**Allocation of a starting point for guided visits to Madrid’s monuments and museums**

**Description of the problem:**

The objective of this studio is to select a set of coordinates to serve as a starting point for guided visits aimed at tourists wanting to explore the monuments of Madrid (Spain).

These tours will be special though. Upon arriving to the starting point, each group of tourists will pick one of three sets of monuments to visit (the best way to group the monuments will also be calculated in this notebook). Thus, before calculating the starting point for the tour (which will be the final goal) we have to cluster the monuments in three groups.

Note that the particular path for each of the tours is not an objective of this study, and it will not be calculated.

**Description of the data:**

For this study, we will need to main sources of data:

On the one hand, we will need a map of Madrid to help us visualize the results of the study and to verify (with the naked eye) that they are reasonable.

On the other hand, we will need the coordinates of Madrid's most centrical monuments. For practical reasons, we will limit the search to the thirty monuments closest to the center of the city.

The aim is to cluster the monuments into three groups. After that, we should calculate the centroid of each group to serve as reference for the calculation of the starting point. And finally, we should be able to calculate a global centroid for the three previously calculated points.

Note that there are simpler ways to calculate the centroid, this one was selected as a way to showcase the knowledge acquired during the course

**Methodology:**

Before acquiring the coordinates of the monuments from foursquare, a study was conducted to ensure that there were enough registers of monuments in foursquare for the city of Madrid. It turned out that there were only 5 registered venues with the category of monuments in a 5 km radius from the center of the city. Thus, it was decided to supplement the monuments category with instances of museums of the capital.

Having an acceptable number of venues, the data was formally stored on a dataframe and manipulated to suit the requirements of the task at hand.

“The company” decided that three separated tours were enough, so three different clustered were created using the k-means method.

With the tours clearly defined, the final step was to calculate the meeting point for the tourist before the beginning of whatever tour they choose. It was calculated using the method of the centroid, using for the calculation the centroid of each of the clusters, to avoid influencing the calculation with the different number of venues in each custer.

**Results:**

As we expected, our starting point has ended up being near the very center of the city.

Though the calculation of the centroids for each cluster ("each tour") helped to mitigate the weight of the two more clouded clusters, the starting point leaves the third cluster in clear disadvantage over the other two.

Leaving aside the clusterization process, the result is a little bit disappointing. In the conclusions, I will try to delve a little bit more on what could have been done to avoid a situation like this.

**Conclusion:**

In retrospect, some other approaches may have yielded better results:

For sure, one of the better alternatives would have been to implement a pathing algorithm to create an optimal route for each cluster, and pick the starting point according to the coordinates of the start of the routes.

Another good alternative would have been to select (leaving aside the pathing problem) the most centric point of each cluster and calculate the centroid based on those three points.