Analysis of various macroeconomic characteristics

Tin Ferković   Luka Ilić   Fani Sentinella-Jerbić   Ana Terović

18 January 2021

Contents

1 INTRODUCTION

Macroeconomics examines economy as a whole and is an integral part in governmental decision making on both a national and global scale. Our dataset contains various macroeconomic features, such as GDP per capita and employment rates, for 66 countries.

The main topics covered are examining distributions of various features and whether there are significant outliers, relation between employment rates and GDP, and which variables explain differences in life expectancy. Considering the current global state, we also tried to focus on features which possibly affect climate change and health expenditure (during a pandemic).

Through the analysis we compared Europe to the rest of the world as well as different European regions between each other.

We used descriptive statistics, t-test, ANOVA and also examined linear dependencies through simple and multivariate regression.

Ending consists of a conclusion about everything we have done and comments on the dataset and possible future work.

1.1 Initial mining

Our dataset requires some initial cleaning before doing any analysis. We have detected some dirty data that had to be inspected.

For example, number of individuals using the Internet per 100 inhabitants:

```
##  [1]  256  948  118   25   37   91  990  104  122  197   64 1080  835  176   72
## [16]   53   47  156   23   36  278  116  374   66 1052 1281  134   50  174  359
## [31]  404  113   87   26 1272 1162   40  199   64  140  783   58  281   39  111
## [46]  104  235  131   71  293   54  143  581  617  587   54   74  611  110  388
## [61]  102   56  102 1513  328  616
```

Such features contained too many nonsensical values so we concluded that it makes sense to remove them from the dataset. It was also necessary to split some of the features whose values were of shape *value*/*value* into separate features as well as convert them to proper types.

```
dataset[dataset == -99] <- NA
```

For the sake of convenience, we also replaced all the cells containing value -99 with NA. Basically, in this dataset, -99 is a replacement for NA. Lastly, we converted to numeric all the values convertible to numeric. For the ones which can't be converted, we generated NA's.

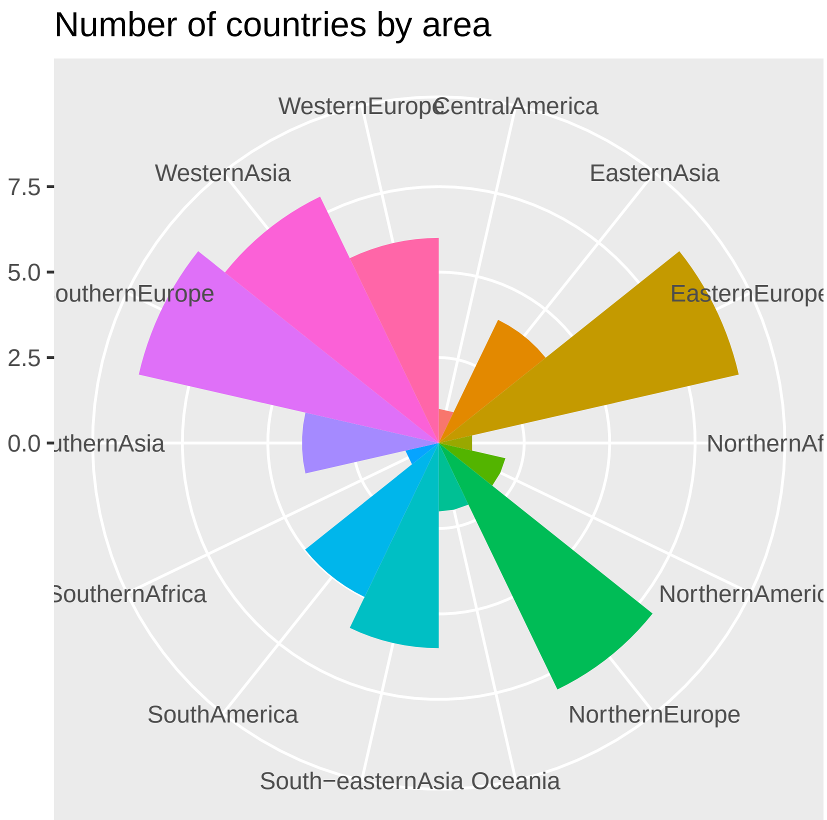
1.2 Descriptive statistics

Descriptive statistics, in short, help describe and understand the features of a specific dataset by giving short summaries about the sample and measures of the data. As a good intro to more complex topics, we are here presenting a general overview of our dataset.

```
##  Number of rows and columns, respectively:  66 96
```

One can see that our dataset is quite peculiar, having more columns than rows. We will keep this in mind through further analysis.

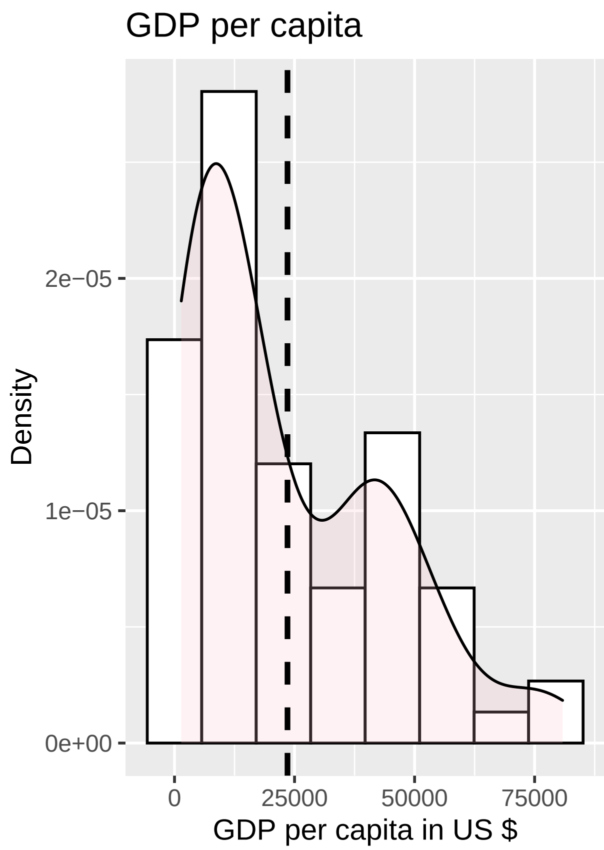
We want to see which parts of the world are represented in our dataset. Here we are using a polar graph for visualization.

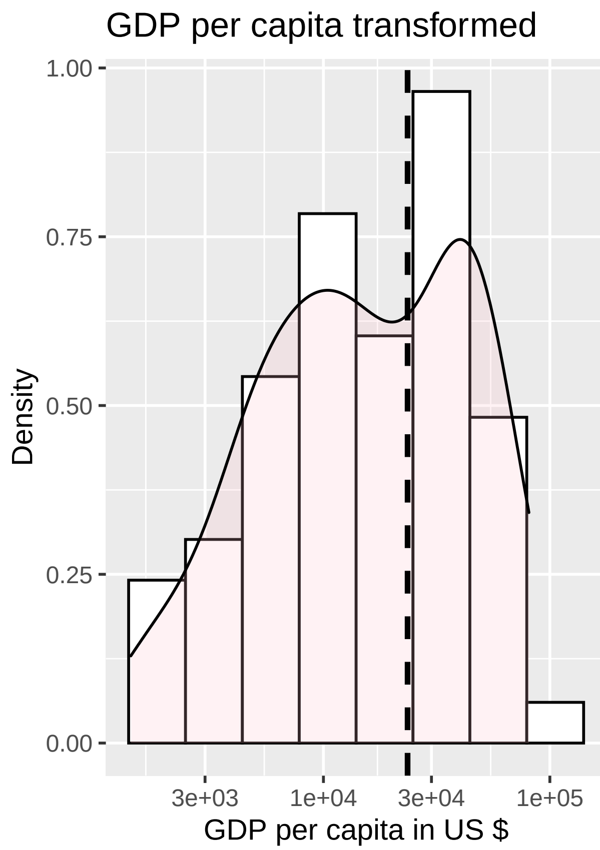


1.2.1 Distributions

Wanting to examine the distributions across all countries, we shall plot multiple histograms with mean as a measure of central tendency as well as the density to get inspiration for further analysis.

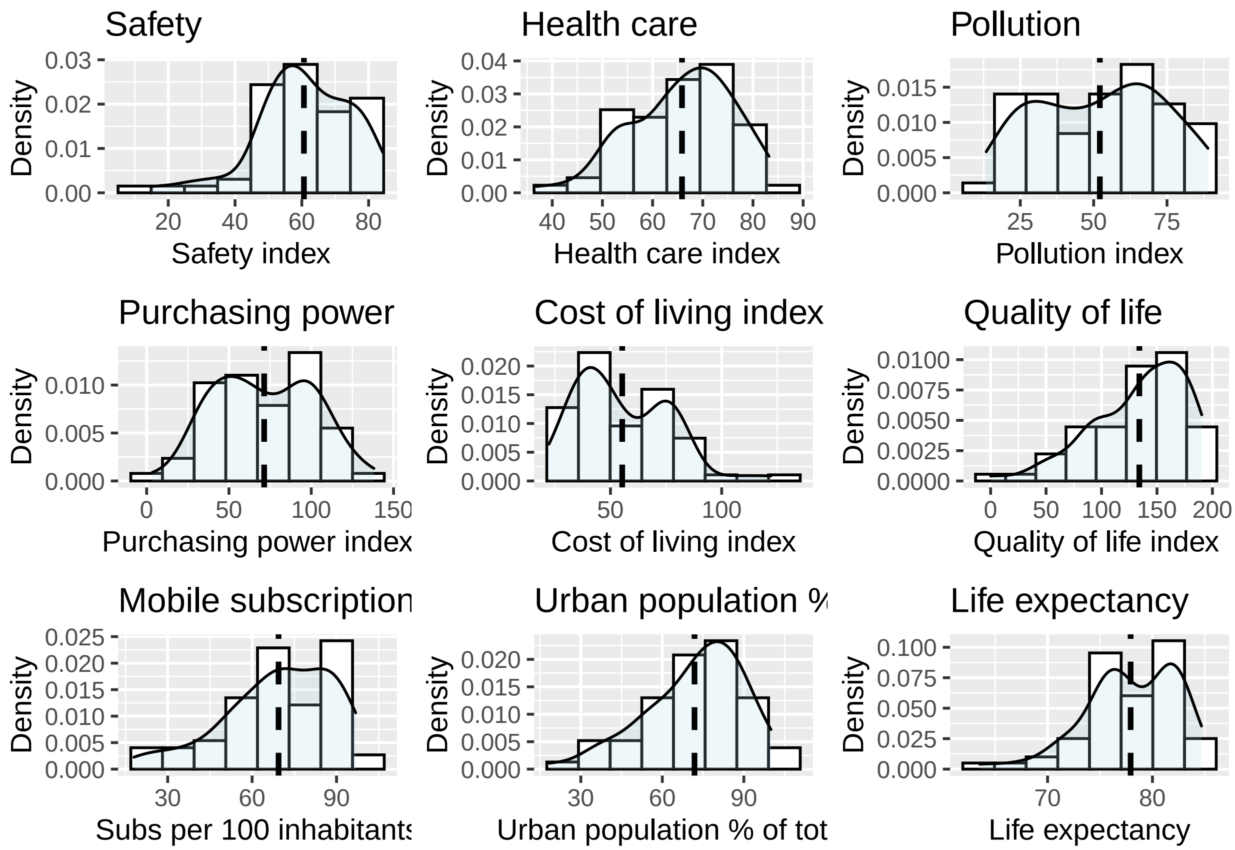
One of the most common indicators of a country's well being is its GDP, so we plot the histogram of GDP *per capita* as it generally delivers more of a prosperity measure than the total GDP. We are expecting to see a smaller number of countries with a large GDP *per capita* and a larger amount of countries with small or average GDP.





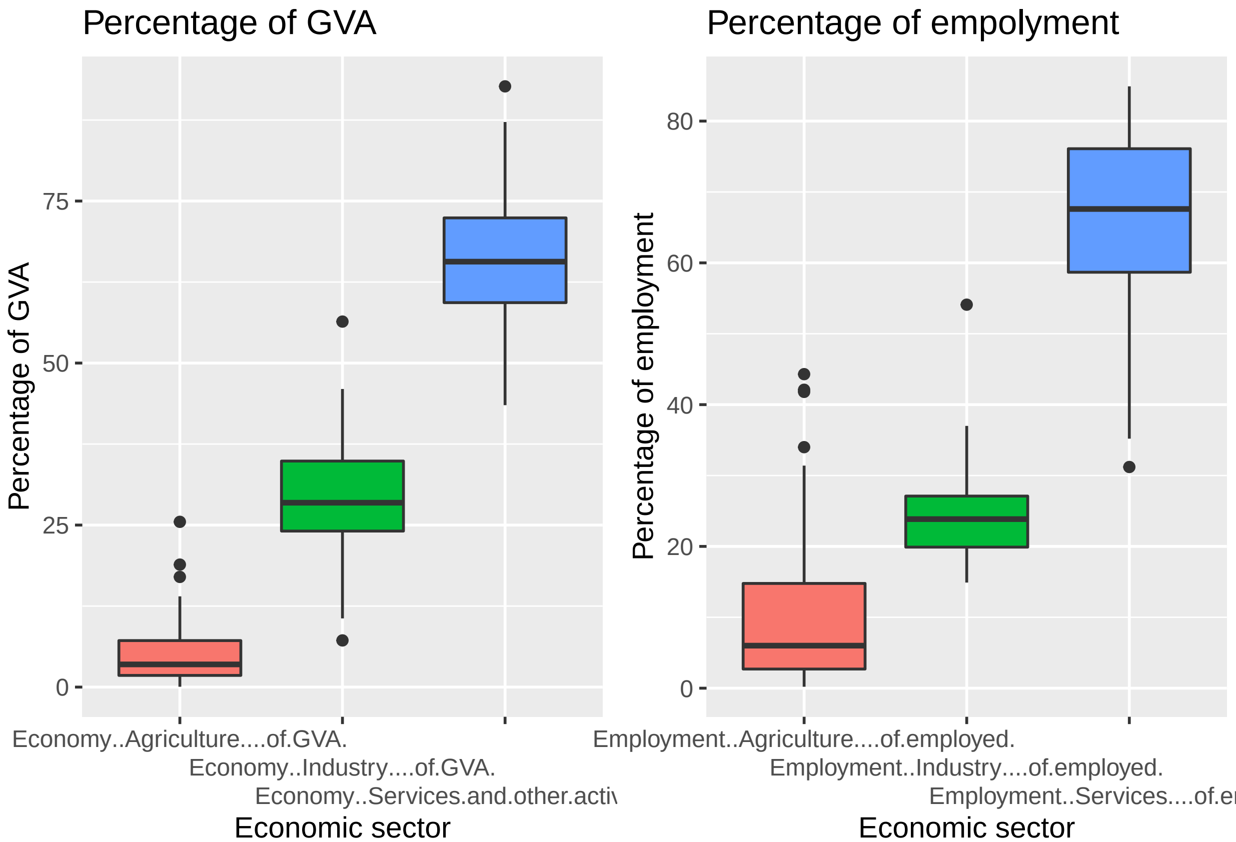
As we have assumed, there are more countries with lower GDP *per capita* than those with a high GDP *per capita*. Considering we have an asymmetrical distribution, we decided to plot the mean as an expression of central tendency as seen on the graph in the shape of a dashed line. However, in order to run any significant tests one should check for normal distribution so we decided to also plot the log transformed histogram. Unfortunately, the distribution was not normal even after transformation.

Furthermore, we have plotted various other features in order to get a grip of the way our data behaves and gain some intuition.



As it can be seen from the graphs, most of the features are not normally distributed. That, along with the fact that there are only 66 countries, makes some statistical hypotheses and conclusions more difficult - or as we see it, more challenging.

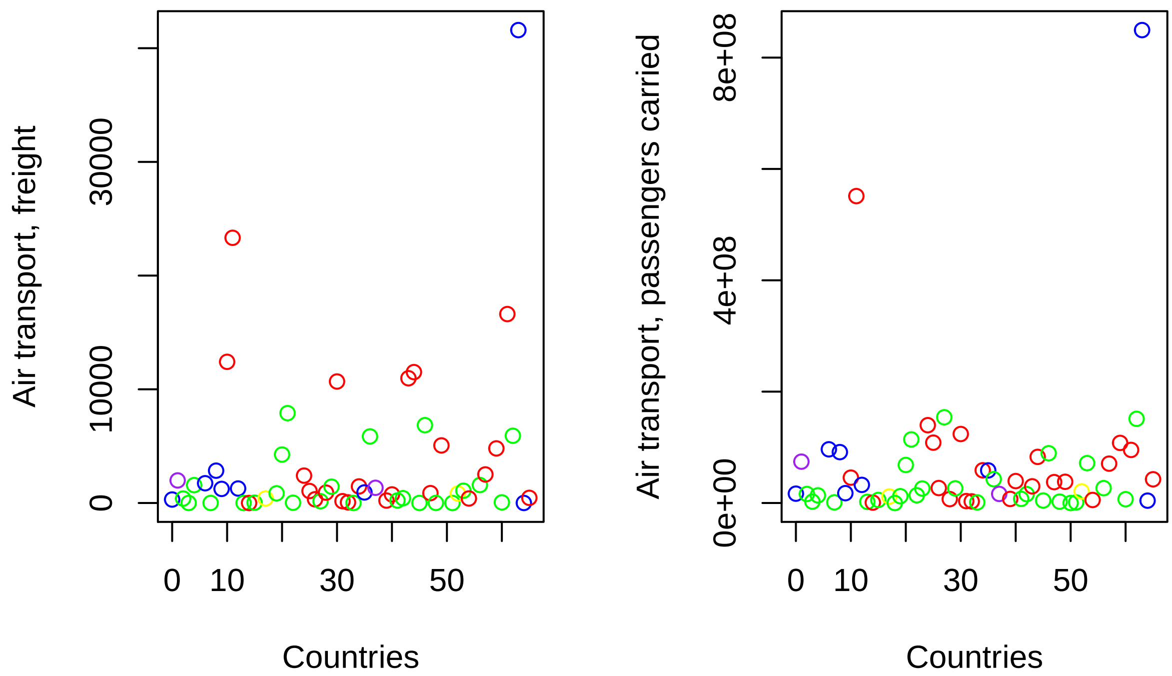
Since we also want to focus on the factors which are somewhat correlated to the GDP throughout our further analysis, we want to check out some candidate variables and their properties. Here we can see the box plot showing contribution to GVA and employment percentage by each sector of economy.



A few interesting notes can be taken from this. Although the agriculture contribution to the GVA is quite small, there seem to be more more people working in the sector than one might expect. Industry seems to have a bigger influence on the GVA considering how many people are working in the field. This can naturally be attributed to the fact that there are factory machines doing the work. We can also see that the employment in the services sector varies the most out of the three sectors.

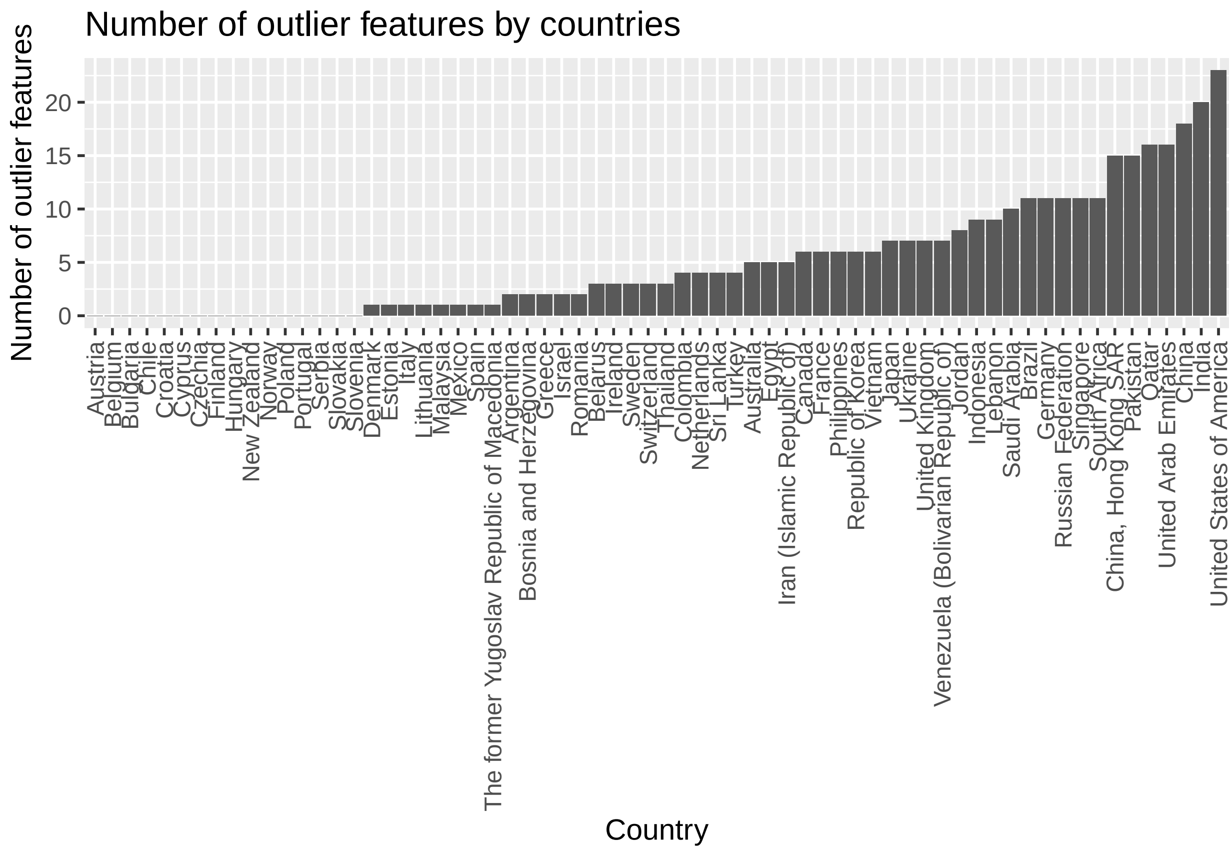
1.2.2 Significant outliers

Interesting part of the initial exploration is finding outlier values. For example, the US is an outlier in all air transport, far surpassing the competition as seen on the following graph:



Qatar is also an outlier worth of mentioning, but it will be later discussed throughout the paper.

We took the next step and automatized the process of finding the outliers in order to see which countries have the most outliers across all parameters.



We can see that the USA has outlier values in most categories out of given countries. One can also notice that Croatia does not have outlier values in any of the given features. Despite that, in the rest of the document it will be colored red in order to see how it is ranks in comparison to other countries.

2 TESTING HYPOTHESIS

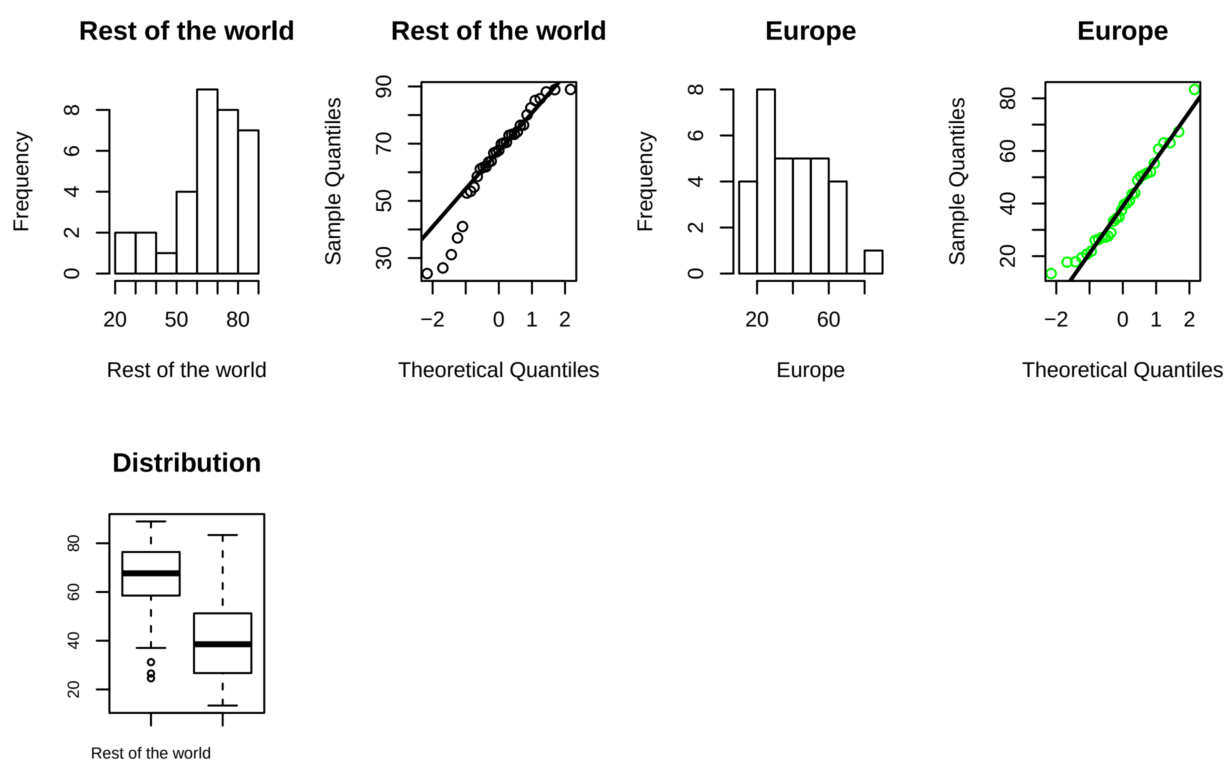
In this section we will test different assumptions using the t-test.
In our t.test function we check the distribution of the data by drawing graphs and we also check whether the variances are equal to be able to conduct the internal R supplied t.test. When making a decision we look at the p-value. If the p-value of the t-test is smaller than confidence interval we can reject the null hypothesis in favor of the alternative hypothesis and if the p-value is bigger we cannot reject null hypothesis.
Lets analyze Europe compared to other world countries while focusing on current world issues.

2.1 Climate change

2.1.1 Assumption: Pollution in Europe is lower than in the rest of the world.

```
myTtest(rest_of_the_world$Pollution.index,europe$Pollution.index, "less", FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = -5.9752, df = 63, p-value = 5.848e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -18.37111
## sample estimates:
## mean of x mean of y
##  39.65656  65.15030
```

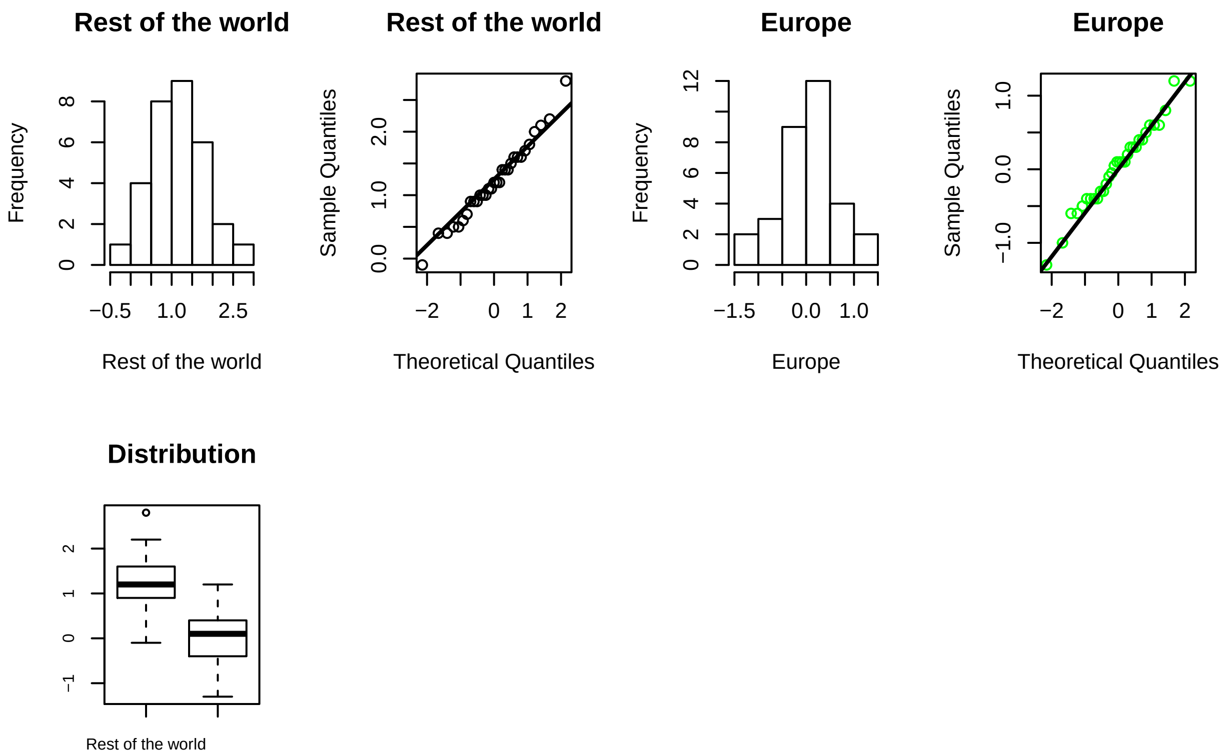


Based on the p-value we can reject the null hypothesis in favor of the alternative hypothesis meaning that the pollution in Europe is lower than that in the rest of the world.

2.1.2 Assumption: Population growth in Europe is lower than in other countries.

```
myTtest(rest_of_the_world$Population.growth.rate..average.annual...,europe$Population.growth.rate..avera
```

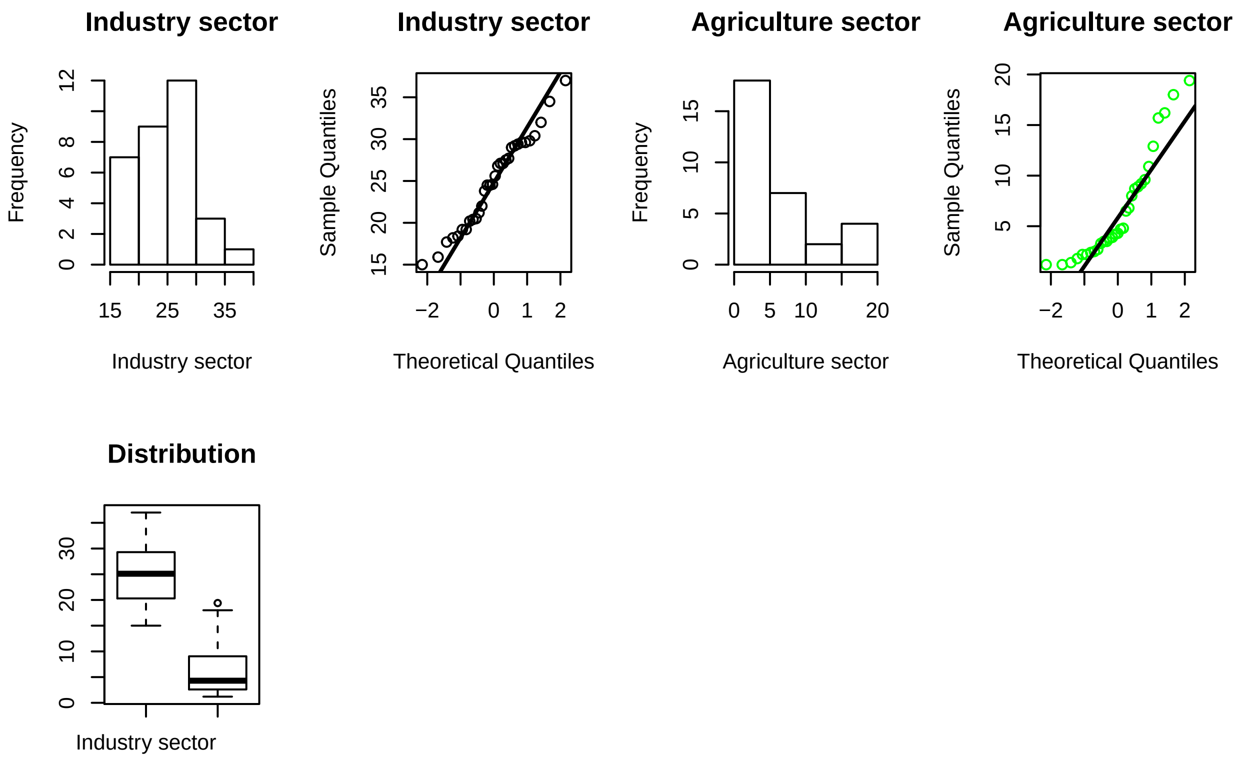
```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = -7.8667, df = 60.293, p-value = 4.013e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.9233406
## sample estimates:
## mean of x mean of y
##  0.040625  1.212903
```



Based on the p-value we can reject the null hypothesis in favor of the alternative hypothesis meaning that the population growth in Europe is slower than that in the rest of the world.

2.1.3 Assumption: Based on the fact that the pollution in Europe is lower than in the rest of the world we are making an asumption that the amount of people working in Agriculture sector is much higher than in the Industry sector.

```
##
##  Paired t-test
##
## data:  europe$Employment..Agriculture....of.employed. and europe$Employment..Industry....of.employed
## t = -13.996, df = 31, p-value = 3.038e-15
## alternative hypothesis: true difference in means is less than 0
## 90 percent confidence interval:
##       -Inf -16.08081
## sample estimates:
## mean of the differences
##               -17.74063
```



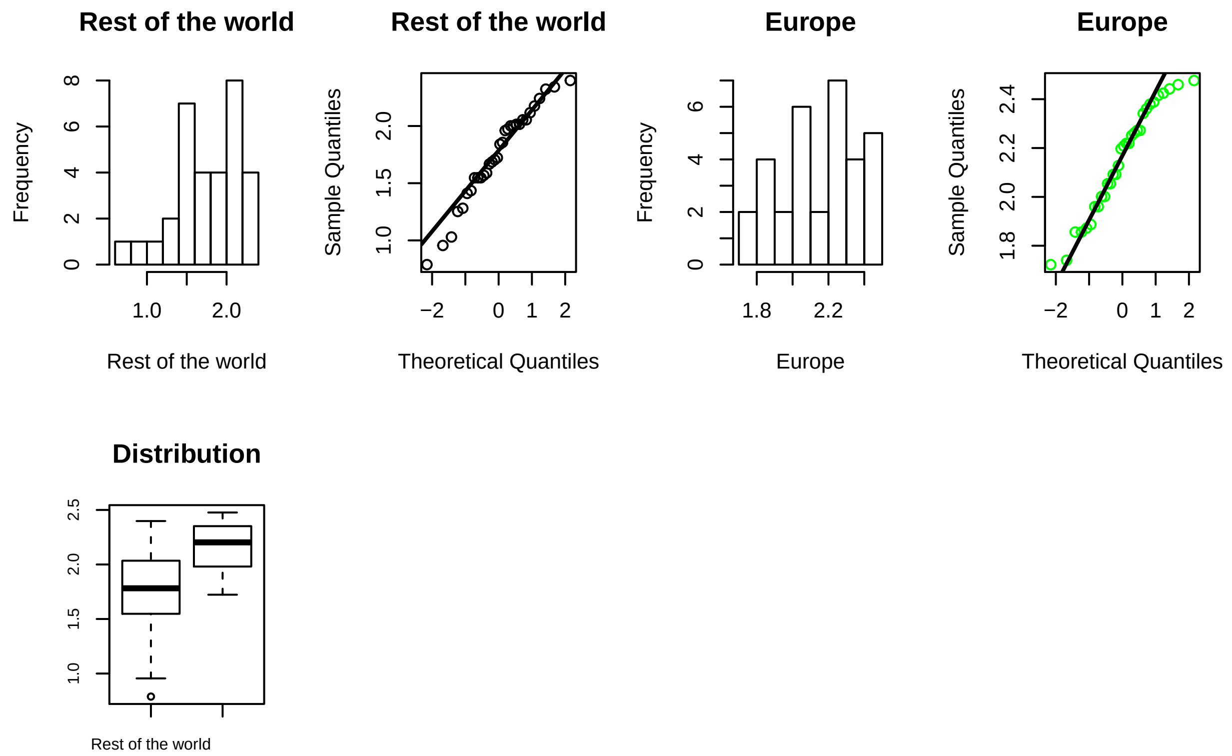
Based on the p-value we can reject the null hypothesis meaning that the amount of people working in Industry sector in Europe is higher than in Agriculture.

2.2 Corona virus

2.2.1 Assumption: Health expenses in Europe are greater than in the rest of the world.

```
myTtest(log(rest_of_the_world$Health..Total.expenditure....of.GDP.),log(europe$Health..Total.expenditur
```

```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = 4.8646, df = 47.629, p-value = 6.474e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2614136       Inf
## sample estimates:
## mean of x mean of y
##  2.152170  1.753167
```

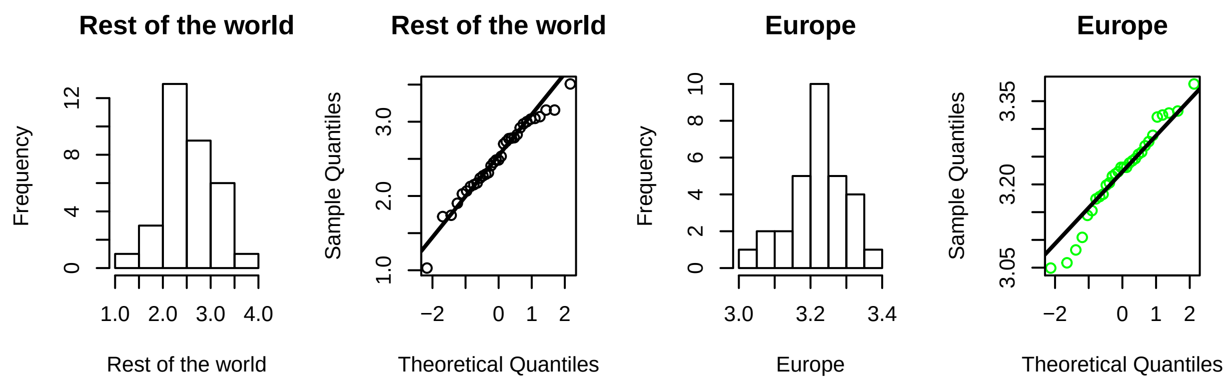


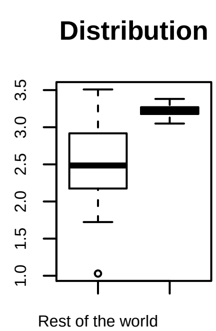
Based on the p-value we can reject the null hypothesis meaning that health expanses in Europe are higher than that in the rest of the world.

2.2.2 Assumption: There are more older people in Europe than in the rest of the world.

```
myTtest(log(rest_of_the_world$Population.age.distribution.60.years....),log(europe$Population.age.distri
```

```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = 7.783, df = 33.727, p-value = 2.472e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5554957       Inf
## sample estimates:
## mean of x mean of y
##  3.221315  2.511592
```





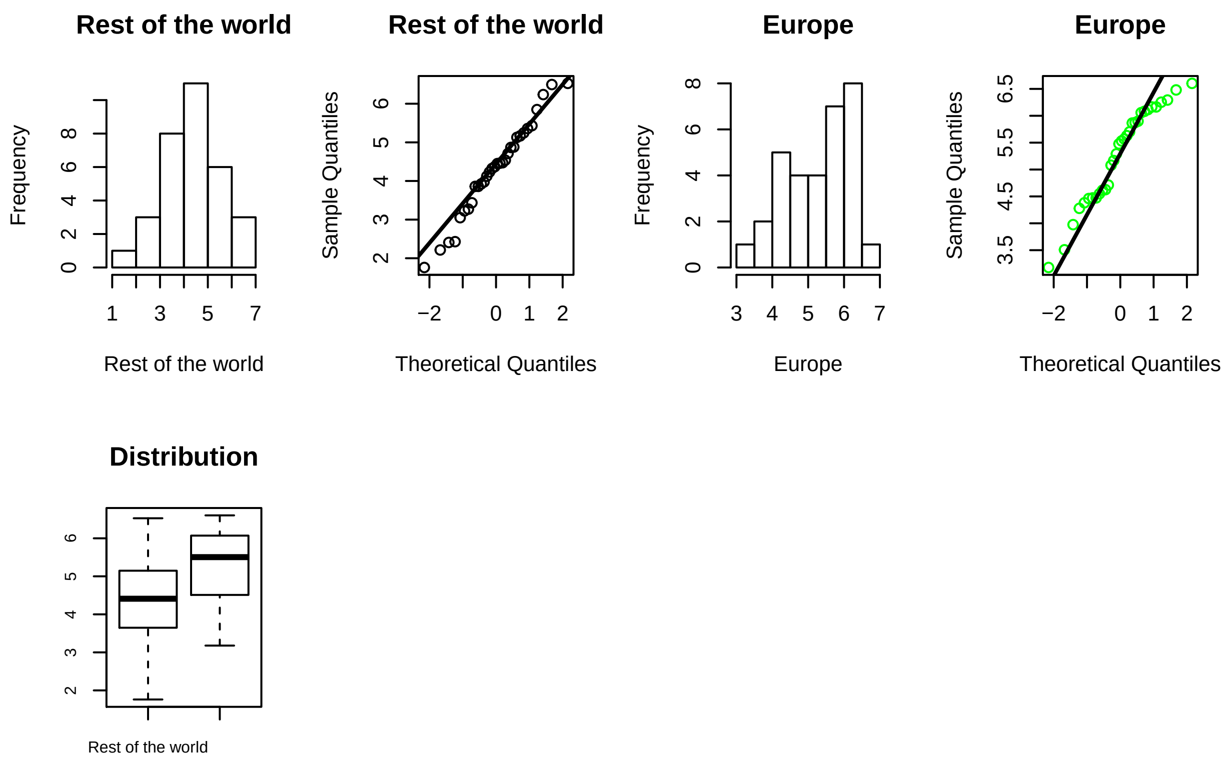
Based on the p-value we can reject the null hypothesis in favor of the alternative hypothesis meaning that there are more older people in Europe than in the rest of the world.

2.3 Gender equality

2.3.1 Assumption: There are more women in parliament in Europe than in the rest of the world.

```
myTtest(sqrt(rest_of_the_world$Seats.held.by.women.in.national.parliaments..),sqrt(europe$Seats.held.by
```

```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = 3.5649, df = 57.453, p-value = 0.0003706
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5019913       Inf
## sample estimates:
## mean of x mean of y
##  5.266447  4.321142
```

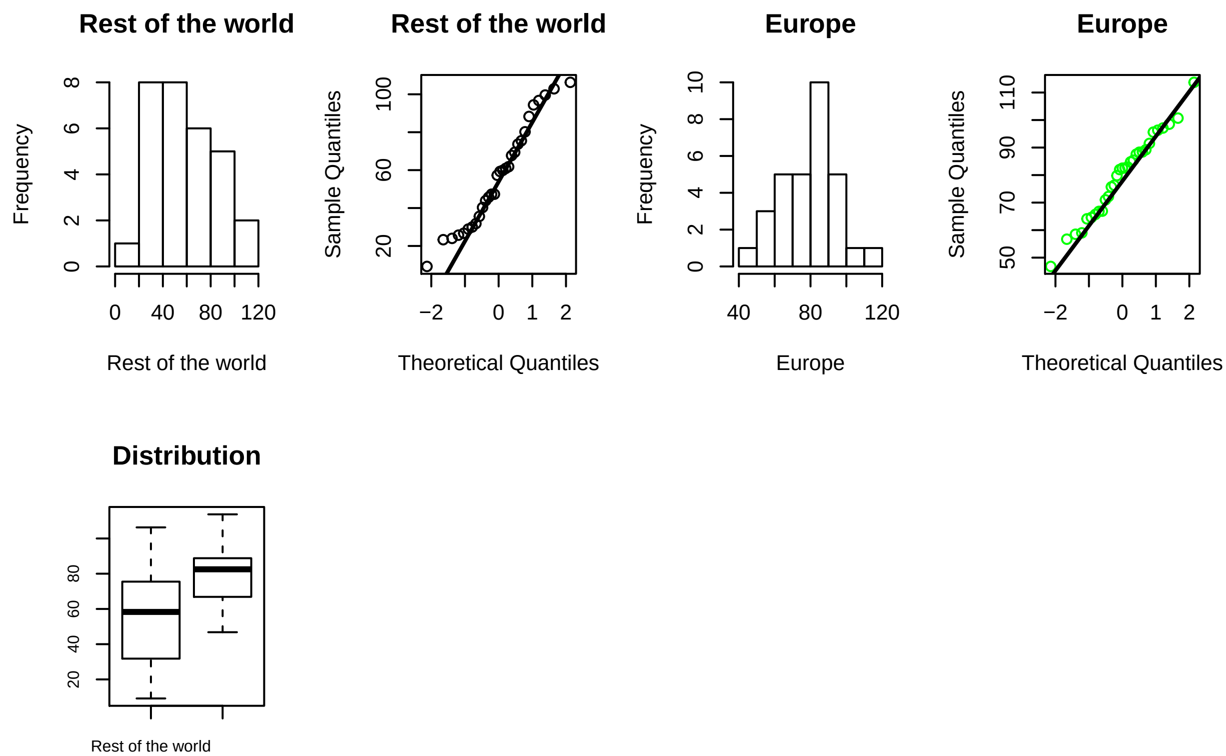


Based on the p-value we can reject the null hypothesis in favor of the alternative hypothesis meaning that there are more women in parliaments in Europe than in the rest of the world.

2.3.2 Assumption: There are more women going to college in Europe than in the rest of the world.

```
myTtest(rest_of_the_world$Education..Tertiary.gross.enrol..ratio..f.per.100.pop..,europe$Education..Ter
```

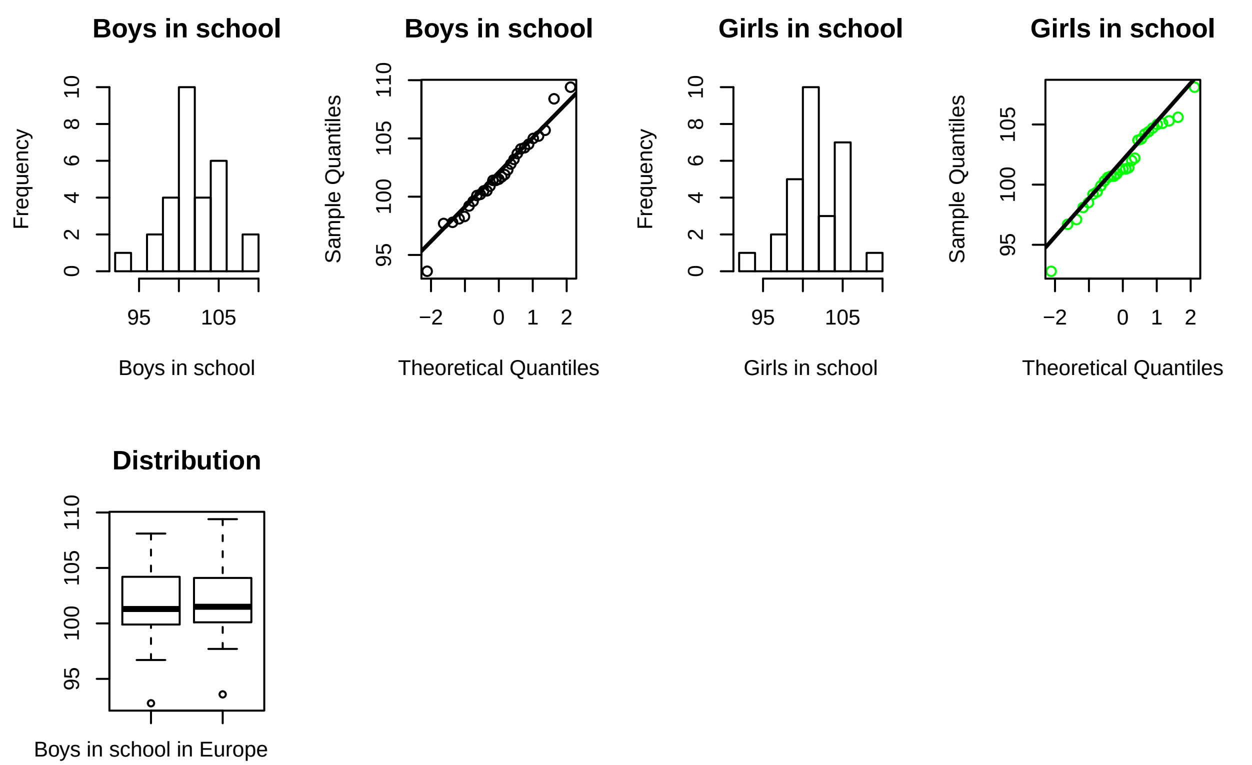
```
##
##  Welch Two Sample t-test
##
## data:  data2 and data1
## t = 3.9772, df = 45.17, p-value = 0.0001247
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.05541      Inf
## sample estimates:
## mean of x mean of y
##  79.71613  57.12000
```



Based on the p-value we can reject the null hypothesis in favor of the alternative hypothesis meaning that there are more women in parliaments in Europe than in the rest of the world.

2.3.3 Assumption: There is no difference in the number of girls compared to the number of boys going to school.

```
##
##  Paired t-test
##
## data:  data1 and data2
## t = 1.4648, df = 28, p-value = 0.07706
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.04840313         Inf
## sample estimates:
## mean of the differences
##                     0.3
```



Based on the p-value we cannot reject the null hypothesis in favor of the alternative hypothesis meaning that there is no difference of boys and girls going to school in Europe.

3 ANOVA

After comparing Europe to the rest of the world we decided to test similar assumptions between European regions.

3.1 Quick introduction

ANOVA is a method for analysing differences between group means in a sample. We presume that the total variance is caused by variability inside each group(result of coincidence) as well as variability between the groups. The latter being the result of differences between group means. Our goal is to determine whether those differences between groups are statistically significant.

For ANOVA to work the following assumptions must be met: * independence between data in samples * normal distribution of data * variance homogeneity between samples

Our goal is to use ANOVA to test whether all European regions have the same GDP per capita mean. First we correct the Region from character to factor and continue to test assumptions above. Independence is implied because these are all separate countries.

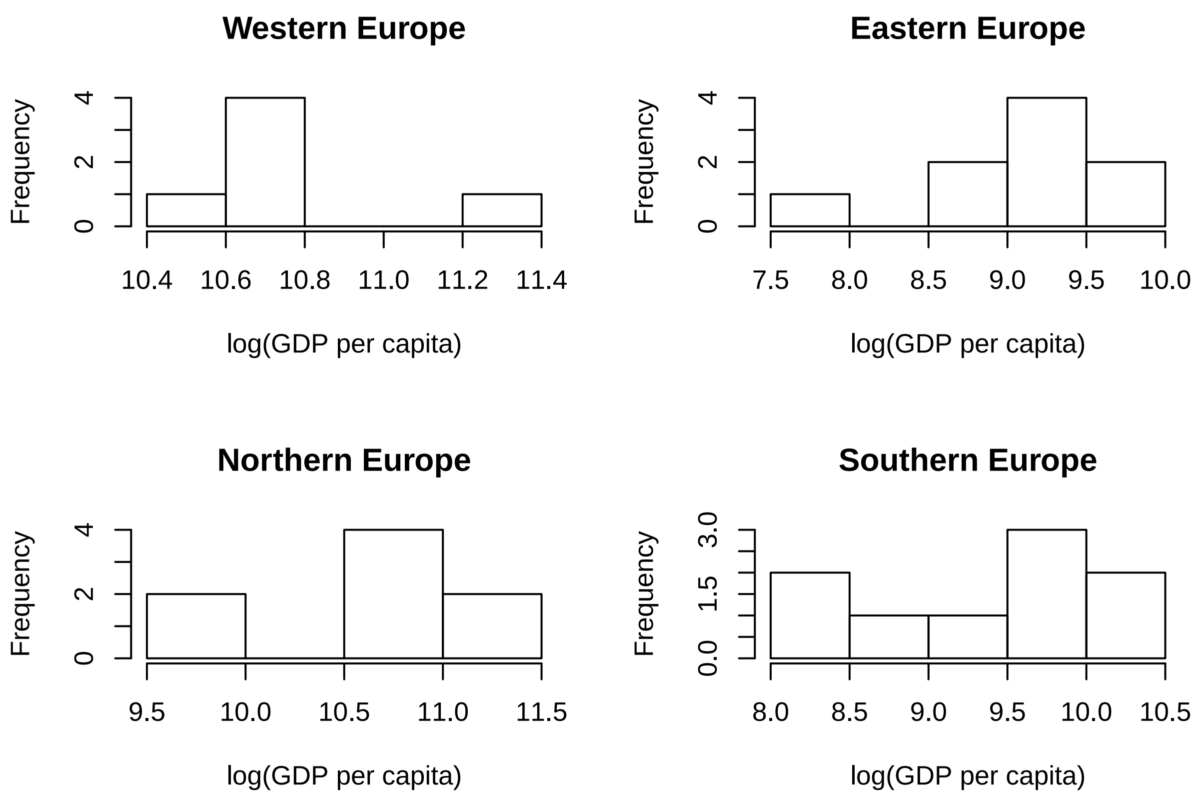
```
europe$Region <- as.factor(europe$Region)
```

3.2 Testing mean assumptions

3.2.1 Assumption: GDP per capita mean is the same across all European regions

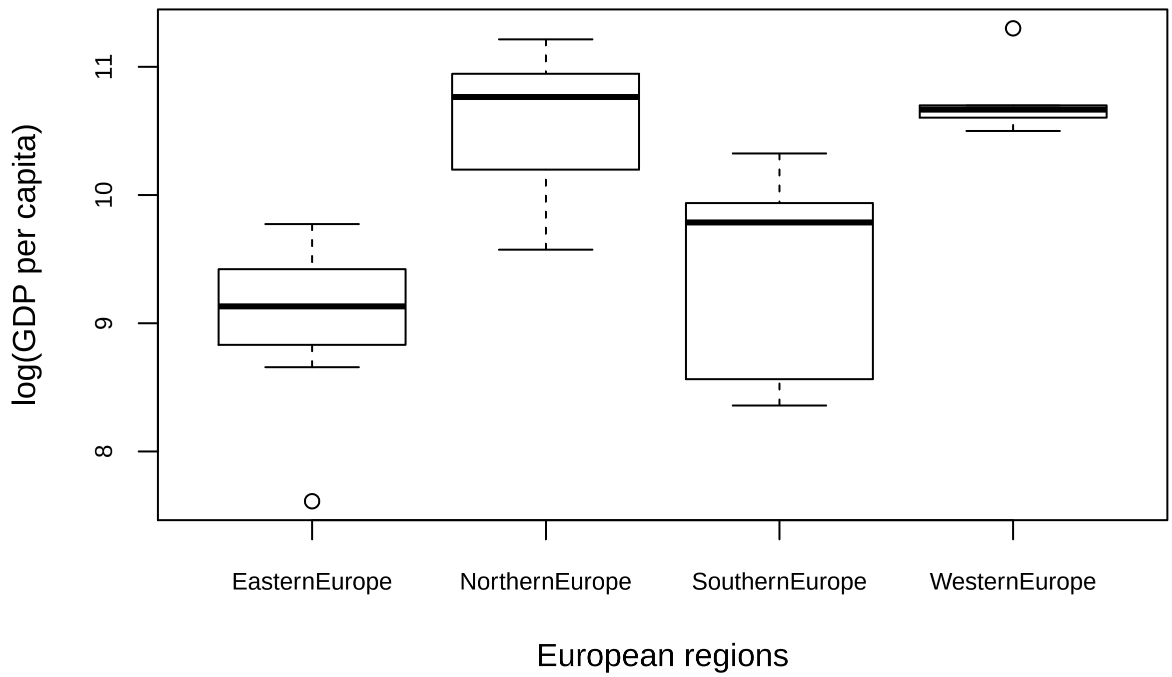
```
myAnovaLogTest(europe$GDP.per.capita..current.US..,"log(GDP per capita)")
```

```
## [1] "Testing normality:"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log(data[europe$Region == "WesternEurope"])
## D = 0.38894, p-value = 0.004962
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log(data[europe$Region == "EasternEurope"])
## D = 0.19434, p-value = 0.4216
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log(data[europe$Region == "NorthernEurope"])
## D = 0.29914, p-value = 0.03371
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  log(data[europe$Region == "SouthernEurope"])
## D = 0.23659, p-value = 0.1551
```




```
## [1] "Testing variance homogeneity:"
##
##  Bartlett test of homogeneity of variances
##
## data:  log(data) by europe$Region
## Bartlett's K-squared = 4.4893, df = 3, p-value = 0.2132
##
## [1] 0.08084323
## [1] 0.435158
## [1] 0.352265
## [1] 0.5884491
```

```
## ANOVA:
```




```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## europe$Region  3  16.02   5.339   13.52 1.23e-05 ***
## Residuals     28  11.06   0.395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All tests, except normality for Western and Northern Europe, are favourable. Our groups are of similar size and knowing that ANOVA is robust with respect to normality for similarly sized groups we proceeded.

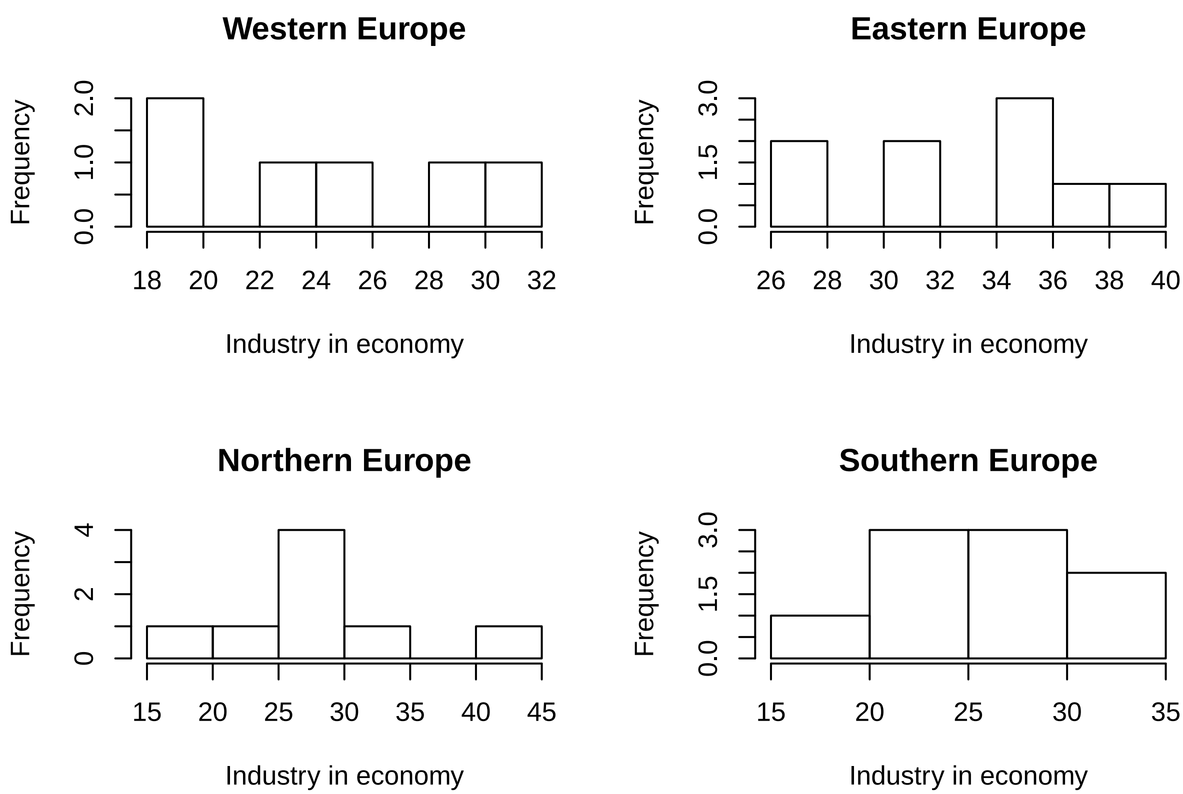
ANOVA showed that the means of GDP per capita for regions are not the same. The same can be seen from the boxplot.

3.2.2 Assumption: Industry makes up the same amount of economy across all European regions

```r
myAnovaTest(europe$Economy..Industry....of.GVA., "Industry in economy")
```

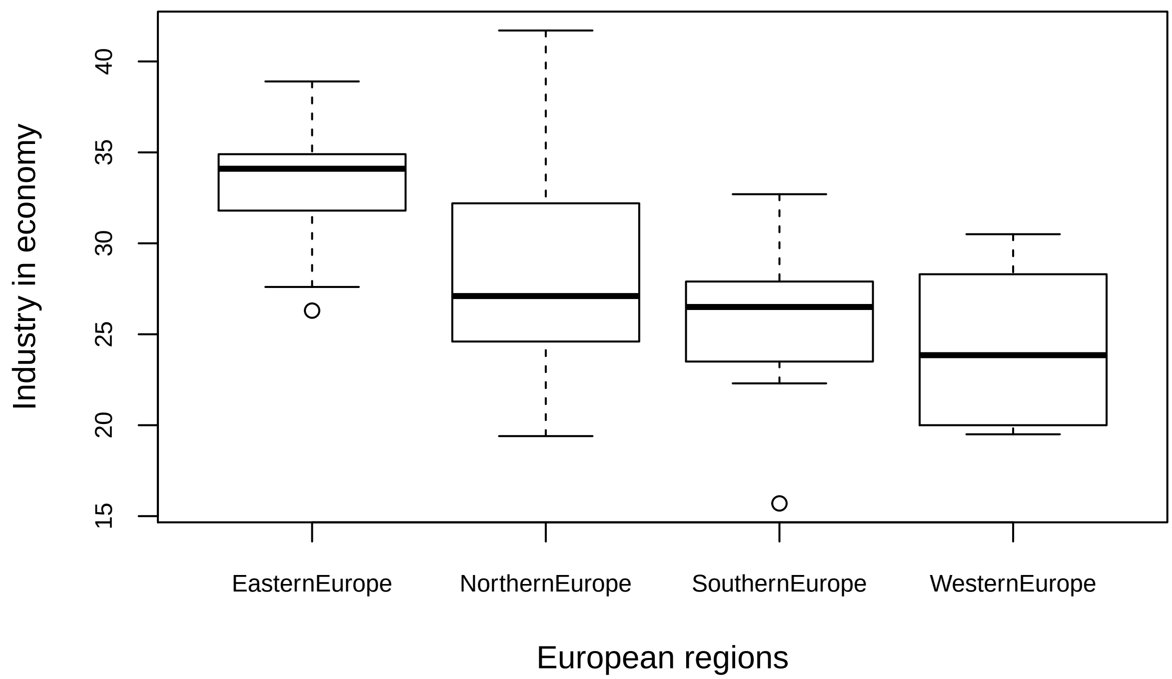
```
## [1] "Testing normality:"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "WesternEurope"])
## D = 0.18181, p-value = 0.7725
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
## data:  (data[europe$Region == "EasternEurope"])
## D = 0.1548, p-value = 0.7722
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "NorthernEurope"])
## D = 0.19445, p-value = 0.4995
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "SouthernEurope"])
## D = 0.14929, p-value = 0.8149
```




```
## [1] "Testing variance homogeneity:"
##
##  Bartlett test of homogeneity of variances
##
## data:  (data) by europe$Region
## Bartlett's K-squared = 2.0584, df = 3, p-value = 0.5604
##
## [1] 20.36267
## [1] 17.80944
## [1] 48.02125
## [1] 26.05694
```

```
## ANOVA:
```




```
##                Df Sum Sq Mean Sq F value Pr(>F)
## europe$Region  3  371.0  123.66   4.389 0.0119 *
## Residuals     28  788.9   28.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All tests are favourable for ANOVA assumptions and we can proceed

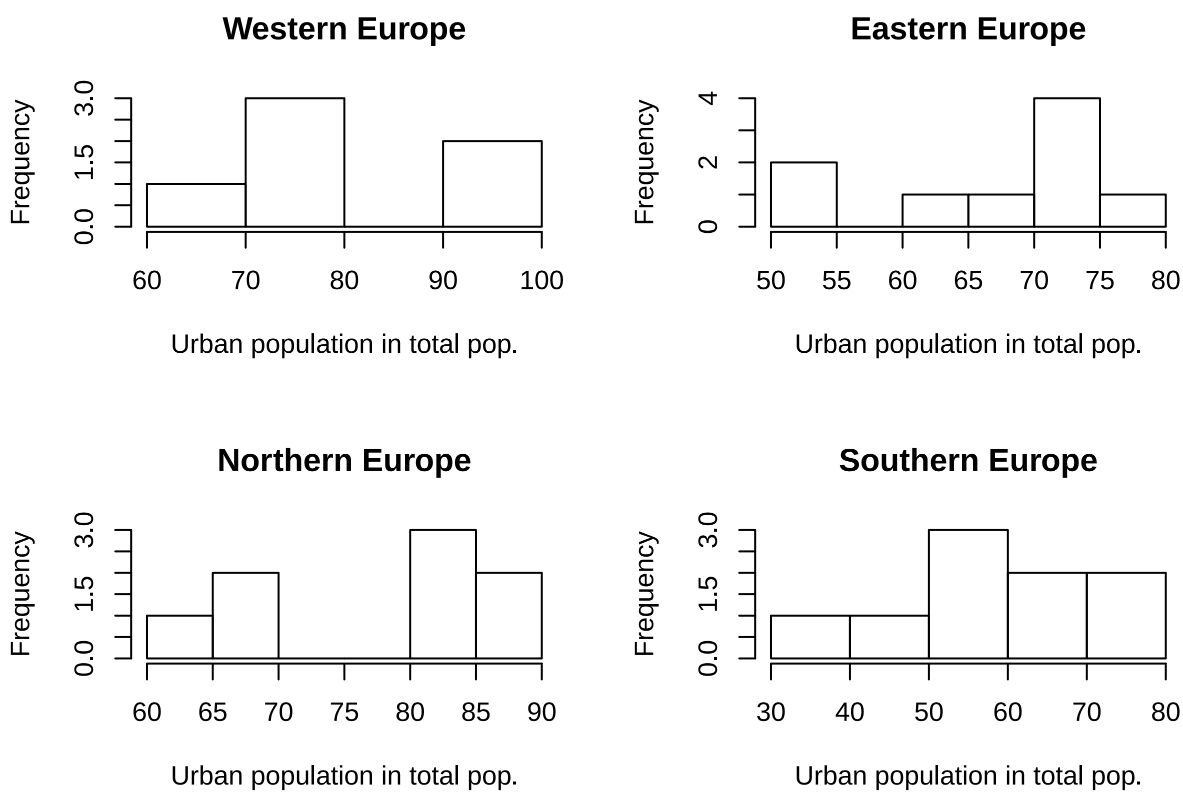
From the box plot, other than Eastern Europe, all regions appear to have similar means and with ANOVA using 1% significance we cannot reject the assumption that all groups have the same means.

3.2.3 Assumption: Mean of Urban population in total population is the same across all European regions

```
myAnovaTest(europe$Urban.population....of.total.population._x, "Urban population in total pop.")
```

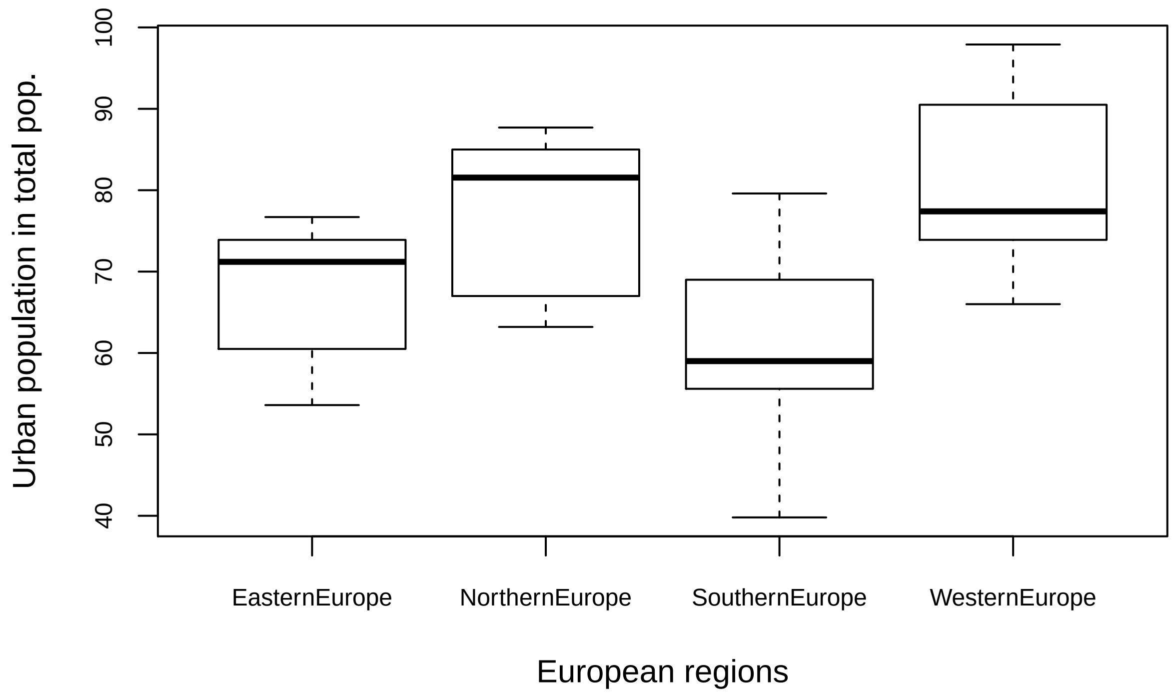
```
## [1] "Testing normality:"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "WesternEurope"])
## D = 0.20129, p-value = 0.6257
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
## data:  (data[europe$Region == "EasternEurope"])
## D = 0.26639, p-value = 0.06498
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "NorthernEurope"])
## D = 0.2544, p-value = 0.1326
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "SouthernEurope"])
## D = 0.12518, p-value = 0.9516
```




```
## [1] "Testing variance homogeneity:"
##
##  Bartlett test of homogeneity of variances
##
## data:  (data) by europe$Region
## Bartlett's K-squared = 1.2334, df = 3, p-value = 0.745
##
## [1] 136.9217
## [1] 78.155
## [1] 96.83143
## [1] 166.2653
```

```
## ANOVA:
```




```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## europe$Region  3   1818   606.0   5.114 0.00601 **
## Residuals     28   3318   118.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

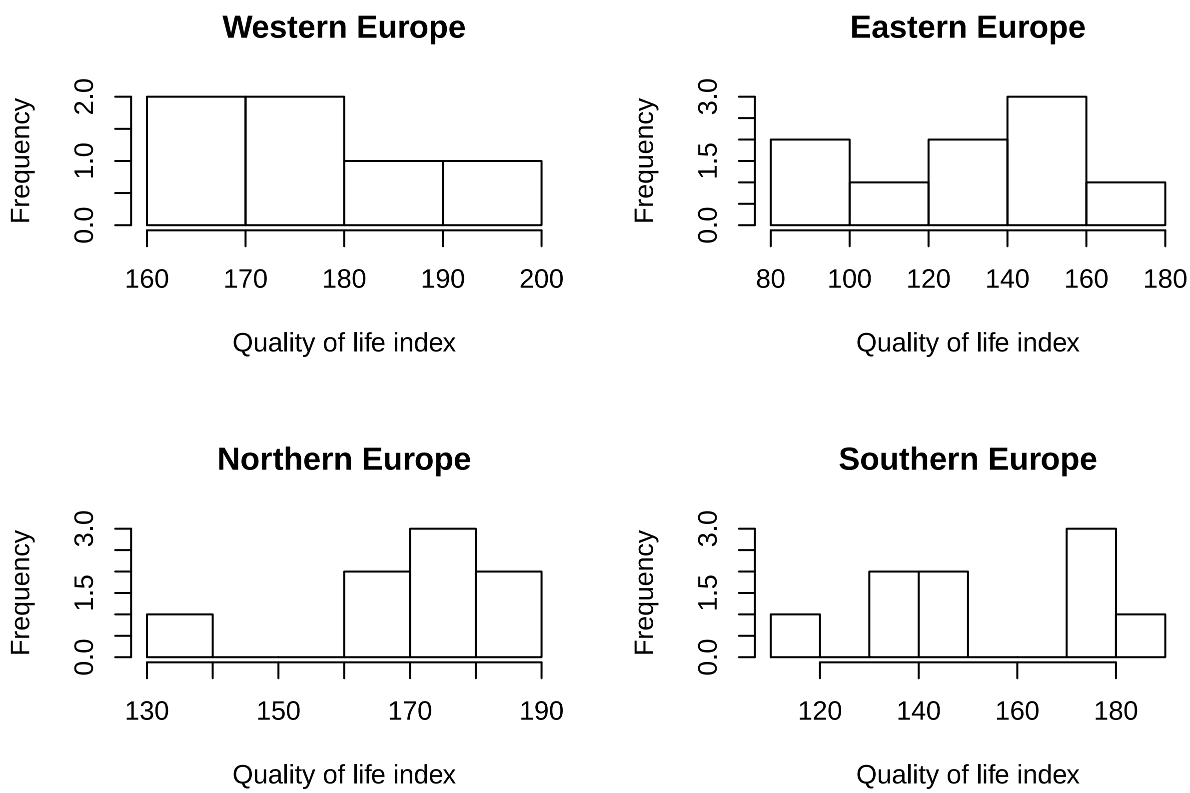
Tests are favourable, the only one raising suspicion is Lilliefors normality test Eastern European region. Our groups are of similar size and knowing that ANOVA is robust with respect to normality for similarly sized groups we proceeded.
From both the box plot and ANOVA we can see rejection of the assumption that all regions have the same part of urban population in total population.

3.2.4 Assumption: Mean of Quality of life index is the same across all European regions

```
myAnovaTest(europe$Quality.Of.Life.Index, "Quality of life index")
```

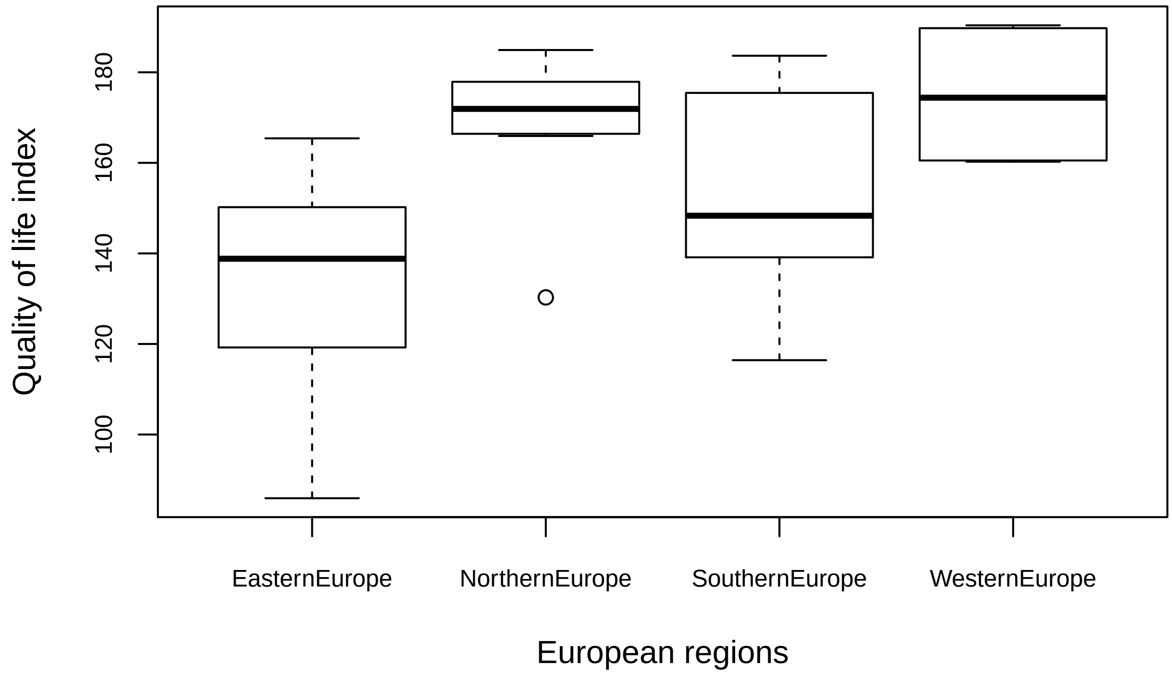
```
## [1] "Testing normality:"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "WesternEurope"])
## D = 0.20061, p-value = 0.6309
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  (data[europe$Region == "EasternEurope"])
## D = 0.26475, p-value = 0.06867
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "NorthernEurope"])
## D = 0.31544, p-value = 0.01856
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "SouthernEurope"])
## D = 0.20158, p-value = 0.3636
```




```
## [1] "Testing variance homogeneity:"
##
##  Bartlett test of homogeneity of variances
##
## data:  (data) by europe$Region
## Bartlett's K-squared = 3.6993, df = 3, p-value = 0.2958
##
## [1] 176.5803
## [1] 792.2492
## [1] 284.5061
```

```
## [1] 554.0526

## ANOVA:
```




```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## europe$Region  3   8933  2977.8   6.111 0.00248 **
## Residuals     28  13645   487.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

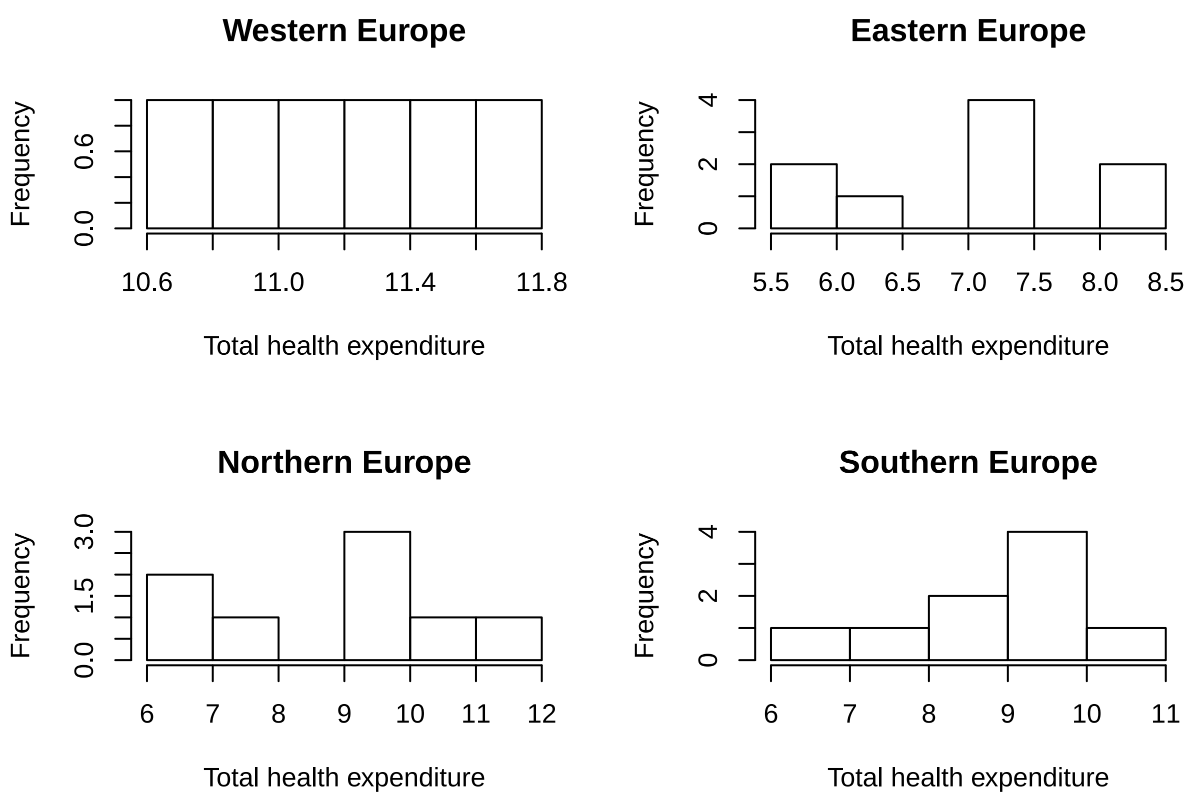
Tests are favourable, the only one raising suspicion is Lilliefors normality test Northern European region. Our groups are of similar size and knowing that ANOVA is robust with respect to normality for similarly sized groups we proceeded.
From both the box plot and ANOVA we can see rejection of the assumption that all regions have the same Quality of life index mean.

3.2.5 Assumption: Health expenditure mean is the same across all European regions

```
myAnovaTest(europe$Health..Total.expenditure....of.GDP., "Total health expenditure")
```

```
## [1] "Testing normality:"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "WesternEurope"])
## D = 0.16667, p-value = 0.8668
##
##
```

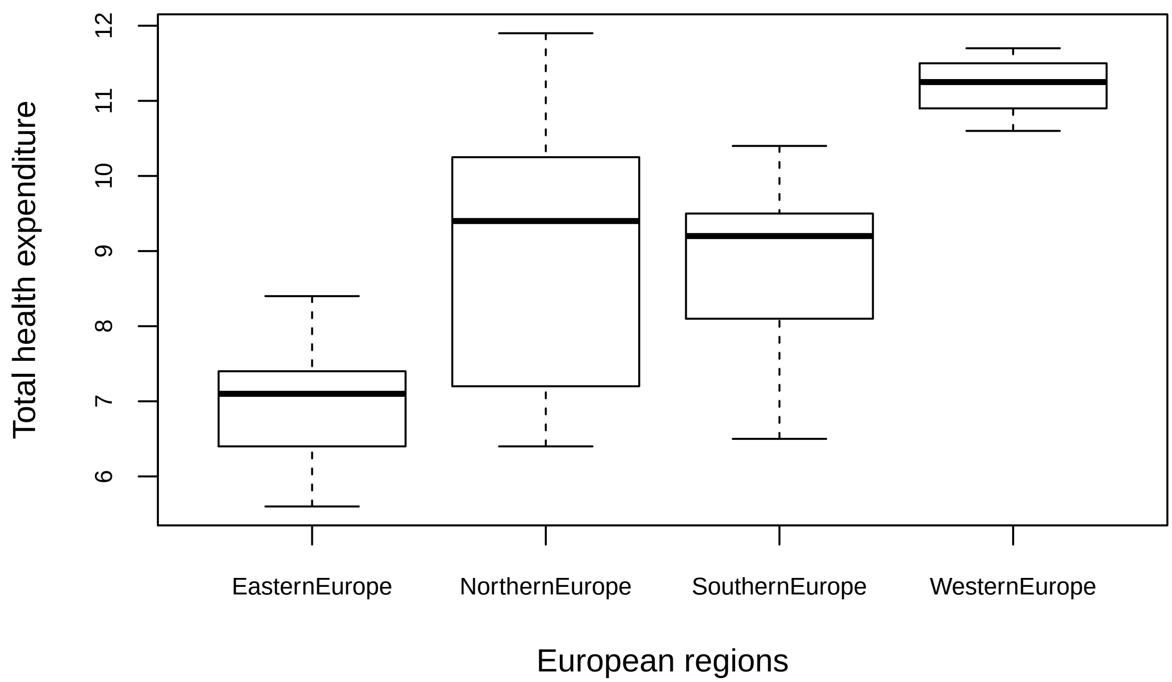
```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "EasternEurope"])
## D = 0.19865, p-value = 0.3866
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "NorthernEurope"])
## D = 0.14546, p-value = 0.8871
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data[europe$Region == "SouthernEurope"])
## D = 0.23115, p-value = 0.1792
```




```
## [1] "Testing variance homogeneity:"
##
##  Bartlett test of homogeneity of variances
##
## data:  (data) by europe$Region
## Bartlett's K-squared = 10.961, df = 3, p-value = 0.01194
##
## [1] 0.16
## [1] 0.9394444
```

```
## [1] 3.8
## [1] 1.353611
```

```
## ANOVA:
```




```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## europe$Region  3  63.33  21.108   12.92 1.76e-05 ***
## Residuals     28  45.74   1.634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Normality tests are favourable, but the variances do not seem to be homogeneous. Our groups are of similar size and knowing that ANOVA is robust with respect to variance homogeneity for similarly sized groups we proceeded.
From both the box plot and ANOVA we can see rejection of the assumption that all regions have the same Health expenditure mean.

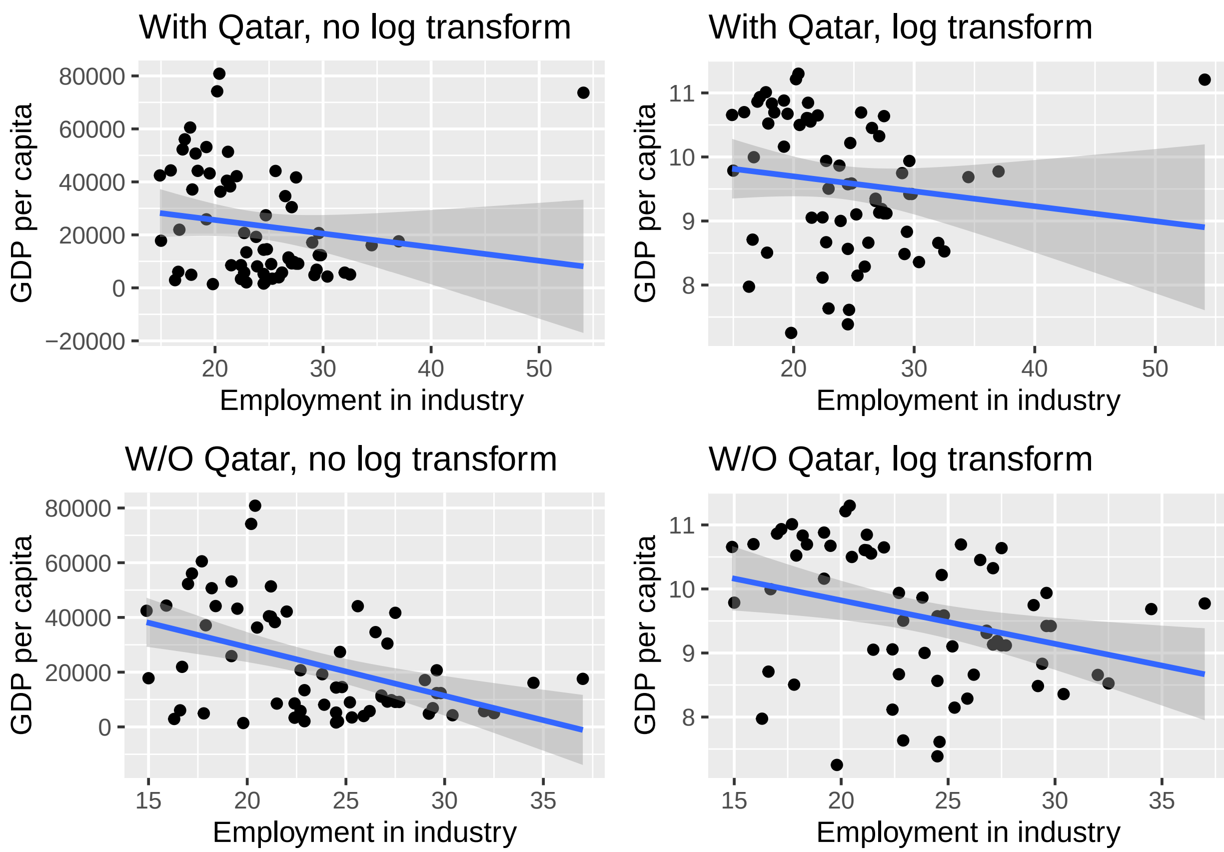
4 LINEAR REGRESSION

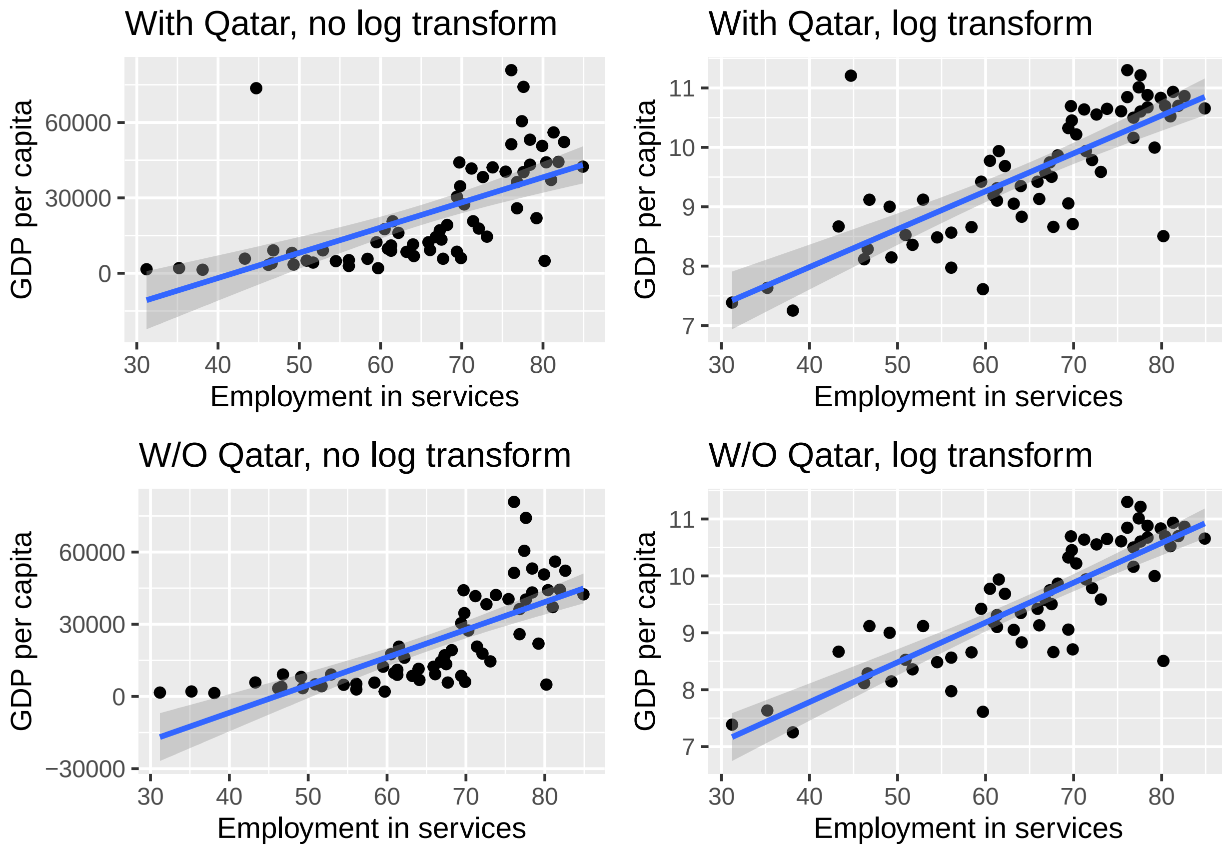
Linear regression is a method of modelling the relationship between a scalar response and one or more variables (regressors).

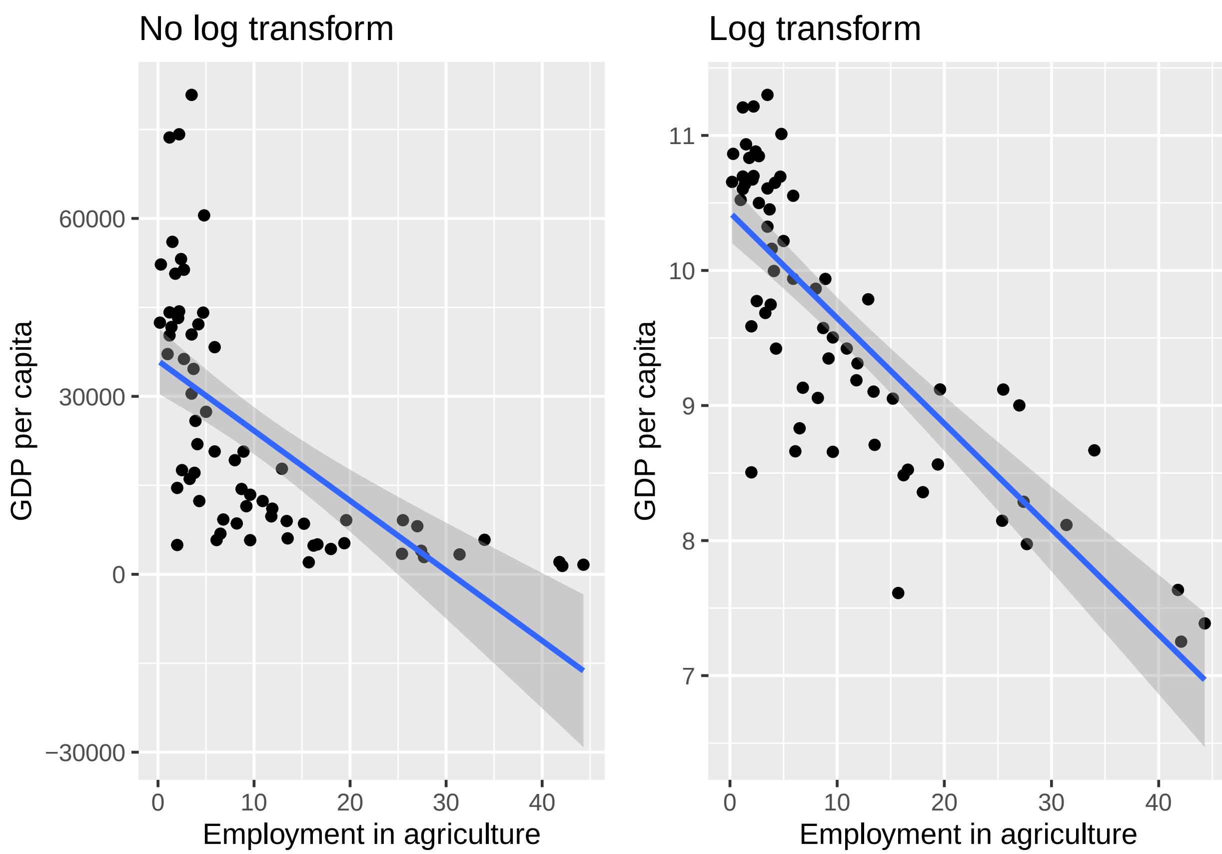
It is mostly used for predicting a value of a variable by using values of some different variable(s). Training is done on a train dataset and testing (predicting) on a never before seen data.

4.1 Predicting GDP per capita with employments per sectors

Let's first visualize the data we're working with.




With Qatar, no log transform
GDP per capita
60000
30000
0
30
40
50
60
70
80
Employment in services
With Qatar, log transform
GDP per capita
11
10
9
8
7
30
40
50
60
70
80
Employment in services
W/O Qatar, no log transform
GDP per capita
60000
30000
0
−30000
30
40
50
60
70
80
Employment in services
W/O Qatar, log transform
GDP per capita
11
10
9
8
7
30
40
50
60
70
80
Employment in services



Qatar is a huge outlier so we decided to remove it and proceed without it.

We need to check if model hypotheses are (too) violated. The most importants things here are hypotheses about regressors (in multivariate regression, regressors shouldn't be mutually correlated) and about residuals (residual normality and homogeneity of variance).

Residual normality can be tested graphically, with Q-Q plot (comparing it to normal distribution line), and statistically with Kolmogorov Smirnov test.

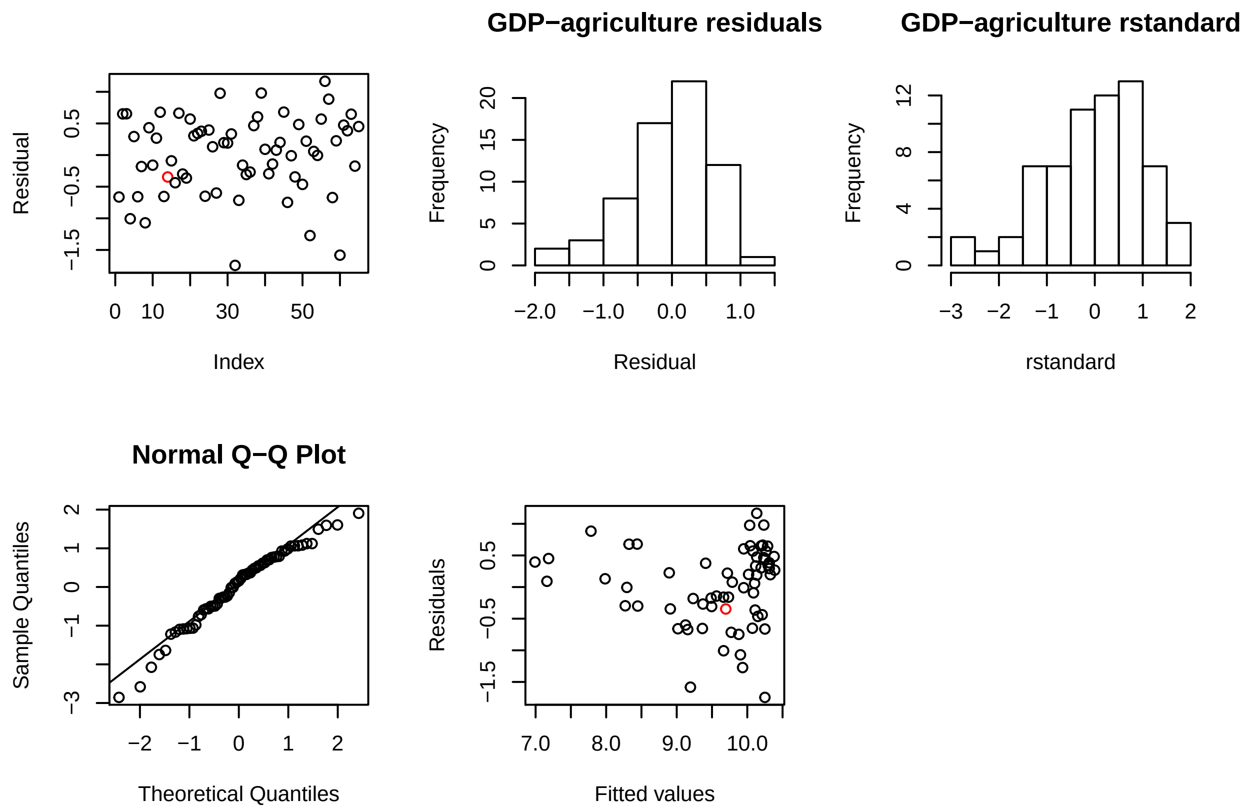
```
gdp.vs.agriculture = lm_GDP_agriculture_without_qatar

par(mfrow=c(2,3))

plot(gdp.vs.agriculture$residuals, ylab = "Residual",col =colorData$Color)

hist((gdp.vs.agriculture$residuals), main = "GDP-agriculture residuals", xlab = "Residual")
hist(rstandard(gdp.vs.agriculture), main = "GDP-agriculture rstandard", xlab = "rstandard")

qqnorm(rstandard(gdp.vs.agriculture))
qqline(rstandard(gdp.vs.agriculture))

plot(gdp.vs.agriculture$fitted.values, gdp.vs.agriculture$residuals, xlab = "Fitted values", ylab = "Re
ks.test(rstandard(gdp.vs.agriculture),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(gdp.vs.agriculture)
```

```
## D = 0.098904, p-value = 0.5165
## alternative hypothesis: two-sided
```

```
lillie.test(rstandard(gdp.vs.agriculture))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(gdp.vs.agriculture)
## D = 0.096929, p-value = 0.1357
```



We can conclude that model hypotheses about residual normality and homogeneity of variance aren't too violated of estimating **GDP per capita** with **employment in agriculture**.

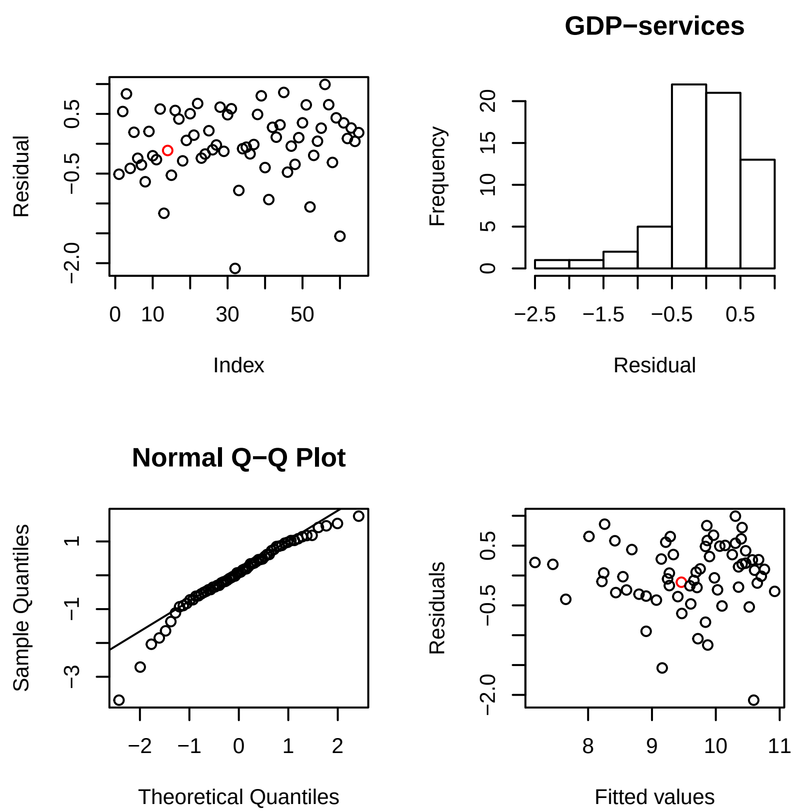
```
gdp.vs.services = lm_GDP_services_without_qatar

par(mfrow=c(2,3))

plot(gdp.vs.services$residuals, ylab = "Residual",col =colorData$Color)

hist((gdp.vs.services$residuals), main = "GDP-services", xlab = "Residual")
hist(rstandard(gdp.vs.services), main = "GDP-services", xlab = "rstandard")

qqnorm(rstandard(gdp.vs.services))
qqline(rstandard(gdp.vs.services))

plot(gdp.vs.services$fitted.values, gdp.vs.services$residuals, xlab = "Fitted values", ylab = "Residual
```

```r
ks.test(rstandard(gdp.vs.services),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(gdp.vs.services)
## D = 0.083334, p-value = 0.7258
## alternative hypothesis: two-sided
```

```r
lillie.test(rstandard(gdp.vs.services))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(gdp.vs.services)
## D = 0.084269, p-value = 0.3019
```



We can conclude that model hypotheses about residual normality and homogeneity of variance aren't too violated for this simple linear regression model of estimating **GDP per capita** with **employment in services**.

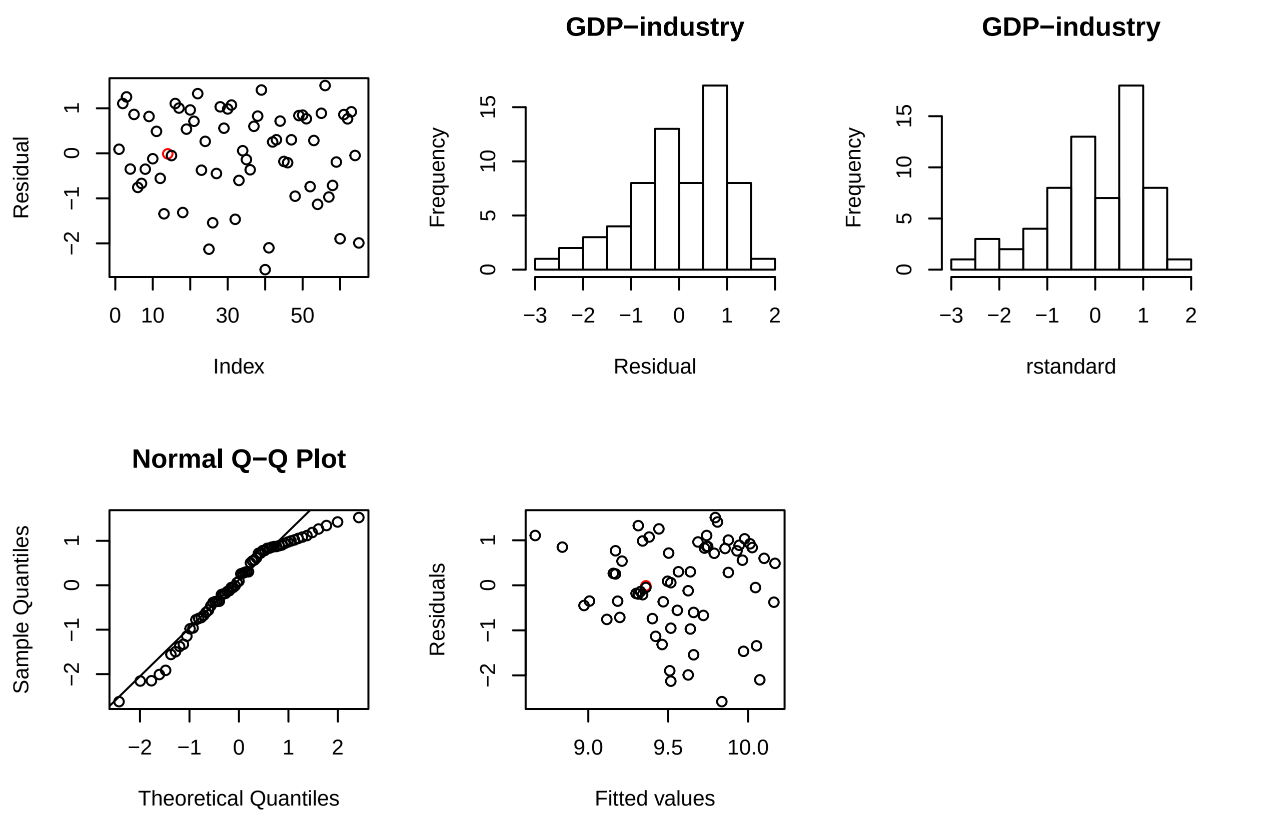
```r
gdp.vs.industry = lm_GDP_industry_without_qatar

par(mfrow=c(2,3))

plot(gdp.vs.industry$residuals, ylab = "Residual",col =colorData$Color)
```

```r
hist((gdp.vs.industry$residuals), main = "GDP-industry", xlab = "Residual")
hist(rstandard(gdp.vs.industry), main = "GDP-industry", xlab = "rstandard")

qqnorm(rstandard(gdp.vs.industry))
qqline(rstandard(gdp.vs.industry))

plot(gdp.vs.industry$fitted.values, gdp.vs.industry$residuals, xlab = "Fitted values", ylab = "Residual

ks.test(rstandard(gdp.vs.industry),'pnorm')
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(gdp.vs.industry)
## D = 0.11819, p-value = 0.2999
## alternative hypothesis: two-sided
```

```r
lillie.test(rstandard(gdp.vs.industry))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(gdp.vs.industry)
## D = 0.1162, p-value = 0.02938
```



Here, the situation is a little bit worse but we can still conclude that model hypotheses about residual normality and homogeneity of variance aren't **too** violated for this simple linear regression model of estimating

GDP per capita with **employment in industry**. However, we should be careful with it.

```r
summary(lm_GDP_agriculture_without_qatar)
```

```
##
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74450 -0.34671  0.09095  0.45050  1.16602
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          10.403707   0.107581   96.70  < 2e-16 ***
## employment_agriculture_without_qatar -0.077032   0.007031  -10.96 3.14e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6189 on 63 degrees of freedom
## Multiple R-squared:  0.6558, Adjusted R-squared:  0.6504
## F-statistic:   120 on 1 and 63 DF,  p-value: 3.144e-16
```

```r
summary(lm_GDP_services_without_qatar)
```

```
##
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_services_without_qatar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08811 -0.26613  0.04271  0.41342  0.99344
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.987580   0.383871   12.99   <2e-16 ***
## employment_services_without_qatar 0.069896   0.005743   12.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5763 on 63 degrees of freedom
## Multiple R-squared:  0.7016, Adjusted R-squared:  0.6969
## F-statistic: 148.1 on 1 and 63 DF,  p-value: < 2.2e-16
```

```r
summary(lm_GDP_industry_without_qatar)
```

```
##
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_industry_without_qatar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58392 -0.60281  0.09032  0.83791  1.50532
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                      11.18004    0.60600  18.449  < 2e-16 ***
## employment_industry_without_qatar -0.06790    0.02514  -2.701  0.00887 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9987 on 63 degrees of freedom
## Multiple R-squared:  0.1038, Adjusted R-squared:  0.08954
## F-statistic: 7.294 on 1 and 63 DF,  p-value: 0.008875
```

We can see that the model in which we use **employment in industry** to estimate **GDP per capita** performs much worse. That is also because model hypotheses in this model weren't really completely satisfied.

```
## Correlation of GDP per capita and employment in agriculture:  -0.6306469
```

```
## Correlation of GDP per capita and employment in services:  0.7299651
```

```
## Correlation of GDP per capita and employment in industry:  -0.4473786
```

We can see that correlation between GDP per capita and employment in industry is somewhat lower than when comparing to employment in agriculture or services.

```
## Correlation of employment in agriculture and industry:  0.1070753
```

```
## Correlation of employment in agriculture and services:  -0.919615
```

```
## Correlation of employment in services and industry:  -0.4890126
```

We can see that it would make no sence to include both employment in **agriculture** and **services** because they're highly correlated.

```
#without services
fit.multi.v1 = lm(log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar + employment_
summary(fit.multi.v1)
```

```
##
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar +
##     employment_industry_without_qatar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01421 -0.25026  0.06972  0.36577  1.02320
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          11.562127   0.347963  33.228  < 2e-16 ***
## employment_agriculture_without_qatar -0.074606   0.006522 -11.439  < 2e-16 ***
## employment_industry_without_qatar    -0.050200   0.014453  -3.473 0.000943 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5708 on 62 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.7026
## F-statistic:  76.6 on 2 and 62 DF,  p-value: < 2.2e-16
```

```
#with services
fit.multi.v2 = lm(log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar + employment_
summary(fit.multi.v2)
```

```
##
```

```
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar +
##     employment_services_without_qatar + employment_industry_without_qatar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99854 -0.25553  0.03134  0.37709  1.03546
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            78.7977   151.3100   0.521    0.604
## employment_agriculture_without_qatar  -0.7474     1.5141  -0.494    0.623
## employment_services_without_qatar     -0.6727     1.5140  -0.444    0.658
## employment_industry_without_qatar     -0.7216     1.5111  -0.478    0.635
##
## Residual standard error: 0.5746 on 61 degrees of freedom
## Multiple R-squared:  0.7128, Adjusted R-squared:  0.6987
## F-statistic: 50.47 on 3 and 61 DF,  p-value: < 2.2e-16
```

We can see that adding a feature which represents percent of employed in service does not contribute to a model and it reduces its Adjusted R^2. What's more, that feature has no sense since, if we know percent of people employed in agriculture and industry, than what's left until 100% is filled with percent of people employed in services. And along with all of that, it's very correlated with one of the regressors used, as stated above.

Now we'll split Region feature into separate dummy variables (we'll omit that chunk of code).

```r
# without northern America
fit.multi.v3 = lm(log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar + employment_
summary(fit.multi.v3)
```

```
##
## Call:
## lm(formula = log(gdp_per_capita_without_qatar) ~ employment_agriculture_without_qatar +
##     employment_industry_without_qatar + westernEurope + easternEurope +
##     northernEurope + southernEurope + southAmerica + centralAmerica +
##     easternAsia + westernAsia + southeasternAsia + southernAsia +
##     southernAfrica + northernAfrica + oceania)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50277 -0.17142  0.03381  0.25894  1.03910
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           11.224680   0.452021  24.832  < 2e-16 ***
## employment_agriculture_without_qatar -0.064485   0.008504  -7.583 8.33e-10 ***
## employment_industry_without_qatar    -0.016605   0.016025  -1.036  0.30520
## westernEurope                         0.046114   0.399887   0.115  0.90867
## easternEurope                        -1.041136   0.429076  -2.426  0.01897 *
## northernEurope                       -0.063259   0.386244  -0.164  0.87058
## southernEurope                       -0.666661   0.402068  -1.658  0.10369
## southAmerica                         -0.945016   0.417728  -2.262  0.02815 *
## centralAmerica                       -0.839267   0.612022  -1.371  0.17653
## easternAsia                          -0.190555   0.429554  -0.444  0.65928
## westernAsia                          -0.792223   0.392754  -2.017  0.04918 *
```

```
## southeasternAsia                      -0.548464   0.443378  -1.237  0.22198
## southernAsia                          -0.833506   0.509148  -1.637  0.10802
## southernAfrica                        -1.735311   0.608145  -2.853  0.00632 **
## northernAfrica                        -1.019845   0.636929  -1.601  0.11576
## oceania                                0.106062   0.487614   0.218  0.82871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4848 on 49 degrees of freedom
## Multiple R-squared:  0.8357, Adjusted R-squared:  0.7854
## F-statistic: 16.62 on 15 and 49 DF,  p-value: 2.819e-14
```

We removed northern America dummy from the model because $N-1$ categorical features are enough to figure out the N-th one. Adding dummy variables adds great boost to our model, insreasing its R^2 and adjusted R^2 significantly. We are aware that now we have some regressors which are not significant and thus not needed but we will not proceed with removing them in this case for the sake of convenience

4.2 Predicting life expectancy

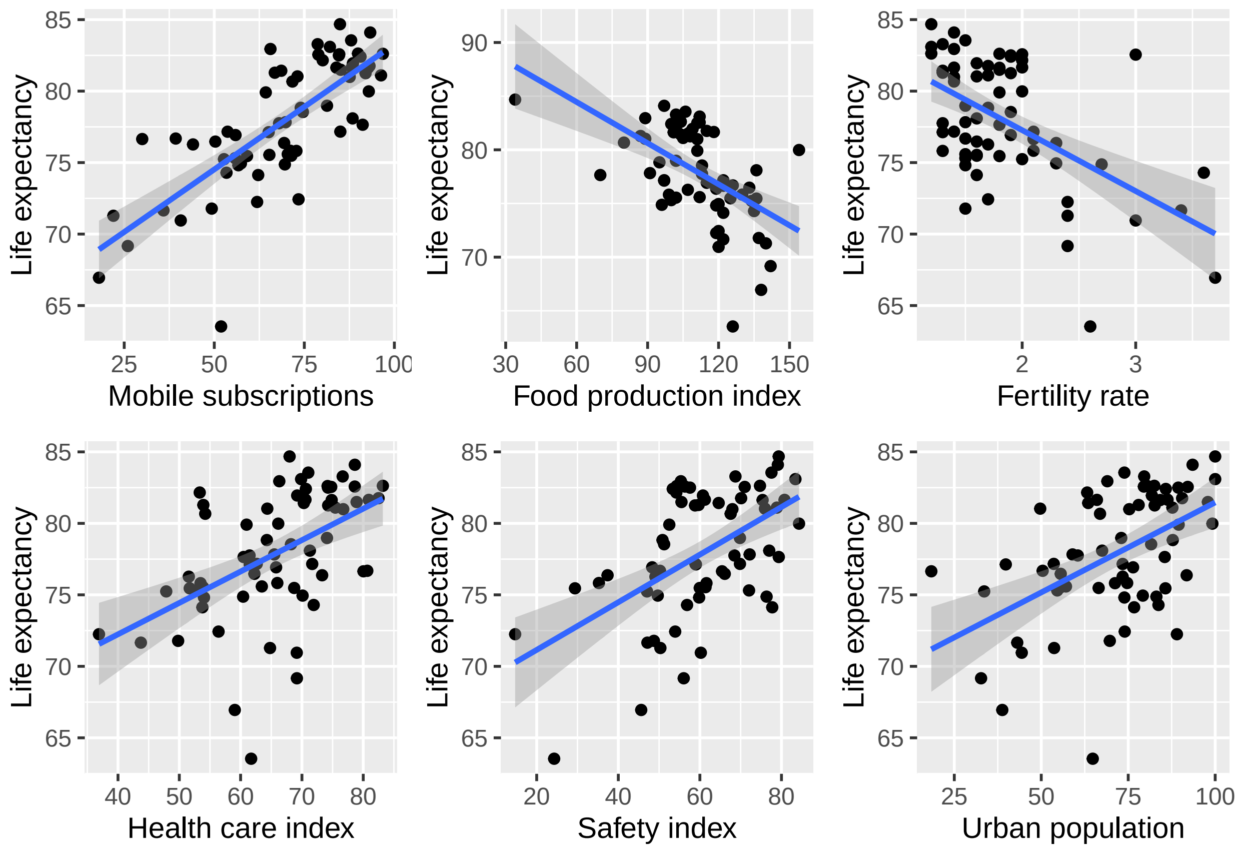
By finding correlation between life expectancy at birth and all the other features in our dataset, we come to mostly intuitive results. It's not a surprise that a higher living standard of a country implicates that a life expectancy at birth will be longer. That's why, from the feature that are highly correlated with life expectancy at birth, we'll try to pick the ones that are more interesting.

For example, **number of mobile cellular subscriptions per 100 inhabitants** is an interesting feature and has positive impact on life expectancy. **Food production index** is negatively correlated with life expectancy because those countries have more developed agriculture, produce more food and perform more physical work, thus leading to a shorter life.

Next up, **fertility rate (total live births per woman)** has a significantly negative impact on life expectancy. In general, countries with lower levels of education and lower quality of life index have that rate higher.

Not very surprisingly, **health care index** has a very positive impact on life expectancy. We'll include it in order to get better results from linear regression. What's more, the countries with higher **safety index** tend to have a longer life expectancy. An interetsing result which mostly contributes to a smaller number of violent and non-natural deaths.

And the last feature which we decided to include is the **percentage of urban population**. This is maybe too correlated with **number of mobile cellular subscriptions per 100 habitants**, having Pearson's correlation coefficient of 0.6699935 but we still decided to include it in the first model. Maybe it will be removed later. Some other features such as education seemed really interesting but contained NA values in some examples so we decided to skip those.




```
lm.life.expectancy = lm(formula = Life.expectancy.at.birth..total..years. ~ Mobile.cellular.subscriptio
summary(lm.life.expectancy)
```

```
##
## Call:
## lm(formula = Life.expectancy.at.birth..total..years. ~ Mobile.cellular.subscriptions..per.100.inhabi
##     Food.production.index..2004.2006.100. + Fertility.rate..total..live.births.per.woman. +
##     Health.Care.Index + Safety.Index + Urban.population....of.total.population._x,
##     data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3243 -1.2886 -0.0033  1.6032  5.0995
##
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                           69.30274    3.51492
## Mobile.cellular.subscriptions..per.100.inhabitants..1  0.07702    0.02770
## Food.production.index..2004.2006.100.                 -0.05141    0.01925
## Fertility.rate..total..live.births.per.woman.         -1.51020    0.65967
## Health.Care.Index                                      0.11649    0.03422
## Safety.Index                                           0.03039    0.02684
## Urban.population....of.total.population._x             0.03145    0.02383
##                                                       t value Pr(>|t|)
## (Intercept)                                            19.717  < 2e-16 ***
## Mobile.cellular.subscriptions..per.100.inhabitants..1   2.781  0.00726 **
## Food.production.index..2004.2006.100.                  -2.671  0.00977 **
## Fertility.rate..total..live.births.per.woman.          -2.289  0.02566 *
## Health.Care.Index                                       3.404  0.00120 **
## Safety.Index                                            1.132  0.26218
## Urban.population....of.total.population._x              1.320  0.19200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.396 on 59 degrees of freedom
## Multiple R-squared:  0.7313, Adjusted R-squared:  0.704
## F-statistic: 26.76 on 6 and 59 DF,  p-value: 3.828e-15
```

After removing **safety index** from regressors, these are the new R^2 and adjusted R^2 that we get.

```
## R^2 =  0.7254778
```

```
## Adjusted R^2 =  0.702601
```

We can conclude that this variable is not too significant and could be removed without many negative circumstances.

And after removing **percent of urban population in total population** from regressors, these are the new R^2 and adjusted R^2 that we get.

```
## R^2 =  0.720206
```

```
## Adjusted R^2 =  0.7018589
```

Again, percent of **urban population in total population** doesn't seem too significant and could be removed in order to make the model more simple.

Let's see what happens if we now remove **fertility rate** from the regressors.

```
## R^2 =  0.6959642
```

```
## Adjusted R^2 =  0.6812527
```

Now, we could say that **fertility rate** does make some significant impact on this model and we might want to **keep** it.

5 LOGISTIC REGRESSION

Logistic regression is a method of machine learning in which we explain the relationship of one binary (doesn't have to be binary, but most of the times it is) dependent variable and one or more ordinal, nominal interval or ratio-level independent variables.

5.1 Predicting if a country is a European one

We would like to be able to predict if a country is in Europe based on some of the variables. First of all, we'll make some assumptions. We think that Europe countries should have a lower **percent of population within the age span 0-14 years**. We also think that they might have a lower percent of **participation of female labour force**, **fertility rate**, **urban population growth rate** and **traffic commute time**.

We think that they should have a higher **health expanditure** and higher **number of woman in parliament**.

```
# eliminate Hong Kong, SAR because it has NA in some row which we need
dataset.logistic.regression = dataset[-c(11), ]
```

We also added a dummy variable representing whether the country is a European one.

```
logreg.model = glm(is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-
summary(logreg.model)
```

```
##
## Call:
## glm(formula = is.europe ~ Labour.force.participation..female.pop +
##     `Population.age.distribution.0-14.years....` + Fertility.rate..total..live.births.per.woman. +
##     Pollution.index + Traffic.commute.time.index + Urban.population.growth.rate..average.annual... +
##     Quality.Of.Life.Index + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments.., data = dataset.logistic.regression)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.54069  -0.14235   0.02848   0.18608   0.54838
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                      4.214846   0.752956   5.598
## Labour.force.participation..female.pop          -0.031800   0.007445  -4.271
## `Population.age.distribution.0-14.years....`    -0.068157   0.019974  -3.412
## Fertility.rate..total..live.births.per.woman.    0.424016   0.185773   2.282
## Pollution.index                                 -0.002745   0.003030  -0.906
## Traffic.commute.time.index                      -0.007883   0.007512  -1.049
## Urban.population.growth.rate..average.annual... -0.077596   0.047353  -1.639
## Quality.Of.Life.Index                           -0.002605   0.001564  -1.666
## Health..Total.expenditure....of.GDP.            -0.033335   0.019716  -1.691
## Seats.held.by.women.in.national.parliaments..    0.007401   0.004257   1.739
##                                                 Pr(>|t|)
## (Intercept)                                     7.12e-07 ***
## Labour.force.participation..female.pop          7.76e-05 ***
## `Population.age.distribution.0-14.years....`     0.00121 **
## Fertility.rate..total..live.births.per.woman.    0.02635 *
## Pollution.index                                  0.36892
## Traffic.commute.time.index                       0.29858
## Urban.population.growth.rate..average.annual...  0.10699
```

```
## Quality.Of.Life.Index                           0.10139
## Health..Total.expenditure....of.GDP.            0.09655 .
## Seats.held.by.women.in.national.parliaments..    0.08771 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07779324)
##
##     Null deviance: 16.2462  on 64  degrees of freedom
## Residual deviance:  4.2786  on 55  degrees of freedom
## AIC: 29.613
##
## Number of Fisher Scoring iterations: 2
```

We can see quite a lot of room for improvement because some regressors are insignificant.

```r
cat("R^2 = ", 1 - logreg.model$deviance/logreg.model$null.deviance, "\n")
```

```
## R^2 =  0.7366375
```

Confusion matrix:

```
##        yHat
##         FALSE TRUE
##   FALSE    30    3
##   TRUE      2   30
```

```
## accuracy:  0.9230769
```

```
## precision:  0.9090909
```

```
## recall:  0.9375
```

```
## specificity:  0.9375
```

```
## f1_score:  0.9230769
```

Previously, we also checked for correlation between regressors but we'll omit that chunk of code.

The countries for which our model gives *false positives* are **Canada**, **Japan** and **Republic of Korea**. The countries for which our model gives *false negatives* are **Ireland** and **Switzerland**.

For Croatia, model confidently claims that it's a European country.

Let's try removing pollution index index from the model and see how it responds.

```r
anova(logreg.model, logreg.model.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Pollution.index +
##     Traffic.commute.time.index + Urban.population.growth.rate..average.annual... +
##     Quality.Of.Life.Index + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments..
## Model 2: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Traffic.commute.time.index +
##     Urban.population.growth.rate..average.annual... + Quality.Of.Life.Index +
##     Health..Total.expenditure....of.GDP. + Seats.held.by.women.in.national.parliaments..
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1        55     4.2786
```

```
## 2        56     4.3425 -1 -0.063846    0.365
```

P-value of Chi-Squared test shows us that there are no significant differences between this and a previous model. Let's go one step further and try removing **traffic commute time index** and see how the model responds.

```
anova(logreg.model, logreg.model.3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Pollution.index +
##     Traffic.commute.time.index + Urban.population.growth.rate..average.annual... +
##     Quality.Of.Life.Index + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments..
## Model 2: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Urban.population.growth.rate..average.annual... +
##     Quality.Of.Life.Index + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments..
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        55     4.2786
## 2        57     4.4269 -2 -0.14829   0.3855
```

Once again, we can see that there are no significant differences between this model and the first one. We'll also try removing **quality of life index**.

```
logreg.model.4 = glm(is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.

anova(logreg.model, logreg.model.4, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Pollution.index +
##     Traffic.commute.time.index + Urban.population.growth.rate..average.annual... +
##     Quality.Of.Life.Index + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments..
## Model 2: is.europe ~ Labour.force.participation..female.pop + `Population.age.distribution.0-14.years
##     Fertility.rate..total..live.births.per.woman. + Urban.population.growth.rate..average.annual...
##     Health..Total.expenditure....of.GDP. + Seats.held.by.women.in.national.parliaments..
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        55     4.2786
## 2        58     4.5314 -3 -0.25278   0.3547
```

Once again, the removal of the **quality of life index** variable shows no significant degradation in our model performance.

```
summary(logreg.model.4)
```

```
##
## Call:
## glm(formula = is.europe ~ Labour.force.participation..female.pop +
##     `Population.age.distribution.0-14.years....` + Fertility.rate..total..live.births.per.woman. +
##     Urban.population.growth.rate..average.annual... + Health..Total.expenditure....of.GDP. +
##     Seats.held.by.women.in.national.parliaments.., data = dataset.logistic.regression)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -0.60749  -0.15227   0.04846   0.18432   0.56586
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                     3.355807   0.563136   5.959
## Labour.force.participation..female.pop         -0.029034   0.007065  -4.109
## `Population.age.distribution.0-14.years....`   -0.067221   0.016251  -4.136
## Fertility.rate..total..live.births.per.woman.   0.378753   0.170688   2.219
## Urban.population.growth.rate..average.annual... -0.094734   0.045634  -2.076
## Health..Total.expenditure....of.GDP.           -0.040034   0.017651  -2.268
## Seats.held.by.women.in.national.parliaments..   0.008324   0.003855   2.159
##                                                Pr(>|t|)
## (Intercept)                                    1.59e-07 ***
## Labour.force.participation..female.pop         0.000126 ***
## `Population.age.distribution.0-14.years....`   0.000115 ***
## Fertility.rate..total..live.births.per.woman.  0.030415 *
## Urban.population.growth.rate..average.annual... 0.042342 *
## Health..Total.expenditure....of.GDP.           0.027063 *
## Seats.held.by.women.in.national.parliaments..  0.034969 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07812778)
##
##     Null deviance: 16.2462  on 64  degrees of freedom
## Residual deviance:  4.5314  on 58  degrees of freedom
## AIC: 27.344
##
## Number of Fisher Scoring iterations: 2
```

Now, all the variables are significant and we will not proceed with new regressor removals.

Confusion matrix:

```
##         yHat
##          FALSE TRUE
##   FALSE    30    3
##   TRUE      2   30
```

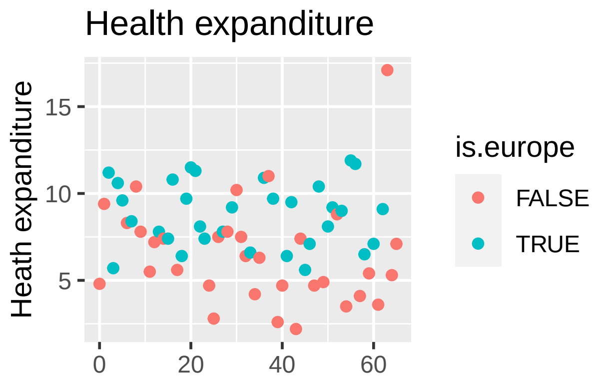
We removed **three** regressors and still got the same result!

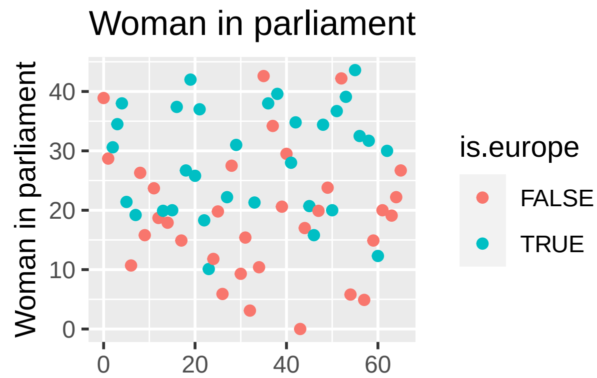
Here we need to say that we trained and tested our model on the same data which is never done.

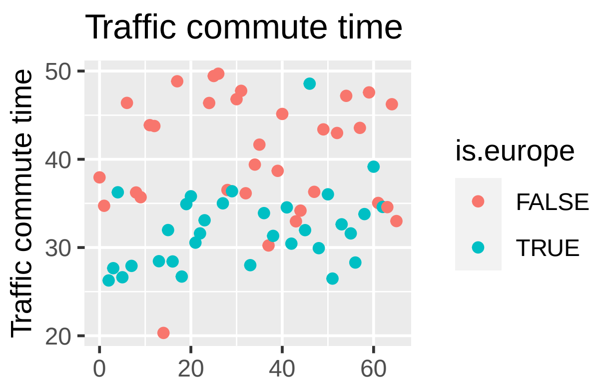
The reason for this was to show basic principles of logistic regression and we also don't have enough examples to split the dataset on train and test.

Now let's graphically check if our assumptions were well made.


Female labour participation
Female labour participation
90
80
70
60
0
20
40
60
is.europe
FALSE
TRUE
Population aged 0−14 years
Population aged 0−14 years
35
30
25
20
15
0
20
40
60
is.europe
FALSE
TRUE
Fertility rate
Fertility rate
3
2
0
20
40
60
is.europe
FALSE
TRUE
Urban population growth rate
Urban population growth rate
6
4
2
0
0
20
40
60
is.europe
FALSE
TRUE







They were indeed!

Even though the scatter plot of, for example, **traffic commute time index** indicates that it should have an impact on predicting whether or not a country is European, that's not really the case in our logistic regression model.

Let's see its correlation with all the other regressors in our model:

```
## Correlation with female labour force participation:  0.4314417
```

```
## Correlation with pop. age distribution 0-14 years:  0.6203202
```

```
## Correlation with fertility rate:  0.4841935
```

```
## Correlation with urban population growth rate:  0.4250364
```

```
## Correlation with health expanditure:  -0.4075308
```

```
## Correlation with number of woman in parliament:  -0.3962299
```

Here we can see that it's also not really directly correlated to any of our used regressors. What's probably the case is that it's described by a combination (or combinations) of them and thus ends up being insignificant.

6 SVM

Another approach that we didn't learn on this subject, but would like to try is **Support Vector Machines**.

It is a supervised machine learning method which is based on taking data points from regressors, putting them into a high-dimensional space (if the number of regressors is high-dimensional, and most of the times it is) and creating a **hyperplane** which is supposed to divide one class from another.

That's why this method is mostly used for two-group classification. The elements from one group should ideally be as far as possible from those from another group.

The hyperplane is chosen by maximizing the margins from both tags. If the number of regressors is n, then out hyperplane is n-dimensional.

6.1 Predicting if a country is a European one

```
svm.classifier = svm(formula = is.europe ~ Labour.force.participation..female.pop + `Population.age.dis
```

Confusion matrix:

```
yPred <- svm.classifier$fitted
tab <- table(dataset.logistic.regression$is.europe, yPred)
tab
```

```
##        yPred
##         FALSE TRUE
##   FALSE    31    2
##   TRUE      1   31
```

We can see that an SVM model gives an even better prediction than the logistic regression one, using the same columns as regressors.

Here we need to say that we trained and tested our model on the same data which is never done.

We don't have enough data to split our dataset and it's only to show how SVM can be used.

7 CONCLUSION

To start with, we would like to reflect on the given dataset. Although there is a great number of features, we have found that it was quite difficult to draw strong conclusions with that small amount of rows. In most places where it was a condition for data to be normally distributed we had to “stretch” the definition because of the small sample and it was also difficult to group data because most groups were simply too small to do anything with.

Next, let’s discuss the results. There are many outliers and certain countries are outliers in many features. One of the countries that stands out is Qatar with a very big outlier number compared to its small size. Another notable outlier is America where it was surprising to find out how extremely low its international trade balance is and how high its international air travel is compared to its pollution index not being that high. Croatia is not an outlier in any of the features. When comparing Europe to other world countries it was good for us, as Europeans, to notice that Europe is well positioned when fighting current world issues. All ANOVA assumptions were rejected which was surprising considering we expected macroeconomic features to be similarly distributed across European regions. When trying to predict which countries are European we found out that Ireland and Switzerland didn’t fit the mold however Canada, Japan and Republic of Korea were predicted to be European. Croatia was predicted to be an European country.

For further research it would be great to have more data and to compare different continents and regions. It would also be interesting to compare countries through different years.

The dataset was fun but small.