

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

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

The screenshot shows the RStudio interface with the following details:

- File:** Practical9.R
- Script:** Practical9.R
- Content:** The script starts with comments about text manipulation on the Titanic dataset using str_sub() and str_split(). It loads necessary libraries (stringr, tidyverse, dplyr), imports the titanic dataset from a CSV file, and prints the first 6 rows. Then, it extracts the last name from the Name column using str_sub(). Finally, it splits the name into parts using str_split().
- Environment:** Shows variables like clean_183, clean_891, name_s, rest_s, S081_P, titani_891, avg_age 29.6991176470588, and avn_fa 32.2042079685746.
- R Resources:** Includes links to Learning R Online, CRAN Task Views, R on StackOverflow, Getting Help with R, and RStudio documentation.
- Manuals:** Includes links to Posit Support, Posit Community Forum for the RStudio IDE, Posit Cheat Sheets, RStudio Packages, Posit Products, and Manuals.

The screenshot shows the RStudio interface with the following details:

- File:** Practical9.R
- Script:** Practical9.R
- Content:** The script continues with str_split() to split the name into parts, assign new columns (LastName_2 for last name and RestName for title + first name), and print the first 6 rows. It then adds a bonus section to split the RestName into Title and First Names. This is followed by a section on splitting titles, another section on separating ticket numbers, and finally a note about ticket prefixes.
- Environment:** Shows the same variables as the previous screenshot.
- R Resources:** Includes links to Learning R Online, CRAN Task Views, R on StackOverflow, Getting Help with R, and RStudio documentation.
- Manuals:** Includes links to Posit Support, Posit Community Forum for the RStudio IDE, Posit Cheat Sheets, RStudio Packages, Posit Products, and Manuals.

Name :- Priya Gupta
Roll no. :- S081

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

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source

Console Terminal Background Jobs

R - R 4.5.2 - ~ /

```
> # R Script: Text Manipulation on TITANIC DATASET
> # Functions: str_sub(), str_split(), separate()
> #

> # Load necessary libraries
library(stringr)
library(tidyverse)
library(dplyr)
>
> # =====
> # 1. IMPORT TITANIC DATASET
> #

> titanic_df <- read.csv(
+ "C:/Users/Priya Gupta/Downloads/S081 Priya Rakesh Gupta - titanic.csv",
+ na.strings = c("", "NA")
+ )
>
> cat("n--- Original Titanic Dataset (First 6 Rows) ---\n")

--- Original Titanic Dataset (First 6 Rows) ---
> print(head(titanic_df))
  PassengerId Survived PClass          Name       Sex Age SibSp Parch Ticket  Fare Cabin Embarked
1           1         0     3    Braund, Mr. Owen Harris   male  22      1     0   A/5 21173 7.250000   S
2           2         1     1  Cumings, Mrs. John Bradley (Florence Brigid Thayer) female  38      1     0   PC 17599 70.233333   C85
3           3         1     3            Heikkinen, Miss. Laina  female  26      0     0  STON/O2. 3101282 7.925000   S
4           4         1     1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35      1     0   133803 53.100000  C123
5           5         0     3            Allen, Mr. William Henry   male  35      0     0   373450 8.050000   S
6           6         0     3            Moran, Mr. James   male  NA      0     0   330877 8.458333 <NA>   Q
>
> # =====
> # 2. USING str_sub() - Extract Last Name from Name Column
> #

> # Extract the position of the comma in each Name
> comma_positions <- str_locate(titanic_df$name, ",")[,1]
>
> # Extract last name (everything before the comma)
> titanic_df$lastName <- str_sub(titanic_df$name, 1, comma_positions - 1)
>
```

Environment History Connections

R Global Environment

Data

clean_ 183 obs. of 12 v...
clean_ 891 obs. of 12 v...
name_ chr [1:891, 1:2] ...
rest_ chr [1:891, 1:1] ...
s081_ 891 obs. of 12 v...
titanic_ 891 obs. of 18 v...

Values

avg_age 29.6991176470588
avn_fa 32.204207965746

Files Plots Packages Help View

Home Find in Topic

R Resources

Learning R Online
CRAN Task Views
R on Stack Overflow
Getting Help with R

RStudio

Posit Support
Posit Community Forum
for the RStudio IDE
Posit Cheat Sheets
RStudio Packages
Posit Products

Manuals

ENG IN 23:16 01-12-2025

Name :- Priya Gupta
Roll no. :- S081

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

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R 4.5.2 - ~/
> # 2. USING str_sub() - Extract Last Name from Name Column
> # =====
>
> # Extract the position of the comma in each Name
> comma_positions <- str_locate(titanic_df\$name, ",")[,1]
>
> # Extract last name (everything before the comma)
> titanic_df\$lastName <- str_sub(titanic_df\$name, 1, comma_positions - 1)
>
> cat("\n--- Extracted Last Names (First 6 Rows) ---\n")
--- Extracted Last Names (First 6 Rows) ---
> print(head(titanic_df\$lastName))
[1] "Braund" "Cumings" "Heikkinen" "Futrelle" "Allen" "Moran"
>
> # 3. USING str_split() - Split Name into Parts
> # =====
>
> # Split based on comma ("LastName, Title FirstName")
> name_split <- str_split(titanic_df\$name, ", ", simplify = TRUE)
>
> # Assign new columns
> titanic_df\$lastName_2 <- name_split[,1] # Last name
> titanic_df\$restName <- name_split[,2] # Title + First name
>
> cat("\n--- Data after str_split() (First 6 Rows) ---\n")
--- Data after str_split() (First 6 Rows) ---
> print(head(titanic_df[, c("Name", "LastName_2", "RestName")]))
Name LastName_2 RestName
1 Braund, Mr. Owen Harris Braund Mr. Owen Harris
2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) Cumings Mrs. John Bradley (Florence Briggs Thayer)
3 Heikkinen, Miss. Laina Heikkinen Miss. Laina
4 Futrelle, Mrs. Jacques Heath (Lily May Peel) Futrelle Mrs. Jacques Heath (Lily May Peel)
5 Allen, Mr. William Henry Allen Mr. William Henry
6 Moran, Mr. James Moran Mr. James
>
> # =====

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R 4.5.2 - ~/
> # =====
> # 4. BONUS: Split the RestName into Title and First Names
> # =====
>
> # Split RestName based on space (e.g., "Mr. Owen Harris")
> rest_split <- str_split(titanic_df\$restName, " ", simplify = TRUE)
>
> # First word = Title (Mr., Mrs., Miss., etc.)
> titanic_df\$title <- rest_split[,1]
>
> cat("\n--- Title Extracted (First 6 Rows) ---\n")
--- Title Extracted (First 6 Rows) ---
> print(head(titanic_df\$title))
[1] "Mr." "Mrs." "Miss." "Mr." "Mr."
>
> # =====
> # 5. Using separate() - Split Ticket into Prefix & Number
> # =====
>
> # Some tickets have prefixes (e.g. "A/5 21171"), some don't
> titanic_df <- titanic_df %>%
+ separate(Ticket, into = c("Ticket_Prefix", "Ticket_Number"),
+ sep = " ", fill = "right")
Warning message:
Expected 2 pieces. Additional pieces discarded in 13 rows [116, 174, 244, 383, 401, 415, 434, 474, 580, 591, 637, 665, 745].
>
> cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
--- After Splitting Ticket (First 6 Rows) ---
> print(head(titanic_df[, c("Ticket_Prefix", "Ticket_Number")]))
Ticket_Prefix Ticket_Number
1 A/5 21171
2 PC 17599
3 STON/O2. 3101282
4 113803 <NA>
5 373450 <NA>
6 330877 <NA>
>

Name :- Priya Gupta
Roll no. :- S081

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

The screenshot shows the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Project dropdown set to (None). The left sidebar has tabs for Source, Console, Terminal, and Background Jobs, with the Source tab selected. The main console area contains R code for data manipulation:

```
R - R 4.5.2 - ~/R
> # 5. Using separate() - Split Ticket into Prefix & Number
> #
> # Some tickets have prefixes (e.g. "A/5 21171"), some don't
> titanic_df <- titanic_df %>%
+   separate(Ticket, into = c("Ticket_Prefix", "Ticket_Number"),
+           sep = " ", fill = "right")
Warning message:
Expected 2 pieces. Additional pieces discarded in 13 rows [116, 174, 244, 383, 401, 415, 434, 474, 580, 591, 637, 665, 745].
> cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
--- After Splitting Ticket (First 6 Rows) ---
> print(head(titanic_df, c("Ticket_Prefix", "Ticket_Number")))
#> #   Ticket_Prefix Ticket_Number
#> 1       A/5          21171
#> 2        PC          17599
#> 3    STON/O2.      3101282
#> 4       113803       <NA>
#> 5      373450       <NA>
#> 6      330877       <NA>
> #
> # 6. OPTIONAL: Extract Deck Letter from Cabin using str_sub()
> #
> titanic_df$Deck <- str_sub(titanic_df$Cabin, 1, 1)
> cat("\n--- Extracted Deck (First 6 Rows) ---\n")
--- Extracted Deck (First 6 Rows) ---
> print(head(titanic_df$Deck))
#> # [1] NA "C" NA "C" NA NA
> #
> # END OF SCRIPT
> #
> |
```

The right pane displays the Global Environment, showing objects like clean_183, clean_891, name_s_, rest_s_, S081_P_891, and titanic_891. It also includes sections for R Resources, Posit Support, and Manuals.

Name :- Priya Gupta
Roll no. :- S081