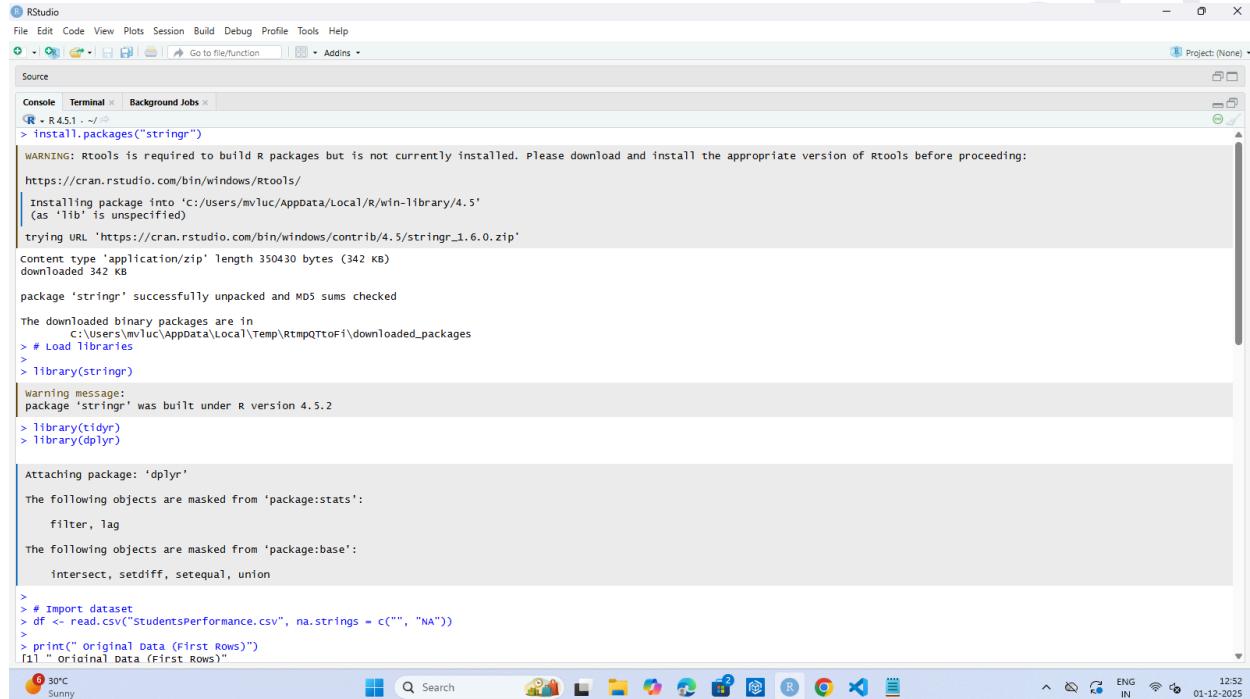


Practical No 9

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

Output :



The screenshot shows the RStudio interface with the following R code in the console:

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Terminal Background Jobs
R 4.5.1 - ~/~
> 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/mvluc/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\mvluc\AppData\Local\Temp\RtmpQztoFI\downloaded_packages
> # Load libraries
>
> library(stringr)
warning message:
package 'stringr' was built under R version 4.5.2
> library(tidyverse)
> 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
>
> # Import dataset
> df <- read.csv("studentsPerformance.csv", na.strings = c("", "NA"))
>
> print(" original data (First Rows)")
[1] " original data (First Rows)"
```

The RStudio interface includes a menu bar, a toolbar with icons for file operations, and a status bar at the bottom showing system information like battery level, network, and date.

SHETH L.U.J. AND SIR M.V. COLLEGE OF ARTS SCIENCE AND COMMERCE

SUBJECT : R Programming

The screenshot shows the RStudio interface with the following R code in the console:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)
Source
Console Terminal Background Jobs
[R - R 4.5.1 - ~]
> # Import dataset
> df <- read.csv("studentsperformance.csv", na.strings = c("", "NA"))
>
> print(" Original Data (First Rows)")
[1] " Original Data (First Rows)"
> print(head(df))
#> gender race.ethnicity parental.level.of.education lunch test.preparation.course math.score reading.score writing.score
#> 1 female group B bachelor's degree standard none 72 72 74
#> 2 female group C some college standard completed 69 90 88
#> 3 female group B master's degree standard none 90 95 93
#> 4 male group A associate's degree free/reduced none 47 57 44
#> 5 male group C some college standard none 76 78 75
#> 6 female group B associate's degree standard none 71 83 78
>
> # 1. Using str_sub(): extract substrings
>
> # Extract first 3 letters of gender
> df$gender_code <- str_sub(df$gender, 1, 3)
>
> # Extract last 2 characters of lunch type
> df$lunch_end <- str_sub(df$lunch, -2, -1)
>
> print("Data After str_sub()")
[1] "Data After str_sub()"
> print(df %>% select(gender, gender_code, lunch, lunch_end) %>% head())
#> gender gender_code lunch lunch_end
#> 1 female fem standard rd
#> 2 female fem standard rd
#> 3 female fem standard rd
#> 4 male mal free/reduced ed
#> 5 male mal standard rd
#> 6 female fem standard rd
>
> # 2. using str_split(): split text into parts
>
> # split 'race/ethnicity' column (e.g., "group A" -> "group", "A")
> split_matrix <- str_split(df$race.ethnicity, "-", simplify = TRUE)
>
> df$race_main <- split_matrix[, 1]
> df$race_group <- split_matrix[, 2]
>
> print("---- Data After str_split() ----")
[1] "---- Data After str_split()----"

```

The screenshot shows the RStudio interface with the following R code in the console:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)
Source
Console Terminal Background Jobs
[R - R 4.5.1 - ~]
#> 1 female fem standard - rd
#> 2 female fem standard rd
#> 3 female fem standard rd
#> 4 male mal free/reduced ed
#> 5 male mal standard rd
#> 6 female fem standard rd
>
> # 2. using str_split(): split text into parts
>
> # split 'race/ethnicity' column (e.g., "group A" -> "group", "A")
> split_matrix <- str_split(df$race.ethnicity, "-", simplify = TRUE)
>
> df$race_main <- split_matrix[, 1]
> df$race_group <- split_matrix[, 2]
>
> print("---- Data After str_split() ----")
[1] "---- Data After str_split()----"
> print(df %>% select(race.ethnicity, race_main, race_group) %>% head())
#> race.ethnicity race_main race_group
#> 1 group B group B
#> 2 group C group C
#> 3 group B group B
#> 4 group A group A
#> 5 group C group C
#> 6 group B group B
>
> # 3. Tidy method: separate()
>
> tidy_df <- df %>%
+   separate(race.ethnicity, into = c("race_word", "race_letter"), sep = "-")
>
> print(" Data After separate()")
[1] " Data After separate()"
> print(tidy_df %>% select(race_word, race_letter) %>% head())
#> race_word race_letter
#> 1 B
#> 2 C
#> 3 B
#> 4 A
#> 5 C
#> 6 B
>

```