HW week 11

w203: Statistics for Data Science

Shan He

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.

```
library(car)
load("Week11.Rdata")
ls()
```

```
## [1] "Data"        "Definitions"
```

- 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.

```
Definitions
```

```
##          Series.Code
## 1     AG.LND.FRST.ZS
## 2  MS.MIL.XPND.GD.ZS
## 3     MS.MIL.XPND.ZS
## 4     NY.GDP.MKTP.CD
## 5     NY.GDP.PCAP.CD
## 6  NY.GDP.PETR.RT.ZS
## 7     MS.MIL.XPRT.KD
## 8  TX.VAL.AGRI.ZS.UN
## 9     MS.MIL.MPRT.KD
## 10    NE.IMP.GNFS.CD
## 11    NE.EXP.GNFS.CD
##                                                          Series.Name
## 1                                  Forest area (% of land area)
## 2                                Military expenditure (% of GDP)
## 3     Military expenditure (% of central government expenditure)
## 4                                             GDP (current US$)
## 5                                  GDP per capita (current US$)
## 6                                         Oil rents (% of GDP)
## 7                    Arms exports (SIPRI trend indicator values)
## 8  Agricultural raw materials exports (% of merchandise exports)
## 9                    Arms imports (SIPRI trend indicator values)
## 10                   Imports of goods and services (current US$)
## 11                   Exports of goods and services (current US$)
```

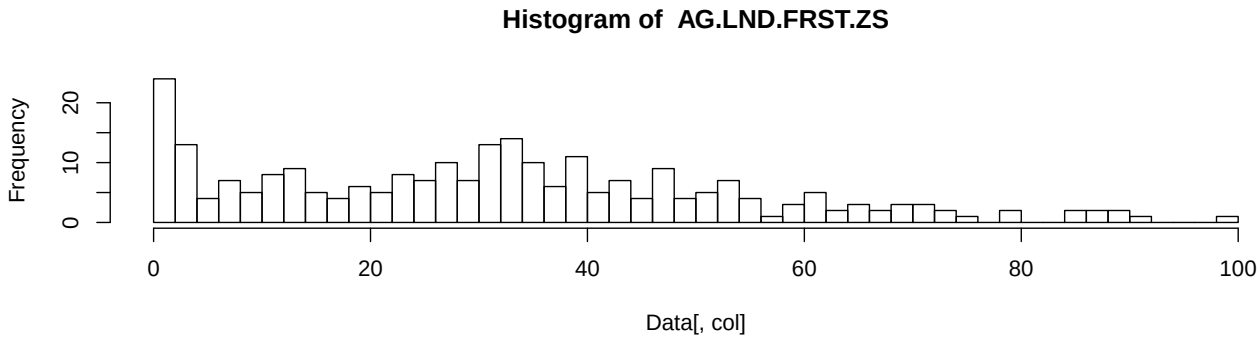
```
columns <- names(Data)
classes <- sapply(Data,class)
columns[classes == 'numeric']
```

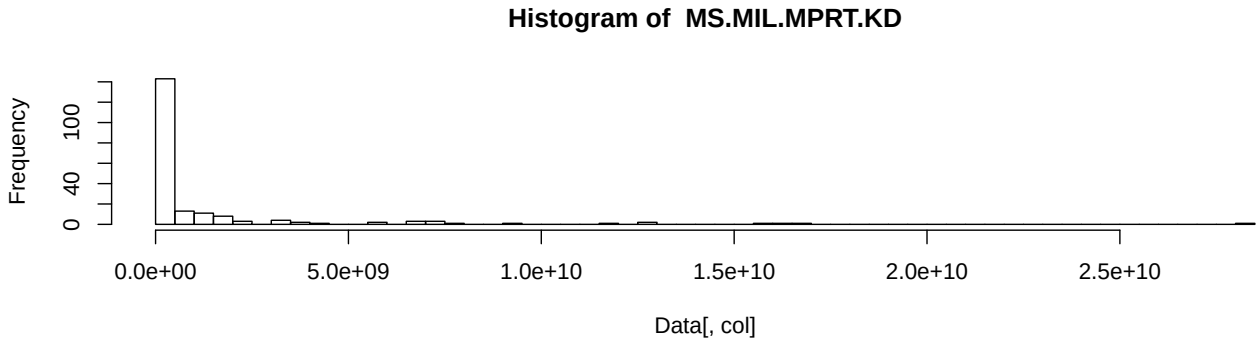
```
##  [1] "AG.LND.FRST.ZS"    "MS.MIL.MPRT.KD"    "MS.MIL.XPND.GD.ZS"
##  [4] "MS.MIL.XPND.ZS"    "MS.MIL.XPRT.KD"    "NE.EXP.GNFS.CD"
##  [7] "NE.IMP.GNFS.CD"    "NY.GDP.MKTP.CD"    "NY.GDP.PCAP.CD"
## [10] "NY.GDP.PETR.RT.ZS" "TX.VAL.AGRI.ZS.UN"
```

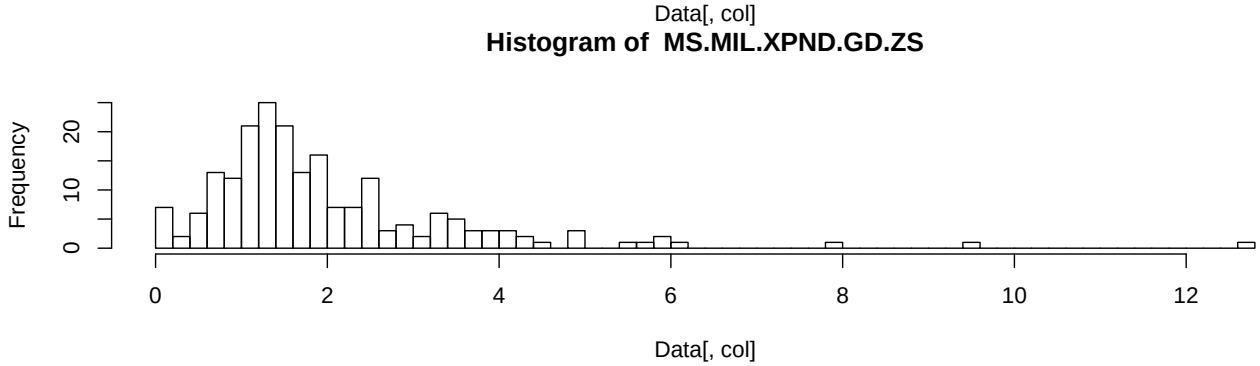
```r
summary(Data)
```

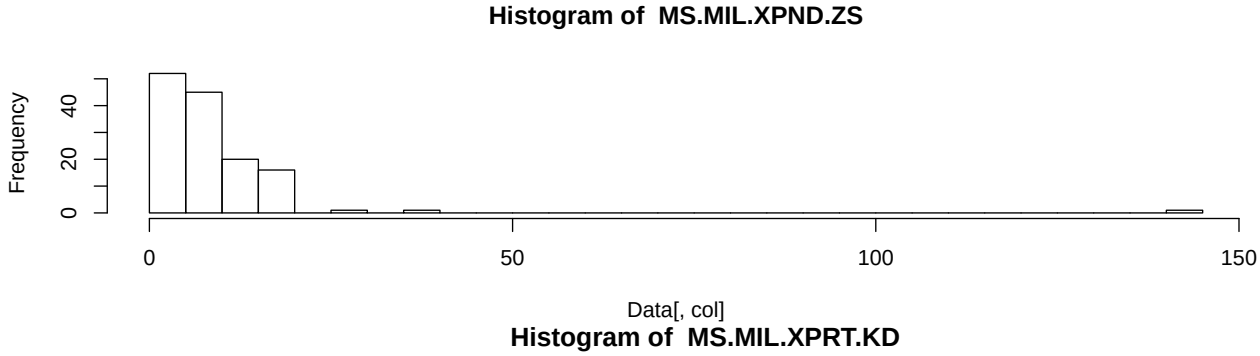
```
##          Country.Name  Country.Code AG.LND.FRST.ZS  MS.MIL.MPRT.KD
##  Afghanistan    :  1   ABW    :  1   Min.   : 0.00   Min.   :0.000e+00
##  Albania        :  1   ADO    :  1   1st Qu.:12.47   1st Qu.:1.081e+07
##  Algeria        :  1   AFG    :  1   Median :31.11   Median :7.458e+07
##  American Samoa:  1   AGO    :  1   Mean   :31.53   Mean   :1.299e+09
##  Andorra        :  1   ALB    :  1   3rd Qu.:46.00   3rd Qu.:7.234e+08
##  Angola         :  1   ARB    :  1   Max.   :98.34   Max.   :2.804e+10
##  (Other)        :258   (Other):258   NA's   :8       NA's   :62
##  MS.MIL.XPND.GD.ZS MS.MIL.XPND.ZS    MS.MIL.XPRT.KD
##  Min.   : 0.000    Min.   :  0.000   Min.   :0.000e+00
##  1st Qu.: 1.115    1st Qu.:  4.074   1st Qu.:1.800e+07
##  Median : 1.535    Median :  6.746   Median :5.733e+07
##  Mean   : 1.997    Mean   :  8.947   Mean   :2.266e+09
##  3rd Qu.: 2.426    3rd Qu.: 10.467   3rd Qu.:1.434e+09
##  Max.   :12.787    Max.   :144.906   Max.   :1.816e+10
##  NA's   :59        NA's   :128       NA's   :186
##  NE.EXP.GNFS.CD      NE.IMP.GNFS.CD      NY.GDP.MKTP.CD
##  Min.   :1.817e+07   Min.   :1.646e+08   Min.   :3.744e+07
##  1st Qu.:3.855e+09   1st Qu.:5.594e+09   1st Qu.:8.998e+09
##  Median :2.823e+10   Median :2.904e+10   Median :5.262e+10
##  Mean   :7.813e+11   Mean   :7.589e+11   Mean   :2.469e+12
##  3rd Qu.:2.894e+11   3rd Qu.:2.892e+11   3rd Qu.:5.396e+11
##  Max.   :2.210e+13   Max.   :2.149e+13   Max.   :7.346e+13
##  NA's   :32          NA's   :32          NA's   :19
##  NY.GDP.PCAP.CD     NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
##  Min.   :   253.4   Min.   : 0.0000   Min.   : 0.00022
##  1st Qu.:  1687.2   1st Qu.: 0.0000   1st Qu.: 0.59231
##  Median :  5785.5   Median : 0.1494   Median : 1.60804
##  Mean   : 14975.8   Mean   : 5.2032   Mean   : 3.47449
##  3rd Qu.: 15065.1   3rd Qu.: 5.0281   3rd Qu.: 3.29650
##  Max.   :154286.4   Max.   :57.7407   Max.   :49.05388
##  NA's   :19         NA's   :24        NA's   :52
```

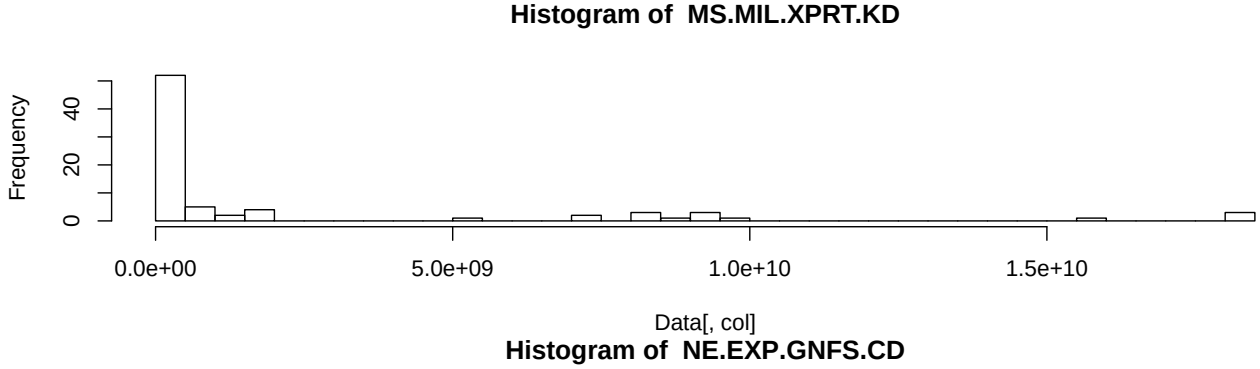
```r
#plot histogram for each numeric variable
for(col in columns[classes == 'numeric'])
{
  hist(Data[,col], breaks = 50, main = paste("Histogram of ", col))
}
```

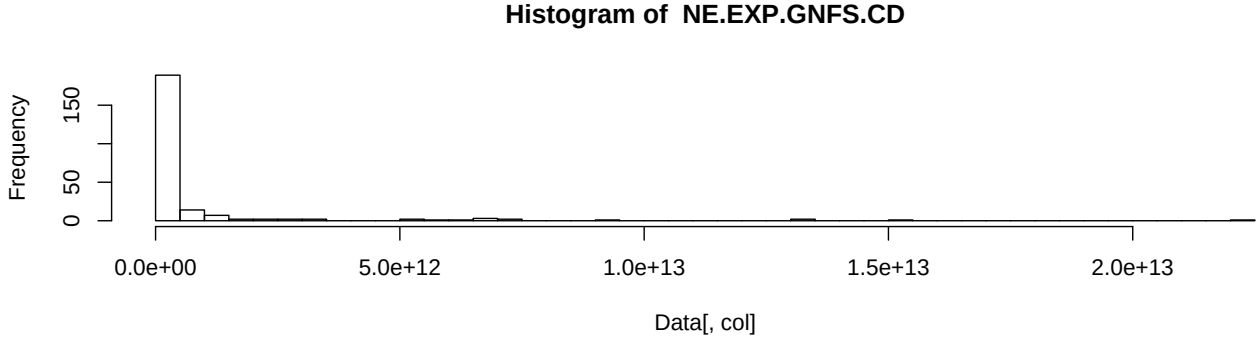


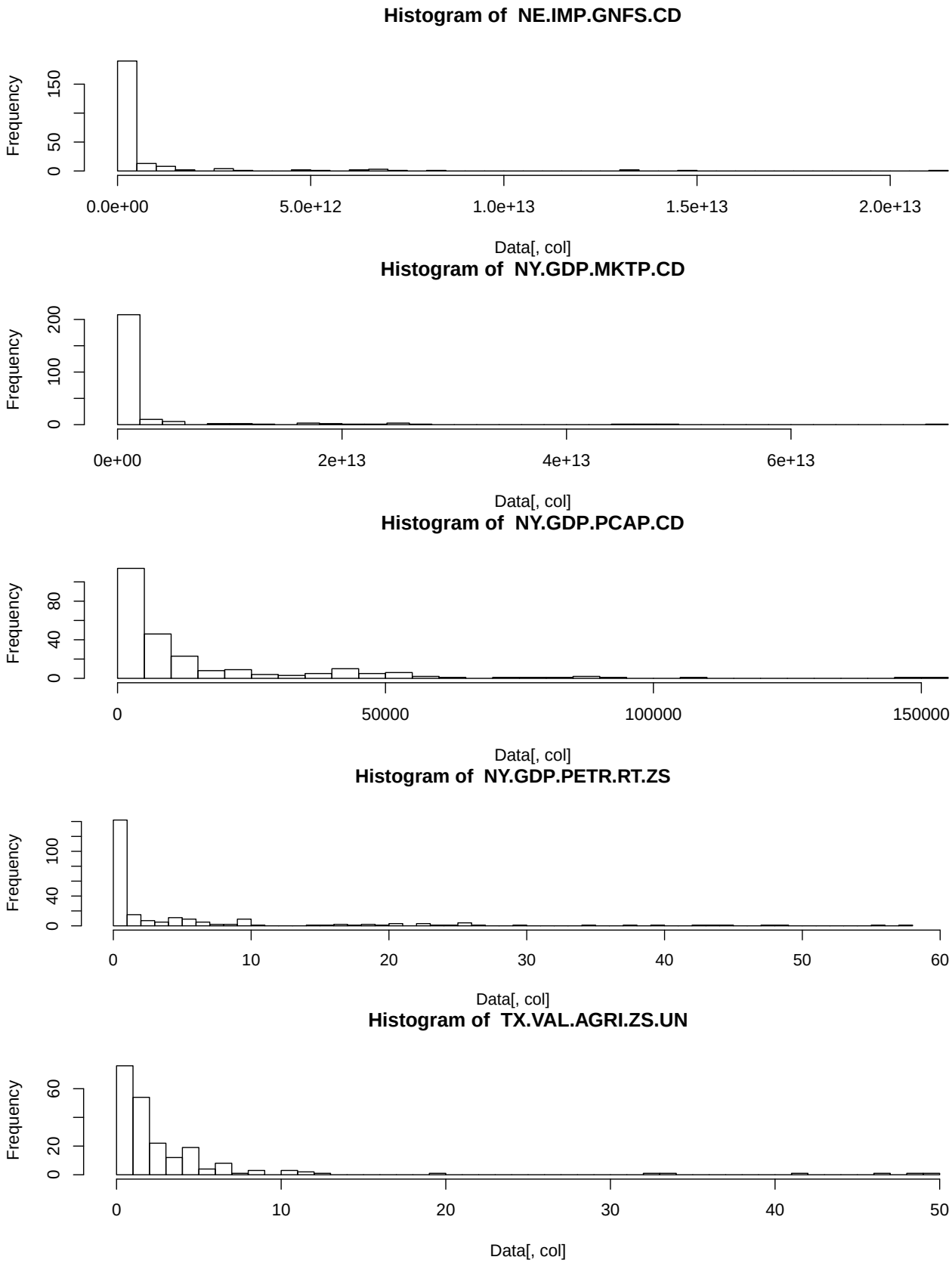

Histogram of MS.MIL.MPRT.KD
Frequency
0
40
100
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
2.5e+10
Data[, col]


Histogram of MS.MIL.XPND.GD.ZS
Frequency
0
10
20
0
2
4
6
8
10
12
Data[, col]


Histogram of MS.MIL.XPND.ZS
Frequency
0
20
40
0
50
100
150
Data[, col]


Histogram of MS.MIL.XPRT.KD
Frequency
0
20
40
0.0e+00
5.0e+09
1.0e+10
1.5e+10
Data[, col]


Histogram of NE.EXP.GNFS.CD
Frequency
0
50
150
0.0e+00
5.0e+12
1.0e+13
1.5e+13
2.0e+13
Data[, col]



We observe some skewness and potential outliers in our data but they seem to be in range. No transformations

is needed at the point.

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

```
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 function looks at dataframe Data (with first two columns excluded) and compute the percentage of non-NA values in the columns. (!is.na() returns 0 for NA ans 1 for non-NA)

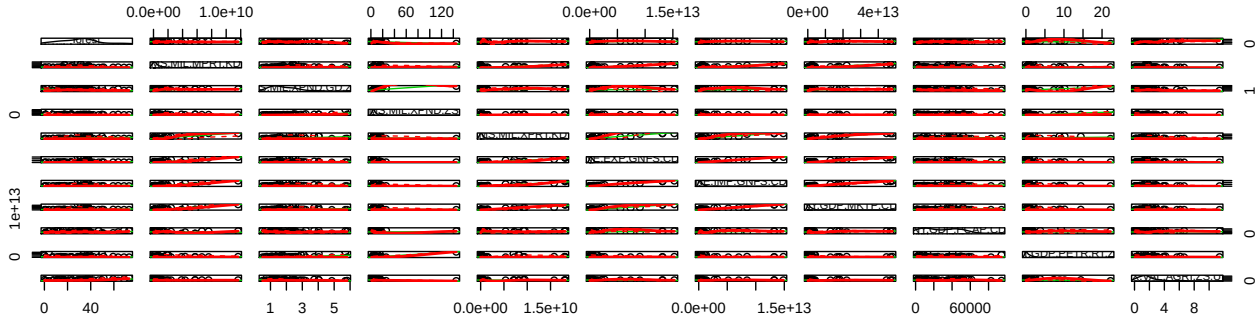
- Can you include both `NE.IMP.GNFS.CD` and `NE.EXP.GNFS.CD` in the same OLS model? Why? Yes, they might be correlated but shouldn't have collinearity with each other.
- Rename the variable named `AG.LND.FRST.ZS` to `forest`. This is going to be our dependent variable.

```
names(Data)[names(Data) == "AG.LND.FRST.ZS"] <- "forest"
```

Decribe a model for that predicts `forest`

- Write a model with two explanatory variables.

```
scatterplotMatrix(Data[, -(1:2)])
```




```
cor(Data[, -(1:2)], use = "complete.obs")
```

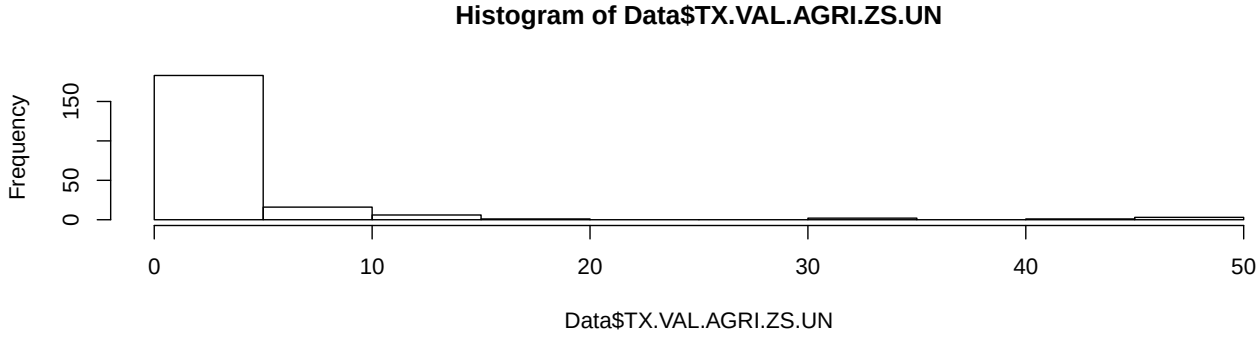
```
##                      forest MS.MIL.MPRT.KD MS.MIL.XPND.GD.ZS
## forest            1.00000000    -0.03998654       -0.25220161
## MS.MIL.MPRT.KD   -0.03998654     1.00000000        0.19155995
## MS.MIL.XPND.GD.ZS -0.25220161     0.19155995        1.00000000
## MS.MIL.XPND.ZS   -0.24280966     0.08337472        0.61711211
## MS.MIL.XPRT.KD    0.14881941     0.73559833        0.24571779
## NE.EXP.GNFS.CD    0.08781793     0.82433388        0.08998635
## NE.IMP.GNFS.CD    0.08486420     0.82757634        0.10165348
## NY.GDP.MKTP.CD    0.08539308     0.82040039        0.15307625
## NY.GDP.PCAP.CD    0.11106271    -0.06158964       -0.11782788
## NY.GDP.PETR.RT.ZS -0.05459529     0.02889363        0.45098282
## TX.VAL.AGRI.ZS.UN  0.38927867    -0.06947298       -0.23266049
##                   MS.MIL.XPND.ZS MS.MIL.XPRT.KD NE.EXP.GNFS.CD
## forest               -0.24280966     0.14881941     0.08781793
## MS.MIL.MPRT.KD        0.08337472     0.73559833     0.82433388
## MS.MIL.XPND.GD.ZS     0.61711211     0.24571779     0.08998635
```

```
## MS.MIL.XPND.ZS        1.00000000    -0.01281551    -0.03256410
## MS.MIL.XPRT.KD       -0.01281551     1.00000000     0.91161535
## NE.EXP.GNFS.CD       -0.03256410     0.91161535     1.00000000
## NE.IMP.GNFS.CD       -0.03098878     0.91677341     0.99886225
## NY.GDP.MKTP.CD       -0.02014183     0.92999254     0.97489084
## NY.GDP.PCAP.CD        0.01723753     0.10576651     0.14709980
## NY.GDP.PETR.RT.ZS     0.70162419     0.11558163    -0.04885716
## TX.VAL.AGRI.ZS.UN    -0.17232007    -0.06781204    -0.07914586
##                   NE.IMP.GNFS.CD NY.GDP.MKTP.CD NY.GDP.PCAP.CD
## forest                0.08486420     0.08539308    0.111062709
## MS.MIL.MPRT.KD        0.82757634     0.82040039   -0.061589639
## MS.MIL.XPND.GD.ZS     0.10165348     0.15307625   -0.117827876
## MS.MIL.XPND.ZS       -0.03098878    -0.02014183    0.017237530
## MS.MIL.XPRT.KD        0.91677341     0.92999254    0.105766507
## NE.EXP.GNFS.CD        0.99886225     0.97489084    0.147099799
## NE.IMP.GNFS.CD        1.00000000     0.98389962    0.149148299
## NY.GDP.MKTP.CD        0.98389962     1.00000000    0.162137440
## NY.GDP.PCAP.CD        0.14914830     0.16213744    1.000000000
## NY.GDP.PETR.RT.ZS    -0.05525580    -0.05063575   -0.004316487
## TX.VAL.AGRI.ZS.UN    -0.07384466    -0.04944996    0.028044168
##                   NY.GDP.PETR.RT.ZS TX.VAL.AGRI.ZS.UN
## forest                -0.054595289        0.38927867
## MS.MIL.MPRT.KD         0.028893630       -0.06947298
## MS.MIL.XPND.GD.ZS      0.450982821       -0.23266049
## MS.MIL.XPND.ZS         0.701624189       -0.17232007
## MS.MIL.XPRT.KD         0.115581635       -0.06781204
## NE.EXP.GNFS.CD        -0.048857161       -0.07914586
## NE.IMP.GNFS.CD        -0.055255804       -0.07384466
## NY.GDP.MKTP.CD        -0.050635754       -0.04944996
## NY.GDP.PCAP.CD        -0.004316487        0.02804417
## NY.GDP.PETR.RT.ZS      1.000000000       -0.08090071
## TX.VAL.AGRI.ZS.UN     -0.080900705        1.00000000
```

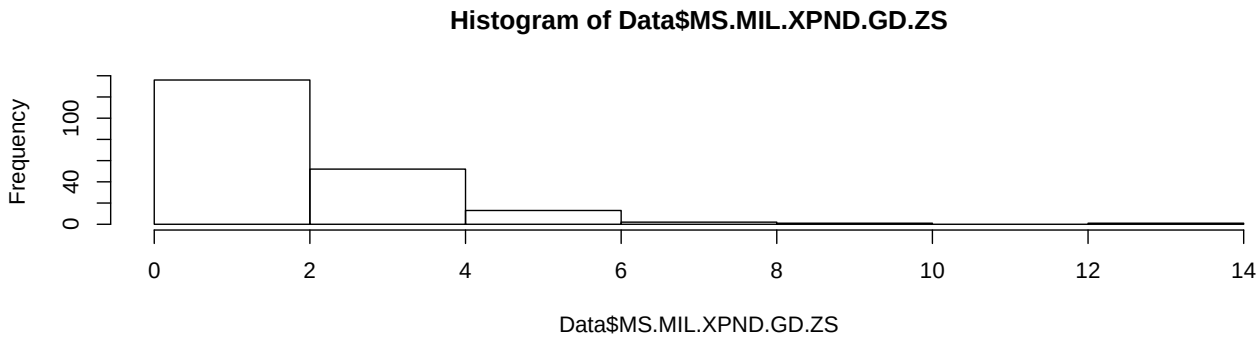
From the analysis above, choose the two variables with the highest correlation.

$$forest = \beta_0 + \beta_1 * TX.VAL.AGRI.ZS.UN + \beta_1 * MS.MIL.XPND.GD.ZS + u$$

```
hist(Data$TX.VAL.AGRI.ZS.UN)
```




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

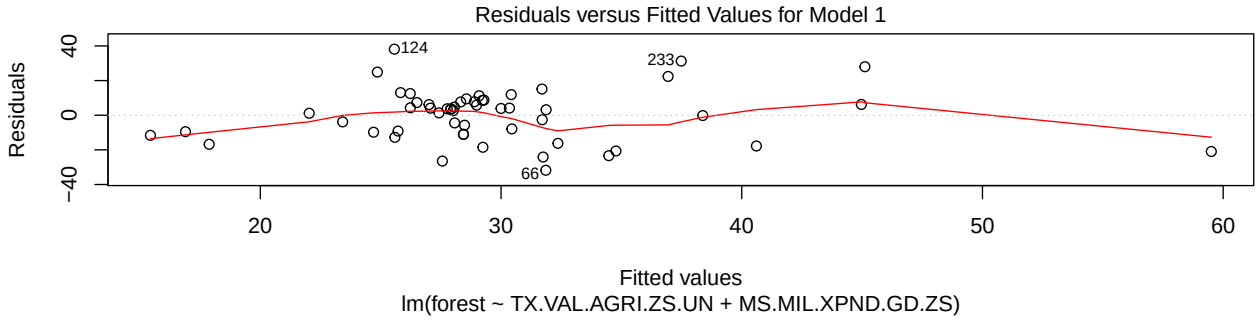


We notice that for both varaibles, they are positively skewed with a few outliers that have higher values.

```
model1 = lm(forest ~ TX.VAL.AGRI.ZS.UN + MS.MIL.XPND.GD.ZS, data = na.omit(Data))
```

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

```
plot(model1, which = 1, "Residuals versus Fitted Values for Model 1")
```




```
mean(model1$residuals)
```

```
## [1] -1.233485e-15
```

The residuals averages about 0, indicating unbiased coefficients.

- How many observations are being used in your analysis?

```
nobs(model1)
```

```
## [1] 54
```

there are 54 observations

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

No, all the NA values are dropped and this is not random that certain coutries have NA values. This could lead to inaccurate inference as we lack certain data points that could potentially have large influences.

- Now add a third variable.

```
#picking from the correlation matrix, we have the third most correlated variable as MS.MIL.XPND.ZS
#MS.MIL.XPND.ZS and MS.MIL.XPND.GD.ZS are possibly correlated but they don't have collinearity
model2 <- lm(forest ~ TX.VAL.AGRI.ZS.UN + MS.MIL.XPND.GD.ZS + MS.MIL.XPND.ZS, data = na.omit(Data))
model2$coefficients
```

```
##       (Intercept) TX.VAL.AGRI.ZS.UN MS.MIL.XPND.GD.ZS    MS.MIL.XPND.ZS
##        27.6510104         2.9215465        -1.3060375        -0.1049471
```

- 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.

```
model3 <- lm(MS.MIL.XPND.ZS ~ TX.VAL.AGRI.ZS.UN + MS.MIL.XPND.GD.ZS, data = na.omit(Data))
model4 <- lm(na.omit(Data)$forest ~ model3$residuals)
model4$coefficients
```

```
##      (Intercept) model3$residuals
##       29.7295488       -0.1049471
```

We got the same coefficient for MS.MIL.XPND.ZS.

- Compare your two models.

```
summary(model1)
```

```
##
## Call:
## lm(formula = forest ~ TX.VAL.AGRI.ZS.UN + MS.MIL.XPND.GD.ZS,
##     data = na.omit(Data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.788 -10.728   3.271   7.741  38.139
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         28.670      5.099   5.623 7.94e-07 ***
## TX.VAL.AGRI.ZS.UN    2.954      1.102   2.679   0.0099 **
## MS.MIL.XPND.GD.ZS   -2.355      1.797  -1.310   0.1961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.14 on 51 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.147
## F-statistic: 5.566 on 2 and 51 DF,  p-value: 0.006511
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = forest ~ TX.VAL.AGRI.ZS.UN + MS.MIL.XPND.GD.ZS +
##     MS.MIL.XPND.ZS, data = na.omit(Data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.882 -10.704   2.774   7.850  37.899
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        27.6510     5.2874   5.230 3.35e-06 ***
## TX.VAL.AGRI.ZS.UN   2.9215     1.1075   2.638   0.0111 *
## MS.MIL.XPND.GD.ZS  -1.3060     2.2607  -0.578   0.5660
## MS.MIL.XPND.ZS     -0.1049     0.1363  -0.770   0.4449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.2 on 50 degrees of freedom
```

```
## Multiple R-squared:  0.1888, Adjusted R-squared:  0.1401
## F-statistic: 3.878 on 3 and 50 DF,  p-value: 0.01434
```

- Do you see an improvement? Explain how you can tell.

Yes, as the R-square value increased from model 1 to model 2.

Make up a country

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

```
a = median(na.omit(Data)$TX.VAL.AGRI.ZS.UN)
b = median(na.omit(Data)$MS.MIL.XPND.GD.ZS)
c = median(na.omit(Data)$MS.MIL.XPND.ZS)
```

- How much forest would this country have?

```
predict(model2,data.frame(TX.VAL.AGRI.ZS.UN = a, MS.MIL.XPND.GD.ZS = b, MS.MIL.XPND.ZS = c))
```

```
##        1
## 29.27143
```

It's predicted to have 29.27% (of land area) forest.

Take away

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

We can't really make any conclusion regarding causality from our analysis since we can't say that all the other vairables besides our predictors remain unchanged when we change our predictors.