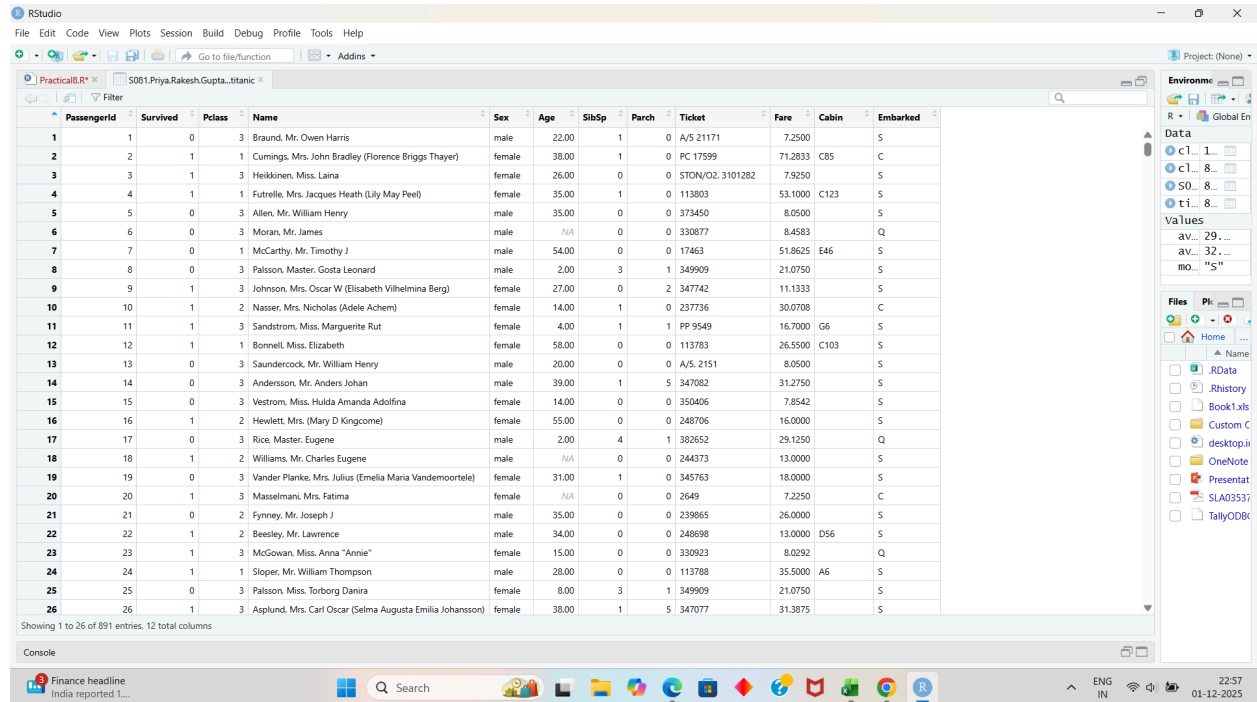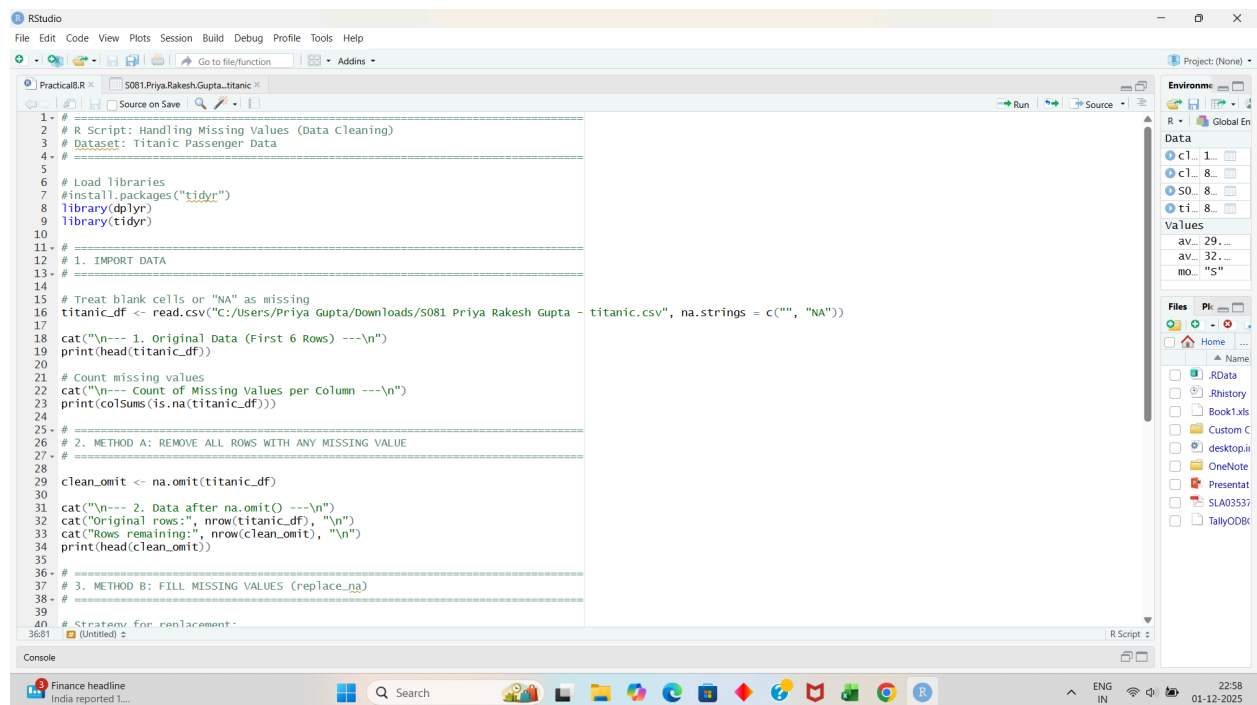# Sheth L.U.J. & Sir M.V. College

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

Name :- Priya Gupta

Roll no. :- S081

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

```r
# 2. METHOD A: REMOVE ALL ROWS WITH ANY MISSING VALUE
# =========

clean_omit <- na.omit(titanic_df)

cat("\n--- 2. Data after na.omit() ---\n")
cat("Original rows:", nrow(titanic_df), "\n")
cat("Rows remaining:", nrow(clean_omit), "\n")
print(head(clean_omit))

# =========
# 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]

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")
print(head(clean_replace))

cat("\n--- Remaining NAs After Cleanup ---\n")
print(colSums(is.na(clean_replace)))
```

Environment panel:
Data
- cl... 1...
- cl... 8...
- S0... 8...
- ti... 8...

Values
- av... 29...
- av... 32...
- mo... "S"

Files panel:
Home
- .Rhistory
- Book1.xls
- Custom C
- desktop.i
- OneNote
- Presentat
- SLA03537
- TallyODBC

Console:

```
> # =========
> # 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 (Florence Briggs Thayer) female  38     1     0    PC 17599 71.2833   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         177           0           0           0           0         687           2
>
```

Environment panel:
Data
- clean_... 183 obs. of 12 v...
- clean_... 891 obs. of 12 v...
- S081.P... 891 obs. of 12 v...
- titani... 891 obs. of 12 v...

Values
- avg_age 29.6991176470588
- avg_fa... 32.2042079685746
- most_c... "S"

Files panel:
Home
- .RData
- .Rhistory
- Book1.xlsx
- Custom Office Templates
- desktop.ini
- OneNote Notebooks
- Presentation.pptx
- SLA035375211408.pdf
- TallyODBC64_9000.dsn

Name :- Priya Gupta

Roll no. :- SO81

Name :- Priya Gupta

Roll no. :- SO81