**Clustering European Cities**

1. **Business Problems**

This project aims to cluster the most important European cities according to its size and the people movement around them. This information could be very useful to explain how Covid 19 has affected each of them in the past or predict how it can affect them in the future. The data that has considered relevant to this purpose is the population in the metropolitan area, the number of flight passengers per year and the number of venues of different categories.

1. **Data Description and Exploration** 
   1. **Data Description**

This section cites the data used in the project, justifies the need to analyze it and explains how it has been obtained.

* Population in the Metropolitan Area
* Number of Flight Passengers:
* Number of Venues of different Categories

**Population in the Metropolitan Area**

Population is of course the first parameter to investigate while analyzing a city size. It is important to consider the whole metropolitan area and not only the city population as people that live around the city normally work or spend time in the city, which contributes to the city. For example, in London 13.614 million people live in the metropolitan area while only 3.231, which represents only a 24%. This fact could be explained by the argument that living in London is very expensive so many people who work there prefer living in the surrounding of the city and displace to London every day.

The population in the metropolitan area has been obtained from New Geography [1].

**Number of Flight Passengers**

Another important parameter to study while analyzing a city movement is the number of passengers that fly to the city per year. It could come from people doing tourism around the city, people doing business trips or even people doing layovers.

The number of passengers flying per year has been obtained from Wikipedia [2].

**Number of Venues of different Categories**

The number of venues in the city is a good representation of the city movement. For example, there will be more restaurants, bars or museums in a busier city. This information will be obtained using the Foursquare API [3].

It could be interested analyzing the number of visits that the venues receive every year, however, Foursquare does not provide this information so it will not be analyzed.

As Foursquare limits the number of venues per call to the API to 100 the procedure followed to compare the number of venues in the city will be to calculate the density of venues in a radius of 500 meters in 9 different points in the city center.

Analyzing the type of venues could be interesting to differentiate people living in the city, people visiting the city for tourism or people travelling for business reasons. Therefore, the venues will be counted using the following queries: Restaurant, Bar, Coffee, Museum, Monument and Hotel.

The city latitude and longitude are also necessary to search venues using Foursquare and to plot the map. They have been obtained from [4].

The data used is summarized in the following table.

|  |  |
| --- | --- |
| City | |
| Coordinates | Latitude |
| Longitude |
| Population | |
| Passengers | |
| Venues | Restaurant |
| Coffee |
| Bar |
| Museum |
| Hotel |
| Monument |

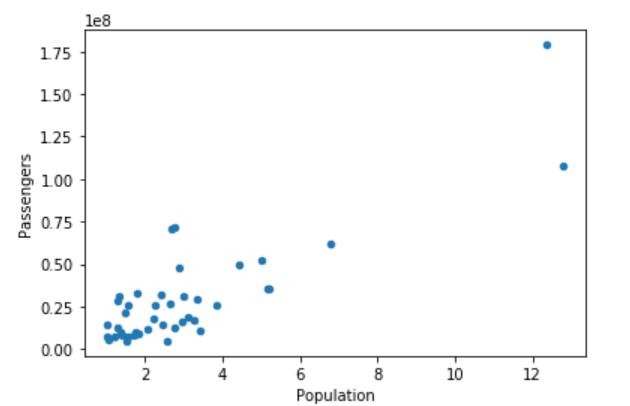
**2.2. Data Exploration**

Once the data has been obtained and cleaned, it is interesting to analyze it to extract some previous hypothesis about it. The following figure shows the first rows of data. It has been normalized so that it can be understood easier.



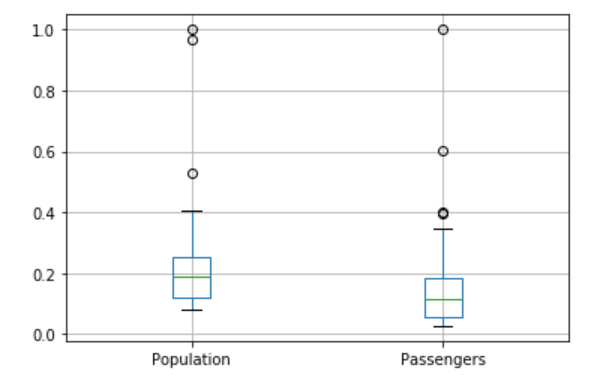
Paris and London are by bar the most populated cities and they are the ones that most flight passengers have. They have many restaurants, bars and coffee shops. However, other cities such as Barcelona and Madrid have more monuments, museums and hotels which could mean that they receive more tourists.

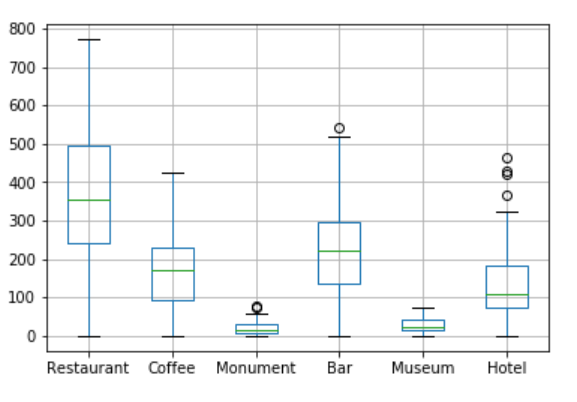
The relation between population and number of passengers is represented in the following figure.



It can be inferred a linear relationship between population and number of passengers. London and Paris can be identified as outlawyers. They are followed by Madrid, Berlin, Milan and Barcelona.

The different columns can be represented using a boxplot, which is shown in the following figures.





1. **Methodology**

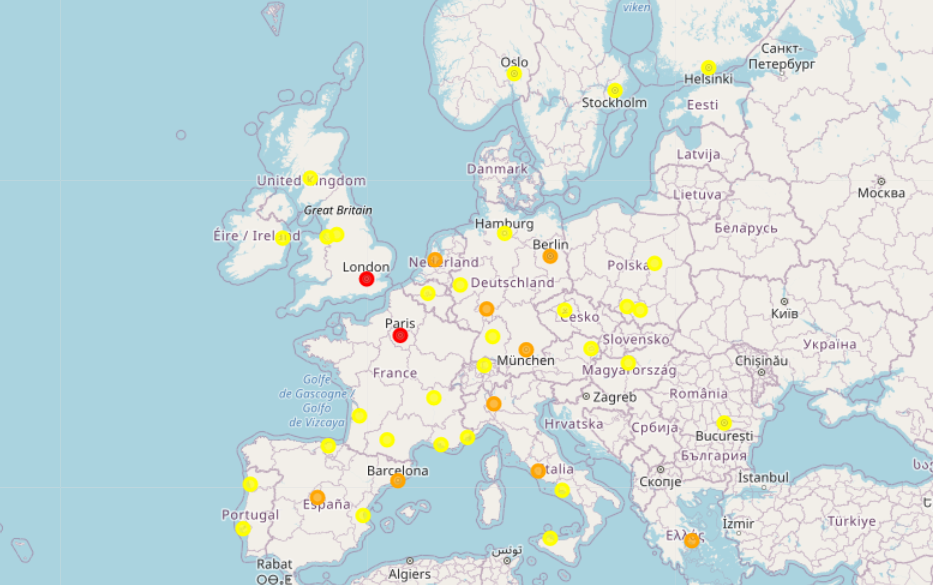
The data previously analyzed will be clustered using K-means. There will be three clusters. It is important to normalize the data before fitting the model. However, not all parameters will be weighed same as every type of venue is not considered as important as population or number of passengers. Therefore, all the venues will be weighed same as population and passengers, which is equivalent to dividing the venues values by the number of categories analyzed 6.

1. **Results**

The results are three clusters of cities, which are represented in the following lines:

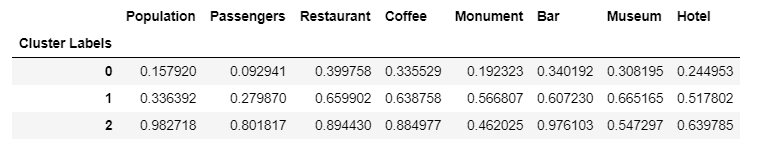
* Cluster 1: London and Paris
* Cluster 2: Madrid, Berlin, Milan, Barcelona, Rome, Athens, Munich, Amsterdam and Frankfurt.
* Cluster 3: Naples, Manchester, Hamburg, Warsaw, Lisbon, Budapest, Stuttgart, Brussels, Katowice, Bucharest, Vienna, Stockholm, Prague, Lyon, Glasgow, Dublin, Marseille, Valencia, Sofia, Dusseldor, Liverpool, Helsinki, Krakóv, Toulouse, Zurich, Oslo, Porto, Bordeaus, Bilbao, Palermo and Nice.

The results are also represented using a map as shown in the following figure.

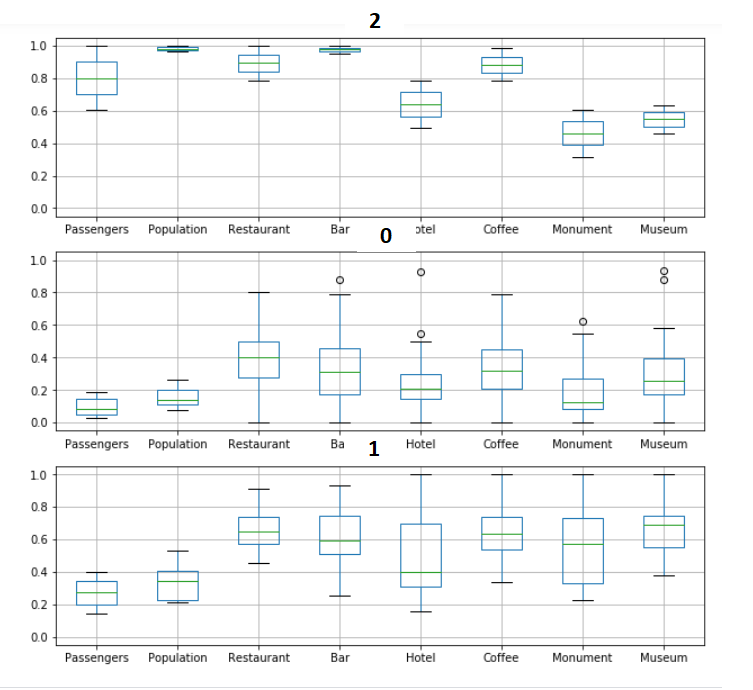


1. **Discussion**

It is interesting to analyze how the clusters have been formed. The following figure represents the mean value of columns for each cluster.



The following figure shows how every parameter is distributed in each cluster.



From the previous table and figure the following observation can be made.

* Cluster 2 has most population and number of passengers. It has most restaurants, bars, coffees and hotels too. London and Paris, which are the most populated cities are in this Cluster.
* Cluster 1 has a quite high number of population and passengers but it is not as high as Cluster 1. It has most monuments and museums. Cities thar receive many tourists such as Madrid and Barcelona are in this Cluster.
* Cluster 0 gathers less populated cities with less venues and flight passenger per year.

1. **Conclusion and Future Lines** 
   1. **Conclusions**

The most important cities in Europe has been clustered into 3 clusters using the method k-means. The parameters population in the metropolitan area, number of flight passenger per year and the number of venues of the categories restaurant, coffee, bar, museum, monument and hotel have been analyzed. It can be concluded that London and Paris are the cities with more people movement. They are followed by other quite big cities and that receive many tourists such as Madrid, Berlin, Milan, Barcelona, Rome, Athens, Munich, Amsterdam and Frankfurt. The rest of cities are smaller and gather less people.

* 1. **Future Lines**

The project could be improved following the steps explained in the lines below.

In the first place, it could be interesting obtain more accurate venues values as follows.

* Calculate the density of venues in larger areas
* Analyze other venues such as offices, schools or universities
* Compare Foursquare results with Google Maps
* Analyze the number of visits

In the second place, the project could be improved by analyzing other methods

* Use different algorithms
* Use different number of clusters

1. **References**

[1] New Geography. Major Metropolitan Areas in Europe. <http://www.newgeography.com/content/003879-major-metropolitan-areas-europe>

[2] Wikipedia. List of Busiest Airports in Europe <https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_Europe>

[3] Foursquare API. <https://developer.foursquare.com/>

[4] Geopy Geocoders