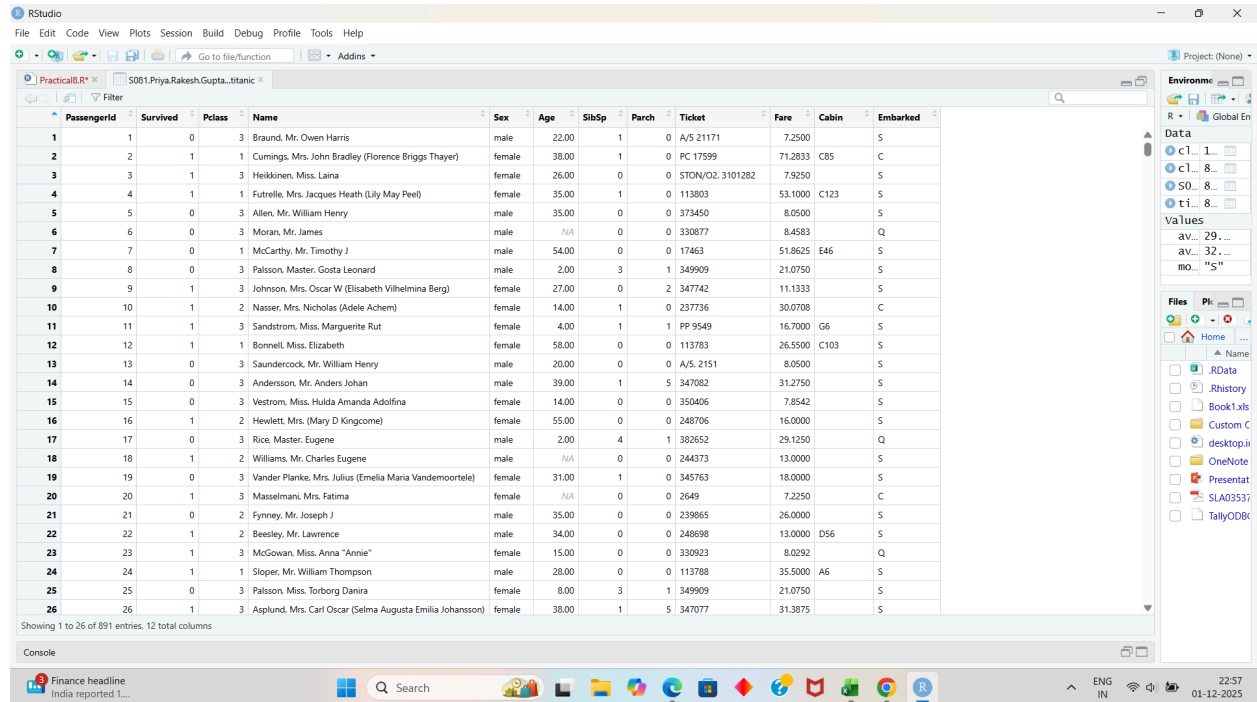


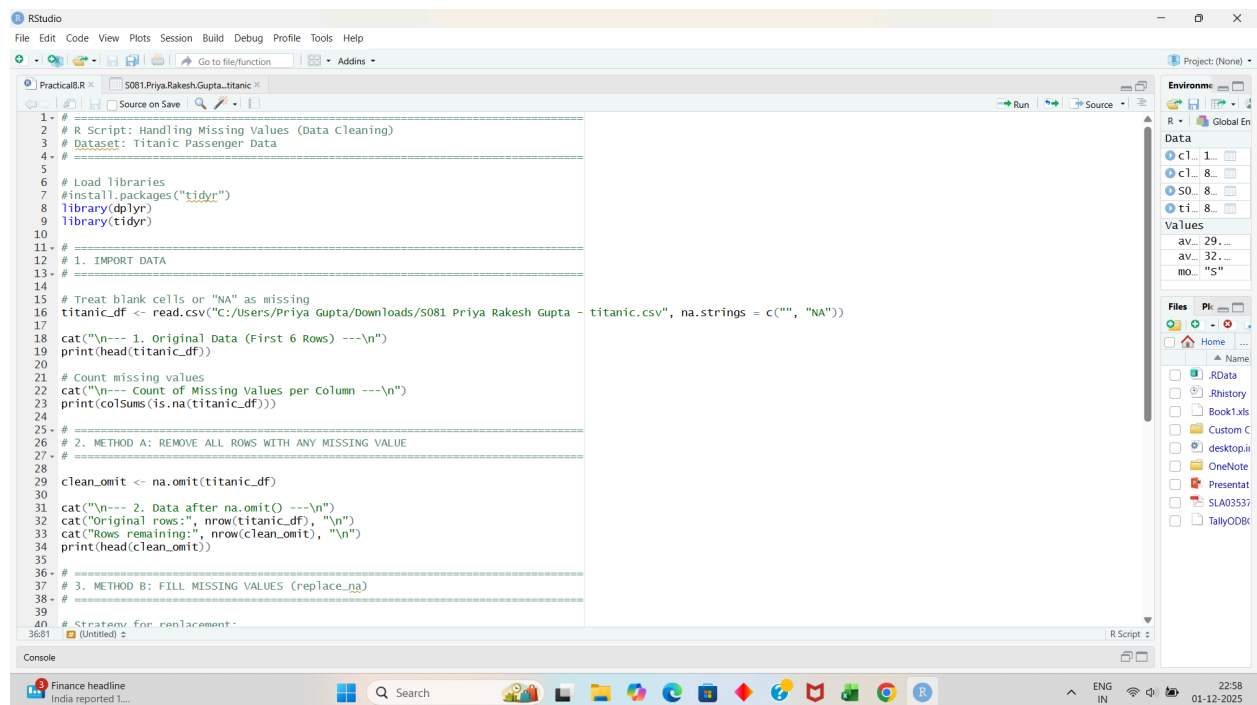
# Sheth L.U.J. & Sir M.V. College

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. The Environment pane on the right shows the dataset as a data frame with 851 rows and 12 columns. The Data pane shows the first few rows of the dataset.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
13	0	3	Saunderscock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.0500		S
14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.2750		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16.0000		S
17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250		Q
18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00	1	0	345763	18.0000		S
20	1	3	Maselmans, Mrs. Fatima	female	NA	0	0	2649	7.2250		C
21	0	2	Fynney, Mr. Joseph J	male	35.00	0	0	239865	26.0000		S
22	1	2	Beesley, Mr. Lawrence	male	34.00	0	0	248698	13.0000	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28.00	0	0	113788	35.5000	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8.00	3	1	349909	21.0750		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.00	1	5	347077	31.3875		S



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

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
S081.Priya.Rakesh.Gupta.titanic
27 # 2. METHOD A: REMOVE ALL ROWS WITH ANY MISSING VALUE
28
29 clean_omit <- na.omit(titanic_df)
30
31 cat("\n--- 2. Data after na.omit() ---\n")
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:
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
3681 (Untitled) R Script
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
S081.Priya.Rakesh.Gupta.titanic
27 # 2. METHOD A: REMOVE ALL ROWS WITH ANY MISSING VALUE
3681 (Untitled) R Script

Console Terminal Background Jobs
R - R 4.5.2 - /-
> # =====
> # R Script: Handling Missing Values (Data Cleaning)
> # Dataset: Titanic Passenger Data
> # =====
> # Load libraries
> #install.packages("tidyr")
> library(dplyr)
> library(tidyr)
> # =====
> # 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 female  38   1     0      PC 17599  53.0000   C85    C
3          3         1      3 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  C123    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
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**

FileEditViewCodePlotsSessionBuildDebugProfileToolsHelp

Go to file/functionAddins

PracticalR.R5081.Priya.Rakesh.Gupta.titanic.R

RunSource

EnvironmentHistoryConnectionsRGlobal EnvironmentData

FilesPlotsPackagesHelpView

HomeName.RDataRhistoryBook1.xlsxCustom Office Templatesdesktop.iniOneNote NotebooksPresentation.pptxSLA035375211408.pdfTallyODBC64\_9000.dsn

23:0011-12-2025

```
# RStudio
# 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          2         1       1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) Female  38   1   0 PC 17599  17.2833   C85        C
4          4         1       1 Futrelle, Mrs. Jacques Heath (Lily May Peel) Female  35   1   0 113803  53.1000  C123        S
7          7         0       1 McCarthy, Mr. Timothy J male  54   0   0 17463  51.8625  E46        S
11         11         1       3 Sandstrom, Miss. Marguerite Rut female  4   1   1 PP 9549 16.7000   G6        S
12         12         1       1 Bonnell, Miss. Elizabeth female  58   0   0 113783 26.5500  C103        S
22         22         1       2 Beesley, Mr. Lawrence male  34   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])
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Addins

Project: (None)

Practical6.R

S081.Priya.Rakesh.Gupta..titanic

Source on Save

Run

Source

27  
3681  
(Untitled)

R Script

Console

Terminal

Background Jobs

```
R - R4.52 > # Replace missing with most common value ("S")
> # Embarked -> Replace missing with mean fare
> # 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
1          0         0       3
2          1         1       1
3          3         1       3
4          4         1       1
5          5         0       3
6          6         0       3

  Name                               Sex  Age  SibSp  Parch    Ticket   Fare Cabin Embarked
1 Braund, Mr. Owen Harris             male  22.0000   1    0      A/5 21171  7.2500  unknown
2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38.0000   1    0      PC 17599  71.2833   C85    C
3 Heikkinen, Miss. Laina              female  26.0000   0    0 STON/O2. 3101282  7.9250  unknown
4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35.0000   1    0    113803  53.1000  C123    C
5 Allen, Mr. William Henry            male  35.0000   0    0    373450  8.0500  unknown
6 Moran, Mr. James                   male  29.6912   0    0    330877  8.4583  unknown

> 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
1           0           0           0           0           0           0           0           0           0           0           0           0
> |
```

Environment

History

Connections

Data

Values

Files

Plots

Packages

Help

View

Home

Name

☐

.RData

☐

Rhistory

☐

Book1.xlsx

☐

desktop.ini

☐

Custom Office Templates

☐

OneNote Notebooks

☐

Presentation.pptx

☐

SLAO35375211408.pdf

☐

TallyO08C64\_9000.dsn

Trending videos

Stranger Things...

Search

ENG IN

23:00

01-12-2025

**Name :- Priya Gupta**

**Roll no. :- S081**