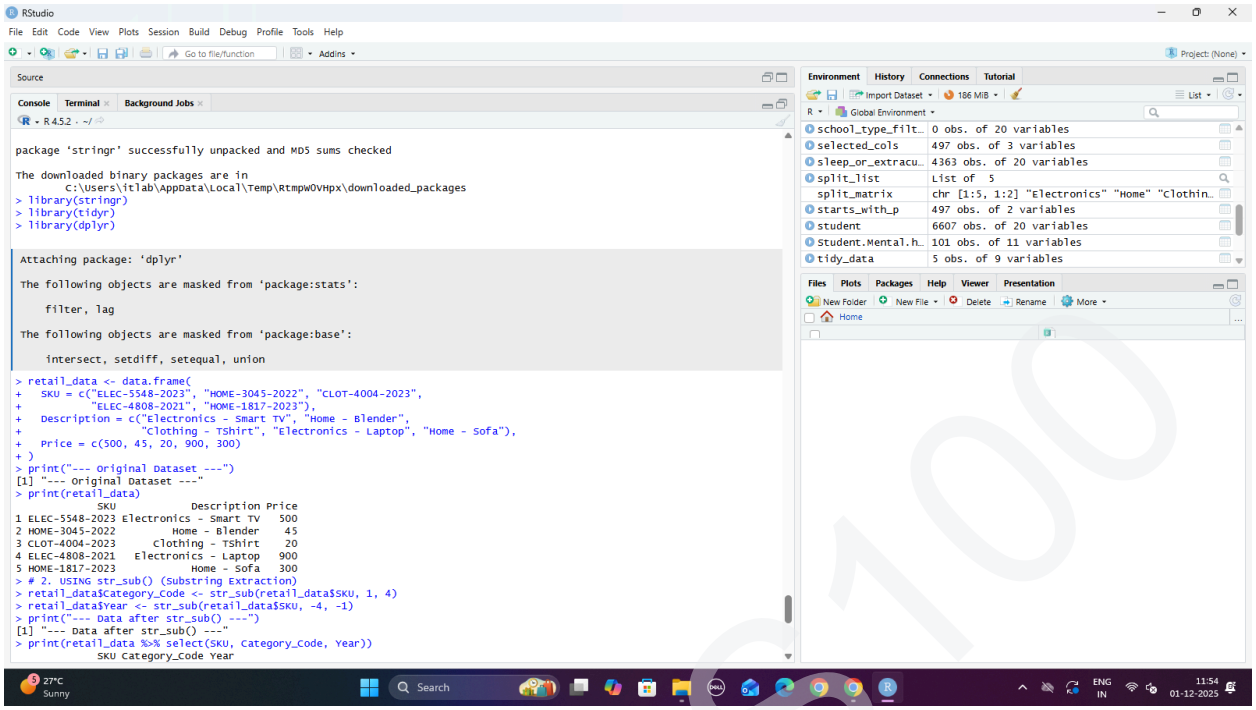
MVLU COLLEGE

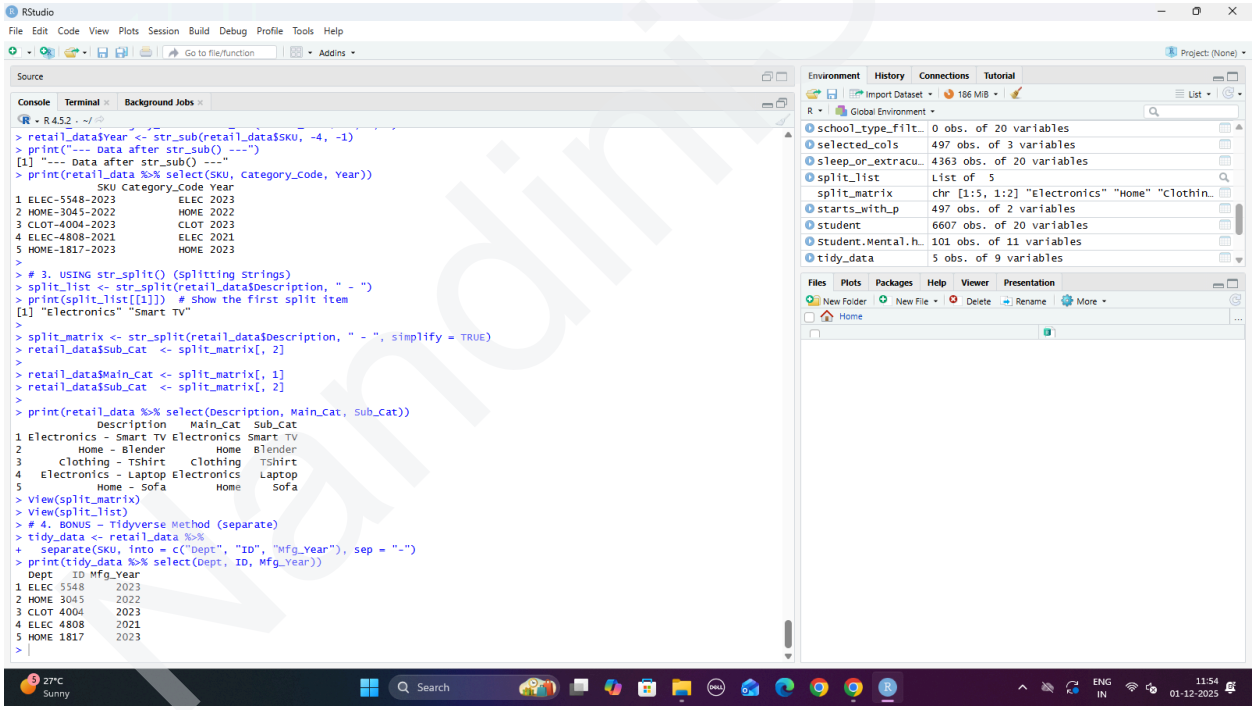
PRACTICAL NO :- 09
AIM :- Performing text manipulation using str_sub(), str_split() (R). import dataset.
CODE :-

```
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





NANDINI PANDIT S100
DATA ANALYSIS WITH SAS/SPSS/R.