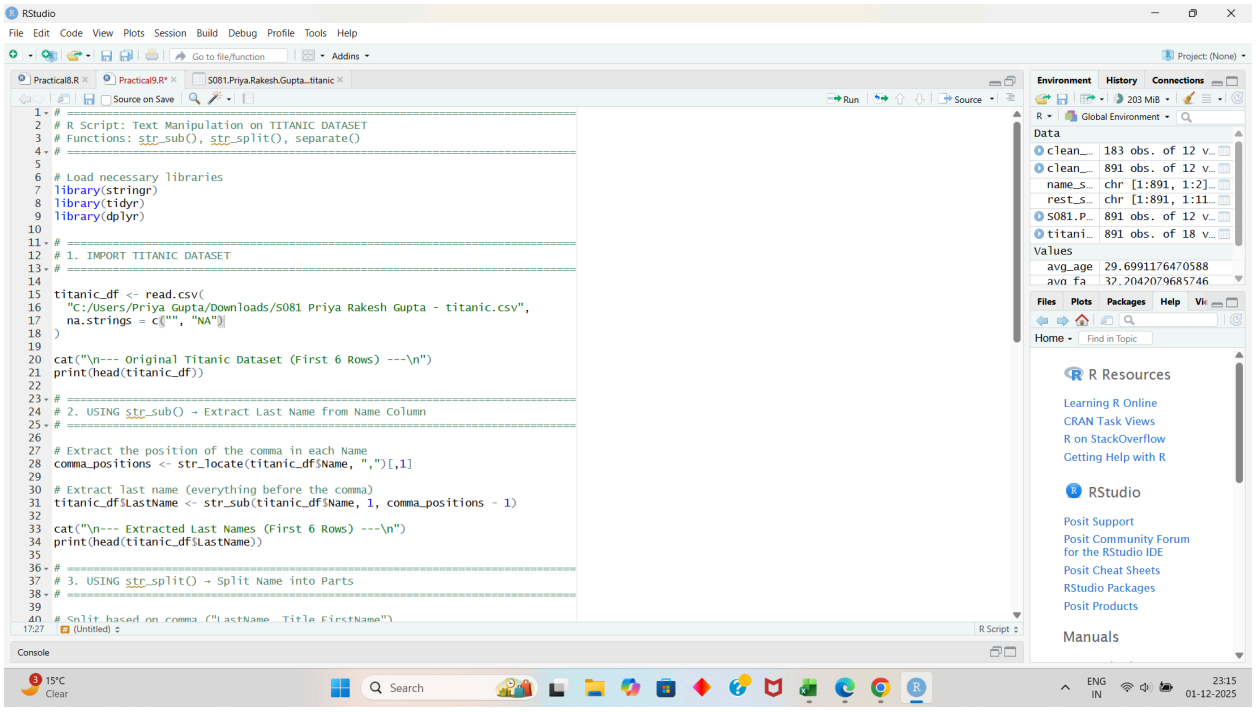
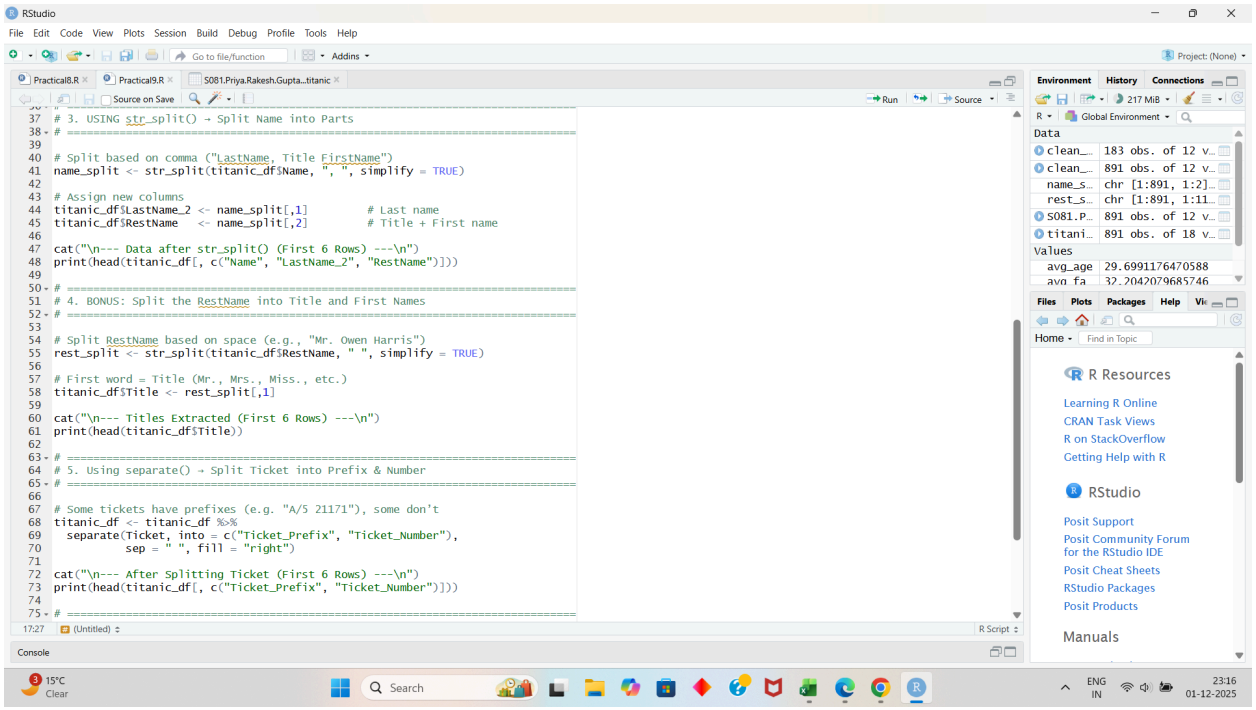
Sheth L.U.J. & Sir M.V. College

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



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

# Load necessary libraries
library(stringr)
library(tidyr)
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")
print(head(titanic_df))

# ================================================================
# 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")
print(head(titanic_df$LastName))

# ================================================================
# 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")
print(head(titanic_df[, c("Name", "LastName_2", "RestName")]))

# ================================================================
# 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")
print(head(titanic_df$Title))

# ================================================================
# 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")

cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
print(head(titanic_df[, c("Ticket_Prefix", "Ticket_Number")]))

# ================================================================
```

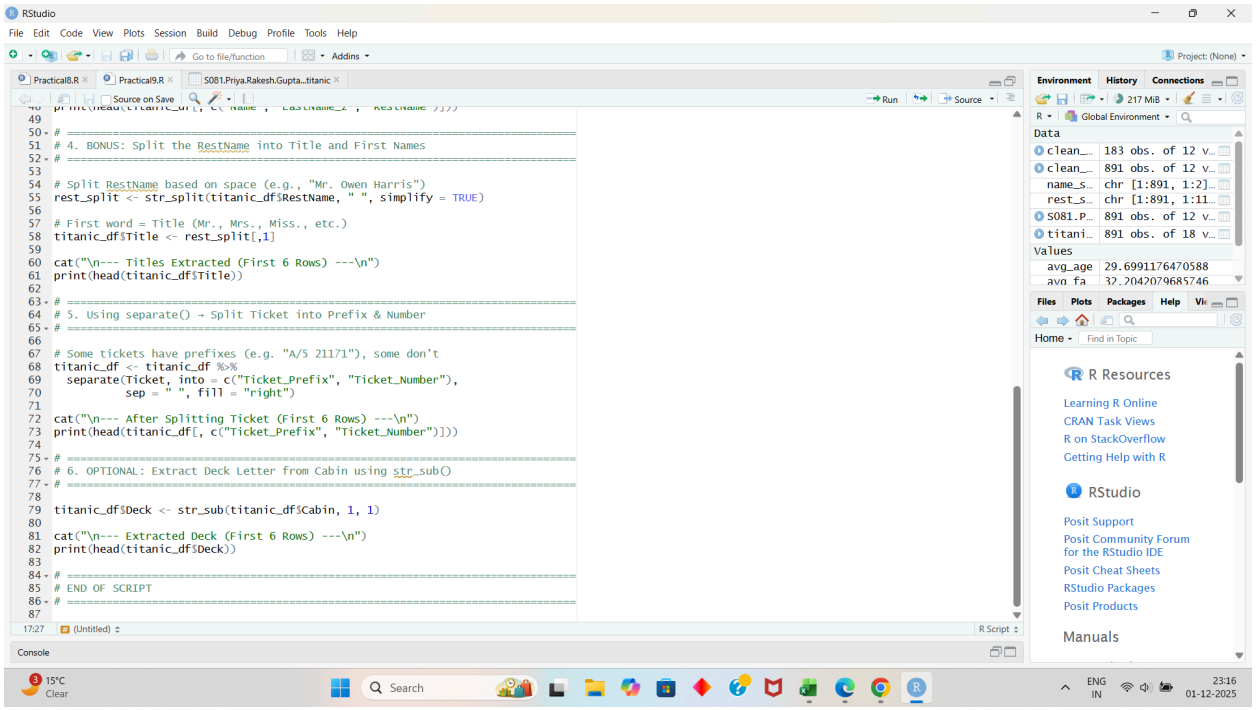


```
# ================================================================
# 3. USING str_split() → Split Name into Parts
# ================================================================

# 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")
print(head(titanic_df[, c("Name", "LastName_2", "RestName")]))

# ================================================================
# 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")
print(head(titanic_df$Title))

# ================================================================
# 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")

cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
print(head(titanic_df[, c("Ticket_Prefix", "Ticket_Number")]))

# ================================================================
```

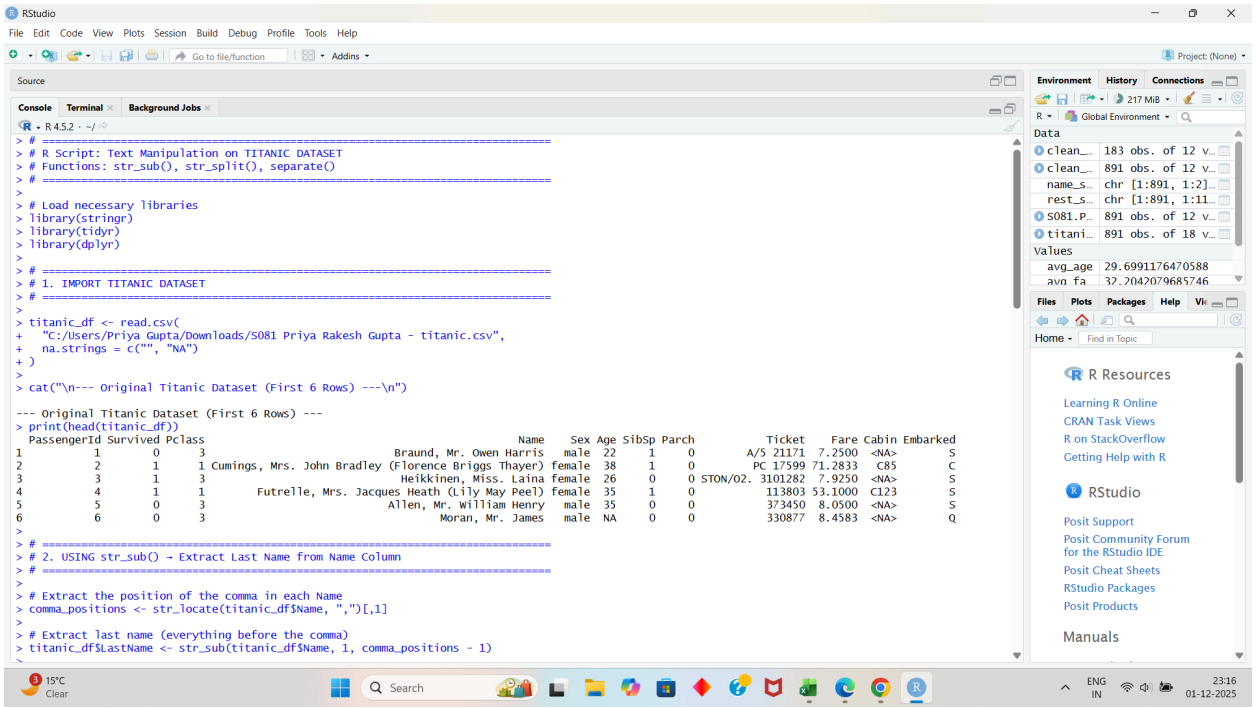
Name :- Priya Gupta

Roll no. :- S081

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



```
# ================================================================
# 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")
print(head(titanic_df$Title))

# ================================================================
# 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")

cat("\n--- After Splitting Ticket (First 6 Rows) ---\n")
print(head(titanic_df[, c("Ticket_Prefix", "Ticket_Number")]))

# ================================================================
# 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")
print(head(titanic_df$Deck))

# ================================================================
# END OF SCRIPT
# ================================================================
```

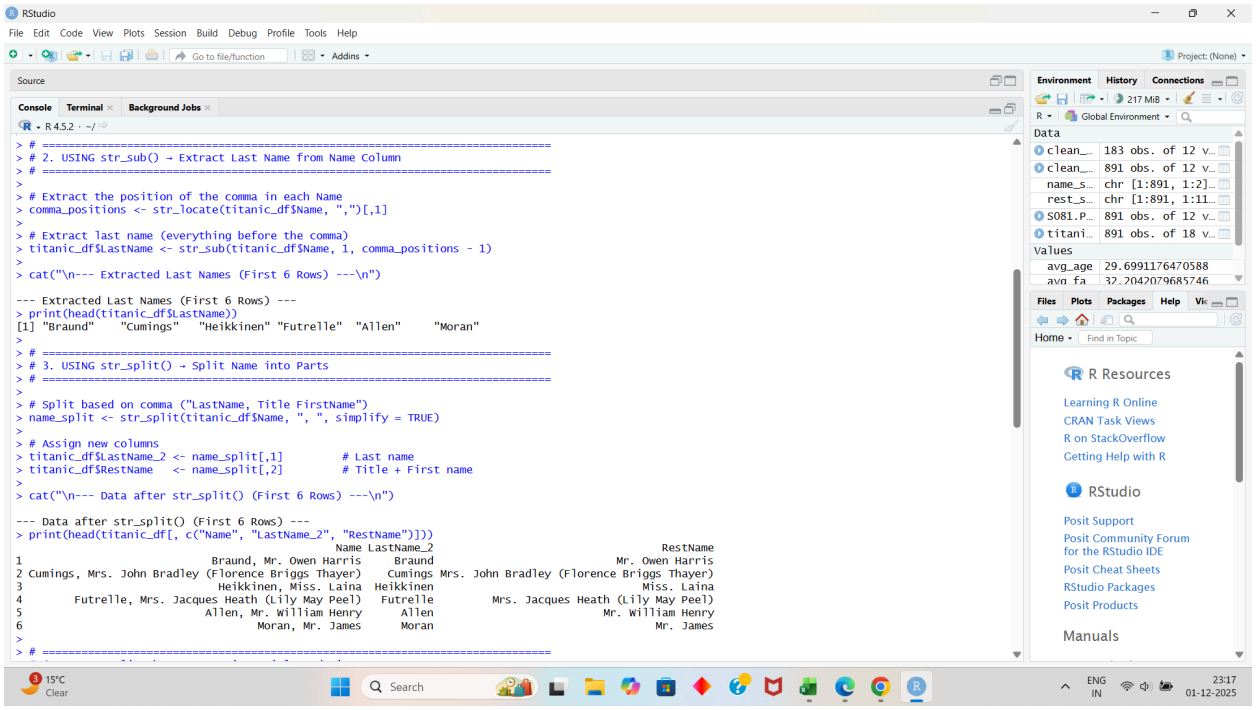


```
> # ================================================================
> # R Script: Text Manipulation on TITANIC DATASET
> # Functions: str_sub(), str_split(), separate()
> # ================================================================
>
> # Load necessary libraries
> library(stringr)
> library(tidyr)
> 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 21171  7.2500  <NA>        S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250  <NA>        S
4           4        1      1         Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0           373450  8.0500  <NA>        S
6           6        0      3                                     Moran, Mr. James   male  NA     0     0           330877  8.4583  <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)
```

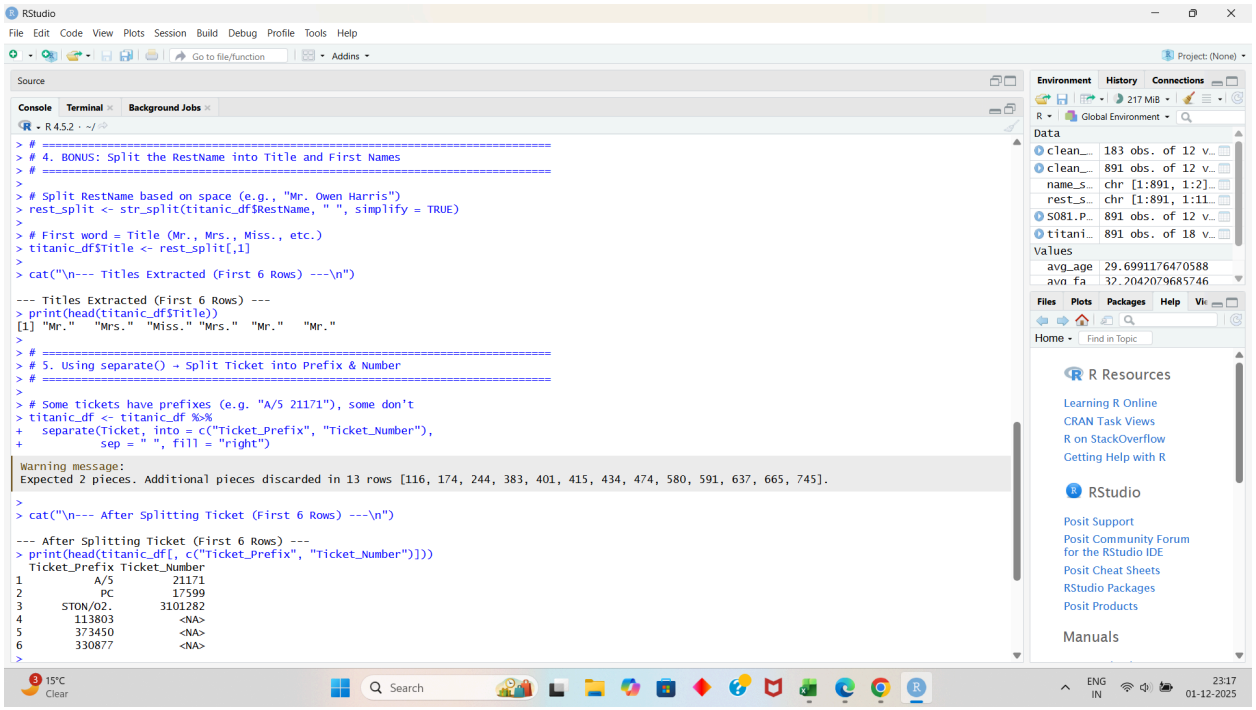
Name :- Priya Gupta

Roll no. :- S081

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



```
> # ================================================================
> # 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
>
> # ================================================================
```

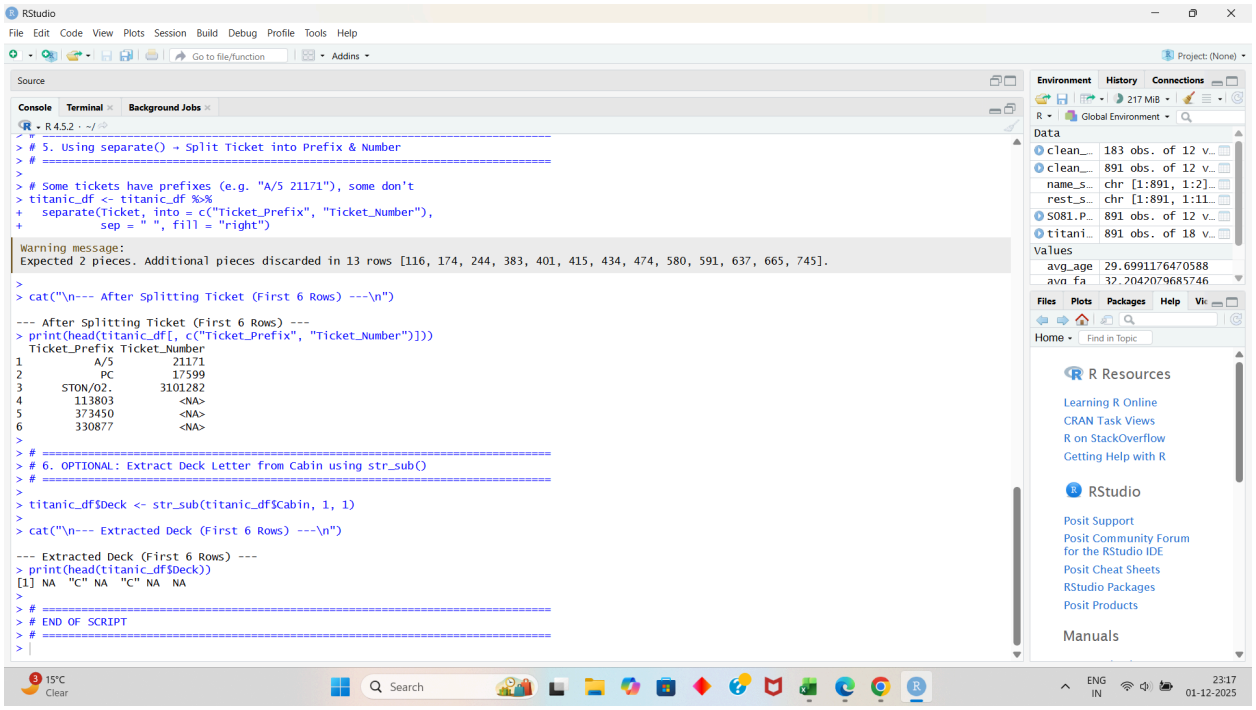


```
> # ================================================================
> # 4. BONUS: Split the RestName into Title and First Names
> # ================================================================
>
> # Split RestName based on space (e.g., "Mr. Owen Harris")
> rest_split <- str_split(titanic_df$RestName, " ", simplify = TRUE)
>
> # First word = Title (Mr., Mrs., Miss., etc.)
> titanic_df$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." "Mrs."  "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



Name :- Priya Gupta

Roll no. :- S081