|  |
| --- |
| Education Enrolment in Europe |
| Data analysis on the education level in European Contries in 2015 |
| Laura Becheru  Group: 1077  CSIE |

Contents

[Data Description 2](#_Toc503526438)

[PCA Analysis 4](#_Toc503526439)

[1. Reading and Observing the Data 4](#_Toc503526440)

[2. Determining the Principal Components 5](#_Toc503526441)

[3. Interpreting the Result 5](#_Toc503526442)

[Conclusion 8](#_Toc503526443)

[Works Cited 9](#_Toc503526444)

[Table of Figures 9](#_Toc503526445)

# Data Description

Despite the fact that many progresses has been made regarding education, there are still countries that are struggling with either a decreasing or a lower school enrolment rate than it was expected. I have chosen this subject after watching a play meant to bring some social awareness regarding the low education level that Romania is facing today. One of the main reasons that have led to an increasing in the number of uneducated people is a low participation rate in school.

In Europe, the lowest participation rates in primary education were registered in Romania (85.8 %), Hungary (91.3 %) and Luxembourg (92.3 %). Despite the resource mobilisation campaigns and political commitments, the share of children attending compulsory education decreased in several EU Member States over the last decade. Romania is the country with the most significant decrease in the past 10 years, almost 10%. (Eurostat, 2015).

Early school leaving is linked to unemployment, social exclusion, poverty and poor health. There are many reasons why some young people give up education and training prematurely: personal or family problems, learning difficulties, or a fragile socio-economic situation. The way the education system is set up, school climate and teacher-pupil relations are also important factors. Since there is not a single reason for early school leaving, there are no easy answers. Policies to reduce early school leaving must address a range of triggers and combine education and social policy, youth work and health related aspects such as drug use or mental and emotional problems. (European Commission, 2016)

For analysing the low education level I have used the following data:

* Pupils and Students in all level of education as % of total age population
* Mean and median income by age and sex - among people age between 18 to 64 years, of any gender, expressed in euro
* Total public expenditure on education as % of GDP, for all levels of education combined
* Unemployment by sex and age – annual average - among people between 25 and 74 years, as percentage of total population
* Self-perceived health by sex, age and educational attainment level - among people in all level of education, of any age and sex, with a good or very good level, measured as percentage
* Total fertility rate
* Gross domestic product per capita current prices
* Gross National Income (GNI) per Capita
* Unemployment by sex and age - annual average

|  |
| --- |
| The data is entirely extracted from [Eurostat Database](http://ec.europa.eu/eurostat/data/database), the model has 30 observations, consisting of European countries, both part of EU and outside EU, and the reference year is 2015. |

**Analysis Approach**

***Principal component analysis (PCA)*** is a statistical procedure to describe a set of multivariate data of possibly correlated variables by relatively few numbers of linearly uncorrelated variables. The uncorrelated variables are created as linear combination of the original variables and in the decreasing order of importance so that the first component explains most of the original variation in the data. Each component is orthogonal to the other component to explain the variation that is not already explained by other components.

With a large number of variables, the variance-covariance matrix may be too large to study and interpret properly. There would be too many pairwise correlations between the variables to consider. Graphical display of data may also not be of particular help in case the data set is very large. To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component. PCA is mostly used as a tool in [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis) and for making [predictive models](https://en.wikipedia.org/wiki/Predictive_modeling).

The purpose of this analysis is to highlight some of the main causes responsible for a lower education level in Romania compared with other European countries basing on the social and economic indicators mentioned above. In order to reach a valid conclusion we should examine which factors have a bigger influence on the level of education. Therefor I have chosen to apply principal component analysis in order to discard the data that does not particularly have a significant relevance for the carried out issue and emphasize the relevant factors used in this analysis. Thereby, instead of 8 variables as I have considered initially, I will try to obtain a smaller number (one between 2 and 4) so they will be uncorrelated. As a result the factor will not be represented on an eight dimension but on 2 or 3 dimension, depending on how many variable are actually relevant. In the new coordinate system, the first axis corresponds to the first principal component, which is the component that explains the greatest amount of the variance in the data.

# PCA **Analysis**

## Reading and Observing the Data

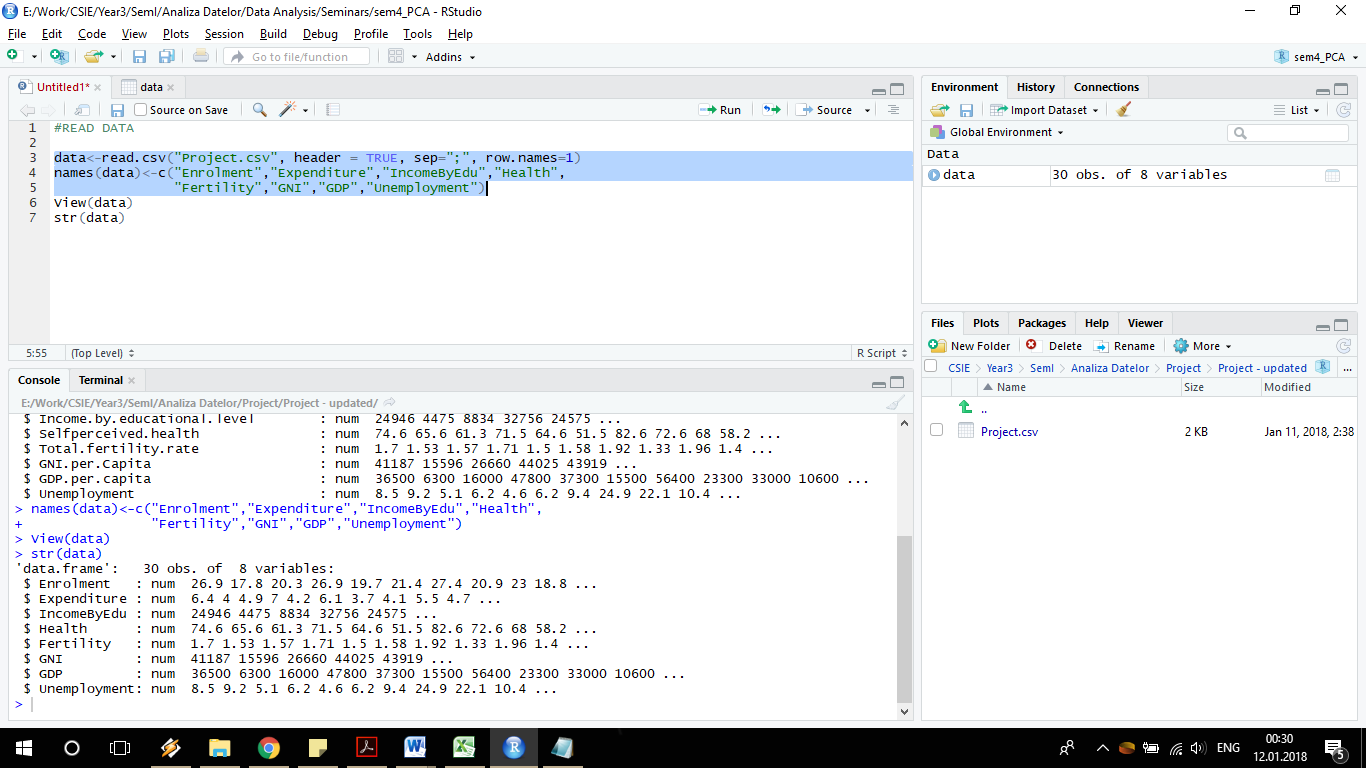
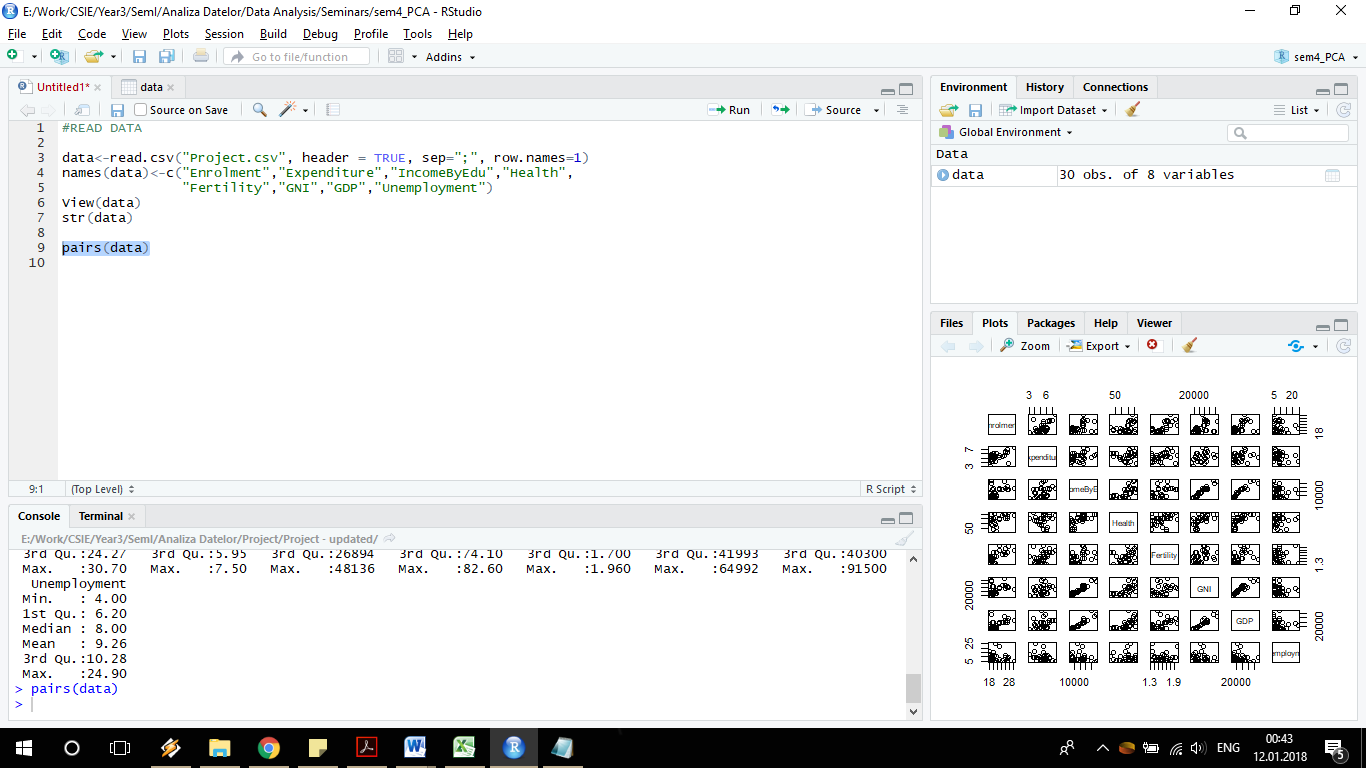
After reading the data in R Studio, the following output will be displayed when checking the internal structure of our object:

Figure 1: Internal Structure of Data

The output states that there are 30 observations of 8 variables, all the variables being numerical.

For emphasizing the correlation between the variables I have represented them through a pairs scatter plot.

From this plot we can observe that some of the variables are quite correlated.

Furthermore a good image about the correlation of the variables is also given by the correlation matrix.

But before computing the correlation matrix is recommended to standardize the data. Since the variables are independent variables and measured at different scales, they do not give equal contribution to the analysis.

Figure 2: Pairs Scatter Plot of Data

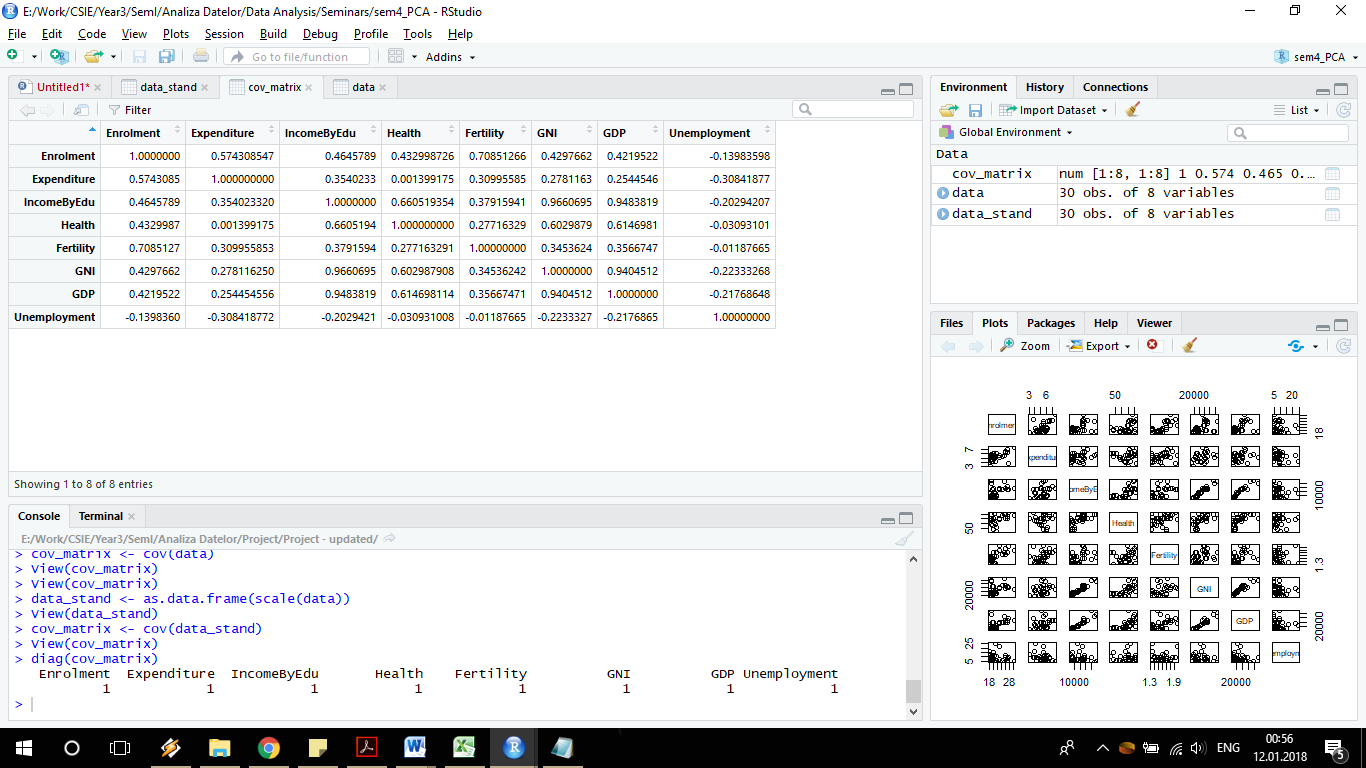


Figure 3: Correlation Matrix

## Determining the Principal Components

In order to determine the Principal Components for this analysis I have used the *prcomp()* function.

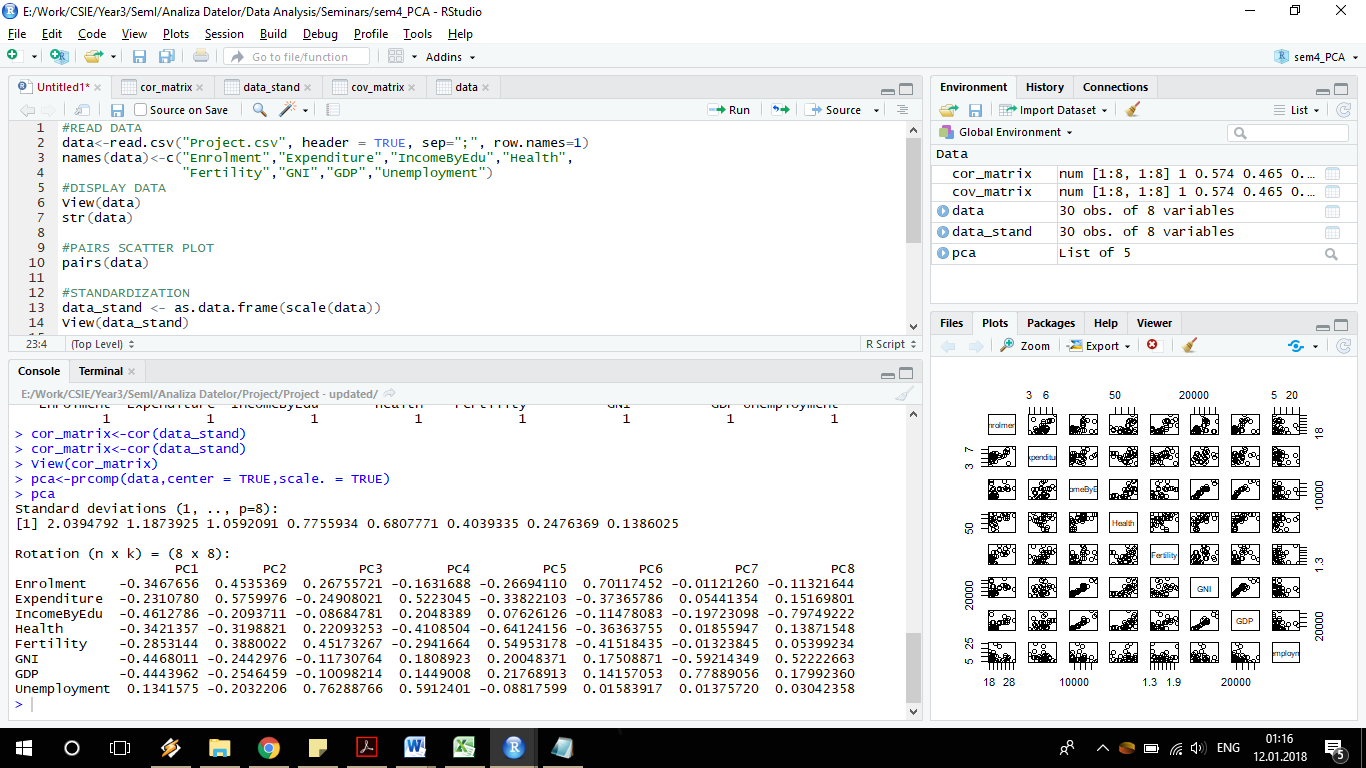


Figure 4: Results of pca

## Interpreting the Result

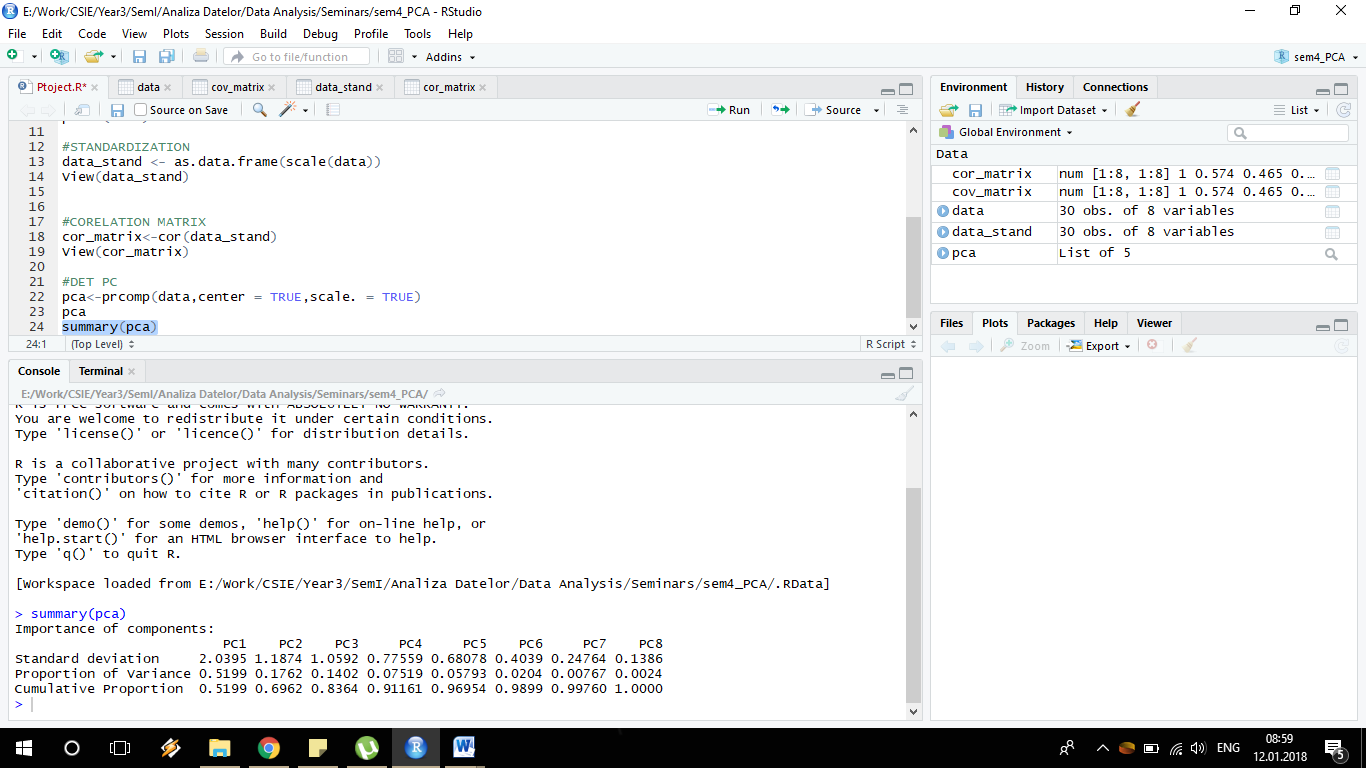
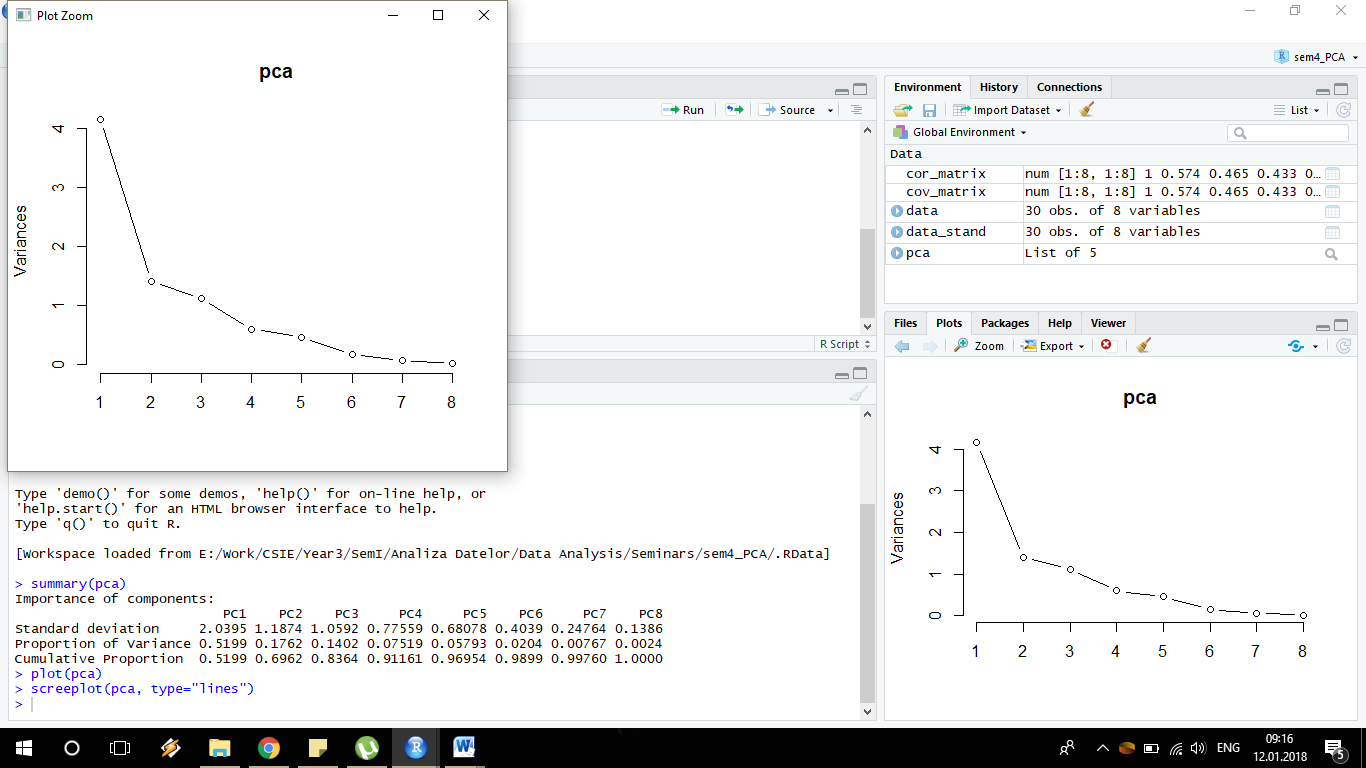


Figure 5: Pca Summary

This summary of this *pca* tells us the standard deviation associated to each component and it also tells us the proportion of the variance that that component comprises of the total variance. In this case, the first component comprises 51.99 % of the total variation, which means that component one is quite strong. Furthermore one can observe that component 2 and component 3 comprises 17.62%, respectively 14.02% of the total variation, whereas the others component have an insignificant proportion of variation. Looking at the Cumulative proportion we can see that the first 3 components explain almost 84% of the variation.

This can also be observed from the following scree plot:

Figure 6: Scree Plot PCA

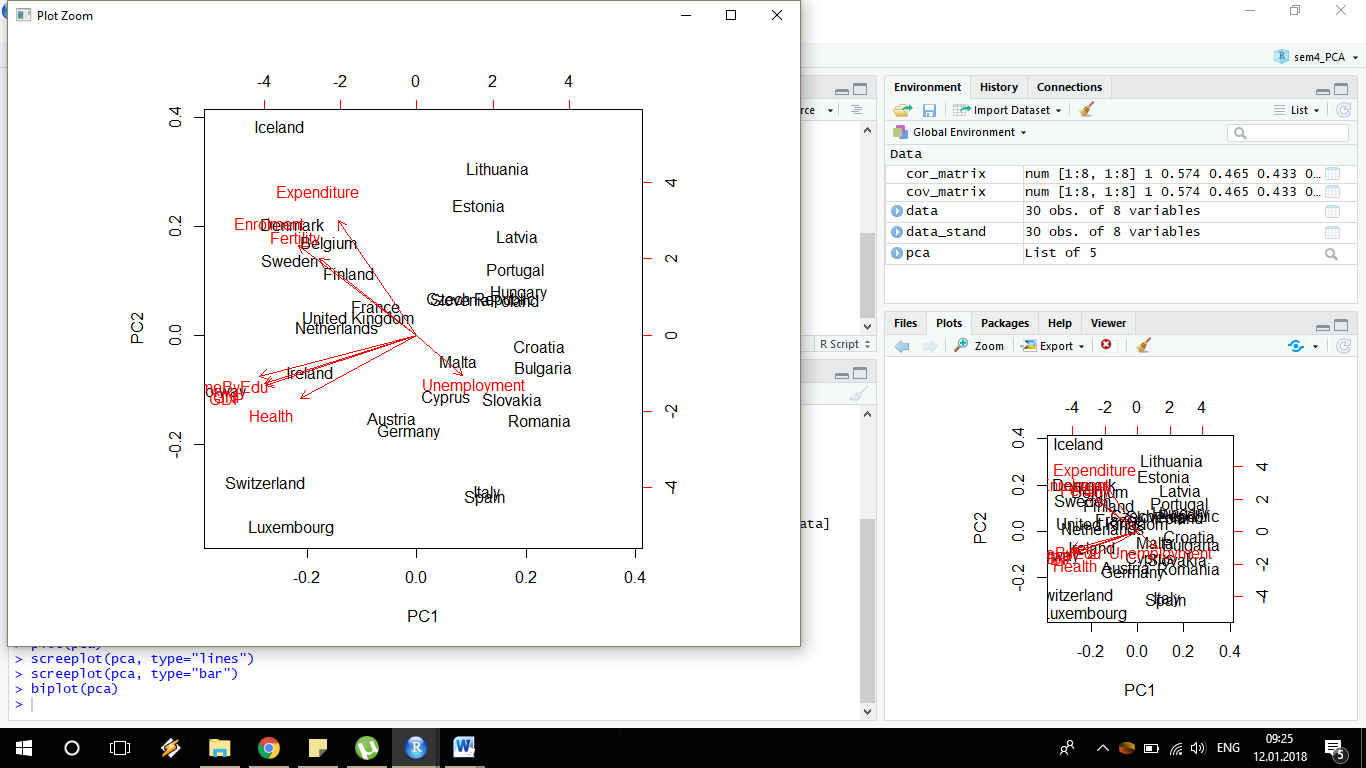
The plot shows how component 1 is really high, component 2 and 3 are smaller but still significant and the others are becoming smaller and smaller.

Figure 7: Biplot PCA

This graph shows the principal components and the direction they are following. From here we can see that the Education Enrollment, Fertility and Government Expenditure on Education negatively explain PC1 and positively PC2 while PC1 is positively explained by Unemployment. GNI, GDP, Income by education level and Self-preserved health are influencing the principal components 1 and 2 negatively.

In the graph below one can see better the observations (countries), and their distribution related to the first and second principal component.

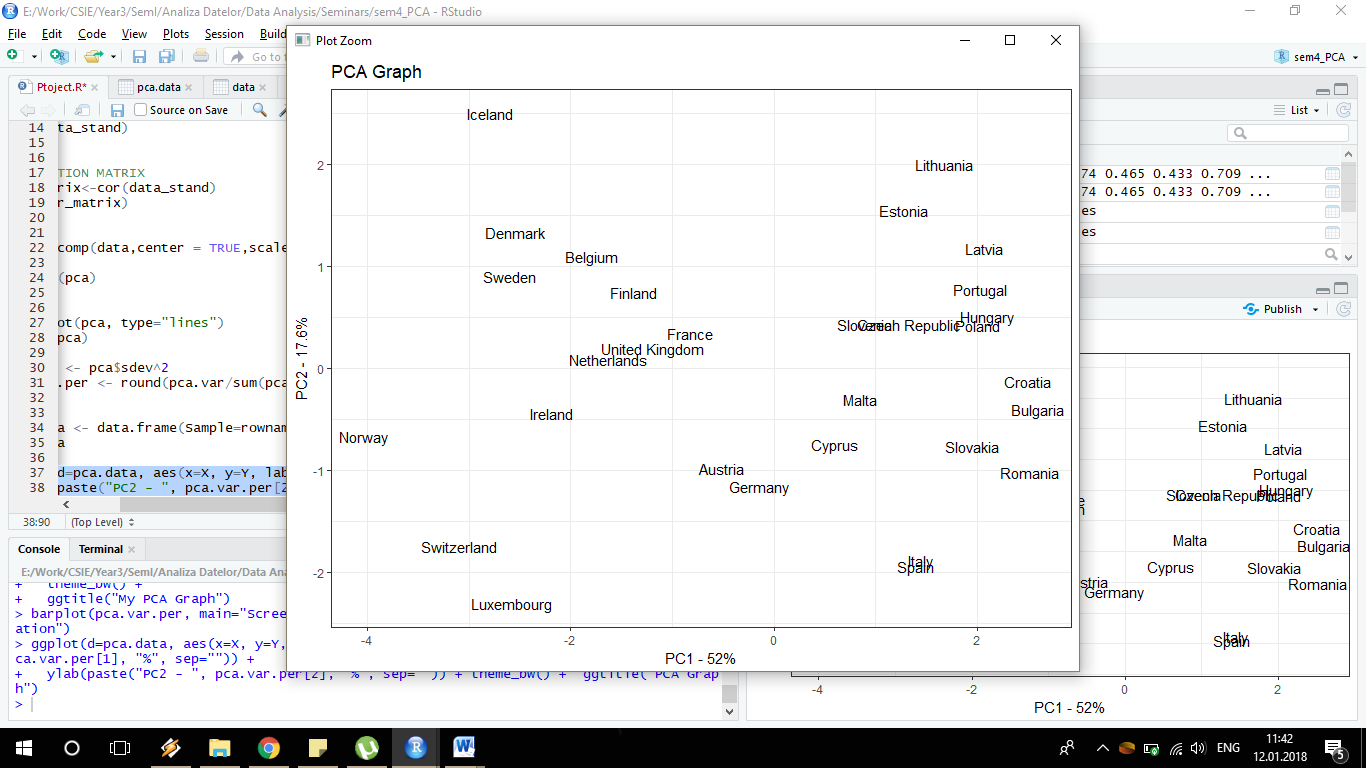


Figure 8: Ggplot PCA

# Conclusion

Education enrolment rate is still a problem in many European countries. Many children live in disadvantaged areas and the access to education is sometimes a challenge for them, the dropout rate is actually increasing in some of the European countries in spite the efforts that Europe has done in order to stop it.

After running a Principle Component Analyses on this dataset, I came to the conclusion that there is a significant correlation between Education Enrollment, Fertility and Government Expenditure on Education, as these variables are negatively explaining the First Principal Component and positively the Second Principal Component. Since the First Principle Component and Second Component has the greatest impact on our model, we can say that the Education Level in the countries used as observation in our data set, is influenced mostly by the number of people enrolled in school, the money that the government is spending on education and the fertility rate. Also, unemployment can have an important contribution, since the Principal Component is positively explained by it.

Low education rate can led to other social and economic difficulties for a country, as missing labour force or increase in mortality rate. European Union is trying to improve education all over Europe. They have a new strategy for decreasing the rate of school dropout under 10% in all the countries in Europe till 2020, whereas this target seems a little non-realistic regarding the higher rate of children leaving school in some countries, which is even worse in some countries since it has increased significantly in the past years.

# Works Cited

European Commission. (2016). *Early school leaving.* http://ec.europa.eu/education/policy/school/early-school-leavers\_en.

Eurostat. (2015). *Being young in Europe today - education.* Eurostat.

# Table of Figures

[Figure 1: Internal Structure of Data 0](file:///E:\Work\CSIE\Year3\SemI\Analiza%20Datelor\Project\Project.docx#_Toc503526067)

[Figure 2: Pairs Scatter Plot of Data 0](file:///E:\Work\CSIE\Year3\SemI\Analiza%20Datelor\Project\Project.docx#_Toc503526068)

[Figure 3: Correlation Matrix 0](#_Toc503526069)

[Figure 4: Results of pca 0](file:///E:\Work\CSIE\Year3\SemI\Analiza%20Datelor\Project\Project.docx#_Toc503526070)

[Figure 5: Pca Summary 0](file:///E:\Work\CSIE\Year3\SemI\Analiza%20Datelor\Project\Project.docx#_Toc503526071)

[Figure 6: Scree Plot PCA 0](file:///E:\Work\CSIE\Year3\SemI\Analiza%20Datelor\Project\Project.docx#_Toc503526072)

[Figure 7: Biplot PCA 0](#_Toc503526073)

[Figure 8: Ggplot PCA 0](#_Toc503526074)