

SHETH L.U.J AND SIR M.V COLLEGE

Subject: Data Analysis with SAS / SPSS / R

Practical no. 9

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

Outputs→

The screenshot displays the RStudio environment. The console window on the left shows the following commands and output:

```
> print("-----S071 Gazala Alam-----")
[1] "-----S071 Gazala Alam-----"
> install.packages("stringr")

Restarting R session...

> 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/gazala/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\gazala\AppData\Local\Temp\RtmpkjbNQJ\downloaded_packages
> 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/gazala/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 Environment pane on the right shows the following objects:

Object	Details
range_cols	200 obs. of 3 variables
sales	1000 obs. of 16 variables
sales_data_sales...	1000 obs. of 14 variables
sales_data_xls	1000 obs. of 14 variables
selected_cols	200 obs. of 3 variables
sorted_by_volume	122 obs. of 7 variables
split_matrix	chr [1:1000, 1:2] "North" "west" "South" "Sou..."
starts_with_d	200 obs. of 3 variables
tidy_sales	1000 obs. of 15 variables

The Files pane at the bottom shows a list of files in the current directory, including 'Leadership it3.pdf', 'My Music', 'My Pictures', 'My Videos', 'NetBeansProjects', 'practical 5 demuli.circ', 'practical 5 multi.circ', 'practical 9.circ', 'sgfdu05m.ini', 'statistic research work.pdf', 'Steps for IP Address.docx', 'The-Life-and-Legacy-of-APJ-Abdul-Kalam (blueprint)', 'Virtual Machines', 'Food_Delivery_Route_Efficiency_Dataset - Copy.csv', and 'sales_data - sales_data.csv'.

SHETH L.U.J AND SIR M.V COLLEGE

Subject: Data Analysis with SAS / SPSS /R

The screenshot shows the RStudio interface with the following components:

- Source:** Contains the R script being executed.
- Console:** Displays the output of the R script, including package loading messages and the first few rows of the 'sales' dataset.
- Environment:** Lists the objects in the global environment, including 'range_cols', 'sales', 'sales_data_sales...', 'sales_data_xls', 'selected_cols', 'sorted_by_volume', 'split_matrix', 'starts_with_d', and 'tidy_sales'.
- Files:** Shows the file explorer with various files and folders.

```
R - R 4.5.2 - ~/R
File Edit Code View Plots Session Build Debug Profile Tools Help

Source
Console Terminal Background Jobs
R - R 4.5.2 - ~/R
The downloaded binary packages are in
C:\Users\gazala\AppData\Local\Temp\RtmpkjbNQJ\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

> # =====
> # 1. IMPORT DATASET
> # =====
> sales <- read_csv("sales_data - sales_data.csv", na.strings = c("", "NA"))
> print("--- Original Dataset ---")
[1] "--- Original Dataset ---"
> print(head(sales))
  Product_ID Sale_Date Sales_Rep Region Sales_Amount Quantity_Sold
1      1052 2023-02-03      Bob North      5053.97           18
2      1093 2023-04-21      Bob West      4384.02           17
3      1015 2023-09-21    David South      4631.23           30
4      1072 2023-08-24      Bob South      2167.94           39
5      1061 2023-03-24    Charlie East      3750.20           13
6      1021 2023-02-11    Charlie West      3761.15           32
  Product_Category Unit_Cost Unit_Price Customer_Type Discount Payment_Method
1      Furniture      152.75      267.22      Returning 0.09      Cash
2      Furniture      3816.39     4209.44      Returning 0.11      Cash
3          Food       261.56       371.40      Returning 0.20      Bank Transfer
4      Clothing     4330.03     4467.75        New 0.02      Credit Card
5      Electronics     637.37     692.71        New 0.08      Credit Card
```

The screenshot shows the RStudio interface with the following components:

- Source:** Contains the R script being executed.
- Console:** Displays the output of the R script, including the result of the 'select' operation.
- Environment:** Lists the objects in the global environment, including 'range_cols', 'sales', 'sales_data_sales...', 'sales_data_xls', 'selected_cols', 'sorted_by_volume', 'split_matrix', 'starts_with_d', and 'tidy_sales'.
- Files:** Shows the file explorer with various files and folders.

```
R - R 4.5.2 - ~/R
File Edit Code View Plots Session Build Debug Profile Tools Help

Source
Console Terminal Background Jobs
R - R 4.5.2 - ~/R
5      Electronics     637.37     692.71        New 0.08      Credit Card
6          Food       900.79     1106.51        New 0.21      Cash
  Sales_Channel Region_and_Sales_Rep
1      Online      North-Bob
2      Retail      West-Bob
3      Retail      South-David
4      Retail      South-Bob
5      Online      East-Charlie
6      Online      West-Charlie
> sales$Category_Code <- str_sub(sales$Product_ID, 1, 4)
> sales$Item_Code <- str_sub(sales$Product_ID, -4, -1)
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(sales %>% select(Product_ID, Category_Code, Item_Code))
  Product_ID Category_Code Item_Code
1      1052          1052      1052
2      1093          1093      1093
3      1015          1015      1015
4      1072          1072      1072
5      1061          1061      1061
6      1021          1021      1021
7      1083          1083      1083
8      1087          1087      1087
9      1075          1075      1075
10     1075          1075      1075
11     1088          1088      1088
12     1100          1100      1100
13     1024          1024      1024
14     1003          1003      1003
15     1022          1022      1022
16     1053          1053      1053
17     1002          1002      1002
18     1088          1088      1088
19     1030          1030      1030
20     1038          1038      1038
21     1002          1002      1002
22     1064          1064      1064
23     1060          1060      1060
```

SHETH L.U.J AND SIR M.V COLLEGE

Subject: Data Analysis with SAS / SPSS /R

The screenshot shows the RStudio interface. The console displays a data table with 184 rows and 4 columns. The Environment pane on the right shows the following objects:

Object	Size	Modified
range_cols	200 obs. of 3 variables	
sales	1000 obs. of 16 variables	
sales_data_sales...	1000 obs. of 14 variables	
sales_data_xls	1000 obs. of 14 variables	
selected_cols	200 obs. of 3 variables	
sorted_by_volume	122 obs. of 7 variables	
split_matrix	chr [1:1000, 1:2] "North" "West" "South" "Sou..."	
starts_with_d	200 obs. of 3 variables	
tidy_sales	1000 obs. of 15 variables	

The screenshot shows the RStudio interface with the following R code executed in the console:

```
[ reached 'max' / getOption("max.print") -- omitted 667 rows ]
> split_matrix <- str_split(sales$Region_and_Sales_Rep, "-", simplify = TRUE)
> sales$Region <- split_matrix[, 1]
> sales$Sales_Rep <- split_matrix[, 2]
> print("--- Data after str_split() (Manual Assignment) ---")
[1] "--- Data after str_split() (Manual Assignment) ---"
> print(sales %>% select(Region_and_Sales_Rep, Region, Sales_Rep))
```

The console displays the resulting data table:

Region_and_Sales_Rep	Region	Sales_Rep
1 North-Bob	North	Bob
2 West-Bob	West	Bob
3 South-David	South	David
4 South-Bob	South	Bob
5 East-Charlie	East	Charlie
6 West-Charlie	West	Charlie
7 West-Bob	West	Bob
8 South-Eve	South	Eve
9 South-David	South	David
10 West-Charlie	West	Charlie
11 North-Eve	North	Eve

SHETH L.U.J AND SIR M.V COLLEGE

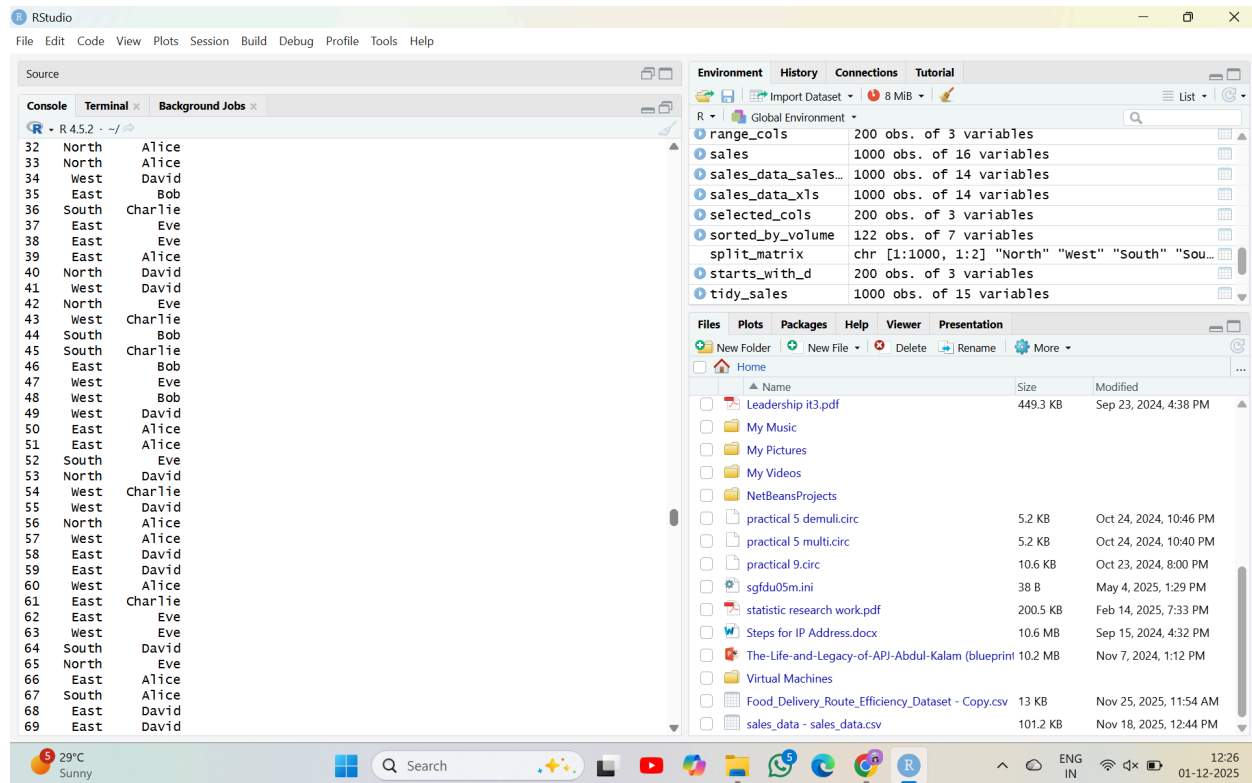
Subject: Data Analysis with SAS / SPSS /R

RStudio interface showing a data frame with columns for Region, Sales_Rep, and Name. The Environment pane shows the Global Environment with various objects like range_cols, sales, sales_data_sales..., sales_data_xls, selected_cols, sorted_by_volume, split_matrix, starts_with_d, and tidy_sales. The Files pane shows a file explorer view of the user's home directory.

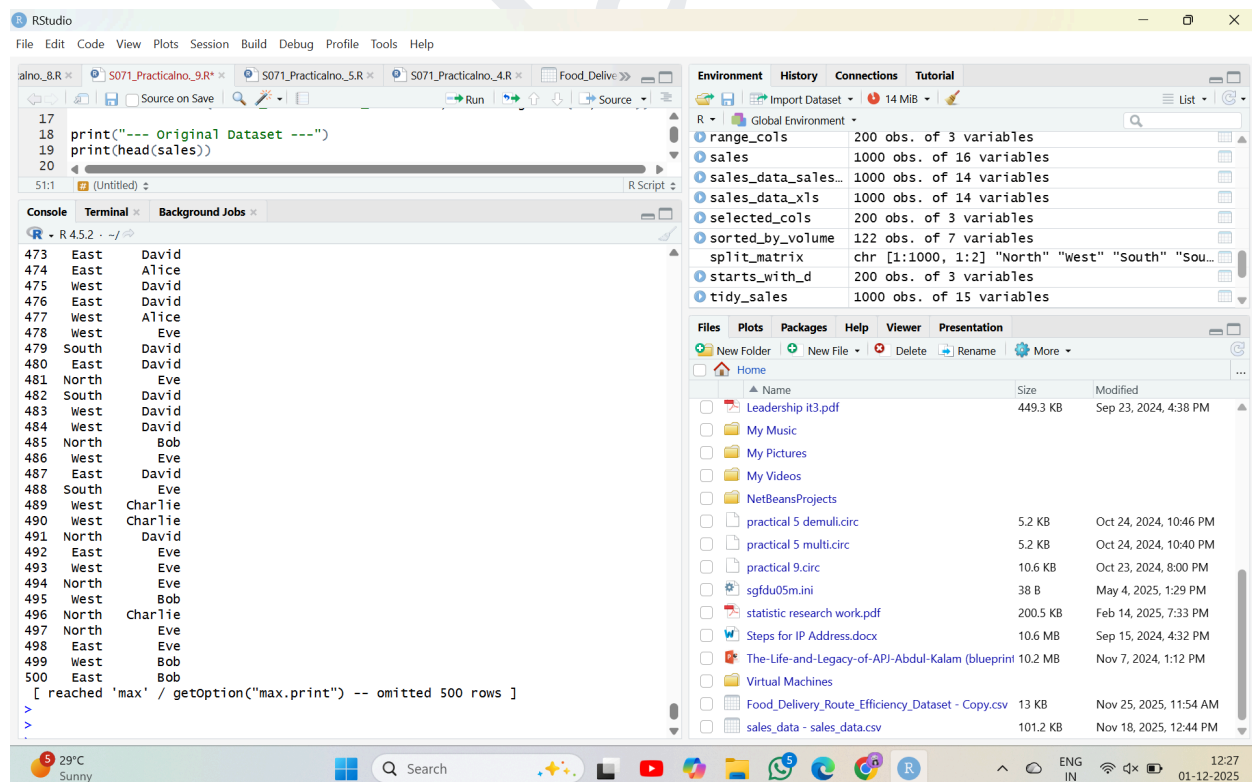
RStudio interface showing the same data frame as the previous screenshot. The Environment pane shows the same objects. The Files pane shows the same file explorer view. The console shows the execution of the tidy_sales object and the separate function.

SHETH L.U.J AND SIR M.V COLLEGE

Subject: Data Analysis with SAS / SPSS /R



The screenshot shows the RStudio interface. The console displays a list of names and regions, with line numbers 32 through 69. The environment pane on the right shows a list of objects: range_cols (200 obs. of 3 variables), sales (1000 obs. of 16 variables), sales_data_sales... (1000 obs. of 14 variables), sales_data_xls (1000 obs. of 14 variables), selected_cols (200 obs. of 3 variables), sorted_by_volume (122 obs. of 7 variables), split_matrix (chr [1:1000, 1:2] "North" "West" "South" "Sou...), starts_with_d (200 obs. of 3 variables), and tidy_sales (1000 obs. of 15 variables). The Files pane shows a directory structure with various files and folders.



The screenshot shows the RStudio interface with a script editor, console, and environment pane. The script editor contains R code for printing the original dataset and its head. The console displays the output of the code, showing a list of names and regions. The environment pane shows the same list of objects as the first screenshot.

```
17 print("--- Original Dataset ---")
18 print(head(sales))
19
20
```

```
473 East David
474 East Alice
475 West David
476 East David
477 West Alice
478 West Eve
479 South David
480 East David
481 North Eve
482 South David
483 West David
484 West David
485 North Bob
486 West Eve
487 East David
488 South Eve
489 West Charlie
490 West Charlie
491 North David
492 East Eve
493 West Eve
494 North Eve
495 West Bob
496 North Charlie
497 North Eve
498 East Eve
499 West Bob
500 East Bob
```

```
[ reached 'max' / getOption("max.print") -- omitted 500 rows ]
>
```