2021/2022

The Best Country to spend your 20's

December, 2021

GROUP Z
ALESSANDRA RECORDARE E20211501
ANA AGOSTINHO R20201591
JOANA ESTROMPA R20201592
LUÍS SILVANO R20201479
PEDRO SOUSA R20201611

INDEX

1 - Summary

Since we are young people, our project has a purpose to find what is “the best country to spend your 20’s”, we found out interesting to know what is the country that can give us the best conditions to live our young years the best way. The main method that we used for research our materials was internet, namely statistical bases from EUROSAT, and then we advanced for our analysis. Our study will focus on the European countries and we had into account the years of 2015 and 2019. For our work we considered important factors like the economic situation, mental and physical health and some other social factors. To achieve the purpose, we have collected, treated and analysed the data fallowed by a Principal components factoring and cluster analysis. The main findings were that, once the majority of the European countries are developed countries most of the countries are good candidates to be the best country to spend your young years but, considering the factors each individual find out important to spend these years the result can be different. The one whose showed the most potential considering our variables and analysis was Czechia in both years (2015/2019), however other countries like Croatia and Austria showed some potential as well.

2 - Introduction

How we want to spend our young years is a question present in our minds for various reasons. Some people give more importance to economic factors, others to health questions and some to security. There are many issues that you can take into account when it comes to your mind. So we found out important combine some factors that we consider important in this subject and try to appoint the best country to spend your young years. Our main focus will be on questions like material deprivation, poverty and social exclusion, education, health and some other economic and social questions that we consider that are the most important factors and which can give us an answer to the big question.

3 - Methodology

3.1 - Data Description

After defining our purpose and the main points of our work, the first thing to do was find variables that go accordingly with our values and which can measure all the important factors that we discuss to be the most important to us to spend our young years. For a more efficient analysis we selected values form 2015 and 2019 which were the years that had more data about all countries and can provide us with a fair comparison. The next table shows our nine variables and also the meaning of each variable for a better interpretation of them.

Acronym	Designation	Description
Mat_Deprevation	Material Deprivation	Number of individuals between 20 and 29 years that are unable to afford desirable or even necessary items to assure their quality of life, per 1000 habitants
Health_Perc	Health Perception	Number of individuals with a great health perception, per 1000 habitants
Leave_Par_house	Leaving the parental household	Average age at which young people leave their parental household
Poverty	Poverty and social exclusion	Number of individuals that finds itself at risk of poverty or social exclusion, per 1000 habitants
House_Overburden	Housing Cost Overburden	Number of individual where their total housing costs represent more than 40 % of disposable income, per 1000 habitants
Employment	Employment	Indivuals between 20 and 29 years old who are employed, per 1000 habitants
Tem_Employ	Temporary Employment	Indivuals between 20 and 29 years old who are employed, per 1000 habitants
Education 0-2	Low Education Attainment	Population between 20 and 29 years old who have an educational level under high school or equal, per 1000 habitants
Dep_Symptoms	Depressive Symptoms	Indivuduals between 20 and 29 years old who have felt developed depressive symptoms in the last year, per 1000 habitants

Table 1 - Variables Descriptions

3.2 - Data Treatment

After collect all the data and merging the datasets that we found necessary for the devolvement of our analysis, we started organizing and transposing them into a table on excel. The next step was do a summary statistic to have a better look into our variables. We found some missing values as the next table shows. However before treating this missing values we must took out the outliers from the dataset and it is what our next point talks about.

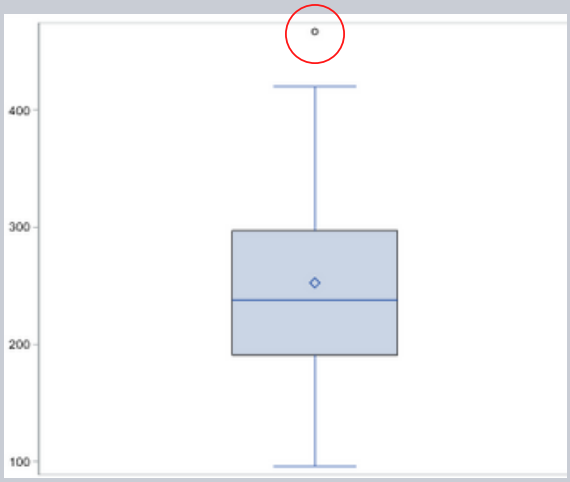
	Mean	Std Dev	Variance	N	N Miss
Mat_Deprevation	85.889	70.402	4956.4	54	0
Health_Perc	460.870	168.103	28258.64	54	0
Leave_Par_House	26.365	3.300	10.8932669	54	0
Poverty	252.556	82.956	6881.69	54	0
House_Overburden	114.759	96.352	9283.7	54	0
Education 0-2	139.278	64.245	4127.41	54	0
Employment	635.481	93.443	8731.54	54	0
Temp_Employment	248.796	154.302	23809.11	54	0
Dep_Symptoms	62.904	22.183	492.0886124	52	2

Table 2 - Summary Statistic of all Variables

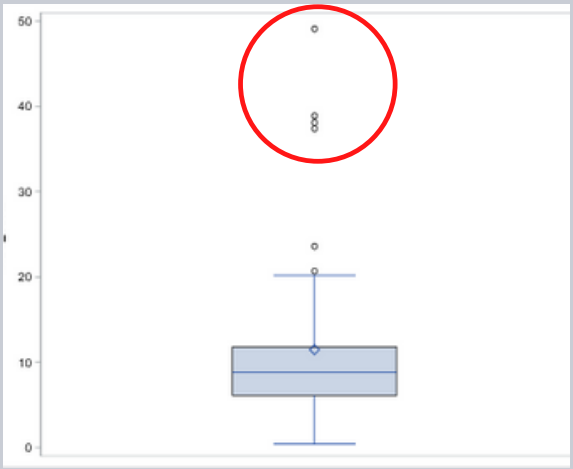
Looking into the summary statistics we realize that there were 2 missing values in the variable "Depression symptoms".

3.2.1 - Data Treatment -Outliers

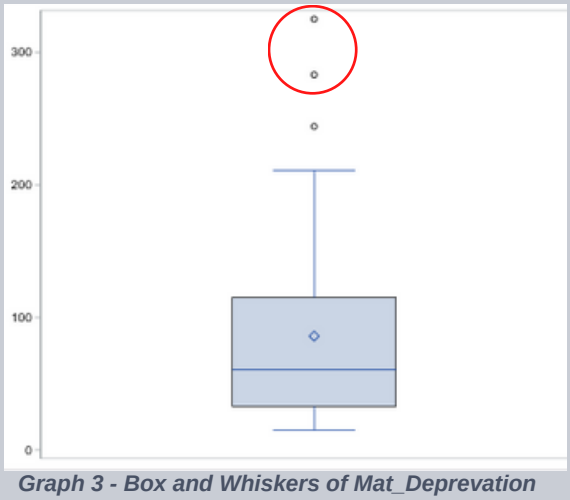
We looked for outliers in the box and whiskers plot (as are showed in the the graphs bellow) and we notice that Greece was an outlier in most of the variables (Material deprivation, Health, Poverty and House overburden), so the decision was taking out Greece. Doing an analysis and performing a cluster analysis we also confirm that Denmark was a outlier because of its really extreme values in the variable House_Overburden, and since its prolongation in the data set affected negatively the analysis, giving a particular cluster to this country and less meaningful factors overall.



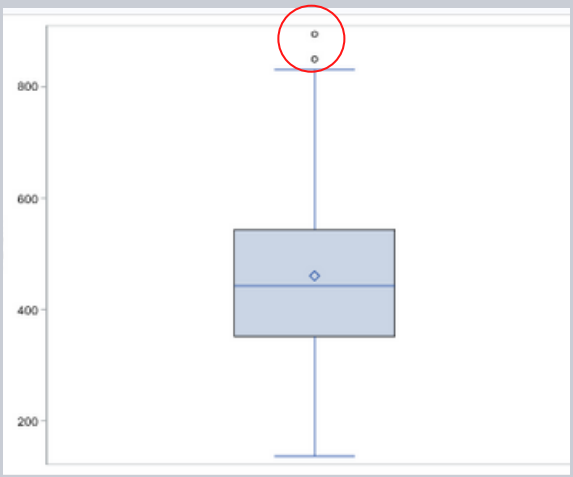
Graph 1 - Box and Whiskers of Poverty



Graph 2 - Box and Whiskers of House_Overburden



Graph 3 - Box and Whiskers of Mat_Deprevation



Graph 4 - Box and Whiskers of Health_Perc

- It is important to emphasize that there were some others "off pattern" countries, with some variables in extreames, but they did not affect the future analysis, so we kept them.

3.2.2 - Data Treatment- Missing Values

With the outliers taken off we can continue and fill in the missing values that we found in the Summary Statistics. To fill this values we used the method to predict the value based on the evolution of each country from 2015 to 2019, in the variable Depressive Symptoms.

The final results were:

- Belgium 2015: 74
- Netherlands 2015: 78

Finally this was the final Summary Statistic of the 9 variables:

	Mean	Std Dev	Variance	N	N Miss
Mat_Deprevation	81.100	65.04	4230.62	50	0
Health_Perc	447.360	152.36	23213.05	50	0
Leave_Par_House	26.464	3.21	10.32	50	0
Poverty	239.420	70.59	4983.02	50	0
House_Overburden	9.124	4.73	22.35	50	0
Education 0-2	137.880	63.91	4085.01	50	0
Employment	641.680	87.01	7571.08	50	0
Temp_Employment	250.880	160.22	25671.09	50	0
Dep_Symptoms	64.120	21.82	475.99	50	0

Table 3 - Summary Statistic of all Variables after treatment

Before continue the forward analysis we notice that all variables had very different variances - Health Perception has a variance of 23213.05, while Leave Parental Householding has a variance of 10.32. Adding that to the fact that Leave Parental Householding is in a different unit than the rest of the variables we had to standardize the data to perform a correct analysis.

3.3 - Factor Analysis

Moving on into the analysis, now we started working with SAS Enterprise Guide, more specific performing Factor Analysis. Factor Analysis is a technique that centres on distinguishing the components that are capable for the correlation between indicators.

Deciding between choosing PCF or PAF, we decided to continue the analysis with PCF, meaning that we are sure that our data does not have much noise, therefore, for our analysis, that we do not have to consider the unique variance. This because we believe that our data comes from reliable, objective and real information sources (like EUROSTAT itself), with small to zero errors in their datasets.

First we checked the Correlation Matrix:

Correlation Matrix

	Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Temp_Employment	Dep_Symptoms
Mat_Deprevation	1								
Health_Perc	0.29713	1							
Leave_Par_House	0.48523	0.24098	1						
Poverty	0.58768	0.08066	-0.02001	1					
House_Overburden	-0.10015	-0.10737	-0.4835	0.42184	1				
Education 0-2	0.12197	-0.30668	0.05343	0.43612	0.29704	1			
Employment	-0.50888	-0.1537	-0.49253	-0.47029	0.08387	-0.23929	1		
Temp_Employment	-0.28162	-0.07144	0.0973	0.18302	0.20981	0.14834	-0.36764	1	
Dep_Symptoms	-0.25174	-0.2974	-0.45653	0.05087	0.36336	0.2709	0.26224	0.19535	1

Table 4 - Correlation Matrix

- Initially we notice 3 highly correlate variables which are Material Deprivation, Poverty and Employment (with values near the 0.5 in absolute terms).

- Then we notice Temporary Employment and Health perception are not strong correlated with the none of the others variables from our dataset, neither between themselves.

- However we can conclude from all correlations values in our dataset that we don´t have a correlation that highlights more than the others.

3.3 - Factor Analysis

After this we wanted to have a better understanding if our data was proper to perform a Factor Analysis, so we checked the KMO matrix (Kaiser-Meyer-Olkin):

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.39511706								
Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Temp_Employment	Dep_Symptoms
0.34123515	0.73461016	0.36391937	0.36928756	0.75350737	0.33188423	0.64144083	0.16197756	0.43314827

Table 5 - KMO with 9 variable

Observing Kaiser-Meyer-Olkin table, we understand we just only have 3 good and adequate factorial models in our dataset (Employment, Health and Housing Overburden), because their values are between (0.6 to 1). We also note that the overall MSA is really bad, near 0.39, so a move has to be done to resolve this problem.

We believe the others factorial models are not the appropriated because the variable Tem_Employment is influencing these variables. So we decide to remove him and only have 8 variables:

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.59903314							
Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Dep_Symptoms
0.58180697	0.68701576	0.55375169	0.47257177	0.69680363	0.5185392	0.7287333	0.7790884

Table 6 - KMO with 8 variables

Removing the variable Temporary Employment, we only have adequate KMO values because Temporary Employment doesn´t satisfied the conditions of Factor Analysis, but now we can continue our analysis, without the variable Temporary Employment it will be easier to understand our final results.

More than that the Overall MSA its better, not the greatest value but it is very close to 0.6. And other values also improve with the removal of the variable. We did notice that Poverty was still very low (close to 0.47) but we decided to keep it, because we consider a good variable to measure our project.

3.3 - Factor Analysis

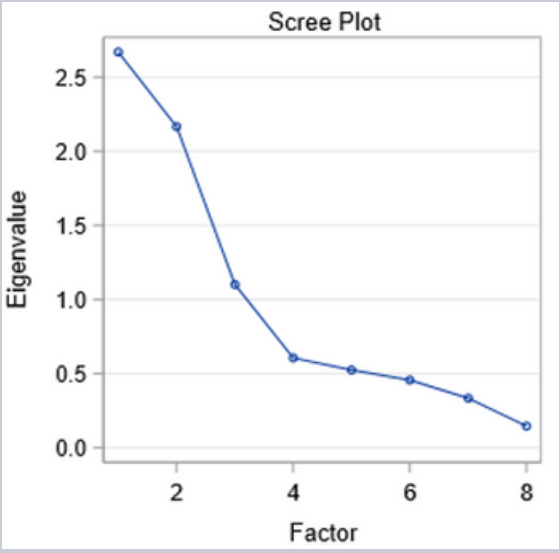
For go further with our analyse, we need decide how many factors we should use. We need to choose the number of factors because it will explain better the correlation between our variables. This decision as based in 3 different criteria:

- Pearson's criteria – cumulative variance should reach 80% or more.
- Kaiser's criteria – eigenvalues should have a value of 1 or higher (with standardized data).
- Scree Plot's criteria – looking to the 'elbow' in the graph.

To answer this we looked for the following SAS outputs:

Eigenvalues of the Correlation Matrix: Total = 8 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.66923983	0.50294139	0.3337	0.3337
2	2.16629844	1.06548173	0.2708	0.6044
3	1.10081671	0.4954384	0.1376	0.742
4	0.60537831	0.08163142	0.0757	0.8177
5	0.5237469	0.06773058	0.0655	0.8832
6	0.45601632	0.12267033	0.057	0.9402
7	0.33334599	0.18818848	0.0417	0.9819
8	0.14515751		0.0181	1

Table 7 - Eigenvalues of the Correlation Matrix



Graph 5 - Scree Plot

Regarding Kaiser criteria, we should keep the first 3 factors because their eigenvalues are greater than one.

According Pearson criteria, we should preserve the first 3 or 4 factors because their cumulative variance reaches the 80% in fourth factor, but is very close with the third factor.

Finally in Scree Plot, we should use 4 factors because it starts the "elbow" in our graph.

3.3 - Factor Analysis

Now the question was choosing between keeping 3 or 4 factors. For this we looked into the Factor Pattern, the Variance Explain by each one, and Root Mean Square Of-Diagonal Residuals (these tables were also taken off the outputs of SAS).

First we saw the Root Mean Square Of -Diagonal Residuals:

Root Mean Square Off-Diagonal Residuals: Overall = 0.08343442							
Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Dep_Symptoms
0.08605075	0.10148139	0.07075492	0.05815157	0.08095689	0.08844584	0.07931144	0.0945112

Table 8 - Root Mean Square Of -Diagonal Residuals for 3 factor solution

Root Mean Square Off-Diagonal Residuals: Overall = 0.07420178							
Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Dep_Symptoms
0.08339543	0.08694849	0.06454036	0.05437386	0.07098631	0.08783944	0.07950097	0.05775113

Table 9 - Root Mean Square Of -Diagonal Residuals for 4 factor solution

As expected the residuals with 4 factors are smaller than with 3 factors, but we can easily notice that the difference between the 2 tables are very smaller.

First, overall the difference was about 0.01, very smaller to consider a 4 factors only for this. Then, the only variables that had their residuals reduced were Health Perception and Depressive Symptoms (this one with a bigger difference, of 0.04). The others variables kept the same, or with a very small decrease.

3.3 - Factor Analysis

After that we looked into the Factor Explain and the Variance Explain by each Factor:

Factor Pattern			
	Factor1	Factor2	Factor3
Poverty	0.84618	0.38935	-0.03776
Mat_Deprevation	0.80433	-0.21043	0.21211
Employment	-0.75634	0.32711	0.05878
House_Overburden	0.2058	0.84706	-0.08532
Dep_Symptoms	-0.16716	0.58814	-0.42865
Leave_Par_House	0.40729	-0.7977	0.03175
Health_Perc	0.249	-0.0562	0.83452
Education 0-2	0.46837	0.17195	-0.71423

Table 9 - Factor Patterns with 3 factor solutions

Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
Mat_Deprevation	0.79708	0.28739	0.13521	0.17141
Health_Perc	0.78126	-0.24941	-0.36143	0.1577
Leave_Par_House	-0.75024	-0.30707	0.15927	0.10914
Poverty	0.41098	0.79993	0.24547	-0.09195
House_Overburden	0.05408	0.73005	-0.47242	0.07607
Education 0-2	-0.35654	0.69632	0.3939	-0.23365
Employment	0.46178	-0.25404	0.69556	0.29727
Dep_Symptoms	-0.59033	0.45272	-0.06438	0.61807

Table 10 - Factor Patterns with 4 factor solutions

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.4525719	2.0353257	1.4484574

Table 11 - Variance Explain with 3 factor solutions

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
2.6692398	2.1662984	1.1008167	0.6053783

Table 12 - Variance Explain with 4 factor solutions

In the table of Factor Pattern we can see that the composition of each factor changed by adding one more factor, both tables are good to perform the analysis but we didn't want to have a unique factor to Employment and another one only for Depressive Symptoms. Having 2 factors only for 2 different variables wasn't a great way to simplify the analysis. So this was the principal reason that lead us to choose 3 factor solutions over 4.

We then proceeded to label the 3 factors, according to their composition:

→ Factor 1: Bad Living Conditions - characterized by high levels of Poverty and Material Deprivation, and low levels of Employment.

3.3 - Factor Analysis

- Factor 2: Rent Overburden - characterized by high levels of House Cost Overburden and Depressive Symptoms, and low levels of Leave the Parental Household

- Factor 3: Health and Education - characterized by high levels of Health Perception and low levels of Low Education Attainment (under or equal high school).

The previous results were obtain using Varimax as the factor rotation technique. But why?

After trying both rotations (Varimax and Quartimax) we notice that the difference between the two was too small to compare both. Since the factor solution was the same when applied both Varimax and Quartimax rotation, we proceeded using Varimax.

So the last to see its the Final Comunalities table:

Final Communality Estimates: Total = 5.936355							
Mat_Deprevation	Health_Perc	Leave_Par_House	Poverty	House_Overburden	Education 0-2	Employment	Dep_Symptoms
0.73621493	0.76158088	0.80320722	0.86904118	0.76714198	0.75907075	0.68251009	0.55758796

Table 13 - Final Communalities Estiamte

As seen in the Final Communalities table, we can observe overall the 3 factors that we got explain all the variables. The 3 finals factors that we got explain was close to 74.125 % of the total variance of the 8 original variables.

Before moving on, we need to mention one thing, as said overall the communalities are great, all above 0.70 or very close to it, as the case of Employment. But the same doesn't happen with depressive symptoms, its values are close to 0.55, which is not bad at all but isn't great either. We decided to keep it because we consider Depressive Symptoms really important for the analysis of the project and to get a good result.

3.4 - Cluster Analysis

After Factor Analysis, we decide to perform Cluster Analysis in order to obtain groups of Countries that are similar to each other with respect to their characteristics.

Cluster Analysis is used for combining observations into groups such that observations in each group are similar to each other and observations of one group are different from the observations of other groups.

Our idea was to apply Cluster Analysis to both standardized data and obtained factors, because comparing their results we can understand if factors are a good representation of the dataset. We didn't use non-standardized data because, also in this technique, the variables that are presented in higher scales may override the effects of other variables with smaller scales since we have to use a distance measure.

The similarity measure that we choose is the Squared-Euclidean Distance since all the variables that we use are metric-variables.

3.4.1 - Variable's clusters - hierarchical method

We start performing hierarchical method on standardized data.
There are various hierarchical methods to choose from that differ in how they calculate the distances. The methods that we apply are 5:

- Centroid
- Single
- Complete
- Average
- Ward

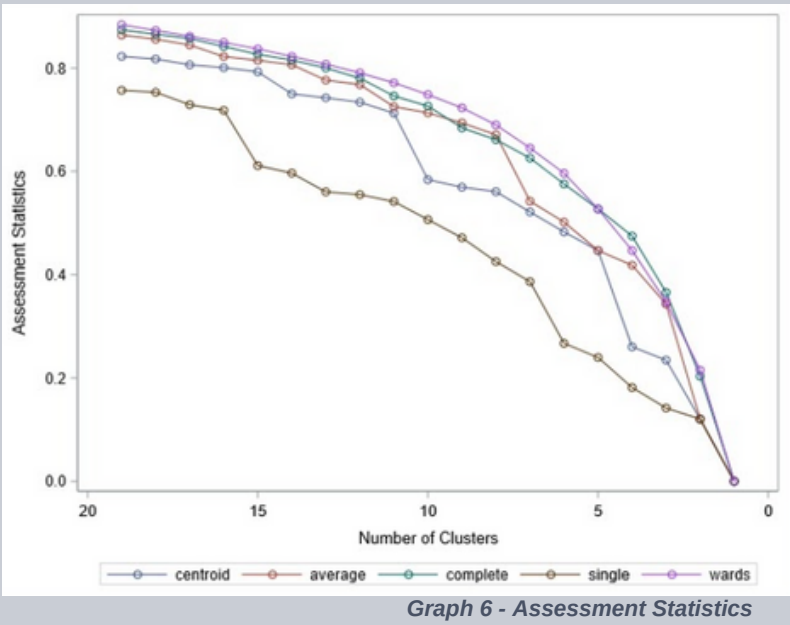
We decided which method was better based on statistics as Root-mean-squared standard deviation, R-Squared and Semi-Partial RSQ, giving more weight to R-Square values, because R-Squared measures the extent to which clusters are different from each other and it can be interpreted as a measure of the proportion of the total variance that is retained in each of the solutions, so we want it to be as close to 1 as possible.

3.4.1 - Variable's clusters - hierarchical method

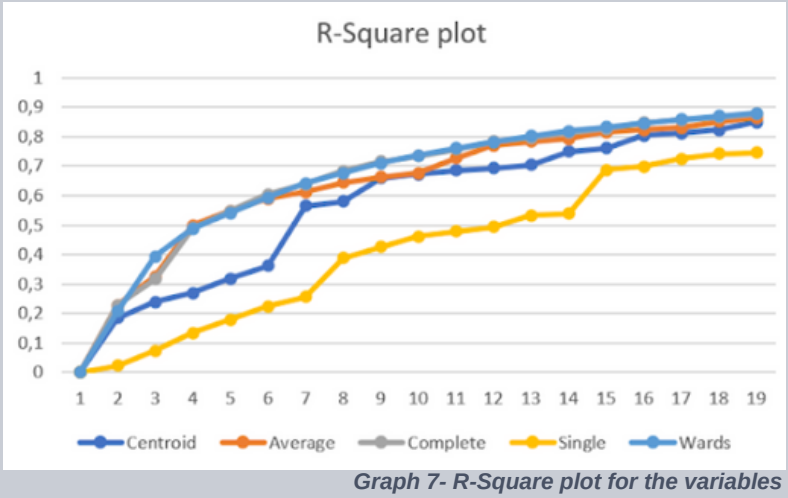
We decided which method was better based on statistics as Root-mean-squared standard deviation, R-Squared and Semi-Partial RSQ, giving more weight to R-Square values, because R-Squared measures the extent to which clusters are different from each other and it can be interpreted as a measure of the proportion of the total variance that is retained in each of the solutions, so we want it to be as close to 1 as possible.

N_Clusters	Centroid	Average	Complete	Single	Wards
1	0.00	0.00	0.00	0.00	0.00
2	0.18	0.23	0.23	0.02	0.21
3	0.24	0.33	0.32	0.07	0.39
4	0.27	0.50	0.49	0.13	0.49
5	0.32	0.55	0.55	0.18	0.54
6	0.36	0.59	0.60	0.23	0.59
7	0.57	0.61	0.64	0.26	0.64
8	0.58	0.64	0.68	0.39	0.68
9	0.66	0.66	0.71	0.43	0.71
10	0.67	0.68	0.73	0.46	0.74
11	0.69	0.73	0.76	0.48	0.76
12	0.69	0.77	0.78	0.49	0.78
13	0.70	0.78	0.80	0.53	0.80
14	0.75	0.79	0.81	0.54	0.82
15	0.76	0.82	0.83	0.69	0.83
16	0.80	0.82	0.85	0.70	0.85
17	0.81	0.83	0.86	0.72	0.86
18	0.82	0.85	0.87	0.74	0.87
19	0.85	0.86	0.88	0.75	0.88

Table 14 - Cluster history for the variable



Graph 6 - Assessment Statistics for the variables



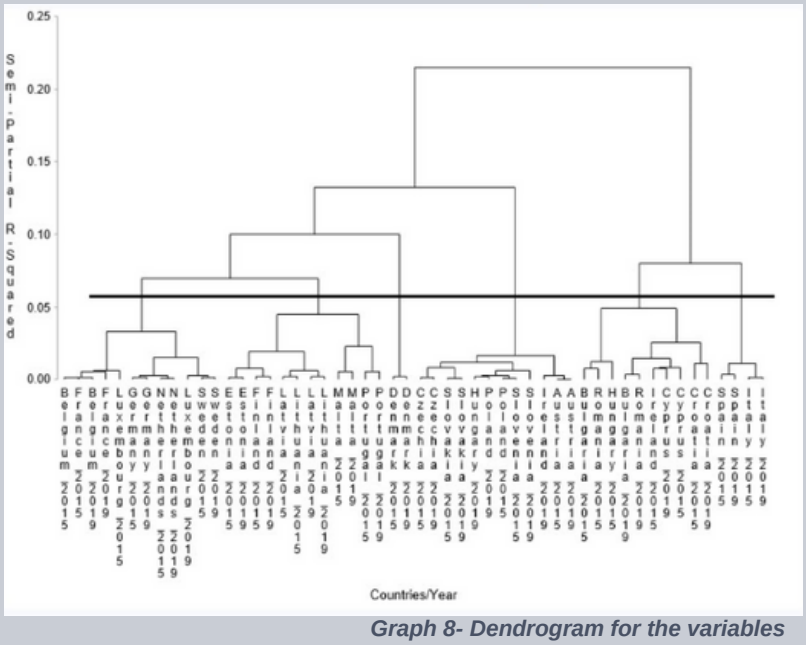
Graph 7- R-Square plot for the variables

3.4.1 - Variable's clusters - hierarchical method

Analysing all statistics, we have seen that the best methods for this data are Complete and Ward, followed by Average and Centroid, to end with the worst method which is the Single linkage method. Between Ward and Complete, we chose to use Ward for our analysis as it returns values for R-Square slightly higher for most cases.

Ward's method forms clusters by maximizing within-cluster homogeneity, using the within-group sum of squares as a measure of homogeneity. Ward's method tends to find clusters that are compact and nearly of equal size and shape.

The next step, after choosing the method to use, was to define the number of clusters. We did it first by looking at the "best cut" in the dendrogram and then analysing it.



Graph 8- Dendrogram for the variables

Looking at the dendrogram, in our opinion, the best cut was for a Semi-Partial R-Square value between 0.05 and 0.07. In this way we obtain 6 clusters.

3.4.1 - Variable's clusters - hierarchical method

N° Clusters	RMSSTD	SPRSQ	RSQ
10	0.82	0.02	0.74
9	0.64	0.03	0.71
8	0.62	0.03	0.68
7	0.71	0.03	0.64
6	0.83	0.05	0.59
5	0.79	0.05	0.54
4	0.72	0.05	0.49
3	0.83	0.09	0.39
2	0.91	0.19	0.21
1	1.00	0.21	0.00

Table 15 - Wards output or the variable

The statistics that we used, as before, are Root-mean-squared standard deviation (RMSSTD), R-Squared (RSQ) and Semi-Partial RSQ (SPRSQ). We want:

- An high value for RSQ (for the reasons mentioned above).
- A low value for RMSSTD, since the smaller the value, the more homogenous the observations with respect to the variables.
- A low value for SPRSQ because it represents the loss of homogeneity.

- In Graph 6 - Assessment statistics we don't have a real elbow, but looking at the table we can see that as the number of clusters increases, the statistics improve, but, at the same time, we don't want too many clusters because they would probably be difficult to interpret. So, looking at the statistics, we decide to consider 7 clusters, but 6 cluster was also a big possibility, even though its high value for RMSSTS. After some discussion and a final agreement we agree in using only 6 clusters.

3.4.2 - Variable's clusters - non-hierarchical method

Often hierarchical methods are used in an exploratory analysis, to perform, later, non-hierarchical methods like k-means that usually have better performance. The disadvantage of k-means is that the number of clusters must be known a priori, so we can use the hierarchical clustering centroid as seeds of k-means.

3.4.2 - Variable's clusters - non-hierarchical method

We run the K-Means algorithm using both 6 number of clusters and we obtain the following results about the composition of each cluster and that the principal statistics of each:

Cluster Number	Countries/Year				
1	Czechia_2015	Czechia_2019	Ireland_2015	Ireland_2019	Croatia_2015
	Croatia_2019	Cyprus_2019	Hungary_2019	Austria_2015	Austria_2019
	Poland_2015	Poland_2019	Slovenia_2015	Slovenia_2019	Slovakia_2015
	Slovakia_2019				
2	Estonia_2019	Finland_2015	Finland_2019	Germany_2015	Germany_2019
	Luxembourg_2019	Netherlands_2015	Netherlands_2019	Sweden_2015	Sweden_2019
3	Belgium_2015	Belgium_2019	Estonia_2015	France_2015	France_2019
	Luxembourg_2015	Portugal_2015	Portugal_2019		
4	Bulgaria_2015	Bulgaria_2019	Cyprus_2015	Hungary_2015	Spain_2015
	Spain_2019	Italy_2015	Italy_2019	Romania_2015	Romania_2019
5	Latvia_2015	Latvia_2019	Lithuania_2015	Lithuania_2019	Malta_2015
	Malta_2019				

Table 16 - Cluster Compostions for the variables

Cluster 1				
Variable	N	Mean	Median	Variance
Mat_Deprevation	16	-0.212	-0.217	0.300
Health_Perc	16	0.716	0.756	0.404
Leave_Par_House	16	0.464	0.260	0.445
Poverty	16	-0.670	-0.679	0.706
House_Overburden	16	-0.499	-0.618	0.274
Education 0-2	16	-0.926	-0.906	0.167
Employment	16	0.061	0.119	0.618
Dep_Symptoms	16	-0.662	-0.601	0.567

Table 17 - Cluster 1 statistics (variables)

Cluster 2				
Variable	N	Mean	Median	Variance
Mat_Deprevation	10	-0.793	-0.847	0.031
Health_Perc	10	-0.517	-0.590	0.178
Leave_Par_House	10	-1.439	-1.374	0.412
Poverty	10	0.331	0.462	0.254
House_Overburden	10	1.264	1.264	1.207
Education 0-2	10	0.068	0.018	0.193
Employment	10	0.845	0.797	0.242
Dep_Symptoms	10	0.820	0.911	0.532

Table 18 - Cluster 2 statistics (variables)

3.4.2 - Variable's clusters - non-hierarchical method

Cluster 3				
Variable	N	Mean	Median	Variance
Mat_Deprevation	8	-0.394	-0.524	0.169
Health_Perc	8	-0.394	-0.075	0.793
Leave_Par_House	8	-0.366	-0.627	0.544
Poverty	8	-0.316	-0.381	0.171
House_Overburden	8	-0.018	0.027	0.217
Education 0-2	8	0.569	0.315	0.651
Employment	8	-0.226	-0.232	0.232
Dep_Symptoms	8	0.923	0.888	0.416

Table 19 - Cluster 3 statistics (variables)

Cluster 4				
Variable	N	Mean	Median	Variance
Mat_Deprevation	10	1.358	1.305	1.584
Health_Perc	10	0.426	0.283	1.109
Leave_Par_House	10	0.768	0.743	0.090
Poverty	10	1.377	1.404	0.322
House_Overburden	10	0.209	0.175	0.435
Education 0-2	10	0.939	0.823	1.093
Employment	10	-1.262	-0.962	0.633
Dep_Symptoms	10	-0.340	-0.785	0.861

Table 20 - Cluster 4 statistics (variables)

Cluster 5				
Variable	N	Mean	Median	Variance
Mat_Deprevation	6	0.150	0.168	0.393
Health_Perc	6	-1.233	-1.328	0.550
Leave_Par_House	6	0.369	0.182	0.410
Poverty	6	-0.639	-0.686	0.310
House_Overburden	6	-1.098	-0.946	0.192
Education 0-2	6	0.033	-0.186	1.037
Employment	6	0.833	0.665	0.509
Dep_Symptoms	6	-0.265	-0.326	0.588

Table 21 - Cluster 5 statistics (variables)

3.4.2 - Variable's clusters - non-hierarchical method

So with this table we were able to finally classify the clusters:

- **Cluster 1** is the one with most observations (16) and contains observations with high negative values for Education 0-2, Poverty, House_Overburden and Depressive_Symptoms and high positive values for Health;

- **Cluster 2** has 10 observations and it's characterized by really high positive values on average for House_Overburden, Employemnt and Dep_Symptoms, while has high negative values for Mat_Deprevation and Leave_Par_House

- **Cluster 3** contains 8 observations that have high values for Dep_Symptoms, and low or neutral values for all the others variables;

- **Cluster 4** has 10 observations with on average high negative values for the variable Employment and high positive values for Mat_Deprevation, Health_Perc, Leave_Par_House, Poverty and Education 0-2;

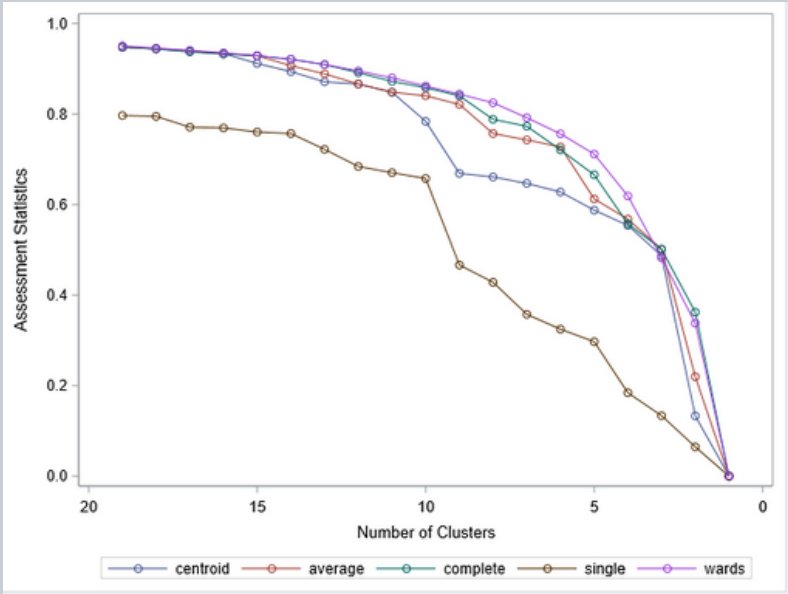
- **Cluster 5** is composed by 6 observations characterized by really high positive values for the variable Employmnet, but also high negative values for the variable Health_Perc, Poverty and House_Overburden;

3.4.3 - Factor's clusters - hierarchical method

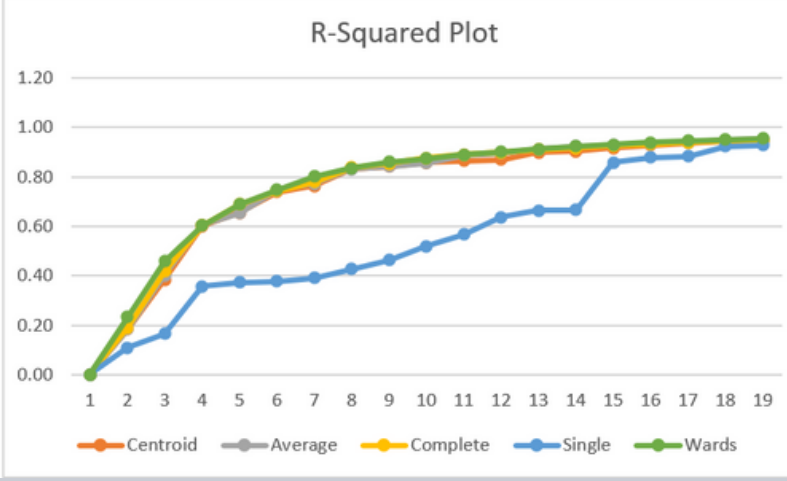
The previous cluster analysis was made for the standardized variables. Now, we are going to do the cluster analysis for the factors that we obtained from Factor's Analysis to compare to the previous results and see if they are similar.

N_Clusters	Centroid	Average	Complete	Single	Wards
1	0.00	0.00	0.00	0.00	0.00
2	0.18	0.18	0.19	0.11	0.23
3	0.38	0.40	0.42	0.17	0.46
4	0.60	0.61	0.60	0.36	0.60
5	0.65	0.65	0.69	0.37	0.69
6	0.74	0.75	0.74	0.38	0.75
7	0.76	0.77	0.78	0.39	0.80
8	0.83	0.83	0.84	0.43	0.84
9	0.84	0.84	0.85	0.46	0.86
10	0.86	0.86	0.88	0.52	0.88
11	0.87	0.89	0.89	0.57	0.89
12	0.87	0.90	0.90	0.64	0.90
13	0.90	0.91	0.91	0.66	0.91
14	0.90	0.92	0.92	0.67	0.92
15	0.92	0.93	0.93	0.86	0.93
16	0.93	0.94	0.93	0.88	0.94
17	0.94	0.94	0.94	0.88	0.95
18	0.95	0.95	0.95	0.92	0.95
19	0.95	0.95	0.95	0.93	0.95

Table 22 - Cluster history for the factors



Graph 9 - Assessment Statistics for the factors

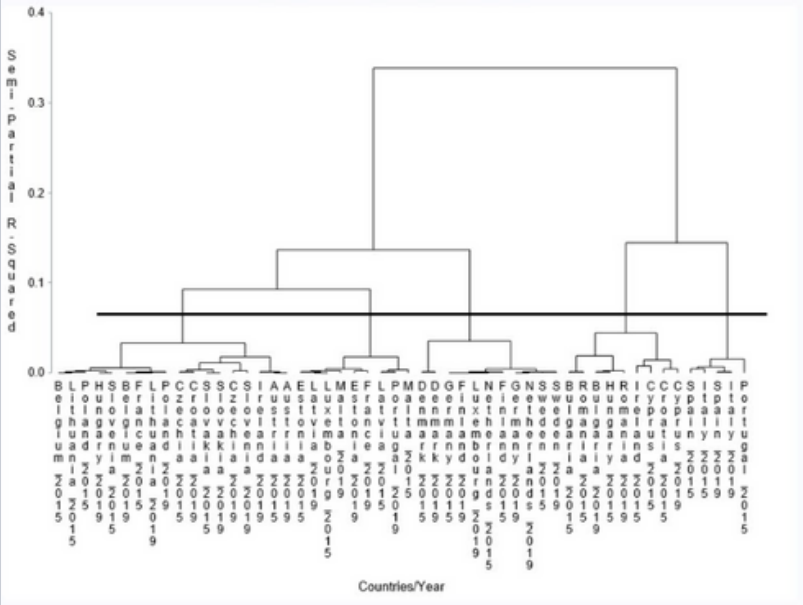


Graph 10 - R-Square plot for the factors

3.4.3 - Factor's clusters - hierarchical method

Analysing all statistics, we have seen that the best methods for this data are Ward, followed by Average, then by Centroid and Complete, and to end with the worst method which is the Single linkage method. We used Ward for our analysis as it returns values for R-Square slightly higher for most cases.

The next step, after choosing the method to use, was to define the number of clusters. We did it first by looking at the "best cut" in the dendrogram and then analysing it.



Graph 11 - Dendrogram for the factors

Looking at the dendrogram, in our opinion the "best cut" was the one that gives us 5 clusters. To be sure in choosing the number of clusters, we have also analysed the statistics as RMSSTD, SPRSQ and RSQ.
As before, we want high values for RSQ and low values for the other 2.

3.4.3 - Factor's clusters - hierarchical method

	RMSSTD	SPRSQ	RSQ
10	0.54	0.01	0.88
9	0.40	0.02	0.86
8	0.59	0.02	0.84
7	0.53	0.03	0.80
6	0.82	0.05	0.75
5	0.55	0.06	0.69
4	0.82	0.09	0.60
3	0.83	0.14	0.46
2	0.90	0.23	0.23
1	1.00	0.23	0.00

Table 23 - Wards output for the factors

After some consideration we consider choosing between 7 or 5 clusters (since with 6 the RMSSTD was one of the highest, we didn't consider). After some discussion we decided to stay with 5 clusters in order to have a better interpretability.

3.4.4 - Factor's clusters - non-hierarchical method

As before we use the centroid of the hierarchical method just performed as seeds for the k-means algorithm. We obtain the clusters below:

Cluster Number	Countries/Year				
1	Belgium_2015	Belgium_2019	Estonia_2015	Estonia_2019	France_2015
	France_2019	Luxembourg_2015	Finland_2015	Finland_2019	
2	Czechia_2015	Czechia_2019	Ireland_2019	Croatia_2019	Cyprus_2019
	Lithuania_2015	Lithuania_2019	Hungary_2019	Austria_2015	Austria_2019
	Poland_2015	Poland_2019	Slovenia_2015	Slovenia_2019	Slovakia_2015
	Slovakia_2019				
3	Latvia_2015	Latvia_2019	Malta_2015	Malta_2019	Portugal_2015
	Portugal_2019	Spain_2015	Spain_2019	Italy_2015	Italy_2019
4	Germany_2015	Germany_2019	Luxembourg_2019	Netherlands_2015	Netherlands_2019
	Sweden_2015	Sweden_2019			
5	Bulgaria_2015	Bulgaria_2019	Ireland_2015	Croatia_2015	Cyprus_2015
	Hungary_2015	Romania_2015	Romania_2019		

Table 24 - Cluster's Compostion for the factors

3.4.4 - Factor's clusters - non-hierarchical method

After performing the k-means algorithm, we analysed the variables statistics of each cluster to give an explanation to each group.

Cluster 1				
Factor	N	Mean	Median	Variance
Bad Living Conditions	9	-0.487	-0.406	0.201
Rent Overburden	9	0.434	0.471	0.078
Health and Education	9	-0.304	-0.136	0.198

Table 25 - Cluster 1 statistics (factors)

Cluster 2				
Factor	N	Mean	Median	Variance
Bad Living Conditions	16	-0.559	-0.544	0.167
Rent Overburden	16	-0.611	-0.656	0.294
Health and Education	16	0.666	0.553	0.398

Table 26 - Cluster 2 statistics (factors)

Cluster 3				
Factor	N	Mean	Median	Variance
Bad Living Conditions	10	0.435	0.402	1.523
Rent Overburden	10	-0.748	-0.904	0.227
Health and Education	10	-1.413	-1.154	0.025

Table 27 - Cluster 3 statistics (factors)

Cluster 4				
Factor	N	Mean	Median	Variance
Bad Living Conditions	7	-0.361	-0.439	0.068
Rent Overburden	7	1.901	1.891	0.081
Health and Education	7	-0.151	-0.175	0.025

Table 28 - Cluster 4 statistics (factors)

Cluster 5				
Factor	N	Mean	Median	Variance
Bad Living Conditions	8	1.438	1.233	0.612
Rent Overburden	8	0.006	0.065	0.410
Health and Education	8	0.909	0.744	0.635

Table 29 - Cluster 5 statistics (factors)

3.4.4 - Factor's clusters - non-hierarchical method

So with this table we were able to finally classify the clusters:

- **Cluster 1** has 9 observations with high positive values for the variable *Bad Living Conditions* and medium/neutral values for other 2 factors;
- **Cluster 2** contains the majority of the observations; this 16 observations are characterized by high negative *Bad Living Conditions* (meaning Good Living Conditions), medium *Rent Overburden* and high values for the factor *Health and Education*;
- **Cluster 3** is formed by 10 observations with high negative values for *Health and Education* and for *Rent Overburden,* and medium/neutral values for the other factor;
- **Cluster 4** contains 7 items with really high positive values (on average) for the factor *Rent Overburden* and medium/neutral values for the other factors;
- **Cluster 5** has 8 items and scores with high positive values (on average) for the factors *Bad Living Conditions* and *Health and Education* and medium/neutral values for *Rent Overburden;*

3.4.5 - Differences between both Cluster Analysis

- After performing the cluster analysis of both variables and factors, our observations were grouped into 5 clusters. What we did was find relationship between the clusters formed by the two analyses. Though they are the same number of clusters, their composition and statistics vary way to much to conclude that the factors are a good representation of the dataset. The two clustering obtained are not exactly the same, but we noted some similarity between the two results.

From now on, we will call k-F the k-th cluster obtained by making the k-means on the factors and k-V the k-th cluster obtained by performing the k-means on the variables.

3.4.5 - Differences between both Cluster Analysis

Cluster 2-V, contains countries that are inside clusters 1-F and 4-F. We can see also that characteristics of 2-V and 4-F are very similar, in fact cluster 2-V it's characterized by high value on average for Hou_Burd and Dep_Syn, while has low values for Par_House; cluster 4-F, instead, contains items with high value for the factor Independence, that mean that has exactly the same characteristics since the factor is formed by that variables.

Cluster 4-V contains the majority of countries contained by 5-F, plus some countries inside the cluster 3-F. Also here, values on average of 4-V and 5-F are pretty similar, in fact they shared low values for Employ and high values for Mat_Dep, Health and Poverty .

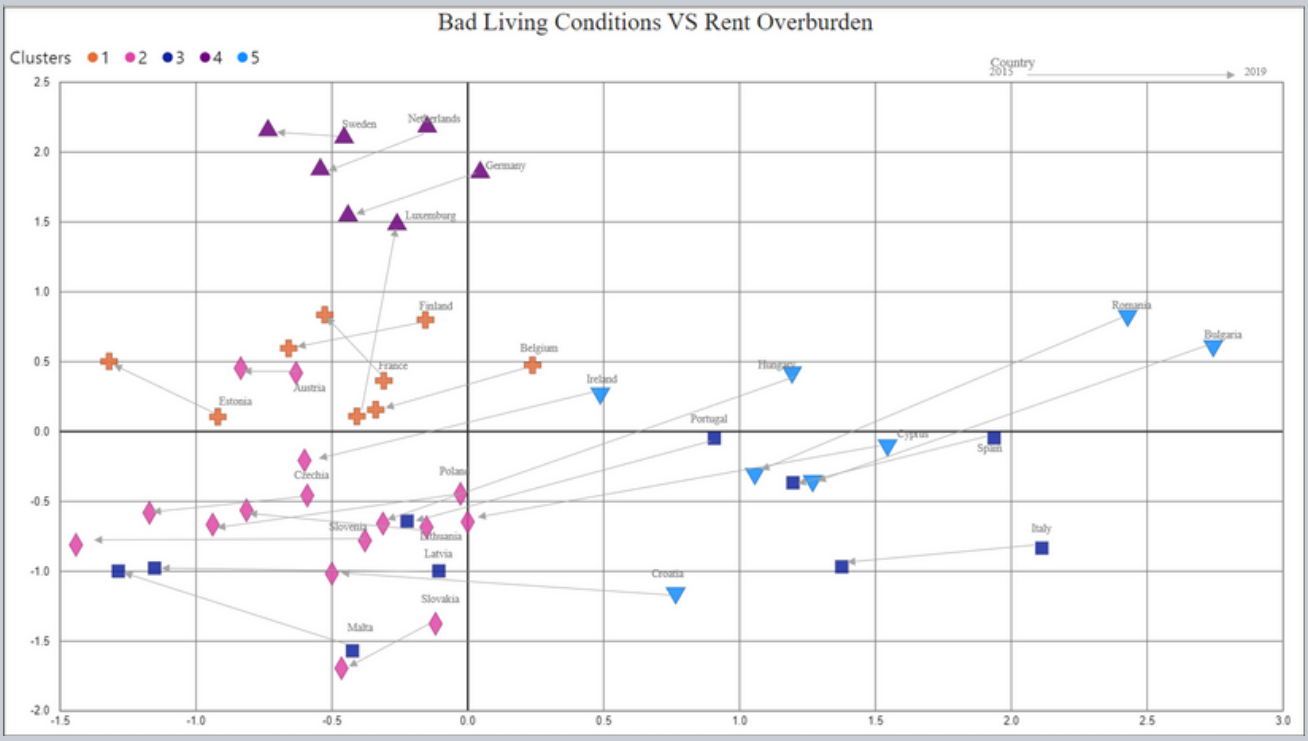
For the rest there are no other close similarities between the clusters obtained through the two analyses.

In both analyses, Portugal_2015 and Portugal_2019 are within the same cluster. In the first case the cluster containing our country is characterized by high values for Dep_Syn that is not a good sign, while in the other analysis is contained by a cluster with low values for Health and high values for Bad Education.

4-Results and Discussion

- We are now getting to the final evaluations about the analysis performed along the hole report. These next steps will conclude our insights and thoughts.

4.1-Factors and Clusters Interpretation



Analysing the particular situation of each cluster on each factor, we can draw some interesting thoughts about it.

The first graph has *Bad Living Conditions* as the X axis and *Rent Overburden* as the Y axis. A special note on the fact that, when evaluating the position of the variables relating to the X axis, the main goal would be to score as negative as possible, since we are working with an already negative factor. There was the goal to score as negative as possible on the Y axis, as well.

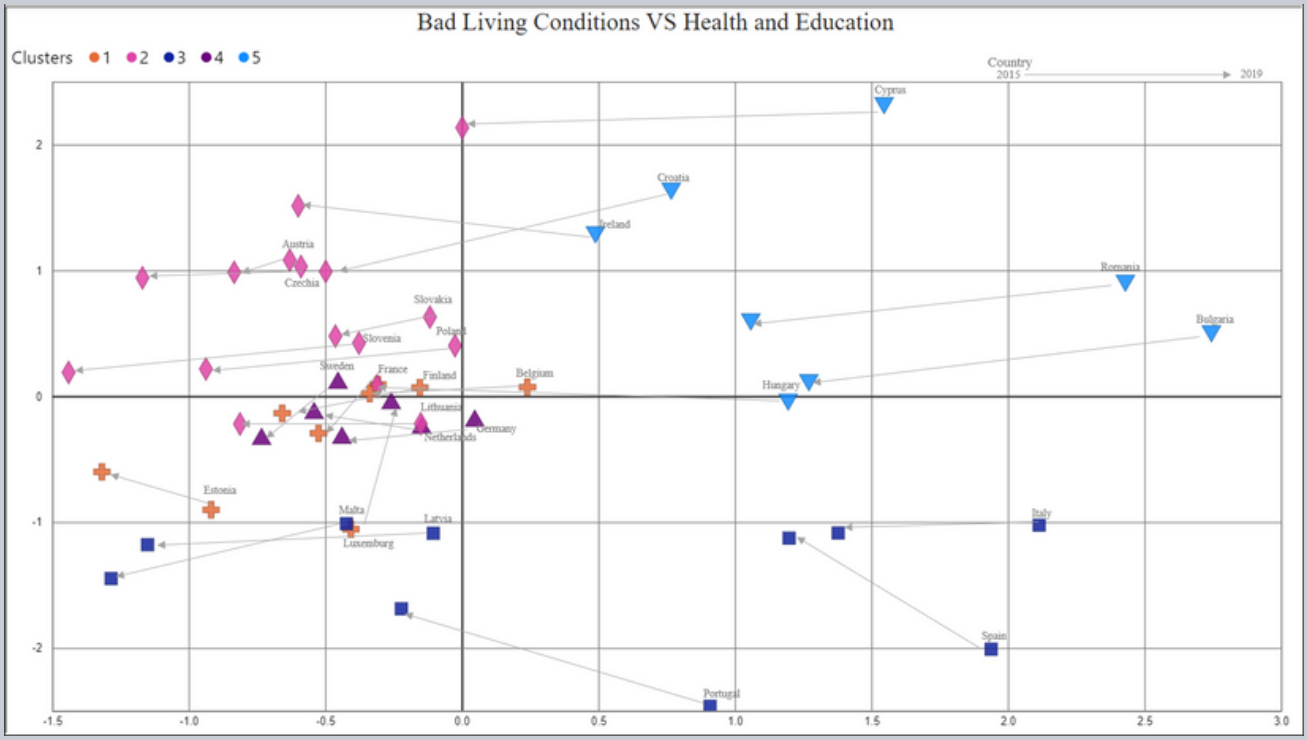
According to these insights, we can easily note that the Cluster 5 and Cluster 4 are the less well placed, since the countries on Cluster 5 have (apparently) *Bad Living Conditions* and the ones on the Cluster 4 have a high *Rent Overburden*.

The Cluster 2 is the cluster that is closer to the desired placement on this graph (explained before), meaning that these countries would have good *Living Conditions* and a low *Rent Overburden*.

It is also interesting to note the general tendency to decrease the *Bad Living Conditions* score, meaning that the overall panorama is that the Living Conditions are getting better. It is not easy to draw any special conclusion about the evolution of the *Rent Overburden* between the years in study.

Finishing with a brief analysis about Portugal, it shows a positive tendency, even though the position in 2015 was not the best - with poorest *Living Conditions* than desired - *Rent Overburden* was slightly bellow the average and it moved both factors on the desired direction.

4.1-Factors and Clusters Interpretation



The desired position in this graph is, once again, to score as low (negative) as possible on the X axis. When evaluating according to the Y axis - Health and Education -, the desired position is as high (positive) as possible.

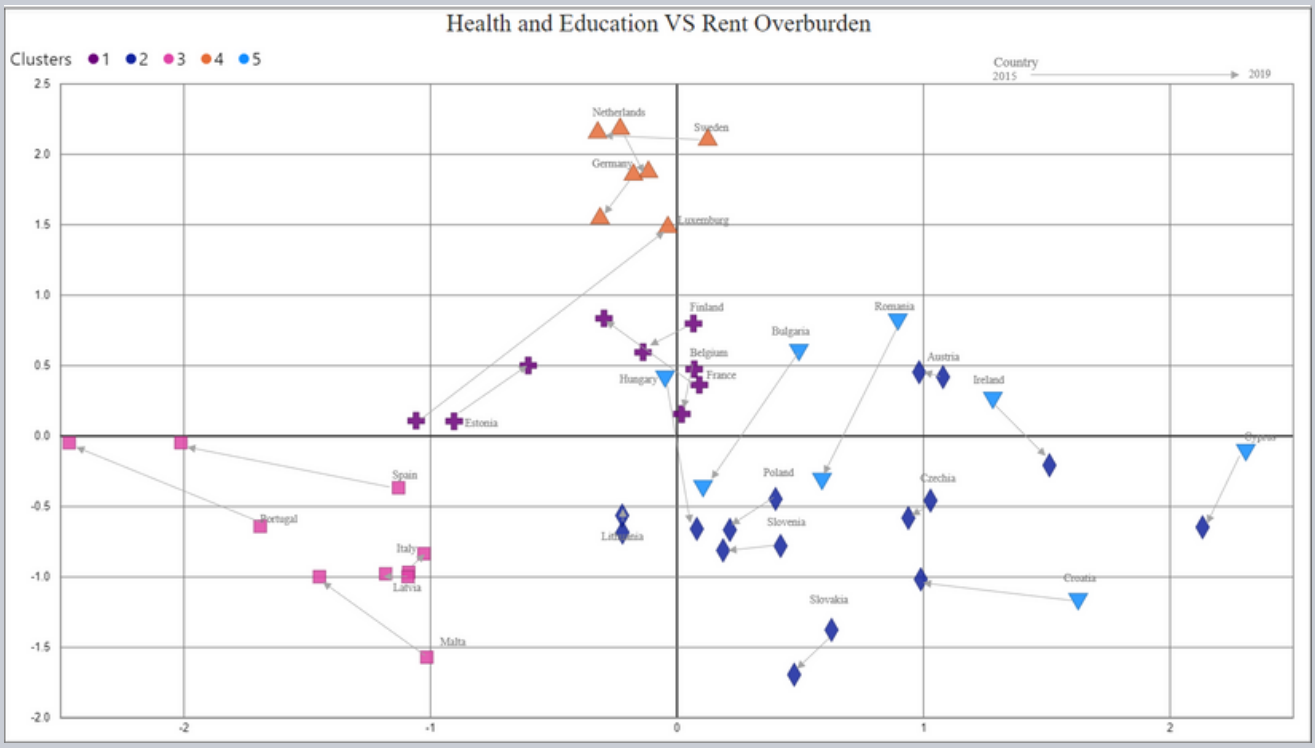
So, we can note that the countries that are poorest positioned are Italy (on both years), Spain (on both years) and Portugal (on 2015). Even knowing that these variables all belong to Cluster 3, it is not correct to say that this cluster is the worst positioned, because it also has countries on a not so undesired position. However, as a Cluster itself, Cluster 5 deserves a note, also, because even though all the variables score higher than the average on Health and Education, we can affirm that they also all score positive on Bad Living Conditions, whish is the opposite to desired, as explained on the previous graph.

The best scored cluster would be, again, Cluster 2: has seen before is in the desired position about Living Conditions and is also above the average on Health and Education (except for a couple of variables: Lithuania on both years).

There is not a general tendency about Health and Education, but it would be preferable if we could see an increasing trend.

A note about Portugal was already made, but it is important to highlight a tendency on the desired direction.

4.1-Factors and Clusters Interpretation



The desired position in this graph is to score as high (positive) as possible on the X axis - Health and Education - and as low (negative) as possible on the Y axis - Rent Overburden.

Cluster 3 would be the worst positioned when concerning about Health and Education, even though all the variables are bellow the average on Rent Overburden. Clusters 1 and 4 also show (overall) Health and Education access bellow the average and they also have a heavy (or at least above the average values of) Rent Overburden.

Clusters 2 and 5 show the best positions on this graph: Cluster 5 includes Hungary_2015 with poorest Health and Education values compared to the average and also highest Rent Overburden than the average, along with Bulgaria_2015, Romania_2015 and Ireland_2015; and Cluster 2 also counts with a couple countries in non desirable positions - Austria has highest Rent Overburden than the average and Lithuania was already mention for having worst Health and Education than the average.

On this graph, Health and Education appear to show a decreasing tendency, what is not desirable. There are some exceptions to this trend, the biggest is Luxembourg.

5 - Conclusion

Finally, after all the analysis and taking into account our variables and the points we consider important when it comes to what is the best country to spend our young years, remembering that we give importance to variables such as health, poverty and social exclusion, house cost overburden, education and others, we decided that the country whose had showed the best potential was Czechia in 2019 and also in 2015.

Czechia scores were high as we wanted, being Bad living conditions and Rent Overburden negative, and Health and Education positive (as we already mention on the factor and cluster analysis that were the goals). Czechia didn't score the highest in all factors or variables, but overall was the one with best result accordingly our criteria

It is important to notice that we were not expecting at first though that Czechia would be our final choice, but considering our variables it makes sense to not be the countries that we assumed first when we thought about the theme (like very developed countries for eg. Germany, France, etc). But this result can be pretty much explain by the factor Rent Overburden, countries well developed, as mentioned, also score really high in this factor as expected.

5 - Conclusion

It is also important to consider that Croatia, for the of 2019, also had great scores but we did not considered the best country because of our priorities: we valorise good living conditions over not having the rent overburdening the net income of young adult.

Austria, both in 2015 and 2019, also had obtained great scores but the rent overburden has showed positive, its important to emphasizing countries like that, because they are good examples of the subjectivity of this theme (the results can be different considering every person priorities). Overall if a person prefers to have its "independence", can leave earlier the the parental household, resulting in housing costs that represents more than 40 % of disposable income.

In conclusion, however the results had been different from what we expected, we can recommend based on our analysis, to a person who values the same factors as we, that Czechia and Croatia are the better options when you are looking for a country which can give you some easy economic independence, good values of health and education.


Croatia's flag


Austria's flag

6 - References

- Eurostat. (2021, October 13). Retrieved from Young people by educational attainment level, sex and age: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=yth_demo_040&lang=en

- Eurostat. (2021, December 17). Retrieved from Severe material deprivation rate by age and sex: https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-127819_QID_6219B722_UID_-3F171EB0&layout=TIME,C,X,0;GEO,L,Y,0;UNIT,L,Z,0;AGE,L,Z,1;SEX,L,Z,2;INDICATORS,C,Z,3;&zSelection=DS-127819AGE,Y16-29;DS-127819SEX,T;DS-127819INDICATORS,OBS_FLAG;DS

- EuroStat. (2021, September 10). Retrieved from Youth employment by sex, age and educational attainment level: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=yth_empl_010&lang=en

- EuroStat. (2021, September 10). Retrieved from Young temporary employees as percentage of the total number of employees, by sex, age and country of birth: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=yth_empl_050&lang=en

- EuroStat. (2021, December 17). Retrieved from Self-perceived health by sex, age and income quintile: https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-051950_QID_-41C6AC4F_UID_-3F171EB0&layout=TIME,L,X,0;GEO,L,Y,0;QUANTILE,L,Z,0;AGE,L,Z,1;SEX,L,Z,2;LEVELS,L,Z,3;UNIT,L,Z,4;INDICATORS,C,Z,5;&zSelection=DS-051950INDICATORS,OBS_FLAG;DS-05195

- EuroStat. (2021, December 17). Retrieved from People at risk of poverty or social exclusion by age and sex: https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-127829_QID_109B0E0E_UID_-3F171EB0&layout=TIME,C,X,0;GEO,L,Y,0;UNIT,L,Z,0;AGE,L,Z,1;SEX,L,Z,2;INDICATORS,C,Z,3;&zSelection=DS-127829UNIT,PC;DS-127829AGE,Y16-29;DS-127829SEX,T;DS-127829INDIC

- EuroStat. (2021, December 17). Retrieved from Housing cost overburden rate by age, sex and poverty status - EU-SILC survey: https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-093640_QID_3D7EA1D7_UID_-3F171EB0&layout=TIME,C,X,0;GEO,L,Y,0;UNIT,L,Z,0;INCGRP,L,Z,1;AGE,L,Z,2;SEX,L,Z,3;INDICATORS,C,Z,4;&zSelection=DS-093640AGE,Y16-29;DS-093640INDICATORS,OBS_FLAG;DS-0

- EuroStat. (2021, October 29). Retrieved from Current depressive symptoms by sex, age and income quintile: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_ehis_mh1i&lang=en

7 - Attachments

```
/*CLUSTERS - Variables

/*1st step - hierarchical methods*/
options center;
%let Dataset= data12;    /*Input SAS Dataset Name*/
%let Variables= Mat_Deprevation Health_Perc Leave_Par_House Poverty House_Overburden 'Education 0-2'n Employment Dep_Symptoms;
%let Algorithm=centroid;     /*Hierarchical Method*/
%let NClus= 10;     /*Define the Number of Clusters*/
%let ID= 'Countries/Year'n;          /*ID or Observation Name Variable*/
PROC CLUSTER data=&Dataset SIMPLE NOEIGEN RMSSTD RSQUARE NOTIE STANDARD METHOD=&Algorithm OUT=Var_HCA_Tree_&Algorithm;
ID &ID;
VAR &Variables;
Run;
PROC TREE DATA=var_HCA_Tree_&Algorithm OUT=Var_HCA_&Algorithm NCLUSTERs=&NCLUS;
ID  &ID;
COPY &Variables;
RUN;

/*Assess hierarchical methods*/
proc sql;
create table var_centroid as
select*
from Var_hca_tree_centroid
where _ncl_<20;
run;
proc sql;
create table var_average as
select*
from Var_hca_tree_average
where _ncl_<20;
run;
proc sql;
create table var_complete as
select*
from Var_hca_tree_complete
where _ncl_<20;
run;
proc sql;
create table var_single as
select*
from Var_hca_tree_single
where _ncl_<20;
run;
proc sql;
create table var_wards as
select*
from Var_hca_tree_wards
where _ncl_<20;
run;
proc sql;
create table var_all_methods as
select distinct a._NCL_, a._RSQ_ as RSQ_centroid, b._RSQ_ as RSQ_average,
                c._RSQ_ as RSQ_complete, d._RSQ_ as RSQ_single, e._RSQ_ as RSQ_wards
from var_centroid as a, var_average as b, var_complete as c, var_single as d, var_wards as e
where a._NCL_=b._NCL_=c._NCL_=d._NCL_=e._NCL_;
run;
ods graphics on;
proc sgplot data=var_all_methods;
    xaxis reverse;
    yaxis label='Assessment Statistics';
    series x=_ncl_ y=RSQ_centroid /markers legendlabel='centroid';
    series x=_ncl_ y=RSQ_average /markers legendlabel='average';
    series x=_ncl_ y=RSQ_complete /markers legendlabel='complete';
    series x=_ncl_ y=RSQ_single /markers legendlabel='single';
    series x=_ncl_ y=RSQ_wards /markers legendlabel='wards';
run;
ods graphics off;

/*2nd step - extract the initial seeds (solution given by hierarchical methods)*/
PROC SORT data=var_hca_wards;
BY CLUSTER;
*standardizing initial seeds;
proc standard data=var_hca_wards mean=0 std=1 out=var_hca_wards_std;
var &Variables;
run;
proc means data=var_hca_wards_std mean nway noprint ;
var &Variables;
by cluster;
output out=var_initial_seeds mean=;
*standardizing original data for input in k-means;
proc standard data=&Dataset mean=0 std=1 out=var_dataset_std;
var &Variables;
run;

/*3rd step - non-hierarchical methods using initial seeds*/
%let NClus=6;
Proc Fastclus data=var_dataset_std SEED=var_Initial_Seeds MAXCLUSTERS=&NCLUS OUT=var_KMeans_Results MAXITER=50 REPLACE=NONE;
ID  &ID;
Var &Variables;
PROC SORT data=var_kmeans_results;
BY CLUSTER;
PROC PRINT data=var_kmeans_results;
BY CLUSTER;
VAR  &ID &Variables;
Run;
Proc means data=var_KMeans_Results N MEAN MEDIAN MODE MIN MAX STD VAR NWAY P10 P90;
Var &Variables;
```

```
/*CLUSTER - Factors

/*1st step - hierarchical methods*/
options center;
%let Dataset= data12;    /*Input SAS Dataset Name*/
%let Variables= 'Economic Power'n Independece 'Health vs Bad Education'n;    /*Variables to use in Cluster Analysis*/
%let Algorithm=wards;    /*Hierarchical Method*/
%let NClus= 6;      /*Define the Number of Clusters*/
%let ID= 'Countries/Year'n;         /*ID or Observation Name Variable*/
PROC CLUSTER data=&Dataset SIMPLE NOEIGEN RMSSTD RSQUARE NOTIE STANDARD METHOD=&Algorithm OUT=HCA_Tree_&Algorithm;
ID &ID;
VAR &Variables;
Run;
PROC TREE DATA=HCA_Tree_&Algorithm OUT=HCA_&Algorithm NCLUSTERs=&NCLUS;
ID  &ID;
COPY &Variables;
RUN;

/*Assess hierarchical methods*/
proc sql;
create table centroid as
select*
from hca_tree_centroid
where _ncl_<20;
run;
proc sql;
create table average as
select*
from hca_tree_average
where _ncl_<20;
run;
proc sql;
create table complete as
select*
from hca_tree_complete
where _ncl_<20;
run;
proc sql;
create table single as
select*
from hca_tree_single
where _ncl_<20;
run;
proc sql;
create table wards as
select*
from hca_tree_wards
where _ncl_<20;
run;
proc sql;
create table all_methods as
select distinct a._NCL_, a._RSQ_ as RSQ_centroid, b._RSQ_ as RSQ_average,
                c._RSQ_ as RSQ_complete, d._RSQ_ as RSQ_single, e._RSQ_ as RSQ_wards
from centroid as a, average as b, complete as c, single as d, wards as e
where a._NCL_=b._NCL_=c._NCL_=d._NCL_=e._NCL_;
run;
ods graphics on;
proc sgplot data=all_methods;
    xaxis reverse;
    yaxis label='Assessment Statistics';
    series x=_ncl_ y=RSQ_centroid /markers legendlabel='centroid';
    series x=_ncl_ y=RSQ_average /markers legendlabel='average';
    series x=_ncl_ y=RSQ_complete /markers legendlabel='complete';
    series x=_ncl_ y=RSQ_single /markers legendlabel='single';
    series x=_ncl_ y=RSQ_wards /markers legendlabel='wards';
run;
ods graphics off;

/*2nd step - extract the initial seeds (solution given by hierarchical methods)*/
PROC SORT data=hca_wards;
BY CLUSTER;
*standardizing initial seeds;
proc standard data=hca_wards mean=0 std=1 out=hca_wards_std;
var &Variables;
run;
proc means data=hca_wards_std mean noway noprint ;
var &Variables;
by cluster;
output out=initial_seeds mean=;
*standardizing original data for input in k-means;
proc standard data=&Dataset mean=0 std=1 out=dataset_std;
var &Variables;
run;

/*3rd step - non-hierarchical methods using initial seeds*/
%let NClus=6;
Proc Fastclus data=dataset_std SEED=Initial_Seeds MAXCLUSTERS=&NCLUS OUT=KMeans_Results MAXITER=50 REPLACE=NONE;
ID  &ID;
Var &Variables;
PROC SORT data=kmeans_results;
BY CLUSTER;
PROC PRINT data=kmeans_results;
BY CLUSTER;
VAR  &ID &Variables;
Run;
Proc means data=KMeans_Results N MEAN MEDIAN MODE MIN MAX STD VAR NWAY P10 P90;
Var &Variables;
by cluster;
output out=KMeans_Statistics mean=;
RUN;
```