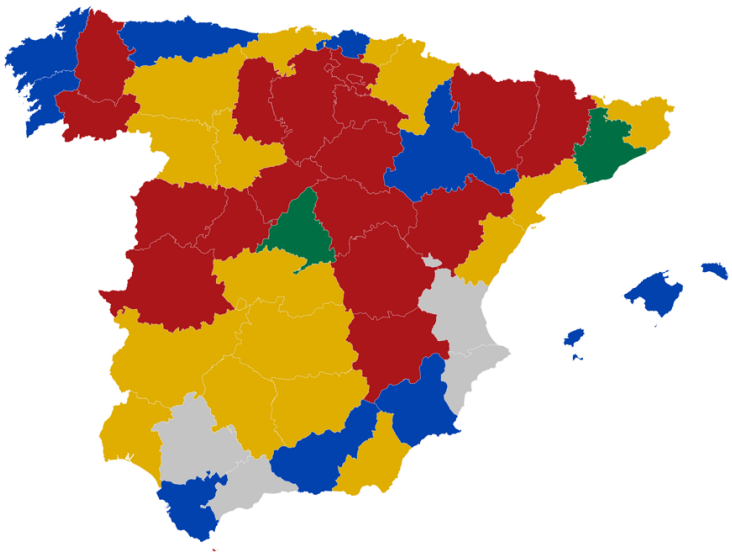
- Strictly Confidential -
Spanish Ministry of Economics

PROVINCES SEGMENTATION



TEAM C

Eun Suk Hong, Eleonora Jimenéz, Qing Loh,
Albert Sehnaoui, Julius von Selchow,
Wilhelm Stahlke IV, Eugen Wettstein

MACHINE LEARNING GROUP ASSIGNMENT
MBD 2020-A

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

At the request of the Spanish Ministry of Economy, Team C has prepared an in-depth analysis of Spain's fifty-two provinces and segmented them across economic and demographic lines. The provided data set, Provinces.xlsx, contains the empirical evidence analyzed. This report will provide a series of regional policy suggestions to best guide the Ministry in applying new policies to improve economic performance and diversify the Spanish economy from an industry and geographic perspective.

The insights gleaned through segmentation led the group to discover five clusters (*See: ANNEX 1, ANNEX 2*) with similarities that underline the basis of our policy suggestions. The main tool employed was Dataiku which utilizes unsupervised learning methodology to find provincial similarities and group them accordingly. The analysis was executed by applying several clustering methods, chief of which was the K-Means approach, supported by the Gaussian Mixture as well as interactive clustering. The latter achieved a deeper output than the Gaussian Mixture. However, the K-Means clustering performed the best and therefore leads the analysis. Through applying the silhouette score, the most relevant clusters can be identified. The silhouette score ranges from negative one to positive one, with a higher value indicating an increasingly well-matched fit of object to cluster compared to other clusters. In this scientific study, five distinct clusters were identified, with the highest silhouette result at 0,5966.

The overarching goal is to identify provinces akin to one another for the purpose of unique policy implementation. Thus, the team delved deep into the clusters generated by Dataiku to rigorously verify observed trends such as population growth/decline, unemployment rates as well as economic drivers such as high levels of tourism and restaurants & bars. A more granular insight of each cluster can be found in the Findings and Conclusion section, where the increase/decrease of each variable is presented in more detail. This exercise assisted in determining if the suggested proposals fit within the context of the Spanish economy.

Team C recommends the Ministry to focus on **Cluster_2 - *Coastal High Potential Areas*** for a combination of maximum economic benefit and highest likelihood of success. The recommendations will be further explored in Section 3, but centre on a plan to attract large companies as well as a start-up culture to the region as a counterbalance to **Cluster_3 - *Rural Areas w/ Stable Income***. The proposed economic policies' ultimate goal is to generate sustainable employment and to target high potential professionals to settle in the regions.

II. DATA ANALYSIS APPROACH

The first step towards cluster creation was review the given data, its descriptive analytics, and the correlation amongst each other. The following was observed:

Outlier handling: Each field's descriptive statistics led the team to discover that the outliers of each field were not actually outliers, but rather the reality of the data distribution across the diverse country. Therefore, they were not excluded.

Unit of measure: The given variables had different units of measure with considerable distance between each other. Consequently, some standardization was considered on the experimental face to determine how the model responded.

Correlation: There was a high correlation between the variables Wholesale Trade Index and Retail Trade Index with other variables in the model (*See: ANNEX 5*). This insight provided possible candidates to eliminate in the experimental phase.

The model was then implemented using the core data by first defining the clustering algorithms for experimentation. Out of the various segmentation analysis algorithms, team C selected K-Means, an extensively used technique for data cluster analysis that uses an iterative refinement method. The decision was reinforced not only by its popularity but also by the algorithim's effectiveness and local fine-tuning capability.

K-Means also has its drawbacks. In this situation, there was concern surrounding sensitivity to outliers and potentially unusable clusters. Through multiple rounds of expirmimentation, it was discovered that the extreme data points were beneficial as a weight factor in cluster creation. For example, Madrid and Barcelona were joined when considering absolute population. Connecting this with the other variables resulted in a logical cluster that merits its own economic policies.

Supplementary experiments using K-Means assisted in determining the optimal outcome for the economic case. Running iterations of the model revealed the ideal number of clusters by using the right variable mix. Additionally, the Wholesale Trade Index and Retail were eliminated as they had minimal impact on the model and were highly correlated with other variables.

Another experiment accounted for different units of measure. Despite theoretically favouring variable standardization, this was subjected to a different interpretation when comparing the dataset's clusters between standardized and non-standardized. The standardized clusters were not easily interpretable within the context of the economic goals and did not result in a logical output. For example, when forgoing standarization of the population variable, the output demonstrated provincial economic behaviors and the recommended policies were thus tailored accordingly.

In total, the model was run approximately ten times. The final selection was a model with a silhouette of 0.5966 trained with twenty-three variables and producing five clusters. The clusters' unique characteristics support their groupings. Moreover, a low variance is observed on the most relevant variables within each cluster as compared to their global variance. This demonstrated an adequate separation between the clusters and homogeneity amongst the cluster members. The decisive factor in selecting this model was it's easy interpretation that permits an understanding of each province's role in the Spanish Economy.

III. FINDINGS AND CONCLUSION

Having established the confidence by combining the two methodologies, descriptions of the five clusters are as follows (*See: ANNEX 1, ANNEX 2*):

Cluster_1 – *Metropolitan Areas* – This cluster consists of the country's two largest provinces: Madrid and Barcelona. With a slightly lower unemployment rate (-3%), the cities have the highest absolute population numbers and diverse industries such as banking, building, pharmaceuticals, etc. Furthermore, as is characteristic of the national economy, tourism and restaurants & bars exceed the national average by 445% and 615%.

Cluster_2 – *Coastal High Potential Area* – Situated in the south/southwest of the country the provinces of Valencia, Sevilla, Malaga and Alicante are known to draw tourists from around the world. Industries such as tourism (+135%) and restaurant & bars (+124%) are supported in great part by foreign residents (+40%). Seasonality and a lack of another dominant industry contribute to a higher than average unemployment rate (+28%) and likely precipitated the significant population decline in Time 2 (-152%). Only textile (+94%) and manufacturing (+19%) stand out, but are expected for the next largest absolute population after Cluster_1.

Cluster_3 – *Rural Areas w/ Stable Income* – The provinces that comprise this group are less populated compared to other parts of Spain and are mostly inland with the exception of Ceuta and Melilla. The region has markedly low tourism figures (-86%) but its relatively low unemployment (-16%) is due in part to its industrial activity underpinned by energy (+42%), inter-industry trade (+10%) and agriculture (+5%).

Cluster_4 – *Semi Developed Residential Areas* – Geographically, this region is characterized by Spanish Islands (Baleares, Canarias) and larger provinces all along coastal areas with the exception of Zaragoza. The region experienced a negative 190% dip during the second population growth rate (2011-2019) in comparison to the whole data set, which likely caused a higher than average unemployment rate (+15%). The

region's economy is currently underpinned by tourism and restaurants & bars, which exceed the national average by 90% and 28% respectively.

Cluster_5 – *Rural Areas w/ Less Stable Income* – This grouping is an amalgamation geographically, but has important trends that unite the provinces. The population is relatively stable but from an employment perspective has seen decreased activity in tourism (-47%), food & bars (-37%) and textiles (-30%).

IV. RECOMMENDATIONS

The main goal from this analysis is to reduce the unemployment rate for **Cluster_2 - *Coastal High Potential Areas*** by leveraging the existing infrastructure and popularity of the provinces to create opportunities that will propel these provinces forward to compete with **Cluster_1 – *Metropolitan Areas***. Provinces in this cluster are: Valencia, Alicante, Sevilla, Málaga and Murcia, all of which have good infrastructure (public transport, airports, train stations, bus stations, etc.) and existing reputation as being top tourist destinations amongst locals and tourists. To reduce the unemployment rate, we need to re-brand these provinces as the new up and coming areas for young professionals and young families to reallocate. Recommendations as follows:

Firstly, invest in the push of high employable companies. Encourage traditional conglomerates to move to these provinces by offering tax breaks, cheaper land, etc. These types of companies are ideal because they can provide a lot of quality jobs appealing to young professionals interested in personal development and career advancement. Fortunately, this cluster has already attracted quite a few companies (see % Manufacturing Industry, % Pharmaceuticals, etc. in *ANNEX 3*). With government encouragement, it will not be a difficult task to attract more employable companies to this cluster.

Secondly, subsidize start-ups to attract a younger population. Start-ups encourage innovation and are particularly popular among younger generations. Start-ups also have a faster growth rate than traditional companies which can help move the gentrification process of these provinces along faster.

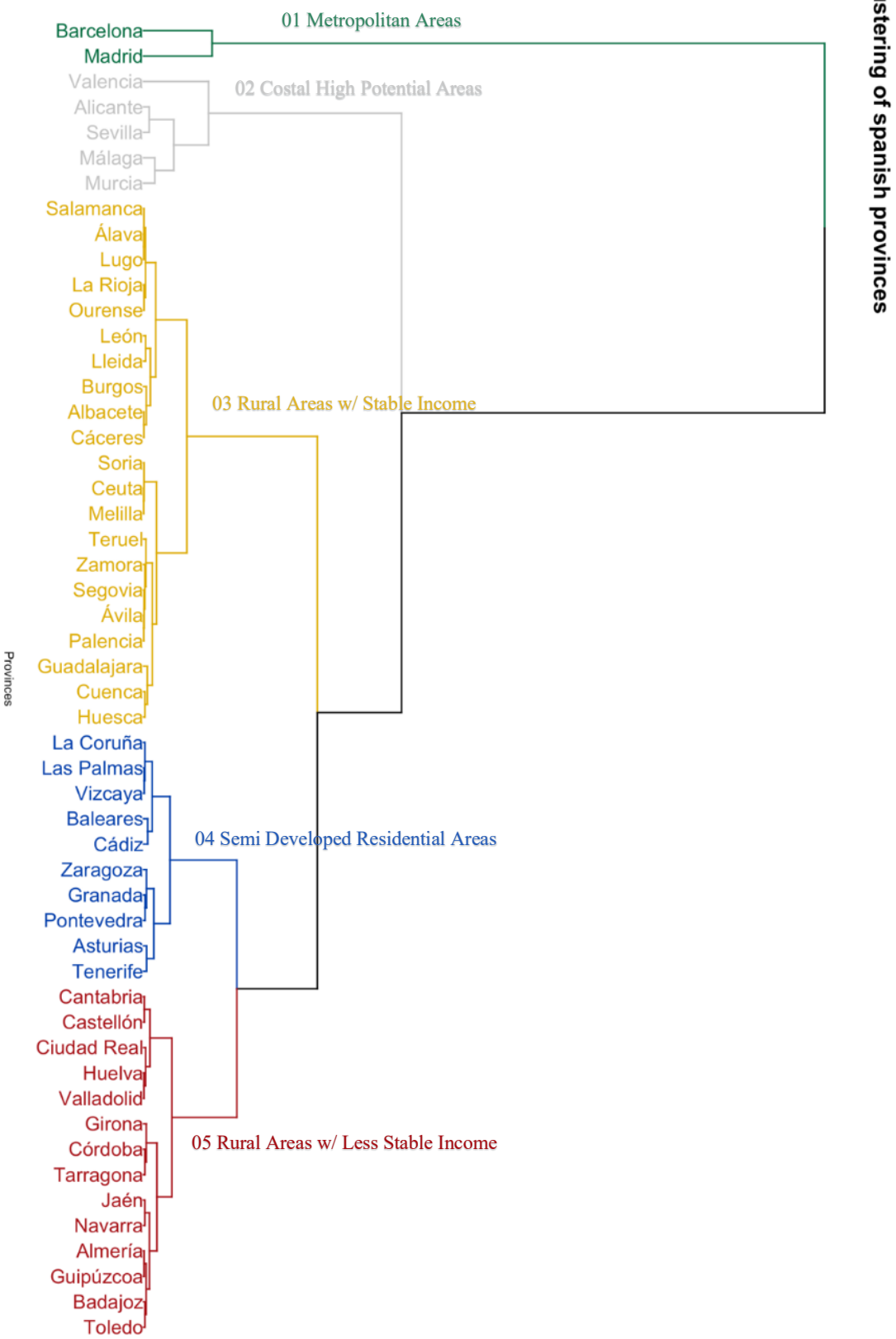
Thirdly, further optimize the quality of life in these provinces. With the predicted influx of young professionals, there needs to be steps taken to improve the liveability in this cluster. This includes more residential options both in central cities as well as suburban areas, inner-province public transportation, better facilities, etc. Proper planning and strategic implementation is needed to support the new developments planned.

The desired outcome is that these provinces will expand their source of revenue stream, which is currently limited to their tourism seasons. It will transform them from a simple tourist destination to economic hubs that encompasses multi-industries and sectors. As more companies open in **Cluster_2 - *Coastal High Potential Areas***, younger generations will have more options than just Madrid and Barcelona as cities with better job prospects and career opportunities. The final determining evaluation method will be a decrease in unemployment rate for this cluster.

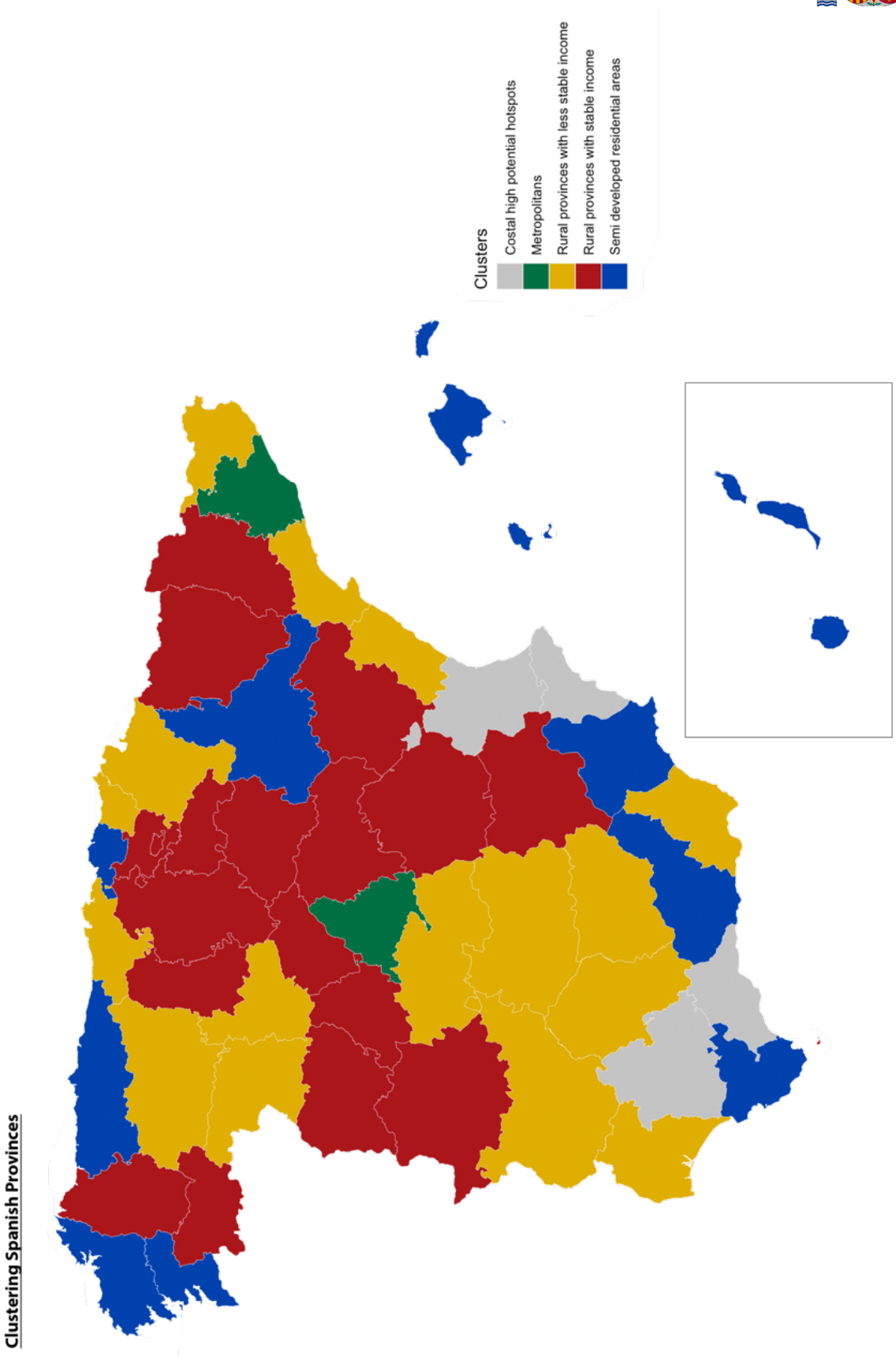
V TECHNICAL ANNEX

- Strictly Confidential -
Spanish Ministry of Economics

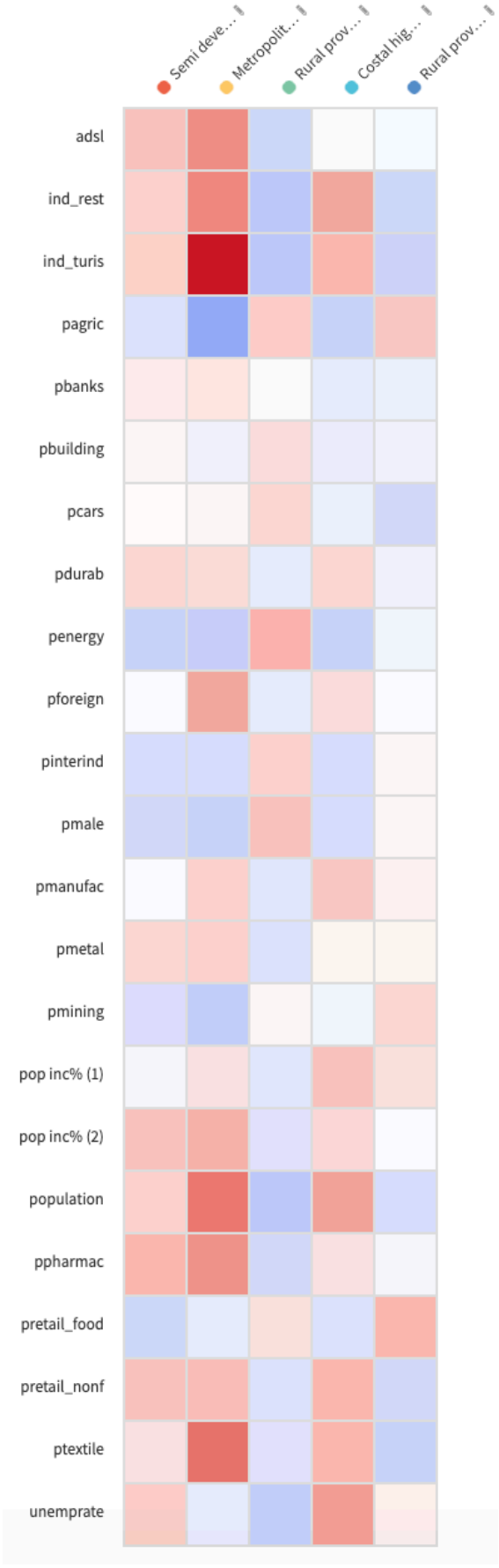
ALL OF THE FOLLOWING VISUALISATIONS ARE PRODUCED BY TEAM C IN THE STATISTICAL TOOL R (ALL RIGHTS RESERVED ®) AND ONLY VALID IN RELATION WITH THIS DOCUMENT



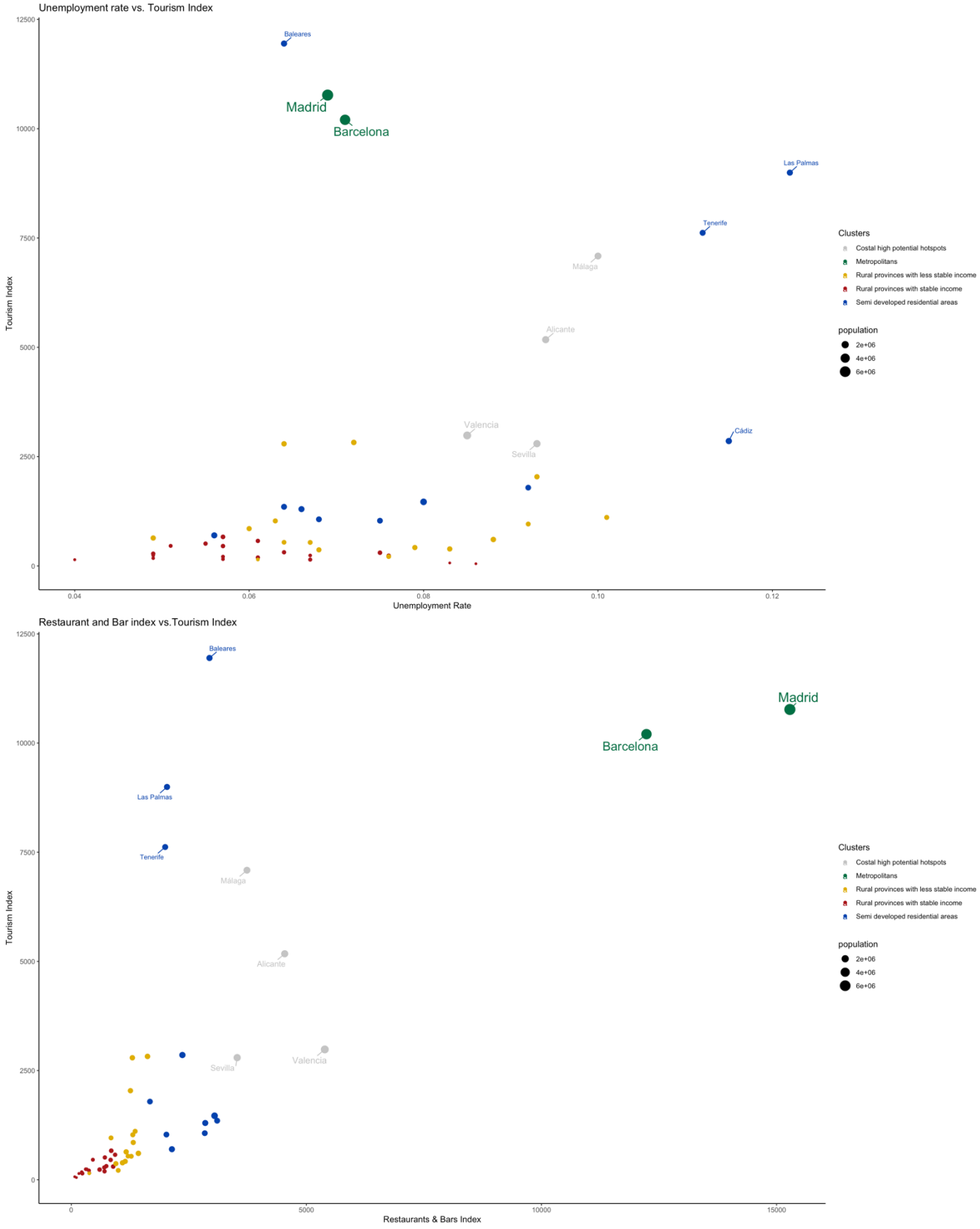
Annex 1: Dendogram of different Clusters



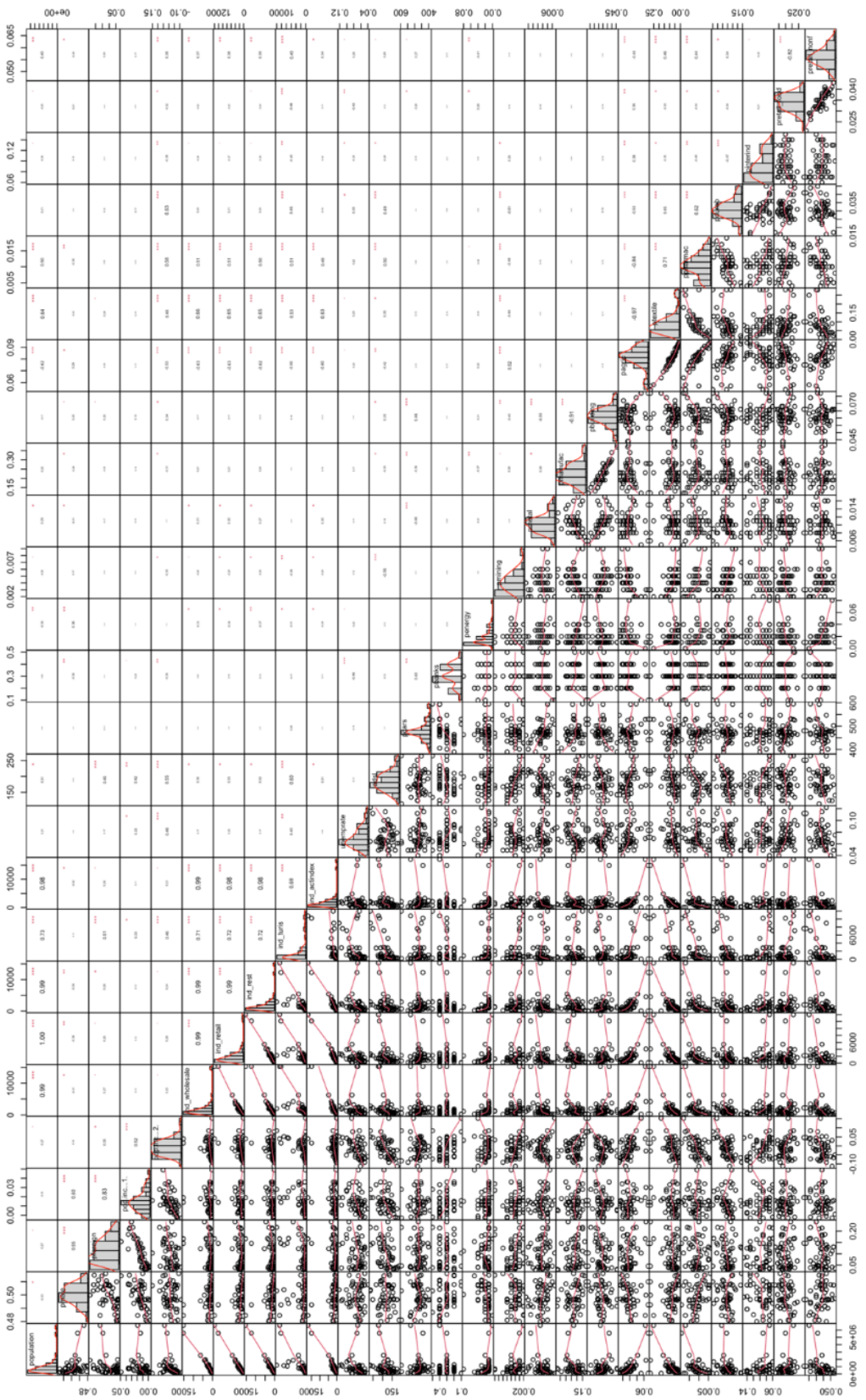
Annex 2: Spain's Map w/ Provinces in the colours of the 5 clusters
Produced by Dataiku, especially coded & constructed in R



Annex 3: Heatmap for correlation of clusters with variables



Annex 4: Scatterplot of cluster showing: population size, tourism index, restaurant&bars index, unemployment rate and relevant provinces labelled



Annex 5: Correlation Map showing correlation variables and their density respectively