AJ Iglesias

Sept 13th, 2019

Code Report

Ideal Location for New Coffee Shop Franchise in Miami

1. Introduction

I love coffee, the stakeholders love coffee, and a heck of a lot of people love coffee. I have been presented with the opportunity to help determine an ideal location in Miami to place the first shop for the new franchise, Crackin’ Coffee. The owners of Crackin' Coffee believe their additions to classic coffee drinks such as caramel macchiato and iced latte will keep customers coming back and staying away from those 'other' guys. Those 'other' guys are popular establishments such as Dunkin Donuts and Starbucks Coffee, and local coffee shops. While being in the vicinity of other popular establishments may cause more people to see Crackin' Coffee it is also unlikely they will pass up the reputable establishmets for some new start up. The owners have expressed their interest in building Crackin' Coffee in a location that still has plenty of potential customers but not many places that sell coffee close by. Using Foursquare location data I’ll help determine the frequency of coffee shops in a particular neighborhood in Miami in order to place the new franchise in the best location to succeed. This study will help the founder(s) of Crackin’ Coffee's first establishment in a neighborhood where the frequency of coffee shops is low, but the population is relatively meaning the nearby competition will not be so daunting to the start up. This location could potentially make or break Crackin' Coffee as a top coffee selling brand, because a profit must be made for the owners in order to build multiple stores in different locations in the future.

2. Data acquisition and cleaning

In order to get neighborhood names and population counts I will use the following webpage: https://en.wikipedia.org/wiki/List_of_communities_in_Miami-Dade_County,_Florida, and I will convert the data into a suitable pandas dataframe while deleting the unnecessary columns from the webpage data. This left a dataframe with 34 rows each represented by a partifucular community in Miami. Once I obtained this data I was able to plot the population numbers for each of these communities which is important to note for our study. I also need Foursquare location data to best solve this problem. I will call the explore endpoint using Foursquare API in python. This will outline various venues in Miami within a specific radius allowing me to take a glance at the whole city to find the best location(s) to place Crackin' Coffee. I will use geopy to get coordinates for each of the communities of Miami-Dade these coordinates will then be combined in to the prior constructed table in order to obtain one complete table with the communities, population, and respective coordinates. Once I obtained the correct coordinates I was able to use the folium library to construct a map in order to get a clear visual of the different communities in Miami.

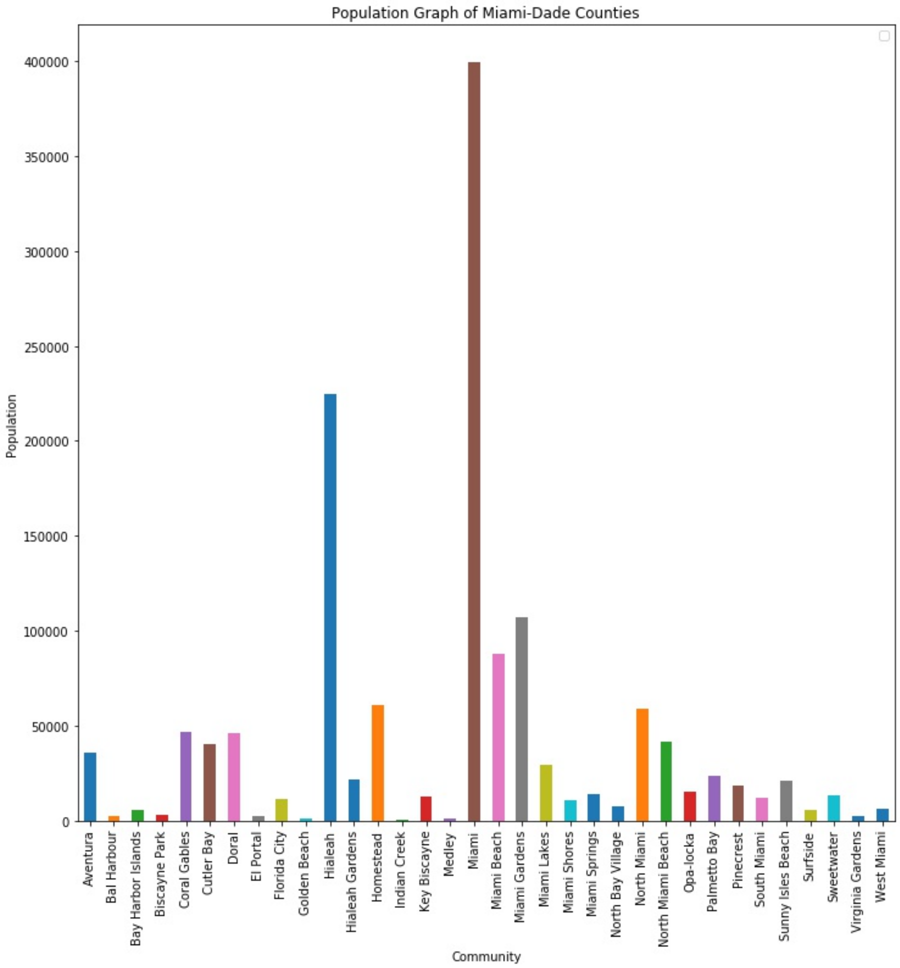
I will then work my way around those locations to get the venues for each area with a limit of 100 returned results. After running the Foursquare API for each community I was left with a dataframe with the shape (1,345, 7), which included the name of the venue, its coordinates and category (i.e. Coffee Shop). In order to construct frequency table for each community, I needed to change the categorical variables (venues) into binary vectors. This is where one-hot encoding comes in, in order to do this I grouped the venue dataframe by 'Community' and then applied .count() in order to change the all the data to numerical data.

Communities that gathered less than 20 venues were dropped due to the lack of venue data that will skew the frequency data tables. Once I made specific calls to the Foursquare API, I generated a top 5 venue frequency table in order to show the shops with the highest frequency in a particular community to make my judgement. I plotted the frequencies in a bar chart for coffee shops, bakeries and cafe because all of these have potential to sell coffee and therefore be a competitor for Crackin' Coffee.

Now we can cluster the particular communities using KMeans algorithm based on most common venues in order to get generalized areas for potential Crackin' Coffee landing spots. I utilized the elbow method in order to determine the most ideal number of clusters for my data which turned out to be 5. Once the algorithm was fit on the clustered dataset I was able to plot another map using folium however this time each community belonged to 1 of the 5 clusters and examined results. Pretty definitively I found the best spot to establish Crackin' Coffee in the city of Miami is Homestead where coffee shops, bakeries and cafes are a rarity therefore meaning I can let the stakeholders know this specific location is an adequate landing spot for their first Crackin' Coffee shop.

3. Methodology

I began my analysis by plotting the population for each community since that will give the audience a fair view of how many potential customers they will have in each respective community. Although these population numbers are from 2010 we can assume for this assignment that population number differences between communities haven't changed much. That is to say that the communities with larger population numbers are still rather large and haven't decreased greatly, if at all.

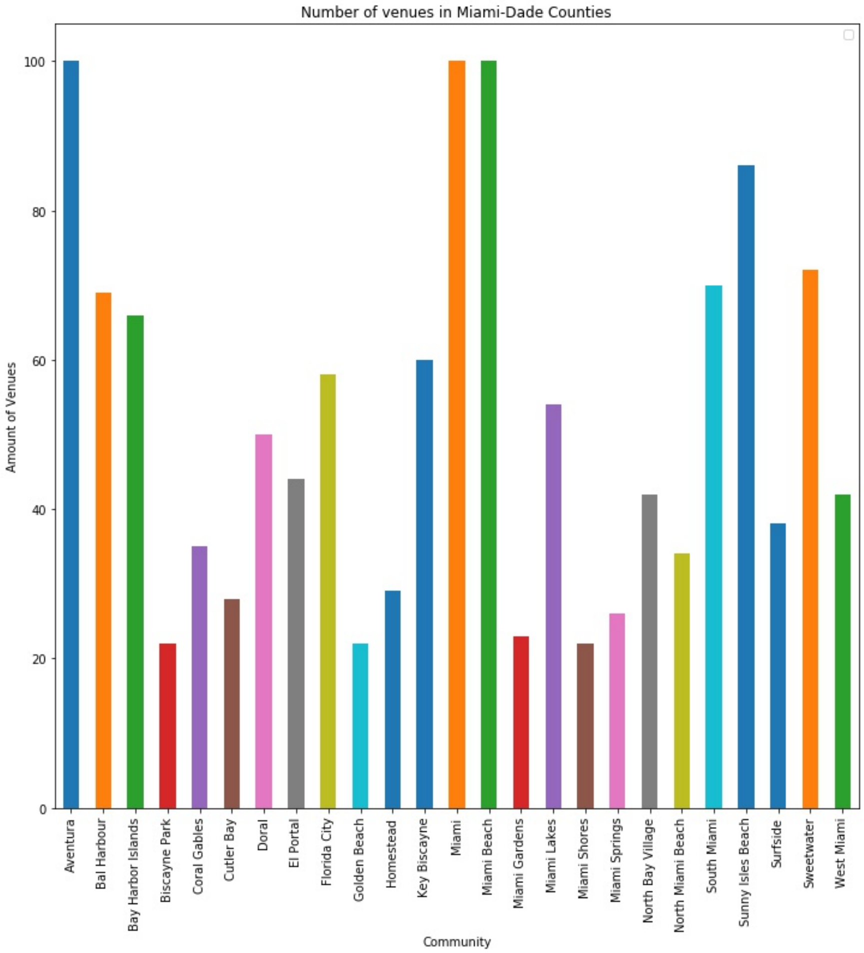


It is worth noting that the above graphing doesn't do much to solve our problem and is presented as a means to give an idea to the stakeholders what population numbers in the prospective communities look like, since it will reflect on potential customer count.

Once population was plotted it was time to take care of the geographical coordinates for all the communities in my table using the geopy library. Unfortunately, the geopy library gave faulty coordinates for Coral Gables, Florida placing the marker in the ocean. To account for this error I used the website https://www.latlong.net website in order to obtain the correct coordinates. Obtaining the correct coordinates was crucial because if faulty coordinates were given to the Foursquare API then the venues found for each location would be incorrect as well.

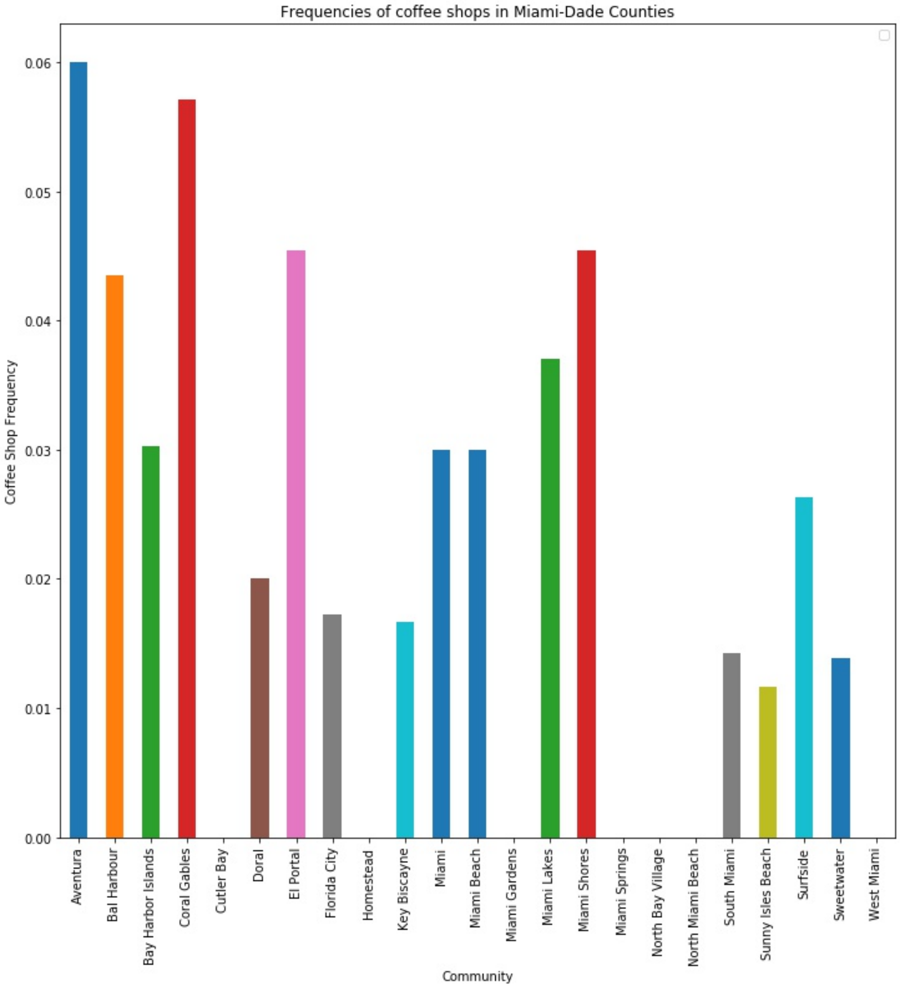
Finally, I merged the latitude and longitude table with the previously crafted table of communities and populations into a new complete table of these values to use for the Foursquare API calls.

Next, I create a function to call the Foursquare API on each community and find venues within a radius = 1000 and limit the return to 100 venues. This function returns venue names, respective latitudes and longitudes and venue category, which in particular is extremely important in my study. Before building the frequency table I removed the communities that returned less than 20 venues by grouping by community based on value counts, resulting in removing 11 communities from the table.

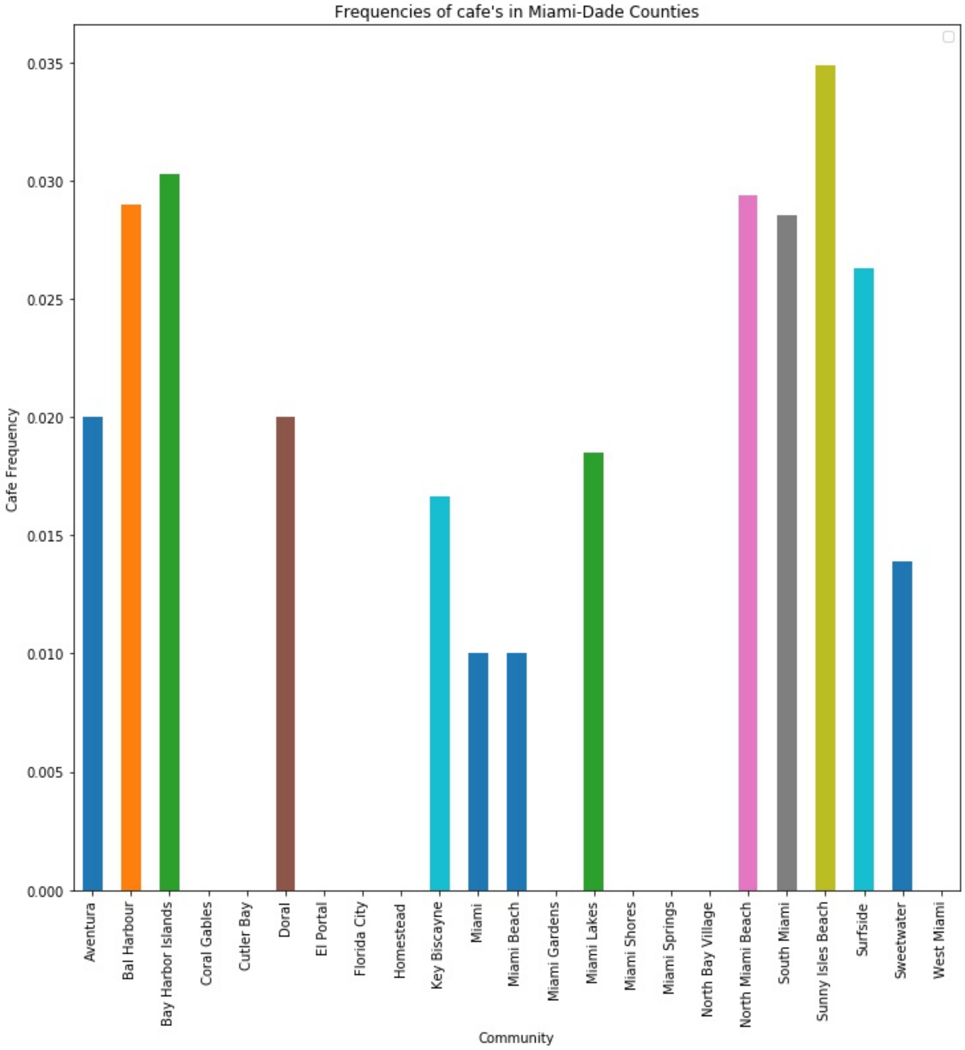


In order to build frequency tables for each of the remaining communities I needed to change the categorical variables into numerical variables. Hence, one-hot encoding needed to be implemented in order to construct binary vectors for the returned venue data. This was done by constructing dummy variables using pandas get_dummies function on the venues table column Venue Category. This makes each category now represented as a binary value, 1 or 0, allowing me to calculate frequencies of each venue throughout the table. I particularly cared about the frequency of coffee shops, café's and bakeries in each community since the end goal is find the best location to place a new coffee franchise and have the least amount of competitors there.

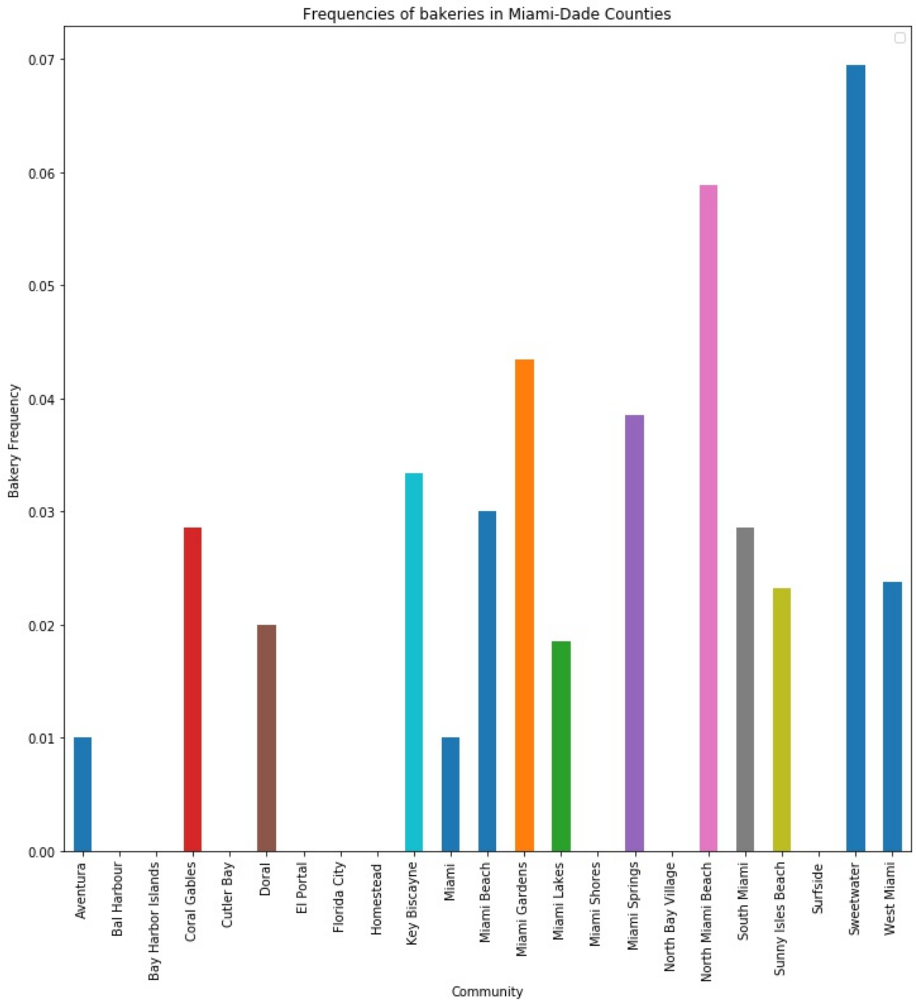
Coffee Shop Frequency:



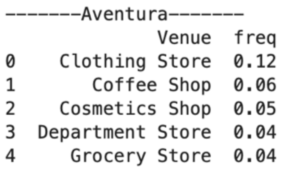
Café Frequency:



Bakery Frequency:



Although the bar charts are sufficient evidence to tell us the frequencies of places that sell coffee in each area, a frequency table of the top 5 most common venues in each area can also supply some additional information to the stakeholders. I created a function that accounts for the top 5 frequency values in the table and combines them with their respective venue for each community. This allowed me to obtain the following frequency table for Aventura,

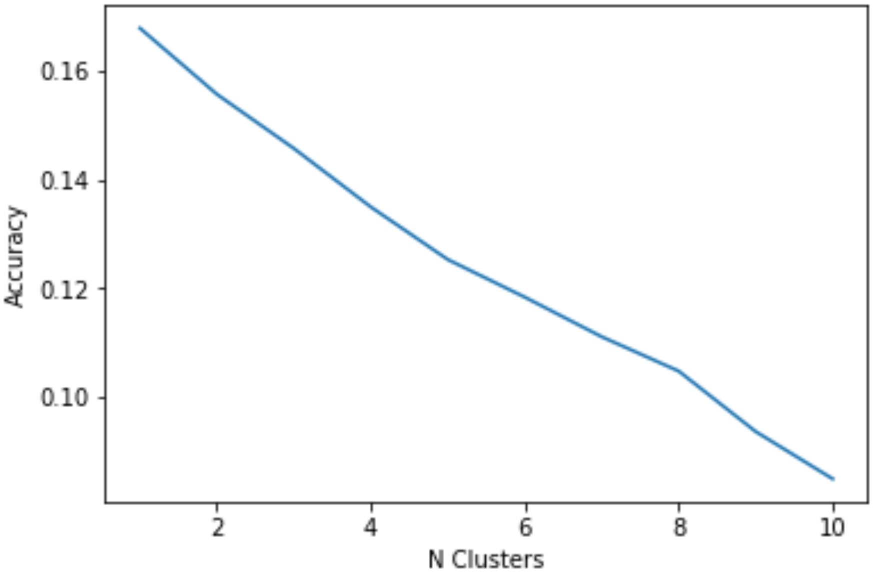


```
-------Aventura-------
                Venue  freq
0     Clothing Store  0.12
1        Coffee Shop  0.06
2     Cosmetics Shop  0.05
3   Department Store  0.04
4      Grocery Store  0.04
```

.

This was done for all communities, establishing the 5 most common venues for each community (*check the code if you want to view other communities)*.

The next steps of my program further preprocess the data for use in KMeans clustering algorithm. I begin by created a new dataframe that places the ten most common venues for each community which will help cluster the data together when using the clustering algorithm. In order to determine how many clusters to use for the Kmeans, I used the elbow method to determine the amount of n_clusters best to use.

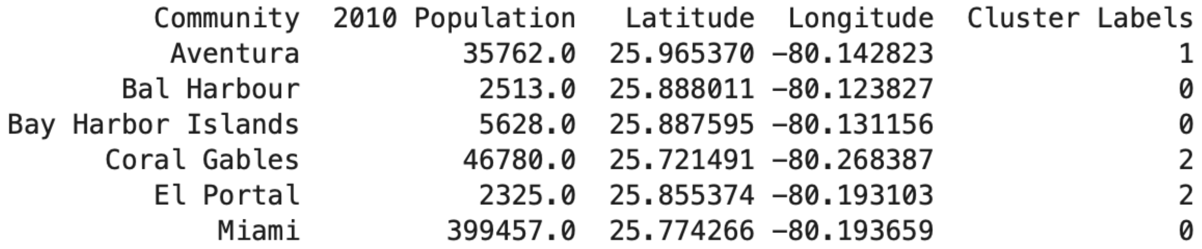


It is tough to see at first glance but we can see the 'elbow' of the graph is at N clusters = 5. Hence, I used 5 clusters when clustering the neighborhoods based on common venues.

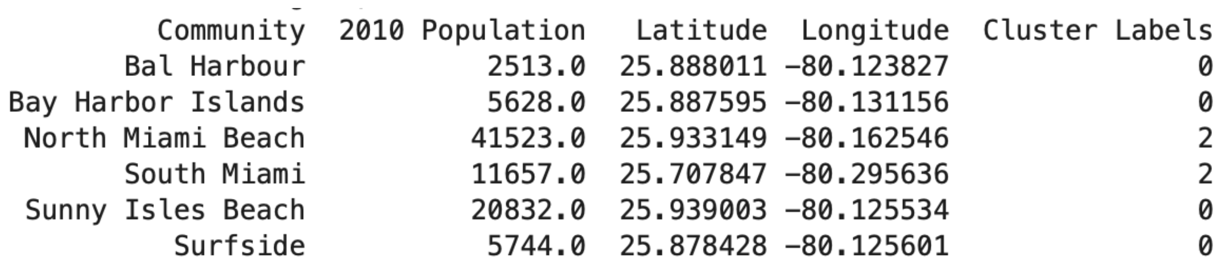
Finally, I finish my analysis by searching through the cluster table for occurrences of the words Coffee Shop, Bakery, or Café. This returned a list of communities that had either of these venue types in their top ten. Using the cluster data I also plotted a similar map to part one of this analysis using folium however this time the plot is done with the communities now in respective clusters (color coded).

4. Results

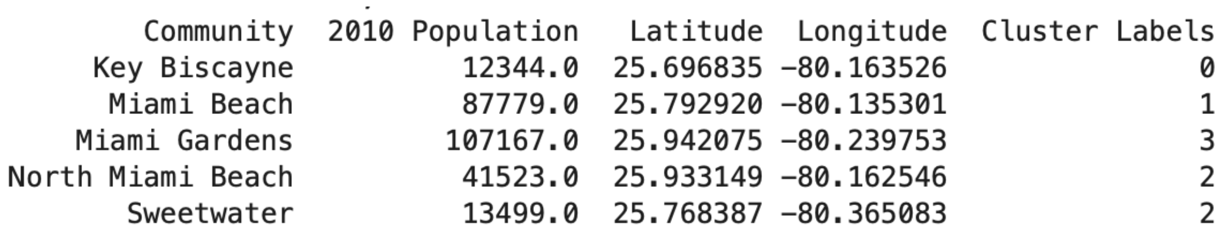
The results of my study were very informative for the stakeholders of that issued this problem to me. Based on my evidence the only cluster in my results that did not contain any coffee shops, bakeries or cafés in their top ten most common venues.



```
          Community  2010 Population   Latitude  Longitude  Cluster Labels
           Aventura          35762.0  25.965370 -80.142823               1
        Bal Harbour           2513.0  25.888011 -80.123827               0
 Bay Harbor Islands           5628.0  25.887595 -80.131156               0
       Coral Gables          46780.0  25.721491 -80.268387               2
          El Portal           2325.0  25.855374 -80.193103               2
              Miami         399457.0  25.774266 -80.193659               0
```

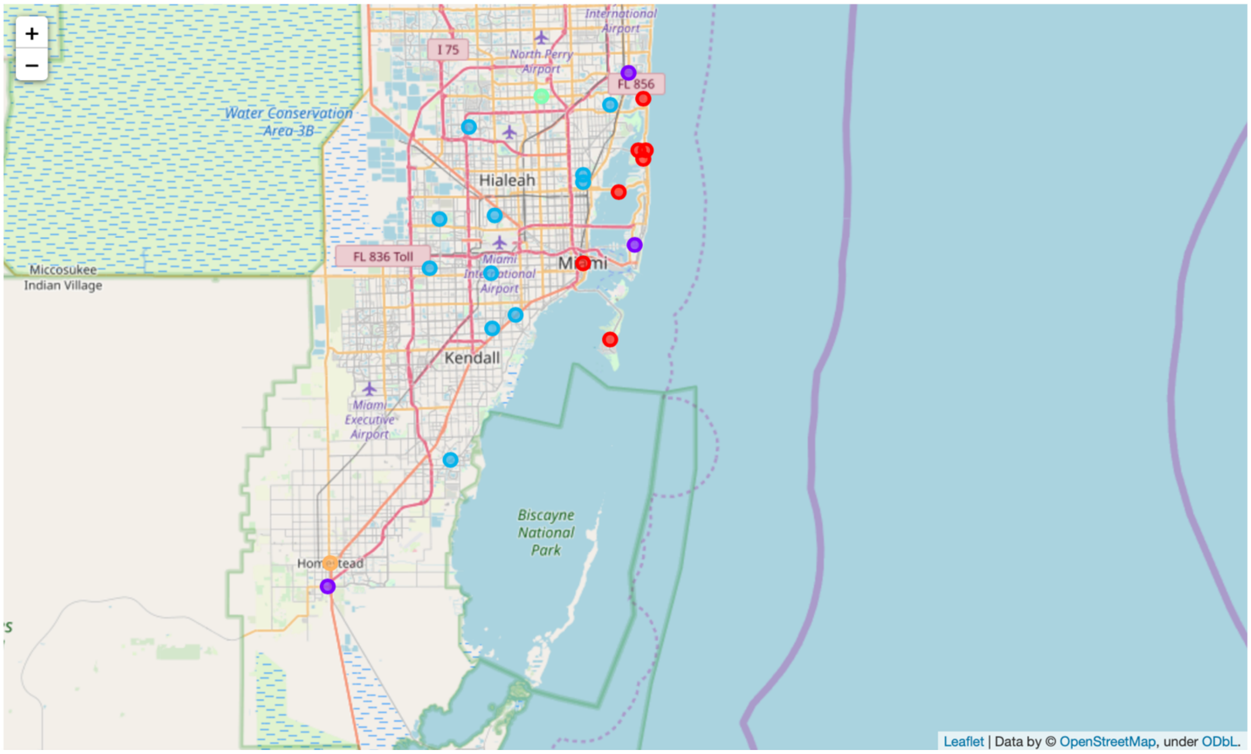


```
          Community  2010 Population   Latitude  Longitude  Cluster Labels
        Bal Harbour           2513.0  25.888011 -80.123827               0
 Bay Harbor Islands           5628.0  25.887595 -80.131156               0
  North Miami Beach          41523.0  25.933149 -80.162546               2
        South Miami          11657.0  25.707847 -80.295636               2
  Sunny Isles Beach          20832.0  25.939003 -80.125534               0
           Surfside           5744.0  25.878428 -80.125601               0
```



```
          Community  2010 Population   Latitude  Longitude  Cluster Labels
       Key Biscayne          12344.0  25.696835 -80.163526               0
        Miami Beach          87779.0  25.792920 -80.135301               1
      Miami Gardens         107167.0  25.942075 -80.239753               3
  North Miami Beach          41523.0  25.933149 -80.162546               2
         Sweetwater          13499.0  25.768387 -80.365083               2
```

Notice cluster label 4 is not included in the results above, even more impressive there was only one location in cluster 4 after running the KMeans algorithm, Homestead. These results efficiently tell the stakeholders that the best location to place a new coffee franchise where the least competition will be present is indeed Homestead, Florida. I have included a cluster map in order to share with the stakeholders in order to orient themselves with this location (Homestead or cluster 4 is denoted with the orange dot).



5. Discussion:

In this particular problem, higher population and lower frequency of competitors in a community made that community an ideal location for Crackin' Coffee's first shop. It turned out that my analysis was definitive in identifying such location rather smoothly. Homestead was proved through both the bar charts as well as the frequency table results after clustering. The KMeans algorithm clustered it (Homestead) alone and it was the only cluster to not have either coffee shop, bakery or café in its top ten most common venue.

The one issue I had during the analysis was the geopy library's coordinates for Coral Gables was incorrect. I ran it multiple times to check if it was just a runtime error or coding issue, however it just returns the wrong coordinates. When plotting it would plot Coral Gables in the ocean based on geopy's coordinates. I will contact geopy's creators to let them know about this error however for this project I simply hardcoded them in from https://www.latlong.net.

6. Conclusion:

In this study, I analyzed the abundance of coffee shops, bakeries, and cafés in the Miami area based on the communities of Miami-Dade county. I identified that these three venue types were efficient evaluators for this study because they all present competition to a new franchise that will mostly sell coffee and breakfast items. I built various frequency tables and graphs in order to effectively present the abundance of these venues in each community. I followed that up by building a classification model that helped classify venue commonality to the community's. These models helped translate geographic data into numerical data in order to help Crackin' Coffee ownership to choose a good starting location for their new coffee franchise. I am confident that these results present stakeholders with the best location to profit the most from their coffee shop and develop a successful franchise for years to come.