Data Science Capstone Project
- The Battle of the Neighborhoods

Relocating to Miami after retirement

By: L. Hao

Introduction

Miami is a metropolis in the southeastern part of the state of Florida. It is the second largest city in the state and seventh largest in the country. It is in Miami-Dade county, FL.

Miami is one of the most popular places to retire nowadays. People choose to spend their retirement life there because of high health care quality, convenient and relaxing life styles, beautiful climate, recreational opportunities all year round, and many more.

In this project, I will use data science skills to help customers who would consider relocating to Miami after retirement to get familiar with the city, as well as to find some areas they might be interested in. As most retired people take major consideration on easy access to hospital facilities, grocery shopping markets, indoor entertainment places and outdoor opportunities, I will focus on the analysis and comparison of different zip code areas for these four specific category groups. A detailed report and effective presentation will be provided to the customers so they can make their final decision based on their individual requirements or interests.

Data

To solve the problem, the following data is needed:

1. List of zips in Miami and related information, such as latitude and longitude coordinates and population of those zips. This information is required to plot the map, get the venue data, and do all the analysis.
2. Use geopy library to get the geographical coordinates of Miami, FL.
3. Use Foursquare API to get venues for each zip in Miami. The venue data in category groups like hospital facilities, grocery facilities, indoor facilities and outdoor facilities is particularly useful for this project. The venue data is also used to perform clustering on different zips.

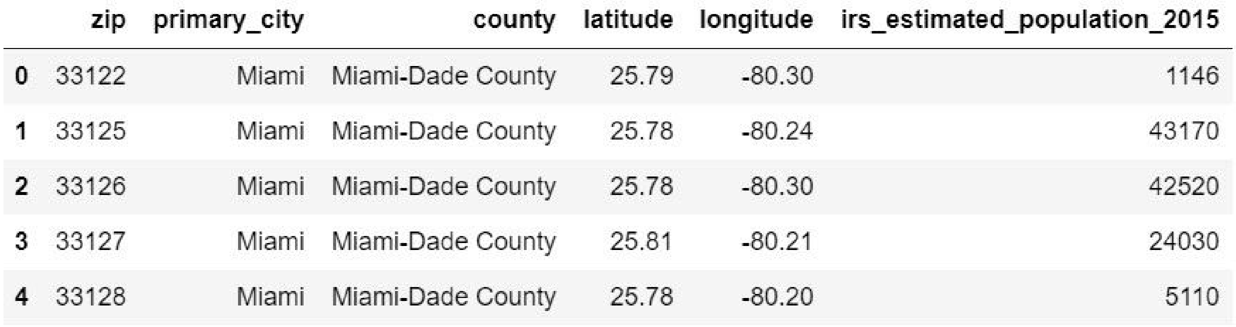
Methodology

I used the following libraries or APIs in this project:

1. pandas for data manipulation and analysis.
2. geopy library to get the geographical coordinates of a location.
3. Python Folium library to visualize geospatial data.
4. Foursquare API to explore venues for different zips.
5. Matplotlib library and associated modules for plotting.
6. Scikit-learn k-means as the method of unsupervised learning to cluster the zips.

1. Download and Explore Dataset

The original data was downloaded from https://www.unitedstateszipcodes.org/. It contains zip codes, primary cities, counties, states, time zones, area codes, latitudes, longitudes, IRS estimated populations in 2015, etc. I put the dataset into a dataframe, and extracted related rows for all the standard zips of the city of Miami, FL with the population more than 0. Then, I only kept some main columns (zip, primary_city, county, latitude, longitude, irs_estimated_populartion_2015).



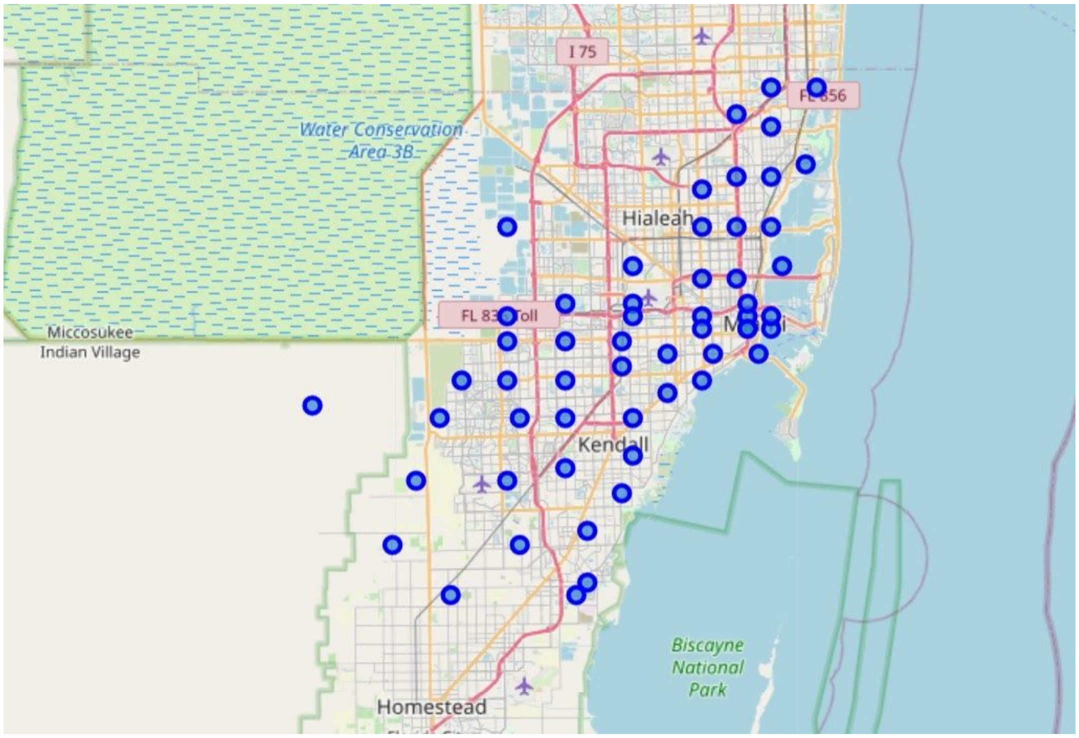
	zip	primary_city	county	latitude	longitude	irs_estimated_population_2015
0	33122	Miami	Miami-Dade County	25.79	-80.30	1146
1	33125	Miami	Miami-Dade County	25.78	-80.24	43170
2	33126	Miami	Miami-Dade County	25.78	-80.30	42520
3	33127	Miami	Miami-Dade County	25.81	-80.21	24030
4	33128	Miami	Miami-Dade County	25.78	-80.20	5110

Use geopy library to get the geographical coordinates of Miami, FL.

The geograpical coordinate of Miami, FL are 25.7741728, -80.19362.

Create the map of Miami.

I used python folium library to create a map of Miami with its zips marked on it.

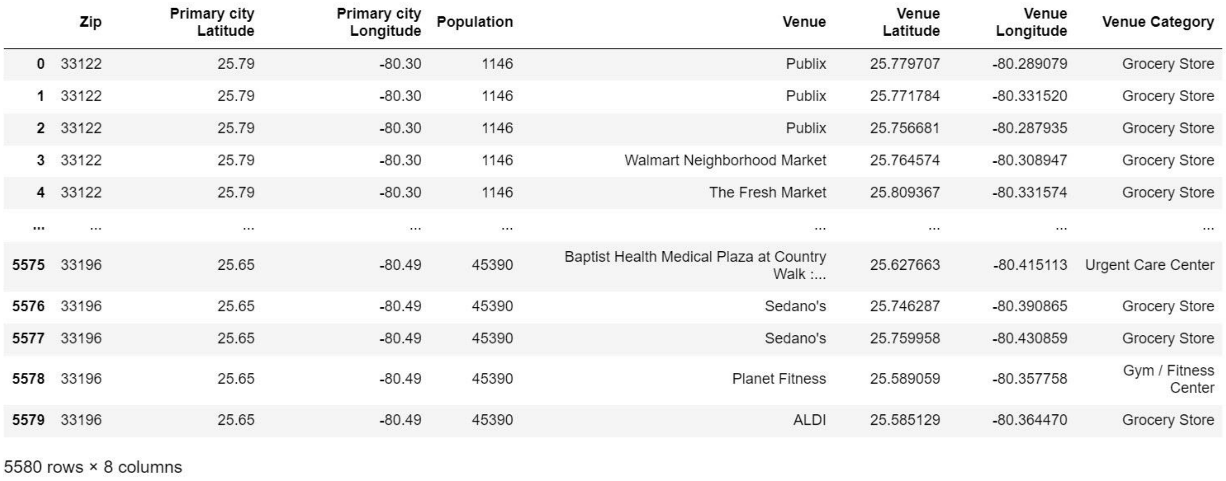


2. Use the Foursquare API to explore each zip

The Foursquare API provides a global database of venue data.

1. Before calling the API, I set up the credentials, category IDs, and version, I set the limit as 100 which means I would get back up to 100 venues close to each zip. Based on the customers' preferences, I looked up the venue category IDs from the Foursquare Venue Category Hierarchy at https://developer.foursquare.com/docs/build-with-foursquare/categories/. The category IDs in this section can be modified for different customers with different requirements.
2. By calling the Foursquare API and then parsing the returned JSON file, I got the top 100 (or less, if there are not as much as 100) venues within a radius of about 10 miles of each zip coordinates.

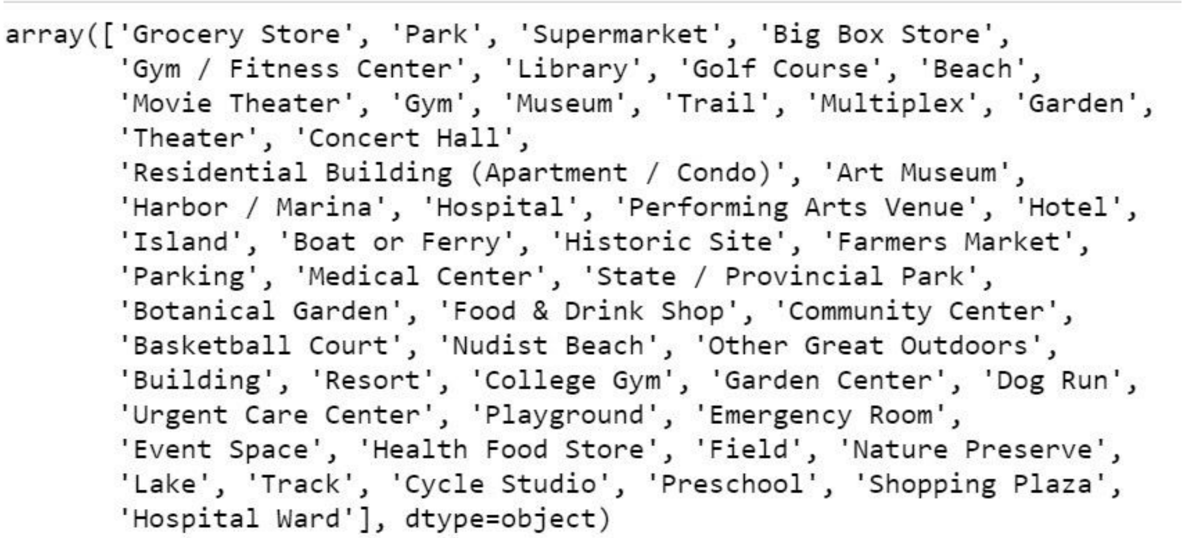
There are totally 5580 venues returned, as shown below:



	Zip	Primary city Latitude	Primary city Longitude	Population	Venue	Venue Latitude	Venue Longitude	Venue Category
0	33122	25.79	-80.30	1146	Publix	25.779707	-80.289079	Grocery Store
1	33122	25.79	-80.30	1146	Publix	25.771784	-80.331520	Grocery Store
2	33122	25.79	-80.30	1146	Publix	25.756681	-80.287935	Grocery Store
3	33122	25.79	-80.30	1146	Walmart Neighborhood Market	25.764574	-80.308947	Grocery Store
4	33122	25.79	-80.30	1146	The Fresh Market	25.809367	-80.331574	Grocery Store
...	...	...	...	...	...	...	...	...
5575	33196	25.65	-80.49	45390	Baptist Health Medical Plaza at Country Walk :...	25.627663	-80.415113	Urgent Care Center
5576	33196	25.65	-80.49	45390	Sedano's	25.746287	-80.390865	Grocery Store
5577	33196	25.65	-80.49	45390	Sedano's	25.759958	-80.430859	Grocery Store
5578	33196	25.65	-80.49	45390	Planet Fitness	25.589059	-80.357758	Gym / Fitness Center
5579	33196	25.65	-80.49	45390	ALDI	25.585129	-80.364470	Grocery Store

5580 rows × 8 columns

There are 53 unique categories:

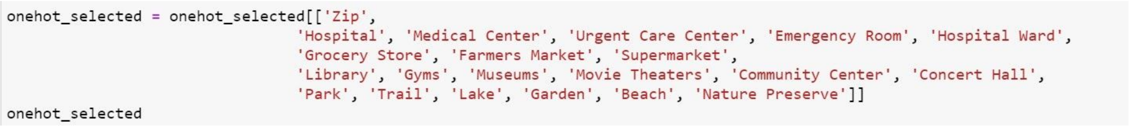


```
array(['Grocery Store', 'Park', 'Supermarket', 'Big Box Store',
       'Gym / Fitness Center', 'Library', 'Golf Course', 'Beach',
       'Movie Theater', 'Gym', 'Museum', 'Trail', 'Multiplex', 'Garden',
       'Theater', 'Concert Hall',
       'Residential Building (Apartment / Condo)', 'Art Museum',
       'Harbor / Marina', 'Hospital', 'Performing Arts Venue', 'Hotel',
       'Island', 'Boat or Ferry', 'Historic Site', 'Farmers Market',
       'Parking', 'Medical Center', 'State / Provincial Park',
       'Botanical Garden', 'Food & Drink Shop', 'Community Center',
       'Basketball Court', 'Nudist Beach', 'Other Great Outdoors',
       'Building', 'Resort', 'College Gym', 'Garden Center', 'Dog Run',
       'Urgent Care Center', 'Playground', 'Emergency Room',
       'Event Space', 'Health Food Store', 'Field', 'Nature Preserve',
       'Lake', 'Track', 'Cycle Studio', 'Preschool', 'Shopping Plaza',
       'Hospital Ward'], dtype=object)
```

3. Analyze the zips in Miami

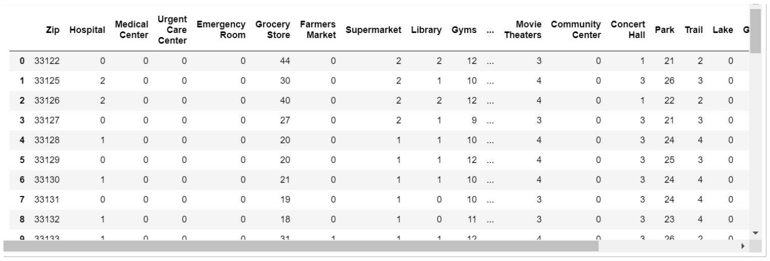
Step 1: Convert venues into a dataframe grouped by zips and show categories of interest.

1. Use one hot encoding and pandas get_dummies function to convert categorical data into numerical data.
2. Combine some similar types of venue categories, like all types of museums as museums, all types of gyms as gyms, and all types of movie theaters as movie theaters.
3. Select only those venue categories of interest.



```
onehot_selected = onehot_selected[['Zip',
                                   'Hospital', 'Medical Center', 'Urgent Care Center', 'Emergency Room', 'Hospital Ward',
                                   'Grocery Store', 'Farmers Market', 'Supermarket',
                                   'Library', 'Gyms', 'Museums', 'Movie Theaters', 'Community Center', 'Concert Hall',
                                   'Park', 'Trail', 'Lake', 'Garden', 'Beach', 'Nature Preserve']]
onehot_selected
```

4. Group by Zip to get a dataframe showing those selected venues for each zip.

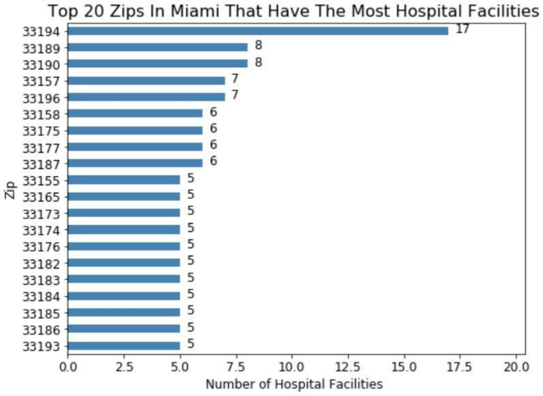


	Zip	Hospital	Medical Center	Urgent Care Center	Emergency Room	Grocery Store	Farmers Market	Supermarket	Library	Gyms	...	Movie Theaters	Community Center	Concert Hall	Park	Trail	Lake	G
0	33122	0	0	0	0	44	0	2	2	12	...	3	0	1	21	2	0	
1	33125	2	0	0	0	30	0	2	1	10	...	4	0	3	26	3	0	
2	33126	2	0	0	0	40	0	2	2	12	...	4	0	1	22	2	0	
3	33127	0	0	0	0	27	0	2	1	9	...	3	0	3	21	3	0	
4	33128	1	0	0	0	20	0	1	1	10	...	4	0	3	24	4	0	
5	33129	0	0	0	0	20	0	1	1	12	...	4	0	3	25	3	0	
6	33130	1	0	0	0	21	0	1	1	10	...	4	0	3	24	4	0	
7	33131	0	0	0	0	19	0	1	0	10	...	3	0	3	24	4	0	
8	33132	1	0	0	0	18	0	1	0	11	...	3	0	3	23	4	0	
9	33133	1	0	0	0	31	1	1	1	12		4	0	3	26	2	0	

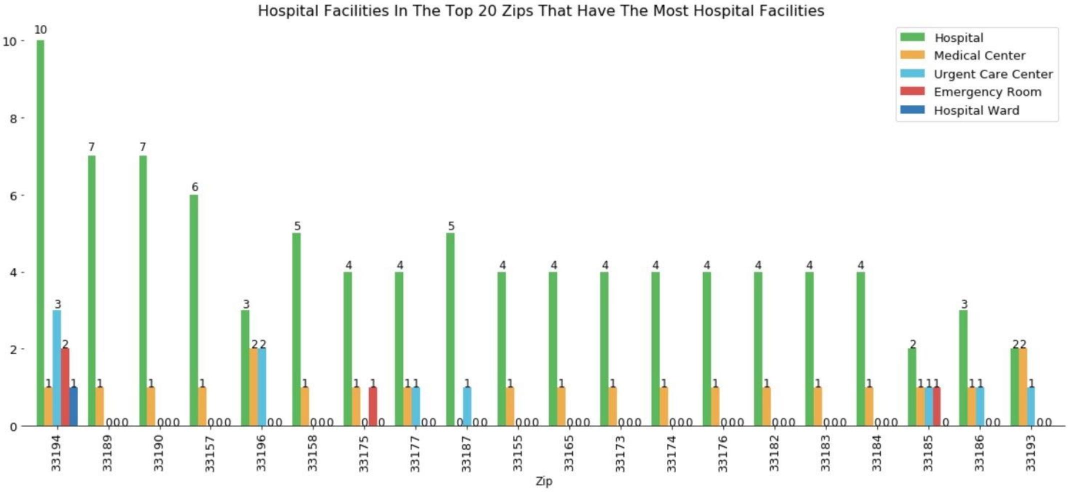
Step 2: Analyze the zip codes and compare them.

1. Define some functions:
   a) a grouping function to extract venues categories in the comparing group;
   b) a sorting function to sort a dataframe by some criteria;
   c) a plotting function to plot a bar chart to show the total counts of the category group for the top zips;
   d) a plotting function to plot a bar chart to show the count of each category in the category group for the top zips.
2. Analyze the 4 category groups (hospital, grocery, indoor, outdoor):

1) Analyze hospital facilities including hospitals, medical centers, urgent care centers, and emergency rooms, by calling the functions defined:
   a) Find the top 10 (or more, if there are ties) zips that have the most total numbers of hospital facilities, and plot two bar charts for each of the zips:
      i) The total number of venues in hospital facilities category group:

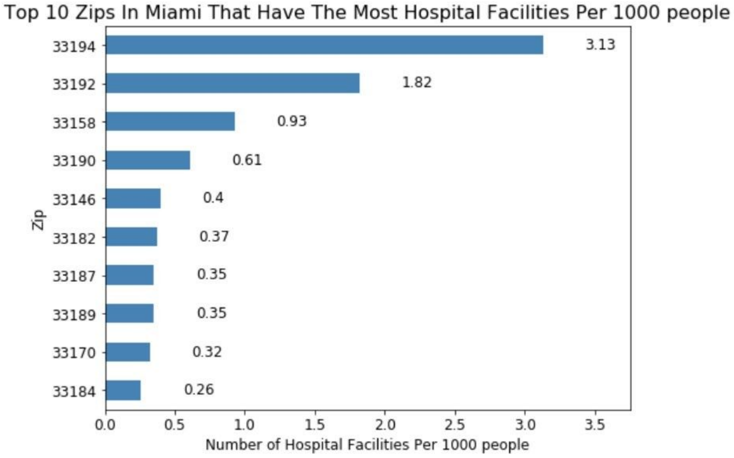


ii) The number of each venue category in the hospital facilities:

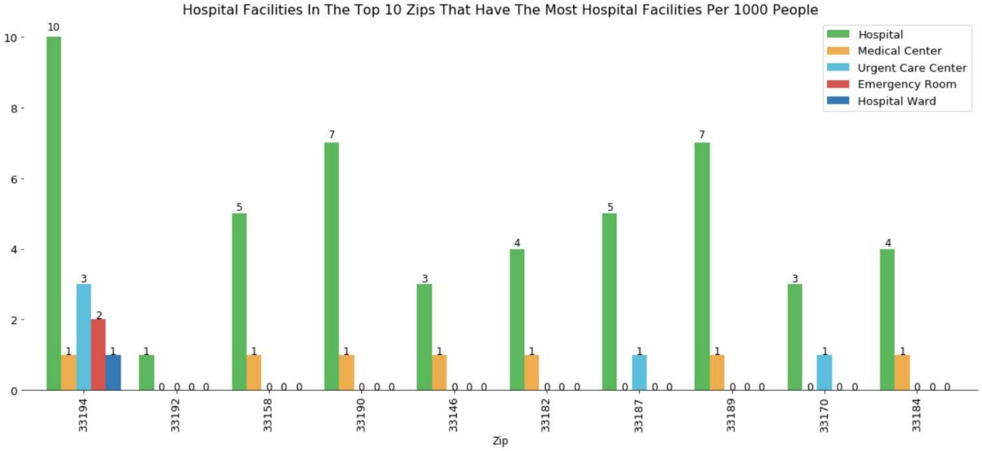


b) Find the top 10 (or more, if there are ties) zips that have the most numbers of hospital facilities per 1000 people, and plot two bar charts:

i) The number of venues per 1000 people in hospital facilities category group:



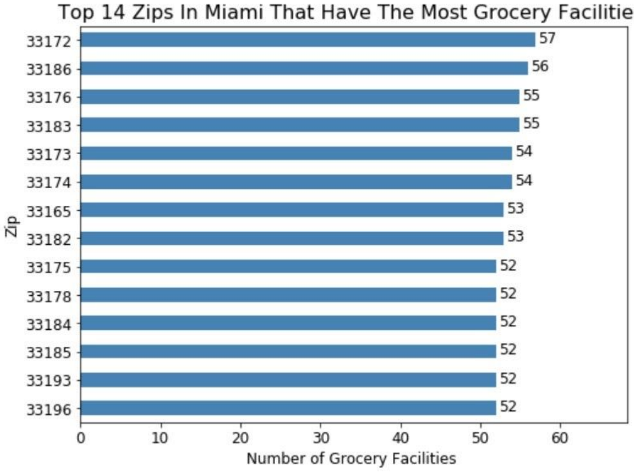
ii) The number of each venue category in the hospital facilities:



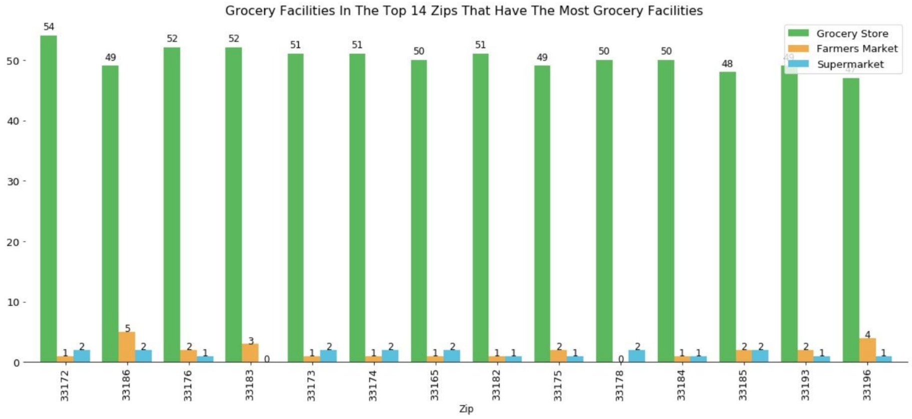
2) Analyze grocery facilities including grocery stores, farmers markets, and supermarkets, by calling the functions defined:

   a) Find the top 10 (or more, if there are ties) zips that have the most total numbers of grocery facilities, and plot two bar charts for each of the zips:

      i) The total number of venues in grocery facilities category group:

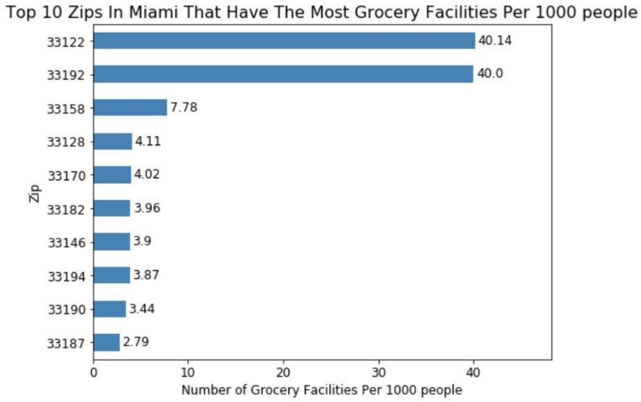


      ii) The number of each venue category in the grocery facilities:

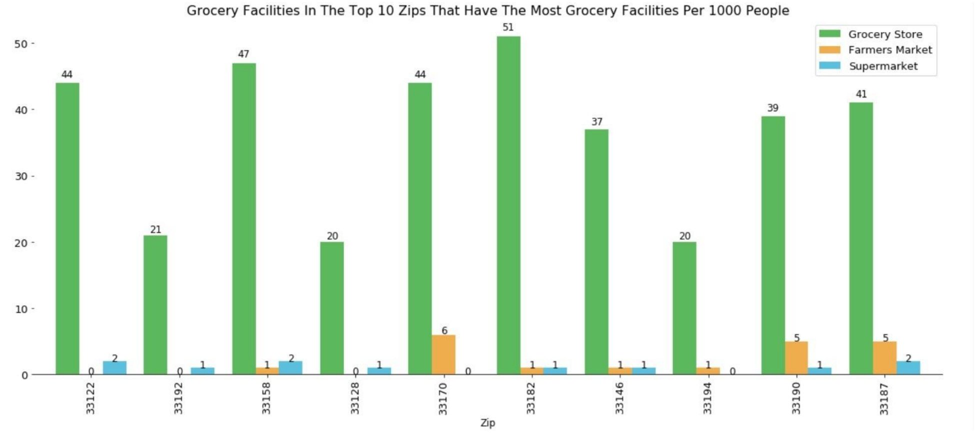


   b) Find the top 10 (or more, if there are ties) zips that have the most numbers of grocery facilities per 1000 people, and plot two bar charts:

      i) The number of venues per 1000 people in grocery facilities category group:



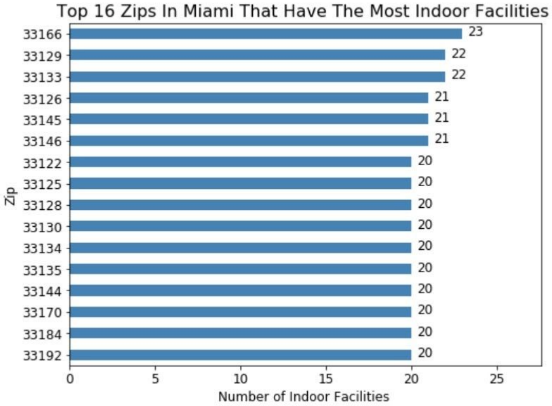
ii) The number of each venue category in the grocery facilities:



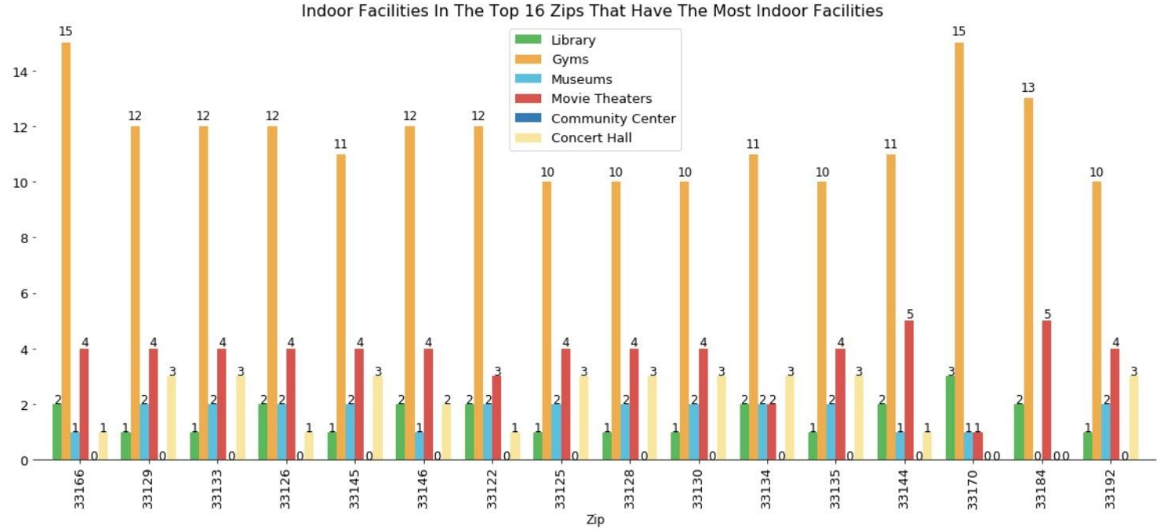
3) Analyze indoor facilities including libraries, gyms, museums, movie theaters, community centers, and concert halls, by calling the functions defined:

a) Find the top 10 (or more, if there are ties) zips that have the most total numbers of indoor facilities, and plot two bar charts for each of the zips:

i) The total number of venues in indoor facilities category group:

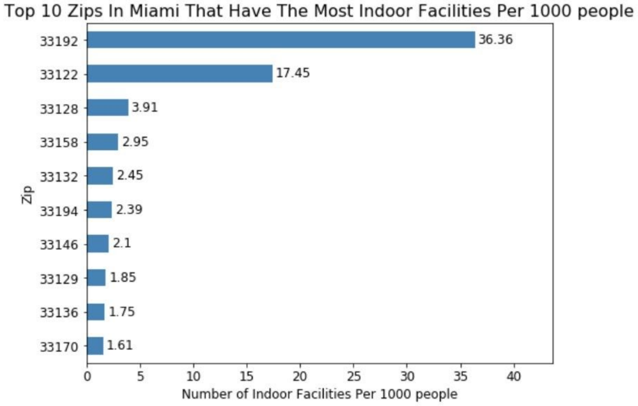


ii) The number of each venue category in the indoor facilities:

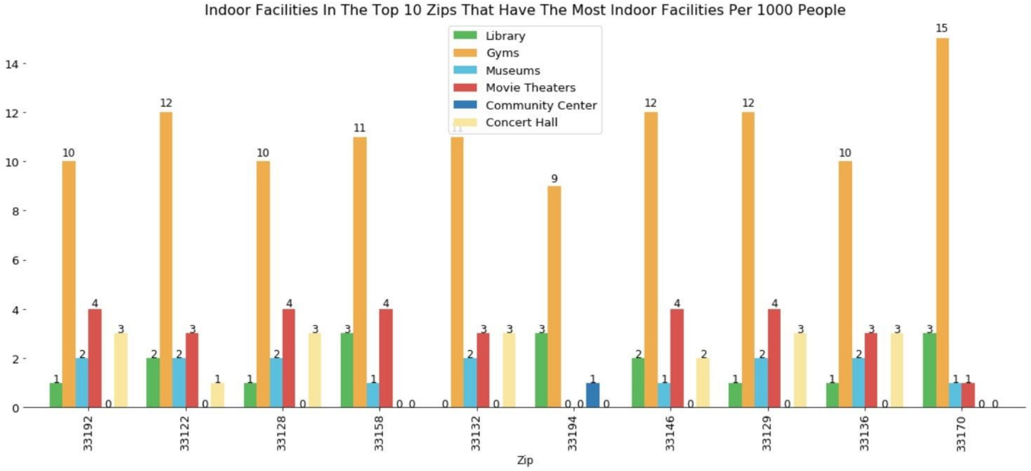


b) Find the top 10 (or more, if there are ties) zips that have the most numbers of indoor facilities per 1000 people, and plot two bar charts:

i) The number of venues per 1000 people in indoor facilities category group:



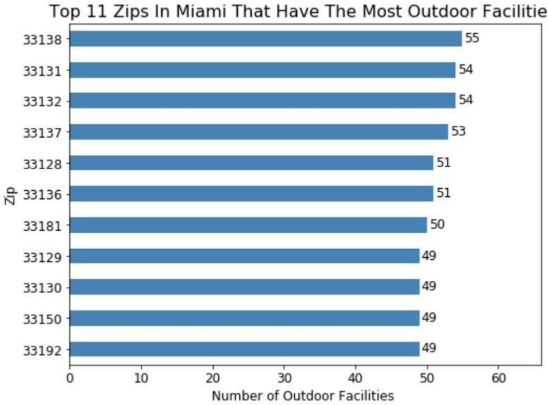
ii) The number of each venue category in the indoor facilities:



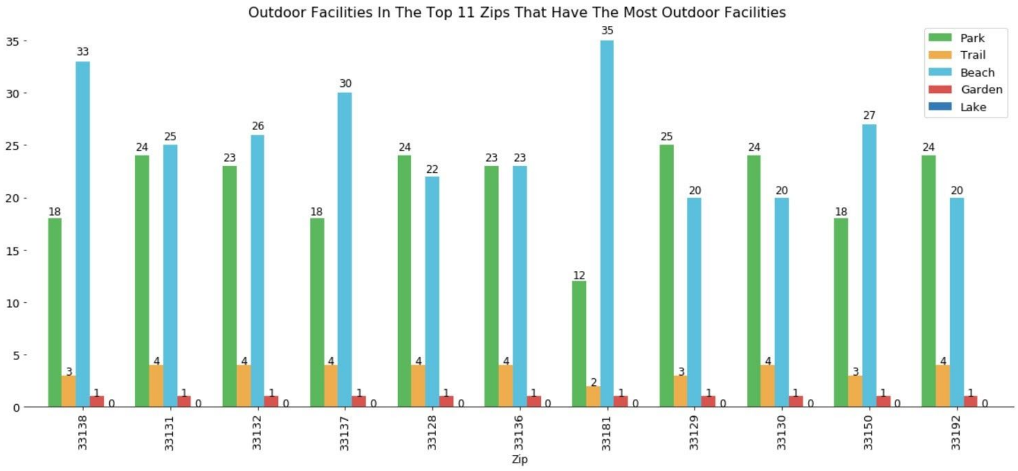
4) Analyze outdoor facilities including parks, trails, gardens, beaches, and lakes, by calling the functions defined:

a) Find the top 10 (or more, if there are ties) zips that have the most total numbers of outdoor facilities, and plot two bar charts for each of the zips:

i) The total number of venues in outdoor facilities category group:

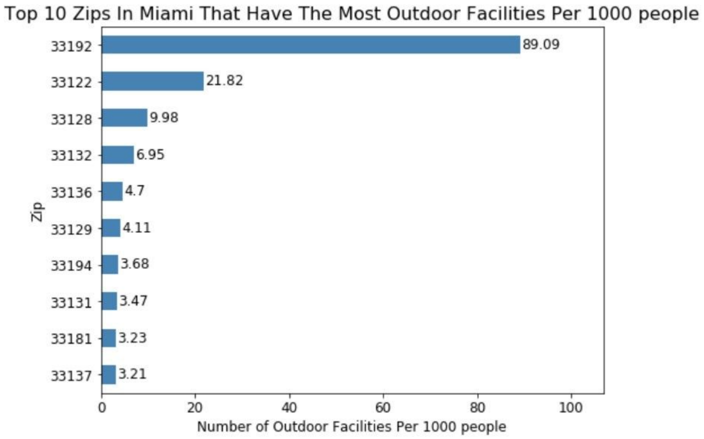


ii) The number of each venue category in the outdoor facilities:

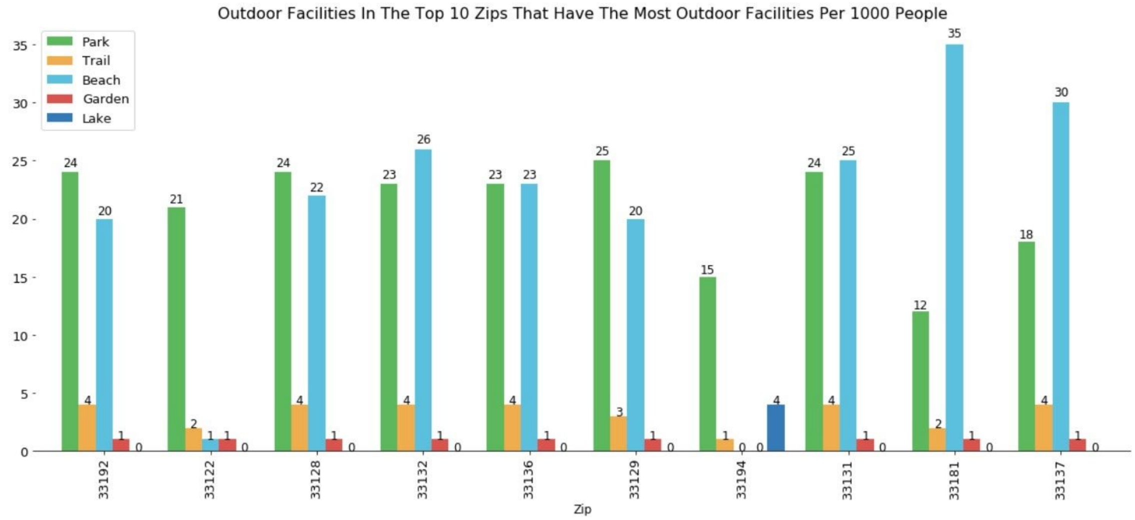


b) Find the top 10 (or more, if there are ties) zips that have the most numbers of outdoor facilities per 1000 people, and plot two bar charts:

i) The number of venues per 1000 people in outdoor facilities category group:



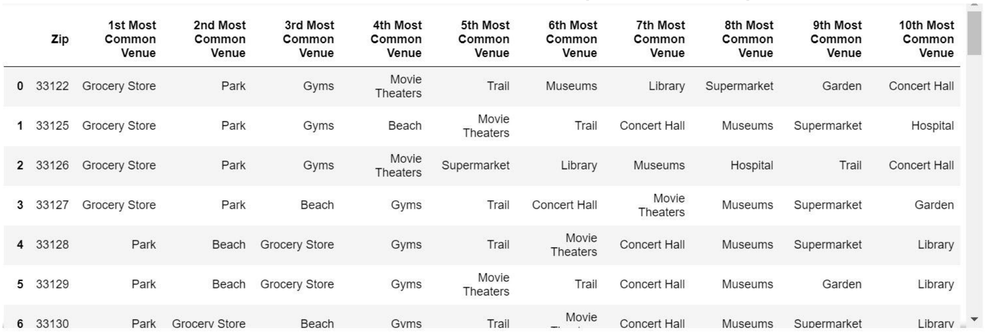
ii) The number of each venue category in the outdoor facilities:



4. Cluster the zips and examine the clusters

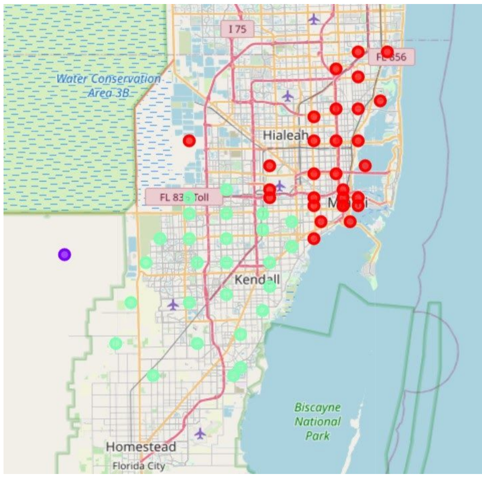
1. Group rows by zips and take the mean of the frequency of occurrence of each category.

2. Find the 10 most common venues for each zip code and put them into a dataframe.

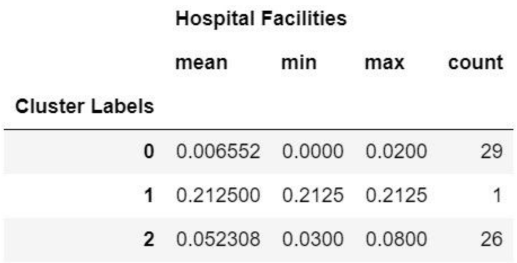


	Zip	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	33122	Grocery Store	Park	Gyms	Movie Theaters	Trail	Museums	Library	Supermarket	Garden	Concert Hall
1	33125	Grocery Store	Park	Gyms	Beach	Movie Theaters	Trail	Concert Hall	Museums	Supermarket	Hospital
2	33126	Grocery Store	Park	Gyms	Movie Theaters	Supermarket	Library	Museums	Hospital	Trail	Concert Hall
3	33127	Grocery Store	Park	Beach	Gyms	Trail	Concert Hall	Movie Theaters	Museums	Supermarket	Garden
4	33128	Park	Beach	Grocery Store	Gyms	Trail	Movie Theaters	Concert Hall	Museums	Supermarket	Library
5	33129	Park	Beach	Grocery Store	Gyms	Movie Theaters	Trail	Concert Hall	Museums	Garden	Library
6	33130	Park	Grocery Store	Beach	Gyms	Trail	Movie	Concert Hall	Museums	Supermarket	Library

3. Run k-means to cluster zips for different category groups, and examine the clusters.
1) Define some functions:
   a) a function to get a clustering dataframe for a category group and a statistical dataframe showing the means, min, max, and number of zips of each cluster.
   b) a function to create a clustering map.
2) Choose the k value.
3) Do clustering for each of the four category groups.
   a) For hospital facilities:
      Clustering map (red for cluster 0, purple for cluster 1, green for cluster 2)



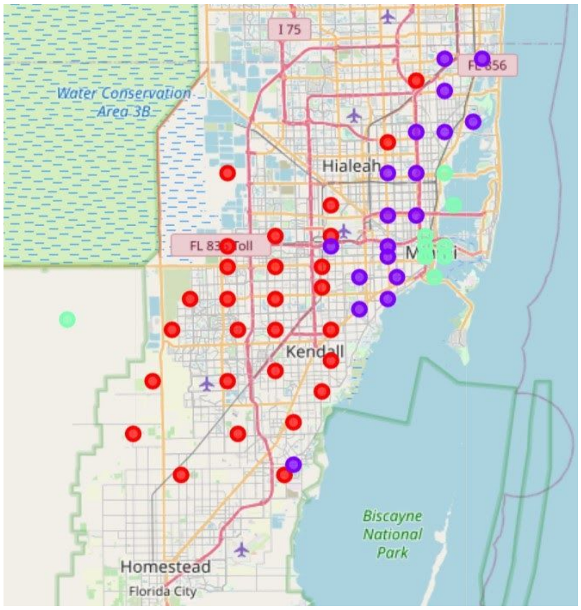
Clustering statistics:



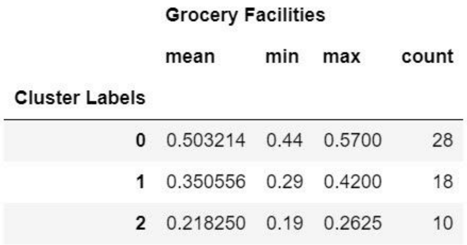
	Hospital Facilities			
	mean	min	max	count
Cluster Labels				
0	0.006552	0.0000	0.0200	29
1	0.212500	0.2125	0.2125	1
2	0.052308	0.0300	0.0800	26

Then, for each cluster, create a dataframe to show the zips and their means.

b) For grocery facilities:
Clustering map (red for cluster 0, purple for cluster 1, green for cluster 2)



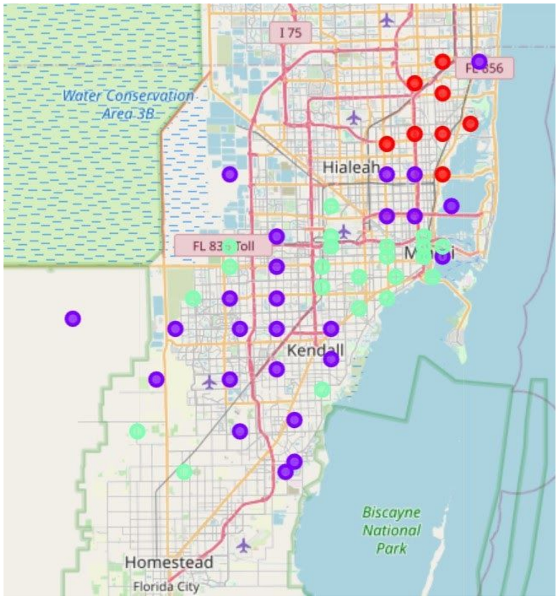
Clustering statistics:



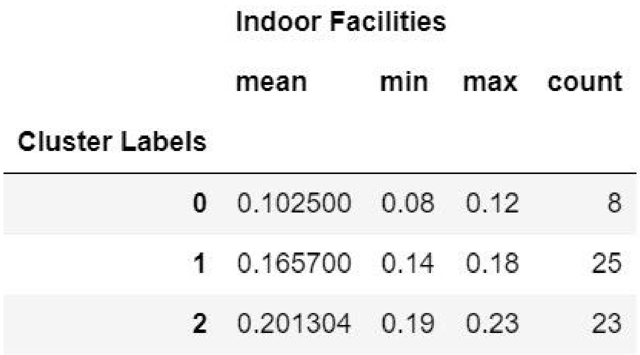
	Grocery Facilities			
	mean	min	max	count
Cluster Labels				
0	0.503214	0.44	0.5700	28
1	0.350556	0.29	0.4200	18
2	0.218250	0.19	0.2625	10

Then for each cluster, create a dataframe to show the zips and their means.

c) For indoor facilities:
Clustering map (red for cluster 0, purple for cluster 1, green for cluster 2)



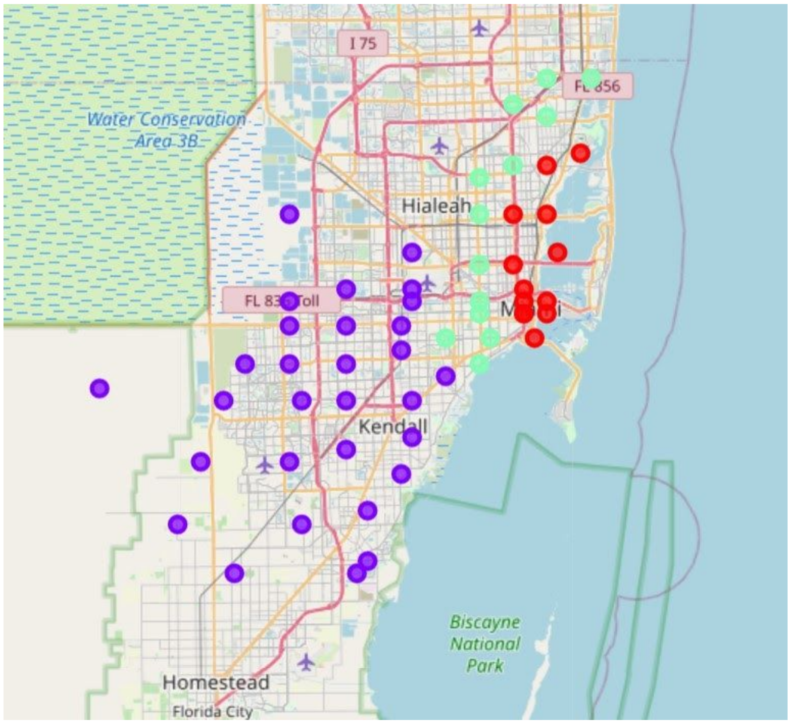
Clustering statistics:



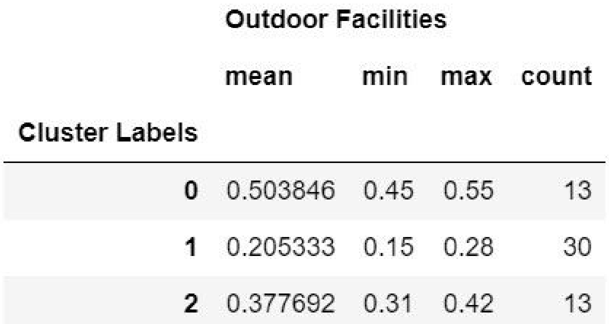
	Indoor Facilities			
	mean	min	max	count
Cluster Labels				
0	0.102500	0.08	0.12	8
1	0.165700	0.14	0.18	25
2	0.201304	0.19	0.23	23

Then, for each cluster, create a dataframe to show the zips and their means.

d) For outdoor facilities:
Clustering map (red for cluster 0, purple for cluster 1, green for cluster 2)



Clustering statistics:

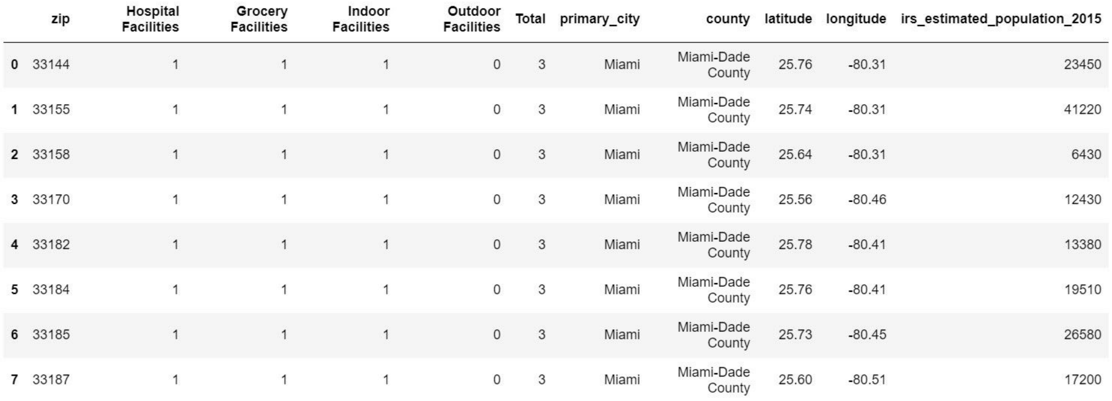


	Outdoor Facilities			
	mean	min	max	count
Cluster Labels				
0	0.503846	0.45	0.55	13
1	0.205333	0.15	0.28	30
2	0.377692	0.31	0.42	13

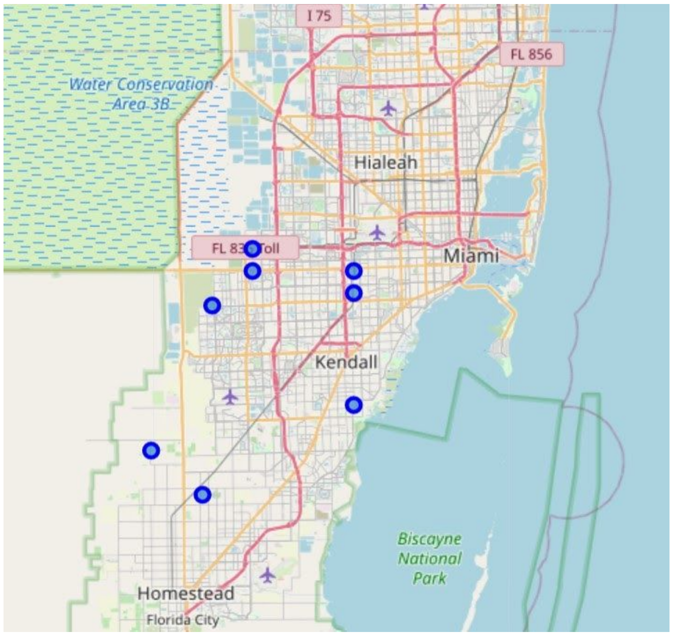
Then, for each cluster, create a dataframe to show the zips and their means.

Result

At the final section, I extracted those zips in most of the clusters with relatively more facilities and created a map showing them.



	zip	Hospital Facilities	Grocery Facilities	Indoor Facilities	Outdoor Facilities	Total	primary_city	county	latitude	longitude	irs_estimated_population_2015
0	33144	1	1	1	0	3	Miami	Miami-Dade County	25.76	-80.31	23450
1	33155	1	1	1	0	3	Miami	Miami-Dade County	25.74	-80.31	41220
2	33158	1	1	1	0	3	Miami	Miami-Dade County	25.64	-80.31	6430
3	33170	1	1	1	0	3	Miami	Miami-Dade County	25.56	-80.46	12430
4	33182	1	1	1	0	3	Miami	Miami-Dade County	25.78	-80.41	13380
5	33184	1	1	1	0	3	Miami	Miami-Dade County	25.76	-80.41	19510
6	33185	1	1	1	0	3	Miami	Miami-Dade County	25.73	-80.45	26580
7	33187	1	1	1	0	3	Miami	Miami-Dade County	25.60	-80.51	17200



As customers are more interested in four category groups related to their retirement lifes (hospital facilities, grocery shopping facilities, indoor entertainment facilities, outdoor facilities), our analysis did clustering on each of these groups. For each group, I clustered those 56 standard zips in Miami into 3 clusters, representing zips with relatively more facilities, relatively fewer facilities, and median number of facilities.

Then I would like to show customers the zips that occur most frequently in the clusters with "relatively more facilities". Please note there is only 1 zip in the "relatively more facilities" cluster for hospital facilities, so I included the cluster with "median number of facilities" for this category group as well. As shown in the dataframe final_zips, there are 8 zips appearing in 3 out of all 4 clusters with relatively more facilities (or median number for hospital facilities). It seems that all of these 8 zips are with high concentration of hospital facilities, grocery facilities, and indoor facilities.

The map marking all these 8 final top zips, as shown above, tells the customers the distribution of the areas.

Discussion

In this project, our major category groups are hospital facilities, grocery facilities, indoor facilities and outdoor facilities. But different customers may have different preferences. With slight changes of the category ids to call the Foursquare API, this project is reusable for other categories, such as restaurants, residences, gas stations, etc. It can also be reused for other cities or counties or areas in the country.

Conclusion

Our analysis could help people who would relocate to Miami for retirement to answer the following questions:

Which zips have the most hospitals, medical centers, urgent care centers, and emergency rooms?

Which zips have the most grocery stores, farmers markets, and supermarkets?

Which zips have the most libraries, gyms, museums, movie theaters, community centers, and concert halls?

Which zips have the most parks, trails, gardens, beaches, and lakes?

Which zips have the highest average facilities per 1000 people for those category groups?

What are the 10 most common venues for each zip code?

If we cluster the zips into 3 clusters for each category group, how is the distribution of all the zips?

Which zips will be in the final list of consideration because they occur most frequently in the clusters with "relatively more facilities"?

And where are they located on the city map?

I used Foursquare API to get the venues for all the standard zips in Miami. I found top zips with the highest numbers of total and average facilities for each category group, and plotted bar charts to demonstrate the numbers. Bar charts are also used to compare each venue in the category group for the top zips. With the k-means clustering method, for each category group, I clustered the zips into 3 clusters based on the concentration, and marked them with different colors on the Miami map. Finally, I found the zip list appearing in most clusters having more facilities.

The final decision on which zips of the city they would move to will be made by customers based on their individual or specific requirements/interests. With minor changes, this analysis can also be used by customers who would relocate to any city in the country with other preferences. Furthermore, I can extend this project to analyze and compare the neighborhoods within each zip on the final list.