Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Project
Marks: 20 Time: 90 minutes
Name: Deepak Thakur
Reg. No: 19-05-5046 Dept: Entomology

1. **Short Questions** **(6*1=06)**
   a) In R, you can use install. Packages () to install a package from CRAN.
   b) To check the structure of an object in R, the function str () is used.
   c) To subset a data frame by selecting specific rows and columns, the [] operator is used.
   d) In R, the summary() function provides a summary of key descriptive statistics
   e) In R, the na.omit () function can be used to remove missing values (NA) from a vector x.
   f) The residuals of a regression model are the differences between the observed values and the fitted values predicted by the model.

2. For the ***iris*** data**:** **(7)**
   a) Calculate descriptive statistics ($\boldsymbol{median \pm SD, mean, CV}$) for each numeric variable in a single table.

```
iris_summary <- data.frame(
  Variable = names(iris)[1:4],
  Median = sapply(iris[, 1:4], median),
  SD = sapply(iris[, 1:4], sd),
  Mean = sapply(iris[, 1:4], mean),
  CV = sapply(iris[, 1:4], sd) / sapply(iris[, 1:4], mean) * 100  )
"Median ± SD"
iris_summary$Median <- paste0(iris_summary$Median, " ± ", iris_summary$SD)
```

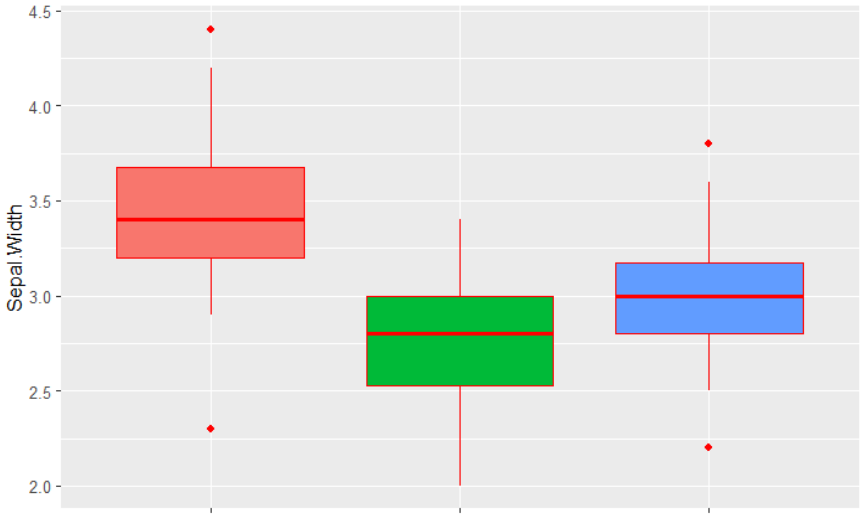
```
# for the table
table_data <- iris_summary[, c("Variable", "Median", "Mean", "CV")]
write.table(table_data, "iris_descriptive_stats.txt", row.names = FALSE, sep = "\t")
print(table_data)
```

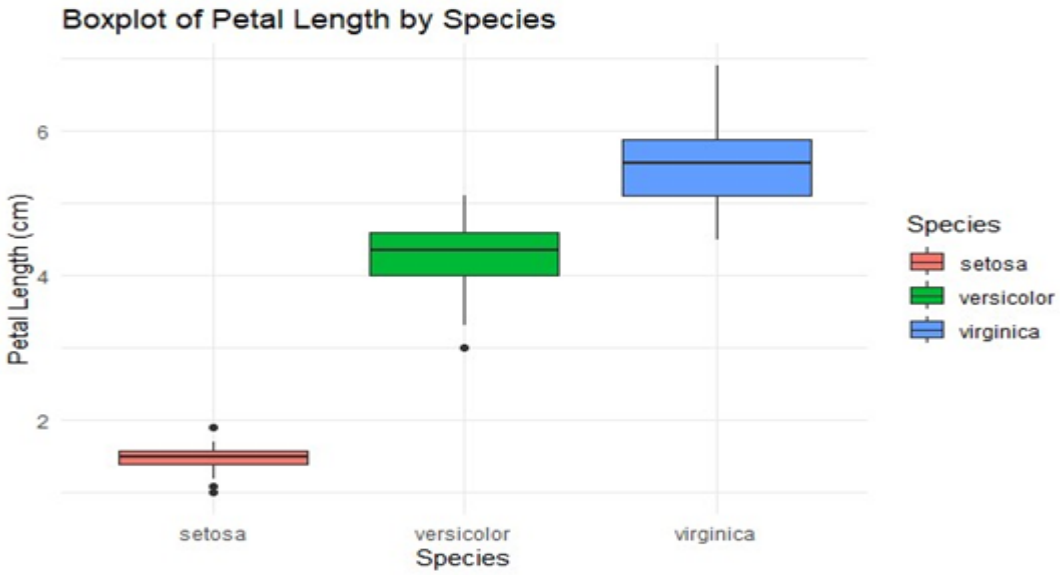
Variable	**Median ± SD**	**Mean**	**CV**
Sepal.Length	5.8 ± 0.83	5.843333	0.1417113
Sepal.Width	3 ± 0.44	3.057333	0.1424642
Petal.Length	4.35 ± 1.77	3.758000	0.4697441
Petal.Width	1.3 ± 0.76	1.199333	0.6355511

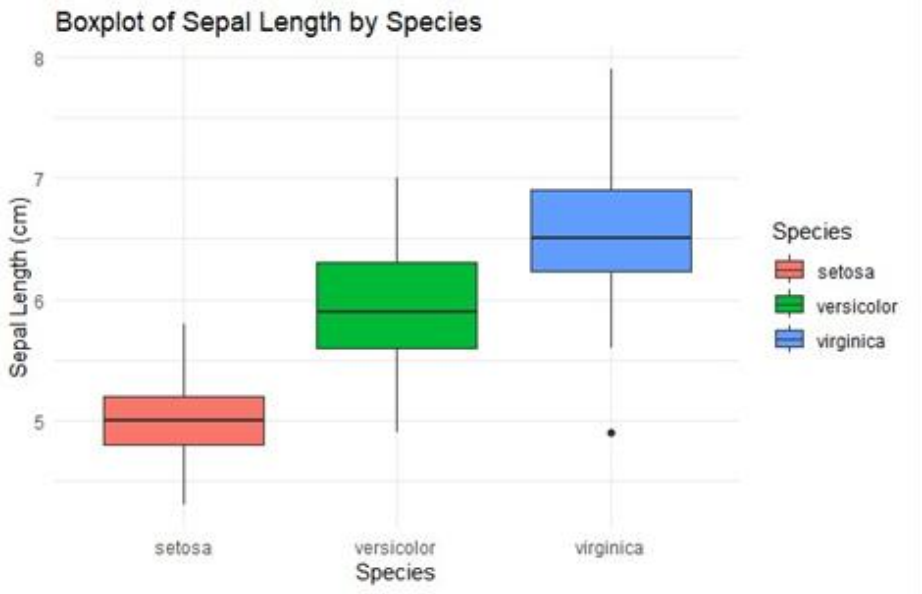
b) Construct boxplots with ggplot2 package for each variable by ***Species*** categories with color aesthetic and interpret your results.

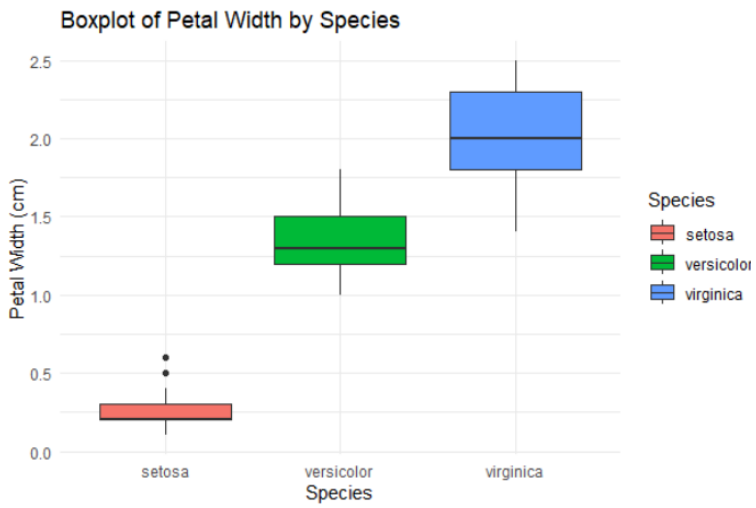
```
data<-iris
library(ggplot2)
library(ggExtra)
ggplot(data= iris,
      aes(x=Species,y= Sepal.Width,fill = Species)) +
geom_boxplot(show.legend = FALSE, colour= "red")
ggplot(data= iris,
    aes(x=Species,y= petal.length,fill = Species)) +
geom_boxplot(show.legend = FALSE, colour= "red")

ggplot(data= iris,
    aes(x=Species,y= Sepal.length,fill = Species)) +
geom_boxplot(show.legend = FALSE, colour= "red")
ggplot(data= iris,
    aes(x=Species,y= petal.width,fill = Species)) +
geom_boxplot(show.legend = FALSE, colour= "red")
```

Box plot sepal width by species









Interpretation

Median: The median Sepal.Width for Setosa is approximately 3.4. IQR: The IQR ranges roughly from 3.2 to 3.8.

Whiskers and Outliers: The whiskers extend from about 2.3 to 4.4. There is one outlier below the lower whisker (around 2.0).

Versicolor:

Median: The median Sepal.Width for Versicolor is approximately 2.8. IQR: The IQR ranges from about 2.5 to 3.0.

Whiskers and Outliers: The whiskers extend from about 2.0 to 3.4. There are no significant outliers outside the whiskers.

Virginica:

Median: The median Sepal.Width for Virginica is approximately 3.0. IQR: The IQR ranges from about 2.7 to 3.4.

Whiskers and Outliers: The whiskers extend from about 2.2 to 3.8. There is one outlier above the upper whisker (around 4.2).

Central Tendency: Setosa has the highest median Sepal.Width, followed by Virginica, and

Versicolor has the lowest.

Spread: Setosa has the widest IQR, indicating greater variability in Sepal.Width. Versicolor has the narrowest IQR, indicating the least variability.

Outliers: Both Setosa and Virginica have outliers, while Versicolor does not have any significant outliers.

Summary:

Setosa: Generally has wider sepals with higher variability and one lower outlier. Versicolor: Has narrower sepals with the least variability and no significant outliers.

Virginica: Has sepal widths that are in between Setosa and Versicolor, with one upper outlier. This boxplot effectively summarizes the distribution of Sepal.Width for the three species, highlighting differences in central tendency, variability, and outliers.

Setosa has wider sepals with higher variability and one lower outlier and upper outlier

Versicolor has narrower sepals with the least variability and no significant outliers.

Virginica has sepal widths that are in between Setosa and Versicolor, with one upper outlier and lower outlier.
Furthermore, **Setosa** has the highest median Sepal. Width, followed by **Virginica,** and **Versicolor** has the lowest. **Setosa** has the widest IQR, indicating greater variability in Sepal. Width**.** **Versicolor** has the narrowest IQR, indicating the least variability.

3. For the provided dataset of "***veg"***, answer the following questions: **(7)**

a) Identify missing values in each variable and impute them using the mean values of the corresponding variables.

```
veg<-read.csv("veg.csv")

str(veg)

summary(veg)

is.na(veg)

table(is.na(veg))

which(is.na(veg))

D<-na.omit(veg)

veg$Length.of.vine..cm.[is.na(veg$Length.of.vine..cm.)]<-mean(veg$Length.of.vine..cm.,na.rm = TRUE)

veg$Length.of.vine.internodes..cm.[is.na(veg$Length.of.vine.internodes..cm.)]<-mean(veg$Length.of.vine.internodes..cm,na.rm = TRUE)

veg$Petiole.length..cm.[is.na(veg$Petiole.length..cm.)]<-mean(veg$Petiole.length..cm.,na.rm = TRUE)

veg$Number.of.branches..main.[is.na(veg$Number.of.branches..main.)]<-mean(veg$Number.of.branches..main.,na.rm = TRUE)

veg$Number.of.days.required.for.maturity[is.na (veg$Number.of.days.required.for.maturity)]<-mean(veg$Number.of.days.required.for.maturity,na.rm = TRUE)

summary(veg)
```

b) Fit a suitable multiple linear regression model for the dataset and interpret your findings.

```
cor(veg)

cor(veg,method=c('pearson')) ### pearson correlation

cor(veg,method=c('spearman')) ### spearman correlation

pairs(veg)

x<-veg[,2]

y<-veg[,5]


#multiple regression

data1<-read.csv ('veg.csv', header=TRUE)

model1<-lm (log~. Data=data1)

summary (model1)
```

Result:

```
lm (formula = Yield.per.plot..kg. ~, data = vegetables)

Residuals:

Min     1Q Median     3Q    Max

-2.747 -0.490 -0.191  0.054 68.808

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept)                    0.90499   1.13057  0.800   0.424

Length.of.vine..cm.                0.25102   0.31664   0.793   0.428

Length.of.vine.internodes..cm.       0.41308   0.26943   1.533   0.126

Petiole.length..cm.                -0.21562   0.11062 -1.949   0.052.

 Number.of.leaves.per.plant          0.09696   0.24164   0.401   0.688

Number.of.branches..main.          -0.07477   0.15906 -0.470   0.639


Number.of.days.required.for.maturity 0.03758   0.19331   0.194   0.846


Number.of.tubers.per.plant          0.16784   0.13101   1.281   0.201
```

Signif.Codes: 0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 3.448 on 408 degrees of freedom

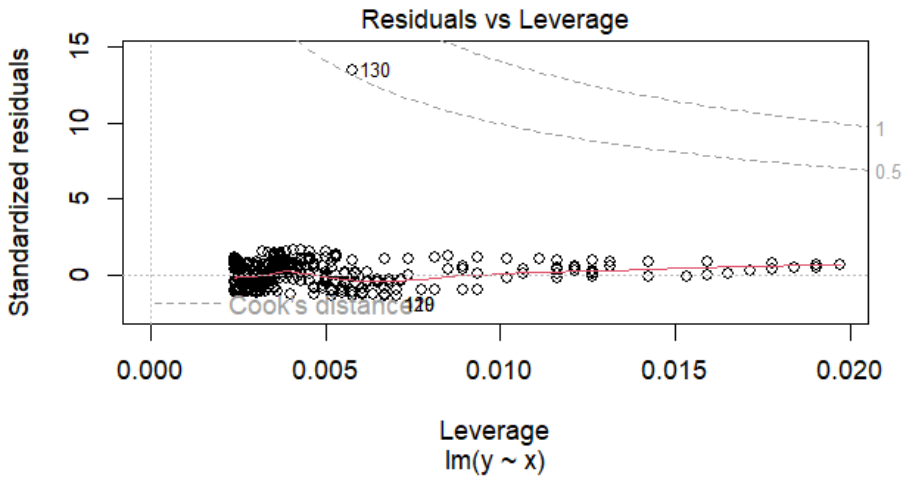
Multiple R-squared: 0.1208, Adjusted R-squared: 0.1057

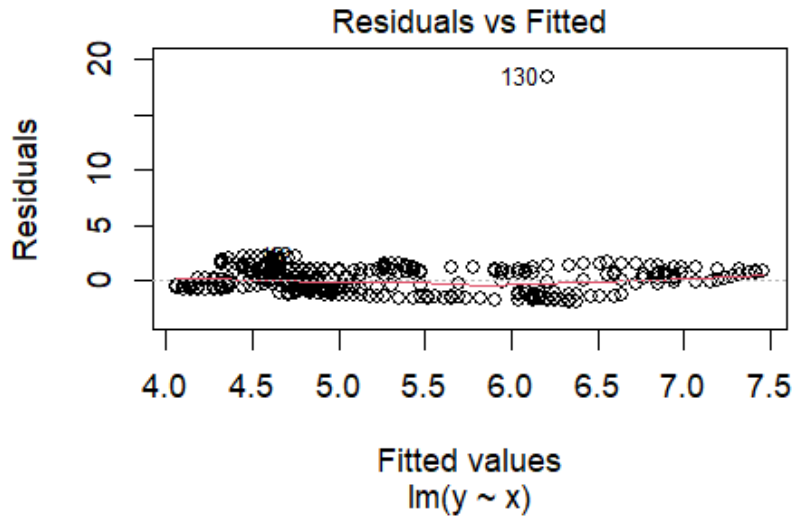
F-statistic: 8.008 on 7 and 408 DF, p-value: 3.976e-0

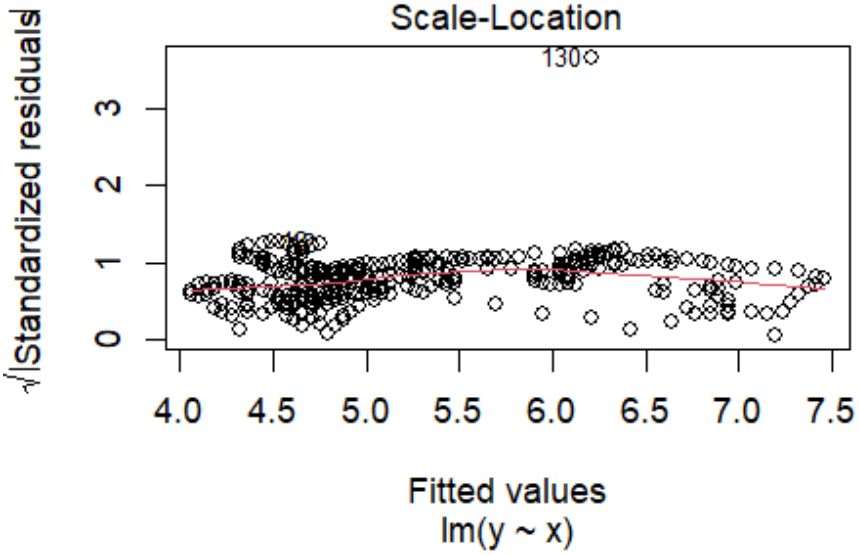
AIC(model1)

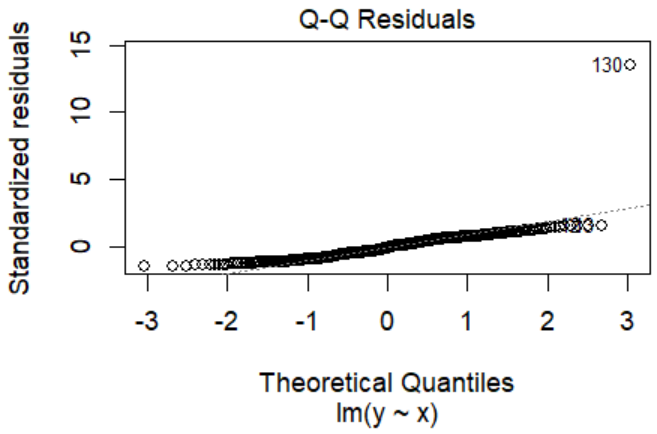
abline(lm(log~.,data=data1))

plot(model1)




Residuals vs Fitted
130
Residuals
20
10
5
0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(y ~ x)


Scale-Location
130
√|Standardized residuals|
3
2
1
0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(y ~ x)



Interpretation

The summary() function provides key insights

Coefficients:

Positive coefficients indicate an increase in the response variable (yield) for an increase in the predictor.

Negative coefficients indicate a decrease in yield.

Statistical Significance:

The Pr(>|t|) column shows p-values. Variables with p-values < 0.05 are statistically significant predictors of yield.

Adjusted R-squared:

Represents the proportion of variance in the dependent variable (yield) explained by the predictors.

A higher value indicates a better fit.

F-statistic:

A low p-value indicates the model is significant