

## PRACTICAL NO 8

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

CODE:

```

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("tidyr")
8 library(dplyr)
9 library(tidyr) # contains replace_na()
10
11 # =====
12 # 1. CREATE AND IMPORT DATASET
13 # =====
14 #
15 # Note: 'na.strings = ""' tells R to treat empty spaces as NA (Missing values).
16 # Read dataset (fixed version - works even if file is not in working directory)
17 retail_df <- read.csv(file.choose(), na.strings = c("", "NA"))
18
19 print("--- 1. Original Data (First 6 Rows) ---")
20 print(head(retail_df))
21
22 # Check how many NAs are in each column
23 print("--- Count of Missing values per column ---")
24 print(colSums(is.na(retail_df)))
25
26 # =====
27 # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
28 # =====
29 # This is the "nuclear option". If a row has even ONE missing value, it is deleted.
30 clean_omit <- na.omit(retail_df)
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
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

```

```

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(colSums(is.na(clean_replace)))
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

```

## OUTPUT:

```

R - R 4.1.2 - ~/
> # Load necessary libraries
> # install.packages("dplyr")
> install.packages("tidyr")

Restarting R session...

> install.packages("tidyr")

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
tidyr  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/tidyr_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 'tidyr' successfully unpacked and MD5 sums checked

```

```

R - 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/lib/x64/rlang.dll to C:/Users/IT-03/Documents/R/win-library/4.1/rlang/lib/x64/rlang.dll
1: Permission denied
2: package 'dplyr' was built under R version 4.1.3

> library(tidyr) # Contains replace_na()

Warning message:
package 'tidyr' 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

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) ---")

```

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

```

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 ~ /
[1] "---- 1. Original Data (First 6 Rows) ----"
> print(head(retail_df))
  InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID Country Category
1 536365 85123A WHITE HANGING HEART T-LIGHT HOLDER 6 12/1/2010 8:26 2.55 17850 United Kingdom WHITE HANGING HEART T-LIGHT HOLDER
2 536365 71053 WHITE METAL LANTERN 6 12/1/2010 8:26 3.39 17850 United Kingdom WHITE METAL LANTERN
3 536365 84406B CREAM CUPID HEARTS COAT HANGER 8 12/1/2010 8:26 2.75 17850 United Kingdom CREAM CUPID HEARTS COAT HANGER
4 536365 84029G KNITTED UNION FLAG HOT WATER BOTTLE 6 12/1/2010 8:26 3.39 17850 United Kingdom KNITTED UNION FLAG HOT WATER BOTTLE
5 536365 84029E RED WOOLLY HOTTIE WHITE HEART. 6 12/1/2010 8:26 3.39 17850 United Kingdom RED WOOLLY HOTTIE WHITE HEART.
6 536365 22752 SET 7 BABUSHKA NESTING BOXES 2 12/1/2010 8:26 7.65 17850 United Kingdom SET 7 BABUSHKA NESTING BOXES

Price Stock Discount
1 2.55 6 NA
2 3.39 6 NA
3 2.75 8 NA
4 3.39 6 NA
5 3.39 6 NA
6 7.65 2 NA

> # 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(retail_df)))
  InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID Country Category Price Stock Discount
0 0 0 1454 0 0 0 135080 0 1454 0 0 541909

> clean_omit <- na.omit(retail_df)
> print("---- 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))
  InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID Country Category Price Stock Discount
<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 %>%
  clean_replace <- retail_df %>%

```

```

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 ~ /
<U rows> (or U-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, ])
  InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID Country Category Price Stock
3 536365 84406B CREAM CUPID HEARTS COAT HANGER 8 12/1/2010 8:26 2.75 17850 United Kingdom CREAM CUPID HEARTS COAT HANGER 2.75 8
Discount
3 FALSE
> print(head(clean_replace))
  InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice CustomerID Country Category Price Stock
1 536365 85123A WHITE HANGING HEART T-LIGHT HOLDER 6 12/1/2010 8:26 2.55 17850 United Kingdom WHITE HANGING HEART T-LIGHT HOLDER
2 536365 71053 WHITE METAL LANTERN 6 12/1/2010 8:26 3.39 17850 United Kingdom WHITE METAL LANTERN
3 536365 84406B CREAM CUPID HEARTS COAT HANGER 8 12/1/2010 8:26 2.75 17850 United Kingdom CREAM CUPID HEARTS COAT HANGER
4 536365 84029G KNITTED UNION FLAG HOT WATER BOTTLE 6 12/1/2010 8:26 3.39 17850 United Kingdom KNITTED UNION FLAG HOT WATER BOTTLE
5 536365 84029E RED WOOLLY HOTTIE WHITE HEART. 6 12/1/2010 8:26 3.39 17850 United Kingdom RED WOOLLY HOTTIE WHITE HEART.
6 536365 22752 SET 7 BABUSHKA NESTING BOXES 2 12/1/2010 8:26 7.65 17850 United Kingdom SET 7 BABUSHKA NESTING BOXES

Price Stock Discount
1 2.55 6 FALSE
2 3.39 6 FALSE
3 2.75 8 FALSE
4 3.39 6 FALSE
5 3.39 6 FALSE
6 7.65 2 FALSE

> # 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
Go to file/function Addins Project: (None)

Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
> install.packages("tidyr")

Restarting R session...
> install.packages("tidyr")

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
tidyr 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/tidyr_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 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/IT-03/Documents/R/win-library/4.1

Type here to search 26°C Sunny 12:27 01-12-2025
```

```
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 - 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/lib/x64/rlang.dll to C:/Users/IT-03/Documents/R/win-library/4.1/rlang/lib/x64/rlang.dll
1: Permission denied
2: package 'dplyr' was built under R version 4.1.3

> library(tidyr) # Contains replace_na()

Warning message:
package 'tidyr' 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[3, ])
print(head(clean_replace))
print("--- Remaining NAs after replacement ---")
print(colSums(is.na(clean_replace)))

Type here to search 26°C Sunny 12:27 01-12-2025
```