

# Sheth L.U.J. & Sir M.V. College

## 8. 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 titanic dataset loaded into a data frame. The data frame has 12 columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The first 26 rows of the dataset are displayed, showing various passengers' details like name, age, sex, and survival status. The RStudio environment pane on the right shows global variables and files.

The screenshot shows the RStudio interface with an R script open. The code is written to handle missing values in the titanic dataset. It starts by loading libraries (tidyverse), then reads the CSV file into a data frame. It prints the first few rows and counts missing values. The script then removes all rows with missing values using the na.omit() function. Finally, it prints the head of the cleaned dataset. The RStudio environment pane on the right shows global variables and files.

```
1 # -----
2 # R Script: Handling Missing Values (Data Cleaning)
3 # Dataset: Titanic Passenger Data
4 # -----
5
6 # Load libraries
7 #install.packages("tidyverse")
8 library(dplyr)
9 library(tidyverse)
10 # -----
11 # 1. IMPORT DATA
12 # -----
13 #
14
15 # Treat blank cells or "NA" as missing
16 titanic_df <- read.csv("C:/Users/Priya Gupta/Downloads/S081 Priya Rakesh Gupta - titanic.csv", na.strings = c("", "NA"))
17
18 cat("----- 1. Original Data (First 6 Rows) -----")
19 print(head(titanic_df))
20
21 # Count missing values
22 cat("----- Count of Missing Values per Column -----")
23 print(colSums(is.na(titanic_df)))
24
25 # -----
26 # 2. METHOD A: REMOVE ALL ROWS WITH ANY MISSING VALUE
27 # -----
28
29 clean_omit <- na.omit(titanic_df)
30
31 cat("----- 2. Data after na.omit() -----")
32 cat("Original rows:", nrow(titanic_df), "\n")
33 cat("Rows remaining:", nrow(clean_omit), "\n")
34 print(head(clean_omit))
35
36 # -----
37 # 3. METHOD B: FILL MISSING VALUES (replace_na)
38 #
39
40 # Strategy for replacement:
```

**Name :- Priya Gupta**  
**Roll no. :- S081**

## Sheth L.U.J. & Sir M.V. College

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

PracticalR.R x S081.Priya.Rakesh.Gupta.titanic x

Source on Save Run Source Environs Global Env

Project: (None)

```
27 # =====
28
29 clean_omit <- na.omit(titanic_df)
30 cat("\n--- 2. Data after na.omit() ---\n")
31 cat("Original rows:", nrow(titanic_df), "\n")
32 cat("Rows remaining:", nrow(clean_omit), "\n")
33 print(head(clean_omit))
34
35 # =====
36 # 3. METHOD B: FILL MISSING VALUES (replace_na)
37 # =====
38 # =====
39
40 # Strategy for replacement:
41 # - Age      - Replace missing ages with mean age
42 # - Cabin    - Replace missing cabin with "Unknown"
43 # - Embarked - Replace missing with most common value ("S")
44 # - Fare     - Replace missing with mean fare
45
46 avg_age <- mean(titanic_df$Age, na.rm = TRUE)
47 avg_fare <- mean(titanic_df$Fare, na.rm = TRUE)
48
49 # Find most common Embarked value
50 most_common_embarked <- names(sort(table(titanic_df$Embarked), decreasing = TRUE))[1]
51
52 clean_replace <- titanic_df %>
53   replace_na(list(
54     Age      = avg_age,
55     Cabin    = "Unknown",
56     Embarked = most_common_embarked,
57     Fare     = avg_fare
58   ))
59
60 cat("\n--- 3. Data after replace_na() ---\n")
61 print(head(clean_replace))
62
63 cat("\n--- Remaining NAs After Cleanup ---\n")
64 print(colSums(is.na(clean_replace)))
65
```

36:81 (Untitled).R Script

Console

Finance headline  
India reported 1...

Search

22:58  
01-12-2025

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Practical8.R S081.PriyaRakeshGupta.titanic

Source Save Run Source R Script

Console Terminal Background Jobs R 4.5.2 - /

```
> # =====
> # R Script: Handling Missing Values (Data Cleaning)
> # Dataset: Titanic Passenger Data
> #
> #
> # Load libraries
> #install.packages("tidyverse")
> library(dplyr)
> library(tidyverse)
> #
> # 1. IMPORT DATA
> #
> #
> # Treat blank cells or "NA" as missing
> titanic_df <- read.csv("C:/Users/Priya Gupta/Downloads/S081 Priya Rakesh Gupta - titanic.csv", na.strings = c("", "NA"))
>
> cat("\n--- 1. Original Data (First 6 Rows) ---\n")
--- 1. Original data (First 6 Rows) ---
> print(head(titanic_df))
  PassengerId Survived Pclass          Name      Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
1           1        0     3    Braund, Mr. Owen Harris   male  22      1     0   A/5 21171 7.2500   <NA>      S
2           2        1     1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38      1     0 PC 17599 71.2833   C85      C
3           3        1     1 Heikkinen, Miss. Laina  female  26      0     0 STON/O2. 3101282 7.9250   <NA>      S
4           4        1     1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35      1     0   113803 53.1000  CL123      S
5           5        0     3            Allen, Mr. William Henry   male  35      0     0   373450  8.0500   <NA>      S
6           6        0     3            Moran, Mr. James    male   NA      0     0   330877  8.4583   <NA>      Q
>
> # Count missing values
> cat("\n--- Count of Missing Values per Column ---\n")
--- Count of Missing Values per Column ---
> print(colsums(is.na(titanic_df)))
  PassengerId Survived Pclass          Name      Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
1           0        0     0          0       0  0  0  0       0     0  687      2
```

**Name :- Priya Gupta**  
**Roll no. :- S081**

# Sheth L.U.J. & Sir M.V. College

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

```

 27 # =====
> # =====
> # 2. METHOD A: REMOVE ALL ROWS WITH ANY MISSING VALUE
> # =====
> 
> clean.omit <- na.omit(titanic_df)
> 
> cat("\n--- 2. Data after na.omit() ---\n")
--- 2. Data after na.omit() ---
> cat("Original rows:", nrow(titanic_df), "\n")
Original rows: 891
> cat("Rows remaining:", nrow(clean.omit), "\n")
Rows remaining: 183
> print(head(clean.omit))
  PassengerId Survived Pclass      Name     Sex   Age SibSp Parch Ticket Fare Cabin Embarked
2           1         1      3 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.0      1    0 PC 17599 71.2833   S
4           1         1      1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 55.0      1    0 313333 53.1000 C123   S
7           0         0      1          McCarthy, Mr. Timothy J male 35.0      0    0 17463 51.8625 E46   S
11          1         1      3 Sandstrom, Miss. Marguerite Rut female 4.0      1    1 PP 9549 16.7000 G6   S
12          1         1      1 Bonnell, Miss. Elizabeth female 58.0      0    0 113783 26.5500 C103  S
22          1         2          Beesley, Mr. Lawrence male 34.0      0    0 248698 13.0000 D56   S
> 
> # =====
> # 3. METHOD B: FILL MISSING VALUES (replace_na)
> # =====
> 
> # Strategy for replacement:
> # - Age      - Replace missing ages with mean age
> # - Cabin    - Replace missing cabin with "Unknown"
> # - Embarked - Replace missing with most common value ("S")
> # - Fare     - Replace missing with mean fare
> 
> avg_age <- mean(titanic_df$Age, na.rm = TRUE)
> avg_fare <- mean(titanic_df$Fare, na.rm = TRUE)
> 
> # Find most common Embarked value
> most_common_embarked <- names(sort(table(titanic_df$Embarked), decreasing = TRUE))[1]

```

Trending videos Stranger Things... Search

ENG IN 23:00 01-12-2025

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source on Save Run Source

```

 27 # =====
> # - Embarked - Replace missing with most common value ("S")
> # - Fare       - Replace missing with mean fare
> 
> avg_age <- mean(titanic_df$Age, na.rm = TRUE)
> avg_fare <- mean(titanic_df$Fare, na.rm = TRUE)
> 
> # Find most common Embarked value
> most_common_embarked <- names(sort(table(titanic_df$Embarked), decreasing = TRUE))[1]
> 
> clean_replace <- titanic_df %>
+   replace_na(list(
+     Age      = avg_age,
+     Cabin    = "Unknown",
+     Embarked = most_common_embarked,
+     Fare     = avg_fare
+   ))
> 
> cat("\n--- 3. Data after replace_na() ---\n")
--- 3. Data after replace_na() ---
> print(head(clean_replace))
  PassengerId Survived Pclass      Name     Sex   Age SibSp Parch Ticket Fare Cabin Embarked
1           1         0      3 Braund, Mr. Owen Harris male 22.00000 1    0 A/5 21371 7.2500 Unknown   S
2           2         1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00000 1    0 PC 17599 71.2833   S
3           3         1      3 Heikkinen, Miss. Laina female 26.00000 0    0 STON/O2 313338 7.9250 Unknown   S
4           4         1      1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00000 1    0 113803 53.1000 C123   S
5           5         0      3 Allen, Mr. William Henry male 35.00000 0    0 373450 8.0500 Unknown   S
6           6         0      3 Moran, Mr. James male 29.69912 0    0 390877 8.4583 Unknown   Q
> 
> cat("\n--- Remaining NAs After Cleanup ---\n")
--- Remaining NAs After Cleanup ---
> print(colsums(is.na(clean_replace)))
PassengerId     Survived      Pclass      Name     Sex   Age SibSp Parch Ticket Fare Cabin Embarked
0             0         0      0         0     0     0     0     0     0     0     0     0     0
> 
```

Trending videos Stranger Things... Search

ENG IN 23:00 01-12-2025

**Name :- Priya Gupta**  
**Roll no. :- S081**