

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

## SUBJECT: Data Analysis with R

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

The top screenshot shows the RStudio interface with the following data frame 'df':

Category	Price	Rating	Stock	Discount
1 NA	5548	1.870322	NA	0
2 NA	3045	4.757798	NA	38
3 NA	4004	NA	In Stock	0
4 NA	4808	1.492085	NA	33
5 NA	1817	NA	Out of Stock	23
6 NA	3522	NA	NA	NA
7 C	667	3.668341	In Stock	41
8 A	7125	4.983998	Out of Stock	7
9 A	2777	2.678384	In Stock	6
10 NA	463	4.636187	NA	3

The bottom screenshot shows the final data frame 'df\_calc' with an added 'Final\_Price' column:

Category	Price	Rating	Stock	Discount	Discount_Amount	Final_Price
1 NA	5548	1.870322	NA	0	0.00	5548.00
2 NA	3045	4.757798	NA	38	1157.10	1887.90
3 NA	4004	0.000000	In Stock	0	0.00	4004.00
4 NA	4808	1.492085	NA	33	1586.64	3221.36
5 NA	1817	0.000000	Out of Stock	23	417.91	1399.09
6 NA	3522	0.000000	NA	0	0.00	3522.00
7 C	667	3.668341	In Stock	41	273.47	393.53
8 A	7125	4.983998	Out of Stock	7	498.75	6626.25
9 A	2777	2.678384	In Stock	6	166.62	2610.38
10 NA	463	4.636187	NA	3	13.89	449.11

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

## SUBJECT: Data Analysis with R

The screenshot displays the R Studio interface with the following components:

- Environment Pane:** Lists data frames: `df` (4362 obs. of 5 variables), `df_calc` (4362 obs. of 7 variables), `df_clean` (4362 obs. of 5 variables), `df_logic` (4362 obs. of 7 variables), `df_text` (4362 obs. of 6 variables), and `final_data` (4362 obs. of 8 variables).
- Files Pane:** Shows a file explorer with various files and folders, including `desktop.ini`, `disease_diagnosis.csv`, `GIS Database`, `ISExpress`, `loan_approval.csv`, `My Music`, `My Pictures`, `My Videos`, `My Web Sites`, `NetBeansProjects`, `Power BI Desktop`, `S103 Cn 1stpdf`, `scala for DS`, and `Sound Recordings`.
- Console:** Contains R code for data cleaning and logic application, followed by a printed table of results.
- Viewer Pane:** Displays a table with 7 columns: `Category`, `Price`, `Rating`, `Stock`, `Discount`, `Quality_Label`, and `Price_Category`. The table shows 10 rows of data.

**Console Code:**

```
R - R 4.5.2 - C:/Users/soham/Desktop/
> df_clean <- df %>%
+   filter(!is.na(Price), !is.na(Rating), !is.na(Stock), !is.na(Discount))
+   mutate(
+     Quality_Label = ifelse(Rating > 4.0, "Top Rated", "Average"),
+     # Let's add a second logic: Is it expensive?
+     Price_Category = ifelse(Price > 4000, "Premium", "Budget")
+   )
> print("--- Method B: Logic Results (Labels) ---")
[1] "--- Method B: Logic Results (Labels) ---"
> print(df_logic %>% select(Rating, Quality_Label, Price, Price_Category))
  Rating Quality_Label Price Price_Category
1  1.870322      Average    5548      Premium
2  4.757798    Top Rated    3045      Budget
3  0.000000      Average    4004      Premium
4  1.492085      Average    4808      Premium
5  0.000000      Average    1817      Budget
6  0.000000      Average    3522      Budget
7  3.668341      Average     667      Budget
8  4.983998    Top Rated    7125      Premium
9  2.678384      Average     2777      Budget
10 4.626187    Top Rated     463      Budget
11 2.947838      Average    1151      Budget
12 4.890750    Top Rated    3772      Budget
13 2.982242      Average    7719      Premium
14 1.270943      Average    8416      Premium
15 0.000000      Average    8530      Premium
16 3.032832      Average    7936      Premium
17 3.479064      Average    9319      Premium
18 4.097464    Top Rated     2066      Budget
19 0.000000      Average    2066      Budget
20 0.000000      Average    1784      Budget
```

**Viewer Table:**

Category	Price	Rating	Stock	Discount	Quality_Label	Price_Category
1 NA	5548	1.870322	NA	0	Average	Premium
2 NA	3045	4.757798	NA	38	Top Rated	Budget
3 NA	4004	0.000000	In Stock	0	Average	Premium
4 NA	4808	1.492085	NA	33	Average	Premium
5 NA	1817	0.000000	Out of Stock	23	Average	Budget
6 NA	3522	0.000000	NA	0	Average	Budget
7 C	667	3.668341	In Stock	41	Average	Budget
8 A	7125	4.983998	Out of Stock	7	Top Rated	Premium
9 A	2777	2.678384	In Stock	6	Average	Budget
10 NA	463	4.626187	NA	1	Top Rated	Budget

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

## SUBJECT: Data Analysis with R

**Environment** History Connections Tutorial

R • Global Environment •

Data

- df 4362 obs. of 5 variables
- df\_calc 4362 obs. of 7 variables
- df\_clean 4362 obs. of 5 variables
- df\_logic 4362 obs. of 7 variables
- df\_text 4362 obs. of 6 variables
- final\_data... 4362 obs. of 8 variables

**Files** Plots Packages Help Viewer Presentation

Home

- rhitory 1 KB Nov 2
- Custom Office Templates
- Database1.acddb 992 KB Oct 1
- desktop.ini 402 B Aug 2
- disease\_diagnosis.csv 189.1 KB Dec 1
- GIS Database
- ISExpress
- loan\_approval.csv 113.1 KB Nov 2
- My Music
- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- S103 Cn 1stpdf 2.6 MB Nov 15
- scala for DS
- Sound Recordings

**Console** Terminal Background Jobs

```
R • R452 - C:\Users\luj\Desktop\
> df$Is_High_Value = ifelse(df$Price > 4000, TRUE, FALSE)
+ Status_Report = paste0("Rating: ", round(Rating, 1), " / Dis: ", discount, "%")
+ )
> print("--- Final Combined Dataset ---")
[1] --- Final combined dataset ---
> print(head(final_dataset))
  Category Price Rating Stock discount Final_Price Is_High_Value Status_Report
1 <NA> 5548 1.870322 <NA> 0 5548.00 TRUE Rating: 1.9 / Dis: 0%
2 <NA> 3045 4.757798 <NA> 38 1887.90 FALSE Rating: 4.8 / Dis: 38%
3 <NA> 4004 0.000000 In Stock 0 4004.00 TRUE Rating: 0 / Dis: 0%
4 <NA> 4808 1.492085 <NA> 33 3221.36 TRUE Rating: 1.5 / Dis: 33%
5 <NA> 1817 0.000000 Out of Stock 23 1399.09 FALSE Rating: 0 / Dis: 23%
6 <NA> 3522 0.000000 <NA> 0 3522.00 TRUE Rating: 0 / Dis: 0%
>
> view(df)
> view(df_calc)
> view(df_clean)
> view(df_logic)
> view(df_text)
> |
```

Showing 1 to 13 of 4,362 entries, 6 total columns

**Environment** History Connections Tutorial

R • Global Environment •

Data

- df 4362 obs. of 5 variables
- df\_calc 4362 obs. of 7 variables
- df\_clean 4362 obs. of 5 variables
- df\_logic 4362 obs. of 7 variables
- df\_text 4362 obs. of 6 variables
- final\_data... 4362 obs. of 8 variables

**Files** Plots Packages Help Viewer Presentation

Home

- rhitory 1 KB Nov 2
- Custom Office Templates
- Database1.acddb 992 KB Oct 1
- desktop.ini 402 B Aug 2
- disease\_diagnosis.csv 189.1 KB Dec 1
- GIS Database
- ISExpress
- loan\_approval.csv 113.1 KB Nov 2
- My Music
- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- S103 Cn 1stpdf 2.6 MB Nov 15
- scala for DS
- Sound Recordings

**Console** Terminal Background Jobs

```
R • R452 - C:\Users\luj\Desktop\
> df$Is_High_Value = ifelse(df$Price > 4000, TRUE, FALSE)
+ Status_Report = paste0("Rating: ", round(Rating, 1), " / Dis: ", discount, "%")
+ )
> print("--- Final Combined dataset ---")
[1] --- Final combined dataset ---
> print(head(final_dataset))
  Category Price Rating Stock discount Final_Price Is_High_Value Status_Report
1 <NA> 5548 1.870322 <NA> 0 5548.00 TRUE Rating: 1.9 / Dis: 0%
2 <NA> 3045 4.757798 <NA> 38 1887.90 FALSE Rating: 4.8 / Dis: 38%
3 <NA> 4004 0.000000 In Stock 0 4004.00 TRUE Rating: 0 / Dis: 0%
4 <NA> 4808 1.492085 <NA> 33 3221.36 TRUE Rating: 1.5 / Dis: 33%
5 <NA> 1817 0.000000 Out of Stock 23 1399.09 FALSE Rating: 0 / Dis: 23%
6 <NA> 3522 0.000000 <NA> 0 3522.00 TRUE Rating: 0 / Dis: 0%
>
> view(df)
> view(df_calc)
> view(df_clean)
> view(df_logic)
> view(df_text)
> |
```

Showing 1 to 18 of 4,362 entries, 8 total columns