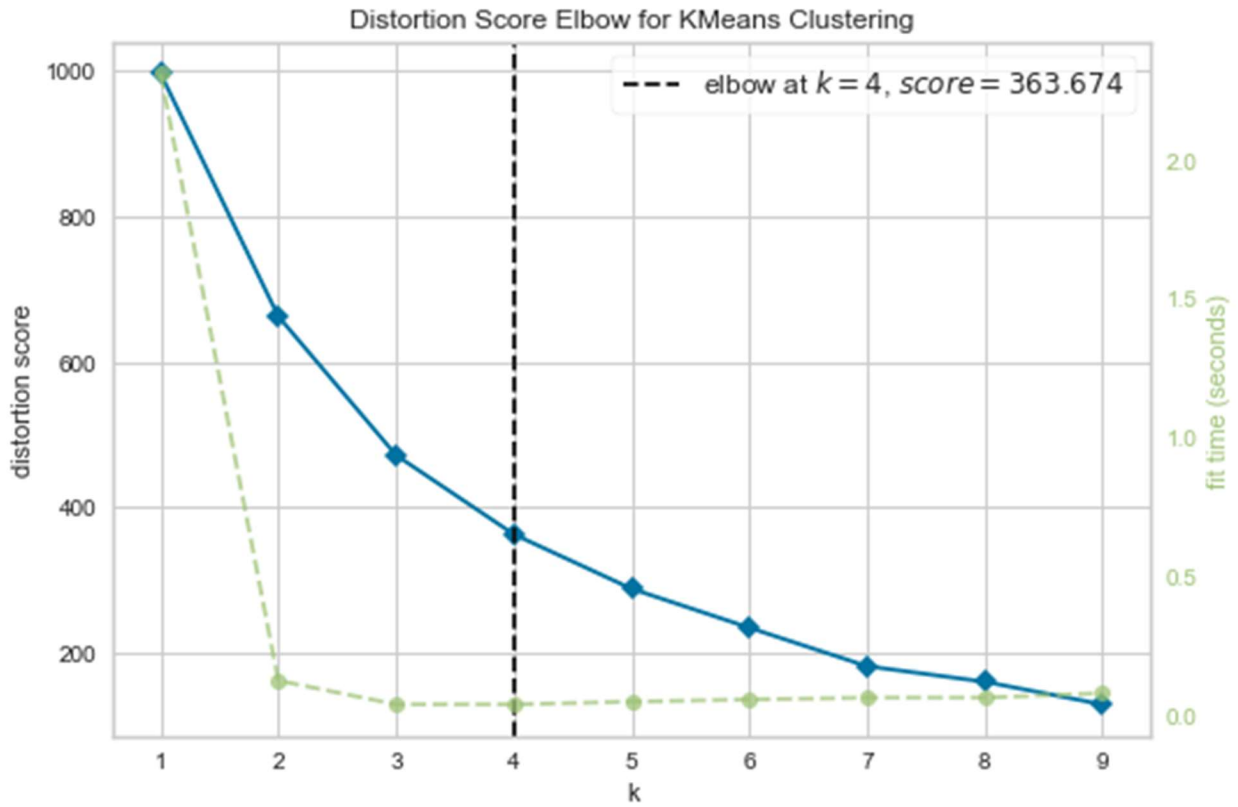
I used Plotly Express to produce not only the above map, but all my other visualizations as well. I chose to use Plotly Express because it allows you to create very nice looking charts, graphs, and maps with only a few lines of code.

The next part of the project was retrieving the venue data for each municipality using Foursqaure API. Because I was working with large areas, I had set the radius search for the call to 9000 meters, or about 5.6 miles. The venue limit for the call was set at the default value of 100 venues. Using the latitude and longitude coordinates for each municipality that were stored in my master GeoDataFrame (I only used one pair of coordinates for each one), I retrieved the top 100 venues for each municipality from the Foursquare API. In order to be able to answer the above questions, I had to pull out venues that were important for my project from the huge pool of venues that Foursquare sent back. The important venue categories I was looking for were

- Tourist venues - These are venues that are likely to attract tourists from other parts of the world to the area. They include beaches, historical landmarks, great outdoor venues (trails, nature preserves, waterfalls, lakes, caves, forests, and other natural features), monuments, museums, and coffee shops.
- Places to stay - Hotels, AirBnBs, Motels, Resorts
- Restaurants

By looking through the list of all the venue categories, I was able to create a filtering and grouping algorithm in my notebook that pulled out each venue that belonged to one of the above groups and placed it into the correct group. I then created a DataFrame that contained the number of venues each municipality had in each of the above groups. All other venues that did not belong to any of the above groups were tossed out.

I was now ready to start clustering the municipalities in a way that will allow me to answer each of the above questions. For the tourist question, I wanted to group the municipalities in a way such that the ones that had tourists venues belonging to the same group would be clustered together. The algorithm that I used for this was the **K-Means Clustering Algorithm**. This is one of the most popular unsupervised machine learning algorithms for grouping data. In order to use the K-Means algorithm, I first had to find the optimal value of k for the clustering. To do this, I used Python's built in **KElbowVisualizer** that was part of the **YellowBrick** package. This program uses the "Elbow Method" to find the optimal k value. The result is shown below.



As we can see, 4 is the optimal value for k. Thus, I used the K-Means algorithm to group the municipalities into 4 clusters for further analysis. After combing through each cluster group and looking at what tourist venues each municipality the group had in common, I was able to come up initially with the following names for each cluster:

- **Cluster 0: "Beaches"**
- **Cluster 1: "Not Much of Anything"**
- **Cluster 2: "Beaches, Great Outdoor Venues"**
- **Cluster 3: "Beaches, Historical Landmarks, Museums, Coffee Shops"**

However, because Cluster 1 contained the vast majority of municipalities, I decided to comb through it again to see if maybe I can split it up a little bit. I found out that a few of the municipalities in Cluster 1 had 3 great outdoor venues, which was the maximum value for that cluster for that tourist venue type. Since none of the municipalities in Cluster 1 had any more than 2 of the other tourist venue types (most didn't even have one venue for any of the other venue types), I decided to separate the municipalities with 3 great outdoor venues from the rest of Cluster 1 and place them into a **fifth cluster, Cluster 4**. After doing this, I now had the following names for each cluster:

- **Cluster 0: "Beaches"**
- **Cluster 1: "Not Much of Anything"**
- **Cluster 2: "Beaches, Great Outdoor Venues"**
- **Cluster 3: "Beaches, Historical Landmarks, Museums, Coffee Shops"**

- **Cluster 4: "Great Outdoor Venues"**

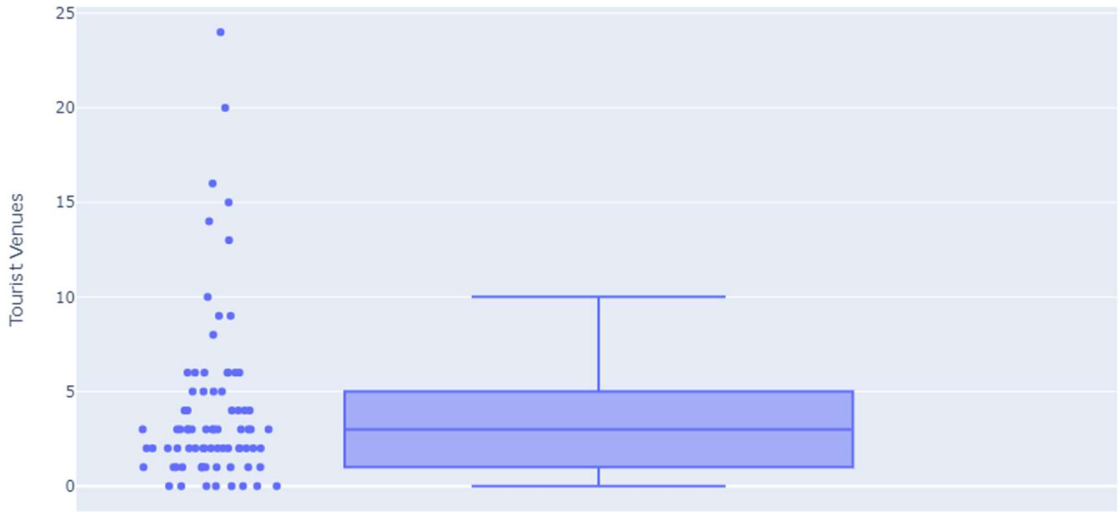
For the Hotel/AirBnB question, I wanted to group the municipalities based on 3 things:

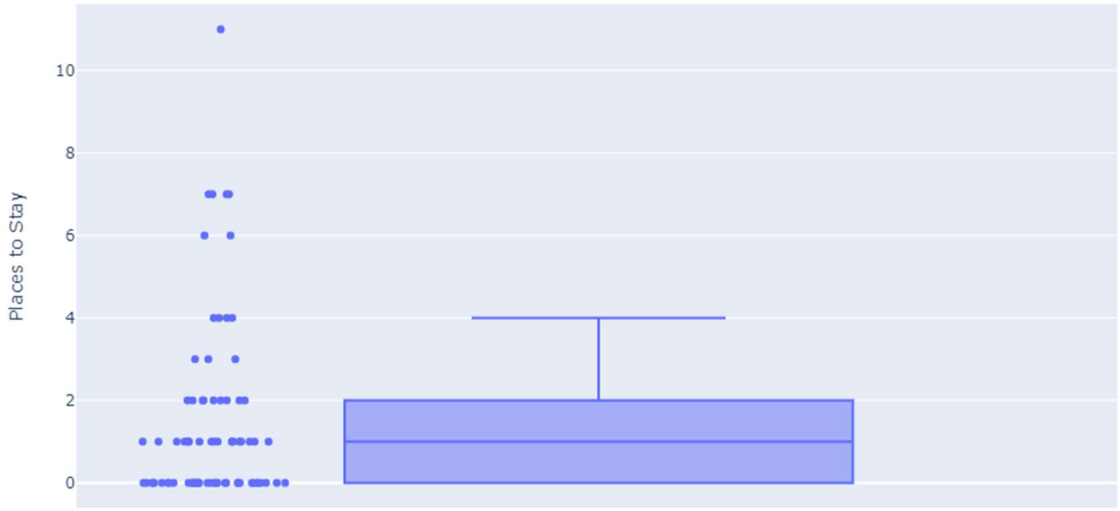
1. The total number of tourist venues
2. The number of places to stay
3. The number of restaurants

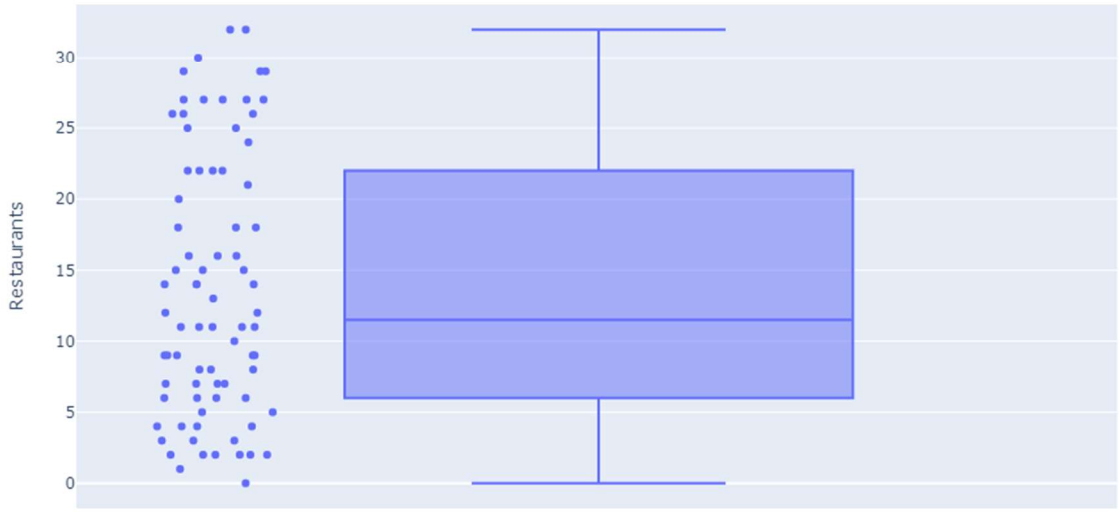
When it comes to deciding on where to open a new hotel or start an AirBnB business, the best places to do so would have the following characteristics:

1. A high number of tourist venues
2. A low number of hotels, AirBnBs, or other places to stay that are already in the area
3. A high number of restaurants

In order to find which places had the above characteristics, I grouped the municipalities using the "binning process". This process divides the data into "bins" based on the numerical values of the data. I used this process to group the municipalities 3 times, first on total number of tourist venues, then on the number of places to stay, and then finally on the number of restaurants. In order to decide on the intervals to use for each bin and to see how the data was distributed for each of the categories, I created boxplots for the number of tourist venues, places to stay, and restaurants.







For the tourist venue data, the boxplot showed that most of the data lied between 0 and 10, with a few outliers above 10, which is the upper fence value. The five-number summary is as follows:

- **Min: 0**
- **Q1: 1**
- **Median: 3**
- **Q3: 5**

- **Max: 10**

Because I wanted the "higher" cluster numbers to reflect the better choices based on the number of tourist venues, I decided to label the clusters as follows:

- **Cluster 0: Less than 3 tourist venues**
- **Cluster 1: 3 or 4 tourist venues**
- **Cluster 2: 5 - 10 tourist venues**
- **Cluster 3: More than 10 tourist venues**

For the places to stay data, the boxplot showed that the overwhelming majority of the data was between 0 and 2 places to stay, with a few lying above 2. The five-number summary is as follows:

- **Min: 0**
- **Q1: 0**
- **Median: 1**
- **Q3: 2**
- **Max: 4**

As with the tourist venue cluster labels, I wanted the higher cluster number labels to reflect the better choice municipalities, which in this case are the ones with fewer places to stay. Thus, I labeled the clusters as follows:

- **Cluster 0: 2 or more places to stay**
- **Cluster 1: Less than 2 places to stay**

The restaurant data boxplot clearly shows a near even distribution of the data, with no outliers. Here is the five-number summary.

- **Min: 0**
- **Q1: 6**
- **Median: 11.5**
- **Q3: 22**
- **Max: 32**

In continuing to follow the same formulation that I used for the first 2 categories, I labeled the clusters as follows:

- **Cluster 0: Less than 6 restaurants**
- **Cluster 1: 6 - 22 restaurants**
- **Cluster 2: More than 22 restaurants**

After completing all 3 groupings, I calculated a numerical value called the "Investability Score", which indicates how good of a choice that location would be for a new hotel or AirBnB based on the data. The formula for calculating the value is right below.

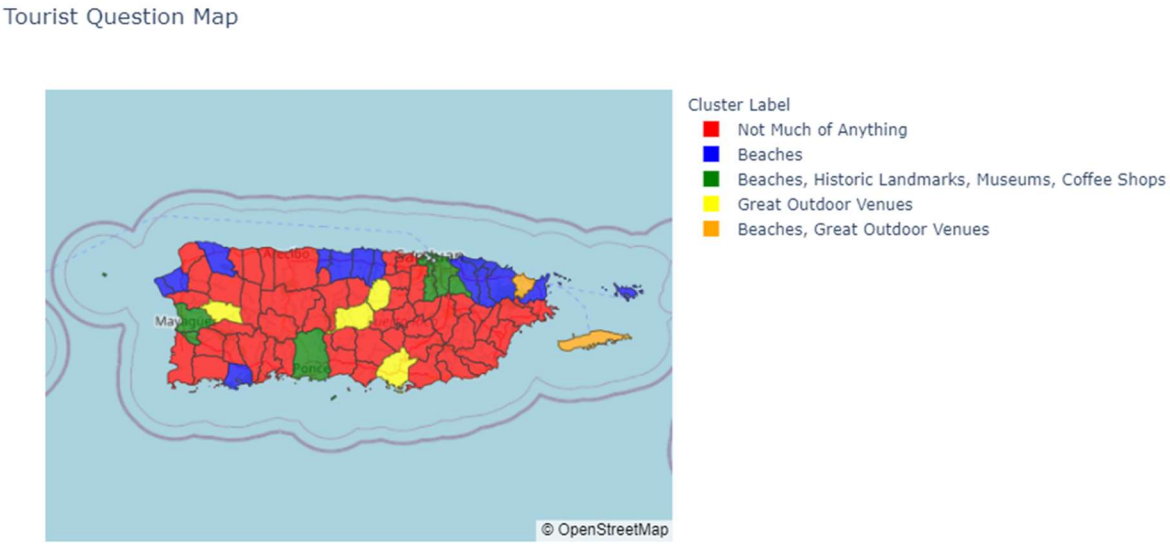
$$Investability\ Score = Tourist\ Venues\ Cluster\ Label + Places\ to\ Stay\ Cluster\ Label + Restaurants\ Cluster\ Label$$

Finally, I binned the municipalities based on Investability Score under the following labels and labeled this as the "Investment Rating" for each municipality.

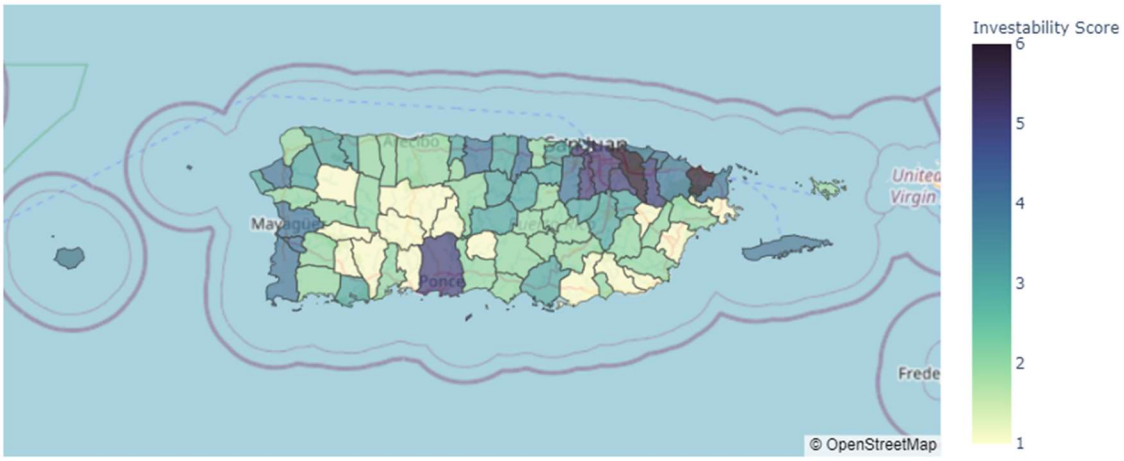
- **Score 5-6: Excellent**
- **Score 4: Good**
- **Score 3: Fair**
- **Score 0-2: Poor**

C. Results

I created choropleth maps for each of the tourist and hotel/AirBnB questions using Plotly Express to display the results of my analysis.



AirBnB Question Map



D. Observations

D.1 Tourist Question

I will make the following recommendations based on the Tourist Question Map and what you are hoping to see to answer this question. The region is listed after the municipality.

1. Beaches
    a. San Juan - Metro
    b. Manatí - North
    c. Carolina - Metro
    d. Culebra - Culebra & Vieques
    e. Vieques - Culebra & Vieques
    f. Isabela - West
    g. Luquillo - East
    h. Rincón - West

2. Historic Landmarks
    a. San Juan - Metro
    b. Cataño - Metro

3. Great Outdoor Venues
    a. Luquillo - East
    b. Vieques - Culebra & Vieques
    c. Orocovis - Central Mountains
    d. Corozal - Central Mountains
    e. Salinas - South
    f. Las Marías - West

4. Museums
    a. Ponce - South

5. Coffee Shops
    a. San Juan - Metro
    b. Cataño - Metro
    c. Mayagüez - West
    d. Ponce - South
    e. Trujillo Alto - Metro
    f. Guaynabo - Metro
    g. Hormigueros - West

D.2 Hotel/AirBnB Question

I can make the following statements based on the Hotel/AirBnB Map above to answer this question.

1. The best municipalities in which to open a new hotel or an AirBnB are the ones that recieved an Investment Rating of "Excellent" (Investability Score of 5 or 6). These include:
    a. San Juan - Metro
    b. Carolina - Metro
    c. Cataño - Metro
    d. Guaynabo - Metro
    e. Trujillo Alto - Metro
    f. Canóvanas - East
    g. Luquillo - East
    h. Ponce - South

2. Any municipality that had an Investability Score of 4 ("Good" Investment Rating) or higher would likely be a great place to open a new hotel or AirBnB. You will have a good chance to get a decent return on your investment with these.
3. Municipalities that scored a 3 (Fair) may be decent options as well if you are willing to take on more risk and can afford to possibly take a loss on your investment without falling into financial trouble. I would consider these as "Hit or Miss" ones when it comes to whether or not you will get a decent return on your investment.
4. Municipalities that scored 2 or lower (Poor) should be avoided. The risk for these is likely way too high for an investment into a new hotel or AirBnB. This is due to one or both of the following statements being true.
    a. There are not enough tourist venues and restaurants in the area that would make tourists want to stay in the area long enough for a hotel or AirBnB to generate enough profit to offset the cost.

b. There are too many other hotels, AirBnBs, and other places to stay that are already in the area, which would mean a lot more competition for the new hotel or AirBnB.

D.3 Limitations of This Project

Unfortunately, there are some limitations to this project when it comes to using it as a tool for the real-world scenario.

1. Land Availability - The availability of land for building a new hotel was not included in this analysis. Although some places scored well for investability, it would not matter at all if no land was available in those areas for building a hotel. Further research into land availability could help with this.
2. The large search radius may have caused some venues to be placed under the wrong municipalities. I saw this and corrected it for some of San Juan's venues that were placed under Cataño, but there may have been many more of these for other places. This is much more of a limitation for the tourist question, however, than it is for the hotel/AirBnB question. The reason for this is that even if the municipality itself has no venues, it may still be a good idea to open a hotel or AirBnB in that area if there are some venues nearby in neighboring municipalities.
3. Some venues may have been **miscategorized** by Foursquare, which would have affected whether or not they were important venues.
4. Only Foursquare API data was used. This is a limitation because of the following:
    a. Not all important venues may be registered with Foursquare, meaning many venues may have been left out of this study that could have affected the results.
    b. The venues returned by Foursquare are affected by the time of the call, meaning if I made the call to Foursquare API at a different time, I may have gotten different results.

E. Conclusion

Puerto Rico is a very fascinating place to go to and has a lot of history and culture. I would highly recommend making a trip there if you are looking to spend some time away in paradise and are looking to experience a different culture. It would be a great idea to learn some Spanish before you visit though, as it is the most common language spoken on the island.

F. Works Cited

[4] *Discover Puerto Rico*. 2021. June 2021.

[3] Foursquare. *Foursquare Developers*. 2021. June 2021.

[2] Google. *Google Maps*. 2021. June 2021.

[1] Hijmans, Robert J and Berkeley. Museum of Vertebrate Zoology University of California. "First-level Administrative Divisions, Puerto Rico, 2015." 2015. *Stanford Earthworks Library.* GeoJSON File. June 2021.

[5] Rivera, Magaly. *Welcome to Puerto Rico*. 2021. June 2021.