Assignment 2

```
library(readxl)
```

```
temp = tempfile(fileext = ".xlsx")
```

```
dataURL <- "https://raw.githubusercontent.com/popester007/data/master/project%20data%20GHI.csv"
```

```
download.file(dataURL, destfile =temp, mode='wb')
```

```
ghIndex = read.csv(dataURL)
```

```
head(ghIndex,51)
```

```
##    Rank                  Country X1992 X2000 X2008 X2017
## 1   119 Central African Republic  52.2  50.9  47.0  50.9
## 2   118                     Chad  62.5  51.9  50.9  43.5
## 3   117             Sierra Leone  57.2  54.7  44.5  38.5
## 4   116               Madagascar  43.9  43.6  36.8  38.3
## 5   115                   Zambia  48.5  52.3  45.0  38.2
## 6   114                    Yemen  43.5  43.4  36.2  36.1
## 7   113                    Sudan    NA    NA    NA  35.5
## 8   112                  Liberia  51.2  48.2  38.9  35.3
## 9   111                    Niger  62.2  52.6  37.0  34.5
## 10  110              Timor-Leste    NA    NA  46.8  34.3
## 11  109                    Haiti  51.6   2.7  42.6  34.2
## 12  108                 Zimbabwe  35.8  40.9  34.5  33.8
## 13  107              Afghanistan  50.2  52.7  37.5  33.3
## 14  106                 Pakistan  42.7  38.2  34.7  32.6
## 15  105                   Angola  65.8  57.5  39.7  32.5
## 16  104                 Ethiopia    NA  56.0  40.2  32.3
## 17  103                   Uganda  41.2  39.2  33.3  32.0
## 18  102                 Djibouti  60.3  46.7  35.1  31.4
## 19  101                    India  46.2  38.2  35.1  31.4
## 20  100                   Rwanda  53.3  56.3  35.6  31.4
## 21   99             Guina-Bissau  44.5  43.1  36.2  30.6
## 22   98               Mozambique  63.6  48.7  31.4  30.5
## 23   97                 Tanzania  42.9  42.4  37.5  28.8
## 24   96               Tajikistan    NA  41.8  32.6  28.7
## 25   95                   Guinea  46.5  44.0  33.4  28.6
## 26   94                     Mali  51.4  44.2  35.1  28.6
## 27   93              North Korea  31.9  40.3  30.7  28.2
## 28   92             Burkina Faso  47.0  47.9  36.4  27.6
## 29   91                 Laos PDR  52.3  48.1  33.4  27.5
## 30   90                   Malawi  58.2  44.6  31.5  27.2
## 31   89               Bangladesh  53.6  37.6  32.2  26.5
## 32   88            Côte d'Ivoire  32.9  32.6  35.1  26.5
## 33   87                  Namibia  35.4  30.8  30.9  25.7
## 34   86                    Congo  39.1  36.0  31.6  25.6
## 35   85                  Nigeria  48.8  41.0  33.7  25.5
## 36   84                Sri Lanka  31.6  26.8  24.2  25.5
## 37   83               Mauritania  39.4  33.6  23.7  25.2
## 38   82                    Benin  44.5  37.5  31.7  24.4
## 39   81                 Botswana  33.8  33.0  30.7  24.4
## 40   80                  Lesotho  26.5  33.2  28.4  24.1
## 41   79                   Gambia  35.2  27.5  23.8  23.2
```

```
## 42   78                  Iraq  21.8  26.5  25.7  22.9
## 43   77               Myanmar  55.6  43.6  30.1  22.6
## 44   76                  Togo  45.8  39.0  28.3  22.5
## 45   75              Cambodia  45.8  43.5  27.1  22.2
## 46   74              Cameroon  40.0  39.6  29.5  22.1
## 47   73             Indonesia  35.0  25.5  28.3  22.0
## 48   72                 Nepal  42.5  36.8  28.9  22.0
## 49   71             Swaziland  24.0  29.9  30.7  21.2
## 50   70                 Kenya  39.1  37.6  29.6  21.0
## 51   69             Guatemala  28.5  27.4  22.2  20.7
```

```
str(ghIndex)
```

```
## 'data.frame':    51 obs. of  6 variables:
##  $ Rank   : int  119 118 117 116 115 114 113 112 111 110 ...
##  $ Country: Factor w/ 51 levels "Afghanistan",..: 9 10 40 27 50 49 42 26 35 46 ...
##  $ X1992  : num  52.2 62.5 57.2 43.9 48.5 43.5 NA 51.2 62.2 NA ...
##  $ X2000  : num  50.9 51.9 54.7 43.6 52.3 43.4 NA 48.2 52.6 NA ...
##  $ X2008  : num  47 50.9 44.5 36.8 45 36.2 NA 38.9 37 46.8 ...
##  $ X2017  : num  50.9 43.5 38.5 38.3 38.2 36.1 35.5 35.3 34.5 34.3 ...
```

```
names(ghIndex)
```

```
## [1] "Rank"    "Country" "X1992"   "X2000"   "X2008"   "X2017"
```

```
summary(ghIndex)
```

```
##       Rank               Country       X1992           X2000
##  Min.   : 69.0   Afghanistan : 1   Min.   :21.80   Min.   : 2.70
##  1st Qu.: 81.5   Angola      : 1   1st Qu.:37.45   1st Qu.:36.00
##  Median : 94.0   Bangladesh  : 1   Median :44.50   Median :41.00
##  Mean   : 94.0   Benin       : 1   Mean   :44.80   Mean   :40.61
##  3rd Qu.:106.5   Botswana    : 1   3rd Qu.:51.90   3rd Qu.:47.90
##  Max.   :119.0   Burkina Faso: 1   Max.   :65.80   Max.   :57.50
##                  (Other)     :45   NA's   :4       NA's   :2
##      X2008           X2017
##  Min.   :22.20   Min.   :20.70
##  1st Qu.:30.25   1st Qu.:24.40
##  Median :33.40   Median :28.60
##  Mean   :33.92   Mean   :29.26
##  3rd Qu.:36.70   3rd Qu.:32.95
##  Max.   :50.90   Max.   :50.90
##  NA's   :1
```

```
names(ghIndex) <- c("Rank", "Country", "1992", "2000", "2008", "2017")
```

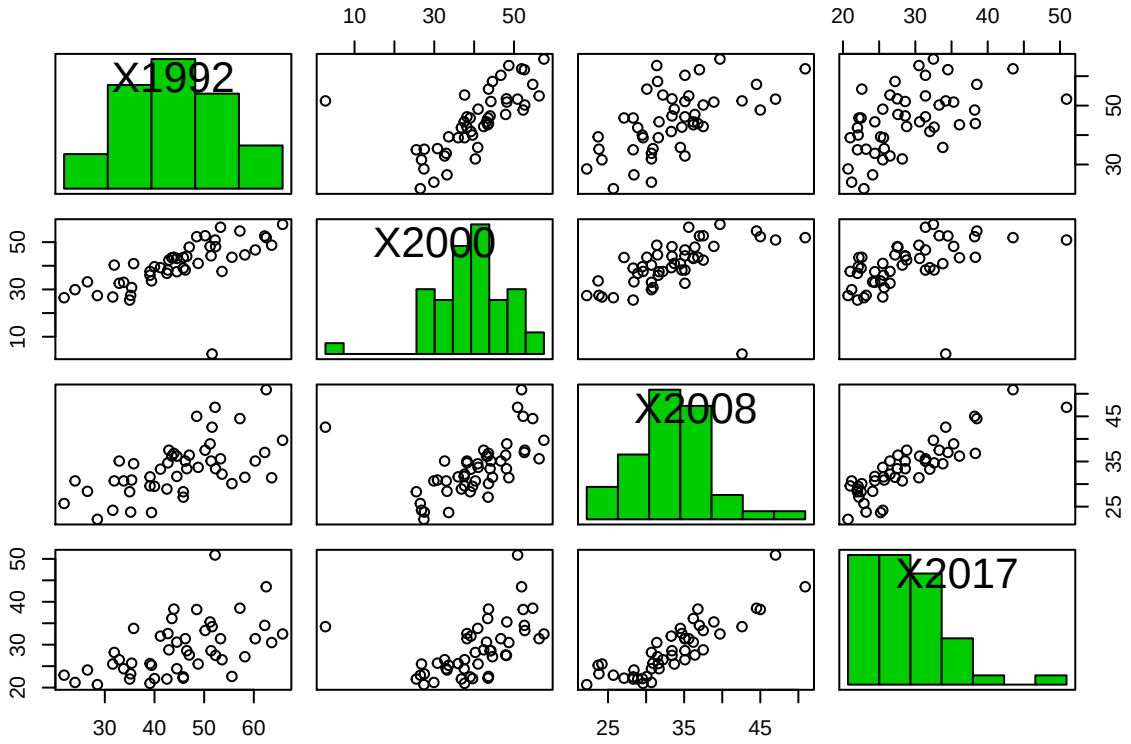
```
names(ghIndex)
```

```
## [1] "Rank"    "Country" "1992"    "2000"    "2008"    "2017"
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
scatterplotMatrix(ghIndex[,c(3:6)], diagonal='histogram',reg.line=NULL, smoother=NULL)
```




```
str(ghIndex)
```

```
## 'data.frame':    51 obs. of  6 variables:
##  $ Rank   : int  119 118 117 116 115 114 113 112 111 110 ...
##  $ Country: Factor w/ 51 levels "Afghanistan",..: 9 10 40 27 50 49 42 26 35 46 ...
##  $ 1992   : num  52.2 62.5 57.2 43.9 48.5 43.5 NA 51.2 62.2 NA ...
##  $ 2000   : num  50.9 51.9 54.7 43.6 52.3 43.4 NA 48.2 52.6 NA ...
##  $ 2008   : num  47 50.9 44.5 36.8 45 36.2 NA 38.9 37 46.8 ...
##  $ 2017   : num  50.9 43.5 38.5 38.3 38.2 36.1 35.5 35.3 34.5 34.3 ...
```

```
summary(ghIndex)
```

```
##       Rank               Country        1992            2000      
##  Min.   : 69.0   Afghanistan : 1   Min.   :21.80   Min.   : 2.70  
##  1st Qu.: 81.5   Angola      : 1   1st Qu.:37.45   1st Qu.:36.00  
##  Median : 94.0   Bangladesh  : 1   Median :44.50   Median :41.00  
##  Mean   : 94.0   Benin       : 1   Mean   :44.80   Mean   :40.61  
##  3rd Qu.:106.5   Botswana    : 1   3rd Qu.:51.90   3rd Qu.:47.90  
##  Max.   :119.0   Burkina Faso: 1   Max.   :65.80   Max.   :57.50  
##                  (Other)     :45   NA's   :4       NA's   :2      
##       2008            2017      
##  Min.   :22.20   Min.   :20.70  
##  1st Qu.:30.25   1st Qu.:24.40  
##  Median :33.40   Median :28.60  
##  Mean   :33.92   Mean   :29.26  
##  3rd Qu.:36.70   3rd Qu.:32.95  
##  Max.   :50.90   Max.   :50.90  
##  NA's   :1
```

Pick one variable to explore: 1992

```
summary(ghIndex$'1992',na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.80   37.45   44.50   44.80   51.90   65.80       4
```

Analyze dispersion

```
sd(ghIndex$'1992',na.rm = T)
```

```
## [1] 10.66472
```

```
library(e1071)
```

Analyze Skewness

```
skewness(ghIndex$'1992',na.rm = T)
```

```
## [1] -0.06774339
```

Analyze Kurtosis

```
kurtosis(ghIndex$'1992',na.rm = T)
```

```
## [1] -0.6329649
```

Next step ?

```
data=ghIndex[is.finite(ghIndex$'1992'),]
```

selecting a variable

```
var=data$'1992'
```

saving mean and sd

```
mnVar=mean(var,na.rm = T)
sdVar=sd(var,na.rm = T)
```

plotting

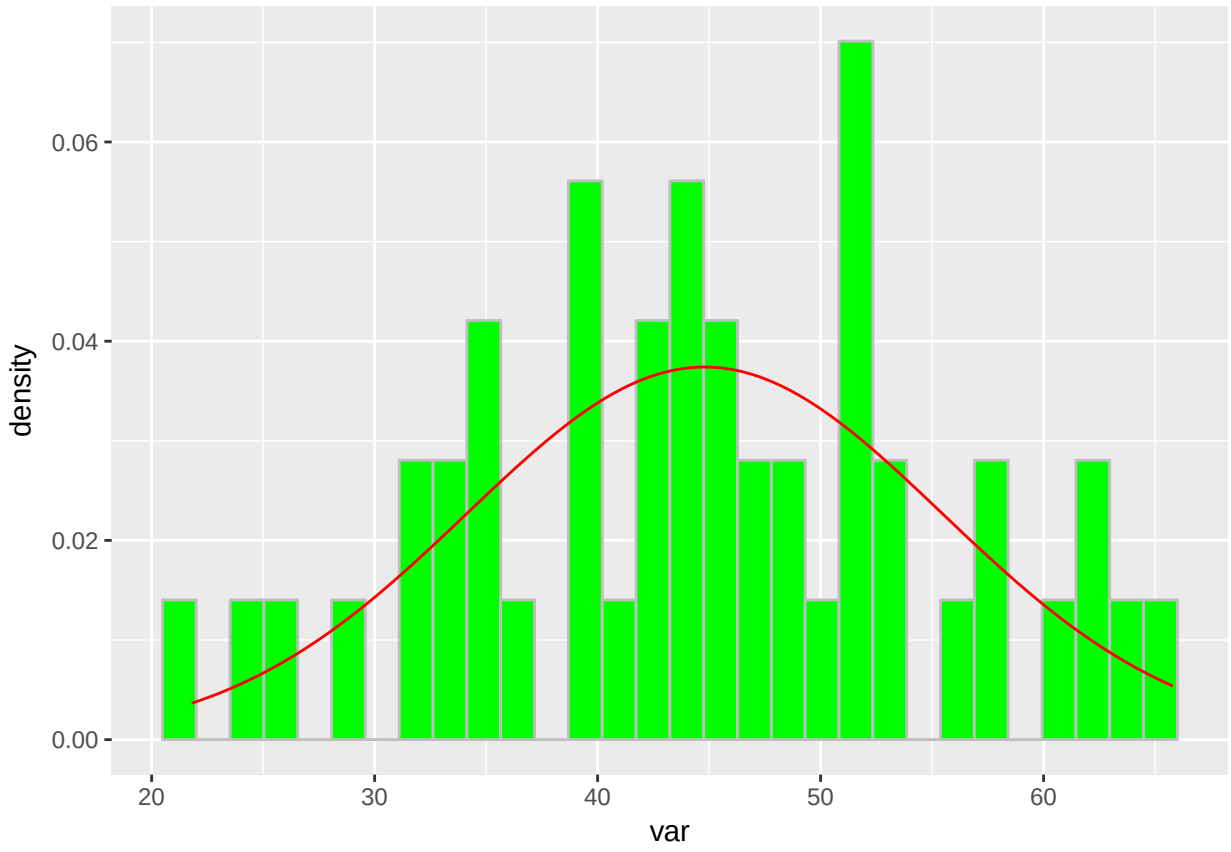
```
library(ggplot2)
base = ggplot(data, aes(x=var))
hist = base + geom_histogram(fill="green",
                             color='grey',
                             aes(y=..density..))
```

next step...?

```
histAndNormal = hist + stat_function(fun=dnorm,
                                     color="red",
                                     args=list(mean=mnVar,sd=sdVar))
```

```
histAndNormal
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



now histogram with central measures

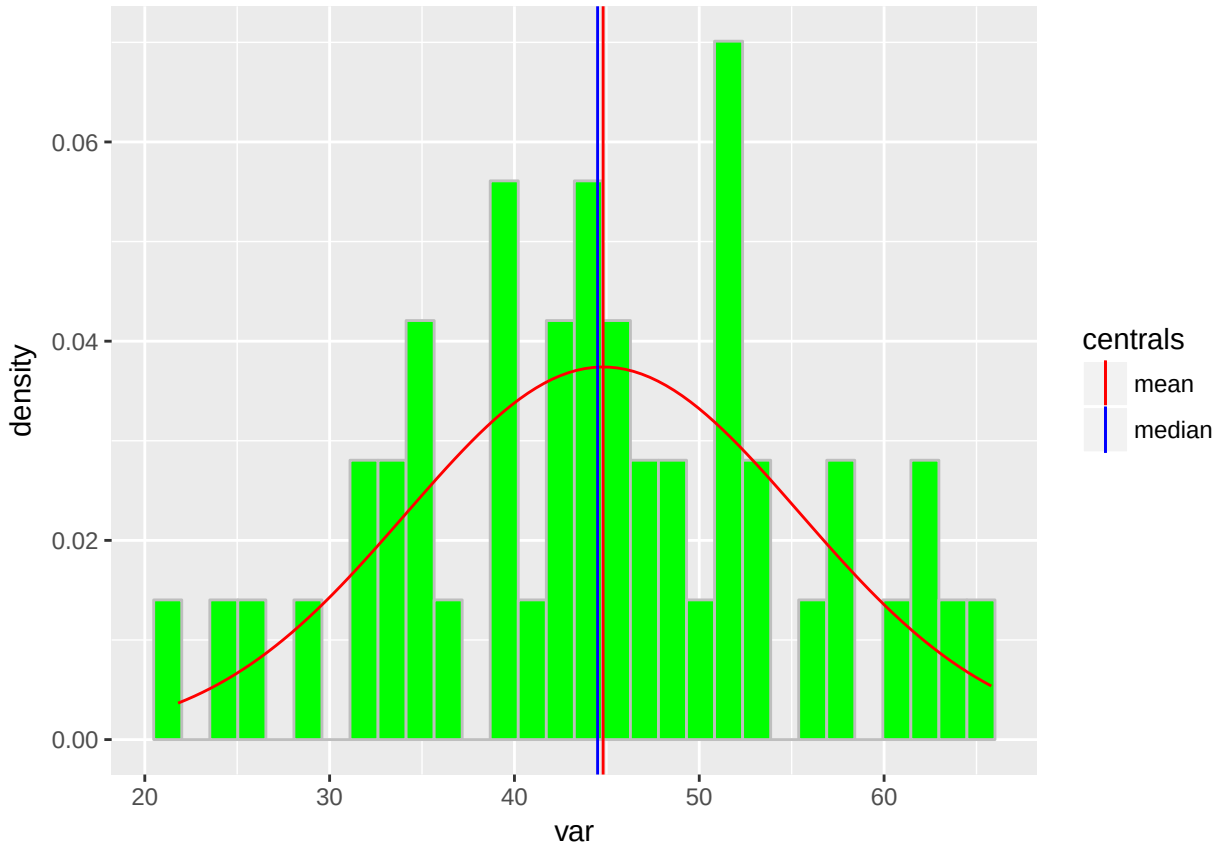
```
mdVar=median(var)
```

now histogram with central measures

```
histAndNormal + geom_vline(aes(xintercept = mnVar, colour="mean"),
                           show.legend = TRUE) +
  geom_vline(aes(xintercept = mdVar,colour="median"),
             show.ledgend = TRUE) +
  scale_color_manual(name = "centrals",
                     values = c(median = "blue", mean = "red"))
```

```
## Warning: Ignoring unknown parameters: show.ledgend
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Now for outliers

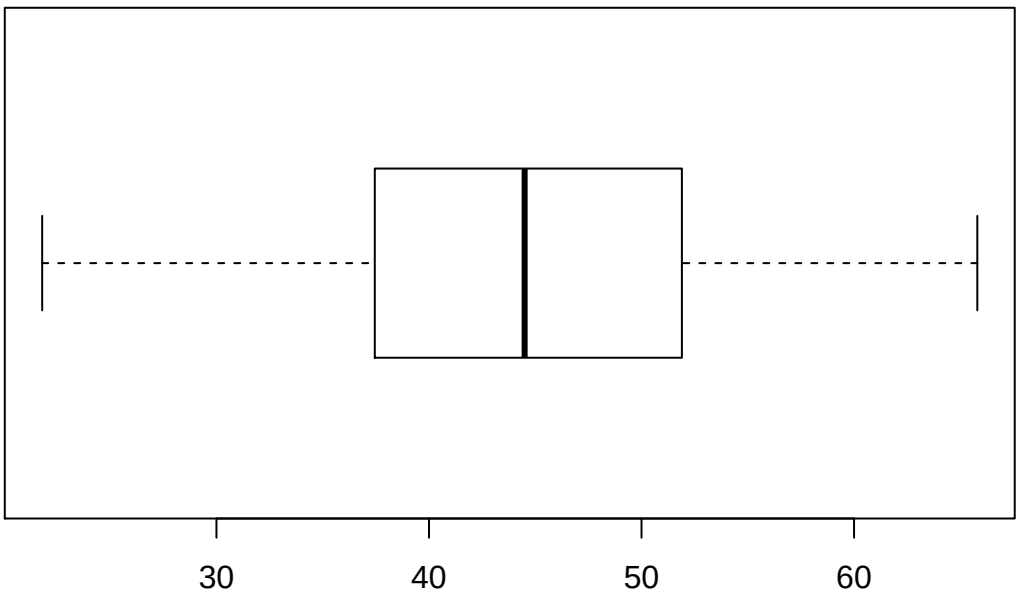
```
ghIndex$'1992'=as.numeric(ghIndex$'1992')
```

```
summary(ghIndex$'1992')
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.80   37.45   44.50   44.80   51.90   65.80       4
```

looking at 25th and 75th percentile

```
(bp=boxplot(ghIndex$'1992',horizontal = T))
```




```
## $stats
##       [,1]
## [1,] 21.80
## [2,] 37.45
## [3,] 44.50
## [4,] 51.90
## [5,] 65.80
##
## $n
## [1] 47
##
## $conf
##          [,1]
## [1,] 41.16976
## [2,] 47.83024
##
## $out
## numeric(0)
##
## $group
## numeric(0)
##
## $names
## [1] "1"
```

```
bp$stats
```

```
##       [,1]
## [1,] 21.80
## [2,] 37.45
## [3,] 44.50
## [4,] 51.90
## [5,] 65.80
```

compute IQR

```
(iqr=IQR(ghIndex$'1992',na.rm = T))
```

```
## [1] 14.45
```

For 75th Q:

```
q75=bp$stats[4]
(capHigh = q75 + iqr*1.5)
```

```
## [1] 73.575
```

For 25th Q:

```
q25=bp$stats[2]
(capLow=q25 - iqr*1.5)
```

```
## [1] 15.775
```

Any value above the High value (73.575) or below the Low value (15.775) is an outlier

```
length(bp$out)
```

```
## [1] 0
```

```
StdDev=sd(ghIndex$'1992',na.rm = T)
Mean=mean(ghIndex$'1992',na.rm = T)
```

```
(lowCapT=Mean-2*StdDev)
```

```
## [1] 23.46843
```

```
(upCapT=Mean+2*StdDev)
```

```
## [1] 66.12732
```

making maps

```
library(utils)
```