

MVLU COLLEGE

PRACTICAL NO. 8

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

The screenshot shows the RStudio interface with a script editor on the left and the Environment pane on the right. The script contains the following code:

```
1 install.packages("tidyr")
2 library(dplyr)
3 library(tidyr)
4 retail_df <- read.csv("C:/Users/itlab/OneDrive/documents/S105/Retail Product - Retail Product.csv", na.strings = c("
5 print("---- 1. Original Data (First 6 Rows) ----")
6 print(head(retail_df))
7 print("---- Count of Missing Values per Column ----")
8 print(colSums(is.na(retail_df)))
9 clean_omit <- na.omit(retail_df)
10 print("---- 2. Data after na.omit() ----")
11 print(paste("Original rows:", nrow(retail_df)))
12 print(paste("Rows remaining:", nrow(clean_omit)))
13 print(head(clean_omit))
14 retail_df$Price <- as.numeric(retail_df$Price)
15 avg_price <- mean(retail_df$Price, na.rm = TRUE)
16 clean_replace <- retail_df %>%
17   replace_na(list(
18     category = "unknown",
19     discount = 0,
20     stock = "check warehouse",
21     price = avg_price
22   ))
23 cat("\n--- 3. Data after replace_na() ---\n")
24 print(clean_replace[3, ])
25 print(head(clean_replace))
26 cat("\n--- Remaining rows after Replacement ---\n")
27 print(colSums(is.na(clean_replace)))
28
```

The Environment pane on the right shows the following objects:

Object	Size
appended	298 obs. of 17 variables
clean_omit	540 obs. of 5 variables
clean_replace	4362 obs. of 5 variables
cleaned_BMW_Sales	99 obs. of 11 variables
cleaned_car_price	199 obs. of 10 variables
df1	99 obs. of 11 variables
df2	199 obs. of 10 variables
dropped_multiple	4573 obs. of 13 variables
dropped_one	4573 obs. of 14 variables
dropped_range	4573 obs. of 11 variables
housing	4573 obs. of 15 variables
merged_full	294 obs. of 17 variables
merged_inner	4 obs. of 17 variables
merged_left	99 obs. of 17 variables
range_cols	4573 obs. of 6 variables
retail_df	4362 obs. of 5 variables
retail_product_Re	4362 obs. of 5 variables
selected_cols	4573 obs. of 3 variables
spotify	4573 obs. of 15 variables
spotify_data_clean	8573 obs. of 15 variables
starts_with_track	4573 obs. of 5 variables

The screenshot shows the RStudio interface with a script editor on the left and the Environment pane on the right. The script contains the following code:

```
> library(readr)
> retail_product_retail_product <- read_csv("S105/Retail Product - Retail Product.csv")
Rows: 4362 Columns: 5
--- Column specification
Delimiter: ","
chr (2): category, stock
dbl (3): Price, Rating, Discount

I use 'spec()' to retrieve the full column specification for this data.
I specify the column types or set 'show_col_types = FALSE' to quiet this message.
> view(retail_product_retail_product)
> 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/itlab/AppData/Local/R/win-library/4.5'
(as 'lib' is unspecified)

also installing the dependencies 'stringi', 'purrr', 'stringr'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/stringi_1.8.7.zip'
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/purrr_1.2.0.zip'
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/stringr_1.6.0.zip'
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/tidyr_1.3.1.zip'
package 'stringi' successfully unpacked and MD5 sums checked
package 'purrr' successfully unpacked and MD5 sums checked
package 'stringr' successfully unpacked and MD5 sums checked
package 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/itlab/AppData/Local/Temp/RtmpQVEAM1/downloaded_packages
> library(dplyr)
> library(tidyr)
> retail_df <- read_csv("Retail Product.csv", na.strings = c("", "NA"))

Error in file(file, "rt") : cannot open the connection

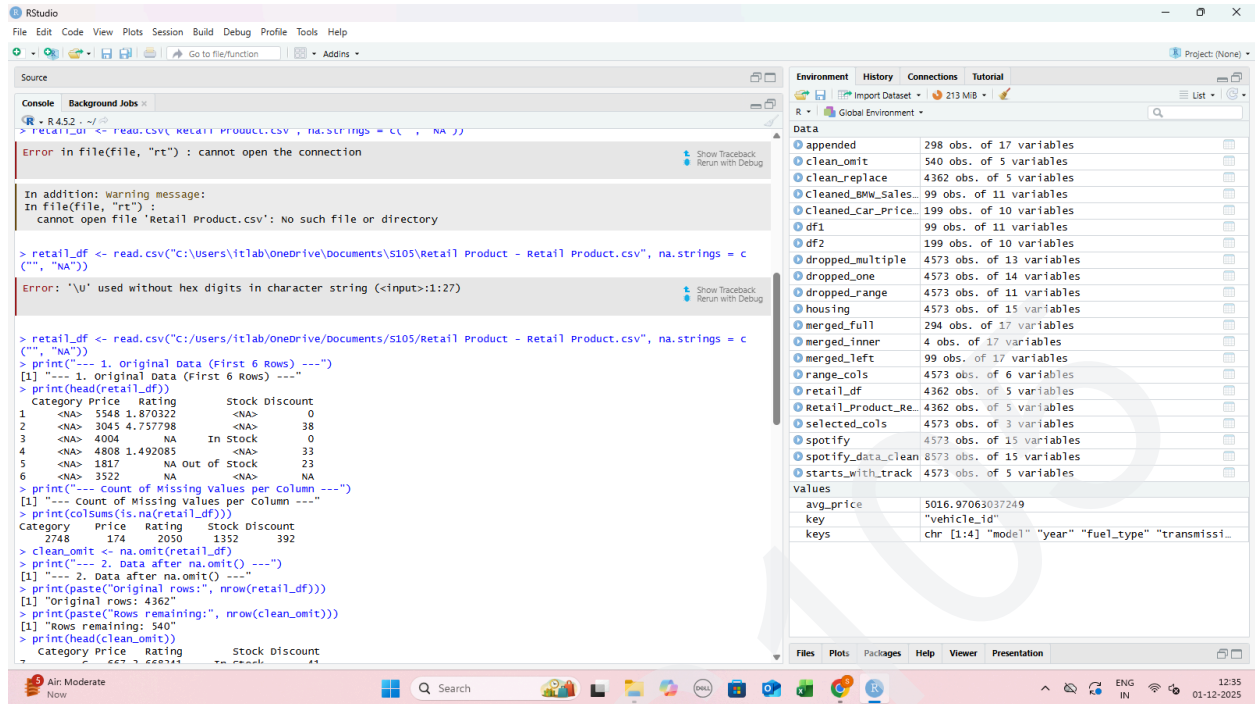
In addition: warning message:
Warning: ...
```

The Environment pane on the right shows the following objects:

Object	Size
appended	298 obs. of 17 variables
clean_omit	540 obs. of 5 variables
clean_replace	4362 obs. of 5 variables
cleaned_BMW_Sales	99 obs. of 11 variables
cleaned_car_price	199 obs. of 10 variables
df1	99 obs. of 11 variables
df2	199 obs. of 10 variables
dropped_multiple	4573 obs. of 13 variables
dropped_one	4573 obs. of 14 variables
dropped_range	4573 obs. of 11 variables
housing	4573 obs. of 15 variables
merged_full	294 obs. of 17 variables
merged_inner	4 obs. of 17 variables
merged_left	99 obs. of 17 variables
range_cols	4573 obs. of 6 variables
retail_df	4362 obs. of 5 variables
retail_product_Re	4362 obs. of 5 variables
selected_cols	4573 obs. of 3 variables
spotify	4573 obs. of 15 variables
spotify_data_clean	8573 obs. of 15 variables
starts_with_track	4573 obs. of 5 variables

MVLU COLLEGE

PRACTICAL NO. 8



The screenshot shows the RStudio interface. The console on the left displays the following R code and its output:

```
> retail_df <- read.csv("rt")
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

> retail_df <- read.csv("C:/Users/itlab/OneDrive/Documents/S105/Retail Product - Retail Product.csv", na.strings = c("", "NA"))
Error: '\u' used without hex digits in character string (<input>:1:27)

> retail_df <- read.csv("C:/Users/itlab/OneDrive/Documents/S105/Retail Product - Retail Product.csv", na.strings = c("", "NA"))
> print("--- 1. Original data (First 6 Rows) ---")
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(retail_df))
  Category Price Rating Stock Discount
1 <NA> 5548 1.870322 <NA> 0
2 <NA> 3045 4.757798 <NA> 38
3 <NA> 4004 NA In Stock 0
4 <NA> 4808 1.492085 <NA> 33
5 <NA> 1817 NA out of Stock 23
6 <NA> 3522 NA <NA> NA
> print("--- Count of Missing values per column ---")
[1] "--- Count of Missing values per column ---"
> print(colSums(is.na(retail_df)))
Category Price Rating Stock Discount
2748 174 2050 1352 392
> 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: 4362"
> print(paste("rows remaining:", nrow(clean_omit)))
[1] "rows remaining: 540"
> print(head(clean_omit))
  Category Price Rating Stock Discount
1 2748 5548 1.870322 0
2 3045 4.757798 38
3 4004 NA In Stock 0
4 4808 1.492085 33
5 1817 NA out of Stock 23
6 3522 NA  NA NA
```

The Environment pane on the right shows the following data objects:

Object	Size
appended	298 obs. of 17 variables
clean_omit	540 obs. of 5 variables
clean_replace	4362 obs. of 5 variables
cleaned_BMW_Sales	99 obs. of 11 variables
cleaned_car_Price	199 obs. of 10 variables
df1	99 obs. of 11 variables
df2	199 obs. of 10 variables
dropped_multiple	4573 obs. of 13 variables
dropped_one	4573 obs. of 14 variables
dropped_range	4573 obs. of 11 variables
housing	4573 obs. of 15 variables
merged_full	294 obs. of 17 variables
merged_inner	4 obs. of 17 variables
merged_left	99 obs. of 17 variables
range_cols	4362 obs. of 6 variables
retail_df	4362 obs. of 5 variables
Retail_Product_Re	4362 obs. of 5 variables
selected_cols	4573 obs. of 3 variables
spotify	4573 obs. of 15 variables
spotify_data_clean	8573 obs. of 15 variables
starts_with_track	4573 obs. of 5 variables

The Files pane at the bottom shows the following files:

File	Size
avg_price	5016.97063037249
key	"vehicle_id"
keys	chr [1:4] "model" "year" "fuel_type" "transmissi...

MVLU COLLEGE

PRACTICAL NO. 8

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Background Jobs
R - R452 - ~/
[1] "Rows remaining: 540"
> print(head(clean_omit))
  Category Price Rating Stock Discount
7        C  667 3.668341 In Stock    41
8        A 7125 4.983998 out of Stock  7
9        A 2777 2.678384 In Stock    6
12       A 3772 4.890750 In Stock    45
16       A 7936 3.032832 In Stock    44
17       B 9319 3.479064 In Stock    28
> avg_price <- mean(retail_df$Price, na.rm = TRUE)
>
> clean_replace <- retail_df %>%
+   replace_na(list(
+     category = "unknown",
+     discount = 0,
+     stock = "check warehouse",
+     price = avg_price
+   ))
Error in `vec_assign()`:
! Can't convert from 'replacePrice' <double> to 'data$Price' <integer> due to loss of precision.
+ Locations: 1
Run `langr::last_trace()` to see where the error occurred.
> 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
+   ))
>
> cat("\n--- 3. Data after replace_na() ---\n")
--- 3. Data after replace_na() ---
> print(clean_replace[, ])
  Category Price Rating Stock Discount
3 unknown  4004    NA In Stock      0
> 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("--- Remaining NAs after replacement ---")
[1] "--- Remaining NAs after replacement ---"
> print(colSums(is.na(clean_replace)))
+ print(colSums(is.na(clean_replace)))
Error: unexpected symbol in:
"print(colSums(is.na(clean_replace))
print"
> cat("\n--- Remaining NAs after Replacement ---\n")
--- Remaining NAs after Replacement ---
> print(colSums(is.na(clean_replace)))
  Category Price Rating Stock Discount
0         0         0    2050         0         0
> |
```

Environment History Connections Tutorial

R - Global Environment

Data

- appended 298 obs. of 17 variables
- clean_omit 540 obs. of 5 variables
- clean_replace 4362 obs. of 5 variables
- cleaned_BMW_Sales 99 obs. of 11 variables
- cleaned_car_Price 199 obs. of 10 variables
- df1 99 obs. of 11 variables
- df2 199 obs. of 10 variables
- dropped_multiple 4573 obs. of 13 variables
- dropped_one 4573 obs. of 14 variables
- dropped_range 4573 obs. of 11 variables
- housing 4573 obs. of 15 variables
- merged_full 294 obs. of 17 variables
- merged_inner 4 obs. of 17 variables
- merged_left 99 obs. of 17 variables
- range_cols 4573 obs. of 6 variables
- retail_df 4362 obs. of 5 variables
- retail_Product_Re 4362 obs. of 5 variables
- selected_cols 4573 obs. of 3 variables
- spotify 4573 obs. of 15 variables
- spotify_data_clean 8573 obs. of 15 variables
- starts_with_track 4573 obs. of 5 variables

Values

avg_price	5016.97063037249
key	"vehicle_id"
keys	chr [1:4] "model" "year" "fuel_type" "transmissi...

Files Plots Packages Help Viewer Presentation

12:35 01-12-2025

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Background Jobs
R - R452 - ~/
> 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
+   ))
>
> cat("\n--- 3. Data after replace_na() ---\n")
--- 3. Data after replace_na() ---
> print(clean_replace[, ])
  Category Price Rating Stock Discount
3 unknown  4004    NA In Stock      0
> 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("--- Remaining NAs after replacement ---")
[1] "--- Remaining NAs after replacement ---"
> print(colSums(is.na(clean_replace)))
+ print(colSums(is.na(clean_replace)))
Error: unexpected symbol in:
"print(colSums(is.na(clean_replace))
print"
> cat("\n--- Remaining NAs after Replacement ---\n")
--- Remaining NAs after Replacement ---
> print(colSums(is.na(clean_replace)))
  Category Price Rating Stock Discount
0         0         0    2050         0         0
> |
```

Environment History Connections Tutorial

R - Global Environment

Data

- appended 298 obs. of 17 variables
- clean_omit 540 obs. of 5 variables
- clean_replace 4362 obs. of 5 variables
- cleaned_BMW_Sales 99 obs. of 11 variables
- cleaned_car_Price 199 obs. of 10 variables
- df1 99 obs. of 11 variables
- df2 199 obs. of 10 variables
- dropped_multiple 4573 obs. of 13 variables
- dropped_one 4573 obs. of 14 variables
- dropped_range 4573 obs. of 11 variables
- housing 4573 obs. of 15 variables
- merged_full 294 obs. of 17 variables
- merged_inner 4 obs. of 17 variables
- merged_left 99 obs. of 17 variables
- range_cols 4573 obs. of 6 variables
- retail_df 4362 obs. of 5 variables
- retail_Product_Re 4362 obs. of 5 variables
- selected_cols 4573 obs. of 3 variables
- spotify 4573 obs. of 15 variables
- spotify_data_clean 8573 obs. of 15 variables
- starts_with_track 4573 obs. of 5 variables

Values

avg_price	5016.97063037249
key	"vehicle_id"
keys	chr [1:4] "model" "year" "fuel_type" "transmissi...

Files Plots Packages Help Viewer Presentation

12:35 01-12-2025