

SHETH L.U.J. AND SIR M.V. COLLEGE

PRACTICAL NO 8

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

CODE:

The image shows two separate sessions of RStudio. Both sessions have tabs for 'S090_PRACICAL6.R', 'S090_PRACICAL7.R', and 'S090_PRACICAL8.R'. The first session (top) contains code for importing a dataset and calculating the number of missing values per column. The second session (bottom) contains code for removing missing values using `na.omit()` and replacing missing values using `replace_na()`.

```
1 # R Script: Handling Missing Values (Data Cleaning)
2 # Dataset: Retail Product Data
3
4
5 # Load necessary libraries
6 # install.packages("dplyr")
7 install.packages("tidyverse")
8 library(dplyr)
9 library(tidyverse) # Contains replace_na()
10
11
12 # =====
13 # 1. CREATE AND IMPORT DATASET
14 #
15
16 # Note: 'na.strings = ""' tells R to treat empty spaces as NA (Missing Values).
17
18 # Read dataset (fixed version - works even if file is not in working directory)
19 retail_df <- read.csv(file.choose(), na.strings = c("", "NA"))
20
21 print("--- 1. Original Data (First 6 Rows) ---")
22 print(head(retail_df))
23
24 # Check how many NAs are in each column
25 print("--- Count of Missing values per column ---")
26 print(coltsums(is.na(retail_df)))
27
28 # =====
29 # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
30 #
31 # This is the "nuclear option". If a row has even ONE missing value, it is deleted.
32
33 clean.omit <- na.omit(retail_df)
34
35 print("--- 2. Data after na.omit() ---")
36 print(paste("Original rows:", nrow(retail_df)))
37 print(paste("Rows remaining:", nrow(clean.omit)))
38 print(head(clean.omit))
39
40 #
41 # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
42 #
43
44 # Strategy:
45 # 1. Category: Fill missing with "Unknown"
46 # 2. Discount: Fill missing with 0
47 # 3. Stock: Fill missing with "Check warehouse"
48 # 4. Price: Fill missing with the Mean (Average) Price
49
50 # Calculate average price (ignoring NAs)
51 avg_price <- mean(retail_df$Price, na.rm = TRUE)
52
53 clean.replace <- retail_df %>%
54   replace_na(list(
55     Category = "Unknown",
56     Discount = 0,
57     Stock = "Check warehouse",
58     Price = avg_price
59   ))
60
61 print("--- 3. Data after replace_na() ---")
62 print(clean.replace[3,])
63 print(head(clean.replace))
64
65 print("--- Remaining NAs after replacement ---")
66 print(coltsums(is.na(clean.replace)))
67
68
69
```

SHETH L.U.J. AND SIR M.V. COLLEGE

OUTPUT:

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.1.2 . ~/
> # Load necessary libraries
> # install.packages("dplyr")
> install.packages("tidyverse")
Restarting R session...
> install.packages("tidyverse")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
<https://cran.rstudio.com/bin/windows/Rtools/>
Installing package into 'C:/Users/IT-03/Documents/R/win-library/4.1'
(as 'lib' is unspecified)
also installing the dependency 'rlang'

There are binary versions available but the source versions are later:
binary source needs_compilation
rlang 1.1.0 1.1.6 TRUE
tidyverse 1.3.0 1.3.1 TRUE

Binaries will be installed
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/rlang_1.1.0.zip'
Content type 'application/zip' length 1710397 bytes (1.6 MB)
downloaded 1.6 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/tidyverse_1.3.0.zip'
Content type 'application/zip' length 1440051 bytes (1.4 MB)
downloaded 1.4 MB

package 'rlang' successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package 'rlang'
Warning: restored 'rlang'
package 'tidyverse' successfully unpacked and MD5 sums checked
Type here to search 26°C Sunny 12:25 01-12-2025

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.1.2 . ~/
The following objects are masked from 'package:stats':
filter, lag
The following objects are masked from 'package:base':
intersect, setdiff, setequal, union

Warning messages:
1: In file.copy(savedcopy, lib, recursive = TRUE) :
problem copying C:/Users/IT-03/Documents/R/win-library/4.1/00LOCK/rlang/libs/x64/rlang.dll to C:/Users/IT-03/Documents/R/win-library/4.1/rlang/libs/x64/rlang.dll
1: Permission denied
2: package 'dplyr' was built under R version 4.1.3

> library(tidyverse) # contains replace_na()
Warning message:
package 'tidyverse' was built under R version 4.1.3

> # Read dataset
> #retail_df <- read.csv("Retail_Product")
> retail_df <- read.csv("Retail_Product.csv", na.strings = c("", "NA"))
Error in file(file, "rt") : cannot open the connection
Show Traceback
Rerun with Debug

In addition: warning message:
In file(file, "rt") :
cannot open file 'Retail_Product.csv': No such file or directory

> print("--- 1. Original Data (First 6 Rows) ---)
Type here to search 26°C Sunny 12:26 01-12-2025

SHETH L.U.J. AND SIR M.V. COLLEGE

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Project: (None)

Source

```
[R - R4.12 - ~/]
[1] "--> original Data (First 6 Rows) -->
> print(head(retail_df))
#> # Check how many NAs are in each column
#> print(" --- count of Missing values per column ---")
[1] " --- count of Missing values per column ---"
> print(colsyms(is.na(retail_df)))
#> clean.omit <- na.omit(retail_df)
#> #--- 2. Data after na.omit() ---
[1] " --- 2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(retail_df)))
[1] "Original rows: 541909"
> print(paste("Rows remaining:", nrow(clean.omit)))
[1] "Rows remaining: 0"
> print(head(clean.omit))
#> # calculate average price (ignoring NAs) to use for filling
#> avg_price <- mean(retail_df$price, na.rm = TRUE)
#> clean_replace <- retail_df %>
#> clean_replace <- na.omit(clean_replace)
#> # calculate average price (ignoring NAs) to use for filling
#> avg_price <- mean(clean_replace$price, na.rm = TRUE)
#> clean_replace <- replace_na(clean_replace, list(
#>   Category = "Unknown",
#>   Discount = 0,
#>   Stock = "Check warehouse",
#>   Price = avg_price
#> ))
#> print(" --- 3. Data after replace_na() ---")
[1] " --- 3. Data after replace_na() ---"
#> # Check row 3 specifically: In original data, Price was NA. Now it should be the average.
> print(clean_replace[3, ])
#> # verify no NAs exist in the columns we cleaned (except Rating, which we didn't touch)
> print(" --- Remaining NAs after replacement ---")
```

RStudio

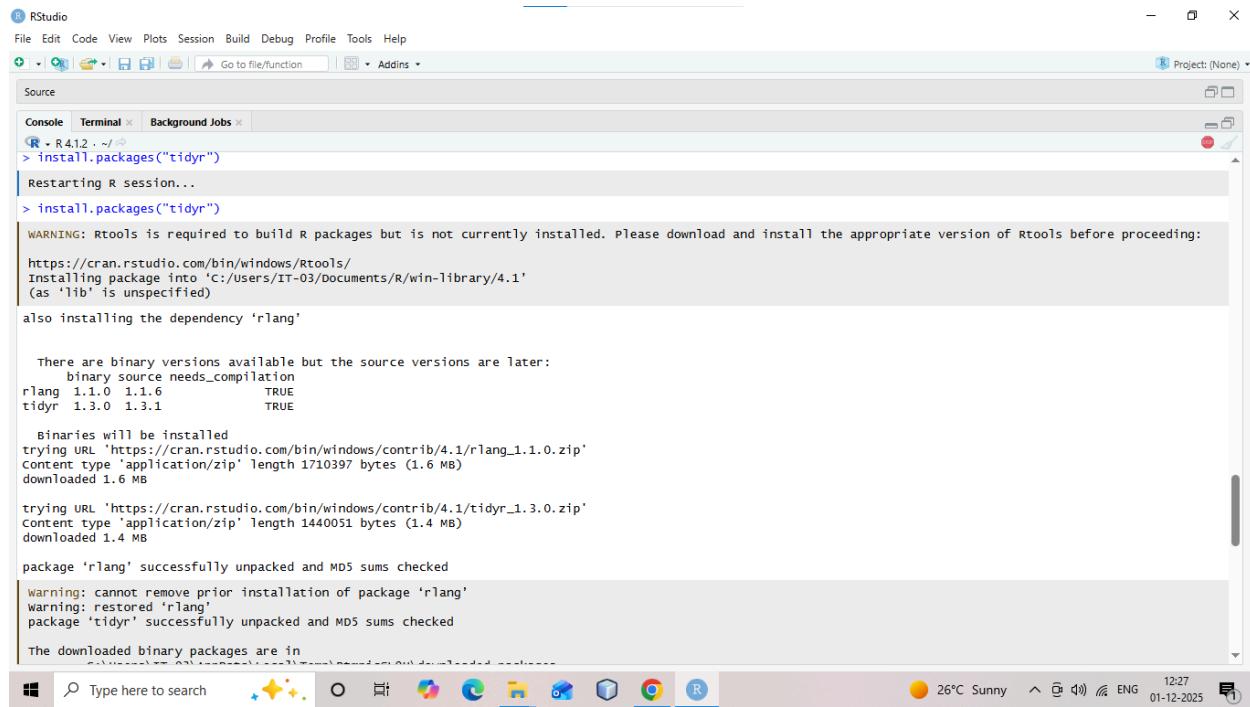
File Edit Code View Plots Session Build Debug Profile Tools Help

Project: (None)

Source

```
[R - R4.12 - ~/]
<0 rows> (or 0-length row.names)
#> # calculate average price (ignoring NAs) to use for filling
#> avg_price <- mean(retail_df$price, na.rm = TRUE)
#> clean_replace <- retail_df %>
#> replace_na(list(
#>   Category = "Unknown",
#>   Discount = 0,
#>   Stock = "Check warehouse",
#>   Price = avg_price
#> ))
#> print(" --- 3. Data after replace_na() ---")
[1] " --- 3. Data after replace_na() ---"
#> # Check row 3 specifically: In original data, Price was NA. Now it should be the average.
> print(clean_replace[3, ])
#> # verify no NAs exist in the columns we cleaned (except Rating, which we didn't touch)
> print(" --- Remaining NAs after replacement ---")
```

SHETH L.U.J. AND SIR M.V. COLLEGE



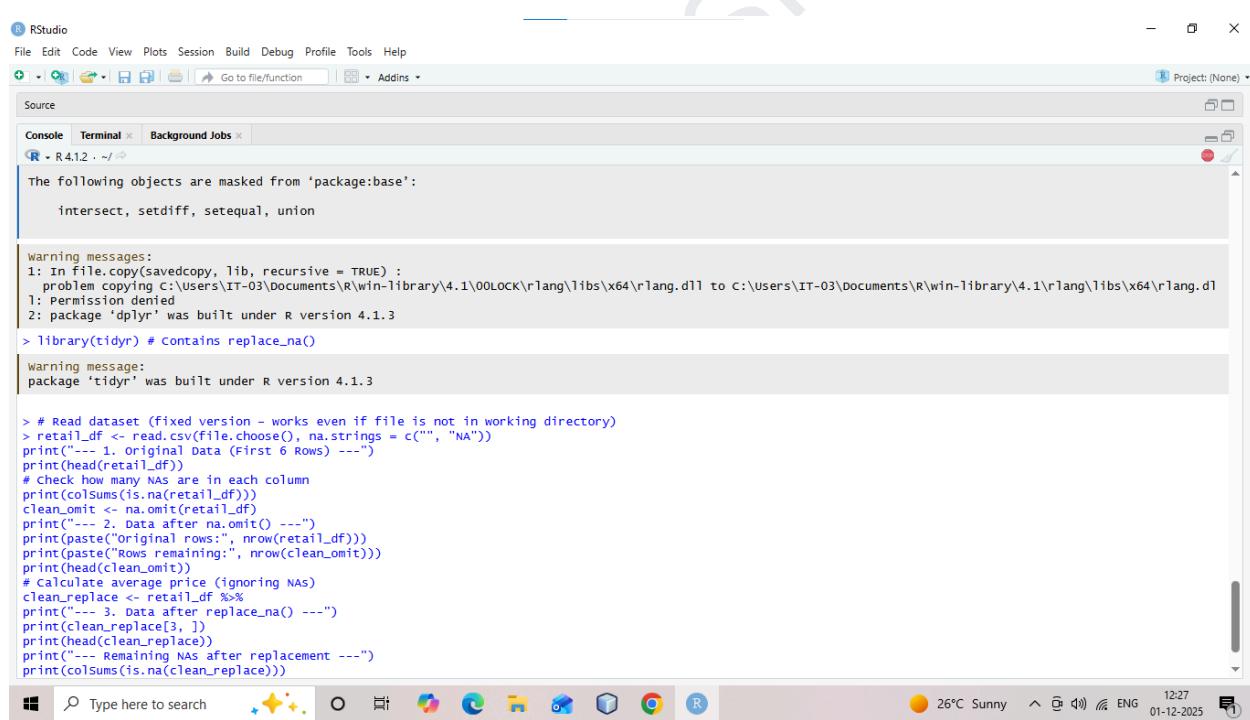
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
> install.packages("tidyverse")
Restarting R session...
> install.packages("tidyverse")
WARNING: Rtools required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
<https://cran.rstudio.com/bin/windows/Rtools/>
Installing package into 'c:/users/IT-03/Documents/R/win-library/4.1'
(as 'lib' is unspecified)
also installing the dependency 'rlang'

There are binary versions available but the source versions are later:
binary source needs compilation
rlang 1.1.0 1.1.6 TRUE
tidyverse 1.3.0 1.3.1 TRUE

Binaries will be installed
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/rlang_1.1.0.zip'
Content type 'application/zip' length 1710397 bytes (1.6 MB)
downloaded 1.6 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/tidyverse_1.3.0.zip'
Content type 'application/zip' length 1440051 bytes (1.4 MB)
downloaded 1.4 MB

package 'rlang' successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package 'rlang'
Warning: restored 'rlang'
package 'tidyverse' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:/Users/IT-03/Documents/R/win-library/4.1
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
26°C Sunny 12:27 01-12-2025



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
The following objects are masked from 'package:base':
intersect, setdiff, setequal, union

Warning messages:
1: In file.copy(savedcopy, lib, recursive = TRUE) :
problem copying c:/users/IT-03/documents/R/win-library/4.1\00LOCK\rlang\libs\x64\rlang.dll to c:/users/IT-03/documents/R/win-library/4.1\rlang\libs\x64\rlang.dll
1: Permission denied
2: package 'dplyr' was built under R version 4.1.3
> library(tidyverse) # contains replace_na()

Warning message:
package 'tidyverse' was built under R version 4.1.3

> # Read dataset (fixed version - works even if file is not in working directory)
> retail_df <- read.csv(file.choose(), na.strings = c("", "NA"))
print("--- 1. original data (First 6 Rows) ---")
print(head(retail_df))
Check how many NAs are in each column
print(colsums(is.na(retail_df)))
clean.omit <- na.omit(retail_df)
print("--- 2. data after na.omit() ---")
print(paste("original rows:", nrow(retail_df)))
print(paste("rows remaining:", nrow(clean.omit)))
print(head(clean.omit))
Calculate average price (ignoring NAs)
clean.replace <- retail_df %>%
print("--- 3. data after replace_na() ---")
print(clean.replace[, 1])
print(head(clean.replace))
print("--- Remaining NAs after replacement ---")
print(colsums(is.na(clean.replace)))
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
26°C Sunny 12:27 01-12-2025