

PRACTICAL NO 8

AIM:- Applying basic data cleaning functions: handling missing values using `na.omit()/replace_na()` in R. import dataset.

OUTPUT:-

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains an R script for data cleaning and import. The script includes comments in Hindi and English, package installation, file path handling, and data reading. The output shows the data is loaded with 200 rows and 8 columns.
- Environment:** Lists the objects created in the global environment, including `id_cols`, `input_file`, `nm`, `numeric_patterns`, `out_file`, `p`, `packages`, `pat`, `pkg`, and `possible_paths`.
- Files:** Shows the file explorer with various files and folders, including `Processed_Global_Mobile_Prices.csv`, `SALARY.xlsx`, `synthetic_freelance_jobs.csv`, `WhatsApp Image 2025-11-30`, `PRACTICAL NO 6.R`, `amazon_merged_jan_feb_big.csv`, `amazon_final_list_big.csv`, `merged_output.csv`, `final_output.csv`, `nyc_flights.csv`, `PRACTICAL NO 7.R`, `Breast_Cancer.csv`, `PRACTICAL NO 8.R`, and `Breast_Cancer_cleaned.csv`.

```
> # ===== Breast Cancer - Auto-cleaning & Missing-value handler (dataset-aware) =====  
> # Usage: keep this script in same session where "Breast_Cancer.csv" is present  
> # It will produce "Breast_Cancer_cleaned.csv" in working directory.  
> # Comments in Hinglish for readability.  
> # =====  
> # ---- 0. packages ----  
> packages <- c("dplyr", "tidyr", "readr", "stringr")  
> for (pkg in packages) {  
+   if (!require(pkg, character.only = TRUE)) {  
+     install.packages(pkg, repos = "https://cloud.r-project.org")  
+     library(pkg, character.only = TRUE)  
+   }  
+ }  
> # ---- 1. Read dataset (use uploaded file if present) ----  
> # The uploaded path noted: "/mnt/data/Breast_Cancer.csv"  
> # But we use a flexible approach: try that path first, else default filename.  
> possible_paths <- c("/mnt/data/Breast_Cancer.csv", "Breast_Cancer.csv", "breast_cancer.csv")  
> input_file <- NULL  
> for (p in possible_paths) {  
+   if (file.exists(p)) { input_file <- p; break }  
+ }  
> if (is.null(input_file)) {  
+   stop("Breast_Cancer.csv not found in expected locations. Place file in working directory.")  
+ }  
> df <- readr::read_csv(input_file, na = c("", "NA"))  
Rows: 200 Columns: 8  
Column specification  
Delimiter: ","  
chr (3): Patient_ID, Tumor_Location, Diagnosis  
dbl (5): Age, Tumor_Size_mm, Mean_Radius, Mean_Texture, Mean_Smoothness  
I use 'spec()' to retrieve the full column specification for this data.  
I Specify the column types or set 'show_col_types = FALSE' to quiet this message.  
> cat("Loaded:", input_file, "\nRows:", nrow(df), "Cols:", ncol(df), "\n\n")  
Loaded: Breast_Cancer.csv  
Rows: 200 Cols: 8
```

Sheth L.U.J. College of Arts And Sir M.V. College of Science and Commerce

Data Analysis with SAS / SPSS /R

The screenshot shows the RStudio interface with the following content:

```
Source
Console Terminal Background Jobs
R 4.3.2 - C:/Users/as993/DATA SET/
> library(kableExtra) # see https://cran.r-project.org/web/packages/kableExtra/index.html
> cat("Loaded:", input_file, "\nRows:", nrow(df), "Cols:", ncol(df), "\n\n")
Loaded: Breast_Cancer.csv
Rows: 200 Cols: 8

> # ---- 2. Quick inspect ----
> cat("Column names:\n"); print(names(df)); cat("\nMissing per column (before):\n")
Column names:      "Age"      "Tumor_Size_mm"  "Tumor_Location"  "Mean_Radius"
[1] "Patient_ID"      "Mean_Smoothness" "Diagnosis"
[6] "Mean_Texture"      "Mean_Smoothness" "Diagnosis"

Missing per column (before):
> print(colSums(is.na(df))); cat("\n")
Patient_ID      Age      Tumor_Size_mm  Tumor_Location  Mean_Radius  Mean_Texture
0              0              0              0              0              0
Mean_Smoothness      Diagnosis
0                    0

> # ---- 3. Helpers ----
> get_mode <- function(x) {
+   x_no <- na.omit(x)
+   if (length(x_no) == 0) return(NA_character_)
+   ux <- unique(x_no)
+   ux[which.max(tabulate(match(x_no, ux)))]
+ }
> is_numeric_like <- function(x) {
+   if (!is.character(x)) return(FALSE)
+   s <- stringr::str_trim(x)
+   # allow decimal and integer numbers, drop empty/NA
+   s <- s[s != "" & !is.na(s)]
+   if (length(s) == 0) return(FALSE)
+   suppressWarnings(all(is.na(as.numeric(s))))
+ }
> convert_numeric_like <- function(dat) {
+   for (nm in names(dat)) {
+     if (is.character(dat[[nm]]) && is_numeric_like(dat[[nm]]) {
+       dat[[nm]] <- as.numeric(stringr::str_trim(dat[[nm]]))
+       message("Converted to numeric-like: ", nm)
+     }
+   }
+ }
```

The Environment pane on the right shows the loaded data and packages:

id_cols	"Patient_ID"
input_file	"Breast_Cancer.csv"
nm	"Diagnosis"
numeric_patterns	chr [1:9] "radius" "texture" "smooth" "perimete_"
out_file	"Breast_Cancer_cleaned.csv"
p	"Breast_Cancer.csv"
packages	chr [1:4] "dplyr" "tidyr" "readr" "stringr"
pat	"fractal"
pkg	"stringr"
possible_paths	chr [1:3] "/mnt/data/Breast_Cancer.csv" "Breast_"

The screenshot shows the RStudio interface with the following content:

```
Source
Console Terminal Background Jobs
R 4.3.2 - C:/Users/as993/DATA SET/
> # ---- 4. Preprocess conversions ----
> df <- convert_numeric_like(df)
> # If there's an 'id' like column, standardize name to Patient_ID (non-destructive)
> id_cols <- names(df)[tolower(names(df)) %in% c("id", "patient_id", "patientid", "pid")]
> if (length(id_cols) > 0) {
+   # rename first matching to Patient_ID
+   names(df)[names(df) == id_cols[1]] <- "Patient_ID"
+   message("Renamed column '", id_cols[1], "' -> 'Patient_ID'")
+ } else {
+   # if no id, create Patient_ID sequentially
+   df$Patient_ID <- paste0("BC", sprintf("%04d", seq_len(nrow(df))))
+   message("No ID column found -> created Patient_ID")
+ }
Renamed column 'Patient_ID' -> 'Patient_ID'

> # ---- 5. Dataset-aware custom rules (only apply if those columns exist) ----
> # - Diagnosis: keep as-is; if missing, mark as "Unknown"
> if ("Diagnosis" %in% names(df)) {
+   df$Diagnosis <- as.character(df$Diagnosis)
+   df$Diagnosis[is.na(df$Diagnosis)] <- "Unknown"
+   message("Diagnosis missing values -> 'Unknown'")
+ }
Diagnosis missing values -> 'Unknown'

> # - Common numeric columns conversions (if present)
> # Typical BC datasets may have columns with names containing 'radius', 'texture', 'smooth', 'perimeter', 'area'
> numeric_patterns <- c("radius", "texture", "smooth", "perimeter", "area", "compactness", "concavity", "symmetry", "fractal")
> for (pat in numeric_patterns) {
+   hits <- grep(pat, names(df), ignore.case = TRUE, value = TRUE)
+   if (length(hits) > 0) {
+     for (h in hits) {
+       if (!is.numeric(df[[h]])) {
+         df[[h]] <- suppressWarnings(as.numeric(df[[h]]))
+       }
+     }
+     message("Checked numeric-like column: ", h)
+   }
+ }
```

The Environment pane on the right shows the updated data and packages:

id_cols	"Patient_ID"
input_file	"Breast_Cancer.csv"
nm	"Diagnosis"
numeric_patterns	chr [1:9] "radius" "texture" "smooth" "perimete_"
out_file	"Breast_Cancer_cleaned.csv"
p	"Breast_Cancer.csv"
packages	chr [1:4] "dplyr" "tidyr" "readr" "stringr"
pat	"fractal"
pkg	"stringr"
possible_paths	chr [1:3] "/mnt/data/Breast_Cancer.csv" "Breast_"

Sheth L.U.J. College of Arts And Sir M.V. College of Science and Commerce

Data Analysis with SAS / SPSS /R

```
DATA SET - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Source
Console Terminal Background Jobs
R 4.3.2 C:/Users/as993/DATA SET/
y ~ fractal ~
> for (pat in numeric_patterns) {
+ hits <- grep(pat, names(df), ignore.case = TRUE, value = TRUE)
+ if (length(hits) > 0) {
+   for (h in hits) {
+     if (is.numeric(df[[h]])) {
+       df[[h]] <- suppressWarnings(as.numeric(df[[h]]))
+     }
+     message("Checked numeric-like column:", h)
+   }
+ }
+ }
+ }

Checked numeric-like column:Mean_Radius
Checked numeric-like column:Mean_Texture
Checked numeric-like column:Mean_Smoothness

> # ---- 6. Imputation strategy ----
> cleaned <- df # copy

> # If Price-like logic not relevant here; we use general rules:
> for (nm in names(cleaned)) {
+ col <- cleaned[[nm]]
+ if (any(is.na(col))) {
+   if (is.numeric(col)) {
+     med <- median(col, na.rm = TRUE)
+     if (is.na(med)) med <- 0
+     cleaned[[nm]][is.na(cleaned[[nm]])] <- med
+     message(sprintf("Numeric imputed (median) for %s -> %s", nm, format(med, digits=6)))
+   } else if (is.logical(col)) {
+     cleaned[[nm]][is.na(cleaned[[nm]])] <- FALSE
+     message(sprintf("Logical imputed (FALSE) for %s", nm))
+   } else { # character/factor
+     modev <- get_mode(col)
+     if (is.na(modev)) modev <- "Unknown"
+     cleaned[[nm]][is.na(cleaned[[nm]])] <- modev
+     message(sprintf("Categorical imputed (mode/'Unknown') for %s -> %s", nm, as.character(modev)))
+   }
+ }
+ }

Environment History Connections Tutorial
R 268 MiB
Global Environment
id_cols "Patient_ID"
input_file "Breast_Cancer.csv"
nm "Diagnosis"
numeric_patterns chr [1:9] "radius" "texture" "smooth" "perimete_
out_file "Breast_Cancer_cleaned.csv"
p "Breast_Cancer.csv"
packages chr [1:4] "dplyr" "tidyr" "readr" "stringr"
pat "fractal"
pkg "stringr"
possible_paths chr [1:3] "/mnt/data/Breast_Cancer.csv" "Breast_

Files Plots Packages Help Viewer Presentation
New Folder New File Delete Rename More
C:/Users/as993/DATA SET
Name Size Modified
PRACTICAL NO 3.R 310 B Nov 23, 2025, 1:23 PM
Processed_Global_Mobile_Prices.csv 89.2 KB Nov 24, 2025, 11:12 PM
SALARY.xlsx 5.3 KB Nov 24, 2025, 6:50 PM
synthetic_freelance_jobs.csv 300.6 KB Nov 24, 2025, 11:08 PM
WhatsApp Image 2025-11-30 at 13:41:44_0156d8c0.jpg 143.3 KB Nov 30, 2025, 3:49 PM
WhatsApp Image 2025-11-30 at 13:41:46_6541518e.jpg 93.5 KB Nov 30, 2025, 4:13 PM
WhatsApp Image 2025-11-30 at 13:41:47_401b4e63.jpg 146.6 KB Nov 30, 2025, 3:14 PM
PRACTICAL NO 6.R 863 B Dec 1, 2025, 11:53 AM
amazon_merged_jan_feb_big.csv 333 B Dec 1, 2025, 11:29 AM
amazon_final_list_big.csv 321 B Dec 1, 2025, 11:29 AM
merged_output.csv 333 B Dec 1, 2025, 11:46 AM
final_output.csv 321 B Dec 1, 2025, 11:46 AM
nyc_flights.csv 6.6 MB Dec 1, 2025, 12:14 PM
PRACTICAL NO 7.R 2.3 KB Dec 1, 2025, 12:21 PM
Breast_Cancer.csv 10.2 KB Dec 1, 2025, 12:52 PM
PRACTICAL NO 8.R 6.5 KB Dec 1, 2025, 12:57 PM
Breast_Cancer_cleaned.csv 10.2 KB Dec 1, 2025, 1:01 PM
```

```
DATA SET - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Source
Console Terminal Background Jobs
R 4.3.2 C:/Users/as993/DATA SET/

Missing per column (after):
Patient_ID 0 Age 0 Tumor_Size_mm 0 Tumor_Location 0 Mean_Radius 0 Mean_Texture 0
Mean_Smoothness 0 Diagnosis 0

> num_summary <- cleaned %>% select(where(is.numeric))
> if (ncol(num_summary) > 0) {
+ cat("Numeric summary (selected):\n"); print(summary(num_summary)); cat("\n")
+ }
Numeric summary (selected):
   Age      Tumor_Size_mm   Mean_Radius   Mean_Texture   Mean_Smoothness
Min. :30.00   Min. :10.00   Min. :10.01   Min. :15.10   Min. :0.07000
1st Qu.:41.75 1st Qu.:21.00 1st Qu.:13.32 1st Qu.:19.88 1st Qu.:0.08975
Median :54.50 Median :33.00 Median :17.21 Median :24.39 Median :0.10750
Mean :54.84 Mean :33.95 Mean :17.36 Mean :24.69 Mean :0.10981
3rd Qu.:68.00 3rd Qu.:47.00 3rd Qu.:21.34 3rd Qu.:29.53 3rd Qu.:0.13025
Max. :80.00 Max. :60.00 Max. :24.87 Max. :34.90 Max. :0.15000

> out_file <- "Breast_Cancer_cleaned.csv"
> readr::write_csv(cleaned, out_file)
> cat("Cleaned file written to:", out_file, "\n")
Cleaned file written to: Breast_Cancer_cleaned.csv

> # ---- 9. Provide quick counts ----
> cat(sprintf("Rows: %d | Columns: %d\n", nrow(cleaned), ncol(cleaned)))
Rows: 200 | Columns: 8
> cat("Diagnosis value counts:\n"); if ("Diagnosis" %in% names(cleaned)) print(table(cleaned$Diagnosis)) else ca
t("No Diagnosis column.\n")
Diagnosis value counts:
   Benign Malignant
   101          99

> cat("\nDone. Agar aur customization chahiye (e.g., specific columns ka different rule, or scaling, or encodin
g), batao.\n")
Done. Agar aur customization chahiye (e.g., specific columns ka different rule, or scaling, or encoding), batao.
>

Environment History Connections Tutorial
R 268 MiB
Global Environment
id_cols "Patient_ID"
input_file "Breast_Cancer.csv"
nm "Diagnosis"
numeric_patterns chr [1:9] "radius" "texture" "smooth" "perimete_
out_file "Breast_Cancer_cleaned.csv"
p "Breast_Cancer.csv"
packages chr [1:4] "dplyr" "tidyr" "readr" "stringr"
pat "fractal"
pkg "stringr"
possible_paths chr [1:3] "/mnt/data/Breast_Cancer.csv" "Breast_

Files Plots Packages Help Viewer Presentation
New Folder New File Delete Rename More
C:/Users/as993/DATA SET
Name Size Modified
PRACTICAL NO 3.R 310 B Nov 23, 2025, 1:23 PM
Processed_Global_Mobile_Prices.csv 89.2 KB Nov 24, 2025, 11:12 PM
SALARY.xlsx 5.3 KB Nov 24, 2025, 6:50 PM
synthetic_freelance_jobs.csv 300.6 KB Nov 24, 2025, 11:08 PM
WhatsApp Image 2025-11-30 at 13:41:44_0156d8c0.jpg 143.3 KB Nov 30, 2025, 3:49 PM
WhatsApp Image 2025-11-30 at 13:41:46_6541518e.jpg 93.5 KB Nov 30, 2025, 4:13 PM
WhatsApp Image 2025-11-30 at 13:41:47_401b4e63.jpg 146.6 KB Nov 30, 2025, 3:14 PM
PRACTICAL NO 6.R 863 B Dec 1, 2025, 11:53 AM
amazon_merged_jan_feb_big.csv 333 B Dec 1, 2025, 11:29 AM
amazon_final_list_big.csv 321 B Dec 1, 2025, 11:29 AM
merged_output.csv 333 B Dec 1, 2025, 11:46 AM
final_output.csv 321 B Dec 1, 2025, 11:46 AM
nyc_flights.csv 6.6 MB Dec 1, 2025, 12:14 PM
PRACTICAL NO 7.R 2.3 KB Dec 1, 2025, 12:21 PM
Breast_Cancer.csv 10.2 KB Dec 1, 2025, 12:52 PM
PRACTICAL NO 8.R 6.5 KB Dec 1, 2025, 12:57 PM
Breast_Cancer_cleaned.csv 10.2 KB Dec 1, 2025, 1:01 PM
```