# Group Assignment #1

## Machine Learning I

**PROFESSOR ALVARO JOSÉ MÉNDEZ LÓPEZ**


## Group D

Adilet Gaparov
Amanda Marques
Benjamín Chumaceiro
David Facusse
Salim Aouzai
Tingting Sun


Dearest Minister of Economy,

First of all, thank you for trusting our group with your request: to find which Spanish provinces, out of 52, are similar based on 28 variables to allow a reduced set of policies to several provinces to be applied simultaneously.

We carefully reviewed the data you shared with us and we conducted the segmentation analysis on it. The analysis led to the identification of 6 clusters and the most significant industry of each were selected to label, describe and compare them. Afterwards recommendations were provided, however, do not hesitate to contact us to discuss more in detail each cluster and potential policies.

Cluster 0 is characterized with small population, low unemployment rate and high proportion of population working on the energy industry. Cluster 1 has the highest population, highest unemployment rate and a greater percentage of people working on textile. Cluster 2 is characterized for discrepancy between Tourism Index and Restaurants & Bars Index, as well as proximity to water. Cluster 3 takes the lead on the mining industry. Clusters 4 and 5 are Madrid and Barcelona that were considered as outliers and analysed separately.

**Approach to analysis**

Data from 52 provinces were stored in an Excel file with 28 different features. They were:

1. 15 features, describing the percentage of population working in a specific industry for that province. The industries are: Energy & Water, Mining, Metal, Manufacturing, Building/Construction, Agriculture-Food, Textile, Pharmaceutical, Durable Good Wholesale Trade, Inter-Industry Wholesale Trade, Other Inter-Industry Wholesale Trade, Retail Trade (Food), Retail Trade (Non-food), Other Retail Trade.
2. 5 features, describing market indices: Wholesale Trade Index, Retail Trade Index, Restaurants & Bars Index, Tourism Index, Global Economic Activity Index
3. 8 features, describing socio-demographic data: percentage of male population, total population, percentage of foreign residents, unemployment rate, population growth rate (average 2004-2009), ADSL service per 1000 inhabitants, number of cars per 1000 inhabitants, number of bank branches per 1000 inhabitants.

After the recognition of the dataset, the approach for the analysis was defined based on the variables that are significant for our segmentation model. We have run the Pearson correlation analysis among features and defined the list of features that are highly correlated with each other (more than 0.7, *Annex -Figure 1. Pearson correlation matrix (implemented in R)*). Correlation analysis revealed that all market indexes are highly correlated with total population. We therefore decided to use only the total population in our clustering analysis.

Given the small size of the dataset and the nature of variables (numeric), we decided to try two clustering algorithms: K-Means and Agglomerative Clustering.

On the first approach, all 15 industry-related attributes and 8 socio-demographic attributes were used to run the clustering. However, when implementing different number of clusters (from 3 to 10) two outliers appeared: Madrid and Barcelona. These two provinces were considered as outliers as they are the two biggest cities in Spain population wise - the difference with the third largest city in terms of

population is approximately 3M habitants. However, after some deeper research, it is clear that the same economic policy cannot be applied to them. In cultural and political terms, Madrid and Barcelona have been historically opposite poles. Therefore, and in order to avoid bias due to population only, we prepared the data and excluded them to different clusters to be treated separately.

Second, after removing Madrid and Barcelona from the original dataset, we re-run both algorithms with the same attributes and concluded that the optimal number of clusters in terms of model quality (Silhouette - 0.55/Inertia - 4.448) and labels quality were 4.

Finally, having the benchmark models to compare with, we tried to optimize the models by analyzing the contribution of each feature. We figured out that from socio-demographic attributes only 3 were relevant to our analysis: percent of foreign population, total population and unemployment rate. Among 15 industry-related attributes, 6 were highly correlated so we moved them from clustering attributes to profiling attributes.

The final list of 12 attributes used in clustering is: total population, percentage of foreigners, unemployment rate, Energy & Water, Mining, Metal, Manufacturing, Agriculture-Food, Durable Good Wholesale Trade, Inter-Industry Wholesale Trade, Retail Trade (Food) and Other Retail Trade. The highest Silhouette value was achieved by K-Means and is equal to 0.550.

To have all attributes on the same numerical scale (from 0 to 1), we applied Min-Max normalization to total population.

**Findings and Recommendations:**

Please refer to a more detailed and graphic explanation on the technical annex for all clusters.
For a visualization of the clusters on the map, refer to Annex, *Figure 2. Visualization of clusters on the map.*

**Cluster 0 (purple color on fig.2 in the annex):** This cluster is characterized by the **energy** industry**.**
In this cluster, provinces have the smallest population (63% smaller than average) and the lowest unemployment rate (15% smaller than average). They are also characterized by a high proportion of energy industry and building activities with an average of 6.29% in this sector. In addition, this cluster has the lowest population growth and global economic activity index between clusters. Our recommendation for this cluster is to incentivize citizens living in cluster 1, 4 and 5 to move to cluster 0, in order to further develop the economy in those provinces and also decrease the unemployment rates in the country overall.

Provinces (20) that fall under this cluster are: Álava, Albacete, Ávila, Burgos, Ceuta, Cuenca, Cáceres, Guadalajara, Huesca, La Rioja, Lleida, Lugo, Melilla, Ourense, Palencia, Salamanca, Segovia, Soria, Teruel and Zamora.

**Refer to annex:** *Figure 3. Unemployment Rate versus Global Economic Activity Index by cluster*

**Cluster 1 (yellow color on fig.2 in the annex):** This cluster is characterized by the **textile** industry**.**
Those are the provinces with the largest population after Madrid and Barcelona and the highest unemployment rate among all clusters. They also have a higher proportion of population in the textile industry (111% greater than average), manufacturing and non-food retail trade. For those provinces, we would decrease taxes or incentivize the textile and retail-non food industries in order to foment

companies growth and decrease unemployment rate. There should also be considered training for unemployed who could potentially move to provinces in Cluster 0 to find a job.

Provinces (4) that fall under this cluster are: Alicante, Málaga, Sevilla, and Valencia.

**Refer to annex:** *Figure 3. Unemployment Rate versus Global Economic Activity Index by cluster*

**Cluster 2 (blue color on fig.2 in the annex):** Characterized mainly by unusual discrepancy between **Tourism Index** and **Restaurants and Bars Index**, as well as proximity to water.
It has the second biggest population among clusters without considering Madrid and Barcelona (third largest if including those cities) and it is the cluster with more activities related to other group of retail. The fact that it is closer to the sea makes this group geographically different. In addition, another unique characteristic of this cluster is that the average index of tourism is considerably higher (3,648) than the average index of restaurants (2,455), whereas for other clusters both indexes were about on the same level. In order to tackle this and maximize the business opportunity we would suggest the promotion of restaurants and food retailers near the beach.
Provinces (11) that fall under this cluster are: Asturias, Baleares, Cádiz, Granada, La Coruña, Las Palmas, Murcia, Pontevedra, Santa Cruz de Tenerife, Vizcaya, Zaragoza.

**Refer to annex:** *Figure 4. Clusters by indexes.*

**Cluster 3 (green color on fig.2 in the annex):** Characterized mainly by the **mining** industry.
In this cluster, provinces have the second lowest population and relatively higher proportion of retail-food (average 3.68%). When compared to other clusters, those provinces are more focused on the mining industry - its population is working on this industry, on average, 13.67% more. In addition, this group is the most balanced between unemployment rate and global economic activity.

The industry that is falling behind on this cluster is the retail of non-food (2.66% smaller than the average in the country). Considering those cities are already on a balance level of employment and global economic activity, our recommendation would be to expand this industry further.

Provinces (15) that fall under this cluster are: Almeria, Badajoz, Cantabria, Castellón, Ciudad Real, Córdoba, Girona, Guipúzcoa, Huelva, Jaén, León, Navarra, Tarragona, Toledo, Valladolid.

**Cluster 4 (black color on fig.2):** Characterized mainly by the **largest** population.
This cluster consists of only Madrid. Being the Spanish capital and the third largest city in Europe, the industries where most of the population is concentrated is textile and manufacturing. It is also remarkable the amount of foreigners living in Madrid - making up to 15% of the population. Our recommendation for Madrid is to foment the inclusion of immigrants in the Spanish economy and also incentivize them to move to cities on Cluster 0, so population is more equally distributed.
**Refer to annex:** *Figure 3. Unemployment Rate versus Global Economic Activity Index by cluster*

**Cluster 5 (red color on fig.2 in the annex):** Characterized mainly by the second **largest** population.
This cluster consists of only Barcelona. Even though its main industries are similar to the ones in Madrid, Barcelona has historical and cultural heritage that differs from the rest of the big cities in Spain. Its rate of visitors has suffered a decrease due to the world financial crisis of 2007 and 2008. Differently from Madrid, our recommendation would be to encourage the development of an entrepreneurship environment.

## Annex

Detailed graphical and support analysis:

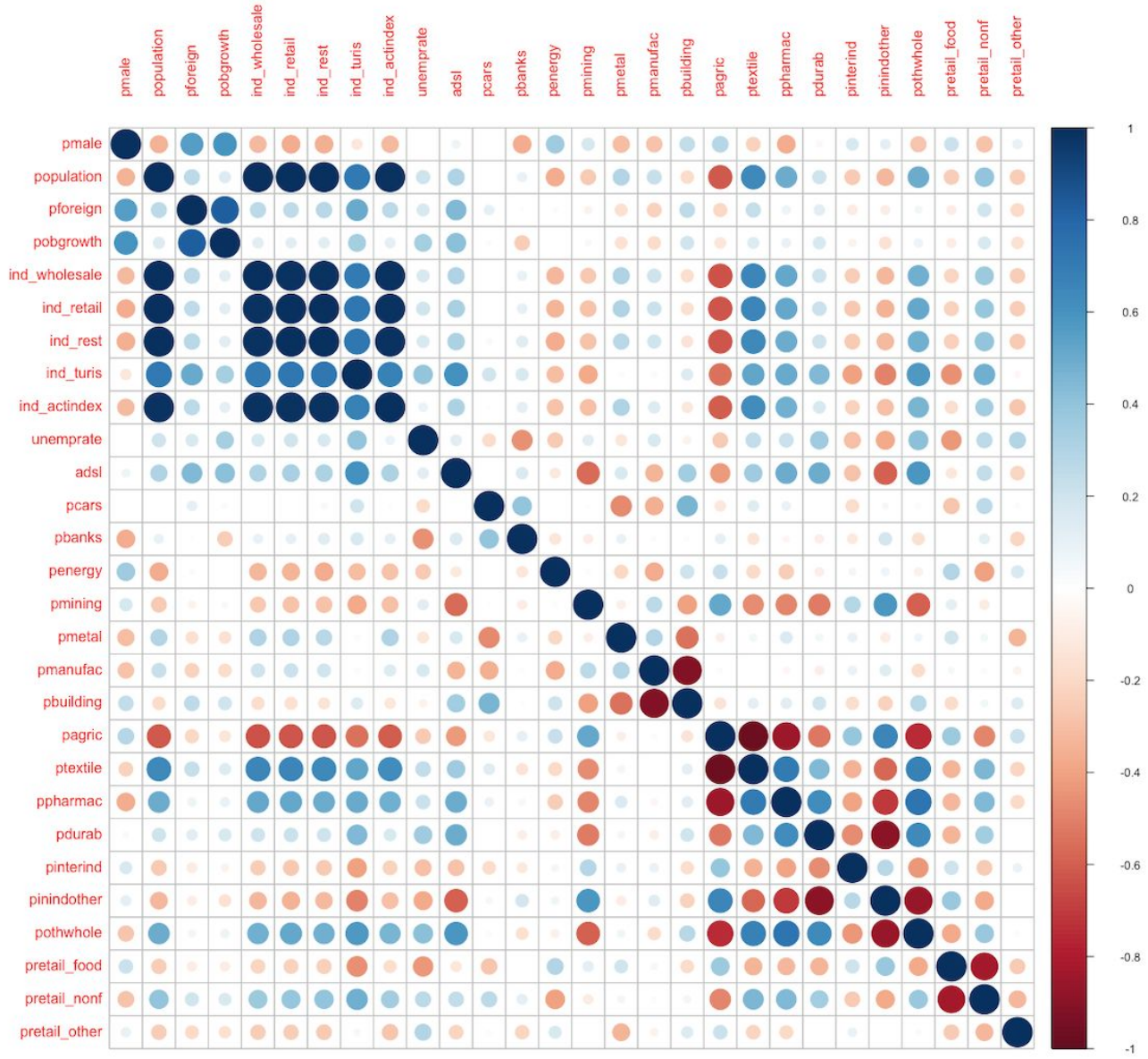


**Figure 1. Pearson correlation matrix (implemented in R)**

High positive correlation between total population and Indexes (upper left corner). High negative correlation between industry-related attributes (lower right corner).

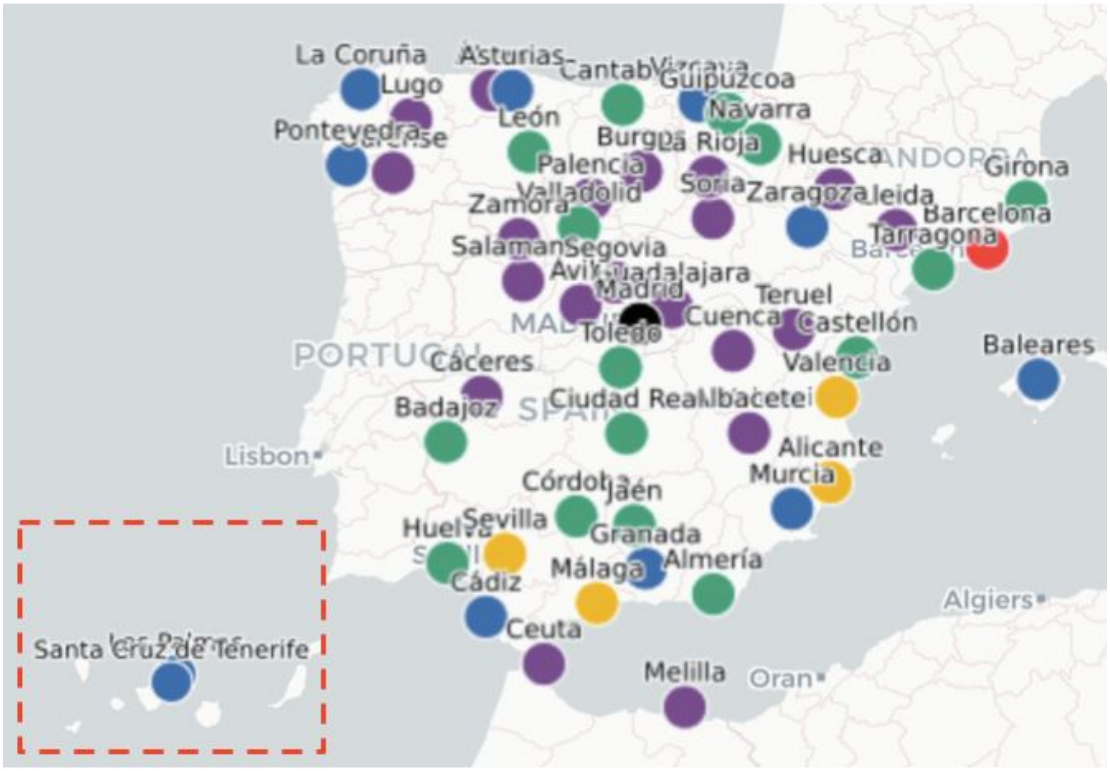


**Figure 2. Visualization of clusters on the map**

Cluster 0 -

Cluster 1 -

Cluster 2 -

Cluster 3 -

Cluster 4 -

Cluster 5 -

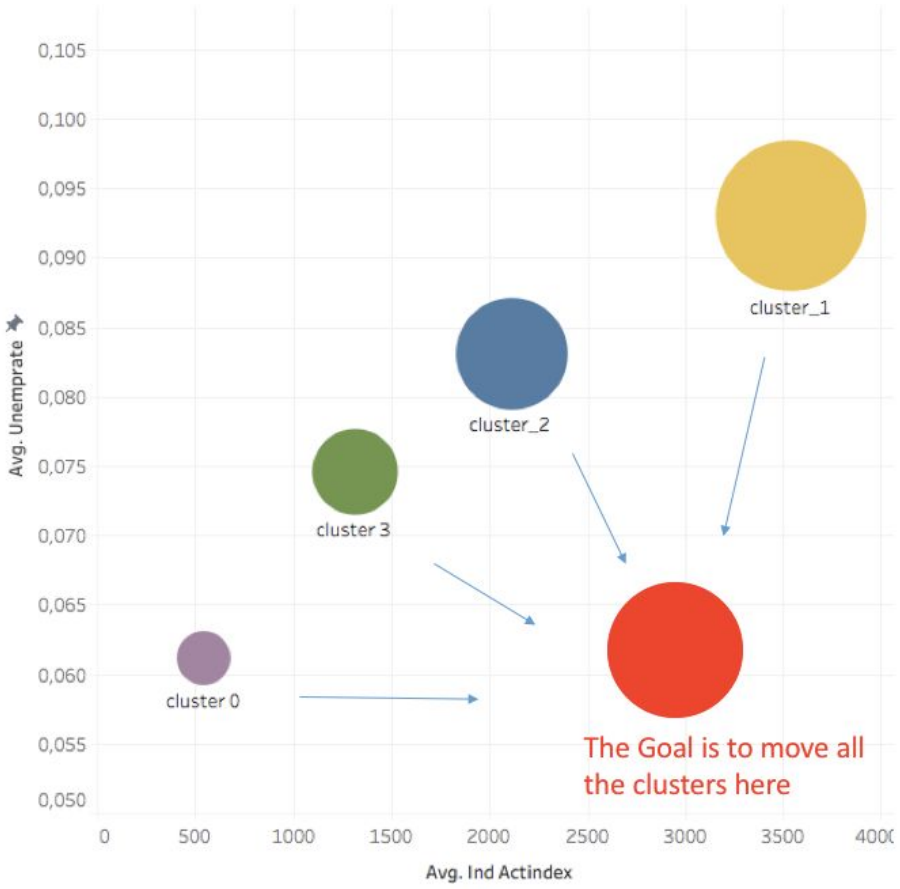


**Figure 3. Unemployment Rate versus Global Economic Activity Index by cluster**

Positive correlation between the unemployment rate and Global Economic Activity Index. Size of the circle represents the average size of the population.

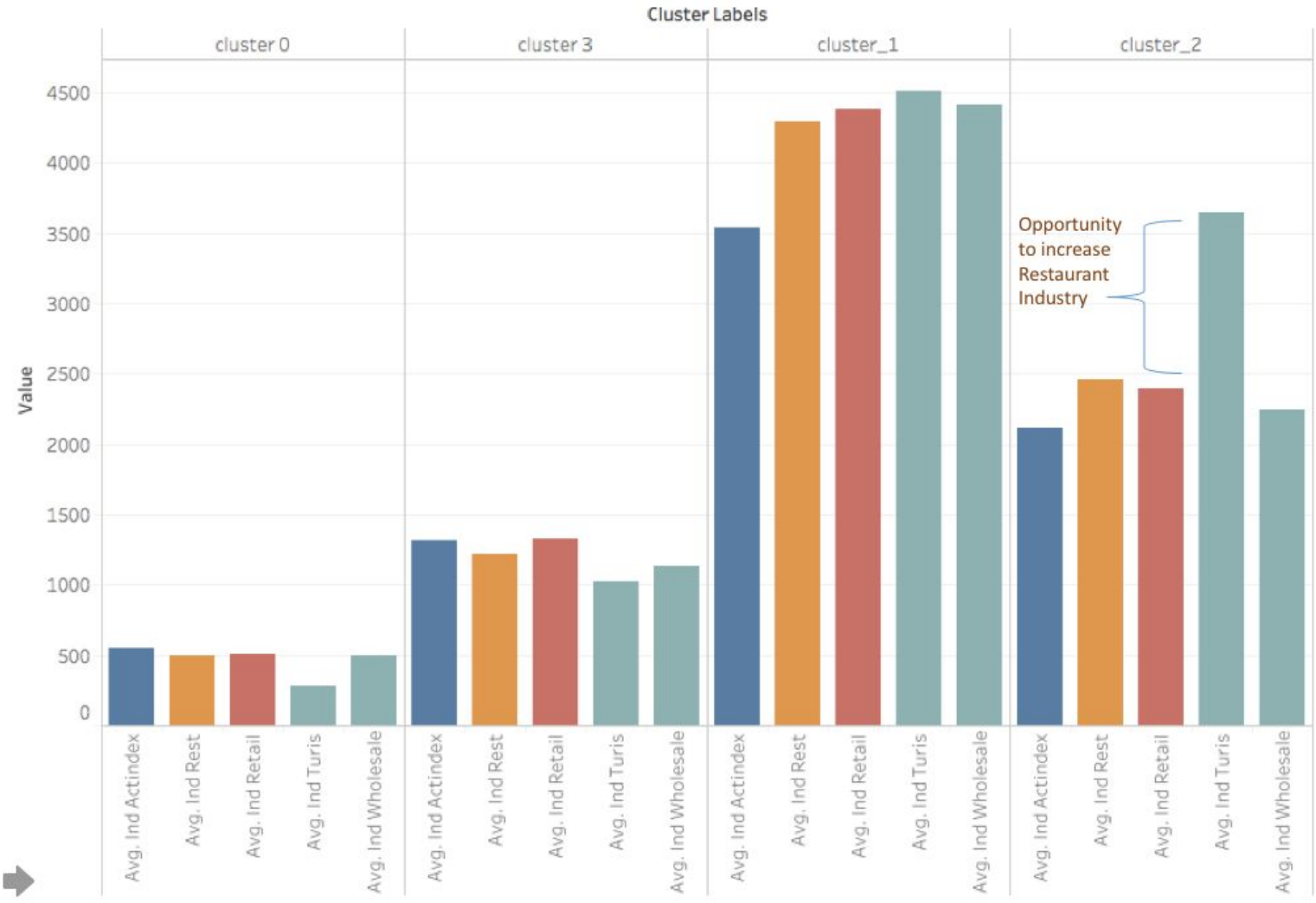


**Figure 4. Clusters by indexes.** Opportunity to improve restaurant and food consumption in tourist areas.

As observed in the graph above, the average index for tourism and Restaurant had a discrepancy in value only on cluster_2. Based on this, we spotted an opportunity to increase the incentives for restauration on provinces in this cluster.

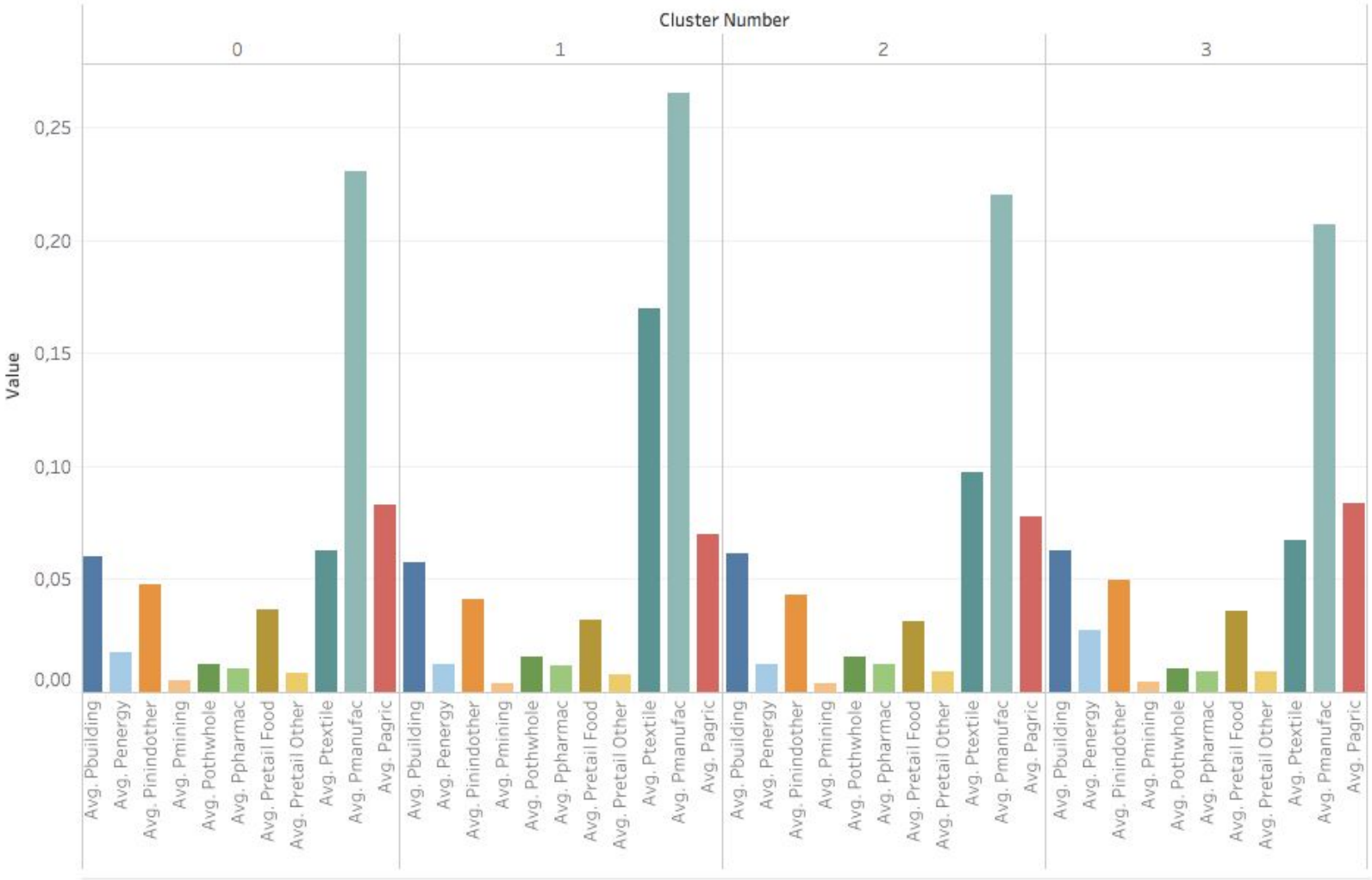


**Figure 5. Overall graphic visualization of all clusters by industries.**

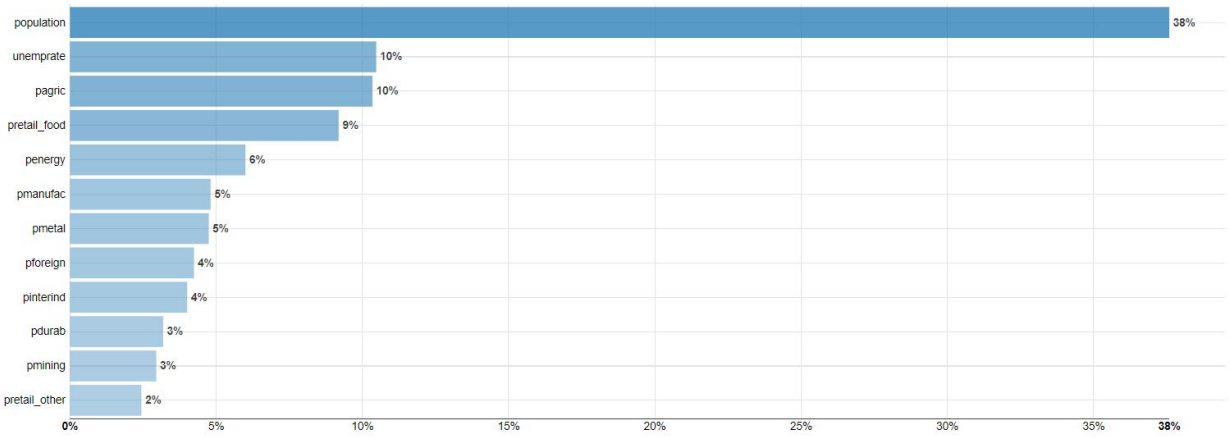


**Figure 6. Variables importance in our K-Means (k=4)**

On this graph produced after the clustering analysis, it is possible to observe the importance of the attributes used. Even after the normalization and exclusion of Madrid and Barcelona population is still considered the most important variable for this dataset.