

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("tidyverse")
library(dplyr)
library(tidyverse) # 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 screenshot shows the RStudio interface with two sessions running. The top session is titled 'R 4.5.2' and contains R code for installing packages and reading a CSV file. The bottom session is also titled 'R 4.5.2' and shows more complex data manipulation and error handling. Both sessions have tabs for 'Source', 'Console', 'Terminal', and 'Background Jobs'. The right side of the interface features the 'Environment' tab, which lists various datasets and their characteristics. The status bar at the bottom indicates battery level, signal strength, and system date.

NANDINI PANDIT S100

DATA ANALYSIS WITH SAS/SPSS/R.

MVLU COLLEGE

The screenshot shows the RStudio interface with the following details:

- File Edit Code View Plots Session Build Debug Profile Tools Help**
- Addins** dropdown menu is open.
- Source** tab is selected.
- Console** tab is active, showing R code and its execution results.
- Background Jobs** tab is visible.
- R 4.5.2 ~** is displayed in the top right.
- Error in `vec_assign()`:** A warning message is present:

```
! Can't convert from 'replace$price' <double> to 'data$price' <integer> due to loss of precision.  
• Locations: 1  
Run rlang:::last_trace() to see where the error occurred.
```
- > rlang:::last_trace()** command was run, showing the call stack:

```
<- vecr:::vec_cast_llosy>  
Error in `vec_assign()`:  
! Can't convert from 'replace$price' <double> to 'data$price' <integer> due to loss of precision.  
• Locations: 1  
---  
Backtrace:  
1. |--retail_df %>% ...  
2. |--tidy:::replace_na(...)  
3. |--tidy:::replace_na.data.frame(...)  
4. |--vctr:::vec_assign(x = col, i = missing, value = value, x_arg = col_arg, value_arg = value_arg)  
> rlang:::last_trace(drop = FALSE)  
<error>/vctr_error_cast_llosy>  
Error in `vec_assign()`:  
! Can't convert from 'replace$price' <double> to 'data$price' <integer> due to loss of precision.  
• Locations: 1  
---  
Backtrace:  
1. |--retail_df %>% ...  
2. |--tidy:::replace_na(...)  
3. |--tidy:::replace_na.data.frame(...)  
4. |--vctr:::vec_assign(x = col, i = missing, value = value, x_arg = col_arg, value_arg = value_arg)  
5. |--vctr (local) <fn>()  
6. |--vctr:::vec_cast.integer.double(...)
```
- Environment**, **History**, **Connections**, **Tutorial** tabs are visible at the top right.
- R** dropdown menu is open, showing the global environment.
- Project (None)** is listed at the top right.
- File Plots Packages Help Viewer Presentation** menu is at the bottom left.
- New Folder**, **New File**, **Delete**, **Rename** buttons are at the bottom left.
- Home** button is highlighted.
- System tray icons** are at the bottom right.
- Taskbar icons** are at the bottom center.

MVLU COLLEGE

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Console Terminal Background Jobs
R 4.5.2 - ~
180 A 8/13,000 Z.380993 Check Warehouse 30
181 A 1269,000 NA In Stock 31
182 Unknown 8069,000 NA In Stock 31
183 D 6751,000 2.337783 Check warehouse 2
184 C 4970,000 2.365365 Check warehouse 5
185 Unknown 2323,000 NA Check warehouse 0
186 Unknown 5016,971 NA Check warehouse 0
187 D 5016,971 4.709545 Out of Stock 0
188 Unknown 8846,000 1.364043 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 9833,000 3.624531 Out of Stock 0
196 Unknown 8409,000 1.076074 In Stock 48
197 Unknown 8070,000 4.220246 Check warehouse 20
198 D 7564,000 2.576792 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 5348 1.870323 Check warehouse 0
2 Unknown 5445 4.737788 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
Import Dataset 170 MB
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