

MVLU COLLEGE

PRACTICAL NO. 9

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

```

install.packages("stringr")
install.packages("tidyverse") # for separating columns after splitting
library(stringr)
library(tidyverse)
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 - s"),
  Price = c(500, 45, 20, 900, 300)
)
print("--- Original Dataset ---")
print(retail_data)
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))
split_list <- str_split(retail_data$Description, " - ")
print(" --- Basic Split output (List format) ---")
print(split_list[[1]])
split_main_cat <- split_matrix[[1]]
retail_data$Sub_Cat <- split_matrix[[2]]
print(" --- Data after str_split() (Manual Assignment) ---")
print(retail_data %>% select>Description, Main_Cat, sub_cat)
tidy_data <- retail_data %>%
  separate(SKU, into = c("Dept", "ID", "Mfg_year"), sep = "-")
print(" --- Bonus: The 'separate' function (easier splitting) ---")
print(tidy_data %>% select(Dept, ID, Mfg_Year))

```

R Script

Console

30°C Sunny 12:54 01-12-2025

```

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\tmp\rtmpQVEam1\downloaded_packages
> install.packages("tidyverse") # for separating columns after splitting
| Restarting R session...
> install.packages("tidyverse")
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 'tidyverse' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
  c:\users\itlab\appdata\local\tmp\RtmpGuAFGr\downloaded_packages
> library(stringr)
> library(tidyverse)
> library(dplyr)

```

R Studio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source Environment History Connections Tutorial Project: (None)

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_omni 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

30°C Sunny 12:54 01-12-2025

MVLU COLLEGE

PRACTICAL NO. 9

RStudio Environment View:

- Data** pane shows objects like appended, clean_omit, clean_replace, cleaned_BMW_Sales, cleaned_car_Price, df1, df2, dropped_multiple, dropped_one, dropped_range, housing, merged_full, merged_inner, merged_left, range_cols, retail_data, retail_df, Retail_Product_Re, selected_cols, split_list, split_matrix, spotify, spotify_data_clean, starts_with_track, tidy_data.
- Values** pane shows avg_price, key, keys.

Source Code (R Script):

```

> 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
> 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$description, " - ")
> print("---- Basic Split Output (List format) ----")
[1] "---- Basic split output (List format) ----"
> print(split_list[[1]])
[1] "Electronics" "Smart TV"
> split_matrix <- str_split(retail_data$description, " - ", simplify = TRUE)
> retail_data$sub_cat <- split_matrix[, 2]
> print("---- Data after str_split() (Manual Assignment) ----")
[1] "---- Data after str_split() (Manual Assignment) ----"
> 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
> tidy_data <- retail_data %>%
+   separate(SKU, into = c("Dept", "ID", "Mfg_Year"), sep = "-")
> print("---- Bonus: The 'separate' function (easier splitting) ----")
[1] "---- Bonus: The 'separate' function (easier splitting) ----"
> print(tidy_data %>% select(dept, ID, Mfg_Year))
  Dept ID Mfg_Year
1 ELEC 1 2023
2 HOME 3045 2022
3 CLOT 4004 2023
4 ELEC 4808 2021
5 HOME 1817 2023
  
```

RStudio Environment View:

- Data** pane shows objects like appended, clean_omit, clean_replace, cleaned_BMW_Sales, cleaned_car_Price, df1, df2, dropped_multiple, dropped_one, dropped_range, housing, merged_full, merged_inner, merged_left, range_cols, retail_data, retail_df, Retail_Product_Re, selected_cols, split_list, split_matrix, spotify, spotify_data_clean, starts_with_track, tidy_data.
- Values** pane shows avg_price, key, keys.

Source Code (R Script):

```

> 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
> 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$description, " - ")
> print("---- Basic Split Output (List format) ----")
[1] "---- Basic split output (List format) ----"
> print(split_list[[1]])
[1] "Electronics" "Smart TV"
> split_matrix <- str_split(retail_data$description, " - ", simplify = TRUE)
> retail_data$sub_cat <- split_matrix[, 2]
> print("---- Data after str_split() (Manual Assignment) ----")
[1] "---- Data after str_split() (Manual Assignment) ----"
> 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
> tidy_data <- retail_data %>%
+   separate(SKU, into = c("Dept", "ID", "Mfg_Year"), sep = "-")
> print("---- Bonus: The 'separate' function (easier splitting) ----")
[1] "---- Bonus: The 'separate' function (easier splitting) ----"
> print(tidy_data %>% select(dept, ID, Mfg_Year))
  Dept ID Mfg_Year
1 ELEC 1 2023
2 HOME 3045 2022
3 CLOT 4004 2023
4 ELEC 4808 2021
5 HOME 1817 2023
  
```