

**Machine Learning I**

**Spanish Ministry of Economy Province Segmentation Study**

Group Assignment 1

MBD – A1 – Group C

**Table of Contents**

[1. Executive Summary 1](#_Toc12472797)

[2. Analysis 1](#_Toc12472798)

[2.1. Approach 1](#_Toc12472799)

[2.2. Solution 1](#_Toc12472800)

[2.3. Conclusions 2](#_Toc12472801)

[3. Recommendations 2](#_Toc12472802)

[3.2. Mid-Population-Sized Manufacturing Provinces 3](#_Toc12472803)

[3.3. Mid-Population-Sized Agricultural and Building Provinces 3](#_Toc12472804)

[3.4. Smallest size and not growing agriculture provinces with the lowest unemployment rate 3](#_Toc12472805)

[3.5. Mid-population-sized touristic provinces, with high unemployment rates 3](#_Toc12472806)

[4. Technical Annexes 5](#_Toc12472807)

[4.1. Annex I – Correlation Analysis 5](#_Toc12472808)

[4.2. Annex II – Variable Significance for Clustering Model 6](#_Toc12472809)

[4.3. Annex III – Detailed Metrics 6](#_Toc12472810)

[4.4. Annex IV – Cluster Heatmap 6](#_Toc12472811)

[4.5. Annex V – Cluster-Provinces Classification 7](#_Toc12472812)

[4.6. Annex VI – Map of the Clusters 8](#_Toc12472813)

[4.7. Annex VII – Cluster Feature Averages 9](#_Toc12472814)

# Executive Summary

In order to identify similar provinces in Spain, different territory features need to be considered. When identifying different groups, the human brain can easily distinguish between most populated cities or the ones with biggest unemployment’s rate, but when identifying similar groups considering a greater number of characteristics, the problem becomes non-trivial, and more complex and powerful analytical tools must be used, such as clustering analysis.

Following the preparation of the data, various demographic and economic variables were used as inputs for a clustering model to identify provinces in Spain that are similar on both of those dimensions. In the end, six variables were used. These variables are the unemployment rate, the ratio of percentage manufacturing industry to percentage building, ratio of percentage other industry trade to percentage other wholesale trade, percentage agriculture-food, the average of all the economic indices, and the population. The output was 5 different clusters, which maximizes the similarity within each group while maximizing the differences between all other group.

To gain further insights and have a clearer distinction regarding the groups, they were labelled based on their similar characteristics. The resulting labels focus on the average population size of each cluster, the main industrial activity, and relevant economic variables. These labels are the largest-population-sized, industrialized textile provinces, the mid-population-sized manufacturing provinces, the mid-population-sized agricultural and building provinces, the smallest-population-sized agriculture provinces with low unemployment, and the mid-population-sized touristic provinces with high unemployment rates. Specific economic policies recommendations are also proposed, based on the strengths and weaknesses of each group, to target problems shared by a group of provinces.

# Analysis

## Approach

The approach to this study first began with studying the available information that has been supplied. This includes segmenting the different variables into categories, including demographic variables, economic variables, and services variables. After this initial study, a correlation analysis of the different features available was conducted to reveal which of the variables are highly correlated, the results of which are presented in Annex 1. Following this several new features were created in to be able to effectively use those variables in the segmentation analysis, to circumvent issues regarding highly correlated variables. These new features were either the ratios of pairs of correlated variables or the averages of variables when more than two features were highly correlated amongst themselves.

After the preparation of the data, the segmentation of the provinces into different groups was conducted using a k-means clustering algorithm run on Dataiku DSS. This clustering technique segments the different provinces based on the similarity across the features selected, therefore selecting the features of interest for segmentation is critical in the clustering process. For each selection of variables to include in the clustering analysis, the number of clusters to be found was varied between 3 and 9 clusters, to ensure that the outcome is significant for the Ministry’s purposes.

## Solution

From the available variables, a combination of demographic and economic variables was selected to identify the clusters, to ensure that segmentation occurs on both of those dimensions. Three new variables, the ratio of percentage manufacturing industry to percentage building and the ratio of percentage other inter-industry trade to percentage other wholesale-trade and the average of all the economic indices, were created to offset the fact that their constituent variables are highly correlated amongst themselves. The other variables used were the unemployment rate, the percentage agriculture-food, and population. This combination lead to the creation of five clusters with a silhouette value of 0.3243, that vary in size from 2 to 25 provinces in each of the clusters. The importance of each of the variables in the creation of the clusters can be found in Annex II, along with the detailed metrics in Annex III and the variable heatmap in Annex IV. The provinces in each cluster can be found in Annex V and a map representation of that information is available in Annex VI.

## Conclusions

Taking into consideration the most significant characteristics of each cluster, the groups were then labelled in order to gain greater insights and to understand their composition more completely:

* **Cluster 1- Largest-population-sized, industrialized textile provinces:** The first group consists of only the provinces of Madrid and Barcelona. Being the most populated ones and the largest contributors to the economic activity of Spain, they form a logical group. They have a considerable number of foreigners, and their unemployment rate is above the mean of all provinces. In terms of industries, the most important sectors for them are manufacturing and textile production.
* **Cluster 2- Mid-population-sized manufacturing provinces:** The middle size manufacturing provinces are: Albacete, Badajoz, Cádiz, Castellón, Ciudad Real, Córdoba Granada Huelva Jaén La Rioja Murcia Sevilla and Toledo. Their economies also have, on average, a higher dependency on agriculture. The building sector is not as significant as in other provinces and the banks per capita is below the average of other Spain provinces.
* **Cluster 3- Mid-population-sized agricultural and building provinces:** This group is the largest group found by the algorithm, containing 25 provinces in total which are listed in Annex V. These provinces have a higher than average dependency on agricultural and building, whereas they have a lower than average dependency on manufacturing is compared to other provinces.
* **Cluster 4- Smallest-population-sized agriculture provinces with low unemployment:** This group contains Cuenca, Lugo, Segovia, Soria and Teruel. They, on average, have the smallest populations and the lowest unemployment rates, thus, the smallest contribution to the Spanish economy with the lowest indices for all areas. In addition, they have higher than average mining, energy and inter-industry trade, both wholesale and others, percentages when compare to others.
* **Cluster 5- Mid-population-sized touristic provinces, with high unemployment rates:** This group contains 7 touristic provinces with mid-sized populations, while having the highest unemployment rates and higher than average foreign population rates. These provinces are more dependent on building and textile industries. They also have the lowest banks per capita ratio.

# Recommendations

Since the clusters differ in their economic performance, demographic trends and existing infrastructure, the most effective economic policies will have to consider the strengths and target the weaknesses of the groups. Thus, it is essential to work in conjunction with regional authorities to ensure successful implementation of the policy. The full breakdown of the figures on which these recommendations are based on is available in Annex VII.

* 1. Largest-Population-Sized, Industrialized Textile Provinces

This group is not dependent on one particular industry, its specialization is equally distributed between manufacturing and textiles, which says that this cluster is more resilient to economic turbulence. However, this group has a relatively high population growth and a medium unemployment rate. In the medium term, this trend may result in a rapid increase in the unemployment rate unless new jobs are created. This is not the most urgent issue, nevertheless it must be addressed before a crisis happens. This can be achieved through indirect measures like retraining programs, decreasing unemployment benefits, etc.

## Mid-Population-Sized Manufacturing Provinces

Policies regarding this group should target unemployment and banking sector. The industries present in this group seem to be rather diverse. This cluster represents the mean of the country by many indicators; thus, it does not require immediate actions in terms of stimulating economic growth. Unemployment has to be targeted via creation of new jobs since the population growth is relatively high, and this problem may be exacerbated in the future if it’s not tackled.

## Mid-Population-Sized Agricultural and Building Provinces

This group includes half of Spain’s provinces, however, the relative contribution to the economy of each province is, on average, the second to last, which means that each province within the cluster isn’t developed enough. The ministry should consider a general stimulating economic policy that will have an impact across all industries. For example, transfers, increase in government procurement, decrease in taxes for building, agriculture, or manufacturing, given that those industries have the highest volumes.

## Smallest size and not growing agriculture provinces with the lowest unemployment rate

The fourth group has, on average, the smallest population size, and a negative population growth rate. This group is very similar to cluster 2, so in the future it may assimilate with it. The main difference is that this group is more specialized in banking, however, the other parameters are very similar, with the lower unemployment rate in cluster 4 potentially explained by the low population growth and size. This means that whatever policy is applied to this group will probably have similar effect on cluster 2. Another possible option is to attempt to stimulate the tourism industry or to try to increase the foreign population, so that it can become the main source of labour in the province in the future.

## Mid-population-sized touristic provinces, with high unemployment rates

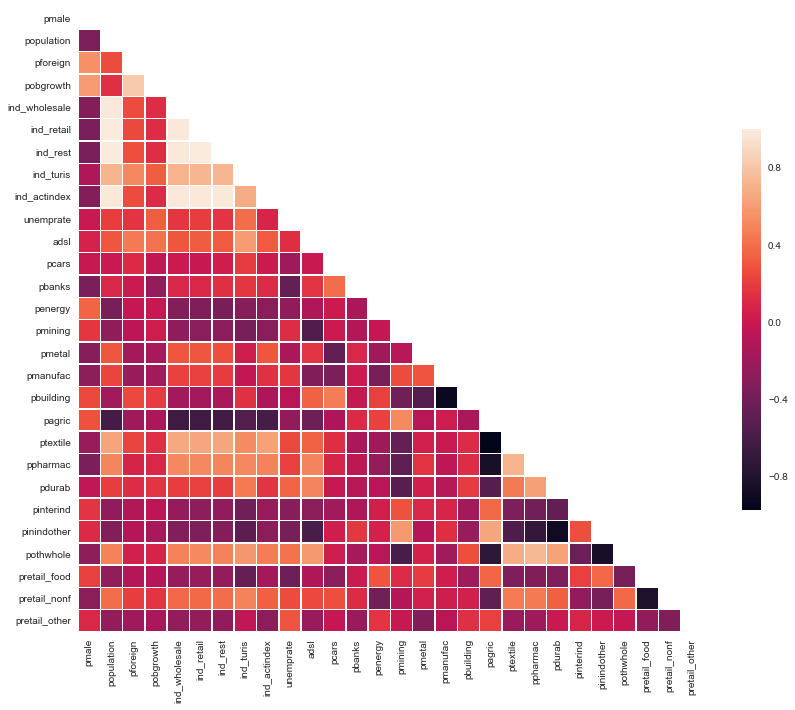
The main issue within this cluster is the rapid population growth and very high unemployment. The government’s main focus of economic policy should be tackling unemployment in the short run. A potential solution could be decreasing unemployment benefits to create an incentive for people to look for a job faster and not rely heavily on unemployment benefits. This could also be achieved through providing companies with subsidies, either directly or via alleviating corporate taxes. These measures can mitigate the problem of unemployment in the short run. However, the main direction of unemployment policy should be direct creation of new jobs, since in the future rapid population growth will create an excessive labour supply that will only aggravate the problem if left unattended.

This cluster has a lowest concentration of banks, which may indicate that financial markets are less developed in this group. Stimulating financial industry could also help invigorate other industries and escape the high dependence on the tourism.

# Technical Annexes

## **Annex I – Correlation Analysis**

Provinces features correlation discovery process, to identify different attributes sharing the same information



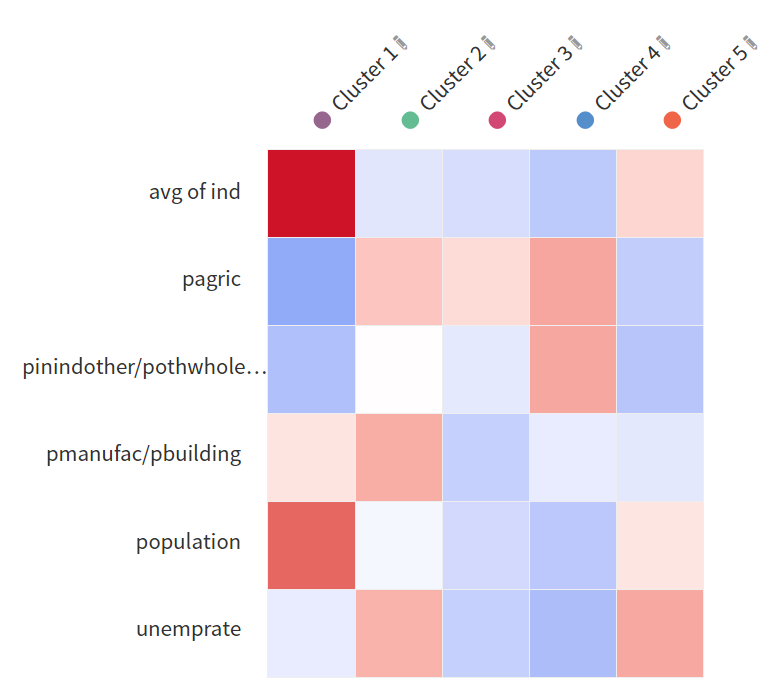
## **Annex II – Variable Significance for Clustering Model**

Understanding the most relevant variables when conforming the groups with similar provinces

## **Annex III – Detailed Metrics**

|  |  |
| --- | --- |
| Measure | Value |
| Silhouette | 0.3243 |
| Inertia | 91.93 |

## **Annex IV – Cluster Heatmap**



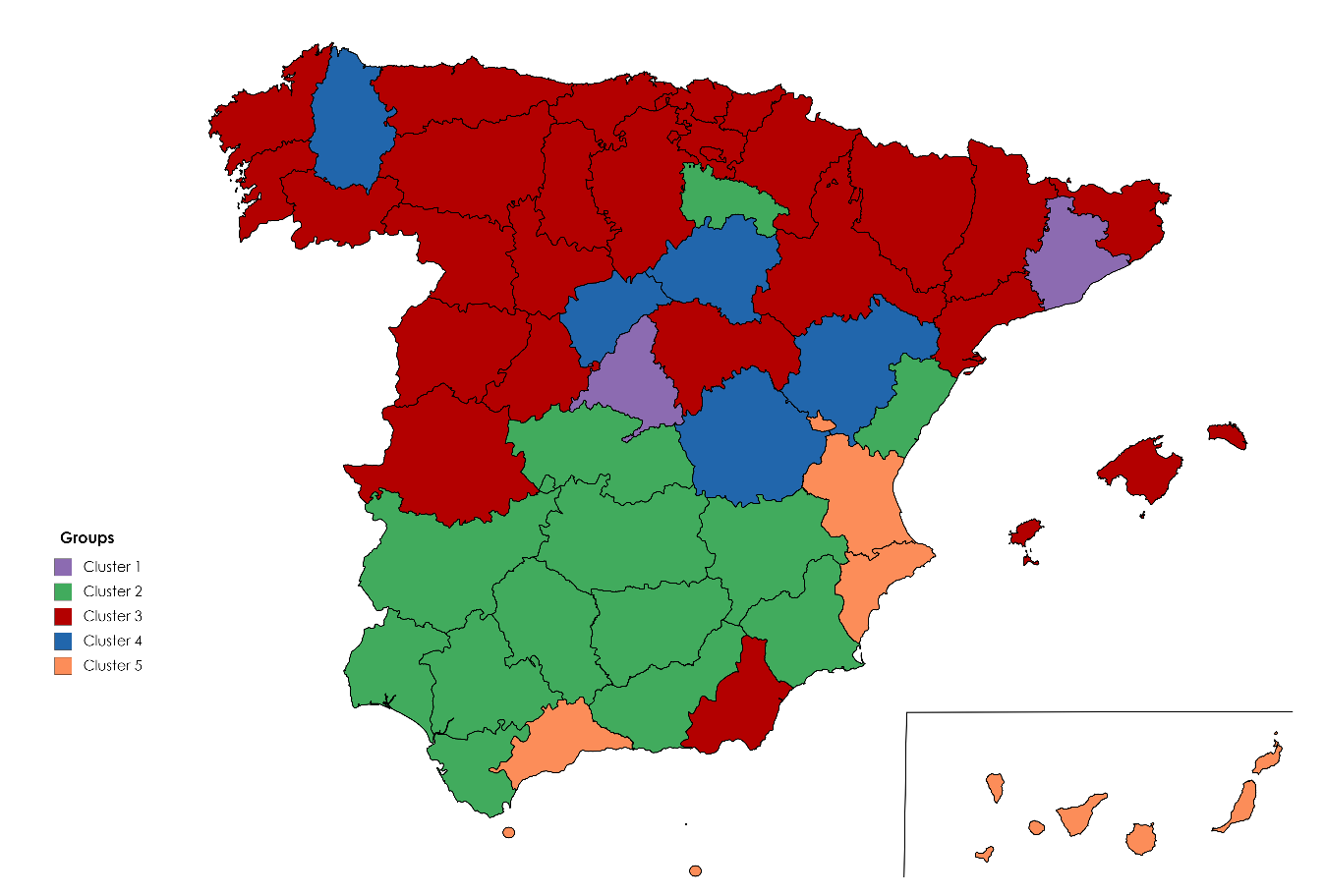
## **Annex V – Cluster-Provinces Classification**

Tabular representation of the results



## **Annex VI – Map of the Clusters**

Visual representation of the results



## **Annex VII – Cluster Feature Averages**

Average of all features for each of the clusters.



