

# Statistical Modelling of High-Dimensional Data

Multidimensional scaling of Spanish provincial employability



facultade de  
informática  
da coruña

This study has been carried out in its entirety by *Jorge Crespo Rivas* ([j.crespo.rivas@udc.es](mailto:j.crespo.rivas@udc.es)) and *Jesús Estévez Amoedo* ([j.esteveza@udc.es](mailto:j.esteveza@udc.es)), students of the Bachelor's Degree in Data Science and Engineering at the University of A Coruña and members of internship group number 3 of the subject Statistical Modelling of High-Dimensional Data.

## Contents.

<b>INTRODUCTION.</b> .....	3
<b>EXERCISE 1.</b> .....	5
<b>EXERCISE 2.</b> .....	10
<b>EXERCISE 3.</b> .....	11
<b>EXERCISE 4.</b> .....	11
<b>EXERCISE 5.</b> .....	11
<b>EXERCISE 6.</b> .....	14
<b>EXERCISE 7.</b> .....	15
<b>EXERCISE 8.</b> .....	16
<b>EXERCISE 9.</b> .....	17
<b>EXERCISE 10.</b> .....	17
<b>EXERCISE 11.</b> .....	18
<b>EXERCISE 12.</b> .....	19
<b>VISUAL REFERENCES.</b> .....	19

## INTRODUCTION.

The study of employability at the subnational level is crucial for understanding labor dynamics, territorial differences in economic opportunities, and the effectiveness of public policies. Analysing how demographic and economic indicators vary between provinces allows us to identify structural patterns (population mobility, levels of activity and unemployment, and differences in macroeconomic performance) that are not observable through partial analyses and that, therefore, require rigorous multivariate treatment.

For this work, a provincial database corresponding to the year 2017 is used, extracted from the National Institute of Statistics and available in the work file. The variables considered summarize demographic and labour market aspects that, when treated together, allow us to capture regional heterogeneity and the complex relationships between the demographic, labour, and economic dimensions, aspects that are not evident through isolated univariate analyses.

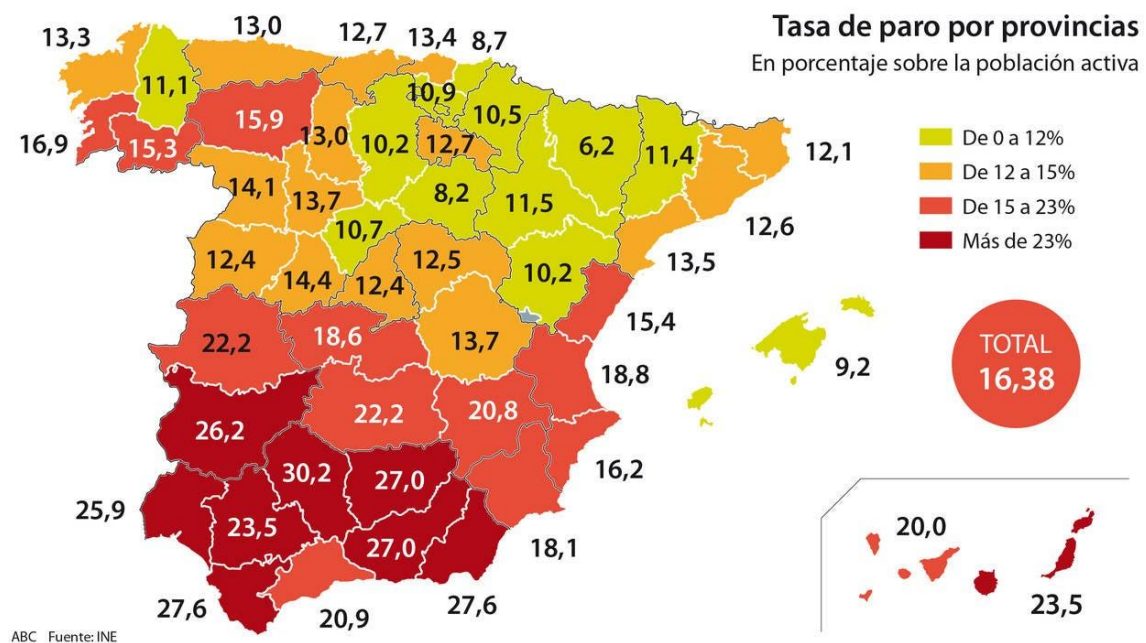


Figure 1: Unemployment rate by province as a percentage of the labour force in 2017. [1]

The central objective of this work is to find a representation in a Euclidean space of reduced dimension in which each point represents a province and the distances between points approximate, as far as possible, a reasonable measure of dissimilarity between the observations. This approach corresponds to the family of multidimensional scaling methods (MDS) or principal coordinates, whose purpose is to facilitate the visualization and interpretation of similarity/difference structures in multivariate data. In addition, the work will address relevant theoretical and practical issues: the possible non-uniqueness of the solution (due to rotations and reflections in space), the effect of unit changes in the variables on the final configuration, and the choice of the representation dimension that balances the result with respect to the dataset. It will also include the graphic representation of the configurations (with identifying labels by province), the exploration of the need for rotations or symmetries to facilitate interpretation, and the extension of the procedure to represent, alternatively, the variables themselves in the same geometric frame.

As such, this research combines a descriptive and computational approach with a rigorous theoretical foundation (always based on the subject matter covered in notes and classes of the subject) to evaluate whether the spatial representation obtained is useful to capture relevant patterns in provincial employability and what limitations emerge from the applied methodology. Throughout this report, the methodological decisions, intermediate calculations, and interpretations derived from the final graphs will be documented, allowing a critical assessment of the applicability of the method to the set of data considered.

## Evolución del paro por provincias

Puntos de diferencia de la tasa de paro de la EPA en el segundo trimestre de 2010 a 2017 comparada con la de 2009

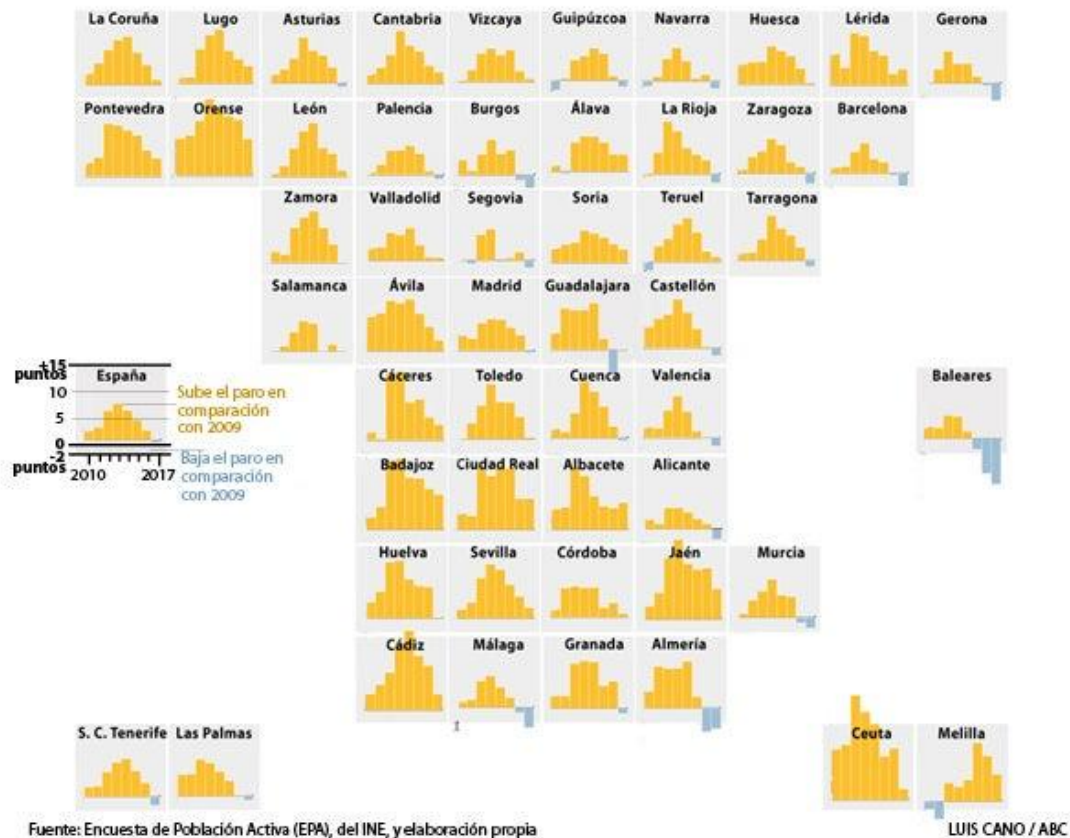


Figure 2: Evolution of differential points in the unemployment rate 2010-2017 vs. 2009. [2]

## EXERCISE 1: Carry out a descriptive study (univariate and multivariate) of the variables considered to be of interest.

### UNIVARIATE DESCRIPTIVE STUDY

To carry out this exercise, we will first visualize the first rows of the data with which we will work.

```
> head(employment)
      inm.1000  em.1000  activos  empleados  en.paro    ipc    pib
Albacete      9.658594 13.084991   57.16    41.55   27.30 101.391 17383.24
Alicante/Alacant 10.258191  9.872011   57.35    43.07   24.91 101.047 17562.79
Almería       12.998179 11.081435   63.52    40.85   35.70 101.246 17287.07
Araba/Alava    15.118920 11.433495   61.09    50.93   16.64 100.306 34053.30
Asturias       6.237551  7.617976   51.63    40.90   20.78 100.756 19505.32
Ávila         16.740323 20.871446   53.30    39.80   25.33 101.999 17622.24
```

Figure 3: Primeras filas del dataset “employment” asignado.

After executing the head function, we observe that the database consists of 7 variables, all of them of a numerical nature (quantitative), being represented in m. 1000 and 1000 per thousand (%), on the other hand, active, employed, and unemployed per hundred (%), and, finally, CPI and GDP indices/indicators.

We will also analyze the summary of the analyzed data.

```
> summary(employment)
      inm.1000  em.1000      activos      empleados      en.paro
Min.   : 6.238   Min.   : 7.080   Min.   :48.16   Min.   :32.78   Min.   :14.05
1st Qu.: 8.816   1st Qu.: 9.094   1st Qu.:55.49   1st Qu.:39.96   1st Qu.:18.42
Median :10.342   Median :11.860   Median :57.75   Median :43.05   Median :23.17
Mean   :11.387   Mean   :12.617   Mean   :57.92   Mean   :43.90   Mean   :24.19
3rd Qu.:13.210   3rd Qu.:14.097   3rd Qu.:60.84   3rd Qu.:48.00   3rd Qu.:29.37
Max.   :27.724   Max.   :26.380   Max.   :65.84   Max.   :53.26   Max.   :42.34

      ipc      pib
Min.   : 99.83   Min.   :14957
1st Qu.:100.72   1st Qu.:17518
Median :100.93   Median :19016
Mean   :100.95   Mean   :20832
3rd Qu.:101.18   3rd Qu.:24376
Max.   :102.00   Max.   :34053
```

Figure 4: Sumario del dataset “employment”.

To carry out the study, we have decided to discard both GDP and CPI as useful variables. The information provided by these indicators could be convenient in other contexts where there is more information on each province, such as population or the average cost of housing, but not in this case.

Analyzing in isolation the variables represented per thousand (using the barplot function), we obtain the following results:

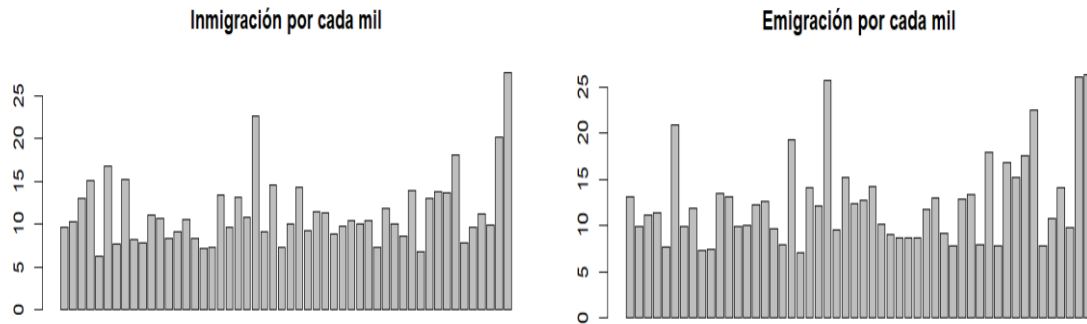


Figure 5: Barplot of the variables immigration and emigration.

From this analysis, it can be seen that, both in terms of immigration and emigration, most provinces do not exceed 20 inhabitants per thousand inhabitants, although emigration tends to be significantly higher. Some atypical provinces show significantly higher than average levels of immigration and emigration; in no case are lower figures recorded, which is consistent, given that rates below 5 per thousand would be extraordinarily low. In practice, the prevailing trend places the vast majority of provinces in an approximate range of 5 to 15 inhabitants per thousand.

Alternatively, and through the use of the boxplot function, we have also analysed the variables represented in percentage in isolation, thus obtaining the following graphs:

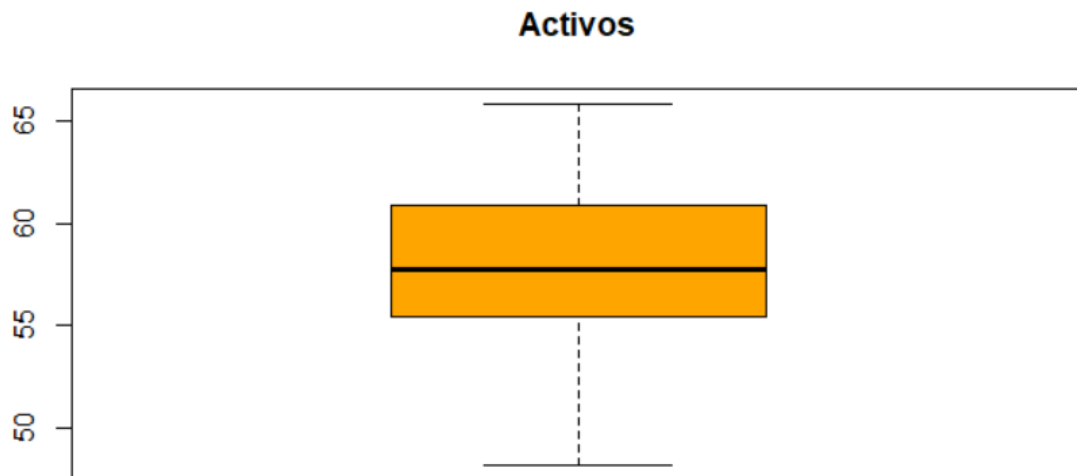


Figure 6: Boxplot de población activa.

For the active, the median is approximately 57.5% and the cash is narrow (covering only 5% of the total), which indicates that the vast majority of provinces have more than half of the active population.

In addition, we have not found atypical, and the mustaches are barely far from the box, which suggests that there are no provinces with a very low or high amount of assets compared to the rest.

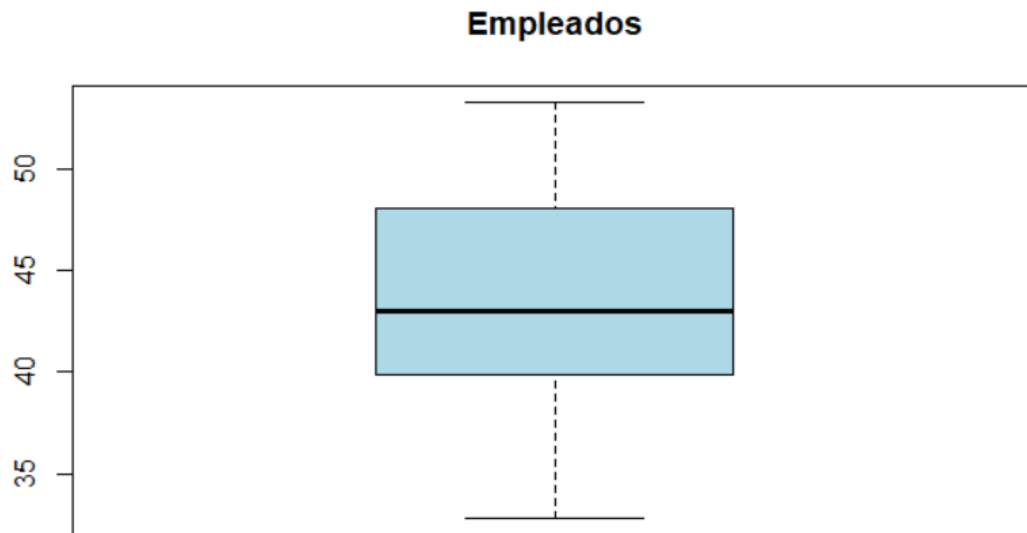


Figure 7: Boxplot of employed population.

For employees, the median with respect to cash is the most asymmetrical of the three analysed (around 43%), although not enough to highlight or take it into account. The whiskers in this case are not particularly distant either.

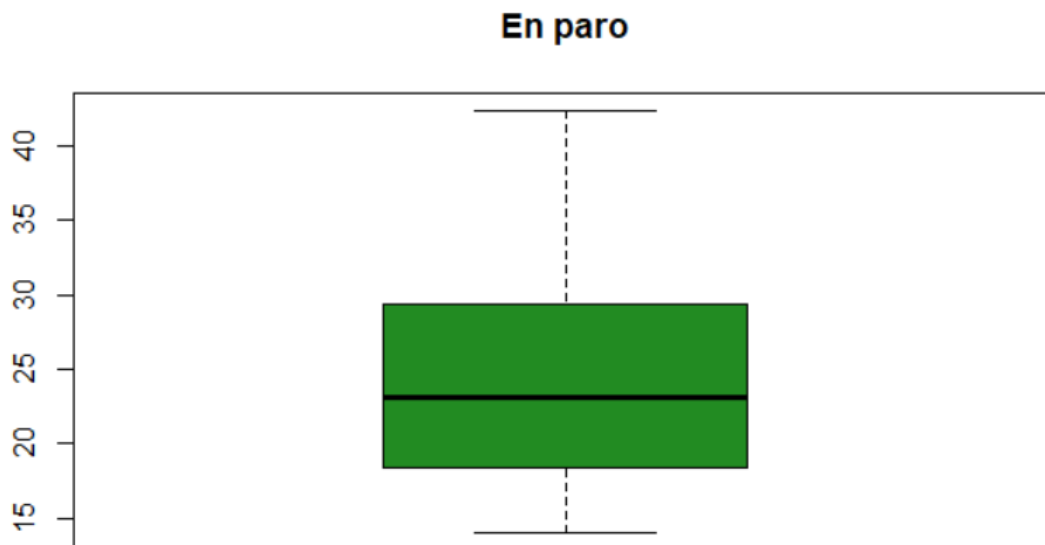


Figure 8: Boxplot of the unemployed population.

In relation to unemployment, the cash register is also narrow and the median symmetrical as in the previous cases. This indicates that most of the data (50% of the observations) are contained among these values, which suggests that the unemployment rate in most provinces is not extremely low or excessively high.

The upper moustache is quite far from the box, which indicates that there are some provinces with a high unemployment rate, and that unemployment in most provinces is a moderate issue.

All these data have been analyzed by performing a univariate descriptive analysis, which is very limiting when interpreting the results. By studying the variables univariately, it is possible that relevant information is being lost that could explain patterns or behaviours (such as unemployment in a region, which could depend on factors that are not reflected if the unemployment rate is only observed in isolation).

Then, given these circumstances, the multivariate descriptive study will be carried out.

## MULTIVARIATE DESCRIPTIVE STUDY

To analyse these data in a multivariate way, we have decided that it could be interesting to make a graph comparing the difference between immigrants/emigrants with the unemployment rate.

After looking at the graph, we see no noticeable type of pattern or relationship between these differences, so, by way of confirmation, we denote that the correlation between both variables is very small.

### Comparación entre Diferencia de Inmigrantes/Emigrantes y Tasa de Paro

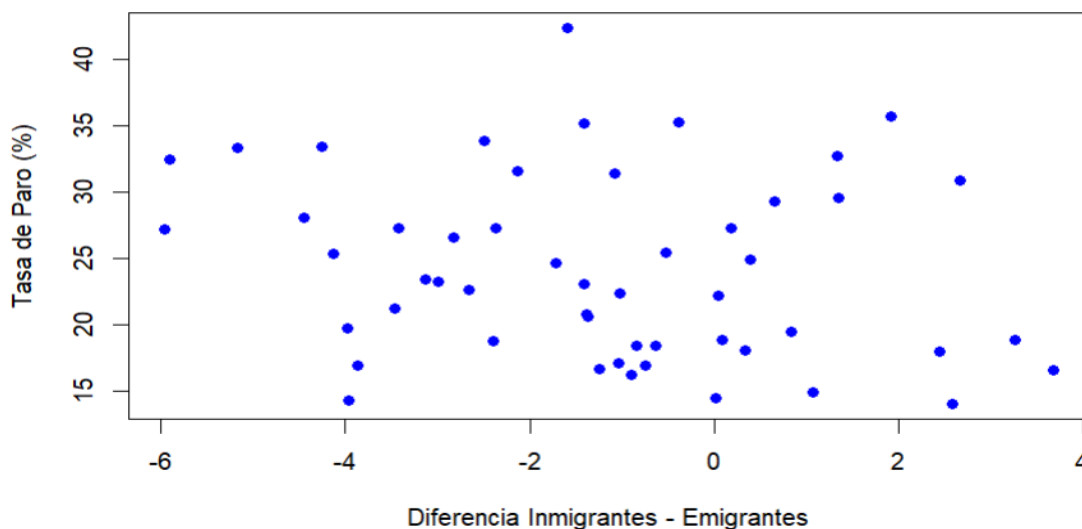


Figure 9: Graph in/ms vs paro tasa.

At first, we thought it would be a good idea to compare these 2 variables, as there could be indications that in the provinces where more people enter than leave, there would be an increase in the demand for work, so unemployment would increase.

As is coherent, the difference between immigrants and emigrants is also not explanatory with respect to the employability rate.

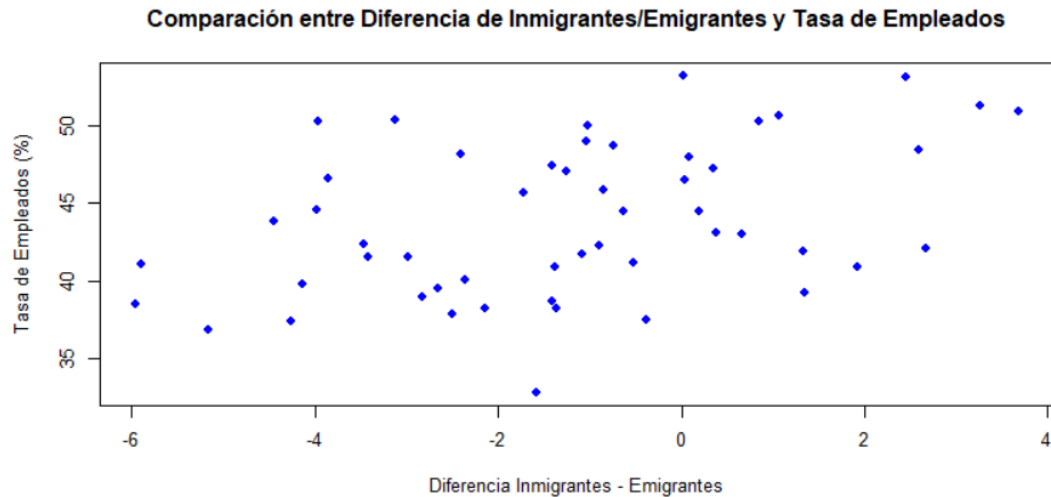


Figure 10: Inm/em vs employment rate graph.

It is because of this that we have decided to include GDP as a variable of interest and compare it with the employment and unemployment rates in search of a relationship. A coherent and expected result from this analysis would be that the provinces with the highest GDP would also have a higher employability rate and a lower unemployment rate, because this index measures the value of all the goods and services produced in the annual economy of each province.

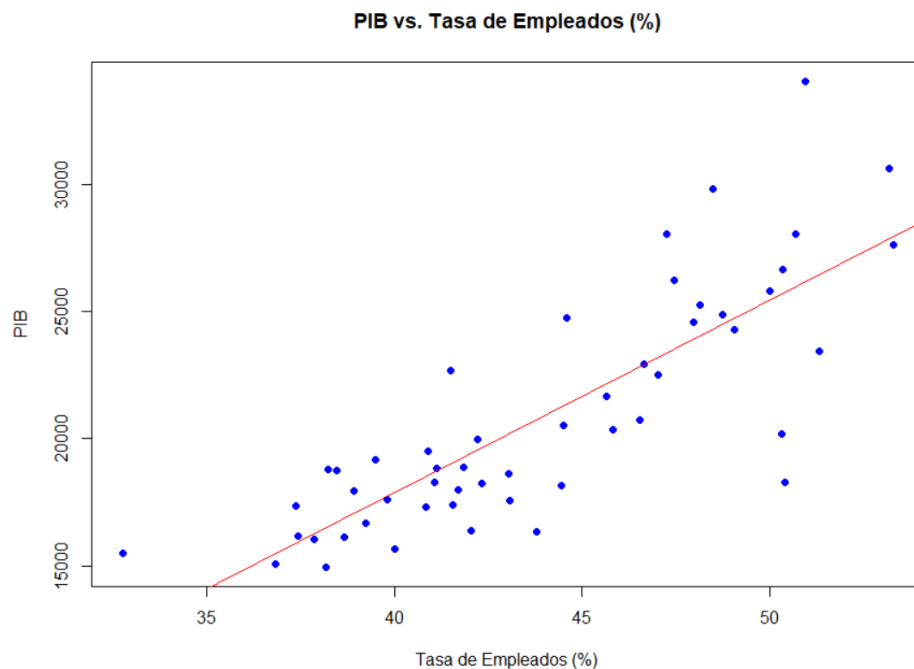


Figure 11: GDP vs employment rate graph.

Indeed, after analysing the graph, we denote an appreciable relationship between GDP and the employability rate. As a general rule, the more GDP a province has, the higher its employment rate, as we had previously estimated.

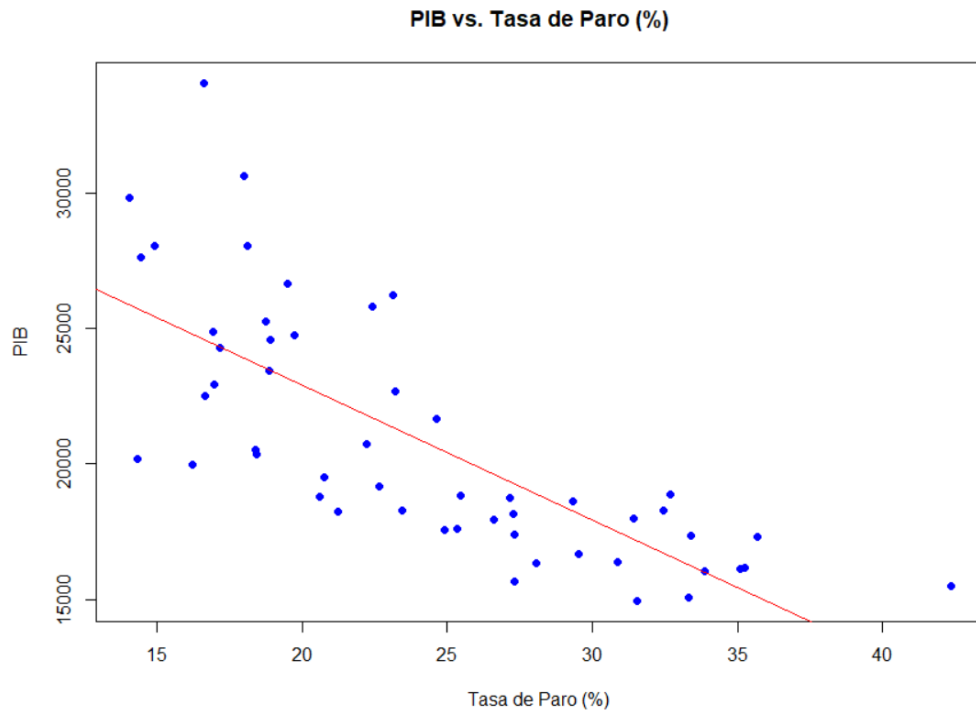


Figure 12: GDP vs unemployment rate graph.

As was also to be expected, lower GDP is generally associated with a lower unemployment rate. We conclude our multivariate study in this way.

## EXERCISE 2: Calculate the distance matrix of the sample observations.

For the calculation of the distance matrix, we have used the StatMatch library and the Mahalanobis distance. dist function of R, thus obtaining the following result:

```
> Distancias.emp
```

	Albacete	Alicante/Alacant	Almería	Araba/Álava	Asturias	Avila	Badajoz	Balears, Illes	Barcelona	Bizkaia	Burgos	Cáceres	Cádiz	Cantabria
Albacete	NA	2.120772	4.680367	4.941293	2.896628	2.833361	1.676642	3.887639	3.815114	3.401468	1.946732	2.213539	4.180445	1.8823911
Alicante/Alacant	2.120772	NA	4.650005	4.802153	2.866744	3.315578	1.830818	2.536960	3.953873	3.804122	3.133052	2.077203	4.558379	1.5264730
Almería	4.680367	4.650005	NA	5.670927	5.339424	5.346923	5.323894	5.778895	6.150243	5.248541	4.986773	5.636751	6.103339	4.6893884
Araba/Álava	4.941293	4.802153	5.670927	NA	4.998245	4.864594	5.170694	4.537181	4.530197	3.281017	3.874138	5.165275	5.068070	4.8888012
Asturias	2.896628	2.866744	5.339424	4.998245	NA	4.392432	3.279631	4.205888	3.414768	2.510000	2.959333	3.307752	4.325154	2.4624308
Avila	2.833361	3.315578	5.346923	4.864594	4.392432	NA	3.210215	4.812575	5.874794	4.875215	3.548651	3.768572	5.446231	3.2355638
Badajoz	1.676642	1.830818	5.323894	5.170694	3.279631	3.210215	NA	3.720847	4.188772	4.031083	3.199424	2.097324	3.696824	2.5297615
Balears, Illes	3.887639	2.536960	5.778895	4.537181	4.205888	4.812575	3.720847	NA	3.450813	4.226292	3.896507	3.087327	5.145998	3.3495556
Barcelona	3.815114	3.953873	6.150243	4.530197	3.414768	5.874794	4.188772	3.450813	NA	2.335999	2.939173	3.228781	4.046503	3.9946391
Bizkaia	3.401468	3.804122	5.248541	3.281017	2.510000	4.875215	4.031083	4.226292	2.335999	NA	2.164400	3.445513	3.999470	3.4353525
Burgos	1.946732	3.133052	4.986773	3.874138	2.959333	3.548651	3.199424	3.896507	2.939173	2.164400	NA	2.979867	4.505723	2.4686486
Cáceres	2.213539	2.077203	5.636751	5.165275	3.307752	3.768572	2.097324	3.087327	3.228781	3.445513	2.979867	NA	3.529411	2.6027443
Cádiz	4.180445	4.558379	6.103339	5.068070	4.325154	5.446231	3.696824	5.145998	4.046503	3.999470	4.505723	3.529411	NA	5.2863271
Cantabria	1.882391	1.526473	4.689388	4.888801	2.462431	3.235564	2.529762	3.349556	3.994639	3.435352	2.468649	2.602744	5.286327	NA
Castellón/Castelló	1.396358	2.533047	3.974411	4.279007	3.489442	3.062083	2.582139	3.932101	3.899864	3.218723	1.686063	3.246882	4.694173	2.0925249
Ciudad Real	1.909632	3.391345	5.270154	4.822178	3.967838	3.010933	2.060910	5.017798	4.623944	3.893437	2.834124	3.325672	3.630331	3.4609348
Córdoba	1.823541	2.679692	4.980056	4.912145	2.817727	4.026749	1.775185	4.028288	3.170341	3.079861	2.693929	1.972884	2.538608	3.0770914
Coruña, A	2.330520	2.047768	4.797944	4.892754	1.358476	3.934641	2.926897	3.600445	3.498093	2.794209	2.542937	2.391105	4.922801	1.2281218
Cuenca	2.432102	3.804714	5.927017	4.647375	3.582877	2.508783	3.057108	5.022357	4.593507	3.662165	2.597188	3.006881	3.970550	3.6201981

Figure 13: Mahalanobis distances between provinces.

Due to the large number of provinces and therefore the large size of the matrix, we have attached only the beginning of the matrix. In addition, we have included in the analysis the minimum and maximum distances of the least and most distant elements from each other, respectively.

```
> min_distancia
[1] 0.9165571
> max_distancia
[1] 7.074022
>
> which(Distancias.emp == min_distancia, arr.ind = TRUE)
      row col
Zaragoza  50  33
Navarra   33  50
> which(Distancias.emp == max_distancia, arr.ind = TRUE)
      row col
Melilla   52   3
Almería   3  52
```

Figure 14: Remaining operations of distances between provinces.

As we can see, using the distance of Mahalanobis, Zaragoza and Navarre are the provinces most similar to each other, while Melilla and Almeria would be the most distant.

### EXERCISE 3: Is there a single solution to the problem?

There is indeed another alternative, for example, Euclidean distance; however, its use would not be entirely appropriate. The variables have high correlations in cases such as inm 1000 and em 1000, in addition to containing variables with great variability, such as GDP. Euclidean distance directly measures the separation between two observations without considering correlations between them or differences in scale, which can be problematic in this context. On the other hand, the Mahalanobis distance is more suitable for these cases, as it takes correlations into account and offers a more robust treatment compared to outliers, which significantly affect the Euclidean metric.

### EXERCISE 4: Would the solution vary substantially if the variables were expressed in other units? How would it vary, if so?

If we focus exclusively on the Mahalanobis distance, we can affirm that the variations in the magnitudes of the variables do not significantly affect the analysis, since this metric incorporates in its formulation the correlations between the different variables. In this way, it allows a more balanced comparison between observations, regardless of the units of measurement or the scale of the data, which is especially useful in contexts where the variables have different orders of magnitude or high interdependence.

On the contrary, when using Euclidean distance, changes in the magnitude of the variables acquire a remarkable relevance. For example, if one of the variables (such as an index) is expressed in millions, its weight would dominate the rest, distorting the results. This type of imbalance can have important consequences in statistical and learning methods, such as cluster analysis or multidimensional scaling, since both depend directly on the distances between observations. Consequently, the choice of distance metric is a key aspect to guarantee the validity and interpretability of the results obtained.

### EXERCISE 5: Perform the procedure to find the requested point configuration, without using R's own function. How are the necessary matrices defined in each step?

The procedure we have developed follows a classic multidimensional scaling (MDS) analysis approach.

The Euclidean distance matrix between the scaled observations is calculated. This produces a matrix  $D$  of size  $n \times n$ , where  $n$  represents the number of observations. Each element  $D_{ij}$  represents the Euclidean distance between the observations  $i$  and  $j$ .

```
> D
```

	Albacete	Alicante/Alacant	Almeria	Araba/Alava	Asturias
Albacete	0.000000	1.160694	2.263781	5.386146	2.716571
Alicante/Alacant	1.160694	0.000000	2.431649	4.757566	2.130888
Almeria	2.263781	2.431649	0.000000	5.579795	4.338550
Araba/Alava	5.386146	4.757566	5.579795	0.000000	5.200729
Asturias	2.716571	2.130888	4.338550	5.200729	0.000000
Ávila	2.968059	3.773376	4.154654	6.623631	4.850359
Badajoz	1.406025	1.776401	2.568240	6.308361	2.624215
Balears, Illes	4.360542	3.581106	4.292057	2.421230	4.604191
Barcelona	5.088451	4.183050	5.193850	2.796876	4.369198
Bizkaia	4.106897	3.276356	4.822610	2.635941	3.008821
Burgos	2.798792	2.363892	3.674836	2.807083	3.274446
Cáceres	1.538067	1.318236	2.795855	5.372331	2.209579
Cádiz	3.748549	3.679257	3.455038	6.959178	4.038579
Cantabria	1.871607	1.391676	3.510317	4.332477	2.120377
Castellón/Castelló	1.605735	1.547757	2.349109	4.121993	3.229254
Ciudad Real	1.369973	2.248069	2.486465	6.361652	3.239400
Córdoba	1.927684	1.902619	2.409573	6.000681	2.658896
Coruña, A	2.361868	1.629135	3.945943	4.443705	1.231355
Cuenca	2.066844	2.822257	3.560046	5.965060	3.640602

Figure 15: First rows and columns of matrix  $D$ .

As in the previous sections, and taking into account the considerably large dimensions of the matrix obtained, it has been decided to present only a representative part of it. Specifically, only the first rows and columns are displayed, which allows you to clearly illustrate the structure and formatting of the data without overloading the document with excessive information. The same approach has been followed for the subsequent captures and representations included in this exercise, to maintain consistency in the presentation of the results and facilitate their understanding while ensuring the readability and relevance of the content displayed.

After this, we transform the distance matrix  $D$  into a matrix  $A$  using the formula  $-\frac{1}{2}D^2$ . This is based on the fact that we now work with matrices of scalar products rather than distances.

```
> A
```

	Albacete	Alicante/Alacant	Almería	Araba/Álava
Albacete	0.0000000	-0.6736048	-2.5623525	-14.505285
Alicante/Alacant	-0.6736048	0.0000000	-2.9564581	-11.317215
Almería	-2.5623525	-2.9564581	0.0000000	-15.567054
Araba/Álava	-14.5052846	-11.3172150	-15.5670535	0.0000000
Asturias	-3.6898803	-2.2703422	-9.4115087	-13.523789
Ávila	-4.4046879	-7.1191835	-8.6305754	-21.936245
Badajoz	-0.9884532	-1.5778009	-3.2979276	-19.897712
Balears, Illes	-9.5071631	-6.4121590	-9.2108772	-2.931178
Barcelona	-12.9461666	-8.7489534	-13.4880382	-3.911258
Bizkaia	-8.4333027	-5.3672540	-11.6287856	-3.474094
Burgos	-3.9166187	-2.7939934	-6.7522109	-3.939859
Cáceres	-1.1828246	-0.8688724	-3.9084019	-14.430968
Cádiz	-7.0258085	-6.7684642	-5.9686429	-24.215077
Cantabria	-1.7514558	-0.9683809	-6.1611630	-9.385180
Castellón/Castelló	-1.2891930	-1.1977759	-2.7591565	-8.495414
Ciudad Real	-0.9384134	-2.5269072	-3.0912550	-20.235308
Córdoba	-1.8579838	-1.8099789	-2.9030219	-18.004089
Coruña, A	-2.7892105	-1.3270406	-7.7852328	-9.873257
Cuenca	-2.1359218	-3.9825666	-6.3369645	-17.790971

Figure 16: First rows and columns of matrix A.

Next, the matrix H is created, which is a square matrix of size  $n \times n$  defined as  $H = I^T$ , where I is the identity matrix of size n, and 1 is the column vector of ones of size n.

```
> H
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.98076923	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[2,]	-0.01923077	0.98076923	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[3,]	-0.01923077	-0.01923077	0.98076923	-0.01923077	-0.01923077	-0.01923077
[4,]	-0.01923077	-0.01923077	-0.01923077	0.98076923	-0.01923077	-0.01923077
[5,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	0.98076923	-0.01923077
[6,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	0.98076923
[7,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[8,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[9,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[10,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[11,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[12,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[13,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[14,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[15,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[16,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[17,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[18,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077
[19,]	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077	-0.01923077

Figure 17: First rows and columns of the H-matrix.

The result obtained is considered coherent, since only two different values are identified in the entire matrix: one located on the main diagonal and another present in the rest of the positions. This setting confirms that the data is correctly centered.

Finally, the matrix B, also known as the Gram matrix, is obtained by applying the centering matrix H to the matrix A, according to the expression  $B = HAH$ .

```
> B
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 2.319516434 0.99469382 1.900465848 -5.550623051 0.54382305 3.44908505
[2,] 0.994693819 1.01708089 0.855142483 -3.013771217 1.31214333 0.08337164
[3,] 1.900465848 0.85514248 6.606120281 -4.469090071 -3.03450348 1.36649947
[4,] -5.550623051 -3.01377122 -4.469090071 15.589806600 -2.65494048 -7.44732677
[5,] 0.543823050 1.31214333 -3.034503480 -2.654940477 6.14789021 -1.99503126
[6,] 3.449085046 0.08337164 1.366499467 -7.447326771 -1.99503126 13.38802938
[7,] 3.303791335 2.06322592 3.137618882 -8.970322249 2.76317934 2.62023984
[8,] -3.944096862 -1.50031056 -1.504509031 9.267033258 -3.12203430 -6.65666694
[9,] -4.701270946 -1.15527551 -3.099840560 10.968783238 0.61413666 -13.03082413
[10,] -3.260144419 -0.84531348 -4.312325444 8.334209723 2.56084364 -8.54241169
[11,] -1.526944856 -1.05553735 -2.219235147 5.084960292 -1.05713682 -1.73142963
[12,] 1.365966422 1.02870082 0.783691083 -5.247031490 2.02185715 1.64581208
[13,] 2.171496233 1.77762277 5.371963751 -8.382627207 2.95643244 -1.56697835
[14,] 0.452394925 0.58425210 -1.814010382 -0.546184414 1.87003765 0.12685825
[15,] 0.573775445 0.01397483 1.247113885 0.002699737 -1.43688645 0.53670681
[16,] 3.757233392 1.51752185 3.747693678 -8.904516455 1.36297722 5.08962212
[17,] 2.345534803 1.74232191 3.443798582 -7.165425471 2.58284109 -0.38189674
[18,] 0.053295115 0.86424726 -2.799425245 -0.395606193 3.99857446 -2.19404064
[19,] 2.635929652 0.13806702 0.578188835 -6.383974590 0.05904861 9.13153361
```

Figure 18: First rows and columns of the Gram matrix.

### EXERCISE 6: What dimension does it make sense to consider for the representation of data? Reason the answer.

After obtaining the eigenvalues and calculating the explained variance (dividing each eigenvalue by the total sum of the eigenvalues, obtaining the proportion of the variance explained by each dimension, and thus standardizing the eigenvalues so that they add up to 1), we calculate the cumulative variance.

```
> lambda
[1] 1.534389e+02 1.022723e+02 6.017464e+01 2.790801e+01 9.866932e+00
[6] 3.302406e+00 3.683196e-02 1.935066e-14 1.153574e-14 1.130905e-14
[11] 1.128121e-14 6.393169e-15 6.050391e-15 4.874027e-15 4.795807e-15
[16] 4.307736e-15 3.506665e-15 3.262556e-15 3.090679e-15 2.738214e-15
[21] 2.314455e-15 1.987000e-15 1.917315e-15 1.753648e-15 1.326380e-15
[26] 1.074171e-15 7.053156e-16 5.275607e-16 -5.741470e-17 -1.296412e-16
[31] -3.162966e-16 -4.204168e-16 -7.272210e-16 -1.045842e-15 -1.116279e-15
[36] -1.274491e-15 -1.356267e-15 -1.386180e-15 -1.528852e-15 -2.382676e-15
[41] -2.422924e-15 -2.619286e-15 -2.691385e-15 -2.942264e-15 -4.503454e-15
[46] -5.432623e-15 -5.661511e-15 -5.884452e-15 -6.477130e-15 -8.302571e-15
[51] -9.912579e-15 -1.761415e-14
```

Figure 19: Lambda.

```
> var_exp
[1] 4.298008e-01 2.864771e-01 1.685564e-01 7.817371e-02 2.763847e-02
[6] 9.250436e-03 1.031708e-04 5.420352e-17 3.231299e-17 3.167800e-17
[11] 3.160004e-17 1.790804e-17 1.694787e-17 1.365274e-17 1.343363e-17
[16] 1.206649e-17 9.822592e-18 9.138811e-18 8.657365e-18 7.670068e-18
[21] 6.483068e-18 5.565825e-18 5.370630e-18 4.912179e-18 3.715351e-18
[26] 3.008883e-18 1.975674e-18 1.477761e-18 -1.608255e-19 -3.631407e-19
[31] -8.859850e-19 -1.177638e-18 -2.037034e-18 -2.929529e-18 -3.126833e-18
[36] -3.570004e-18 -3.799068e-18 -3.882858e-18 -4.282499e-18 -6.674161e-18
[41] -6.786903e-18 -7.336935e-18 -7.538893e-18 -8.241635e-18 -1.261472e-17
[46] -1.521743e-17 -1.585857e-17 -1.648306e-17 -1.814322e-17 -2.325650e-17
[51] -2.776633e-17 -4.933935e-17
```

Figure 20: Accumulated variance.

By means of the cumulative calculation of the proportions, represented by the variable `var_exp_acum`, the total fraction of variance explained as the different dimensions are progressively incorporated into the analysis was obtained. This procedure allows for a detailed assessment of how each additional component contributes to the explanation of the total variability of the data, providing a more complete view of the relative importance of each dimension and facilitating the determination of the optimal number of components to be retained for an adequate representation of the original information.

```
> var_exp_acum
[1] 0.4298008 0.7162778 0.8848342 0.9630079 0.9906464 0.9998968 1.0000000 1.0000000
[9] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[17] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[25] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[33] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[41] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[49] 1.0000000 1.0000000 1.0000000 1.0000000
```

Figure 21: Fraction of total variance explained by incorporating dimensions.

As can be seen in the capture, considering only one dimension explains approximately 42% of the total variance, while incorporating a second dimension increases to 72%, continuing this increase progressively until the maximum value of 1 is reached. After analysing the evolution of the cumulative explained variance, it has been determined that three dimensions are sufficient to carry out an adequate representation of the data, since they provide an explanation of 88.48% of the total variance, a value that is considered satisfactory for the analysis carried out.

### EXERCISE 7: What would be the coordinates of the points obtained?

To obtain the coordinates of the points obtained, we calculated the eigenvectors of matrix B, thus extracting the 3 most relevant ones.

```
> autovec
      [,1]      [,2]      [,3]
[1,] -0.106378063 -0.006769027 -0.026683923
[2,] -0.048465976 -0.053279148  0.005229943
[3,] -0.100962708  0.041264754  0.234931703
[4,]  0.295302903  0.032982047  0.013377909
[5,] -0.046050503 -0.196123433 -0.133457241
[6,] -0.178322483  0.222198946 -0.222971673
[7,] -0.181066264 -0.098945173  0.021935787
```

Figure 22: Self-vectors of matrix B.

```
> vec12
[1] -0.106378063 -0.048465976 -0.100962708
[7] -0.181066264  0.205215496  0.245161213
[13] -0.216067167  0.015556061  0.012833763
[19] -0.161411697  0.228469767  0.167997710
```

Figure 23: Autovectors of matrix B.

After this, we scale the selected eigenvector by multiplying it by the square root of the corresponding eigenvalue. This scaling ensures that the projected coordinates preserve the original distances between the points as much as possible.

```
> result
[1] -1.3177098 -0.6003502 -1.2506295  3.6579302 -0.5704296 -2.2088885 -2.2428758
[8]  2.5420135  3.0368228  2.1659836  1.4158819 -1.3753073 -2.6764336  0.1926936
[15]  0.1589724 -2.4197066 -1.8824316  0.1925028 -1.9994139  2.8300652  2.0809951
[22] -1.6691935  0.8951389 -1.9394745  1.3437513 -2.3675056 -1.8001269  2.9647089
[29] -0.2411370  3.3564283 -0.4751507 -0.0603866  2.6423682 -1.6573782 -0.5301548
[36] -0.6047769 -0.9370143  1.6997233 -1.2726201 -0.5801228  1.4935615 -0.5330224
[43]  0.6932937  1.2780993  0.4603000 -1.2314392  0.8717854  0.8265708 -1.4217536
[50]  1.3828951 -1.0334280 -1.2836240
```

Figure 24: Result of multidimensional climbing.

In this way, the coordinates corresponding to the different points in the first dimension of the transformed space are obtained, the result of the process of dimensional reduction previously described. These coordinates reflect the relative position of each observation within this main axis, allowing us to interpret how the data are distributed according to the component that explains the highest proportion of variance. In other words, this first dimension concentrates most of the relevant information in the dataset, offering a simplified but meaningful representation of the underlying structure of the dataset.

### EXERCISE 8: Graphically represent the proposed solution, including labels that identify each observation.

This graphical representation allows the visualization of the distances between the different observations in a three-dimensional space, generated by the scatterplot3d function from the previously calculated coordinates.

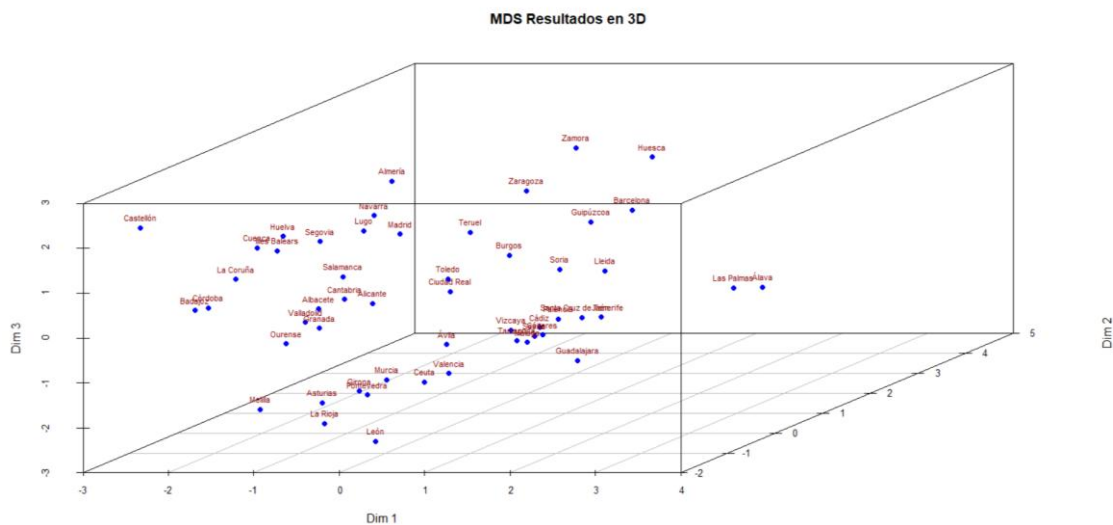


Figure 25: Representation of multidimensional scaling with tags.

Likewise, the labels associated with each observation have been represented on a two-dimensional plane, maintaining correspondence with the points shown in the graph. Although the representation is adequate and correctly reflects the spatial structure of the data, it is observed that some elements cannot be clearly distinguished, which is mainly due to the superposition of points in certain areas of three-dimensional space.

**EXERCISE 9: Is it necessary to apply any rotation or symmetry? Why? If any rotation or symmetry has been made, graphically represent the new solution with its corresponding labels.**

As mentioned above, the initial representation of the data was not the most appropriate for a clear and precise interpretation of the observations. In order to improve the three-dimensional visualization, we proceeded to rotate the axes using the `cbind` function.

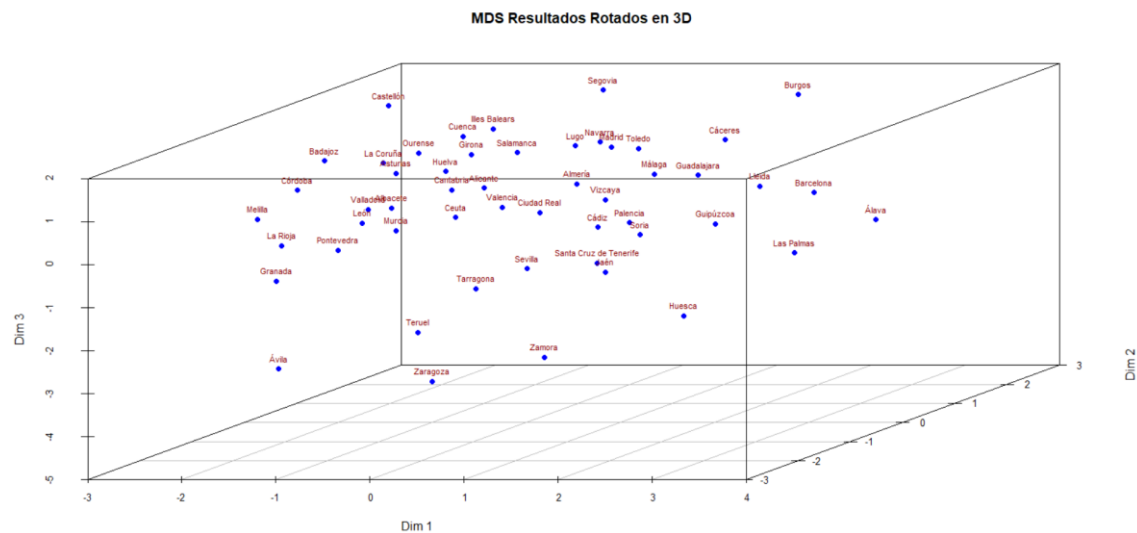


Figure 26: Representation of multidimensional scaling with rotation.

After experimenting with various possible combinations, a configuration was identified that offered a better overview, allowing the different observations to be distinguished more clearly and avoiding overlap between them. In this new arrangement, the X-axis remains unchanged, while the Y-axis now occupies the position that previously corresponded to the Z-axis, and the latter (Z-axis) now takes the place of the old Y-axis, albeit with the sign reversed, which significantly improves the readability of the graph.

**EXERCISE 10: Apply the method directly using the R function itself, and check that the results correspond to what was obtained by following the previous steps.**

After carrying out a multidimensional scaling using the specific function offered by RStudio, it can be verified that the previously developed procedure has been fully correct.

```
> cmdscale(D, k=3)
```

	[,1]	[,2]	[,3]
Albacete	-1.3177098	-0.06845502	-0.20699336
Alicante/Alacant	-0.6003502	-0.53881082	0.04056988
Almería	-1.2506295	0.41730953	1.82241952
Araba/Álava	3.6579302	0.33354670	0.10377554
Asturias	-0.5704296	-1.98339185	-1.03525867
Ávila	-2.2088885	2.24709293	-1.72964281
Badajoz	-2.2428758	-1.00063031	0.17016097
Balears, Illes	2.5420135	0.57772414	1.10524847
Barcelona	3.0368228	-1.49368419	1.35932817
Bizkaia	2.1659836	-1.61331527	-0.10284265
Burgos	1.4158819	0.19054916	-0.44680907
Cáceres	-1.3753073	-0.28081055	0.09636320
Cádiz	-2.6764336	-1.37971011	2.18304641
Cantabria	0.1926936	-0.64141436	-1.02952158
Castellón/Castelló	0.1589724	0.13017661	0.05929346
Ciudad Real	-2.4197066	-0.40959929	-0.03292583
Córdoba	-1.8824316	-1.24542488	0.97122474
Coruña, A	0.1925028	-1.46430395	-0.97737213
Cuenca	-1.9994139	1.18755958	-1.19605882
Gipuzkoa	2.8300652	-1.48102593	-0.55116379
Girona	2.0809951	0.77973046	0.97449769
Granada	-1.6691935	-0.18218428	1.45876682
Guadalajara	0.8951389	4.27366136	1.05783341
Huelva	-1.9394745	-0.91861001	1.19214912
Huesca	1.3437513	1.09992120	-0.91279156
Jaén	-2.3675056	-0.90909543	0.45877654
León	-1.8001269	-0.23242519	-2.04022208
Lleida	2.9647089	0.96813410	-0.20353640
Lugo	-0.2411370	-1.14117828	-1.75151720
Madrid	3.3564283	-0.15085849	0.52812665
Málaga	-0.4751507	-0.63082351	1.78473386
Murcia	-0.0603866	-0.82791853	1.07146020
Navarra	2.6423682	-0.82462299	-0.40901975
Ourense	-1.6573782	-0.67943337	-2.17000967
Palencia	-0.5301548	-0.29034102	-1.07758145
Palmas, Las	-0.6047769	-0.47608106	1.60922590
Pontevedra	-0.9370143	-1.43023736	-0.38119917
Rioja, La	1.6997233	0.16337876	-0.32647917

Figure 27: Results of multidimensional scaling with RStudio's cmdscale function.

The coordinates obtained through this function coincide exactly with those calculated in our previous analysis, which confirms the validity and consistency of the steps performed. This result supports the precision of the applied method and demonstrates that the transformations carried out have been adequate to faithfully represent the relationships between the observations within multidimensional space.

### EXERCISE 11: Perform the same procedure to find the configuration of points in a space, but now to represent the variables (not the observations).

To carry out a multidimensional scaling on the variables, it is necessary to work directly on the correlation matrix. By applying this procedure, it has been possible to observe the existence of a

direct relationship between the Euclidean distance matrix calculated from the standardized variables and their own correlation matrix.

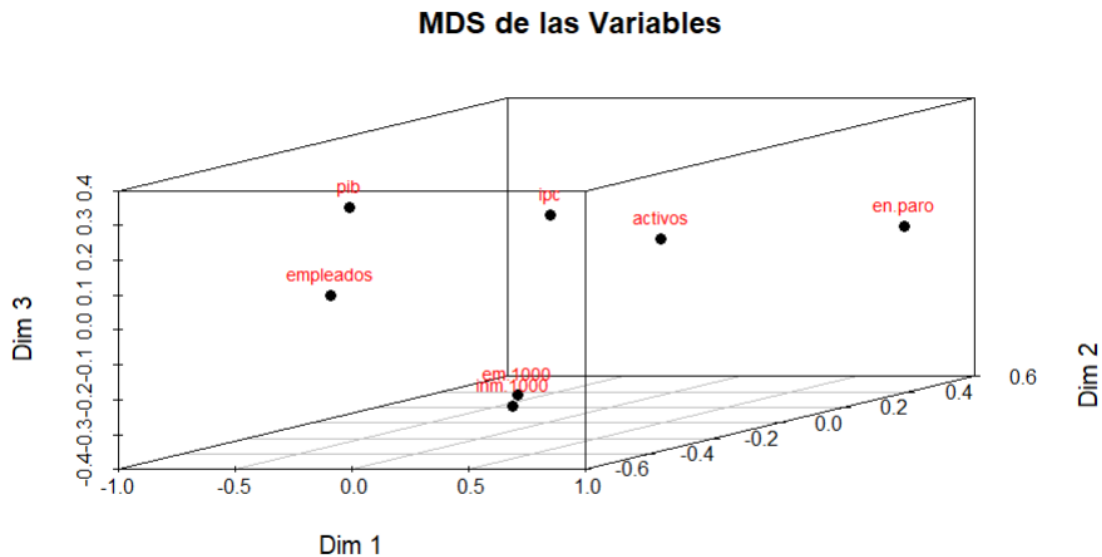


Figure 28: Results of multidimensional scaling with R's cdm-scale function.

With this in mind, it has been decided again to select three dimensions in order to facilitate a clearer and more understandable visualization of the data. As can be seen in the resulting representation, the emigration rate and the immigration rate are highly correlated with each other, while the rest of the variables show a more distant relationship, which allows the structure of interdependencies within the dataset to be identified more intuitively.

**EXERCISE 12:** In view of the results obtained throughout the year, can it be said that the application of the method to this dataset has been useful? Justify the answer.

In short, it can be firmly concluded that the exploration and analysis of the data have been fully satisfactory. The application of the methodology followed has made it possible to obtain a clear and understandable visualization of the information, while at the same time facilitating the identification of significant correlations between the different variables and the observation of how changes in their magnitudes affect the results. This approach has therefore provided a deeper understanding of the structure of the data and the relationships between its components, confirming the effectiveness of the process carried out.

## VISUAL REFERENCES.

[1] <https://www.ine.es>

[2] [https://www.abc.es/economia/abci-mapa-paro-espana-peores-y-mejores-provincias-para-encontrar-trabajo-201710261116\\_noticia.html?utm\\_source=chatgpt.com](https://www.abc.es/economia/abci-mapa-paro-espana-peores-y-mejores-provincias-para-encontrar-trabajo-201710261116_noticia.html?utm_source=chatgpt.com)