Università degli Studi di Milano

MSc in Data Science for Economics LM – DATA



Statistical learning project

The efficiency of sleep: an analysis for its improvement (Supervised)

How European Countries react to their environments (Unsupervised)

Michela Mazzaglia

12851A

ACADEMIC YEAR 2022-2023

Contents

[Abstract 2](#_Toc139976708)

[1.1 Our data 4](#_Toc139976709)

[1.2 Pre-processing 6](#_Toc139976710)

[1.3 Exploratory data analysis 8](#_Toc139976711)

[2.1 Inference analysis 15](#_Toc139976712)

[2.2 Stepwise selection 18](#_Toc139976713)

[2.3 Ridge and Lasso regression 19](#_Toc139976714)

[2.4 Models’ evaluation 21](#_Toc139976715)

[3.1 Robust regression 22](#_Toc139976716)

[3.2 Decision Tree 24](#_Toc139976717)

[3.3 Random Forest 25](#_Toc139976718)

[3.4 Conclusions of the supervised part 28](#_Toc139976719)

[4.1 The dataset 31](#_Toc139976720)

[4.2 Pre-processing and correlation 35](#_Toc139976721)

[5.1 PCA 40](#_Toc139976722)

[5.2 Clustering 42](#_Toc139976723)

[5.3 Conclusions of the unsupervised part 47](#_Toc139976724)

[References 48](#_Toc139976725)

# Abstract

Social health is always at the centre of the attention of world’s countries. Especially nowadays, after the spread of covid-19, it has been confirmed that every aspect of social health is particularly important. Therefore, many analyses have been conducted and this work’s aim is exactly to point out the effect of people’s habits on the efficiency of sleep and how the environment affect some European countries mental health. Sleeping well and the amount of sleep a person gains after the night can affect the person’s reactions during the day, their productivity and many other factors. This work is divided into two parts in which different techniques were implemented. The first part focuses on the analysis of sleep efficiency to detect ways for its improvement in people's lives. It could be useful to understand sleeping problems and how they are influenced by people’s habits. The second part focuses on some European countries and characteristics of their environment to study how it affects social health, in other words how citizens perceive their mental health looking at features as Trust in the legal system, Depressive symptoms, Life satisfaction and others. The project follows a study of the different variables to understand the nature of the data, explanations and comments on what achieved through the statistical analysis conducted and, to conclude, the use of a few machine learning and statistical learning methods with relative interpretation (Robust Regression, Stepwise selection, Lasso, Decision Tree, Random Forest, PCA and Cluster analysis).

Supervised learning project

The efficiency of sleep: an analysis for its improvement

Chapter 1

# 1.1 Our data

The dataset is a collection of sleeping characteristics and people’s habits, it also contains information about a group of test subjects and their sleep patterns. There are 452 records, each representing a test subject which is identified by a unique “Subject ID”. Their age and gender are also recorded as features of the individual. Between the sleeping features there are recorded: the Bedtime and Wakeup time, that indicates when each subject goes to bed and wake up each day; Sleep duration which records the total amount of time each subject slept in hours; Rem sleep percentage, Deep sleep percentage and Light sleep percentage features indicate the amount of time each subject spent in each stage of sleep; Awakenings feature represents the number of time each subject wakes up during the night. On the people’s habits side there can be seen Caffeine consumption, Alcohol consumption, regarding the 24 hours prior to bedtime, their Smoking status and Exercise frequency. The dataset provided was collected as part of a study conducted in the UK by a research team at the University of Oxfordshire.

The goal is to conduct inference and predict the target variable. All the variables that are going to be used as features are cited as follows:

* ID = a unique identifier for each test subject
* Age = age of the test subject
* Gender = male or female
* Bedtime = the time the test subject goes to bed each night
* Wakeup time = the time the test subject wakes up each morning
* Sleep duration = the total amount of time the test subject slept (in hours)
* REM sleep percentage = the percentage of total sleep time spent in REM sleep
* Deep sleep percentage = the percentage of total sleep time spent in deep sleep
* Light sleep percentage = the percentage of total sleep time spent in light sleep
* Awakenings = the number of times the test subject wakes up during the night
* Caffeine consumption = the amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
* Alcohol consumption = the amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
* Smoking status = whether the test subject smokes or not
* Exercise frequency = the number of times the test subject exercises each week

Our target variable is instead the following:

* Sleep efficiency = a measure of the proportion of time in bed spent asleep

## Pre-processing

Firstly, since the dataset presented both qualitative and quantitative variables, some adjustments as converting categorical data into numeric were done. It regards the variable Smoking status (Yes, No) and the variable Gender (Female, Male).

'data.frame': 452 obs. of 15 variables:

$ ID : int 1 2 3 4 5 6 7 8 9 10 ...

$ Age : int 65 69 40 40 57 36 27 53 41 11 ...

$ Gender : chr "Female" "Male" "Female" "Female" ...

$ Bedtime : chr "2021-03-06 01:00:00" "2021-12-05 02:00:00" "2021-05-25 21:30:00" "2021-11-03 02:30:00" ...

$ Wakeup.time : chr "2021-03-06 07:00:00" "2021-12-05 09:00:00" "2021-05-25 05:30:00" "2021-11-03 08:30:00" ...

$ Sleep.duration : num 6 7 8 6 8 7.5 6 10 6 9 ...

$ Sleep.efficiency : num 0.88 0.66 0.89 0.51 0.76 0.9 0.54 0.9 0.79 0.55 ...

$ REM.sleep.percentage : int 18 19 20 23 27 23 28 28 28 18 ...

$ Deep.sleep.percentage : int 70 28 70 25 55 60 25 52 55 37 ...

$ Light.sleep.percentage: int 12 53 10 52 18 17 47 20 17 45 ...

$ Awakenings : num 0 3 1 3 3 0 2 0 3 4 ...

$ Caffeine.consumption : num 0 0 0 50 0 NA 50 50 50 0 ...

$ Alcohol.consumption : num 0 3 0 5 3 0 0 0 0 0 ...

$ Smoking.status : chr "Yes" "Yes" "No" "Yes" ...

$ Exercise.frequency : num 3 3 3 1 3 1 1 3 1 0 ...

Age Gender Sleep.duration Sleep.efficiency Deep.sleep.percentage

Min. : 9.00 Length:452 Min. : 5.000 Min. :0.5000 Min. :18.00

1st Qu.:29.00 Class :character 1st Qu.: 7.000 1st Qu.:0.6975 1st Qu.:48.25

Median :40.00 Mode :character Median : 7.500 Median :0.8200 Median :58.00

Mean :40.29 Mean : 7.466 Mean :0.7889 Mean :52.82

3rd Qu.:52.00 3rd Qu.: 8.000 3rd Qu.:0.9000 3rd Qu.:63.00

Max. :69.00 Max. :10.000 Max. :0.9900 Max. :75.00

Awakenings Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency

Min. :0.000 Min. : 0.00 Min. :0.000 Length:452 Min. :0.000

1st Qu.:1.000 1st Qu.: 0.00 1st Qu.:0.000 Class :character 1st Qu.:0.000

Median :1.000 Median : 25.00 Median :0.000 Mode :character Median :2.000

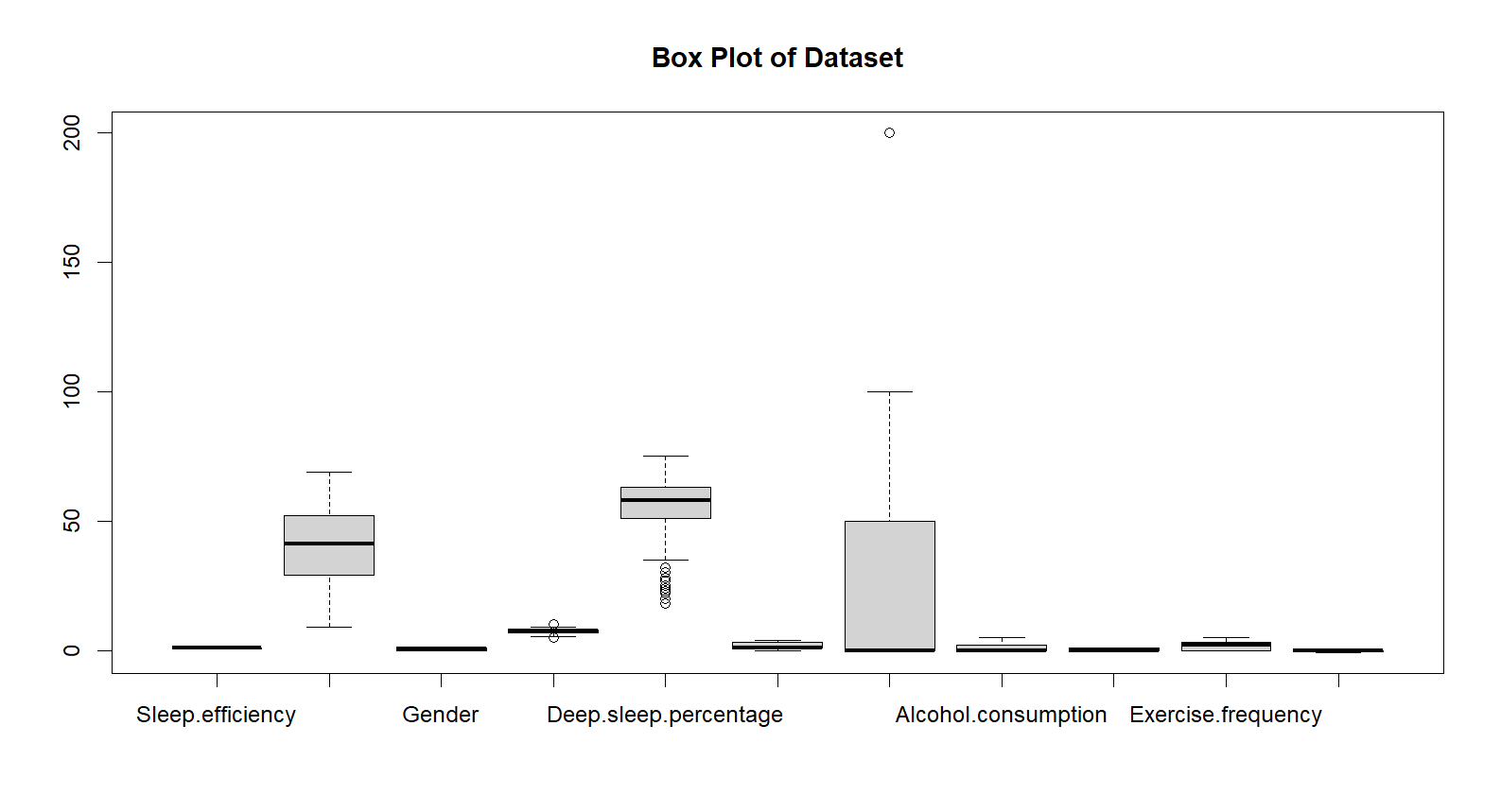
Mean :1.641 Mean : 23.65 Mean :1.174 Mean :1.791

3rd Qu.:3.000 3rd Qu.: 50.00 3rd Qu.:2.000 3rd Qu.:3.000

Max. :4.000 Max. :200.00 Max. :5.000 Max. :5.000

NA's :20 NA's :25 NA's :14 NA's :6

They were both converted into dummy variables with values {0, 1}. Some other adjustments involved excluding from the selection of the variable’s columns Wakeup time and Bedtime, since they were considered not important for the final goal. Even the column ID is rather useless in terms of the objective of this work. As next step, missing values were detected and since they are difficult to handle, they were removed. The dataset has now 388 observations and an analysis of each variable is required to identify potential outliers.

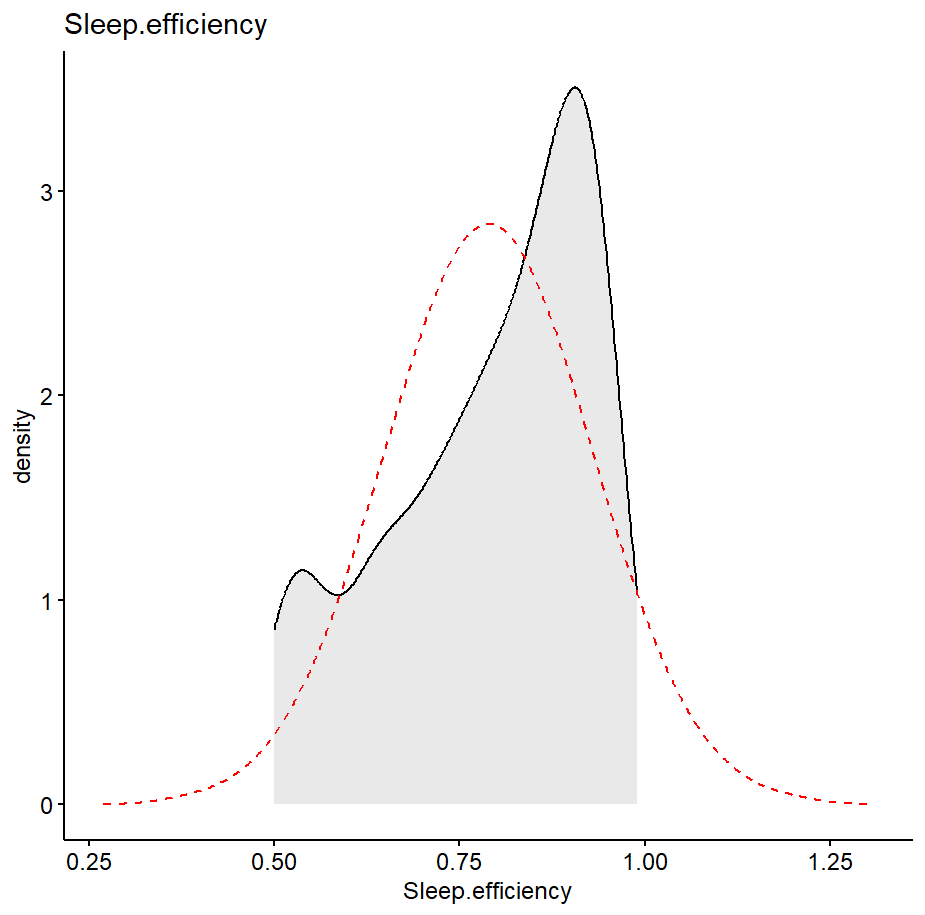


The plot above shows that there are outliers, especially regarding Deep sleep percentage. It is possible to spot some also for Caffeine consumption. The presence of outliers is not necessarily a problem, but it requires careful consideration and appropriate handling. Later in this work there will be techniques to handle them.

## 1.3 Exploratory data analysis

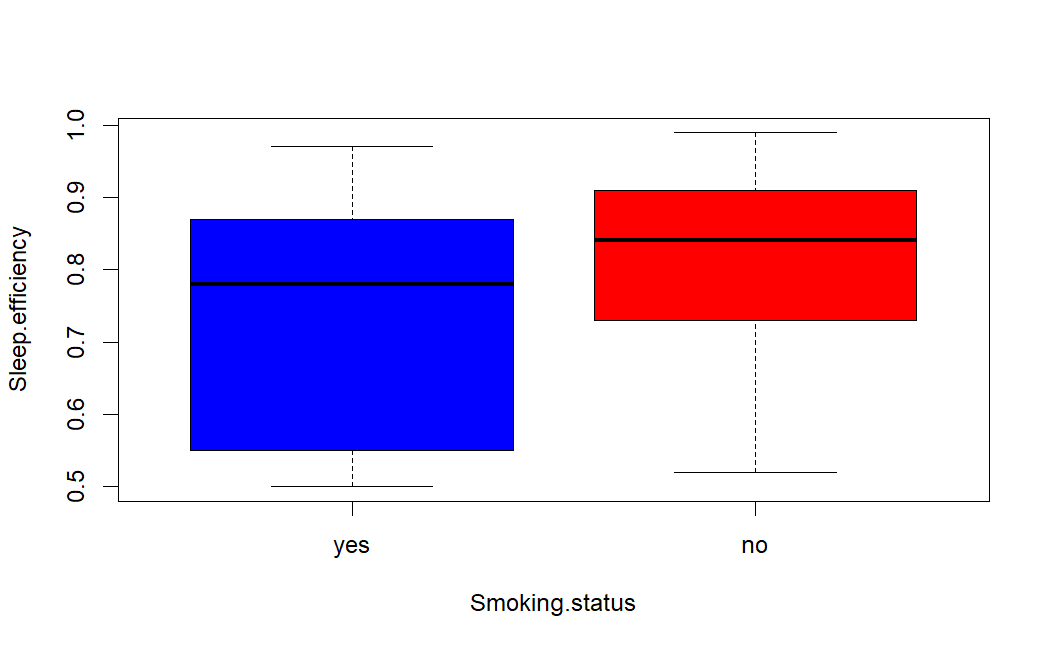
In this part of the work the study was focused on the description of the target variable Sleep efficiency. Normality was checked with the use of the Shapiro-wilk test and it was clear that its p-value was really low meaning that the distribution was not a Gaussian distribution, so we had to reject the null hypothesis of the test.

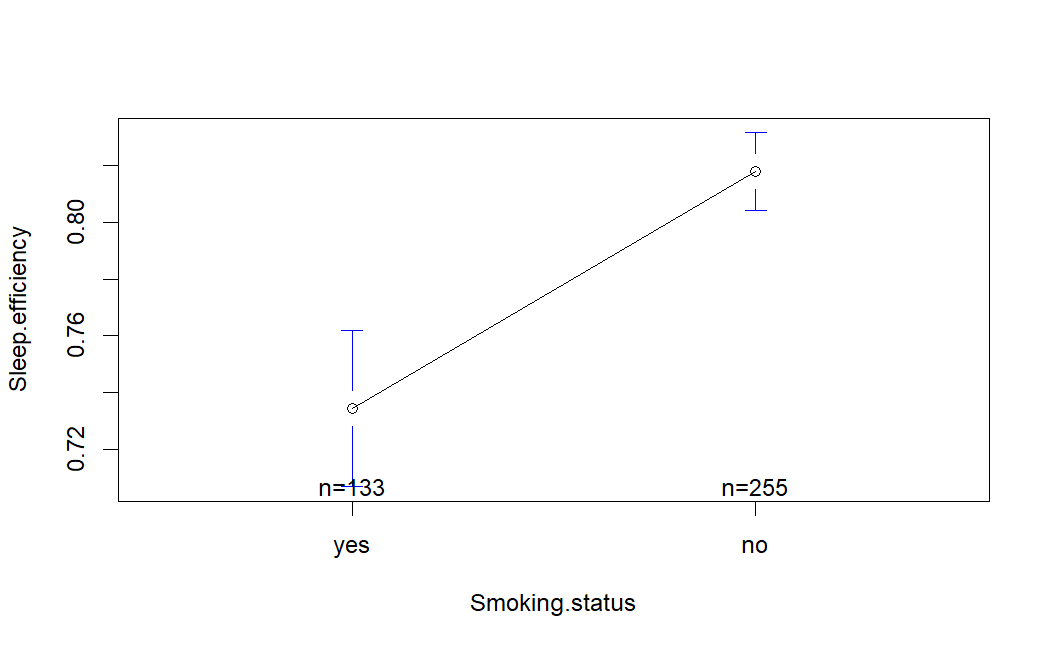
It is clear thanks to the following plot that the distribution is skewed:



However, the logarithm was applied to try to improve its distribution. It is possible to see a slight improvement with the logarithm. Next, some relationships between sleep efficiency and people’s habits were analysed.

Does sleep efficiency change a lot if the individual is a smoker or not?

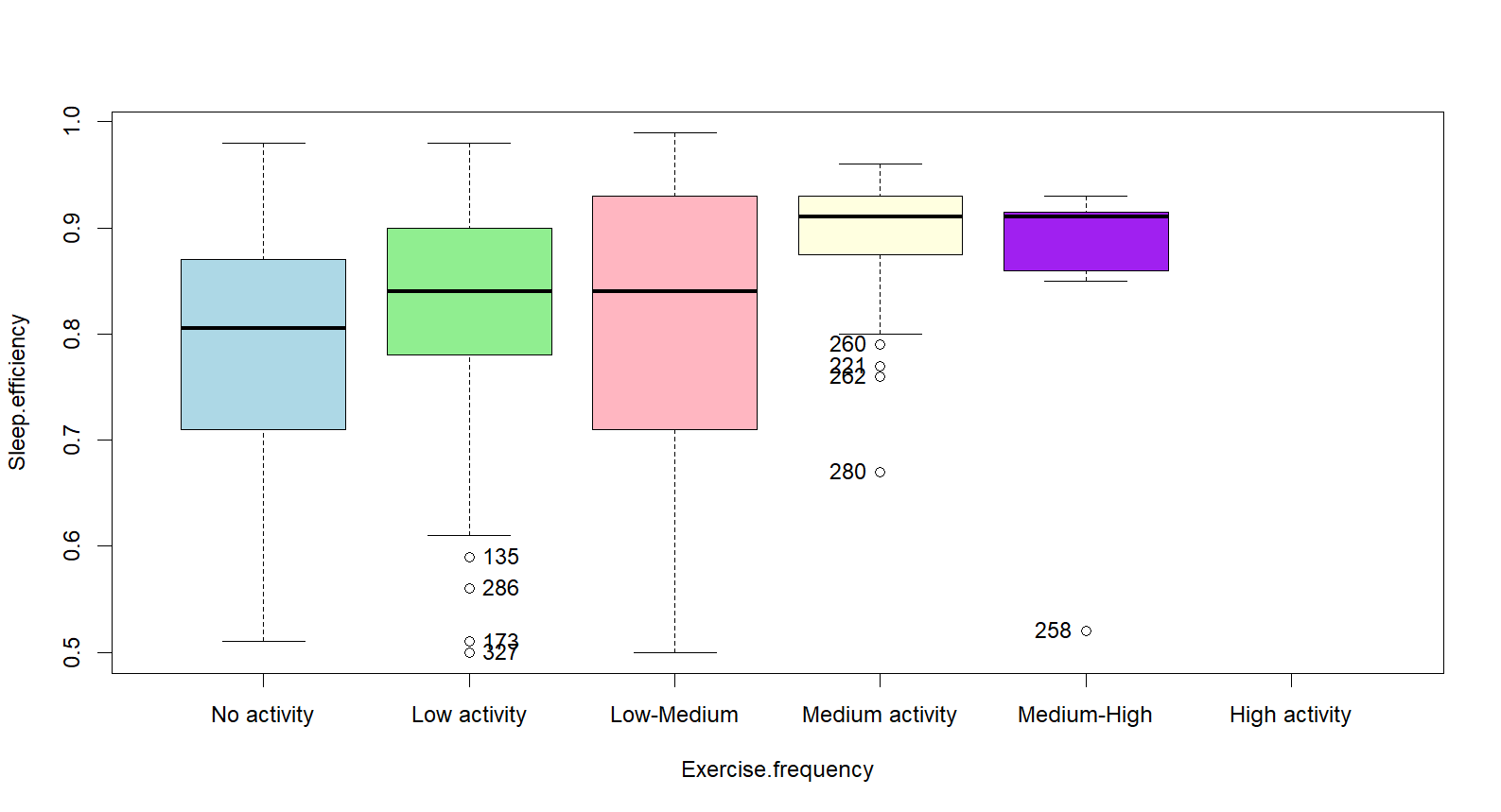




The summary shows that if the person is a smoker, then the mean is lower than if the individual does not smoke (0.73 against 0.81). The standard deviation is bigger meaning that the data varies more and there is a significant difference in numbers since the individuals that smokes are n = 133 and those who do not are n = 255, meaning that there is majority of people who do not smoke in the collected data. The smokers are also those that have a higher variability in Sleep efficiency.

Does sleep efficiency differ significantly for the level of exercise frequency?

To answer this question, it was necessary tounderstand the minimum and maximum level of Exercise frequency. The first one is indicated by 0, the maximum instead by 5. This scale was renamed from 0 as ‘No activity’ and to continue, as shown in the following plot.



Medium activity was registered as the one with the highest mean (0.89), while the one presenting more values was Low-Medium with n = 113. High activity was registered as NA values with a n = 110, that is the reason for its missing boxplot. The plot clearly tells that Sleep efficiency differs for Exercise frequency: the F-statistic obtained was 5.17, larger than the critical value. Moreover, Medium-High was registered as the exercise frequency that takes a wider range of values and differ from the others. From the analysis of their relationship, it is confirmed that individuals with higher sleep efficiency are the ones with higher exercise frequency. This may indicate the beneficial effects of physical activity on the efficiency of sleep.

Does Sleep efficiency depend linearly on the sleep duration?

Call:

lm(formula = Sleep.efficiency ~ Sleep.duration, data = a\_cons)

Residuals:

Min 1Q Median 3Q Max

-0.29058 -0.08911 0.03089 0.11531 0.20236

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.811186 0.058650 13.831 <2e-16 \*\*\*

Sleep.duration -0.002944 0.007817 -0.377 0.707

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1359 on 386 degrees of freedom

Multiple R-squared: 0.0003673, Adjusted R-squared: -0.002222

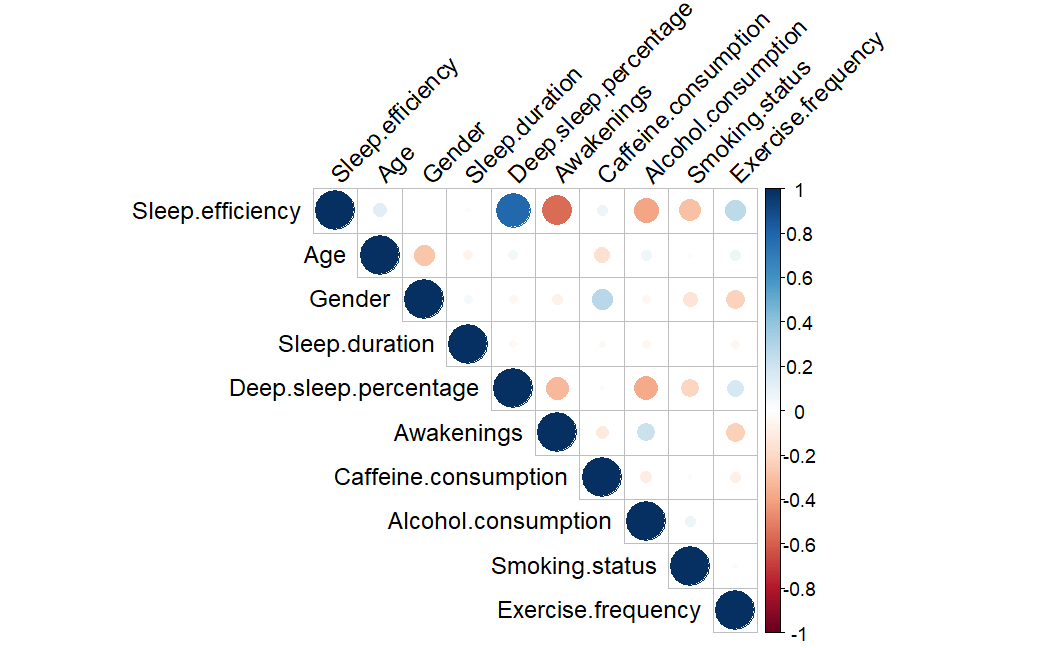
F-statistic: 0.1418 on 1 and 386 DF, p-value: 0.7067

Looking at the p-value, it is not smaller than 0.05 which means that it is not significantly different and does not suggest a linear relationship between the two variables. The R-squared is also very low, if it had been large, it would have suggested a stronger linear relationship.

Here the following scatterplot that represent it: Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

In addition, to better understand all the variables in the dataset an analysis of Sleep efficiency by age was conducated which led to acknowledge a significant drop in sleep efficiency in individuals between 30 and 40 years old, compared to its high value before individuals turned 30. This may be due to childbirth, parenting responsibilities, work stress or other factors. Number of awakenings is also a really important variable. Its relationship with Sleep efficiency shows that people with less awakenings are also the ones having a higher sleep efficiency and deep sleep percentage. The last two variables analysed with Sleep efficiency were Caffeine consumption and Alcohol consumption. In both it is obvious that a minor consumption, or even no consumption, of these two products lead to a higher sleep efficiency. As next step, a correlation map was displayed. From the mixed correlation with Sleep efficiency, the most correlated variable is Deep sleep percentage, followed by a negative correlation with Awakenings. It has also a negative correlation with Caffeine consumption and Alcohol consumption, it indicates a negative relationship or inverse association between the target variable and the other variables. In other words, as the values of the target variable increase, the values of the other variables tend to decrease, and vice versa. The following plot shows the correlation:



Chapter 2

# 2.1 Inference analysis

In order to conduct a regression analysis for Sleep efficiency, certain variables of the dataset were selected due to the presence of too much correlation and to avoid afterwards multicollinearity that could affect the reliability of the work. The data was split into training set (76 observations) and test set (312 observations), respectively 20% and 80% of the total observations of the dataset.

Call:

lm(formula = log(Sleep.efficiency) ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-0.39392 -0.06784 0.01462 0.07889 0.30884

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.1041402 0.0741634 -1.404 0.161284

Age 0.0019651 0.0005663 3.470 0.000596 \*\*\*

Gender -0.0068500 0.0162153 -0.422 0.673003

Sleep.duration -0.0112491 0.0085256 -1.319 0.188016

Caffeine.consumption 0.0002032 0.0002767 0.734 0.463331

Alcohol.consumption -0.0378847 0.0046431 -8.159 9.09e-15 \*\*\*

Smoking.status -0.1125359 0.0159359 -7.062 1.13e-11 \*\*\*

Awakenings -0.0574063 0.0057042 -10.064 < 2e-16 \*\*\*

Exercise.frequency 0.0158644 0.0053321 2.975 0.003163 \*\*

---

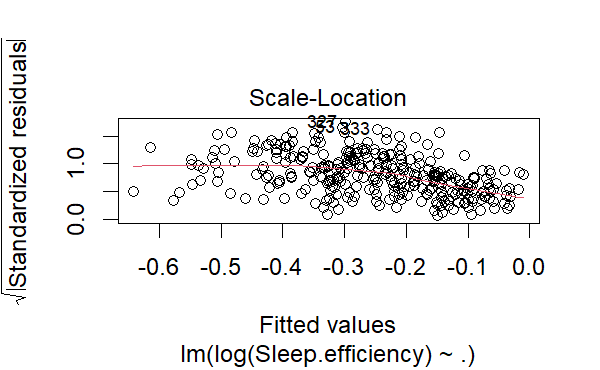
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1288 on 303 degrees of freedom

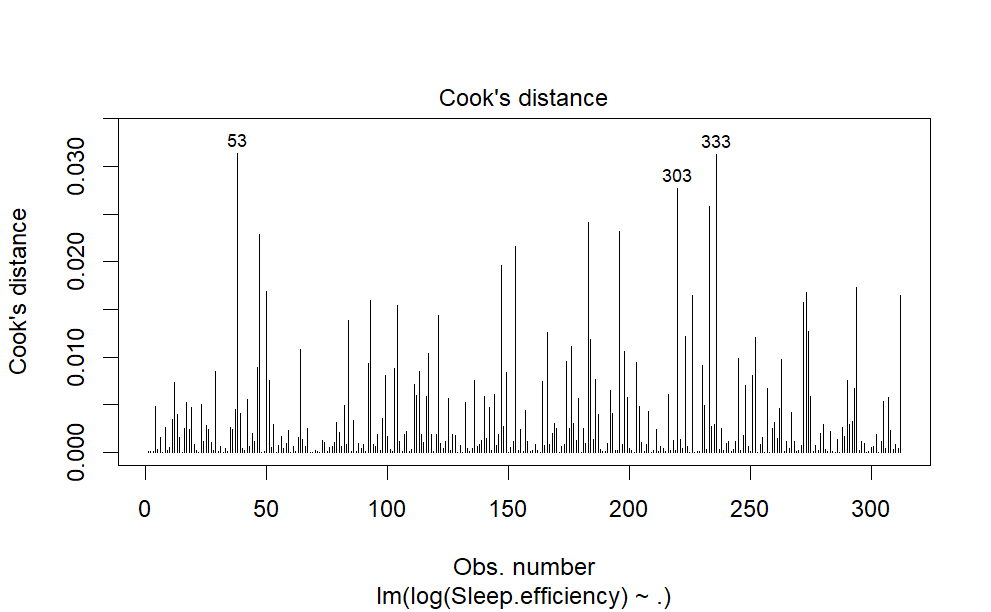
Multiple R-squared: 0.5143, Adjusted R-squared: 0.5015

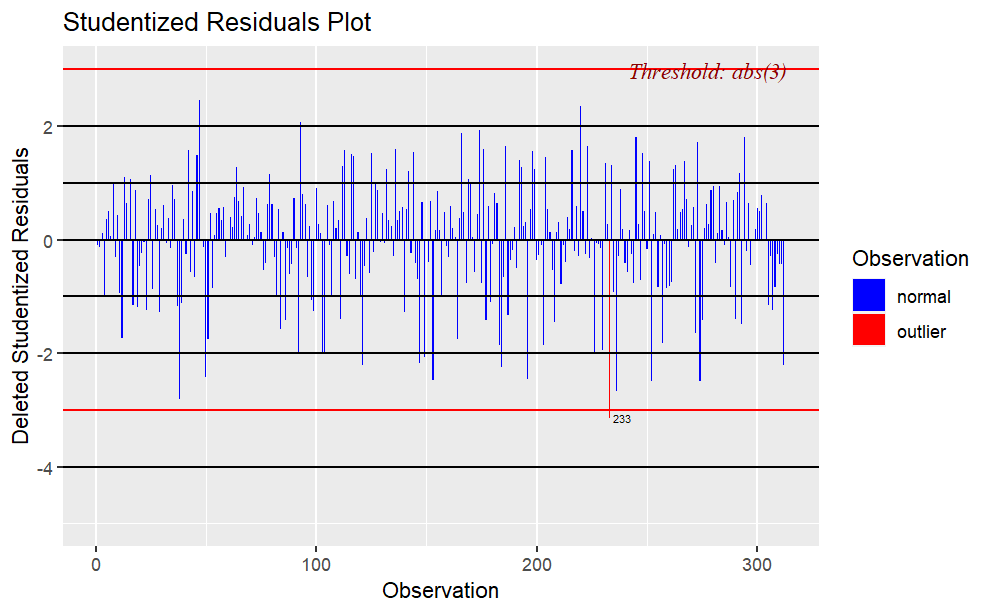
F-statistic: 40.1 on 8 and 303 DF, p-value: < 2.2e-16

A regression of Sleep efficiency on all the other variables in the dataset was developed. As results, the formula displayed a low p-value which suggests that the regression model was statistically significant and provides a better fit to the data. The R-squared was 0.5143, it indicates that approximately 51.43% of the variation in the logarithm of Sleep efficiency can be explained by the predictor variables. As most statistically significant variables there can be highlighted: Age, Alcohol consumption, Smoking status, Awakenings and Exercise frequency. In order to avoid multicollinearity, as stated before, the dataset was cleaned and checked using the *vif* function in R. All the variables resulted false which indicates that the correlation with Sleep efficiency is not too high. Non- normality of the residuals and heteroscedasticity (non-constant error variance) were detected. The hypothesis of Normality and Homoscedasticity had to be rejected. Although, Normality is not a problem since the sample of the dataset is big enough to use the properties of the Central Limit Theorem. Heteroscedasticity will be dealt instead with proper techniques as the Robust Regression.



Looking at the residual’s distribution, it is possible to notice with the Shapiro-wilk normality test that there is a good fit to the normal distribution. The output of the studentized Breusch-Pagan test detected heteroscedasticity in the data. The low p-value provides evidence to reject the null hypothesis of homoscedasticity. Cook’s distance and studentized residuals were applied to check for outliers:





Leverages and outliers were detected with unsurprisingly some outliers. This led to choose as method of analysis the Robust regression which can take care of these problems.

# 2.2 Stepwise selection

A stepwise selection was performed on the train data:

Call:

lm(formula = log(Sleep.efficiency) ~ Age + Alcohol.consumption +

Smoking.status + Awakenings + Exercise.frequency, data = train)

Residuals:

Min 1Q Median 3Q Max

-0.38752 -0.06496 0.01441 0.08165 0.31341

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.189790 0.027230 -6.970 1.96e-11 \*\*\*

Age 0.002018 0.000547 3.688 0.000267 \*\*\*

Alcohol.consumption -0.037663 0.004619 -8.154 9.15e-15 \*\*\*

Smoking.status -0.111572 0.015673 -7.119 7.80e-12 \*\*\*

Awakenings -0.057547 0.005640 -10.204 < 2e-16 \*\*\*

Exercise.frequency 0.016216 0.005130 3.161 0.001728 \*\*

---

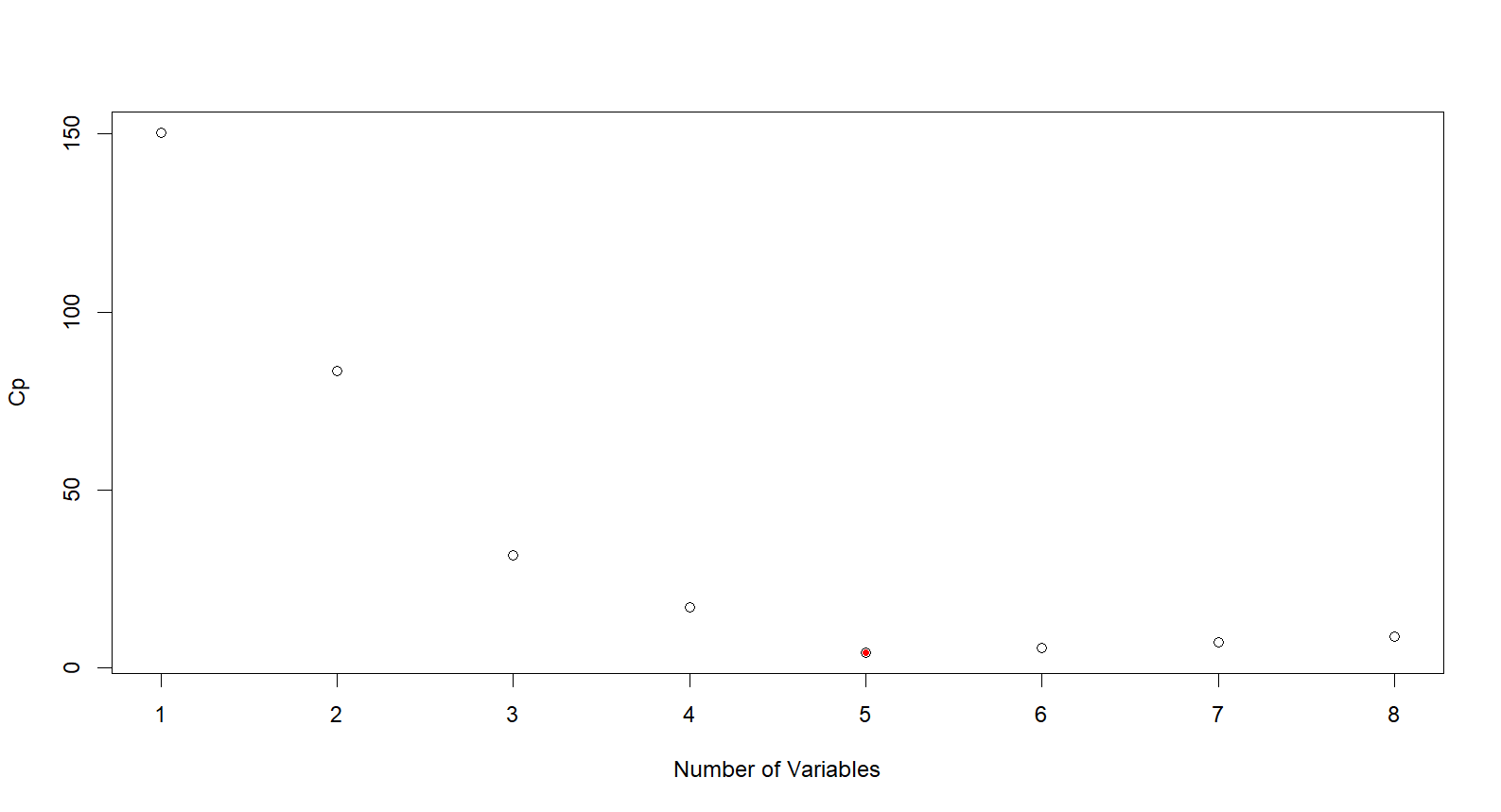
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1287 on 306 degrees of freedom

Multiple R-squared: 0.5103, Adjusted R-squared: 0.5023

F-statistic: 63.77 on 5 and 306 DF, p-value: < 2.2e-16

Here it is possible to observe the variables that were selected as most significant for the analysis. The model selects automatically the best variables performing both a forward and a backward stepwise selection. The number of variables drastically decreased. The model assumptions are the same of the normal regression one, as seen by the same tests that were conducted on the stepwise model. The following plot shows the best number of variables selected by the Stepwise selection model:



To conclude, the best number was 5. The analysis was ran also using other techniques as the Ridge and Lasso regression.

# 2.3 Ridge and Lasso regression

Using the Ridge regression model the value of lambda obtained was of -2 around 0.01 approximately, which explains a more constrained model. The number of selected variables was 8, more than the Stepwise selection method. The ones selected by the Ridge model were:

s1

(Intercept) 0.8399097334

Age 0.0008527556

Gender -0.0017711619

Sleep.duration -0.0034296217

Caffeine.consumption 0.0001464648

Alcohol.consumption -0.0145099611

Smoking.status -0.0492058342

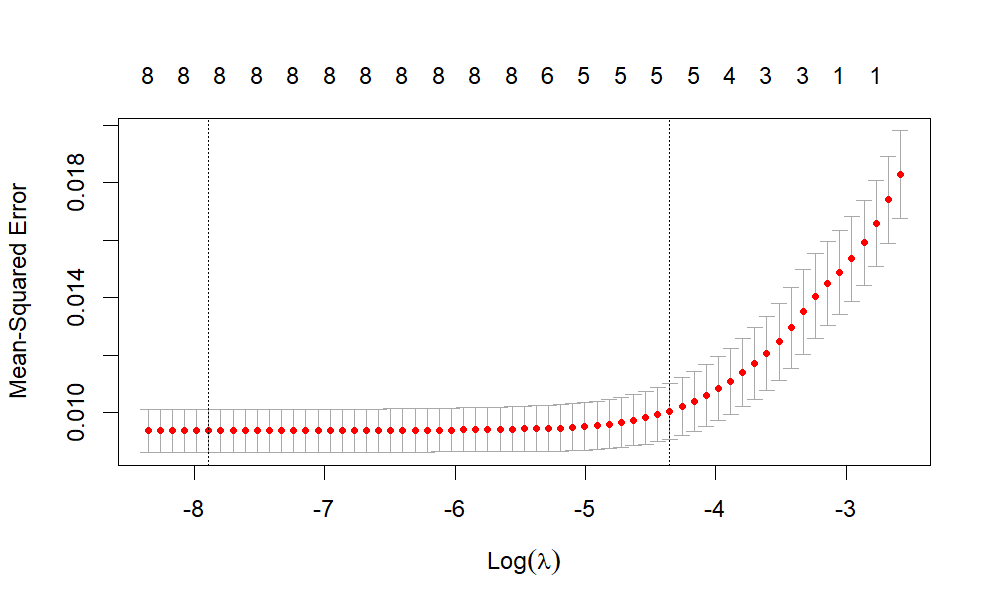
Awakenings -0.0297170306

Exercise.frequency 0.0105773543

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

On the other hand, the Lasso model was applied. This method differentiates from the other two. As it is possible to see from the following graph the lambda chosen was between -5 and -4. The number of chosen variables changed depending on the values of lambda. In this case, the value selected was 5, as in the result of the stepwise selection. Based on these overall results then, the best number of variables that must be selected is 5. Next, we evaluate the performance of these models.



# 2.4 Models’ evaluation

Now that all the models were assessed the next step is proceeding in comparing their performance to choose the best one that could predict and that works better for the dataset. In order to achieve that, RMSE (root-mean-square deviation) of OLS, Stepwise, Ridge and Lasso were computed with the cross-validation technique. These were the values:

| **Model** | **RMSE** |  |  |  |
| --- | --- | --- | --- | --- |
| OLS | 0.2002482 |  |  |  |
| Stepwise | 1.0743037 |  |  |  |
| Lasso | 0.1011143 |  |  |  |
| Ridge | 0.1186134 |  |  |  |

A lower RMSE indicates a better prediction accuracy, in this case is the LASSO. It is followed by the ridge and then the OLS model, as a kind of ranking between them. The stepwise selection positions as the last one.

Chapter 3

# 3.1 Robust regression

After having assigned the best model, the study focused on the application of a Robust Regression with the aim to deal with the outliers of the dataset. As method of use the Bisquare weights were chosen. They show the first 10 observations that are considered as worst outliers scenario.

Dependent variable:

------------------------------------------------------------------------------------

OLS OLS robust

linear

Baseline OLS Stepwise Lasso Robust

(1) (2) (3) (4)

---------------------------------------------------------------------------------------------------------

Age 0.002\*\*\* 0.002\*\*\* 0.001\*\*\* 0.001\*\*\*

(0.001) (0.001) (0.0005) (0.0004)

Gender -0.01 -0.003 0.0003

(0.01) (0.01) (0.01)

Sleep.duration -0.01 -0.01 -0.01

(0.01) (0.01) (0.01)

Caffeine.consumption 0.0002 0.0002 0.0001

(0.0002) (0.0002) (0.0001)

Alcohol.consumption -0.04\*\*\* -0.04\*\*\* -0.03\*\*\* -0.03\*\*\*

(0.01) (0.005) (0.004) (0.004)

Smoking.status -0.11\*\*\* -0.11\*\*\* -0.08\*\*\* -0.07\*\*\*

(0.02) (0.02) (0.01) (0.01)

Awakenings -0.06\*\*\* -0.06\*\*\* -0.04\*\*\* -0.05\*\*\*

(0.01) (0.01) (0.004) (0.004)

Exercise.frequency 0.02\*\*\* 0.02\*\*\* 0.01\*\*\* 0.01\*\*\*

(0.005) (0.005) (0.004) (0.003)

Constant -0.10 -0.19\*\*\* 0.90\*\*\* 0.90\*\*\*

(0.07) (0.03) (0.05) (0.05)

---------------------------------------------------------------------------------------------------------

Observations 312 312 312 312

R2 0.51 0.51 0.53

Adjusted R2 0.50 0.50 0.52

Residual Std. Error 0.13 (df = 303) 0.13 (df = 306) 0.09 (df = 303) 0.08 (df = 303)

F Statistic 40.10\*\*\* (df = 8; 303) 63.77\*\*\* (df = 5; 306) 42.62\*\*\* (df = 8; 303)

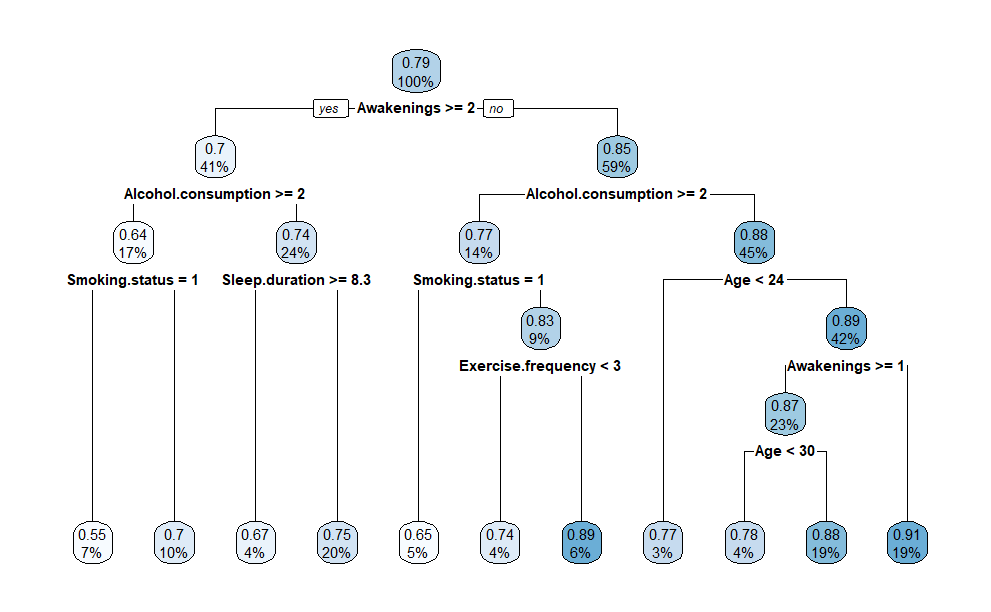
=========================================================================================================

It is possible to notice from this table that the four models’ performance do not differ too much. The significant variables are almost the same for each model. There seems to be a certain share of which variables are the most significant ones among all the models. In conclusion, it can be affirmed that the age of the individual’s is one of the most important variables in explaining the variability of sleep efficiency. Smoking status is the variable that affect the efficiency of sleep the most, followed by awakenings and alcohol consumption. Exercise frequency appears to be the one apart from the age having a positive relationship with sleep efficiency.

# 3.2 Decision Tree

The previous part of the study focused on the inference, now perhaps the analysis should be moved towards models with less interpretability and more predictability. To pursue that two different alternative methods were induced. These techniques could capture non-linearity, present in our dataset, and interaction between our variables. The first method is the Decision Tree. A tree model was built using the train data, predictions and metrics were defined. As metrics RMSE, R-squared and MAE were chosen. Here following their estimates:

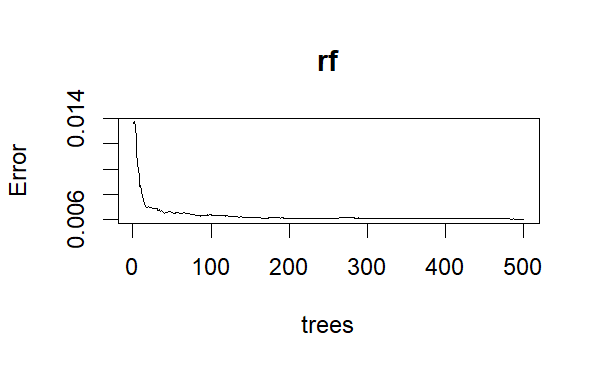
| **metric**  <chr> | **.estimator**  <chr> | **.estimate**  <dbl> |  |  |
| --- | --- | --- | --- | --- |
| rmse | standard | 0.10152112 |  |  |
| rsq | standard | 0.51950920 |  |  |
| mae | standard | 0.07337634 |  |  |



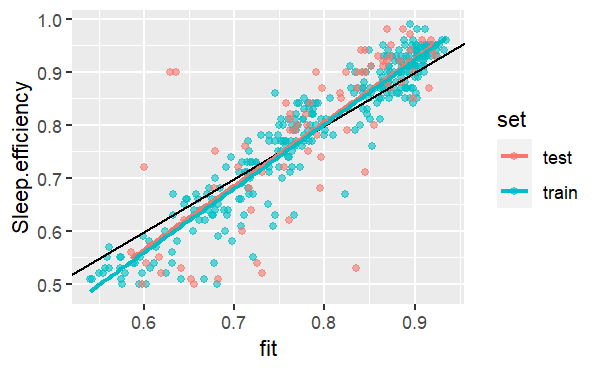
It is clear from the above table that the R-squared of 0.51 can be considered reasonably good, it suggests that a 51% of the variance in the dependent variable is accounted for by the independent variables in the model. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individiual differences have equal weight. The MRSE is the standard deviation of the residuals (prediction errors), it measures how spread the residuals are. In the tree plot above it is possible to spot Awakenings as the best discriminant variable, followed by Alcohol consumption. Almost all the variables were taken into consideration by the previous analysed models, apart from Sleep duration that now is taken into account.

# 3.3 Random Forest

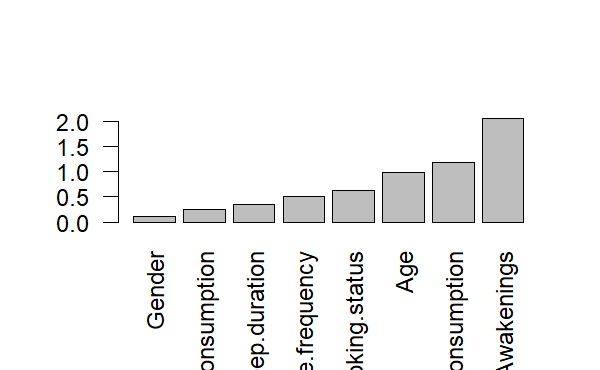
Tree based models can better perceive nonlinear effects, for this reason Random Forest was chosen to be applied. It combines multiple decision trees to make prediction through the bootstrapping process, which follows a random selection of observations from the original dataset to make the new subsets trees. It directly selects a random sample of features and then predictions are constructed. The final prediction of the Random Forest is the combination of predictions from all individual trees. The following plot shows the number of trees that were generated by the algorithm:



Random Forest can improve accuracy compared to a single decision tree and it is less sensitive to noise data. It also can measure the importance of each variable in the prediction process. The correlation between the training set and the test set is showed here:



Here the following variables of importance table:



The variable that was considered the most important by the algorithm was Awakenings, followed by Alcohol consumption and other variables as Age that were still considered significant by the other models. Regarding the differences between Tree based models and Linear models in this case, it can be said that they are not so valuable and relevant. The models kind of perform the same way.

Lastly, a table to show the comparison between all the models is displayed:

| **Model**  <chr> | **RMSE**  <dbl> | **R2**  <dbl> |  |  |
| --- | --- | --- | --- | --- |
| OLS | 1.06631253 | 0.5142884 |  |  |
| Stepwise | 1.06553718 | 0.5102914 |  |  |
| Lasso | 0.09802278 | 0.5294765 |  |  |
| Ridge | 0.11607959 | 0.5294765 |  |  |
| Tree | 0.10152112 | 0.5163844 |  |  |
| RandomForest | 0.09420890 | 0.5835419 |  |  |

# 3.4 Conclusions of the supervised part

From the last findings, the results of the Random Forest were the better ones in terms of R2, meaning that tree-based models can predict better target variables. In order to increase sleep efficiency, it was clear that individuals must decrease the times of awakenings during their sleep, since it was recorded as most important variable among all the others.

Concerning performance, Linear models and Tree based models do not differ much. The Lasso was the one that performed best, as well as the Ridge, in MRSE instead the models with a smaller value have a better predictive accuracy, like Random Forest, followed by the Lasso and Decision Tree. It was not expected that the Lasso provided more accurate predictions compared to the Decision Tree, because of the assumption that they work best with nonlinear data. Probably it is due to the feature selection of the Lasso that help reduce overfitting and improve prediction accuracy.

Unsupervised learning project

How European Countries react to their environments

Chapter 4

# 4.1 The dataset

This part of the work focuses on the application of some unsupervised learning techniques to try to discover some interesting insights. The dataset is different from the one chosen for the supervised part. It was retrieved from the official website and data warehouses of Eurostat, indeed it regards some European Countries. The variables were selected personally, based on the goal to identify different groups of countries from how they perceive their overall environment and its effects on citizens’ mental health. As follows a description of the data.

Countries:

European Union – 27 countries (from 2020)

Belgium

Bulgaria

Czechia

Denmark

Germany

Estonia

Ireland

Greece

Spain

France

Croatia

Italy

Cyprus

Latvia

Lithuania

Luxembourg

Hungary

Malta

Netherlands

Austria

Poland

Portugal

Romania

Slovenia

Slovakia

Finland

Sweden

Iceland

Liechtenstein

Norway

Switzerland

United Kingdom

Montenegro

North Macedonia

Albania

Serbia

Turkey

These are the 39 countries chosen for the analysis. The dataset takes the form of cross-sectional data since all the variables were chosen from the same year (2021). Here attached the related variables:

Life expectancy at birth by sex = defined as the mean number of years still to be lived by a person at birth -, if subjected throughout the rest of his or her life to the current mortality conditions.

Share of people with good or very good perceived health by sex = subjective measure on how people judge their health in general on a scale from "very good" to "very bad". It is expressed as the share of the population aged 16 or over perceiving itself to be in "good" or "very good" health. The data stem from the EU Statistics on Income and Living Conditions (EU SILC). Indicators of perceived general health have been found to be a good predictor of people’s future health care use and mortality.

Exposure to air pollution by particulate matter (source EEA) = measures the population weighted annual mean concentration of particulate matter at urban background stations in agglomerations. Fine and coarse particulates (PM10), i.e. particulates whose diameters are less than 10 micrometers, can be carried deep into the lungs where they can cause inflammation and exacerbate the condition of people suffering heart and lung diseases. Fine particulates (PM2.5) are those whose diameters are less than 2.5 micrometers. They are therefore a subset of the PM10 particles. Their deleterious health impacts are more serious than PM10 as they can be drawn further into the lungs and may be more toxic.

Self-perceived health by sex, age and income quintile = measured in percentage

Depressive symptoms = measured in percentage

Average rating of life satisfaction = measured in rating from 0 to 10

Employment rates = measured in percentage

Trust in the legal system = measured in rating from 0 to 10

Crime, violence or vandalism in the area = measured in percentage

Social support = persons who have someone to ask for help by sex, age and educational attainment level, measured in percentage

The dataset target variable was Self-perceived health, which was removed to apply the unsupervised learning techniques. After having uploaded the dataset some data pre-processing was applied and then, two unsupervised techniques were used: PCA and Hierarchical Clustering.

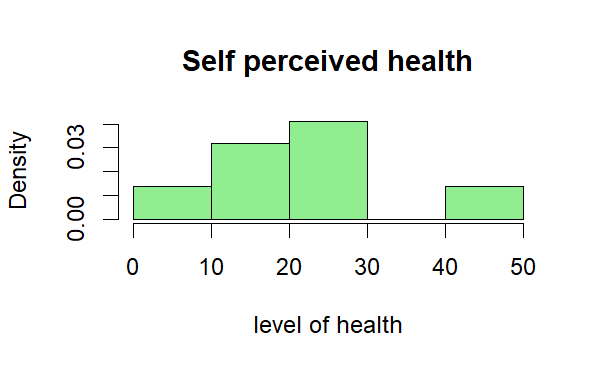
# 4.2 Pre-processing and correlation

The dataset presented some NA values, the first and last columns were removed since they were not relevant for the analysis. This is a sample of the dataset after the cleaning:

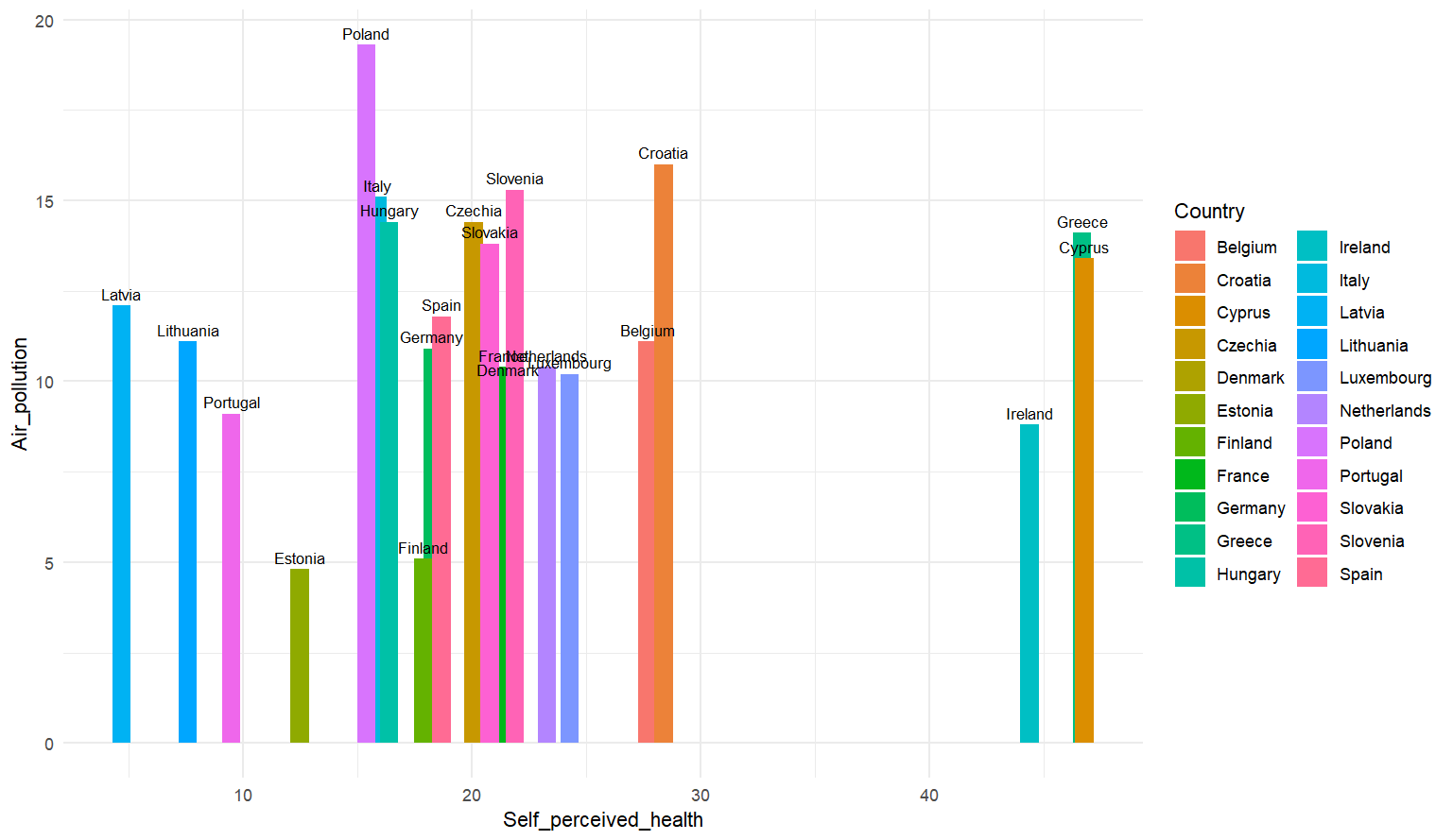
Immagine che contiene testo, numero, schermata, Carattere

Descrizione generata automaticamente

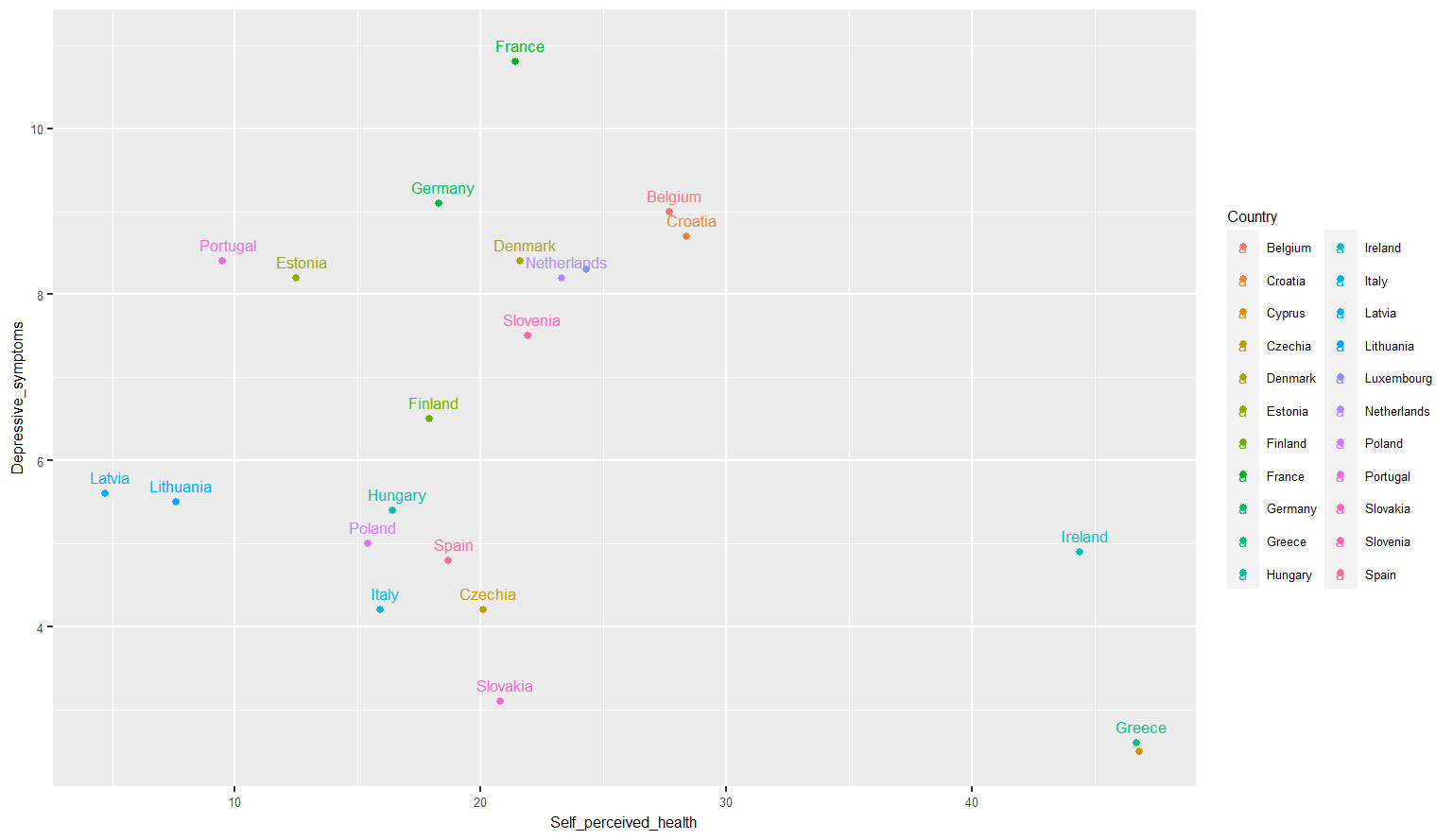
Since self-perceived health is our target, it was taken for a closer analysis. It is possible to notice in the following histogram that higher density of Self perceived health regards values between 20 and 30. Only few countries have a self-perception between 40 and 50, conversely to the values between 30 and 40 which are null. This shows that there is a gap, the distribution of the data is not continuous, and it is probably due to certain values that were not included in the analysis after the remotion of some data to obtain an accurate analysis and avoid biased results. If NA values are present, they can lead to misleading conclusions and leaving them could compromise the integrity of the dataset.



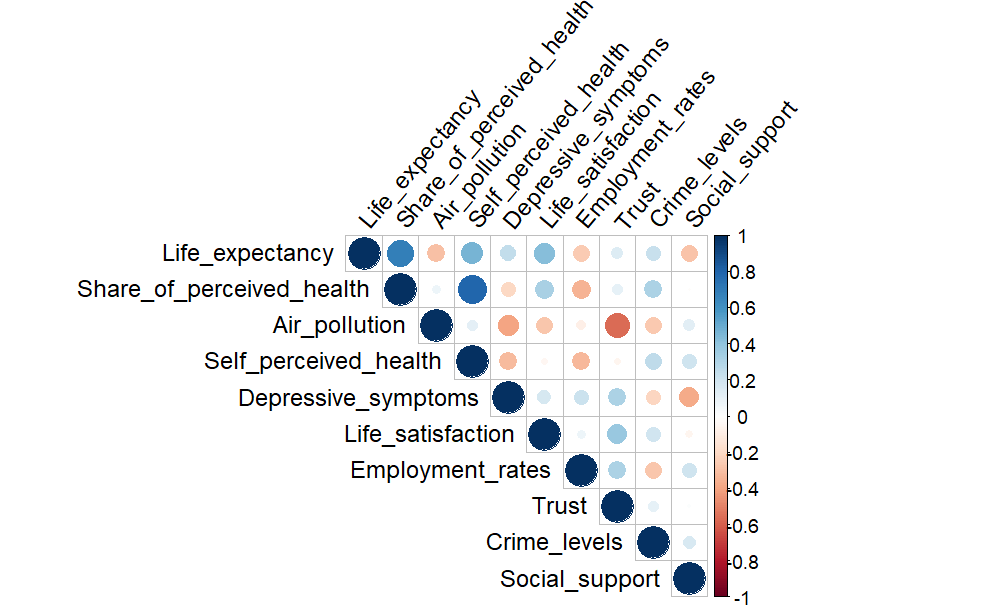
In the second plot, whereas Self-perceived health is examined in relation with Air pollution to explore their potential association. It is possible to distinguish that the countries with the highest health perception, more than 40, are Cyprus, Greece and Ireland. On the other hand, the values of Air pollution for these top 3 countries are different from each other. Ireland settled as best countries having the highest self-perceived health and less percentage of air pollution than the other countries. As worst country for highest percentage of Air pollution there is Poland, distancing itself from all the other European countries. However, at the lowest in terms of Self-perceived health there is Latvia followed by Lithuania. It is challenging to pinpoint specific causes, but it could be because of various factors, for example: socioeconomic conditions, including poverty, income inequality, unemployment rates, the quality and accessibility of healthcare services, health-related behaviours, such as smoking, alcohol consumption, physical activity, and factors like depression, anxiety regarding the mental health of people, as well as air pollution.



For this reason, here it is displayed a plot showing the relationship between depressive symptoms and self-perceived health. It is clearly that the top three countries highlighted before are confirming their position, having low values of depressive symptoms. French people are the ones suffering most of depression, as it is possible to determine by its position, which is quite surprising. The reason behind it could rely on the access to healthcare services and how these symptoms are detected and reported, but it could also come from societal and lifestyle factors, for instance work-related stress and social support networks.



To conclude, life satisfaction levels were observed: Belgium, Czechia and Denmark got the highest levels which means that the overall satisfaction perceived by their citizens are positive and regard work, family, ambitions and other life aspects. Outliers were identified with the use of boxplots. The boxplots show that there are few outliers, 1 reported from Employment rates and 3 from Self perceived health, which is our target variable, so it is going to be removed to process the unsupervised techniques. The next step was to identify the correlation between all the variables. It is noticeable at first glance that Life expectancy has a certain correlation with the share of perceived health and this one with Self perceived health.



Chapter 5

# 5.1 PCA

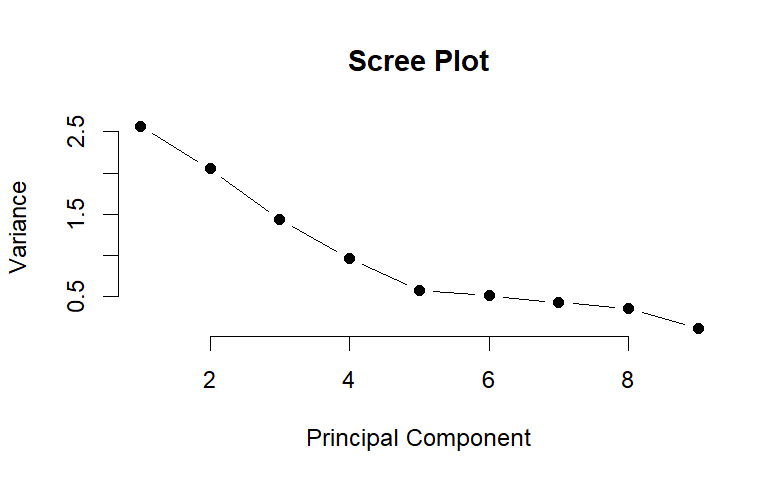
The first unsupervised technique that was implemented is the PCA. As previously stated, the target variable was removed from the dataset. The dataset presented 22 rows and 10 columns, mean and standard deviation of all the variables were computed. PCA was performed with the *prcomp* function, and the data was transformed into a new set of variables, the principal components, which capture the maximum amount of variation in the data. The first component explains the most variability and the second the second most variability and so on. From performing PCA The Netherlands, Finland, Denmark, Luxembourg got closer values in Life satisfaction and Trust. It is known that these countries are relatively prosperous countries with high standards of living. They have strong economies, high income levels, social welfare systems, strong governance and institutions that may contribute to higher life satisfaction and trust in the legal system. Estonia, Germany, France and Portugal seemed closer in terms of Depressive symptoms and Employment rates. These countries could probably share same labour market conditions and similarities in the design and effectiveness of social support programs, and other factors.

The following plot shows the biplot, in other words the result of the PCA:

Immagine che contiene testo, diagramma, linea, Diagramma

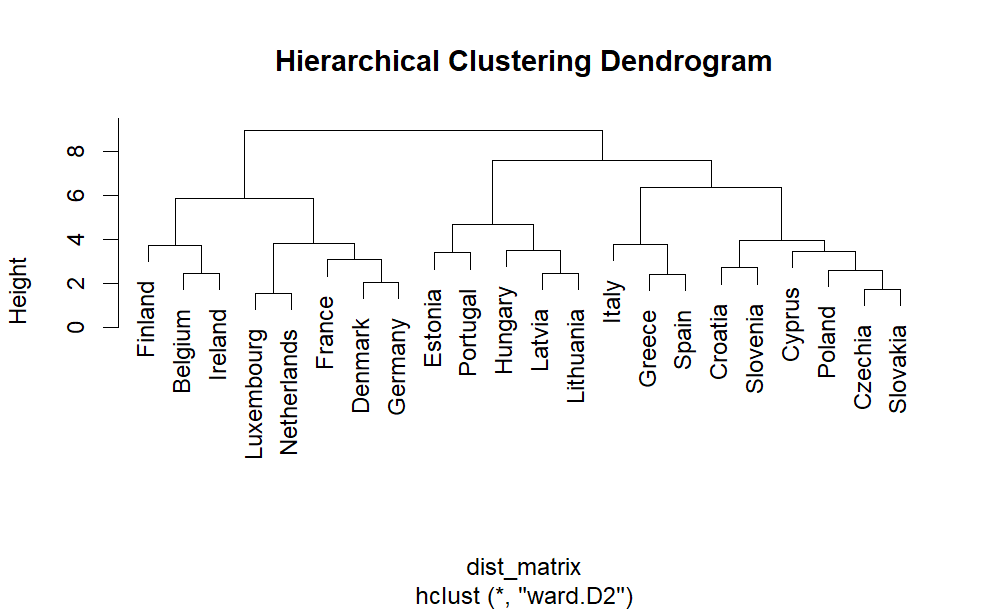
Descrizione generata automaticamente

Next, the proportion of variance explained was computed and from the plot it is clear how from the sixth principal component the proportion decreased until it got to zero. The scree plot perfectly explains this relationship between the principal components and the variance:



# 5.2 Clustering

In this last step of the project the other unsupervised technique is applied: Cluster analysis. In order to perform Clustering, the categorical variable Country was reintroduced in the dataset, since the algorithm can deal with both quantitative and qualitative variables. It was decided to standardise the distances to better ensure the predictability of the analysis. The distances were computed with the Euclidean method. It quantifies the distance between two points in a multidimensional space and in clustering it is used to address similarity and dissimilarity between data points or observations. This means that it determines how distant or different are two points from each other in terms of their feature values. The Ward linkage method was implemented as it is a hierarchical agglomerative clustering method used to build dendrograms that represent the structure and relationships between the data point. The algorithm proceeds in identifying each data point as an individual cluster and then it creates groups that have the smallest increase in total within cluster variance. Ward’s method gives more consideration to dissimilarities between clusters.



The result is compact and well separated clusters as it is possible to see in the above plot. Other distances were used to see the differences in clustering the data points such as Manhattan and Minkowski. Two methods were also applied: the complete linkage and the average linkage. The first one creates clusters determining the maximum distance between two data points and it is based on the most dissimilar pair of points found between them. The second one creates clusters based on the average of all pairwise distances between the points of one cluster and the points of the other one. This last one, in particular, can handle better variations in within-cluster similarities, since it is less sensitive to noise. The following plot shows its clustering performance:

Immagine che contiene testo, diagramma, Carattere, schermata

Descrizione generata automaticamente

Up till here only dissimilarities measures were presented. A similarity one that was used is the Gower’s index. It can handle better a combination of numerical and categorical values. Finally, the number of clusters was computed:

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, diagramma, Piano, Disegno tecnico

Descrizione generata automaticamente

# 5.3 Conclusions of the unsupervised part

As final findings, it can be said that both PCA and hierarchical clustering performed well in representing the relationships between various countries based on their features and in representing the hierarchy of groups. PCA helped with identifying the most influential variables in the dataset as Life expectancy and to visualise the relationships of the data in a reduced dimensional space. Clustering revealed patterns of similarity and dissimilarity among the countries. Clusters that are close to each other on the dendrogram have higher similarity, for instance Italy, Spain and Greece or Luxembourg and The Netherlands. All countries everyone would expect to position closer to each other, due to their geographical proximity and share of customs and lifestyle habits. Although, some other discoveries were achieved. A personally ambiguous cluster is represented by Estonia and Portugal. They are both European countries, albeit located to opposite ends of the continent, and the reason of this group could come from socio-economic factors or similarities in social aspects.

# References

Link dataset for Supervised part: <https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency>

Link Eurostat website for Unsupervised part:

<https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_MH1I$DV_464/default/table?lang=en>

This work is based on the lectures of the module Statistical Learning provided by Professor Silvia Salini from The University of Milan and some contents of the book ‘An Introduction to Statistical Learning’ by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Thank you for reading.