Kevin Pan

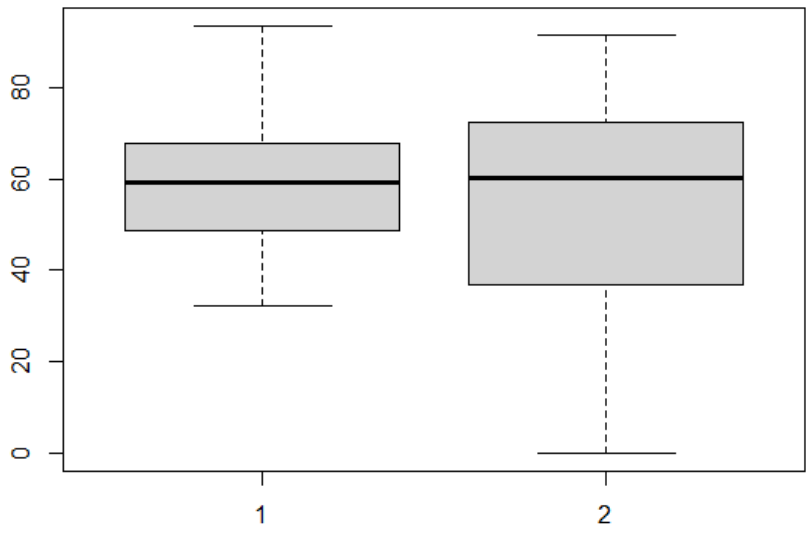
Data Analytics (Level 6000)

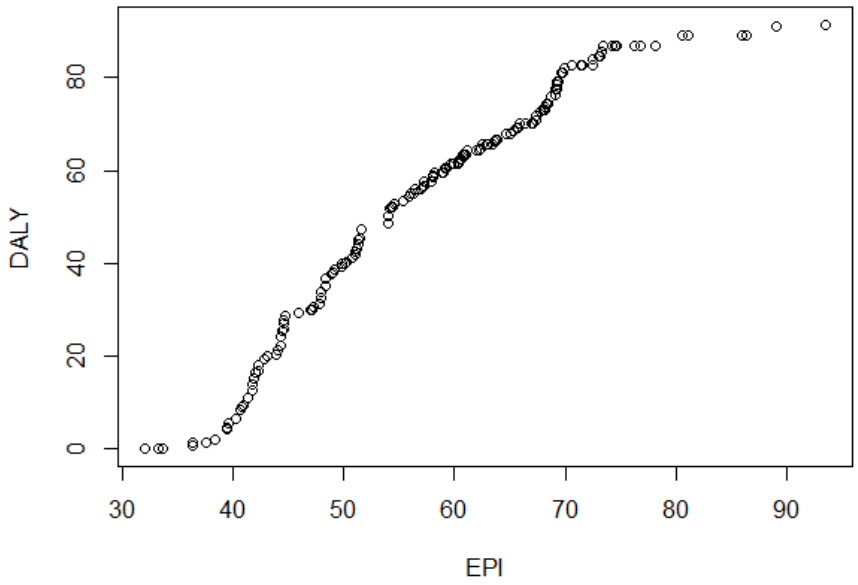
Lab2

Due: March. 04, 2021

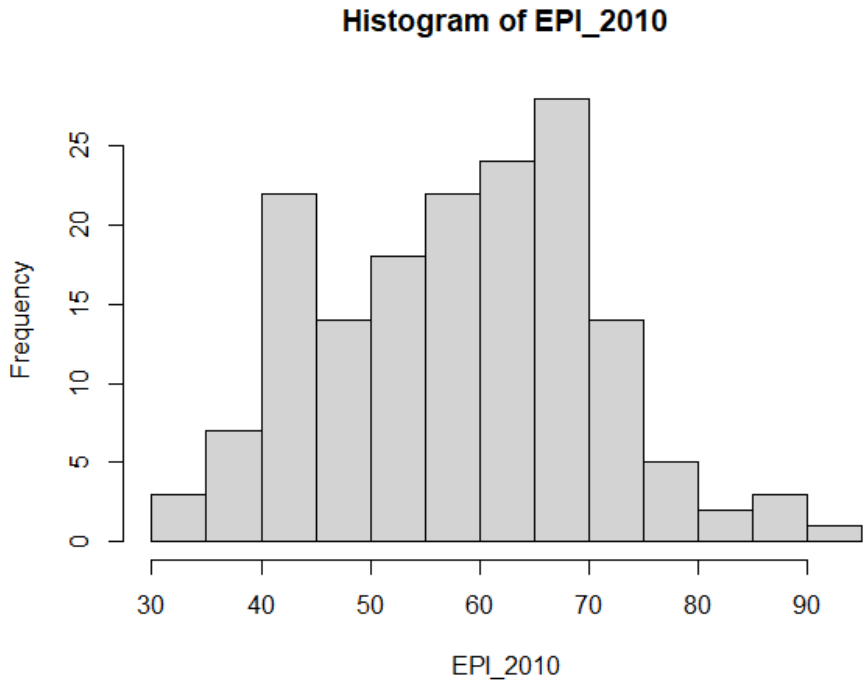
Part 1

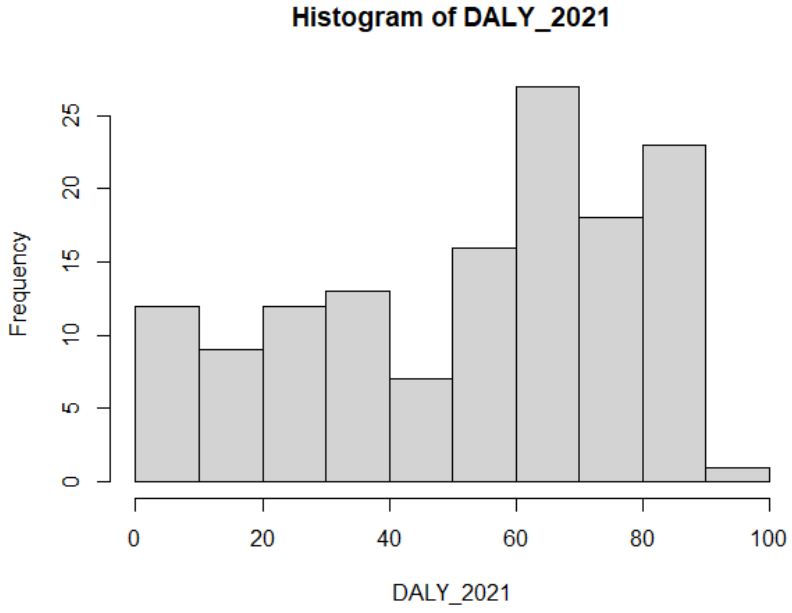
```
> EPI_data <- read.csv('EPI_data.csv')
> # viewing the data in another window
> View(EPI_data)
> # set the default object
> attach(EPI_data)
> fix(EPI_data)
> # assign values and filter out the null objects
> EPI <- EPI_data$EPI[!is.na(EPI)]
> DALY <- EPI_data$DALY[!is.na(DALY)]
> # filters out NA values, new array
> summary(EPI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.10   48.60   59.20   58.37   67.60   93.50
> summary(DALY)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   37.19   60.35   53.94   71.97   91.50
> # summary statistic
> mean(EPI, na.rm = "TRUE")
[1] 58.37055
> median(EPI, na.rm = "TRUE")
[1] 59.2
> mean(DALY, na.rm = "TRUE")
[1] 53.94313
> median(DALY, na.rm = "TRUE")
[1] 60.35
> # box plot and qqplot
> boxplot(EPI, DALY)
> qqplot(EPI, DALY)
```






```
> ######################## for the 2010 EPI dataset ####################
> EPI_data_2010 <- read.csv('2010EPI_data.csv',skip=1)
> attach(EPI_data_2010)
> fix(EPI_data_2010)
> # assign values and filter out the null objects
> EPI_2010 <- EPI_data_2010$EPI[!is.na(EPI)]
> DALY_2021 <- EPI_data_2010$DALY[!is.na(DALY)]
> mean(EPI_2010, na.rm = "TRUE")
[1] 58.37055
> median(EPI_2010, na.rm = "TRUE")
[1] 59.2
> mean(DALY_2021, na.rm = "TRUE")
[1] 53.62466
> median(DALY_2021, na.rm = "TRUE")
[1] 60.35
> # histograms
> hist(EPI_2010)
> hist(DALY_2021)
```


Histogram of EPI_2010
Frequency
0
5
10
15
20
25
30
40
50
60
70
80
90
EPI_2010


Histogram of DALY_2021
Frequency
0
5
10
15
20
25
0
20
40
60
80
100
DALY_2021

Part 2

```
> ################### exercise1 ###################
> multiReg <- read.csv("dataset_multipleRegression.csv")
> attach(multiReg)
> View(multiReg)
> # using the linear model
> linearModel <- lm(ROLL ~ UNEM + HGRAD)
> UNEM = c(7)
> HGRAD = c(90000)
> INC = c(25000)
> UNEM_HGRAD <- data.frame(UNEM,HGRAD)
> predictedRoll1 <- predict(linearModel, UNEM_HGRAD, interval='prediction')
> predictedRoll1
       fit      lwr      upr
1 81437.04 68082.31 94791.78
> UNEM_HGRAD_INC <- data.frame(UNEM, HGRAD, INC)
> predictedRoll2 <- predict(linearModel, UNEM_HGRAD_INC, interval='prediction')
> predictedRoll2
       fit      lwr      upr
1 81437.04 68082.31 94791.78
> detach(multiReg)
```

```
> ################### exercise2 ###################
> library(class)
> abalone <- read.csv("abalone.csv")
> attach(abalone)
> View(abalone)
> abalone$Rings <- as.numeric(abalone$Rings)
> abalone$Sex <- NULL
> ind <- sample(2,nrow(abalone), replace=TRUE, prob = c(0.8, 0.2))
> trainData <- abalone[ind==1,]
> testData <- abalone[ind==2,]
> # using the knn model
> KNNpred <- knn(train=trainData[1:7],test=testData[1:7],cl=trainData$Rings,k=55)
> table(KNNpred)
KNNpred
  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  27  29
  0   0  14  32  62  93 163 171 196 113  14  10   1   6   1   0   0   0   0   0   0   0   0   0   0   0
> detach(abalone)
```

```
> ################### exercise3 ###################
> library(ggplot2)
> View(iris)
> sapply(iris[,-5],var)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.6856935    0.1899794    3.1162779    0.5810063
> k.max<-10
> wss <- sapply(1:k.max,function(k){kmeans(iris[,3:4],k,nstart=20,iter.max=1000)$tot.withinss})
> wss
 [1] 550.895333  86.390220  31.371359  19.465989  13.916909  11.025145   9.185076   7.615402   6.45649
> plot(1:k.max,wss,type="b",xlab = "Number of clusters(k)",ylab="within cluster sum of squares")
> icluster<-kmeans(iris[,3:4],3,nstart=20)
> table(icluster$cluster,iris$Species)

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