Assignment 1

Ishaan Yadav

Q1.Solve questions or ‘airquality-dataset’ Import/use airquality data

df<-airquality  
head(df)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

How many observations are there in the given data?

nrow(df)

## [1] 153

Print last two rows of the data.

tail(df,2)

## Ozone Solar.R Wind Temp Month Day  
## 152 18 131 8.0 76 9 29  
## 153 20 223 11.5 68 9 30

1. What is the value of Ozone in 47th row?

df[47,1]

## [1] 21

1. How many values are missing in Ozone column?

sum(is.na(df$Ozone))

## [1] 37

1. What is the mean of Ozone column excluding missing values?

mean(df$Ozone,na.rm=T)

## [1] 42.12931

1. What is the mean of “Temp” when “Month” is equal to 6?

#subset1<-subset(df,Month==6)  
#mean(subset1$Temp)  
mean(subset(df,Month==6)$Temp)

## [1] 79.1

1. Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?

subset2<-subset(df,(Ozone>31)&(Temp>90))  
mean(subset2$Solar.R)

## [1] 212.8

1. What was the maximum ozone value in the month of May (i.e. Month is equal to 5)?

max(subset(df,Month==5)$Ozone,na.rm = T)

## [1] 115

Second question

1. Write a single R code to display the answers the following questions

Case Study: Hair Eye color Data set 1. How many people have brown eye color?

df1<-read.csv("C:/Users/mardi/Downloads/Expt1-Datset\_2\_hair\_eye\_color\_csv.csv")  
head(df1)

## Person.No. Hair.Color Eye.Color  
## 1 1 Blonde Blue  
## 2 2 Blonde Blue  
## 3 3 Blonde Green  
## 4 4 Blonde Brown  
## 5 5 Blonde Brown  
## 6 6 Blonde Black

nrow(subset(df1,Eye.Color=="Brown"))

## [1] 10

1. How many people have Blonde hair?

nrow(subset(df1,Hair.Color=="Blonde"))

## [1] 6

1. How many Brown haired people have Black eyes?

nrow(subset(df1,Hair.Color=="Brown" & Eye.Color=="Black"))

## [1] 2

1. What is the percentage of people with Green eyes?

Green=nrow(subset(df1,Eye.Color=="Green"))  
Total=nrow(df1)  
Ans=(Green/Total)\*100  
Ans

## [1] 10

1. What percentage of people have red hair and Blue eyes?

Both=nrow(subset(df1,Eye.Color=="Blue" & Hair.Color=="Red"))  
Total=nrow(df1)  
Answer=(Both/Total)\*100  
Answer

## [1] 5

1. Write a single R code to display the answers for the following questions

Case study: Germination Data Set

df2<-read.csv("C:/Users/mardi/Downloads/Expt1\_ Dataset3\_germination\_csv.csv")  
head(df2)

## Box water\_amt germinated  
## 1 Uncovered 1 22  
## 2 Uncovered 1 25  
## 3 Uncovered 1 27  
## 4 Uncovered 1 23  
## 5 Uncovered 2 41  
## 6 Uncovered 2 46

1. What is the average number of seeds germinated for the uncovered boxes with level of watering equal to 4?

WaterLevel=subset(df2,water\_amt==4 & Box=="Uncovered")  
List1=sum(WaterLevel[,3])/nrow(WaterLevel)  
List1

## [1] 78

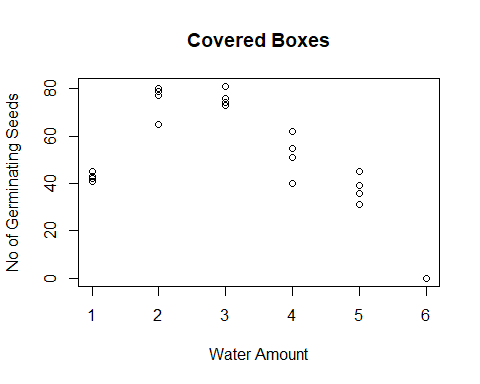
1. What is the median value seeds germinated for covered boxes?

Covered=subset(df2,Box=="Covered")  
FinalAnswer=median(Covered[,3],na.rm=T)  
FinalAnswer

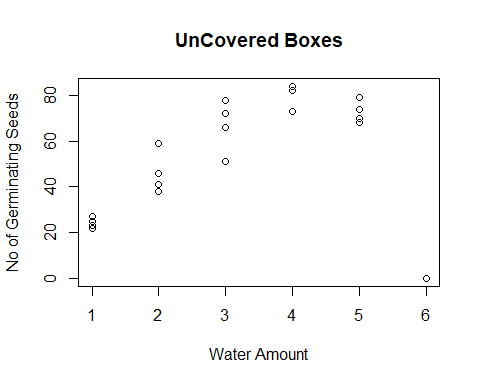
## [1] 45

Establish conclusions on the basis of available data and write them in the conclusion part. a. Association of levels of watering with the number of germinating seeds in case of covered boxes as well as uncovered boxes.

Subset1=subset(df2,Box=="Covered")  
Subset2=subset(df2,Box=="Uncovered")  
plot(Subset1$water\_amt,Subset1$germinated,xlab="Water Amount",ylab="No of Germinating Seeds",main="Covered Boxes")



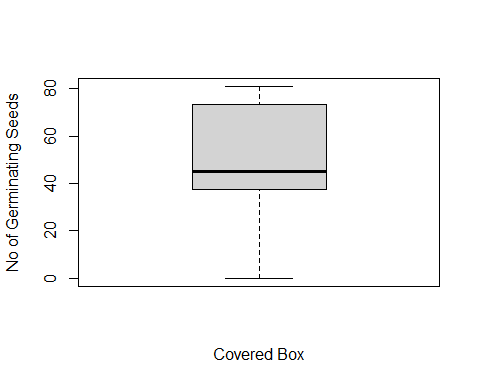
plot(Subset2$water\_amt,Subset2$germinated,xlab="Water Amount",ylab="No of Germinating Seeds",main="UnCovered Boxes")



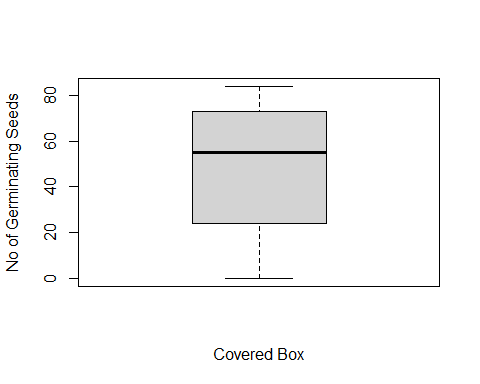
# Conclusion :-

1.For covered boxes, as the water level increases from 1 to 3, the number of germinating seed increase, while as water level further increases from 4 to 6, the number of germinating seeds decrease.  
2. For uncovered boxes, as the water level increases from 1 to 4, the number of germinating seed increase, while as water level further increases from 5 to 6, the number of germinating seeds decrease.  
3.More number of seeds germinate in uncovered box than covered box

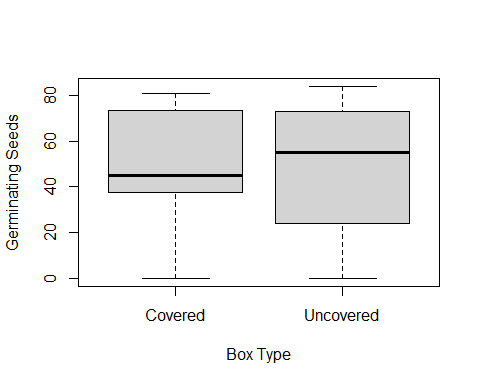
boxplot(Subset1$germinated,xlab="Covered Box",ylab="No of Germinating Seeds")



boxplot(Subset2$germinated,xlab="Covered Box",ylab="No of Germinating Seeds")



boxplot(df2$germinated~df2$Box,xlab="Box Type",ylab="Germinating Seeds")



# conclusion:-

As compared to covered boxes, the median and total number of germinating seeds in uncovered boxes is more

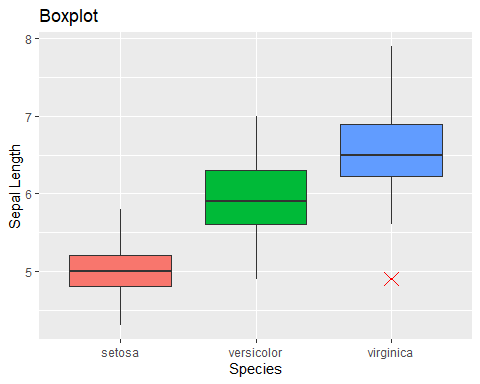
1. Write a single R code :
2. To display the Boxplot for sepal length of iris data set as shown below
3. To display the Scatter plot for murders data set present in “dslabs” package as shown below.

Give proper title, x,y axis label etc. to each plot.

df3=iris  
head(df3)

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

library(ggplot2)  
ggplot(data=df3,mapping=aes(x=Species,y=Sepal.Length,fill=Species))+geom\_boxplot(outlier.colour = "red",outlier.shape = 4,outlier.size = 4,outlier.stroke = 1,show.legend = F)+ggtitle("Boxplot")+xlab("Species")+ylab("Sepal Length")



library(dslabs)  
ggplot(data=murders,mapping=aes(x=population/1000000,y=total,color=region,label=abb))+geom\_point()+scale\_x\_log10()+scale\_y\_log10()+xlab("POPULATION")+ylab("TOTAL")+ggtitle("SCATTERPLOT")+geom\_text(color="black",size=2.5,nudge\_x=0.08)

