Alessia Di Giovanni

Fertility rate among countries

A SUPERVISED AND UNSUPERVISED ANALYSIS

Immagine di Francesco Ciccolella

Dataset

The average number of babies born to women during their reproductive years

Fertility rate

Number of days of paid maternity leave

Maternity days

% of women who are currently using, or whose sexual partner is currently using, at least one method of contraception, regardless of the method used

Use of contraceptive

Islam, Christianity, Buddhism, Hinduism

Religion

Number of years of education

Years of school

Average time a human being is expected to live

Life expectancy

Dataset

(recorded+unrecorded) alcohol per capita (15+) consumption

Consumption of alcohol

Share of the labor force that is without work but available for and seeking employment

Unemployment female rate

Index from 1 to 100 by Freedom in the World

Freedom

The difference between the number of persons entering and leaving a country during the year per 1,000 persons

Migration rate

Measure of the total income generated by a country's residents

Gross National Income

log GNI

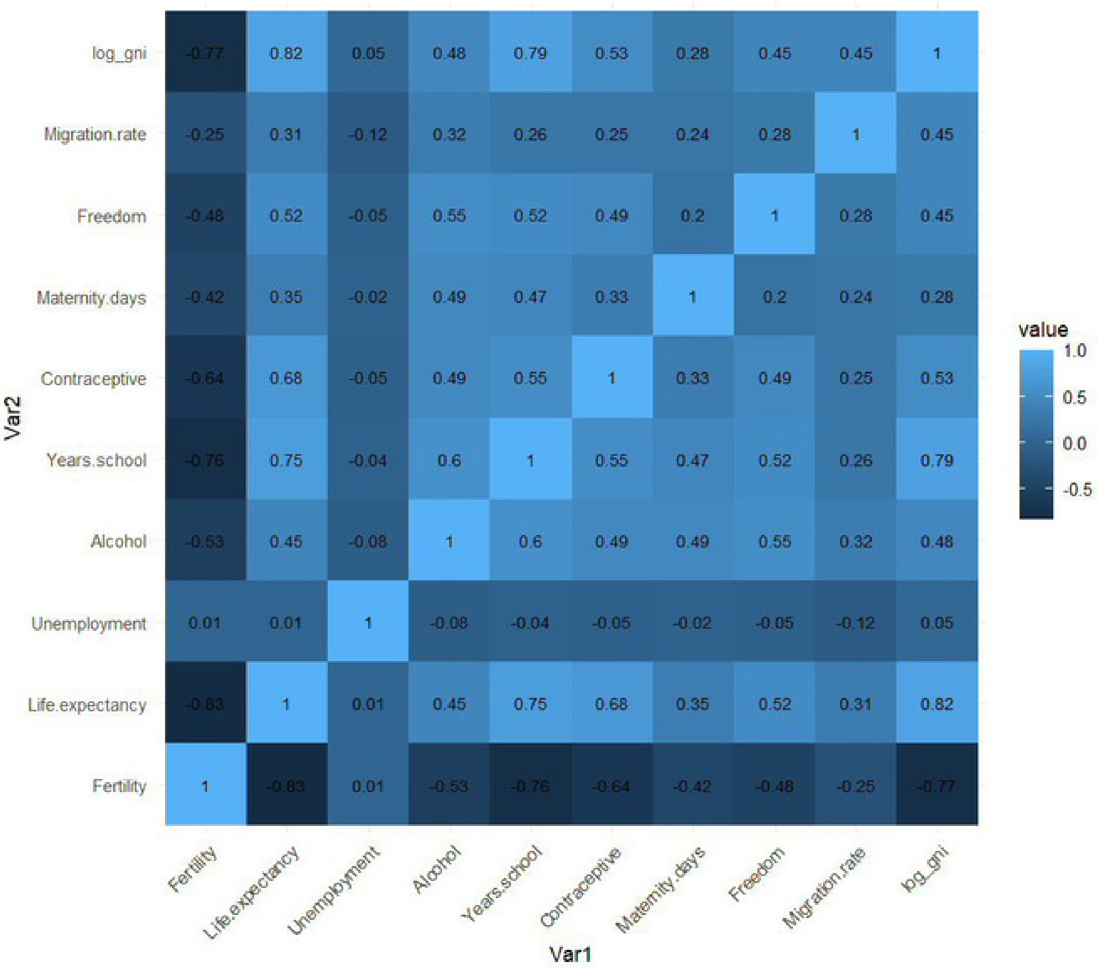
1. Scrape from different sources
2. Merge all datasets
3. Eliminate rows with a lot of missing values
4. Fill the missing values with the mean of predictors

↓

143 different Countries!

Construction of dataset

EDA: correlation matrix

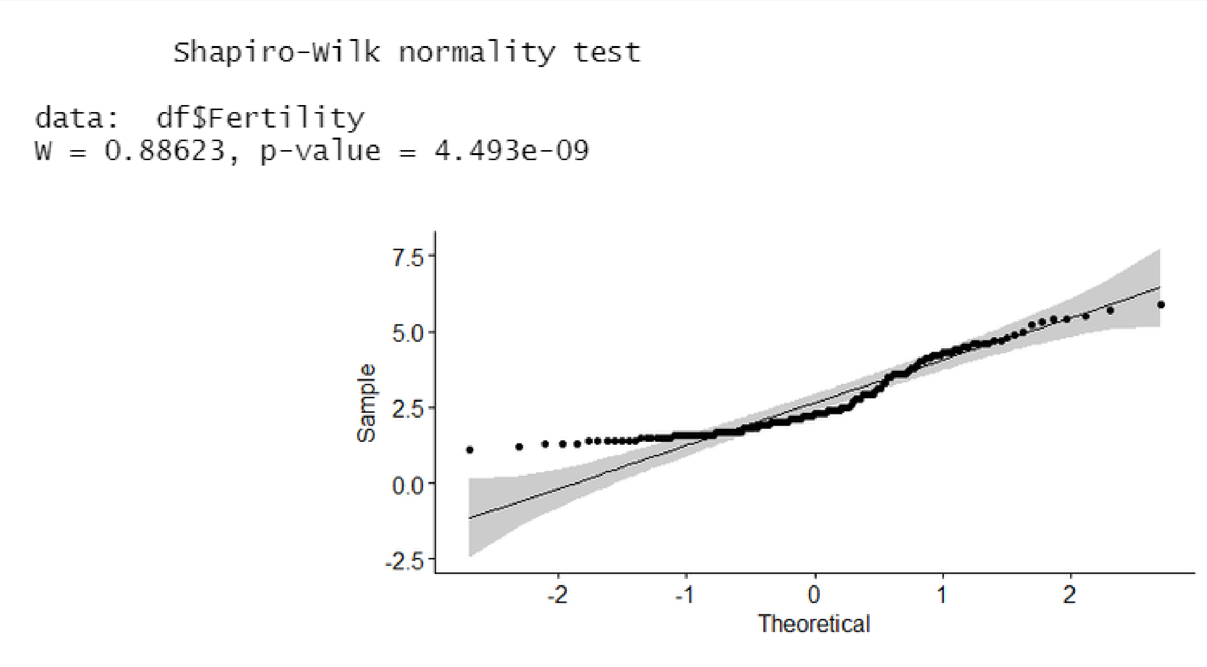


Var2	Fertility	Life.expectancy	Unemployment	Alcohol	Years.school	Contraceptive	Maternity.days	Freedom	Migration.rate	log_gni
log_gni	-0.77	0.82	0.05	0.48	0.79	0.53	0.28	0.45	0.45	1
Migration.rate	-0.25	0.31	-0.12	0.32	0.26	0.25	0.24	0.28	1	0.45
Freedom	-0.48	0.52	-0.05	0.55	0.52	0.49	0.2	1	0.28	0.45
Maternity.days	-0.42	0.35	-0.02	0.49	0.47	0.33	1	0.2	0.24	0.28
Contraceptive	-0.64	0.68	-0.05	0.49	0.55	1	0.33	0.49	0.25	0.53
Years.school	-0.76	0.75	-0.04	0.6	1	0.55	0.47	0.52	0.26	0.79
Alcohol	-0.53	0.45	-0.08	1	0.6	0.49	0.49	0.55	0.32	0.48
Unemployment	0.01	0.01	1	-0.08	-0.04	-0.05	-0.02	-0.05	-0.12	0.05
Life.expectancy	-0.83	1	0.01	0.45	0.75	0.68	0.35	0.52	0.31	0.82
Fertility	1	-0.83	0.01	-0.53	-0.76	-0.64	-0.42	-0.48	-0.25	-0.77

SUPERVISED LEARNING

Multiple linear regression

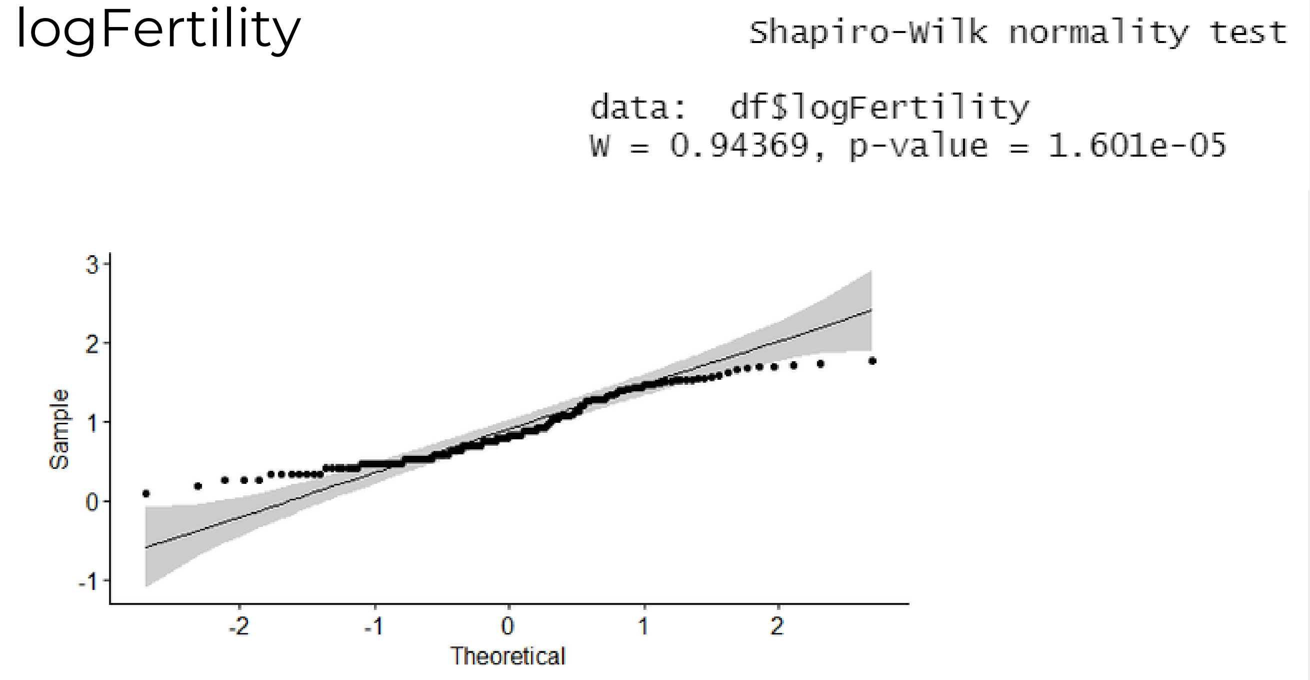
Target variable



```
        Shapiro-Wilk normality test

data:  df$Fertility
W = 0.88623, p-value = 4.493e-09
```

Fertility

logFertility

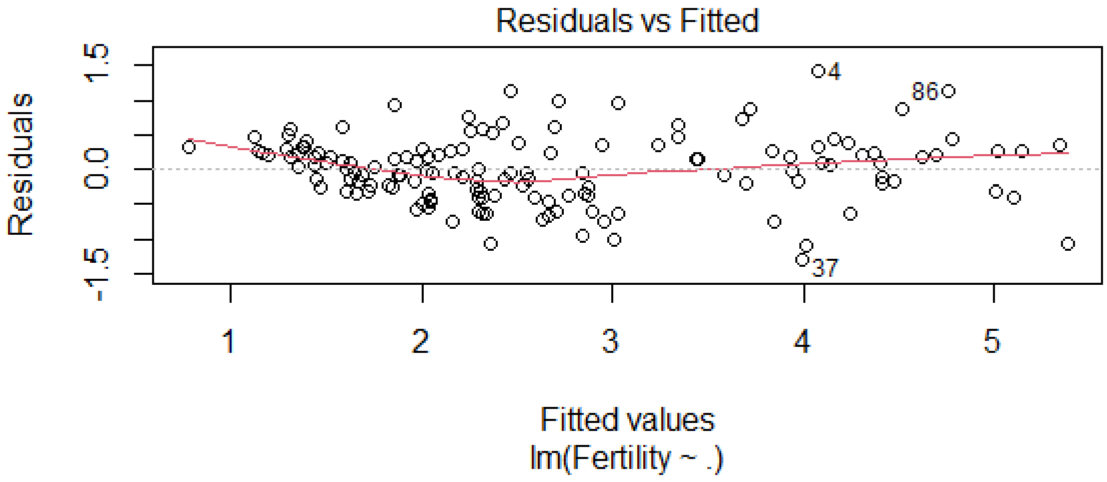


```
        Shapiro-Wilk normality test

data:  df$logFertility
W = 0.94369, p-value = 1.601e-05
```

SUPERVISED LEARNING

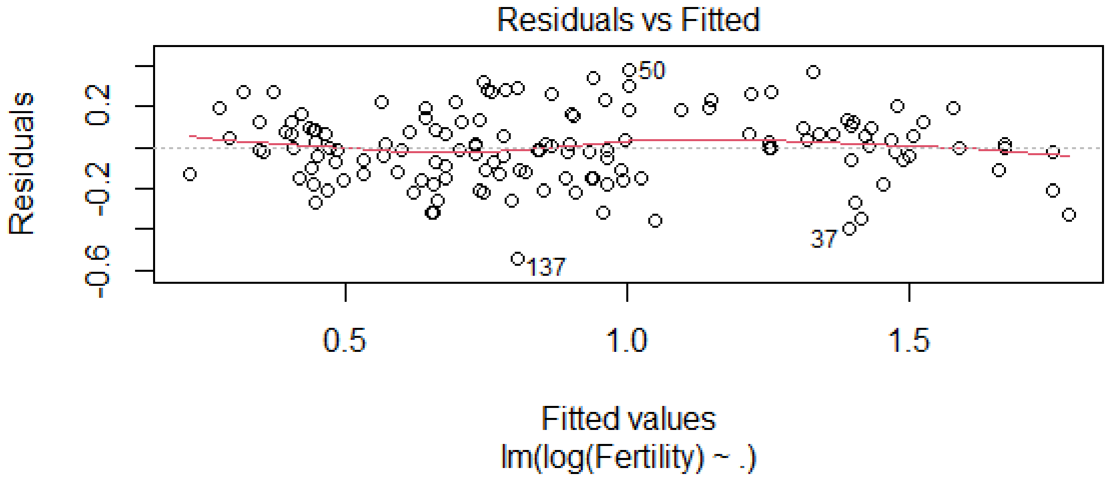
Multiple linear regression

Linear relationship between the preditcors and the response



Fertility

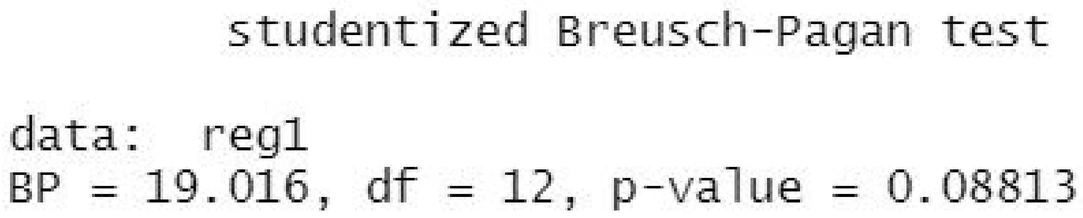
logFertility



SUPERVISED LEARNING

Multiple linear regression

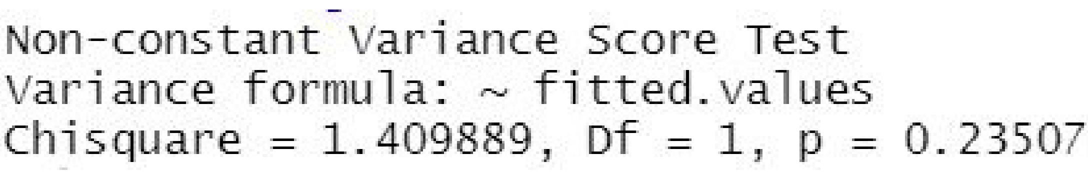
Homoscedasticity / Heteroscedasticity



```
           studentized Breusch-Pagan test

data:  reg1
BP = 19.016, df = 12, p-value = 0.08813
```

Breusch-Pagan test

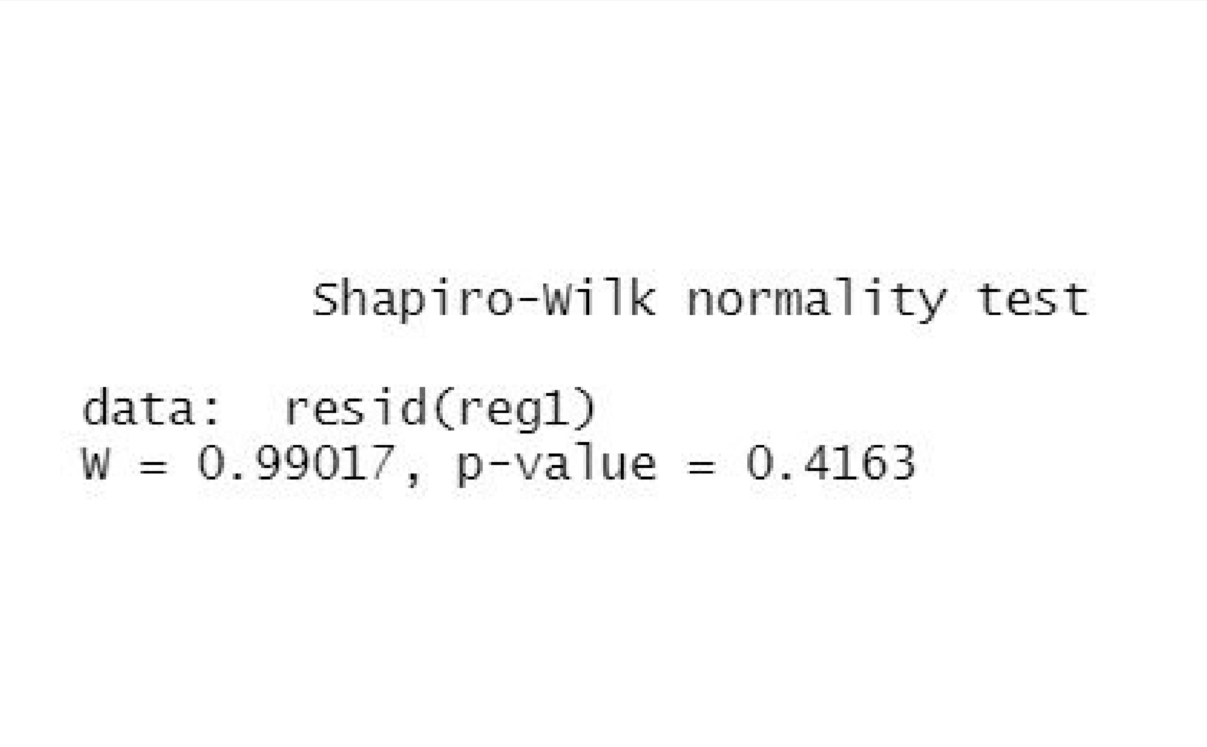
Non-constant variance score test



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.409889, Df = 1, p = 0.23507
```

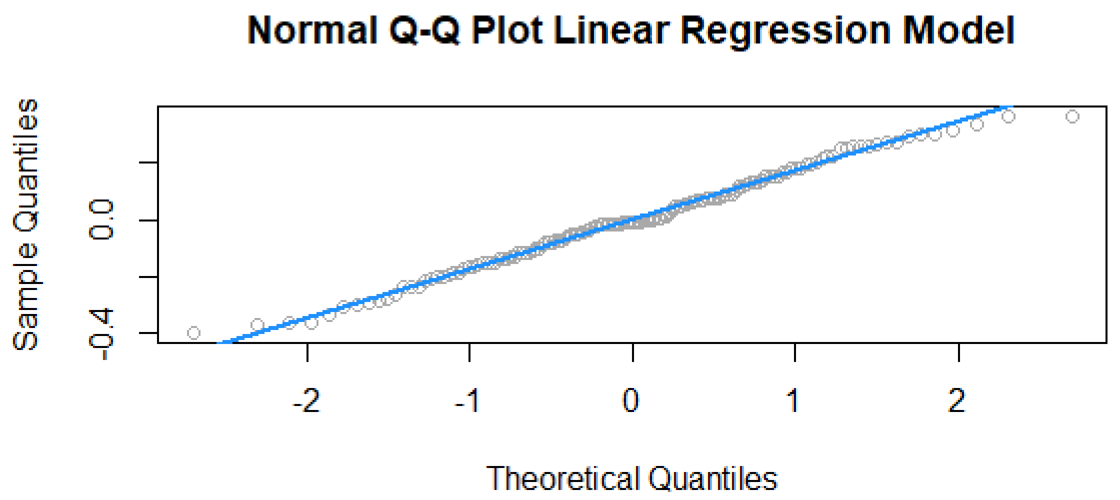
Multiple linear regression

Normality of the residuals



```
        Shapiro-Wilk normality test

data:  resid(reg1)
W = 0.99017, p-value = 0.4163
```

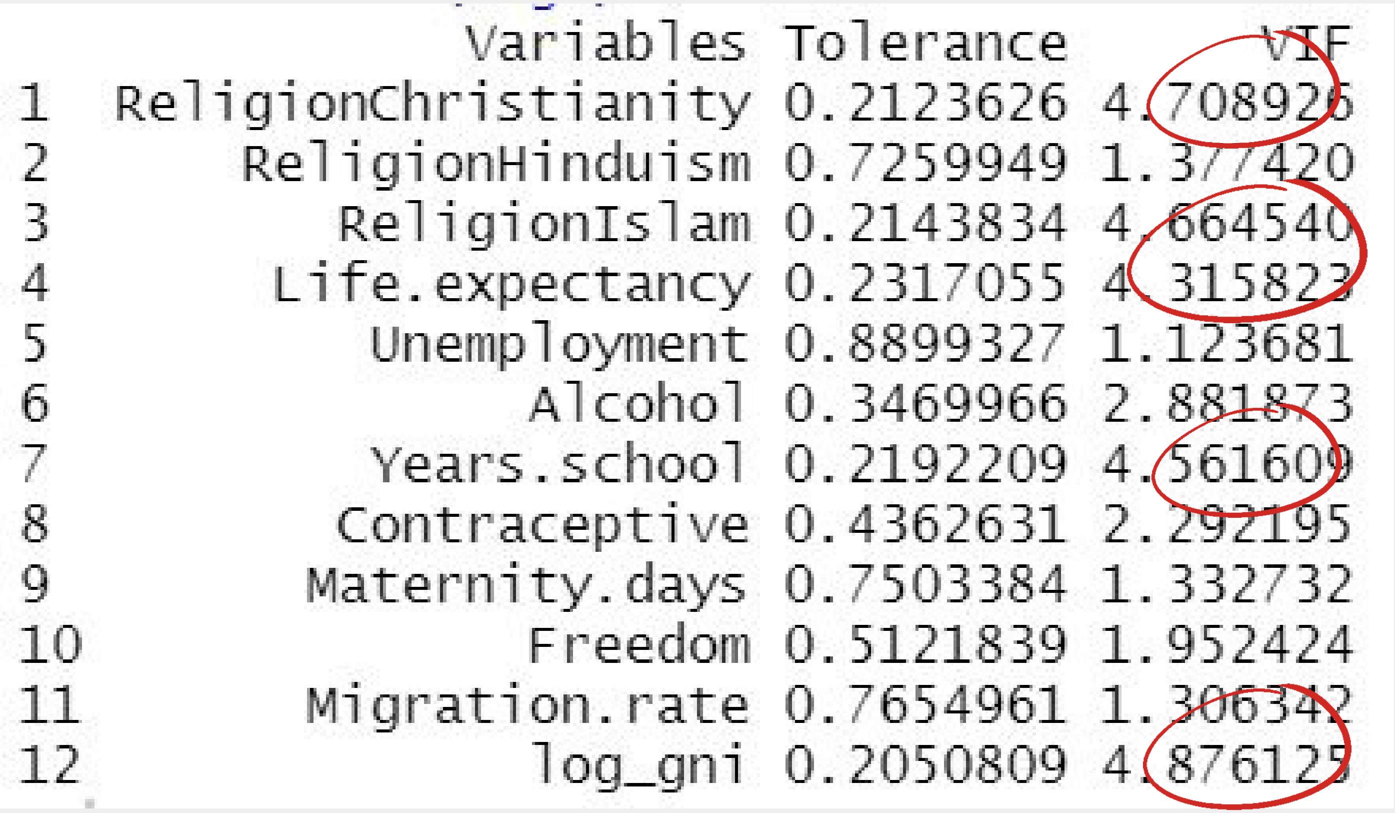
✓



SUPERVISED LEARNING

Multiple linear regression

Multicollinearity

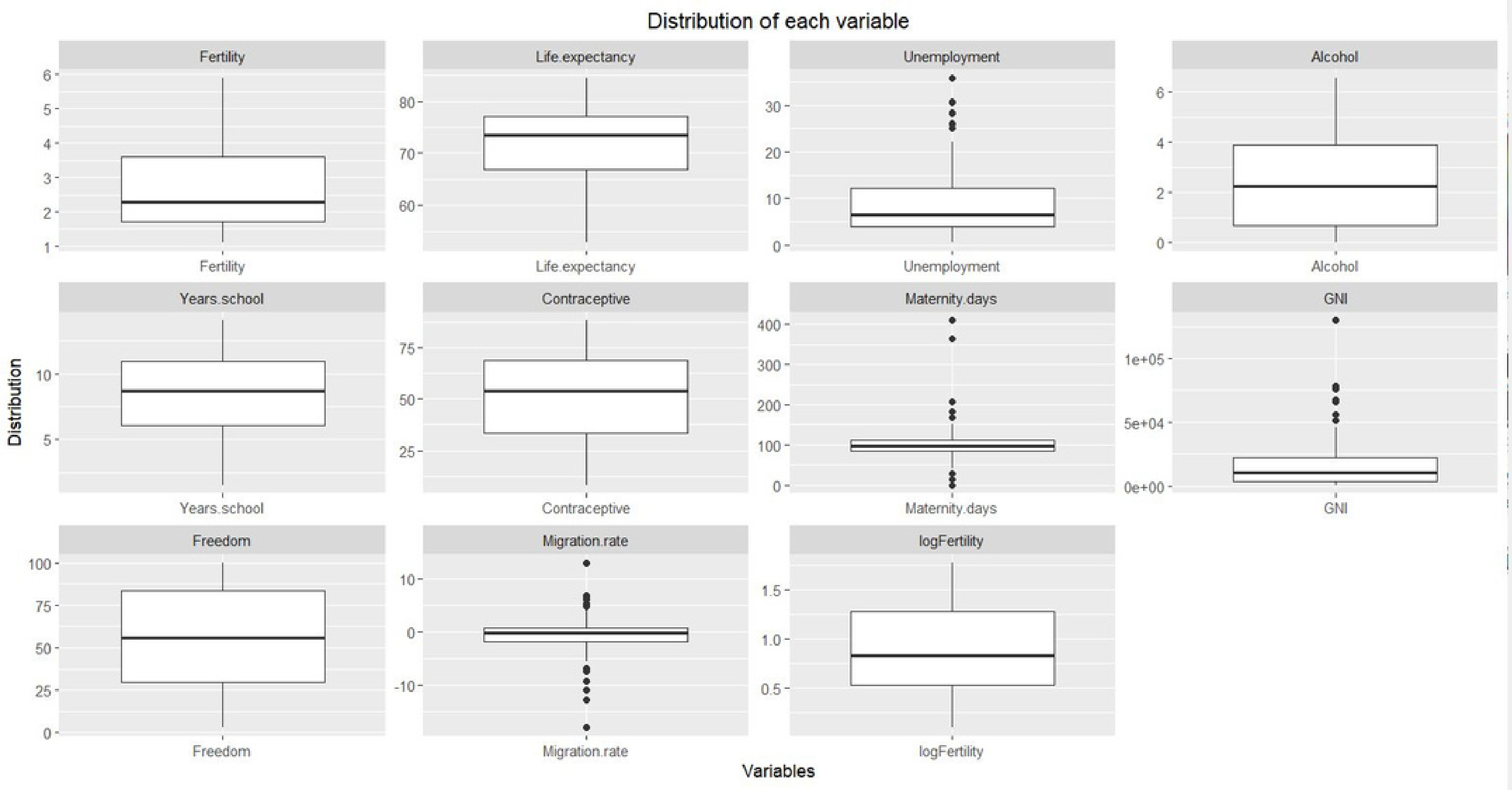


```
             Variables Tolerance      VIF
1  ReligionChristianity 0.2123626 4.708926
2       ReligionHinduism 0.7259949 1.377420
3          ReligionIslam 0.2143834 4.664540
4        Life.expectancy 0.2317055 4.315823
5           Unemployment 0.8899327 1.123681
6                Alcohol 0.3469966 2.881873
7           Years.school 0.2192209 4.561609
8          Contraceptive 0.4362631 2.292195
9         Maternity.days 0.7503384 1.332732
10               Freedom 0.5121839 1.952424
11        Migration.rate 0.7654961 1.306342
12               log_gni 0.2050809 4.876125
```

SUPERVISED LEARNING

Multiple linear regression

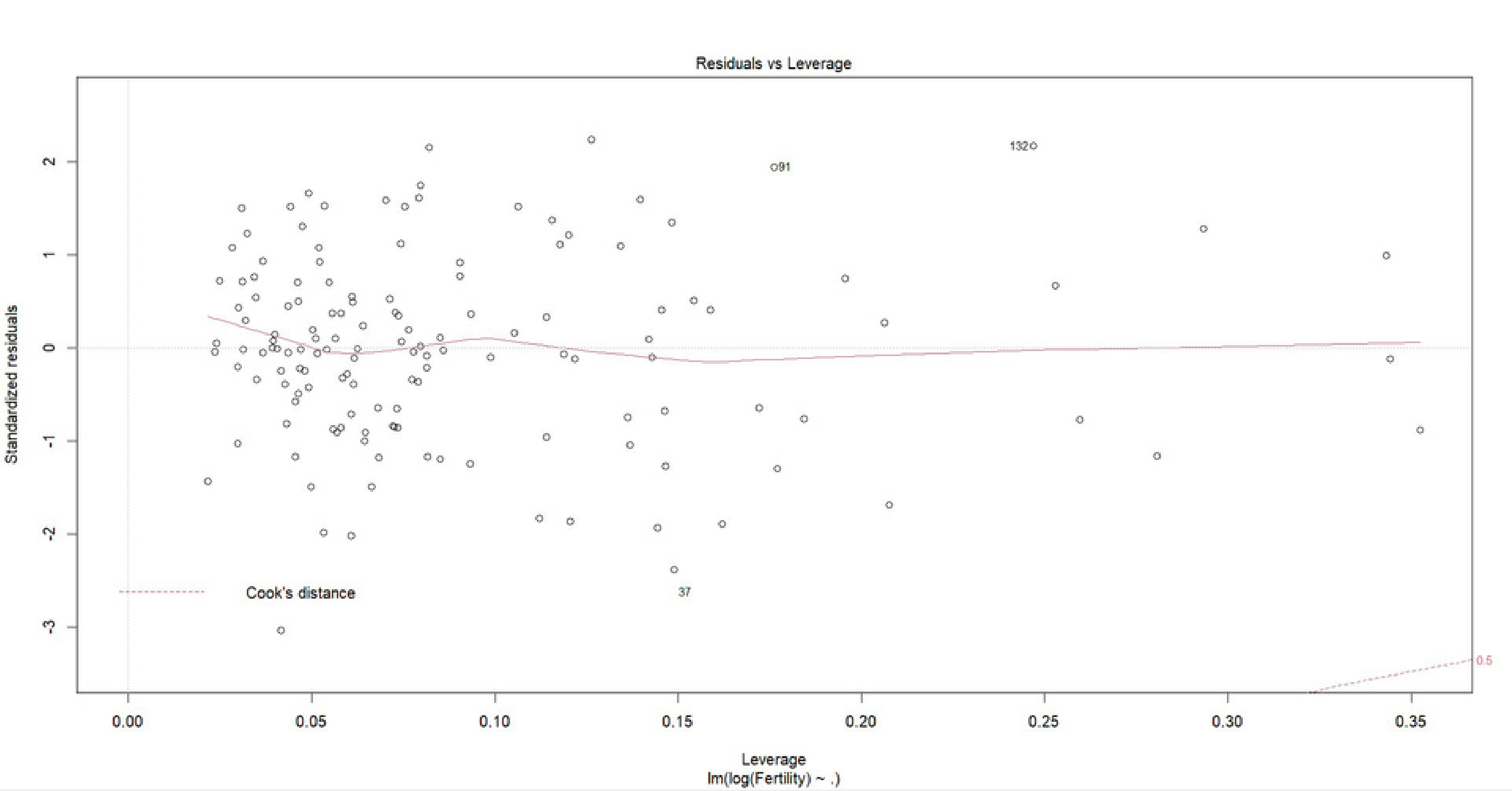
Outliers and leverage

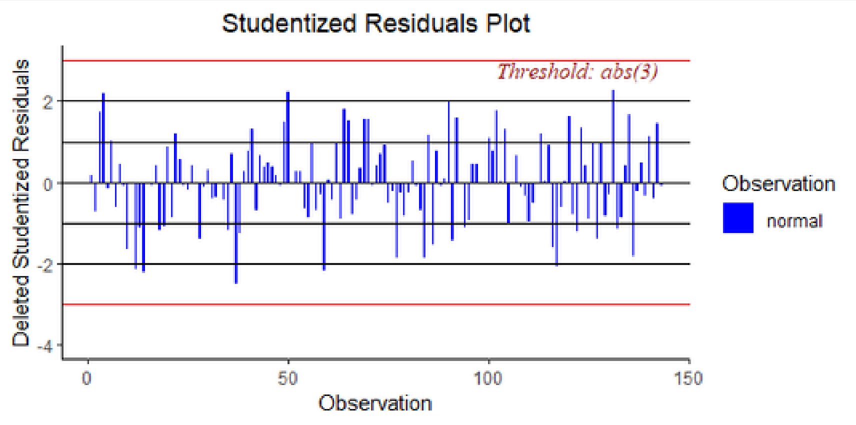


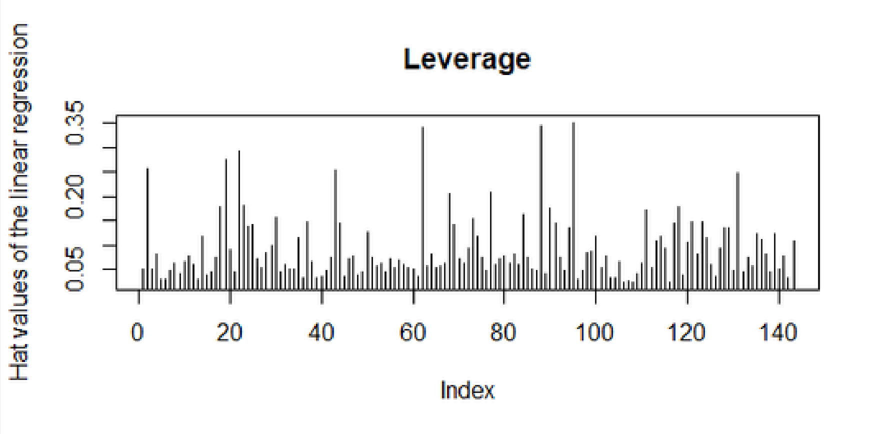
SUPERVISED LEARNING

Multiple linear regression

Outliers and leverage



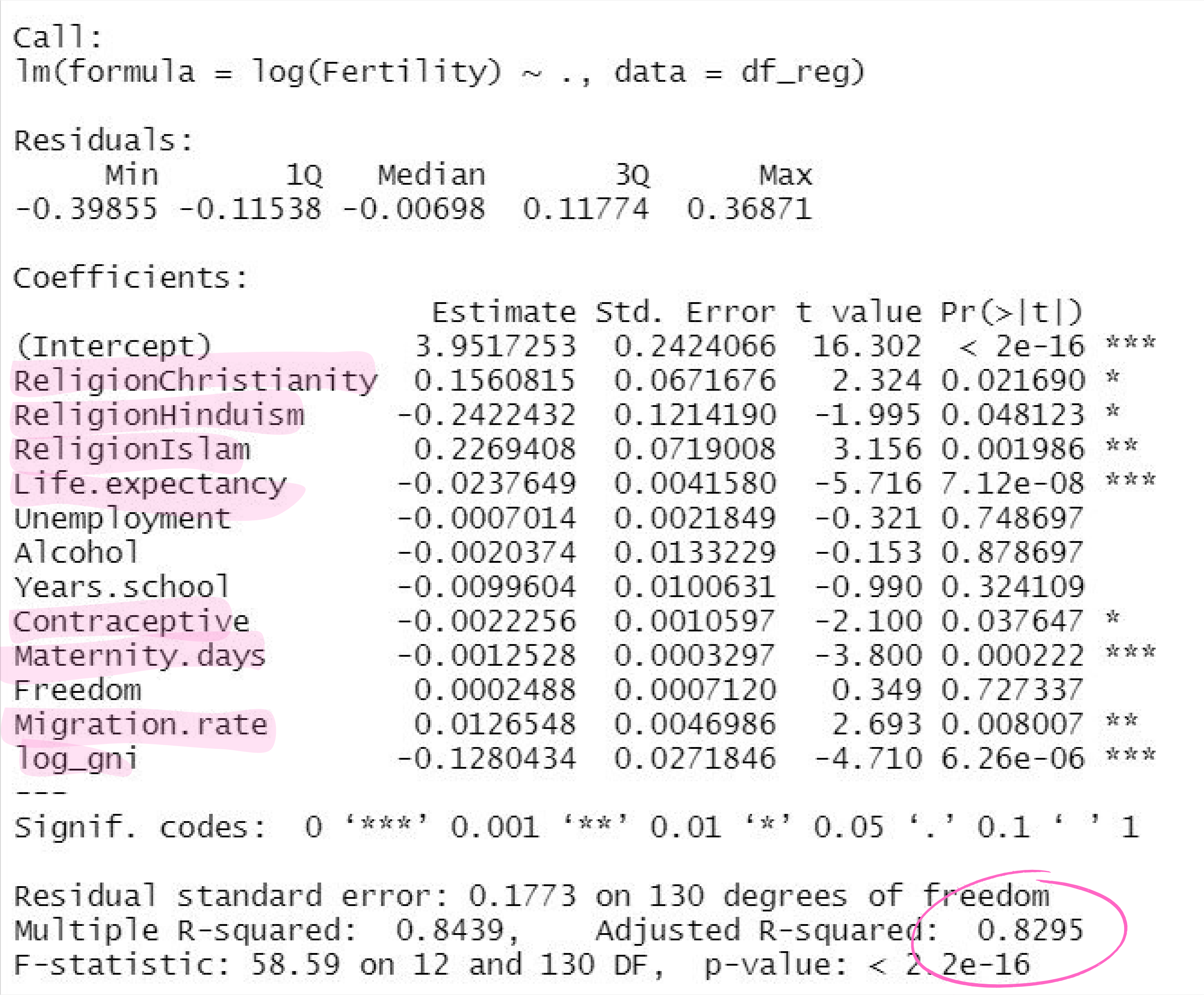




SUPERVISED LEARNING

Results

Multiple linear regression



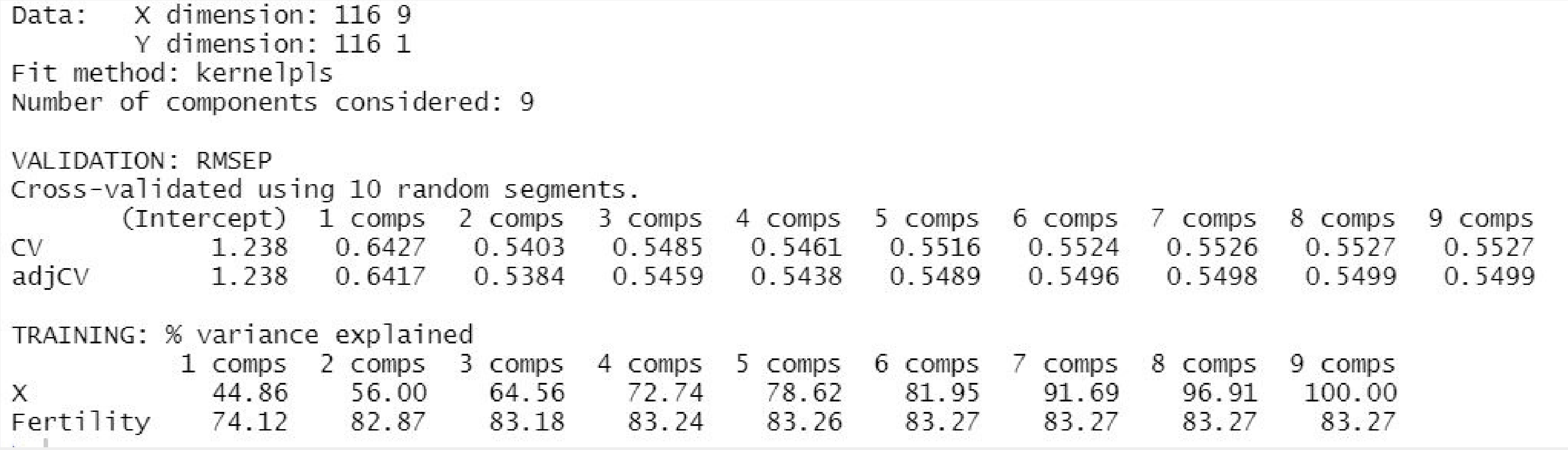
```
Call:
lm(formula = log(Fertility) ~ ., data = df_reg)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.39855 -0.11538 -0.00698  0.11774  0.36871 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.9517253  0.2424066  16.302  < 2e-16 ***
ReligionChristianity  0.1560815  0.0671676   2.324 0.021690 *  
ReligionHinduism     -0.2422432  0.1214190  -1.995 0.048123 *  
ReligionIslam         0.2269408  0.0719008   3.156 0.001986 ** 
Life.expectancy      -0.0237649  0.0041580  -5.716 7.12e-08 ***
Unemployment         -0.0007014  0.0021849  -0.321 0.748697    
Alcohol              -0.0020374  0.0133229  -0.153 0.878697    
Years.school         -0.0099604  0.0100631  -0.990 0.324109    
Contraceptive        -0.0022256  0.0010597  -2.100 0.037647 *  
Maternity.days       -0.0012528  0.0003297  -3.800 0.000222 ***
Freedom               0.0002488  0.0007120   0.349 0.727337    
Migration.rate        0.0126548  0.0046986   2.693 0.008007 ** 
log_gni              -0.1280434  0.0271846  -4.710 6.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1773 on 130 degrees of freedom
Multiple R-squared:  0.8439,	Adjusted R-squared:  0.8295 
F-statistic: 58.59 on 12 and 130 DF,  p-value: < 2.2e-16
```

K-FOLD CROSS VALIDATION TO TEST THE PERFORMANCE

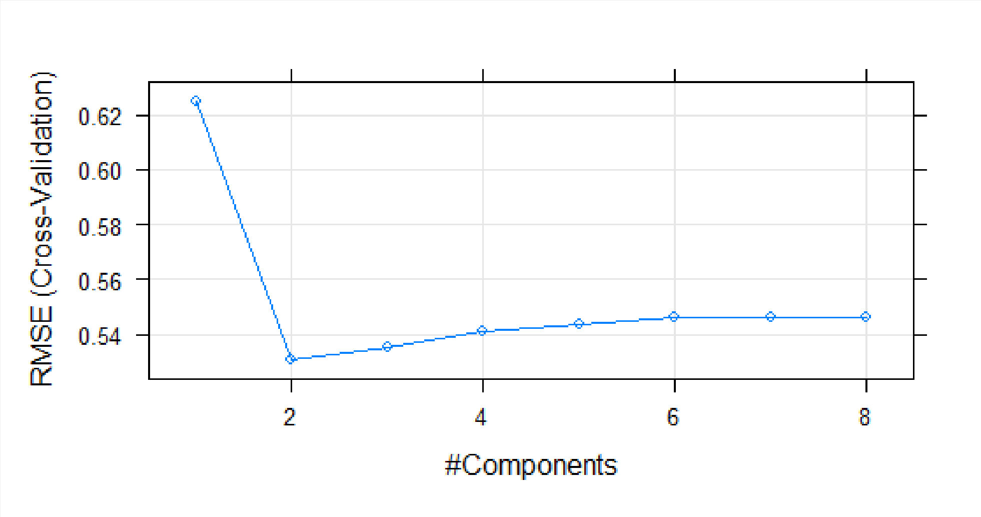
RMSE: 0.19

SUPERVISED LEARNING

Partial least square regression



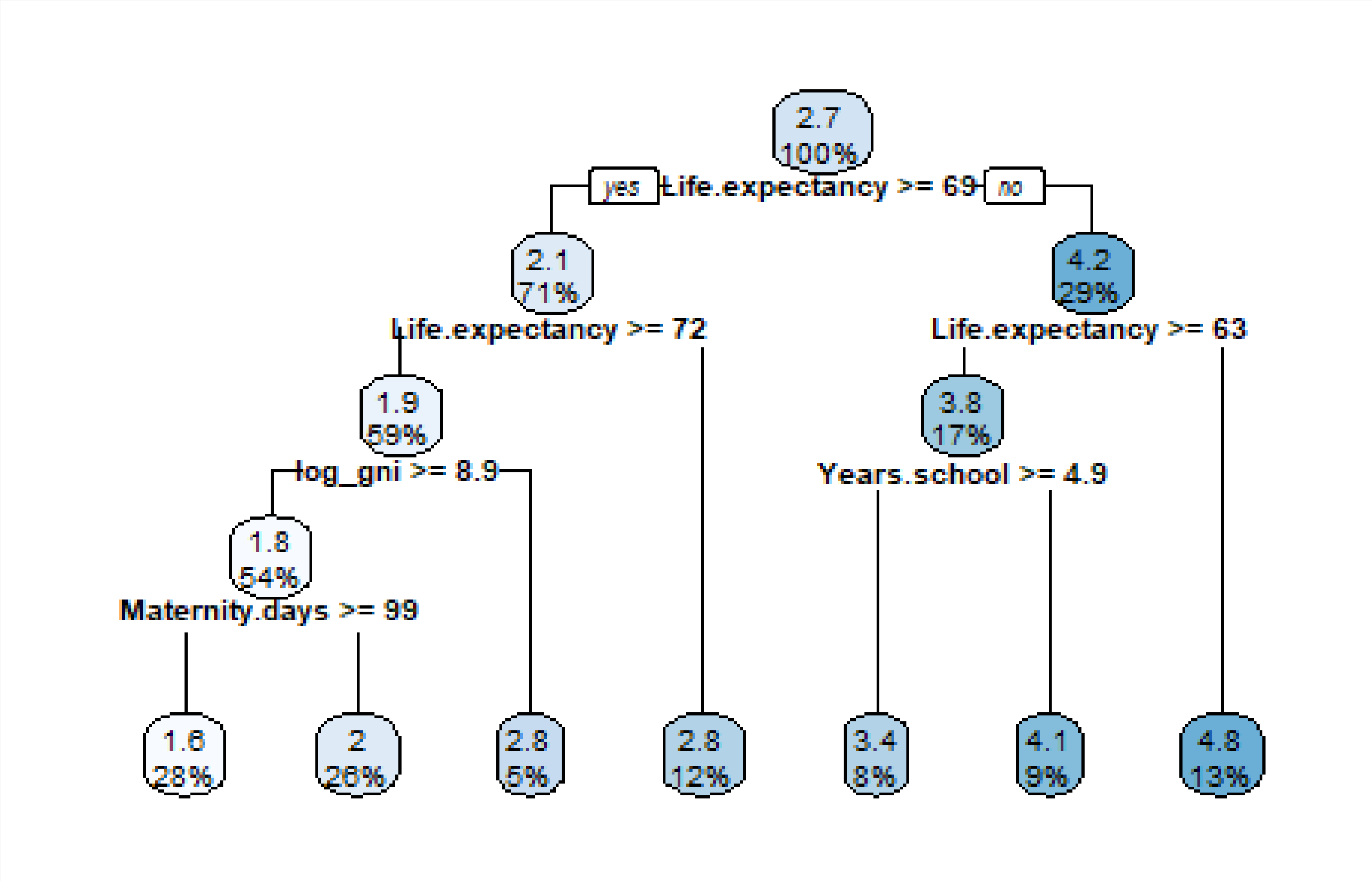
```
Data:    X dimension: 116 9
         Y dimension: 116 1
Fit method: kernelpls
Number of components considered: 9

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV           1.238   0.6427   0.5403   0.5485   0.5461   0.5516   0.5524   0.5526   0.5527   0.5527
adjCV        1.238   0.6417   0.5384   0.5459   0.5438   0.5489   0.5496   0.5498   0.5499   0.5499

TRAINING: % variance explained
           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X            44.86    56.00    64.56    72.74    78.62    81.95    91.69    96.91   100.00
Fertility    74.12    82.87    83.18    83.24    83.26    83.27    83.27    83.27    83.27
```



RMSE: 0.64

SUPERVISED LEARNING

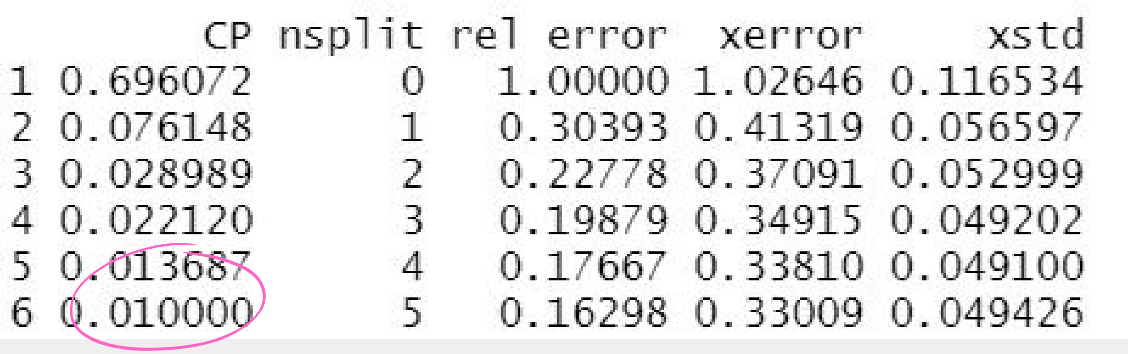
Regression trees



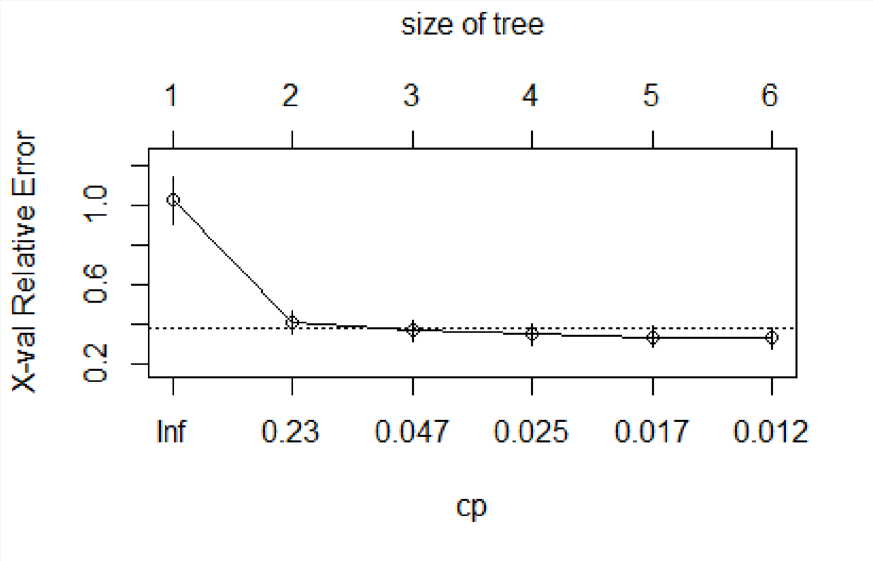
RMSE: 0.70

SUPERVISED LEARNING

Improving the performance: prune regression tree



```
        CP nsplit rel error  xerror     xstd
1 0.696072      0   1.00000 1.02646 0.116534
2 0.076148      1   0.30393 0.41319 0.056597
3 0.028989      2   0.22778 0.37091 0.052999
4 0.022120      3   0.19879 0.34915 0.049202
5 0.013687      4   0.17667 0.33810 0.049100
6 0.010000      5   0.16298 0.33009 0.049426
```



But can we try with 3 splits?

RMSE: 0.69

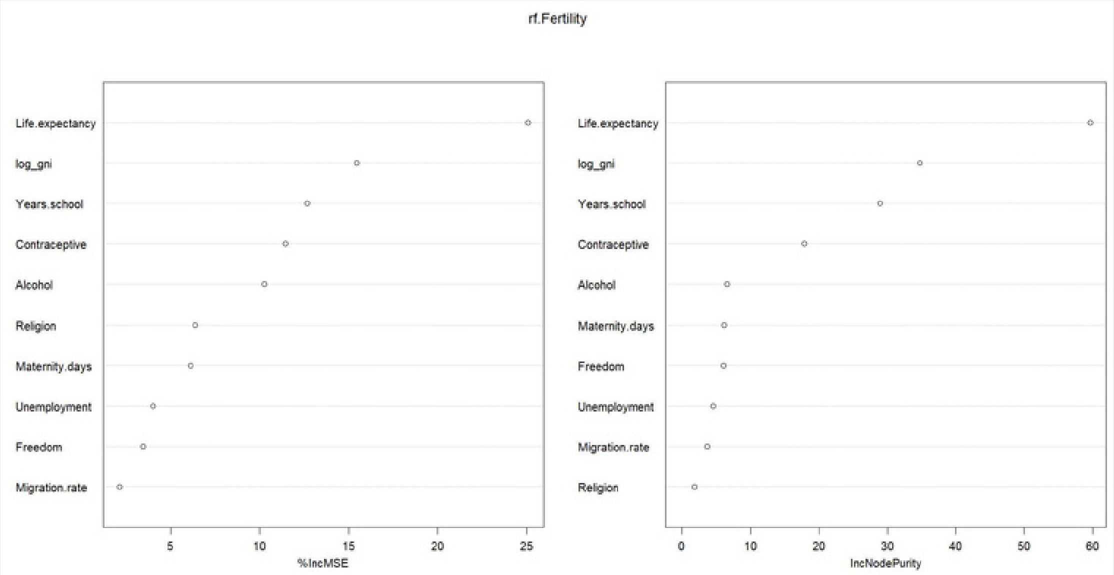
SUPERVISED LEARNING

Improving the performance: bagging and random forest

RMSE: 0.64

3 variables in each random split

RMSE: 0.55



UNSUPERVISED LEARNING

Clustering

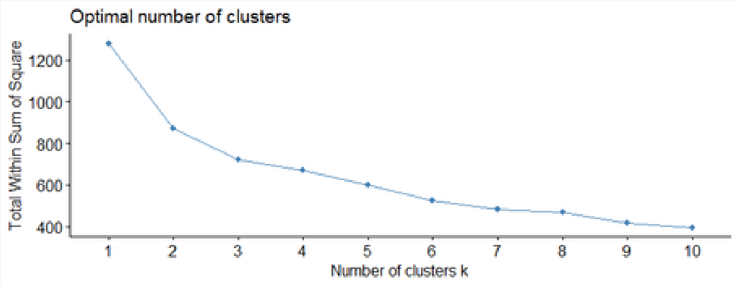
K-means on numerical variables

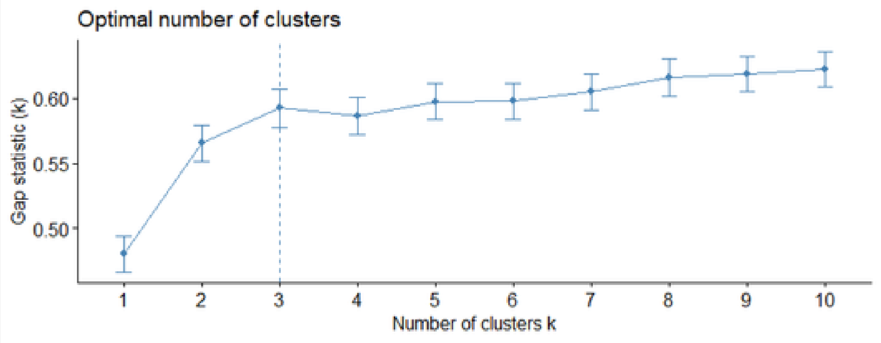
1) Variables are standardised

2) Fertility rate is not considered

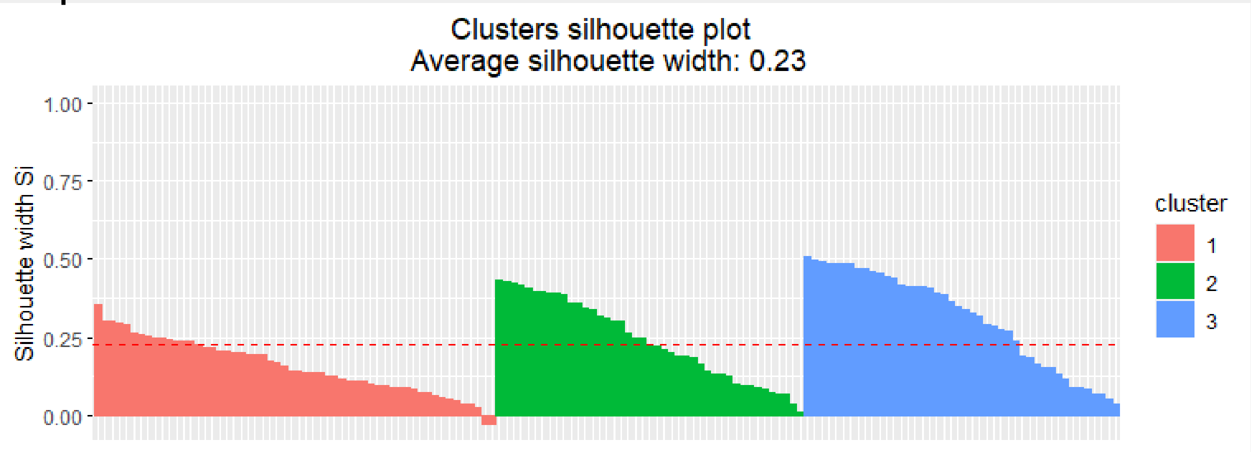
3) Choice of k!

3



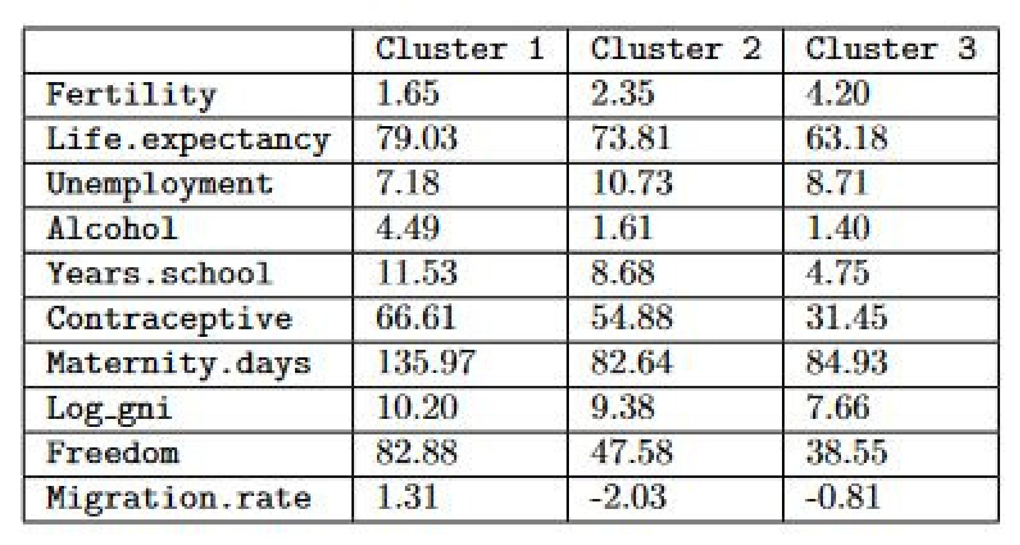


4) Evaluate the performance



Clustering

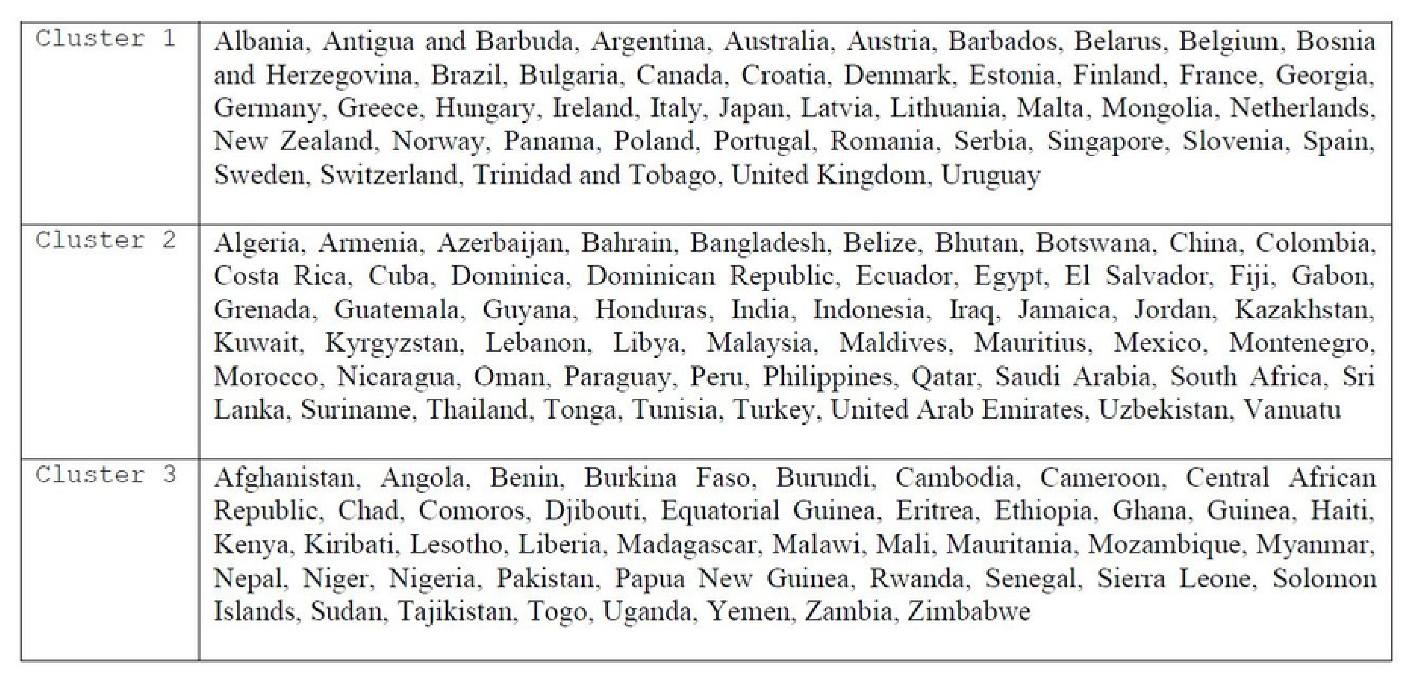
K-means on numerical variables



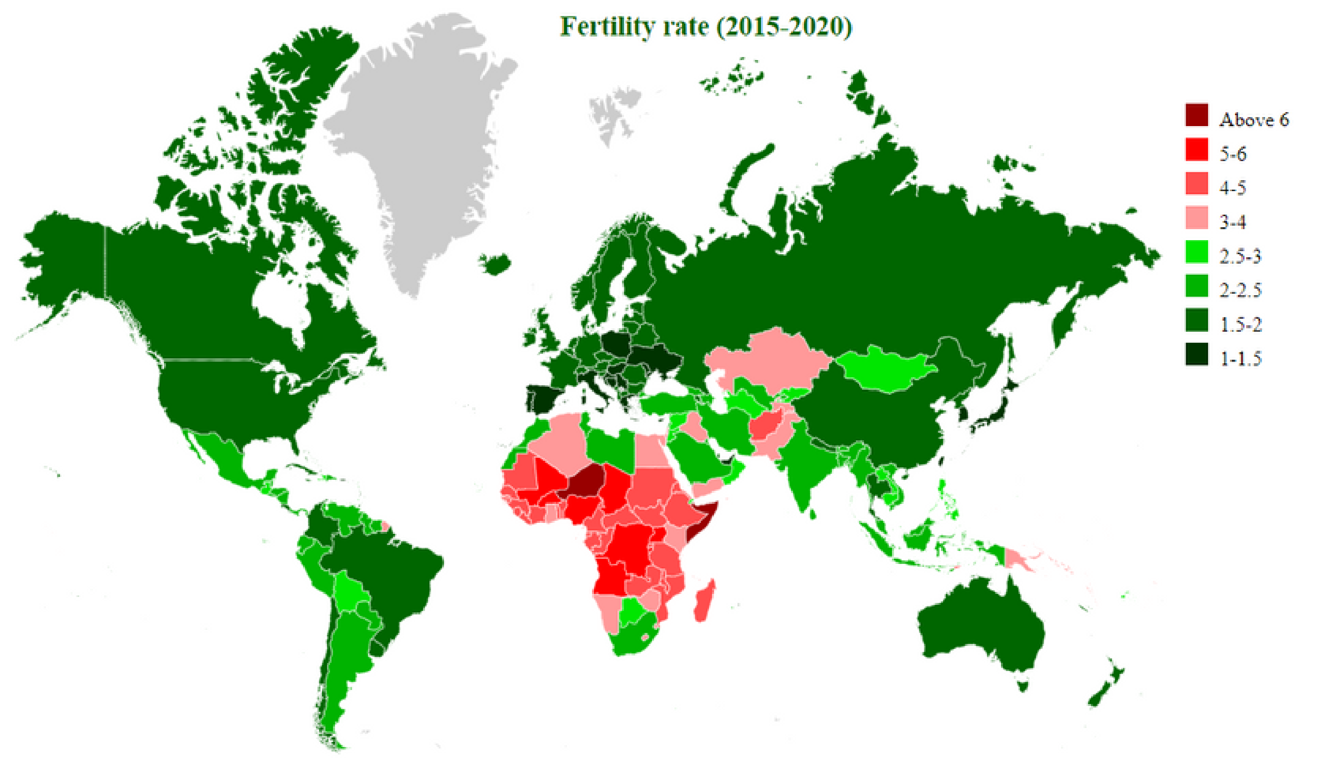
	Cluster 1	Cluster 2	Cluster 3
Fertility	1.65	2.35	4.20
Life.expectancy	79.03	73.81	63.18
Unemployment	7.18	10.73	8.71
Alcohol	4.49	1.61	1.40
Years.school	11.53	8.68	4.75
Contraceptive	66.61	54.88	31.45
Maternity.days	135.97	82.64	84.93
Log_gni	10.20	9.38	7.66
Freedom	82.88	47.58	38.55
Migration.rate	1.31	-2.03	-0.81

Mean values of variables

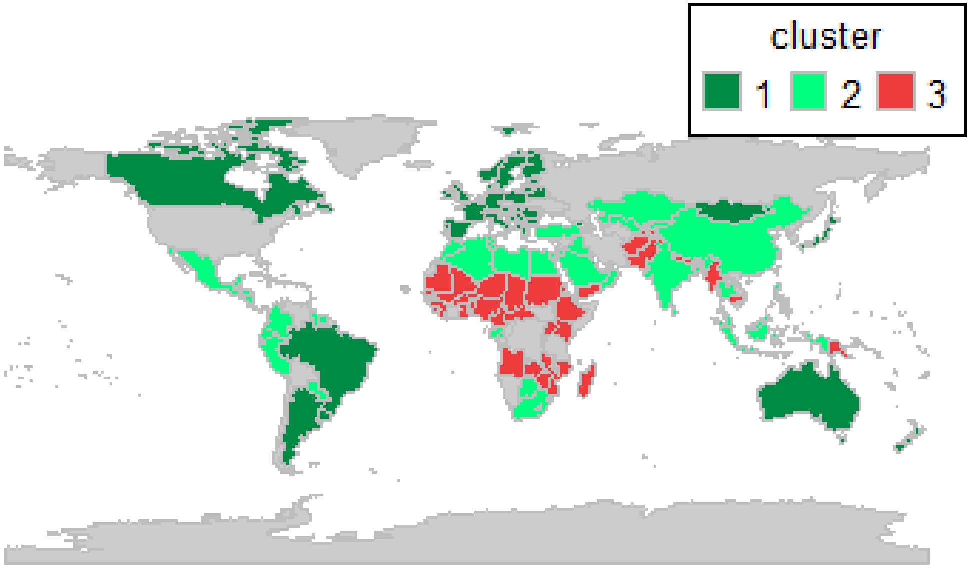
Countries



Cluster 1	Albania, Antigua and Barbuda, Argentina, Australia, Austria, Barbados, Belarus, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Canada, Croatia, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, Malta, Mongolia, Netherlands, New Zealand, Norway, Panama, Poland, Portugal, Romania, Serbia, Singapore, Slovenia, Spain, Sweden, Switzerland, Trinidad and Tobago, United Kingdom, Uruguay
Cluster 2	Algeria, Armenia, Azerbaijan, Bahrain, Bangladesh, Belize, Bhutan, Botswana, China, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Fiji, Gabon, Grenada, Guatemala, Guyana, Honduras, India, Indonesia, Iraq, Jamaica, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Libya, Malaysia, Maldives, Mauritius, Mexico, Montenegro, Morocco, Nicaragua, Oman, Paraguay, Peru, Philippines, Qatar, Saudi Arabia, South Africa, Sri Lanka, Suriname, Thailand, Tonga, Tunisia, Turkey, United Arab Emirates, Uzbekistan, Vanuatu
Cluster 3	Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Ghana, Guinea, Haiti, Kenya, Kiribati, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nepal, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Solomon Islands, Sudan, Tajikistan, Togo, Uganda, Yemen, Zambia, Zimbabwe



https://statisticstimes.com/demographics/countries-by-fertility-rate.php



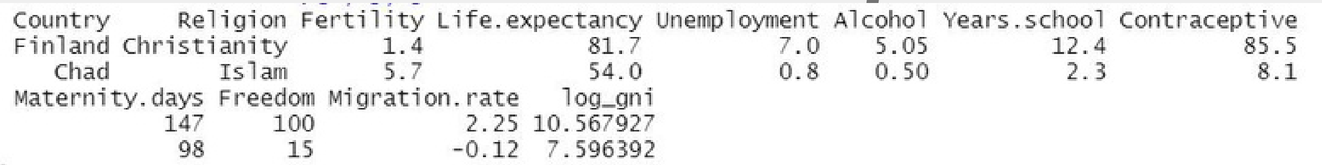
UNSUPERVISED LEARNING

Clustering

Gower distance - numerical and categorical variables

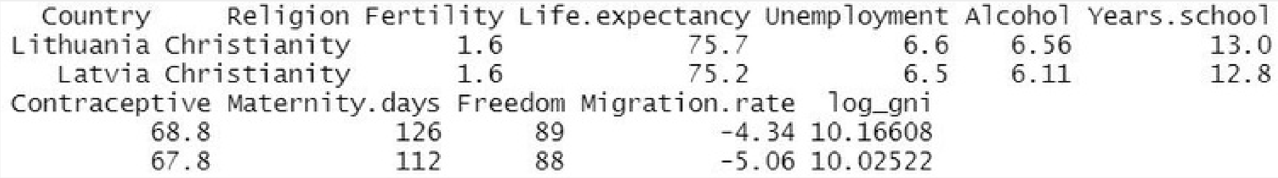
Distance matrix using gower distance

Most dissimilar pair



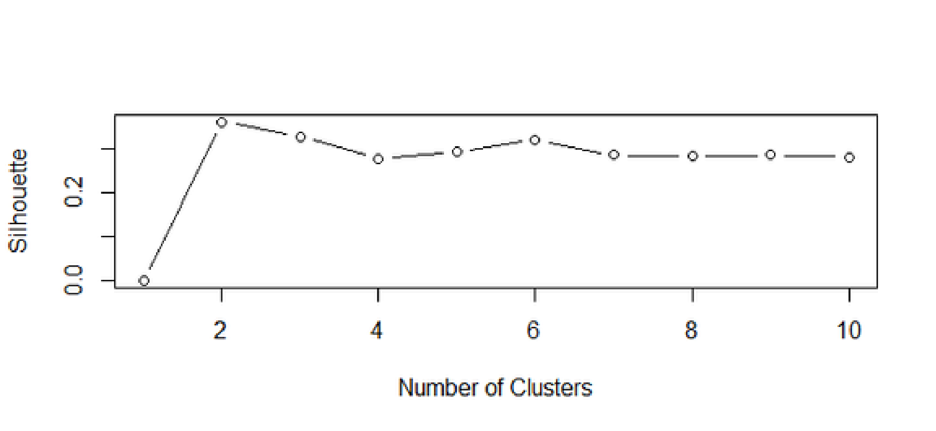
```
Country      Religion Fertility Life.expectancy Unemployment Alcohol Years.school Contraceptive
Finland Christianity       1.4            81.7          7.0    5.05         12.4          85.5
   Chad         Islam       5.7            54.0          0.8    0.50          2.3           8.1
Maternity.days Freedom Migration.rate   log_gni
           147     100           2.25 10.567927
            98      15          -0.12  7.596392
```

Most similar pair



```
  Country     Religion Fertility Life.expectancy Unemployment Alcohol Years.school
Lithuania Christianity       1.6            75.7          6.6    6.56         13.0
   Latvia Christianity       1.6            75.2          6.5    6.11         12.8
Contraceptive Maternity.days Freedom Migration.rate  log_gni
         68.8            126      89          -4.34 10.16608
         67.8            112      88          -5.06 10.02522
```

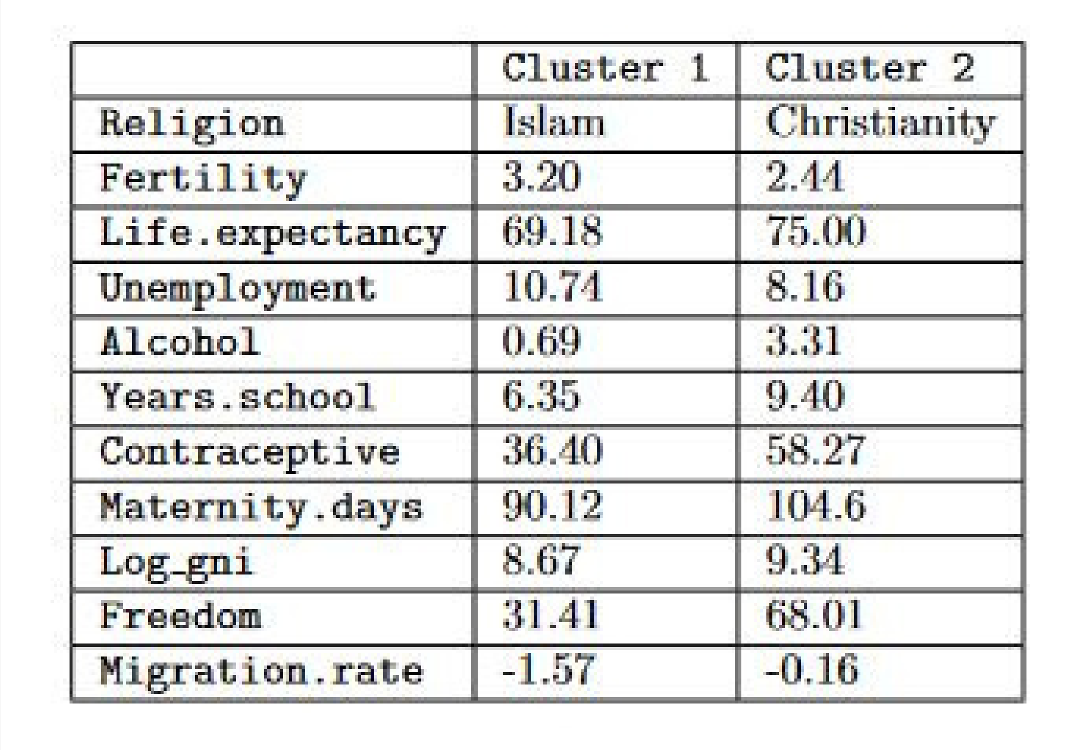
PAM ALGORITHM (partitioning and medoids)



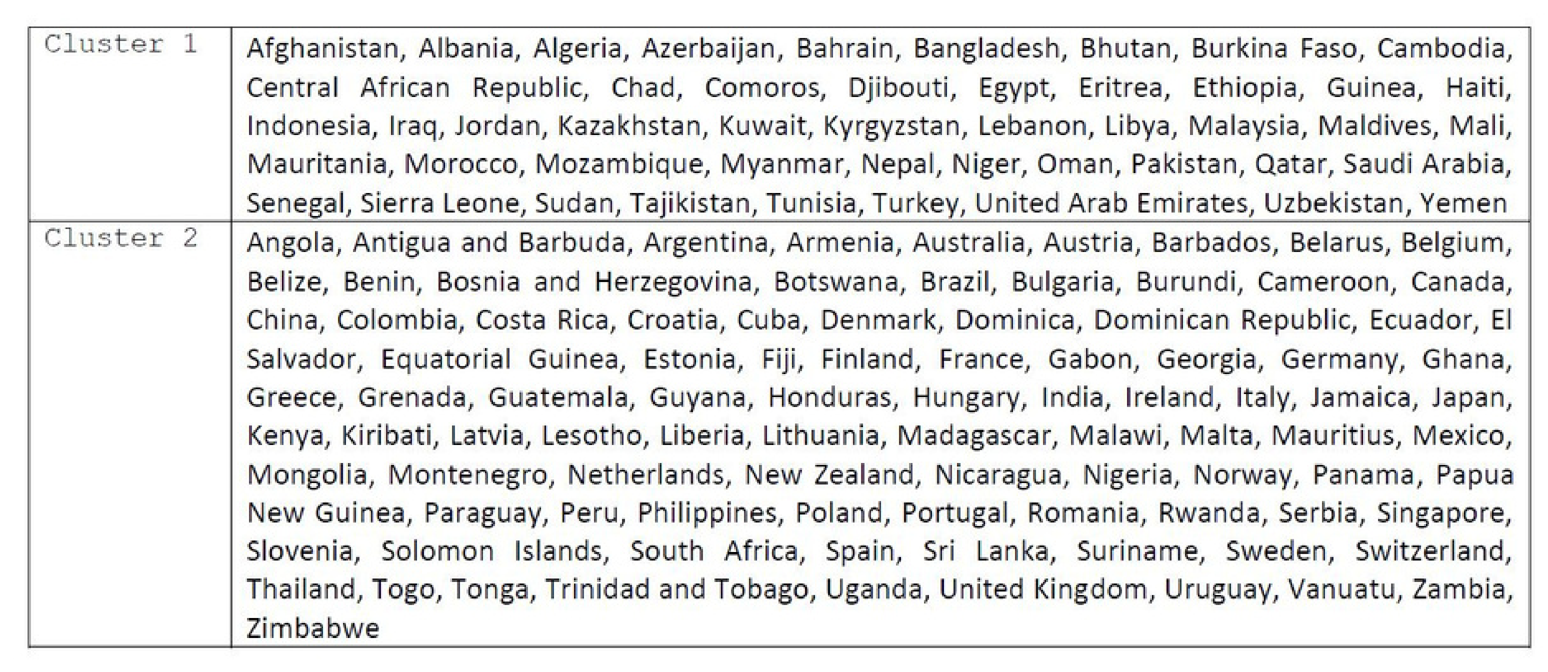
Clustering

Gower distance - numerical and categorical variables

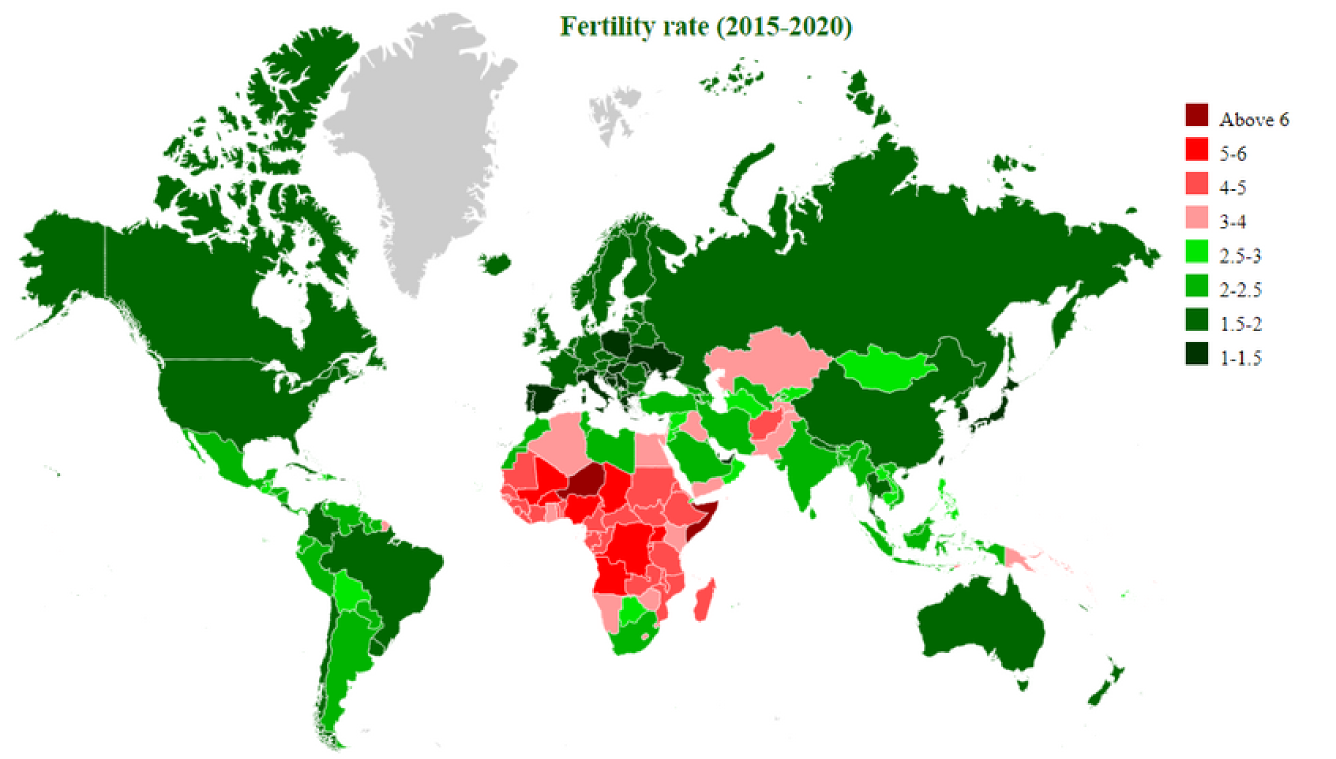
UNSUPERVISED LEARNING



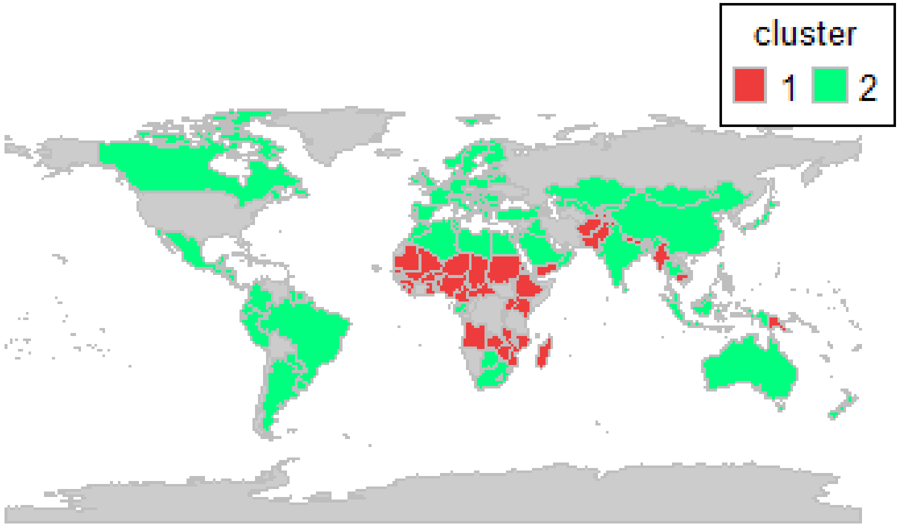
	Cluster 1	Cluster 2
Religion	Islam	Christianity
Fertility	3.20	2.44
Life.expectancy	69.18	75.00
Unemployment	10.74	8.16
Alcohol	0.69	3.31
Years.school	6.35	9.40
Contraceptive	36.40	58.27
Maternity.days	90.12	104.6
Log_gni	8.67	9.34
Freedom	31.41	68.01
Migration.rate	-1.57	-0.16



Cluster 1	Afghanistan, Albania, Algeria, Azerbaijan, Bahrain, Bangladesh, Bhutan, Burkina Faso, Cambodia, Central African Republic, Chad, Comoros, Djibouti, Egypt, Eritrea, Ethiopia, Guinea, Haiti, Indonesia, Iraq, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Libya, Malaysia, Maldives, Mali, Mauritania, Morocco, Mozambique, Myanmar, Nepal, Niger, Oman, Pakistan, Qatar, Saudi Arabia, Senegal, Sierra Leone, Sudan, Tajikistan, Tunisia, Turkey, United Arab Emirates, Uzbekistan, Yemen
Cluster 2	Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Barbados, Belarus, Belgium, Belize, Benin, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burundi, Cameroon, Canada, China, Colombia, Costa Rica, Croatia, Cuba, Denmark, Dominica, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Estonia, Fiji, Finland, France, Gabon, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guyana, Honduras, Hungary, India, Ireland, Italy, Jamaica, Japan, Kenya, Kiribati, Latvia, Lesotho, Liberia, Lithuania, Madagascar, Malawi, Malta, Mauritius, Mexico, Mongolia, Montenegro, Netherlands, New Zealand, Nicaragua, Nigeria, Norway, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Rwanda, Serbia, Singapore, Slovenia, Solomon Islands, South Africa, Spain, Sri Lanka, Suriname, Sweden, Switzerland, Thailand, Togo, Tonga, Trinidad and Tobago, Uganda, United Kingdom, Uruguay, Vanuatu, Zambia, Zimbabwe



https://statisticstimes.com/demographics/countries-by-fertility-rate.php



Thanks for the attention!