

## MVLU COLLEGE

### PRACTICAL NO :- 08

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

**CODE :-**

```
install.packages("tidyr")
library(dplyr)
library(tidyr) # Contains replace_na()

retail_df <- read.csv("C:/Users/itlab/Downloads/S100/Retail Product.csv", na.strings = c("", "NA"))

print("--- 1. Original Data (First 6 Rows) ---")
print(head(retail_df))

print("--- Count of Missing Values per Column ---")
print(colSums(is.na(retail_df)))

# 2. METHOD A: REMOVE MISSING VALUES (na.omit)
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))

# 3. METHOD B: REPLACE MISSING VALUES (replace_na)

retail_df$Price <- as.numeric(retail_df$Price)
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() ---")
print(clean_replace[3,])
print(head(clean_replace))

print("--- Remaining NAs after replacement ---")
print(colSums(is.na(clean_replace)))
```

# MVLU COLLEGE

The image displays two screenshots of the RStudio interface, showing the process of installing packages, reading a CSV file, and cleaning data.

**Top Screenshot:**

- Console:** Shows the installation of the `tidyr` package and its dependencies (`stringr`, `purrr`, `stringi`). It then reads a CSV file named `retail_df` and displays the first few rows of the data frame.
- Environment:** Lists the objects in the global environment, including `clean_replace`, `data_A`, `data_B`, `data_main`, `data_new_users`, `dropped_multiple`, `dropped_one`, `dropped_range`, and `final_dataset`.

**Bottom Screenshot:**

- Console:** Shows the removal of missing values from the `retail_df` data frame using the `na.omit` function. It then displays the first few rows of the cleaned data frame.
- Environment:** Lists the objects in the global environment, including `clean_replace`, `data_A`, `data_B`, `data_main`, `data_new_users`, `dropped_multiple`, `dropped_one`, `dropped_range`, and `final_dataset`.

# MVLU COLLEGE

The image displays two screenshots of the RStudio interface, showing R code execution and environment variables.

**Top Screenshot:**

The console shows the following code and error messages:

```
> clean_replace <- retail_df %>%
+   replace_na(list(
+     Category = "unknown",
+     Discount = 0,
+     Stock = "check warehouse",
+     Price = avg_price
+   ))
Error in `vec_assign()` :
! Can't convert from 'replacePrice' <double> to 'data$Price' <integer> due to loss of precision.
• Locations: 1
Run `rlang::last_trace()` to see where the error occurred.

> rlang::last_trace()
<error/vctrs_error_cast_lossy>
Error in `vec_assign()` :
! Can't convert from 'replacePrice' <double> to 'data$Price' <integer> due to loss of precision.
• Locations: 1
Backtrace:
 1. |retail_df %>% ...
 2. |tidyr::replace_na(...)
 3. |tidyr::replace_na.data.frame(...)
 4. |vctrs::vec_assign(x = col, i = missing, value = value, x_arg = col_arg, value_arg = value_arg)
Run `rlang::last_trace(drop = FALSE)` to see 12 hidden frames.
> rlang::last_trace(drop = FALSE)
<error/vctrs_error_cast_lossy>
Error in `vec_assign()` :
! Can't convert from 'replacePrice' <double> to 'data$Price' <integer> due to loss of precision.
• Locations: 1
Backtrace:
 1. |retail_df %>% ...
 2. |tidyr::replace_na(...)
 3. |tidyr::replace_na.data.frame(...)
 4. |vctrs::vec_assign(x = col, i = missing, value = value, x_arg = col_arg, value_arg = value_arg)
 5. |vctrs (local) ~>()
 6. |vctrs::vec_cast.integer.double(...)
```

The Environment pane shows the following variables:

Variable	Size
clean_replace	4362 obs. of 5 variables
data_A	500 obs. of 3 variables
data_B	500 obs. of 5 variables
data_main	500 obs. of 10 variables
data_new_users	2 obs. of 7 variables
dropped_multiple	497 obs. of 6 variables
dropped_one	497 obs. of 7 variables
dropped_range	497 obs. of 5 variables
final_dataset	502 obs. of 11 variables

**Bottom Screenshot:**

The console shows the following code and output:

```
> retail_df$Price <- as.numeric(retail_df$Price)
Error in retail_df$Price <- as.numeric(retail_df$Price) :
could not find function ">$<-"

> retail_df$Price <- as.numeric(retail_df$Price)
> 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() ---"
> print(clean_replace)
  Category Price Rating Stock Discount
1 unknown 5548.000 1.870322 check warehouse 0
2 unknown 3045.000 4.757798 check warehouse 38
3 unknown 4004.000 NA In Stock 0
4 unknown 4808.000 1.492085 check warehouse 33
5 unknown 1817.000 NA Out of Stock 23
6 unknown 3522.000 NA check warehouse 0
7 C 667.000 3.668341 In Stock 41
8 A 7125.000 4.983998 Out of Stock 7
9 A 2777.000 2.678384 In Stock 6
10 unknown 463.000 4.626187 check warehouse 3
11 A 1151.000 2.947838 check warehouse 0
12 A 3772.000 4.890750 In Stock 45
13 unknown 7719.000 2.982242 In Stock 4
14 C 8416.000 1.270943 check warehouse 29
15 B 8530.000 NA check warehouse 10
16 A 7936.000 3.032832 In Stock 44
17 B 9319.000 3.479064 In Stock 28
18 unknown 5016.971 4.097464 check warehouse 40
19 D 2066.000 NA check warehouse 30
20 B 1784.000 NA check warehouse 12
21 unknown 5589.000 NA check warehouse 39
22 C 5016.971 NA Out of Stock 25
23 unknown 4176.000 2.264051 In Stock 42
24 D 407.000 2.774270 In Stock 44
```

The Environment pane shows the same variables as the top screenshot.

# MVLU COLLEGE

```
R - R 4.5.2 - ~/R
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Source
Console Terminal Background Jobs

180 A 8/13.000 2.380993 Check Warehouse 30
181 A 1269.000 NA In Stock 38
182 unknown 8069.000 NA In Stock 31
183 D 6751.000 2.337783 Check Warehouse 2
184 C 4970.000 2.365565 Check Warehouse 5
185 unknown 2323.000 NA Check Warehouse 0
186 unknown 5016.971 NA Check Warehouse 0
187 A 5515.000 4.709545 out of Stock 0
188 unknown 8846.000 1.364083 Check Warehouse 4
189 unknown 2584.000 NA out of Stock 1
190 C 7911.000 2.765739 out of Stock 37
191 D 6799.000 NA out of Stock 29
192 unknown 4620.000 4.340548 Check Warehouse 10
193 unknown 9424.000 1.076074 out of Stock 0
194 unknown 8409.000 1.696636 out of Stock 29
195 unknown 9063.000 3.624531 out of Stock 0
196 unknown 8405.000 1.218936 In Stock 48
197 unknown 8070.000 4.220246 Check Warehouse 20
198 D 7564.000 2.576793 In Stock 35
199 D 8555.000 2.381393 out of Stock 22
200 unknown 1870.000 NA In Stock 32
[ reached 'max' / getoption("max.print") -- omitted 4162 rows ]
> print(head(clean_replace))
  Category Price Rating Stock Discount
1 unknown 5548 1.870322 Check Warehouse 0
2 unknown 3045 4.757798 Check Warehouse 38
3 unknown 4004 NA In Stock 0
4 unknown 4808 1.492085 Check Warehouse 33
5 unknown 1817 NA out of Stock 23
6 unknown 3522 NA Check Warehouse 0
>
> print(colSums(is.na(clean_replace)))
+ print(colSums(is.na(clean_replace)))
Error: unexpected symbol in:
"print(colSums(is.na(clean_replace)))
print"
>
> print(colSums(is.na(clean_replace)))
Category Price Rating Stock Discount
0 0 2050 0 0
> |
```

Environment History Connections Tutorial

R - Global Environment

- clean\_replace 4362 obs. of 5 variables
- data\_A 500 obs. of 3 variables
- data\_B 500 obs. of 5 variables
- data\_main 500 obs. of 10 variables
- data\_new\_users 2 obs. of 7 variables
- dropped\_multiple 497 obs. of 6 variables
- dropped\_one 497 obs. of 7 variables
- dropped\_range 497 obs. of 5 variables
- final\_dataset 502 obs. of 11 variables

Files Plots Packages Help Viewer Presentation

New Folder New File Delete Rename More

Home