**Venues data analysis of Romania and “luxury” clustering**

**A. Introduction**

A.1. Description & Disscusion of the Background

Romania is a country located at the crossroads of Central, Eastern, and Southeastern Europe. With a total area of 238,397 square kilometres (92,046 square miles), Romania is the 12th-largest country in Europe and the 7th most populous member state of the European Union, having approximately 20 million inhabitants. Its capital and largest city is Bucharest. Other major urban areas include: Cluj Napoca, Timișoara, Iași, Constanța, Craiova, Brașov, and Galați.

Romania is divided into 41 counties and the municipality of Bucharest. Each county is subdivided further into cities and communes, which have their own mayor and local council. There are a total of 320 cities and 2,861 communes in Romania.

However, there are some discrepancies between the counties when it comes to living comfort. It is known that the north-west counties (Botosani, Vaslui, Bacau) are some of the poorest in Romania, and counties like Cluj Napoca, Timisoara, Constanta are among the wealthiest. Some of the facts which describes the comfort level of a county would be the venues around, considering a high amount of specific restaurants, hotels and maybe fancy pubs corresponding to a wealthy county, and a small amount of these, finding only supermarkets and simple bars would correspond to a poor county.

Considering this information, we will try to confirm it by handling data about Romania and processing it.

A.2. Data Description

We’ll consider for this study data from different sources, like:

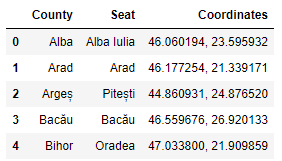
- First-level Administrative Divisions of the Romania from Spatial Data Repository of NYU. The .json file has the borders of the all counties of Romania. This data will be used to create choropleth map of the “comfort level” of the counties of Romania.

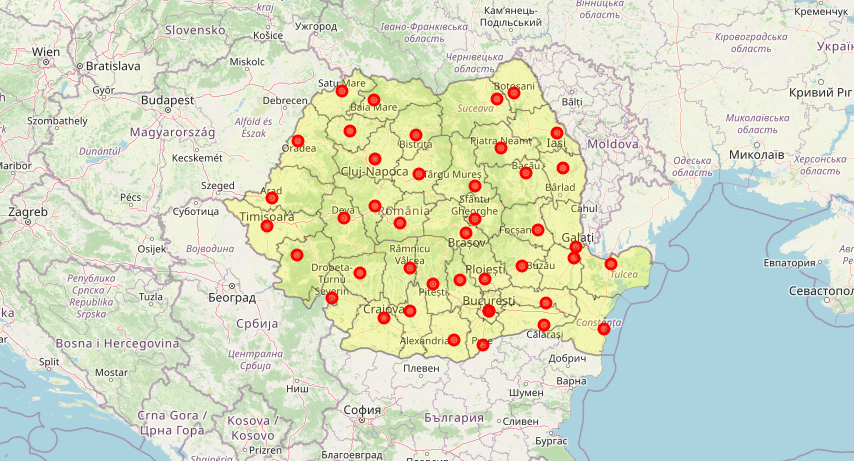
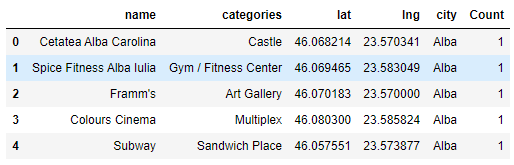
- Forsquare API to get the most common venues of given county of Romania.

- Google Maps, using ‘Search Nearby’ option to get the center coordinates of the each Borough. [5].

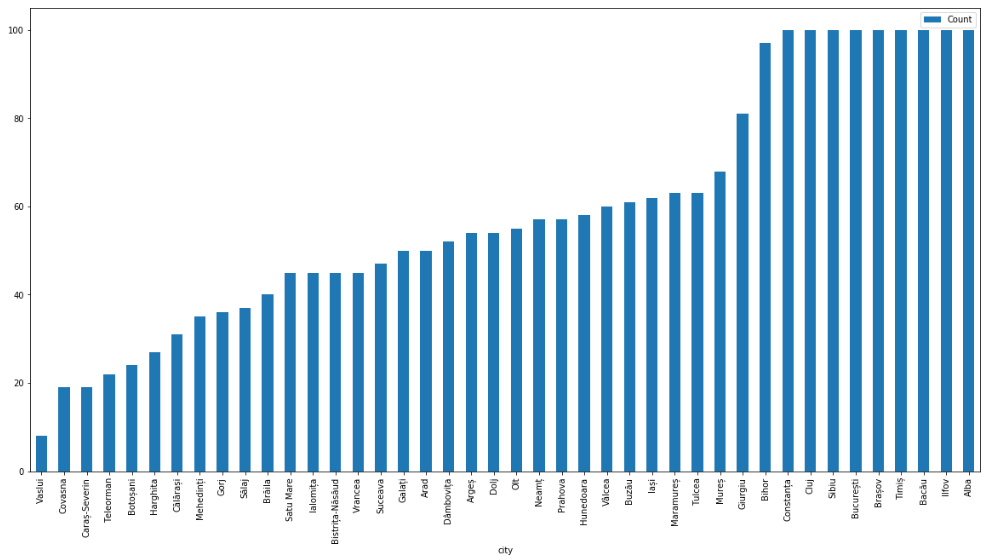
**A. Methodology**

As storage, same repository where the code is located is used for the database (counties.csv file containing the name, the county seat and coordinates). My master data which has the main components *Borough, Average House Price, Latitude* and *Longitude* informations of the city.

Using python’s folium library we can visualize geographic details of Romania and its counties and we can create a map. The main data used here used are the latitude and longitude values of each county seat:

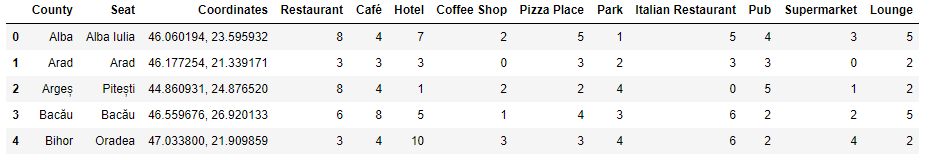
Forwards, we used the Foursquare API to explore the county seats. The configuration used is: limit of 100 venues and the radius of 15000 meters for each seat from their given latitude and longitude information. Below you can observe the head of the list Venues name, category, latitude and longitude information from Foursquare’s API.

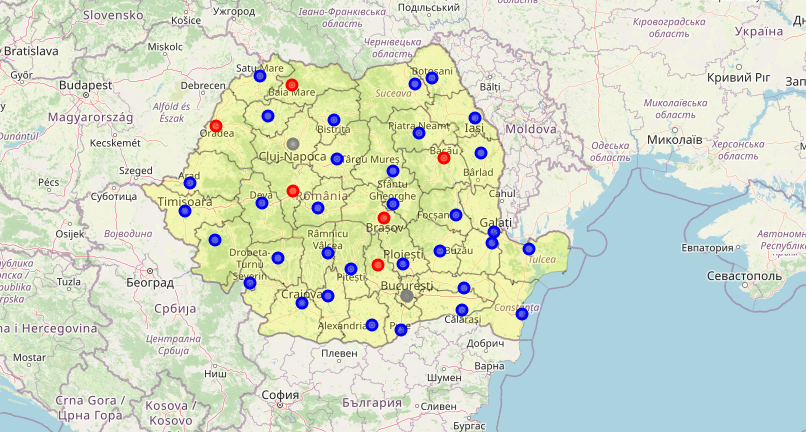
Analyzing the data, we can observe that some of the counties have reached the limit of 100 venues provided by Foursquare’s API, meanwhile some of them struggle with lower than 20 venues. Hence, we have the first insight from the data, the information we already knew: counties considered wealthy (like the capital, Buchares, Brasov, Constanta) have a high number of venues, while counties considered poor (like Vaslui and Botosani) are located in the bottom of the list. One important factor is also that the Foursquare database doesn’t have much data for the poorest counties, and that’s because of the lack of places of interest and tourists.

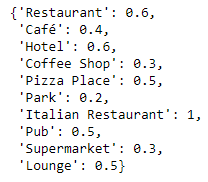
Also, the result doesn’t mean include all the possible results in the counties. Actually, it depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each county (the seat). 

The next step is to wrangle this data in order to make it ready for a clustering algorithm to be used. For this, we’ve counted the amount of items in each venue category, for the whole dataset, sorted them and chose the top 10 venues. After that, a matrix with each county and the amount of times that venue is found in the county is created. This would be a good final data to be fetched to the ML algorithm.

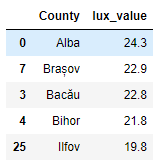
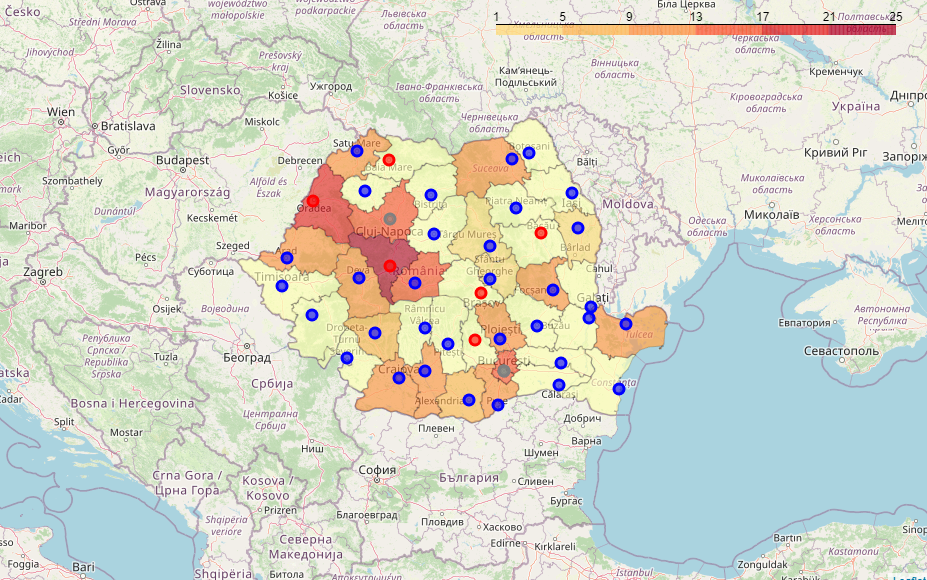




Considering this data, next we’ve used an unsupervised learning **K-means algorithm** to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning. We tried to group the counties in 3 counties.

Next, to gather the ‘luxury’ level, we’ve manually created a weight vector for each of the 10 most common venues:

Calculating the dot product between the amount of venues for each county and the weight vector, we obtain the ‘luxury level’:

In a choropleth map, it would look like this:

Conclusion: