

SHETH L.U.J. AND SIR M.V. COLLEGE

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

The screenshot shows the RStudio interface with the following components:

- Source:** Contains the R script code for importing and cleaning the dataset.
- Console:** Displays the output of the R commands, including warnings about Rtools and the successful installation of the 'tidyr' package.
- Environment:** Lists the objects in the global environment, including 'merged_data', 'my_data', 'new_rows', 'product_sales_dataset', 'range_cols', 'region_filter', 'robot_related', 'selected_cols', and 'sleep_df'.
- Files:** Shows the file explorer with various files and folders, including 'Practical_No_4s_RR', 'Practical_No_5_RR', 'Practical_No_5B_RR', 'Processed_Student_Mental_Health.csv', 'S093_CHETAN_MANDAVKAR_FEATURE_ENGINEERING_SUBMISSION - Colab', 'S093_Chetan_Mandavkar_FP_Soft_Copy_Documentation_[1].docx', 'SHRUTI_MANDAVKAR_EVS_REPORT.pdf', 'Structure and functions of pancreas Shruti Mandavkar.pptx', 'Virtual Machines', 'WhatsApp Image 2025-11-25 at 19:51:57_2b9447da.jpg', 'Attendance_Prediction.csv', 'product_sales_dataset_final.csv', 'Practical_No_6_RR', 'synthetic_fraud_dataset.csv', 'Practical_No_7_RR', 'dataset_2191_sleep.csv', and 'Practical_No_8_RR'.

```
> install.packages("tidyr")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and
install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into 'C:/Users/CHETAN/AppData/Local/R/win-library/4.5'
(as 'lib' is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/tidyr_1.3.1.zip'

Content type 'application/zip' length 1276404 bytes (1.2 MB)
downloaded 1.2 MB

package 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/CHETAN/AppData/Local/Temp/RtmpSg817p/downloaded_packages
> library(dplyr)
> library(tidyr)
> sleep_df <- read.csv("dataset_2191_sleep.csv", na.strings = c("", "NA"))
> print("---- 1. Original Data (First 6 Rows) ----")
[1] "---- 1. Original Data (First 6 Rows) ----"
> print(head(sleep_df))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index
1    6654.000    5712.0    38.6      645           3             5
2     1.000      6.6      4.5      42           3             1
3     3.385     44.5     14         60           1             1
4     0.920      5.7      7         25           5             2
5    2547.000    4603.0     69      624           3             5
6    10.550     179.5     27      180           4             4
  danger_index total_sleep
1             3           3
2             3           8
3             3          12
4             3          16
5             4           9
6             4           9
> print("---- Count of Missing Values per Column ----")
```

The screenshot shows the RStudio interface with the following components:

- Source:** Contains the R script code for cleaning the dataset, including removing rows with missing values and replacing missing values with the mean.
- Console:** Displays the output of the R commands, including the count of missing values and the successful cleaning of the dataset.
- Environment:** Lists the objects in the global environment, including 'merged_data', 'my_data', 'new_rows', 'product_sales_dataset', 'range_cols', 'region_filter', 'robot_related', 'selected_cols', and 'sleep_df'.
- Files:** Shows the file explorer with various files and folders, including 'Practical_No_4s_RR', 'Practical_No_5_RR', 'Practical_No_5B_RR', 'Processed_Student_Mental_Health.csv', 'S093_CHETAN_MANDAVKAR_FEATURE_ENGINEERING_SUBMISSION - Colab', 'S093_Chetan_Mandavkar_FP_Soft_Copy_Documentation_[1].docx', 'SHRUTI_MANDAVKAR_EVS_REPORT.pdf', 'Structure and functions of pancreas Shruti Mandavkar.pptx', 'Virtual Machines', 'WhatsApp Image 2025-11-25 at 19:51:57_2b9447da.jpg', 'Attendance_Prediction.csv', 'product_sales_dataset_final.csv', 'Practical_No_6_RR', 'synthetic_fraud_dataset.csv', 'Practical_No_7_RR', 'dataset_2191_sleep.csv', and 'Practical_No_8_RR'.

```
> print("---- Count of Missing Values per Column ----")
[1] "---- Count of Missing Values per Column ----"
> print(colSums(is.na(sleep_df)))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index
0             0             0             0             0             0             0
  danger_index total_sleep
0             0             0
> clean_omit <- na.omit(sleep_df)
> print("---- 2. Data after na.omit() ----")
[1] "---- 2. Data after na.omit() ----"
> print(paste("Original rows:", nrow(sleep_df)))
[1] "Original rows: 62"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 62"
> print(head(clean_omit))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index
1    6654.000    5712.0    38.6      645           3             5
2     1.000      6.6      4.5      42           3             1
3     3.385     44.5     14         60           1             1
4     0.920      5.7      7         25           5             2
5    2547.000    4603.0     69      624           3             5
6    10.550     179.5     27      180           4             4
  danger_index total_sleep
1             3           3
2             3           8
3             3          12
4             3          16
5             4           9
6             4           9
> clean_replace <- sleep_df %>%
+   replace_na(list(
+     body_weight = mean(sleep_df$body_weight, na.rm = TRUE),
+     brain_weight = mean(sleep_df$brain_weight, na.rm = TRUE),
+     max_life_span = mean(sleep_df$max_life_span, na.rm = TRUE),
+     gestation_time = mean(sleep_df$gestation_time, na.rm = TRUE),
+     predation_index = mean(sleep_df$predation_index, na.rm = TRUE),
+     sleep_exposure_index = mean(sleep_df$sleep_exposure_index, na.rm = TRUE),
+     danger_index = mean(sleep_df$danger_index, na.rm = TRUE),
+     total_sleep = mean(sleep_df$total_sleep, na.rm = TRUE)
+   ))
```

NAME:- CHETAN MANDAVKAR

ROLL NO. S093

SUBJECT:- Data Analysis with SAS / SPSS / R

SHETH L.U.J. AND SIR M.V. COLLEGE

```
R - R4.5.2 ~ / ~
> clean_replace <- sleep_df %>%
+   replace_na(list(
+     body_weight = mean(sleep_df$body_weight, na.rm = TRUE),
+     brain_weight = mean(sleep_df$brain_weight, na.rm = TRUE),
+     max_life_span = mean(sleep_df$max_life_span, na.rm = TRUE),
+     gestation_time = mean(sleep_df$gestation_time, na.rm = TRUE),
+     predation_index = mean(sleep_df$predation_index, na.rm = TRUE),
+     sleep_exposure_index = mean(sleep_df$sleep_exposure_index, na.rm = TRUE),
+     danger_index = mean(sleep_df$danger_index, na.rm = TRUE),
+     total_sleep = mean(sleep_df$total_sleep, na.rm = TRUE)
+   ))

Warning messages:
1: In mean.default(sleep_df$max_life_span, na.rm = TRUE) :
  argument is not numeric or logical: returning NA
2: In mean.default(sleep_df$gestation_time, na.rm = TRUE) :
  argument is not numeric or logical: returning NA
3: In mean.default(sleep_df$total_sleep, na.rm = TRUE) :
  argument is not numeric or logical: returning NA

> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> print(head(clean_replace))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index
1    6654.000    5712.0      38.6         645             3              5
2     1.000         6.6         4.5          42             3              1
3     3.385         44.5        14          60             1              1
4     0.920         5.7         7          25             5              2
5    2547.000    4603.0        69         624             3              5
6     10.550     179.5        27         180             4              4
  danger_index total_sleep
1           3           3.3
2           3           8.3
3           1          12.5
4           3          16.5
5           4           3.9
6           4           9.8

> print("--- Remaining NAs after replacement ---")
[1] "--- Remaining NAs after replacement ---"
> print(colSums(is.na(clean_replace)))
```

```
R - R4.5.2 ~ / ~
> print(colSums(is.na(clean_replace)))
  body_weight brain_weight max_life_span gestation_time
0           0           0             0             0
  predation_index sleep_exposure_index danger_index total_sleep
0           0             0             0             0

> print(colSums(is.na(clean_replace)))
  body_weight brain_weight max_life_span gestation_time
0           0           0             0             0
  predation_index sleep_exposure_index danger_index total_sleep
0           0             0             0             0

> |
```

NAME:- CHETAN MANDAVKAR

ROLL NO. S093

SUBJECT:- Data Analysis with SAS / SPSS / R