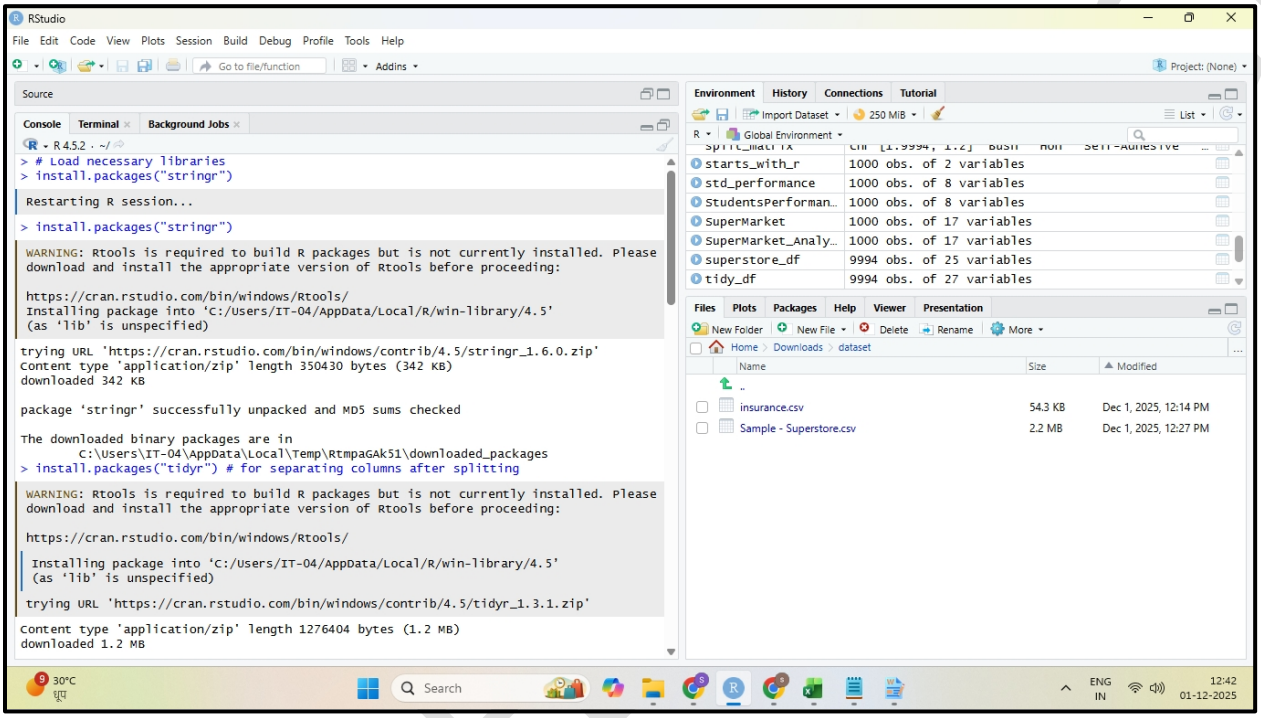
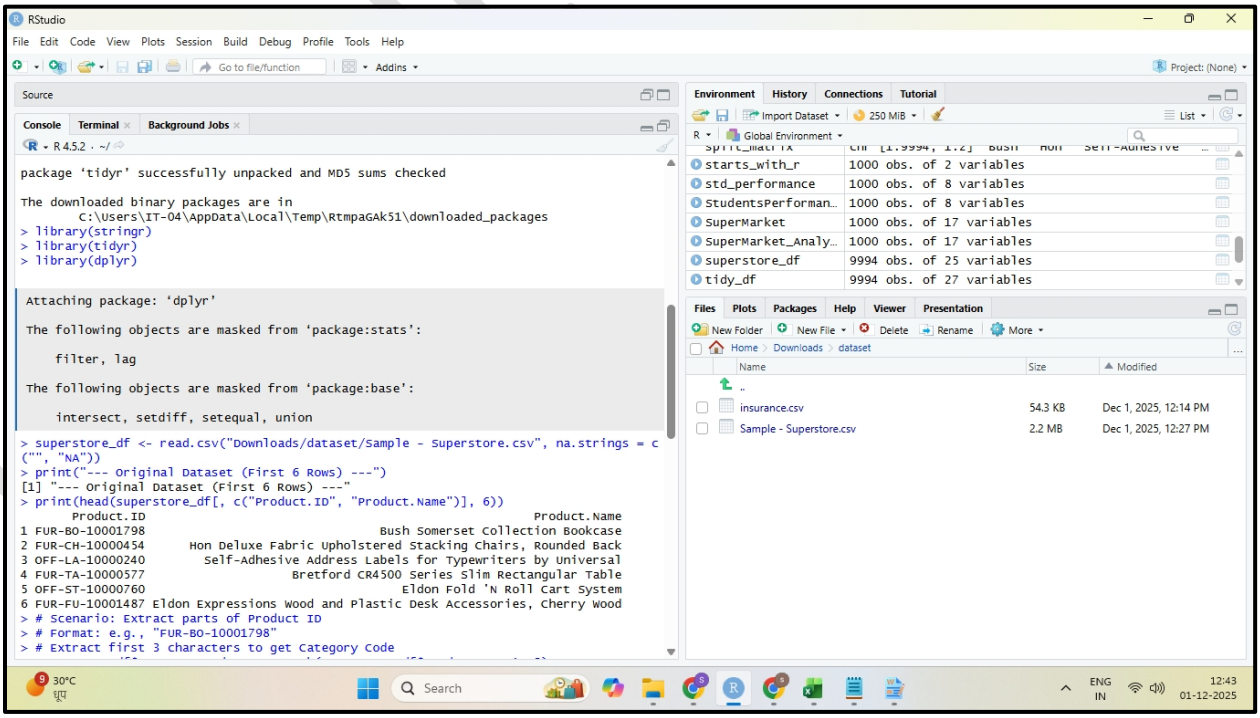
Sheth L.U.J. College of Arts & Sir M.V. College Of Science & Commerce
SUBJECT NAME: Data Analysis with SAS / SPSS/R

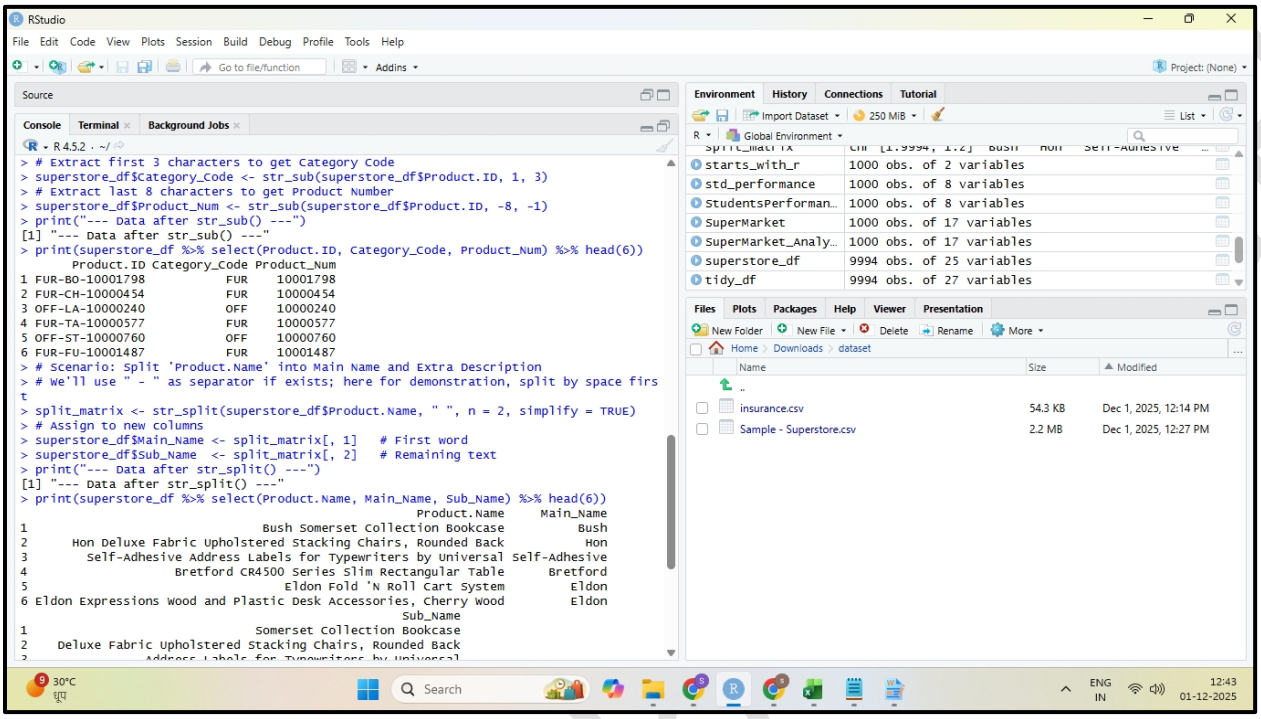
PRACTICAL NO : 9

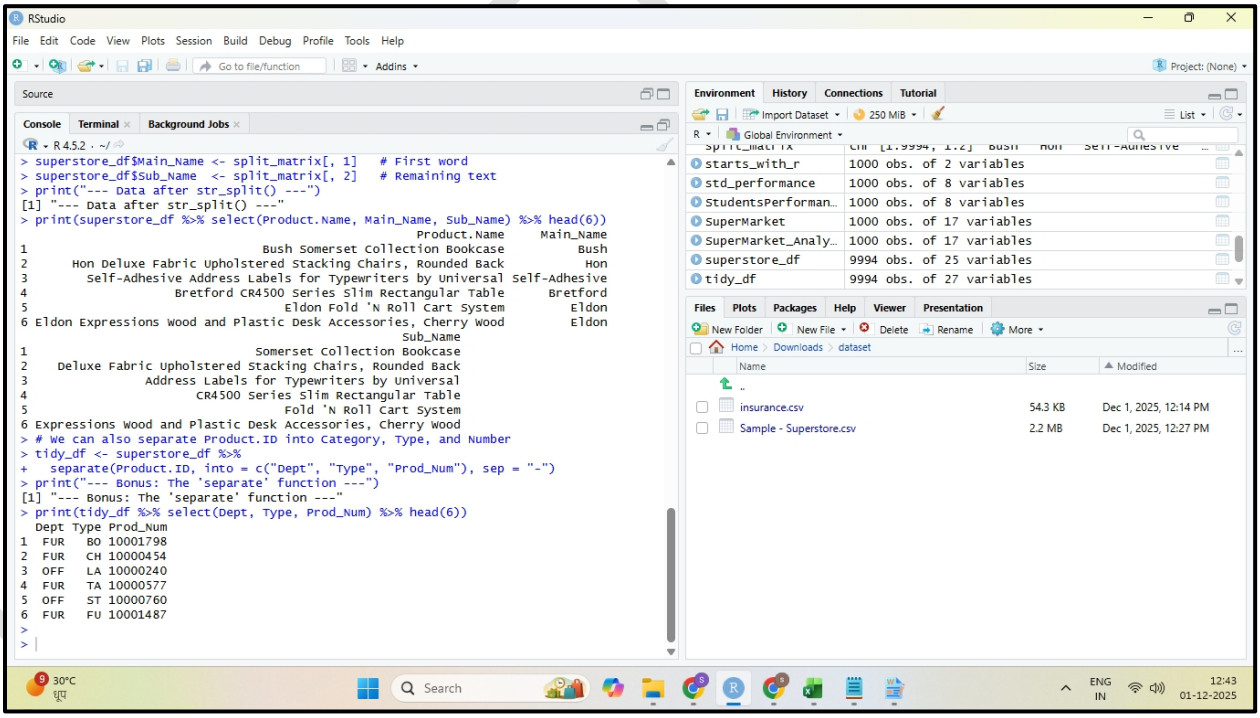
AIM : Performing text manipulation using str_sub(), str_split() (R). import dataset.



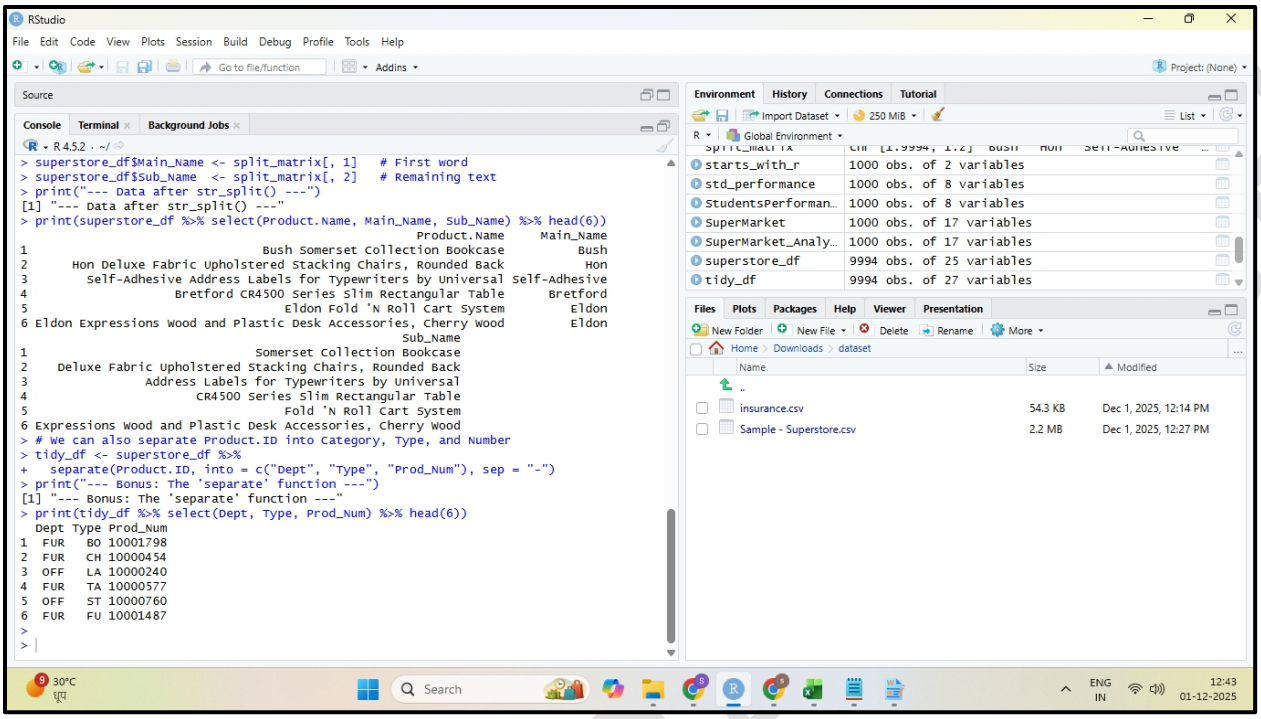


DHEERAJ SINGH
S118


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.5.2 · ~/
> # Extract first 3 characters to get Category Code
> superstore_df$Category_Code <- str_sub(superstore_df$Product.ID, 1, 3)
> # Extract last 8 characters to get Product Number
> superstore_df$Product_Num <- str_sub(superstore_df$Product.ID, -8, -1)
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(superstore_df %>% select(Product.ID, Category_Code, Product_Num) %>% head(6))
Product.ID Category_Code Product_Num
1 FUR-BO-10001798 FUR 10001798
2 FUR-CH-10000454 FUR 10000454
3 OFF-LA-10000240 OFF 10000240
4 FUR-TA-10000577 FUR 10000577
5 OFF-ST-10000760 OFF 10000760
6 FUR-FU-10001487 FUR 10001487
> # Scenario: Split 'Product.Name' into Main Name and Extra Description
> # We'll use " - " as separator if exists; here for demonstration, split by space firs
t
> split_matrix <- str_split(superstore_df$Product.Name, " ", n = 2, simplify = TRUE)
> # Assign to new columns
> superstore_df$Main_Name <- split_matrix[, 1] # First word
> superstore_df$Sub_Name <- split_matrix[, 2] # Remaining text
> print("--- Data after str_split() ---")
[1] "--- Data after str_split() ---"
> print(superstore_df %>% select(Product.Name, Main_Name, Sub_Name) %>% head(6))
Product.Name Main_Name
1 Bush Somerset Collection Bookcase Bush
2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back Hon
3 Self-Adhesive Address Labels for Typewriters by Universal Self-Adhesive
4 Bretford CR4500 Series Slim Rectangular Table Bretford
5 Eldon Fold 'N Roll Cart System Eldon
6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood Eldon
Sub_Name
1 Somerset Collection Bookcase
2 Deluxe Fabric Upholstered Stacking Chairs, Rounded Back
Environment History Connections Tutorial
Import Dataset 250 MiB
Global Environment
starts_with_r 1000 obs. of 2 variables
std_performance 1000 obs. of 8 variables
StudentsPerforman... 1000 obs. of 8 variables
SuperMarket 1000 obs. of 17 variables
SuperMarket_Analy... 1000 obs. of 17 variables
superstore_df 9994 obs. of 25 variables
tidy_df 9994 obs. of 27 variables
Files Plots Packages Help Viewer Presentation
New Folder New File Delete Rename More
Home Downloads dataset
Name Size Modified
insurance.csv 54.3 KB Dec 1, 2025, 12:14 PM
Sample - Superstore.csv 2.2 MB Dec 1, 2025, 12:27 PM
30°C
Search
ENG IN
12:43 01-12-2025


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
R 4.5.2 · ~/
> superstore_df$Main_Name <- split_matrix[, 1] # First word
> superstore_df$Sub_Name <- split_matrix[, 2] # Remaining text
> print("--- Data after str_split() ---")
[1] "--- Data after str_split() ---"
> print(superstore_df %>% select(Product.Name, Main_Name, Sub_Name) %>% head(6))
Product.Name Main_Name
1 Bush Somerset Collection Bookcase Bush
2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back Hon
3 Self-Adhesive Address Labels for Typewriters by Universal Self-Adhesive
4 Bretford CR4500 Series Slim Rectangular Table Bretford
5 Eldon Fold 'N Roll Cart System Eldon
6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood Eldon
Sub_Name
1 Somerset Collection Bookcase
2 Deluxe Fabric Upholstered Stacking Chairs, Rounded Back
3 Address Labels for Typewriters by Universal
4 CR4500 Series Slim Rectangular Table
5 Fold 'N Roll Cart System
6 Expressions Wood and Plastic Desk Accessories, Cherry Wood
> # We can also separate Product.ID into Category, Type, and Number
> tidy_df <- superstore_df %>%
+ separate(Product.ID, into = c("Dept", "Type", "Prod_Num"), sep = "-")
> print("--- Bonus: The 'separate' function ---")
[1] "--- Bonus: The 'separate' function ---"
> print(tidy_df %>% select(Dept, Type, Prod_Num) %>% head(6))
Dept Type Prod_Num
1 FUR BO 10001798
2 FUR CH 10000454
3 OFF LA 10000240
4 FUR TA 10000577
5 OFF ST 10000760
6 FUR FU 10001487
Environment History Connections Tutorial
Import Dataset 250 MiB
Global Environment
starts_with_r 1000 obs. of 2 variables
std_performance 1000 obs. of 8 variables
StudentsPerforman... 1000 obs. of 8 variables
SuperMarket 1000 obs. of 17 variables
SuperMarket_Analy... 1000 obs. of 17 variables
superstore_df 9994 obs. of 25 variables
tidy_df 9994 obs. of 27 variables
Files Plots Packages Help Viewer Presentation
New Folder New File Delete Rename More
Home Downloads dataset
Name Size Modified
insurance.csv 54.3 KB Dec 1, 2025, 12:14 PM
Sample - Superstore.csv 2.2 MB Dec 1, 2025, 12:27 PM
30°C
Search
ENG IN
12:43 01-12-2025

Sheth L.U.J. College of Arts & Sir M.V. College Of Science & Commerce
SUBJECT NAME: Data Analysis with SAS / SPSS/R



S118 DHEERAJ SINGH

DHEERAJ SINGH
S118