

SEGMENTATION REPORT

CLUSTERING SPAIN



By: T. Bachmann, A. De Roni, A. Grullón, R. Mejía, F. Mansilla, U. Varol.



CONTENTS

EXECUTIVE SUMMARY PAGE 2 **TECHNICAL WORKFLOW** PAGE 3 **CLUSTER MEANING** PAGE 4 RECOMMENDATIONS PAGE 5 **APPENDIX** PAGE 6

EXECUTIVE SUMMARY

ABSTRACT AND OVERVIEW

The current report examines the similarity among the Spanish provinces as a requirement to apply a reduced set of policies in the Spanish territory. By using a 7-cluster segmentation, we are able to recommend tax reforms and economic stimuli to improve the quality of life in these areas.

The data mining analysis is performed based on the "Cross-Industry Standard Process for Data Mining" (CRISP-DM), an iterative methodology which combines business and technical efforts to extract insights from data.

The Methodology starts with the definition of the business objective: the application of policies for different cities by creating groups with similar features. The data mining goals seek for a high heterogeneity between groups (separation) and a high homogeneity within-groups (cohesion).

An initial exploration of the data was performed to understand and extract meaningful insights from the dataset, containing all 52 provinces of Spain and 25 explanatory variables (socio – demographic, behavioristic and business-related variables).

During the data preparation, a correlation analysis determined which variables would be excluded for clustering purposes to avoid including redundant information in the study. The technique selected to perform the segmentation analysis is "distance-based" where multivariate distances are measured, and clusters are formed by "neighbors' proximity". The data team decided to use the "K-Means" algorithm to define the model and the evaluation of results was assessed with the tool "Dataiku".

RECOMMENDATIONS

Our main findings include high unemployment across different clusters, as well as low cross regional trade. Our suggestions include stimulating the tourism industry, the establishment of special economic areas. We also recommend relaxing tax policies across different areas while, increasing taxes in wealthier regions to support policies for other regions.

PROPOSED SOLUTIONS

The iterative process led to a 7-cluster segmentation. To assess the technical quality of the clusters, the trade-off between the following characteristics was taken into account: separation between clusters (with at least a specific characteristic to distinguish that group from any other), cohesion (checking dispersion of variables), discriminatory power of features for clusters, and relative size of clusters.

- Highly Industrial Provinces
- Rural Agricultural Provinces
- Upper-Middle Class Provinces with Stable Economy
- Highly Touristic Multi-Cultural Business Capitals
- Agriculture Based Trade Provinces
- Touristic Provinces with Population Growth and High Unemployment
- Ciudades Autónomas

However, the most critical criteria to define the clusters is the interpretability of them to lead to a correct decision making. In that sense, one cluster was further divided to achieve a better interpretation of results and hence, a better quality of the model, obtaining the 7 final clusters that we show in the graph below.



01 TECHNICAL WORKFLOW

As mentioned in the executive summary, the CRISP-DM methodology is the standard procedure that TeamB© follows in order to deal with ML projects. In this section, a more detailed explanation of the different phases will be covered to understand the final result of the segmentation of Spanish provinces.

1 BUSINESS UNDERSTANDING

The Ministry of Economy in Spain wants to find out which provinces are similar to apply a set of policies to several provinces. The business success criteria was defined as the usefulness of the segmentation results to apply the mentioned policies. The assessment of this criteria will be held by the data with cooperation analysts members of Ministry the Economy. A clustering technique potential was selected as the better fits solution that the requirements.

3 DATA PREPARATION

Some actions were required prior to running the model. First of all, standardizing the variables as their units were differing considerably. Formatting and merging of datasets were done as a matter of including the external data source. For that purpose, the tool "Dataiku" was chosen due to its performance on these tasks, and its usefulness in other stages of the methodology.

2 DATA UNDERSTANDING

The exploration of the data prior to the definition of further steps is crucial to ensure the right interpretation of the available information. The data quality report analyzed the coverage, keys across tables, check of missing values, possible outliers and plausibility of values. At this point, the team performed a correlation test among variables to avoid including redundant information that could have a negative impact in the cluster outcome. (See Appendix Correlation Analysis). This allowed the reduction of variables to be used and lead to more interpretable insights of the groups. After some iterations, a new dataset was added to the analysis including the location of the different cities to help with the visualization of the results, without affecting the clusters of the model.

4 MODELING

The algorithm selected to define the clusters is the "K-means", being the most popular, offering good cluster solutions for wellspaced clusters. Other algorithms were used to compare the results, such as "Two-step" algorithm, but the data analysts found a better interpretation of the clusters with a better cluster solution for the main purpose of the Ministry of Economy with "K-means". From a technical point of view, the separation between clusters was evaluated with the Silhouette coefficient, whereas the cohesion within groups was assessed with the distribution of the variables inside each cluster (as a major rule, dispersion means non-explanation). Several iterations were required until the final potential solution was computed, by introducing different variables that could lead to a higher Silhouette and at the same time allowing a meaningful explanation of the clusters. Besides, only solutions where every cluster had at least a distinctive variable to define it were taken into account. The trade-off between all these characteristics led to a 7-cluster model with a Silhouette of 0,267.

5 EVALUATION

The definition of the meaning of clusters was the most restrictive criteria to come up with the final suggested solution. As a technical validation criterion, the team set a different initialization of the centroids for the algorithm (which can lead to different results in "K-means") and almost the same result was obtained for the chosen iteration. The ministry will decide whether to move towards the deployment of the model or to specify new requirements to analyze.

3

02 CLUSTER MEANING

The following section serves as an explanatory outcome of the seven groups obtained from the model.

MI HIGHLY INDUSTRIAL PROVINCES

Cluster 1 is differentiated with higher manufacturing, mining and metal industry activities. Albacete having the biggest industrial parks; La Rioja, Sevilla (incl. aircraft manufacturing); Cordoba, Valencia (incl. textiles, chemicals, metals) having high production volumes in automotive and metal-mechanic and advanced manufacturing, Granada being the largest producer of iron ore and having agriculture-based manufacturing activities (e.g. sugar refining), Jaen being the home to olive oil production highlight this clusters' industrial centricity.

RURAL AGRICULTURAL PROVINCES

Existence of agriculture and lack of adsl plays important role in discrimination of this cluster from the others, especially from Cluster 5. This cluster includes provinces such as Almería, Avila, Badajoz, Murcia, Ourense, where agro-livestock is the main activity with related processing industries built around (e.g. dairy, meat, timber); and with scarce industrial activity compared to the other provinces, limited level of service and administarial jobs.

UPPER-MIDDLE CLASS PROVINCES WITH STABLE ECONOMY

Cluster 3 includes provinces such as Lleida, Girona, Baleares, Valladolid (main employment source is the service sector, with fair volume of manufacturing) and Cantabria (with a circulating economy, mainly focused on service sector). Provinces in this cluster exhibit dynamic economies yet have stable income with focus on different sectors. This has led the banking sector develop parallel to this trend; of which Cantabria hosts the leading organization in the Spanish banking industry, Banco Santander.

M HIGHLY TOURISTIC MULTI-CULTURAL BUSINESS CAPITALS

Cluster 4 includes two most populated, multi-cultural urban capitals with continuing immigration: "Highly Touristic Multi-Cultural Business Capitals" - Madrid and Barcelona which are "the value creators" by far in Spain, as pointed out by the "Global Economic Activity Index". These capitals' economic activity depend more on consumption (supported by the food and non-food retail indicators), with the multi-national businesses requiring high-skilled jobs, as well as tourism and service sectors.

AGRICULTURE BASED TRADE PROVINCES

Cluster 5 include provinces such as Ciudad Real, Zamora, Guadalajara, Huesca, Teruel with economy highly dependent on the agro-livestock activities. Yet, having raw materials for other sectors with agricultural activities, and some unique products (woolen, wines, etc.) makes the inter-industry trade an important driver for these provinces.

TOURISTIC PROVINCES WITH POPULATION GROWTH AND HIGH UNEMPLOYMENT

Cluster 6 includes Cádiz, Las Palmas, Málaga and Tenerife whose economies depend on tourism and services sectors where industrial production and others hold less opportunities. They are subject to seasonal employment in parallel to the demand for the growing tourism. Although tourism creates demand for the job market; the influx of past emigrants and other citizens keep the unemployment rate high. This can be briefly observed from the differentiation of this cluster in tourism index, population increase in the last decade, and unemployment rate.

CIUDADES AUTÓNOMAS

"Ciudades Autónomas" cluster consists of Ceuta and Melilla. These provinces have the highest population density among other provinces. Economy is mostly driven by construction, fishing and cross-border trade, considering the borders with Morocco and the presence of ports. Although presenting employment opportunities; having daily workers available to fulfill the demand in the job market from Morocco and increase in population are not helping unemployment rate to go down in this cluster.

03 RECOMMENDATIONS

In this section, the data analyst team explains some suggestions for the application of policies with the clusters obtained.

- Our suggestion would be to launch marketing campaigns to attract more tourists to some of the more history-rich destinations such as Toledo and Sevilla. In other provinces, there is a need to develop new industries. Places such as Jaen and Grenada are primarily based on agriculture but have no chance of competing with the national hubs from clusters 1 and 4. Therefore, implementing tax benefits for corporations building new factories in the area may help inn attracting more businesses and creating new jobs. Nonetheless, La Rioja's region is famous for its wine yards, thus developing that may create new jobs in that particular area.
- The value proposition should be based on typical, locally sourced foods and wines, which are already present and strong in these areas. Once the hospitality sector starts to benefit from this, it would make sense to raise taxes on both banks, which will experience more cash influxes, as well as the newly flourishing hospitality
- Our suggestion is to develop policies aimed both at creating new financial jobs in the area by incentivizing banks and other financial intermediaries to open new offices and hire locally. If these financial institutions were to prosper, this may also raise the global trade activity index, making these areas more prosperous, at which point taxes should be raised.
- Raising taxes in these two major cities may help in funding all the more expansive policies needed to develop the local economies of those clusters which appear to be less fortunate.
- We propose focusing the previous cluster on a kind of agriculture that stimulates gastronomic tourism, while developing this one into a powerhouse capable of feeding Spain and profit from exporting its fresh goods. Lowering taxes on exports in this area and raising them in the other would be a good start.
- Our suggestions include considering the idea of building new business parks and establishing special economic zones. This would not only inject money into the areas by hiring local building contractors, which are already strongly present, but also attract new business looking for tax and trade benefits. Furthermore, special trade zones would also help in stimulating both inter industry trade as well as the global activity index.
 - Ameliorating security in this area is pivotal, we suggest both increasing policing activity as well as launching social rehabilitation programs aimed at helping illegal migrants, especially the younger ones, to find a job and becoming productive members of society.
- On the other hand, once the cities are made safer, unemployment and a stagnating local economy must be dealt with. For that, we advise to launch social programs meant to train unemployed people to work in the hospitality sector. Coupling this effort with campaigns aimed at developing the image of the two enclaves as touristic attractions. There have been discussions for building a new airport in the past, which did not succeed; perhaps a new hub would help in attracting visitors from both Europe and North Africa. Also, with Morocco recently starting to build a 90 hectares trade zone near Ceuta, there is an opportunity to strongly develop the global trade participation of the area and create new jobs.

04 APPENDIX

The Appendix includes technical information related to the project. It consists of the following sections:

- Correlation Analysis
- Distribution of variables within clusters
- Clusters and list of provinces

CORRELATION OF VARIABLES (1/2)

				pop inc%	pop inc%							
	population	pmale	pforeign	(1)	(2)	ind_wholesale	ind_retail	ind_rest	ind_turis	ind_actindex	unemprate	adsl
population	1											
pmale	-0,33464	1										
pforeign	0,269351	0,551577	1									
pop inc% (1)	0,151315	0,59668	0,82997	1		ı						
pop inc% (2)	0,272195	0,176068	0,346034	0,521053	1		ı					
ind_wholesale	0,990402	-0,31241	0,267168	0,126402	0,2512	1						
ind_retail	0,996294	-0,36279	0,251441	0,122225	0,264361	0,98975	1					
ind_rest	0,993726	-0,35906	0,275148	0,129353	0,240701	0,985424	0,993805	1				
ind_turis	0,728747	-0,12262	0,507308	0,332947	0,455946	0,708966	0,724828	0,724208	1		ı	
ind_actindex	0,982403	-0,3152	0,260241	0,114718	0,232879	0,985016	0,98338	0,982789	0,677875	1		ı
unemprate	0,207333	0,001956	0,165843	0,330404	0,462124	0,168786	0,198614	0,166047	0,397764	0,085432	1	
adsl	0,313019	0,071507	0,447176	0,416765	0,550463	0,303411	0,325251	0,315196	0,603665	0,314178	0,129349	1
pcars	0,011963	-0,00905	0,107438	-0,02653	0,044023	0,008317	-0,00883	0,037199	0,19627	0,010701	-0,18828	-0,00125
pbanks	0,094124	-0,36116	0,01112	-0,24626	-0,35901	0,094671	0,104435	0,139567	0,167962	0,113508	-0,45996	0,152598
penergy	-0,3627	0,356798	-0,02168	-0,00585	0,000633	-0,3253	-0,33641	-0,36781	-0,30644	-0,28917	-0,2501	-0,11576
pmanufac	0,217164	-0,28064	-0,22251	-0,18328	-0,3128	0,207414	0,210012	0,196856	-0,02886	0,147541	0,168413	-0,33371
U_Mining &	0 007007	0.00006	0.00674	0.4077	0.0004.0	0.247024	0.000446	0.005440	0.00700	0.450000	0.46045	0.00404
Metal	0,227037	-0,29026	-0,22671	-0,1877	-0,30913	0,217921	0,220446	0,206113	-0,02739	0,158832	0,16015	-0,32194
pbuilding	-0,17058	0,246504	0,251287	0,194366	0,335301	-0,17257	-0,16901	-0,14326	0,145233	-0,12599	-0,06705	0,349552
pagric	-0,6176	0,2851	-0,20246	-0,13442	-0,53494	-0,63359	-0,62923	-0,62432	-0,54771	-0,6014	-0,25125	-0,42498
ptextile	0,640904	-0,22043	0,235618	0,141322	0,484756	0,659294	0,650672	0,647769	0,525183	0,627745	0,246648	0,351986
ppharmac 	0,500605	-0,36226	0,079558	0,092935	0,579473	0,512084	0,511936	0,49951	0,509371	0,487227	0,219327	0,496717
pdurab	0,213989	-0,03587	0,124842	0,167121	0,627219	0,202	0,214909	0,202498	0,445363	0,161036	0,352402	0,493796
pinterind	-0,26078	0,163706	-0,10992	-0,06179	-0,38233	-0,24008	-0,26725	-0,25563	-0,40381	-0,22311	-0,29648	-0,28662
pretail_food	-0,2501	0,214636	-0,09704	-0,09678	-0,3215	-0,21992	-0,22884	-0,23747	-0,45677	-0,17272	-0,4332	-0,12224
_pretail_nonf	0,4007	-0,28701	0,198749	0,167181	0,376234	0,369764	0,380898	0,393133	0,48545	0,341346	0,263234	0,248008

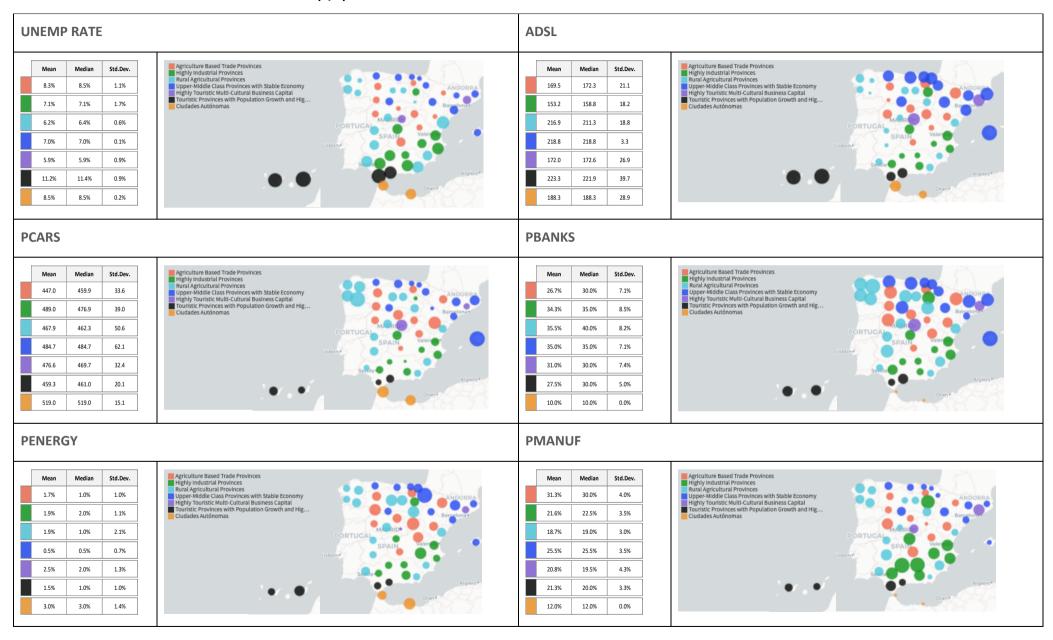
CORRELATION OF VARIABLES (2/2)

					U_Mining								
	pcars	pbanks	penergy	pmanufac	& Metal	pbuilding	pagric	ptextile	ppharmac	pdurab	pinterind	pretail_food	pretail_nonf
population													
pmale													
pforeign													
pop inc% (1)													
pop inc% (2)													
ind_wholesale													
ind_retail													
ind_rest													
ind_turis													
ind_actindex													
unemprate													
adsl													
pcars	1												
pbanks	0,396057	1		ı									
penergy	0,007863	-0,13544	1										
pmanufac	-0,35942	0,019662	-0,36701	1									
U_Mining &													
Metal	-0,37528	0,02341	-0,37074	0,999127	1								
pbuilding	0,46417	-0,01448	0,206219	-0,91328	-0,92461	1		ı					
pagric	-0,12607	0,108613	0,223521	0,025005	0,021117	-0,14103	1		l				
ptextile	0,130593	-0,14062	-0,19562	-0,00129	0,001192	0,114594	-0,97396	1		l			
ppharmac 	0,083262	-0,04119	-0,24434	-0,04823	-0,04084	0,126622	-0,84328	0,708722	1		ı		
pdurab · · · ·	-0,00257	-0,08533	-0,08477	-0,08702	-0,08359	0,19041	-0,52973	0,445106	0,621415	1		ı	
pinterind	-0,18675	-0,11139	0,050193	0,086263	0,089065	-0,17868	0,382123	-0,34526	-0,39873	-0,46523	1		
pretail_food	-0,27878	0,009266	0,292278	0,041476	0,049337	-0,18502	0,360782	-0,33432	-0,32748	-0,3384	0,207139	1	
pretail_nonf	0,269193	0,113745	-0,40613	0,025331	0,027249	0,056736	-0,48391	0,458489	0,444972	0,342491	-0,2529	-0,82397	1

DISTRIBUTION OF VARIABLES WITHIN CLUSTERS (1/3)



DISTRIBUTION OF VARIABLES WITHIN CLUSTERS (2/3)



DISTRIBUTION OF VARIABLES WITHIN CLUSTERS (3/3)



CLUSTERS AND LIST OF PROVINCES

Highly Industrial Provinces	AlbaceteAlicanteCórdoba	 Granada Jaén La Rioja	SevillaToledoValencia	
Rural Agricultural Provinces	AlmeríaÁvilaBadajozBurgos	CáceresCastellónHuelvaLa Coruña	LugoMurciaOurensePalencia	PontevedraSoria
Upper-Middle Class Provinces with Stable Economy	Asturias Baleares Cantabria	GironaGuipúzcoaLleida	NavarraTarragonaValladolid	VizcayaZaragoza
Highly Touristic Multi- Cultural Business Capitals	Barcelona Madrid			
Agriculture Based Trade Provinces	Álava Ciudad Real Cuenca	GuadalajaraHuescaLeón	SalamancaSegoviaTeruel	• Zamora
Touristic Provinces with Population Growth and High Unemployment	Cádiz Las Palmas Málaga	• Tenerife		
Ciudades Autónomas	• Ceuta • Melilla			