

PRACTICAL NO : 9

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

RStudio interface showing the installation of 'stringr' and 'tidyr' packages. The console output includes:

```
R - R4.5.2 - ~/
> # Load necessary libraries
> install.packages("stringr")

Restarting R session...

> 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/IT-04/AppData/Local/R/win-library/4.5'
(as 'lib' is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/stringr_1.6.0.zip'
Content type 'application/zip' length 350430 bytes (342 KB)
downloaded 342 KB

package 'stringr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/IT-04/AppData/Local/Temp/RtmpagAk51/downloaded_packages
> install.packages("tidyr") # for separating columns after splitting

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/IT-04/AppData/Local/R/win-library/4.5'
(as 'lib' is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/tidyr_1.3.1.zip'
Content type 'application/zip' length 1276404 bytes (1.2 MB)
downloaded 1.2 MB
```

The Environment pane on the right shows the following datasets:

Dataset Name	Observations	Variables
starts_with_r	1000	2
std_performance	1000	8
StudentsPerforman..	1000	8
SuperMarket	1000	17
SuperMarket_Analy..	1000	17
superstore_df	9994	25
tidy_df	9994	27

RStudio interface showing the loading of 'stringr', 'tidyr', and 'dplyr' libraries. The console output includes:

```
package 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:/Users/IT-04/AppData/Local/Temp/RtmpagAk51/downloaded_packages
> library(stringr)
> library(tidyr)
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':
  filter, lag

The following objects are masked from 'package:base':
  intersect, setdiff, setequal, union

> superstore_df <- read.csv("Downloads/dataset/Sample - Superstore.csv", na.strings = c
(" ", "NA"))
> print("--- Original Dataset (First 6 Rows) ---")
[1] "--- Original Dataset (First 6 Rows) ---"
> print(head(superstore_df[, c("Product.ID", "Product.Name")], 6))
  Product.ID Product.Name
1 FUR-BO-10001798 Bush Somerset collection Bookcase
2 FUR-CH-10000454 Hon Deluxe Fabric upholstered Stacking Chairs, Rounded Back
3 OFF-LA-10000240 Self-Adhesive Address Labels for Typewriters by Universal
4 FUR-TA-10000577 Bretford CR4500 Series Slim Rectangular Table
5 OFF-ST-10000760 Eldon Fold 'N Roll Cart System
6 FUR-FU-10001487 Eldon Expressions wood and Plastic Desk Accessories, Cherry wood

> # Scenario: Extract parts of Product ID
> # Format: e.g., "FUR-BO-10001798"
> # Extract first 3 characters to get Category Code
```

The Environment pane on the right shows the following datasets:

Dataset Name	Observations	Variables
starts_with_r	1000	2
std_performance	1000	8
StudentsPerforman..	1000	8
SuperMarket	1000	17
SuperMarket_Analy..	1000	17
superstore_df	9994	25
tidy_df	9994	27

```

R - R4.5.2 ~ /
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
> # 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 first
> 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 Sub_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

```

Environment: 250 MiB

Object	Class	Dimensions
split_matrix	chr	[1, 9994, 1, 2]
starts_with_r	tbl_df	1000 obs. of 2 variables
std_performance	tbl_df	1000 obs. of 8 variables
StudentsPerforman...	tbl_df	1000 obs. of 8 variables
SuperMarket	tbl_df	1000 obs. of 17 variables
SuperMarket_Analy...	tbl_df	1000 obs. of 17 variables
superstore_df	tbl_df	9994 obs. of 25 variables
tidy_df	tbl_df	9994 obs. of 27 variables

Files: 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

```

R - R4.5.2 ~ /
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Background Jobs
> 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 Sub_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: 250 MiB

Object	Class	Dimensions
split_matrix	chr	[1, 9994, 1, 2]
starts_with_r	tbl_df	1000 obs. of 2 variables
std_performance	tbl_df	1000 obs. of 8 variables
StudentsPerforman...	tbl_df	1000 obs. of 8 variables
SuperMarket	tbl_df	1000 obs. of 17 variables
SuperMarket_Analy...	tbl_df	1000 obs. of 17 variables
superstore_df	tbl_df	9994 obs. of 25 variables
tidy_df	tbl_df	9994 obs. of 27 variables

Files: 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

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for splitting a dataset into main and sub names, and then tidying the resulting data frame.
- Console:** Shows the output of the R code, including the structure of the data frames and the results of the 'tidy' function.
- Environment:** Lists the objects in the global environment, including 'split_matrix', 'starts_with_r', 'std_performance', 'StudentsPerforman...', 'SuperMarket', 'SuperMarket_Analy...', 'superstore_df', and 'tidy_df'.
- Files:** Shows the file explorer with 'insurance.csv' (54.3 KB) and 'Sample - Superstore.csv' (2.2 MB) in the 'Downloads' folder.

```

R - R4.5.2 - ~/
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R4.5.2 - ~/
> 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 Sub_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
  
```