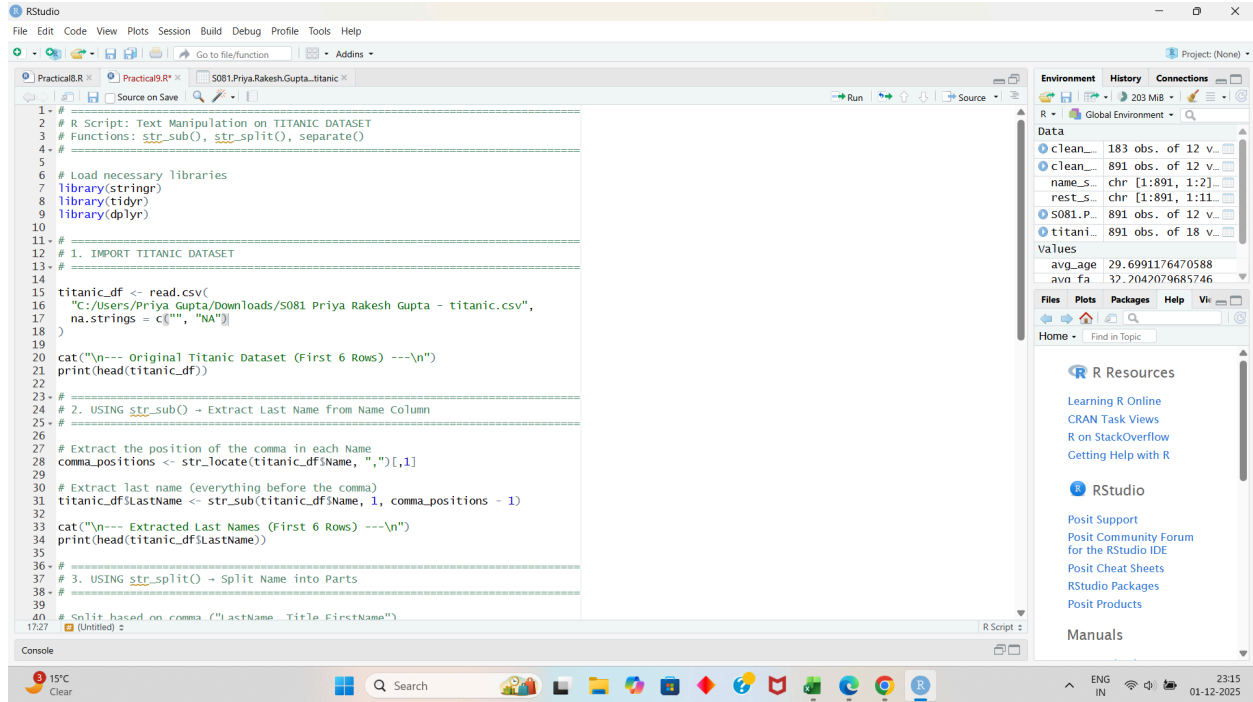


Sheth L.U.J. & Sir M.V. College

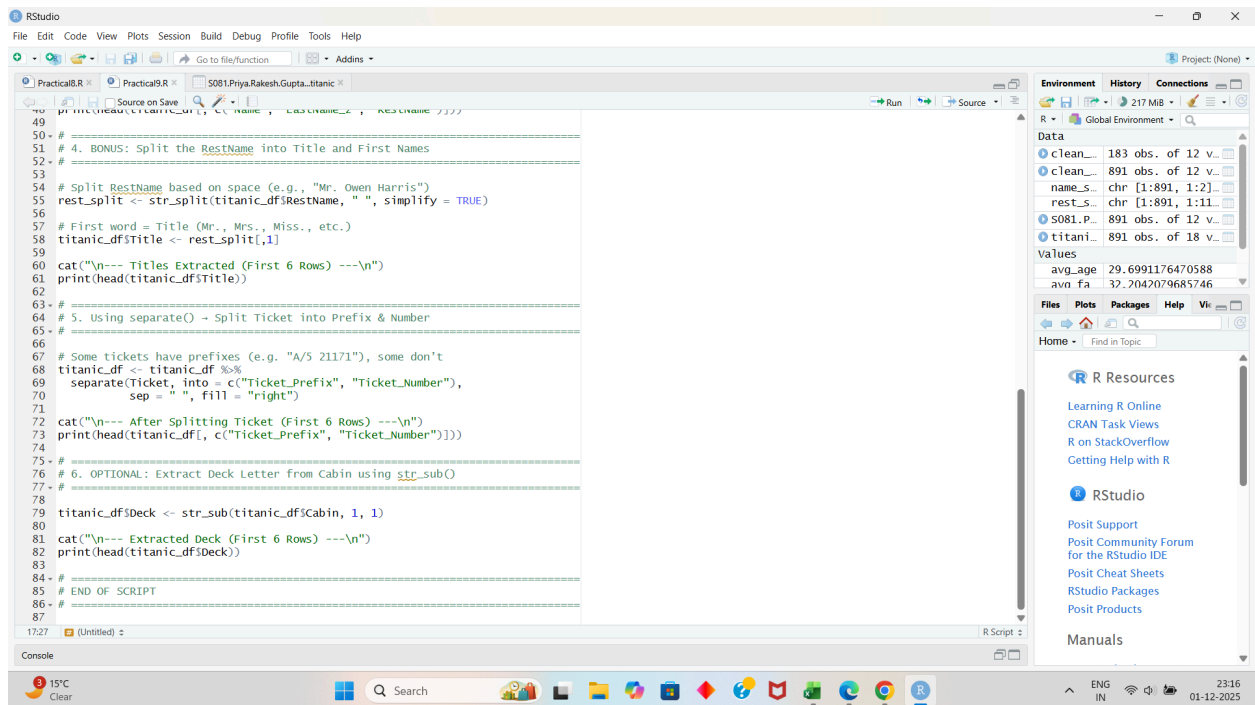
9. Performing text manipulation using `str_sub()`, `str_split()` (R). import dataset.



```
1 # R Script: Text Manipulation on TITANIC DATASET
2 # Functions: str_sub(), str_split(), separate()
3
4 # Load necessary libraries
5 library(stringr)
6 library(tidyverse)
7 library(dplyr)
8
9 # 1. IMPORT TITANIC DATASET
10
11 # Import the dataset
12 titanic_df <- read_csv(
13   "C:/Users/Priya Gupta/Downloads/S081 Priya Rakesh Gupta - titanic.csv",
14   na.strings = c("", "NA"))
15
16 cat("\n--- Original Titanic Dataset (First 6 Rows) ---\n")
17 print(head(titanic_df))
18
19 # 2. USING str_sub() -> Extract Last Name from Name Column
20
21 # Extract the position of the comma in each Name
22 comma_positions <- str_locate(titanic_df$Name, ",")[,1]
23
24 # Extract last name (everything before the comma)
25 titanic_df$lastName <- str_sub(titanic_df$Name, 1, comma_positions - 1)
26
27 cat("\n--- Extracted Last Names (First 6 Rows) ---\n")
28 print(head(titanic_df$lastName))
29
30 # 3. USING str_split() -> Split Name into Parts
31
32 # Split based on comma ("LastName, Title FirstName")
33 name_split <- str_split(titanic_df$Name, ",", simplify = TRUE)
34
35 # Assign new columns
36 titanic_df$lastName_2 <- name_split[,1] # Last name
37 titanic_df$restName <- name_split[,2] # Title + First name
38
39 cat("\n--- Data after str_split() (First 6 Rows) ---\n")
40 print(head(titanic_df[, c("Name", "lastName_2", "restName")]))
41
42 # 4. BONUS: Split the RestName into Title and First Names
43
44 # Split RestName based on space (e.g., "Mr. Owen Harris")
45 rest_split <- str_split(titanic_df$restName, " ", simplify = TRUE)
46
47 # First word = Title (Mr., Mrs., Miss., etc.)
48 titanic_df$title <- rest_split[,1]
49
50 cat("\n--- Titles Extracted (First 6 Rows) ---\n")
51 print(head(titanic_df$title))
52
53 # 5. Using separate() -> Split Ticket into Prefix & Number
54
55 # Some tickets have prefixes (e.g. "A/5 21171"), some don't
56 titanic_df <- titanic_df %>%
57   separate(ticket, into = c("Ticket_Prefix", "Ticket_Number"),
58     sep = " ", fill = "right")
59
60 cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
61 print(head(titanic_df[, c("Ticket_Prefix", "Ticket_Number")]))
```

Name :- Priya Gupta
Roll no. :- S081

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



The screenshot shows the RStudio interface with a script file named 'S081.Priya.Rakesh.Gupta_titanic.R'. The code performs the following steps:

- Prints the first 6 rows of the 'titanic' dataset.
- 4. BONUS: Splits the 'RestName' column into 'Title' and 'First Names' using `str_split` and `rest_split`.
- 5. Uses `separate` to split the 'Ticket' column into 'Prefix' and 'Number'.
- 6. OPTIONAL: Extracts the deck letter from the 'Cabin' column using `str_sub`.

The console shows the output of the first 6 rows of the dataset:

```
cat("\n--- Original Titanic Dataset (First 6 Rows) ---\n")
print(head(titanic_df))
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	<NA>	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	<NA>	C
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	<NA>	S
6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583	<NA>	Q

Name :- Priya Gupta
Roll no. :- S081

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

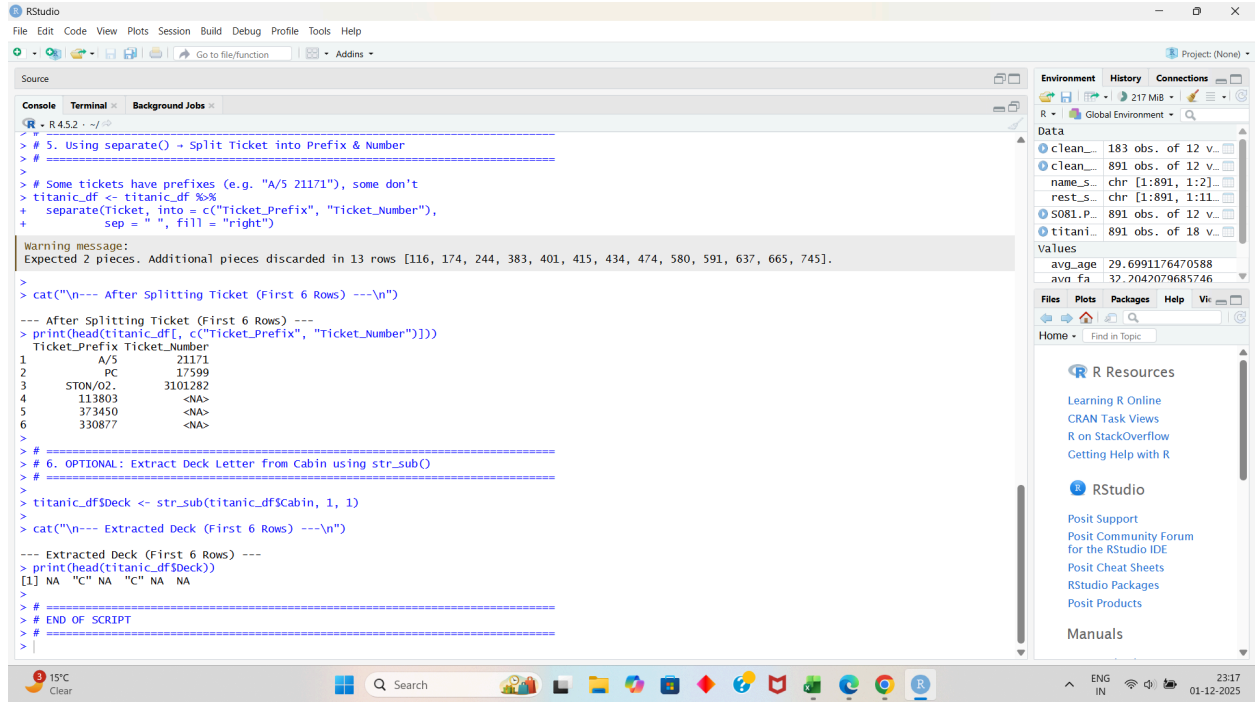
```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R 4.5.2 - ~/ -
> # =====
> # 2. USING str_sub() -> Extract Last Name from Name Column
> # =====
> # Extract the position of the comma in each Name
> comma_positions <- 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 - 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--- Titles Extracted (First 6 Rows) ---\n")
--- Titles Extracted (First 6 Rows) ---
> print(head(titanic_df$title))
[1] "Mr." "Mrs." "Miss." "Mr." "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/OZ. 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 displays the RStudio IDE interface. The main console window shows the following R code and its output:

```
> # 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 Environment pane on the right shows the following objects:

- `clean_`: 183 obs. of 12 v...
- `clean_`: 891 obs. of 12 v...
- `name_s`: chr [1:891, 1:2]...
- `rest_s`: chr [1:891, 1:11]...
- `S081.P`: 891 obs. of 12 v...
- `titani_`: 891 obs. of 18 v...

The Values pane shows the following summary statistics:

Variable	Value
avg_age	29.6991176470588
ava_fa	32.7042079685746

The RStudio interface also includes a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help), a toolbar, and a status bar at the bottom showing the system temperature (15°C) and time (23:17 on 01-12-2025).

Name :- Priya Gupta

Roll no. :- S081