Output of RMD from Week01_IntroductionR

Shun Zhu

Wednesday, March 04, 2015

VIDEO 02: GETTING STARTED IN R

```
# Basic Calculations
8*6
```

```
## [1] 48
```

```
2^16
```

```
## [1] 65536
```

```
# you will see plus sign and wait for you completing the command
    # you Coud complete the command or hit Excape
2^
8*6
```

```
## [1] 1536
```

```
8*10
```

```
## [1] 80
```

```
# Functions could take several arguments
    # Build-in functions
    # Install packages
sqrt(2)
```

```
## [1] 1.414214
```

```
abs(-65)
```

```
## [1] 65
```

```
# Get help of any functions
?sqrt
```

```
## starting httpd help server ... done
```

```
# Variables
    # 1. DO not use spaces in variable names
        # using a mix of capital and lowercase letters
    # 2. Do not start variable names with a number.
    # 3. Case Sensitive
SquareRoot2 = sqrt(2)
SquareRoot2
```

```
## [1] 1.414214
```

```
HoursYear <- 365*24
HoursYear
```

```
## [1] 8760
```

```
#  List all of the variables that you've created in your current R session
ls()
```

```
## [1] "HoursYear"   "SquareRoot2"
```

VIDEO 03: VECTORS AND DATA FRAMES

```
# 01. c() indicates combining same objects in Columns
c(2,3,5,8,13)
```

```
## [1]  2  3  5  8 13
```

```
# 02
Country = c("Brazil", "China", "India", "Switzerland", "USA")
LifeExpectancy = c(74, 76, 65, 83, 79)

# 03 Automatically Convert numbers into string
c("Brazil", 74, "China", 76)
```

```
## [1] "Brazil" "74"     "China"  "76"
```

```
# 04
Country[1]
```

```
## [1] "Brazil"
```

```
LifeExpectancy[3]
```

```
## [1] 65
```

```
# 05
Sequence = seq(1, 100, 2)

# 06 Maintain the type of original object
CountryData = data.frame(Country, LifeExpectancy)
CountryData
```

```
##       Country LifeExpectancy
## 1      Brazil             74
## 2       China             76
## 3       India             65
## 4 Switzerland             83
## 5         USA             79
```

```
# 07 Column Comibine
    # R will just combine the vectors in the order they're typed
Population = c(199000, 1390000, 1240000, 1997, 318000)
CountryData = cbind(CountryData, Population)
CountryData
```

```
##       Country LifeExpectancy Population
## 1      Brazil             74     199000
## 2       China             76    1390000
## 3       India             65    1240000
## 4 Switzerland             83       1997
## 5         USA             79     318000
```

```
# 08. Add new observations with row combining

Country = c("Australia", "Greece")
LifeExpectancy = c(81, 82)
Population = c(23050, 11125)
NewCountryData = data.frame(Country, LifeExpectancy, Population)

AllCountryData = rbind(CountryData, NewCountryData)
AllCountryData
```

```
##       Country LifeExpectancy Population
## 1      Brazil             74     199000
## 2       China             76    1390000
## 3       India             65    1240000
## 4 Switzerland             83       1997
## 5         USA             79     318000
## 6   Australia             81      23050
## 7      Greece             82      11125
```

VIDEO 04: LOADING DATA FILES

```
# 1. Change working directory that has WHO.csv(World Health Organization)

getwd()
```

```
## [1] "F:/SkyDrive/Studying/MIT_COURSES/15-071-TheAnalyticsEdge/lecture/week01"
```

```
setwd("F:/SkyDrive/Studying/MIT_COURSES/15-071-TheAnalyticsEdge/lecture/dataset")

# 2. Loading csv files
WHO = read.csv("WHO.csv")


# 3. str function shows the structure of the data
#  the name of the country
# the region the country is in
# the population in thousands
```

```r
# the percentage of the population under 15
# the percentage of the population over 60
# the fertility rate or average number of children per woman
# the life expectancy in years
# the child mortality rate which is the number of children who die by age five per 1,000 births
# the number of cellular subscribers per 100 population
# the literacy rate among adults aged greater than or equal to 15
# the gross national income per capital
# the percentage of male children enrolled in primary school
# he percentage of female children enrolled in primary school
str(WHO)
```

```
## 'data.frame':    194 obs. of  13 variables:
##  $ Country                      : Factor w/ 194 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Region                       : Factor w/ 6 levels "Africa","Americas",..: 3 4 1 4 1 2 2 4 6 4 ...
##  $ Population                   : int  29825 3162 38482 78 20821 89 41087 2969 23050 8464 ...
##  $ Under15                      : num  47.4 21.3 27.4 15.2 47.6 ...
##  $ Over60                       : num  3.82 14.93 7.17 22.86 3.84 ...
##  $ FertilityRate                : num  5.4 1.75 2.83 NA 6.1 2.12 2.2 1.74 1.89 1.44 ...
##  $ LifeExpectancy               : int  60 74 73 82 51 75 76 71 82 81 ...
##  $ ChildMortality               : num  98.5 16.7 20 3.2 163.5 ...
##  $ CellularSubscribers          : num  54.3 96.4 99 75.5 48.4 ...
##  $ LiteracyRate                 : num  NA NA NA NA 70.1 99 97.8 99.6 NA NA ...
##  $ GNI                          : num  1140 8820 8310 NA 5230 ...
##  $ PrimarySchoolEnrollmentMale  : num  NA NA 98.2 78.4 93.1 91.1 NA NA 96.9 NA ...
##  $ PrimarySchoolEnrollmentFemale: num  NA NA 96.4 79.4 78.2 84.5 NA NA 97.5 NA ...
```

```r
# 4. summary function gives a numerical summary of each of our variables
summary(WHO)
```

```
##                 Country                       Region     Population
##  Afghanistan        :  1   Africa                 :46   Min.   :      1
##  Albania            :  1   Americas               :35   1st Qu.:   1696
##  Algeria            :  1   Eastern Mediterranean:22   Median :   7790
##  Andorra            :  1   Europe                 :53   Mean   :  36360
##  Angola             :  1   South-East Asia        :11   3rd Qu.:  24535
##  Antigua and Barbuda:  1   Western Pacific        :27   Max.   :1390000
##  (Other)            :188
##     Under15          Over60      FertilityRate   LifeExpectancy
##  Min.   :13.12   Min.   : 0.81   Min.   :1.260   Min.   :47.00
##  1st Qu.:18.72   1st Qu.: 5.20   1st Qu.:1.835   1st Qu.:64.00
##  Median :28.65   Median : 8.53   Median :2.400   Median :72.50
##  Mean   :28.73   Mean   :11.16   Mean   :2.941   Mean   :70.01
##  3rd Qu.:37.75   3rd Qu.:16.69   3rd Qu.:3.905   3rd Qu.:76.00
##  Max.   :49.99   Max.   :31.92   Max.   :7.580   Max.   :83.00
##                                  NA's   :11
##  ChildMortality    CellularSubscribers  LiteracyRate        GNI
##  Min.   :  2.200   Min.   :  2.57      Min.   :31.10   Min.   :  340
##  1st Qu.:  8.425   1st Qu.: 63.57      1st Qu.:71.60   1st Qu.: 2335
##  Median : 18.600   Median : 97.75      Median :91.80   Median : 7870
##  Mean   : 36.149   Mean   : 93.64      Mean   :83.71   Mean   :13321
##  3rd Qu.: 55.975   3rd Qu.:120.81      3rd Qu.:97.85   3rd Qu.:17558
##  Max.   :181.600   Max.   :196.41      Max.   :99.80   Max.   :86440
```

```
##                   NA's   :10          NA's   :91      NA's   :32
##  PrimarySchoolEnrollmentMale PrimarySchoolEnrollmentFemale
##  Min.   : 37.20              Min.   : 32.50
##  1st Qu.: 87.70              1st Qu.: 87.30
##  Median : 94.70              Median : 95.10
##  Mean   : 90.85              Mean   : 89.63
##  3rd Qu.: 98.10              3rd Qu.: 97.90
##  Max.   :100.00              Max.   :100.00
##  NA's   :93                  NA's   :93
```

```
# 5. The subset function takes two arguments. The first is the data frame we want to take a subset of,

WHO_Europe = subset(WHO, Region == "Europe")
str(WHO_Europe)
```

```
## 'data.frame':    53 obs. of  13 variables:
##  $ Country                      : Factor w/ 194 levels "Afghanistan",..: 2 4 8 10 11 16 17 22 26 43
##  $ Region                       : Factor w/ 6 levels "Africa","Americas",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Population                   : int  3162 78 2969 8464 9309 9405 11060 3834 7278 4307 ...
##  $ Under15                      : num  21.3 15.2 20.3 14.5 22.2 ...
##  $ Over60                       : num  14.93 22.86 14.06 23.52 8.24 ...
##  $ FertilityRate                : num  1.75 NA 1.74 1.44 1.96 1.47 1.85 1.26 1.51 1.48 ...
##  $ LifeExpectancy               : int  74 82 71 81 71 71 80 76 74 77 ...
##  $ ChildMortality               : num  16.7 3.2 16.4 4 35.2 5.2 4.2 6.7 12.1 4.7 ...
##  $ CellularSubscribers          : num  96.4 75.5 103.6 154.8 108.8 ...
##  $ LiteracyRate                 : num  NA NA 99.6 NA NA NA NA 97.9 NA 98.8 ...
##  $ GNI                          : num  8820 NA 6100 42050 8960 ...
##  $ PrimarySchoolEnrollmentMale  : num  NA 78.4 NA NA 85.3 NA 98.9 86.5 99.3 94.8 ...
##  $ PrimarySchoolEnrollmentFemale: num  NA 79.4 NA NA 84.1 NA 99.2 88.4 99.7 97 ...
```

```
# 6. save this new data frame, WHO_Europe, to a csv file
write.csv(WHO_Europe, "WHO_Europe.csv")

# 7. Removing variables in the working space to save used space
ls()
```

```
##  [1] "AllCountryData" "Country"        "CountryData"    "HoursYear"
##  [5] "LifeExpectancy" "NewCountryData" "Population"     "Sequence"
##  [9] "SquareRoot2"    "WHO"            "WHO_Europe"
```

```
rm(WHO_Europe)
```

VIDEO 05: DATA ANALYSIS

```
# access a variable in a data frame
WHO$under15
```

```
## NULL
```

```r
# Basic data analysis

mean(WHO$Under15)
```

```
## [1] 28.73242
```

```r
sd(WHO$Under15)
```

```
## [1] 10.53457
```

```r
summary(WHO$Under15)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.12   18.72   28.65   28.73   37.75   49.99
```

```r
# There's a country with only 13.12% (minimun) of the population under 15. Let's see which one it is.

which.min(WHO$Under15)
```

```
## [1] 86
```

```r
WHO$Country[86]
```

```
## [1] Japan
## 194 Levels: Afghanistan Albania Algeria Andorra ... Zimbabwe
```

```r
which.max(WHO$Under15)
```

```
## [1] 124
```

```r
WHO$Country[124]
```

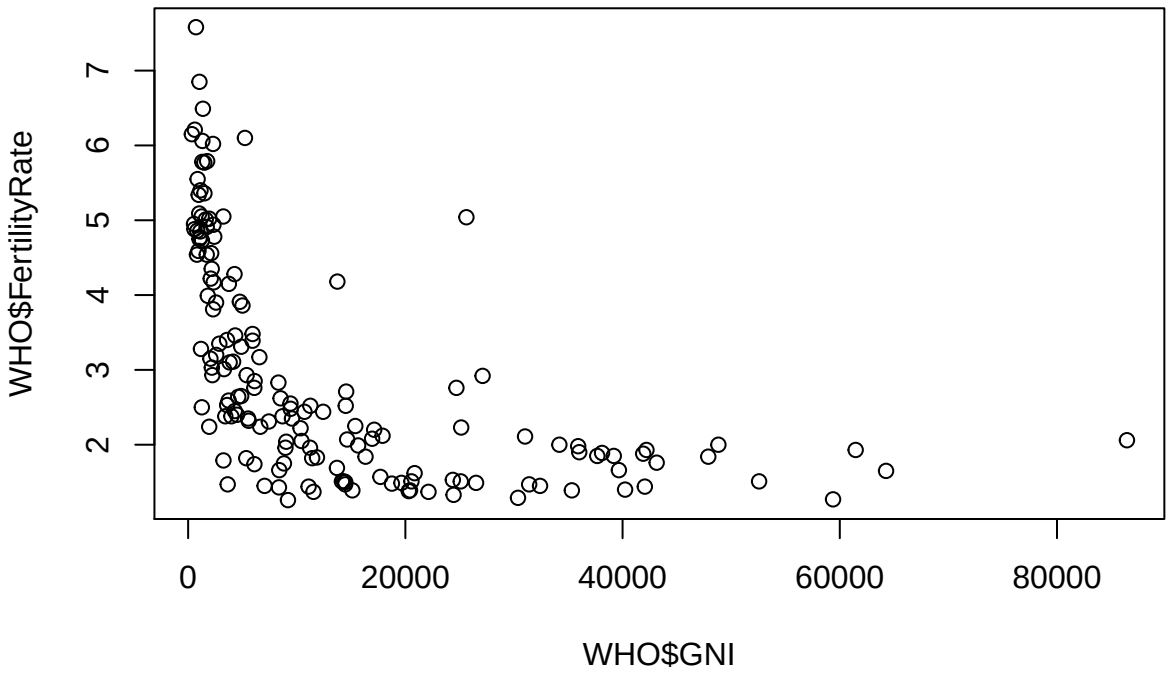
```
## [1] Niger
## 194 Levels: Afghanistan Albania Algeria Andorra ... Zimbabwe
```

```r
sort(WHO$Under15)
```

```
##   [1] 13.12 13.17 13.28 13.53 14.04 14.04 14.16 14.18 14.41 14.51 14.56
##  [12] 14.57 14.60 14.62 14.79 14.91 14.92 14.98 14.98 15.00 15.05 15.10
##  [23] 15.13 15.20 15.20 15.25 15.45 15.69 16.35 16.37 16.42 16.45 16.48
##  [34] 16.52 16.58 16.71 16.88 16.89 17.16 17.21 17.46 17.54 17.62 17.66
##  [45] 17.95 18.26 18.26 18.47 18.64 18.95 18.99 19.01 19.63 20.16 20.17
##  [56] 20.26 20.34 20.71 20.73 21.33 21.38 21.54 21.62 21.64 21.95 21.98
##  [67] 22.05 22.25 22.87 23.22 23.68 23.94 24.19 24.31 24.42 24.56 24.90
##  [78] 25.15 25.28 25.46 25.70 25.75 25.96 25.96 25.96 26.00 26.65 26.96
##  [89] 27.05 27.42 27.53 27.78 27.83 27.85 28.03 28.53 28.65 28.65 28.84
## [100] 28.88 28.90 29.02 29.03 29.18 29.27 29.43 29.45 29.53 29.69 30.10
## [111] 30.10 30.10 30.10 30.17 30.21 30.29 30.53 30.57 30.61 30.61 30.61
```

```
## [122] 30.62 31.23 31.25 32.78 33.37 33.72 33.75 34.13 34.31 34.40 34.53
## [133] 35.23 35.35 35.35 35.58 35.61 35.72 35.75 35.81 36.59 36.75 36.77
## [144] 37.33 37.37 37.88 38.05 38.37 38.49 38.59 38.95 40.07 40.22 40.24
## [155] 40.37 40.51 40.72 40.80 41.48 41.48 41.55 41.60 41.74 41.89 42.17
## [166] 42.28 42.37 42.37 42.46 42.72 42.95 43.06 43.08 43.10 43.29 43.54
## [177] 43.56 44.20 44.23 44.85 45.11 45.38 45.44 45.66 45.90 46.33 46.73
## [188] 47.14 47.35 47.42 47.58 48.52 48.54 49.99
```

```
# Scatterplot
plot(WHO$GNI, WHO$FertilityRate)
```




```
# Subsetting
Outliers = subset(WHO, GNI > 10000 & FertilityRate > 2.5)

#  The number of rows in the dataset
nrow(Outliers)
```

```
## [1] 7
```

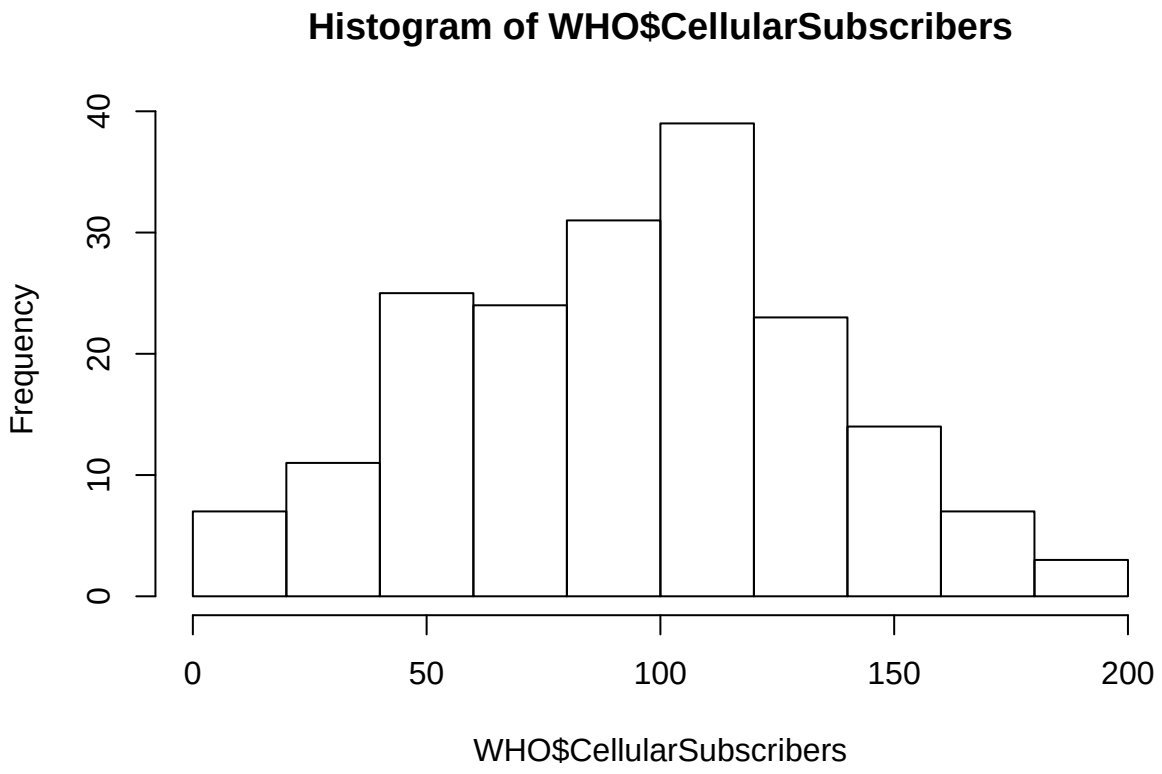
```
# Extract a few variables from a data set
Outliers[c("Country","GNI","FertilityRate")]
```

```
##               Country   GNI FertilityRate
## 23           Botswana 14550          2.71
```

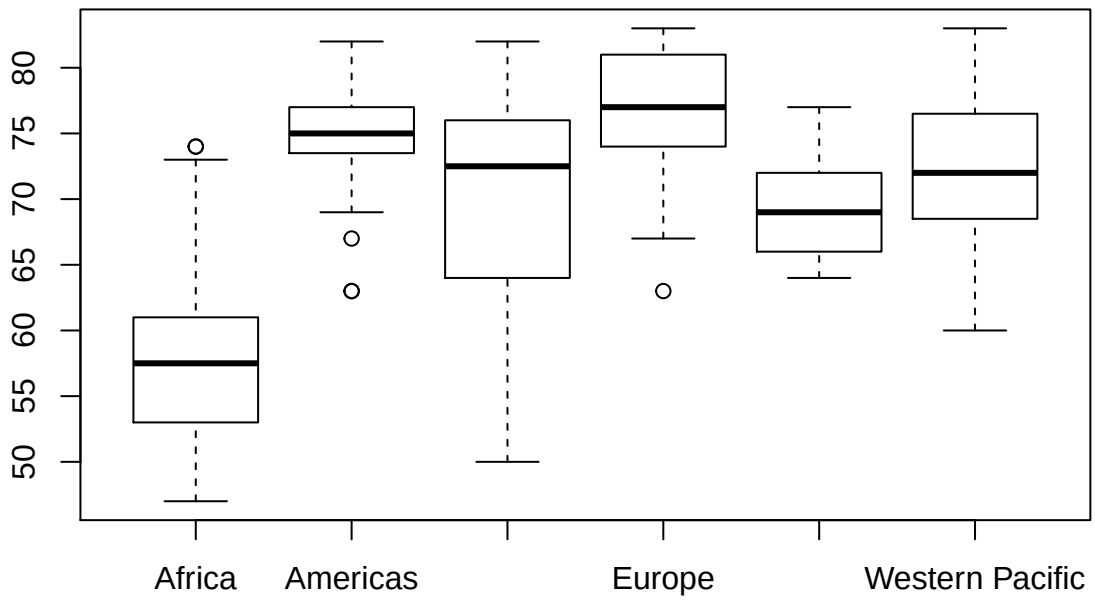
```
## 56  Equatorial Guinea 25620          5.04
## 63               Gabon 13740          4.18
## 83              Israel 27110          2.92
## 88          Kazakhstan 11250          2.52
## 131             Panama 14510          2.52
## 150       Saudi Arabia 24700          2.76
```

VIDEO 06 : PLOTS AND SUMMARY TABLES

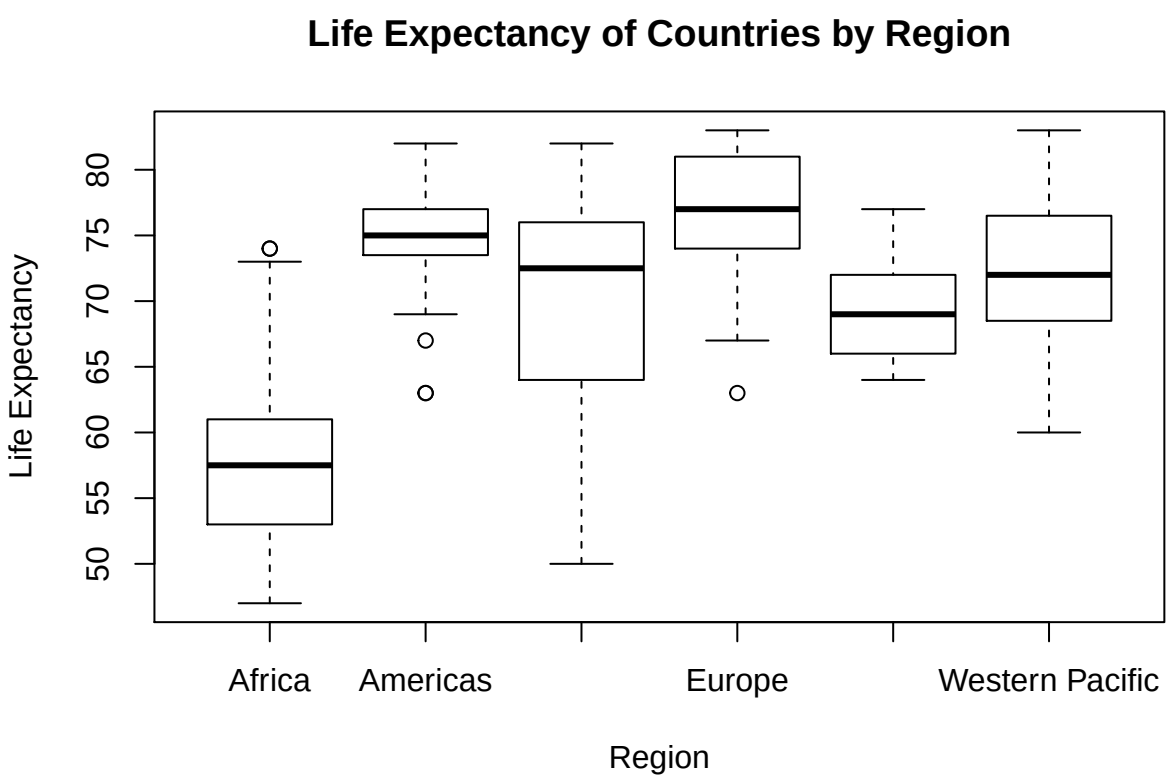
```
# Histograms
hist(WHO$CellularSubscribers)
```




```
# Boxplot sorted by Region
# Outliers are defined by first computing the difference between the first and third quartiles, or the
boxplot(WHO$LifeExpectancy ~ WHO$Region)
```




```
boxplot(WHO$LifeExpectancy ~ WHO$Region, xlab = "Region", ylab = "Life Expectancy", main = "Life Expecta
```




```
# Summary Tables
table(WHO$Region)
```

```
##
##                Africa              Americas Eastern Mediterranean
##                    46                    35                    22
##                Europe       South-East Asia       Western Pacific
##                    53                    11                    27
```

```
tapply(WHO$Over60, WHO$Region, mean)
```

```
##                Africa              Americas Eastern Mediterranean
##              5.220652             10.943714              5.620000
##                Europe       South-East Asia       Western Pacific
##             19.774906              8.769091             10.162963
```

```
tapply(WHO$LiteracyRate, WHO$Region, min)
```

```
##                Africa              Americas Eastern Mediterranean
##                    NA                    NA                    NA
##                Europe       South-East Asia       Western Pacific
##                    NA                    NA                    NA
```

```r
tapply(WHO$LiteracyRate, WHO$Region, min, na.rm=TRUE)
```

```
##                Africa              Americas Eastern Mediterranean
##                  31.1                  75.2                  63.9
##                Europe       South-East Asia       Western Pacific
##                  95.2                  56.8                  60.6
```

VIDEO 07: SAVING WITH SCRIPT FILES