

MVLU COLLEGE PRACTICAL NO. 8

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

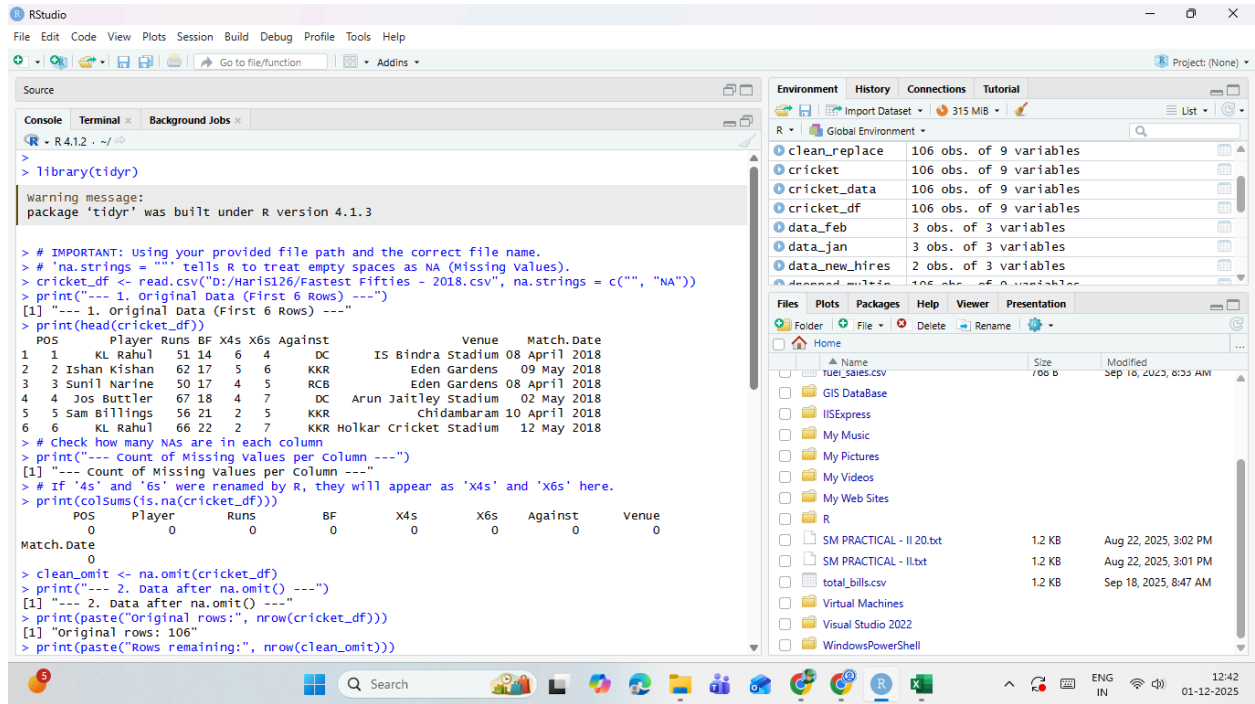
Code:

```
1 # Load necessary libraries
2 library(dplyr)
3 library(tidyr)
4
5 # =====
6 # 1. CREATE AND IMPORT DATASET
7 # =====
8
9 # IMPORTANT: Using your provided file path and the correct file name.
10 # 'na.strings = ""' tells R to treat empty spaces as NA (Missing Values).
11 cricket_df <- read.csv("D:/Haris126/Fastest Fifties - 2018.csv", na.strings = c("", "NA"))
12
13 print("---- 1. Original Data (First 6 Rows) ----")
14 print(head(cricket_df))
15
16 # Check how many NAs are in each column
17 print("---- Count of Missing values per Column ----")
18 # If '4s' and '6s' were renamed by R, they will appear as 'x4s' and 'x6s' here.
19 print(colSums(is.na(cricket_df)))
20
21 # =====
22 # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
23 # =====
24 # This is the "nuclear option". If a row has even ONE missing value, it is deleted.
25
26 clean_omit <- na.omit(cricket_df)
27
28 print("---- 2. Data after na.omit() ----")
29 print(paste("Original rows:", nrow(cricket_df)))
30 print(paste("Rows remaining:", nrow(clean_omit)))
31 print(head(clean_omit))
32
33
34
35
36
37 # =====
38 # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
39 # =====
40 # This is the "surgical option". we fill missing values with logical defaults.
41 # Strategy:
42 # 1. Venue: Fill missing with "unknown Venue" (Categorical)
43 # 2. x4s, x6s: Fill missing with 0 (Assumption: Numeric, No data = 0 boundaries)
44 # 3. Runs: Fill missing with the Median Runs (Median is often preferred over Mean for runs/sc
45
46 # Calculate median runs (ignoring NAs) to use for filling
47 median_runs <- median(cricket_df$Runs, na.rm = TRUE)
48
49 clean_replace <- cricket_df %>%
50   replace_na(list(
51     Venue = "unknown Venue",
52     x4s = 0, # Assuming R renamed '4s' to 'x4s'
53     x6s = 0, # Assuming R renamed '6s' to 'x6s'
54     Runs = median_runs
55   ))
56
57 print("---- 3. Data after replace_na() ----")
58 print(head(clean_replace))
59
60 # Verify no NAs exist in the columns we cleaned
61 print("---- Remaining NAs after replacement ----")
62 print(colSums(is.na(clean_replace)))
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

Output:

Hariprasad Vishwakarma
S126
SYCS

MVLU COLLEGE PRACTICAL NO. 8



```
R - R 4.1.2 - ~/
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)

Source
Console Terminal Background Jobs
R - R 4.1.2 - ~/
> 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:/Hariprasad/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))
  POS Player Runs BF X4s X6s Against Venue Match.Date
1 1 KL Rahul 51 14 6 4 DC IS Bindra Stadium 08 April 2018
2 2 Ishan Kishan 62 17 5 6 KKR Eden Gardens 09 May 2018
3 3 Sunil Narine 50 17 4 5 RCB Eden gardens 08 April 2018
4 4 Jos Buttler 67 18 4 7 DC Arun Jaitley Stadium 02 May 2018
5 5 Sam Billings 56 21 2 5 KKR Chidambaram 10 April 2018
6 6 KL Rahul 66 22 2 7 KKR Holkar Cricket Stadium 12 May 2018

> # Check how many NAs are in each column
> print("--- Count of Missing values per column ---")
[1] "Count of Missing values per column"
> # If '4s' and '6s' were renamed by R, they will appear as 'x4s' and 'x6s' here.
> print(colsums(is.na(cricket_df)))
  POS Player Runs BF X4s X6s Against Venue
0 0 0 0 0 0 0 0

Match.Date
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: 106"
> print(paste("Rows remaining:", nrow(clean_omit)))

Environment History Connections Tutorial
R - Global Environment
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
deleted_data 106 obs. of 9 variables

Files Plots Packages Help Viewer Presentation
Home
Name Size Modified
true_sales.csv 700 B Sep 10, 2023, 6:53 AM
GIS DataBase
IISExpress
My Music
My Pictures
My Videos
My Web Sites
R
SM 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
```

Hariprasad Vishwakarma
S126
SYCS

MVLU COLLEGE PRACTICAL NO. 8

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)

Source
Console Terminal Background Jobs
R - R 4.1.2 ~ /

> 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"
> print(head(clean_omit))
  POS Player Runs BF X4s X6s Against Venue Match.Date
1  1  KL Rahul  51 14  6  4    DC    IS Bindra Stadium 08 April 2018
2  2 Ishan Kishan 62 17  5  6    KKR    Eden Gardens 09 May 2018
3  3 Sunil Narine 50 17  4  5    RCB    Eden Gardens 08 April 2018
4  4 Jos Buttler 67 18  4  7    DC    Arun Jaitley Stadium 02 May 2018
5  5 Sam Billings 56 21  2  5    KKR    Chidambaram 10 April 2018
6  6 KL Rahul 66 22  2  7    KKR    Holkar Cricket Stadium 12 May 2018

> # 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() ---")
[1] "--- 3. Data after replace_na() ---"
> print(head(clean_replace))
  POS Player Runs BF X4s X6s Against Venue Match.Date
1  1  KL Rahul  51 14  6  4    DC    IS Bindra Stadium 08 April 2018
2  2 Ishan Kishan 62 17  5  6    KKR    Eden Gardens 09 May 2018
3  3 Sunil Narine 50 17  4  5    RCB    Eden Gardens 08 April 2018
4  4 Jos Buttler 67 18  4  7    DC    Arun Jaitley Stadium 02 May 2018
5  5 Sam Billings 56 21  2  5    KKR    Chidambaram 10 April 2018
6  6 KL Rahul 66 22  2  7    KKR    Holkar Cricket Stadium 12 May 2018
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)

Source
Console Terminal Background Jobs
R - R 4.1.2 ~ /

2  2 Ishan Kishan 62 17  5  6    KKR    Eden Gardens 09 May 2018
3  3 Sunil Narine 50 17  4  5    RCB    Eden Gardens 08 April 2018
4  4 Jos Buttler 67 18  4  7    DC    Arun Jaitley Stadium 02 May 2018
5  5 Sam Billings 56 21  2  5    KKR    Chidambaram 10 April 2018
6  6 KL Rahul 66 22  2  7    KKR    Holkar Cricket Stadium 12 May 2018

> # 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() ---")
[1] "--- 3. Data after replace_na() ---"
> print(head(clean_replace))
  POS Player Runs BF X4s X6s Against Venue Match.Date
1  1  KL Rahul  51 14  6  4    DC    IS Bindra Stadium 08 April 2018
2  2 Ishan Kishan 62 17  5  6    KKR    Eden Gardens 09 May 2018
3  3 Sunil Narine 50 17  4  5    RCB    Eden Gardens 08 April 2018
4  4 Jos Buttler 67 18  4  7    DC    Arun Jaitley Stadium 02 May 2018
5  5 Sam Billings 56 21  2  5    KKR    Chidambaram 10 April 2018
6  6 KL Rahul 66 22  2  7    KKR    Holkar Cricket Stadium 12 May 2018

> # 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)))
  POS Player Runs BF X4s X6s Against Venue
0      0      0      0      0      0      0
Match.Date
0

>
|
```

Hariprasad Vishwakarma
S126
SYCS