APRIL 5, 2020

APPLIED DATA SCIENCE CAPSTONE

FINAL PROJECT

TABLE OF CONTENTS

Problem Statement

Many people throughout the united states of America internally emigrate to Florida. In fact, Florida is the number one state on receiving internal emigrants every year. Also, a significant percentage of the population of the new emigrants (emigrants from another country) prefer to land in Florida in the hope to either find a job or start a new business. Unfortunately, most people had a business idea but did not know which county in Florida had a real need for their business and thus their business had better potential to flourish. Another fact is, most of these emigrants target Central Florida because it's less crowded than other metropolitan areas which make the housing market reasonable and nice lifestyle is achievable. For this reason, another study will focus only on Central Florida.

In our project, we attempt to help the new emigrants (both internal and external) by clustering the venue categories together. We will supply the model with all venue categories and all the cities of central Florida. After examining the clustering, we should be able to generate a set of recommendations to the new emigrants which should increase their chances to succeed with their business ideas.

Data Description

In our project we will need to get a list of all the cities in central Florida which we are going to get from Wikipedia, and specifically from : https://en.wikipedia.org/wiki/List_of_municipalities_in_Florida. We are going to retrieve the data and then perform scraping using BeautifulSoup library to extract the information we need. We will then get the co-ordinate of each of Central Florida cities from World map and specifically: https://www.mapsofworld.com/usa/states/florida/lat-long.html. We will join the two data sets together and then prepare a query (or set of queries) to get Venue information from FourSquare. After getting the information from FourSquare, we will prepare the data for clustering. Please keep into consideration that although the model looks similar to the exercises we had for Toronto and New York, our project is different, because, the data supplied to the cluster is Venue Category revenue and not the cities.

Data required for the model to Run

Data	Data Source
Latitude and longitude of Florida	https://www.mapsofworld.com/usa/states/florida/lat-long.html
Venue and frequency per FL county	Foursquare
Cities list	https://en.wikipedia.org/wiki/List_of_municipalities_in_Florida.

Output/Cluster data

After the model generates the cluster, we will merge it and visualize the output to present the best neighborhood for each venue category. The data should look similar to this:

Venue Category	Mostly exist 1								Neighborhood n
	• NUM • Freq • Num o venue/person								

Methodology

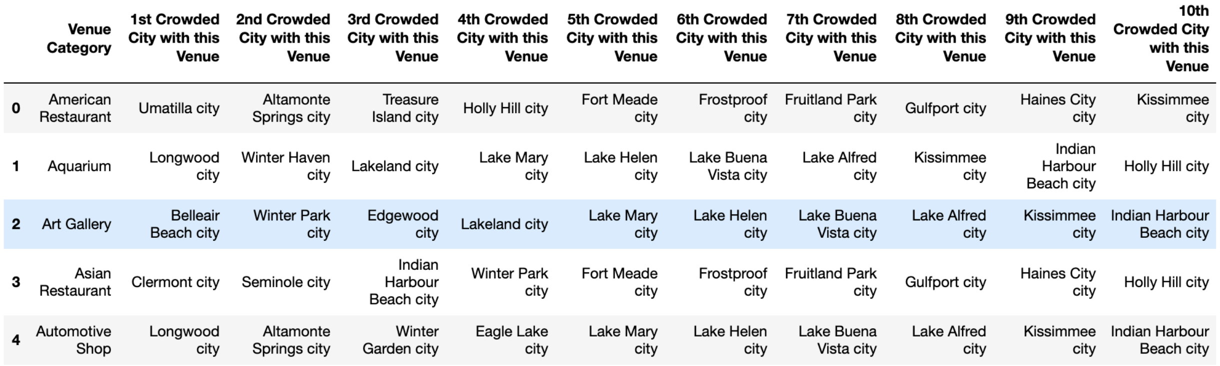
Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

With above data, I used the Kmean clustering technique to resolve the problem. Combine with FourSquare API which provides how many venue categories in different central Florida cities. We used five (5) clusters because five cluster looked reasonable based on the information provided.

Before building the matrix, I have to prepare the required data and apply some data analysis.

```
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(CentralFL_grouped_clustering)
```

CentralFL was grouped for the cluster by the venue category. Remember, we need mainly need the cluster to find the similarities between the categories and not like the other examples we had the neighborhood. That is why we had focused, and grouped our dataset around the venue category. The result set looks like that:



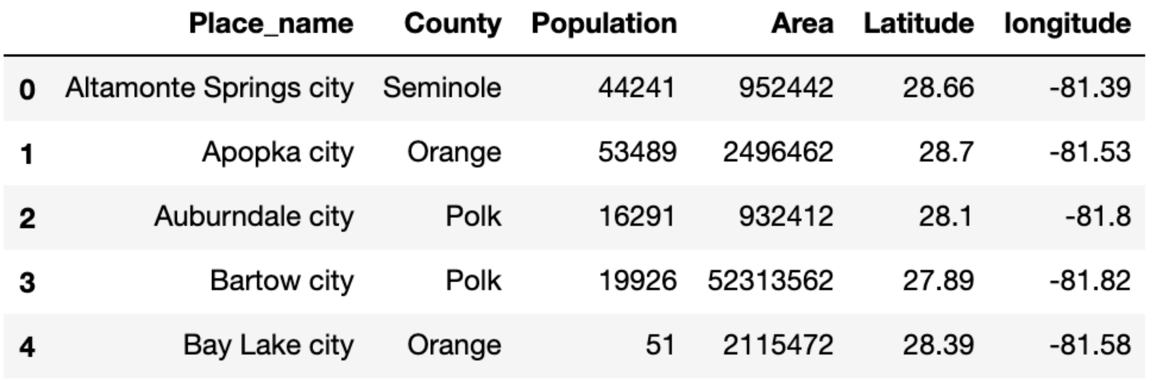
	Venue Category	1st Crowded City with this Venue	2nd Crowded City with this Venue	3rd Crowded City with this Venue	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
0	American Restaurant	Umatilla city	Altamonte Springs city	Treasure Island city	Holly Hill city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Kissimmee city
1	Aquarium	Longwood city	Winter Haven city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
2	Art Gallery	Belleair Beach city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city
3	Asian Restaurant	Clermont city	Seminole city	Indian Harbour Beach city	Winter Park city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Holly Hill city
4	Automotive Shop	Longwood city	Altamonte Springs city	Winter Garden city	Eagle Lake city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city

The above was extracted using the head () method of the data frame.

Data was cleansed based on the following on the following factors:

- The list of cities we got from Wikipedia had many fields that were not needed. The first par of the cleansing was to drop them from the data frame which we created specifically to store the information we received from Wikipedia. Namely the fields are:
    - 'Year' – or when the city was created.

    - 'Label' – City or town
    - 'Government' – how is it presented in the government.
- All the above are out of scope of our study.
- The fields that we kept need some data cleansing, like removing undesired characters and new lines.
- For the cities coordinates, we had to add the word city after each record to make sure the merge between the two data sets [Cities list, and co-orindates are successful)
- The result looked like the below:



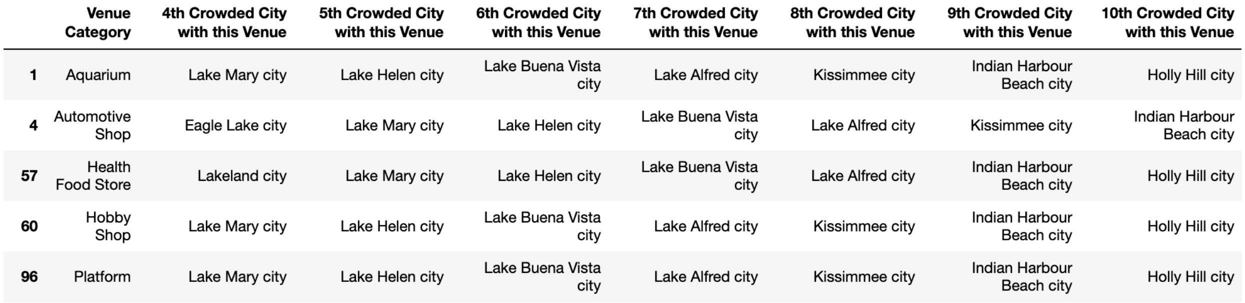
	Place_name	County	Population	Area	Latitude	longitude
0	Altamonte Springs city	Seminole	44241	952442	28.66	-81.39
1	Apopka city	Orange	53489	2496462	28.7	-81.53
2	Auburndale city	Polk	16291	932412	28.1	-81.8
3	Bartow city	Polk	19926	52313562	27.89	-81.82
4	Bay Lake city	Orange	51	2115472	28.39	-81.58

Results

After running the Kmeans model. We got 5 clusters. The clusters are centered around the venues categories. The clustered are here below:

Cluster 1

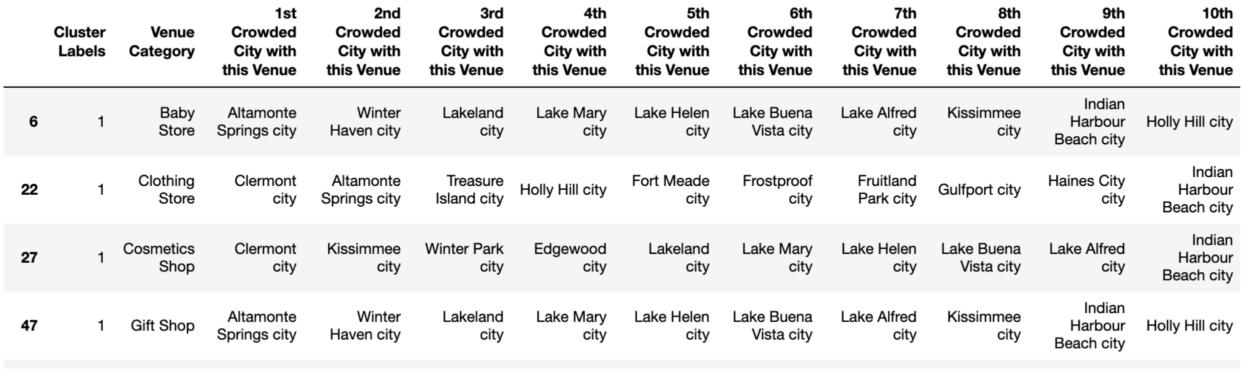
Index 0 but actually the first cluster



	Venue Category	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
1	Aquarium	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
4	Automotive Shop	Eagle Lake city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city
57	Health Food Store	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Indian Harbour Beach city	Holly Hill city
60	Hobby Shop	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
96	Platform	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city

Cluster 2

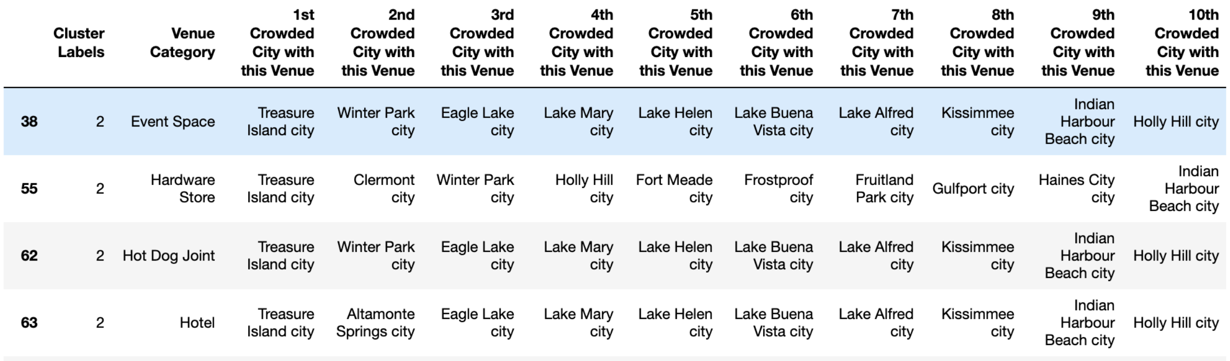
Index 1 but actually the second cluster



	Cluster Labels	Venue Category	1st Crowded City with this Venue	2nd Crowded City with this Venue	3rd Crowded City with this Venue	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
6	1	Baby Store	Altamonte Springs city	Winter Haven city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
22	1	Clothing Store	Clermont city	Altamonte Springs city	Treasure Island city	Holly Hill city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Indian Harbour Beach city
27	1	Cosmetics Shop	Clermont city	Kissimmee city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Indian Harbour Beach city
47	1	Gift Shop	Altamonte Springs city	Winter Haven city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city

Cluster 3

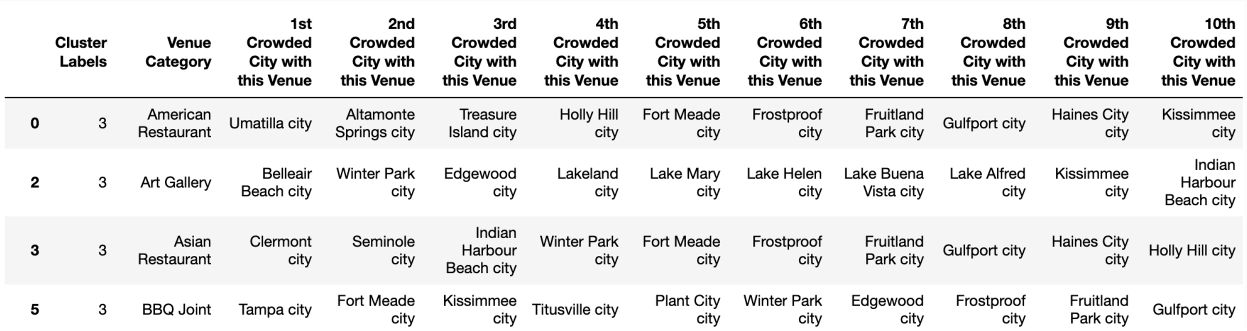
Index 2 but actually the third cluster



	Cluster Labels	Venue Category	1st Crowded City with this Venue	2nd Crowded City with this Venue	3rd Crowded City with this Venue	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
38	2	Event Space	Treasure Island city	Winter Park city	Eagle Lake city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
55	2	Hardware Store	Treasure Island city	Clermont city	Winter Park city	Holly Hill city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Indian Harbour Beach city
62	2	Hot Dog Joint	Treasure Island city	Winter Park city	Eagle Lake city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city
63	2	Hotel	Treasure Island city	Altamonte Springs city	Eagle Lake city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city	Holly Hill city

Cluster 4

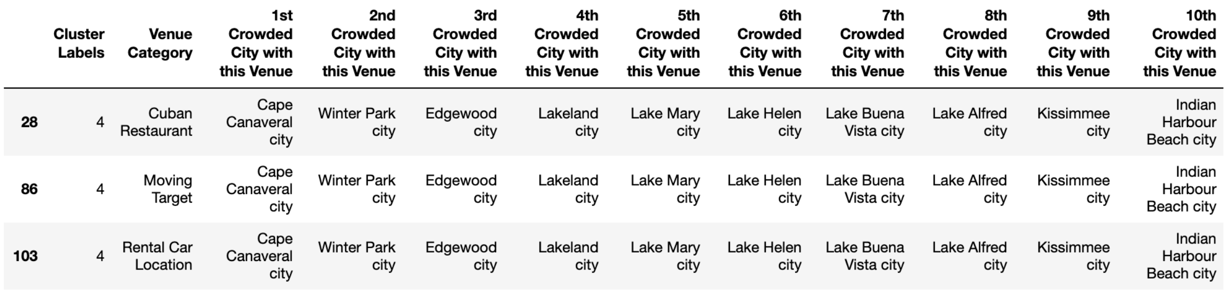
Index 3 but actually the fourth cluster



	Cluster Labels	Venue Category	1st Crowded City with this Venue	2nd Crowded City with this Venue	3rd Crowded City with this Venue	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
0	3	American Restaurant	Umatilla city	Altamonte Springs city	Treasure Island city	Holly Hill city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Kissimmee city
2	3	Art Gallery	Belleair Beach city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city
3	3	Asian Restaurant	Clermont city	Seminole city	Indian Harbour Beach city	Winter Park city	Fort Meade city	Frostproof city	Fruitland Park city	Gulfport city	Haines City city	Holly Hill city
5	3	BBQ Joint	Tampa city	Fort Meade city	Kissimmee city	Titusville city	Plant City city	Winter Park city	Edgewood city	Frostproof city	Fruitland Park city	Gulfport city

Cluster 5

Index 4 but actually the fifth cluster



	Cluster Labels	Venue Category	1st Crowded City with this Venue	2nd Crowded City with this Venue	3rd Crowded City with this Venue	4th Crowded City with this Venue	5th Crowded City with this Venue	6th Crowded City with this Venue	7th Crowded City with this Venue	8th Crowded City with this Venue	9th Crowded City with this Venue	10th Crowded City with this Venue
28	4	Cuban Restaurant	Cape Canaveral city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city
86	4	Moving Target	Cape Canaveral city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city
103	4	Rental Car Location	Cape Canaveral city	Winter Park city	Edgewood city	Lakeland city	Lake Mary city	Lake Helen city	Lake Buena Vista city	Lake Alfred city	Kissimmee city	Indian Harbour Beach city

Conclusion

In this study, I analyzed the possible different clusters for Venue categories in Central Florida. The intention of this study is the help below who are planning to move to Florida or Central Florida specifically to choose the right city from the get-go. Using this study people can either decide which business they should focus on given they know which city they want to live in. Or chose a city to land to and live in if they have a business idea and plan to move to central florida.