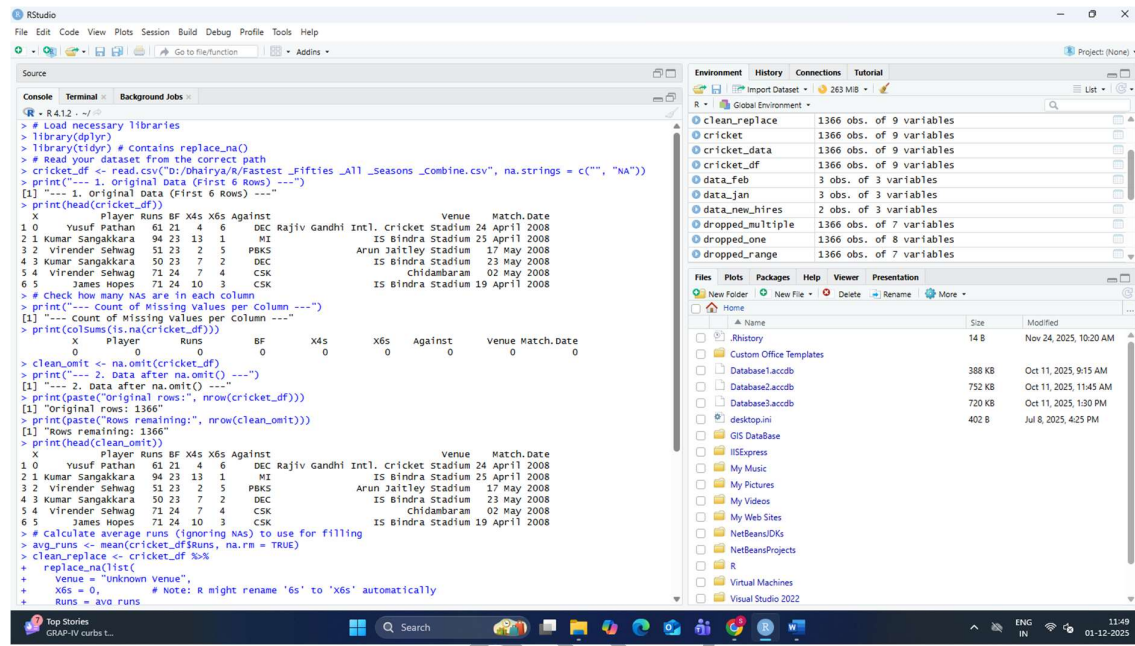


SHETH L.U.J AND SIR M.V. COLLEGE
SUBJECT NAME: **Data Analysis with SAS / SPSS / R**
Practical No. 8

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

Output-



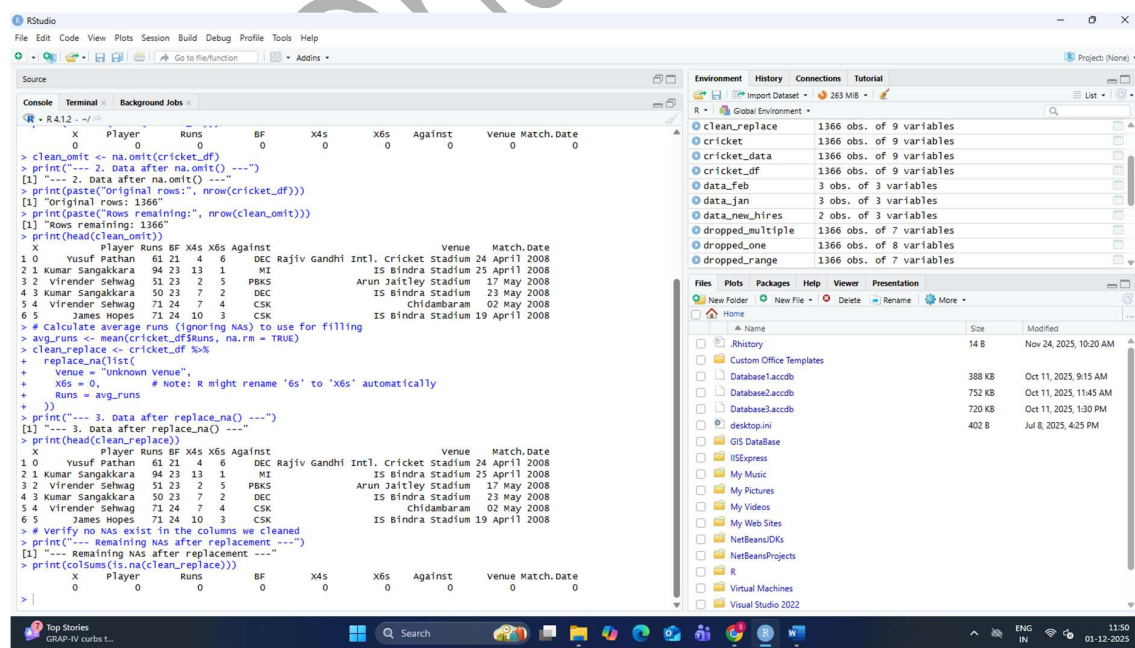
```
R> # Load necessary libraries
> library(dplyr)
> library(tidyverse) # contains replace_na()
> # Read your dataset from the correct path
> cricket_df <- read.csv("D:/Dhairya/R/FASTEST_Fifties_All_Seasons_combine.csv", na.strings = c("", "NA"))
> print("---- 1. original data (first 6 rows) ----")
[1] "---- 1. original data (first 6 rows) ----"
> print(head(cricket_df))
  X      Player Runs BF X4s X6s Against Venue Match.Date
1 0 Yusuf Pathan 61 21 4 6 DEC Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 1 Kumar Sangakkara 94 23 13 1 MI IS Bindra Stadium 25 April 2008
3 2 Virender Sehwag 51 23 2 5 PBKS Arun Jaitley Stadium 17 May 2008
4 3 Kumar Sangakkara 50 23 7 2 DEC IS Bindra Stadium 23 May 2008
5 4 Virender Sehwag 71 24 7 4 CSK Chidambaram 02 May 2008
6 5 James Hopes 71 24 10 3 CSK IS Bindra Stadium 19 April 2008

> # Check how many NAs are in each column
> print("---- Count of Missing values per column ----")
[1] "---- Count of Missing values per column ----"
> print(colSums(is.na(cricket_df)))
  X      Player Runs BF X4s X6s Against Venue Match.Date
0      0      0      0      0      0      0      0

> 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: 1366"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 1366"
> print(head(clean_omit))
  X      Player Runs BF X4s X6s Against Venue Match.Date
1 0 Yusuf Pathan 61 21 4 6 DEC Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 1 Kumar Sangakkara 94 23 13 1 MI IS Bindra Stadium 25 April 2008
3 2 Virender Sehwag 51 23 2 5 PBKS Arun Jaitley Stadium 17 May 2008
4 3 Kumar Sangakkara 50 23 7 2 DEC IS Bindra Stadium 23 May 2008
5 4 Virender Sehwag 71 24 7 4 CSK Chidambaram 02 May 2008
6 5 James Hopes 71 24 10 3 CSK IS Bindra Stadium 19 April 2008

> # Calculate average runs (ignoring NAs) to use for filling
> avg_runs <- mean(cricket_df$Runs, na.rm = TRUE)
> clean_replace <- cricket_df %>%
+   replace_na(list(
+     Venue = "Unknown Venue",
+     X6s = 0, # Note: R might rename '6s' to 'X6s' automatically
+     Runs = avg_runs
+   ))
> print("---- 3. data after replace_na() ----")
[1] "---- 3. data after replace_na() ----"
> print(head(clean_replace))
  X      Player Runs BF X4s X6s Against Venue Match.Date
1 0 Yusuf Pathan 61 21 4 6 DEC Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 1 Kumar Sangakkara 94 23 13 1 MI IS Bindra Stadium 25 April 2008
3 2 Virender Sehwag 51 23 2 5 PBKS Arun Jaitley Stadium 17 May 2008
4 3 Kumar Sangakkara 50 23 7 2 DEC IS Bindra Stadium 23 May 2008
5 4 Virender Sehwag 71 24 7 4 CSK Chidambaram 02 May 2008
6 5 James Hopes 71 24 10 3 CSK IS Bindra Stadium 19 April 2008

> # Verify no NAs exist in the columns we cleaned
> print("---- Remaining NAs after replacement ----")
[1] "---- Remaining NAs after replacement ----"
> print(colSums(is.na(clean_replace)))
  X      Player Runs BF X4s X6s Against Venue Match.Date
0      0      0      0      0      0      0      0
```



```
R> # Verify no NAs exist in the columns we cleaned
> print("---- Remaining NAs after replacement ----")
[1] "---- Remaining NAs after replacement ----"
> print(colSums(is.na(clean_replace)))
  X      Player Runs BF X4s X6s Against Venue Match.Date
0      0      0      0      0      0      0      0
```