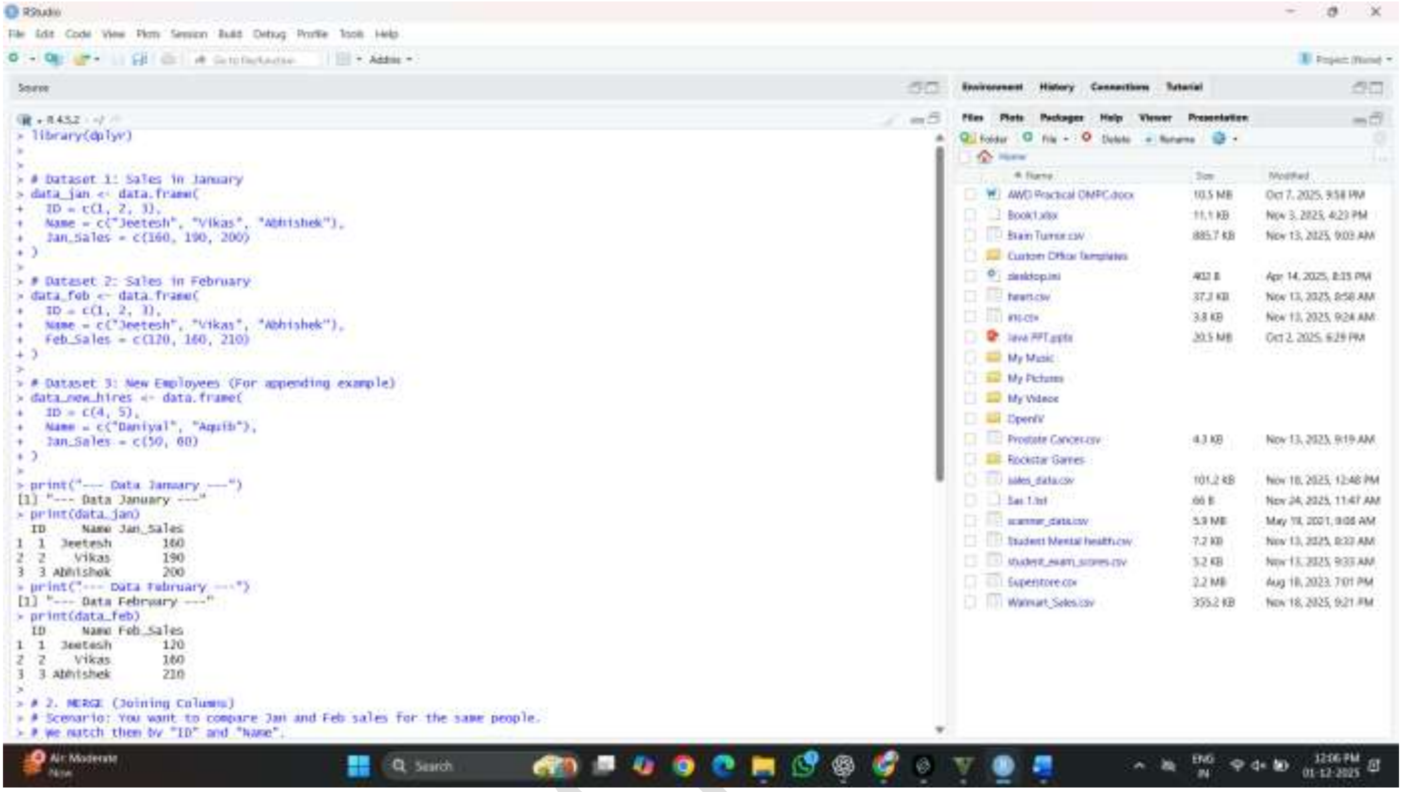
# Sheth L.U.J & Sir M.V College
# SAS/SSPS/R Programming
# Practical No. 6 to 10

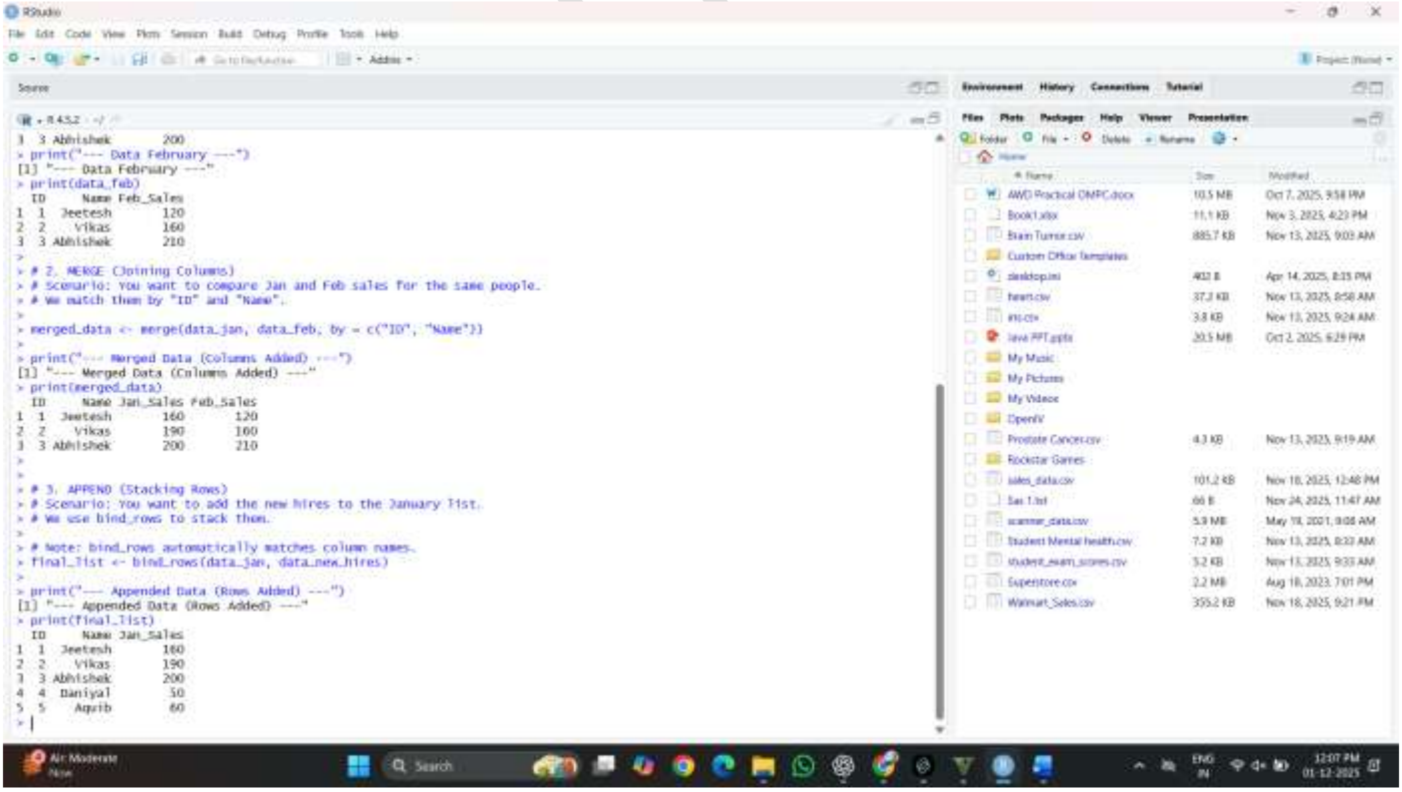
Aim: 6. Combining and appending datasets using merge() or bind_rows() in R.

Output:



```
> library(dplyr)
>
>
> # Dataset 1: Sales in January
> data_jan <- data.frame(
+   ID = c(1, 2, 3),
+   Name = c("Jeetesh", "Vikas", "Abhishek"),
+   Jan_Sales = c(160, 190, 200)
+ )
>
> # Dataset 2: Sales in February
> data_feb <- data.frame(
+   ID = c(1, 2, 3),
+   Name = c("Jeetesh", "Vikas", "Abhishek"),
+   Feb_Sales = c(120, 160, 210)
+ )
>
> # Dataset 3: New Employees (For appending example)
> data_new_hires <- data.frame(
+   ID = c(4, 5),
+   Name = c("Daniyal", "Aquib"),
+   Jan_Sales = c(50, 60)
+ )
>
> print("--- Data January ---")
[1] "--- Data January ---"
> print(data_jan)
  ID     Name Jan_Sales
1  1  Jeetesh       160
2  2    Vikas       190
3  3 Abhishek       200
> print("--- Data February ---")
[1] "--- Data February ---"
> print(data_feb)
  ID     Name Feb_Sales
1  1  Jeetesh       120
2  2    Vikas       160
3  3 Abhishek       210
>
> # 2. MERGE (Joining Columns)
> # Scenario: You want to compare Jan and Feb sales for the same people.
> # We match them by "ID" and "Name".
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |



```
3  3 Abhishek       200
> print("--- Data February ---")
[1] "--- Data February ---"
> print(data_feb)
  ID     Name Feb_Sales
1  1  Jeetesh       120
2  2    Vikas       160
3  3 Abhishek       210
>
> # 2. MERGE (Joining Columns)
> # Scenario: You want to compare Jan and Feb sales for the same people.
> # We match them by "ID" and "Name".
>
> merged_data <- merge(data_jan, data_feb, by = c("ID", "Name"))
>
> print("--- Merged Data (Columns Added) ---")
[1] "--- Merged Data (Columns Added) ---"
> print(merged_data)
  ID     Name Jan_Sales Feb_Sales
1  1  Jeetesh       160       120
2  2    Vikas       190       160
3  3 Abhishek       200       210
>
>
> # 3. APPEND (Stacking Rows)
> # Scenario: You want to add the new hires to the January list.
> # We use bind_rows to stack them.
>
> # Note: bind_rows automatically matches column names.
> final_list <- bind_rows(data_jan, data_new_hires)
>
> print("--- Appended Data (Rows Added) ---")
[1] "--- Appended Data (Rows Added) ---"
> print(final_list)
  ID     Name Jan_Sales
1  1  Jeetesh       160
2  2    Vikas       190
3  3 Abhishek       200
4  4  Daniyal        50
5  5    Aquib        60
> 
```

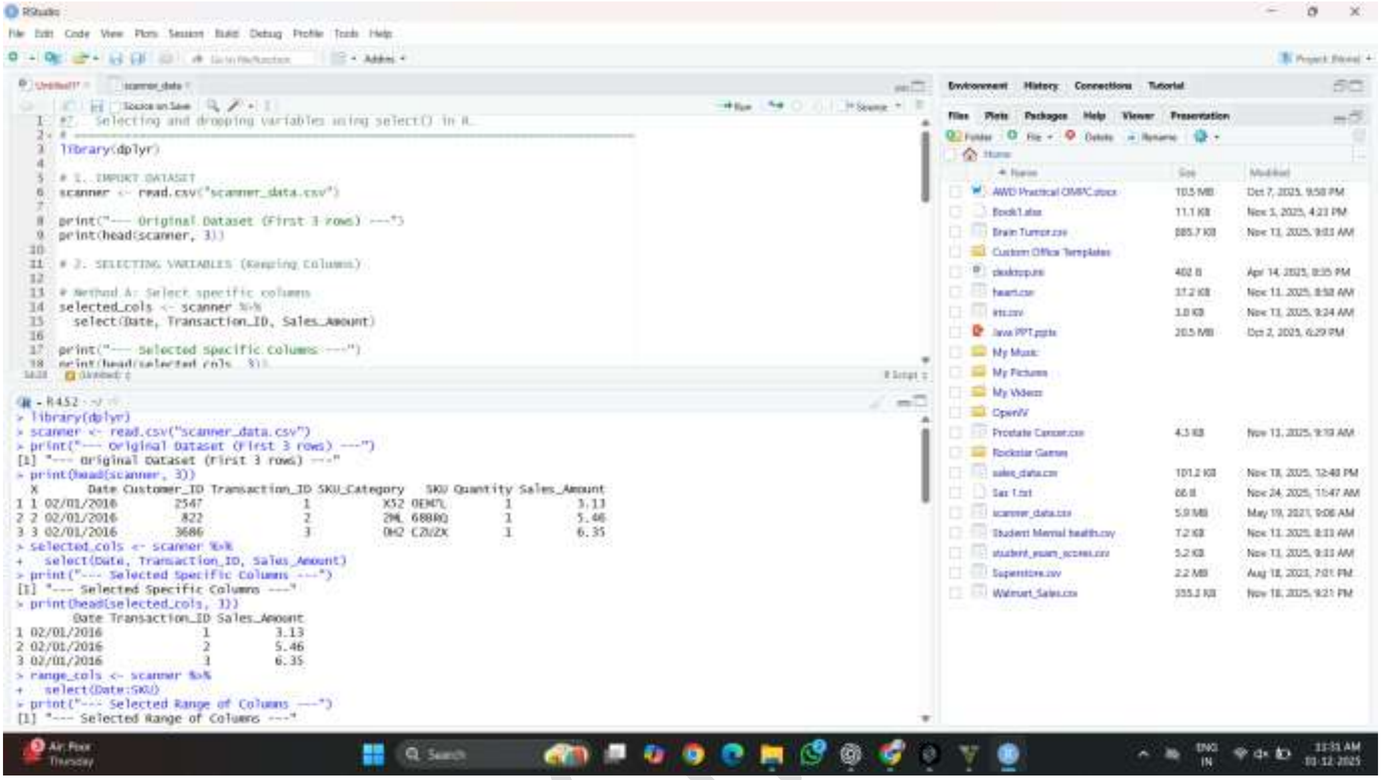
| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |

Jeetesh Shirganoor
S125

# Sheth L.U.J & Sir M.V College
# SAS/SSPS/R Programming
# Practical No. 6 to 10

Aim: 7. Selecting and dropping variables using select() in R. import dataset.

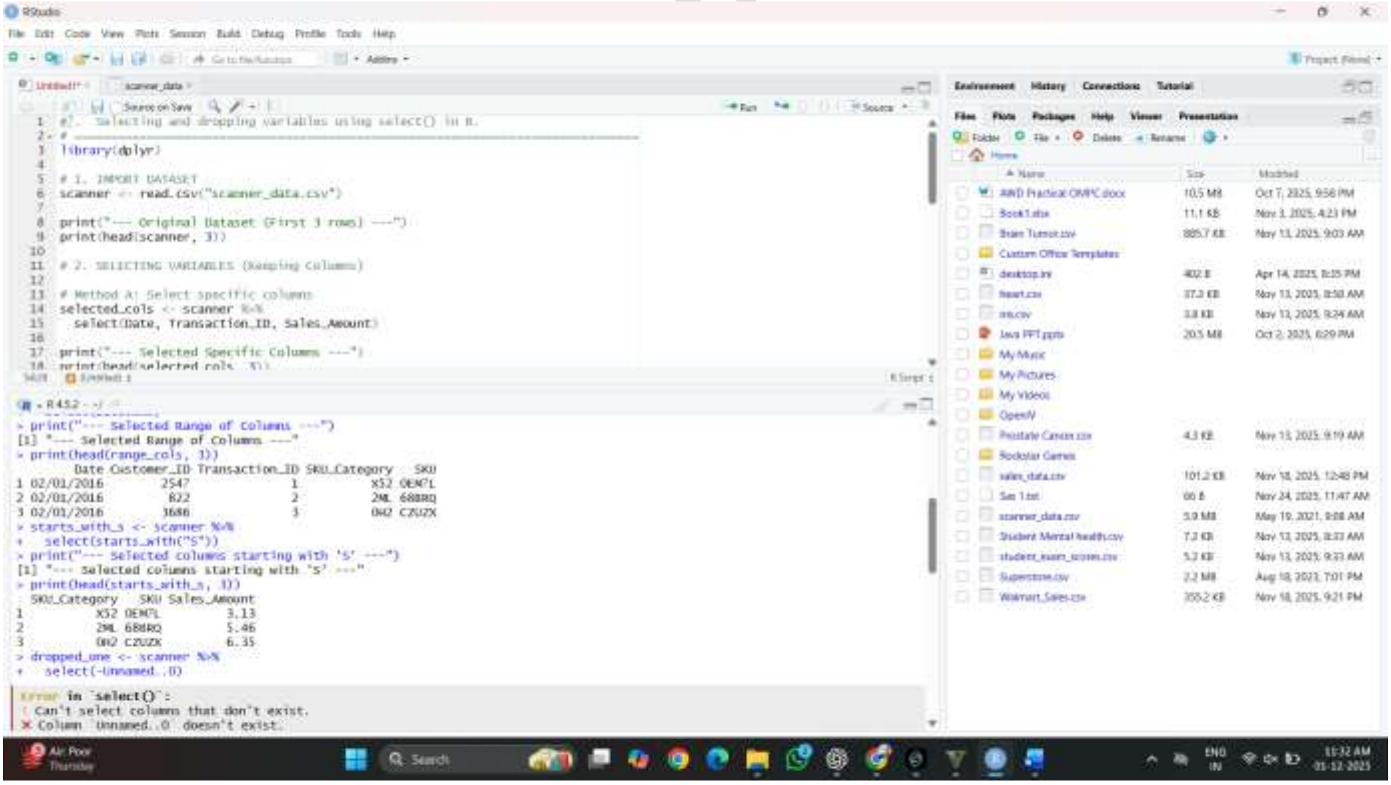
Output:



```
#7.  Selecting and dropping variables using select() in R.
# ------------------------------------------------------------
library(dplyr)

# 1. IMPORT DATASET
scanner <- read.csv("scanner_data.csv")

print("--- Original Dataset (First 3 rows) ---")
print(head(scanner, 3))

# 2. SELECTING VARIABLES (Keeping Columns)

# Method A: Select specific columns
selected_cols <- scanner %>%
  select(Date, Transaction_ID, Sales_Amount)

print("--- Selected Specific Columns ---")
print(head(selected_cols, 3))
```

```
> library(dplyr)
> scanner <- read.csv("scanner_data.csv")
> print("--- Original Dataset (First 3 rows) ---")
[1] "--- Original Dataset (First 3 rows) ---"
> print(head(scanner, 3))
  X       Date Customer_ID Transaction_ID SKU_Category    SKU Quantity Sales_Amount
1 1 02/01/2016        2547              1          X52 0EM7L        1         3.13
2 2 02/01/2016         822              2          2ML 68BRQ        1         5.46
3 3 02/01/2016        3686              3          0H2 CZUZX        1         6.35
> selected_cols <- scanner %>%
+   select(Date, Transaction_ID, Sales_Amount)
> print("--- Selected Specific Columns ---")
[1] "--- Selected Specific Columns ---"
> print(head(selected_cols, 3))
        Date Transaction_ID Sales_Amount
1 02/01/2016              1         3.13
2 02/01/2016              2         5.46
3 02/01/2016              3         6.35
> range_cols <- scanner %>%
+   select(Date:SKU)
> print("--- Selected Range of Columns ---")
[1] "--- Selected Range of Columns ---"
```

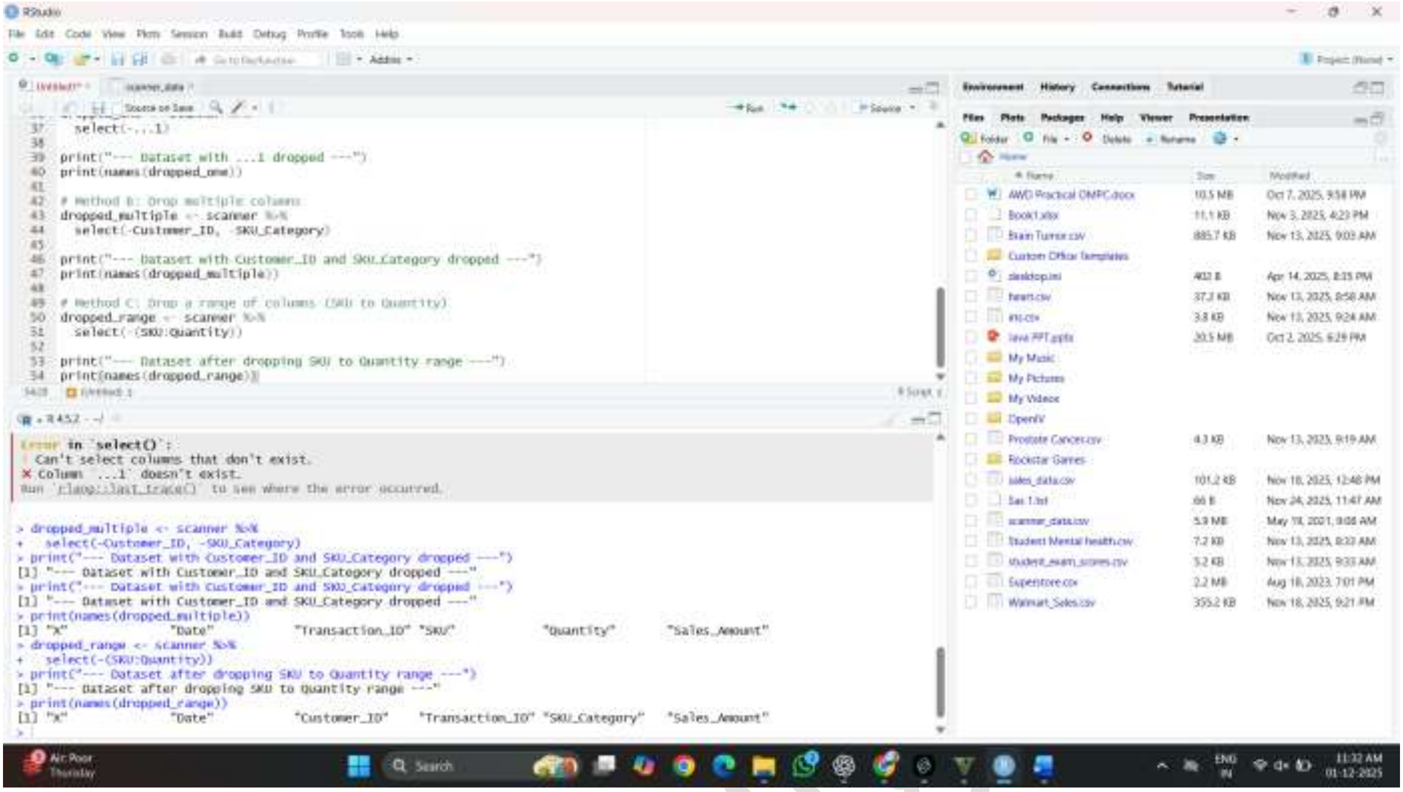
| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:50 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:10 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:06 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |



```
#7.  Selecting and dropping variables using select() in R.
# ------------------------------------------------------------
library(dplyr)

# 1. IMPORT DATASET
scanner <- read.csv("scanner_data.csv")

print("--- Original Dataset (First 3 rows) ---")
print(head(scanner, 3))

# 2. SELECTING VARIABLES (Keeping Columns)

# Method A: Select specific columns
selected_cols <- scanner %>%
  select(Date, Transaction_ID, Sales_Amount)

print("--- Selected Specific Columns ---")
print(head(selected_cols, 3))
```

```
> print("--- Selected Range of Columns ---")
[1] "--- Selected Range of Columns ---"
> print(head(range_cols, 3))
        Date Customer_ID Transaction_ID SKU_Category    SKU
1 02/01/2016        2547              1          X52 0EM7L
2 02/01/2016         822              2          2ML 68BRQ
3 02/01/2016        3686              3          0H2 CZUZX
> starts_with_s <- scanner %>%
+   select(starts_with("S"))
> print("--- Selected columns starting with 'S' ---")
[1] "--- Selected columns starting with 'S' ---"
> print(head(starts_with_s, 3))
  SKU_Category    SKU Sales_Amount
1          X52 0EM7L         3.13
2          2ML 68BRQ         5.46
3          0H2 CZUZX         6.35
> dropped_one <- scanner %>%
+   select(-Unnamed..0)
Error in `select()`:
! Can't select columns that don't exist.
✖ Column `Unnamed..0` doesn't exist.
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:50 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:10 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
scanner_data
Source on Save
Run
Source
R 4.5.2 · ~/
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Project: (None)
Search

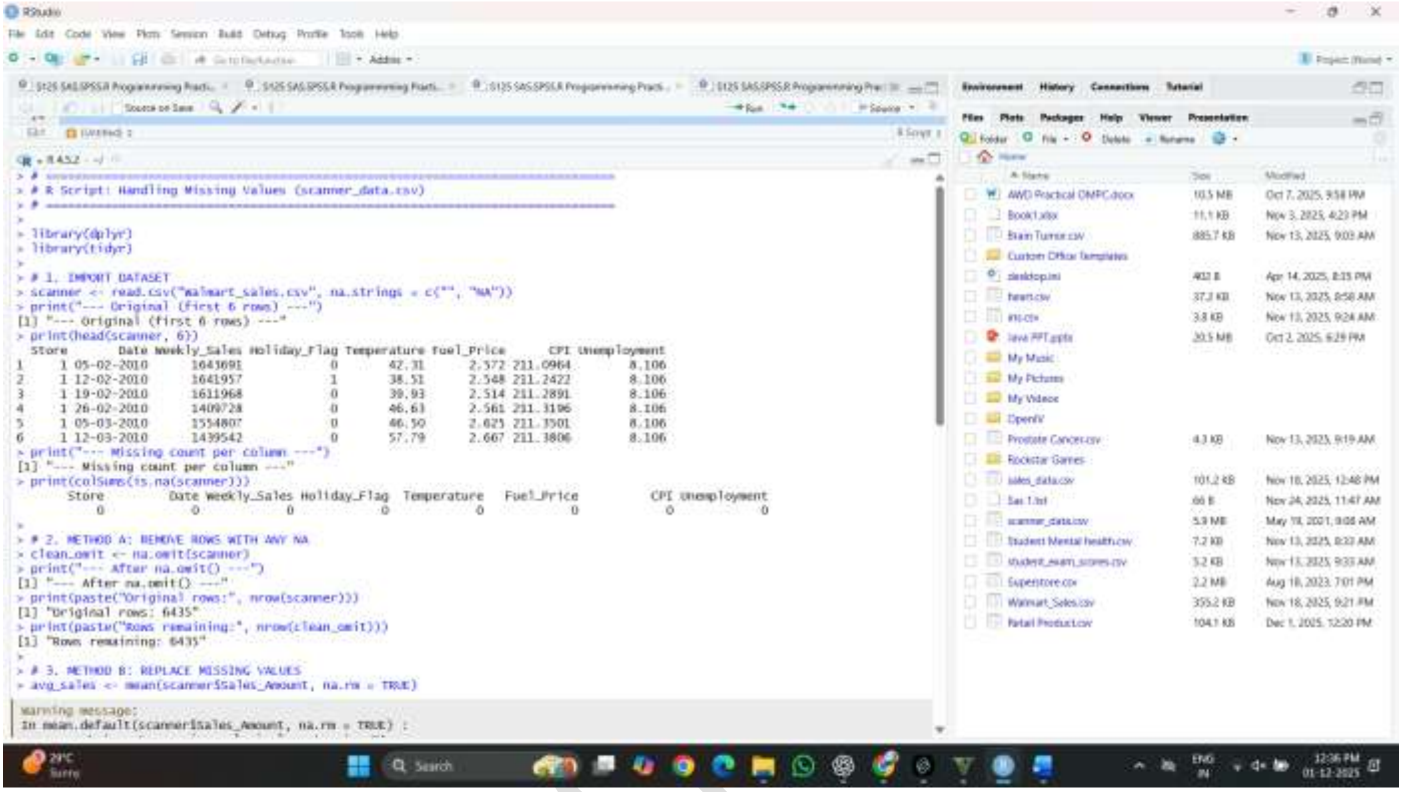
```
  select(-...1)

print("--- Dataset with ...1 dropped ---")
print(names(dropped_one))

# Method B: Drop multiple columns
dropped_multiple <- scanner %>%
  select(-Customer_ID, -SKU_Category)

print("--- Dataset with Customer_ID and SKU_Category dropped ---")
print(names(dropped_multiple))

# Method C: Drop a range of columns (SKU to Quantity)
dropped_range <- scanner %>%
  select(-(SKU:Quantity))

print("--- Dataset after dropping SKU to Quantity range ---")
print(names(dropped_range))
```

```
Error in `select()`:
! Can't select columns that don't exist.
✖ Column `...1` doesn't exist.
Run `rlang::last_trace()` to see where the error occurred.

> dropped_multiple <- scanner %>%
+   select(-Customer_ID, -SKU_Category)
> print("--- Dataset with Customer_ID and SKU_Category dropped ---")
[1] "--- Dataset with Customer_ID and SKU_Category dropped ---"
> print("--- Dataset with Customer_ID and SKU_Category dropped ---")
[1] "--- Dataset with Customer_ID and SKU_Category dropped ---"
> print(names(dropped_multiple))
[1] "X"              "Date"           "Transaction_ID" "SKU"            "Quantity"       "Sales_Amount"
> dropped_range <- scanner %>%
+   select(-(SKU:Quantity))
> print("--- Dataset after dropping SKU to Quantity range ---")
[1] "--- Dataset after dropping SKU to Quantity range ---"
> print(names(dropped_range))
[1] "X"              "Date"           "Customer_ID"    "Transaction_ID" "SKU_Category"   "Sales_Amount"
>
```

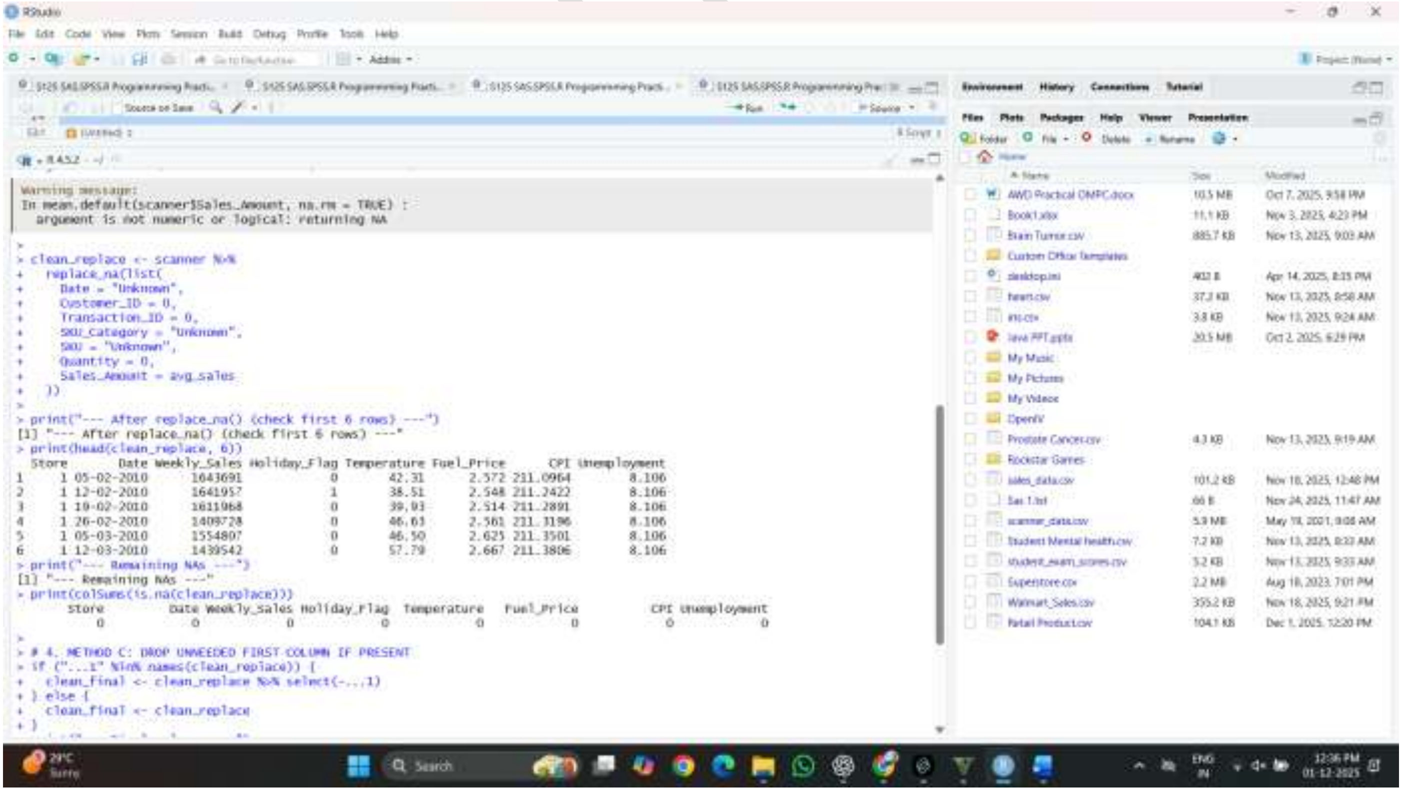
| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |

Aim: 8. Applying basic data cleaning functions: handling missing values using na.omit()/replace_na() in R. import dataset.

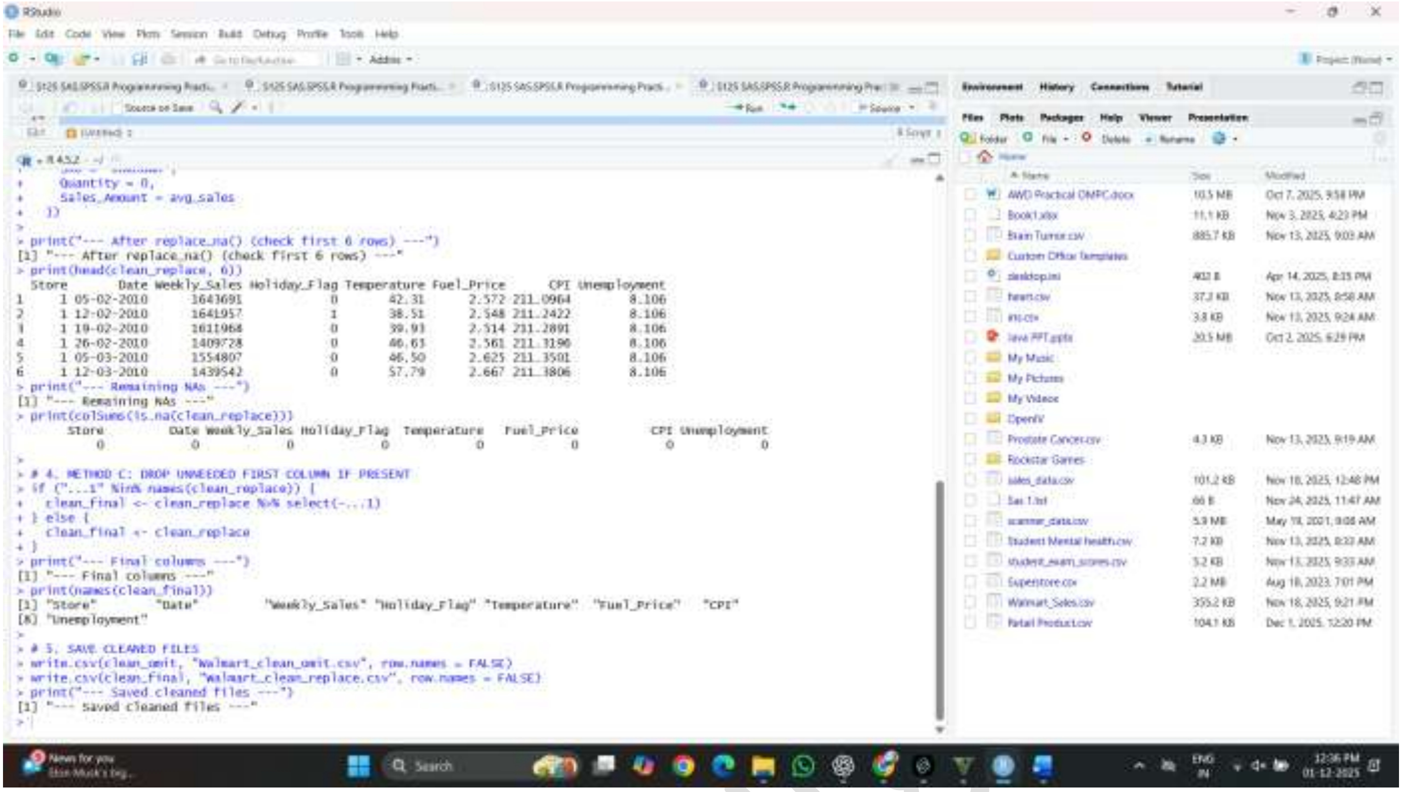
Output:



```
> # ==========================================================
> # R Script: Handling Missing Values (scanner_data.csv)
> # ==========================================================
>
> library(dplyr)
> library(tidyr)
>
> # 1. IMPORT DATASET
> scanner <- read.csv("walmart_sales.csv", na.strings = c("", "NA"))
> print("--- Original (first 6 rows) ---")
[1] "--- Original (first 6 rows) ---"
> print(head(scanner, 6))
  Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI Unemployment
1     1 05-02-2010      1643691            0       42.31      2.572 211.0964        8.106
2     1 12-02-2010      1641957            1       38.51      2.548 211.2422        8.106
3     1 19-02-2010      1611968            0       39.93      2.514 211.2891        8.106
4     1 26-02-2010      1409728            0       46.63      2.561 211.3196        8.106
5     1 05-03-2010      1554807            0       46.50      2.625 211.3501        8.106
6     1 12-03-2010      1439542            0       57.79      2.667 211.3806        8.106
> print("--- Missing count per column ---")
[1] "--- Missing count per column ---"
> print(colSums(is.na(scanner)))
       Store         Date Weekly_Sales Holiday_Flag  Temperature   Fuel_Price          CPI Unemployment
           0            0            0            0            0            0            0            0
>
> # 2. METHOD A: REMOVE ROWS WITH ANY NA
> clean_omit <- na.omit(scanner)
> print("--- After na.omit() ---")
[1] "--- After na.omit() ---"
> print(paste("Original rows:", nrow(scanner)))
[1] "Original rows: 6435"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 6435"
>
> # 3. METHOD B: REPLACE MISSING VALUES
> avg_sales <- mean(scanner$Sales_Amount, na.rm = TRUE)
Warning message:
In mean.default(scanner$Sales_Amount, na.rm = TRUE) :
```



```
Warning message:
In mean.default(scanner$Sales_Amount, na.rm = TRUE) :
  argument is not numeric or logical: returning NA
>
> clean_replace <- scanner %>%
+   replace_na(list(
+     Date = "Unknown",
+     Customer_ID = 0,
+     Transaction_ID = 0,
+     SKU_Category = "Unknown",
+     SKU = "Unknown",
+     Quantity = 0,
+     Sales_Amount = avg_sales
+   ))
>
> print("--- After replace_na() (check first 6 rows) ---")
[1] "--- After replace_na() (check first 6 rows) ---"
> print(head(clean_replace, 6))
  Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI Unemployment
1     1 05-02-2010      1643691            0       42.31      2.572 211.0964        8.106
2     1 12-02-2010      1641957            1       38.51      2.548 211.2422        8.106
3     1 19-02-2010      1611968            0       39.93      2.514 211.2891        8.106
4     1 26-02-2010      1409728            0       46.63      2.561 211.3196        8.106
5     1 05-03-2010      1554807            0       46.50      2.625 211.3501        8.106
6     1 12-03-2010      1439542            0       57.79      2.667 211.3806        8.106
> print("--- Remaining NAs ---")
[1] "--- Remaining NAs ---"
> print(colSums(is.na(clean_replace)))
       Store         Date Weekly_Sales Holiday_Flag  Temperature   Fuel_Price          CPI Unemployment
           0            0            0            0            0            0            0            0
>
> # 4. METHOD C: DROP UNNEEDED FIRST COLUMN IF PRESENT
> if ("...1" %in% names(clean_replace)) {
+   clean_final <- clean_replace %>% select(-...1)
+ } else {
+   clean_final <- clean_replace
+ }
```



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
S125 SAS.SPSS.R Programming Practi...
Source on Save
Run
Source
Untitled1
R 4.5.2 · ~/
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Search

```
+     Quantity = 0,
+     Sales_Amount = avg_sales
+   ))
>
> print("--- After replace_na() (check first 6 rows) ---")
[1] "--- After replace_na() (check first 6 rows) ---"
> print(head(clean_replace, 6))
  Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      CPI Unemployment
1     1 05-02-2010      1643691            0       42.31      2.572 211.0964        8.106
2     1 12-02-2010      1641957            1       38.51      2.548 211.2422        8.106
3     1 19-02-2010      1611968            0       39.93      2.514 211.2891        8.106
4     1 26-02-2010      1409728            0       46.63      2.561 211.3196        8.106
5     1 05-03-2010      1554807            0       46.50      2.625 211.3501        8.106
6     1 12-03-2010      1439542            0       57.79      2.667 211.3806        8.106
> print("--- Remaining NAs ---")
[1] "--- Remaining NAs ---"
> print(colSums(is.na(clean_replace)))
       Store         Date Weekly_Sales Holiday_Flag  Temperature   Fuel_Price          CPI Unemployment 
           0            0            0            0            0            0            0            0 
>
> # 4. METHOD C: DROP UNNEEDED FIRST COLUMN IF PRESENT
> if ("...1" %in% names(clean_replace)) {
+   clean_final <- clean_replace %>% select(-...1)
+ } else {
+   clean_final <- clean_replace
+ }
> print("--- Final columns ---")
[1] "--- Final columns ---"
> print(names(clean_final))
[1] "Store"        "Date"         "Weekly_Sales" "Holiday_Flag" "Temperature"  "Fuel_Price"   "CPI"         
[8] "Unemployment"
>
> # 5. SAVE CLEANED FILES
> write.csv(clean_omit, "Walmart_clean_omit.csv", row.names = FALSE)
> write.csv(clean_final, "Walmart_clean_replace.csv", row.names = FALSE)
> print("--- Saved cleaned files ---")
[1] "--- Saved cleaned files ---"
> 
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical DMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |
| Retail Product.csv | 104.1 KB | Dec 1, 2025, 12:20 PM |

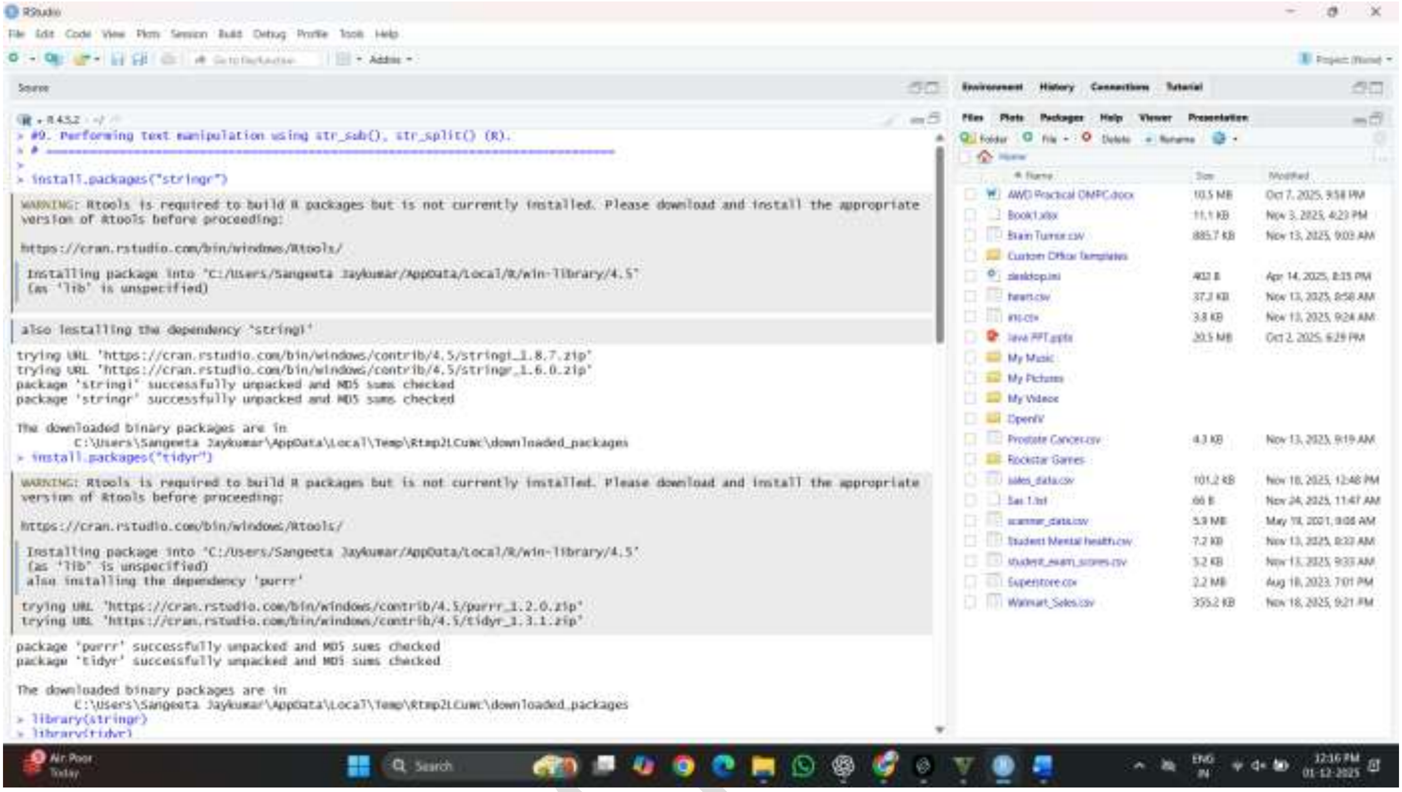
Sheth L.U.J & Sir M.V College

SAS/SSPS/R Programming

Practical No. 6 to 10

Aim: 9. Performing text manipulation using  str_sub(), str_split() (R). import dataset.

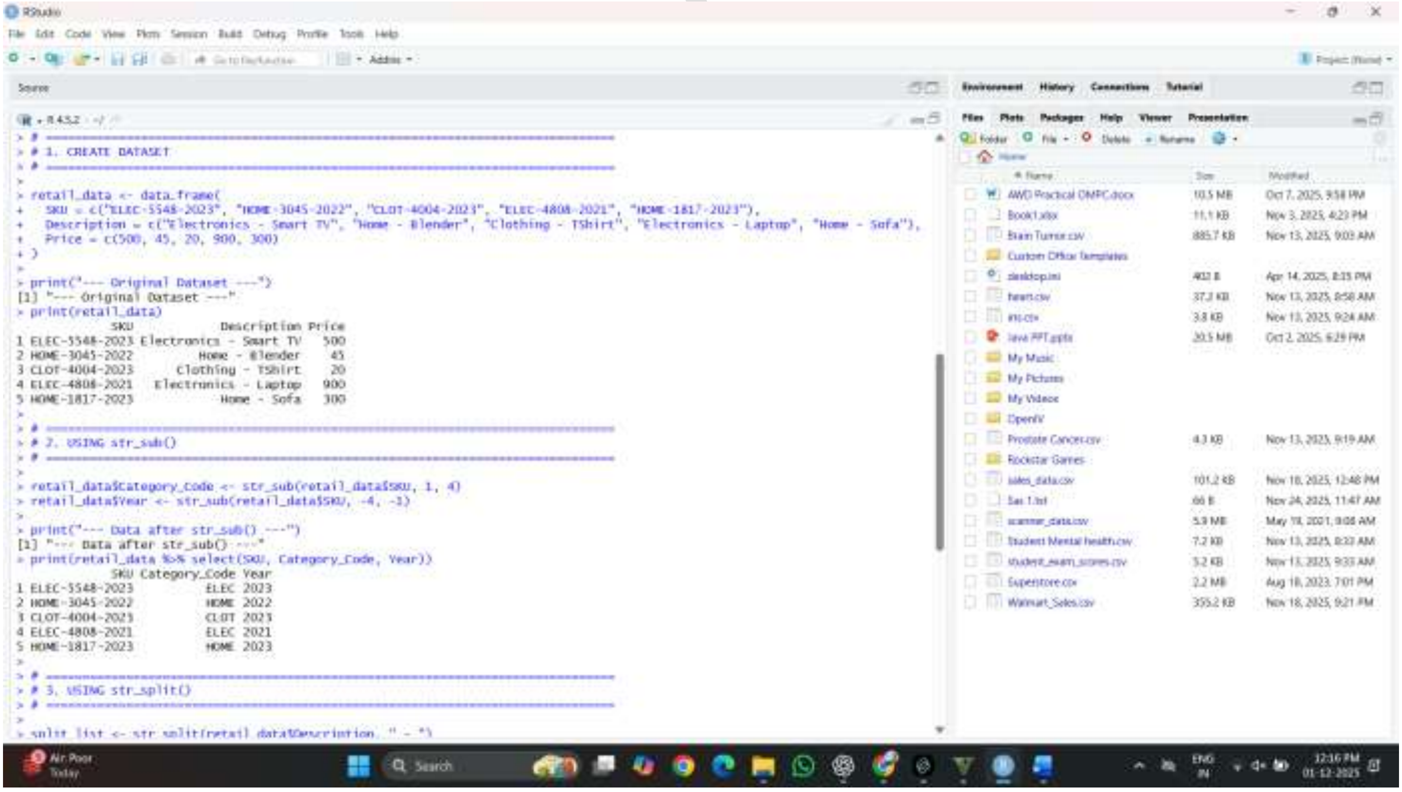
Output:



```
> #9. Performing text manipulation using str_sub(), str_split() (R).
> # ------------------------------------------------------------------
>
> install.packages("stringr")
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/Sangeeta Jaykumar/AppData/Local/R/win-library/4.5'
 (as 'lib' is unspecified)

also installing the dependency 'stringi'

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/stringr_1.6.0.zip'
package 'stringi' successfully unpacked and MD5 sums checked
package 'stringr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Sangeeta Jaykumar\AppData\Local\Temp\Rtmp2LCuwc\downloaded_packages
> 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/Sangeeta Jaykumar/AppData/Local/R/win-library/4.5'
 (as 'lib' is unspecified)
 also installing the dependency 'purrr'

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/tidyr_1.3.1.zip'
package 'purrr' successfully unpacked and MD5 sums checked
package 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Sangeeta Jaykumar\AppData\Local\Temp\Rtmp2LCuwc\downloaded_packages
> library(stringr)
> library(tidyr)
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Source
R 4.5.2
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Air Poor
Today
Search
ENG
IN
12:16 PM
01-12-2025



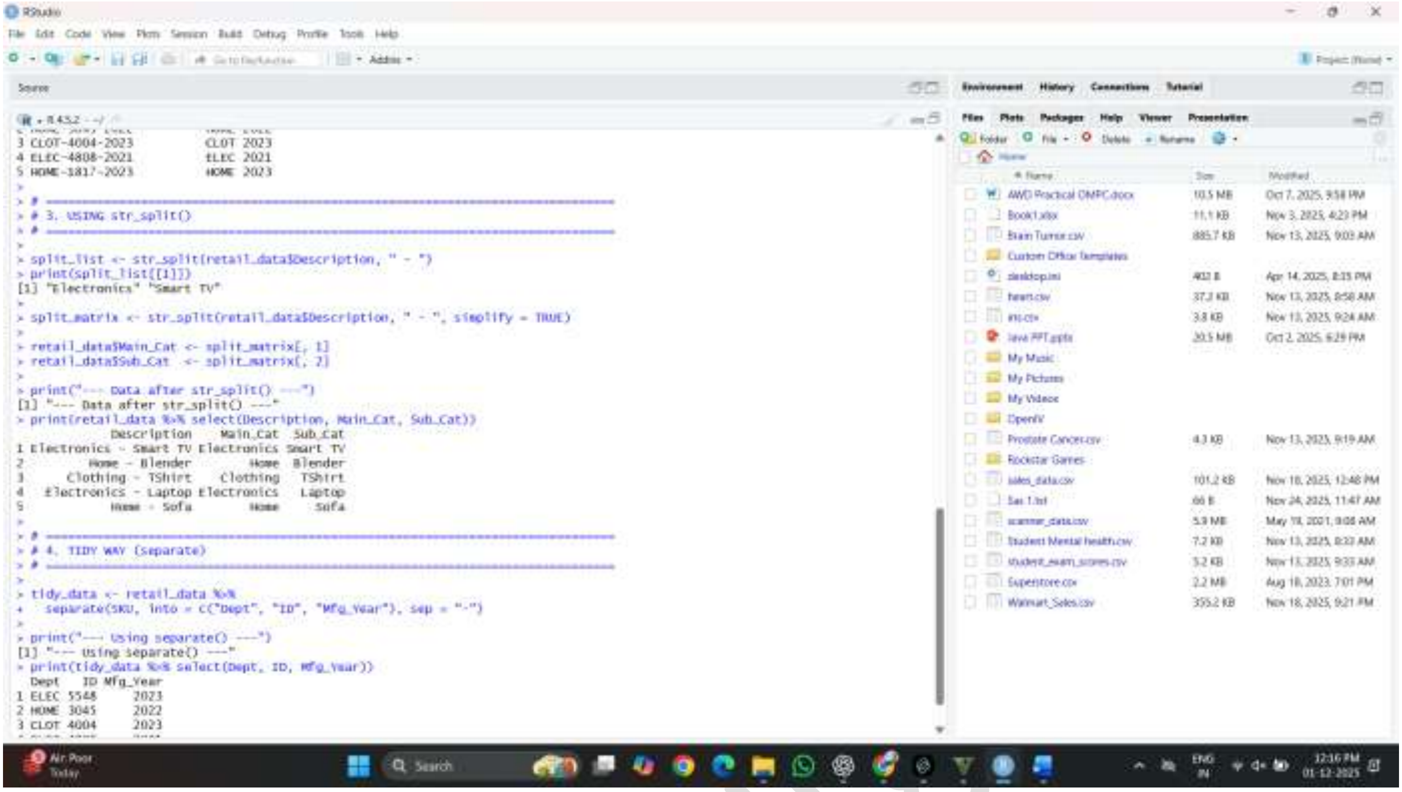
```
> # ------------------------------------------------------------------
> # 1. CREATE DATASET
> # ------------------------------------------------------------------
>
> retail_data <- data.frame(
+   SKU = c("ELEC-5548-2023", "HOME-3045-2022", "CLOT-4004-2023", "ELEC-4808-2021", "HOME-1817-2023"),
+   Description = c("Electronics - Smart TV", "Home - Blender", "Clothing - TShirt", "Electronics - Laptop", "Home - Sofa"),
+   Price = c(500, 45, 20, 900, 300)
+ )
>
> print("--- Original Dataset ---")
[1] "--- Original Dataset ---"
> print(retail_data)
             SKU            Description Price
1 ELEC-5548-2023 Electronics - Smart TV   500
2 HOME-3045-2022         Home - Blender    45
3 CLOT-4004-2023      Clothing - TShirt    20
4 ELEC-4808-2021   Electronics - Laptop   900
5 HOME-1817-2023            Home - Sofa   300
>
> # ------------------------------------------------------------------
> # 2. USING str_sub()
> # ------------------------------------------------------------------
>
> retail_data$Category_Code <- str_sub(retail_data$SKU, 1, 4)
> retail_data$Year <- str_sub(retail_data$SKU, -4, -1)
>
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(retail_data %>% select(SKU, Category_Code, Year))
             SKU Category_Code Year
1 ELEC-5548-2023          ELEC 2023
2 HOME-3045-2022          HOME 2022
3 CLOT-4004-2023          CLOT 2023
4 ELEC-4808-2021          ELEC 2021
5 HOME-1817-2023          HOME 2023
>
> # ------------------------------------------------------------------
> # 3. USING str_split()
> # ------------------------------------------------------------------
>
> split_list <- str_split(retail_data$Description, " - ")
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Source
R 4.5.2
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Air Poor
Today
Search
ENG
IN
12:16 PM
01-12-2025

Jeetesh Shirganoor

S125



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.5.2 · ~/
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Search
Air Poor
Today
ENG
IN
12:16 PM
01-12-2025

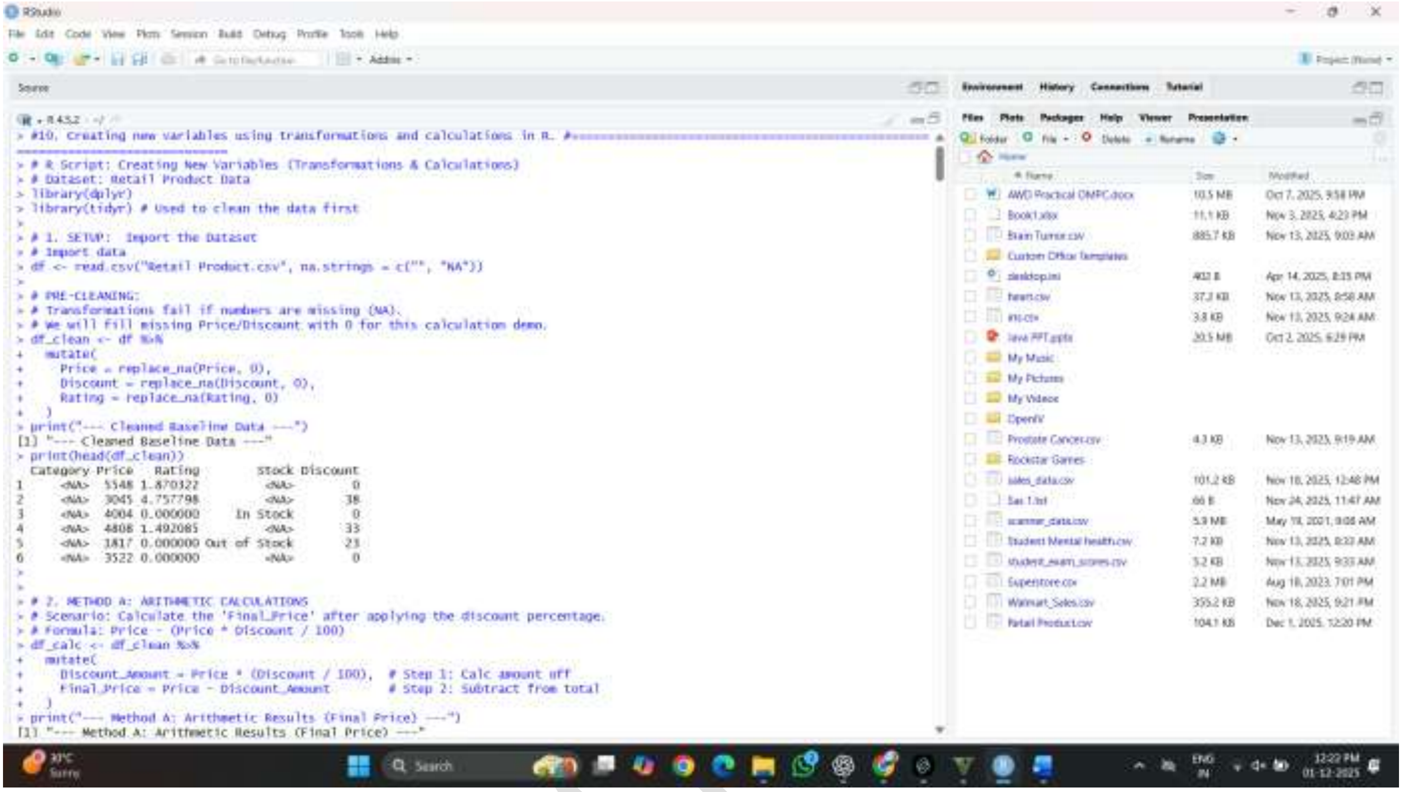
```
3 CLOT-4004-2023          CLOT 2023
4 ELEC-4808-2021          ELEC 2021
5 HOME-1817-2023          HOME 2023
> 
> # ================================================================
> # 3. USING str_split()
> # ================================================================
> 
> split_list <- str_split(retail_data$Description, " - ")
> print(split_list[[1]])
[1] "Electronics" "Smart TV"
> 
> split_matrix <- str_split(retail_data$Description, " - ", simplify = TRUE)
> 
> retail_data$Main_Cat <- split_matrix[, 1]
> retail_data$Sub_Cat  <- split_matrix[, 2]
> 
> print("--- Data after str_split() ---")
[1] "--- Data after str_split() ---"
> print(retail_data %>% select(Description, Main_Cat, Sub_Cat))
            Description    Main_Cat  Sub_Cat
1 Electronics - Smart TV Electronics Smart TV
2        Home - Blender        Home  Blender
3      Clothing - TShirt    Clothing   TShirt
4   Electronics - Laptop Electronics   Laptop
5           Home - Sofa        Home     Sofa
> 
> # ================================================================
> # 4. TIDY WAY (separate)
> # ================================================================
> 
> tidy_data <- retail_data %>%
+   separate(SKU, into = c("Dept", "ID", "Mfg_Year"), sep = "-")
> 
> print("--- Using separate() ---")
[1] "--- Using separate() ---"
> print(tidy_data %>% select(Dept, ID, Mfg_Year))
  Dept   ID Mfg_Year
1 ELEC 5548     2023
2 HOME 3045     2022
3 CLOT 4004     2023
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical DMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |

# Sheth L.U.J & Sir M.V College
# SAS/SSPS/R Programming
# Practical No. 6 to 10

Aim: 10. Creating new variables using transformations and calculations in R. import dataset.

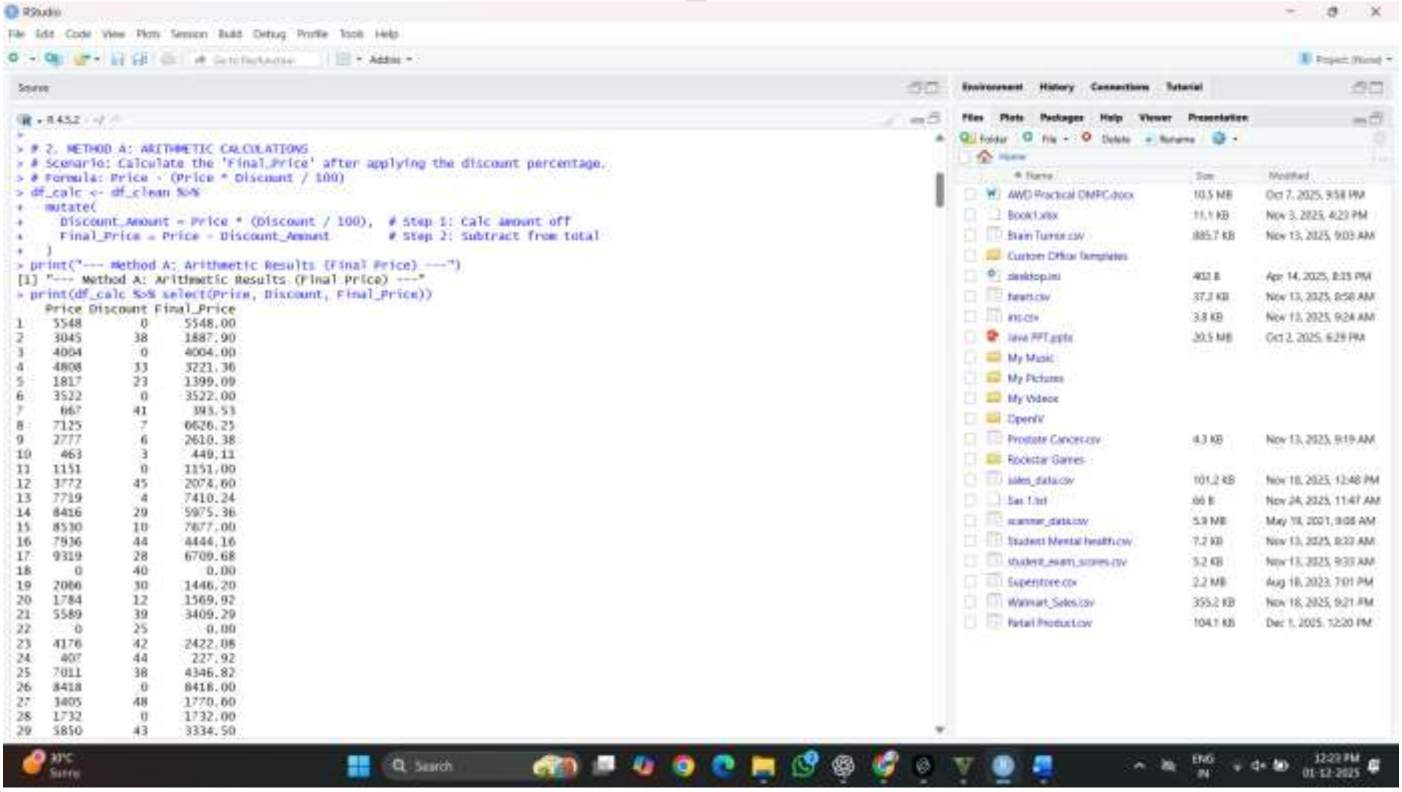
Output:



```
> #10. Creating new variables using transformations and calculations in R. #==========================
==============================
> # R Script: Creating New Variables (Transformations & Calculations)
> # Dataset: Retail Product Data
> library(dplyr)
> library(tidyr) # Used to clean the data first
>
> # 1. SETUP:  Import the Dataset
> # Import data
> df <- read.csv("Retail Product.csv", na.strings = c("", "NA"))
>
> # PRE-CLEANING:
> # Transformations fail if numbers are missing (NA).
> # We will fill missing Price/Discount with 0 for this calculation demo.
> df_clean <- df %>%
+   mutate(
+     Price = replace_na(Price, 0),
+     Discount = replace_na(Discount, 0),
+     Rating = replace_na(Rating, 0)
+   )
> print("--- Cleaned Baseline Data ---")
[1] "--- Cleaned Baseline Data ---"
> print(head(df_clean))
  Category Price   Rating        Stock Discount
1     <NA>  5548 1.870322         <NA>        0
2     <NA>  3045 4.757798         <NA>       38
3     <NA>  4004 0.000000     In Stock        0
4     <NA>  4808 1.492085         <NA>       33
5     <NA>  1817 0.000000 Out of Stock       23
6     <NA>  3522 0.000000         <NA>        0
>
>
> # 2. METHOD A: ARITHMETIC CALCULATIONS
> # Scenario: Calculate the 'Final_Price' after applying the discount percentage.
> # Formula: Price - (Price * Discount / 100)
> df_calc <- df_clean %>%
+   mutate(
+     Discount_Amount = Price * (Discount / 100),  # Step 1: Calc amount off
+     Final_Price = Price - Discount_Amount         # Step 2: Subtract from total
+   )
> print("--- Method A: Arithmetic Results (Final Price) ---")
[1] "--- Method A: Arithmetic Results (Final Price) ---"
```

| Name | Size | Modified |
|---|---|---|
| AWD Practical OMPC.docx | 10.5 MB | Oct 7, 2025, 9:58 PM |
| Book1.xlsx | 11.1 KB | Nov 3, 2025, 4:23 PM |
| Brain Tumor.csv | 885.7 KB | Nov 13, 2025, 9:03 AM |
| Custom Office Templates | | |
| desktop.ini | 402 B | Apr 14, 2025, 8:35 PM |
| heart.csv | 37.2 KB | Nov 13, 2025, 8:58 AM |
| iris.csv | 3.8 KB | Nov 13, 2025, 9:24 AM |
| Java PPT.pptx | 20.5 MB | Oct 2, 2025, 6:29 PM |
| My Music | | |
| My Pictures | | |
| My Videos | | |
| OpenIV | | |
| Prostate Cancer.csv | 4.3 KB | Nov 13, 2025, 9:19 AM |
| Rockstar Games | | |
| sales_data.csv | 101.2 KB | Nov 18, 2025, 12:48 PM |
| Sas 1.txt | 66 B | Nov 24, 2025, 11:47 AM |
| scanner_data.csv | 5.9 MB | May 19, 2021, 9:08 AM |
| Student Mental health.csv | 7.2 KB | Nov 13, 2025, 8:33 AM |
| student_exam_scores.csv | 5.2 KB | Nov 13, 2025, 9:33 AM |
| Superstore.csv | 2.2 MB | Aug 18, 2023, 7:01 PM |
| Walmart_Sales.csv | 355.2 KB | Nov 18, 2025, 9:21 PM |
| Retail Product.csv | 104.1 KB | Dec 1, 2025, 12:20 PM |

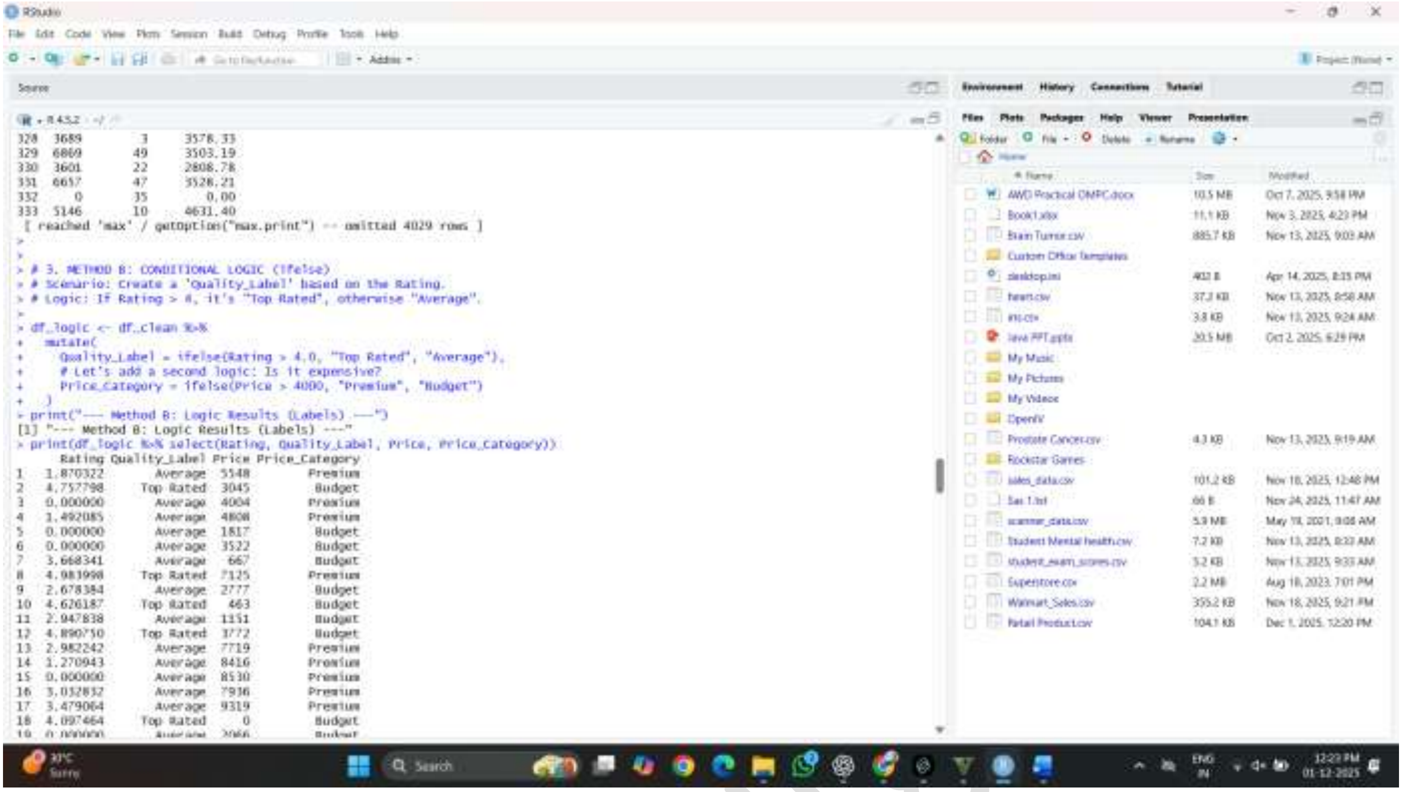
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Source
R 4.5.2
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Project: (None)
30°C Sunny
Search
ENG IN
12:22 PM 01-12-2025



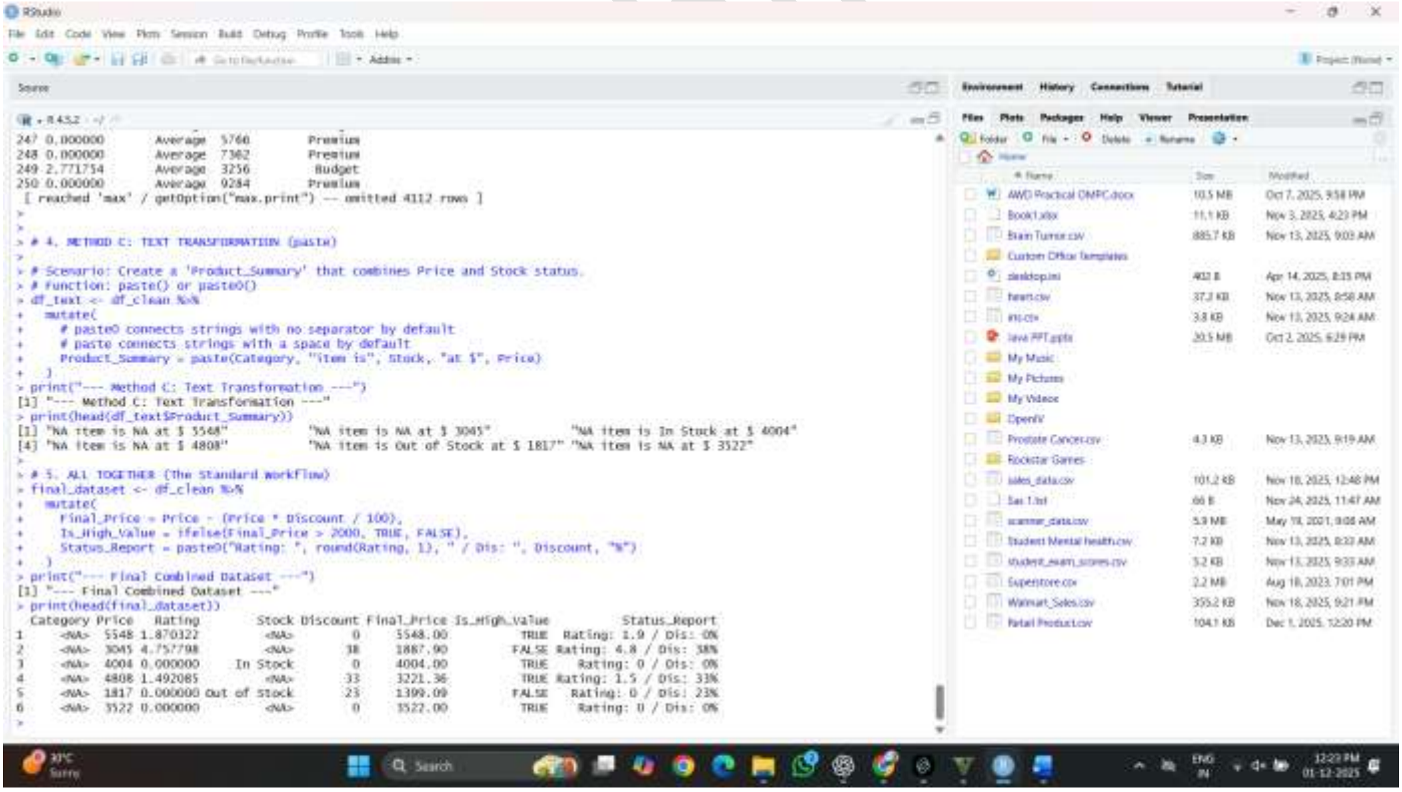
```
>
> # 2. METHOD A: ARITHMETIC CALCULATIONS
> # Scenario: Calculate the 'Final_Price' after applying the discount percentage.
> # Formula: Price - (Price * Discount / 100)
> df_calc <- df_clean %>%
+   mutate(
+     Discount_Amount = Price * (Discount / 100),  # Step 1: Calc amount off
+     Final_Price = Price - Discount_Amount         # Step 2: Subtract from total
+   )
> print("--- Method A: Arithmetic Results (Final Price) ---")
[1] "--- Method A: Arithmetic Results (Final Price) ---"
> print(df_calc %>% select(Price, Discount, Final_Price))
   Price Discount Final_Price
1   5548        0     5548.00
2   3045       38     1887.90
3   4004        0     4004.00
4   4808       33     3221.36
5   1817       23     1399.09
6   3522        0     3522.00
7    667       41      393.53
8   7125        7     6626.25
9   2777        6     2610.38
10   463        3      449.11
11  1151        0     1151.00
12  3772       45     2074.60
13  7719        4     7410.24
14  8416       29     5975.36
15  8530       10     7677.00
16  7936       44     4444.16
17  9319       28     6709.68
18     0       40        0.00
19  2066       30     1446.20
20  1784       12     1569.92
21  5589       39     3409.29
22     0       25        0.00
23  4176       42     2422.08
24   407       44      227.92
25  7011       38     4346.82
26  8418        0     8418.00
27  3405       48     1770.60
28  1732        0     1732.00
29  5850       43     3334.50
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
Source
R 4.5.2
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
Home
Project: (None)
30°C Sunny
Search
ENG IN
12:23 PM 01-12-2025

Jeetesh Shirganoor
S125



```
328  3689          3      3578.33
329  6869         49      3503.19
330  3601         22      2808.78
331  6657         47      3528.21
332     0         35         0.00
333  5146         10      4631.40
 [ reached 'max' / getOption("max.print") -- omitted 4029 rows ]
>
>
> # 3. METHOD B: CONDITIONAL LOGIC (ifelse)
> # Scenario: Create a 'Quality_Label' based on the Rating.
> # Logic: If Rating > 4, it's "Top Rated", otherwise "Average".
>
> df_logic <- df_clean %>%
+   mutate(
+     Quality_Label = ifelse(Rating > 4.0, "Top Rated", "Average"),
+     # Let's add a second logic: Is it expensive?
+     Price_Category = ifelse(Price > 4000, "Premium", "Budget")
+   )
> print("--- Method B: Logic Results (Labels) ---")
[1] "--- Method B: Logic Results (Labels) ---"
> print(df_logic %>% select(Rating, Quality_Label, Price, Price_Category))
      Rating Quality_Label Price Price_Category
1   1.870322       Average  5548        Premium
2   4.757798     Top Rated  3045         Budget
3   0.000000       Average  4004        Premium
4   1.492085       Average  4808        Premium
5   0.000000       Average  1817         Budget
6   0.000000       Average  3522         Budget
7   3.668341       Average   667         Budget
8   4.983998     Top Rated  7125        Premium
9   2.678384       Average  2777         Budget
10  4.626187     Top Rated   463         Budget
11  2.947838       Average  1151         Budget
12  4.890750     Top Rated  3772         Budget
13  2.982242       Average  7719        Premium
14  1.270943       Average  8416        Premium
15  0.000000       Average  8530        Premium
16  3.032832       Average  7936        Premium
17  3.479064       Average  9319        Premium
18  4.097464     Top Rated     0         Budget
19  0.000000       Average  2066         Budget
```



```
247 0.000000       Average  5766        Premium
248 0.000000       Average  7362        Premium
249 2.771754       Average  3256         Budget
250 0.000000       Average  9284        Premium
 [ reached 'max' / getOption("max.print") -- omitted 4112 rows ]
>
>
> # 4. METHOD C: TEXT TRANSFORMATION (paste)
>
> # Scenario: Create a 'Product_Summary' that combines Price and Stock status.
> # Function: paste() or paste0()
> df_text <- df_clean %>%
+   mutate(
+     # paste0 connects strings with no separator by default
+     # paste connects strings with a space by default
+     Product_Summary = paste(Category, "item is", Stock, "at $", Price)
+   )
> print("--- Method C: Text Transformation ---")
[1] "--- Method C: Text Transformation ---"
> print(head(df_text$Product_Summary))
[1] "NA item is NA at $ 5548"           "NA item is NA at $ 3045"           "NA item is In Stock at $ 4004"
[4] "NA item is NA at $ 4808"           "NA item is Out of Stock at $ 1817" "NA item is NA at $ 3522"
>
> # 5. ALL TOGETHER (The Standard Workflow)
> final_dataset <- df_clean %>%
+   mutate(
+     Final_Price = Price - (Price * Discount / 100),
+     Is_High_Value = ifelse(Final_Price > 2000, TRUE, FALSE),
+     Status_Report = paste0("Rating: ", round(Rating, 1), " / Dis: ", Discount, "%")
+   )
> print("--- Final Combined Dataset ---")
[1] "--- Final Combined Dataset ---"
> print(head(final_dataset))
  Category Price   Rating        Stock Discount Final_Price Is_High_Value         Status_Report
1     <NA>  5548 1.870322         <NA>        0     5548.00          TRUE  Rating: 1.9 / Dis: 0%
2     <NA>  3045 4.757798         <NA>       38     1887.90         FALSE Rating: 4.8 / Dis: 38%
3     <NA>  4004 0.000000     In Stock        0     4004.00          TRUE    Rating: 0 / Dis: 0%
4     <NA>  4808 1.492085         <NA>       33     3221.36          TRUE Rating: 1.5 / Dis: 33%
5     <NA>  1817 0.000000 Out of Stock       23     1399.09         FALSE   Rating: 0 / Dis: 23%
6     <NA>  3522 0.000000         <NA>        0     3522.00          TRUE    Rating: 0 / Dis: 0%
>
```