

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))
```

NANDINI PANDIT S100

DATA ANALYSIS WITH SAS/SPSS/R.

MVLU COLLEGE

The image displays two screenshots of the RStudio interface, showing R code execution and the environment panel.

Top Screenshot:

- Console:** Shows the installation of the 'stringr' package and the loading of 'stringr', 'tidyr', and 'dplyr' libraries. It then creates a data frame 'retail_data' with columns: SKU, Description, and Price. The data is printed, showing 5 rows of product information.
- Environment:** Lists the loaded packages and their object counts: 'school_type_filt' (0 obs. of 20 variables), 'selected_cols' (497 obs. of 3 variables), 'sleep_or_extrac...' (4363 obs. of 20 variables), 'split_list' (List of 5), 'split_matrix' (chr [1:5, 1:2] "Electronics" "Home" "Clothin..."), 'starts_with_p' (497 obs. of 2 variables), 'student' (6607 obs. of 20 variables), 'student.Mental.h...' (101 obs. of 11 variables), and 'tidy_data' (5 obs. of 9 variables).

Bottom Screenshot:

- Console:** Continues the data manipulation. It uses 'str_sub()' to extract parts of the SKU and Description. It then uses 'str_split()' to split the Description into a list and a matrix. Finally, it uses 'tidy_data' and 'separate()' to split the SKU into 'Dept', 'ID', and 'Mfg_Year'.
- Environment:** Shows the updated environment with the same packages as the top screenshot, but with 'tidy_data' now having 5 obs. of 9 variables.