Fig. 6.1 Describe

```
> describe(date)
                vars  n     mean       sd   median  trimmed      mad      min       max     range  skew kurtosis       se
Country_name*      1 42    21.50    12.27    21.50    21.50    15.57     1.00     42.00     41.00  0.00    -1.29     1.89
Country_Code*      2 42    21.50    12.27    21.50    21.50    15.57     1.00     42.00     41.00  0.00    -1.29     1.89
EMP_VULN_MA        3 42    32.76    25.58    23.34    30.47    25.18     0.67     90.65     89.98  0.62    -0.87     3.95
EMP_VULN_FE        4 42    37.49    33.70    25.06    34.94    28.57     0.96     97.19     96.23  0.55    -1.35     5.20
SL_EMP_WORK_MA     5 42    62.68    25.13    70.79    64.37    26.01     7.97     97.36     89.39 -0.49    -1.04     3.88
SL_EMP_WORK_FE     6 42    60.37    33.55    72.38    62.63    30.02     2.38     97.58     95.20 -0.51    -1.40     5.18
GDP_EMP            7 42 39170.61 32520.54 33260.30 35667.58 29936.71  1614.05 109422.23 107808.18  0.84    -0.46  5018.03
UEM_TOTL_MA        8 42     6.13     3.46     5.81     5.89     3.41     0.30     14.80     14.50  0.55    -0.17     0.53
UEM_TOTL_FE        9 42     7.87     4.64     6.83     7.36     4.87     2.23     21.53     19.30  0.91     0.25     0.72
UEM_1524_MA       10 42    14.58     8.09    13.06    14.18     8.26     2.06     33.07     31.01  0.40    -0.77     1.25
UEM_1524_FE       11 42    17.36    10.52    14.56    16.56    11.13     3.67     44.39     40.72  0.61    -0.67     1.62
SRV_EMPL_MA       12 42    52.36    13.44    54.30    53.22    17.43    18.17     76.64     58.46 -0.43    -0.34     2.07
SRV_EMPL_FE       13 42    67.70    24.87    79.19    70.16    17.39    17.89     96.22     78.33 -0.69    -1.04     3.84
IND_EMPL_MA       14 42    25.45     8.63    26.02    25.42     7.24     8.07     42.10     34.03  0.02    -0.55     1.33
IND_EMPL_FE       15 42    10.60     4.87     9.75    10.28     3.62     1.11     22.35     21.24  0.62     0.05     0.75
AGR_EMPL_MA       16 42    22.19    19.25    17.69    20.28    20.65     0.09     70.97     70.89  0.70    -0.61     2.97
AGR_EMPL_FE       17 42    21.70    24.80     7.41    18.56    10.34     0.02     76.79     76.78  0.82    -0.88     3.83
```

Fig. 6.2 Covariance of the chosen indicators-Apply

```
> apply(c2,2,cv)
   EMP_VULN_MA    EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE        GDP_EMP    UEM_TOTL_MA    UEM_TOTL_FE    UEM_1524_MA    UEM_1524_FE    SRV_EMPL_MA
      78.08947       89.89389       40.08763       55.57176       83.02280       56.46841       58.95082       55.50210       60.58458       25.67280
   SRV_EMPL_FE    IND_EMPL_MA    IND_EMPL_FE    AGR_EMPL_MA    AGR_EMPL_FE
      36.73486       33.88565       45.97842       86.76641      114.31121
```

Fig. 6.3 Summary

```
> summary(date)
          Country_name  Country_Code  EMP_VULN_MA      EMP_VULN_FE      SL_EMP_WORK_MA   SL_EMP_WORK_FE      GDP_EMP        UEM_TOTL_MA      UEM_TOTL_FE
 Afghanistan       : 1   ABW     : 1   Min.   : 0.666   Min.   : 0.957   Min.   : 7.972   Min.   : 2.381   Min.   :  1614   Min.   : 0.296   Min.   : 2.232
 Albania           : 1   AFG     : 1   1st Qu.:11.932   1st Qu.: 7.949   1st Qu.:40.008   1st Qu.:28.352   1st Qu.: 13477   1st Qu.: 3.511   1st Qu.: 4.084
 American_Samoa    : 1   AGO     : 1   Median :23.345   Median :25.059   Median :70.786   Median :72.376   Median : 33260   Median : 5.812   Median : 6.827
 Andorra           : 1   ALB     : 1   Mean   :32.763   Mean   :37.487   Mean   :62.678   Mean   :60.370   Mean   : 39171   Mean   : 6.125   Mean   : 7.868
 Angola            : 1   AND     : 1   3rd Qu.:53.943   3rd Qu.:66.541   3rd Qu.:83.246   3rd Qu.:89.933   3rd Qu.: 52994   3rd Qu.: 7.968   3rd Qu.:10.997
 Antigua_and_Barbuda: 1   ARB     : 1   Max.   :90.651   Max.   :97.190   Max.   :97.357   Max.   :97.583   Max.   :109422   Max.   :14.795   Max.   :21.530
 (Other)           :36   (Other):36
  UEM_1524_MA      UEM_1524_FE      SRV_EMPL_MA      SRV_EMPL_FE      IND_EMPL_MA      IND_EMPL_FE      AGR_EMPL_MA      AGR_EMPL_FE
 Min.   : 2.062   Min.   : 3.670   Min.   :18.17   Min.   :17.89   Min.   : 8.072   Min.   : 1.114   Min.   : 0.089   Min.   : 0.015
 1st Qu.: 8.366   1st Qu.: 8.229   1st Qu.:42.50   1st Qu.:44.67   1st Qu.:20.739   1st Qu.: 8.162   1st Qu.: 4.176   1st Qu.: 1.980
 Median :13.065   Median :14.564   Median :54.30   Median :79.19   Median :26.021   Median : 9.754   Median :17.692   Median : 7.412
 Mean   :14.579   Mean   :17.357   Mean   :52.36   Mean   :67.70   Mean   :25.455   Mean   :10.603   Mean   :22.189   Mean   :21.697
 3rd Qu.:20.602   3rd Qu.:25.687   3rd Qu.:60.95   3rd Qu.:87.26   3rd Qu.:30.637   3rd Qu.:12.907   3rd Qu.:34.118   3rd Qu.:41.486
 Max.   :33.069   Max.   :44.391   Max.   :76.64   Max.   :96.22   Max.   :42.101   Max.   :22.354   Max.   :70.974   Max.   :76.792

>
```

**The first indicator**subject to analysis is represented by the vulnerability of employment among men (EMP_VULN_MA). It can be seen that the minimum is 0.67, and the maximum is 90.65. The difference between the two values, also called the amplitude, has a value of 89.98 and is sufficiently large compared to the average which has a value of 32.763.

On average, a percentage of 32,763 men are part of the population vulnerable to employment. Also, 25% of the male population has a degree of vulnerability lower than 11,932 (Q1) and 75% of the population has a degree of vulnerability to employment lower than 53,943 (Q3).

Taking into account the standard deviation, we can state that the share of the population vulnerable to employment deviates from the average with a value of 25.5. This value is one

quite high, close to the average, which suggests that the data corresponding to this indicator are scattered

The value for skewness, also known as the asymmetry coefficient, is 0.62. Being a positive value, slightly above 0, we can say that we are dealing with a slight asymmetry to the right, where small data predominates.

The value for kurtosis, also known as the flattening coefficient, is -0.87. This value is less than 3, which means that the data distribution is platykurtic.

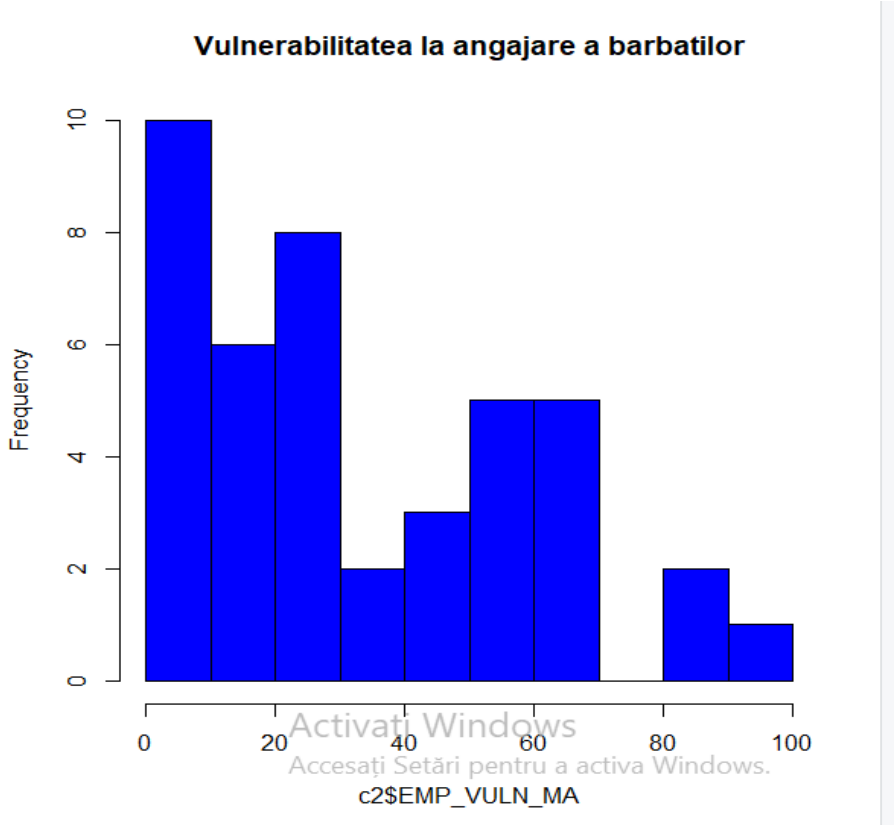
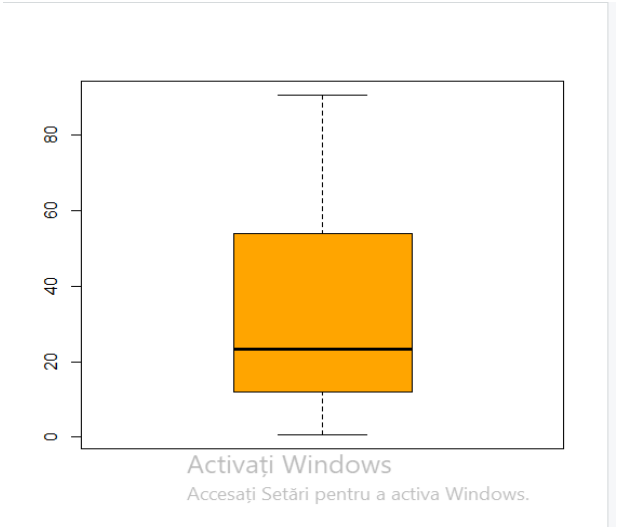



Fig. 6.4 Histogram of men's employment vulnerability

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 78.08947. This value is very high and does not approach 0, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




According to the Boxplot, it can be seen that they are not outliers for the first indicator.

Fig. 6.5 Boxplot EMP_VULN_MA

**The second indicator**subject to analysis is represented by the vulnerability of employment among women (EMP_VULN_FE). It can be seen that the minimum is 0.96, and the maximum 97.19. The difference between the two values, also called the amplitude, has a value of 96.23 and is sufficiently large compared to the average which has a value of 37.49.

On average, a percentage of 37.49 women are part of the population vulnerable to employment. Also, 25% of the female population has a degree of vulnerability lower than 7,949 (Q1) and 75% of the population has a degree of vulnerability to employment lower than 66,541 (Q3).

Taking into account the standard deviation, we can state that the share of the population vulnerable to employment deviates from the average with a value of 33.7. This value is quite high, very close to the average, which suggests that the data corresponding to this indicator are scattered.

The value for skewness is 0.55. Being a positive value, slightly above 0, we can say that we are dealing with a slight asymmetry to the right, where small data predominates.

The value for kurtosis is -1.35. This value is less than 3, which means that the data distribution is platykurtic.

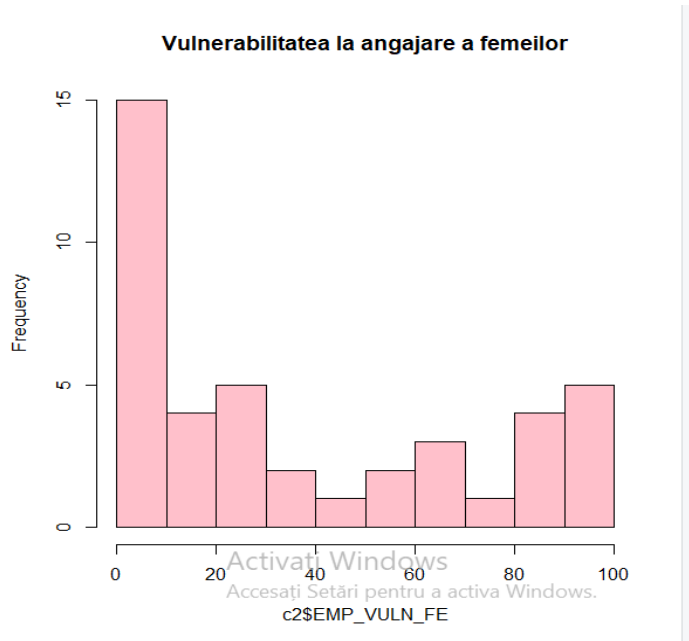
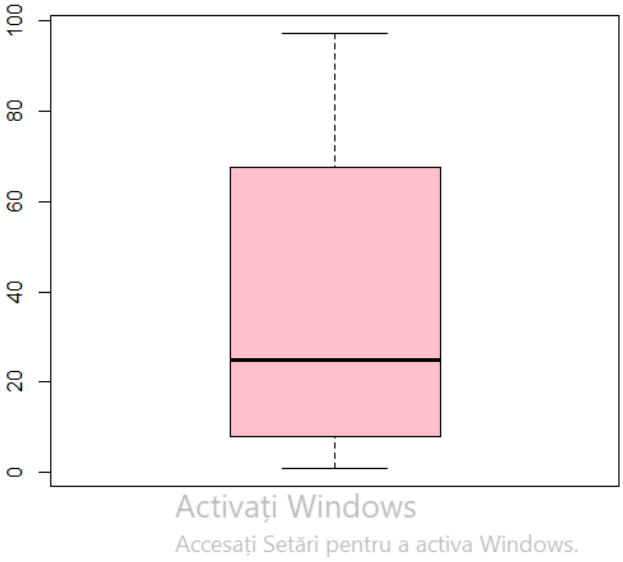


Fig. 6.6. Histogram of women's employment vulnerability

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 89.89389. This value is very high and does not approach 0, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.






According to the boxplot, there are no outliers for the second indicator.

Fig. 6.7- Boxplot EMP_VULN_FE

**The third indicator**subject to analysis is represented by the salaried male population with a work contract (SL_EMP_WORK_MA). It can be seen that the minimum is 7.97, and the maximum is 97.36. The difference between the two values, also called the amplitude, has the value of 89.39 and in the case of this indicator, it is not that big compared to the average of 62.68

On average, a percentage of 62.68 men are part of the salaried population with an employment contract. Also, the level of salaried workers for 25% of the observations is lower than 40,008 (Q1) and the level of salaried persons for 75% of the observations is lower than 83,246 (Q3).

Taking into account the standard deviation, we can state that the share of the salaried population deviates from the average with a value of 25.13. This value is not as close to the average as the other two, which suggests that the data corresponding to this indicator are less scattered than the other two indicators, but still have a significant level of dispersion.

The value for skewness is -0.49. Being a negative value, we can say that we are dealing with an asymmetry to the left, where big data predominates.

The value for kurtosis is -1.04. This value is less than 3, which means that the data distribution is platykurtic.

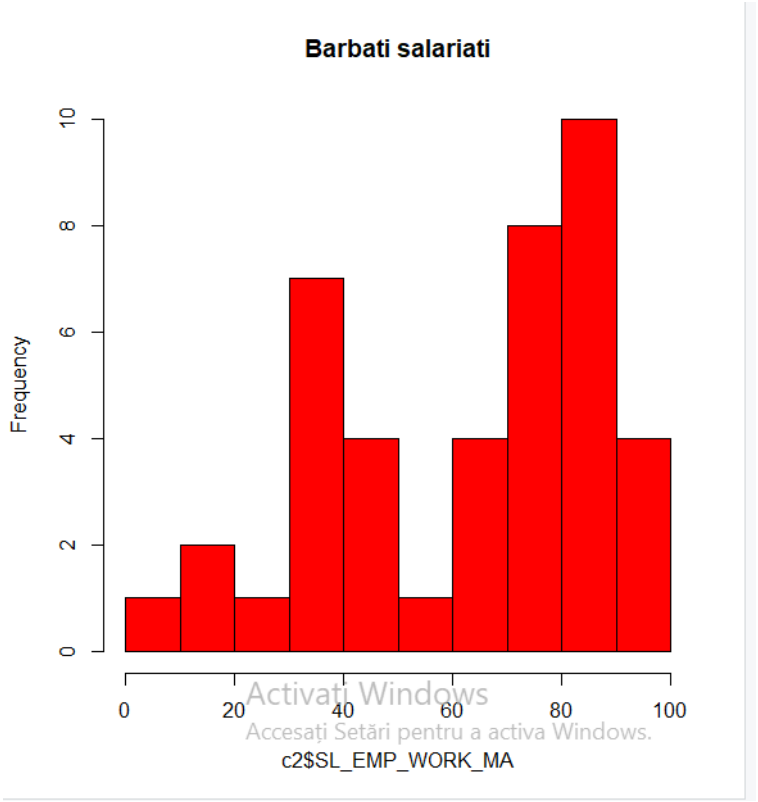
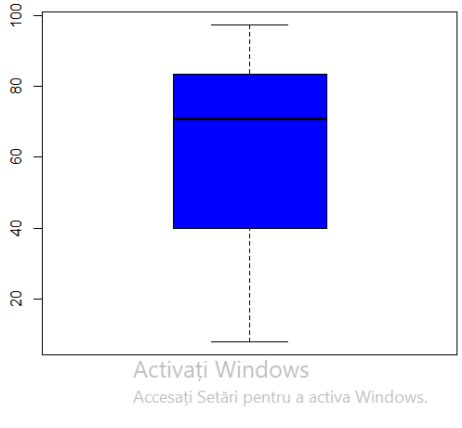



Fig. 6.8. Histogram of employed men

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 40.087. This value is high and does not approach 0, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




According to the boxplot, there are no outliers

for the third indicator

Fig. 6.9- Boxplot SL_EMP_WORK_MA

**Indicator 4**subject to analysis is represented by the salaried female population with a work contract (SL_EMP_WORK_FE). It can be seen that the minimum is 2.38, and the maximum 97.58. The difference between the two values, also called the amplitude, has the value of 95.20 and in the case of this indicator, it is not that big compared to the average of 60.37.

On average, a percentage of 60.37 women are part of the salaried population with an employment contract. Also, the level of salaried women for 25% of observations is lower than 28,352 (Q1) and the level of salaried women for 75% of observations is lower than 89,933 (Q3).

Taking into account the standard deviation, we can state that the share of the salaried population deviates from the average with a value of 33.55. This value is not very close to the average, which suggests that the data corresponding to this indicator are less scattered compared to the first two indicators, but still have a significant level of dispersion.

The value for skewness is -0.51. Being a negative value, we can say that we are dealing with an asymmetry to the left, where big data predominates.

The value for kurtosis is -1.40. This value is less than 3, which means that the data distribution is platykurtic.

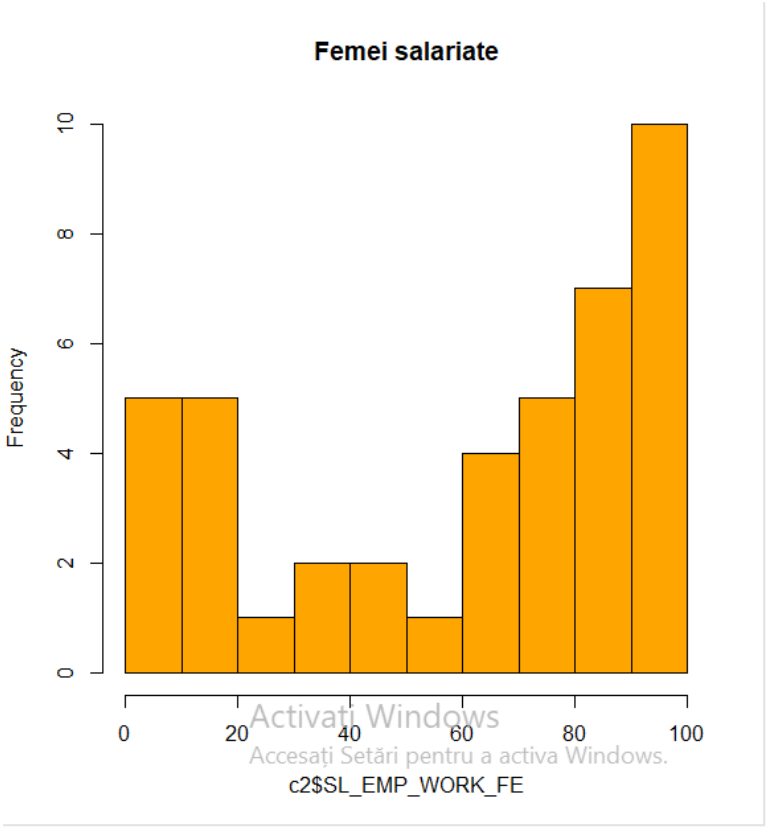
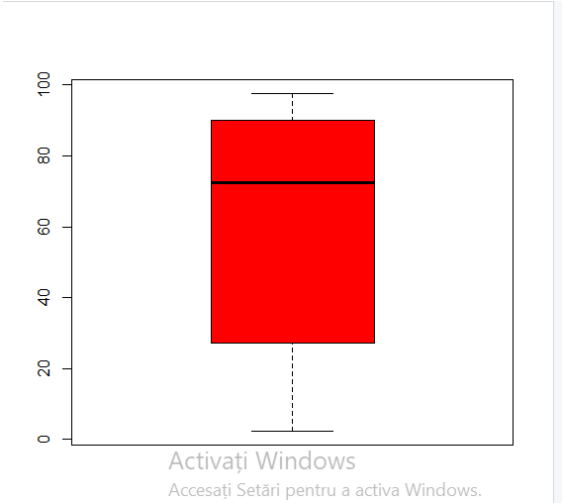



Fig. 6.10. Histogram of female employees

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 55.57176. This value is high and does not approach 0, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




According to the boxplot, there are no outliers

for the fourth indicator

Fig. 6.7- Boxplot SL_EMP_WORK_MA

**Indicator 5**is the GDP corresponding to each employed person (GDP_EMP). It can be seen that the minimum is 1614.05, and the maximum 109422.23. The difference between the two values, also called the amplitude, has the value of 107808.18 and in the case of this indicator, it is a high value compared to the average 39170.61.

On average, one employee corresponds to a GDP of $39,170.61. Also, 25% of the employees have a corresponding GDP from the observations is lower than 13477 (Q1) and for 75% of the employees corresponds a GDP lower than 52994 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 32520.54. This value is very close to the average, which suggests that the data corresponding to this indicator are scattered.

The value for skewness is 0.84. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small data predominates.

The value for kurtosis is -0.46. This value is less than 3, which means that the data distribution is platykurtic.

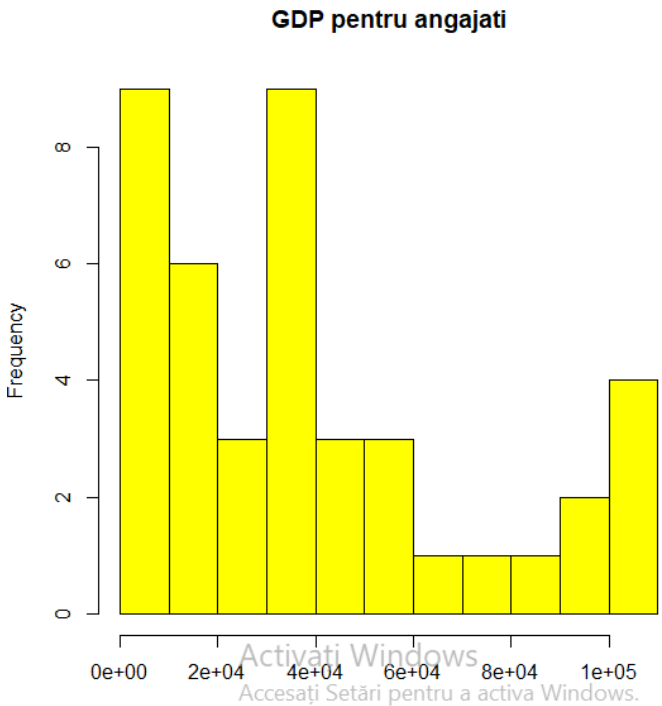
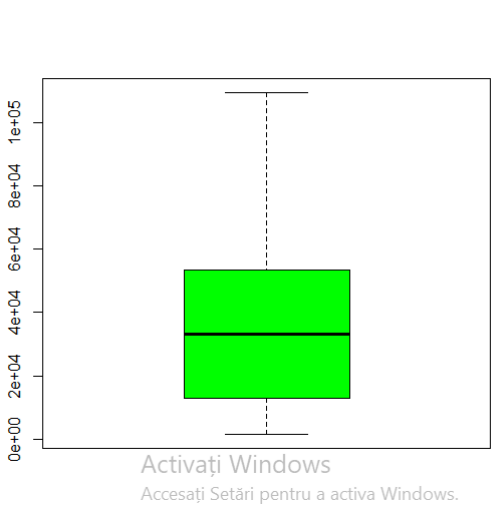



Fig. 6.10. GDP employed

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 83.02280. This value is high and does not approach 0, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




According to the boxplot, there are no outliers

for the fourth indicator

Fig. 6.7- Boxplot GDP_EMP

**Indicator 6**it is represented by the unemployed male population (UEM_TOTL_MA). ). It can be seen that the minimum is 0.30, and the maximum 14.80. The difference between the two values, also called the amplitude, has the value of 14.50 and in the case of this indicator, it is not a large value compared to the average of 6.13.

On average, the unemployed male population is 6.13. Also, the unemployment level for 25% of the male population is lower than 3,511 (Q1) and for 75% of men, the unemployment level is lower than 7,968 (Q3)

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 3.46. This value is very close to the average, which suggests that the data corresponding to this indicator are scattered.

The value for skewness is 0.55. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small data predominates.

The value for kurtosis is -0.17. This value is less than 3, which means that the distribution of the data is platykurtic.

The coefficient of variation (Fig. 6.2) is 56.46841. This value is high and does not approach 0, far exceeding the 35% threshold, which increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.

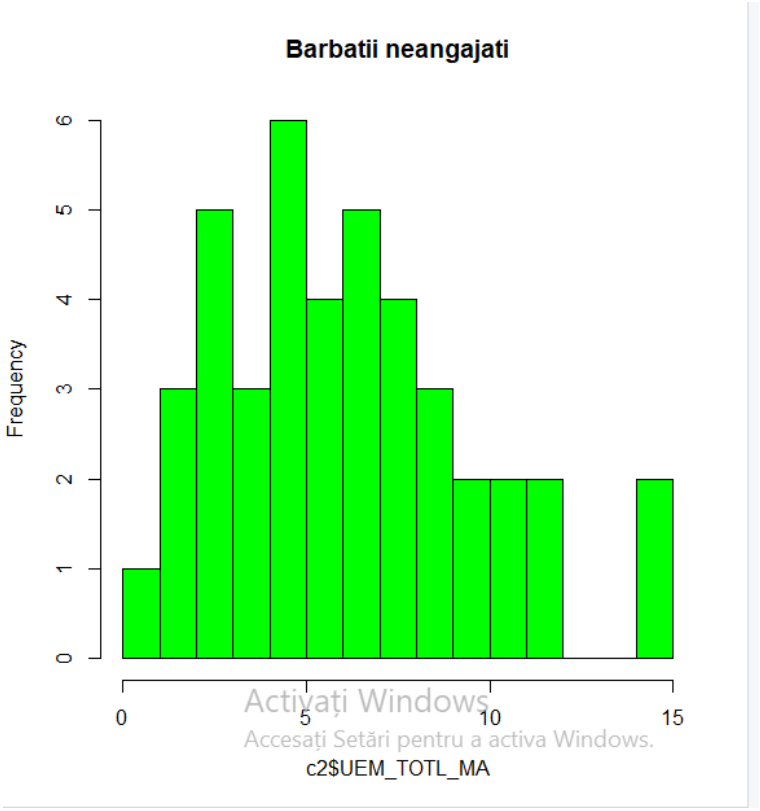
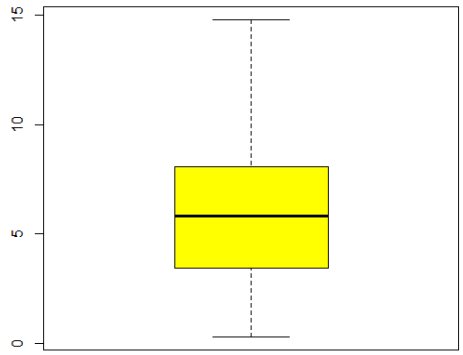



Fig. 6.10. Unemployed men

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

The coefficient of variation (Fig. 6.2) is 56.46841. This value is high and does not approach 0 and exceeds 35%, an aspect that increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.






According to the boxplot, there are no outliers

for the fifth indicator

Fig. 6.11- Boxplot UEM_TOTL_MA

**Indicator 7**is represented by the unemployed female population (UEM_TOTL_FE). It can be seen that the minimum is 2.23, and the maximum 21.53. The difference between the two values, also called the amplitude, has the value of 19.30 and in the case of this indicator, it is quite a large value compared to the average of 7.87.

On average, the unemployed female population is 7.87. Also, the level of unemployed women for 25% of the female population is lower than 4,084 (Q1) and for 75% of the female population, the unemployment level is lower than 10,997 (Q3)

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 4.64. This value is quite close to the average, which suggests that the data corresponding to this indicator are scattered. Compared to other previously studied indicators, these data corresponding to this indicator are less scattered.

The value for skewness is 0.91. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small data predominates.

The value for kurtosis is 0.25. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 58.95. This value is high and does not approach 0, far exceeding the threshold of 35%, which increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.

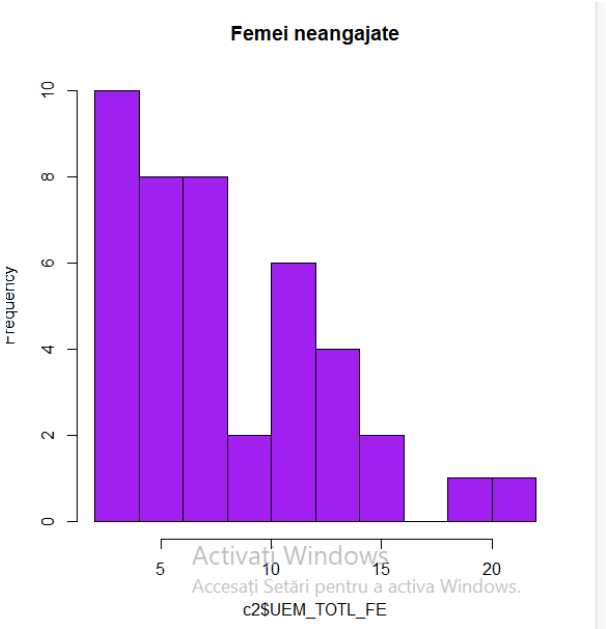
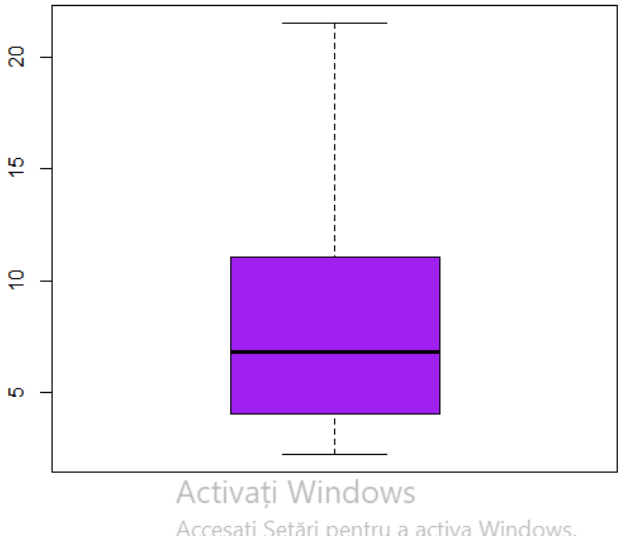



Fig. 6.12 Histogram of the unemployed female population




According to the boxplot, there are no outliers

nor for indicator 7.

Fig. 6.13- Boxplot UEM_TOTL_FE

**Indicator 8**it is represented by the young male population, aged between 15 and 24, not employed, but available for employment (UEM_1524_MA). ). It can be seen that the minimum is 2.06, and the maximum 33.07. The difference between the two values, also called amplitude, has the value of 31.01 and in the case of this indicator, it is not such a large value compared to the average of 14.58.

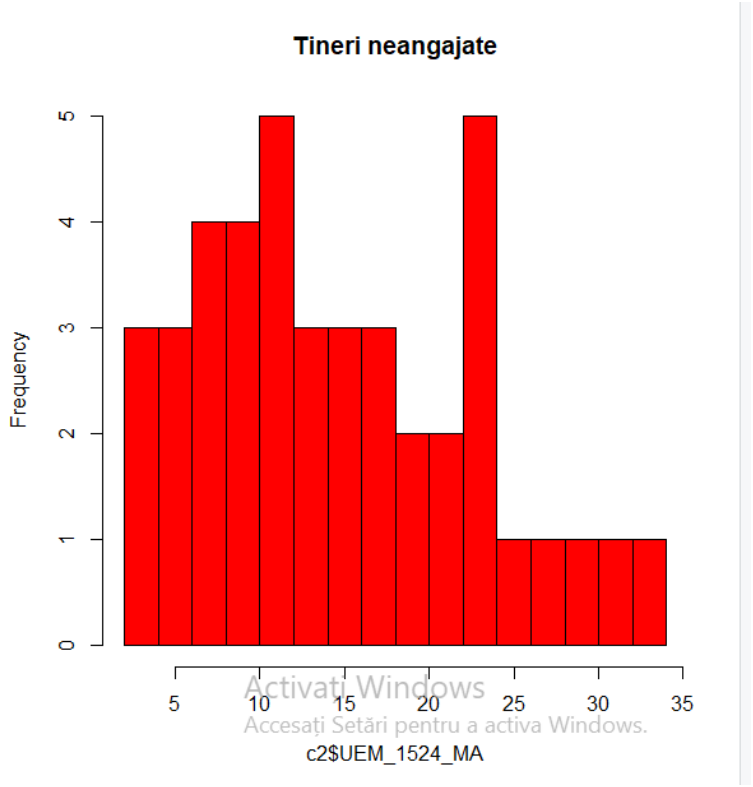
On average, the young unemployed male population is 14.58. Also, the level of unemployed youth for 25% of the young male population is lower than 8.36 (Q1) and for 75% of the young male population, the unemployment level is lower than 20.602 (Q3)

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 8.09. This value is close to the average value, representing more than half of its value, an aspect that suggests in this case that the data are quite scattered, and very less uniform.

The value for skewness is 0.40. Being a positive value, and not very far from 0, we can say that we are dealing with a slight asymmetry to the right, where small values predominate.

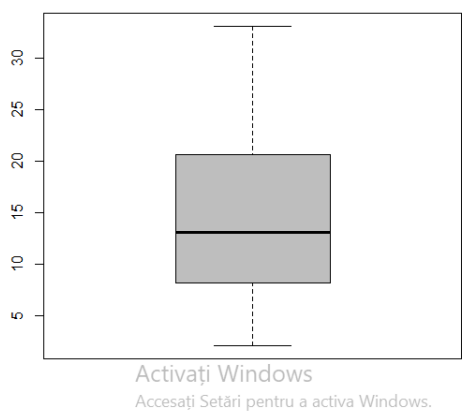
The value for kurtosis is -0.77. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 55.50210. This value is high and does not approach 0, far exceeding the threshold of 35%, which increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




6.14. Histogram of the young unemployed population, but looking for a job

From the histogram above, we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing scattered values and not gathered around the average, where obvious fluctuations of the values are visible.




According to the boxplot, there are no outliers

nor for indicator 8.

Fig. 6.15- Boxplot UEM_1524_MA

**Analysis of the 9th indicator**includes the analysis of the data that are part of the category of the young female population, aged between 15-24 years who are not employed, but who are fit and looking for a job (UEM_1524_FE). Note that the minimum is 3.67, and the maximum 44.39. The amplitude has the value of 40.72, being a rather big difference between this and the average which has the value of 17.36

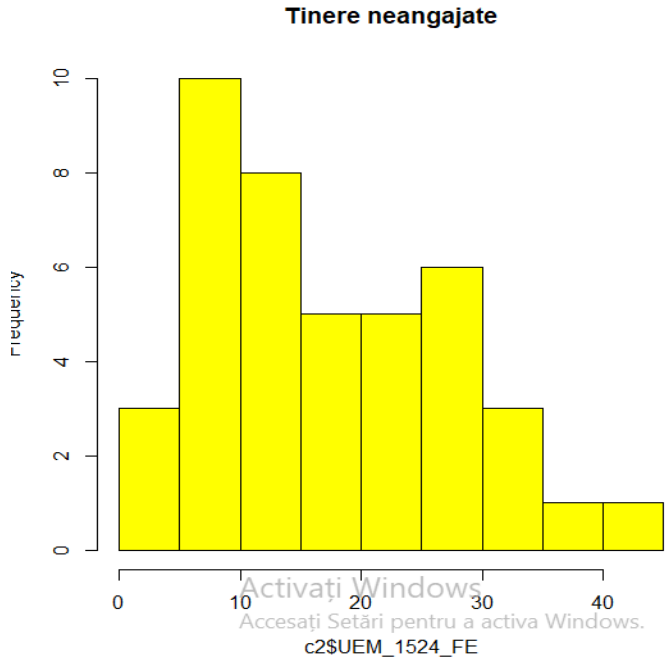
On average, the young unemployed female population is 17.36. Also, the level of unemployed young women for 25% of the young female population is lower than 8,229 (Q1) and for 75% of the young female population, the unemployment level is lower than 25,687 (Q3)

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 10.52. This value is very close to the average value, an aspect that also suggests in this case that the data are quite scattered, and very less uniform.

The value for skewness is 0.61. Being a positive value, and not very far from 0, we can say that we are dealing with a slight asymmetry to the right, where small values predominate.

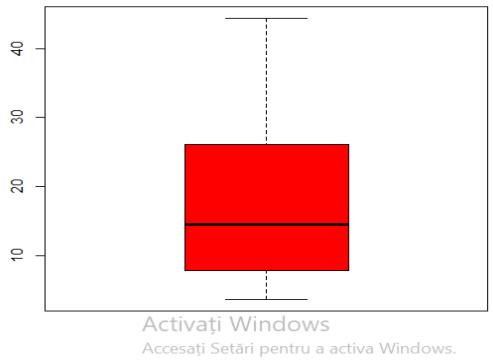
The value for kurtosis is -0.67. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 60,684. This value is high and does not approach 0, far exceeding the threshold of 35%, which increases the fact that the average is not representative for this indicator, and the data series is not homogeneous.




6.16. Histogram of the young female population not employed, but looking for a job

From the histogram above, we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing scattered values and not gathered around the average, where obvious fluctuations of the values can be seen




According to the boxplot, there are no outliers

nor for the indicator

Fig. 6.17- Boxplot UEM_1524_FE

**Indicator 10**used in the analysis of the degree of development of the labor force is represented by men who work in the tertiary sector, that of services. It is known that a high degree of the population working in this sector indicates for the respective country a specific economic development. Note that the minimum is 18.7, and the maximum 76.64. The amplitude has the value of 58.46, being a small difference this time compared to the average of 52.36.

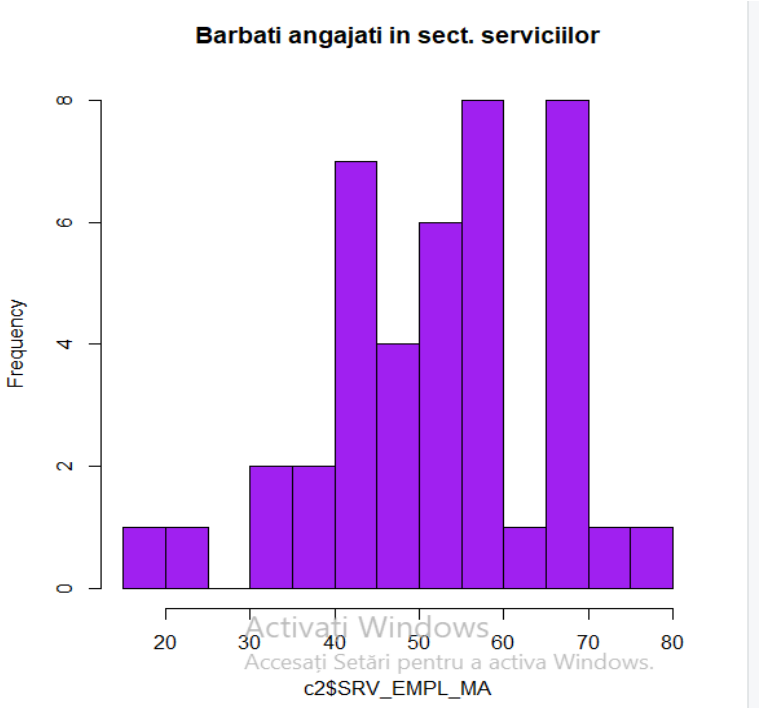
On average, the male population working in the service sector is 52.36. Also, the level of men employed in the service sector for 25% of the male population is less than 42.50 (Q1) and for 75% of the male population, the level of people employed in the service sector is less than 60.95 (Q3)

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 13.44. This value is quite close to the average value, an aspect that suggests that the degree of dispersion for these data is not as high as those analyzed previously. In other words, the male population working in the service sector introduces a slightly more uniform data set.

The value for skewness is -0.43. Being a negative value, and not very far from 0, we can say that we are dealing with a slight asymmetry to the left, where large values predominate. This aspect indicates that the countries where the number of men in the service sector is higher prevail in the data analysis.

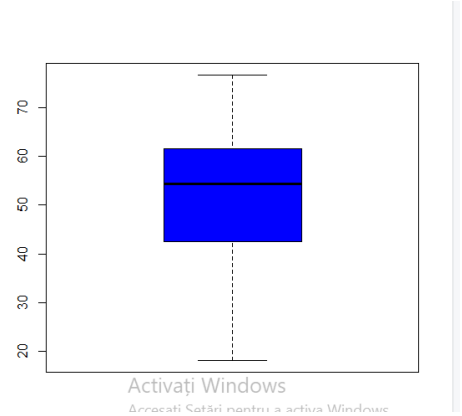
The value for kurtosis is -0.34. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 25.67280. This value is quite small and this time does not exceed the 35% threshold, an aspect that increases the fact that the average is representative for this indicator, and the data series is homogeneous.




6.18. Histogram of the male population in the service sector

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing slightly dispersed values and which gather to a certain extent in the average area, where there are obvious fluctuations of the values from very high, up to very low values, depending on the particularities of the analyzed countries. It is certain that the histogram increases the idea that large values predominate.




According to the boxplot, there are no outliers

nor for the indicator

Fig. 6.19- Boxplot SRV_EMPL_MA

**Indicator 11**which helps the proposed analysis is represented by the female population that works in the field of services, i.e. in the tertiary sector. (SRV_EMPL_MA). Note that the minimum is 17.86, and the maximum 96.22. The amplitude has the value of 78.33, being a small difference again this time compared to the average of 67.70.

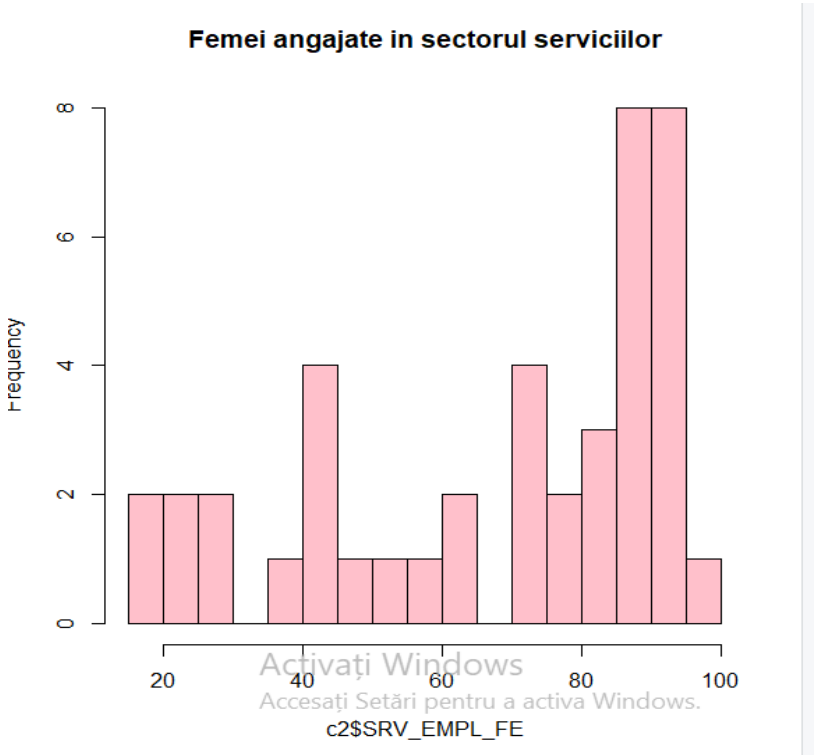
On average, the male population working in the service sector is 67.70. Also, the level of women employed in the service sector for 25% of the female population is lower than 44.67 (Q1) and for 75% of the female population, the level of people employed in the service sector is lower than 87.26 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 24.87. This value is quite close to the average value, an aspect that suggests that the degree of dispersion for these data is not as high as those analyzed previously. In other words, the female population, as well as the male population, who work in the service sector introduces a slightly more uniform data set.

The value for skewness is -0.69. Being a negative value, we can say that we are dealing with an asymmetry to the left, where large values predominate. This aspect indicates that the countries where the number of women in the service sector is higher prevail in the data analysis.

The value for kurtosis is -1.04. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 36.734. This value is slightly higher than the 35% threshold, an aspect that increases the fact that the average is not representative for this indicator, and the data series is inhomogeneous.




6.20. Histogram of the young female population employed in the service sector

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing slightly dispersed values and not gathered around the average, not having a central tendency and obvious fluctuations of the values from very high values are visible , up to very low values, depending on the particularities of the analyzed countries. It is certain that the histogram increases the idea that large values predominate.

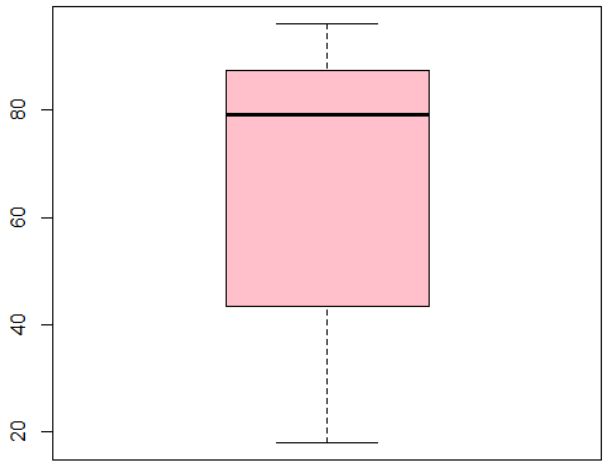





Fig. 6.21- Boxplot SRV_EMPL_FE

According to the boxplot, there are no outliers

nor for the indicator

**Indicator 12**represents the male population working in the industrial sector, that is, in the secondary sector. Note that the minimum is 8.07, and the maximum 42.10. The amplitude has the value of 34.03, being a small difference compared to the average of 25.45.

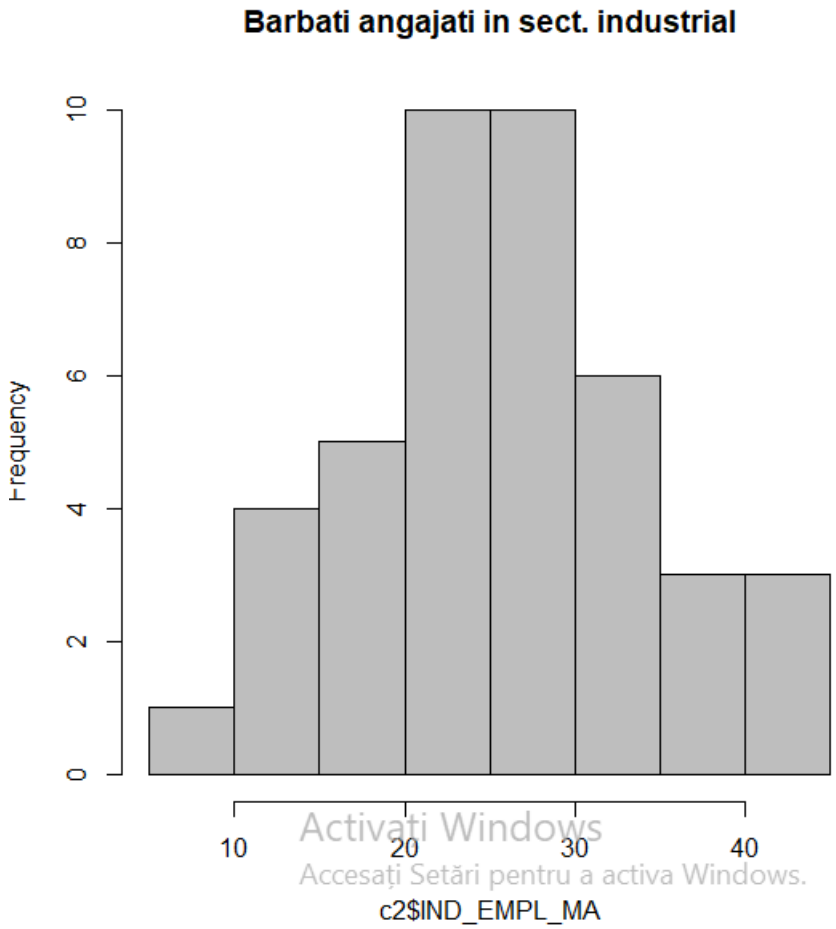
On average, the male population working in the industrial sector is 25.45. Also, the level of men employed in the industrial sector for 25% of the male population is lower than 20,739 (Q1) and for 75% of the female population, the level of people employed in the service sector is lower than 30,637 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 8.63. This value is quite close to the average value, an aspect that suggests that the degree of dispersion for these data is not as high as those analyzed previously. In other words, the male population introduces a slightly more uniform data set.

The value for skewness is 0.02. Being a positive value, which goes very slightly above 0, we can say that we are dealing with a very slight asymmetry to the right, where small values predominate. This aspect means that the data analysis predominates in countries where the share of men working in the industrial sector is lower, but it does not differ much, since the asymmetry is very small.

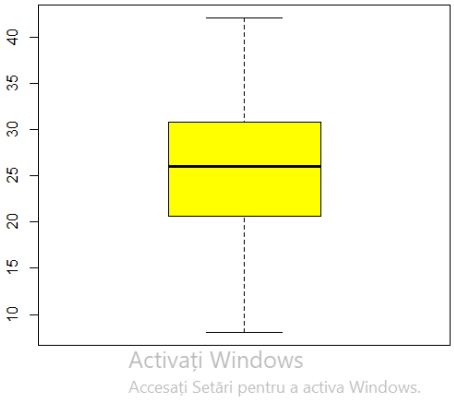
The value for kurtosis is -0.55. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 33.88565. This value does not exceed the 35% threshold, but it is still close to the limit. This aspect indicates that the average is representative and the data are homogeneous.




6.22. Histogram of men employed in the industrial field

As can be seen, the histogram reflects the analysis made through the indicators from the descriptive statistics. Thus, here it can be observed how the larger data slightly predominates, that is, not in many countries the share of men working in the industry is higher. The homogeneity of the data, their central tendency, is also emphasized through the prism of this histogram, which in shape is close to the Gauss Curve.




According to the boxplot, there are no outliers

nor for the indicator

Fig. 6.23- Boxplot IND_EMPL_MA

**Indicator 13**subject to analysis is represented by the female population working in the industrial sector. Note that the minimum is 1.11, and the maximum 22.35. The amplitude has the value of 21.24, which is quite a big difference compared to the average of 10.60.

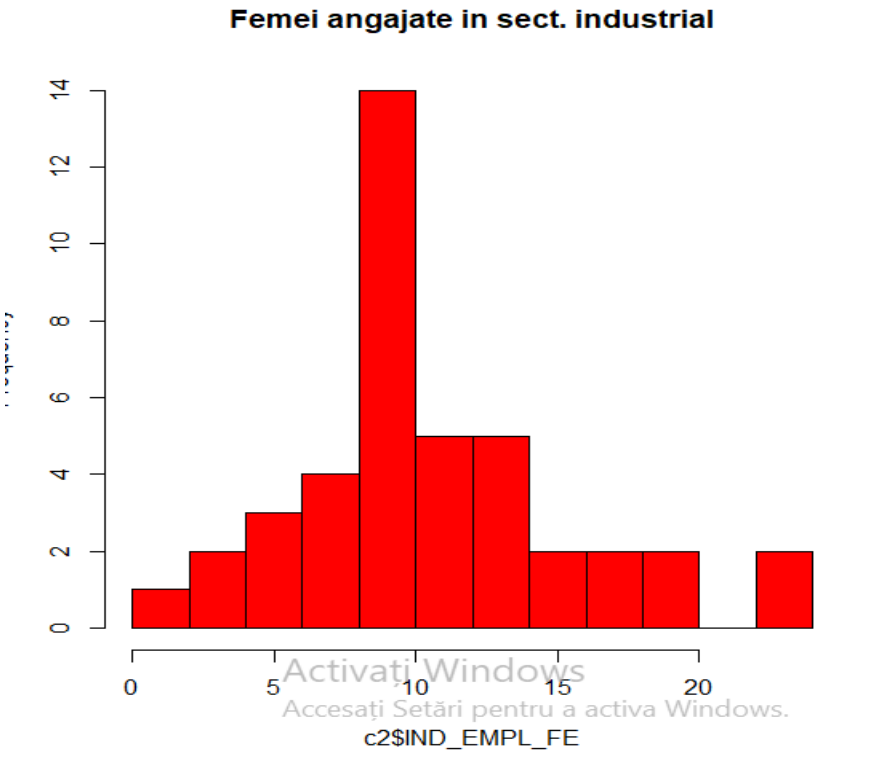
On average, the female population working in the industrial sector is 10.60. Also, the level of women employed in the industrial sector for 25% of the female population is lower than 8,162 (Q1) and for 75% of the female population, the level of people employed in the industrial field is lower than 12,907 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 4.87. This value is not very close to the average value, being less than half of its value, an aspect that suggests that the degree of dispersion for these data is not as high as those previously analyzed.

The value for skewness is 0.62. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small values predominate. This aspect indicates that there is a predominance of countries where the share of women working in the industry is small.

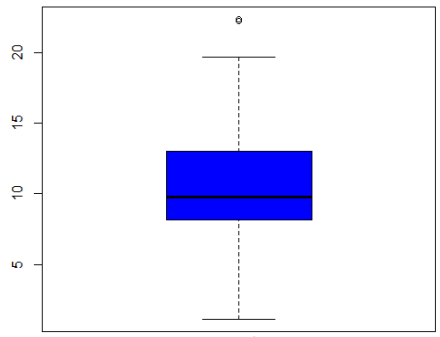
The value for kurtosis is 0.05. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 45.97. This value exceeds the 35% threshold, so the average is not representative and the data are not homogeneous.




6.24. Histogram of women employed in the industrial field

The histogram increases the analysis done previously through the prism of the indicators used in the realization of the descriptive statistics. Thus, one can easily observe the platykurtic distribution, which is interrupted, however, by higher values from the middle of the interval [5:10] to 10. This aspect suggests the fact that the level of women working in the industrial sector is very low , this aspect being obvious due to the conditions and the physical effort required in this sector where men generally predominate. The central tendency is not respected, the data are not homogeneous, so they do not gather around the average.




According to the boxplot, there is an outlier, aspect also visible from the histogram.



Fig. 6.25- Boxplot IND_EMPL_FE

**Indicator 14**it is represented by the male population working in the primary sector, i.e. agriculture. It is known that the share of the population working in this sector can be an economic indicator for the respective country, since a large share working in agriculture is specific to poorly developed countries, where the main activity of the population is agriculture.

Note that the minimum is 0.09, and the maximum is 70.97. The amplitude has the value of 70.89, which is a big difference compared to the average of 22.19.

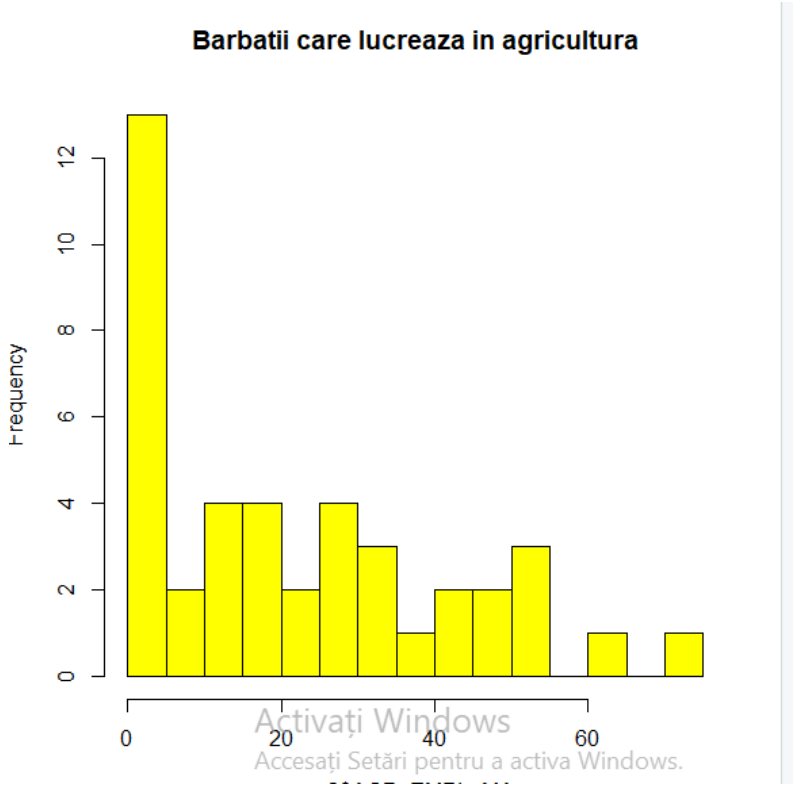
On average, the male population working in the agricultural sector is 22.19. Also, the level of men employed in the agricultural sector for 25% of the male population is lower than 4,176 (Q1) and for 75% of the male population, the level of people employed in agriculture is lower than 34,118 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 19.25. This value is very close to the average value, which suggests that the degree of dispersion for these data is high.

The value for skewness is 0.70. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small values predominate. This aspect indicates that there are countries where the share of men working in agriculture is low.

The value for kurtosis is -0.61. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 86.76. This value far exceeds the threshold of 35%, and does not approach 0, as would be the case with homogeneous data, so the average is not representative and thus, the data are not homogeneous.




6.26. Histogram of men employed in agriculture

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing dispersed values and not gathered around the average.

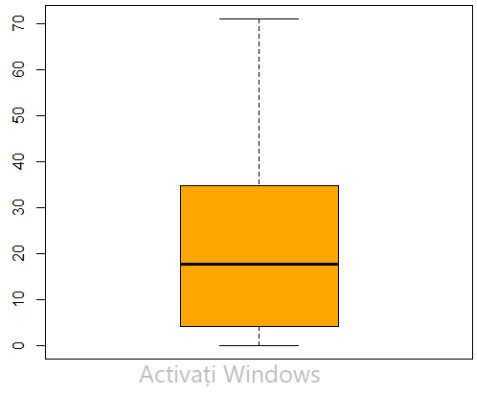



Fig. 6.27- Boxplot AGR_EMPL_MA

According to the boxplot, there are no outliers

nor for the indicator

**Indicator 15**it is represented by the female population working in agriculture. population is agriculture. Note that the minimum is 0.02, and the maximum 76.79. The amplitude has the value of 76.78, which is a big difference compared to the average of 21.70.

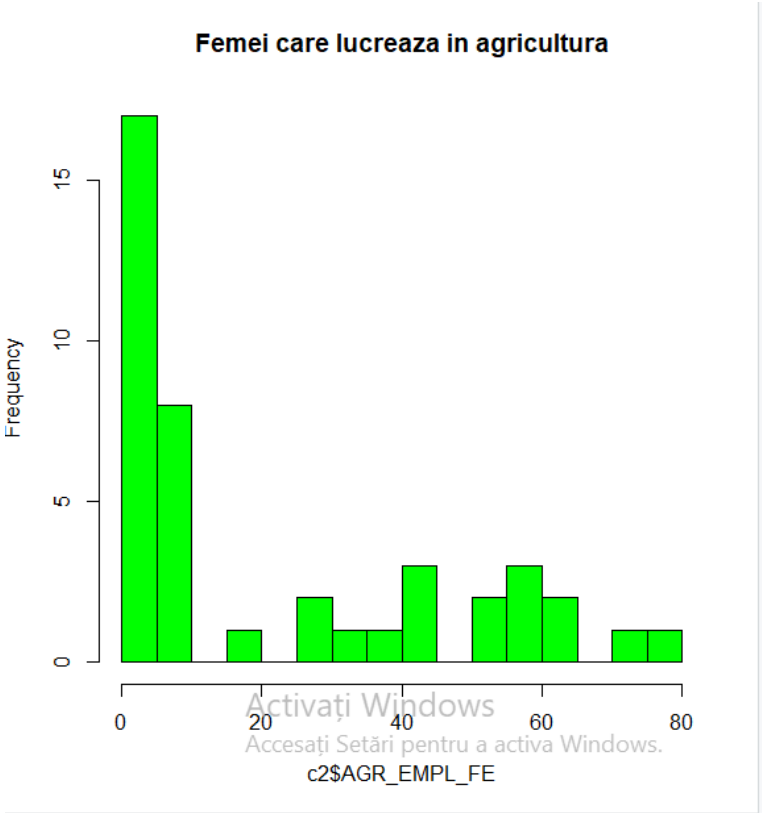
On average, the female population working in the agricultural sector is 21.70. Also, the level of women employed in the agricultural sector for 25% of the female population is lower than 1,980 (Q1) and for 75% of the female population, the level of people employed in agriculture is lower than 41,486 (Q3).

Taking into account the standard deviation, we can state that the data of this indicator deviates from the average with a value of 24.80. This value exceeds the average value, which suggests that the degree of dispersion for these data is very high.

The value for skewness is 0.82. Being a positive value, we can say that we are dealing with an asymmetry to the right, where small values predominate. This aspect indicates that the countries where the proportion of women working in agriculture is low prevail.

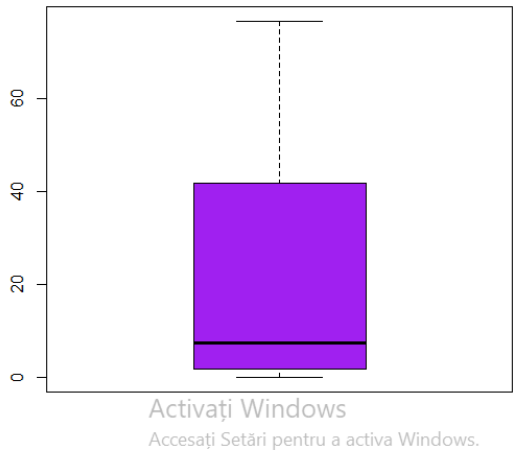
The value for kurtosis is -0.88. This value is lower than 3, which means that the distribution of the data in this case is also platykurtic.

The coefficient of variation (Fig. 6.2) is 114.31. This value far exceeds the threshold of 35%, and does not approach 0, as would be the case with homogeneous data, so the average is not representative and thus, the data are not homogeneous.




6.28. Histogram of women employed in agriculture

From the histogram above we can see that the graphic part confirms the analysis of the previously obtained values, the distribution of the data series really showing very dispersed values and not at all gathered around the average.



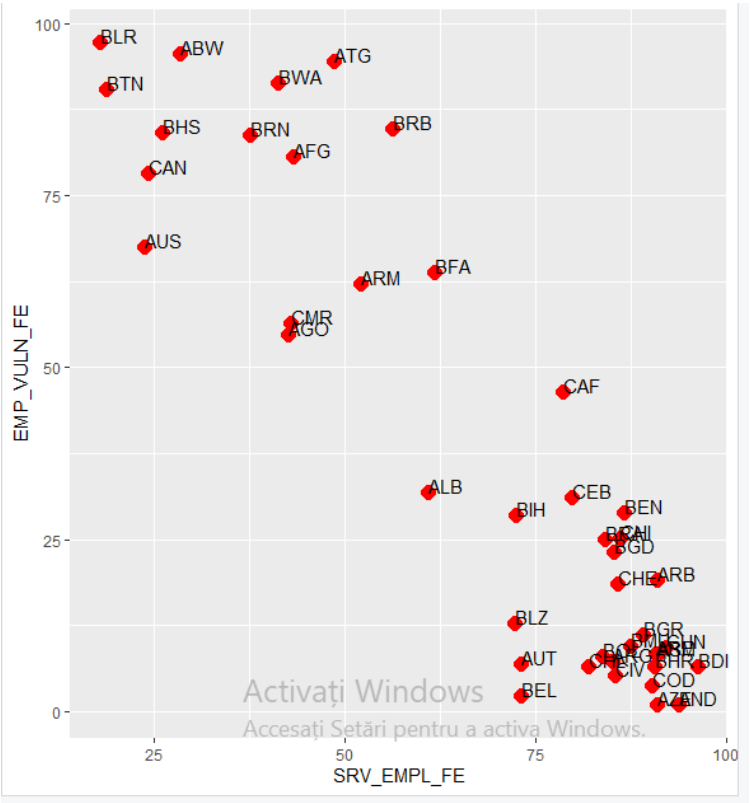

According to the boxplot, there are no outliers

nor for the indicator

Fig. 6.29- Boxplot AGR_EMPL_FE

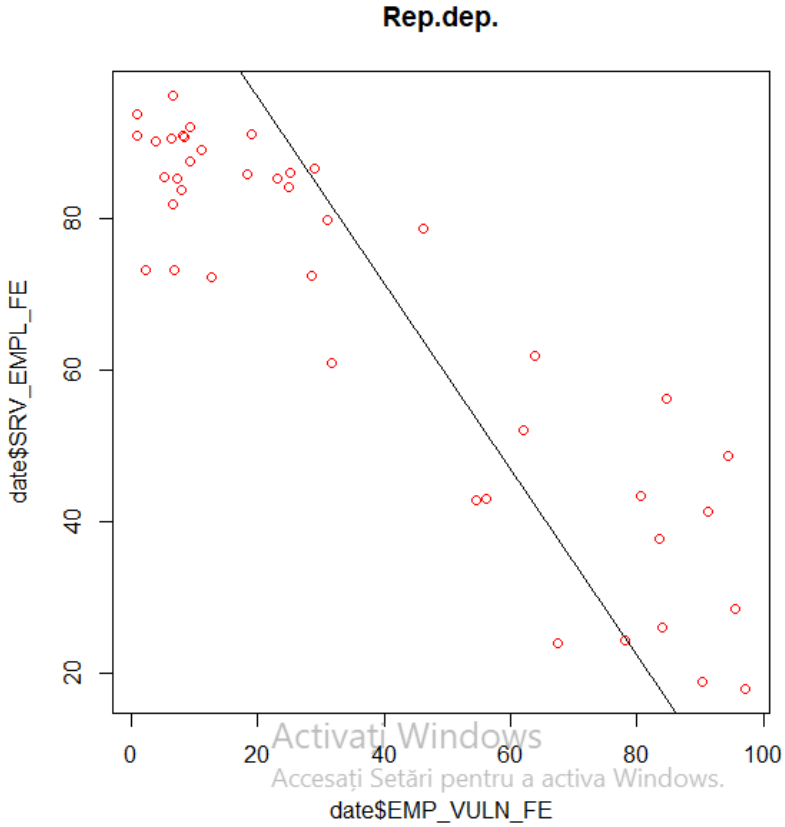
**6. Graphical representations between two variables**

I chose to make the graphic representation between the variable of employment vulnerability among women (EMP_VULN_FE) and the female population working in the field of services (SRV_EMPL_FE), since they were among the two most correlated variables.



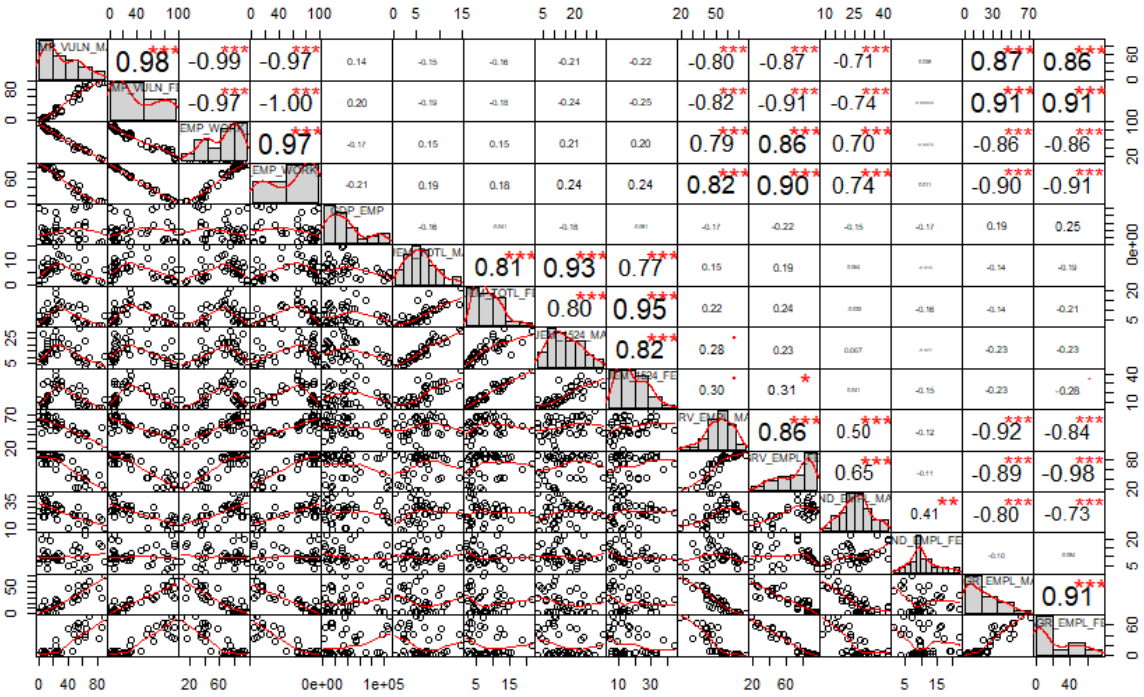

7.1. Vulnerability to employment among women influenced by the population women who work in the field of services

According to the graph presented above, it can be seen that most of the countries under analysis (BDI, COD, BGR, etc.) are gathered in a cloud of points in the lower right part of the graph. This aspect suggests the fact that for most of the countries under analysis, the number of women working in the service sector is very high, so the vulnerability of the female population to employment is also very low, as they are negatively correlated. Thus, from the point of view of an analysis of the labor force, we can say that for the female population, these countries are more stable and more secure in finding a job. The countries on the upper left side of the graph are those countries for which the vulnerability of employment the work is very big since the share of the female population working in the field of services for them is quite small. This aspect can suggest a weak degree of desolation of the respective countries and it can be assumed either that women work in other sectors of the economy, such as the agricultural sector, or they do not work at all.




7.2. The regression graph between EMP_VULN_FE and SRV_EMPL_FE

By inverting the coordinates between them, it is possible that the graph looks like a mirror. More precisely, for countries where vulnerability is very low, the share of women employed in the tertiary sector is very high and you can see the predominant cloud of points in the upper left corner, and for countries where vulnerability is very high, the number of women in this field it is smaller. The line drawn on this graph increases the fact that the 2 variables are correlated because the distance between the points is quite small. Also, the connection is indirect.



7.3. Performance Analytics

In the image above it can be seen that on the main diagonal are histograms with probability densities in which you can see the shape of the distribution of each data set specific to the indicators used in the analysis (The shape of the distribution was described in point 6). Below the main diagonal are the plots between the variables, taken two by two, and the red lines are dependent on each other. In the upper right, above the main diagonal, are the correlation coefficients. The bigger the writing and the more stars next to the coefficients, the stronger the correlation. The highest correlation is marked with asterisks (***).

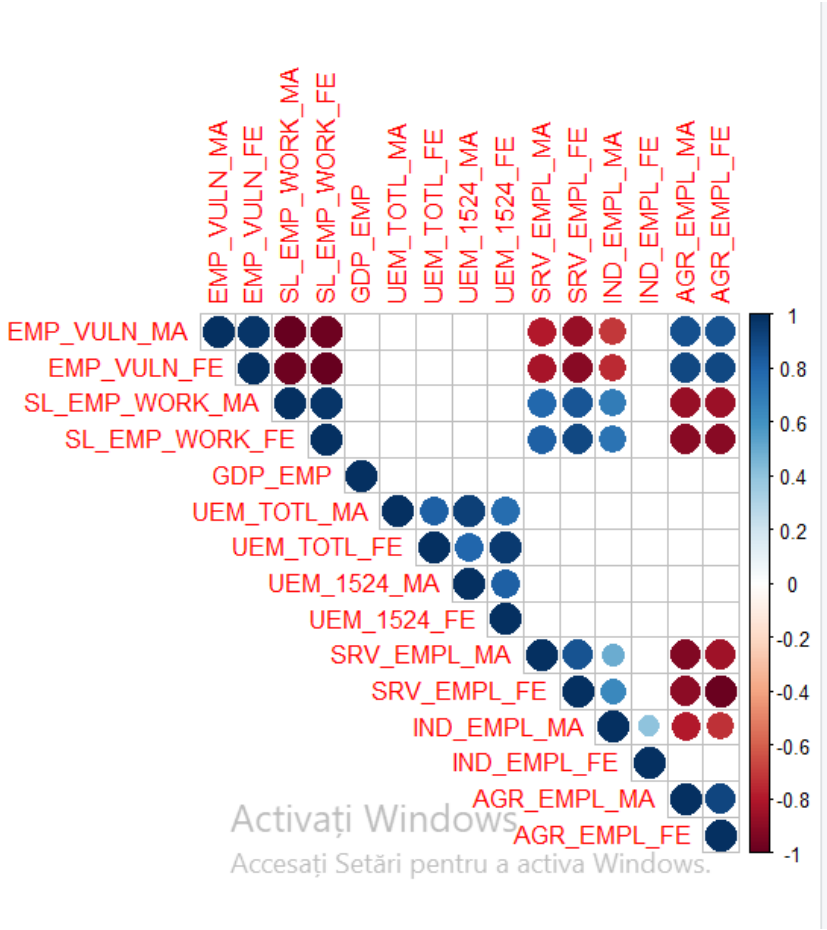
## 7. Correlation matrix and covariance matrix

The correlation matrix indicates the degree of connection and the type of connection between two variables. The higher the correlation coefficient, the stronger the correlation and the data influence each other, being thus interdependent. The sign indicates the type of connection, which can be direct, when the coefficient has a positive value, or indirect, when the coefficient has a negative value.

For example, the correlation between the number of women working in the service sector and the vulnerability of employment among women has a correlation of -0.97. This aspect indicates that they depend on each other, and the relationship is inverse, being very strongly correlated. When the number of women employed in the service sector increases, vulnerability decreases and vice versa.

```
> cor(c2)
               EMP_VULN_MA   EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE      GDP_EMP UEM_TOTL_MA UEM_TOTL_FE UEM_1524_MA UEM_1524_FE SRV_EMPL_MA SRV_EMPL_FE
EMP_VULN_MA     1.00000000  0.9765876444   -0.994409346    -0.97353360  0.14251397 -0.15283610 -0.16096901 -0.21242932 -0.21660441  -0.7990734  -0.8657256
EMP_VULN_FE     0.97658764  1.0000000000   -0.972694957    -0.99861174  0.19847048 -0.19027839 -0.18278456 -0.24346815 -0.24848990  -0.8217112  -0.9073324
SL_EMP_WORK_MA -0.99440935 -0.9726949574    1.000000000     0.97448832 -0.17068096  0.14762818  0.14640118  0.20932792  0.20359257   0.7870234   0.8554699
SL_EMP_WORK_FE -0.97353360 -0.9986117407    0.974488321     1.00000000 -0.20972813  0.19086230  0.17876645  0.24405719  0.24354654   0.8186371   0.9027432
GDP_EMP         0.14251397  0.1984704808   -0.170680956    -0.20972813  1.00000000 -0.16164283 -0.04145996 -0.18207677 -0.08082018  -0.1702308  -0.2196522
UEM_TOTL_MA    -0.15283610 -0.1902783872    0.147628178     0.19086230 -0.16164283  1.00000000  0.81046713  0.93498199  0.76809404   0.1534912   0.1874597
UEM_TOTL_FE    -0.16096901 -0.1827845557    0.146401185     0.17876645 -0.04145996  0.81046713  1.00000000  0.79665367  0.95159079   0.2155776   0.2369765
UEM_1524_MA    -0.21242932 -0.2434681543    0.209327917     0.24405719 -0.18207677  0.93498199  0.79665367  1.00000000  0.81962006   0.2842708   0.2328857
UEM_1524_FE    -0.21660441 -0.2484898966    0.203592571     0.24354654 -0.08082018  0.76809404  0.95159079  0.81962006  1.00000000   0.2983726   0.3052626
SRV_EMPL_MA    -0.79907339 -0.8217111672    0.787023447     0.81863710 -0.17023079  0.15349122  0.21557763  0.28427084  0.29837256   1.0000000   0.8621821
SRV_EMPL_FE    -0.86572564 -0.9073323577    0.855469859     0.90274321 -0.21965221  0.18745975  0.23697654  0.23288568  0.30526257   0.8621821   1.0000000
IND_EMPL_MA    -0.70725601 -0.7410452515    0.699234418     0.73976414 -0.14952945  0.06392326 -0.03165648  0.06700240  0.04134964   0.4985819   0.6469305
IND_EMPL_FE     0.02840236  0.0004471167   -0.007490503     0.01117855 -0.16533076 -0.01495784 -0.15523395 -0.02741231 -0.15183778  -0.1168321  -0.1119782
AGR_EMPL_MA     0.87472642  0.9056669619   -0.862719559    -0.90294666  0.18583571 -0.13579586 -0.13632012 -0.22847769 -0.22682793  -0.9215111  -0.8917552
AGR_EMPL_FE     0.86252243  0.9097389463   -0.856348993    -0.90742248  0.25275637 -0.18503479 -0.20711483 -0.22813863 -0.27625744  -0.8415913  -0.9807386
               IND_EMPL_MA   IND_EMPL_FE AGR_EMPL_MA AGR_EMPL_FE
EMP_VULN_MA    -0.70725601  0.0284023554   0.8747264   0.86252243
EMP_VULN_FE    -0.74104525  0.0004471167   0.9056670   0.90973895
SL_EMP_WORK_MA  0.69923442 -0.0074905029  -0.8627196  -0.85634899
SL_EMP_WORK_FE  0.73976414  0.0111785517  -0.9029467  -0.90742248
GDP_EMP        -0.14952945 -0.1653307574   0.1858357   0.25275637
UEM_TOTL_MA     0.06392326 -0.0149578355  -0.1357959  -0.18503479
UEM_TOTL_FE    -0.03165648 -0.1552339545  -0.1363201  -0.20711483
UEM_1524_MA     0.06700240 -0.0274123068  -0.2284777  -0.22813863
UEM_1524_FE     0.04134964 -0.1518377814  -0.2268279  -0.27625744
SRV_EMPL_MA     0.49858193 -0.1168321321  -0.9215111  -0.84159127
SRV_EMPL_FE     0.64693053 -0.1119782084  -0.8917552  -0.98073862
IND_EMPL_MA     1.00000000  0.4088197222  -0.7960889  -0.72906663
IND_EMPL_FE     0.40881972  1.0000000000  -0.1015899  -0.08427493
AGR_EMPL_MA    -0.79608889 -0.1015898778   1.0000000   0.91417761
AGR_EMPL_FE    -0.72906663 -0.0842749302   0.9141776   1.00000000
> |
```

Fig. 8.1 The correlation matrix




8.2. Visual representation of the correlation matrix

For a better visualization of the correlations, this visual representation is used. The color blue indicates a direct connection, and red indicates an indirect connection. The darker the color, the higher the correlation.

Covariance is the measure of common variation between two variables. Practically, it shows the amount of information that is repeated both in one variable and in another. So, taking as an example the covariance between EMP_VULN_FE and GDP_EMP, the covariance between the two is 2.175*e+05, which is very small. This aspect suggests that the redundancy between the two is very low.

```
> cov(c2)
              EMP_VULN_MA   EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE       GDP_EMP   UEM_TOTL_MA  UEM_TOTL_FE   UEM_1524_MA   UEM_1524_FE   SRV_EMPL_MA
EMP_VULN_MA       654.564228  8.419720e+02  -6.392468e+02  -8.356010e+02  1.185745e+05 -1.352479e+01   -19.102768    -43.978006    -58.275205   -274.790600
EMP_VULN_FE       841.972047  1.135588e+03  -8.235959e+02  -1.128961e+03  2.175023e+05 -2.217830e+01   -28.571149    -66.389197    -88.056095   -372.193322
SL_EMP_WORK_MA   -639.246752 -8.235959e+02   6.313271e+02   8.214398e+02 -1.394665e+05  1.282995e+01    17.062776     42.559771     53.793462    265.799367
SL_EMP_WORK_FE   -835.601021 -1.128961e+03   8.214398e+02   1.125496e+03 -2.288159e+05  2.214729e+01    27.818633     66.253438     85.919990    369.149579
GDP_EMP        118574.485023  2.175023e+05  -1.394665e+05  -2.288159e+05  1.057585e+09 -1.818204e+04 -6254.094963 -47913.419660 -27638.717573 -74410.610551
UEM_TOTL_MA       -13.524792 -2.217830e+01   1.282995e+01   2.214729e+01 -1.818204e+04  1.196347e+01    13.002948     26.168372     27.937226     7.135939
UEM_TOTL_FE       -19.102768 -2.857115e+01   1.706278e+01   2.781863e+01 -6.254095e+03  1.300295e+01    21.515715     29.901460     46.416083    13.440662
UEM_1524_MA       -43.978006 -6.638920e+01   4.255977e+01   6.625344e+01 -4.791342e+04  2.616837e+01    29.901460     65.477166     69.742560    30.918351
UEM_1524_FE       -58.275205 -8.805609e+01   5.379346e+01   8.591999e+01 -2.763872e+04  2.793723e+01    46.416083     69.742560    110.581059    42.173372
SRV_EMPL_MA      -274.790600 -3.721933e+02   2.657994e+02   3.691496e+02 -7.441061e+04  7.135939e+00    13.440662     30.918351     42.173372   180.666714
SRV_EMPL_FE      -550.843056 -7.604106e+02   5.345685e+02   7.531953e+02 -1.776498e+05  1.612531e+01    27.337243     46.866141     79.833540   288.210336
IND_EMPL_MA      -156.075831 -2.153961e+02   1.515420e+02   2.140662e+02 -4.194375e+04  1.907086e+00    -1.266553      4.676470      3.750551    57.804079
IND_EMPL_FE         3.542417  7.345145e-02  -9.175027e-01   1.828212e+00 -2.621081e+04 -2.522124e-01    -3.510216     -1.081334     -7.783759    -7.655447
AGR_EMPL_MA       430.867998  5.875904e+02  -4.173426e+02  -5.832165e+02  1.163544e+05 -9.042966e+00   -12.174011    -35.594640    -45.923265   -238.470955
AGR_EMPL_FE       547.300674  7.603377e+02  -5.336511e+02  -7.550242e+02  2.038634e+05 -1.587309e+01   -23.826955    -45.785000    -72.049970   -280.556118
              SRV_EMPL_FE   IND_EMPL_MA   IND_EMPL_FE   AGR_EMPL_MA   AGR_EMPL_FE
EMP_VULN_MA     -550.84306   -156.075831  3.542417e+00    430.867998     547.30067
EMP_VULN_FE     -760.41062   -215.396142  7.345145e-02    587.590410     760.33775
SL_EMP_WORK_MA   534.56851    151.541951 -9.175027e-01   -417.342586    -533.65114
SL_EMP_WORK_FE   753.19526    214.066169  1.828212e+00   -583.216545    -755.02419
GDP_EMP       -177649.80241 -41943.754042 -2.621081e+04 116354.376747 203863.38789
UEM_TOTL_MA        16.12531      1.907086 -2.522124e-01     -9.042966     -15.87309
UEM_TOTL_FE        27.33724     -1.266553 -3.510216e+00    -12.174011     -23.82695
UEM_1524_MA        46.86614      4.676470 -1.081334e+00    -35.594640     -45.78500
UEM_1524_FE        79.83354      3.750551 -7.783759e+00    -45.923265     -72.04997
SRV_EMPL_MA       288.21034     57.804079 -7.655447e+00   -238.470955    -280.55612
SRV_EMPL_FE       618.50459    138.775240 -1.357607e+01   -426.985334    -604.92870
IND_EMPL_MA       138.77524     74.398770  1.719032e+01   -132.202692    -155.96586
IND_EMPL_FE       -13.57607     17.190317  2.376503e+01     -9.534885     -10.18936
AGR_EMPL_MA      -426.98533   -132.202692 -9.534885e+00    370.673654     436.52175
AGR_EMPL_FE      -604.92870   -155.965862 -1.018936e+01    436.521753     615.11866
>
```

8.3. The covariance matrix

## 8. Grouping of variables

For the grouping of variables, I chose to group by the GDP_EMP variable, which is the GDP specific to salaried persons. Thus, 5 employed people have a corresponding GDP between (5000;10000], 33 people have a corresponding GDP higher than 10000 and 4 people have a GDP lower than 5000. In the given analysis, the specific GDP values thus prevail to each employee greater than $10,000, and the fewest employees contribute a GDP of less than $5,000.

```
> table(c2$MC_categorii)

GDP intre (5000;10000]  GDP mai mare de 10000    GDP mai mic de 5000
                     5                     33                      4
>
```

Fig. 9.1 Grouping of variables according to the GDP_EMP variable

## 9. Data standardization

After data standardization, it can be observed that the standard deviation for each indicator used in the analysis is 1, this being a property of standardized data.

```
> apply(date_std,2,sd)
   EMP_VULN_MA    EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE        GDP_EMP    UEM_TOTL_MA    UEM_TOTL_FE    UEM_1524_MA    UEM_1524_FE    SRV_EMPL_MA
             1              1              1              1              1              1              1              1              1              1
   SRV_EMPL_FE    IND_EMPL_MA    IND_EMPL_FE    AGR_EMPL_MA    AGR_EMPL_FE
             1              1              1              1              1
>
```

Fig. 10.1. Standard deviation for standardized data

After data standardization, it can be observed that the average for each variable takes very small values, extremely close to 0, thus respecting another property of standardized data.

```
> apply(date_std,2,mean)
   EMP_VULN_MA    EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE        GDP_EMP    UEM_TOTL_MA    UEM_TOTL_FE    UEM_1524_MA    UEM_1524_FE    SRV_EMPL_MA
 4.430773e-17  -1.253544e-17   9.591059e-17   1.044035e-16   4.756163e-17   4.478529e-17  -8.094473e-17  -4.937250e-17  -6.726184e-17   4.740545e-17
   SRV_EMPL_FE    IND_EMPL_MA    IND_EMPL_FE    AGR_EMPL_MA    AGR_EMPL_FE
 1.050721e-16  -7.468927e-17   1.415129e-16  -6.802078e-18  -5.285098e-17
>
```

Fig. 10.2 Mean for standardized data

Another property of standardized data is that the correlation matrix is equal to the covariance matrix, and this property is also respected in the case of the data used for labor force analysis, as can be seen from the matrices below:

```
> cor(date_std)
               EMP_VULN_MA   EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE      GDP_EMP UEM_TOTL_MA UEM_TOTL_FE UEM_1524_MA UEM_1524_FE SRV_EMPL_MA SRV_EMPL_FE
EMP_VULN_MA     1.00000000  0.9765876444   -0.994409346    -0.97353360  0.14251397 -0.15283610 -0.16096901 -0.21242932 -0.21660441  -0.7990734  -0.8657256
EMP_VULN_FE     0.97658764  1.0000000000   -0.972694957    -0.99861174  0.19847048 -0.19027839 -0.18278456 -0.24346815 -0.24848990  -0.8217112  -0.9073324
SL_EMP_WORK_MA -0.99440935 -0.9726949574    1.000000000     0.97448832 -0.17068096  0.14762818  0.14640118  0.20932792  0.20359257   0.7870234   0.8554699
SL_EMP_WORK_FE -0.97353360 -0.9986117407    0.974488321     1.00000000 -0.20972813  0.19086230  0.17876645  0.24405719  0.24354654   0.8186371   0.9027432
GDP_EMP         0.14251397  0.1984704808   -0.170680956    -0.20972813  1.00000000 -0.16164283 -0.04145996 -0.18207677 -0.08082018  -0.1702308  -0.2196522
UEM_TOTL_MA    -0.15283610 -0.1902783872    0.147628178     0.19086230 -0.16164283  1.00000000  0.81046713  0.93498199  0.76809404   0.1534912   0.1874597
UEM_TOTL_FE    -0.16096901 -0.1827845557    0.146401185     0.17876645 -0.04145996  0.81046713  1.00000000  0.79665367  0.95159079   0.2155776   0.2369765
UEM_1524_MA    -0.21242932 -0.2434681543    0.209327917     0.24405719 -0.18207677  0.93498199  0.79665367  1.00000000  0.81962006   0.2842708   0.2328857
UEM_1524_FE    -0.21660441 -0.2484898966    0.203592571     0.24354654 -0.08082018  0.76809404  0.95159079  0.81962006  1.00000000   0.2983726   0.3052626
SRV_EMPL_MA    -0.79907339 -0.8217111672    0.787023447     0.81863710 -0.17023079  0.15349122  0.21557763  0.28427084  0.29837256   1.0000000   0.8621821
SRV_EMPL_FE    -0.86572564 -0.9073323577    0.855469859     0.90274321 -0.21965221  0.18745975  0.23697654  0.23288568  0.30526257   0.8621821   1.0000000
IND_EMPL_MA    -0.70725601 -0.7410452515    0.699234418     0.73976414 -0.14952945  0.06392326 -0.03165648  0.06700240  0.04134964   0.4985819   0.6469305
IND_EMPL_FE     0.02840236  0.0004471167   -0.007490503     0.01117855 -0.16533076 -0.01495784 -0.15523395 -0.02741231 -0.15183778  -0.1168321  -0.1119782
AGR_EMPL_MA     0.87472642  0.9056669619   -0.862719559    -0.90294666  0.18583571 -0.13579586 -0.13632012 -0.22847769 -0.22682793  -0.9215111  -0.8917552
AGR_EMPL_FE     0.86252243  0.9097389463   -0.856348993    -0.90742248  0.25275637 -0.18503479 -0.20711483 -0.22813863 -0.27625744  -0.8415913  -0.9807386
               IND_EMPL_MA   IND_EMPL_FE AGR_EMPL_MA AGR_EMPL_FE
EMP_VULN_MA    -0.70725601  0.0284023554   0.8747264   0.86252243
EMP_VULN_FE    -0.74104525  0.0004471167   0.9056670   0.90973895
SL_EMP_WORK_MA  0.69923442 -0.0074905029  -0.8627196  -0.85634899
SL_EMP_WORK_FE  0.73976414  0.0111785517  -0.9029467  -0.90742248
GDP_EMP        -0.14952945 -0.1653307574   0.1858357   0.25275637
UEM_TOTL_MA     0.06392326 -0.0149578355  -0.1357959  -0.18503479
UEM_TOTL_FE    -0.03165648 -0.1552339545  -0.1363201  -0.20711483
UEM_1524_MA     0.06700240 -0.0274123068  -0.2284777  -0.22813863
UEM_1524_FE     0.04134964 -0.1518377814  -0.2268279  -0.27625744
SRV_EMPL_MA     0.49858193 -0.1168321321  -0.9215111  -0.84159127
SRV_EMPL_FE     0.64693053 -0.1119782084  -0.8917552  -0.98073862
IND_EMPL_MA     1.00000000  0.4088197222  -0.7960889  -0.72906663
IND_EMPL_FE     0.40881972  1.0000000000  -0.1015899  -0.08427493
AGR_EMPL_MA    -0.79608889 -0.1015898778   1.0000000   0.91417761
AGR_EMPL_FE    -0.72906663 -0.0842749302   0.9141776   1.00000000
```

Fig. 10.3 Correlation matrix of standardized data

```
> cov(date_std)
              EMP_VULN_MA   EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE     GDP_EMP UEM_TOTL_MA UEM_TOTL_FE UEM_1524_MA UEM_1524_FE SRV_EMPL_MA SRV_EMPL_FE
EMP_VULN_MA     1.00000000  0.9765876444   -0.994409346    -0.97353360  0.14251397 -0.15283610 -0.16096901 -0.21242932 -0.21660441  -0.7990734  -0.8657256
EMP_VULN_FE     0.97658764  1.0000000000   -0.972694957    -0.99861174  0.19847048 -0.19027839 -0.18278456 -0.24346815 -0.24848990  -0.8217112  -0.9073324
SL_EMP_WORK_MA -0.99440935 -0.9726949574    1.000000000     0.97448832 -0.17068096  0.14762818  0.14640118  0.20932792  0.20359257   0.7870234   0.8554699
SL_EMP_WORK_FE -0.97353360 -0.9986117407    0.974488321     1.00000000 -0.20972813  0.19086230  0.17876645  0.24405719  0.24354654   0.8186371   0.9027432
GDP_EMP         0.14251397  0.1984704808   -0.170680956    -0.20972813  1.00000000 -0.16164283 -0.04145996 -0.18207677 -0.08082018  -0.1702308  -0.2196522
UEM_TOTL_MA    -0.15283610 -0.1902783872    0.147628178     0.19086230 -0.16164283  1.00000000  0.81046713  0.93498199  0.76809404   0.1534912   0.1874597
UEM_TOTL_FE    -0.16096901 -0.1827845557    0.146401185     0.17876645 -0.04145996  0.81046713  1.00000000  0.79665367  0.95159079   0.2155776   0.2369765
UEM_1524_MA    -0.21242932 -0.2434681543    0.209327917     0.24405719 -0.18207677  0.93498199  0.79665367  1.00000000  0.81962006   0.2842708   0.2328857
UEM_1524_FE    -0.21660441 -0.2484898966    0.203592571     0.24354654 -0.08082018  0.76809404  0.95159079  0.81962006  1.00000000   0.2983726   0.3052626
SRV_EMPL_MA    -0.79907339 -0.8217111672    0.787023447     0.81863710 -0.17023079  0.15349122  0.21557763  0.28427084  0.29837256   1.0000000   0.8621821
SRV_EMPL_FE    -0.86572564 -0.9073323577    0.855469859     0.90274321 -0.21965221  0.18745975  0.23697654  0.23288568  0.30526257   0.8621821   1.0000000
IND_EMPL_MA    -0.70725601 -0.7410452515    0.699234418     0.73976414 -0.14952945  0.06392326 -0.03165648  0.06700240  0.04134964   0.4985819   0.6469305
IND_EMPL_FE     0.02840236  0.0004471167   -0.007490503     0.01117855 -0.16533076 -0.01495784 -0.15523395 -0.02741231 -0.15183778  -0.1168321  -0.1119782
AGR_EMPL_MA     0.87472642  0.9056669619   -0.862719559    -0.90294666  0.18583571 -0.13579586 -0.13632012 -0.22847769 -0.22682793  -0.9215111  -0.8917552
AGR_EMPL_FE     0.86252243  0.9097389463   -0.856348993    -0.90742248  0.25275637 -0.18503479 -0.20711483 -0.22813863 -0.27625744  -0.8415913  -0.9807386
              IND_EMPL_MA   IND_EMPL_FE AGR_EMPL_MA AGR_EMPL_FE
EMP_VULN_MA    -0.70725601  0.0284023554   0.8747264   0.86252243
EMP_VULN_FE    -0.74104525  0.0004471167   0.9056670   0.90973895
SL_EMP_WORK_MA  0.69923442 -0.0074905029  -0.8627196  -0.85634899
SL_EMP_WORK_FE  0.73976414  0.0111785517  -0.9029467  -0.90742248
GDP_EMP        -0.14952945 -0.1653307574   0.1858357   0.25275637
UEM_TOTL_MA     0.06392326 -0.0149578355  -0.1357959  -0.18503479
UEM_TOTL_FE    -0.03165648 -0.1552339545  -0.1363201  -0.20711483
UEM_1524_MA     0.06700240 -0.0274123068  -0.2284777  -0.22813863
UEM_1524_FE     0.04134964 -0.1518377814  -0.2268279  -0.27625744
SRV_EMPL_MA     0.49858193 -0.1168321321  -0.9215111  -0.84159127
SRV_EMPL_FE     0.64693053 -0.1119782084  -0.8917552  -0.98073862
IND_EMPL_MA     1.00000000  0.4088197222  -0.7960889  -0.72906663
IND_EMPL_FE     0.40881972  1.0000000000  -0.1015899  -0.08427493
AGR_EMPL_MA    -0.79608889 -0.1015898778   1.0000000   0.91417761
AGR_EMPL_FE    -0.72906663 -0.0842749302   0.9141776   1.00000000
>
```

Fig. 10.4. The covariance matrix of the standardized data

## 10. Principal Components Analysis

### 11.1. The eigenvalues of the covariance matrix

The eigenvalues identified in the analysis of the main components are those in the figure below, with the following explanations:

- The valp column contains the eigenvalues for each variable used in the analysis. The
- percentA column shows the percentage of information that each component retains in the analysis.
- The cumulative information percentage can be found on the percentageC column.

```
> V=zapsmall(data.frame(valp,procentA,procentC))
> V
          valp procentA  procentC
Comp.1  8.20437 54.69578  54.69578
Comp.2  3.30837 22.05580  76.75159
Comp.3  1.32824  8.85492  85.60651
Comp.4  0.89558  5.97056  91.57707
Comp.5  0.41238  2.74918  94.32625
Comp.6  0.31003  2.06684  96.39309
Comp.7  0.24742  1.64949  98.04259
Comp.8  0.17431  1.16204  99.20463
Comp.9  0.06442  0.42950  99.63413
Comp.10 0.03277  0.21846  99.85258
Comp.11 0.01671  0.11138  99.96397
Comp.12 0.00519  0.03457  99.99854
Comp.13 0.00022  0.00146 100.00000
Comp.14 0.00000  0.00000 100.00000
Comp.15 0.00000  0.00000 100.00000
>
```

Fig. 11.1.1 Own values and percentage of information

## 11.2. Criteria for choosing the number of main components

### A) Kaiser criterion

According to the Kaiser criterion, it is observed that the eigenvalues that are greater than 1 are three in number, and this means that in the analysis of the principal components, we will use 3 principal components.

### B) Cumulative percentage criterion

According to the coverage percentage criterion, we will keep in the analysis the 3 main components which together have a cumulative percentage of 85.606% data.

### C) Screeplot (Slope Criterion)

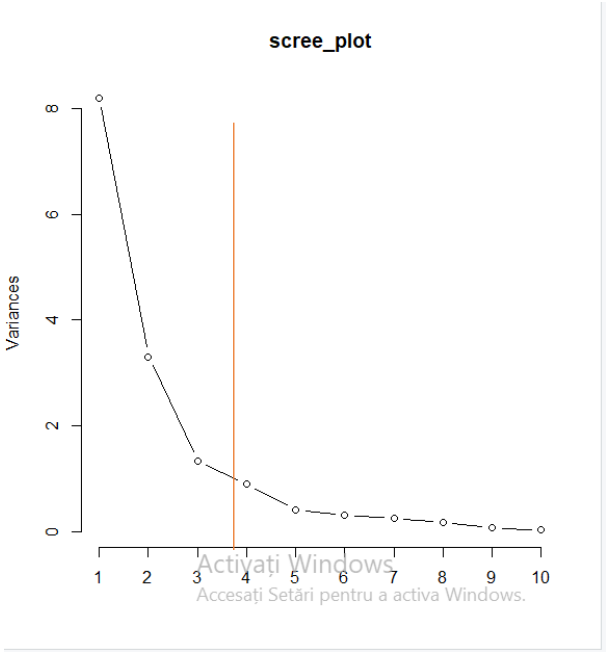



Fig. 11.2.1

The third criterion is the slope criterion, which graphically represents what the previous criteria showed in numbers. A line parallel to the ordinate Oy is drawn where, to the right of this cut, the graph has a slope close to 0. In this case, this line is drawn between 3 and 4. The number of the main components is the first integer to the left of the cut. Thus, from the figure above it can be seen that there are 3 main components.

## 11.3. The eigenvectors of the covariance matrix

These are all the eigenvectors used in the analysis of the main components and for the calculation of the general shape of the main components, the values from the table below will be taken over the columns.

Fig. 11.3.1 Eigenvectors from ACP

```
> e$vectors#=loadings #vectorii proprii
             [,1]        [,2]        [,3]         [,4]        [,5]        [,6]         [,7]        [,8]        [,9]       [,10]        [,11]        [,12]
 [1,] -0.329855912  0.09008853  0.08807722  0.047300597  0.31779690 -0.03759548 -0.264398370 -0.06499450 -0.03705218  0.41502082 -0.098128402  0.641075854
 [2,] -0.339199578  0.07705584  0.04365567  0.011798367  0.21866742 -0.05029809 -0.113401396  0.02539316  0.02447582 -0.52650126  0.096881432  0.331228958
 [3,]  0.327666818 -0.09450041 -0.06662445 -0.023591319 -0.34665026  0.03249500  0.321820638  0.03813712 -0.02191510 -0.39476457  0.119290849  0.627154937
 [4,]  0.338615199 -0.07813308 -0.03234689 -0.004415985 -0.22712663  0.04235629  0.134740944 -0.03213930 -0.02122318  0.56395148 -0.114462078  0.286015465
 [5,] -0.084366349 -0.02987020 -0.42124207 -0.880119631  0.04494855 -0.12804251 -0.003242261 -0.13396855 -0.04988448  0.02754583  0.001369298  0.017598957
 [6,]  0.109674643  0.47725265  0.15928320 -0.048541715 -0.29499530 -0.29708125 -0.271201888 -0.30003554  0.41372040  0.06870099  0.466559757  0.007543745
 [7,]  0.110790823  0.49117639 -0.04498840 -0.094733773  0.14934903  0.41188625  0.171299827  0.02663292  0.51269209 -0.08799205 -0.491759474  0.027621890
 [8,]  0.132226215  0.47157926  0.12811676 -0.014934199 -0.14222955 -0.46781205 -0.063095760  0.06595939 -0.50904066 -0.12018392 -0.471510582 -0.013803456
 [9,]  0.135611986  0.47267305 -0.03900382 -0.068346595  0.23915765  0.38654569  0.128253626  0.27480145 -0.43073067  0.10917962  0.505563403 -0.011977382
[10,]  0.305949772 -0.03051478 -0.17337676  0.101150730  0.47900475 -0.46164892  0.174614700  0.22076527  0.21495462  0.02802569  0.086434047  0.012258481
[11,]  0.328394178 -0.04632473 -0.12487438  0.097748718  0.22540250  0.13220205 -0.233323731 -0.46165789 -0.14516456 -0.12112923 -0.036338420  0.005464553
[12,]  0.259118473 -0.16730163  0.33337866 -0.279062793 -0.11229840  0.23447763 -0.603882416  0.38870343  0.02681040 -0.08422915 -0.067334255  0.035182422
[13,]  0.007980635 -0.10212637  0.77699634 -0.323107732  0.21291206 -0.05525144  0.421478642 -0.18478751 -0.00736796  0.02096023  0.033189978 -0.026701247
[14,] -0.329683797  0.09625740 -0.02831461  0.054404820 -0.28409533  0.21725184  0.148628486 -0.32826644 -0.16210043  0.01819564 -0.030148407 -0.024216081
[15,] -0.330865911  0.06652597 -0.02750991 -0.034510106 -0.26787862 -0.12169897  0.151117186  0.49923731  0.14701017  0.11732664  0.029889086 -0.000250747
             [,13]         [,14]         [,15]
 [1,]  0.3198789182  9.195794e-05 -4.895265e-05
 [2,] -0.6451431701 -3.835455e-05  1.046987e-04
 [3,]  0.2976415640  6.799121e-05 -4.918669e-05
 [4,] -0.6262606559 -2.613469e-05  1.051700e-04
 [5,]  0.0015728546  2.396707e-06 -2.453161e-06
 [6,]  0.0018316354  6.638327e-06  5.714274e-06
 [7,]  0.0166795973 -1.699344e-05 -9.751354e-06
 [8,] -0.0019718028 -8.211999e-06 -7.131285e-06
 [9,] -0.0152094260  2.113717e-05  1.058837e-05
[10,]  0.0065574583 -5.319815e-01  7.563340e-02
[11,] -0.0051212164  9.871185e-02  6.943670e-01
[12,]  0.0040763221 -3.413777e-01  4.852799e-02
[13,] -0.0007007692  1.934629e-02  1.361130e-01
[14,] -0.0062582130 -7.620028e-01  1.083205e-01
[15,]  0.0054868754  9.843988e-02  6.924731e-01
>
```

The specific eigenvectors of the principal components are as follows

Fig. 11.3.2. The eigenvectors of the 3 comp. main

```
> c2 #Am retinut primii vectori proprii pentru cele 3 componente principale
         Comp.1    Comp.2    Comp.3
 [1,]  4.411673 -1.749942  1.294171
 [2,]  3.026396  1.260815 -1.102388
 [3,]  1.196285  3.661043  1.682635
 [4,] -0.948533  2.685479 -0.350203
 [5,] -2.473525 -1.966173 -2.405409
 [6,] -2.575237  1.166267 -0.336947
 [7,] -2.269634 -1.226215 -0.582290
 [8,] -2.070877 -1.991717 -0.612383
 [9,]  2.314845 -0.117549 -1.411386
[10,] -2.628825 -0.530897 -0.441014
[11,]  4.422324 -1.550431  1.958221
[12,]  2.968440 -0.456157  0.237857
[13,] -2.092581 -1.517916  2.557512
[14,] -2.712578 -3.112345  0.001266
[15,] -3.162869  1.888713 -1.037627
[16,] -2.008270 -1.837160  1.969354
[17,] -0.818119  1.249143 -1.474930
[18,]  1.966309 -1.478382 -0.002986
[19,] -1.864472  2.480437  0.185205
[20,] -2.757594  1.417680 -0.189826
[21,] -3.318993  1.508193 -0.493871
[22,]  4.488059 -0.678747 -0.199361
[23,] -1.623504  4.795974  0.279832
[24,]  6.291033  1.057581 -0.085830
[25,] -1.563450 -1.736961  2.129601
[26,] -2.103263 -1.889745 -0.528099
[27,] -2.460814 -0.541280  0.909394
[28,] -1.341591 -0.080651 -0.535660
[29,]  4.214321 -1.197060 -1.920960
[30,]  4.103257 -1.115013 -0.901062
[31,]  5.607516 -0.186558 -1.467804
[32,]  2.564471  2.041105  2.383428
[33,]  0.018925  0.972355  0.577376
[34,]  4.195942 -0.588205  0.757844
[35,] -1.401246  1.831150  1.085981
[36,] -1.688555  0.733434  0.032456
[37,] -1.369870  1.028760 -0.021462
[38,] -1.028114 -2.475803 -0.611759
[39,] -2.724936  0.201233 -0.424699
[40,] -2.376257 -2.678680  0.829330
[41,]  2.054293  2.456205 -1.009918
[42,] -2.460379 -1.731981 -0.723592
>
```

10.4. Factor matrix with principal components

According to the factor matrix, component 1 correlates strongly positively with EMP_VULN_MA, EMP_VULN_FE and also correlates strongly but negatively with SL_EMP_WORK_MA, SL_EMP_WORK_FE, SRV_EMPL_MA, SRV_EMPLFE, IND_EMPL_MA, AGR_EMPL_MA, AGR_EMPL_FE.

At the same time, component 2 correlates strongly positively with the following variables: UEM_TOTL_MA, UEM_TOTL_FE, UEM_1524_MA, UEM_1524_FE.

Component 3 correlates strongly only with the variable IND_EMPL_FE.

Fig.11.4.1 Factor matrix

```
> corFact
                  Comp.1     Comp.2     Comp.3
EMP_VULN_MA     0.9448151  0.1638614  0.1015084
EMP_VULN_FE     0.9715784  0.1401563  0.0503129
SL_EMP_WORK_MA -0.9385448 -0.1718861 -0.0767842
SL_EMP_WORK_FE -0.9699045 -0.1421157 -0.0372796
GDP_EMP         0.2416528 -0.0543307 -0.4854782
UEM_TOTL_MA    -0.3141440  0.8680714  0.1835727
UEM_TOTL_FE    -0.3173411  0.8933972 -0.0518488
UEM_1524_MA    -0.3787391  0.8577521  0.1476536
UEM_1524_FE    -0.3884370  0.8597416 -0.0449516
SRV_EMPL_MA    -0.8763401 -0.0555031 -0.1998154
SRV_EMPL_FE    -0.9406282 -0.0842597 -0.1439168
IND_EMPL_MA    -0.7421999 -0.3043038  0.3842163
IND_EMPL_FE    -0.0228591 -0.1857569  0.8954823
AGR_EMPL_MA     0.9443221  0.1750819 -0.0326323
AGR_EMPL_FE     0.9477081  0.1210036 -0.0317049
>
```

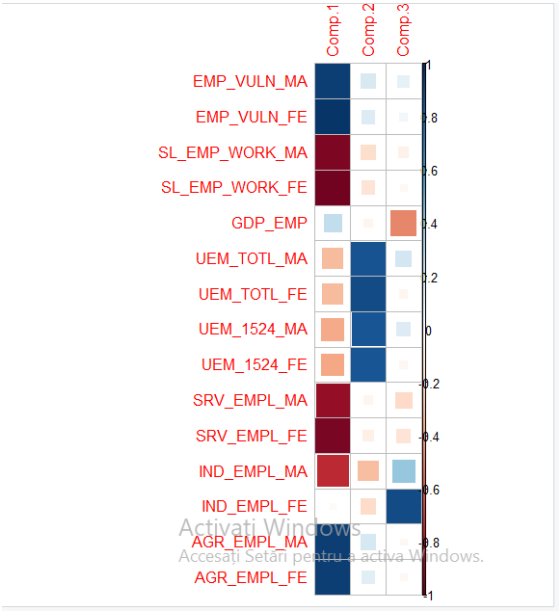




Fig. 11.4.2 Visual representation of the factor matrix

Analyzing in more detail the method of choosing the main components, we notice that the eigenvalue for component 3 is very close to 1, more precisely with the value of 1.32, and the level of information brought by it is a low one, of 8.85%. In addition, the most important argument why we choose not to consider the first 3 main components, but only the first 2, is the information provided by the factor matrix, where we observe that only one indicator correlates strongly with this component.

Therefore, the 2 components have the following names:

- Component 1 (z1) will be called**the vulnerability component of employment.**
- Component 2 (z2) will be called**component of the unemployed population.**

11.5. The general shape of the main components

z1=0.944*EMP_VULN_MA+0.971*EMP_VULN_FE-0.938*SL_EMP_WORK_MA-0.969*SL_EMP_WORK_
0.022*IND_EMPL_FE-0.944*AGR_EMPL_MA+0.947*AGR_EMPL_FE

z2=0.163*EMP_VULN_MA+0.140*EMP_VULN_FE-0.171*SL_EMP_WORK_MA-0.142*SL_EMP_WORK_
0.054*GDP_EMP+0.868*UEM_TOTL_MA+0.893*UEM_TOTL_FE+0.857*UEM_1524_
MA+0.859*UEM_1524_FE-0.055*SRV_EMPL_MA-0.084*SRV_EMPL_FE-0.304*IND_EMPL_MA-0.185*IND_

0.175*AGR_EMPL_MA+0.121*AGR_EMPL_FE

z3=0.101*EMP_VULN_MA+0.050*EMP_VULN_FE-0.076*SL_EMP_WORK_MA-0.037*SL_EMP_WORK_

0.384*IND_EMPL_MA+0.895*IND_EMPL_FE-0.032*AGR_EMPL_MA-0.031*AGR_EMPL_FE

11.6. Main scores

To calculate the main scores in the formulas with which we calculated the eigenvectors, we enter the standardized values of the initial indicators on the lines.

There are 42 observations and 15 indicators, which means that the total number of scores will be 42*15=630 scores

Example scores:

z11

```
> z1 #4.358837
          [,1]
[1,] 4.358837
```

z12

```
> z2
          [,1]
[1,] -1.728984
>
```

Main score for the first main component:

Comp.1=0.944*2.0805+0.971*1.7237+0.938*1.9655+0.969*1.6718-
0.241*1.0533+0.314*1.4655+0.317*1.1743+0.378*1.5469+0.388*1.3015*0.096+0.8076
940*1.5791+0.742*0.8-0.022*1.0751-0.944*0.3652+0.947*1.3721

11.7. Representation of observations in the main plan


Fig. 11.7.1 Obs. representation. In the main plane z1 and z2

For this graph, the areas where a cloud of points is formed, but also the points outside it, are interpreted. Thus, it is observable that the cloud of points agglomerates between the values z1= approximately -2 and z2=approximately -2. This aspect suggests the fact that for these countries BLZ, COD, AND, BEL, ARG, BMD the vulnerability of employment is low, so obviously unemployment in the respective countries is also low. A low employment vulnerability rate indicates that the population finds jobs quite easily, therefore unemployment takes very low values, so there are not many vulnerable groups from this point of view.

Another smaller point cloud is CAN, BHS,ATG,ABW,BRN etc. The vulnerability of employment is higher for them, that is, among the population, there are more people prone to not finding a job, more groups vulnerable to this aspect, but despite all this, the unemployment rate is low, a fact that indicates a high availability of the labor force and an economic development of the respective countries, since, although the preponderance of vulnerable groups is high, they are still employed in large numbers.

The other countries outside the cloud of points, such as AFG, BWA, CMR, have a high vulnerability of the employment of the labor force, which in turn induces a high unemployment rate, a situation, as can be seen, found in most of the countries subject to this analysis.

In the case of the BIH country (Bosnia & Herzegovina) vulnerability is low, but unemployment is high, an aspect that denotes a very poor development of the labor force, since, although it does not predominate

groups vulnerable to employment, the unemployment rate is high, so people find it difficult to find a job.

11.8. Biplot and the circle of correlations

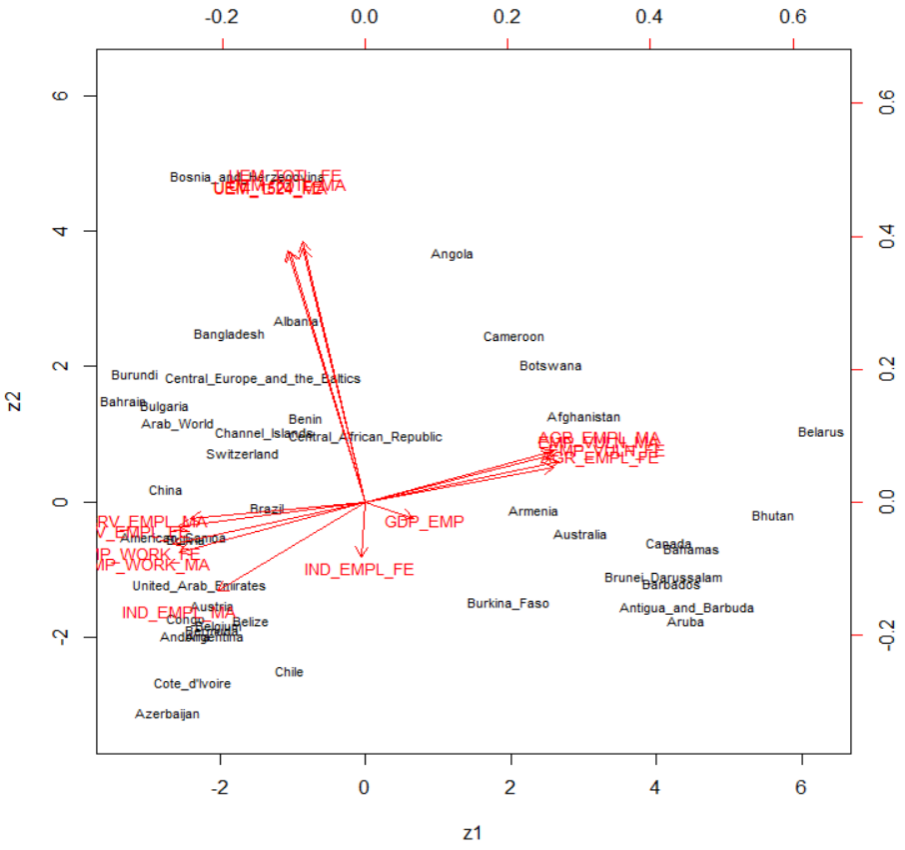



Fig.11.8.1 Biplot for ACP

This representation allows us to see those countries that are similar from the point of view of the two main components, which are the countries that register high values for certain indicators and also the correlations between the initial variables, identified by the angle between them. The smaller the angle, the more correlated the data.

The main component 1 is correlated with IND_EMPL_MA, SERV_EMP_MA, SERV_EMP_FE, with SL_EMP_WORK_MA, SL_EMP_WO RK_FE, IND_EMPL_MA, and these are highly correlated because the angle between them is very small and positively correlated with AGR_EMP_MA, AGR_EMP_FE, which are also highly correlated, and z2 is negatively correlated , but very weak with IND_EMPL_FE, GDP_EMP, but these are not as correlated as the others, because the angle between them is smaller and positively correlated with UNEM_TOTAL_FE, UNEM_1524_MA and UNEM_TOT_MA, which are also correlated.

From the point of view of the two main components, Afghanistan and Armenia are similar, where vulnerability is high and so is unemployment. They are similar

as well as Albania, Central African Republic, Benin, etc., where vulnerability is low and unemployment is high. Again, similar are China, the United Arab Emirates, etc. where both vulnerability and unemployment are low.

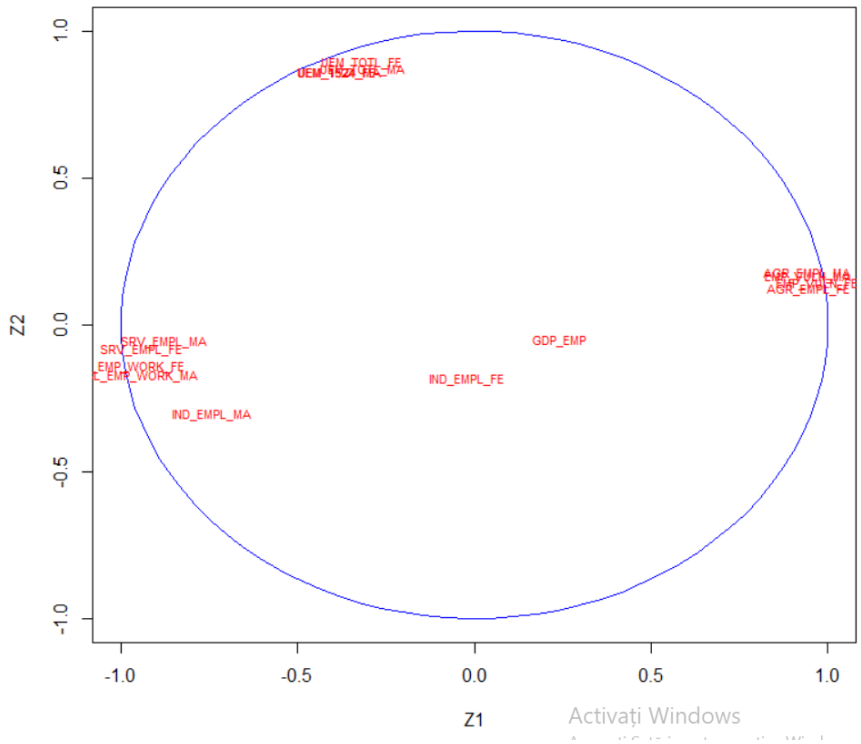



Fig.11.8.1 Circle of correlations for ACP

The circle of correlations has the same data interpretation as the Biplot analyzed above. The difference between these two graphs is visual. In the case of the circle of correlations, the proximity of the indicators to the circumference of the circle is taken into account: more precisely, the closer the indicators are to the circumference, the higher the correlation with the main components. However, the closer the indicators are to each other, the higher the correlation between them.

## 11.9. PCA function

The table below presents a summary for the analysis of the main components. It can be seen that the eigenvalues of the correlation matrix are displayed, i.e. the variants of the main components. For each country under analysis, the distance to the centroid of the cloud of points (dist), the coordinates on each dimension (dim) and the quality of representation on each axis (cos2) are calculated. The contribution of a variable to the definition of the main component is calculated as the ratio between the cos2 value and the sum of all the cos2 values associated with that component.

For example, in the data analysis for this project, the contribution of the EMP_VULN_MA indicator is 10,880. This is calculated in the form: 0.893/ (0.893+0.944+0.881+0.941+0.058+0.099+0.101+0.143+0.151+0.768).

Fig. 11.9.1 Output PCA

```
> cp1=PCA(date_std)
> summary(cp1)

Call:
PCA(X = date_std)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14  Dim.15
Variance               8.204   3.308   1.328   0.896   0.412   0.310   0.247   0.174   0.064   0.033   0.017   0.005   0.000   0.000   0.000
% of var.             54.696  22.056   8.855   5.971   2.749   2.067   1.649   1.162   0.429   0.218   0.111   0.035   0.001   0.000   0.000
Cumulative % of var.  54.696  76.752  85.607  91.577  94.326  96.393  98.043  99.205  99.634  99.853  99.964  99.999 100.000 100.000 100.000

Individuals (the 10 first)
                  Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                |  5.449 | -4.412  5.648  0.655 | -1.750  2.204  0.103 |  1.294  3.002  0.056 |
2                |  3.942 | -3.026  2.658  0.589 |  1.261  1.144  0.102 | -1.102  2.178  0.078 |
3                |  4.385 | -1.196  0.415  0.074 |  3.661  9.646  0.697 |  1.683  5.075  0.147 |
4                |  3.396 |  0.949  0.261  0.078 |  2.685  5.190  0.625 | -0.350  0.220  0.011 |
5                |  4.212 |  2.474  1.776  0.345 | -1.966  2.782  0.218 | -2.405 10.372  0.326 |
6                |  3.057 |  2.575  1.925  0.710 |  1.166  0.979  0.146 | -0.337  0.204  0.012 |
7                |  2.800 |  2.270  1.495  0.657 | -1.226  1.082  0.192 | -0.582  0.608  0.043 |
8                |  3.442 |  2.071  1.245  0.362 | -1.992  2.855  0.335 | -0.612  0.672  0.032 |
9                |  3.052 | -2.315  1.555  0.575 | -0.118  0.010  0.001 | -1.411  3.571  0.214 |
10               |  2.817 |  2.629  2.006  0.871 | -0.531  0.203  0.036 | -0.441  0.349  0.025 |

Variables (the 10 first)
                 Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
EMP_VULN_MA    | -0.945 10.880  0.893 |  0.164  0.812  0.027 |  0.102  0.776  0.010 |
EMP_VULN_FE    | -0.972 11.506  0.944 |  0.140  0.594  0.020 |  0.050  0.191  0.003 |
SL_EMP_WORK_MA |  0.939 10.737  0.881 | -0.172  0.893  0.030 | -0.077  0.444  0.006 |
SL_EMP_WORK_FE |  0.970 11.466  0.941 | -0.142  0.610  0.020 | -0.037  0.105  0.001 |
GDP_EMP        | -0.242  0.712  0.058 | -0.054  0.089  0.003 | -0.485 17.744  0.236 |
UEM_TOTL_MA    |  0.314  1.203  0.099 |  0.868 22.777  0.754 |  0.184  2.537  0.034 |
UEM_TOTL_FE    |  0.317  1.227  0.101 |  0.893 24.125  0.798 | -0.052  0.202  0.003 |
UEM_1524_MA    |  0.379  1.748  0.143 |  0.858 22.239  0.736 |  0.148  1.641  0.022 |
UEM_1524_FE    |  0.388  1.839  0.151 |  0.860 22.342  0.739 | -0.045  0.152  0.002 |
SRV_EMPL_MA    |  0.876  9.361  0.768 | -0.056  0.093  0.003 | -0.200  3.006  0.040 |
>
```

In synthesizing the information presented above, the graphic method is also used.

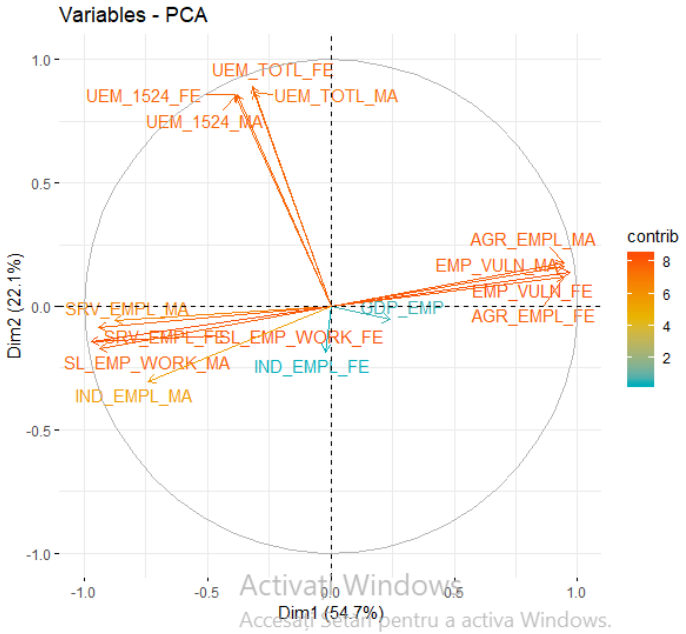


Fig. 11.9.2 Representation of the variables in the space of the new axes

The color of the indicators represented on this graphic indicates the value of their contribution to the formation of the main components. Thus, most of the indicators are colored orange and indicate a very large contribution, and some are yellow, indicating a slightly smaller contribution than the one previously mentioned. It can be seen that the GDP_EMP and IND_EMP_FE indicators do not contribute much to the formation of the main components, being colored blue.

Also, the closer the indicators are to the circumference of the circle, the more important they are in the interpretation of the main components. Therefore, all the indicators are very close to the circumference of the circle, with the exception of the indicators IND_EMPL_FE and GDP_EMP.

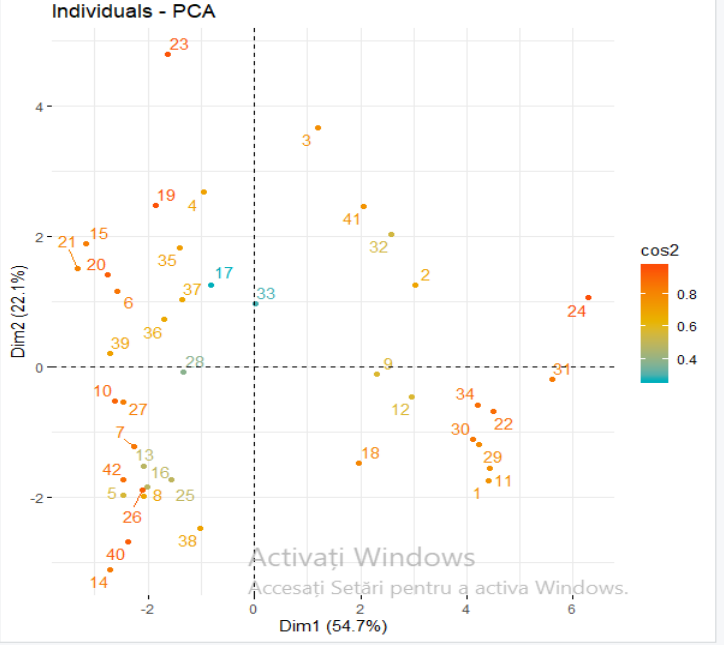



Fig. 11.9.3 The quality of the representation of the indicators

This graphic shows the quality of the representation of the indicators. Thus, from this point of view, the quality of representation for most countries is high, being colored orange, but for 2 countries, the quality is low.

### 11. Correspondence analysis

Correspondence analysis represents a method of dimensionality reduction. This is considered the equivalent of principal component analysis for categorical variables.

The data used for this category are:

| | AGR_EMPL_FE | IND_EMPL_FE | SRV_EMPL_FE |
|---|---|---|---|
| Aruba | 55.728 | 15.844 | 28.427 |
| Afghanistan | 55.529 | 1.114 | 43.357 |
| Angola | 42.289 | 15.015 | 42.696 |
| Albania | 29.863 | 9.187 | 60.950 |
| Andorra | 0.128 | 6.063 | 93.809 |
| Arab_World | 0.015 | 8.975 | 91.010 |
| United_Arab_Emirates | 1.629 | 7.576 | 90.795 |
| Argentina | 3.628 | 11.195 | 85.177 |
| Armenia | 41.935 | 5.976 | 52.089 |
| American_Samoa | 0.693 | 8.452 | 90.855 |
| Antigua_and_Barbuda | 33.778 | 17.630 | 48.592 |
| Australia | 59.432 | 16.737 | 23.830 |
| Austria | 4.472 | 22.354 | 73.174 |
| Azerbaijan | 0.047 | 8.973 | 90.980 |
| Burundi | 0.440 | 3.342 | 96.217 |
| Belgium | 7.209 | 19.658 | 73.133 |
| Benin | 5.233 | 8.137 | 86.630 |
| Burkina_Faso | 28.252 | 9.938 | 61.811 |
| Bangladesh | 4.230 | 10.577 | 85.193 |
| Bulgaria | 1.790 | 9.185 | 89.024 |
| Bahrain | 0.430 | 9.028 | 90.542 |
| Bahamas | 64.807 | 9.107 | 26.086 |
| Bosnia_and_Herzegovina | 17.858 | 9.706 | 72.437 |
| Belarus | 74.985 | 7.128 | 17.887 |
| Belize | 8.529 | 19.191 | 72.280 |
| Bermuda | 2.552 | 9.984 | 87.464 |
| Bolivia | 3.225 | 13.006 | 83.769 |
| Brazil | 5.251 | 10.708 | 84.041 |
| Barbados | 40.140 | 3.622 | 56.238 |
| Brunei_Darussalam | 51.859 | 10.480 | 37.662 |
| Bhutan | 76.792 | 4.385 | 18.823 |
| Botswana | 36.485 | 22.222 | 41.292 |
| Central_African_Republic | 7.616 | 13.824 | 78.560 |
| Canada | 63.014 | 12.609 | 24.377 |
| Central_Europe_and_the_Baltics | 8.165 | 12.013 | 79.822 |
| Switzerland | 4.472 | 9.801 | 85.727 |
| Channel_Islands | 5.691 | 8.236 | 86.072 |
| Chile | 7.624 | 10.478 | 81.898 |
| China | 1.059 | 6.852 | 92.089 |
| Cote_d'Ivoire | 0.869 | 13.630 | 85.500 |
| Cameroon | 52.520 | 4.555 | 42.925 |
| Congo | 0.992 | 8.819 | 90.189 |

```
> 
```

AGR_EMPL_FE = the female population working in the agricultural sector

IND_EMPL_FE= the female population working in the industry sector

SRV_EMPL_FE=female population working in the service sector

To see if the variables are dependent, the chi square is calculated.

```
> chisq

        Pearson's Chi-squared test

data:  dd
X-squared = 1628.9, df = 82, p-value < 2.2e-16

> 
```

It can be observed, thus, that it has the value of 1628.9 with 82 degrees of freedom which is significant from a statistical point of view, because the P-value has a value very close to 0 (P-value<2.2e-16). Therefore, it was shown that there is a connection between lines and columns and the correspondence analysis can be applied.

The Variance column represents the eigenvalues representative of the two dimensions, and the bottom one represents the percentage of variation, followed by the cumulative one. In other words, dimension 1 takes 93.70 of the information, and dimension 2 takes 6.30.

The contribution of the two dimensions to inertia is explained by the calculation formula ctr1*variance1+ctr2*variance2.

Column Ctr shows the contribution of each country to dimension 1, respectively 2. For example, to dimension 1, Aruba contributes with 4.95, and to dimension 2 with 3.13.

Fig. 12.1 Output summary analysis of correspondences

```
Eigenvalues
                       Dim.1  Dim.2
Variance               0.36   0.02
% of var.             93.70   6.30
Cumulative % of var.  93.70 100.00

Rows (the 10 first)
                        Iner*1000   Dim.1   ctr   cos2   Dim.2   ctr   cos2
Aruba                |      18.75 |  0.87  4.95  0.96 |  0.18  3.13  0.04 |
Afghanistan          |      16.67 |  0.78  4.00  0.87 | -0.30  8.78  0.13 |
Angola               |       7.29 |  0.53  1.86  0.93 |  0.15  2.16  0.07 |
Albania              |       0.94 |  0.19  0.25  0.95 | -0.04  0.19  0.05 |
Andorra              |       7.97 | -0.56  2.04  0.93 | -0.15  2.29  0.07 |
Arab_World           |       7.13 | -0.54  1.94  0.99 | -0.06  0.33  0.01 |
United_Arab_Emirates |       6.50 | -0.51  1.72  0.96 | -0.10  1.04  0.04 |
Argentina            |       4.66 | -0.44  1.28  1.00 |  0.01  0.02  0.00 |
Armenia              |       5.83 |  0.47  1.47  0.91 | -0.15  2.06  0.09 |
American_Samoa       |       6.83 | -0.53  1.84  0.98 | -0.08  0.55  0.02 |

Columns
                        Iner*1000    Dim.1    ctr    cos2    Dim.2    ctr    cos2
AGR_EMPL_FE          |     276.76 |   1.13  76.01   1.00 |  -0.05   2.30   0.00 |
IND_EMPL_FE          |      21.88 |  -0.02   0.01   0.00 |   0.45  89.39   1.00 |
SRV_EMPL_FE          |      89.18 |  -0.36  23.98   0.98 |  -0.05   8.32   0.02 |
> |
```

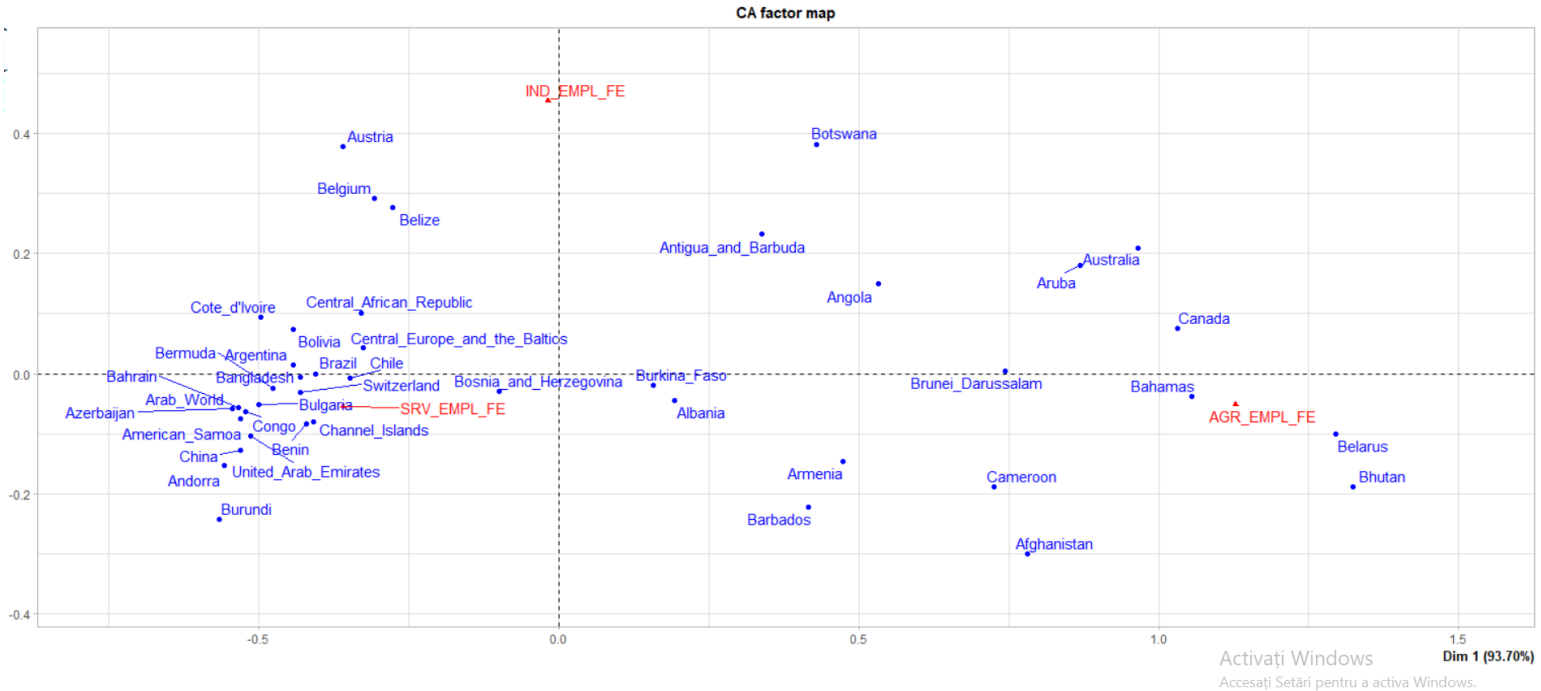


Fig. 12.2 Representation in main coordinates

The interpretation of this graph is done by studying the grouping of the decks. For example, in this analysis, we can state that in most of the countries under analysis, such as Bulgaria, Switzerland, Brazil, Chile, the Arab World, Congo, etc., women work in the field of services. Women do not work in industry, this aspect being self-evident, due to the very difficult working conditions, this being a sector in which men predominate.

In countries such as the Bahamas, Belarus, Bhutan, the majority of women work in the agricultural sector.

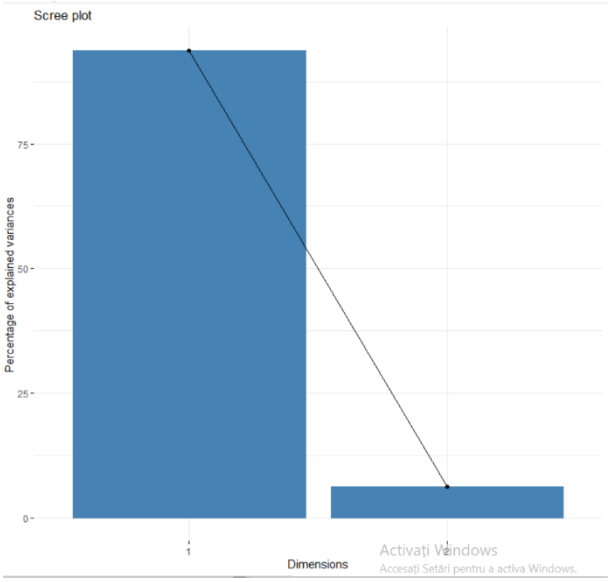



Fig. 12.3 Write plot

This Scree Plot is interpreted similarly to the ACP, so it can be seen that 2 dimensions will be used in the analysis.

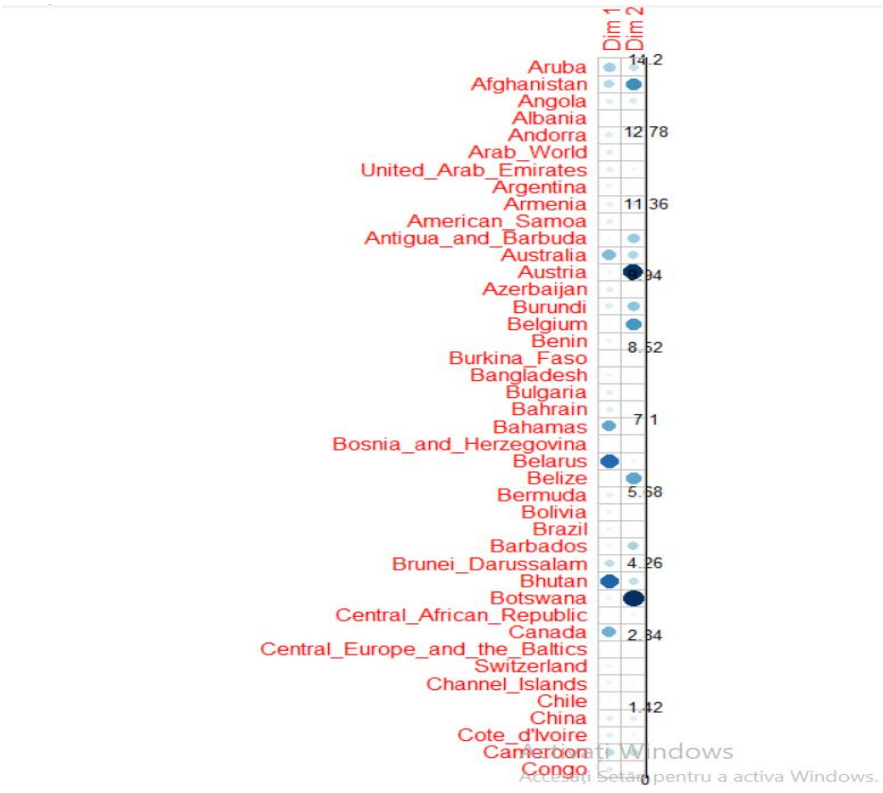


Fig. 12.4 The contribution of each line to the dimensions

From the figure above, it can be seen that in dimension 1, Bhutan and Belarus contribute more strongly, and in dimension 2, Austria and Botswana contribute more strongly, but also Afghanistan.

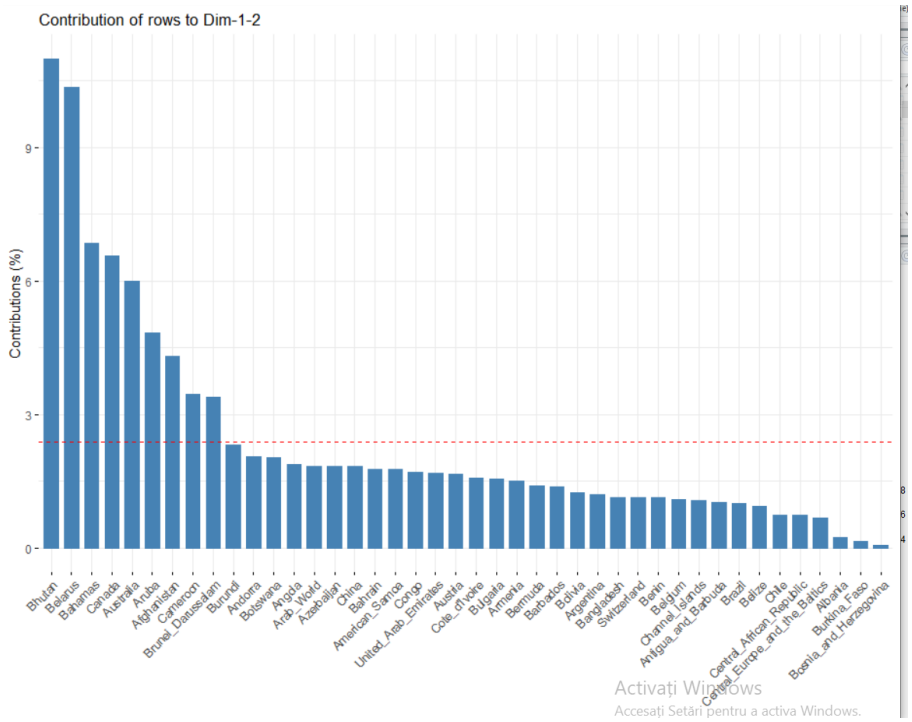



Fig. 12.5 The total contribution, to dimension 1 and 2 of the line

This graphic shows the total contribution, from dimension 1 and dimension 2 of each line. Thus, it is visible that Bhutan has the largest contribution, while Bosnia and Herzegovina has the smallest.

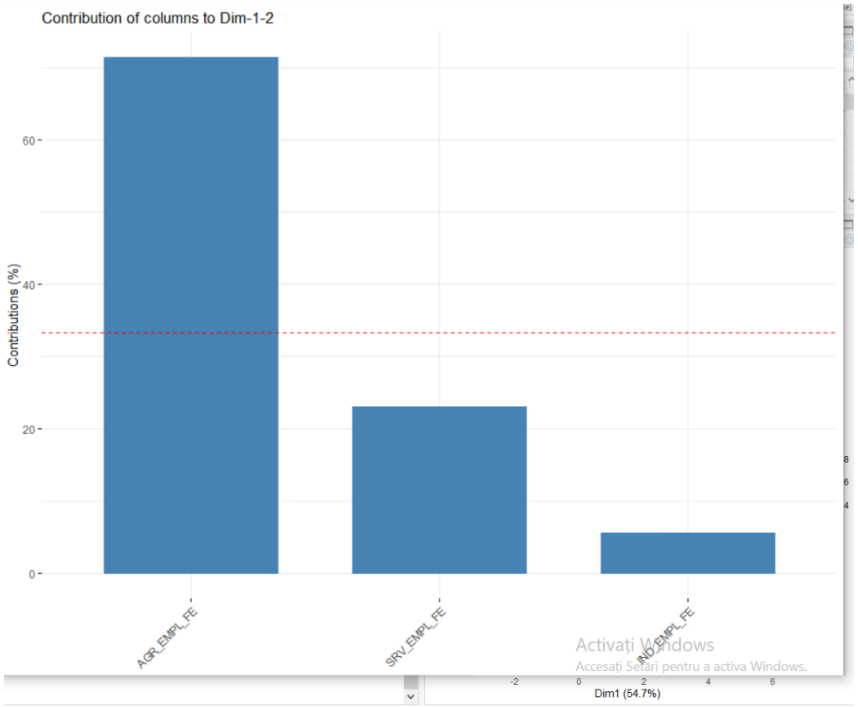


Fig. 12.6 The contribution of each column to the dimension. 1 and 2 added

Therefore, it can be seen from this image that the AGR_EMPL_FE indicator has the largest contribution to the two dimensions, being followed by SRV_EMPL_FE, and as expected, IND_EMPL_FE.

## II. Data Analysis - Part Two

### 1. Cluster analysis

Cluster analysis aims to identify the belonging of some forms or objects (units, phenomena, events, actions, processes, etc.) to certain classes**. Form**it actually represents quantification**the main features in the form of a vector**, and the cluster is formed **from the totality of objects whose characteristics are similar and which are significantly different from the characteristics of objects that form other classes**. At the same time, cluster analysis is a form of unsupervised/uncontrolled recognition.

The hierarchical clustering methods are: simple aggregation method, complete aggregation method, average aggregation method and Ward's method.

Regarding the evaluation of the distances between clusters, these are the Ward method, the centroid method, the nearest neighbor method, the farthest neighbor method, and the average distance between pairs method.

| | Aruba | Afghanistan | Angola | Albania | Andorra | Arab_World | United_Arab_Emirates | Argentina | Armenia | American_Samoa | Antigua_and_Barbuda | Australia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 4.8005008 | | | | | | | | | | | |
| Angola | 6.9091087 | 4.6690327 | | | | | | | | | | |
| Albania | 7.7157586 | 5.2077098 | 3.9939473 | | | | | | | | | |
| Andorra | 8.2844844 | 7.1749797 | 7.9797804 | 5.4558183 | | | | | | | | |
| Arab_World | 7.9405135 | 6.0742345 | 5.1655779 | 3.0473133 | 4.0646152 | | | | | | | |
| United_Arab_Emirates | 7.1634474 | 5.9570131 | 6.4732682 | 4.8080316 | 2.9292413 | 2.7838560 | | | | | | |
| Argentina | 7.5905922 | 6.8619125 | 7.0169778 | 5.2205270 | 2.2381597 | 3.6263189 | 2.5373202 | | | | | |
| Armenia | 4.7364260 | 3.2237882 | 5.1789382 | 4.7968952 | 5.3613955 | 5.1924612 | 5.0985804 | 4.8905099 | | | | |
| American_Samoa | 7.5816870 | 6.1199564 | 6.0830140 | 4.2759135 | 3.0569821 | 1.9809419 | 0.9761414 | 2.5672220 | 5.2067567 | | | |
| Antigua_and_Barbuda | 2.3669414 | 4.8567633 | 6.6014691 | 7.6775124 | 8.4707205 | 7.9815962 | 7.2580574 | 7.5601006 | 4.8819349 | 7.6411936 | | |
| Australia | 4.2881309 | 4.6767800 | 5.3132464 | 5.4131093 | 6.3455669 | 6.2317481 | 6.2094336 | 5.5651058 | 2.6557591 | 6.2517152 | 4.6045201 | |
| Austria | 7.1065410 | 7.1584110 | 6.2526448 | 5.4904485 | 5.0981288 | 4.2796179 | 3.4298009 | 3.6691739 | 6.2668683 | 3.3951961 | 6.8198181 | 6.0704639 |
| Azerbaijan | 7.8379880 | 7.4311275 | 8.3625300 | 6.2962816 | 3.7688038 | 4.8067020 | 2.8685163 | 3.3190799 | 6.4736535 | 3.3427441 | 7.7656139 | 7.1905913 |
| Burundi | 8.8363514 | 6.2653137 | 5.7675794 | 3.5082295 | 4.8373617 | 2.1142475 | 3.3807388 | 4.9326505 | 6.2388285 | 2.8546930 | 8.8647328 | 7.5432627 |
| Belgium | 7.6334919 | 7.3147732 | 6.5243174 | 5.6495092 | 4.7428029 | 4.5618543 | 3.6808859 | 2.8875411 | 6.0310816 | 3.6456679 | 7.1671492 | 5.9563030 |
| Benin | 7.2161556 | 5.1702836 | 5.2921321 | 2.5944861 | 4.0109148 | 3.0424786 | 3.9908024 | 4.3505346 | 4.1381013 | 3.7385076 | 7.0982062 | 5.1972383 |
| Burkina_Faso | 3.7958891 | 3.8638899 | 5.7356407 | 5.4155949 | 5.3529183 | 5.2903528 | 4.5250941 | 4.4468237 | 2.2958066 | 4.8665481 | 3.5452544 | 3.2222060 |
| Bangladesh | 7.9106073 | 5.6357849 | 3.8325338 | 2.2571348 | 5.2979116 | 1.7447811 | 3.9477739 | 4.7674295 | 5.2315230 | 3.2807916 | 7.7403515 | 6.1318090 |
| Bulgaria | 8.1974416 | 6.1323416 | 5.0016670 | 3.1676252 | 4.3780548 | 0.9061902 | 2.8813176 | 3.9673280 | 5.5319185 | 2.0161624 | 8.1889054 | 6.5099075 |
| Bahrain | 8.7422697 | 6.6474251 | 5.7305165 | 3.9188998 | 4.6756780 | 2.1733565 | 3.2494373 | 4.7472711 | 6.3761908 | 2.5314494 | 8.8720209 | 7.2794479 |
| Bahamas | 2.9131482 | 3.0079682 | 6.1648584 | 6.8126200 | 7.7865031 | 7.5877605 | 6.8909058 | 7.4027416 | 3.8033578 | 7.2413499 | 3.1174983 | 3.8967115 |
| Bosnia_and_Herzegovina | 9.4098378 | 6.4136541 | 3.8047548 | 2.6953358 | 7.3752438 | 4.2405102 | 6.3377046 | 7.0094600 | 6.7077589 | 5.7023822 | 9.0940430 | 7.3678991 |
| Belarus | 4.6661985 | 3.7808202 | 6.1814182 | 7.7098583 | 9.7366720 | 9.0136631 | 8.8732693 | 9.1668398 | 4.8861487 | 9.0904055 | 4.1832726 | 5.0666693 |
| Belize | 6.7340249 | 6.7087580 | 6.2038902 | 5.2430585 | 4.8014214 | 4.1828875 | 3.3217619 | 3.1549920 | 5.5874476 | 3.3363203 | 6.3438627 | 5.5496841 |

| | Austria | Azerbaijan | Burundi | Belgium | Benin | Burkina_Faso | Bangladesh | Bulgaria | Bahrain | Bahamas | Bosnia_and_Herzegovina | Belarus | Belize |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | | | | | | | | | | | | | |
| Angola | | | | | | | | | | | | | |
| Albania | | | | | | | | | | | | | |
| Andorra | | | | | | | | | | | | | |
| Arab_World | | | | | | | | | | | | | |
| United_Arab_Emirates | | | | | | | | | | | | | |
| Argentina | | | | | | | | | | | | | |
| Armenia | | | | | | | | | | | | | |
| American_Samoa | | | | | | | | | | | | | |
| Antigua_and_Barbuda | | | | | | | | | | | | | |
| Australia | | | | | | | | | | | | | |
| Austria | | | | | | | | | | | | | |
| Azerbaijan | 3.7272541 | | | | | | | | | | | | |
| Burundi | 5.3253260 | 5.2848783 | | | | | | | | | | | |
| Belgium | 1.9022947 | 3.7344014 | 5.6777809 | | | | | | | | | | |
| Benin | 5.3857616 | 5.5371566 | 3.5619074 | 5.5232852 | | | | | | | | | |
| Burkina_Faso | 5.1136564 | 5.2252144 | 6.3824971 | 4.9898274 | 4.6473786 | | | | | | | | |
| Bangladesh | 4.7730310 | 5.9783992 | 2.4349991 | 5.0917075 | 2.9346752 | 5.5197063 | | | | | | | |
| Bulgaria | 4.2258235 | 5.0134940 | 1.9329903 | 4.5989059 | 3.3829193 | 5.6584351 | 1.7031306 | | | | | | |
| Bahrain | 4.6974391 | 5.4136817 | 2.1029391 | 5.3716707 | 3.9638631 | 6.5601075 | 2.7355315 | 1.5413048 | | | | | |
| Bahamas | 7.2808070 | 7.7490339 | 8.1441451 | 7.5625013 | 6.3431403 | 3.6767720 | 7.3539146 | 7.6824120 | 8.1267896 | | | | |
| Bosnia_and_Herzegovina | 6.7468806 | 8.0710836 | 3.9900340 | 6.9341856 | 4.3639999 | 7.3377293 | 2.7265373 | 4.0399801 | 4.5455424 | 8.4091184 | | | |
| Belarus | 9.1811401 | 9.9272826 | 9.5089232 | 9.1815538 | 7.6049358 | 5.2613141 | 8.3373589 | 9.1220116 | 9.8074081 | 3.0689532 | 8.7736356 | | |
| Belize | 1.3652729 | 3.2178040 | 5.3794610 | 1.5704982 | 5.1304804 | 4.2850572 | 4.7607671 | 4.2868450 | 5.0751408 | 6.8530397 | 6.7753436 | 8.6440575 | |

| | Bermuda | Bolivia | Brazil | Barbados | Brunei_Darussalam | Bhutan | Botswana | Central_African_Republic | Canada | Central_Europe_and_the_Baltics | Switzerland |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | | | | | | | | | | | |
| Angola | | | | | | | | | | | |
| Albania | | | | | | | | | | | |
| Andorra | | | | | | | | | | | |
| Arab_World | | | | | | | | | | | |

This matrix represents the matrix of Euclidean distances calculated between the observations in the set of standardized data under analysis. For example, the Euclidean distance between Afghanistan and Aruba is 4.800501.

```
> d_std[1] #4.800501
[1] 4.800501
>
```

Fig. 1.1.2 Ranking of standardized data

```
> clust_std

Call:
hclust(d = d_std, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 42

> summary(clust_std)###########
             Length Class  Mode
merge        82     -none- numeric
height       41     -none- numeric
order        42     -none- numeric
labels       42     -none- character
method        1     -none- character
call          3     -none- call
dist.method   1     -none- character
>
```

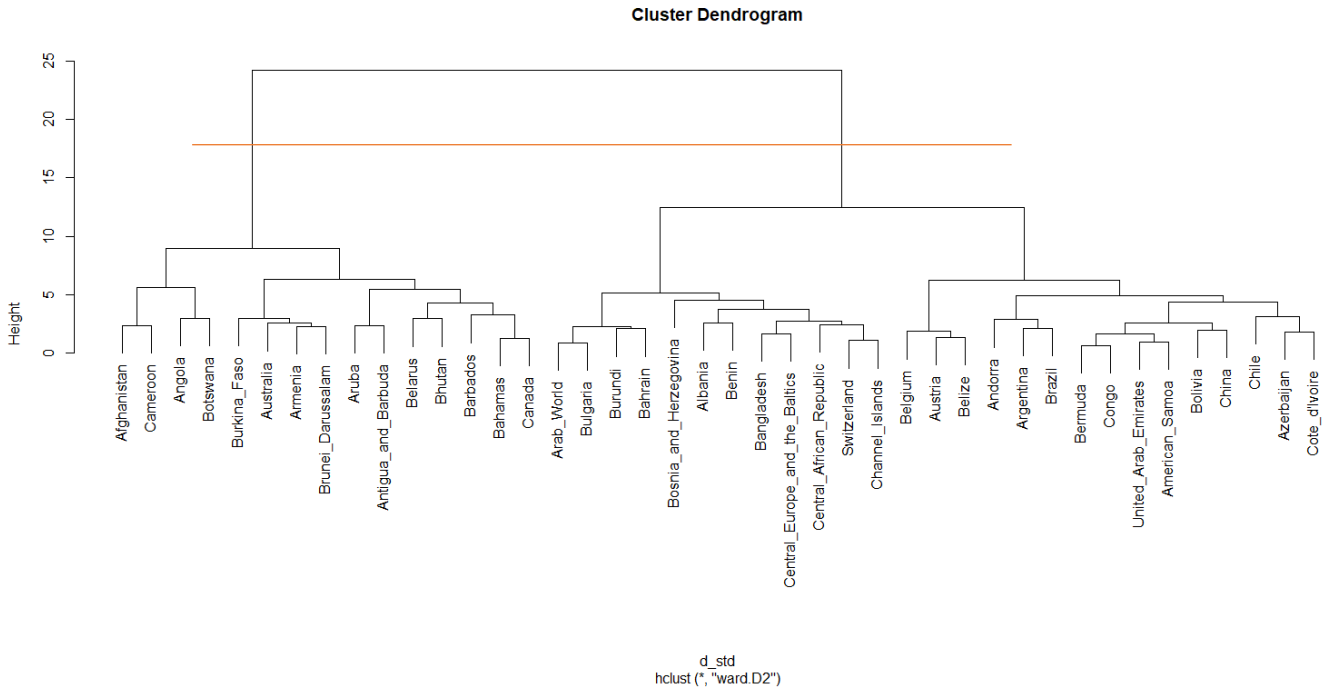
The method used in ranking the data is Ward, for a data set with 42 observations in which the Euclidean distance is determined.

Fig. 1.1.3 Classification stages and aggregation distances for standardized data

```
> cbind(clust_std$merge,clust_std$height)
      [,1] [,2]       [,3]
 [1,]  -26  -42  0.6444639
 [2,]   -6  -20  0.9061902
 [3,]   -7  -10  0.9761414
 [4,]  -36  -37  1.0868580
 [5,]  -22  -34  1.3001733
 [6,]  -13  -25  1.3652729
 [7,]  -19  -35  1.6596759
 [8,]    1    3  1.6733371
 [9,]  -14  -40  1.8241565
[10,]  -16    6  1.8535020
[11,]  -27  -39  1.9683746
[12,]  -15  -21  2.1029391
[13,]   -8  -28  2.1417701
[14,]    2   12  2.2429957
[15,]   -9  -30  2.2818564
[16,]   -2  -41  2.3499678
[17,]   -1  -11  2.3669414
[18,]  -33    4  2.4456331
[19,]  -12   15  2.5599912
[20,]    8   11  2.5600322
[21,]   -4  -17  2.5944861
[22,]    7   18  2.7387788
[23,]   -5   13  2.8746389
[24,]  -24  -31  2.9442577
[25,]   -3  -32  2.9689113
[26,]  -18   19  3.0041954
[27,]  -38    9  3.1610037
[28,]  -29    5  3.2706890
[29,]   21   22  3.7640081
[30,]   24   28  4.2978409
[31,]   20   27  4.3605711
[32,]  -23   29  4.5461958
[33,]   23   31  4.8827991
[34,]   14   32  5.1235329
[35,]   17   30  5.4959859
[36,]   16   25  5.6176841
[37,]   10   33  6.2541255
[38,]   26   35  6.3555135
[39,]   36   38  8.9835114
[40,]   34   37 12.4666413
[41,]   39   40 24.2471297
```

The figure above shows how observations, classes or observations with classes are joined in the ranking by classes and the aggregation distances between them. In the first two columns, you can find the observations or the classes that come together. The values with + represent the classes, and the values with - represents the observations. The third column consists of the aggregation distances between the two mentioned in the first two columns. Thus, it can be seen from the figure above that in step 1, observation 26 and observation 42 join at the aggregation distance of 0.6444639. At step 8, classes 1 and 3 join with the aggregation distance of 1.6733371. At step 18, observation 33 joined class 4 at a distance of 2.4456331.

Fig. 1.1.4. Dendrogram for standardized data




The dendrogram, also called the classification tree, helps to establish the number of classes determined after the cluster analysis. To determine this number, the dendrogram had to be analyzed from top to bottom and where the aggregation distance is greater, a horizontal line is drawn. The number of intersection points represents the number of clusters.

So, in the figure above, you can see that the drawn horizontal line intersects the dendrogram in two points, so the number of clusters determined in this analysis is 2.

Fig. 1.1.5 Determination of the number of clusters

```
> clust_std$height[41]-clust_std$height[40]
[1] 11.78049
> clust_std$height[40]-clust_std$height[39]
[1] 3.48313
> clust_std$height[39]-clust_std$height[38]
[1] 2.627998
> clust_std$height[38]-clust_std$height[37]
[1] 0.101388
> clust_std$height[37]-clust_std$height[36]
[1] 0.6364414
> clust_std$height[36]-clust_std$height[35]
[1] 0.1216982
```

There are 42 observations in the analysis. As can be seen in the figure presented above, the first calculated difference is the largest (11.78049), which confirms the determination of two clusters, an aspect also mentioned in the dendrogram analysis.

Fig. 1.1.6. Proposals for the number of clusters

```
> res<-NbClust(date_std, distance = "euclidean", min.nc=2, max.nc=5,
+              method = "ward.D2", index = "all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 12 proposed 2 as the best number of clusters
* 8 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
>
```

The figure above represents the output of a function with the help of which several numbers of clusters are proposed into which the data used in the analysis could be divided. Thus, 12 proposals are for creating 2 clusters, 8 proposals for creating 3 clusters, 1 proposal for 4 clusters and 2 proposals for 5 clusters. Therefore, according to the majority rule, the best number of clusters is 2.

Fig. 1.1.7 Proposals for the number of clusters and the value of the indicators corresponding to each number of clusters

```
*****************************************************************
> res$All.index
      KL      CH Hartigan    CCC    Scott Marriot    TrCovW   TraceW    Friedman  Rubin Cindex     DB Silhouette   Duda Pseudot2  Beale Ratkowsky     Ball Ptbiserial   Frey
2 3.7337 36.6264  12.7742 1.1065 112.5324       0 1533.0086 321.0384  8781816570 1.9157 0.4156 1.0162     0.4217 0.5899  17.3808 6.9042    0.4340 160.5192     0.7484 1.9949
3 1.8761 29.7850   7.7531 0.9291 185.1595       0  717.1729 243.3298  8912176770 2.5274 0.4301 1.0874     0.3186 0.6933   5.7516 4.2430    0.4277  81.1099     0.6130 0.2240
4 2.1255 25.7119   4.1988 0.9056 233.1508       0  473.9627 202.9780  9764074234 3.0299 0.5232 1.2534     0.2861 0.7041   3.7820 3.9060    0.3942  50.7445     0.6361 0.6923
5 0.9167 21.8732   4.4332 0.2889 290.3260       0  401.3733 182.7818 12136112309 3.3647 0.5007 1.2655     0.2636 0.6945   5.7183 4.2184    0.3647  36.5564     0.6176 1.0536
  McClain   Dunn Hubert SDindex Dindex   SDbw
2  0.4967 0.4313 0.0035  0.7926 2.6469 0.6686
3  1.2004 0.2843 0.0034  0.8447 2.2779 0.4929
4  1.4129 0.3765 0.0037  0.8489 2.1160 0.5370
5  1.6359 0.3788 0.0040  0.9429 2.0020 0.4648
> res$Best.nc
                     KL      CH Hartigan    CCC  Scott Marriot   TrCovW  TraceW   Friedman   Rubin Cindex     DB Silhouette   Duda PseudoT2 Beale Ratkowsky    Ball PtBiserial
Number_clusters 2.0000  2.0000   3.0000 2.0000  3.000       3   3.0000  3.0000          5  4.0000 2.0000 2.0000     2.0000 3.0000   3.0000    NA     2.000  3.0000     2.0000
Value_Index     3.7337 36.6264   5.0211 1.1065 72.627       0 815.8358 37.3568 2372038075 -0.1677 0.4156 1.0162     0.4217 0.6933   5.7516    NA     0.434 79.4092     0.7484
                  Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters 2.0000  2.0000 2.0000      0  2.0000      0 5.0000
Value_Index     1.9949  0.4967 0.4313      0  0.7926      0 0.4648
>
```

The figure above illustrates one value of an indicator, respectively one proposal for the number of clusters in which the data should be divided. For example, KL proposes a number of clusters of 2, this aspect being augmented by the indicator values, as its indicator for the number of 2 clusters is the highest (3.7337), compared to the values for another number of clusters.

## Silhouette chart

To study the separation distances between the obtained classes, it is used the Silhouette chart. The Silhouette graph illustrates the degree of proximity of each object in a cluster to objects in neighboring classes and provides a visual representation to determine the number of classes. Silhouette coefficients are located between [-1,1], and coefficients close to 1 indicate that the object is at a considerable distance from neighboring classes. A value close to 0 indicates that the object is very close or even on the surface of separation between two classes. As for the negative values, they suggest that there is a possibility that the respective objects have been wrongly classified.

Thus, the figure below suggests that for the data used in this analysis, the number of 2 clusters is optimal compared to a larger number of clusters, since most of the coefficients are close to 1 and very few values are close to 0. In the analyzed figure, no negative values are observed, which indicates that there are no objects wrongly included in the two clusters.

The average for cluster 2 is 0.46, which is slightly higher than the average of the cluster 1, which is 0.36. It should be mentioned that the number of centralized objects also differs: cluster 1 contains 15 objects, while the second one contains 27 objects. The distribution of observations in the two clusters is balanced, there being no large discrepancies between the number of elements between the two clusters. The value of the average silhouette width indicator is 0.42.

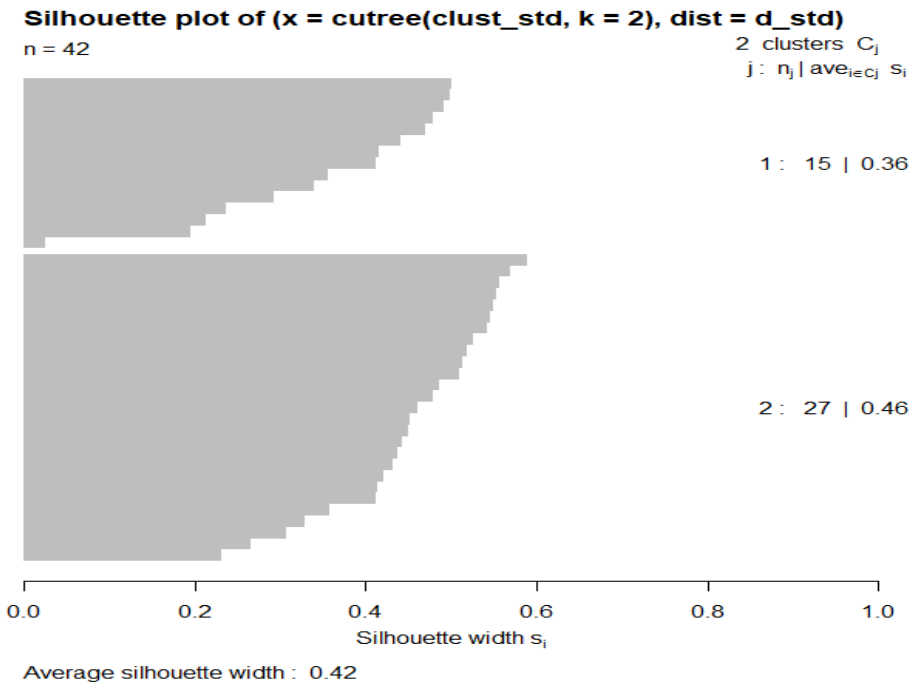



Fig. 1.1.8 Silhouette for standardized data

Fig. 1.1.9 Value of Silhouette coefficients

```
> si4_std
      cluster neighbor  sil_width
 [1,]       1        2 0.41106630
 [2,]       1        2 0.35651448
 [3,]       1        2 0.02426994
 [4,]       2        1 0.30738352
 [5,]       2        1 0.43245484
 [6,]       2        1 0.54889475
 [7,]       2        1 0.55575690
 [8,]       2        1 0.46087478
 [9,]       1        2 0.29192963
[10,]       2        1 0.58999462
[11,]       1        2 0.41616925
[12,]       1        2 0.33937474
[13,]       2        1 0.43670957
[14,]       2        1 0.44227381
[15,]       2        1 0.50919574
[16,]       2        1 0.41281762
[17,]       2        1 0.32868142
[18,]       1        2 0.21297436
[19,]       2        1 0.45241581
[20,]       2        1 0.55298625
[21,]       2        1 0.51891379
[22,]       1        2 0.50084567
[23,]       2        1 0.26525291
[24,]       1        2 0.49925595
[25,]       2        1 0.41302831
[26,]       2        1 0.52604799
[27,]       2        1 0.56944780
[28,]       2        1 0.45013390
[29,]       1        2 0.44022093
[30,]       1        2 0.46894138
[31,]       1        2 0.47972337
[32,]       1        2 0.23603176
[33,]       2        1 0.23048694
[34,]       1        2 0.49115633
[35,]       2        1 0.42014348
[36,]       2        1 0.51313899
[37,]       2        1 0.47895569
[38,]       2        1 0.35740993
[39,]       2        1 0.54566355
[40,]       2        1 0.48653868
[41,]       1        2 0.19454591
[42,]       2        1 0.54201872
attr(,"Ordered")
[1] FALSE
attr(,"call")
silhouette.default(x = cutree(clust_std, k = 2), dist = d_std)
attr(,"class")
[1] "silhouette"
>
```

This figure represents the actual values of the Silhouette coefficients represented graphically in Fig. 1.1.8. Thus, it can be seen that the country with indicator 3, which is found in cluster 1, has a value very close to 0, more precisely 0.024, which indicates that this country is on the border between the two clusters.

b) Main components

Fig. 1.1.10 Matrix of Euclidean distances for principal components

```
> d_z #matricea distantelor pentru componentele principale
                          Aruba Afghanistan    Angola   Albania   Andorra Arab_World United_Arab_Emirates Argentina   Armenia American_Samoa Antigua_and_Barbuda
Afghanistan           4.0898833
Angola                6.3062183   4.1069156
Albania               7.1490380   4.2889993 3.1119770
Andorra               7.8191820   6.5084907 7.8641704 5.3091755
Arab_World            7.7447878   5.6544791 4.9524741 2.2258364 3.7551379
United_Arab_Emirates  6.9595450   5.8739900 6.4052930 4.1352784 1.9780981  2.4243673
Argentina             6.7614251   6.0664047 6.9206324 4.8171109 1.8378575  3.2098447           0.7914557
Armenia               3.7922871   1.5816679 5.0101803 4.4308818 5.2281904  5.1687040           4.7889442 4.8358524
American_Samoa        7.3529284   5.9690175 6.0591842 3.6299726 2.4378232  1.7011953           0.7952639 1.5731076 5.0549327
Antigua_and_Barbuda   0.6934552   4.3839526 6.1353674 7.2192698 8.1710975  7.8494646           7.1653058 6.9974563 4.2247926      7.5176108
Australia             2.2073977   2.1789001 4.7094854 5.0555284 6.2355419  5.8047401           5.3575226 5.3362504 1.8060579      5.6387784           2.5041699
Austria               6.6298712   6.8789637 6.1970695 5.2375743 4.9976630  3.9768942           3.1582895 3.2051818 6.0941434      3.2020190           6.5424910
Azerbaijan            7.3676797   7.2992057 7.9990622 6.0704323 2.6763670  4.2941549           2.0234182 1.4297384 6.0199142      2.6204008           7.5614858
Burundi               8.7207100   6.2213705 5.4353664 2.4516673 4.1480315  1.1654151           3.2723044 4.0535182 5.8455218      2.5486583           8.8508733
Belgium               6.4559395   6.6619003 6.3703708 5.1920763 4.4013246  3.8289751           2.6367504 2.5871168 5.7511557      2.8109233           6.4369926
Benin                 6.6342417   3.8625409 4.4548068 1.8289561 3.7342228  2.0950745           3.0051792 3.5800229 3.4186761      2.7415649           6.8619514
Burkina_Faso          2.7813975   3.1361870 5.4633278 5.0945661 5.0716562  5.2660549           4.2828019 4.1150635 1.9892045      4.7122021           3.1438089
Bangladesh            7.6495697   5.2024962 3.6061560 1.0805776 5.1821398  1.5826793           3.8068988 4.5474071 5.1735216      3.1693089           7.6756290
Bulgaria              7.9771218   5.8576375 4.9164974 2.2148895 4.0546230  0.3436675           2.7170419 3.5034442 5.4386368      1.9689158           8.0607073
Bahrain               8.5776302   6.3792986 5.4552428 2.6506077 4.0546291  0.8334935           2.9301812 3.7176884 5.9350649      2.1533727           8.6773217
Bahamas               1.8395466   2.5911032 5.7629388 6.3951001 7.4153736  7.3015858           6.7906385 6.7017999 2.5508450      7.1225199           2.3279418
Bosnia_and_Herzegovina 8.9610980  6.0024544 3.3477066 2.3036315 7.3252775  3.8027603           6.1178021 6.8606804 6.5202395      5.4686233           8.9244542
Belarus               3.6494634   3.4252808 5.9884844 7.4250433 9.5572495  8.8704916           8.8739633 8.9161104 4.3529424      9.0671542           3.8042011
Belize                6.0332588   6.3638903 6.0790101 5.1073992 4.6311003  3.9416116           2.8484925 2.8001533 5.4956580      3.0327675           5.9911309
                      Australia   Austria Azerbaijan   Burundi   Belgium     Benin Burkina_Faso Bangladesh  Bulgaria   Bahrain   Bahamas Bosnia_and_Herzegovina
```

This matrix represents the matrix of Euclidean distances calculated for the components main purposes. For example, the Euclidean distance between Afghanistan and Aruba is 4.089883 3.

```
> d_z[1]
[1] 4.089883
>
```

Fig. 1.1.11 Hierarchy of the main components

```
> clust_z

Call:
hclust(d = d_z, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 42

> summary(clust_z)
            Length Class  Mode
merge       82     -none- numeric
height      41     -none- numeric
order       42     -none- numeric
labels      42     -none- character
method       1     -none- character
call         3     -none- call
dist.method  1     -none- character
>
```

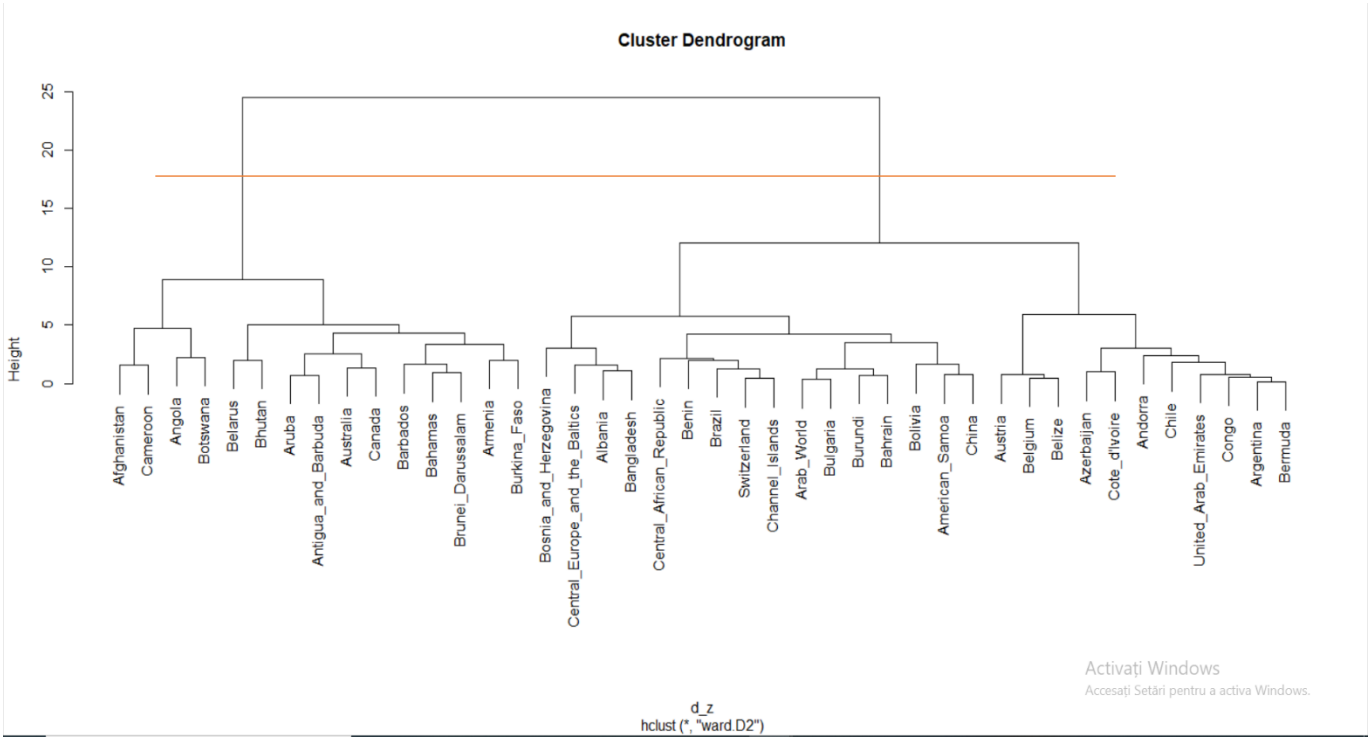
The method used in ranking the data is Ward, for a data set with 42 observations in which the Euclidean distance is calculated.

Fig. 1.1.12 Classification stages and aggregation distances for main components

```
> cbind(clust_z$merge,clust_z$height) #etapele clasificarii si distanta de agregare
      [,1] [,2]       [,3]
 [1,]   -8  -26  0.1362017
 [2,]   -6  -20  0.3436675
 [3,]  -36  -37  0.4378179
 [4,]  -16  -25  0.4833053
 [5,]  -42    1  0.5246625
 [6,]  -15  -21  0.6817925
 [7,]   -1  -11  0.6934552
 [8,]  -10  -39  0.7385918
 [9,]  -13    4  0.7524319
[10,]   -7    5  0.7945212
[11,]  -22  -30  0.9114741
[12,]  -14  -40  0.9934119
[13,]   -4  -19  1.0805776
[14,]    2    6  1.2228342
[15,]  -28    3  1.2925486
[16,]  -12  -34  1.3396209
[17,]   -2  -41  1.5435330
[18,]  -35   13  1.6044058
[19,]  -29   11  1.6230324
[20,]  -27    8  1.6289199
[21,]  -38   10  1.7984925
[22,]  -17   15  1.9691130
[23,]  -24  -31  1.9811438
[24,]   -9  -18  1.9892045
[25,]  -33   22  2.1797295
[26,]   -3  -32  2.2332132
[27,]   -5   21  2.4033055
[28,]    7   16  2.5465527
[29,]   12   27  3.0079956
[30,]  -23   18  3.0293584
[31,]   19   24  3.3434620
[32,]   14   20  3.5082384
[33,]   25   32  4.2630442
[34,]   28   31  4.3033759
[35,]   17   26  4.6826916
[36,]   23   34  5.0451813
[37,]   30   33  5.7671426
[38,]    9   29  5.9483016
[39,]   35   36  8.8935701
[40,]   37   38 12.0468803
[41,]   39   40 24.5210570
>
```

The figure above, as in the case of standardized data, shows how they came together the observations between them, the observations with the classes or the classes between them and the aggregation distances between them. It is visible that at step 1, observations 8 and 26 joined with the aggregation distance of 0.1362017. At step 5, observation 42 joined class 1 at the aggregation distance of 0.5246625. At step 14, classes 2 and 6 joined at the aggregation distance of 1.2228342.

Fig. 1.1.13. Dendrogram for principal components




The dendrogram above, as well as the one analyzed in the case of standardized data (Fig. 1.1.4), suggests the creation of two clusters, since the drawn horizontal line intersects the dendogram in two points.

Fig. 1.1.14. Determining the number of clusters for the main components

```
> clust_z$height[41]-clust_z$height[40]
[1] 12.47418
> clust_z$height[40]-clust_z$height[39]
[1] 3.15331
> clust_z$height[39]-clust_z$height[38]
[1] 2.945269
> clust_z$height[38]-clust_z$height[37]
[1] 0.1811589
> clust_z$height[37]-clust_z$height[36]
[1] 0.7219614
> clust_z$height[36]-clust_z$height[35]
[1] 0.3624897
>
```

In the analysis there are 42 observations and 3 main components. As can be seen in the figure presented above, the first calculated difference is the largest (12.47418), which confirms the determination of two clusters, an aspect also mentioned in the dendrogram analysis.

Fig. 1.1.15 Proposals for the number of clusters for the main components

```
> res<-NbClust(scoruri, distance = "euclidean", min.nc=2, max.nc=5,
+              method = "ward.D2", index = "all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 10 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 7 proposed 4 as the best number of clusters
* 3 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
>
```

The figure above represents the output of a function with the help of which several numbers of clusters are proposed into which the data used in the analysis could be divided. Thus, this time, 10 proposals are to create 2 clusters, 4 proposals to create 3 clusters, 7 proposals for 4 clusters and 3 proposals for 5 clusters. Therefore, according to the majority rule, the best number of clusters in this situation is 2.

Fig. 1.1.16. Proposals for the number of clusters and the values of the indicators corresponding to each number

```
*****************************************************************
> res$All.index
       KL      CH Hartigan    CCC    Scott Marriot   TrCovW   TraceW Friedman  Rubin Cindex     DB Silhouette   Duda Pseudot2  Beale Ratkowsky     Ball Ptbiserial   Frey
2  4.8115 50.3840  17.4730 1.2992  86.7923 1352949 8644.724 238.6799   6.8769 2.2596 0.3793 0.8608     0.4966 0.5059  24.4190 1.5989    0.2310 119.3399     0.7528 1.6296
3  1.2534 43.8096  12.1860 1.0783 120.1619 1375341 3478.954 166.1162   8.3297 3.2466 0.3789 0.9056     0.4088 0.5693   9.8343 1.1959    0.3395  55.3721     0.6427 0.2759
4 46.4001 41.3073   6.1745 1.3467 153.6193 1102368 1625.456 126.5684  12.1578 4.2611 0.4106 1.0066     0.3720 0.3699  15.3316 2.6101    0.3305  31.6421     0.6597 1.2749
5  0.0417 36.5696   6.6702 0.6691 170.1759 1161303 1238.037 108.8773  12.6945 4.9535 0.3967 0.8483     0.3818 0.6402   7.8697 0.8932    0.3553  21.7755     0.6331 1.0640
  McClain   Dunn Hubert SDindex Dindex   SDbw
2  0.4335 0.3727 0.0032  0.6014 2.2008 0.5640
3  0.9775 0.1176 0.0033  0.7150 1.7941 0.4496
4  1.1142 0.1454 0.0036  0.6472 1.6015 0.2431
5  1.2686 0.1454 0.0036  0.6740 1.4257 0.1672
> res$Best.nc
                    KL     CH Hartigan    CCC   Scott  Marriot  TrCovW  TraceW Friedman   Rubin Cindex     DB Silhouette   Duda PseudoT2  Beale Ratkowsky    Ball PtBiserial
Number_clusters  4.0000  2.000   4.0000 4.0000  4.0000      4.0    3.00  3.0000    4.000  4.0000 3.0000 5.0000     2.0000 2.0000    2.000 2.0000    5.0000  3.0000     2.0000
Value_Index     46.4001 50.384   6.0115 1.3467 33.4574 331907.8 5165.77 33.0159    3.828 -0.3221 0.3789 0.8483     0.4966 0.5059   24.419 1.5989    0.3553 63.9679     0.7528
                  Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters 2.0000  2.0000 2.0000      0  2.0000      0 5.0000
Value_Index     1.6296  0.4335 0.3727      0  0.6014      0 0.1672
>
```

The figure above illustrates one value of an indicator, respectively one proposal for the number of clusters in which the data should be divided. For example, KL proposes a number of 4 clusters this time, this aspect being augmented by the indicator values, as its indicator for the number of 4 clusters is the highest (46.4001), compared to the values for another number of clusters.

## Silhouette chart

The figure below suggests that for the data used in this analysis, the number of 2 clusters is optimal compared to a larger number of clusters, since most of the coefficients are close to 1 and very few values are close to 0 or even 0. In the analyzed figure, no negative values are observed, which indicates that there are no objects wrongly included in the two clusters. The plot is almost identical to the Silhouette plot from the standardized data analysis.

The average for cluster 2 is 0.52, which is very slightly higher than the average cluster 1, which is 0.46. It should be mentioned that the number of centralized objects also differs: cluster 1 contains 15 objects, while the second one contains 27 objects. The distribution of observations in the two clusters is balanced, there being no large discrepancies between the number of elements between the two clusters. The value of the average silhouette width indicator is 0.5.

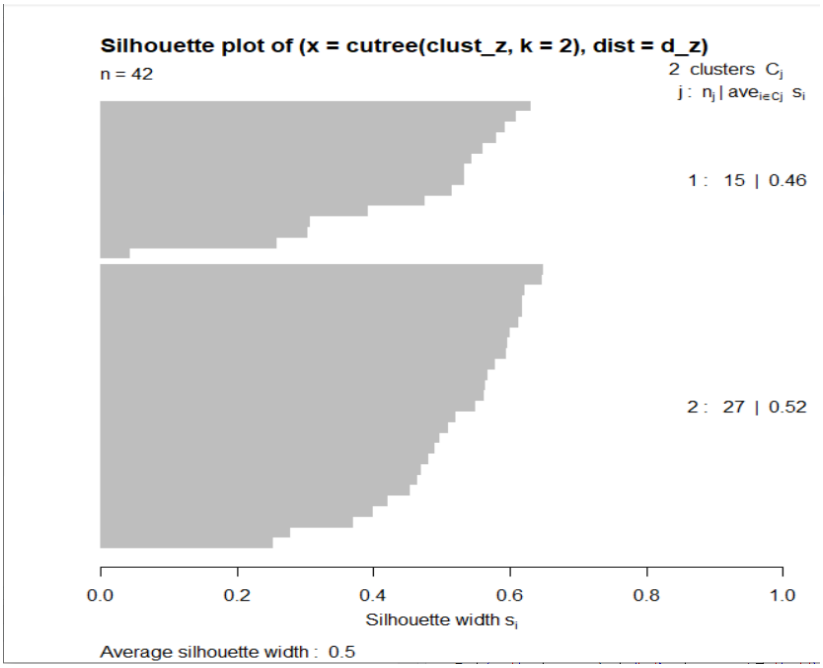



Fig. 1.1.17. Principal Components Silhouette Chart

Fig. 1.1.18. The value of the Silhouette coefficients

```
> si4_z
      cluster neighbor  sil_width
 [1,]       1        2 0.54368016
 [2,]       1        2 0.47570627
 [3,]       1        2 0.04263965
 [4,]       2        1 0.37015831
 [5,]       2        1 0.48016856
 [6,]       2        1 0.62163784
 [7,]       2        1 0.61743304
 [8,]       2        1 0.56825555
 [9,]       1        2 0.39171991
[10,]       2        1 0.64637260
[11,]       1        2 0.51416240
[12,]       1        2 0.53362156
[13,]       2        1 0.46916577
[14,]       2        1 0.51029946
[15,]       2        1 0.56338811
[16,]       2        1 0.49692266
[17,]       2        1 0.40009935
[18,]       1        2 0.30644634
[19,]       2        1 0.49044589
[20,]       2        1 0.61179968
[21,]       2        1 0.59979851
[22,]       1        2 0.63035284
[23,]       2        1 0.27826336
[24,]       1        2 0.56109309
[25,]       2        1 0.45370704
[26,]       2        1 0.57910184
[27,]       2        1 0.61753476
[28,]       2        1 0.56189149
[29,]       1        2 0.53240877
[30,]       1        2 0.59223139
[31,]       1        2 0.57970401
[32,]       1        2 0.30384027
[33,]       2        1 0.25254603
[34,]       1        2 0.60867248
[35,]       2        1 0.46471487
[36,]       2        1 0.59466516
[37,]       2        1 0.54890232
[38,]       2        1 0.42027022
[39,]       2        1 0.64828526
[40,]       2        1 0.52040844
[41,]       1        2 0.25896846
[42,]       2        1 0.59661131
```

This figure represents the actual values of the Silhouette coefficients represented graphically in Fig. 1.1.17. Thus, it can be seen that the country with indicator 3 has a value very close to 0, more precisely 0.0426, which indicates that this country is in cluster 1, but very close to the border with the second cluster.

**Conclusion**

Performing cluster analysis both for standardized data and for components main, it is concluded that the number of clusters identified in this analysis is 2 in both situations.

**1.2. Average aggregation method**

Fig. 1.2.1 Ranking of standardized data

```
> clust_3

Call:
hclust(d = d_std, method = "average")

Cluster method   : average
Distance         : euclidean
Number of objects: 42

> summary(clust_3)
            Length Class  Mode
merge       82     -none- numeric
height      41     -none- numeric
order       42     -none- numeric
labels      42     -none- character
method       1     -none- character
call         3     -none- call
dist.method  1     -none- character
>
```

Fig. 1.2.2. Hierarchy for principal components

```
> clust_2 = hclust(d_z, method = "average")
> clust_2

Call:
hclust(d = d_z, method = "average")

Cluster method   : average
Distance         : euclidean
Number of objects: 42

> summary(clust_2)
            Length Class  Mode
merge       82     -none- numeric
height      41     -none- numeric
order       42     -none- numeric
labels      42     -none- character
method       1     -none- character
call         3     -none- call
dist.method  1     -none- character
>
```

This time, the method used is that of average aggregation (Avergare), just as it is visible in the two figures above, for the same number of observations, 42 and a number of 3 main components and calculating the Euclidean distance.

Fig. 1.2.3. Classification stages and aggregation distances for standardized data

```
> cbind(clust_3$merge,clust_3$height)
      [,1] [,2]      [,3]
 [1,]  -26  -42 0.6444639
 [2,]   -6  -20 0.9061902
 [3,]   -7    1 0.9660028
 [4,]  -36  -37 1.0868580
 [5,]  -22  -34 1.3001733
 [6,]  -13  -25 1.3652729
 [7,]  -10    3 1.3863925
 [8,]  -19  -35 1.6596759
 [9,]  -16    6 1.7363964
[10,]  -14  -40 1.8241565
[11,]  -21    2 1.8573307
[12,]  -27    7 1.8857765
[13,]  -39   12 2.0094458
[14,]  -15   11 2.0500590
[15,]    4    8 2.0695316
[16,]   -8  -28 2.1417701
[17,]   -9  -30 2.2818564
[18,]   -2  -41 2.3499678
[19,]  -33   15 2.3503233
[20,]   -1  -11 2.3669414
[21,]  -12   17 2.4876974
[22,]   -4  -17 2.5944861
[23,]   14   19 2.6517160
[24,]   -5   16 2.6746789
[25,]   10   13 2.7795653
[26,]  -18   21 2.7983447
[27,]  -38   25 2.8221558
[28,]  -29    5 2.8889717
[29,]  -24  -31 2.9442577
[30,]   -3  -32 2.9689113
[31,]   24   27 3.1070237
[32,]   22   23 3.1295204
[33,]   20   28 3.2991673
[34,]    9   31 3.4735715
[35,]   26   33 3.7940031
[36,]  -23   32 3.9087034
[37,]   29   35 4.1092175
[38,]   34   36 4.2935336
[39,]   18   30 4.3654559
[40,]   37   39 4.9668082
[41,]   38   40 6.6720751
>
```

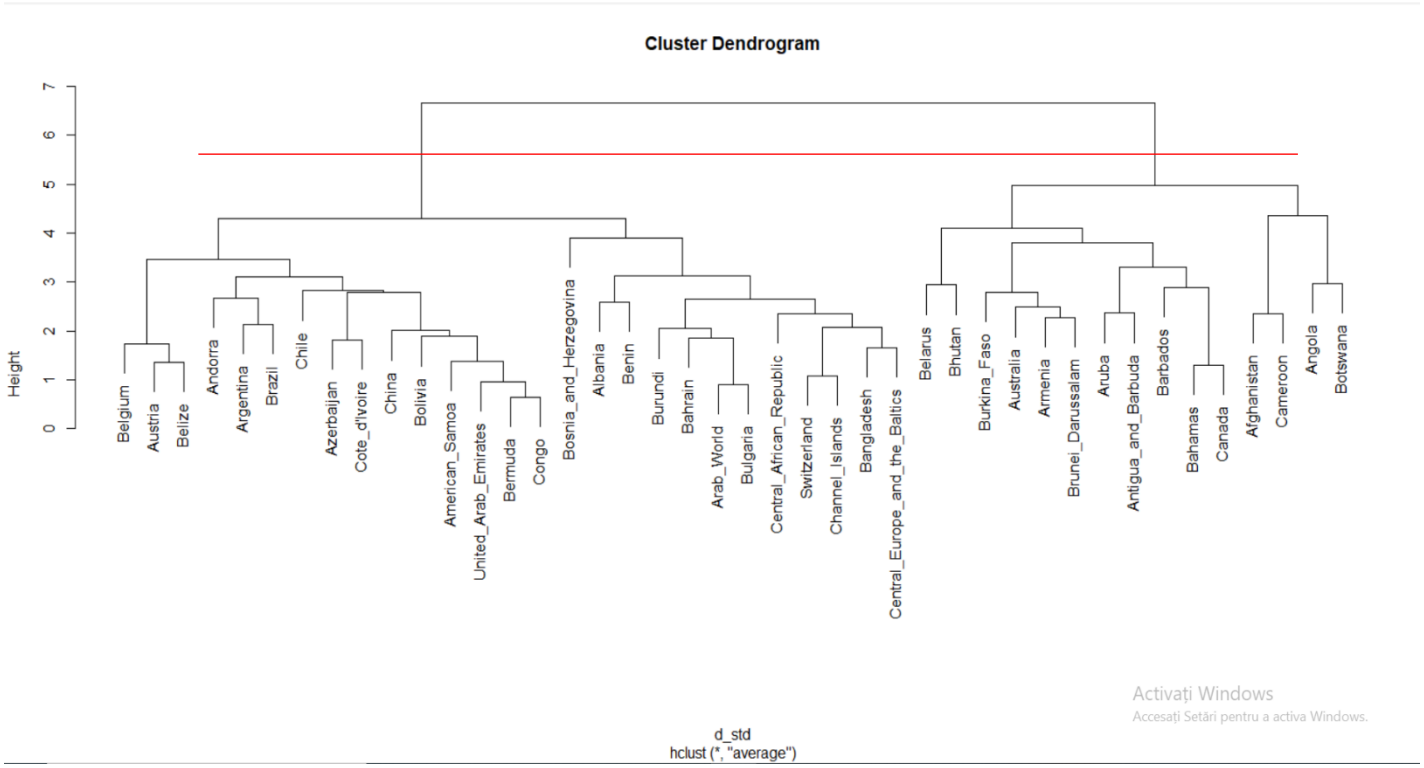
According to the figure above, it can be noted that observations 26 and 42 join at step 1 with an aggregation distance of 0.6444639. At step 3, observation 7 joins with class 1 at an aggregation distance of 0.9660028, and at step 15, classes 4 and 8 join with a distance of 2.0695316.

Fig. 1.2.4 Classification stages and aggregation distances for main components

```
> cbind(clust_2$merge,clust_2$height)
      [,1] [,2]      [,3]
 [1,]   -8  -26 0.1362017
 [2,]   -6  -20 0.3436675
 [3,]  -36  -37 0.4378179
 [4,]  -42    1 0.4589055
 [5,]  -16  -25 0.4833053
 [6,]   -7    4 0.6787903
 [7,]  -15  -21 0.6817925
 [8,]   -1  -11 0.6934552
 [9,]  -13    5 0.6946967
[10,]  -10  -39 0.7385918
[11,]  -22  -30 0.9114741
[12,]    2    7 0.9237184
[13,]  -14  -40 0.9934119
[14,]   -4  -19 1.0805776
[15,]  -28    3 1.1373563
[16,]  -12  -34 1.3396209
[17,]  -29   11 1.4239261
[18,]  -38    6 1.4391274
[19,]  -27   10 1.4551701
[20,]  -35   14 1.4672283
[21,]   -2  -41 1.5435330
[22,]   15   19 1.6784968
[23,]   13   18 1.7451986
[24,]  -31   17 1.7993356
[25,]    8   16 1.8910478
[26,]   12   22 1.9408783
[27,]   -9  -18 1.9892045
[28,]  -33   20 2.1184640
[29,]   -3  -32 2.2332132
[30,]   -5   23 2.2540555
[31,]  -17   28 2.2673650
[32,]   26   31 2.4362730
[33,]   24   25 2.5666713
[34,]   27   33 2.8078711
[35,]    9   30 2.9789360
[36,]  -24   34 3.4529498
[37,]   21   29 3.5593289
[38,]   32   35 3.6505152
[39,]   36   37 4.3622324
[40,]  -23   38 5.2345277
[41,]   39   40 6.3798455
>
```

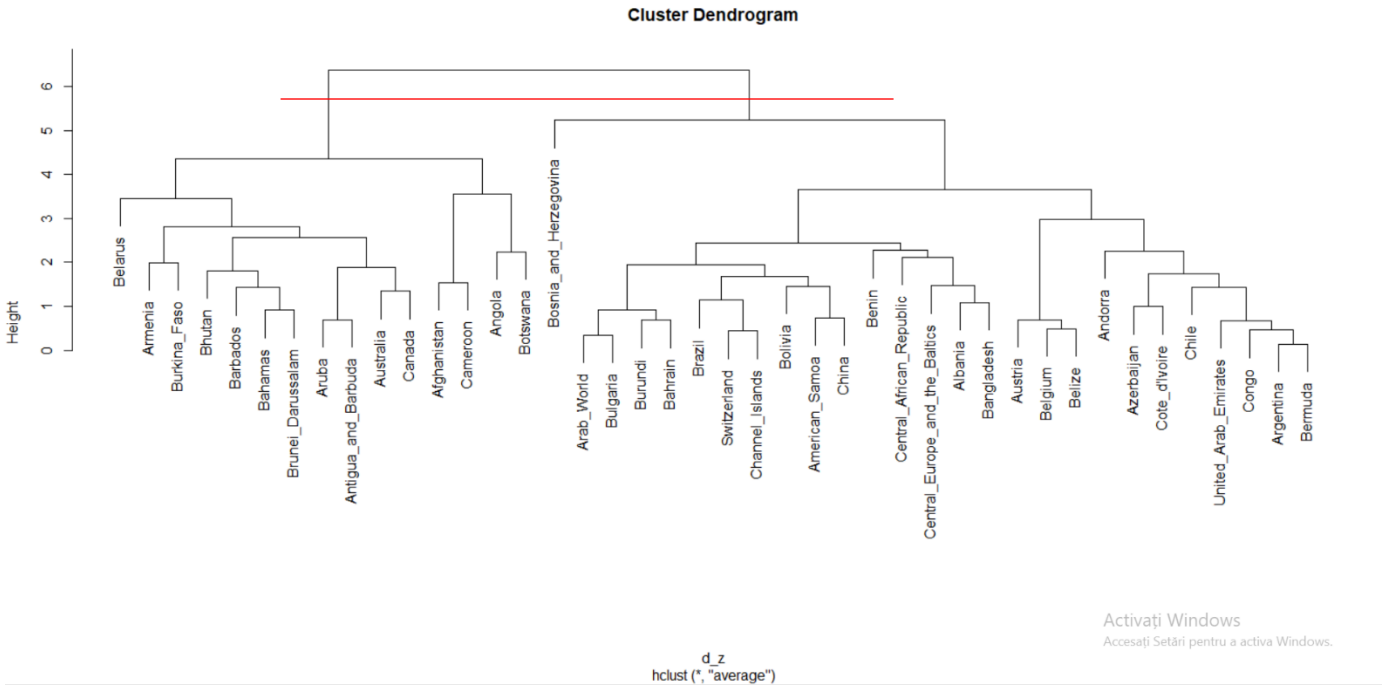
In the case of the main components, the figure above indicates that in step 1, observations 8 and 26 joined at a distance of 0.1362017. At step 4, observation 42 joined class 1 at a distance of 0.4589055. At step 12, classes 2 and 7 joined at a distance of 0.9237184.

Fig. 1.2.5. Dendrogram for standardized data




According to the figure above, the number of clusters resulting from this analysis is 2, because the horizontally drawn line intersects the dendrogram in two points.

Fig. 1.2.6 Dendrogram for principal components




Just as in the case of standardized data, and in the case of the cluster analysis for the main components using the average aggregation method, the number of formed classes is 2, since the horizontally drawn line intersects the dendrogram in 2 points.

Fig. 1.2.7. Determining the number of clusters for standardized data

```
> clust_3$height[41]-clust_3$height[40] #n=42 -> cea mai mare diferenta => 2 clustere
[1] 1.705267
> clust_3$height[40]-clust_3$height[39]
[1] 0.6013522
> clust_3$height[39]-clust_3$height[38]
[1] 0.07192235
> clust_3$height[38]-clust_3$height[37]
[1] 0.1843161
> clust_3$height[37]-clust_3$height[36]
[1] 0.2005141
> clust_3$height[36]-clust_3$height[35]
[1] 0.1147003
>
```

As can be seen in the figure presented above, the first calculated difference is the largest (1.705267), which confirms the determination of two clusters, an aspect also mentioned in the dendrogram analysis.

Fig. 1.2.8. Determining the number of clusters for main components

```
> clust_2$height[41]-clust_2$height[40] #n=42
[1] 1.145318
> clust_2$height[40]-clust_2$height[39]
[1] 0.8722953
> clust_2$height[39]-clust_2$height[38]
[1] 0.7117172
> clust_2$height[38]-clust_2$height[37]
[1] 0.09118623
> clust_2$height[37]-clust_2$height[36]
[1] 0.1063791
> clust_2$height[36]-clust_2$height[35]
[1] 0.4740138
>
```

As can be seen in the figure presented above, in the case of the analysis for the main components, the first calculated difference is also the largest (1.145318), which confirms the determination of two clusters, an aspect also mentioned in the dendrogram analysis.

Fig. 1.2.9. Proposals for the number of clusters for standardized data

```
> res<-NbClust(date_std, distance = "euclidean", min.nc=2, max.nc=5,method = "average", index = "all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 10 proposed 2 as the best number of clusters
* 5 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 4 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
>
```

The figure above represents the output of a function with the help of which several numbers of clusters are proposed into which the data used in the analysis could be divided, in the case of using standardized data. Thus, 10 proposals are for creating 2 clusters, 5 proposals for creating 3 clusters, 4 proposals for 4 clusters and 4 proposals for 5 clusters. Therefore, according to the majority rule, the best number of clusters is 2.

Fig. 1.2.10. Proposals for the number of clusters for principal components

```
> res2<-NbClust(scoruri, distance = "euclidean", min.nc=2, max.nc=5,
+               method = "average", index = "all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 6 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
>
```

The figure above represents the same output as in figure 1.2.9, only this time this is the analysis for the main components. Thus, 9 proposals are for creating 2 clusters, 7 proposals for creating 3 clusters, 1 proposal for 4 clusters and 6 proposals for 5 clusters. Therefore, according to the majority rule, the best number of clusters is 2.

Fig. 1.2.11. Proposals for the number of clusters and initiator values for standardized data

```
> res$All.index
      KL      CH Hartigan     CCC    Scott Marriot    TrCovW   TraceW    Friedman  Rubin Cindex     DB Silhouette   Duda Pseudot2   Beale Ratkowsky     Ball
2 9.6799 36.6264   5.7504  1.1065 112.5324       0 1533.0086 321.0384  8781816570 1.9157 0.4156 1.0162     0.4217 0.6933   5.7516  4.2430    0.4340 160.5192
3 4.3273 23.2256   2.3230 -0.8688 152.3137       0 1123.5571 280.6866  9435742278 2.1911 0.3978 1.2240     0.3319 0.3124   4.4024 15.1560    0.3983  93.5622
4 0.0724 16.7398  15.7743 -2.6529 182.5191       0 1056.0769 264.9074  9974503431 2.3216 0.3968 0.9443     0.3022 0.5899  17.3808  6.9042    0.3641  66.2269
5 5.6966 21.1388   3.5715 -0.0482 264.3193       0  406.2696 187.1989 10412591816 3.2853 0.5183 1.0614     0.2745 0.7586   2.8643  2.9583    0.3653  37.4398
  Ptbiserial   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2     0.7484 1.2662  0.4967 0.4313 0.0035  0.6755 2.6469 0.6686
3     0.7401 2.5142  0.6030 0.3727 0.0036  0.8204 2.4849 0.6689
4     0.7369 1.1249  0.6148 0.3727 0.0036  0.6670 2.3858 0.3859
5     0.6356 0.5277  1.4384 0.3765 0.0037  0.6264 2.0168 0.3370
> res$Best.nc
                KL      CH Hartigan    CCC   Scott Marriot   TrCovW  TraceW  Friedman   Rubin Cindex     DB Silhouette   Duda PseudoT2 Beale Ratkowsky   Ball
Number_clusters 2.0000  2.0000   4.0000 2.0000  5.0000       3   5.0000  3.0000         3  3.0000 4.0000 4.0000     2.0000 2.0000   2.0000    NA     2.000  3.000
Value_Index     9.6799 36.6264  13.4512 1.1065 81.8002       0 649.8073 24.5726 653925708 -0.1449 0.3968 0.9443     0.4217 0.6933   5.7516    NA     0.434 66.957
                PtBiserial   Frey McClain   Dunn Hubert SDindex Dindex  SDbw
Number_clusters     2.0000 4.0000  2.0000 2.0000       0  5.0000      0 5.000
Value_Index         0.7484 1.1249  0.4967 0.4313       0  0.6264      0 0.337
>
```

The figure above illustrates one value of an indicator, respectively one proposal for the number of clusters in which the data should be divided. For example, KL proposes a number of 2 clusters, this aspect being augmented by the values of the indicators, as its indicator for the number of 2 clusters is the highest (9.6799), compared to the values for another number of clusters.

Fig. 1.2.12. Proposals for the number of clusters and the initiator values for the main components

```
**********************************************************************
> res2$All.index
       KL      CH Hartigan     CCC    Scott Marriot   TrCovW   TraceW Friedman  Rubin Cindex     DB Silhouette   Duda Pseudot2  Beale Ratkowsky     Ball Ptbiserial
2  2.4168 50.3840   4.6002  1.2992  86.7923 1352949 8644.724 238.6799   6.8769 2.2596 0.3793 0.8608     0.4966 0.9804   0.4993 0.0327    0.2310 119.3399     0.7528
3 12.0292 29.6295   8.8381 -1.6263  95.4918 2474613 6717.606 214.0618   7.1818 2.5195 0.4303 0.6951     0.4043 0.5693   9.8343 1.1959    0.2684  71.3539     0.7678
4  0.0508 26.4786  20.5539 -2.3304 120.1964 2443040 3925.528 174.5140  10.1282 3.0904 0.4538 0.8247     0.3702 0.4988  24.1109 1.6419    0.2864  43.6285     0.7582
5 16.4563 34.7985   3.9657  0.2513 170.8354 1143211 1425.931 113.2552  13.9909 4.7620 0.4264 0.8904     0.3627 0.2515   5.9507 3.3769    0.3049  22.6510     0.6607
    Frey McClain   Dunn Hubert SDindex Dindex   SDbw
2 0.5389  0.4335 0.3727 0.0032  0.6166 2.2008 0.5640
3 1.1506  0.4743 0.3408 0.0033  0.5110 2.0837 0.4204
4 1.0426  0.5679 0.2606 0.0035  0.6798 1.8910 0.3879
5 0.4845  1.1687 0.1586 0.0037  0.7248 1.5093 0.2816
> res2$Best.nc
                 KL      CH Hartigan    CCC  Scott  Marriot   TrCovW   TraceW Friedman  Rubin Cindex     DB Silhouette   Duda PseudoT2  Beale Ratkowsky   Ball
Number_clusters  5.0000  2.000   5.0000 2.0000  5.000       3    4.000   3.0000   5.0000 3.0000 2.0000 3.0000     2.0000 2.0000   2.0000 2.0000    5.0000  3.000
Value_Index     16.4563 50.384  16.5882 1.2992 50.639 -1153237 2792.079 -14.9297   3.8627 0.3111 0.3793 0.6951     0.4966 0.9804   0.4993 0.0327    0.3049 47.986
                PtBiserial Frey McClain   Dunn Hubert SDindex Dindex   SDbw
Number_clusters     3.0000     1  2.0000 2.0000       0   3.000      0 5.0000
Value_Index         0.7678    NA  0.4335 0.3727       0   0.511      0 0.2816
>
```

The figure above illustrates one value of an indicator, respectively one proposal for the number of clusters in which the data should be divided. For example, KL proposes a number of 5 clusters, this aspect being augmented by the values of the indicators, as its indicator for the number of 5 clusters is the highest (16.4563), compared to the values for another number of clusters.

### Silhouette chart for standardized data

The figure below suggests that for the data used in this analysis, the number of 2 clusters is optimal compared to a larger number of clusters, since most of the coefficients

approaches 1 and very few values are close to 0, which indicates a large distance from the other cluster. No negative values can be observed in the analyzed figure, which indicates that there are no objects wrongly included in the two clusters.

The average for cluster 2 is 0.46, which is slightly higher than the average of the cluster 1, which is 0.36. It should be mentioned that the number of centralized objects also differs: cluster 1 contains 15 objects, while the second one contains 27 objects. The distribution of observations in the two clusters is balanced, there being no large discrepancies between the number of elements between the two clusters. The value of the average silhouette width indicator is 0.42.

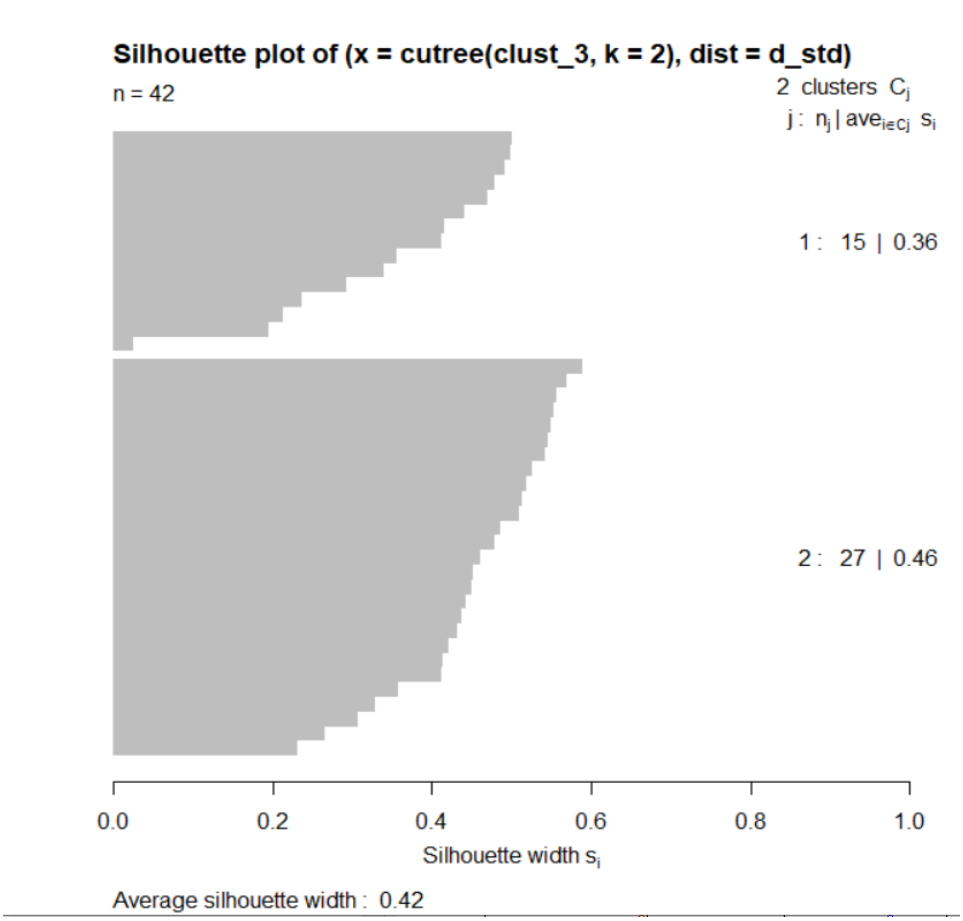



Fig.1.2.13 Silhouette chart for standardized data

Fig. 1.2.14. Silhouette coefficient values for standardized data

```
> si4_std2
      cluster neighbor  sil_width
 [1,]       1        2 0.41106630
 [2,]       1        2 0.35651448
 [3,]       1        2 0.02426994
 [4,]       2        1 0.30738352
 [5,]       2        1 0.43245484
 [6,]       2        1 0.54889475
 [7,]       2        1 0.55575690
 [8,]       2        1 0.46087478
 [9,]       1        2 0.29192963
[10,]       2        1 0.58999462
[11,]       1        2 0.41616925
[12,]       1        2 0.33937474
[13,]       2        1 0.43670957
[14,]       2        1 0.44227381
[15,]       2        1 0.50919574
[16,]       2        1 0.41281762
[17,]       2        1 0.32868142
[18,]       1        2 0.21297436
[19,]       2        1 0.45241581
[20,]       2        1 0.55298625
[21,]       2        1 0.51891379
[22,]       1        2 0.50084567
[23,]       2        1 0.26525291
[24,]       1        2 0.49925595
[25,]       2        1 0.41302831
[26,]       2        1 0.52604799
[27,]       2        1 0.56944780
[28,]       2        1 0.45013390
[29,]       1        2 0.44022093
[30,]       1        2 0.46894138
[31,]       1        2 0.47972337
[32,]       1        2 0.23603176
[33,]       2        1 0.23048694
[34,]       1        2 0.49115633
[35,]       2        1 0.42014348
[36,]       2        1 0.51313899
[37,]       2        1 0.47895569
[38,]       2        1 0.35740993
[39,]       2        1 0.54566355
[40,]       2        1 0.48653868
[41,]       1        2 0.19454591
[42,]       2        1 0.54201872
attr(,"ordered")
[1] FALSE
attr(,"call")
silhouette.default(x = cutree(clust_3, k = 2), dist = d_std)
attr(,"class")
[1] "silhouette"
> 
```

The figure above shows the values of the Silhouette coefficients represented graphically in figure 1.2.13. Therefore, it is emphasized by the actual visualization of the values that most are close to 1 and only one value is very close to 0, this being value number 3 whose coefficient has the value of 0.024. This observation belongs to class 1, but the value very close to 0 of the coefficient indicates that it is very close to the border between the two clusters.

### Principal Components Silhouette Chart

The figure below also increases this time that for the data used in this analysis, the number of 2 clusters is optimal compared to a larger number of clusters, since most of the coefficients are close to 1 and very few values are close to 0, which indicates a large distance from the other cluster. No negative values can be observed in the analyzed figure, which indicates that there are no objects wrongly included in the two clusters.

The average for cluster 2 is 0.52, which is slightly higher than the average of the cluster 1, which is 0.46. It should be mentioned that the number of centralized objects also differs: cluster 1 contains 15 objects, while the second one contains 27 objects. The distribution of observations in the two clusters is balanced, there being no large discrepancies between the number of elements between the two clusters. The value of the average silhouette width indicator is 0.5.

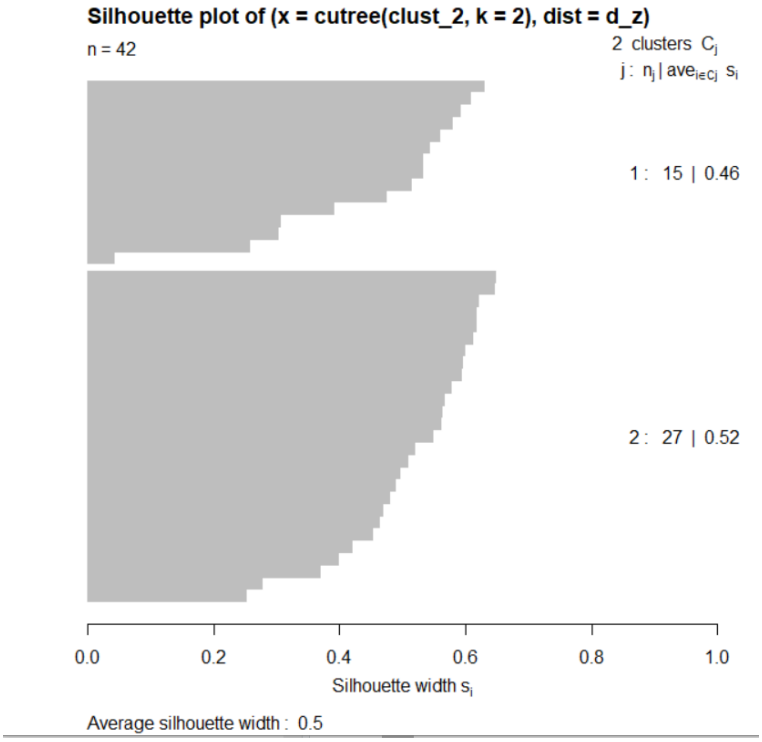



Fig. 1.2.15. Principal Components Silhouette Chart

Fig. 1.2.16. Silhouette coefficient values for principal components

```
> si4_z2
      cluster neighbor  sil_width
 [1,]       1        2 0.54368016
 [2,]       1        2 0.47570627
 [3,]       1        2 0.04263965
 [4,]       2        1 0.37015831
 [5,]       2        1 0.48016856
 [6,]       2        1 0.62163784
 [7,]       2        1 0.61743304
 [8,]       2        1 0.56825555
 [9,]       1        2 0.39171991
[10,]       2        1 0.64637260
[11,]       1        2 0.51416240
[12,]       1        2 0.53362156
[13,]       2        1 0.46916577
[14,]       2        1 0.51029946
[15,]       2        1 0.56338811
[16,]       2        1 0.49692266
[17,]       2        1 0.40009935
[18,]       1        2 0.30644634
[19,]       2        1 0.49044589
[20,]       2        1 0.61179968
[21,]       2        1 0.59979851
[22,]       1        2 0.63035284
[23,]       2        1 0.27826336
[24,]       1        2 0.56109309
[25,]       2        1 0.45370704
[26,]       2        1 0.57910184
[27,]       2        1 0.61753476
[28,]       2        1 0.56189149
[29,]       1        2 0.53240877
[30,]       1        2 0.59223139
[31,]       1        2 0.57970401
[32,]       1        2 0.30384027
[33,]       2        1 0.25254603
[34,]       1        2 0.60867248
[35,]       2        1 0.46471487
[36,]       2        1 0.59466516
[37,]       2        1 0.54890232
[38,]       2        1 0.42027022
[39,]       2        1 0.64828526
[40,]       2        1 0.52040844
[41,]       1        2 0.25896846
[42,]       2        1 0.59661131
attr(,"ordered")
[1] FALSE
attr(,"call")
silhouette.default(x = cutree(clust_2, k = 2), dist = d_z)
attr(,"class")
[1] "silhouette"
>
```

The figure above shows the values of the Silhouette coefficients represented graphically in figure 1.2.15. Therefore, it is emphasized by the actual visualization of the values that most are close to 1 and only one value is very close to 0, this being value number 3 whose coefficient has the value of 0.042. This observation belongs to class 1, but the value very close to 0 of the coefficient indicates that it is very close to the border between the two clusters.

**Conclusion**

Using the average aggregation method, it is observed that the number of clusters that are formed is 2, both in the case of standardized data and of the main components. In addition, both the Ward method and the mean aggregation method show similar results, and although the dendrograms differ from a visual point of view, the other results indicate the formation of two clusters, an aspect that emphasizes the correctness of the formation of two clusters and no more.

## 2. The K-means partitioning algorithm

K-means represents one of the partitional clustering methods by which the number of clusters is supposed to be known.

Steps of the K-means algorithm:

1) A number of k objects are randomly chosen from the data set, these being representative for each class. These k objects represent the initial environments.
2) The second step refers to the allocation of objects in each k class, that is, each observation will be distributed to a class, minimizing the distance to the centroid of the class. In other words, each object will be distributed in the class where the distance to the centroid is the smallest.

3) In the third step, the averages or centroids are recalculated.
4) at the fourth step, the algorithm from step 2 is repeated, repeating steps 2 and 3 until there are no more variations in the allocation of objects by class.

Fig. 2.1. Output K-std

```
> k_std
K-means clustering with 2 clusters of sizes 27, 15

Cluster means:
  EMP_VULN_MA EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE    GDP_EMP UEM_TOTL_MA UEM_TOTL_FE UEM_1524_MA UEM_1524_FE SRV_EMPL_MA SRV_EMPL_FE IND_EMPL_MA IND_EMPL_FE
1  -0.6499042  -0.6831396      0.6495465      0.6834645 -0.1734080   0.1698776   0.2275433   0.1995525   0.2657922   0.5729962   0.6692919   0.5109442  0.02034593
2   1.1698276   1.2296513     -1.1691838     -1.2302360  0.3121344  -0.3057796  -0.4095779  -0.3591945  -0.4784260  -1.0313932  -1.2047255  -0.9196996 -0.03662268
  AGR_EMPL_MA AGR_EMPL_FE
1  -0.6289389  -0.6751312
2   1.1320900   1.2152362

Clustering vector:
                      Aruba                 Afghanistan                      Angola                     Albania                     Andorra
                          2                           2                           2                           1                           1
                 Arab_World         United_Arab_Emirates                   Argentina                     Armenia              American_Samoa
                          1                           1                           1                           2                           1
          Antigua_and_Barbuda                   Australia                     Austria                  Azerbaijan                     Burundi
                          2                           2                           1                           1                           1
                    Belgium                       Benin                Burkina_Faso                  Bangladesh                    Bulgaria
                          1                           1                           2                           1                           1
                    Bahrain                     Bahamas      Bosnia_and_Herzegovina                     Belarus                      Belize
                          1                           2                           1                           2                           1
                    Bermuda                     Bolivia                      Brazil                    Barbados          Brunei_Darussalam
                          1                           1                           1                           2                           2
                     Bhutan                    Botswana      Central_African_Republic                      Canada Central_Europe_and_the_Baltics
                          2                           2                           1                           2                           1
                Switzerland              Channel_Islands                       Chile                       China               Cote_d'Ivoire
                          1                           1                           1                           1                           1
                   Cameroon                       Congo
                          2                           1

Within cluster sum of squares by cluster:
[1] 189.4818 131.5565
 (between_SS / total_SS =  47.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
```

In the figure above it can be seen that within the K-means algorithm, using the standardized data, choosing a number of 2 clusters according to the cluster analysis carried out previously, the 42 observations are divided as follows: 27 observations in a class , 15 in the other.

Cluster means represent the centroids calculated the first time, i.e. the objects chosen the first time, at step 1.

Clustering vectors represent the section that indicates each observation in which the class is distributed following the choice of random objects. Therefore, according to step 2 of the algorithm presented above, their allocation in classes was made based on the smallest distance to the centroid calculated above.

For example, Aruba, Afghanistan, Angola are in class 2, while Albania, Andorra, Arab World are part of class 1.

Fig. 2.2 Output K_z

```
> k_z
K-means clustering with 2 clusters of sizes 27, 15

Cluster means:
         Z1          Z2          Z3
1 -1.993525 -0.05032238 -0.007868932
2  3.588344  0.09058028  0.014164078

Clustering vector:
                        Aruba                   Afghanistan                        Angola                       Albania                       Andorra
                            2                             2                             2                             1                             1
                   Arab_World           United_Arab_Emirates                     Argentina                       Armenia                American_Samoa
                            1                             1                             1                             2                             1
           Antigua_and_Barbuda                     Australia                       Austria                    Azerbaijan                      Burundi
                            2                             2                             1                             1                             1
                      Belgium                         Benin                  Burkina_Faso                    Bangladesh                      Bulgaria
                            1                             1                             2                             1                             1
                      Bahrain                       Bahamas          Bosnia_and_Herzegovina                       Belarus                        Belize
                            1                             2                             1                             2                             1
                      Bermuda                       Bolivia                        Brazil                      Barbados           Brunei_Darussalam
                            1                             1                             1                             2                             2
                       Bhutan                      Botswana         Central_African_Republic                        Canada Central_Europe_and_the_Baltics
                            2                             2                             1                             2                             1
                  Switzerland                Channel_Islands                         Chile                         China                 Cote_d'Ivoire
                            1                             1                             1                             1                             1
                     Cameroon                         Congo
                            2                             1

Within cluster sum of squares by cluster:
[1] 146.85398  91.82592
 (between_SS / total_SS =  55.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> |
```

The figure above has the same interpretation as figure 2.1, but this time the main components are used. Thus, out of the 42 observations, they are divided as follows: 27 observations in one class, 15 in the other.

Cluster means represents the centroids calculated the first time, i.e. the objects chosen the first time, at step 1. 2 centroids are calculated, because the number of clusters used in this algorithm is 2, according to the cluster analysis carried out previously.

Clustering vectors represent the section that indicates each observation in which the class is distributed following the choice of random objects. Therefore, according to step 2 of the algorithm presented above, their allocation in classes was made based on the smallest distance to the centroid calculated above.

And in the case of the main components, Aruba, Afghanistan, Angola are in class 2, while Albania, Andorra, Arab World are part of class 1.

The graphic below illustrates the visual representation of the two formed classes.It can be seen that between 0.5 and 1 the graph is divided into two: cluster 1 (red) and cluster 2 (blue), which means that along the x-axis, the first cluster has low values regarding the vulnerability of the male population to Employment. On the other hand, the second class has medium to high values along the variable x axis. Along the y-axis, that of the male population working in the field of services, the values are high for the first class and low for the second. Therefore, from the point of view of the two axes, it can be interpreted that the 42 observations are divided into two clusters, one of which represents the developed countries in terms of employment, this being class 1, and the other represents the countries less developed in terms of employment,

In this graph, the Central African Republic is distinguished by the fact that it is further away from the other elements in the cluster and is closer to the border between the two clusters. This fact represents the fact that the value of the Silhouette coefficient for this country is very close to 0, just as it was mentioned in the analysis of the Silhouette graph.

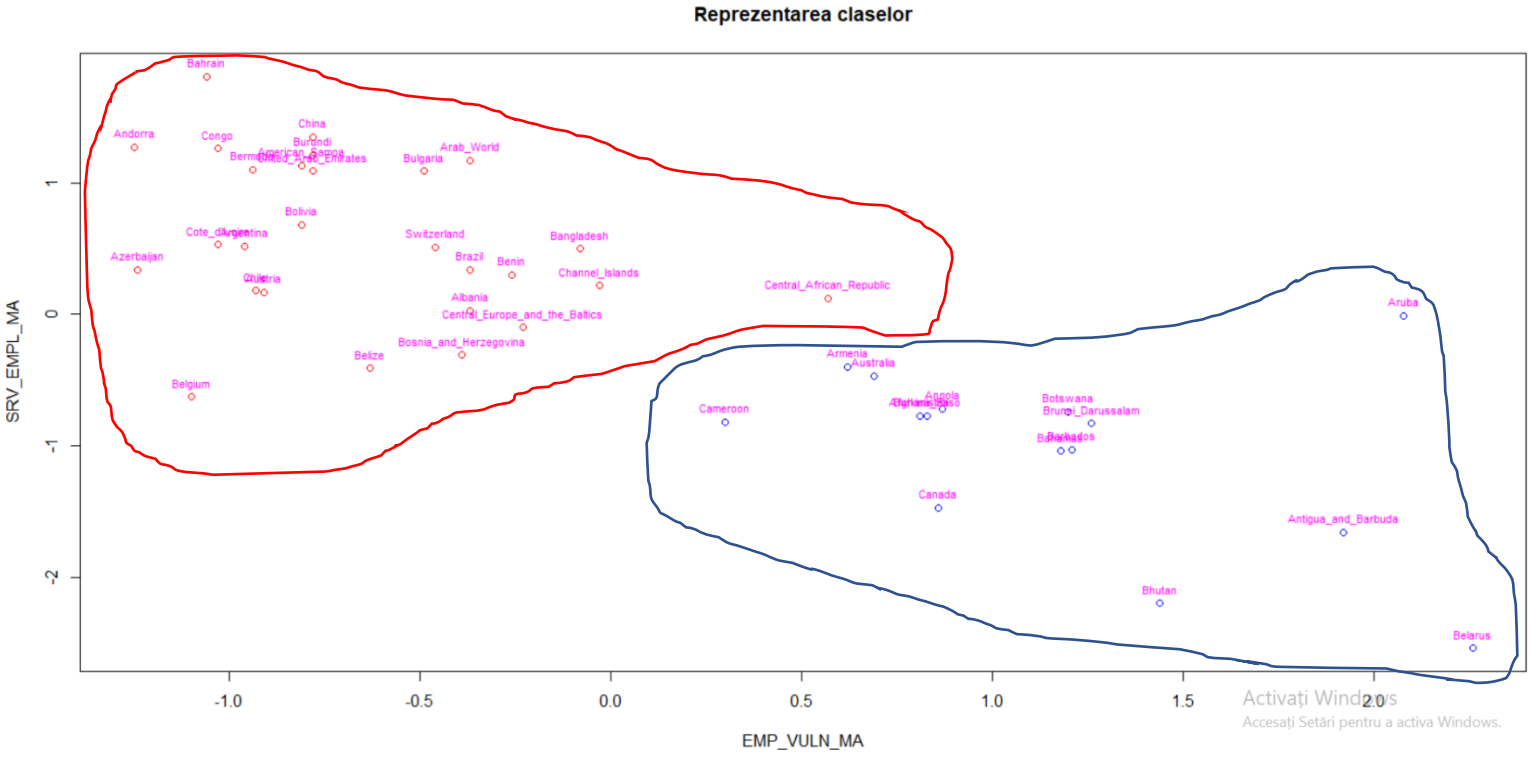



Fig. 2.3. Representation of classes for standardized data

The graph below illustrates the visual representation of the two classes formed for the main components.It can be seen that between 0 and 0.5 the graph is divided into two: cluster 1 (red) and cluster 2 (blue), which means that along the x-axis, the first cluster has low values in terms of the population's vulnerability to occupation workforce (Z1). On the other hand, the second class has quite large values along the axis of the variable x. Along the y-axis, that of the unemployed population (Z2), the values are low for the first class and for the second class vary from medium to high for the second. Therefore, from the point of view of the two axes, it can be interpreted that the 42 observations are divided into two clusters, one of which represents the developed countries in terms of employment, this being class 1, and the other represents the countries less developed in terms of employment,

Also, from an economic point of view, analyzing this graph through the prism of the two axes (Z1 and Z2), the countries in cluster 1 are the countries where the variability of employment is low, and the rate of the unemployed population varies, but low values predominate. For cluster 2, the variability is high, and high values also predominate in terms of the unemployed population. This aspect increases the idea that the first cluster gathers developed countries from the point of view of employment, and the second the less developed countries from this point of view.

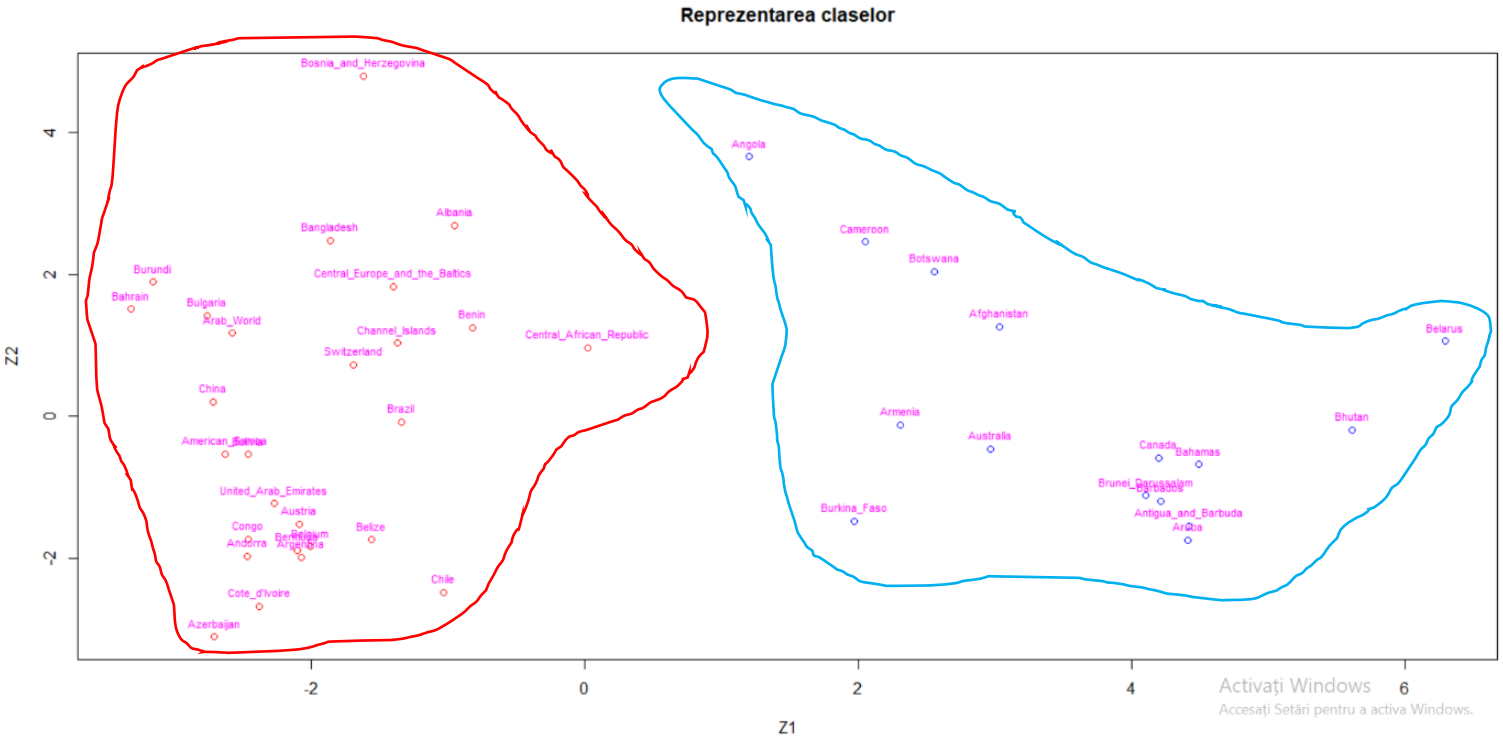



Fig. 2.4. Representation of classes for main components

**Decomposition of variability**

In the case of an efficient clustering, the intra-class variability must be as low as possible, and the inter-class variability must be as high as possible.

Fig. 2.5. Variability decomposition for standardized data

```
> variab_std
     spat_std spaw_std spab_std r_cls_std
[1,]      615 321.0384 293.9616  0.915659
>
```

In the case of standardized data, the variability between classes has a fairly high value (293.9616), which indicates that between them, the classes are quite heterogeneous. Despite this fact, the intra-class variability (321.0384) is higher than the inter-class one, which means that the data are not very homogeneous, but the difference between inter-class and intra-class variability is quite small.

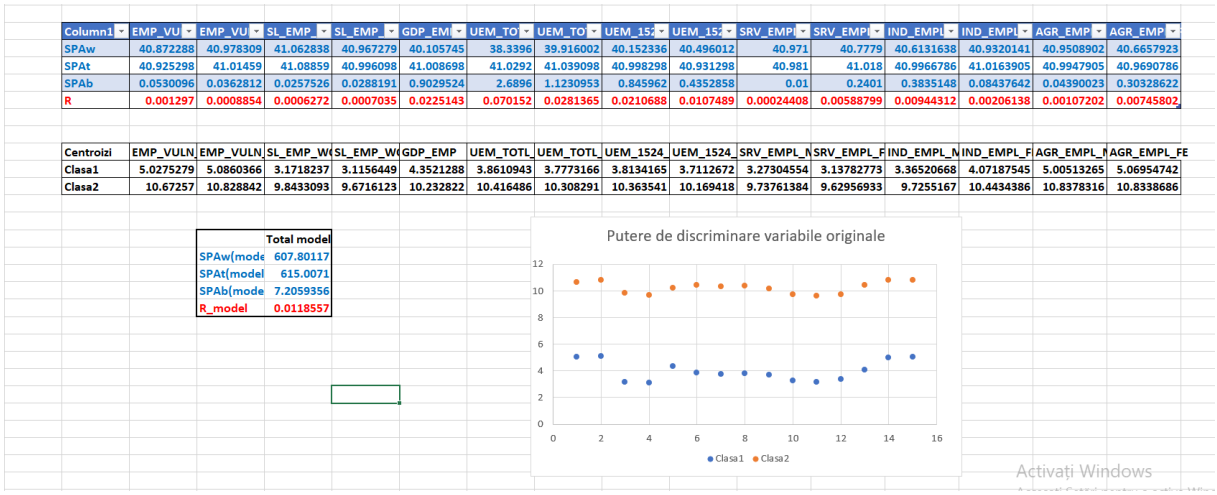
Fig. 2.6. Variability decomposition for principal components

```
> variab_z
      spat_z   spaw_z   spab_z r_cls_z
[1,] 539.321 238.6799 300.6411  1.2596
>
```

In the case of the main components, the variability between classes has a high value (300.6411), which indicates that between them, the classes are heterogeneous. This value is higher than in the case of standardized data, an aspect that also indicates greater heterogeneity. The variability within the class (238.6799) is smaller than the interclass, which means that the data are quite homogeneous, although the difference between the interclass and intraclass variability is quite small.

R is higher for the principal components than for the standardized data, which suggests that in this case, the criterion is more satisfactory than in the case of the standardized data. In this case, the criterion is more satisfactory, because the data are not correlated 2 by 2.

Fig. 2.7. Excel variability decomposition



| Column1 | EMP_VU | EMP_VU | SL_EMP_ | SL_EMP_ | GDP_EMI | UEM_TO | UEM_TO | UEM_15 | UEM_15 | SRV_EMPI | SRV_EMPI | IND_EMPL | IND_EMPL | AGR_EMP | AGR_EMP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SPAw | 40.872288 | 40.978309 | 41.062838 | 40.967279 | 40.105745 | 38.3396 | 39.916002 | 40.152336 | 40.496012 | 40.971 | 40.7779 | 40.6131638 | 40.9320141 | 40.9508902 | 40.6657923 |
| SPAt | 40.925298 | 41.01459 | 41.08859 | 40.996098 | 41.008698 | 41.0292 | 41.039098 | 40.998298 | 40.931298 | 40.981 | 41.018 | 40.9966786 | 41.0163905 | 40.9947905 | 40.9690786 |
| SPAb | 0.0530096 | 0.0362812 | 0.0257526 | 0.0288191 | 0.9029524 | 2.6896 | 1.1230953 | 0.845962 | 0.4352858 | 0.01 | 0.2401 | 0.3835148 | 0.08437642 | 0.04390023 | 0.30328622 |
| R | 0.001297 | 0.0008854 | 0.0006272 | 0.0007035 | 0.0225143 | 0.070152 | 0.0281365 | 0.0210688 | 0.0107489 | 0.00024408 | 0.00588799 | 0.00944312 | 0.00206138 | 0.00107202 | 0.00745802 |

| Centroizi | EMP_VULN | EMP_VULN | SL_EMP_W | SL_EMP_W | GDP_EMP | UEM_TOTL | UEM_TOTL | UEM_1524_ | UEM_1524_ | SRV_EMPL_N | SRV_EMPL_F | IND_EMPL_N | IND_EMPL_F | AGR_EMPL_M | AGR_EMPL_FE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clasa1 | 5.0275279 | 5.0860366 | 3.1718237 | 3.1156449 | 4.3521288 | 3.8610943 | 3.7773166 | 3.8134165 | 3.7112672 | 3.27304554 | 3.13782773 | 3.36520668 | 4.07187545 | 5.00513265 | 5.06954742 |
| Clasa2 | 10.67257 | 10.828842 | 9.8433093 | 9.6716123 | 10.232822 | 10.416486 | 10.308291 | 10.363541 | 10.169418 | 9.73761384 | 9.62956933 | 9.7255167 | 10.4434386 | 10.8378316 | 10.8338686 |

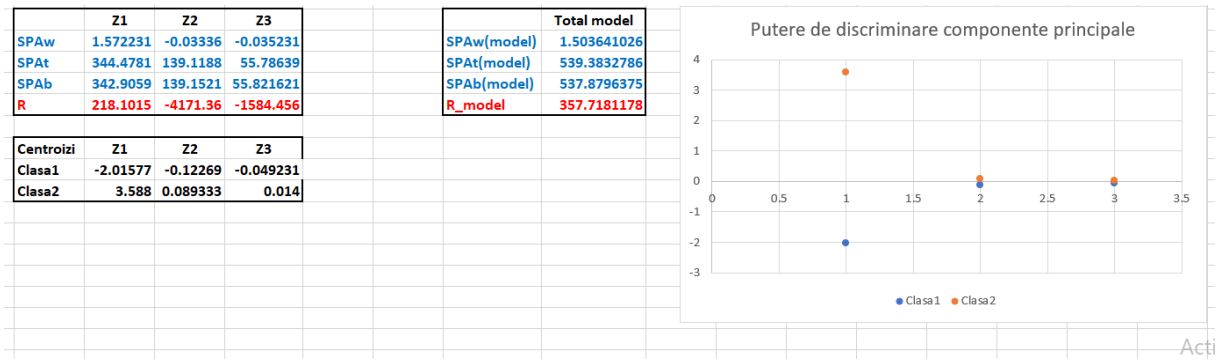
| | Total model |
|---|---|
| SPAw(mode | 607.80117 |
| SPAt(model | 615.0071 |
| SPAb(mode | 7.2059356 |
| R_model | 0.0118557 |





Decomposing the variability in Excel shows that the variability between classes (SPAb) is very small, and the intraclass (SPAw) is very high, this aspect indicating a small heterogeneity between classes and a small homogeneity within the class. Also, the value of R is also very small.

Fig. 2.8. Excel variability decomposition



| | Z1 | Z2 | Z3 |
|---|---|---|---|
| SPAw | 1.572231 | -0.03336 | -0.035231 |
| SPAt | 344.4781 | 139.1188 | 55.78639 |
| SPAb | 342.9059 | 139.1521 | 55.821621 |
| R | 218.1015 | -4171.36 | -1584.456 |

| Centroizi | Z1 | Z2 | Z3 |
|---|---|---|---|
| Clasa1 | -2.01577 | -0.12269 | -0.049231 |
| Clasa2 | 3.588 | 0.089333 | 0.014 |

| | Total model |
|---|---|
| SPAw(model) | 1.503641026 |
| SPAt(model) | 539.3832786 |
| SPAb(model) | 537.8796375 |
| R_model | 357.7181178 |





Decomposing the variability in Excel for the main components shows that the variability between classes (SPAb) is high, and the intraclass (SPAw) is very small, this aspect indicating a high heterogeneity between classes and a high homogeneity within the class. Also, the value of R is also very high.

Making a comparison between the two decompositions, it is visible again this time that the analysis of the main components is more meaningful, precisely because the data are not correlated two by two and thus errors in the analysis can be avoided.

### 3. Discriminant analysis

Discriminant analysis is a supervised recognition technique, because the class membership is known in advance. This aims to find a classification model for

to make further predictions. This method aims to find a rule by which to separate the two clusters as best as possible.

In order to perform discriminant analysis, the sample must be divided into two data sets: the training set (contains more of the amount of information) the test set (contains a smaller amount of the sample).

**Bayes**

Fig.3.1 Training set

| | Z1 | Z2 | Z3 | df[, 5] |
|---|---|---|---|---|
| Aruba | 4.41167344 | -1.74994230 | 1.294171377 | cls2 |
| Afghanistan | 3.02639633 | 1.26081477 | -1.102387549 | cls2 |
| Angola | 1.19628520 | 3.66104308 | 1.682635493 | cls2 |
| Albania | -0.94853306 | 2.68547928 | -0.350202573 | cls1 |
| Andorra | -2.47352505 | -1.96617327 | -2.405408729 | cls1 |
| Arab_World | -2.57523691 | 1.16626678 | -0.336946732 | cls1 |
| United_Arab_Emirates | -2.26963413 | -1.22621532 | -0.582289747 | cls1 |
| Argentina | -2.07087743 | -1.99171670 | -0.612382640 | cls1 |
| Armenia | 2.31484457 | -0.11754921 | -1.411386470 | cls2 |
| American_Samoa | -2.62882474 | -0.53089669 | -0.441014353 | cls1 |
| Antigua_and_Barbuda | 4.42232350 | -1.55043104 | 1.958221035 | cls2 |
| Australia | 2.96843971 | -0.45615686 | 0.237856821 | cls2 |
| Austria | -2.09258105 | -1.51791626 | 2.557512098 | cls1 |
| Azerbaijan | -2.71257813 | -3.11234468 | 0.001266369 | cls1 |
| Burundi | -3.16286871 | 1.88871259 | -1.037627388 | cls1 |
| Belgium | -2.00827021 | -1.83716037 | 1.969354325 | cls1 |
| Benin | -0.81811899 | 1.24914314 | -1.474929684 | cls1 |
| Burkina_Faso | 1.96630869 | -1.47838224 | -0.002986466 | cls2 |
| Bangladesh | -1.86447189 | 2.48043726 | 0.185205384 | cls1 |
| Bulgaria | -2.75759372 | 1.41768032 | -0.189825832 | cls1 |
| Bahrain | -3.31899262 | 1.50819337 | -0.493870580 | cls1 |
| Bahamas | 4.48805894 | -0.67874716 | -0.199360697 | cls2 |
| Bosnia_and_Herzegovina | -1.62350437 | 4.79597410 | 0.279831983 | cls1 |

Showing 1 to 24 of 42 entries

This represents the training set used in the prediction in which all 42 observations used in the analysis are entered.

Fig. 3.2. The test set

```
> test_z
                          Z1         Z2         Z3
Congo_Rep.         0.5301350  2.2640163  0.5761229
Colombia          -0.3373864 -2.0921258 -0.2876505
Comoros           -1.2231804  1.0861213 -0.1982109
Cabo_Verde         0.4920709 -1.6312034 -0.5798869
Costa_Rica        -0.7087435 -1.6712030  0.2639938
Caribbean_small_states  2.4701445  3.0362463  0.3970896
Cuba              -1.2230400 -0.9918517 -0.1714581
>
```

This is the test set used in the discriminant analysis. As can be seen, it contains a smaller amount of information than the training set, being made up of outliers.

To be able to use the Bayesian discriminant, the independence of the events and the normality of the variables must be taken into account.

According to the graphs below, the normal distribution is represented in red, and the distribution for Z is presented in blue. Analyzing the 6 graphs, it can be seen that most of the distributions for Z are similar to the normal distribution (graph 1,2,5,6). This aspect indicates the normality of the variables and suggests that the Bayesian discriminant can be used.

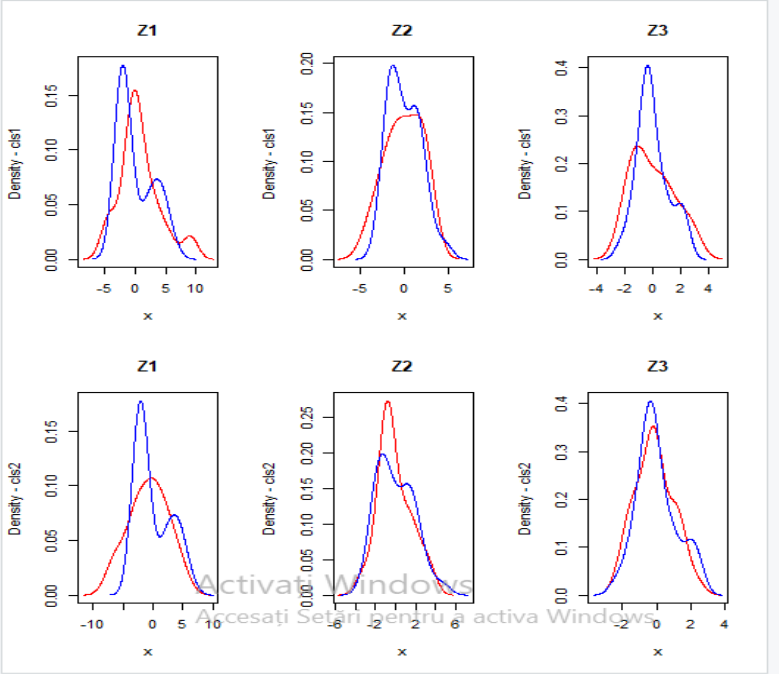



Fig. 3.3. Verification of Bayes normal distribution

Fig. 3.4. Output Bayes

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = train_z[, -4], y = train_z[, 4])

A-priori probabilities:
train_z[, 4]
     cls1      cls2
0.3571429 0.6428571

Conditional probabilities:
            Z1
train_z[, 4]      [,1]      [,2]
        cls1  3.588344 1.4309603
        cls2 -1.993525 0.7713966

            Z2
train_z[, 4]        [,1]     [,2]
        cls1  0.09058028 1.638350
        cls2 -0.05032238 1.972712

            Z3
train_z[, 4]         [,1]     [,2]
        cls1  0.014164078 1.351723
        cls2 -0.007868932 1.077768

> |
```

The figure above represents the output of the NaiveBayes function where A-priori probabilities represent the a priori probabilities ( ( = ))which can be estimated by means of the weight of entities belonging to class k. For the first cluster, the probability is approximately 36%, and for the second class, the probability is approximately 64%, which indicates that the variables in the second class predominate. The conditional probabilities represent the probability densities for each Z separately in the respective class. For example, for Z1, the probability density for house 1 is 3.588 and 1.43, and for class 2 -1.99 and 0.77.

Fig. 3.5. The prediction for the training set

```
> table(predict(bayes_z, train_z), train_z[,4])

      cls1 cls2
  cls1   15    0
  cls2    0   27
>
```

The figure above shows how the variables are classified in each class. On the main diagonal it can be seen that in class 1 there are 15 variables and in class 2 there are 27 variables. Values of 0 indicate that there are no variables incorrectly classified in any of the classes.

Fig. 3.6. The prediction for the test set

```
> predict(bayes_z, test_z)
[1] cls1 cls2 cls2 cls1 cls2 cls1 cls2
Levels: cls1 cls2
> #se clasifica fiecare obs intr-o clasa #clasa 1 15 ob clasa 2 27 ob
> table(predict(bayes_z, test_z))

cls1 cls2
   3    4
;
> rownames(test_z)
[1] "Congo_Rep."            "Colombia"              "Comoros"               "Cabo_Verde"            "Costa_Rica"            "Caribbean_small_states"
[7] "Cuba"
>
```

In the case of the test set, consisting of 7 outliers, it can be observed that 3 of them were distributed in class 1 and 4 were distributed in class 2. And in the case of outliers, the number of variables in class 2 predominates.

## LDA

The training set for LDA consists of columns 1:13 of the standardized data.

The test set for Lda consists of columns 1:13 of the standardized outlier data.

Fig. 3.7. Output LDA standardized data

```
> ad_std
Call:
lda(train_std[, 1:13], grouping = clase_std)

Prior probabilities of groups:
        1         2
0.3571429 0.6428571

Group means:
  EMP_VULN_MA EMP_VULN_FE SL_EMP_WORK_MA SL_EMP_WORK_FE    GDP_EMP UEM_TOTL_MA UEM_TOTL_FE UEM_1524_MA UEM_1524_FE SRV_EMPL_MA SRV_EMPL_FE IND_EMPL_MA IND_EMPL_FE
1   1.1698276   1.2296513     -1.1691838     -1.2302360  0.3121344  -0.3057796  -0.4095779  -0.3591945  -0.4784260  -1.0313932  -1.2047255  -0.9196996 -0.03662268
2  -0.6499042  -0.6831396      0.6495465      0.6834645 -0.1734080   0.1698776   0.2275433   0.1995525   0.2657922   0.5729962   0.6692919   0.5109442  0.02034593

Coefficients of linear discriminants:
                    LD1
EMP_VULN_MA      3.06143476
EMP_VULN_FE     -3.56792590
SL_EMP_WORK_MA   1.92029006
SL_EMP_WORK_FE -0.02636550
GDP_EMP         -0.08391744
UEM_TOTL_MA     -0.75339573
UEM_TOTL_FE      1.42488117
UEM_1524_MA     -0.00912929
UEM_1524_FE     -0.29991536
SRV_EMPL_MA     -0.36248028
SRV_EMPL_FE      1.40885364
IND_EMPL_MA      0.08635531
IND_EMPL_FE      0.26455390
> |
```

The figure below shows again the probabilities of belonging to the two classes: for class 1 the probability is approximately 36%, and for class 2 it is approximately 64%. Group means represent the centroids for each class. For example, the centroids for class 1 are: 1.16, 1.22, etc., and for class 2 they are: -0.64, -0.63, etc.

Using the coefficients of the discriminant functions, we write the following general form:

* the 15 indicators will be marked with i from 1 to 13.

General form:

3.06i1-3.57i2+1.92i3-0.02i4-0.08i5-0.75i6+1.42i7-0.009i8-0.29i9-0.36i10+1.40i11+0.086i12+0.264i13

Fig. 3.8. ConfMatrix standardized data

```
          Predicted
Original  1  2
    cls1 15  0
    cls2  0 27
> |
```

The figure above illustrates the degree of accuracy, from which it can be seen that 15 observations are in class 1 and 27 in class 2 and no observation is wrongly assigned. Thus, the degree of accuracy is 1 and the degree of error, which is equal to 1-degree of accuracy, is equal to 0.

Fig. 3.9. LDA prediction for the training set for standardized data

```
> predict_std$class  #vector de clase : clasele de apartenenta a celor 34 de companii folosind modelul ad_std
 [1] 1 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 1 2 2 2 1 2 1 2 2 2 2 2 1 1 1 2 1 2 2 2 2 2 2 1 2
Levels: 1 2
> table(predict_std$class)  #vector de clase : clasele de apartenenta a celor 34 de companii folosind modelul ad_std

 1  2
15 27
>
```

And in the case of the LDA classifier for standardized data, it can be seen that 15 observations went to class 1 and 27 went to class 2.

Fig. 3.10. Prediction for the test set standardized data

```
> predict_std = predict(ad_std,test_std[,1:13])
> predict_std
$class
[1] 2 2 2 2 2 1 2
Levels: 1 2

$posterior
                                1            2
Congo_Rep.              6.565980e-07 9.999993e-01
Colombia                7.626759e-09 1.000000e+00
Comoros                 7.080170e-09 1.000000e+00
Cabo_Verde              9.922999e-08 9.999999e-01
Costa_Rica              4.383567e-08 1.000000e+00
Caribbean_small_states  1.000000e+00 6.127186e-22
Cuba                    8.987064e-10 1.000000e+00

$x
                              LD1
Congo_Rep.               0.844846
Colombia                 1.458969
Comoros                  1.469219
Cabo_Verde               1.105308
Costa_Rica               1.217920
Caribbean_small_states -7.849992
Cuba                     1.753729

>
```

The figure above shows the allocation of outliers in classes for the standardized data, so that 6 variables are in class 2 and 1 in class 1.

Fig. 3.11. Output LDA principal components

```
> ad_z
Call:
lda(train_z[, 1:3], grouping = cls)

Prior probabilities of groups:
        1         2
0.3571429 0.6428571

Group means:
         Z1          Z2           Z3
1  3.588344  0.09058028  0.014164078
2 -1.993525 -0.05032238 -0.007868932

Coefficients of linear discriminants:
           LD1
Z1 -0.95664319
Z2 -0.05988534
Z3 -0.02332455
>
```

And in the case of the main components, the probability of belonging to class 1 is approximately 36% and the probability of belonging to class 2 is 64%. Group means represent the centroids for the principal components for each class. For example, the centroids for class 1 are 3.58, 0.09, 0.01, and the centroids for class 2 are: -1.99, -0.05, - 0.007.

The general form with the coefficients of the discriminant functions:

- 0.95z1-0.05z2-0.02z3

Fig. 3.12. LDA prediction for the training set for Z

```
> predict_z = predict(ad_z,train_z[,1:3])
> predict_z$class
 [1] 1 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 1 2 2 2 1 2 1 2 2 2 2 1 1 1 1 2 1 2 2 2 2 2 2 1 2
Levels: 1 2
> table(predict_z$class)

 1  2
15 27
>
```

In the figure above, it can be seen that from the training set, 15 observations are in class 1 and 27 observations in class 2, as in the case of the other discriminant.

Fig. 3.13. ConfMatrix for main components

```
> confMatrix_z #grad de accuratete
          Predicted
Original  1  2
    cls1 15  0
    cls2  0 27
>
```

The figure above illustrates the degree of accuracy, from which it can be seen that 15 observations are in class 1 and 27 in class 2 and no observation is wrongly assigned. Thus, the degree of accuracy is 1 and the degree of error, which is equal to 1-degree of accuracy, is equal to 0.

Fig. 3.14. The test set prediction for Z

```
> predict_z
$class
[1] 2 2 2 2 2 1 2
Levels: 1 2

$posterior
                                1           2
Congo_Rep.              2.377290e-01 0.762271037
Colombia                8.184382e-04 0.999181562
Comoros                 2.465231e-05 0.999975348
Cabo_Verde              5.998162e-02 0.940018383
Costa_Rica              1.501327e-04 0.999849867
Caribbean_small_states  9.998750e-01 0.000125048
Cuba                    1.272200e-05 0.999987278

$x
                                LD1
Congo_Rep.              -0.6561692
Colombia                 0.4547554
Comoros                  1.1097277
Cabo_Verde              -0.3595255
Costa_Rica               0.7719376
Caribbean_small_states -2.5541355
Cuba                     1.2334095

>
```

The figure above shows the allocation of outliers in classes for main components, so that 6 variables are in class 2 and 1 in class 1.