Data Analytics: Lab 2

Brendan Donnelly

October 2, 2020

Lab 2 part 1

Measures of central tendancy for EPI,DALY vars

```
library(ggplot2)
EPI<-read.csv("/Users/donneb/Documents/DataAnalytics/EPI_data.csv")
summary(EPI$EPI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.10   48.60   59.20   58.37   67.60   93.50      68
```

```
fivenum(EPI$EPI, na.rm = T)
```

```
## [1] 32.1 48.6 59.2 67.6 93.5
```

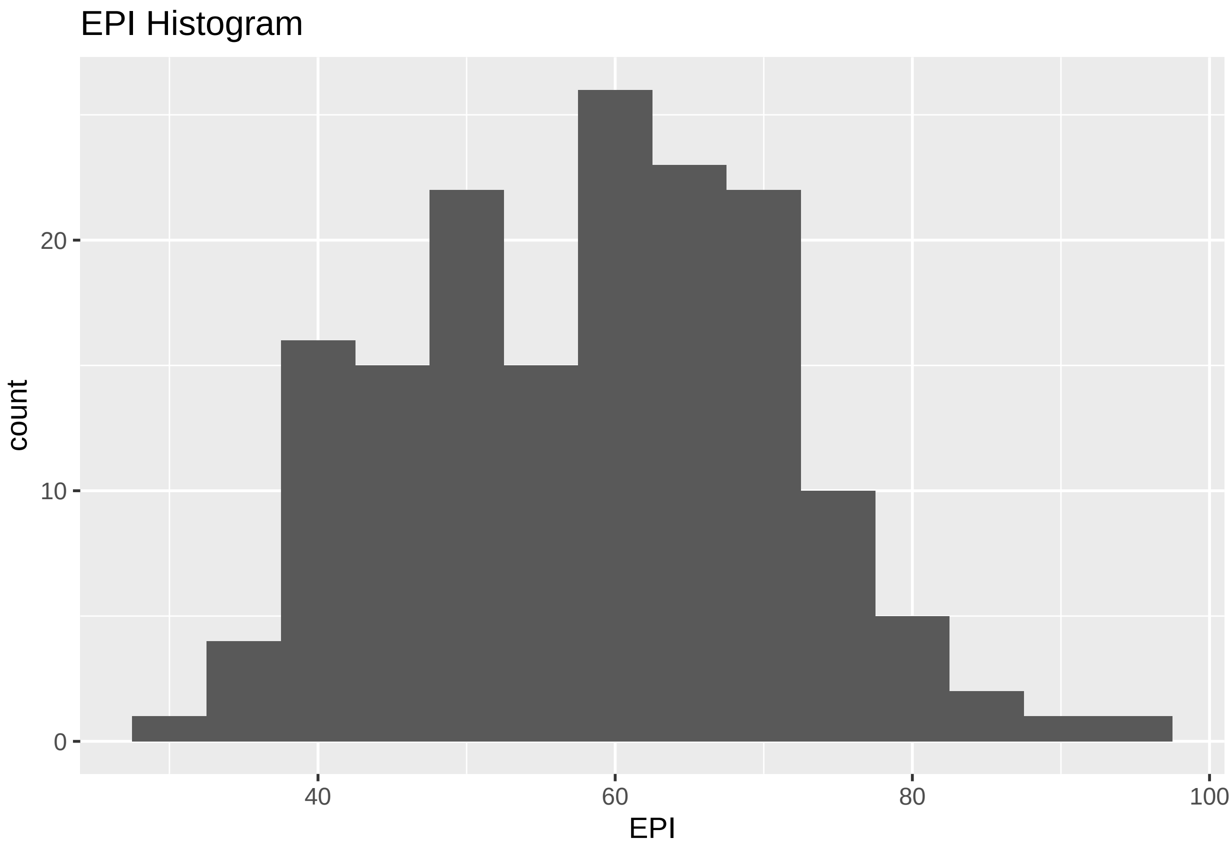
```
summary(EPI$DALY)#stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   37.19   60.35   53.94   71.97   91.50      39
```

```
fivenum(EPI$DALY, na.rm = T)
```

```
## [1]  0.000 36.955 60.350 72.320 91.500
```

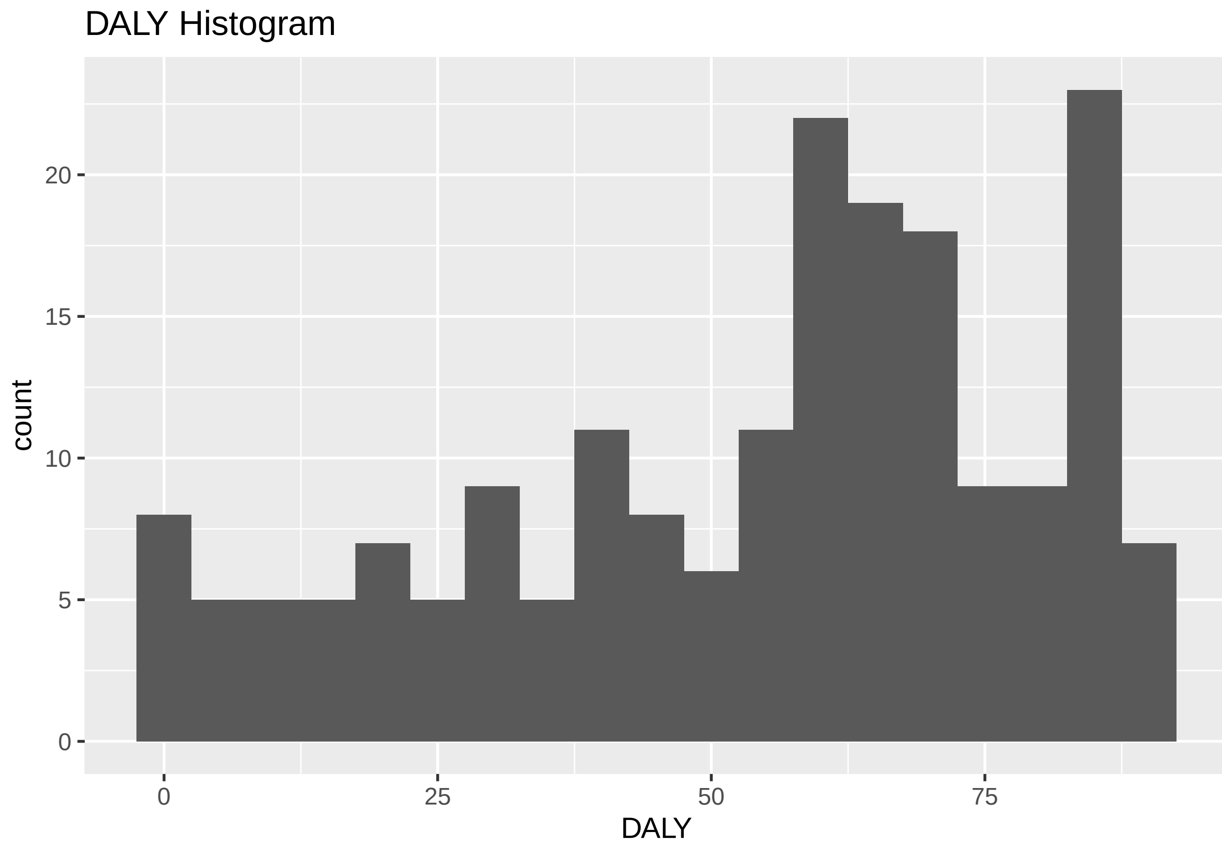
```
# Histograms of both vars
histEPI<- ggplot(EPI, aes(x=EPI)) + geom_histogram(binwidth = 5, na.rm=TRUE)
histEPI +labs(title = "EPI Histogram")
```




```r
histDALY<- ggplot(EPI, aes(x=DALY)) + geom_histogram(binwidth = 5, na.rm=TRUE, title = "Daly Histogram")
```

```
## Warning: Ignoring unknown parameters: title
```

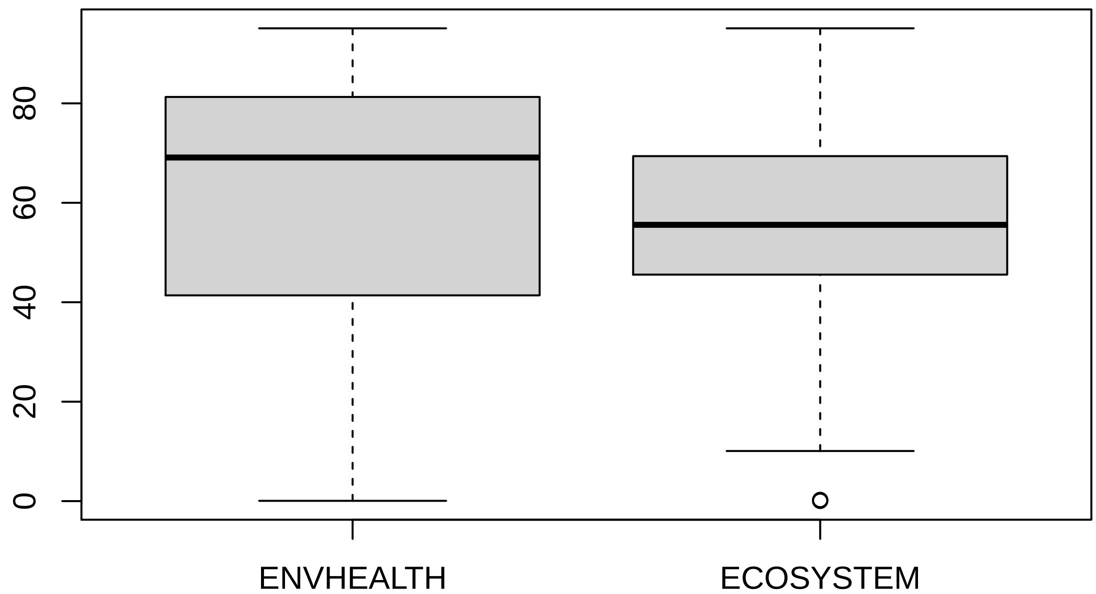
```r
histDALY +labs(title = "DALY Histogram")
```



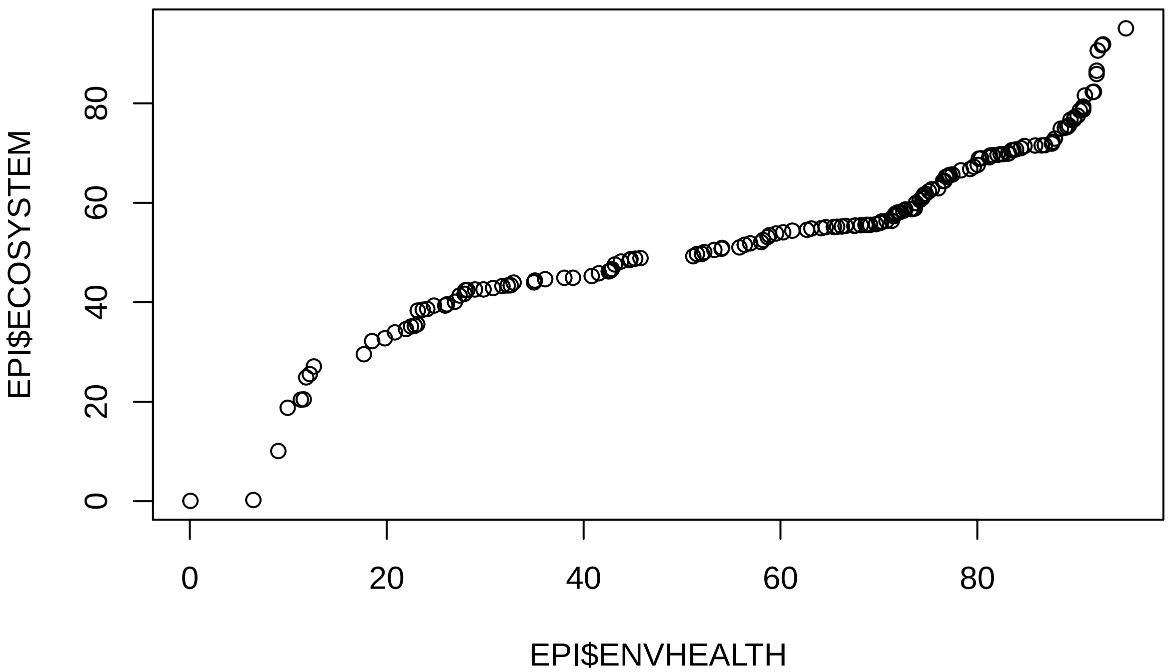
Comparing ENVHEALTH and ECOSYSTEM's Relationship

boxplot,normal distribution plot comparing

```
#boxplot
boxplot(EPI$ENVHEALTH, EPI$ECOSYSTEM, names = c('ENVHEALTH','ECOSYSTEM'))
```

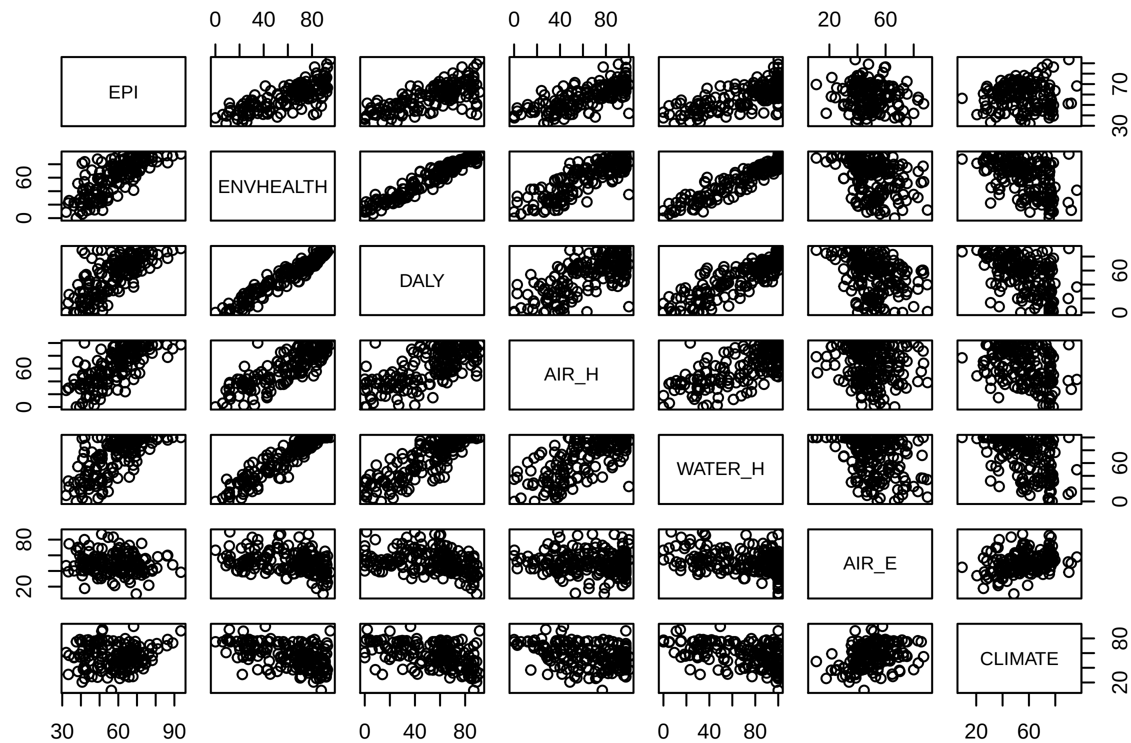



```
#normal dist plots
qqplot(EPI$ENVHEALTH, EPI$ECOSYSTEM)
```



Determining Most Important Factor in EPI Regression

```
#getting a feel for relationships
plot(EPI[c(14,15,17,18,19,20,26)])
```

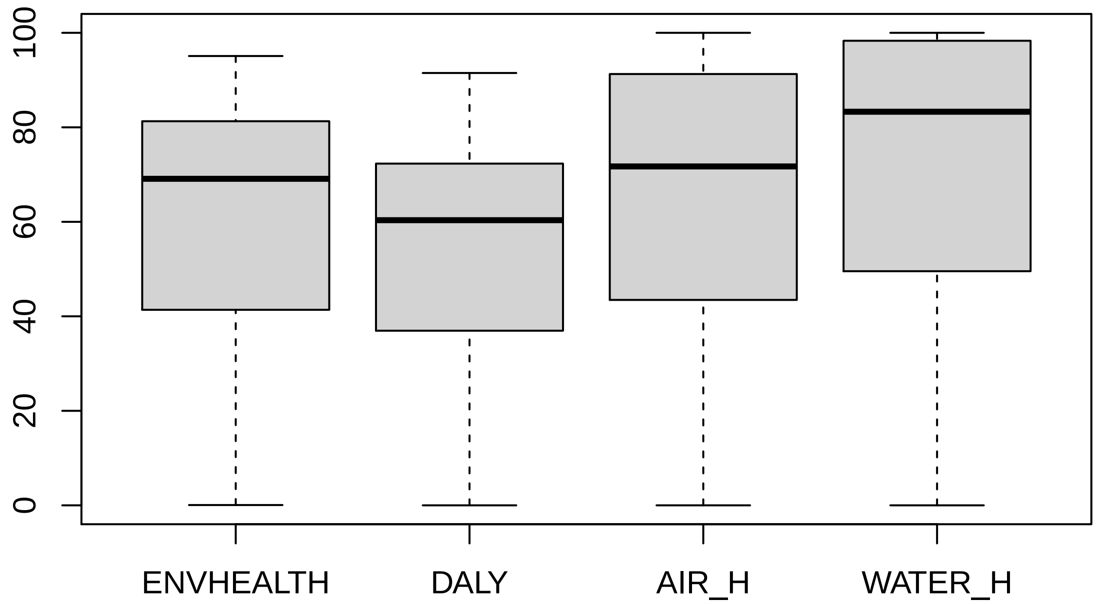


Linear and Least Squares ENVHEALTH

```r
ENVHEALTH <- EPI$ENVHEALTH
DALY <- EPI$DALY
AIR_H<- EPI$AIR_H
WATER_H<- EPI$WATER_H

# spread of all linear regression vars, w/ ENVHEALTh 1

boxplot(ENVHEALTH,DALY,AIR_H,WATER_H, names = c("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
##  -2.673e-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
## -0.0072734 -0.0027299  0.0001145  0.0021423  0.0055205
##
## Coefficients:
##               Estimate Std. Error   t value Pr(>|t|)
## (Intercept) -2.673e-05  6.377e-04    -0.042    0.967
## DALY         5.000e-01  1.922e-05 26020.669   <2e-16 ***
## AIR_H        2.500e-01  1.273e-05 19645.297   <2e-16 ***
## WATER_H      2.500e-01  1.751e-05 14279.903   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003097 on 178 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.983e+09 on 3 and 178 DF,  p-value: < 2.2e-16
```

```r
cENVH<-coef(lmENVH)
cENVH
```

```
##   (Intercept)          DALY         AIR_H       WATER_H
## -2.673362e-05  5.000401e-01  2.499968e-01  2.499781e-01
```

```r
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 = "pred")
```

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

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

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

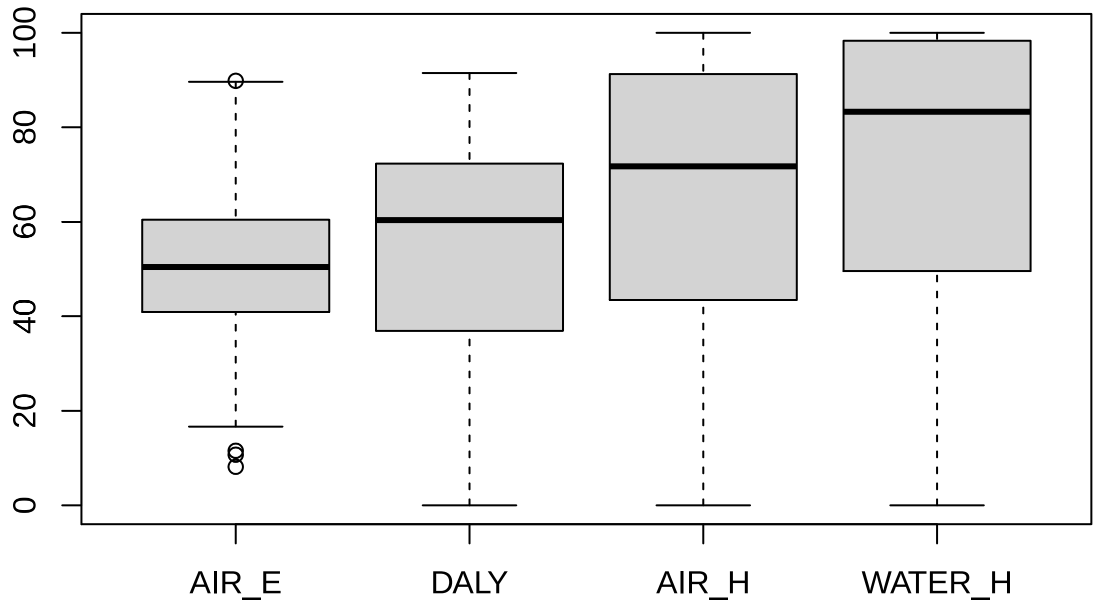
DALY had the largest impact on the regression function compared to AIR_H, and WATER_H to determine ENVHEALTH. Predictions did not turn out well due to a resulting error in row counts in EPI vs. the NEW dataset will try to fix.

regression on AIR_E

```r
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)

AIR_E <- EPI$AIR_E

boxplot(AIR_E,DALY,AIR_H,WATER_H, names = c("AIR_E", "DALY", "AIR_H", "WATER_H"))
```




```
lmAIR_E<-lm(AIR_E~DALY+AIR_H+WATER_H)
lmAIR_E
```

```
##
## Call:
## lm(formula = AIR_E ~ DALY + AIR_H + WATER_H)
##
## Coefficients:
## (Intercept)         DALY        AIR_H      WATER_H
##     59.2903      -0.1248       0.1686      -0.1798
```

```
summary(lmAIR_E)
```

```
##
## Call:
## lm(formula = AIR_E ~ DALY + AIR_H + WATER_H)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.708  -7.328  -1.739   8.117  38.182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.29025    2.55759  23.182  < 2e-16 ***
## DALY        -0.12482    0.07707  -1.620  0.10710
## AIR_H        0.16863    0.05104   3.304  0.00115 **
## WATER_H     -0.17982    0.07021  -2.561  0.01126 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.42 on 178 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.1803, Adjusted R-squared:  0.1664
## F-statistic: 13.05 on 3 and 178 DF,  p-value: 9.654e-08
```

```r
cAIR_E<-coef(lmAIR_E)

pENV<- predict(lmAIR_E,NEW,interval = "prediction")
```

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

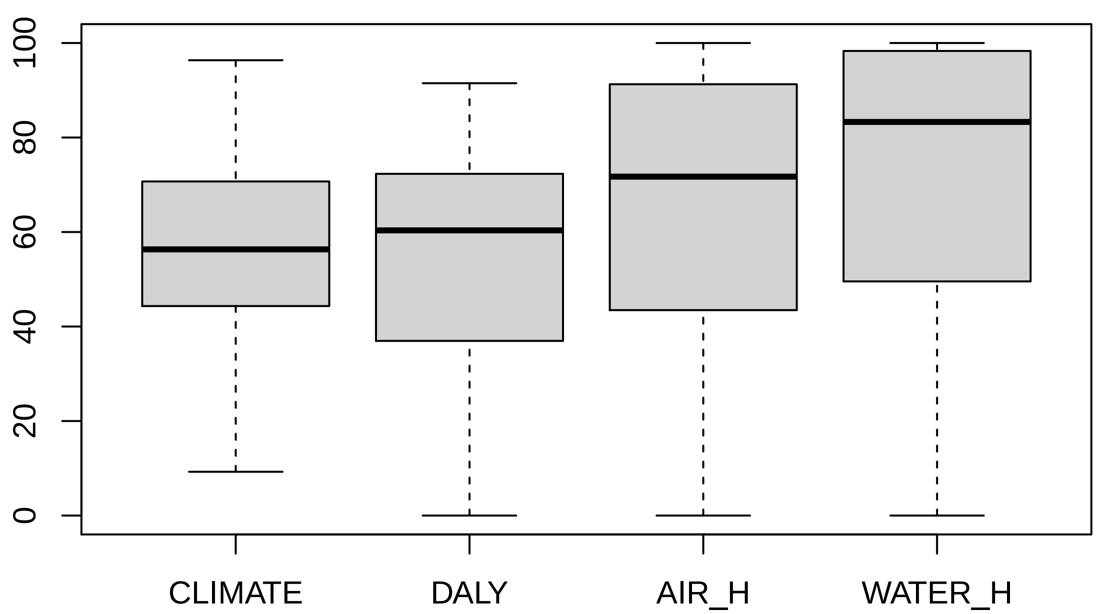
```r
cENV<- predict(lmAIR_E,NEW,interval = "confidence")
```

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

In this regression the variable AIR_H had the largest pull compared to WATER_H and DALY in Determining AIR_E however the intercept had the strongest indicating a bad regression model

regression on CLIMATE

```r
CLIMATE <- EPI$CLIMATE
boxplot(CLIMATE,DALY,AIR_H,WATER_H, names = c("CLIMATE", "DALY", "AIR_H", "WATER_H"))
```



```r
lmCLIMATE<-lm(CLIMATE~DALY+AIR_H+WATER_H)
lmCLIMATE
```

```
##
## Call:
## lm(formula = CLIMATE ~ DALY + AIR_H + WATER_H)
##
## Coefficients:
## (Intercept)         DALY        AIR_H      WATER_H
##     75.3487      -0.1732       0.0181      -0.1538
```

```r
summary(lmCLIMATE)
```

```
##
## Call:
## lm(formula = CLIMATE ~ DALY + AIR_H + WATER_H)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.578  -9.768   1.165   9.164  44.434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 75.34874    3.01412  24.999   <2e-16 ***
## DALY        -0.17323    0.09050  -1.914   0.0573 .
## AIR_H        0.01810    0.05919   0.306   0.7602
## WATER_H     -0.15385    0.08161  -1.885   0.0611 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.15 on 168 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.255,  Adjusted R-squared:  0.2417
## F-statistic: 19.17 on 3 and 168 DF,  p-value: 9.704e-11
```

```r
cCLIMATE<-coef(lmCLIMATE)

pENV<- predict(lmCLIMATE,NEW,interval = "prediction")
```

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

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

```
## Warning: 'newdata' had 19 rows but variables found have 231 rows
```

In this regression the variables DALY,AIR_H,and WATER_H all were insignificant variables in the regression model to determine climate

Lab 2 part 2

Exercise 1: Regression

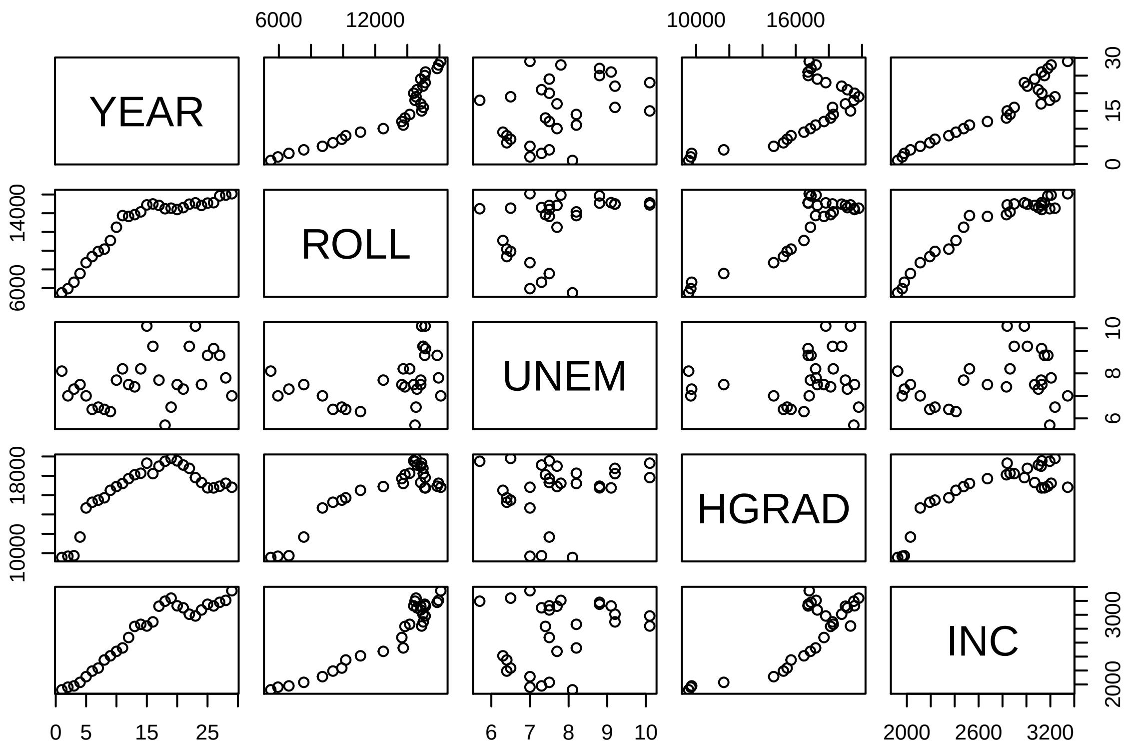
Data Exploration

```r
mult_reg<-read.csv("/Users/donneb/Documents/DataAnalytics/dataset_multipleRegression.csv")
#EDA of data set
```

```r
head(mult_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
plot(mult_reg[])
```

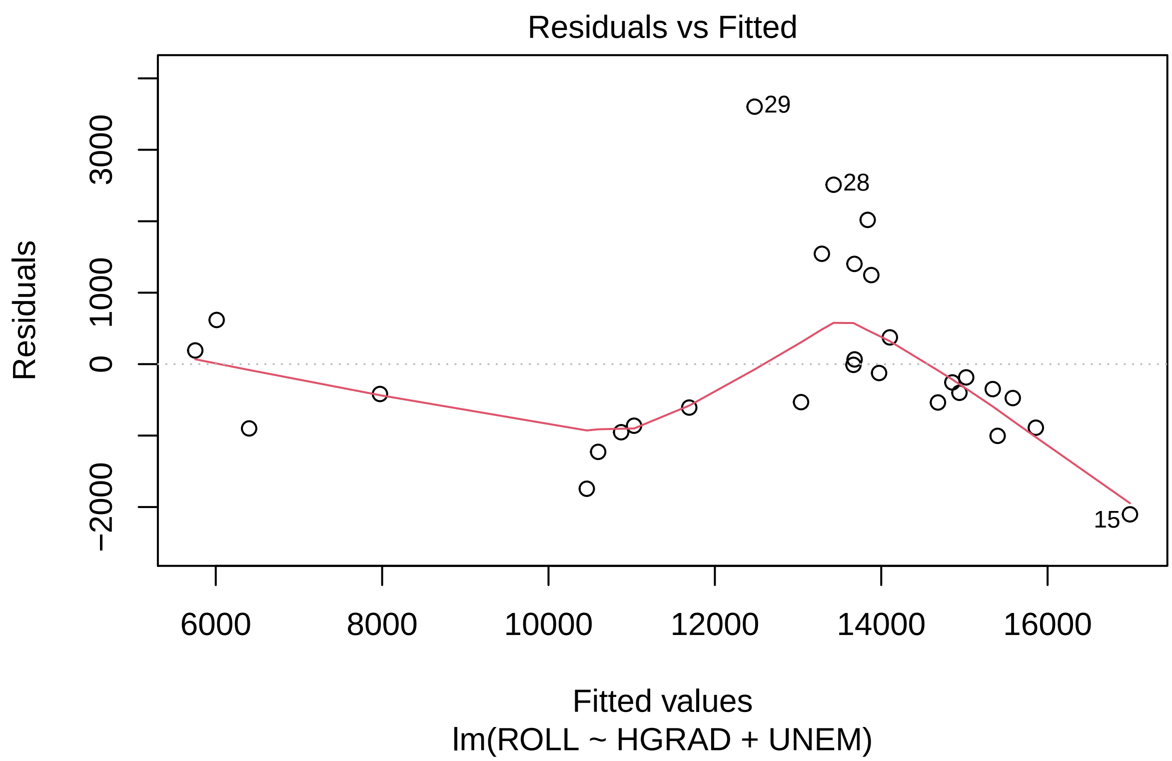


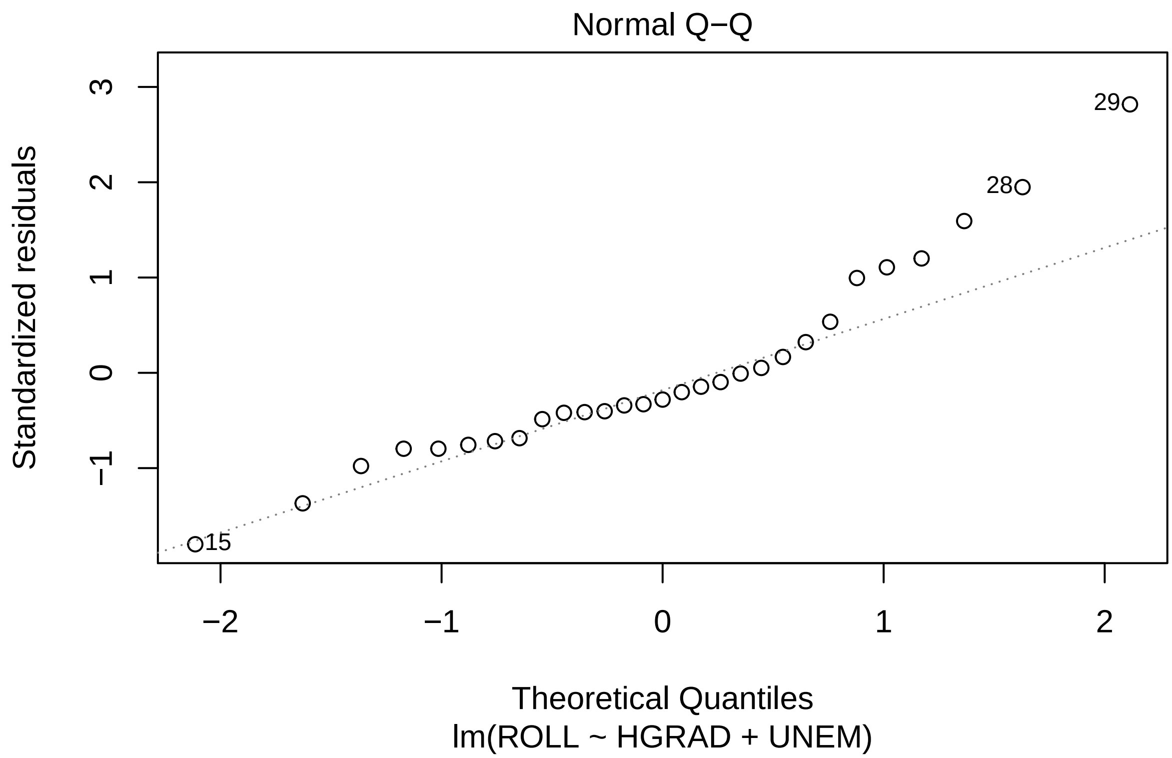
regression model 1: Exploring regression model 1 factoring HGRAD, UNEM

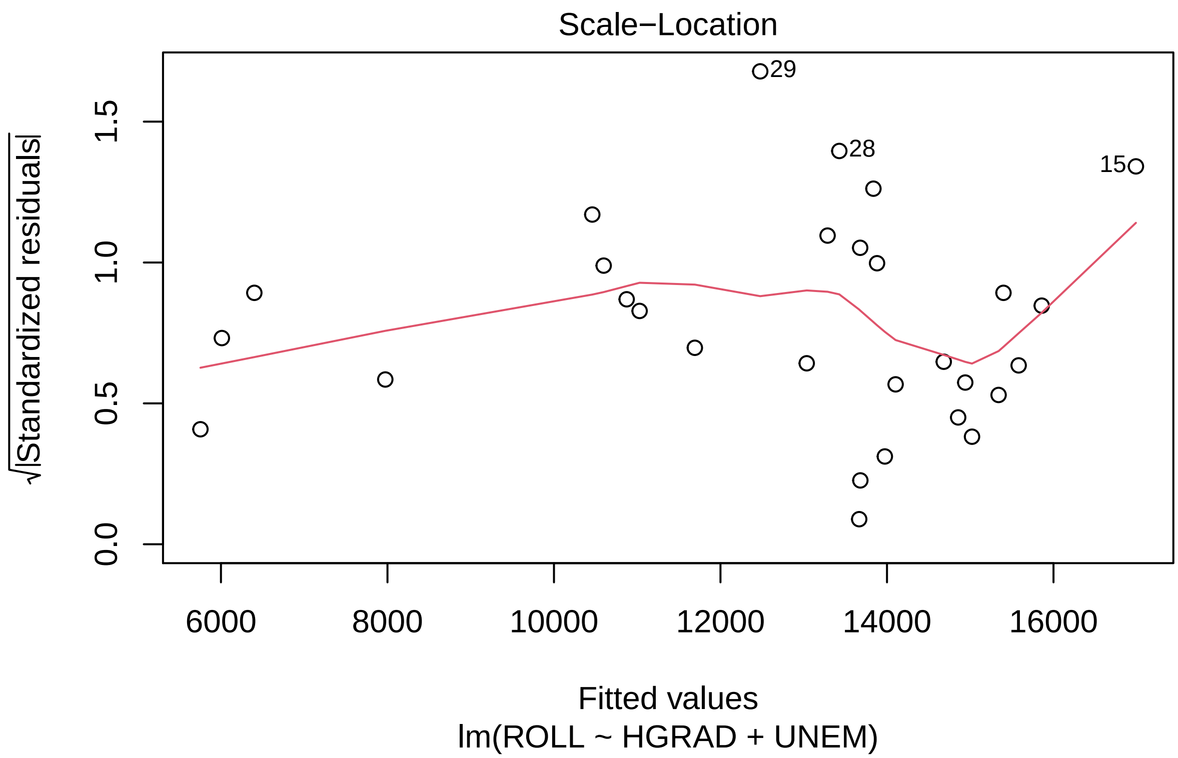
```r
#Exploring regression model 1 factoring HGRAD, UNEM
lmROLL<- lm(ROLL~HGRAD+UNEM, data = mult_reg)
lmROLL
```

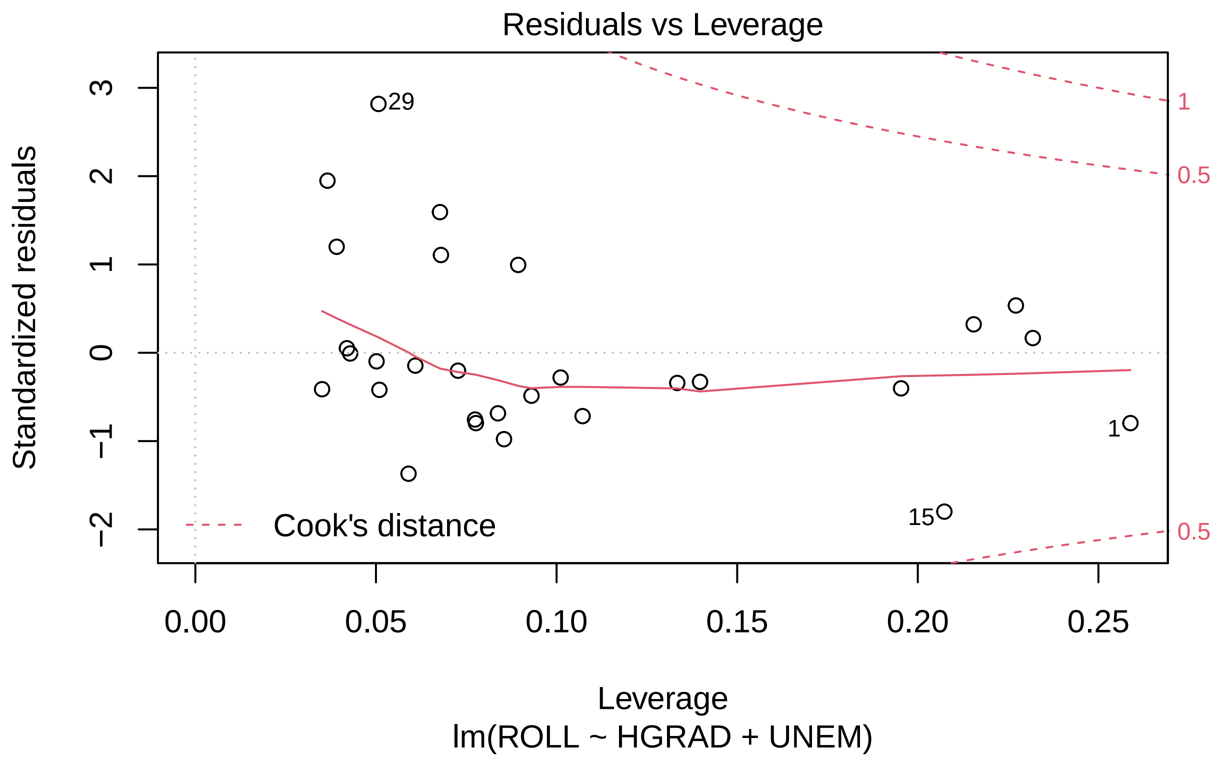
```
##
## Call:
## lm(formula = ROLL ~ HGRAD + UNEM, data = mult_reg)
##
## Coefficients:
## (Intercept)        HGRAD         UNEM
##  -8255.7511       0.9423     698.2681
```

```r
#will only plot lm residuals and all for this example
plot(lmROLL)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
29
28
15
−2
−1
0
1
2
Theoretical Quantiles
lm(ROLL ~ HGRAD + UNEM)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
29
28
15
6000
8000
10000
12000
14000
16000
Fitted values
lm(ROLL ~ HGRAD + UNEM)




```
summary(lmROLL)
```

```
##
## Call:
## lm(formula = ROLL ~ HGRAD + UNEM, data = mult_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2102.2  -861.6  -349.4   374.5  3603.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.256e+03  2.052e+03  -4.023  0.00044 ***
## HGRAD        9.423e-01  8.613e-02  10.941 3.16e-11 ***
## UNEM         6.983e+02  2.244e+02   3.111  0.00449 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1313 on 26 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8373
## F-statistic: 73.03 on 2 and 26 DF,  p-value: 2.144e-11
```

```
lmROLL$coefficients
```

```
##   (Intercept)         HGRAD          UNEM
## -8255.7510591     0.9422769   698.2681316
```

prediction for model 1 w/ UNEM=7%, HGRAD=90,000

```r
#prediction 1 based on model 1
pred_nextyear1 <-predict(lmROLL, newdata=data.frame(HGRAD = 90000 , UNEM =.07 ))
pred_nextyear1
```

```
##        1
## 76598.04
```

new model factoring HGRAD,UNEM,INC

```r
#Exploring regression model 1 factoring HGRAD, UNEM
lmROLL2<- lm(ROLL~HGRAD+UNEM+INC, data = mult_reg)
lmROLL2
```

```
##
## Call:
## lm(formula = ROLL ~ HGRAD + UNEM + INC, data = mult_reg)
##
## Coefficients:
## (Intercept)        HGRAD         UNEM          INC
##  -9153.2545       0.4065     450.1245       4.2749
```

```r
summary(lmROLL2)
```

```
##
## Call:
## lm(formula = ROLL ~ HGRAD + UNEM + INC, data = mult_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1148.84  -489.71    -1.88   387.40  1425.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.153e+03  1.053e+03  -8.691 5.02e-09 ***
## HGRAD        4.065e-01  7.602e-02   5.347 1.52e-05 ***
## UNEM         4.501e+02  1.182e+02   3.809 0.000807 ***
## INC          4.275e+00  4.947e-01   8.642 5.59e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 670.4 on 25 degrees of freedom
## Multiple R-squared:  0.9621, Adjusted R-squared:  0.9576
## F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

prediction factoring per capita income (INC = 25,000)

```r
#prediction 1 based on model 1
pred_nextyear2 <-predict(lmROLL2, newdata=data.frame(HGRAD = 90000 , UNEM =.07, INC = 25000 ))
pred_nextyear2
```

```
##        1
## 134333.2
```

Exercise 2

```r
# summary of data set
abalone<-read.csv("/Users/donneb/Documents/DataAnalytics/abalone.csv")
colnames(abalone) <- c("sex", "length", 'diameter', 'height', 'whole_weight', 'shucked_wieght', 'viscera
summary(abalone)
```

```
##      sex                length          diameter          height      
##  Length:4177        Min.   :0.075   Min.   :0.0550   Min.   :0.0000  
##  Class :character   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150  
##  Mode  :character   Median :0.545   Median :0.4250   Median :0.1400  
##                     Mean   :0.524   Mean   :0.4079   Mean   :0.1395  
##                     3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650  
##                     Max.   :0.815   Max.   :0.6500   Max.   :1.1300  
##   whole_weight    shucked_wieght   viscera_wieght    shell_weight   
##  Min.   :0.0020   Min.   :0.0010   Min.   :0.0005   Min.   :0.0015  
##  1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300  
##  Median :0.7995   Median :0.3360   Median :0.1710   Median :0.2340  
##  Mean   :0.8287   Mean   :0.3594   Mean   :0.1806   Mean   :0.2388  
##  3rd Qu.:1.1530   3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290  
##  Max.   :2.8255   Max.   :1.4880   Max.   :0.7600   Max.   :1.0050  
##      rings       
##  Min.   : 1.000  
##  1st Qu.: 8.000  
##  Median : 9.000  
##  Mean   : 9.934  
##  3rd Qu.:11.000  
##  Max.   :29.000
```

```r
str(abalone)
```

```
## 'data.frame':    4177 obs. of  9 variables:
##  $ sex           : chr  "M" "M" "F" "M" ...
##  $ length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
##  $ diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
##  $ height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
##  $ whole_weight  : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ shucked_wieght: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ viscera_wieght: num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ shell_weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
##  $ rings         : int  15 7 9 10 7 8 20 16 9 19 ...
```

```r
summary(abalone$rings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   8.000   9.000   9.934  11.000  29.000
```

grouping by age rings

```r
# age rings
abalone$rings <- as.numeric(abalone$rings)
abalone$rings <- cut(abalone$rings, br=c(-1,8,11,35), labels = c("young", 'adult', 'old'))
abalone$rings <- as.factor(abalone$rings)

summary(abalone$rings)
```

```
## young adult   old
##  1407  1810   960
```

Copying dataset,removing non numeric for KNN, and normalizing

```
aba<- abalone
aba$sex <-NULL

#  normalize the data using min max normalization
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

aba[1:7] <- as.data.frame(lapply(aba[1:7], normalize))
summary(aba$shucked_wieght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1244  0.2253  0.2410  0.3369  1.0000
```

Its KNN Time!

```
ind <- sample(2, nrow(aba), replace=TRUE, prob=c(0.7, 0.3))

#test,train
KNNtrain <- aba[ind==1,]
KNNtest <- aba[ind==2,]

# set k to sqrt(2918) ~ 54.02 round up to 55
library(class)
KNNpred<- knn(train = KNNtrain[1:7], test = KNNtest[1:7], cl = KNNtrain$rings, k=55)
KNNpred
```

```
##    [1] adult old   adult young adult adult adult old   young adult adult young
##   [13] adult adult young young adult adult adult old   old   old   adult young
##   [25] old   adult adult old   adult adult young young adult adult old   young
##   [37] young old   old   young young old   young young adult old   adult old
##   [49] adult young old   adult adult young adult young old   young adult adult
##   [61] young young young young adult young young young young old   adult old
##   [73] old   adult adult old   adult young old   adult old   adult adult adult
##   [85] adult old   young young young young young young young young adult young
##   [97] young young young adult adult adult young adult adult adult adult adult
##  [109] old   adult old   old   old   adult adult young young adult adult adult
##  [121] adult adult adult adult young young old   old   old   adult adult young
##  [133] adult adult old   old   old   adult adult adult adult adult old   adult
##  [145] young young young young young young young young adult young adult adult
##  [157] adult young young adult adult adult adult young young adult old   adult
##  [169] young young adult adult adult adult young adult young adult adult old
##  [181] young young young young adult young adult young young adult young adult
##  [193] young young young adult young young adult adult adult adult adult young
##  [205] young adult young young young young young young young young young adult
##  [217] adult old   old   adult old   adult adult adult adult old   old   adult
##  [229] old   old   old   old   old   old   old   adult old   old   adult young
```

```
##  [241] adult adult young young young young young adult young adult adult adult
##  [253] adult adult adult adult adult adult adult adult old   young young young
##  [265] young young adult young young young young young young young young young
##  [277] young young young young adult young adult adult young adult adult adult
##  [289] adult adult adult adult adult adult adult adult adult adult adult adult
##  [301] adult adult old   old   young young young young young young young young
##  [313] young young young young young young young young adult young adult young
##  [325] young adult adult adult adult adult adult adult adult adult adult adult
##  [337] adult adult adult adult adult adult adult adult old   adult adult adult
##  [349] adult old   young young young young young young young young young young
##  [361] young young young young young young young young young adult young young
##  [373] young young adult adult adult adult adult adult adult adult adult adult
##  [385] adult adult adult adult adult adult adult adult adult adult adult adult
##  [397] adult adult adult adult adult adult adult adult adult adult adult adult
##  [409] adult adult adult adult adult young young young young young young young
##  [421] young young young young young young young young young adult adult adult
##  [433] adult adult adult adult adult adult adult adult old   adult adult young
##  [445] young young young young young young young young young young adult young
##  [457] young young adult adult adult adult adult adult adult adult adult adult
##  [469] old   adult adult old   adult adult adult adult adult adult adult adult
##  [481] adult adult adult adult adult adult adult adult adult adult adult adult
##  [493] adult adult adult adult adult adult adult adult adult adult adult adult
##  [505] adult adult adult adult old   old   young adult adult adult adult adult
##  [517] adult adult adult adult adult adult old   adult adult adult young young
##  [529] young young young young young adult adult young adult adult adult young
##  [541] adult adult adult adult adult adult adult adult adult adult adult adult
##  [553] adult adult adult adult adult adult adult adult adult adult adult adult
##  [565] adult adult old   adult adult young young young young young young young
##  [577] young adult young adult adult adult adult adult adult adult adult adult
##  [589] young young young young young young young adult adult adult adult adult
##  [601] adult old   young adult young young adult old   young young young adult
##  [613] young old   young adult young young young adult young young old   old
##  [625] adult old   adult young adult young young young adult young old   old
##  [637] young young old   young old   adult adult young adult adult old   young
##  [649] young young young young adult adult adult old   old   adult adult adult
##  [661] old   old   old   old   adult adult adult adult young old   adult adult
##  [673] adult adult adult adult old   adult adult adult old   adult adult old
##  [685] adult old   young adult young old   adult old   young young adult young
##  [697] young young adult young young old   young adult old   adult adult young
##  [709] adult old   adult old   young adult young young young adult adult adult
##  [721] adult adult adult adult old   adult old   young young adult old   young
##  [733] young young young adult adult adult adult adult adult adult adult adult
##  [745] old   adult adult young young young young young young young young young
##  [757] adult young adult adult adult adult adult adult adult adult adult adult
##  [769] adult adult adult adult adult young young adult young adult adult adult
##  [781] adult adult adult adult adult adult adult adult adult adult adult adult
##  [793] adult adult adult young young young young young young young young adult
##  [805] young young adult adult adult adult adult adult adult adult adult adult
##  [817] adult adult adult adult adult adult adult adult adult adult adult adult
##  [829] adult young young young young adult young adult adult adult adult adult
##  [841] adult adult adult adult adult adult adult adult young young young adult
##  [853] young adult adult adult old   adult adult adult adult adult adult adult
##  [865] adult adult adult adult adult adult adult adult adult adult adult adult
##  [877] adult adult adult adult adult old   young young adult adult adult adult
```

```
##  [889] adult adult adult young young young young adult adult adult adult adult
##  [901] adult adult adult adult adult adult adult adult old   adult adult young
##  [913] adult young young adult adult adult young young young young young young
##  [925] adult adult adult adult young young old   adult old   adult old   adult
##  [937] young young adult adult adult old   adult young old   young young adult
##  [949] adult young adult old   young adult young adult old   adult adult adult
##  [961] old   young young adult adult adult adult adult adult old   young adult
##  [973] old   adult adult adult adult adult adult adult old   old   young young
##  [985] adult old   adult adult young adult adult old   young old   young young
##  [997] adult young adult young young young adult adult young young young young
## [1009] young young young adult adult old   adult adult young young young young
## [1021] young young young young young adult adult adult adult young young young
## [1033] young young adult adult adult adult adult adult adult adult adult adult
## [1045] young young young young young young young young young young adult adult
## [1057] adult adult adult adult adult adult adult adult adult adult adult adult
## [1069] young young young young adult old   adult adult adult adult adult adult
## [1081] young young young young young adult adult adult adult adult adult adult
## [1093] adult adult adult adult adult adult adult adult adult adult adult adult
## [1105] adult adult young young young young adult adult adult young adult adult
## [1117] adult adult adult adult adult adult adult adult adult adult adult adult
## [1129] young young adult adult young young old   young adult young adult young
## [1141] adult old   adult old   adult young young young old   adult old   old
## [1153] adult old   young adult adult adult young old   adult young young adult
## [1165] adult adult adult young young young young young young adult adult adult
## [1177] adult adult adult young young young adult adult young young young young
## [1189] young young adult adult adult adult adult adult young adult young young
## [1201] adult adult adult adult adult adult adult young young adult adult adult
## [1213] adult adult adult adult adult adult adult adult adult adult adult adult
## [1225] adult young young young young young adult young adult adult
## Levels: young adult old
```

```r
table(KNNpred)
```

```
## KNNpred
## young adult   old
##   415   698   121
```

Exercise 3 - KNN exploration

```r
library(ggplot2)
iris_copy = iris
#drop species column
iris_copy$Species = NULL
head(iris_copy)
```

```
##   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
str(iris_copy)
```

```
## 'data.frame':    150 obs. of  4 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
```

```r
summary(iris_copy)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```r
sapply(iris_copy[,-5], var)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width 
##    0.6856935    0.1899794    3.1162779    0.5810063
```

```r
iris_copy[,3:4]
```

```
##     Petal.Length Petal.Width
## 1            1.4         0.2
## 2            1.4         0.2
## 3            1.3         0.2
## 4            1.5         0.2
## 5            1.4         0.2
## 6            1.7         0.4
## 7            1.4         0.3
## 8            1.5         0.2
## 9            1.4         0.2
## 10           1.5         0.1
## 11           1.5         0.2
## 12           1.6         0.2
## 13           1.4         0.1
## 14           1.1         0.1
## 15           1.2         0.2
## 16           1.5         0.4
## 17           1.3         0.4
## 18           1.4         0.3
## 19           1.7         0.3
## 20           1.5         0.3
## 21           1.7         0.2
## 22           1.5         0.4
## 23           1.0         0.2
## 24           1.7         0.5
## 25           1.9         0.2
## 26           1.6         0.2
## 27           1.6         0.4
## 28           1.5         0.2
## 29           1.4         0.2
## 30           1.6         0.2
```

```
## 31          1.6         0.2
## 32          1.5         0.4
## 33          1.5         0.1
## 34          1.4         0.2
## 35          1.5         0.2
## 36          1.2         0.2
## 37          1.3         0.2
## 38          1.4         0.1
## 39          1.3         0.2
## 40          1.5         0.2
## 41          1.3         0.3
## 42          1.3         0.3
## 43          1.3         0.2
## 44          1.6         0.6
## 45          1.9         0.4
## 46          1.4         0.3
## 47          1.6         0.2
## 48          1.4         0.2
## 49          1.5         0.2
## 50          1.4         0.2
## 51          4.7         1.4
## 52          4.5         1.5
## 53          4.9         1.5
## 54          4.0         1.3
## 55          4.6         1.5
## 56          4.5         1.3
## 57          4.7         1.6
## 58          3.3         1.0
## 59          4.6         1.3
## 60          3.9         1.4
## 61          3.5         1.0
## 62          4.2         1.5
## 63          4.0         1.0
## 64          4.7         1.4
## 65          3.6         1.3
## 66          4.4         1.4
## 67          4.5         1.5
## 68          4.1         1.0
## 69          4.5         1.5
## 70          3.9         1.1
## 71          4.8         1.8
## 72          4.0         1.3
## 73          4.9         1.5
## 74          4.7         1.2
## 75          4.3         1.3
## 76          4.4         1.4
## 77          4.8         1.4
## 78          5.0         1.7
## 79          4.5         1.5
## 80          3.5         1.0
## 81          3.8         1.1
## 82          3.7         1.0
## 83          3.9         1.2
## 84          5.1         1.6
```

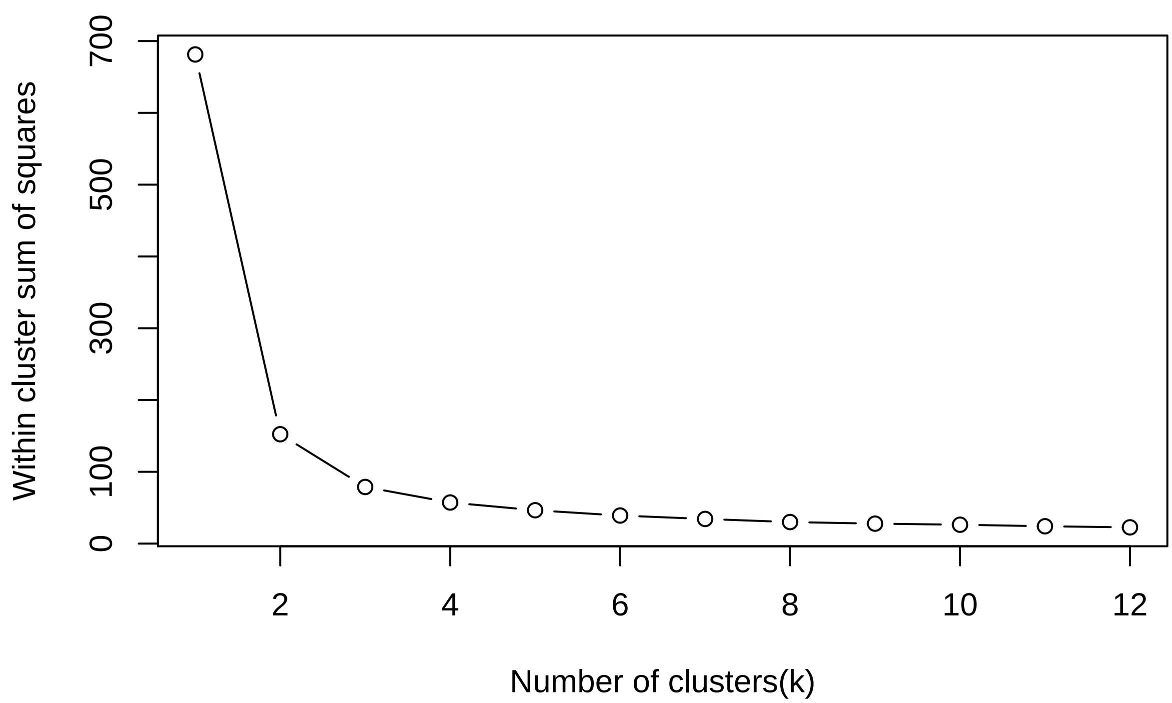
```
## 85            4.5          1.5
## 86            4.5          1.6
## 87            4.7          1.5
## 88            4.4          1.3
## 89            4.1          1.3
## 90            4.0          1.3
## 91            4.4          1.2
## 92            4.6          1.4
## 93            4.0          1.2
## 94            3.3          1.0
## 95            4.2          1.3
## 96            4.2          1.2
## 97            4.2          1.3
## 98            4.3          1.3
## 99            3.0          1.1
## 100           4.1          1.3
## 101           6.0          2.5
## 102           5.1          1.9
## 103           5.9          2.1
## 104           5.6          1.8
## 105           5.8          2.2
## 106           6.6          2.1
## 107           4.5          1.7
## 108           6.3          1.8
## 109           5.8          1.8
## 110           6.1          2.5
## 111           5.1          2.0
## 112           5.3          1.9
## 113           5.5          2.1
## 114           5.0          2.0
## 115           5.1          2.4
## 116           5.3          2.3
## 117           5.5          1.8
## 118           6.7          2.2
## 119           6.9          2.3
## 120           5.0          1.5
## 121           5.7          2.3
## 122           4.9          2.0
## 123           6.7          2.0
## 124           4.9          1.8
## 125           5.7          2.1
## 126           6.0          1.8
## 127           4.8          1.8
## 128           4.9          1.8
## 129           5.6          2.1
## 130           5.8          1.6
## 131           6.1          1.9
## 132           6.4          2.0
## 133           5.6          2.2
## 134           5.1          1.5
## 135           5.6          1.4
## 136           6.1          2.3
## 137           5.6          2.4
## 138           5.5          1.8
```

```
## 139          4.8         1.8
## 140          5.4         2.1
## 141          5.6         2.4
## 142          5.1         2.3
## 143          5.1         1.9
## 144          5.9         2.3
## 145          5.7         2.5
## 146          5.2         2.3
## 147          5.0         1.9
## 148          5.2         2.0
## 149          5.4         2.3
## 150          5.1         1.8
```

```
#setting seeds & kmeans function
set.seed(300)
k.max <- 12
wss<- sapply(1:k.max,function(k){kmeans(iris_copy[],k,nstart = 20,iter.max = 1000)$tot.withinss})
wss
```

```
##  [1] 681.37060 152.34795  78.85144  57.22847  46.44618  39.05498  34.29823
##  [8]  29.98894  27.78609  26.29643  24.13389  22.62722
```

```
plot(1:k.max,wss, type= "b", xlab = "Number of clusters(k)", ylab = "Within cluster sum of squares")
```




```
icluster <- kmeans(iris_copy[,3:4],3,nstart = 20)
correct_table<- table(iris[,5],icluster$cluster)
correct_table
```

```
##
##              1  2  3
##   setosa      0 50  0
##   versicolor 48  0  2
##   virginica   4  0 46
```

The resulting clusters that were created under this KNN clustering were not entirely correct. The 2nd group 100% matched the setosa species However the versicolor was split between the 1st and 3rd group with 48 in the 3rd and 2 in the 1st. In addition, the the virginica clustering was split 46 in the 1st group, and 4 in the 3rd group. This indicates room for improvement and perhaps parameter adjustments

Exercise 4

sample values sample_n

```r
library(dplyr)
```

```
##
## 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[14],5)
```

```
##    EPI
## 1   NA
## 2 58.0
## 3   NA
## 4 37.6
## 5 67.1
```

```r
sample_n(EPI[17],5)
```

```
##    DALY
## 1 82.81
## 2 60.35
## 3 73.01
## 4    NA
## 5 18.16
```

sample_frac

```r
sample_frac(EPI[14],.1)
```

```
##     EPI
## 1    NA
## 2  41.0
## 3  44.6
## 4  62.2
## 5    NA
```

```
## 6  73.2
## 7  47.0
## 8  63.4
## 9  60.4
## 10 54.3
## 11   NA
## 12 59.1
## 13 59.0
## 14 65.9
## 15 66.4
## 16 63.7
## 17 48.9
## 18 72.5
## 19   NA
## 20 76.8
## 21 68.2
## 22 60.4
## 23   NA
```

```
sample_frac(EPI[17],.1)
```

```
##     DALY
## 1     NA
## 2     NA
## 3  79.20
## 4  70.31
## 5  44.18
## 6  69.04
## 7   5.81
## 8  56.74
## 9  30.28
## 10 73.01
## 11 63.34
## 12 55.08
## 13 18.16
## 14 64.40
## 15 91.50
## 16    NA
## 17 67.82
## 18 66.64
## 19 55.08
## 20 65.50
## 21    NA
## 22 64.40
## 23 64.40
```

arrange by descending

```
by_DALY <- EPI %>% group_by(DALY)
by_DALY <- by_DALY %>% arrange(desc(DALY), .by_group = TRUE)
by_EPI <- EPI %>% group_by(EPI)
by_EPI <- by_EPI %>% arrange(desc(EPI), .by_group = TRUE)
```

head of desc outputs

```r
head(by_DALY)
```

```
## # A tibble: 6 x 160
## # Groups:   DALY [3]
##    code ISO3V10 Country EPI_regions GEO_subregion GDPCAP07 Population07 Landarea
##   <int> <chr>   <chr>   <chr>       <chr>            <dbl>        <dbl>    <dbl>
## 1     4 AFG     Afghan~ South Asia  South Asia         NA            NA   634925.
## 2    24 AGO     Angola  Sub-Sahara~ Southern Afr~    4875.     17554585  1251896.
## 3   562 NER     Niger   Sub-Sahara~ Western Afri~     597.     14195085. 1157232.
## 4   694 SLE     Sierra~ Sub-Sahara~ Western Afri~     691.      5420400    72617.
## 5   430 LBR     Liberia Sub-Sahara~ Western Afri~     350.      3627285    96166.
## 6   466 MLI     Mali    Sub-Sahara~ Western Afri~    1023.     12334168. 1248146.
## # ... with 152 more variables: PopulationDensity <dbl>, Landlock <int>,
## #   No_surface_water <int>, Desert <int>, High_Population_Density <int>,
## #   EPI <dbl>, ENVHEALTH <dbl>, ECOSYSTEM <dbl>, DALY <dbl>, AIR_H <dbl>,
## #   WATER_H <dbl>, AIR_E <dbl>, WATER_E <dbl>, BIODIVERSITY <dbl>,
## #   FORESTRY <dbl>, FISHERIES <dbl>, AGRICULTURE <dbl>, CLIMATE <dbl>,
## #   DALY_pt <dbl>, ACSAT_pt <dbl>, ACSAT_pt_imp <int>, WATSUP_pt <dbl>,
## #   WATSUP_pt_imp <int>, INDOOR_pt <dbl>, PM10_pt <dbl>, SO2_pt <dbl>,
## #   NOX_pt <dbl>, NMVOC_pt <dbl>, OZONE_pt <dbl>, WQI_pt <dbl>,
## #   WQI_pt_imp <int>, WQI_pt_GEMS.station.data <dbl>, WSI_pt <dbl>,
## #   WATSTR_pt <dbl>, PACOV_pt <dbl>, MPAEEZ_pt <dbl>, AZE_pt <dbl>,
## #   FORGRO_pt <dbl>, FORCOV_pt <dbl>, MTI_pt <dbl>, EEZTD_pt <dbl>,
## #   AGWAT_pt <dbl>, AGSUB_pt <dbl>, AGPEST_pt <dbl>, GHGCAP_pt <dbl>,
## #   GHGCAP_pt_imp <int>, GHGIND_pt <dbl>, CO2KWH_pt <dbl>, CO2KWH_pt_imp <int>,
## #   DALY_raw <int>, ACSAT_raw <dbl>, ACSAT_raw_imp <int>, WATSUP_raw <dbl>,
## #   WATSUP_raw_imp <int>, INDOOR_raw <dbl>, PM10_raw <dbl>, OZONE_raw <dbl>,
## #   WQI_raw <dbl>, WQI_raw_imp <int>, WQI_raw_GEMS.station.data <dbl>,
## #   SO2_raw <dbl>, NOX_raw <dbl>, NMVOC_raw <dbl>, WSI_raw <dbl>,
## #   WATSTR_raw <dbl>, PACOV_raw <dbl>, AZE_raw <dbl>, MPAEEZ_raw <dbl>,
## #   FORGRO_raw <dbl>, FORCOV_raw <dbl>, MTI_raw <dbl>, EEZTD_raw <dbl>,
## #   AGWAT_raw <dbl>, AGSUB_raw <dbl>, AGPEST_raw <int>, GHGCAP_raw <dbl>,
## #   GHGCAP_raw_imp <int>, GHGIND_raw <dbl>, CO2KWH_raw <dbl>,
## #   CO2KWH_raw_imp <int>, DALY_w <dbl>, ACSAT_w <dbl>, WATSUP_w <dbl>,
## #   INDOOR_w <dbl>, PM10_w <dbl>, OZONE_w <dbl>, SO2_w <dbl>, NOX_w <dbl>,
## #   NMVOC_w <dbl>, WSI_w <dbl>, WATSTR_w <dbl>, PACOV_w <dbl>, AZE_w <dbl>,
## #   MPAEEZ_w <dbl>, FORGRO_w <dbl>, FORCOV_w <dbl>, MTI_w <dbl>, EEZTD_w <dbl>,
## #   AGWAT_w <dbl>, AGSUB_w <dbl>, ...
```

```r
head(by_EPI)
```

```
## # A tibble: 6 x 160
## # Groups:   EPI [6]
##    code ISO3V10 Country EPI_regions GEO_subregion GDPCAP07 Population07 Landarea
##   <int> <chr>   <chr>   <chr>       <chr>            <dbl>        <dbl>    <dbl>
## 1   694 SLE     Sierra~ Sub-Sahara~ Western Afri~     691.      5420400    72617.
## 2   140 CAF     Centra~ Sub-Sahara~ Central Afri~     674.      4343405.  622868.
## 3   478 MRT     Maurit~ Sub-Sahara~ Western Afri~    1820.      3120981. 1036905.
## 4    24 AGO     Angola  Sub-Sahara~ Southern Afr~    4875.     17554585  1251896.
## 5   768 TGO     Togo    Sub-Sahara~ Western Afri~     777.      6300495    57277.
## 6   562 NER     Niger   Sub-Sahara~ Western Afri~     597.     14195085. 1157232.
## # ... with 152 more variables: PopulationDensity <dbl>, Landlock <int>,
## #   No_surface_water <int>, Desert <int>, High_Population_Density <int>,
```

```
## #   EPI <dbl>, ENVHEALTH <dbl>, ECOSYSTEM <dbl>, DALY <dbl>, AIR_H <dbl>,
## #   WATER_H <dbl>, AIR_E <dbl>, WATER_E <dbl>, BIODIVERSITY <dbl>,
## #   FORESTRY <dbl>, FISHERIES <dbl>, AGRICULTURE <dbl>, CLIMATE <dbl>,
## #   DALY_pt <dbl>, ACSAT_pt <dbl>, ACSAT_pt_imp <int>, WATSUP_pt <dbl>,
## #   WATSUP_pt_imp <int>, INDOOR_pt <dbl>, PM10_pt <dbl>, SO2_pt <dbl>,
## #   NOX_pt <dbl>, NMVOC_pt <dbl>, OZONE_pt <dbl>, WQI_pt <dbl>,
## #   WQI_pt_imp <int>, WQI_pt_GEMS.station.data <dbl>, WSI_pt <dbl>,
## #   WATSTR_pt <dbl>, PACOV_pt <dbl>, MPAEEZ_pt <dbl>, AZE_pt <dbl>,
## #   FORGRO_pt <dbl>, FORCOV_pt <dbl>, MTI_pt <dbl>, EEZTD_pt <dbl>,
## #   AGWAT_pt <dbl>, AGSUB_pt <dbl>, AGPEST_pt <dbl>, GHGCAP_pt <dbl>,
## #   GHGCAP_pt_imp <int>, GHGIND_pt <dbl>, CO2KWH_pt <dbl>, CO2KWH_pt_imp <int>,
## #   DALY_raw <int>, ACSAT_raw <dbl>, ACSAT_raw_imp <int>, WATSUP_raw <dbl>,
## #   WATSUP_raw_imp <int>, INDOOR_raw <dbl>, PM10_raw <dbl>, OZONE_raw <dbl>,
## #   WQI_raw <dbl>, WQI_raw_imp <int>, WQI_raw_GEMS.station.data <dbl>,
## #   SO2_raw <dbl>, NOX_raw <dbl>, NMVOC_raw <dbl>, WSI_raw <dbl>,
## #   WATSTR_raw <dbl>, PACOV_raw <dbl>, AZE_raw <dbl>, MPAEEZ_raw <dbl>,
## #   FORGRO_raw <dbl>, FORCOV_raw <dbl>, MTI_raw <dbl>, EEZTD_raw <dbl>,
## #   AGWAT_raw <dbl>, AGSUB_raw <dbl>, AGPEST_raw <int>, GHGCAP_raw <dbl>,
## #   GHGCAP_raw_imp <int>, GHGIND_raw <dbl>, CO2KWH_raw <dbl>,
## #   CO2KWH_raw_imp <int>, DALY_w <dbl>, ACSAT_w <dbl>, WATSUP_w <dbl>,
## #   INDOOR_w <dbl>, PM10_w <dbl>, OZONE_w <dbl>, SO2_w <dbl>, NOX_w <dbl>,
## #   NMVOC_w <dbl>, WSI_w <dbl>, WATSTR_w <dbl>, PACOV_w <dbl>, AZE_w <dbl>,
## #   MPAEEZ_w <dbl>, FORGRO_w <dbl>, FORCOV_w <dbl>, MTI_w <dbl>, EEZTD_w <dbl>,
## #   AGWAT_w <dbl>, AGSUB_w <dbl>, ...
```

mutate

```r
#should have done this a while ago
EPI_copy<- EPI

#mutate functions
mutate(EPI_copy, double_EPI = EPI * 2)
mutate(EPI_copy, double_DALY = DALY * 2)
```

EPI,DALY mean

```r
EPI %>%
  summarize(EPI_mean = mean(EPI, na.rm= TRUE))
```

```
##   EPI_mean
## 1 58.37055
```

```r
EPI %>%
  summarize(DALY_mean = mean(DALY, na.rm= TRUE))
```

```
##   DALY_mean
## 1  53.94313
```