```r
#Lab 2
#Kehan Wang
#661983342

#Lab2 Part1
rm(list=ls())
setwd("H:/RPI/Spring 2020/Data Analytics/Assignment 2")

EPI_data <- read.csv("2010EPI_data.csv",skip=1)
attach(EPI_data)
dim(EPI_data)
```

```
## [1] 65467   160
```

```r
#Remove null values
EPI_data <- EPI_data[1:163,]

#head(EPI_data)
#tail(EPI_data)
#summary(EPI_data)
#The above code will generate huge results, so I comment them.

#Lab2a Measures of Central Tendency
summary(EPI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.10   48.60   59.20   58.37   67.60   93.50   65304
```

```r
names(table(EPI))[which(table(EPI)==max(table(EPI)))]
```

```
## [1] "44.6" "51.3"
```

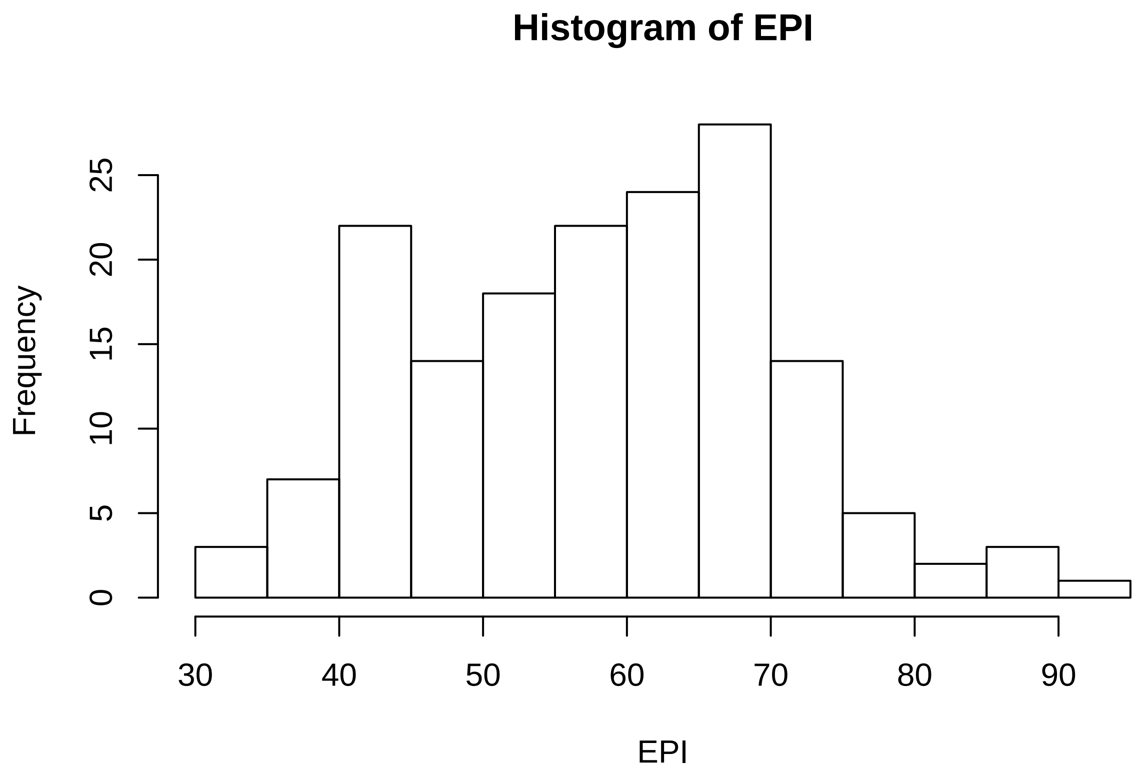
```r
# From the result we can see that mean of EPI is 58.37,
# median of EPI is 59.20, mode of EPI are 44.6 and 51.3.
summary(DALY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   32.44   60.35   53.62   73.01   91.50   65304
```

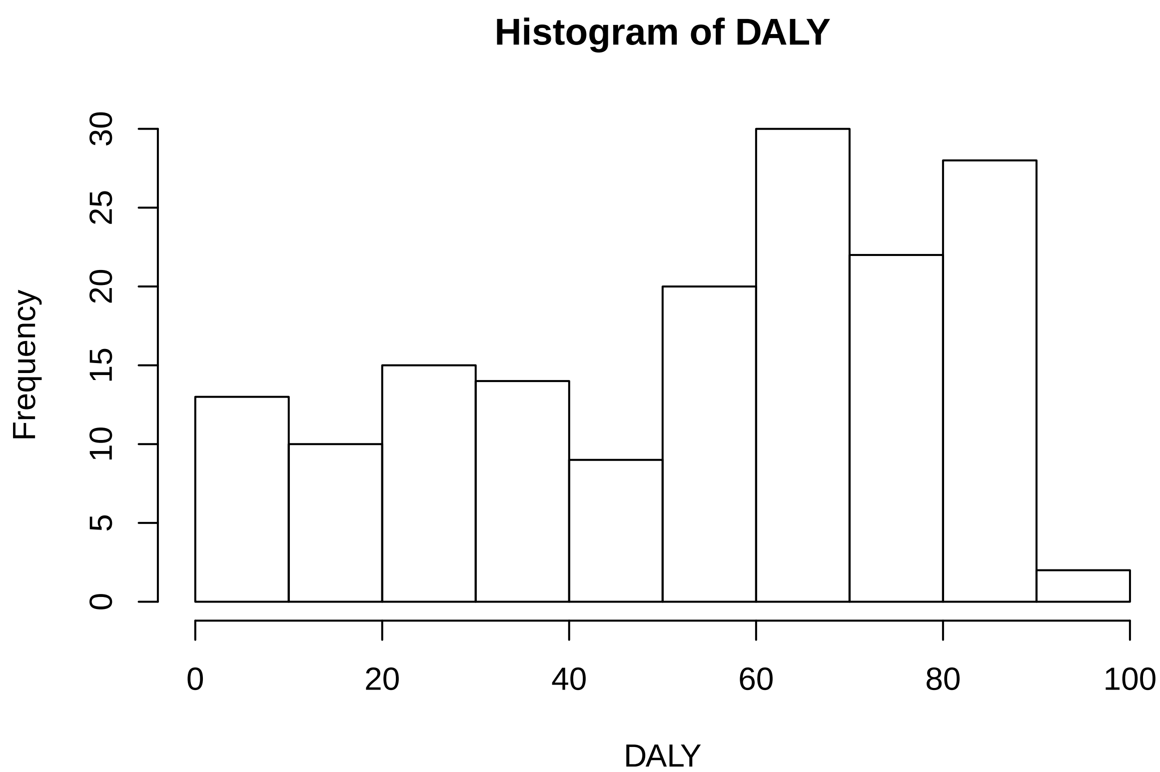
```r
names(table(EPI))[which(table(DALY)==max(table(DALY)))]
```

```
## [1] "62"
```

```r
# From the result we can see that mean of DALY is 53.62,
# median of DALY is 60.35, mode of DALY is 62.

#Lab2a Generate the Histogram for EPI and DALY variables
hist(EPI)
```


Histogram of EPI
Frequency
0
5
10
15
20
25
30
40
50
60
70
80
90
EPI

```r
hist(DALY)
```




```r
#Lab2a Dplyr exercise
#Using sample_n() function in dplyr, get 5 random data points from EPI, DALY
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2

##
## 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
```

```r
sample_n(EPI_data, 5)$EPI
```

```
## [1] 64.6 54.0 44.3 50.1 89.1
```

```r
sample_n(EPI_data, 5)$DALY
```

```
## [1] 82.81 70.31 36.49 74.45 69.04
```

```r
#Using sample_frac() function in dplyr, get 10% random data points from EPI, DALY
sample_frac(EPI_data, 0.1)$EPI
```

```
##  [1] 66.4 65.7 62.2 73.1 69.2 51.3 67.8 56.3 50.8 72.5 78.1 73.2 63.5 62.5 55.9
## [16] 59.1
```

```r
sample_frac(EPI_data, 0.1)$DALY
```

```
##  [1] 54.28 29.17 73.01 14.03 27.06 86.86 61.32 44.18 52.74 27.75 60.35 61.32
## [13] 57.61 63.34 89.10  4.43
```

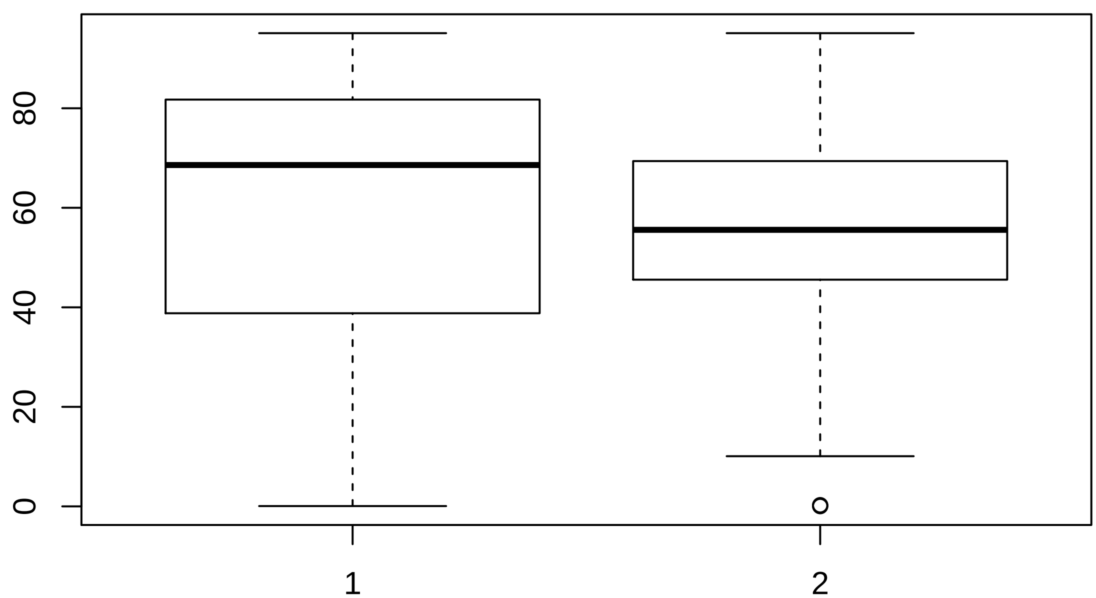
```r
#Use the arrange() and desc() functions to arrange values in the descending order in the EPI and DALY
new_decs_EPI <- arrange(EPI_data, desc(EPI))$EPI
new_decs_DALY <- arrange(EPI_data, desc(DALY))$DALY

#Using the mutate() function, create new columns: double_EPI and double_DALY where multiplying the valu
#mutate(EPI_data, double_EPI = EPI*2)
#mutate(EPI_data, double_DALY = DALY*2)
# The above code will generate huge volumes of results, so I commented them.

#Using the summarise() function along with the mean() function to find the mean for EPI and DALY
summarise(EPI_data, avg_EPI = mean(EPI, na.rm = TRUE))
```

```
##    avg_EPI
## 1 58.37055
```

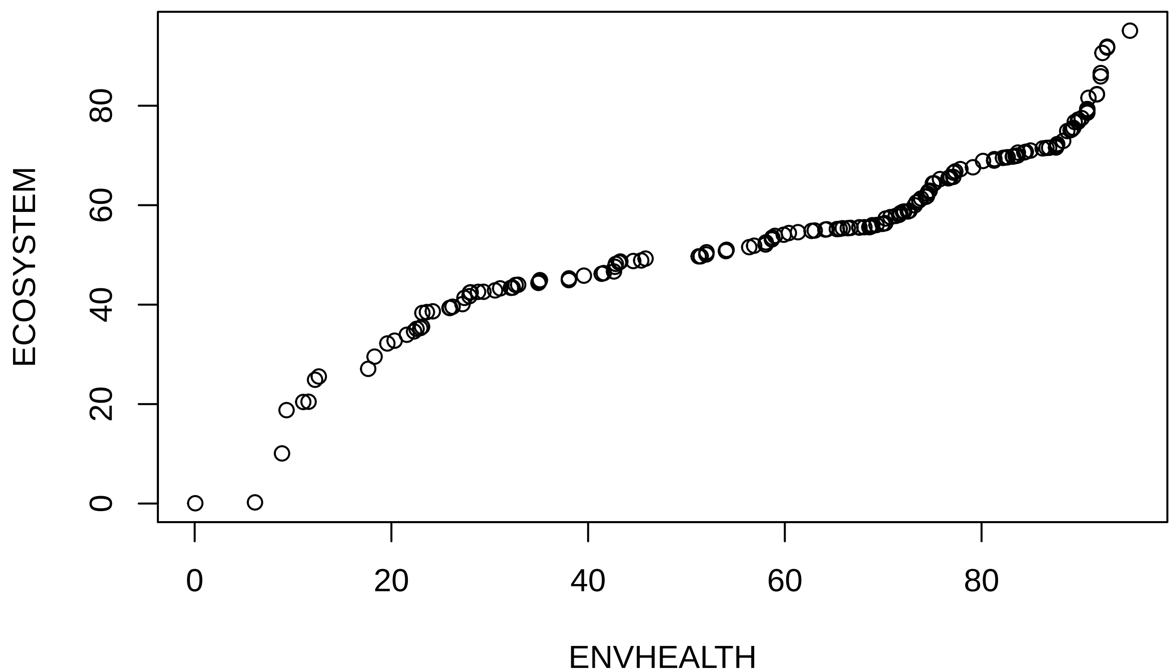
```r
summarise(EPI_data, avg_DALY = mean(DALY, na.rm = TRUE))
```

```
##   avg_DALY
## 1 53.62466
```

```r
boxplot(ENVHEALTH,ECOSYSTEM)
```



```r
qqplot(ENVHEALTH,ECOSYSTEM)
```




```r
#Lab2b Regression Exercise
#I choose Europe Region
EPI_data_new <- subset(EPI_data, EPI_regions == "Europe")

#I limited EPI_regions, so I could remove it from the dataset.
#For Cuntry and GEO_subregion, they have high correlation with EPI_regions,
#so I removed them as well.
#Similarily, code and ISO3V10 are useless, removed.
EPI_data_new <- EPI_data_new[, 6:160]
#convert EPI1 dataset into numeric
EPI1 <- sapply(EPI_data_new,as.numeric)
#make correlation table
corr <- round(cor(EPI1), 2)
```

```
## Warning in cor(EPI1): the standard deviation is zero
```

```r
corr <- data.frame(corr)
corr$EPI
```

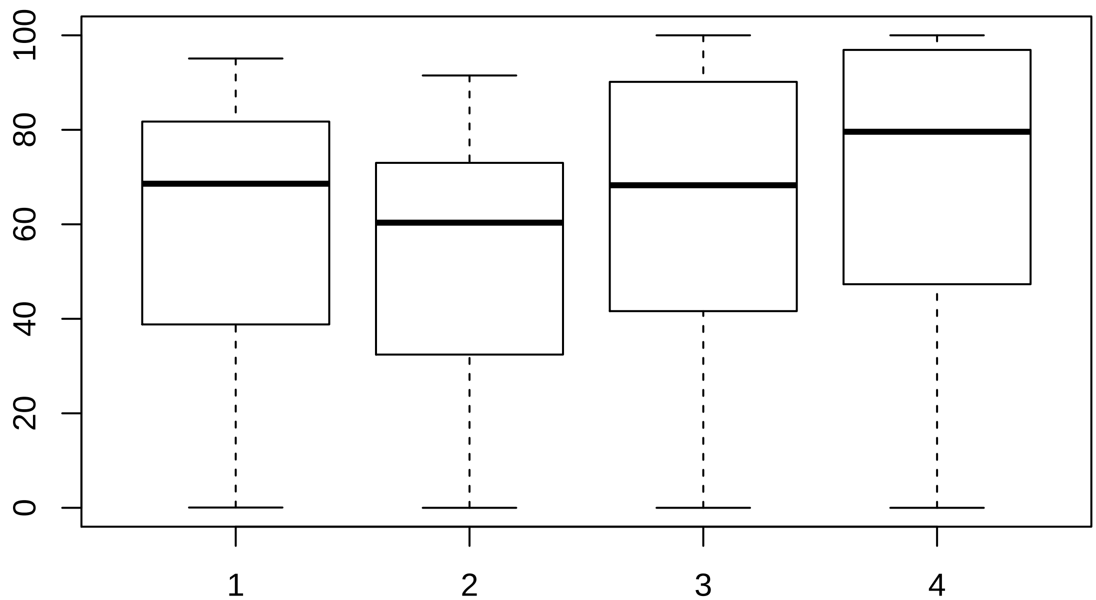
```
##   [1]  0.36  0.35  0.28 -0.39  0.21    NA    NA -0.14  1.00  0.50  0.92  0.40
##  [13]  0.48  0.22  0.12  0.43  0.29  0.17 -0.16  0.03  0.88  0.40  0.20 -0.06
##  [25]  0.22 -0.22  0.28 -0.21  0.25 -0.04 -0.13 -0.08 -0.11  0.11  0.12 -0.33
##  [37]  0.12  0.26 -0.01 -0.01 -0.17 -0.18 -0.30 -0.17 -0.02 -0.07  0.13  0.55
##  [49]    NA  0.61  0.78    NA -0.36  0.20 -0.06  0.22 -0.22  0.10 -0.45  0.09
##  [61] -0.11  0.11  0.12 -0.07  0.10  0.13 -0.30 -0.35  0.26 -0.01 -0.02  0.19
```

```
##  [73]  0.11  0.23  0.06 -0.48  0.48  0.13 -0.52    NA -0.59 -0.64    NA -0.40
##  [85]  0.20  0.22  0.10 -0.45  0.08 -0.25  0.04  0.13 -0.30 -0.39  0.26 -0.01
##  [97] -0.02  0.19  0.11  0.23  0.06 -0.30  0.07  0.13 -0.59 -0.39 -0.78 -0.40
## [109] -0.45  0.08 -0.25  0.04  0.13 -0.39 -0.02 -0.30 -0.59 -0.39 -0.78    NA
## [121]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [133]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [145]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
```

```
# It can be seen that the biggest positive coefficient
#of EPI is ECOSYSTEM, which is 0.92.
# To confirm, I make a regression of the first 20 variables,
#since it includes ECOSYSTEM as well as the variables has meaning
#from their name, such as BIODIVERSITY, Desert and etc.
EPI1 <- as.data.frame(EPI1)
EPI2 <- EPI1[, 1:20]
fit <- lm(EPI ~ ., data = EPI2)
summary(fit)
```

```
##
## Call:
## lm(formula = EPI ~ ., data = EPI2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.040925 -0.014322 -0.000236  0.014043  0.059520
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.069e-01  1.200e+00  -0.089    0.930
## GDPCAP07               -8.365e-04  6.430e-04  -1.301    0.218
## Population07           -1.937e-04  1.980e-04  -0.978    0.347
## Landarea               -7.129e-08  5.893e-08  -1.210    0.250
## PopulationDensity07    -9.181e-06  1.885e-04  -0.049    0.962
## Landlock               -3.649e-02  3.015e-02  -1.210    0.249
## No_surface_water               NA         NA      NA       NA
## Desert                         NA         NA      NA       NA
## High_Population_Density -3.065e-02  3.604e-02  -0.850    0.412
## ENVHEALTH               5.994e-01  3.049e+00   0.197    0.847
## ECOSYSTEM               4.993e-01  6.496e-04 768.633   <2e-16 ***
## DALY                   -4.619e-02  1.524e+00  -0.030    0.976
## AIR_H                  -2.485e-02  7.624e-01  -0.033    0.975
## WATER_H                -2.437e-02  7.627e-01  -0.032    0.975
## AIR_E                   4.669e-04  1.273e-03   0.367    0.720
## WATER_E                -1.511e-04  1.318e-03  -0.115    0.911
## BIODIVERSITY            1.540e-04  4.602e-04   0.335    0.744
## FORESTRY               -7.302e-04  1.192e-02  -0.061    0.952
## FISHERIES              -4.307e-04  3.056e-04  -1.409    0.184
## AGRICULTURE            -6.696e-05  1.053e-03  -0.064    0.950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03423 on 12 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.065e+05 on 17 and 12 DF,  p-value: < 2.2e-16
```

```r
#From the model we can see that P value of ECOSYSTEM  is very small,
#about 0, so the effect of ECOSYSTEM on EPI is very significant.
#Also,it has the positive coefficient and the value is pretty considerable.
#Thus, the single most important factor in increasing the EPI
#in Europe is ECOSYSTEM.

#Linear and least-squares
boxplot(ENVHEALTH,DALY,AIR_H,WATER_H)
```




```r
lmENVH<-lm(ENVHEALTH~DALY+AIR_H+WATER_H)
lmENVH
```

```
##
## Call:
## lm(formula = ENVHEALTH ~ DALY + AIR_H + WATER_H)
##
## Coefficients:
## (Intercept)         DALY        AIR_H      WATER_H
##  -1.458e-05    5.000e-01    2.500e-01    2.500e-01
```

```r
summary(lmENVH)
```

```
##
## Call:
## lm(formula = ENVHEALTH ~ DALY + AIR_H + WATER_H)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0073210 -0.0027069 -0.0000915  0.0022285  0.0053404
##
## Coefficients:
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -1.458e-05  6.520e-04    -0.022    0.982
## DALY         5.000e-01  1.988e-05 25147.716   <2e-16 ***
## AIR_H        2.500e-01  1.276e-05 19593.273   <2e-16 ***
## WATER_H      2.500e-01  1.816e-05 13764.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003015 on 159 degrees of freedom
##   (65304 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.77e+09 on 3 and 159 DF,  p-value: < 2.2e-16
```

```r
cENVH<-coef(lmENVH)

#Predict
DALYNEW<-c(seq(5,95,5))
AIR_HNEW<-c(seq(5,95,5))
WATER_HNEW<-c(seq(5,95,5))
NEW<-data.frame(DALYNEW,AIR_HNEW,WATER_HNEW)
pENV<- predict(lmENVH,NEW,interval="prediction")
```

```
## Warning: 'newdata' had 19 rows but variables found have 65467 rows
```

```r
cENV<- predict(lmENVH,NEW,interval="confidence")
```

```
## Warning: 'newdata' had 19 rows but variables found have 65467 rows
```

```r
#Repeat for
#AIR_E
corr$AIR_E
```

```
##   [1]  0.01  0.13 -0.03 -0.06  0.18    NA    NA -0.53  0.12 -0.30  0.27 -0.43
##  [13]  0.27 -0.25  1.00  0.64  0.10 -0.14 -0.28  0.21  0.07 -0.43 -0.34 -0.18
##  [25] -0.10  0.08 -0.16 -0.48  0.86  0.85  0.67  0.48  0.55 -0.39  0.57 -0.33
##  [37] -0.10 -0.01  0.36 -0.26  0.14 -0.22 -0.37  0.07 -0.36  0.26  0.10 -0.22
##  [49]    NA  0.15  0.30    NA  0.46 -0.34 -0.18 -0.10  0.08 -0.32 -0.32 -0.17
##  [61]  0.55 -0.39  0.57 -0.65 -0.66 -0.69 -0.42 -0.39 -0.01 -0.26  0.10 -0.20
##  [73] -0.10  0.12  0.01 -0.45 -0.14  0.10  0.22    NA -0.11 -0.43    NA  0.43
##  [85] -0.34 -0.10 -0.32 -0.33 -0.48 -0.86 -0.85 -0.67 -0.42 -0.45 -0.01 -0.26
##  [97]  0.08 -0.20 -0.10  0.13  0.01 -0.61 -0.26  0.10  0.21  0.21 -0.23  0.43
## [109] -0.33 -0.48 -0.86 -0.85 -0.67 -0.45  0.08 -0.61  0.21  0.21 -0.23    NA
## [121]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [133]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [145]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
```

```r
EPI3 <- EPI1[, 10:30]
fit <- lm(AIR_E ~ ., data = EPI3)
summary(fit)
```

```
##
## Call:
## lm(formula = AIR_E ~ ., data = EPI3)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65891 -0.50652  0.03048  0.48799  2.45339
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.393e+01  1.009e+02  -0.931 0.376255
## ENVHEALTH     -4.646e+02  2.020e+02  -2.300 0.047022 *
## ECOSYSTEM      4.611e-01  3.029e-01   1.522 0.162319
## DALY           2.041e+02  2.439e+02   0.837 0.424242
## AIR_H          1.161e+02  5.050e+01   2.298 0.047134 *
## WATER_H       -1.173e+03  3.944e+02  -2.974 0.015605 *
## WATER_E        1.074e-01  8.670e-02   1.239 0.246684
## BIODIVERSITY  -4.204e-02  5.485e-02  -0.767 0.462982
## FORESTRY       9.459e-01  8.974e-01   1.054 0.319332
## FISHERIES      2.429e-02  2.130e-02   1.140 0.283578
## AGRICULTURE   -1.447e-01  5.711e-02  -2.534 0.032013 *
## CLIMATE       -4.170e-01  2.691e-01  -1.550 0.155633
## DALY_pt        2.780e+01  2.257e+02   0.123 0.904684
## ACSAT_pt       6.444e+02  1.977e+02   3.260 0.009843 **
## ACSAT_pt_imp  -7.881e+00  1.965e+00  -4.010 0.003064 **
## WATSUP_pt      6.451e+02  1.977e+02   3.263 0.009787 **
## WATSUP_pt_imp  9.422e+00  3.194e+00   2.950 0.016219 *
## INDOOR_pt      8.754e-02  1.244e-01   0.704 0.499383
## PM10_pt       -7.014e-02  1.342e-02  -5.227 0.000544 ***
## SO2_pt         3.387e-01  5.070e-02   6.680 9.06e-05 ***
## NOX_pt         1.182e-01  1.059e-01   1.115 0.293569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.89 on 9 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9737
## F-statistic: 54.69 on 20 and 9 DF,  p-value: 4.587e-07
```

```r
#Similarly, the single most important factor in increasing the AIR_E
#in Europe is SO2_pt.

#CLIMATE
corr$CLIMATE
```

```
##   [1]  0.09  0.14  0.26 -0.33  0.00    NA    NA -0.07  0.88  0.21  0.91  0.14
##  [13]  0.43 -0.08  0.07  0.25  0.04  0.16 -0.15  0.13  1.00  0.14 -0.08  0.07
##  [25] -0.08 -0.05  0.11 -0.25  0.13 -0.11 -0.08  0.00 -0.18  0.13  0.06 -0.17
##  [37]  0.22  0.03  0.19 -0.04 -0.16 -0.16 -0.09 -0.15 -0.08 -0.02  0.10  0.79
##  [49]    NA  0.58  0.77    NA -0.09 -0.08  0.07 -0.08 -0.05  0.00 -0.41  0.03
##  [61] -0.18  0.13  0.06  0.06  0.14  0.11 -0.13 -0.13  0.03 -0.04  0.14  0.28
##  [73]  0.24  0.05  0.10 -0.33  0.48  0.10 -0.78    NA -0.61 -0.63    NA -0.14
##  [85] -0.08 -0.08  0.00 -0.42  0.00 -0.13  0.11  0.08 -0.13 -0.17  0.03 -0.04
##  [97]  0.14  0.28  0.24  0.06  0.10 -0.21  0.02  0.10 -0.78 -0.39 -0.78 -0.14
## [109] -0.42  0.00 -0.13  0.11  0.08 -0.17  0.14 -0.21 -0.78 -0.39 -0.78    NA
## [121]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [133]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## [145]    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
```

```r
fit <- lm(CLIMATE ~ ., data = EPI3)
summary(fit)
```

```
##
## Call:
## lm(formula = CLIMATE ~ ., data = EPI3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91650 -0.89815  0.09395  0.85595  2.24838
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      73.11071  113.70282   0.643 0.536261
## ENVHEALTH      -224.57396  269.96502  -0.832 0.427006
## ECOSYSTEM         1.11477    0.04125  27.027 6.29e-10 ***
## DALY            414.52561  241.99475   1.713 0.120873
## AIR_H            56.04943   67.47596   0.831 0.427652
## WATER_H        -520.58830  586.01204  -0.888 0.397469
## AIR_E            -0.50512    0.32595  -1.550 0.155633
## WATER_E          -0.15235    0.08989  -1.695 0.124344
## BIODIVERSITY     -0.16516    0.02918  -5.660 0.000309 ***
## FORESTRY         -0.34111    1.04070  -0.328 0.750584
## FISHERIES        -0.02494    0.02366  -1.054 0.319303
## AGRICULTURE      -0.15777    0.06328  -2.493 0.034236 *
## DALY_pt        -302.30582  227.28846  -1.330 0.216220
## ACSAT_pt        288.27711  306.59006   0.940 0.371629
## ACSAT_pt_imp     -5.77679    3.05470  -1.891 0.091173 .
## WATSUP_pt       288.62844  306.74338   0.941 0.371301
## WATSUP_pt_imp     8.69967    3.98664   2.182 0.056969 .
## INDOOR_pt         0.02159    0.14044   0.154 0.881189
## PM10_pt          -0.04549    0.02550  -1.784 0.108128
## SO2_pt            0.18987    0.12060   1.574 0.149867
## NOX_pt           -0.12632    0.11703  -1.079 0.308500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.081 on 9 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9804
## F-statistic: 73.41 on 20 and 9 DF,  p-value: 1.25e-07
```

```r
#Similarly, the single most important factor in increasing the CLIMATE
#in Europe is ECOSYSTEM.

#Exercise 1: Regression
Reg <- read.csv("dataset_multipleRegression.csv")
head(Reg)
```

```
##   YEAR ROLL UNEM HGRAD  INC
## 1    1 5501  8.1  9552 1923
## 2    2 5945  7.0  9680 1961
## 3    3 6629  7.3  9731 1979
## 4    4 7556  7.5 11666 2030
## 5    5 8716  7.0 14675 2112
```

```
## 6    6 9369  6.4 15265 2192
```

```r
dim(Reg)
```

```
## [1] 29  5
```

```r
fit1 <- lm(ROLL ~ UNEM + HGRAD, data = Reg)
new1 <- data.frame(UNEM = 7.0, HGRAD = 90000)
ROLL1 <- predict(fit1, newdata = new1)
ROLL1
```

```
##        1
## 81437.04
```

```r
fit2 <- lm(ROLL ~ UNEM + HGRAD + INC, data = Reg)
new2 <- data.frame(UNEM = 7.0, HGRAD = 90000, INC = 25000)
ROLL2 <- predict(fit2, newdata = new2)
ROLL2
```

```
##        1
## 137452.6
```

```r
#Exercise 2: Classification
ab <- read.csv("abalone.csv")
head(ab)
```

```
##   Sex Length Diameter Height Whole.weight Shucked.weight Viscera.weight
## 1   M  0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2   M  0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3   F  0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4   M  0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5   I  0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6   I  0.425    0.300  0.095       0.3515         0.1410         0.0775
##   Shell.weight Rings
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

```r
dim(ab)
```

```
## [1] 4177    9
```

```r
ab$Rings <- as.numeric(ab$Rings)
ab$Rings <- cut(ab$Rings, br=c(-1,8,11,35), labels = c("young", 'adult', 'old'))
ab$Rings <- as.factor(ab$Rings)
ab$Sex <- NULL
ab[1:7] <- scale(ab[1:7])

set.seed(1)
ind <- sample(2, nrow(ab), replace=TRUE, prob=c(0.7, 0.3))
KNNtrain <- ab[ind==1,]
KNNtest <- ab[ind==2,]
k = sqrt(nrow(KNNtrain))

library(class)
KNNpred <- knn(train = KNNtrain[1:7], test = KNNtest[1:7], cl = KNNtrain$Rings, k = k)
```

```
#Result
KNNpred
```

```
##   [1] young young old   young young young young young adult adult adult adult
##  [13] young young young young young young old   young young adult adult adult
##  [25] adult old   adult adult adult old   adult adult adult adult young young
##  [37] old   young young young young adult young young adult old   old   old
##  [49] adult adult young young old   adult adult adult adult adult adult young
##  [61] old   old   young old   young old   old   young adult young adult young
##  [73] young young old   adult young young young adult young adult young young
##  [85] old   adult old   adult young young young young adult young old   young
##  [97] old   adult adult adult adult adult old   adult adult adult old   adult
## [109] adult adult adult young adult young adult young adult adult adult adult
## [121] adult adult adult adult adult old   old   old   adult young old   old
## [133] old   young young old   adult young young adult adult adult adult adult
## [145] adult adult old   adult adult young young young young young young adult
## [157] adult young young young old   adult young young old   adult adult adult
## [169] young adult adult old   adult adult young adult adult adult young young
## [181] old   young young old   adult young young young adult young adult young
## [193] old   young young old   old   old   adult old   old   young old   adult
## [205] young young young young adult young young young young young young old
## [217] adult adult old   old   old   adult adult young adult adult old   young
## [229] adult old   adult adult old   adult old   old   old   young old   old
## [241] adult adult young old   young young young young young young young young
## [253] young young young adult young young adult adult adult adult adult adult
## [265] adult adult adult adult adult adult adult adult adult young young young
## [277] young young young young young young young young young young young young
## [289] young young adult young young young young young young adult adult young
## [301] adult adult adult adult adult adult adult adult adult adult adult adult
## [313] adult adult old   old   adult young young young young young young young
## [325] young young adult young young adult young young young adult adult adult
## [337] young adult adult adult adult adult adult adult adult adult adult adult
## [349] adult adult adult adult adult adult adult adult adult adult adult adult
## [361] adult adult adult adult adult old   young young young young young young
## [373] young young young young young young young young young young young young
## [385] young adult young adult adult old   adult adult young adult adult adult
## [397] adult adult adult adult adult adult adult adult adult adult adult adult
## [409] adult adult adult adult adult adult adult adult old   adult adult adult
## [421] adult adult adult adult adult adult young young young young young young
## [433] young young young young young young adult adult adult adult adult adult
## [445] adult adult adult adult adult adult adult adult adult old   adult adult
## [457] adult adult old   adult adult adult young young young young young young
## [469] young young young young young young young young adult young young young
## [481] young young adult young young adult old   adult adult adult adult adult
## [493] young adult adult adult adult adult adult adult adult adult adult adult
## [505] adult adult adult adult adult adult adult old   adult adult adult adult
## [517] adult adult adult adult adult adult adult adult adult adult old   old
## [529] adult adult adult adult adult adult adult adult adult young young young
## [541] adult adult adult adult adult adult adult adult adult adult young young
## [553] young young young young young young young young adult young adult adult
## [565] adult young adult adult adult adult adult young adult adult adult adult
## [577] adult adult adult adult adult adult adult adult adult adult adult adult
## [589] adult adult adult adult adult adult adult old   adult adult adult old
## [601] adult young young young young young young young young young young adult
```

```
##  [613] adult young young adult young young old   young young young adult adult
##  [625] adult adult adult adult adult old   adult young adult old   old   young
##  [637] young young young young young young adult young young old   adult young
##  [649] old   adult old   adult young old   old   adult old   young old   adult
##  [661] young young old   young old   old   old   young young adult young young
##  [673] old   adult adult old   old   adult adult adult old   old   adult adult
##  [685] adult young adult young adult young adult adult old   adult old   adult
##  [697] old   adult adult young old   adult adult adult young adult adult adult
##  [709] adult adult old   adult old   old   young young adult young young old
##  [721] old   young adult adult young young young young adult old   adult young
##  [733] young adult old   old   old   young young young young young young young
##  [745] young young young old   adult adult old   old   old   young young young
##  [757] young adult young adult adult adult adult adult adult old   adult old
##  [769] adult young young young young young young young young young young adult
##  [781] young adult adult adult adult adult adult adult adult adult adult adult
##  [793] young young young young young young young adult adult adult adult adult
##  [805] adult adult adult adult adult adult adult adult adult young adult young
##  [817] young young young young young young old   adult young adult adult adult
##  [829] adult adult adult adult adult adult adult adult adult adult adult adult
##  [841] adult young young young young adult adult adult adult adult adult adult
##  [853] adult adult adult young young young young young young young young adult
##  [865] adult adult adult adult adult adult adult adult adult adult adult adult
##  [877] adult adult adult adult adult adult adult adult adult adult adult adult
##  [889] adult young young adult adult adult adult adult adult adult old   young
##  [901] young young young young young young old   adult adult adult adult old
##  [913] adult adult adult adult old   adult adult adult adult adult adult young
##  [925] young adult young adult adult young young young young adult adult adult
##  [937] adult adult young adult young young young adult young young adult adult
##  [949] old   adult young adult young young young adult adult old   old   young
##  [961] old   adult adult old   young young adult adult adult adult young young
##  [973] adult young adult adult young old   adult adult adult adult adult adult
##  [985] adult old   old   adult young adult adult young young young old   adult
##  [997] young old   young old   young young young young young young young young
## [1009] young adult young young young adult old   adult young adult young young
## [1021] young adult adult adult adult adult adult young young young young young
## [1033] young adult adult adult adult adult adult adult adult adult adult adult
## [1045] young young young young adult young young adult young young adult adult
## [1057] adult adult adult adult adult adult adult young young young young young
## [1069] young young young young young young young young young adult adult adult
## [1081] adult adult adult adult adult adult adult young young young young adult
## [1093] adult adult adult adult adult adult young young young young young young
## [1105] young young adult adult adult adult adult adult adult adult adult adult
## [1117] adult adult adult adult adult adult adult adult adult adult adult adult
## [1129] adult adult adult adult adult adult adult young adult young adult adult
## [1141] young adult adult adult adult young adult adult adult adult adult adult
## [1153] adult adult adult adult adult adult adult adult adult old   old   young
## [1165] young young young young adult adult adult old   adult adult young adult
## [1177] adult old   young old   adult adult old   young adult young young adult
## [1189] old   adult adult adult adult adult old   young old   young old   old
## [1201] young adult adult adult young young young adult young young adult adult
## [1213] adult adult young young adult young young young young young young adult
## [1225] adult adult adult adult adult adult adult adult adult young young young
## [1237] young adult adult adult adult adult adult adult adult adult adult adult
## [1249] adult adult adult young young adult adult old   adult adult old   young
```

```
## [1261] young young adult young young
## Levels: young adult old
```

```r
table(KNNpred)
```

```
## KNNpred
## young adult   old
##   458   671   136
```

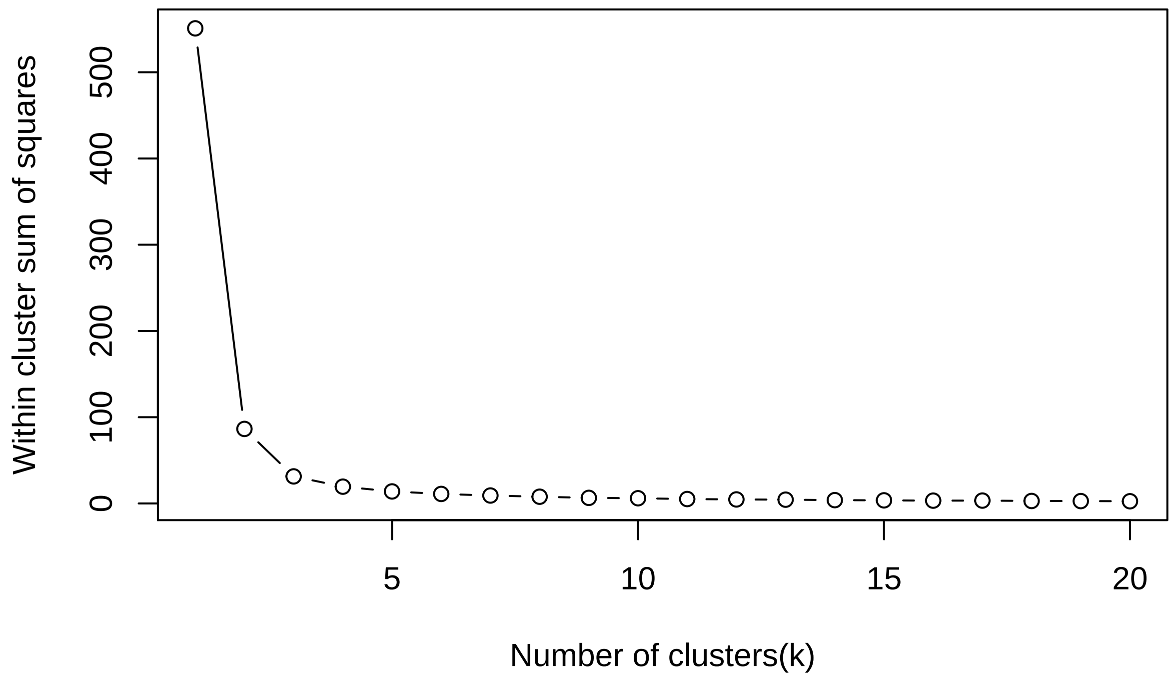
```r
table(KNNtest[,8], KNNpred, dnn = list('Actual', 'Predict'))
```

```
##         Predict
## Actual   young adult old
##   young   346    96   2
##   adult    95   400  40
##   old      17   175  94
```

```r
#Exercise 3: Clustering
ir <- iris[, -5]
head(ir)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

```r
#Method
#k.max <- 1000
#wss<- sapply(1:k.max,function(k){kmeans(ir,k)$tot.withinss})
#The above codes generate error.
#I cannot make k to 1000 because we only have 150 observations #in dataset iris.

#So I limit k to 20
k.max <- 20
wss<- sapply(1:k.max,function(k){kmeans(iris[,3:4],k,nstart = 20,iter.max = 20)$tot.withinss})
plot(1:k.max,wss, type= "b", xlab = "Number of clusters(k)", ylab = "Within cluster sum of squares")
```




```
#From the plot I can infer that when k =3,
#within cluster sum of squares becomes vary small
#and does not change anymore, so I choose k =3.

#Then I try maximum iteration equals 1000
set.seed(1)
icluster <- kmeans(ir,3, iter.max = 1000)
table(iris[,5], icluster$cluster, dnn = list('Acutual','Predict') )
```

```
##               Predict
## Acutual         1  2  3
##   setosa        0  0 50
##   versicolor 48  2  0
##   virginica  14 36  0
```

```
# In the table we can see that most of the observations
# have been clustered correctly.
# The model predict 50 setosa and actual has 50 setosa
# and all of them are predicted accurately.
# The model predict 50 versicolor just as Acutual.
#However 2 of the versicolor have been put in the cluster
#with most of them are virginica
# Similarly, 14 of the verginica have been put in cluster 1
#which mostly has versicolor.
```