Analyzing the cities of South Florida using unsupervised learning - K-means clustering

Rama Naidu Yedla

March 2019

1. Introduction

1.1. Background

South Florida (colloquially and locally known as SoFlo) is a geographic and cultural region that generally comprises Florida's southernmost counties, and is the fourth most populous (6.69 million) urban agglomeration in the United States. It is one of Florida's three most common "directional" regions, the others being Central Florida and North Florida. It includes the populous Miami metropolitan area, the Everglades, the Florida Keys, and other localities. South Florida is the only part of the continental United States with a tropical climate.

1.2. Problem

I want to start a business in South Florida and need to be associated or get in partnership with the companies in various cities. I am from Boca Raton, and my idea is to start with Boca and then expand to other cities in South Florida. The target businesses are usually B2C, and my product will help these businesses get more customers for them.

The idea behind this project is to find out the most common venues within Boca Raton and then find the cities that are alike with Boca. The algorithm to be used is k-means clustering so I can find out the cities that share common characteristics with Boca. As the first step, I need to find out the top venue categories and top venues in Boca Raton and to expand my market; I need to find out the cities that are like Boca Raton.

1.3. Interest

In addition to my personal preference of understanding the cities in Florida cities, this study can be of immense help to the people who want to start business, increase the sales of the existing business and to those who want to explore South Florida for recreational and research purposes

2. Data acquisition and cleaning

2.1. Data sources

i. Data about the list of South Florida Cities obtained from Wikipedia: This is used to get the list of cities for consideration
ii. Coordinates of these cities from my Github repo obtained from google maps: This is used to drop pins at these cities and search venues around
iii. Coordinates of Boca Raton to find out the venues, obtained from google maps: This is used to search venues around Boca
iv. Foursquare API to fetch venues and popular venues: To get the list of venue categories and the venue names basing the lat long values

2.2. Data cleaning and preparation

There is no much cleaning process involved given the type of data - most of it is structured. As a first step, after the data is obtained from the sources, cities are appended with the coordinates and then matched with exact city names obtained from Wikipedia.

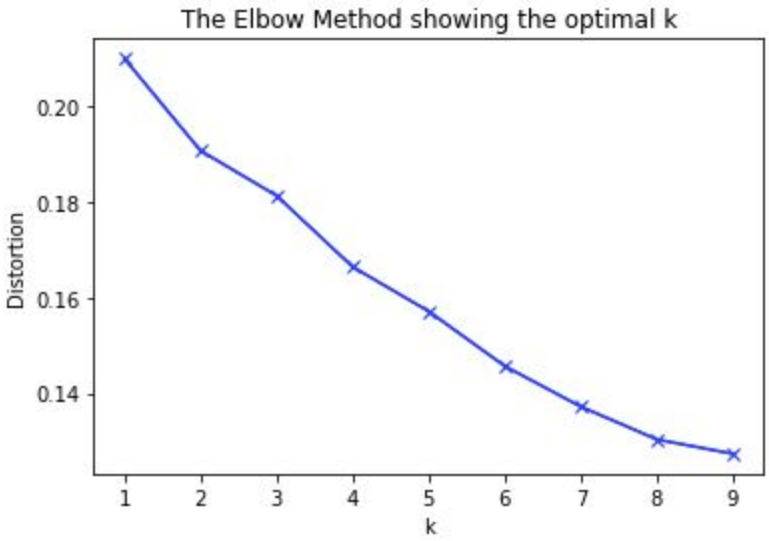
The columns with the year data is removed as we have no requirement with the population per year.

3. Clustering

3.1. Finding Optimal K

With this in place let's see the top venue categories by different cities. Showing the top 5 cities for you and their top 10 venue categories. From the initial impression I can see that the infamous places or venues fall in the category of restaurants. This is just a visual interpretation of the first five cities and their venues. Taking you to the next steps, k-means clustering and finding cities alike.

Starting with the idea of finding the number of clusters to be included i.e. the value of k. I have tried with the elbow method and it seems the data is uniform and the elbow is not seen in the chart.



Silhouette Analysis

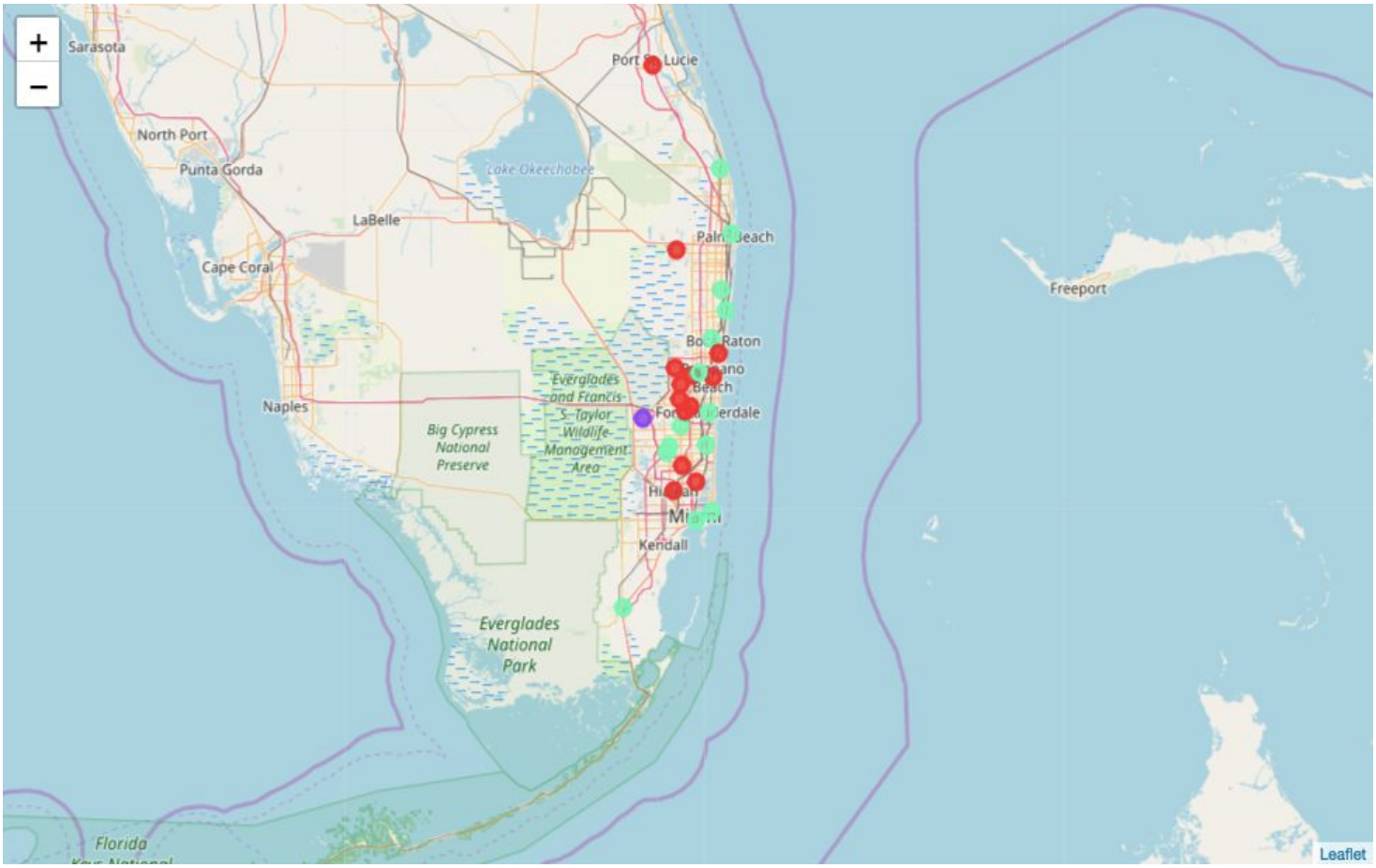
```
For n_clusters = 2 The average silhouette_score is : 0.35324224445705965
For n_clusters = 3 The average silhouette_score is : 0.3740381846991288
For n_clusters = 4 The average silhouette_score is : 0.06175955525547112
For n_clusters = 5 The average silhouette_score is : 0.0669426176397799
For n_clusters = 6 The average silhouette_score is : 0.05211305398965616
For n_clusters = 7 The average silhouette_score is : 0.0576606284854901
For n_clusters = 8 The average silhouette_score is : 0.04502509525120853
For n_clusters = 9 The average silhouette_score is : 0.050732872417508304
```

I have tried with Silhouette Analysis and the optimal is found to be 3 and the k-means clustering algorithm is ran and the below clusters were found

3.2. Clusters found

i. Bar and Restaurants
ii. Fast Food
iii. Business services

Below are the clusters and the most famous venues by cluster:


Sarasota
North Port
Punta Gorda
LaBelle
Cape Coral
Naples
Lake Okeechobee
Port St. Lucie
Palm Beach
Boca Raton
Boynton Beach
Fort Lauderdale
Everglades and Francis S. Taylor Wildlife Management Area
Big Cypress National Preserve
Miami
Kendall
Everglades National Park
Florida Keys National
Freeport
Leaflet

Cluster 1

	City	Clusters	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	Miami	0	Italian Restaurant	Hotel	Steakhouse	Argentinian Restaurant	Coffee Shop	Latin American Restaurant	Japanese Restaurant	Gym / Fitness Center	Residential Building (Apartment / Condo)	Mediterranean Restaurant
2	Fort Lauderdale	0	American Restaurant	Sushi Restaurant	Italian Restaurant	Seafood Restaurant	Bar	Ice Cream Shop	Art Museum	Mexican Restaurant	Sandwich Place	Salon / Barbershop
4	Pembroke Pines	0	Burger Joint	Furniture / Home Store	Pharmacy	Pizza Place	Coffee Shop	Restaurant	Lingerie Store	Fast Food Restaurant	Shopping Mall	Latin American Restaurant
5	Hollywood	0	Bar	Peruvian Restaurant	Sandwich Place	Cocktail Bar	Sushi Restaurant	Mexican Restaurant	Latin American Restaurant	Coffee Shop	Pharmacy	Pizza Place
6	Miramar	0	Food Truck	BBQ Joint	Latin American Restaurant	Donut Shop	Basketball Court	Italian Restaurant	Performing Arts Venue	Spa	Gym / Fitness Center	Gymnastics Gym
9	West Palm Beach	0	Bar	Asian Restaurant	Wine Bar	Mexican Restaurant	French Restaurant	Restaurant	Middle Eastern Restaurant	Farmers Market	Park	Italian Restaurant
11	Davie	0	Coffee Shop	Bar	Hookah Bar	Baseball Field	Deli / Bodega	Dessert Shop	Pub	College Rec Center	Seafood Restaurant	Fast Food Restaurant
12	Miami Beach	0	Hotel	Lounge	Italian Restaurant	Beach	Bar	Bakery	Performing Arts Venue	Coffee Shop	Spa	Pharmacy
15	Boca Raton	0	American Restaurant	Department Store	Clothing Store	Cosmetics Shop	Lingerie Store	Jewelry Store	Burger Joint	Furniture / Home Store	Restaurant	Steakhouse
17	Boynton Beach	0	Department Store	Bakery	American Restaurant	Mexican Restaurant	Pet Store	Furniture / Home Store	Gym	Jewish Restaurant	Spa	Bookstore
20	Delray Beach	0	Pizza Place	Italian Restaurant	New American Restaurant	American Restaurant	Seafood Restaurant	Burger Joint	Dessert Shop	Mediterranean Restaurant	Coffee Shop	Thai Restaurant
21	Homestead	0	Mexican Restaurant	American Restaurant	Pizza Place	Donut Shop	Thai Restaurant	Grocery Store	Sandwich Place	Big Box Store	Restaurant	Fried Chicken Joint

Cluster 2

	City	Clusters	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
1	Hialeah	1	Grocery Store	Cuban Restaurant	Discount Store	South American Restaurant	Soccer Stadium	Theater	Sandwich Place	Dive Shop	Chinese Restaurant	Fast Food Restaurant
3	Port St. Lucie	1	Pizza Place	Gym / Fitness Center	Sandwich Place	Chinese Restaurant	Fast Food Restaurant	Discount Store	Shipping Store	Business Service	Rest Area	Pharmacy
7	Coral Springs	1	Italian Restaurant	Asian Restaurant	Indian Restaurant	Fast Food Restaurant	Pizza Place	Grocery Store	Theater	Spa	Sandwich Place	Fried Chicken Joint
8	Miami Gardens	1	Fast Food Restaurant	Fried Chicken Joint	Pizza Place	Athletics & Sports	Caribbean Restaurant	Strip Club	Seafood Restaurant	Bakery	Clothing Store	Shoe Store
10	Pompano Beach	1	Fast Food Restaurant	Convenience Store	Intersection	Pizza Place	Donut Shop	Mediterranean Restaurant	Café	Flea Market	Food Truck	Fried Chicken Joint
13	Plantation	1	Park	Light Rail Station	Coffee Shop	Fast Food Restaurant	Liquor Store	South American Restaurant	Seafood Restaurant	Lawyer	Flower Shop	Sandwich Place
14	Sunrise	1	Mobile Phone Shop	Coffee Shop	Caribbean Restaurant	Latin American Restaurant	Hotel	Big Box Store	American Restaurant	Rental Car Location	Colombian Restaurant	Discount Store
16	Deerfield Beach	1	Seafood Restaurant	Fast Food Restaurant	Thrift / Vintage Store	Sandwich Place	Donut Shop	Mobile Phone Shop	Pharmacy	Burger Joint	Shipping Store	Salon / Barbershop
18	Lauderhill	1	Caribbean Restaurant	Business Service	Shoe Store	Discount Store	Indian Restaurant	Big Box Store	Intersection	Toll Plaza	Laundromat	Yoga Studio
22	Tamarac	1	Gym	Grocery Store	Pharmacy	Furniture / Home Store	Fast Food Restaurant	Liquor Store	Donut Shop	Lawyer	Food & Drink Shop	Food Truck

Cluster 3

	City	Clusters	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
19	Weston	2	Business Service	Home Service	Golf Course	Park	Department Store	Eastern European Restaurant	Flea Market	Fast Food Restaurant	Farmers Market	Falafel Restaurant

3.3. Finding the top 10 venues by the city

I used the K Means algorithm as part of this clustering study. When I tested with Silhouette Analysis, I set the optimum k value to 3. Based on the study it can be found that the below cities are like Boca Raton.

	City	Clusters	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	Miami	0	Italian Restaurant	Hotel	Steakhouse	Argentinian Restaurant	Coffee Shop	Latin American Restaurant	Japanese Restaurant	Gym / Fitness Center	Residential Building (Apartment / Condo)	Mediterranean Restaurant
2	Fort Lauderdale	0	American Restaurant	Sushi Restaurant	Italian Restaurant	Seafood Restaurant	Bar	Ice Cream Shop	Art Museum	Mexican Restaurant	Sandwich Place	Salon / Barbershop
4	Pembroke Pines	0	Burger Joint	Furniture / Home Store	Pharmacy	Pizza Place	Coffee Shop	Restaurant	Lingerie Store	Fast Food Restaurant	Shopping Mall	Latin American Restaurant
5	Hollywood	0	Bar	Peruvian Restaurant	Sandwich Place	Cocktail Bar	Sushi Restaurant	Mexican Restaurant	Latin American Restaurant	Coffee Shop	Pharmacy	Pizza Place
6	Miramar	0	Food Truck	BBQ Joint	Latin American Restaurant	Donut Shop	Basketball Court	Italian Restaurant	Performing Arts Venue	Spa	Gym / Fitness Center	Gymnastics Gym
9	West Palm Beach	0	Bar	Asian Restaurant	Wine Bar	Mexican Restaurant	French Restaurant	Restaurant	Middle Eastern Restaurant	Farmers Market	Park	Italian Restaurant
11	Davie	0	Coffee Shop	Bar	Hookah Bar	Baseball Field	Deli / Bodega	Dessert Shop	Pub	College Rec Center	Seafood Restaurant	Fast Food Restaurant
12	Miami Beach	0	Hotel	Lounge	Italian Restaurant	Beach	Bar	Bakery	Performing Arts Venue	Coffee Shop	Spa	Pharmacy
15	Boca Raton	0	American Restaurant	Department Store	Clothing Store	Cosmetics Shop	Lingerie Store	Jewelry Store	Burger Joint	Furniture / Home Store	Restaurant	Steakhouse
17	Boynton Beach	0	Department Store	Bakery	American Restaurant	Mexican Restaurant	Pet Store	Furniture / Home Store	Gym	Jewish Restaurant	Spa	Bookstore
20	Delray Beach	0	Pizza Place	Italian Restaurant	New American Restaurant	American Restaurant	Seafood Restaurant	Burger Joint	Dessert Shop	Mediterranean Restaurant	Coffee Shop	Thai Restaurant
21	Homestead	0	Mexican Restaurant	American Restaurant	Pizza Place	Donut Shop	Thai Restaurant	Grocery Store	Sandwich Place	Big Box Store	Restaurant	Fried Chicken Joint
25	Jupiter	0	Sandwich Place	Pizza Place	Grocery Store	Pharmacy	Breakfast Spot	Fast Food Restaurant	Donut Shop	Italian Restaurant	American Restaurant	Convenience Store
27	Coconut Creek	0	Convenience Store	Italian Restaurant	Donut Shop	Chinese Restaurant	Business Service	Burger Joint	Pool	Bistro	Tennis Court	Dance Studio

4. Conclusion

The analysis, unsupervised learning algorithm is used to find out the common clusters in the cities of South Florida. The study was started with getting data from the sources and then linking it to the coordinates of the cities. Then the clustering algorithm is used, to understand the clusters and find the cities that are alike.

Finding cities alike will be of great help to find out the target markets and most common venues will be leveraged to explore business opportunities

5. Feature Directions

Using Silhoutte analysis optimal K was found and then proceeded to the next steps in the study. However, the number of K might change with different data points added / removed from the study. As this study is reliant on the FourSquare API changes in the API will affect the reproducibility of this study. The distortion is also found to be the smallest at the K 3.

Given the current market trends, the study has responded in a way it is now. With the change in the people preferences and the establishment of new businesses, the impact of this study will be vulnerable and has to be reproduced before making a decision. However, I can profoundly believe that the reliance of this study can be extended to at least 6 months from the date mentioned above.