

MVLU COLLEGE

PRACTICAL NO. 9

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

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for installing packages, loading data, and performing text manipulation.
- Environment Pane:** Displays the objects created in the R session.

```
1 install.packages("stringr")
2 install.packages("tidyr") # for separating columns after splitting
3 library(stringr)
4 library(tidyverse)
5 library(dplyr)
6 retail_data <- data.frame(
7   SKU = c("ELEC-5548-2023", "HOME-3045-2023", "CLOT-4004-2023", "ELEC-4808-2021", "HOME-1817-2023"),
8   Description = c("Electronics - Smart TV", "Home - Blender", "Clothing - Tshirt", "Electronics - Laptop", "Home - Sofa"),
9   Price = c(500, 45, 20, 900, 300)
10 )
11 print("--- Original Dataset ---")
12 print(retail_data)
13 retail_data$category_code <- str_sub(retail_data$SKU, 1, 4)
14 retail_data$year <- str_sub(retail_data$SKU, -4, -1)
15 print("--- Data after str_sub() ---")
16 print(retail_data %>% select(SKU, category_code, year))
17 split_list <- str_split(retail_data$Description, " - ")
18 print("--- Basic Split output (List format) ---")
19 print(split_list[[1]])
20 split_matrix <- str_split(retail_data$Description, " - ", simplify = TRUE)
21 retail_data$main_cat <- split_matrix[, 1]
22 retail_data$sub_cat <- split_matrix[, 2]
23 print("--- Data after str_split() (Manual Assignment) ---")
24 print(retail_data %>% select(Description, main_cat, sub_cat))
25 tidy_data <- retail_data %>%
26   separate(SKU, into = c("dept", "ID", "Mfg_Year"), sep = "-")
27 print("--- Bonus: The 'separate' function (easier splitting) ---")
28 print(tidy_data %>% select(dept, ID, Mfg_Year))
29
30
```

Environment Pane:

| Object | Variables |
|--------------------|---|
| appended | 298 obs. of 17 variables |
| clean_omit | 540 obs. of 5 variables |
| clean_replace | 4362 obs. of 5 variables |
| Cleaned_BMW_Sales | 99 obs. of 11 variables |
| Cleaned_Car_Price | 199 obs. of 10 variables |
| df1 | 99 obs. of 11 variables |
| df2 | 199 obs. of 10 variables |
| dropped_multiple | 4573 obs. of 13 variables |
| dropped_one | 4573 obs. of 14 variables |
| dropped_range | 4573 obs. of 11 variables |
| housing | 4573 obs. of 15 variables |
| merged_full | 294 obs. of 17 variables |
| merged_inner | 4 obs. of 17 variables |
| merged_left | 99 obs. of 17 variables |
| range_cols | 4573 obs. of 6 variables |
| retail_data | 5 obs. of 7 variables |
| retail_df | 4362 obs. of 5 variables |
| Retail_Product_Re | 4362 obs. of 5 variables |
| selected_cols | 4573 obs. of 3 variables |
| split_list | List of 5 |
| split_matrix | chr [1:5, 1:2] "Electronics" "Home" "Clothing"... |
| spotify | 4573 obs. of 15 variables |
| spotify_data_clean | 8573 obs. of 15 variables |
| starts_with_track | 4573 obs. of 5 variables |
| tidy_data | 5 obs. of 9 variables |

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for installing packages.
- Console:** Displays the output of the R commands.

```
> install.packages("stringr")
> install.packages("tidyr")
```

Console Output:

```
R 4.5.2 ~\
> 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/itlab/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/itlab/AppData/Local/Temp/RtmpQVEAM1/downloaded_packages
> install.packages("tidyr") # for separating columns after splitting
Restarting R session...
> install.packages("tidyr")
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/itlab/AppData/Local/R/win-library/4.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.5/tidyr_1.3.1.zip'
Content type 'application/zip' length 1276404 bytes (1.2 MB)
downloaded 1.2 MB
package 'tidyr' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:/Users/itlab/AppData/Local/Temp/RtmpGUAfGr/downloaded_packages
> library(stringr)
> library(tidyr)
> library(dplyr)
```

The screenshot shows an RStudio session. The console output is as follows:

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
[Icons] Go to file/function Addins
Source
Console Background Jobs
R - R452 - ~/R
> library(stringr)
> library(tidyr)
> 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 - T-Shirt", "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 - T-Shirt 20
4 ELEC-4808-2021 Electronics - Laptop 900
5 HOME-1817-2023 Home - Sofa 300
> 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-5548-2023 ELEC 2023
2 HOME-3045-2022 HOME 2022
3 CLOT-4004-2023 CLOT 2023
4 ELEC-4808-2021 ELEC 2021
5 HOME-1817-2023 HOME 2023
> split_list <- str_split(retail_data$description, " - ")
Environment History Connections Tutorial
R - Global Environment -
Data
  appended 298 obs. of 17 variables
  clean_omit 540 obs. of 5 variables
  clean_replace 4362 obs. of 5 variables
  cleaned_bmw_sales 99 obs. of 11 variables
  cleaned_car_price 199 obs. of 10 variables
  df1 99 obs. of 11 variables
  df2 199 obs. of 10 variables
  dropped_multiple 4573 obs. of 13 variables
  dropped_one 4573 obs. of 14 variables
  dropped_range 4573 obs. of 11 variables
  housing 4573 obs. of 15 variables
  merged_full 294 obs. of 17 variables
  merged_inner 4 obs. of 17 variables
  merged_left 99 obs. of 17 variables
  range_cols 4573 obs. of 6 variables
  retail_data 5 obs. of 7 variables
  retail_df 4362 obs. of 5 variables
  Retail_Product_Re 4362 obs. of 5 variables
  selected_cols 4573 obs. of 3 variables
  split_list List of 5
  split_matrix chr [1:5, 1:2] "Electronics" "Home" "Clothing..."
  spotify 4573 obs. of 15 variables
  spotify_data_clean 8573 obs. of 15 variables
  starts_with_track 4573 obs. of 5 variables
  tidy_data 5 obs. of 9 variables
  values
    avg_price 5016.97063037249
    key "vehicle_id"
    keys chr [1:4] "model" "year" "fuel_type" "transmissi..."
Files Plots Packages Help Viewer Presentation
```

SIYA S105
R PROGRAMMING