> DataSet <- read.csv("E:\\Aiub\\11TH SEMESTER MID TERM\\INTRODUCTION TO DATASCIENCE\\Project\\Hachibur Rahman\\Depression Student Dataset.csv",

+ header = TRUE, na.strings = c("", "NA", "NaN", "NULL"))

>

> head(DataSet)

Gender Age Academic.Pressure Study.Satisfaction Sleep.Duration Dietary.Habits Have.you.ever.had.suicidal.thoughts.. Study.Hours Financial.Stress

1 Male 28 2 4 7-8 hours Moderate Yes 9 2

2 Male 28 4 5 5-6 hours Healthy Yes 7 1

3 Male 25 1 3 5-6 hours Unhealthy Yes 10 4

4 Male 23 1 4 More than 8 hours Unhealthy Yes 7 2

5 Female 31 1 5 More than 8 hours Healthy Yes 4 2

6 Male 19 4 4 5-6 hours Unhealthy Yes 1 4

Family.History.of.Mental.Illness Depression

1 Yes No

2 Yes No

3 No Yes

4 Yes No

5 Yes No

6 Yes Yes

> str(DataSet)

'data.frame': 502 obs. of 11 variables:

$ Gender : chr "Male" "Male" "Male" "Male" ...

$ Age : int 28 28 25 23 31 19 34 20 33 33 ...

$ Academic.Pressure : num 2 4 1 1 1 4 4 4 1 4 ...

$ Study.Satisfaction : num 4 5 3 4 5 4 2 1 4 3 ...

$ Sleep.Duration : chr "7-8 hours" "5-6 hours" "5-6 hours" "More than 8 hours" ...

$ Dietary.Habits : chr "Moderate" "Healthy" "Unhealthy" "Unhealthy" ...

$ Have.you.ever.had.suicidal.thoughts..: chr "Yes" "Yes" "Yes" "Yes" ...

$ Study.Hours : int 9 7 10 7 4 1 6 3 10 10 ...

$ Financial.Stress : int 2 1 4 2 2 4 2 4 3 1 ...

$ Family.History.of.Mental.Illness : chr "Yes" "Yes" "No" "Yes" ...

$ Depression : chr "No" "No" "Yes" "No" ...

> summary(DataSet)

Gender Age Academic.Pressure Study.Satisfaction Sleep.Duration

Length:502 Min. :18.00 Min. :1.000 Min. :1.000 Length:502

Class :character 1st Qu.:22.00 1st Qu.:2.000 1st Qu.:2.000 Class :character

Mode :character Median :26.50 Median :3.000 Median :3.000 Mode :character

Mean :26.24 Mean :3.004 Mean :3.076

3rd Qu.:30.00 3rd Qu.:4.000 3rd Qu.:4.000

Max. :34.00 Max. :5.000 Max. :5.000

Dietary.Habits Have.you.ever.had.suicidal.thoughts.. Study.Hours Financial.Stress

Length:502 Length:502 Min. : 0.000 Min. :1.000

Class :character Class :character 1st Qu.: 3.000 1st Qu.:2.000

Mode :character Mode :character Median : 7.000 Median :3.000

Mean : 6.404 Mean :2.928

3rd Qu.:10.000 3rd Qu.:4.000

Max. :12.000 Max. :5.000

Family.History.of.Mental.Illness Depression

Length:502 Length:502

Class :character Class :character

Mode :character Mode :character

> categorical\_cols <- c("Gender", "Sleep.Duration", "Dietary.Habits",

+ "Have.you.ever.had.suicidal.thoughts..",

+ "Family.History.of.Mental.Illness", "Depression")

>

> DataSet[categorical\_cols] <- lapply(DataSet[categorical\_cols], as.factor)

>

> # Confirm structure

> str(DataSet)

'data.frame': 502 obs. of 11 variables:

$ Gender : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 1 1 1 2 ...

$ Age : int 28 28 25 23 31 19 34 20 33 33 ...

$ Academic.Pressure : num 2 4 1 1 1 4 4 4 1 4 ...

$ Study.Satisfaction : num 4 5 3 4 5 4 2 1 4 3 ...

$ Sleep.Duration : Factor w/ 4 levels "5-6 hours","7-8 hours",..: 2 1 1 4 4 1 4 4 4 3 ...

$ Dietary.Habits : Factor w/ 3 levels "Healthy","Moderate",..: 2 1 3 3 1 3 2 1 2 3 ...

$ Have.you.ever.had.suicidal.thoughts..: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 1 2 ...

$ Study.Hours : int 9 7 10 7 4 1 6 3 10 10 ...

$ Financial.Stress : int 2 1 4 2 2 4 2 4 3 1 ...

$ Family.History.of.Mental.Illness : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 1 2 1 1 ...

$ Depression : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 2 2 1 2 ...

>

> cat\_cols <- names(DataSet)[sapply(DataSet, is.factor)]

> for (col in cat\_cols) {

+ cat("\n---", col, "---\n")

+ print(table(DataSet[[col]], useNA = "ifany"))

+ }

--- Gender ---

Female Male

235 267

--- Sleep.Duration ---

5-6 hours 7-8 hours Less than 5 hours More than 8 hours

123 128 123 128

--- Dietary.Habits ---

Healthy Moderate Unhealthy

161 172 169

--- Have.you.ever.had.suicidal.thoughts.. ---

No Yes

242 260

--- Family.History.of.Mental.Illness ---

No Yes

265 237

--- Depression ---

No Yes

250 252

|  |
| --- |
| > dup\_count <- sum(duplicated(DataSet))  > cat("Duplicate Rows:", dup\_count, "\n")  Duplicate Rows: 0  >  > if (dup\_count > 0) {  + DataSet <- DataSet[!duplicated(DataSet), ]  + }  >  > # 2. Convert all possible missing-like values to NA  > DataSet[DataSet == ""] <- NA  > DataSet[DataSet == "NA"] <- NA  > DataSet[DataSet == "NaN"] <- NA  > DataSet[DataSet == "NULL"] <- NA  >  > cat("\nMissing Values per Column:\n")  Missing Values per Column:  > print(colSums(is.na(DataSet)))  Gender Age  0 0  Academic.Pressure Study.Satisfaction  0 0  Sleep.Duration Dietary.Habits  0 0  Have.you.ever.had.suicidal.thoughts.. Study.Hours  0 0  Financial.Stress Family.History.of.Mental.Illness  0 0  Depression  0  >  > if ("Age" %in% colnames(DataSet)) {  + invalid\_age <- DataSet$Age < 18 | DataSet$Age > 120  + cat("\nInvalid Age Count:", sum(invalid\_age, na.rm = TRUE), "\n")  + DataSet$Age[invalid\_age] <- NA # convert invalid to NA  + }  Invalid Age Count: 0  >  > for (col in colnames(DataSet)) {  + if (is.numeric(DataSet[[col]])) {  + # Replace NA with mean (average)  + mean\_val <- mean(DataSet[[col]], na.rm = TRUE)  + DataSet[[col]][is.na(DataSet[[col]])] <- mean\_val  + } else {  + # Replace NA with most frequent value (mode)  + freq\_val <- names(sort(table(DataSet[[col]]), decreasing = TRUE))[1]  + DataSet[[col]][is.na(DataSet[[col]])] <- freq\_val  + }  + }  >  > cat("\nMissing Values after cleaning:\n")  Missing Values after cleaning:  > print(colSums(is.na(DataSet)))  Gender Age  0 0  Academic.Pressure Study.Satisfaction  0 0  Sleep.Duration Dietary.Habits  0 0  Have.you.ever.had.suicidal.thoughts.. Study.Hours  0 0  Financial.Stress Family.History.of.Mental.Illness  0 0  Depression  0  > |
|  |
| |  | | --- | | > | |

> numeric\_cols <- names(DataSet)[sapply(DataSet, is.numeric)]

>

> cat("\nOutlier Detection:\n")

Outlier Detection:

> for (col in numeric\_cols) {

+ Q1 <- quantile(DataSet[[col]], 0.25, na.rm = TRUE)

+ Q3 <- quantile(DataSet[[col]], 0.75, na.rm = TRUE)

+ IQR\_val <- Q3 - Q1

+ lower <- Q1 - 1.5 \* IQR\_val

+ upper <- Q3 + 1.5 \* IQR\_val

+ outlier\_count <- sum(DataSet[[col]] < lower | DataSet[[col]] > upper, na.rm = TRUE)

+ cat("Outliers in", col, ":", outlier\_count, "\n")

+ }

Outliers in Age : 0

Outliers in Academic.Pressure : 0

Outliers in Study.Satisfaction : 0

Outliers in Study.Hours : 0

Outliers in Financial.Stress : 0

>

> cat("\nMeasure of Central Tendency:\n")

Measure of Central Tendency:

> for (col in numeric\_cols) {

+ mode\_val <- as.numeric(names(sort(table(DataSet[[col]]), decreasing = TRUE))[1])

+ cat(col,

+ "\n Mean:", mean(DataSet[[col]], na.rm = TRUE),

+ "\n Median:", median(DataSet[[col]], na.rm = TRUE),

+ "\n Mode:", mode\_val, "\n\n")

+ }

Age

Mean: 26.24104

Median: 26.5

Mode: 28

Academic.Pressure

Mean: 3.003984

Median: 3

Mode: 3

Study.Satisfaction

Mean: 3.075697

Median: 3

Mode: 4

Study.Hours

Mean: 6.404382

Median: 7

Mode: 10

Financial.Stress

Mean: 2.928287

Median: 3

Mode: 1

> cat("\nMeasure of Spread:\n")

Measure of Spread:

> for (col in numeric\_cols) {

+ cat(col,

+ "\n SD:", sd(DataSet[[col]], na.rm = TRUE),

+ "\n Variance:", var(DataSet[[col]], na.rm = TRUE),

+ "\n IQR:", IQR(DataSet[[col]], na.rm = TRUE),

+ "\n Range:", diff(range(DataSet[[col]], na.rm = TRUE)), "\n\n")

+ }

Age

SD: 4.896501

Variance: 23.97572

IQR: 8

Range: 16

Academic.Pressure

SD: 1.390007

Variance: 1.93212

IQR: 2

Range: 4

Study.Satisfaction

SD: 1.37349

Variance: 1.886474

IQR: 2

Range: 4

Study.Hours

SD: 3.742434

Variance: 14.00581

IQR: 7

Range: 12

Financial.Stress

SD: 1.425053

Variance: 2.030775

IQR: 2

Range: 4

> normalize <- function(x) {

+ return((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)))

+ }

>

> DataSet[numeric\_cols] <- lapply(DataSet[numeric\_cols], normalize)

>

> cat("\nAfter Normalization (0–1 scale):\n")

After Normalization (0–1 scale):

> summary(DataSet[numeric\_cols])

Age Academic.Pressure Study.Satisfaction Study.Hours Financial.Stress

Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.2500 1st Qu.:0.250 1st Qu.:0.2500 1st Qu.:0.2500 1st Qu.:0.2500

Median :0.5312 Median :0.500 Median :0.5000 Median :0.5833 Median :0.5000

Mean :0.5151 Mean :0.501 Mean :0.5189 Mean :0.5337 Mean :0.4821

3rd Qu.:0.7500 3rd Qu.:0.750 3rd Qu.:0.7500 3rd Qu.:0.8333 3rd Qu.:0.7500

Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.0000

>

> set.seed(123)

>

> train\_index <- sample(1:nrow(DataSet), size = 0.7 \* nrow(DataSet))

>

> train\_data <- DataSet[train\_index, ]

> test\_data <- DataSet[-train\_index, ]

>

> cat("Training Data Rows:", nrow(train\_data), "\n")

Training Data Rows: 351

> cat("Testing Data Rows:", nrow(test\_data), "\n")

Testing Data Rows: 151

>

set.seed(123)

> sample\_data <- DataSet[sample(1:nrow(DataSet), 100), ] rows

>

> View(sample\_data)

> head(sample\_data, 10)

Gender Age Academic.Pressure Study.Satisfaction Sleep.Duration Dietary.Habits

415 Female 0.0625 0.00 1.00 More than 8 hours Moderate

463 Female 0.0625 0.00 0.25 More than 8 hours Healthy

179 Female 0.8125 0.00 0.25 7-8 hours Moderate

14 Male 0.4375 0.00 0.00 5-6 hours Moderate

195 Female 0.0000 0.75 0.00 More than 8 hours Healthy

426 Male 0.5000 1.00 0.75 Less than 5 hours Unhealthy

306 Female 0.3125 0.25 0.25 More than 8 hours Healthy

118 Male 0.4375 0.50 1.00 7-8 hours Unhealthy

299 Female 0.9375 0.75 1.00 7-8 hours Healthy

229 Male 0.4375 0.75 0.25 Less than 5 hours Moderate

Have.you.ever.had.suicidal.thoughts.. Study.Hours Financial.Stress

415 No 0.16666667 0.00

463 No 0.58333333 1.00

179 Yes 0.75000000 0.75

14 Yes 1.00000000 0.50

195 No 0.83333333 0.00

426 Yes 0.91666667 0.00

306 No 0.08333333 0.00

118 Yes 0.16666667 0.00

299 No 1.00000000 0.75

229 No 0.08333333 0.50

Family.History.of.Mental.Illness Depression

415 Yes No

463 No No

179 No No

14 Yes Yes

195 Yes Yes

426 Yes Yes

306 Yes No

118 No No

299 Yes No

229 Yes Yes