

MVLU COLLEGE

PRACTICAL NO :- 09

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

CODE :-

```
install.packages("stringr")
library(stringr)
library(tidyr)
library(dplyr)

retail_data <- data.frame(
  SKU = c("ELEC-5548-2023", "HOME-3045-2022", "CLOT-4004-2023",
          "ELEC-4808-2021", "HOME-1817-2023"),
  Description = c("Electronics - Smart TV", "Home - Blender",
                 "Clothing - TShirt", "Electronics - Laptop", "Home - Sofa"),
  Price = c(500, 45, 20, 900, 300)
)
print("--- Original Dataset ---")
print(retail_data)
```

```
# 2. USING str_sub() (Substring Extraction)
retail_data$Category_Code <- str_sub(retail_data$SKU, 1, 4)
retail_data$Year <- str_sub(retail_data$SKU, -4, -1)

print("--- Data after str_sub() ---")
print(retail_data %>% select(SKU, Category_Code, Year))
```

```
# 3. USING str_split() (Splitting Strings)
split_list <- str_split(retail_data$Description, " - ")
print("--- Basic Split Output (List format) ---")
print(split_list[[1]]) # Show the first split item
```

```
split_matrix <- str_split(retail_data$Description, " - ", simplify = TRUE)
retail_data>Main_Cat <- split_matrix[, 1]
retail_data$Sub_Cat <- split_matrix[, 2]

print("--- Data after str_split() ---")
print(retail_data %>% select(Description, Main_Cat, Sub_Cat))
```

```
# 4. BONUS — Tidyverse Method (separate)
tidy_data <- retail_data %>%
  separate(SKU, into = c("Dept", "ID", "Mfg_Year"), sep = "-")

print("--- Bonus: The 'separate' function output ---")
print(tidy_data %>% select(Dept, ID, Mfg_Year))
```

MVLU COLLEGE

```
package 'stringr' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
  C:\users\itlab\AppData\Local\Temp\Rtmpw0vhpx\downloaded_packages
> library(stringr)
> 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

> retail_data <- data.frame(
+   SKU = c("ELEC-5548-2023", "HOME-3045-2022", "CLOT-4004-2023",
+         "ELEC-4808-2021", "HOME-1817-2023"),
+   Description = c("Electronics - Smart TV", "Home - Blender",
+                 "Clothing - Tshirt", "Electronics - Laptop", "Home - Sofa"),
+   Price = c(500, 45, 20, 900, 300)
+ )
> print("--- original dataset ---")
[1] "--- Original Dataset ---"
> print(retail_data)
      SKU          Description Price
1 ELEC-5548-2023 Electronics - Smart TV 500
2 HOME-3045-2022     Home - Blender 45
3 CLOT-4004-2023    Clothing - Tshirt 20
4 ELEC-4808-2021 Electronics - Laptop 900
5 HOME-1817-2023     Home - Sofa 300
> # 2. USING str_sub() (Substring Extraction)
> retail_data$Category.Code <- str_sub(retail_data$SKU, 1, 4)
> retail_data$Year <- str_sub(retail_data$SKU, -4, -1)
> print(" --- Data after str_sub() ---")
[1] " --- Data after str_sub() ---"
> print(retail_data %>% select(SKU, category_code, Year))
#> SKU Category.Code Year
#> 1 ELEC 2023
#> 2 HOME 2022
#> 3 CLOT 2023
#> 4 ELEC 2021
#> 5 HOME 2023
>
> # 3. USING str_split() (Splitting strings)
> split_list <- str_split(retail_data$Description, " ")
> print(split_list[[1]]) # Show the first split item
[1] "Electronics" "Smart TV"
>
> split_matrix <- str_split(retail_data$Description, " ", simplify = TRUE)
> retail_data$main_cat <- split_matrix[, 1]
> retail_data$sub_cat <- split_matrix[, 2]
>
> print(retail_data %>% select>Description, main_cat, sub_cat)
#> Description main_cat sub_cat
#> 1 Electronics - Smart TV Electronics Smart TV
#> 2 Home - Blender     Home   Blender
#> 3 Clothing - Tshirt Clothing   Tshirt
#> 4 Electronics - Laptop Electronics Laptop
#> 5 Home - Sofa       Home     Sofa
> view(split_matrix)
> view(split_list)
> # 4. BONUS - Tidyverse Method (separate)
> tidy_data <- retail_data %>%
+   separate(SKU, into = c("Dept", "ID", "Mfg_Year"), sep = "-")
> print(tidy_data %>% select(Dept, ID, Mfg_Year))
#> Dept ID Mfg_Year
#> 1 ELEC 5548 2023
#> 2 HOME 3045 2022
#> 3 CLOT 4004 2023
#> 4 ELEC 4808 2021
#> 5 HOME 1817 2023
```