MOVING TO ATLANTA, GEORGIA. ANALYSIS OF THE JOURNEY.

Capstone Project: The Battle of Neighborhoods

Introduction

The main problem is the difficulty that people have to take decision when they want to migrate to other state or countries. In this particular case I will migrate to Atlanta, Georgia in some months so creating this analysis will help me to take a better decision selecting a City in Atlanta, rent an apartment, select a good school and university for my daughter and son.

Two important points to consider in the selection is that Fulton County in Atlanta should be used as a cardinal point to start the analysis around the cities. Second point is that the city should be near to Restaurants, clothes places and transportations.

For that reason, in the solution to develop analysis of features for a people migrating to Atlanta and search a best city as a comparative analysis. In that case showing list of venues available by cities which allow to show the bunch of categories to consider when rent a house. The features include median housing price and better school according to ratings, to taking decision. Also provide a plot information with index in schools, universities and house prices in Atlanta cities.

Data Section

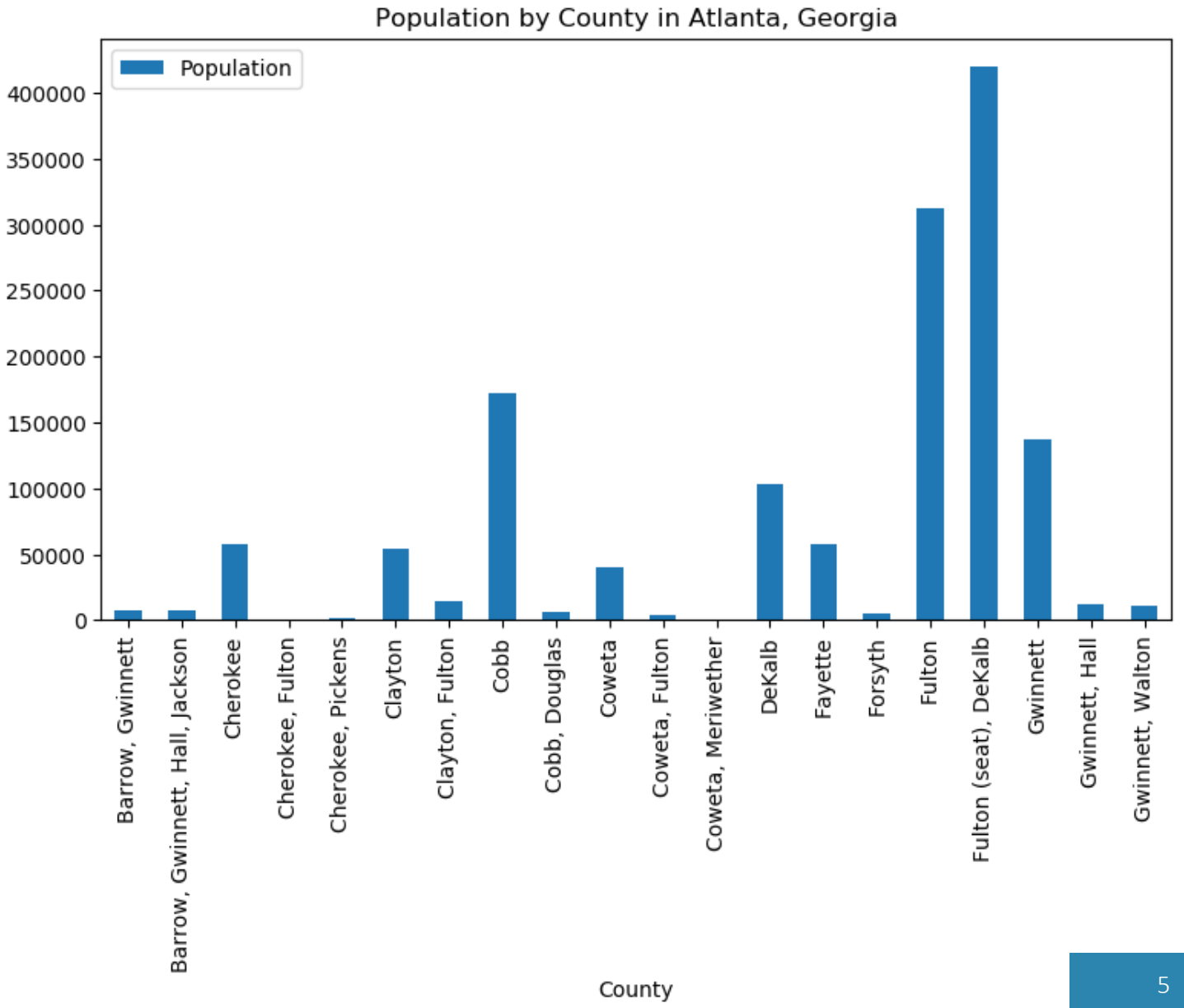
- *1.- Cities and Counties Data:*
- I found the list of Cities and Counties from Wikipedia : https://en.wikipedia.org/wiki/List_of_municipalities_in_Georgia_(U.S._state) . I
- I scrapped the data from the wiki, then cleaned and reduced to be applied creating choropleth map .
-
- *2.- Coordinates for Georgia Cities:*
- I created own table with Coordinates (Latitude and Longitude) using Google Maps. That info will be merged with data from Cities and Counties. [2]
- In this case , since the Geocoder is not allowing to retrieve information related coordinates , we wil create extract coordinates directly using Google Map.
-

- *3.- Foursquare API Data:*
- I used "**Foursquare API**" to extract data from the most common venues of each city in Atlanta. It will help us to provide the options in the comparative analysis. [3]
- In this case we will use the credentials already created to consume the venues.
-
- *4.-School, Colleges, University ranking and price of house rental Data*
- I created own tables for School, Colleges, University Ranking extracting from different websites It will help us to provide ranking analysis. [4]
- There are not too many public datas related to Education and rental prices of apartments. Therefore, I will create own tables to collect that information for our analysis.

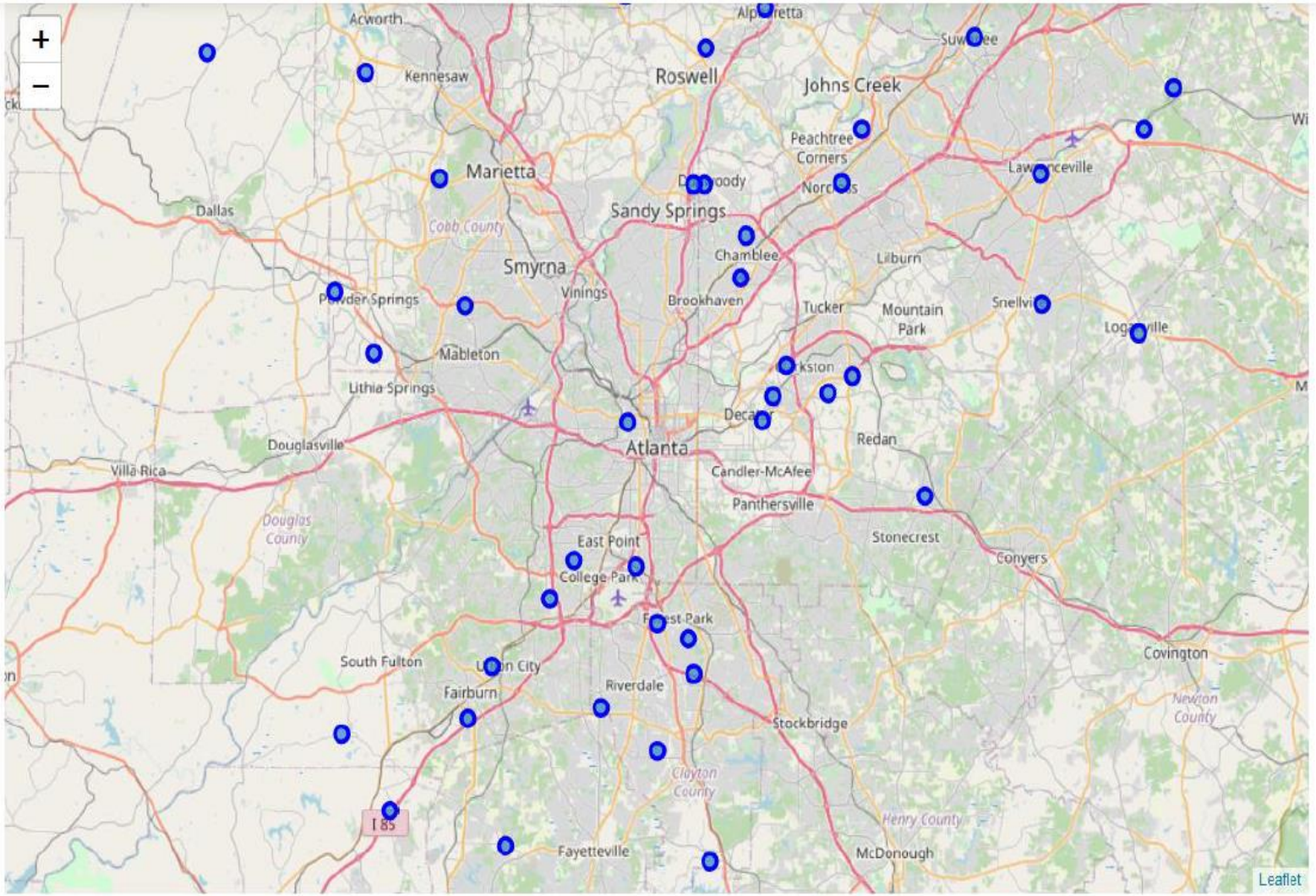
Methodology

To compare the similarities of cities, cities were explored, segmented and grouped into clusters to find similarities between cities in Atlanta, Georgia. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

- First at all , we start collecting cities in “Georgia” and we plot the information so results can give us a view of how the population is currently in Atlanta.



- Using our coordinates from Atlanta and applying our code with folium libraries, we build the map indicating the coordinates from each city in Atlanta and how they are placed to show a view how they are distributed and to have a start picture when comparing with the clusters that will create when K-means will be applied



- We define our Foursquare API credentials and we started exploring cities to obtain their corresponding venues and categorize them.
- With the venues categorized we can display the TOP 10 most common venues in each city that will allow us to identify which venues follow the requirements of the problem.

Display the TOP 10 Most Common venues near by City

```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['City']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
city_venues_sorted = pd.DataFrame(columns=columns)
city_venues_sorted['City'] = georgia_grouped['City']

for ind in np.arange(georgia_grouped.shape[0]):
    city_venues_sorted.iloc[ind, 1:] = return_most_common_venues(georgia_grouped.iloc[ind, :], num_top_venues)

city_venues_sorted.head()
```

	City	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	Acworth	Park	Baseball Field	Garden	Discount Store	Farm	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop
1	Alpharetta	Clothing Store	American Restaurant	New American Restaurant	Coffee Shop	Café	Fast Food Restaurant	Asian Restaurant	Mexican Restaurant	Sushi Restaurant	Deli / Bodega
2	Atlanta	Art Gallery	Coffee Shop	Gas Station	Pizza Place	Gym	American Restaurant	Asian Restaurant	Intersection	Trail	Restaurant
3	Auburn	Discount Store	Gym / Fitness Center	Convenience Store	Pharmacy	Breakfast Spot	Gas Station	Fast Food Restaurant	Train Station	Factory	Farm
4	Austell	Moving Target	Home Service	Food & Drink Shop	Dessert Shop	Business Service	Food Court	Food	Fondue Restaurant	Fabric Shop	Food Service

- Next, we can apply Machine Learning – K-means that help us to make the clustering of the cities . This comparative analysis will be complemented with the visualization of the map and how the cities has been distributed by cluster.
- With clustering applied we have defined 6 clusters which will help us to group the different categories and cities .As we can see in the following evidence :

Cluster Cities

apply k-means to cluster the city into 5 clusters and add in a dataframe.

```
1]: from sklearn.cluster import KMeans
    import sklearn.cluster.k_means_
    km = KMeans(n_clusters=3, init='k-means++', max_iter=100, n_init=1,
      verbose=True)
```

```
2]: kclusters = 6
    georgia_grouped_clustering = georgia_grouped.drop('City', 1)
    kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(georgia_grouped_clustering)
    print(kmeans.labels_[0:10])
    print(len(kmeans.labels_))
```

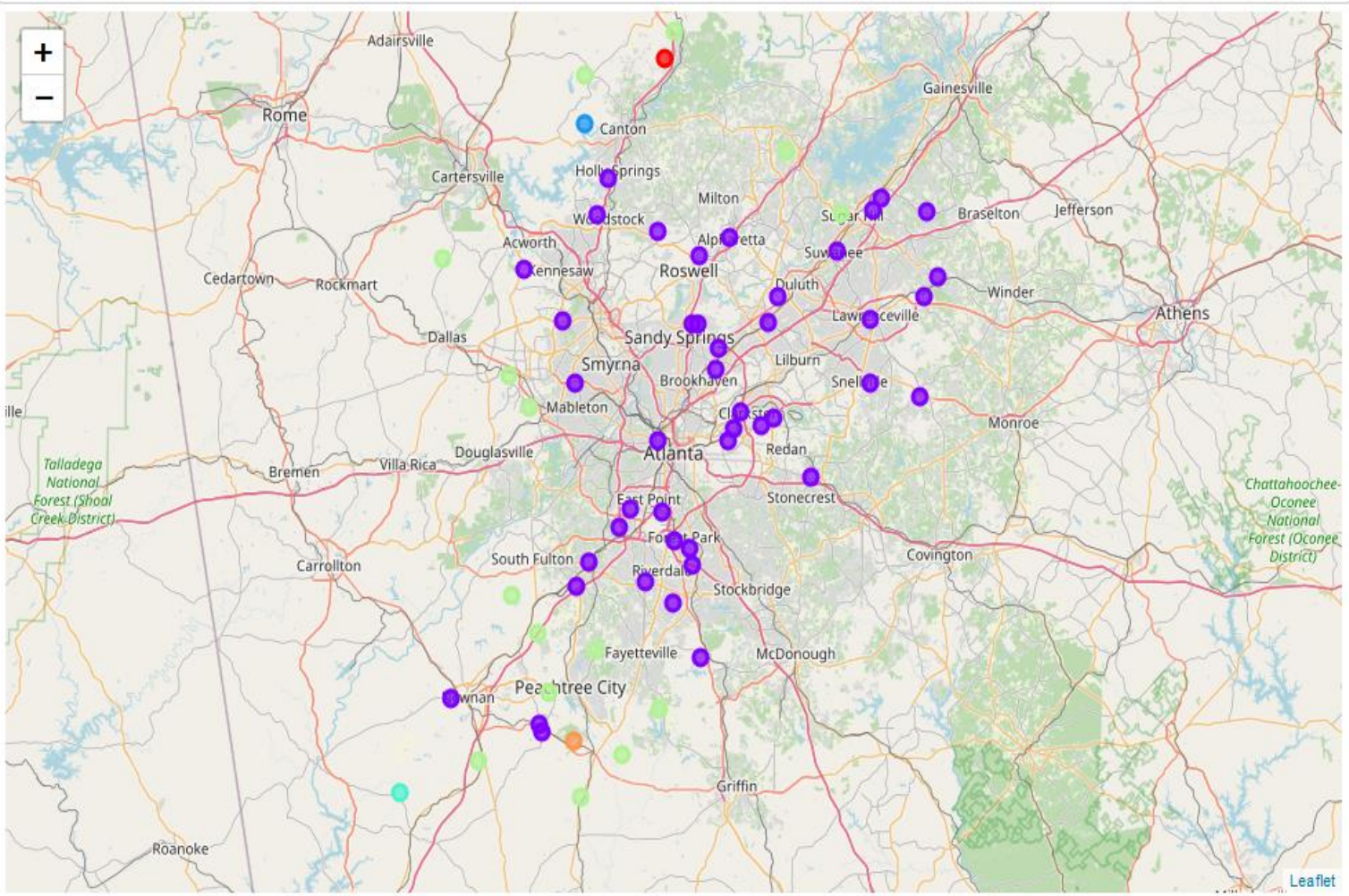
```
[4 1 1 1 4 1 0 1 1 4]
63
```

```
4]: city_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
    georgia_merged = df_atlanta
    georgia_merged = georgia_merged.join(city_venues_sorted.set_index('City'), on='City')
    georgia_merged.dropna(subset=["Cluster Labels"], axis=0, inplace=True)
    georgia_merged.head(10) # check the last columns!
    #print(georgia_merged.dtypes)
```

t[34]:

	index	City	County	Population	Land	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	1	Acworth	Cobb	20425	8.24	34.041467	-84.848962	4.0	Park	Baseball Field	Garden	Discount Store	Farm	Farmers Market
1	13	Alpharetta	Fulton	57551	26.91	34.073757	-84.280203	1.0	Clothing Store	American Restaurant	New American Restaurant	Coffee Shop	Café	Fast Food Restaurant

- Showing our Map with Atlanta cities clustered based in the comparative analysis and similarities of venues categories.



- Finally, we proceed to examine the six clusters defined and for example we see that the Cluster 1 is which group all cities with similarities in venus like restaurants, clothes places. Also after clustering we execute some visualization for schools and universities raitings, and rent prices houses that help us to complement the final analysis

Examine Clusters

Cluster 0

6]:
```
df_cl0= georgia_merged.loc[georgia_merged['Cluster Labels'] == 0, georgia_merged.columns[[1] + list(range
(5, georgia_merged.shape[1]))]]
df_cl0.head()
```

6]:

	City	Latitude	Longitude	Cluster Labels	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 M Comm Venu
6	Ball Ground	34.341301	-84.408877	0.0	Boutique	Construction & Landscaping	Women's Store	Flea Market	Factory	Farm	Farmers Market	Fast F Resta

Cluster 1

7]:
```
df_cl1=georgia_merged.loc[georgia_merged['Cluster Labels'] == 1, georgia_merged.columns[[1] + list(range
(5, georgia_merged.shape[1]))]]
df_cl1.head()
```

7]:

	City	Latitude	Longitude	Cluster Labels	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 Mos Commo Venue
1	Alpharetta	34.073757	-84.280203	1.0	Clothing Store	American Restaurant	New American Restaurant	Coffee Shop	Café	Fast Food Restaurant	Asian Restaura

Results Section

After the analysis done, we can determine that the list of possibilities available in each cluster allow to select and determine which options we have when we want to select a particular place.

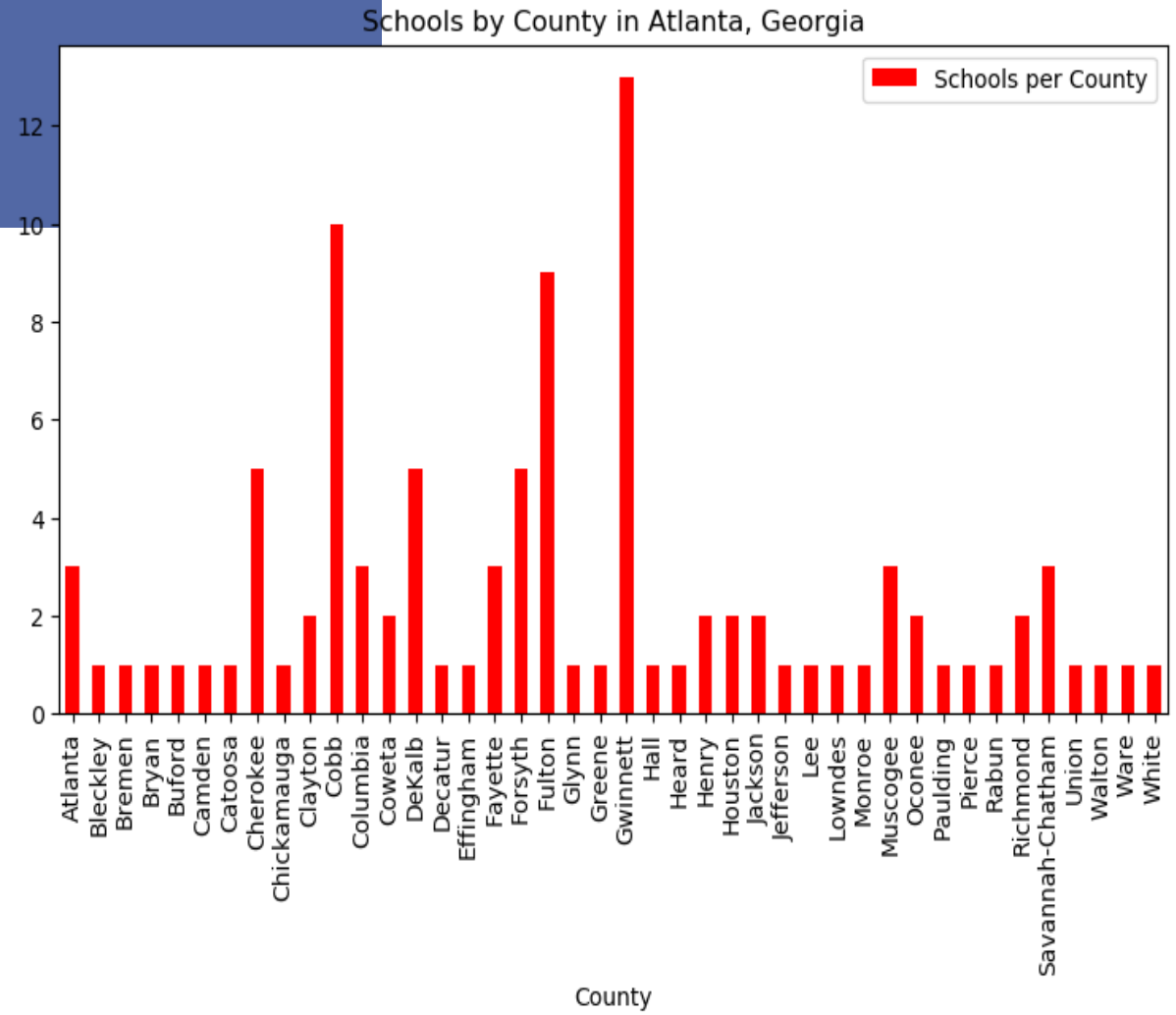
Definitively count with the collection of venues allow to select and consider which will be the most important per each person in the moment to decide what city choose to live.

Finally, our results show that the Cluster 1, follow the initial requirements. All Cities in CLUSTER 1 have Restaurants, Clothes Places, Parks and transportation.

On the other hand, the plots visualize , the view of the ranking of the schools and universities and house prices that will complement the variables to consider in the decision time.

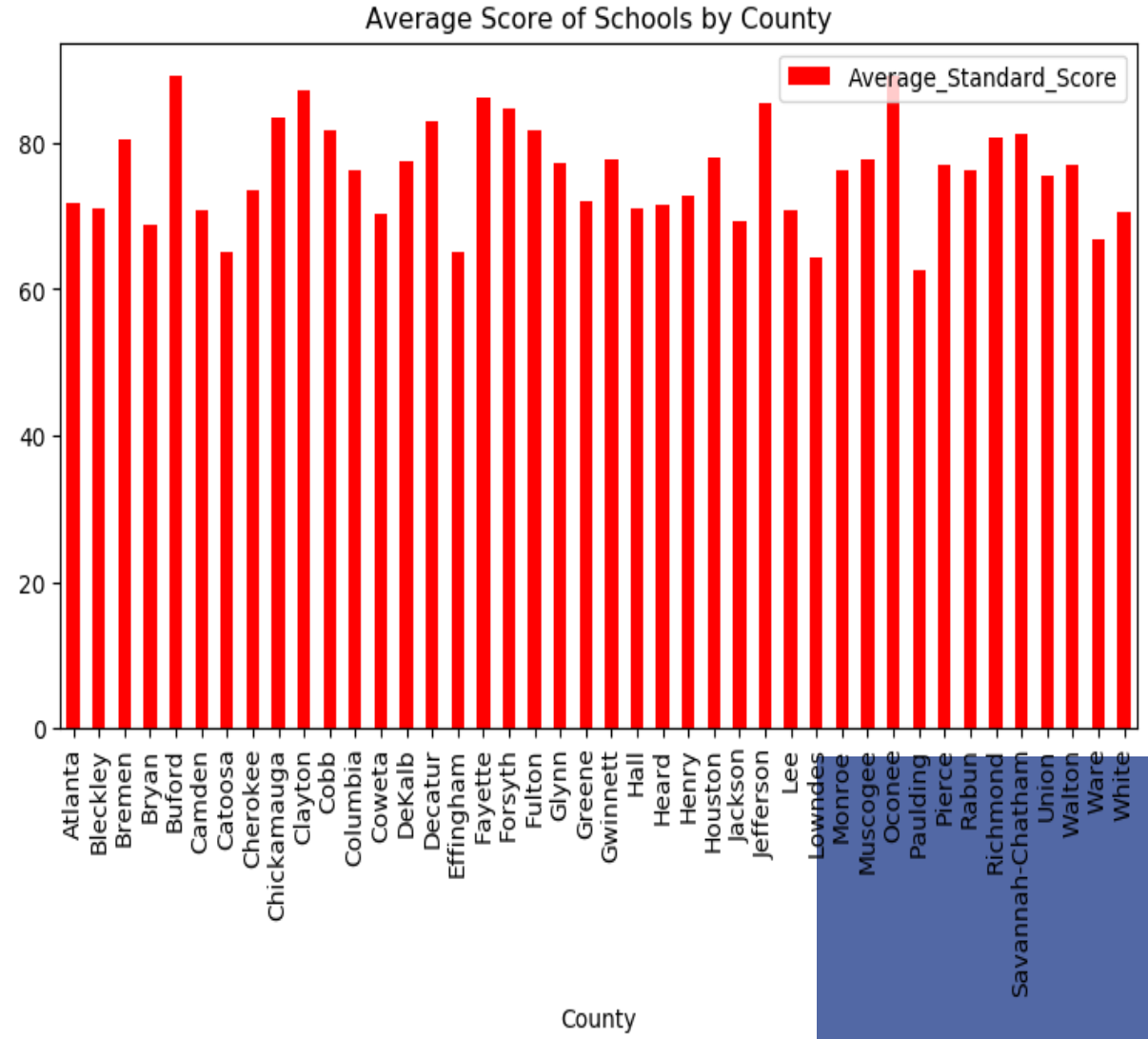
	index	City	County	Population	Land	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
31	259	Kennesaw	Cobb	29783	9.44	34.026161	-84.687084	1.0	Sandwich Place	Mexican Restaurant	BBQ Joint	Pizza Place
38	297	Marietta	Cobb	56579	23.08	33.948371	-84.612413	1.0	Fast Food Restaurant	Breakfast Spot	Trail	Hardware Store
52	416	Sandy Springs	Fulton	93853	37.64	33.943358	-84.352816	1.0	Italian Restaurant	Mexican Restaurant	Pizza Place	Nail Salon

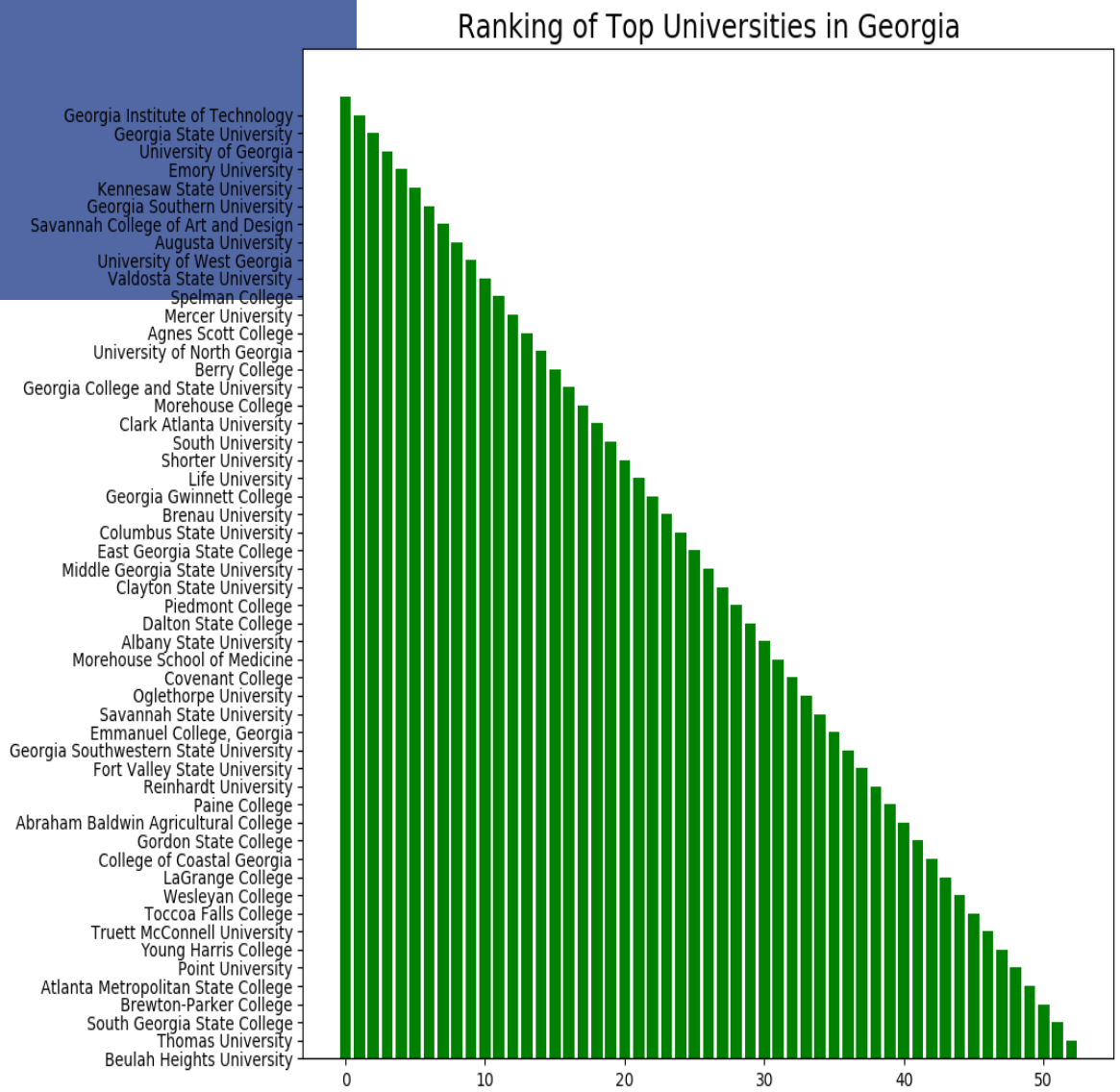
- Reviewing the Cluster 1, we see that 3 cities are the most valuables since are located in Fulton or Cobb county(next to Fulton) , schools in those cities are in the top average of ranking in schools and universities too. The rent price of house is approx. 1200 US$.
- Also, the plots allow to have a view of the ranking of the schools and universities and house prices that will complement the variables to consider in the decision time



Schools by county in Atlanta, Georgia

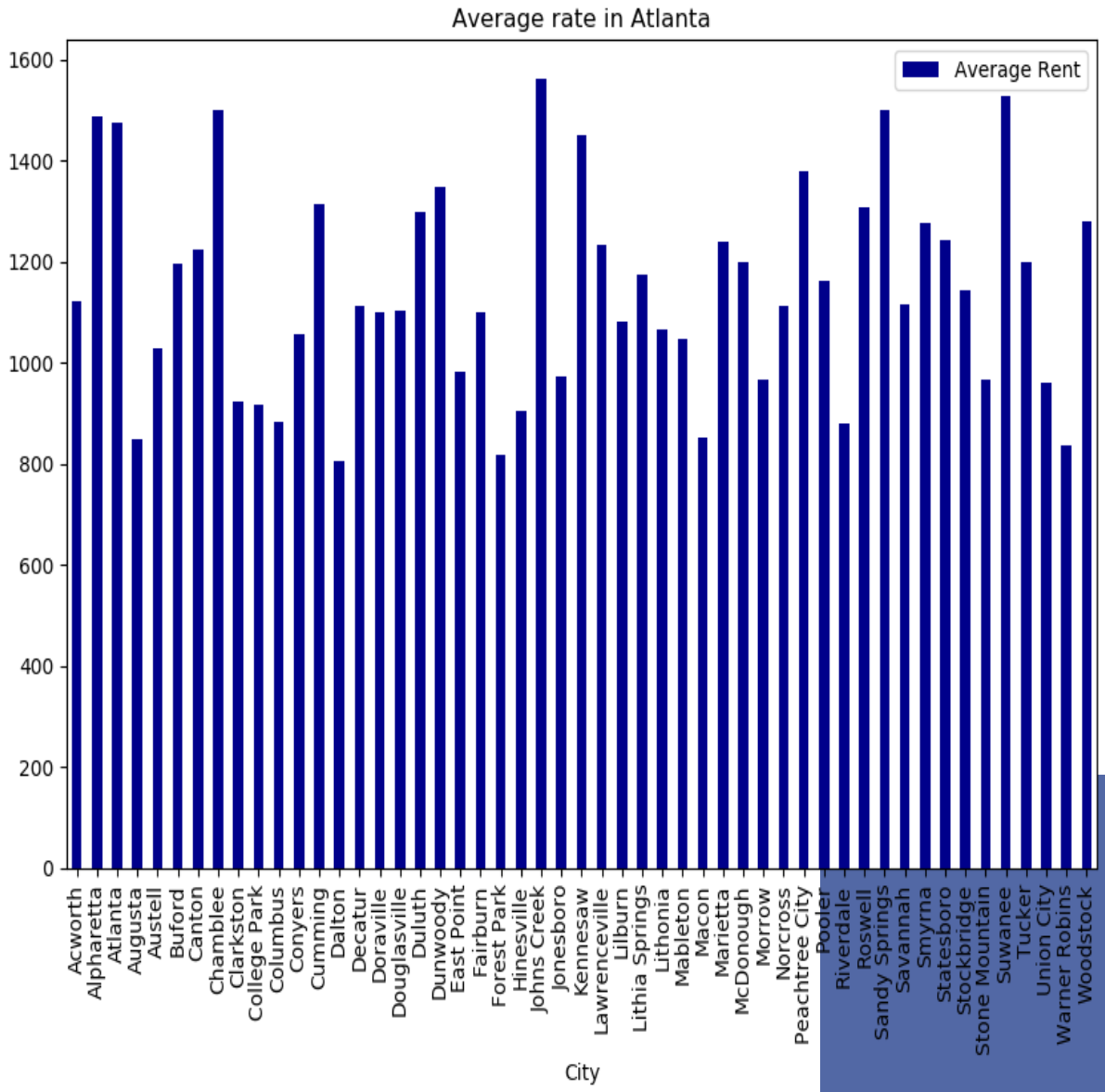
Average score of schools by county





Ranking of top universities in Georgia

Average rent rate in Atlanta



Discussion Section

As we can review results, clustering helps us to split the bunch of options of venues obtained. And together with the information of schools, universities and house prices for rent give a better landscape of opportunities.

I think it is a good start point applying all the knowledge obtained in all this course and I am conscious we know that our analysis can be enriched considering more variables like crime data, traffic in the city, etc. Data that some cases is difficult to obtain but with the topics learned we will be able to improve always.

Conclusion

We can conclude that we have achieved the goal of this project.

Following the analysis learned in the course and applying it using K-means cluster algorithm that helped us to group all the venues provided by Foursquare and make the comparative analysis what help us to determine the options that will be considered in the decision making.

As we show in our Results section, CLUSTER 1 is which follow the requirements defined in the Problem Definition section. This consider initial requirements about the categories of the venues (Restaurants, Clothes places, Parks) and location sitting in Fulton County or next to it. All this analysis was complemented with the plot information.

Particularly, analyzing option in Cluster 1, I most included by Cities as Marietta, Kennesaw and Sandy Springs since those cities are located in Fulton and/or Cobb Counties (last one next to Fulton). And these are in the top average of the school and universities. Rent prices of houses are approx. 1200 US$ which is the average of those cities.

Finally, just say thanks for all what we learn in this path!