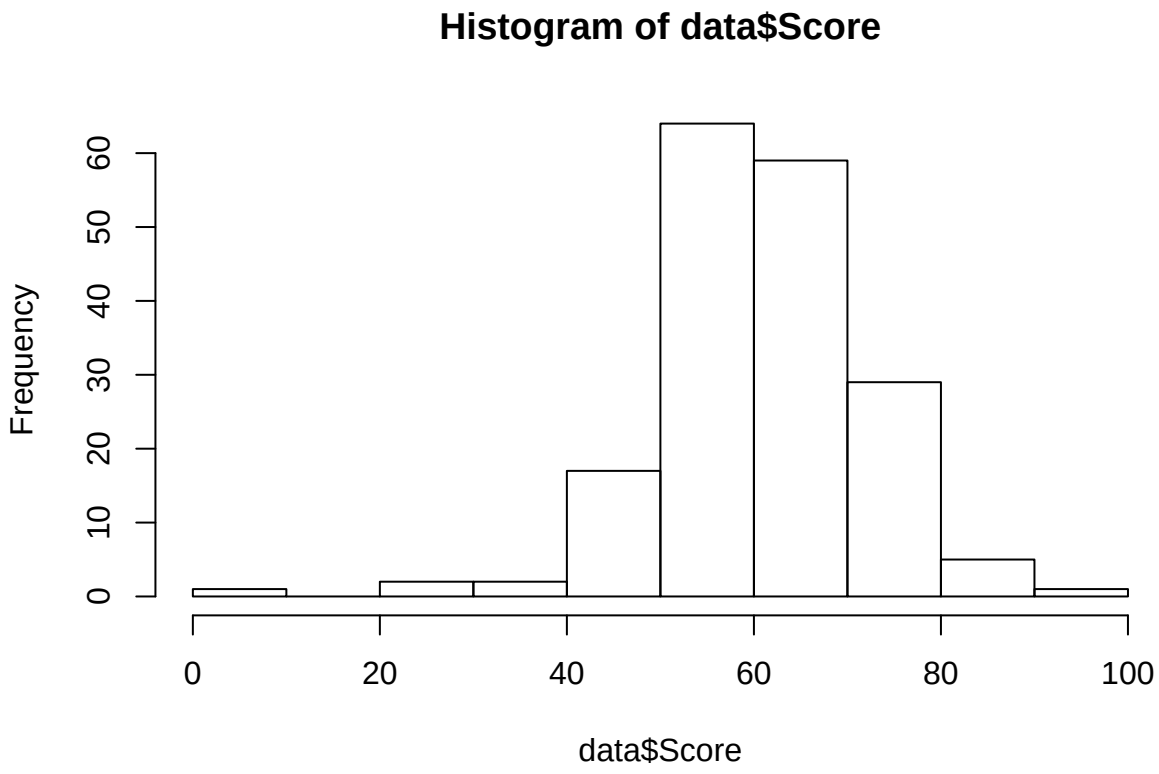
# outliers

## Exploratory Analysis

The first basic step to do once the data is loaded is a basic **exploratory analysis**. We can see that our dataset contains 186 rows (countries) with 27 numerical features and 1 categorical indicator (geographical region).

Instead of going through all the variables we can focus on two of the most interesting ones, `Score` and `Region`.

```
data <- read.delim("MVA-dataset.csv", sep=",",row.names = 1, header = TRUE)
hist(data$Score)
```




We can see how the Index of Economic Freedom (`Score`) roughly follows a $\mathcal{N}(60, 125)$. On the low extreme we have the outlier `North Korea` while, on the other hand, `Hong Kong` is the most *economically free* country.

```
table(data$Region)
```

```
##
##                     Americas                 Asia-Pacific
##                           32                           43
##                       Europe Middle East and North Africa
##                           45                           18
##           Sub-Saharan Africa
##                           48
```

Countries are divided in 4 `Region`s with a similar number of components. (except `Middle East & N Africa`).

So, now that we have some general information, lets continue with pre-processing.

## Missing values

Our dataset contains many missing values and to decide how to deal with them we used three different methods: kNN, random forest and mice. All of those generated different results, but after analyzing and comparing the imputed values with the real values that we found on the internet, we decided that the method which performs better is the kNN. Before imputing we decided to delete some variables becuase more than 60% of their information were missing values.

```
library(DMwR)
```

```
## Warning: package 'DMwR' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: grid
```

```
knn.imput <-knnImputation(dataset, k=1, scale = T)
```

## Outlier detection

After imputing missing values comes the outlier detection. To detect the outliers, we used two different methods to be sure about the result. The first method used, is the one we have seen during our classes, Mahalanobis distance.

```
## Warning: package 'chemometrics' was built under R version 3.5.3
```

```
## Loading required package: rpart
```

```
## Warning: package 'rpart' was built under R version 3.5.3
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## $md
##             Afghanistan                 Albania                 Algeria
##                4.076533                3.910110                4.062490
##                  Angola               Argentina                 Armenia
##                4.428757                3.740111                4.756203
##               Australia                 Austria              Azerbaijan
##                4.322788                3.401732                4.399406
##                 Bahamas                 Bahrain              Bangladesh
##                6.460593                5.592262                4.603689
##                Barbados                 Belarus                 Belgium
##                4.666542                5.369522                3.963309
##                  Belize                   Benin                   Bhutan
##                3.573001                4.400564                4.649188
##                 Bolivia      BosniaHerzegovina                Botswana
```

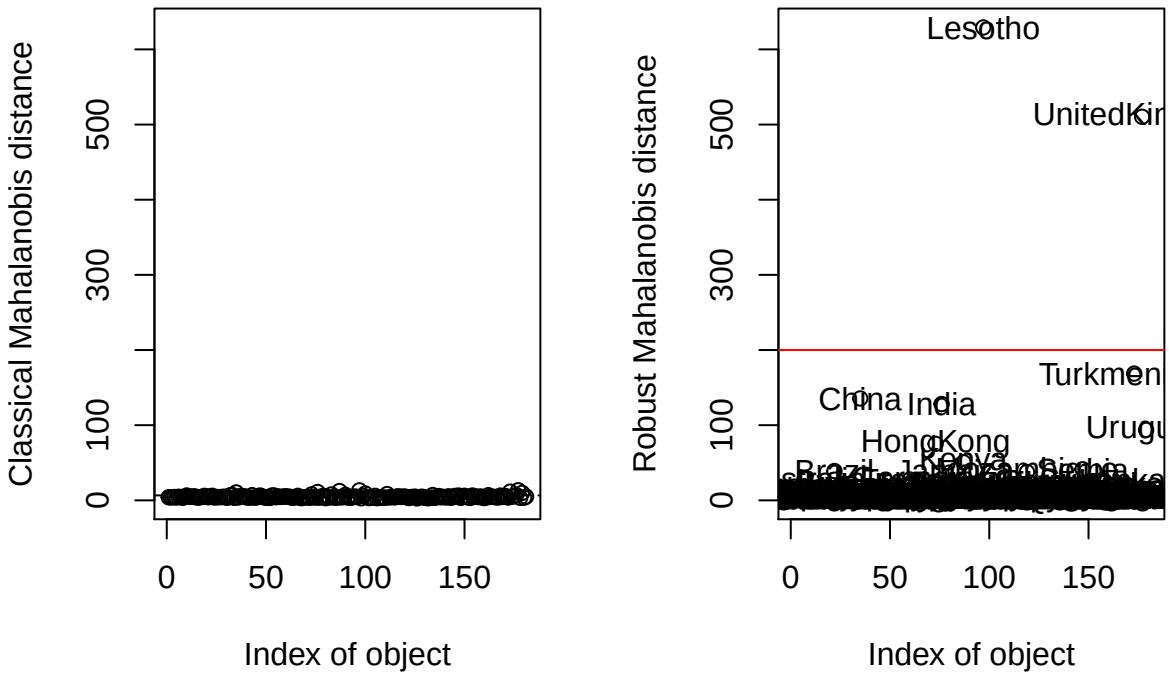
```
##            5.964052            5.185743            4.124946
##              Brazil              Brunei            Bulgaria
##            5.371850            6.015520            3.753417
##         BurkinaFaso               Burma             Burundi
##            4.435209            4.053771            4.473788
##            Cambodia            Cameroon              Canada
##            4.402435            4.900629            3.539825
##           CaboVerde CentralAfricanRepublic                Chad
##            3.809415            6.762451            7.198552
##               Chile               China            Colombia
##            3.355693           10.539345            3.748396
##             Comoros DemcoraticRepublicCongo      RepublicCongo
##            5.766695            5.373480            5.463298
##          CostaRica         CoeDivoire             Croatia
##            4.055243            3.688669            3.603996
##                Cuba              Cyprus       CzechRepublic
##            6.917202            3.741564            3.686150
##             Denmark            Djibouti            Dominica
##            5.341982            6.453585            5.307692
##   DominicanRepublic             Ecuador               Egypt
##            3.986699            3.440704            4.305758
##          ElSalvador     EquatorialGuinea             Eritrea
##            3.190977            6.362269            6.499814
##             Estonia            Eswatini            Ethiopia
##            4.157176            5.276009            4.510219
##                Fiji             Finland              France
##            4.666299            5.057025            5.348621
##               Gabon              Gambia             Georgia
##            4.841436            4.754366            3.405711
##             Germany               Ghana              Greece
##            3.852291            4.161522            5.971160
##           Guatemala              Guinea       GuineaBissau
##            3.424443            2.961457            4.184983
##              Guyana               Haiti            Honduras
##            3.723276            4.631153            3.359413
##            HongKong             Hungary             Iceland
##            7.933718            4.125112            3.727688
##               India           Indonesia                Iran
##           11.076700            3.551209            5.249269
##             Ireland              Israel               Italy
##            5.825901            2.926763            4.490246
##             Jamaica               Japan              Jordan
##            3.707357            7.864113            3.992263
##          Kazakhstan               Kenya            Kiribati
##            3.444307            3.805183           12.374610
##          SouthKorea               Kosovo              Kuwait
##            3.855682            4.693929            7.351538
##       KyrgyzRepublic                Laos              Latvia
##            3.798709            3.953370            3.426015
##             Lebanon             Lesotho             Liberia
##            4.750076            5.043259            4.353327
##               Libya           Lithuania          Luxembourg
##           13.309362            2.748302            5.916895
##               Macau            Macedonia          Madagascar
```

```
##               9.019243                4.193179                2.991466
##                 Malawi                Malaysia                Maldives
##               3.657350                3.689923                5.964458
##                   Mali                   Malta               Mauritania
##               2.844413                3.717793                4.147866
##              Mauritius                  Mexico              Micronesia
##               3.332603                3.596345                8.378890
##                Moldova                Mongolia              Montenegro
##               3.796087                4.518339                4.257587
##                Morocco              Mozambique                 Namibia
##               4.037417                5.211233                5.194628
##                  Nepal             Netherlands             NewZealand
##               4.502952                4.824392                5.065368
##              Nicaragua                   Niger                 Nigeria
##               3.977784                3.028148                4.404567
##                 Norway                    Oman                Pakistan
##               4.445849                4.932565                2.624462
##                 Panama           PapuaNewGuinea                Paraguay
##               3.710640                4.637497                4.681525
##                   Peru             Philippines                  Poland
##               3.307648                3.025124                2.872799
##               Portugal                   Qatar                 Romania
##               3.747820                7.062858                3.388687
##                 Russia                  Rwanda              SaintLucia
##               4.386750                5.688177                3.994628
##  SaintVincentGrenadines                   Samoa         SaoTomePrincipe
##               4.728536                4.137050                3.939329
##            SaudiArabia                 Senegal                  Serbia
##               5.385517                3.482284                4.084599
##             Seychelles              SierraLeone               Singapore
##               5.839550                5.780271                6.329840
##               Slovakia                Slovenia           SolomonIslands
##               3.326132                4.271195                6.055311
##            SouthAfrica                   Spain                SriLanka
##               4.792991                3.592118                3.263436
##                  Sudan                Suriname                  Sweden
##               6.424885                5.305491                4.760496
##            Switzerland                  Taiwan              Tajikistan
##               4.343609                5.193734                3.714111
##               Tanzania                Thailand              TimorLeste
##               3.058031                3.851199                6.871135
##                   Togo                   Tonga           TrinidadTobago
##               4.532104                4.702610                4.511308
##                Tunisia                  Turkey            Turkmenistan
##               4.000642                3.810599                6.339664
##                 Uganda                 Ukraine      UnitedArabEmirates
##               4.535877                4.652444                5.523002
##          UnitedKingdom            UnitedStates                 Uruguay
##               3.711229               11.545725                4.866504
##             Uzbekistan                 Vanuatu               Venezuela
##               5.723677                7.089468               13.366989
##                Vietnam                   Yemen                  Zambia
##               3.397024                9.961033                3.576378
##                Zimbabwe
```

```
##               4.937158
##
## $rd
##             Afghanistan                 Albania                 Algeria
##                5.330469                3.644337                5.399042
##                  Angola               Argentina                 Armenia
##               10.748932                9.084330                4.920898
##               Australia                 Austria              Azerbaijan
##               28.292766                5.805317                4.734830
##                 Bahamas                 Bahrain              Bangladesh
##               10.967522               12.119688               16.768765
##                Barbados                 Belarus                 Belgium
##                5.056643                5.238579                8.522374
##                  Belize                   Benin                  Bhutan
##                3.364319                5.069048                5.067651
##                 Bolivia       BosniaHerzegovina                Botswana
##                5.327325                4.289061                3.676057
##                  Brazil                  Brunei                 Bulgaria
##               39.016474               11.686070                3.959985
##              BurkinaFaso                   Burma                 Burundi
##                4.366164                5.351695                4.837627
##                Cambodia                Cameroon                  Canada
##                4.325798                5.026221               13.113369
##               CaboVerde  CentralAfricanRepublic                    Chad
##                3.776420               11.123447               12.101360
##                   Chile                   China                Colombia
##                4.024193              135.544420                9.114742
##                 Comoros DemcoraticRepublicCongo          RepublicCongo
##                9.331145               14.374356                5.447914
##               CostaRica             CoeDivoire                 Croatia
##                4.846224                4.661379                3.872893
##                    Cuba                  Cyprus           CzechRepublic
##                8.770669                5.219614                4.082677
##                Denmark                Djibouti                Dominica
##               10.532986                8.851618                4.812449
##       DominicanRepublic                 Ecuador                   Egypt
##                4.363176                4.102362               10.171041
##              ElSalvador       EquatorialGuinea                 Eritrea
##                3.644283                8.869010                7.980424
##                Estonia                Eswatini                Ethiopia
##                4.198578                4.710771               10.012165
##                   Fiji                 Finland                  France
##                4.902395                9.009560               27.715130
##                  Gabon                  Gambia                 Georgia
##                5.387696                5.372305                3.676339
##                Germany                   Ghana                  Greece
##               25.664089                4.668923                7.410568
##               Guatemala                  Guinea            GuineaBissau
##                5.074069                3.962385                4.788816
##                  Guyana                   Haiti                Honduras
##                4.140214                4.818578                3.560545
##                HongKong                 Hungary                 Iceland
##               75.169062                4.297696                5.599460
##                   India               Indonesia                    Iran
```

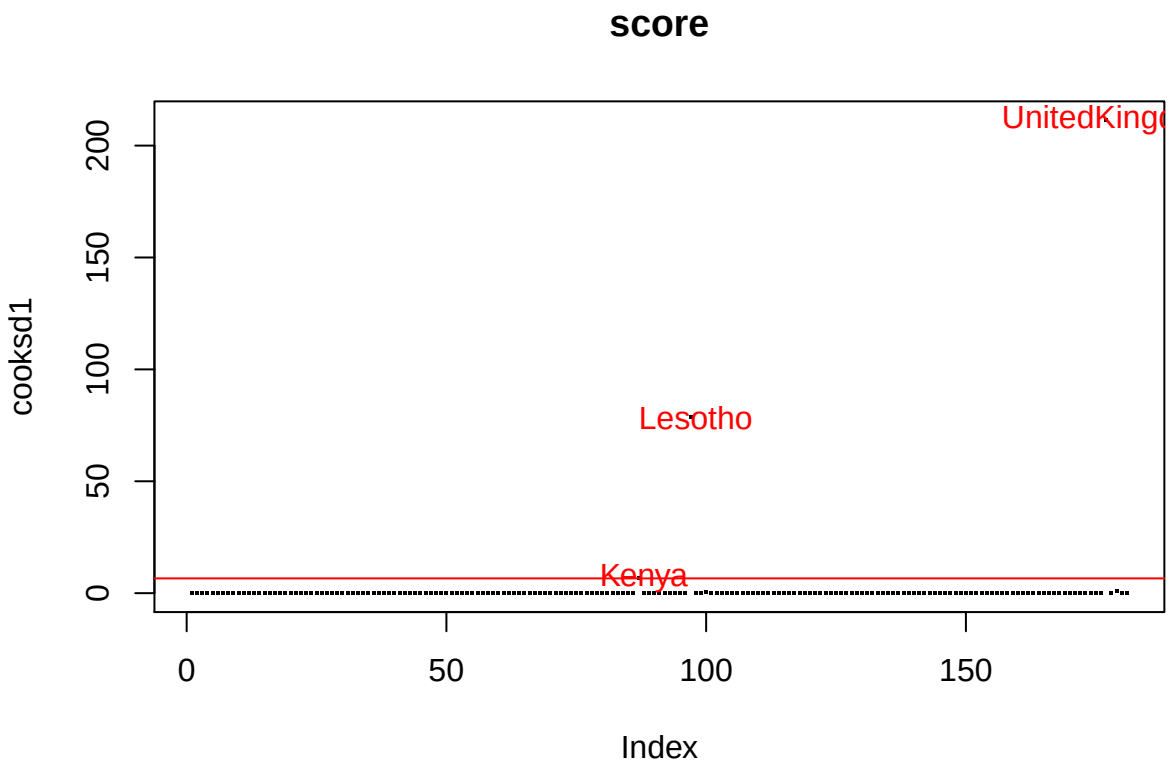
```
##              128.101370              21.800146              11.146908
##                 Ireland                 Israel                  Italy
##              20.071500              11.953661              13.390219
##                 Jamaica                  Japan                 Jordan
##               4.263722              39.964462               4.074817
##              Kazakhstan                  Kenya               Kiribati
##               4.508073               4.596220              50.661834
##              SouthKorea                 Kosovo                 Kuwait
##              11.866348               4.847181              15.702144
##           KyrgyzRepublic                   Laos                 Latvia
##               4.688346               4.311822               4.132863
##                 Lebanon                Lesotho                Liberia
##               5.259652               5.123021               4.735820
##                   Libya              Lithuania             Luxembourg
##             629.191895               3.831985              12.341631
##                   Macau              Macedonia             Madagascar
##              18.269264               4.565588               3.817251
##                  Malawi               Malaysia               Maldives
##               3.679765               5.396640               9.229490
##                    Mali                  Malta             Mauritania
##               2.938879               5.149879               4.085505
##                Mauritius                 Mexico             Micronesia
##               3.839809              16.445580              13.677917
##                 Moldova               Mongolia             Montenegro
##               4.473662               4.659594               4.884317
##                 Morocco             Mozambique                Namibia
##               4.416195               5.910726               4.730284
##                   Nepal            Netherlands            NewZealand
##               4.932786              37.754489               4.936455
##                Nicaragua                  Niger                Nigeria
##               3.699953               3.057354              18.012197
##                  Norway                   Oman                Pakistan
##              13.533930              10.189867              19.423211
##                  Panama          PapuaNewGuinea                Paraguay
##               4.412262               5.074723               5.252238
##                    Peru             Philippines                  Poland
##               4.191690               9.225057               6.110937
##                Portugal                  Qatar                 Romania
##               4.586602              19.954677               3.996869
##                  Russia                 Rwanda             SaintLucia
##              22.201770               4.982457               3.573710
##  SaintVincentGrenadines                  Samoa         SaoTomePrincipe
##               4.482443               4.290443               4.333549
##              SaudiArabia                Senegal                 Serbia
##              18.107451               3.473503               4.239349
##              Seychelles             SierraLeone              Singapore
##               5.587866               5.869889              41.371924
##                Slovakia               Slovenia          SolomonIslands
##               3.558869               5.084269               7.547027
##              SouthAfrica                  Spain               SriLanka
##               6.353619              10.493137               3.923291
##                   Sudan               Suriname                 Sweden
##              13.892528               8.122980              10.242812
##              Switzerland                 Taiwan             Tajikistan
```

```
##               27.092796               10.860313                4.263488
##                Tanzania                Thailand              TimorLeste
##                4.870057                7.161563                12.715165
##                    Togo                   Tonga           TrinidadTobago
##                4.532129                5.308188                5.305223
##                 Tunisia                  Turkey              Turkmenistan
##                3.470757                12.805518                8.902852
##                  Uganda                 Ukraine        UnitedArabEmirates
##                4.852945                7.117494                12.255042
##           UnitedKingdom           UnitedStates                 Uruguay
##               17.512613              168.632076                5.059063
##              Uzbekistan                 Vanuatu               Venezuela
##                7.719878                10.656406              510.321828
##                 Vietnam                   Yemen                  Zambia
##               11.917378               93.702408                3.633806
##                Zimbabwe
##                5.555268
##
## $cutoff
## [1] 6.572253
```




We have the plots of both: Classican and Robust Mahalanodis distance. From the result obtained, we have a cutoff vale of 6.57, which means the number of the observations we have to remove is very big. Considering that this is very ricky situation we performed another method called Cooks distance. This method was not part of our schedule but yet we decided to include it. Cook's method implements a fitting model and then measures the influence of every observation on the fitted response values.

```r
cooksdata <- knn.imput[1:27]
mod1 <- lm(Score ~ ., data=cooksdata)
cooksd1 <- cooks.distance(mod1)
plot(cooksd1, pch=".", cex=2, main="score")  # plot cook's distance
abline(h = 4*mean(cooksd1, na.rm=T), col="red")  # add cutoff line
text(x=1:length(cooksd1)+1, y=cooksd1, labels=ifelse(cooksd1>4*mean(cooksd1, na.rm=T),row.names(data),"
```




Now the outliers are more clear. If we compare both methods, we see that they have in common United Kingdom and Lesotho, so these are the options we have to analyze. Lesotho is a country in South Africa, so it might be reasonable to consider it as oulier but regarding United Kingdom we have our own doubts. Our dataset is about economical indexes, so instead of considering UK as outlier, we might be facing a country with very high economical standards.