

SHETH L.U.J AND SIR M.V. COLLEGE
SUBJECT NAME: Data Analysis with SAS / SPSS / R
Practical No. 9

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

Output-

The screenshot shows the RStudio interface with the following content:

```
R - R4.1.2 ~ /
> library(stringr)
> library(tidyverse)
> library(dplyr)
> # =====
> # 1. IMPORT DATASET
> # =====
> # using your corrected file path
> cricket <- read.csv("D:/Dhairya/R/FASTEST_Fifties_All_Seasons_combine.csv")
> print("---- Original Dataset (First 3 rows) ----")
[1] "---- Original Dataset (First 3 rows) ----"
> print(head(cricket %>% select(player, venue, match.date), 3))
  player      venue match.date
1 Yusuf Pathan Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 Kumar Sangakkara IS Bindra Stadium 25 April 2008
3 Virender Sehwag Arun Jaitley Stadium 17 May 2008
> # Extract the first 3 characters
> cricket$venue_code <- str_sub(cricket$venue, 1, 3)
> # Extract the last 4 characters from Match Date (to get the Year quickly)
> cricket$year_extracted <- str_sub(cricket$match.date, -4, -1)
> print("---- Data after str_sub() ----")
[1] "---- Data after str_sub() ----"
> print(head(cricket %>% select(venue, venue_code, year_extracted), 3))
  venue venue_code year_extracted
1 Rajiv Gandhi Intl. Cricket Stadium Raj 2008
2 IS Bindra Stadium IS 2008
3 Arun Jaitley Stadium Aru 2008
> # simplify = TRUE returns a matrix so we can grab specific columns easily
> split_matrix <- str_split(cricket$match.date, " ", simplify = TRUE)
> # Assign the split parts to new columns
> cricket$day <- split_matrix[, 1] # First part (e.g., "24")
> cricket$month <- split_matrix[, 2] # Second part (e.g., "April")
> print("---- Data after str_split() ----")
[1] "---- Data after str_split() ----"
> print(head(cricket %>% select(match.date, day, month), 3))
  match.date day month
1 24 April 2008 24 April
2 25 April 2008 25 April
3 17 May 2008 17 May
> tidy_cricket <- cricket %>%
+ separate(match.date, into = c("date_day", "date_month", "date_year"), sep = " ")
> print("---- Bonus: The 'separate' function ----")
[1] "---- Bonus: The 'separate' function ----"
> print(head(tidy_cricket %>% select(date_day, date_month, date_year), 3))
  date_day date_month date_year
1 24 April 2008
2 25 April 2008
3 17 May 2008
```

The screenshot shows the continuation of the RStudio output with the following content:

```
[1] "---- Data after str_split() ----"
> print(head(cricket %>% select(match.date, day, month), 3))
  match.date day month
1 24 April 2008 24 April
2 25 April 2008 25 April
3 17 May 2008 17 May
> tidy_cricket <- cricket %>%
+ separate(match.date, into = c("date_day", "date_month", "date_year"), sep = " ")
> print("---- Bonus: The 'separate' function ----")
[1] "---- Bonus: The 'separate' function ----"
> print(head(tidy_cricket %>% select(date_day, date_month, date_year), 3))
  date_day date_month date_year
1 24 April 2008
2 25 April 2008
3 17 May 2008
```