HW week 11

w203: Statistics for Data Science

Tako Hisada

11/25/2017

1. Get familiar with the data

You receive a data set from World Bank Development Indicators.

Load the data using `load` and see what is loaded by using `ls()`. You should see `Data` which is the data frame including data, and `Descriptions` which is a data frame that includes variable names.

```
load('Week11.Rdata')
ls(Data)
```

```
##  [1] "AG.LND.FRST.ZS"    "Country.Code"      "Country.Name"
##  [4] "MS.MIL.MPRT.KD"    "MS.MIL.XPND.GD.ZS" "MS.MIL.XPND.ZS"
##  [7] "MS.MIL.XPRT.KD"    "NE.EXP.GNFS.CD"    "NE.IMP.GNFS.CD"
## [10] "NY.GDP.MKTP.CD"    "NY.GDP.PCAP.CD"    "NY.GDP.PETR.RT.ZS"
## [13] "TX.VAL.AGRI.ZS.UN"
```

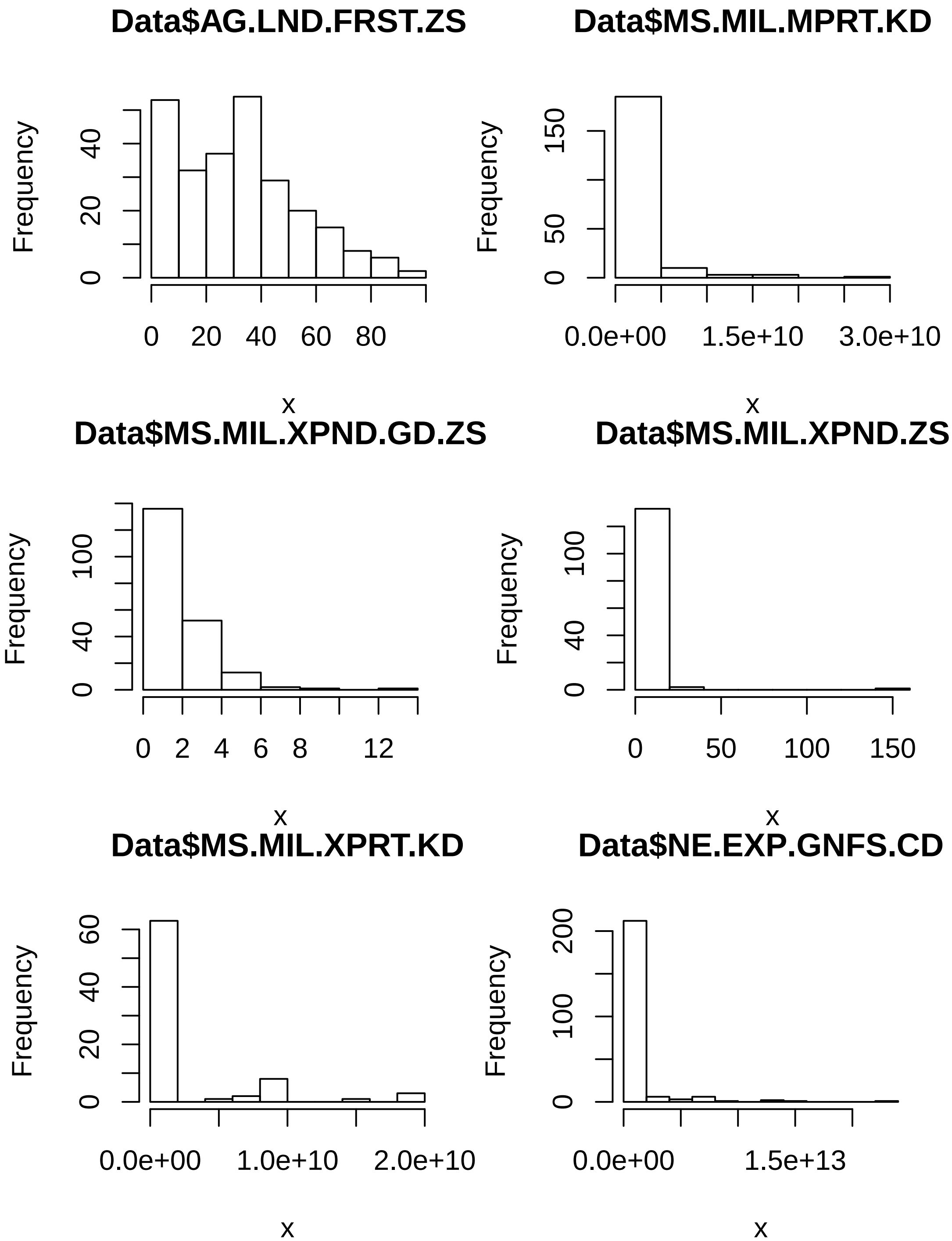
```
ls(Definitions)
```

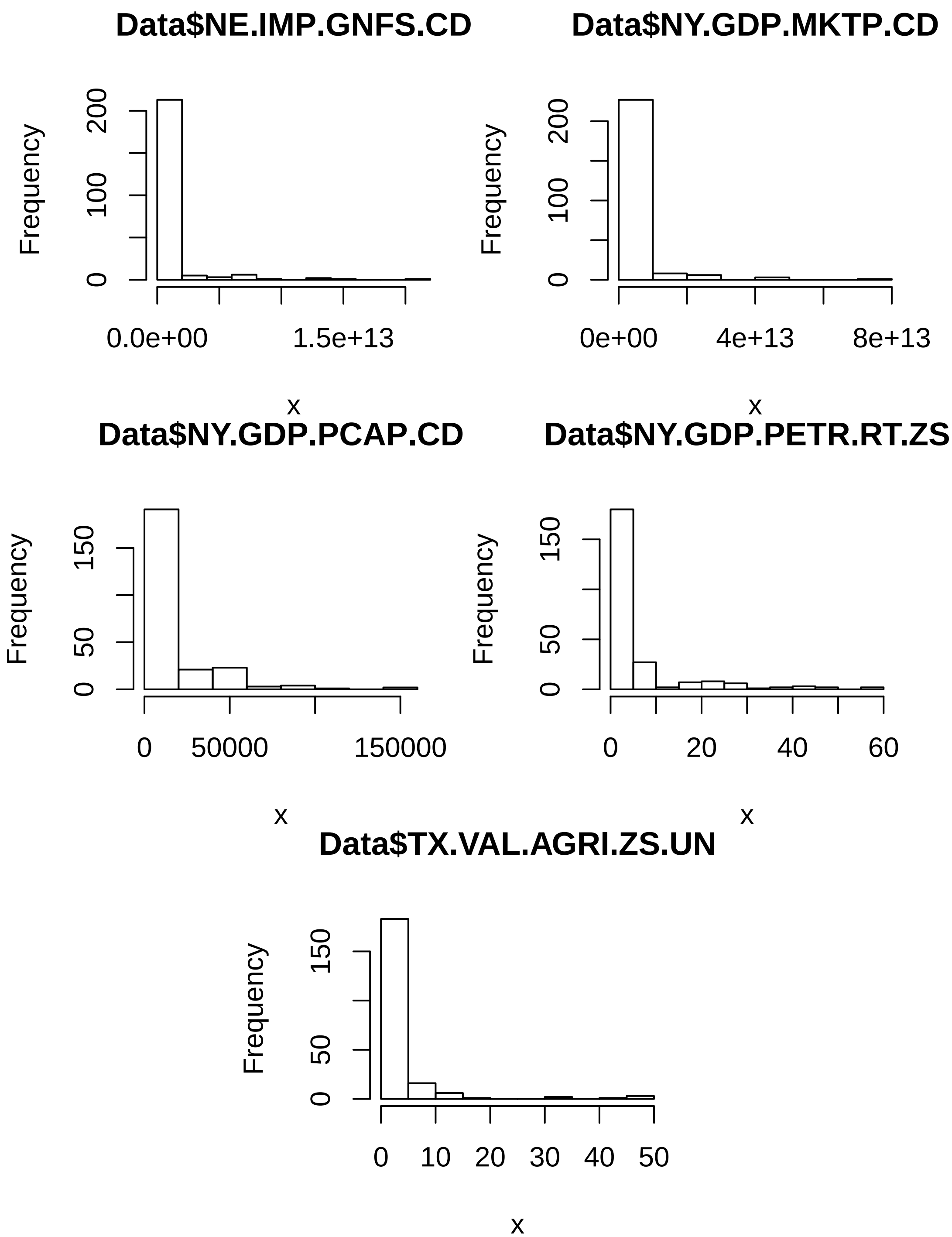
```
## [1] "Series.Code" "Series.Name"
```

Look at the variables, read their descriptions, and take a look at their histograms. Think about the transformations that you may need to use for these variables in the section below.

```
displayHist <- function(x, apply_log=FALSE) {
  title <- deparse(substitute(x))
  if(apply_log) {
    x <- log(x)
  }
  hist(x, main=title)
}
displayHist(Data$AG.LND.FRST.ZS)
displayHist(Data$MS.MIL.MPRT.KD)
displayHist(Data$MS.MIL.XPND.GD.ZS)
displayHist(Data$MS.MIL.XPND.ZS)
displayHist(Data$MS.MIL.XPRT.KD)
displayHist(Data$NE.EXP.GNFS.CD)
displayHist(Data$NE.IMP.GNFS.CD)
displayHist(Data$NY.GDP.MKTP.CD)
displayHist(Data$NY.GDP.PCAP.CD)
```

```r
displayHist(Data$NY.GDP.PETR.RT.ZS)
displayHist(Data$TX.VAL.AGRI.ZS.UN)
```



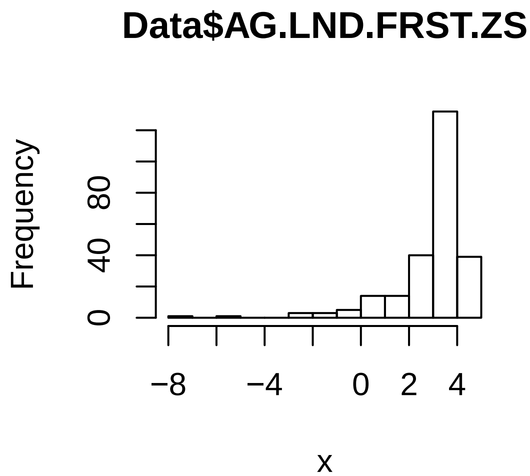


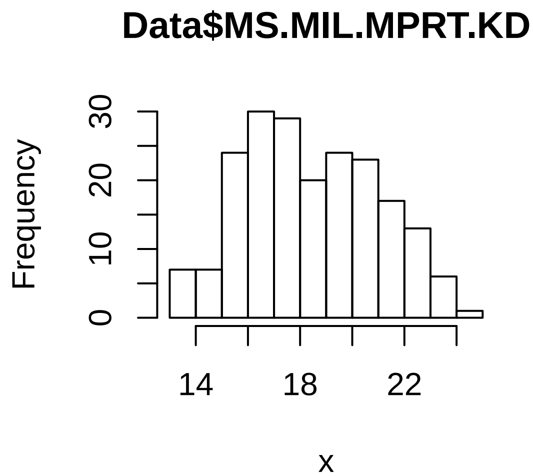
Also most of these variables contain very large numbers as they represent national-level data such as GDP for a given country and it might be helpful to reduce them to in terms of USD billions, etc.

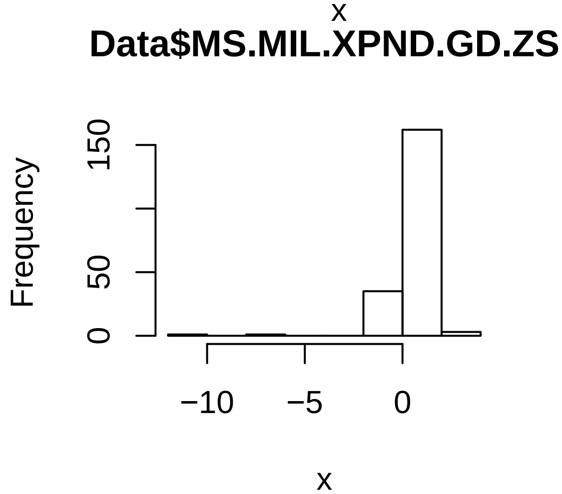
Furthre more, aside from Data$AG.LND.FRST.ZS, all the histograms look positively skewed. It may make sense to apply log() depending on the analysis we need to conduct.

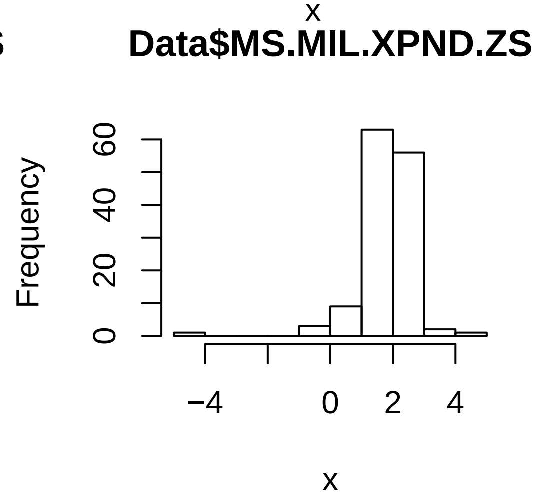
```
displayHist(Data$AG.LND.FRST.ZS, T)
displayHist(Data$MS.MIL.MPRT.KD, T)
displayHist(Data$MS.MIL.XPND.GD.ZS, T)
displayHist(Data$MS.MIL.XPND.ZS, T)
```

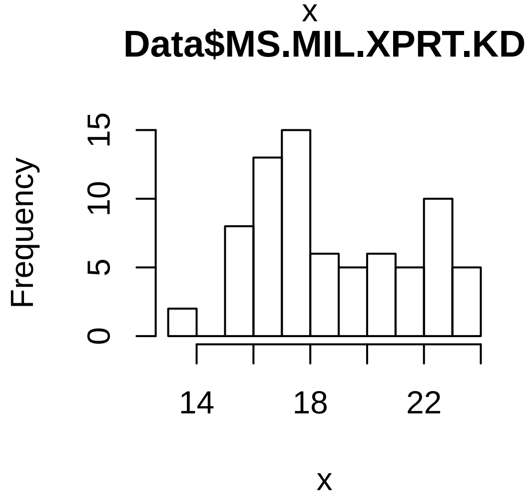
```r
displayHist(Data$MS.MIL.XPRT.KD, T)
displayHist(Data$NE.EXP.GNFS.CD, T)
displayHist(Data$NE.IMP.GNFS.CD, T)
displayHist(Data$NY.GDP.MKTP.CD, T)
displayHist(Data$NY.GDP.PCAP.CD, T)
displayHist(Data$NY.GDP.PETR.RT.ZS, T)
displayHist(Data$TX.VAL.AGRI.ZS.UN, T)
```

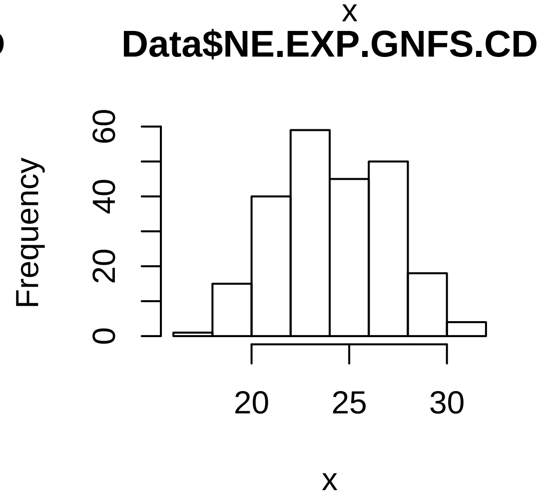


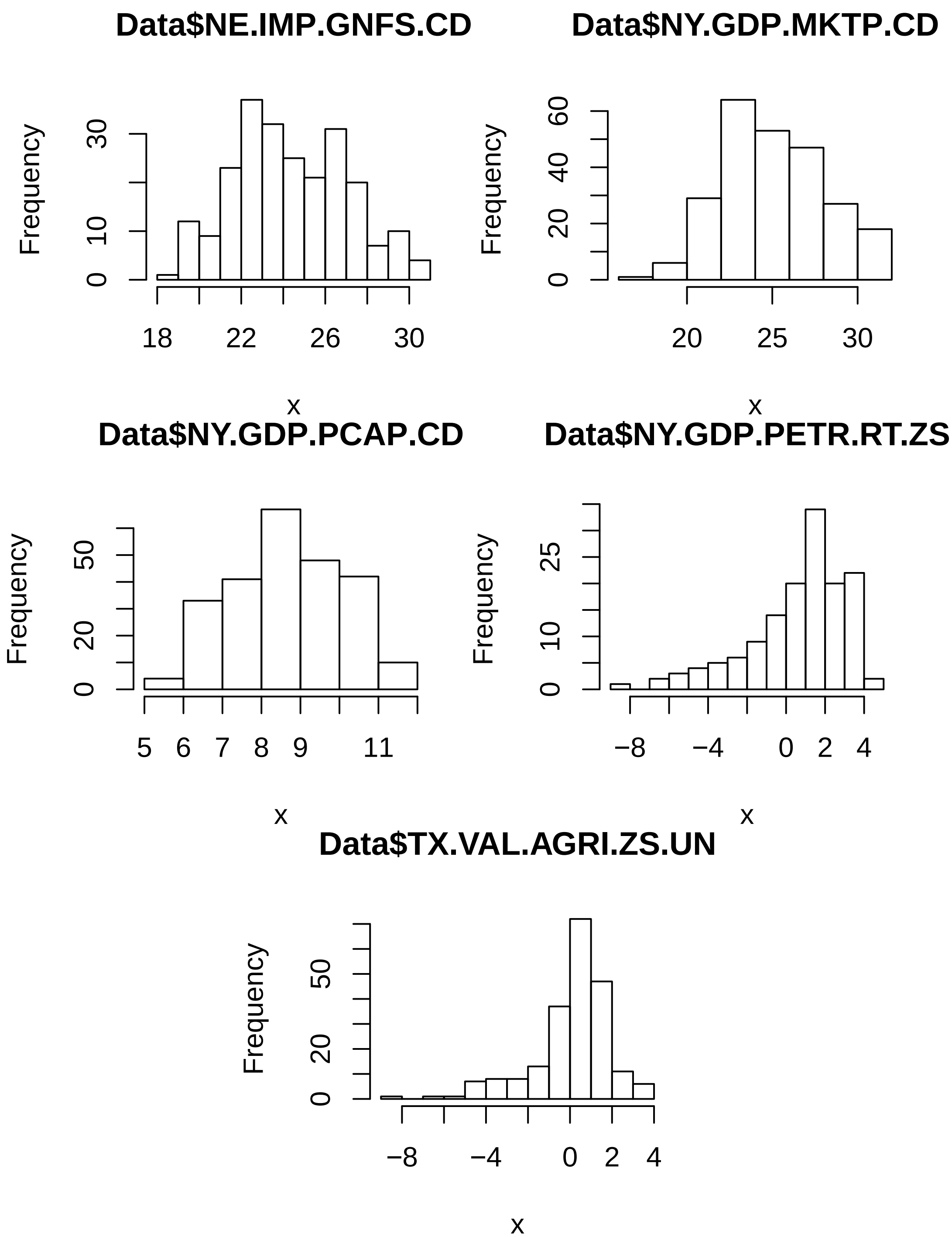












The hisograms appear more normally distributed with log() applied.

Run: `apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )` and explain what it is showing.

```r
apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )
```

```
##     AG.LND.FRST.ZS     MS.MIL.MPRT.KD MS.MIL.XPND.GD.ZS     MS.MIL.XPND.ZS
```

```
##         0.9696970         0.7651515         0.7765152         0.5151515
##    MS.MIL.XPRT.KD    NE.EXP.GNFS.CD    NE.IMP.GNFS.CD    NY.GDP.MKTP.CD
##         0.2954545         0.8787879         0.8787879         0.9280303
##    NY.GDP.PCAP.CD NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
##         0.9280303         0.9090909         0.8030303
```

This calculates the inversed mean of all columns in the dataframe Data except for the first and second columns.

Data[, -(1:2)] specifies all rows from all columns in the datafrmae except for the first and second columns.

is.na() determines if the R object passed in is NA or not and returns TRUE if NA, FALSE if not. "!"" negates that and hence reverses the values.

apply() applies mean() to to the output of !is.na(Data[,-(1:2)]) by column. Because the output of !is.na(Data[,-(1:2)]) is Boolean, the averages computed will be numbers between 0 and 1.

Can you include both NE.IMP.GNFS.CD and NE.EXP.GNFS.CD in the same OLS model? Why?

```r
summary(Data$NE.EXP.GNFS.CD)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 1.817e+07 3.855e+09 2.823e+10 7.813e+11 2.894e+11 2.210e+13        32
```

```r
summary(Data$NE.IMP.GNFS.CD)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 1.646e+08 5.594e+09 2.904e+10 7.589e+11 2.892e+11 2.149e+13        32
```

```r
Data$NE.EXP.GNFS.CD/Data$NE.IMP.GNFS.CD
```

```
##   [1] 0.15293360 0.62132237 1.17408722        NaN        NaN 1.35854723
##   [7] 0.78780689 1.34476093 1.06258703 0.57560734 0.82025580 0.97504507
##  [13] 1.06504788 1.86822512 0.78129566 1.55863052 0.72490103 0.79588914
##  [19] 0.97004045 1.01373568 0.95453484 0.70212853 1.61236844 0.61670941
##  [25] 1.15219941 0.58928792 0.95619547 0.87898071        NaN 2.41936308
##  [31] 0.98824763 0.76465967 0.22940000 0.51112559 0.91903134 0.70147045
##  [37] 0.94925487 0.94498164        NaN 0.39225561 1.02467480 0.85965928
##  [43]        NaN 1.04183283 1.13383310 0.82441043 0.28957092 0.85731524
##  [49] 1.21020249 0.91038149 1.14219162 1.01732644 1.21275419        NaN
##  [55] 0.97561114 1.07164104 1.12621449        NaN 0.69817616 0.77023173
##  [61] 0.96736613 1.07009043 1.10639361 1.10706923 0.91963570 0.70642037
##  [67] 0.59547160 1.32189623 0.43376754 1.04638219 0.42158315 1.06870061
##  [73] 1.06923374 1.08970412 1.07368855 1.05146228 0.86146512 0.88717998
##  [79] 0.98838075 0.91348745 0.93429978        NaN 1.80661213 0.65167102
##  [85] 0.69703002 1.15619112 0.75199053        NaN 0.85809991        NaN
##  [91] 0.60115167        NaN 0.69810540 0.55240983 0.69954884 0.64472254
##  [97] 0.32513093 0.71808710 1.02858019 0.71382285 1.01003899 1.08180321
## [103] 1.04393725 1.15838079 1.02790528 1.01631636 0.69878815 0.87440634
## [109] 0.83624761 1.01675962 1.30013232 1.47932607 1.20202075        NaN
## [115] 1.05443535 1.03241562 0.59623470 0.92210901 0.63234290 1.48054650
## [121] 0.54847520 0.11042064        NaN 1.08887912 0.36252815 2.53768946
## [127] 0.51423473 0.82408551 1.13745996 0.95671932 0.92969547 0.93789407
## [133] 0.95133155 0.80718185 0.74290770 0.42401990 0.33600488 1.04204777
## [139]        NaN 1.00011307 1.02757088 0.57033069 0.89203428 1.20068297
```

```
## [145] 2.64970483 0.71233482 0.72204913 0.75648077 1.16482643 1.15269991
## [151] 0.75385551 1.04720902        NaN 0.71970345 0.81367058 0.96224832
## [157]        NaN 1.32379670 1.01916689 1.02888328 1.03736448 0.53231134
## [163]        NaN 0.77668392 0.65660943 0.71601900 0.44235967        NaN
## [169] 0.72150907        NaN 0.28068189 1.14305946        NaN 1.04732314
## [175] 0.68029524 0.49969743 1.61440000 0.82705192        NaN 1.36574887
## [181]        NaN 0.98891314 1.55753160 1.45733985 0.74931280 0.66371450
## [187] 0.73576190 0.87777461        NaN 1.06303524 1.02180164 0.89034214
## [193] 1.00564768 0.95239721 0.99571548 1.09883888 1.38160448 2.33875735
## [199] 0.92851305 1.34251916 0.45257723 0.55842120        NaN        NaN
## [205] 1.50976469 0.59175405 0.74981511 0.86858476 0.53976443 1.14782526
## [211]        NaN 1.02270355 1.07378093 1.37002110 0.82850202 0.23519847
## [217] 0.97838534 0.80017697 0.80017697 0.97435123 1.04423353 0.68635423
## [223] 0.71049627 0.81176061        NaN 0.46323277 0.99264243 0.99298914
## [229] 0.99264243 0.74511619 1.02104842 0.78136918 1.11190493 1.19681085
## [235]        NaN 0.28779522 0.63387097 1.05188222 0.07432355 0.72390445
## [241] 0.33566350 1.49513843 0.85298598 0.82164173 1.68919162        NaN
## [247]        NaN 0.61371174 0.88779853 1.26026688 0.93494566 0.80139233
## [253] 1.08606133 0.94387512 0.92813400 0.92891124 1.19544873 1.00209720
## [259]        NaN 0.29907332 1.02829503        NaN 1.19922107 0.52918180
```

Yes, `NE.IMP.GNFS.CD` and `NE.EXP.GNFS.CD` should be included as they are not exactly linearly related although their histograms are very similar and hence does not violate MLR Assumption 3 of no perfect collinearity.

Rename the variable named `AG.LND.FRST.ZS` to `forest.` This is going to be our dependent variable.

```
Data$forest. <- Data$AG.LND.FRST.ZS
```

Defined our dependent variable forest.

2. Decribe a model for that predicts `forest`

Write a model with two explanatory variables.

Create a residuals versus fitted values plot and assess whether your coefficients are unbiased.

I am picking NY.GDP.MKTP.CD and NE.EXP.GNFS.CD as my independent variables.

```
summary(Data$forest.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.00   12.47   31.11   31.53   46.00   98.34       8
```

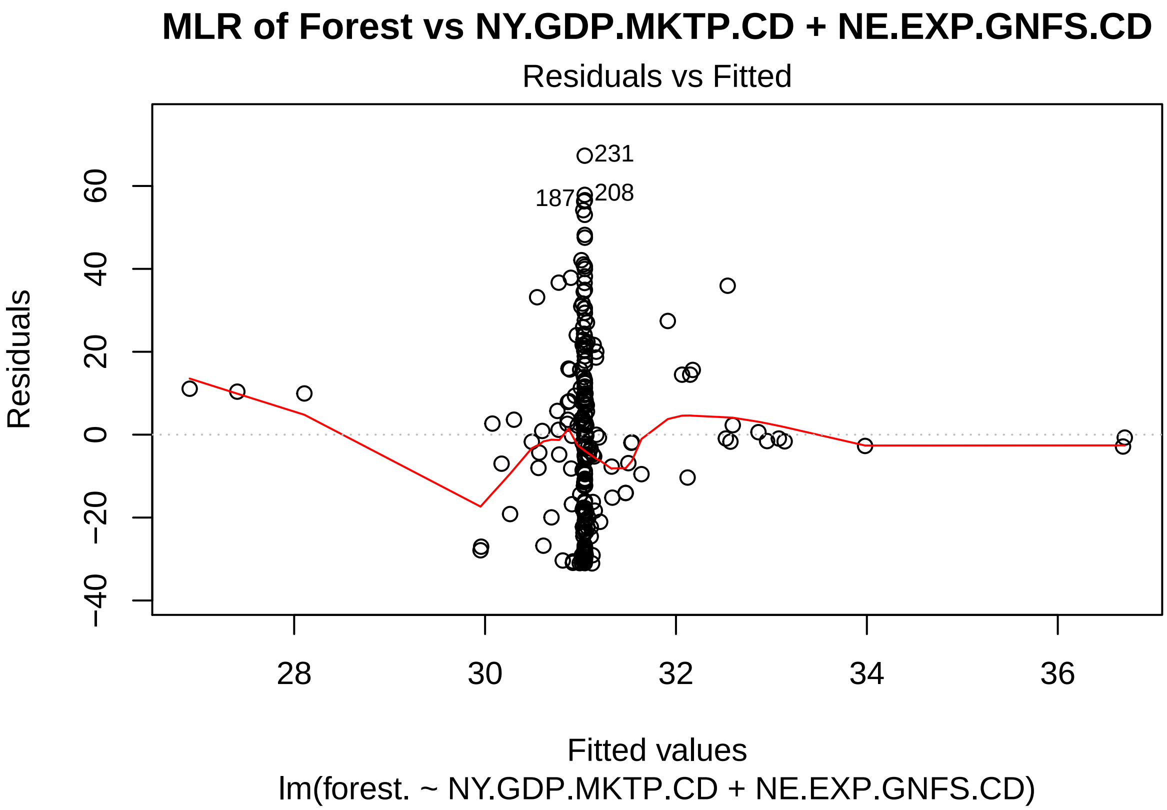
```
summary(Data$NY.GDP.MKTP.CD)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## 3.744e+07 8.998e+09 5.262e+10 2.469e+12 5.396e+11 7.346e+13        19
```

```
summary(Data$NE.EXP.GNFS.CD)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## 1.817e+07 3.855e+09 2.823e+10 7.813e+11 2.894e+11 2.210e+13        32
```

```r
model = lm(forest. ~ NY.GDP.MKTP.CD + NE.EXP.GNFS.CD, data = Data)
plot(model, which = 1, main = "MLR of Forest vs NY.GDP.MKTP.CD + NE.EXP.GNFS.CD")
```



Looking at the plot and the model, we know that the model has a linear relationship (MLR.1). We do not know much about how World Bank collects data and there may be some countries that are underrepresented which makes our assumption of random sampling a little bit questionable (MLR.2). Looking at the coefficients, we can say that the independent variables do not have perfect collinearity (MLR.3). The residuals a little bit high on the left-hand side of the plot due to the outliers however for the most part are staying pretty closely to 0 (MLR. 4). We do not have any reason from the data given that we do not have homoskedasticity in this model. (MLR. 5). Therefore we can conclude that our coefficients are relatively unbiased although we may want to look further into how the data has been collected.

How many observations are being used in your analysis?

```r
n_data = length(Data$forest.)
n_observations = length(model$fitted.values)
```

There are 228 observations that are used in the model while there were 264 data points available in the data frame provided.

Are the countries that are dropping out dropping out by random chance? If not, what would this do to our inference?

```r
Data$Country.Name[is.na(Data$forest.)]
```

```
## [1] Curacao                   Hong Kong SAR, China
```

```
## [3] Kosovo                    Macao SAR, China
## [5] Monaco                    Not classified
## [7] Sint Maarten (Dutch part) South Sudan
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
Data$Country.Name[is.na(Data$NY.GDP.MKTP.CD)]
```

```
##  [1] American Samoa            British Virgin Islands
##  [3] Cayman Islands            Channel Islands
##  [5] Curacao                   French Polynesia
##  [7] Gibraltar                 Guam
##  [9] Korea, Dem. People's Rep. Nauru
## [11] New Caledonia             Northern Mariana Islands
## [13] Not classified            San Marino
## [15] Sint Maarten (Dutch part) St. Martin (French part)
## [17] Syrian Arab Republic      Turks and Caicos Islands
## [19] Virgin Islands (U.S.)
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

```
Data$Country.Name[is.na(Data$NE.EXP.GNFS.CD)]
```

```
##  [1] American Samoa            Andorra
##  [3] British Virgin Islands    Cayman Islands
##  [5] Channel Islands           Curacao
##  [7] Djibouti                  French Polynesia
##  [9] Gibraltar                 Greenland
## [11] Guam                      Isle of Man
## [13] Korea, Dem. People's Rep. Liechtenstein
## [15] Marshall Islands          Micronesia, Fed. Sts.
## [17] Monaco                    Myanmar
## [19] Nauru                     New Caledonia
## [21] Northern Mariana Islands  Not classified
## [23] Papua New Guinea          San Marino
## [25] Sao Tome and Principe     Sint Maarten (Dutch part)
## [27] St. Martin (French part)  Syrian Arab Republic
## [29] Turks and Caicos Islands  Tuvalu
## [31] Virgin Islands (U.S.)     Yemen, Rep.
## 267 Levels:  Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

We are omitting countries with NA values for the variables we are using. For the variable Data$forest, we see that the countries with limited land such as HK, Macao and Monaco tend to have NA. For Data$NY.GDP.MKTP.CD and Data$NE.EXP.GNFS.CD, we are seeing a very similar set of countries are being omitted such as small nations such as Nauru and San Marino, non-sovereign territories such as Guam, New Caledonia as well as countries that are currently in active conflicts such as Syrian Arab Republic. This may introduce biases into our analysis by undermining our MLR.2 assumption of random sampling.

Now add a third variable.

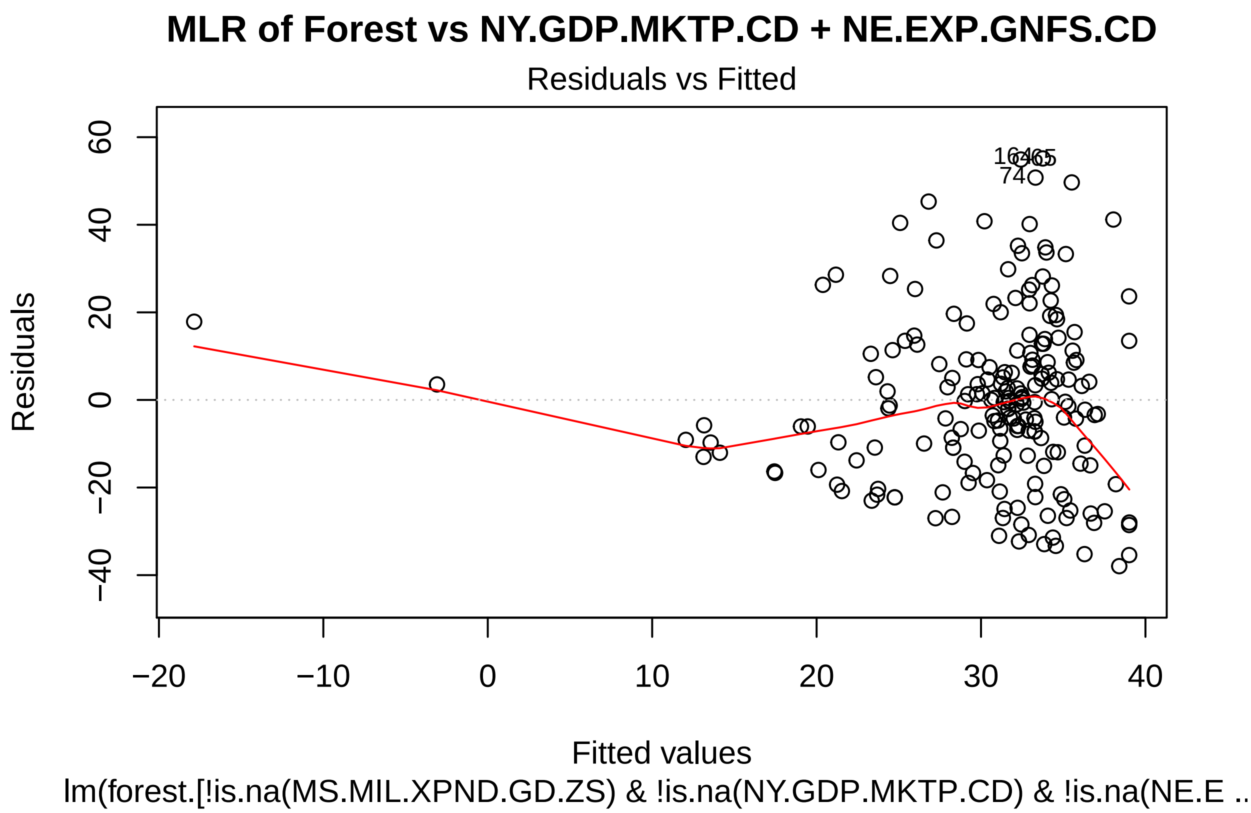
I will add Data$MS.MIL.XPND.GD.ZS as my third variable.

```
summary(Data$MS.MIL.XPND.GD.ZS)
```

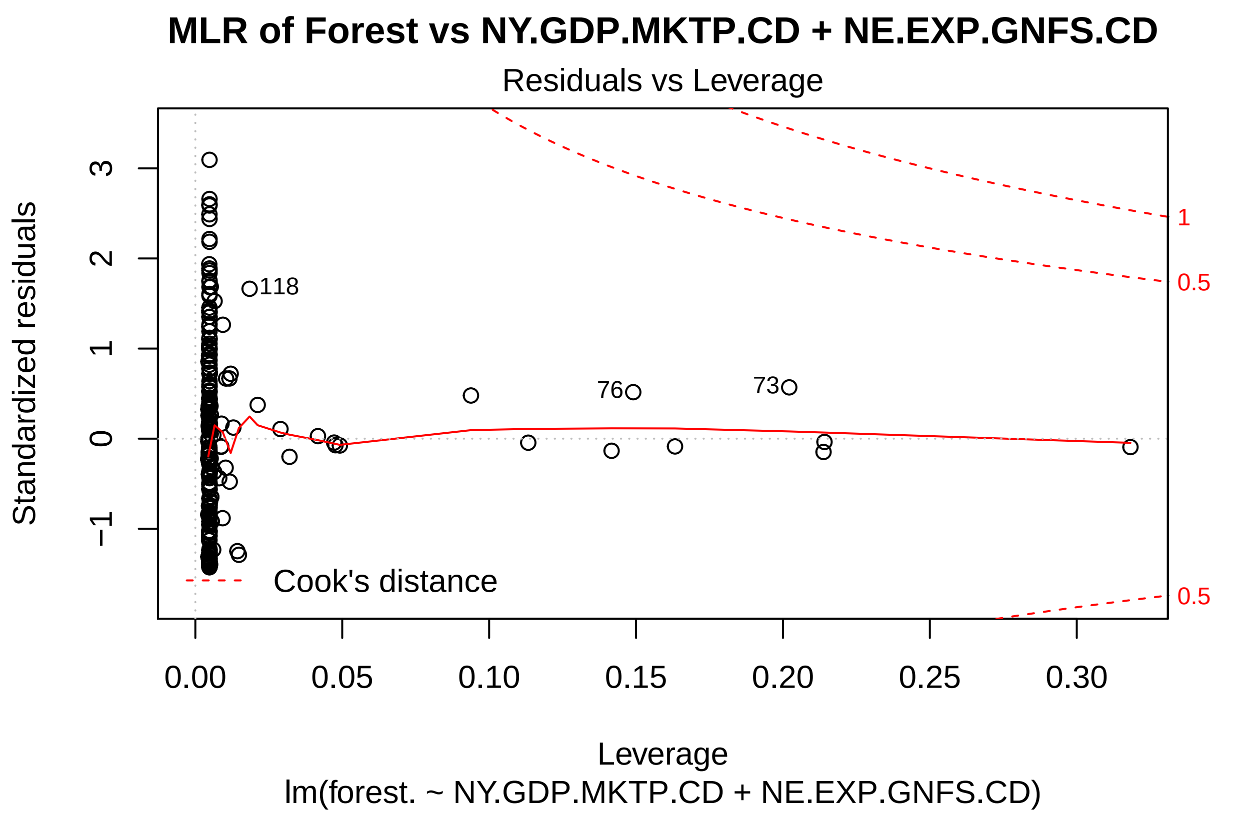
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.115   1.535   1.997   2.426  12.787      59
```

Show how you would use the regression anatomy formula to compute the coefficient on your third variable. First, regress the third variable on your first two variables and extract the residuals. Next, regress forest on the residuals from the first stage.

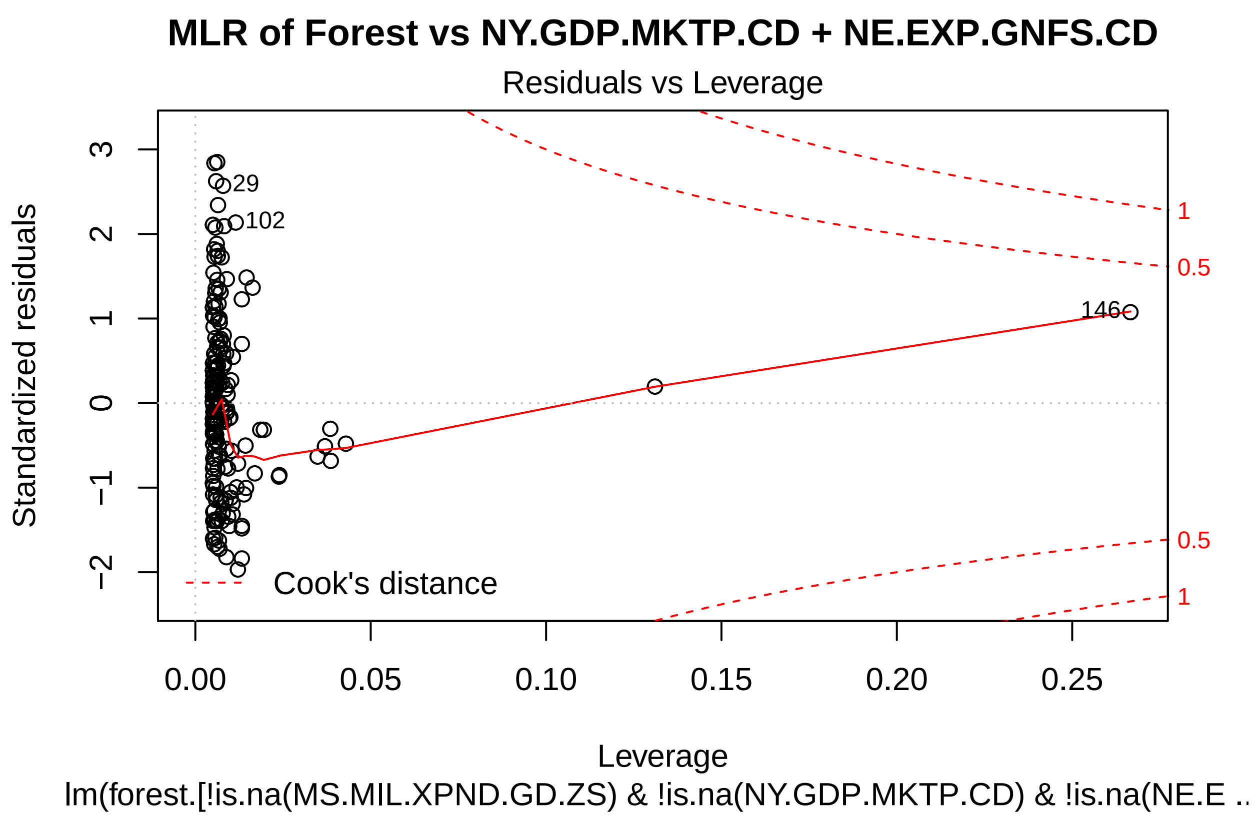
```r
# Regress the 3rd variable on the first 2
model2 <- lm(MS.MIL.XPND.GD.ZS ~ NY.GDP.MKTP.CD + NE.EXP.GNFS.CD, data = Data)
x_3 <- model2$residuals

# Now regrest forest on the residuals
# We have to omit NAs
model3 <- lm(
  forest.[!is.na(MS.MIL.XPND.GD.ZS) & !is.na(NY.GDP.MKTP.CD) & !is.na(NE.EXP.GNFS.CD)] ~ x_3,
  data = Data)
plot(model3, which = 1, main = "MLR of Forest vs NY.GDP.MKTP.CD + NE.EXP.GNFS.CD")
```




```r
plot(model, which = 5, main = "MLR of Forest vs NY.GDP.MKTP.CD + NE.EXP.GNFS.CD")
```




```r
plot(model3, which = 5, main = "MLR of Forest vs NY.GDP.MKTP.CD + NE.EXP.GNFS.CD")
```



```r
rsquare_model <- summary(model)$r.square
rsquare_model3 <- summary(model3)$r.square
AIC(model)
```

```
## [1] 2057.373
```

```r
AIC(model3)
```

```
## [1] 1757.721
```

Compare your two models. Do you see an improvement? Explain how you can tell.

Compared to the first model, we are seeing extrme outliers with large leverages on the left-hand side of the plot and the overall plots are more widely scattered on the x-axis and we are observing relatively large residuals on the right hand-side of the plot as well. Looking at Cook's distance, although neither one reaches Cook's distance value of 0.5, the second model actually approaches nearer to the dotted lines, indicating the presence of highly influential outliers.

However, our R-square value has shown a significant increase from 0.001 to 0.107. This is reasonable and expected as we have introduced an additional independent variable. AIC is consistent with the R-squared result and is showing that our second model is a better model fit.

3. Make up a country

Make up a country named `Mediland` which has every indicator set at the median value observed in the data.

```r
# Modify factors before inserting new row
levels(Data$Country.Name) <- c(levels(Data$Country.Name), 'Mediland')
levels(Data$Country.Code) <- c(levels(Data$Country.Code), 'MED')

# Insert new rot for Mediland into dataframe Data
medians <- apply(Data[,-(1:2)], MARGIN= 2, median, na.rm = TRUE )
medians <- c('Mediland', 'MED', medians)
n = length(Data$Country.Code)
Data <- rbind(Data[1:n,],medians,Data[-(1:n),])
```

How much forest would this country have?

```r
median_forest = Data$forest.[Data$Country.Name == 'Mediland']
```

Mediland would have 31.105% forest area.

4. Take away

What is the causal story, if any, that you can take away from the above analysis? Explain why.

The R-squared value of our second model was 0.107 indicating the proportion that the explanatory variables Data$NY.GDP.MKTP.CD, Data$NE.EXP.GNFS.CD and Data$MS.MIL.XPND.GD.ZS together were responsible for in the variation of Data$forest. Interstingly, adding Data$MS.MIL.XPND.GD.ZS increased the number considerably, suggesting the amount of military expenditure having a high impact on the forest area. Overall, however, there are numerous factors that go into determining how much forest area a given country has and this model only explains a relatively small portion of it. Also, as we have seen in Q2, there may be some countries that tend to be omitted and hence this may not constitute random sampling which may make our OLS biased.