Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Quiz Exam
Marks: 20 Time: 90 minutes
Name: Ferdous Ahmed Sagor
Reg. No: 2018-05-4783 Dept: AER

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.package ()…. () 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(data) …. () 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 ($\boldsymbol{median} \pm \boldsymbol{SD}, \boldsymbol{mean}, \boldsymbol{CV}$) for each numeric variable in a single table.

Answer: R code:

```
data(iris)

calc_stats <- function(x) {
  mean_val <- mean(x, na.rm = TRUE)
  median_val <- median(x, na.rm = TRUE)
  sd_val <- sd(x, na.rm = TRUE)
  cv_val <- sd_val / mean_val * 100  # Coefficient of Variation in percentage
  return(c(mean = mean_val, median = median_val, SD = sd_val, CV = cv_val))
}

numeric_cols <- iris[, 1:4]  # Selecting only the numeric columns

stats_table <- t(apply(numeric_cols, 2, calc_stats))
stats_table <- as.data.frame(stats_table)
stats_table$Variable <- rownames(stats_table)
rownames(stats_table) <- NULL

print(stats_table)
```

Table:

	Median±**SD**	Mean	CV
Sepal.Length	5.80 ± 0.8280661	5.843333	14.17113
Sepal.Width	3.00± 0.4358663	3.057333	14.25642
Petal.Length	4.35± 1.7652982	3.758000	46.97441
Petal.Width	1.30± 0.7622377	1.199333	63.55511

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

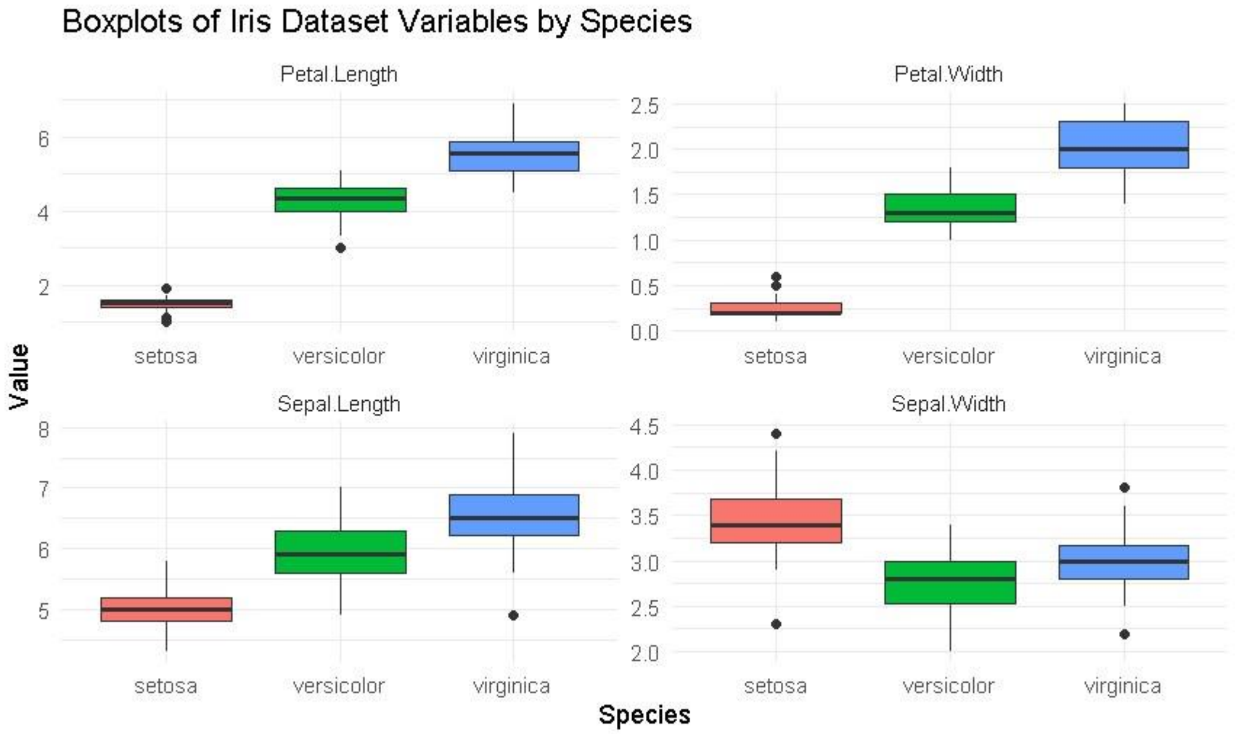
Answer: R code:

```
library(ggplot2)
library(tidyr)

iris_long <- iris %>%
  pivot_longer(cols = -Species, names_to = "Variable", values_to = "Value")

ggplot(iris_long, aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  facet_wrap(~Variable, scales = "free") +
  labs(
    title = "Boxplots of Iris Dataset Variables by Species",
    x = "Species",
    y = "Value"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

Results:



Interpretation:

Sepal Length:

Setosa: Smaller sepal lengths compared to other species with having outliers.

Versicolor and Virginica: Show overlap but with Virginica generally having outliers.

Variability in values increases across species.

Sepal Width:

Setosa: Displays higher and more consistent sepal widths with having outliers.

Versicolor and Virginica: Overlap more, with no clear distinction.

Petal Length:

Clear separation between species.

Setosa: Shorter petal lengths.

Virginica: Longest petals with Versicolor in between with having outliers.

Petal Width:

Similar trends as petal length with distinct groupings by species.

Setosa: Narrower petals with having outliers.

Virginica: Widest petals with having outliers

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: R code:

```
vegetables <- read.csv("1734953626384_vegitables.csv")

str(vegetables)
colSums(is.na(vegetables))



vegetables_imputed <- vegetables
vegetables_imputed[] <- lapply(vegetables_imputed, function(x) {
  if (is.numeric(x)) {
    x[is.na(x)] <- mean(x, na.rm = TRUE)
  }
  return(x)
})

colSums(is.na(vegetables_imputed))

Result: Length.of.vine..cm.      Length.of.vine.internodes..cm.
                          0                           0
            Petiole.length..cm.         Number.of.leaves.per.plant
                          0                           0
      Number.of.branches..main. Number.of.days.required.for.maturity
                          0                           0
     Number.of.tubers.per.plant               Yield.per.plot..kg.
                          0                           0
```

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

Answer: R code:

```
str(vegetables)


model <- lm(Yield.per.plot..kg. ~ ., data = vegetables)
```

```
summary(model)
```

```
par(mfrow = c(2, 2))
plot(model)
```

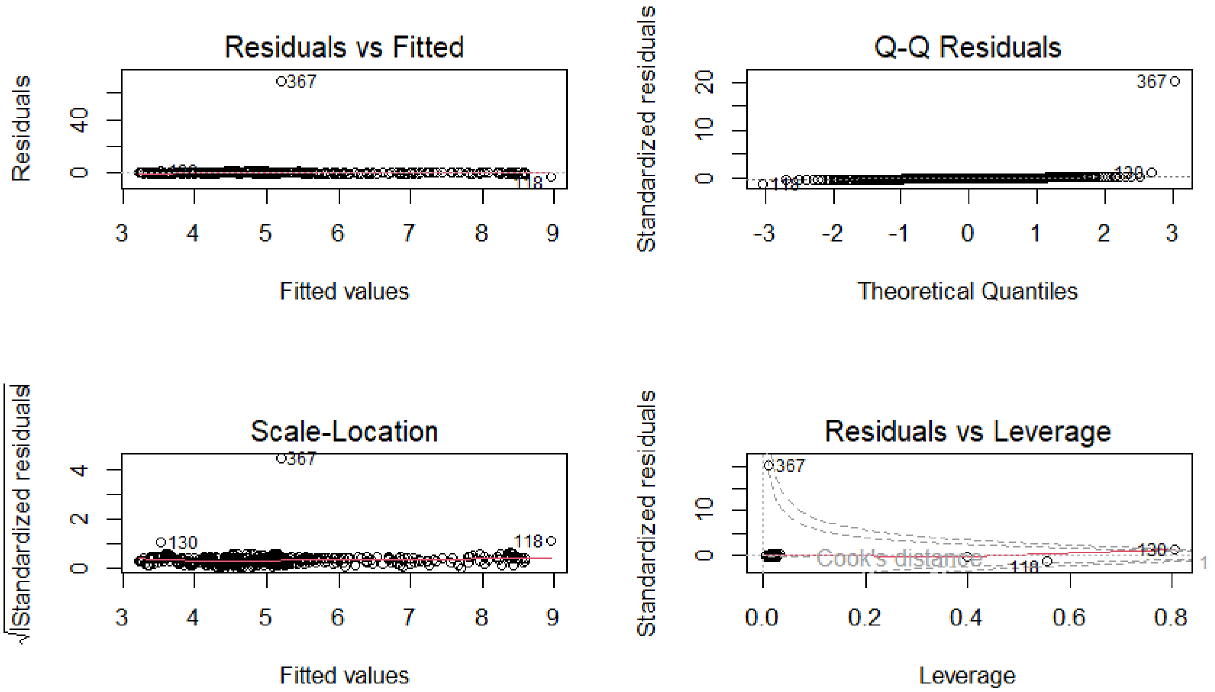
Result:

```
Call:
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-09
```



Interpreting the Findings:

The summary() function provides key insights:

Coefficients:

- The table lists the estimated coefficients for each predictor variable.
- 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:

- Tests whether the model provides a better fit than an intercept-only model. A low p-value indicates the model is significant.