Regional development of Spain using segmentation analysis

Prepared by
Paliz MUNGKALADUNG

Index

Executive summary

Spain is a highly decentralized country in terms of regional economic development. Therefore economic policies should be tailored to the particular situation of each province. The ability of grouping the 52 Spanish provinces according to their similarities would allow for more efficient and effective policy creation and a targeted application.

Main objective

The Ministry of Economy concentrates on determining the structural reform priorities for similar regions in Spain and reducing regulatory differences. On this regard, segmentation analysis has been applied to group all 52 provinces in several similar clusters based on key variables that characterize their social and economic development through k-means cluster analysis.

Key findings and recommendations:

The k-means cluster analysis helped find similarities between provinces and gave a better understanding of the whole economic landscape of Spain. Therefore, our recommendation is to differentiate the policies to be applied based on these three clusters, as they represent different needs and possible areas for development:

- The policies for cluster “Big Cities & Tourism” focus on balanced growth of other industries.
- The policies for cluster “Unemployment issues” focus on strengthening labour market policies.
- The policies for cluster “Population ageing” focus on encouraging birth rate.

The methodology and conclusions from our analysis

The analysis is based on the reports of the Ministry of Economy containing the information about population, unemployment rates, industry indicators and etc. in these regions. To assign the provinces to groups based on the mentioned features, I used the nonparametric method known as k-means clustering.

Data Exploration

The Data:
Our dataset included demographic and economic information of the 52 Spanish Provinces.

Segmentation

Clustering:
Through the implementation of Machine Learning we investigated non-apparent patterns in the Data that would reveal clusters that are not so obvious

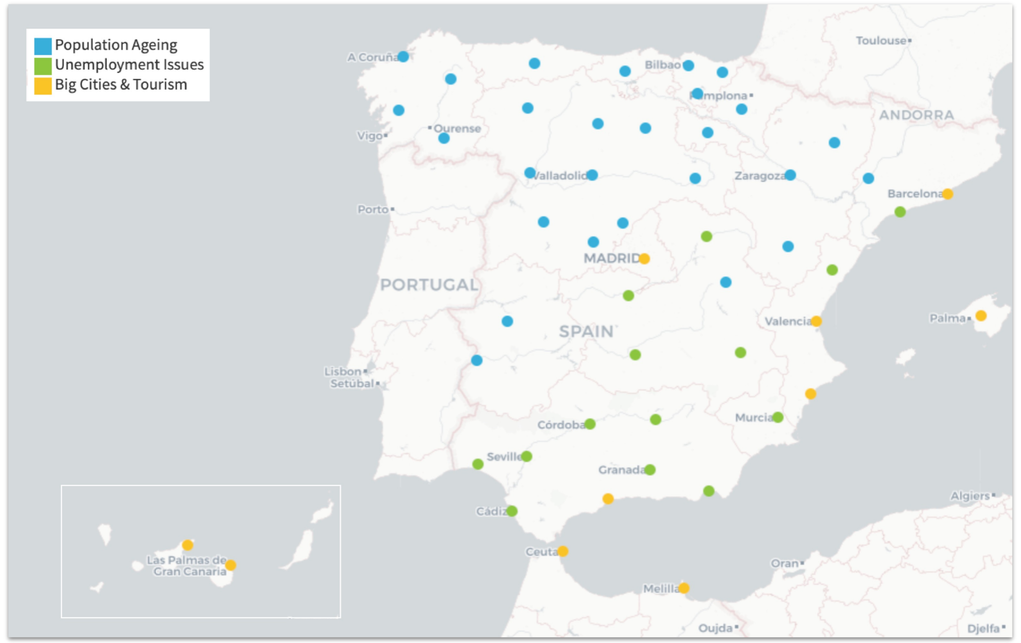
Cluster Analysis

Insights:
We investigated each cluster for its peculiarities and also the differences between each clusters

We ensured that each group is significantly different from the other in order to allow for a differentiation in policies

Final Recommendations

Conclusions:
Based on the distinctive features of each cluster we defined recommendations in terms of policies



Three main clusters:

Cluster "Big Cities & Tourism":
- mainly characterized by tourism;
- has the largest population;
- has strong presence in all indicators except agriculture.

Cluster "Unemployment Issues":
- mainly located in the south of Spain;
- characterized by manufacturing and agriculture;
- has a high unemployment rate and the highest population growth among the three clusters.

Cluster "Population Ageing":
- mainly located in the north of Spain;
- is more industry-driven with the largest presence of banks;
- has the smallest population and population growth.

Recommendations

Findings	Key recommendations
Unemployment issues	
Cluster “Unemployment Issues” has the highest unemployment rate which may lead the increase of labour migration to other regions.	• Reduce dualism in labour market • Develop vocational education and training to ensure that the skills meet current market needs • Implement profiling tools and specialization of counsellors • Incentives and subsidies to attract industry in the area
Deep concentration on tourism industry	
Cluster “Big Cities & Tourism” has the highest percentage of tourism sector in comparison with other clusters.	• Undertake structural reforms towards balanced growth of other industries • Promote financing sources to businesses for diversified sectors • Introduce special tax regime incentives for new businesses unrelated to this sector
Population ageing	
Cluster “Population Ageing” has the smallest population and additionally the smallest population growth among the three clusters	To slow down the aging of the population, it is advisable to stimulate the birth rate by: • introducing long-standing policy to support child-bearing and parenthood • providing various subsidies such as a one-time “childbirth bonus” and income tax exemptions. • Create incentives to develop telecom infrastructures in order to level this cluster with the rest of the country.

ANNEX

ANNEX

Dataiku DSS and DSS Notebook are used for this analysis to manipulate the dataset and to confirm the result. Steps of my analysis are described in the following slides.

1 - Performed data cleaning and EDA - Exploratory Data Analysis with the given dataset “province.xlsx”.

ROWS	52
DUPLICATES	0
FEATURES	26
CATEGORICAL	4
NUMERICAL	21
TEXT	1

Table 1: Data cleaning summary

- The data set at hand was already in good hands
- Contained economic and demographic data for each province
- The population is mainly distributed in the large and touristic cities mainly Madrid, Barcelona and Valencia

Fig. 1: Descriptive Analysis of some numerical Features

2 - Selected the important features

To select the important features, I generated a Heatmap of Correlation Matrix on DSS Notebook. It supported me in understanding the data and to consider these important features to proceed with the next steps. In parallel, I also ran trail tests on K-means clustering in Dataiku DSS to determine the important features. Finally, I decided to go with 14 interesting features which are 'population', 'pobgrowth', 'ind_turis', 'ind_rest', 'adsl', 'pbanks', 'penergy', 'pmanufac', 'pagric', 'ppharmac', 'pdurab', 'ptextile', 'pretail_nonf', 'unemprate'.

I proceeded by first excluding highly correlated features and thus kept the ones covering the information which best describes the demography, industry sector, and economy of provinces in Spain.

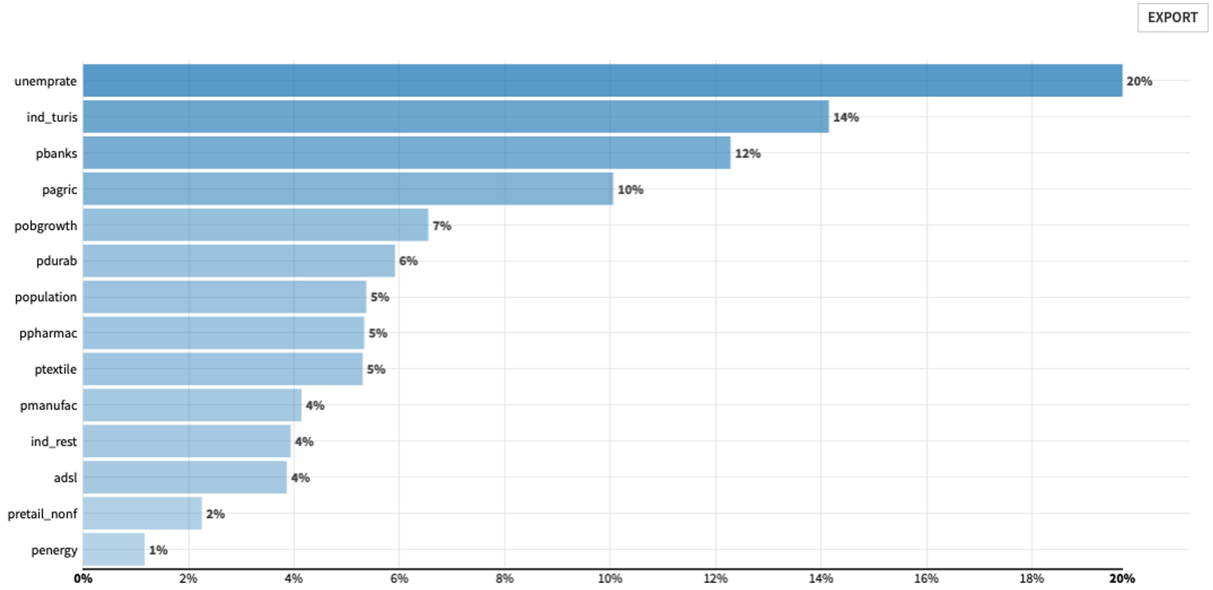


Fig. 2: Variable importance (14 selected features in Dataiku DSS))

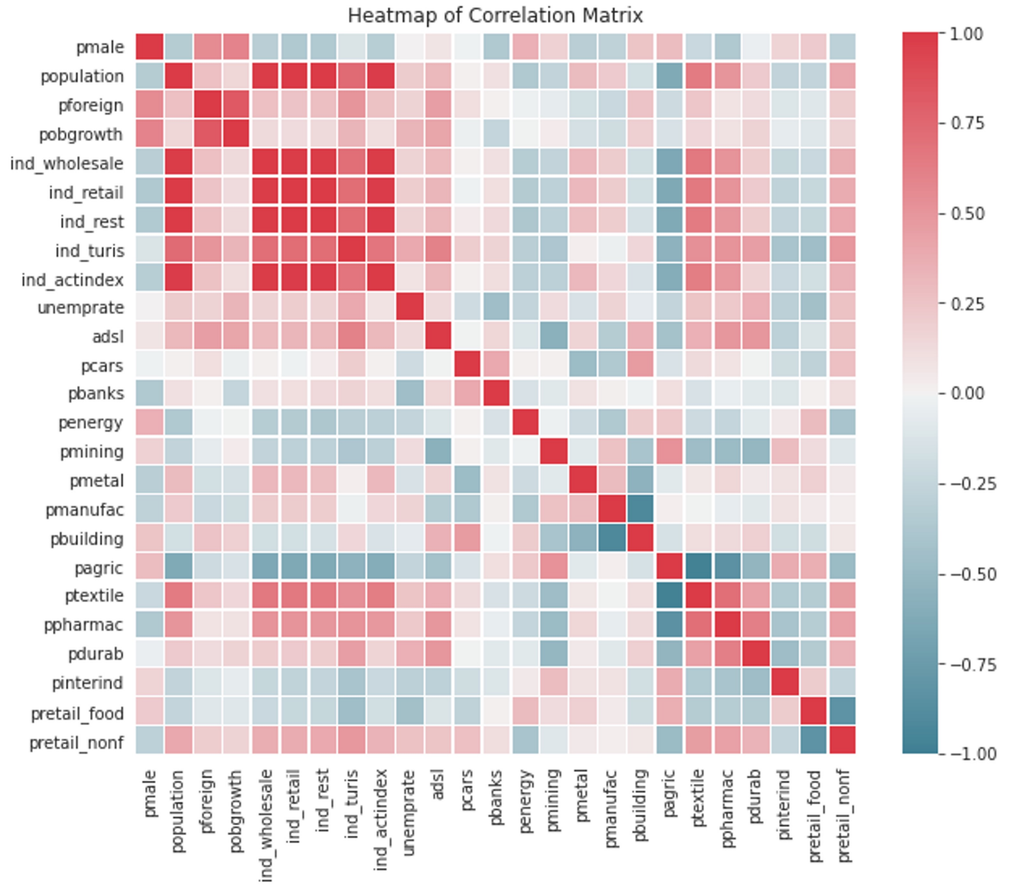


Fig. 3: Heatmap of Correlation Matrix (visualized all 26 features in DSS Notebook)

3 - Run K-means clustering

To define the optimal number of clusters, I iteratively ran the K-means algorithm in DSS Notebook and then ran both an Elbow and a Silhouette Analysis.

I found that the elbow was approximately 3, thus I opted for 3 clusters.

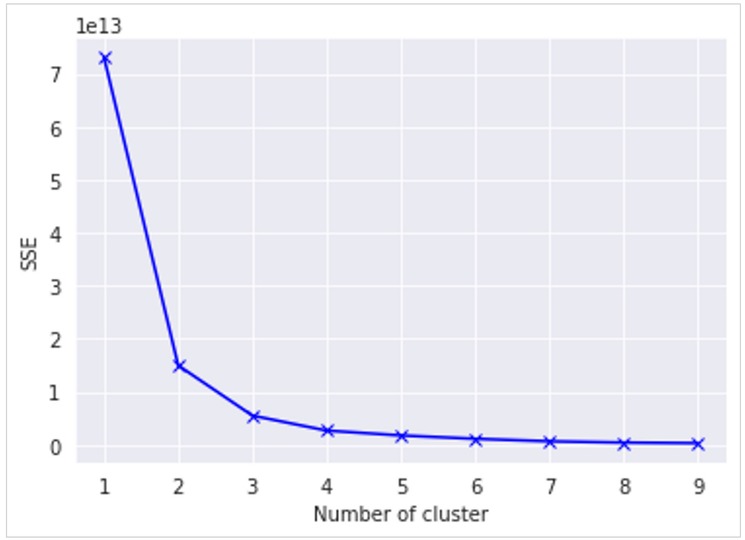


Fig. 4: Elbow plot from DSS Notebook

At the same time in Dataiku DSS, the best silhouette score is at k = 2.

Observing our results with K = 2 I obtain two clusters as reflected in our map::

- Agriculture & Industry Areas - Cluster_0 (41 provinces)
- Big Cities & Tourism - cluster_1 (11 provinces)

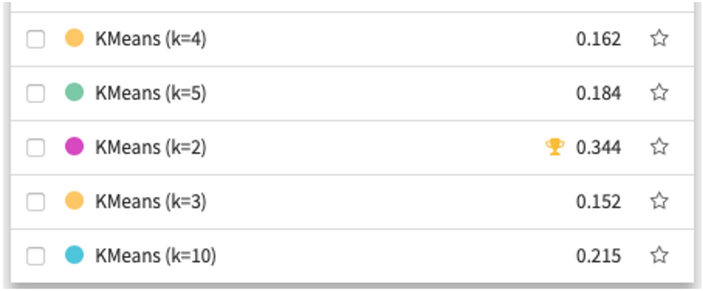


Fig. 5: Silhouette scores from Dataiku DSS

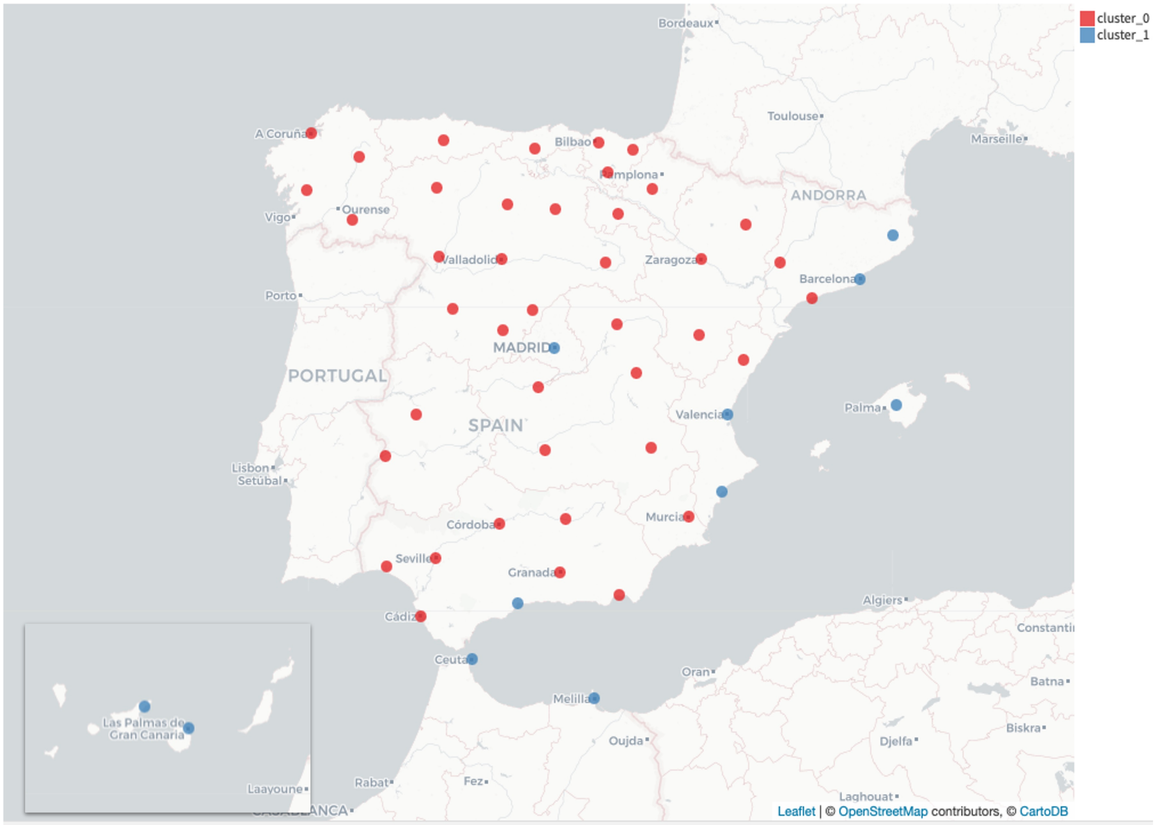


Fig. 6: Geography map plot of 2 clusters

It becomes clear that the 11 provinces are very particular compared to the rest of Spain, however I believe that more differentiation exists within the 41 Provinces of the Agriculture Cluster and therefore I believe it would not be optimal to apply the same policies to this entire group. Hence, it was decided to go for k = 3.

Without proper rescaling, the result of k = 3 (Fig. 7) separates the two metropolis (Madrid and Barcelona) from the Tourism cluster. To control this, I manage preprocessing in DSS Notebook by running MaxMinScaler algorithm.

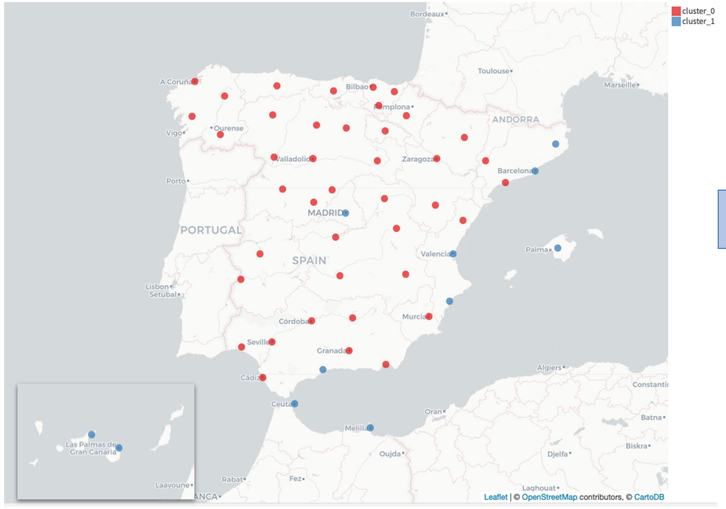


Fig. 7: Geography map plot of 2 clusters

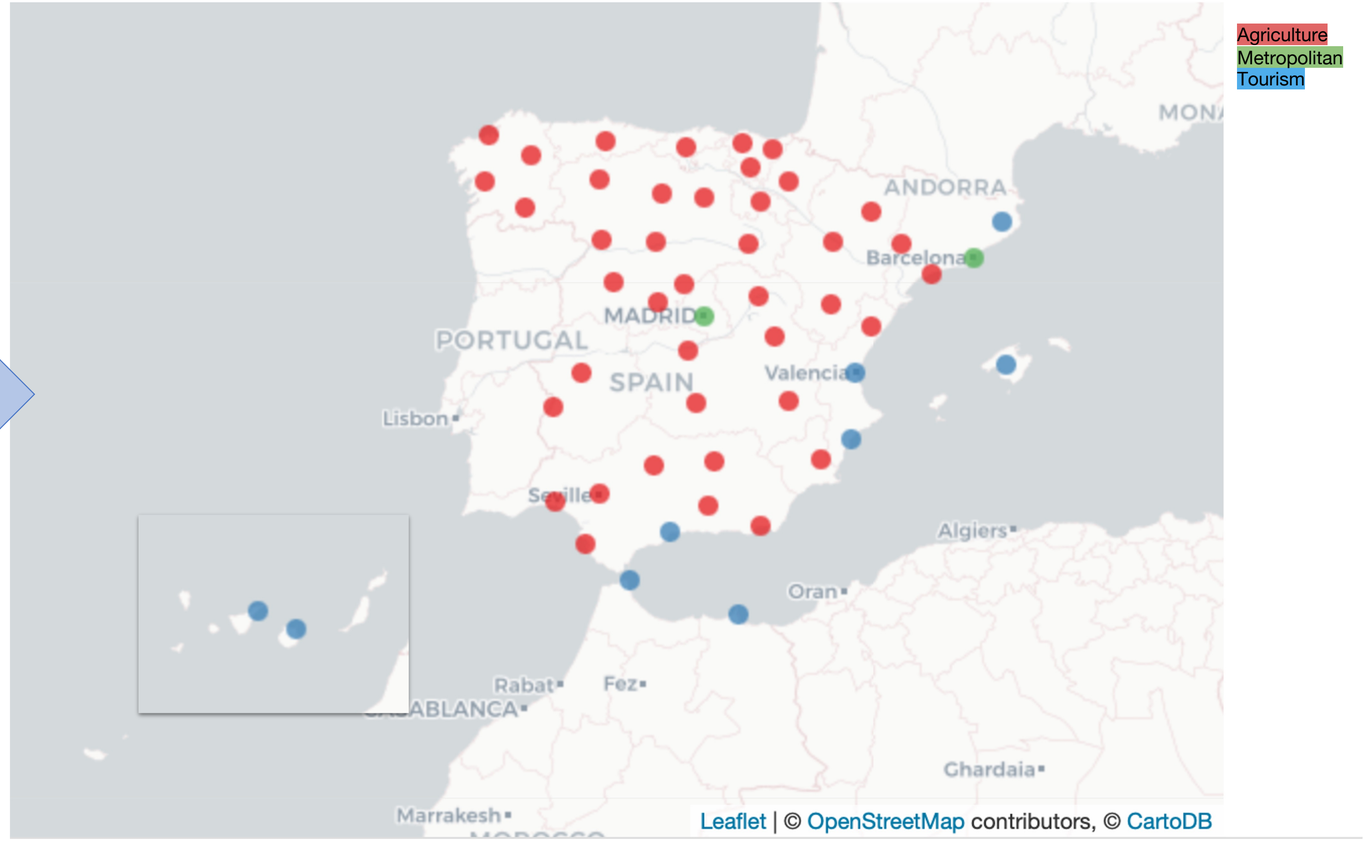


Fig. 8: Geography map plot of 3 clusters, only Madrid and Barcelona are grouped in cluster_2

By applying the model with the new prepared dataset, I obtained the most satisfactory result with the following 3 clusters.

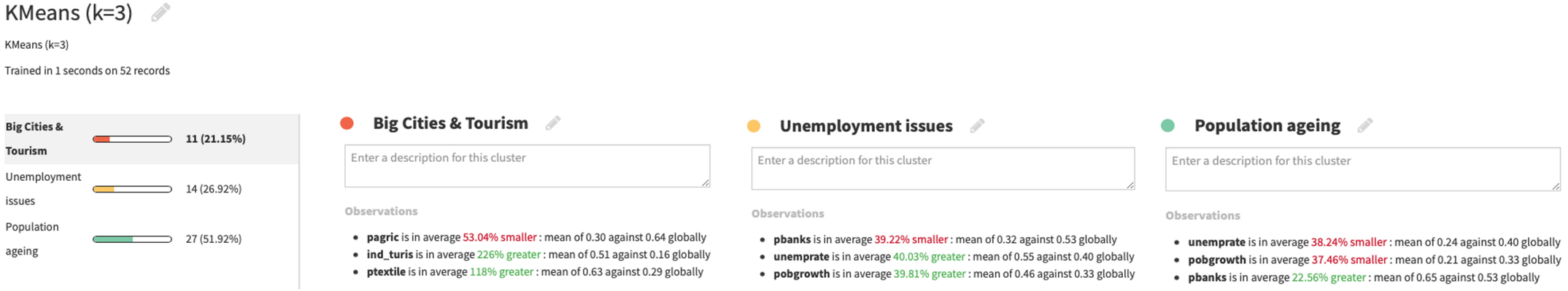


Fig. 9: Summary of clusters in Dataiku DSS

From Dataiku DSS, I can compare the numerical features of our three clusters as per highlighted in the summary. With the help of graphs and box plots of the relevant features, I was able to gather deeper insights in the clusters and therefore explain them.

Additionally, I generated the table of mean values and box plots of our selected features to support the description and thus the understanding of each cluster.

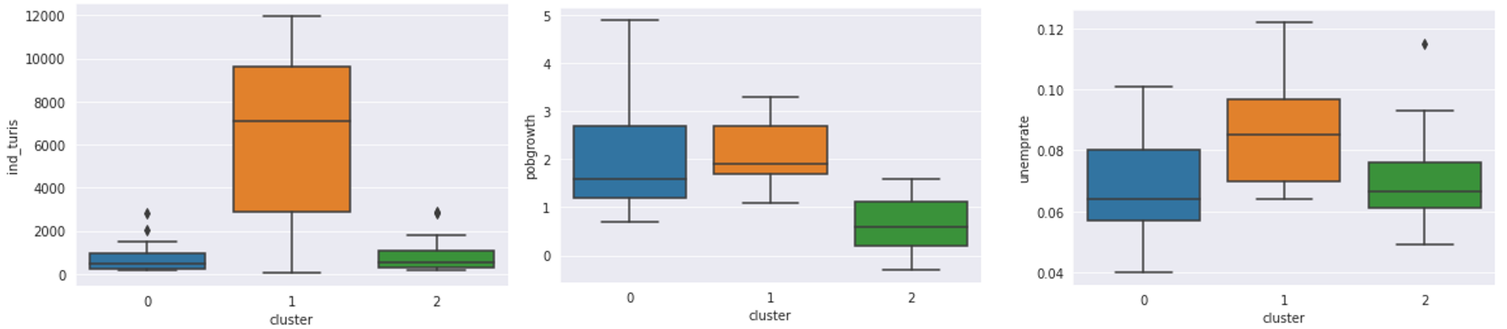


Fig. 10: Box plot the selected features, obtained from DSS Notebook

Legend:
Cluster 0 : Unemployment Issues
Cluster 1: Big Cities & Tourism
Cluster 2: Population ageing

4 - Conclusion and recommendations

Applying our clusters on the map supported us in gaining even better understanding and in the generation of our recommendations.

Cluster_0 : 14 provinces → Unemployment issues
Cluster_1 : 11 provinces → Big Cities & Tourism
Cluster_2 : 27 provinces → Population Aging

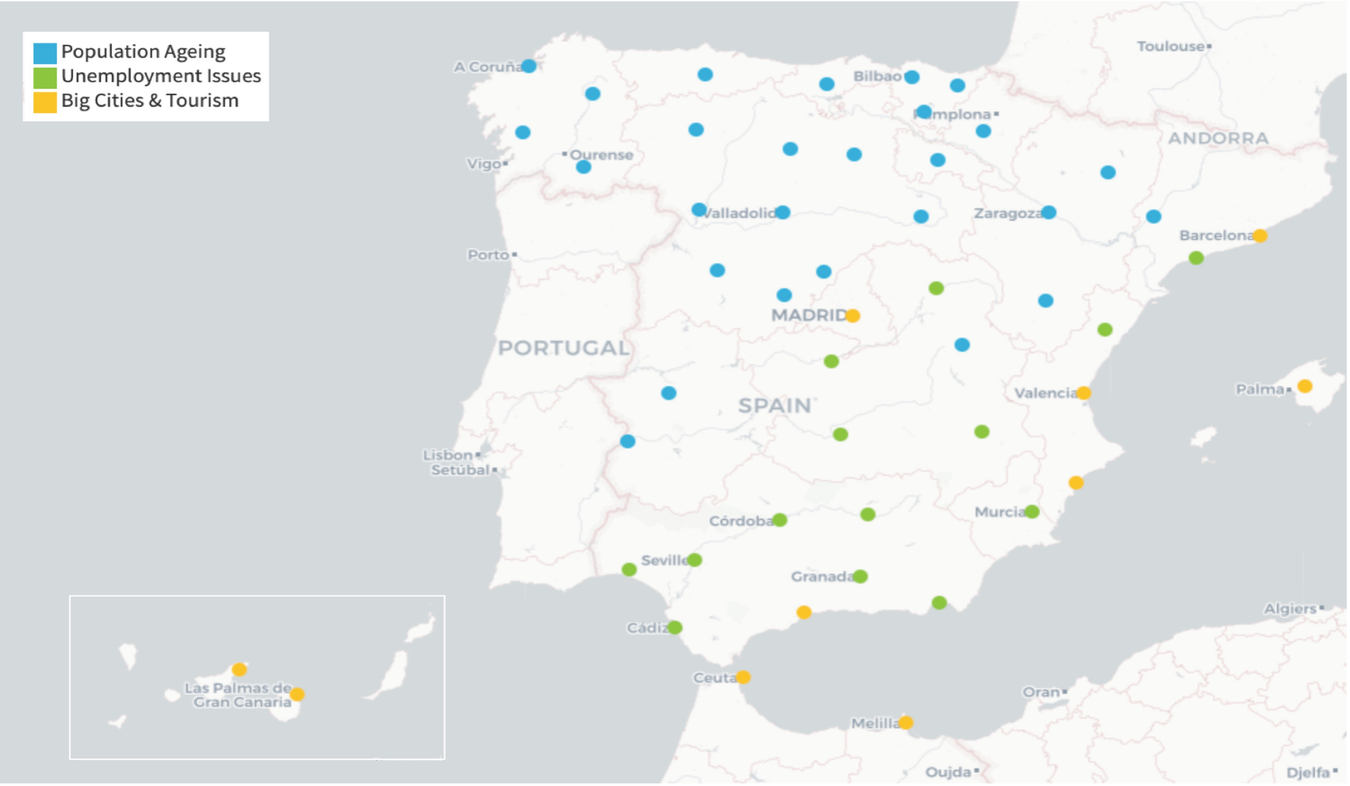


Fig. 11: Geography map plot of the most satisfactory result of 3 clusters

Cluster_0 (n=14)	Cluster_1 (n=11)	Cluster_2 (n=27)
Albacete	Alicante	Álava
Almería	Baleares	Asturias
Cádiz	Barcelona	Ávila
Castellón	Ceuta	Badajoz
Ciudad Real	Girona	Burgos
Córdoba	Las Palmas	Cáceres
Granada	Madrid	Cantabria
Guadalajara, Spain	Melilla	Cuenca
Huelva	Málaga	Guipúzcoa
Jaén	Santa Cruz de Tenerife	Huesca
Murcia	Valencia	La Coruña
Sevilla		La Rioja
Tarragona		León
Toledo		Lleida
		Lugo
		Navarra
		Ourense
		Palencia
		Pontevedra
		Salamanca
		Segovia
		Soria
		Teruel
		Valladolid
		Vizcaya
		Zamora
		Zaragoza

Table 2: 3 clusters of provinces in Spain

References

1. https://www.oecd.org/economy/surveys/Spain-2018-OECD-overview-economic-survey.pdf
2. https://read.oecd-ilibrary.org/economics/oecd-economic-surveys-spain-2017/executive-summary_eco_surveys-esp-2017-2-en#page3
3. ttps://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/spain_2014_2020/spain2014_urban_execsumm_en.pdf
4. https://www.oecd.org/economy/growth/Spain-country-note-going-for-growth-2021.pdf

1. Spain Picture: https://3npro2k2jyx3gj3xd33s373n-wpengine.netdna-ssl.com/wp-content/uploads/2019/09/Spain2020.jpg