

# MVLU COLLEGE

## PRACTICAL NO. 8

**Aim:- Applying basic data cleaning functions: handling missing values using na.omit()/replace\_na() in R. import dataset.**

**Code:**

The image shows two side-by-side sessions of RStudio. Both sessions have the following tabs open: R PRACTICAL 6.R, Untitled1\*, R PRACTICAL 7.R\*, R PRACTICAL 8.R\*, Retail.Product, and Fastest.Fifties...2018. The left session contains code for importing a dataset and removing missing values using na.omit(). The right session contains code for replacing missing values using replace\_na(). Both sessions show a file browser on the right with various CSV files and R scripts, and a taskbar at the bottom.

```

# Load necessary libraries
library(dplyr)
library(tidyverse)

# 1. CREATE AND IMPORT DATASET
# IMPORTANT: Using your provided file path and the correct file name.
# "na.strings = "" tells R to treat empty spaces as NA (Missing values).
cricket_df <- read.csv("D:/Haris126/Fastest Fifties - 2018.csv", na.strings = c("", "NA"))

print("--- 1. original data (First 6 Rows) ---")
print(head(cricket_df))

# Check how many NAs are in each column
print("--- Count of Missing Values per Column ---")
# If '4s' and '6s' were renamed by R, they will appear as 'x4s' and 'x6s' here.
print(coltsums(is.na(cricket_df)))

# 2. METHOD A: REMOVE MISSING VALUES (na.omit)
# This is the "nuclear option". If a row has even ONE missing value, it is deleted.
clean_omit <- na.omit(cricket_df)

print("--- 2. Data after na.omit() ---")
print(paste("Original rows:", nrow(cricket_df)))
print(paste("Rows remaining:", nrow(clean_omit)))
print(head(clean_omit))

```

  

```

# Print original rows
print(paste("Original rows:", nrow(cricket_df)))
print(paste("Rows remaining:", nrow(clean_omit)))
print(head(clean_omit))

# 3. METHOD B: REPLACE MISSING VALUES (replace_na)
# This is the "surgical option". We fill missing values with logical defaults.
# Strategy:
# 1. Venue: Fill missing with "Unknown venue" (Categorical)
# 2. X4s, X6s: Fill missing with 0 (Assumption: Numeric, No data = 0 boundaries)
# 3. Runs: Fill missing with the Median Runs (Median is often preferred over Mean for runs/sc

# Calculate median runs (ignoring NAs) to use for filling
median_runs <- median(cricket_df$Runs, na.rm = TRUE)

clean_replace <- cricket_df %>%
  replace_na(list(
    Venue = "Unknown Venue",
    X4s = 0, # Assuming R renamed '4s' to 'X4s'
    X6s = 0, # Assuming R renamed '6s' to 'X6s'
    Runs = median_runs
  ))

print("--- 3. Data after replace_na() ---")
print(head(clean_replace))

# Verify no NAs exist in the columns we cleaned
print("--- Remaining NAs after replacement ---")
print(coltsums(is.na(clean_replace)))

```

**Output:**

Hariprasad Vishwakarma  
S126  
SYCS

# MVLU COLLEGE

## PRACTICAL NO. 8

The screenshot shows the RStudio interface with the following details:

- File Edit Code View Plots Session Build Debug Profile Tools Help**
- Source** tab selected.
- Console** tab shows R code and its output:

```
R - R 4.1.2 - ~/R
> library(tidyverse)
Warning message:
package 'tidyverse' was built under R version 4.1.3
> # IMPORTANT: Using your provided file path and the correct file name.
> # 'na.strings = "",' tells R to treat empty spaces as NA (missing values).
> cricket_df <- read.csv("D:/Haris126/Fastest Fifties - 2018.csv", na.strings = c("", "NA"))
> print("--- 1. Original Data (First 6 Rows) ---")
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(cricket_df))
#> # Check how many NAs are in each column
#> print("--- Count of Missing values per Column ---")
[1] "# Count of Missing Values per Column ---"
#> # If 'x4s' and 'x6s' were renamed by R, they will appear as 'x4s' and 'x6s' here.
#> print(colsums(is.na(cricket_df)))
#> # Clean up the data
#> clean.omit <- na.omit(cricket_df)
#> print("--- 2. Data after na.omit() ---")
[1] "--- 2. Data after na.omit() ---"
#> print(paste("Original rows:", nrow(cricket_df)))
[1] "Original rows: 106"
#> print(paste("Rows remaining:", nrow(clean.omit)))
[1] "Rows remaining: 106"
```

- Environment** tab shows the following objects:

  - clean\_replace 106 obs. of 9 variables
  - cricket 106 obs. of 9 variables
  - cricket\_data 106 obs. of 9 variables
  - cricket\_df 106 obs. of 9 variables
  - data\_feb 3 obs. of 3 variables
  - data\_jan 3 obs. of 3 variables
  - data\_new\_hires 2 obs. of 3 variables
  - deceased\_males 106 obs. of 9 variables

- Files** tab shows the project structure:

  - Home
  - Name      Size      Modified
  - tue\_sales.csv   766 B     Sep 18, 2025, 8:35 AM
  - GIS DataBase
  - IISExpress
  - My Music
  - My Pictures
  - My Videos
  - My Web Sites
  - R
  - 5M PRACTICAL - II 20.txt   1.2 KB     Aug 22, 2025, 3:02 PM
  - SM PRACTICAL - II.txt   1.2 KB     Aug 22, 2025, 3:01 PM
  - total\_bills.csv   1.2 KB     Sep 18, 2025, 8:47 AM
  - Virtual Machines
  - Visual Studio 2022
  - WindowsPowerShell

- System Status Bar**: ENG IN 12:42 01-12-2025

MVLU COLLEGE  
PRACTICAL NO. 8

The screenshot shows an RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and a search bar labeled "Search".
- Source Editor:** Displays R code for data cleaning and analysis. The code includes:
  - Reading a CSV file named "cricket\_df".
  - Creating a copy named "clean\_omit" by omitting rows where certain columns are missing.
  - Printing the first few rows of the original and cleaned datasets.
  - Calculating the number of rows remaining after cleaning.
  - Replacing missing values (NA) in the dataset.
  - Reordering columns and printing the final cleaned dataset.
- Environment Tab:** Shows the global environment with objects like "clean\_replace", "cricket", "cricket\_data", "cricket\_df", etc., each with their respective dimensions.
- Files Tab:** Shows the file structure, including CSV files like "tue\_sales.csv" and "total\_bill.csv", and various Microsoft Office files.
- Help and Footer:** Includes links for "Import Dataset", "Connections", "Tutorial", and "Project: (None)". The footer shows the date and time as "01-12-2025 12:43".

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Project dropdown set to '(None)'. The left sidebar has tabs for Source, Console, Terminal, and Background Jobs, with 'Console' selected. The main console area displays an R script session. The code starts by reading a CSV file 'tue\_sales.csv' into a data frame 'clean\_replace'. It then calculates median runs for each player and replaces missing values ('NA') with these medians. The session continues with printing the data frame, verifying no NA values remain, and printing column names. The right sidebar shows the Global Environment pane with objects like 'clean\_replace', 'cricket', etc., and a file browser pane showing files in the current directory, including 'R' and 'SM PRACTICAL' subfolders.

Hariprasad Vishwakarma  
S126  
SYCS