Coursera IBM Data Science Certificate

Capstone Project

The Best Neighborhood for a New Resident of Miami

Courtney T Green

June 2020

Introduction

For many people throughout the United States and beyond, Miami is a dream vacation spot filled with exciting nightlife, beautiful beaches, intriguing art galleries and posh hotels. However, Miami is more than just a hot holiday getaway. Being the cultural, economic and financial center of southern Florida makes Miami an attractive place for young professionals to relocate. Not only is the city of Miami ranked third-richest in the United States, it is seventh in terms of “business activity, human capital, information exchange, cultural experience, and political engagement”. But Miami is also the sixth most densely populated major city in the States, having a population of over 460,000 people living in the city and 6.1 million in the metropolitan area. For this reason, a new resident has a daunting task of choosing where exactly in the city they should live.

Business Problem

The objective of this project is to analyze and cluster the twenty-five neighborhoods of Miami based on venues and location to give a new or prospective resident a clear idea of which neighborhoods they may be interested in living. This project seeks to answer the question: If I am moving to Miami and prefer living near x, y, and z, which neighborhoods should I consider?

Data

To give adequate solutions to the problem, I will need the following data:

- List of major neighborhoods in Miami, FL, as currently defined by the city of Miami
- Latitude and Longitude coordinates of these neighborhoods
- Venue data for each neighborhood

Sources of data and methods of extraction:

The Wikipedia page “List of neighborhoods in Miami” (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Miami) contains a table listing the major neighborhoods in Miami, totaling twenty-five. It also gives the geographical coordinates of these neighborhoods. I will use web scraping to extract this table from Wikipedia into a Jupyter notebook and convert it to a Pandas dataframe using Python. The table also provides 2010 population and sub-neighborhood information, but that data is not necessary for the scope of this project.

I will gather the venues data for the neighborhoods using Foursquare API. With a considerable database of over 105 million places, Foursquare will be able to give me a large number of venues present in each neighborhood, which I will then use to determine a general characteristic of these neighborhoods.

Methodology

Downloading and Exploring the Dataset

Fortunately, there is a Wikipedia page containing a table listing all twenty-five major neighborhoods in Miami, along with each neighborhood's respective demonym, population according to the 2010 census, population density, sub-neighborhoods and geographical coordinates[1]. Using Python requests and BeautifulSoup, the table was scraped from the webpage, converted from an HTML table into a Pandas dataframe, and cleaned.

To clean the dataframe, first the unnecessary columns were dropped, leaving only the neighborhood names and coordinates. Next, the 'Coordinates' column was split into two separate columns, appropriately named 'Longitude' and 'Latitude'. The 'Coordinates' column was then removed. While checking the resultant dataframe, I noticed that the neighborhood 'Health District' does not have coordinates. Upon further research, I learned that it is actually more of a subset of Downtown, where there is a hub of hospitals, health centers and research facilities. Because this does not seem relevant to the problem of where to live (the Health District is more a hub of workplaces than of residences; plus, if you choose Downtown, you will be near the Health District anyway), I chose to delete this row from the dataframe. By deleting 'NaN' rows, this will remove the Health District and the totals row at the bottom of the dataframe.

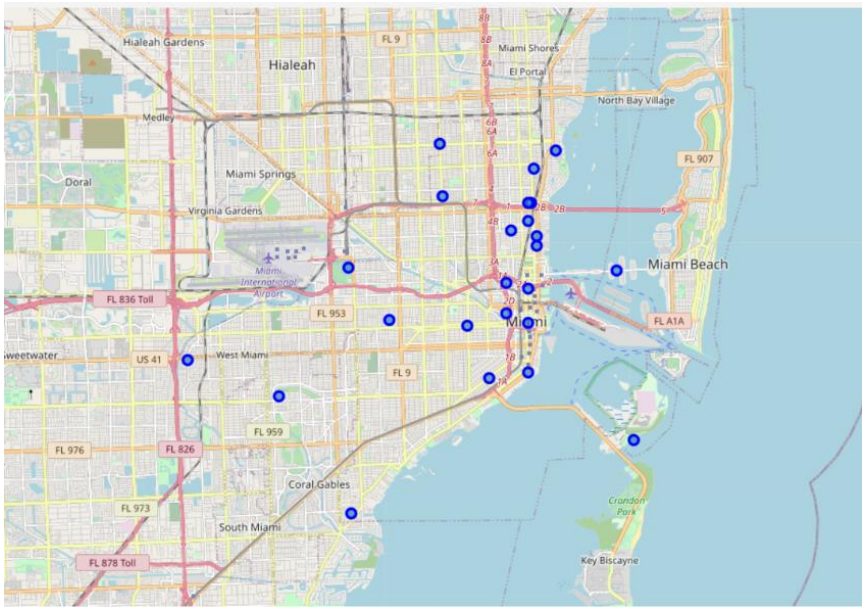
The negative sign on the longitudes was used as the delimiter to split the 'Coordinates' column, therefore it is now missing. The negative sign was added to each longitude value, then both the 'Latitude' and 'Longitude' columns were

[1] https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Miami

changed to type float (instead of the original string), so that they could be used later to create the map.

Exploring and Analyzing the Neighborhoods

After the dataset is complete, a map of Miami with each of its neighborhoods was created to get a preliminary idea of how it looks.



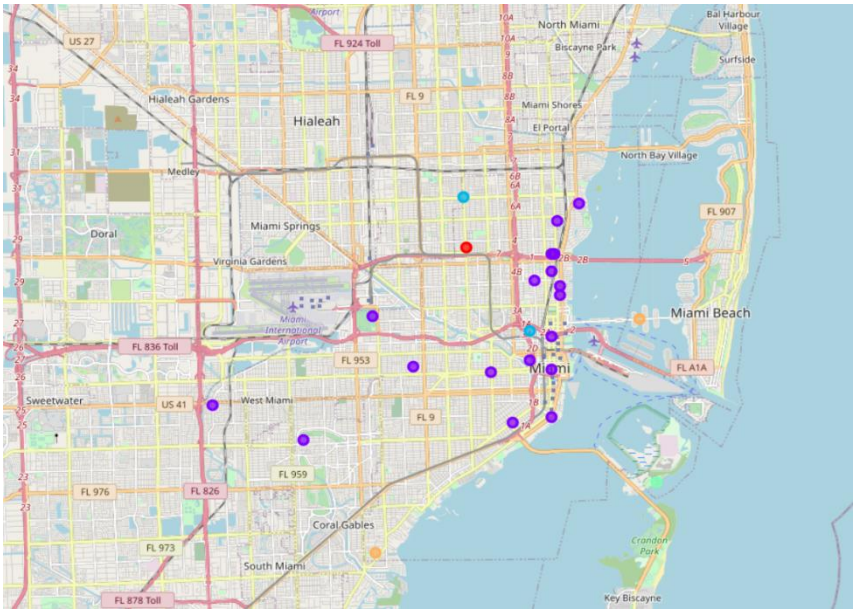
Next, Foursquare was used to explore the neighborhoods further. A function was created to pull the venues data from each neighborhood using each neighborhood's latitude and longitude. It returned the name of the venue, its coordinates, and its category. The number of venues pulled for each neighborhood was limited to 50, with a radius equal to 500. Each neighborhood was run through the function and the results were saved to a new dataframe. Foursquare found 459 total venues for the 24 neighborhoods. While checking how many venues were returned for each neighborhood, I saw that the results were less than ideal. In fact, only half of the neighborhoods had more than ten venues and a quarter of the neighborhoods returned less than five venues. As a result, I

chose to only focus on the top three most common category types for each neighborhood.

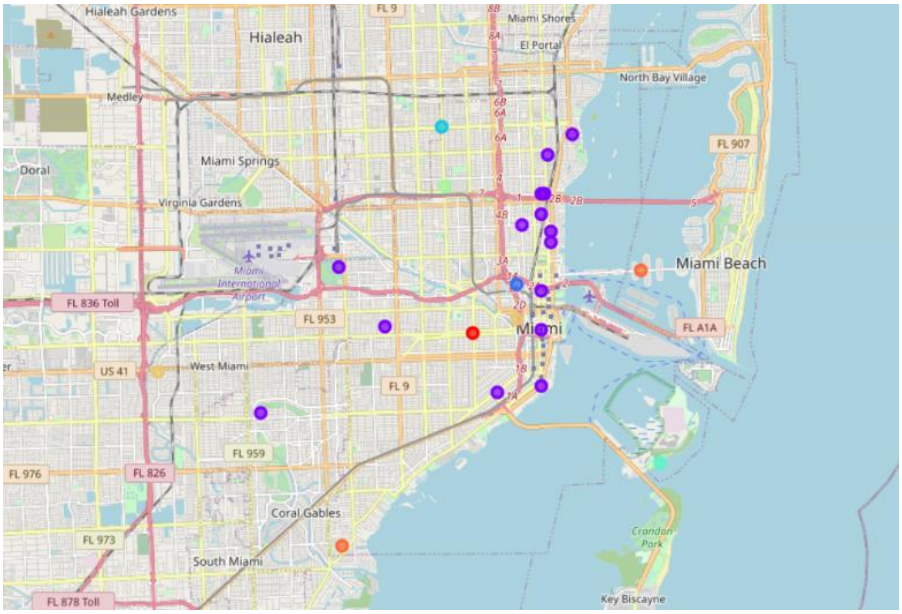
One hot encoding was used to create a dataframe with dummy variable columns showing which categories of venues were present in each neighborhood. Next, the rows of the dataframe were grouped by neighborhood and the mean of the frequency of occurrence of each category was listed. Finally, each neighborhood's name and top three most common category types were pulled into a new dataframe.

Clustering the Neighborhoods

K-Means clustering was used to segment and cluster the neighborhoods. First, the neighborhoods were clustered into five groups, after which a new merged dataframe was created, showing every neighborhood's name, coordinates, cluster label and top three most common venue types. The results were visualized on a map.



As can be seen on the map, this clustering is not very good. There are a disproportionate number of neighborhoods in Cluster 2 (purple markers). To remedy this, I repeated the process, using eight clusters instead of five, hoping to break up the large Cluster 2. In the end, Cluster 2 is still a very large cluster, but it does have fewer neighborhoods than before.



Examining the Clusters

Finally, each cluster was examined individually. A table was created for each cluster, showing which neighborhoods were included and the top three venue types for each. Using these tables, a final table naming and describing each cluster was created. While examining the clusters, I noticed that clusters 3 and 4 were very similar and I thought should have been merged, so I joined them into a single group for the results.

Results

The table below shows the chosen names and a short description of the clusters. As previously stated, Cluster 2 is very large and contains fifteen neighborhoods.

Cluster	Name	Description
1	'Latin Love'	Little Havana – Latin Restaurants
2	'Tourist Town'	City Center – Hotels, Coffee and Jewelry Stores
3 & 4	'Southern Comfort'	Overtown & Liberty City – Soul Food and Wings Restaurants
5	'Fun in the Sun'	Virginia Key – Beaches
6	'Family First'	Allapattah – Convenience and Department Stores (map shows Elementary and Middle Schools)
7	'Happy as a Clam'	Flagami & Lummus Park – Seafood Restaurants
8	'Come Sail Away'	Coconut Grove & Venetian Islands – Boat/Ferry and Parks

A new or prospective resident of Miami can look at this table and choose a neighborhood in which to live based on what type of environment and venues they want nearby. For example, an African-American person can deduce that there is a large black community in Overtown and Liberty City neighborhoods because of the type of restaurants there. A family with children might consider moving to Allapattah, which appears to be more residential and shows schools nearby on the map.

Discussion

A major problem with the project was a lack of data. There were very few FourSquare entries for several neighborhoods in Miami. With so few venue results, it is difficult to cluster the neighborhoods appropriately.

Besides a lack of data, the issue with clustering these neighborhoods could also be a result of the diversity of the area. With no particular venue category being prevalent in any of the neighborhoods, there is not enough difference between the neighborhoods to split them into separate clusters. Miami is known for its diversity in cultures, restaurants, and activities; therefore, it makes sense that there is very little concentration of a particular type of venue. In the future, maybe the very center of Miami could be examined by itself and some venue categories merged into one to increase the frequency of venue categories in various neighborhoods.

Conclusion

In this project, I analyzed the various venues present in each major neighborhood in Miami to give a prospective or new resident an idea of which neighborhood(s) they would be comfortable moving to. By clustering the twenty-four major neighborhoods in Miami using k-Means and FourSquare API, I was able to make a simple table showing which neighborhoods contain specific venues which can then be used to deduce the general environment of the neighborhood.