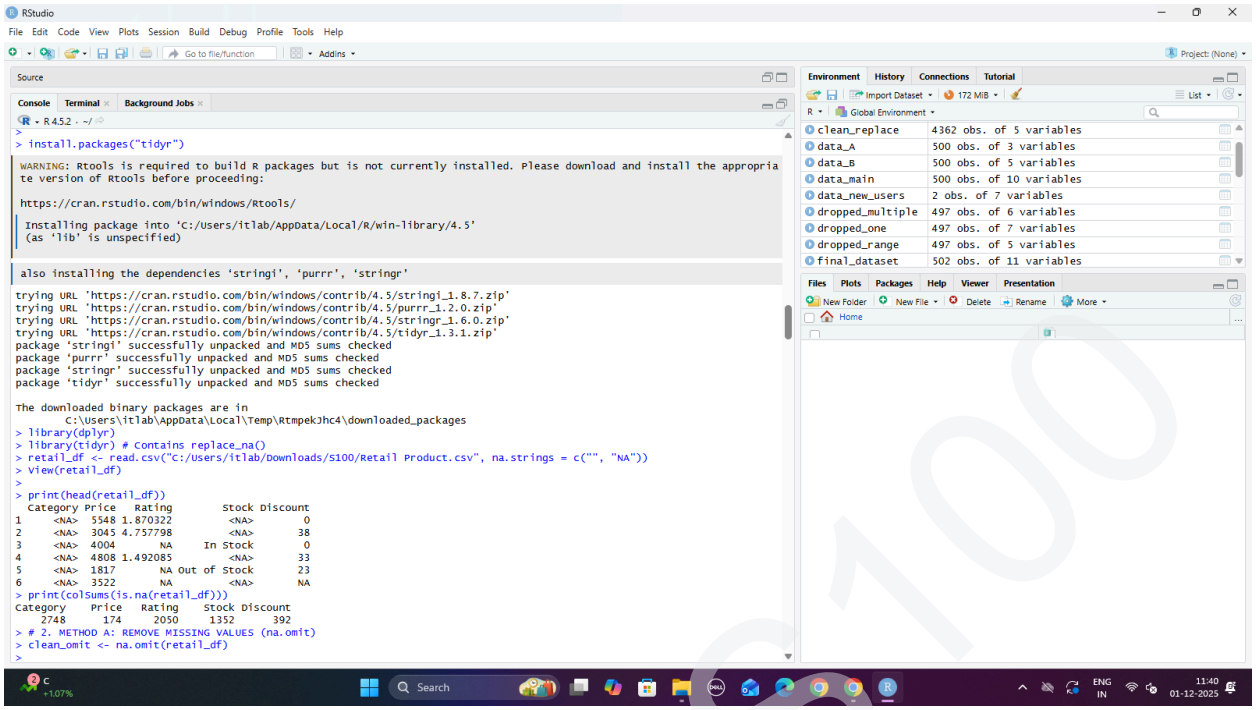
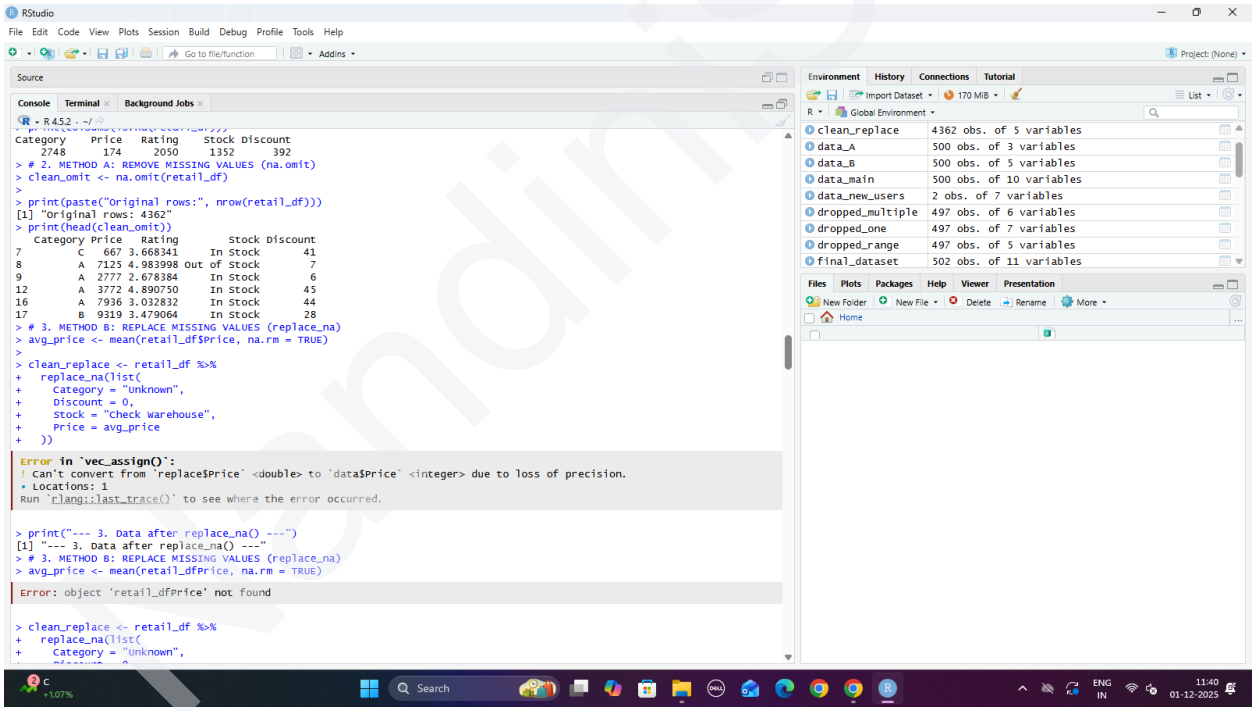
PRACTICAL NO :- 08
AIM :- Applying basic data cleaning functions: handling missing values using na.omit()/replace_na() in R. import dataset.
CODE :-

```
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



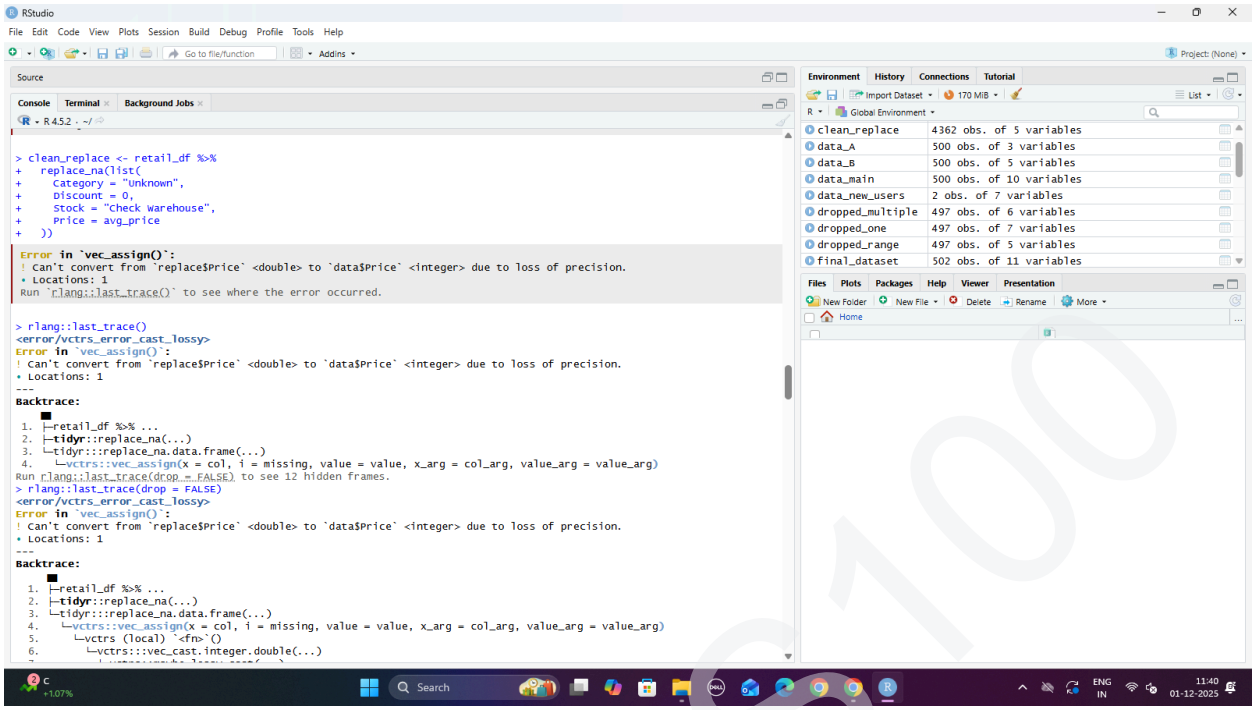
```
> install.packages("tidyr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropria
te version of Rtools before proceeding:

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\RtmpekJhc4\downloaded_packages
> library(dplyr)
> library(tidyr) # Contains replace_na()
> retail_df <- read.csv("C:/Users/itlab/Downloads/S100/Retail Product.csv", na.strings = c("", "NA"))
> View(retail_df)
>
> 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(colSums(is.na(retail_df)))
Category    Price   Rating    Stock Discount
    2748      174     2050     1352      392
> # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
> clean_omit <- na.omit(retail_df)
>
```

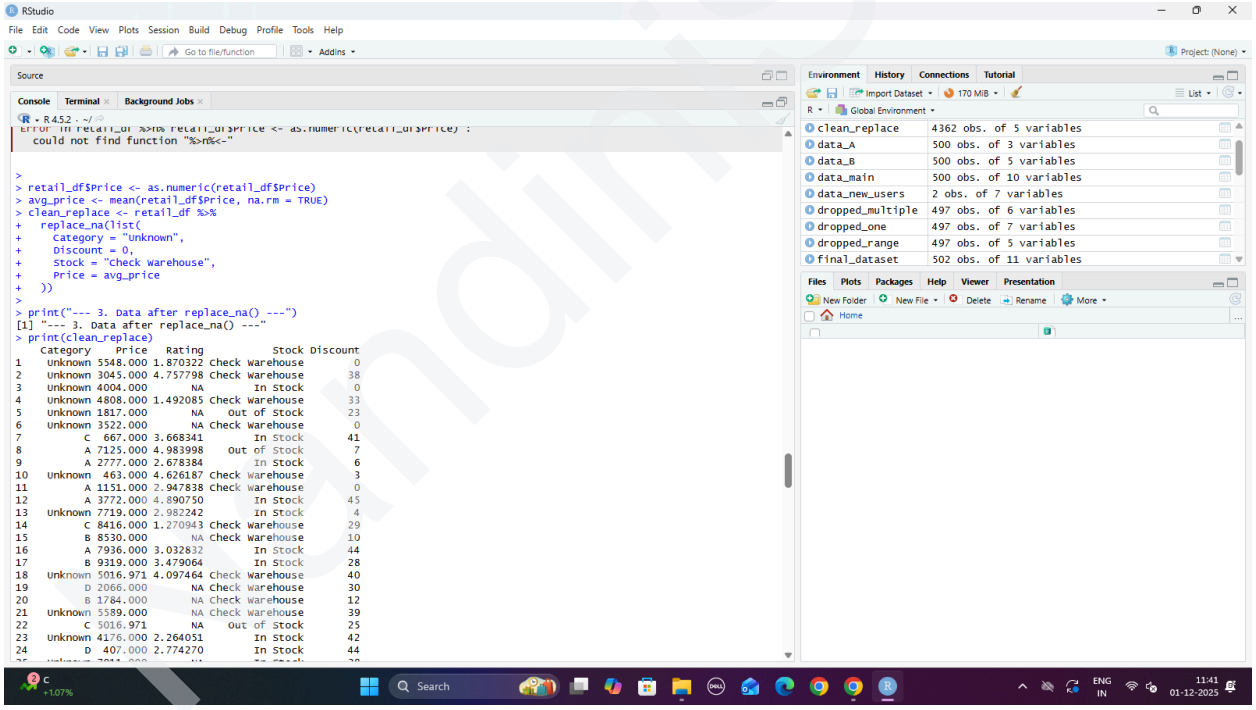


```
Category    Price   Rating    Stock Discount
    2748      174     2050     1352      392
> # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
> clean_omit <- na.omit(retail_df)
>
> print(paste("Original rows:", nrow(retail_df)))
[1] "Original rows: 4362"
> print(head(clean_omit))
  Category Price    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
> # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
> avg_price <- mean(retail_df$Price, na.rm = TRUE)
>
> clean_replace <- retail_df %>%
+   replace_na(list(
+     Category = "Unknown",
+     Discount = 0,
+     Stock = "Check Warehouse",
+     Price = avg_price
+   ))
Error in `vec_assign()`:
! Can't convert from `replace$Price` <double> to `data$Price` <integer> due to loss of precision.
• Locations: 1
Run `rlang::last_trace()` to see where the error occurred.

> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
> avg_price <- mean(retail_dfPrice, na.rm = TRUE)
Error: object 'retail_dfPrice' not found

> clean_replace <- retail_df %>%
+   replace_na(list(
+     Category = "Unknown",
```

NANDINI PANDIT S100
DATA ANALYSIS WITH SAS/SPSS/R.

MVLU COLLEGE

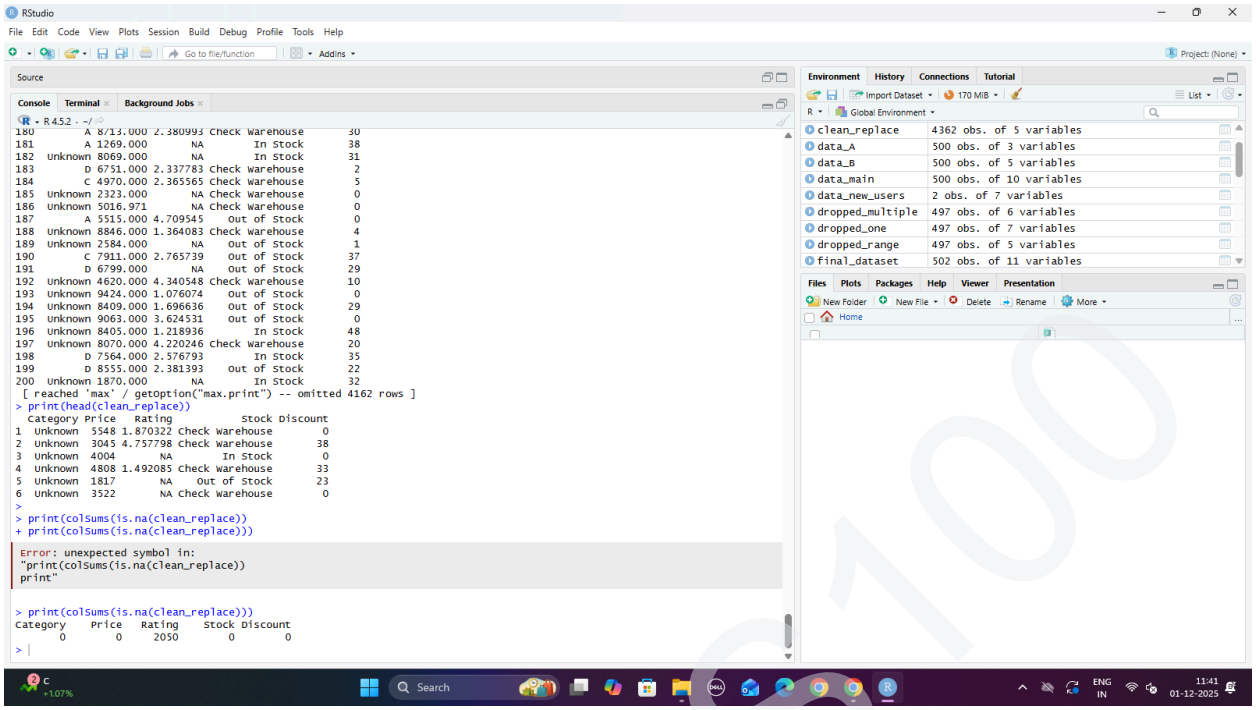




NANDINI PANDIT S100

DATA ANALYSIS WITH SAS/SPSS/R.

MVLU COLLEGE



NANDINI PANDIT S100
DATA ANALYSIS WITH SAS/SPSS/R.