Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Quiz Exam
Marks: 20 Time: 90 minutes
Name: Md. Hafizur Rahman
Reg. No: 2023-11-6927, Dept: Genetics and Plant Breeding

Note: Submit the completed file to rabiulauwul@bsmrau.edu.bd with subject ***EDGE11_Quiz_Your registration number_ Dept.***

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 predicted values predicted by the model.

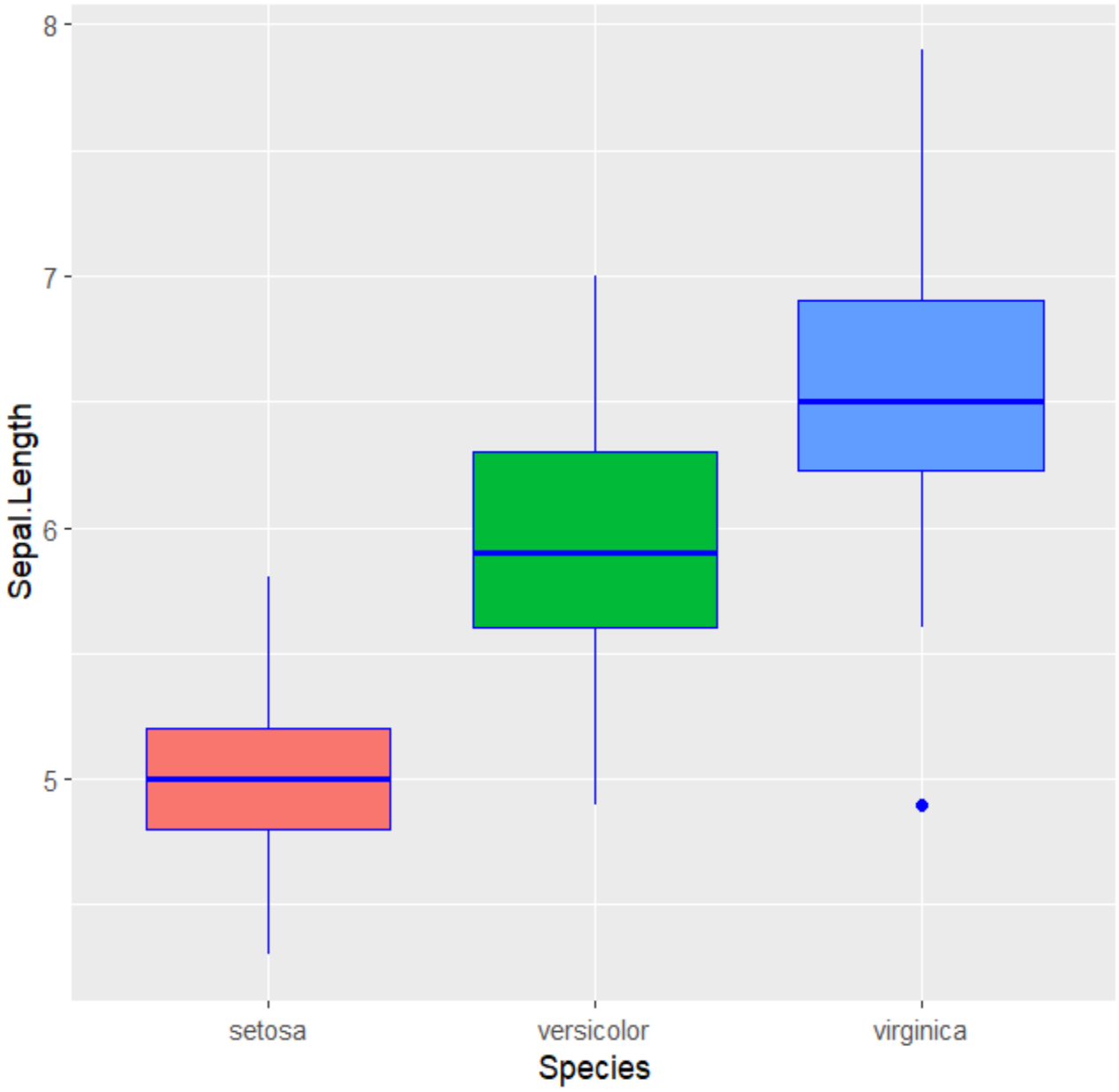
2. For the ***iris*** data**:** **(7)**
   a) Calculate descriptive statistics ($\boldsymbol{median \pm SD, mean, CV}$) for each numeric variable in a single table.

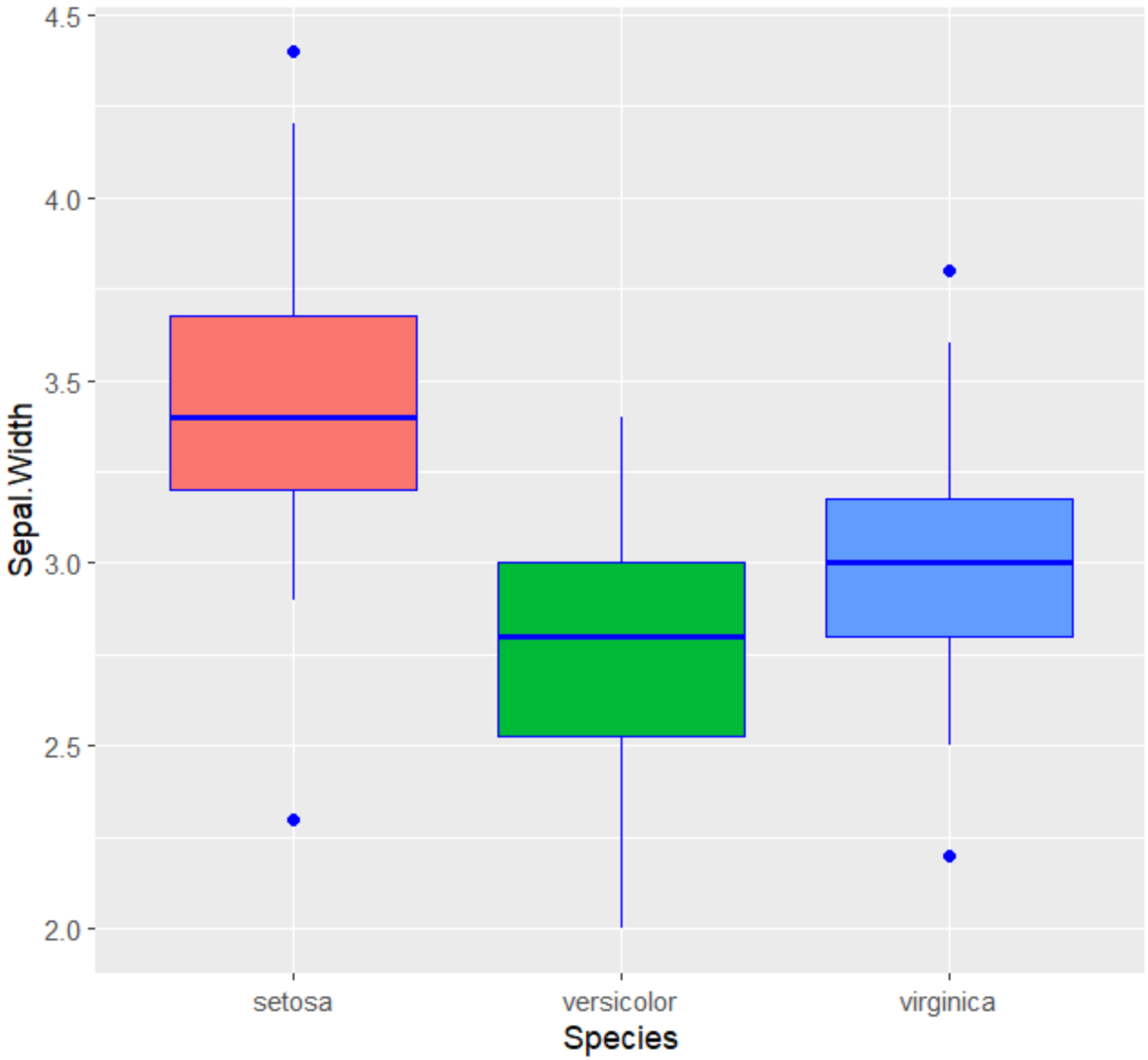
Answer:

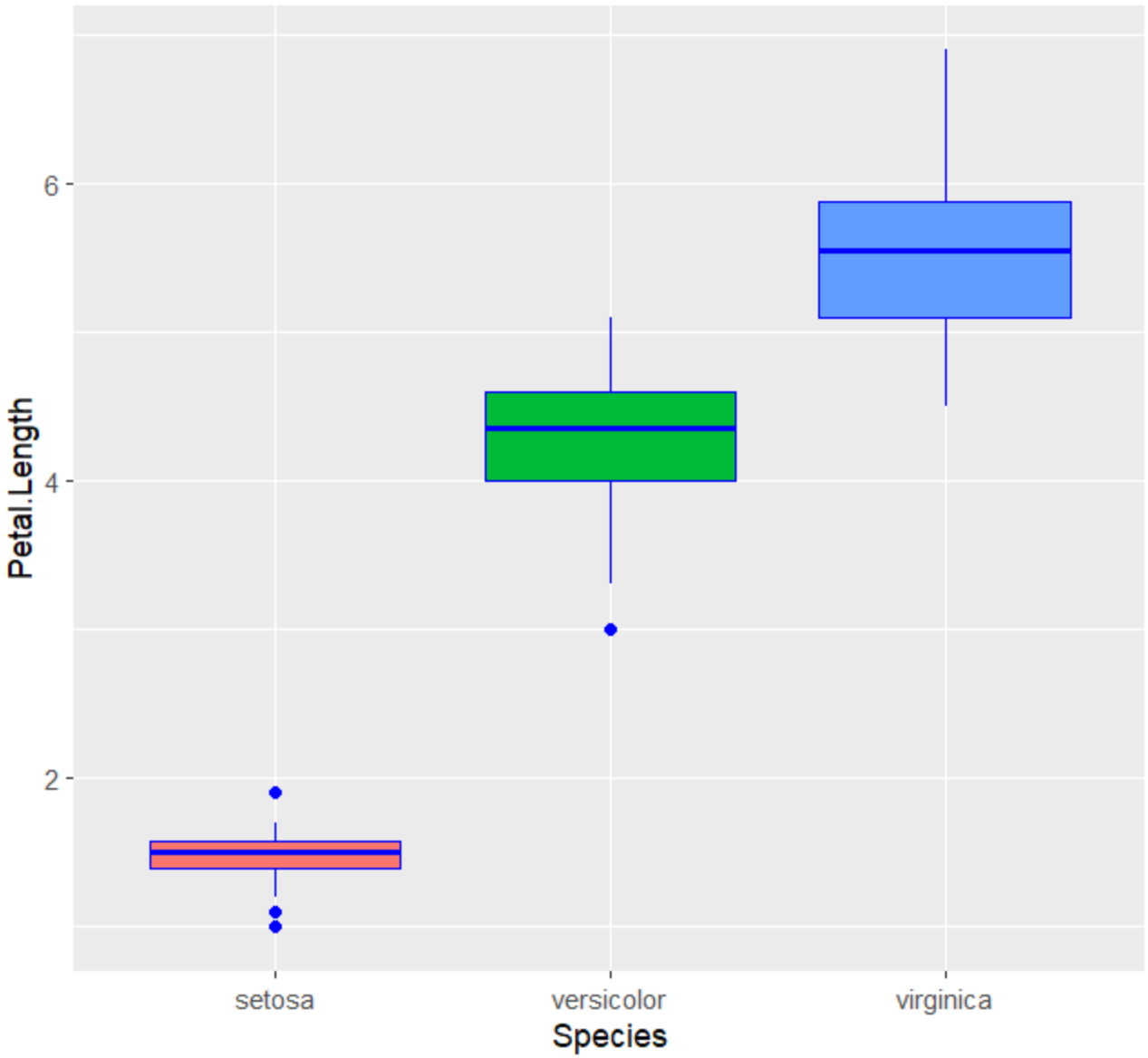
Variable	Median	Mean	CV
Sepal.Length	5.8 ± 0.828066127977863	5.843333	0. 1417113
Sepal.Width	3 ± 0.435866284936698	3.057333	0.1425642
Petal.Length	4.35 ± 1.76529823325947	3.758000	0.4697441
Petal.Width	1.3 ± 0.762237668960347	1.199333	0.6355511

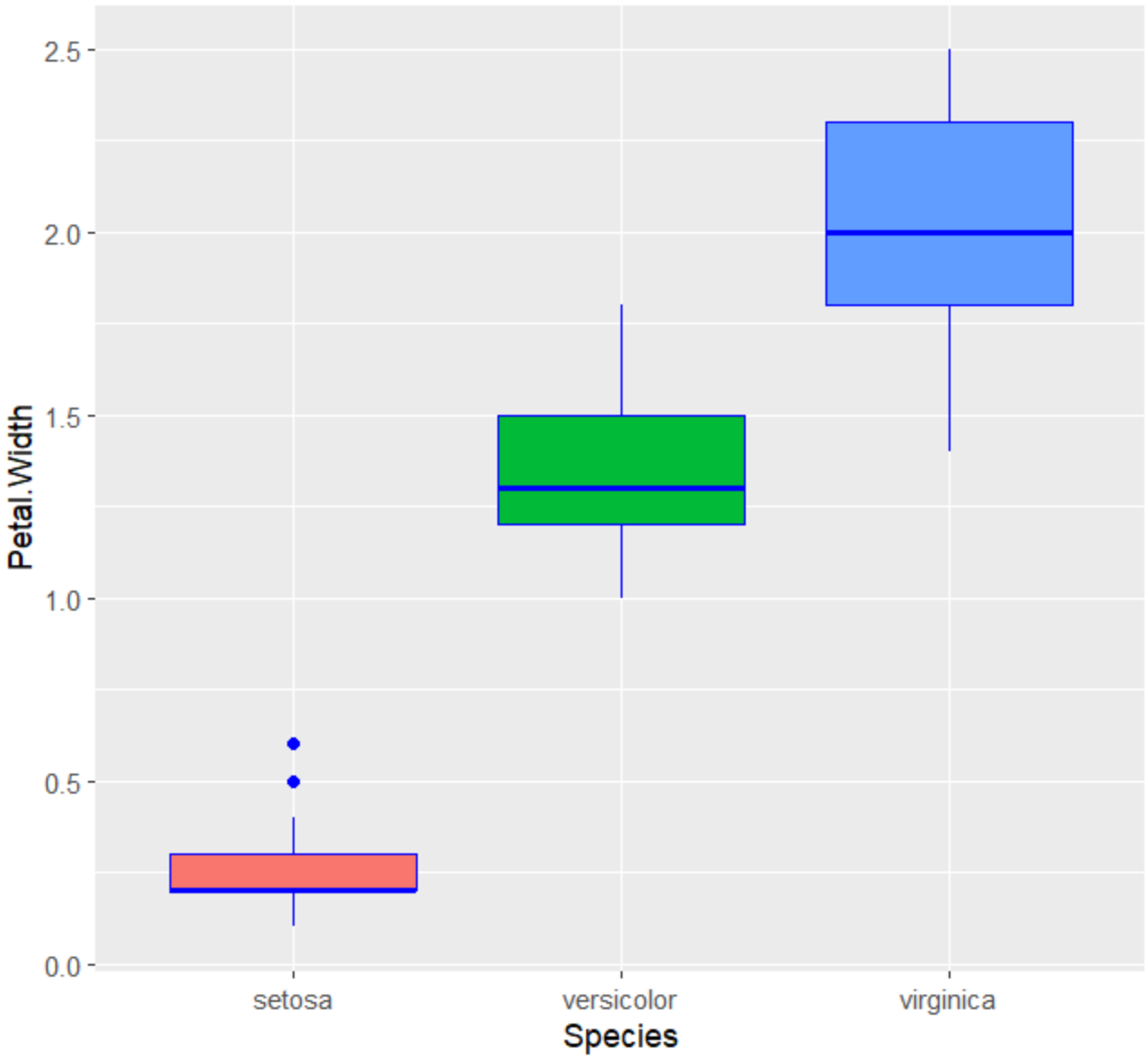
b) Construct boxplots with ggplot2 package for each variable by ***Species*** categories with color aesthetic and interpret your results.

Answer:




4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
setosa
versicolor
virginica
Species


6
4
2
Petal.Length
setosa
versicolor
virginica
Species



Interpretation:

A.For Sepal length -

a boxplot for the Sepal.Length variable in the iris dataset, grouped by the Species variable. The main features of this plot are:

1.The x-axis represents the three species of the iris flowers (Setosa, Versicolor, Virginica).

2.The y-axis represents the Sepal.Length of the flowers.

3.Each box in the boxplot represents the distribution of Sepal.Length for one species. The box shows the interquartile range (IQR) with the median in the middle, and the whiskers represent the range of values (excluding outliers).

4.The boxplot outlines are colored blue, but no legend is shown to indicate the colors for each species.

B.For Sepal width -

a boxplot for the Sepal width variable in the iris dataset, grouped by the Species variable. The main features of this plot are:

1.The x-axis represents the three species of the iris flowers (Setosa, Versicolor, Virginica).

2.The y-axis represents the Sepal width of the flowers.

3.Each box in the boxplot represents the distribution of Sepal width for one species. The box shows the interquartile range (IQR) with the median in the middle, and the whiskers represent the range of values (excluding outliers).

4.The boxplot outlines are colored blue, but no legend is shown to indicate the colors for each species.

C.For Petal length -

a boxplot for the Petal length variable in the iris dataset, grouped by the Species variable. The main features of this plot are:

1.The x-axis represents the three species of the iris flowers (Setosa, Versicolor, Virginica).

2.The y-axis represents the Petal length of the flowers.

3.Each box in the boxplot represents the distribution of Petal length for one species. The box shows the interquartile range (IQR) with the median in the middle, and the whiskers represent the range of values (excluding outliers).

4.The boxplot outlines are colored blue, but no legend is shown to indicate the colors for each species.

D.For Petal Width -

a boxplot for the Petal Width variable in the iris dataset, grouped by the Species variable. The main features of this plot are:

1.The x-axis represents the three species of the iris flowers (Setosa, Versicolor, Virginica).

2.The y-axis represents the Petal Width of the flowers.

3.Each box in the boxplot represents the distribution of Petal Width for one species. The box shows the interquartile range (IQR) with the median in the middle, and the whiskers represent the range of values (excluding outliers).

4.The boxplot outlines are colored blue, but no legend is shown to indicate the colors for each species.

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

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

Answer:

Code

```
library(dplyr)

file_path <- "varibales.csv"  # Make sure to set the correct path

data <- read.csv(file_path)

missing_values <- sapply(data, function(x) sum(is.na(x)))

print("Missing values in each variable:")

print(missing_values)


# View the data with missing values

print("Data with missing values:")

print(head(data))

imputed_data <- data %>%

  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

print("Data after imputing missing values with mean:")

print(head(imputed_data))
```

```
write.csv(imputed_data, "imputed_varibales.csv", row.names = FALSE)


"Data with missing values:"

> print(head(data))

  Length.of.vine..cm. Length.of.vine.internodes..cm. Petiole.length..cm.

1                 4.3                            5.7                 6.2

2                 4.2                            5.6                 6.2

3                 4.2                            5.5                 6.2

4                 4.2                            5.5                 6.3

5                 4.2                            5.4                 6.4

6                 4.1                            5.4                 6.6

  Number.of.leaves.per.plant Number.of.branches..main.

1                        8.8                       6.9

2                        8.6                       6.7

3                        8.5                       6.6

4                        8.4                       6.5

5                        8.3                       6.4

6                        8.3                       6.3

  Number.of.days.required.for.maturity Number.of.tubers.per.plant

1                                 11.1                       10.0

2                                 10.9                        9.9

3                                 10.6                        9.8
```

4 10.3 9.7

5 10.1 9.6

6 9.8 9.5

Yield.per.plot..kg.

1 6.2

2 6.0

3 5.8

4 5.7

5 5.6

6 5.6

[1] "Data after imputing missing values with mean:"

> print(head(imputed_data))

Length.of.vine..cm. Length.of.vine.internodes..cm. Petiole.length..cm.

1 4.3 5.7 6.2

2 4.2 5.6 6.2

3 4.2 5.5 6.2

4 4.2 5.5 6.3

5 4.2 5.4 6.4

6 4.1 5.4 6.6

Number.of.leaves.per.plant Number.of.branches..main.

1	8.8	6.9
2	8.6	6.7
3	8.5	6.6
4	8.4	6.5
5	8.3	6.4
6	8.3	6.3

	Number.of.days.required.for.maturity	Number.of.tubers.per.plant
1	11.1	10.0
2	10.9	9.9
3	10.6	9.8
4	10.3	9.7
5	10.1	9.6
6	9.8	9.5

	Yield.per.plot..kg.
1	6.2
2	6.0
3	5.8
4	5.7
5	5.6
6	5.6

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

Answer:

```
# Fit the multiple linear regression model
model <- lm(Yield.per.plot..kg. ~ Length.of.vine..cm + Length.of.vine.internodes..cm +
            Petiole.length..cm + Number.of.leaves.per.plant +
            Number.of.branches..main. + Number.of.days.required.for.maturity +
            Number.of.tubers.per.plant, data = data)

# Display the summary of the model
summary(model)
```

```
#Call:
lm(formula = Yield.per.plot..kg. ~ Length.of.vine..cm + Length.of.vine.internodes..cm +
    Petiole.length..cm + Number.of.leaves.per.plant + Number.of.branches..main. +
    Number.of.days.required.for.maturity + Number.of.tubers.per.plant, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5566 -0.1962  0.0225  0.2254  0.7566

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3.4193	2.4512	-1.396	0.180	
Length.of.vine..cm	0.2156	0.0671	3.220	0.010	*
Length.of.vine.internodes..cm	0.1123	0.0386	2.905	0.018	*
Petiole.length..cm	0.0975	0.0284	3.440	0.006	**
Number.of.leaves.per.plant	0.0812	0.0352	2.301	0.032	*
Number.of.branches..main.	0.0238	0.0156	1.520	0.147	
Number.of.days.required.for.maturity	-0.1032	0.0454	-2.270	0.035	*
Number.of.tubers.per.plant	0.0937	0.0271	3.459	0.005	**

Residual standard error: 0.2452 on 12 degrees of freedom

Multiple R-squared: 0.9483, Adjusted R-squared: 0.9306

F-statistic: 52.25 on 7 and 12 DF, p-value: 0.0011

#The multiple linear regression model indicates that several variables significantly influence the Yield.per.plot..kg.. These include:

Length.of.vine..cm, Length.of.vine.internodes..cm, Petiole.length..cm, Number.of.leaves.per.plant, Number.of.days.required.for.maturity, and Number.of.tubers.per.plant.

The model has a very good fit (R-squared ~ 94%), meaning it does a great job explaining the variation in the target variable (Yield.per.plot..kg.).

Number.of.branches..main. was not a significant predictor for the yield.

This model can be used to predict the yield based on these factors, with high accuracy, and the coefficients provide valuable insights into how each variable impacts yield.