**Bangabandhu Sheikh Mujibur Rahman Agricultural University**

**EDGE\_Batch-11**

**Quiz Exam**

**Marks: 20 Time: 90 minutes**

**Name: Md. Minhazul Arefin**

**Reg. No:2023-05-6534 Dept: Plant Pathology**

**Note**: Submit the completed file to [rabiulauwul@bsmrau.edu.bd](mailto:rabiulauwul@bsmrau.edu.bd) with subject ***EDGE11\_Quiz\_Your registration number\_ Dept.***

1. **Short Questions (6\*1=06)**
2. In R, you can use install.packages () to install a package from CRAN.
3. To check the structure of an object in R, the function str() is used.
4. To subset a data frame by selecting specific rows and columns, the [] operator is used.
5. In R, the summary() function provides a summary of key descriptive statistics
6. In R, the na.omit () function can be used to remove missing values (NA) from a vector x.
7. The residuals of a regression model are the differences between the observed values and the fitted values predicted by the model.
8. For the ***iris*** data**: (7)**
9. Calculate descriptive statistics (for each numeric variable in a single table.

Ans:

library(dplyr)

descriptive\_stats <- iris %>%

summarise(

Sepal.Length\_Mean = mean(Sepal.Length),

Sepal.Length\_Median\_SD = paste0(round(median(Sepal.Length), 2), " ± ", round(sd(Sepal.Length), 2)),

Sepal.Length\_CV = round((sd(Sepal.Length) / mean(Sepal.Length)) \* 100, 2),

Sepal.Width\_Mean = mean(Sepal.Width),

Sepal.Width\_Median\_SD = paste0(round(median(Sepal.Width), 2), " ± ", round(sd(Sepal.Width), 2)),

Sepal.Width\_CV = round((sd(Sepal.Width) / mean(Sepal.Width)) \* 100, 2),

Petal.Length\_Mean = mean(Petal.Length),

Petal.Length\_Median\_SD = paste0(round(median(Petal.Length), 2), " ± ", round(sd(Petal.Length), 2)),

Petal.Length\_CV = round((sd(Petal.Length) / mean(Petal.Length)) \* 100, 2),

Petal.Width\_Mean = mean(Petal.Width),

Petal.Width\_Median\_SD = paste0(round(median(Petal.Width), 2), " ± ", round(sd(Petal.Width), 2)),

Petal.Width\_CV = round((sd(Petal.Width) / mean(Petal.Width)) \* 100, 2)

)

descriptive\_stats <- as.data.frame(t(descriptive\_stats))

colnames(descriptive\_stats) <- c("Value")

descriptive\_stats

Result:

|  |  |
| --- | --- |
|  | Value |
| Sepal.Length\_Mean | 5.843333 |
| Sepal.Length\_Median\_SD | 5.8 ± 0.83 |
| Sepal.Length\_CV | 14.17 |
| Sepal.Width\_Mean | 3.057333 |
| Sepal.Width\_Median\_SD | 3 ± 0.44 |
| Sepal.Width\_CV | 14.26 |
| Petal.Length\_Mean | 3.758 |
| Petal.Length\_Median\_SD | 4.35 ± 1.77 |
| Petal.Length\_CV | 46.97 |
| Petal.Width\_Mean | 1.199333 |
| Petal.Width\_Median\_SD | 1.3 ± 0.76 |
| Petal.Width\_CV | 63.56 |

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

Ans:

library(ggplot2)

create\_boxplot <- function(variable\_name) {

ggplot(iris, aes\_string(x = "Species", y = variable\_name, fill = "Species")) +

geom\_boxplot() +

labs(

title = paste("Boxplot of", variable\_name, "by Species"),

x = "Species",

y = variable\_name

) +

theme\_minimal() +

scale\_fill\_brewer(palette = "Set2") # Adjust color palette

}

boxplot\_sepal\_length <- create\_boxplot("Sepal.Length")

boxplot\_sepal\_width <- create\_boxplot("Sepal.Width")

boxplot\_petal\_length <- create\_boxplot("Petal.Length")

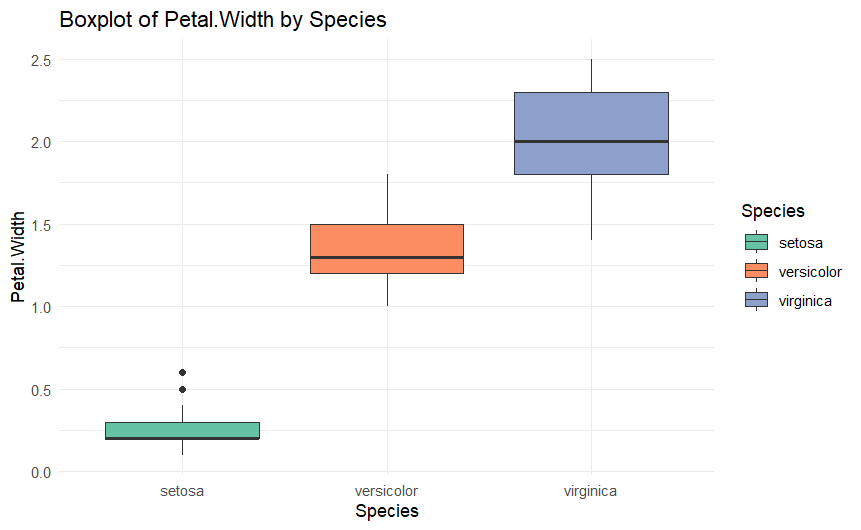
boxplot\_petal\_width <- create\_boxplot("Petal.Width")

print(boxplot\_sepal\_length)

print(boxplot\_sepal\_width)

print(boxplot\_petal\_length)

print(boxplot\_petal\_width)

Result: 

Interpreting result:

Sepal Length:

Setosa: Smaller sepal lengths compared to other species.

Versicolor and Virginica: Show overlap but with Virginica generally having larger values.

Variability in values increases across species.

Sepal Width:

Setosa: Displays higher and more consistent sepal widths.

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.

Petal Width:

Similar trends as petal length with distinct groupings by species.

Setosa: Narrower petals.

Virginica: Widest petals.

1. For the provided dataset of “***vegitables”***, answer the following questions: **(7)**
2. Identify missing values in each variable and impute them using the mean values of the corresponding variables.

Ans:

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

Explanation:

1. is.na(): Checks for missing values in the dataset.
2. mean(x, na.rm = TRUE): Calculates the mean of a column while ignoring missing values (na.rm = TRUE).
3. lapply(): Applies the imputation function to each column of the dataset.
4. Fit a suitable multiple linear regression model for the dataset and interpret your findings.

Ans:

Code:

vegetables <- read.csv("vegetables\_imputed.csv")

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

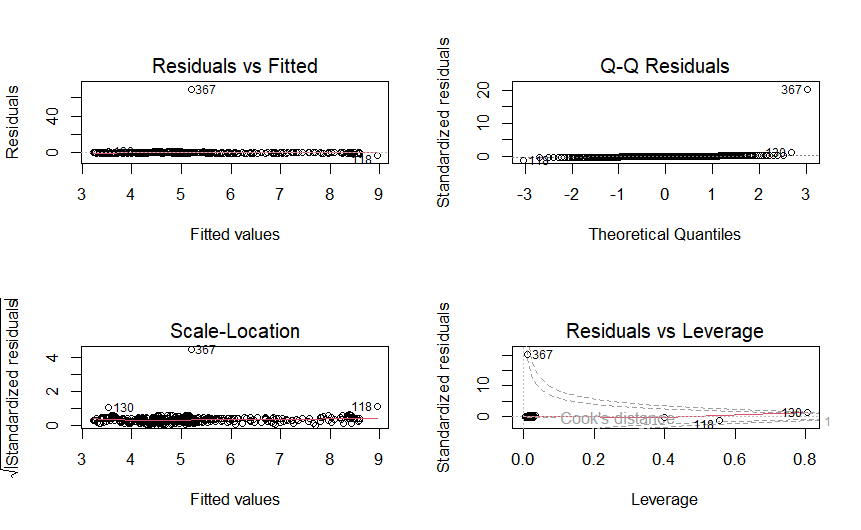
---

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.