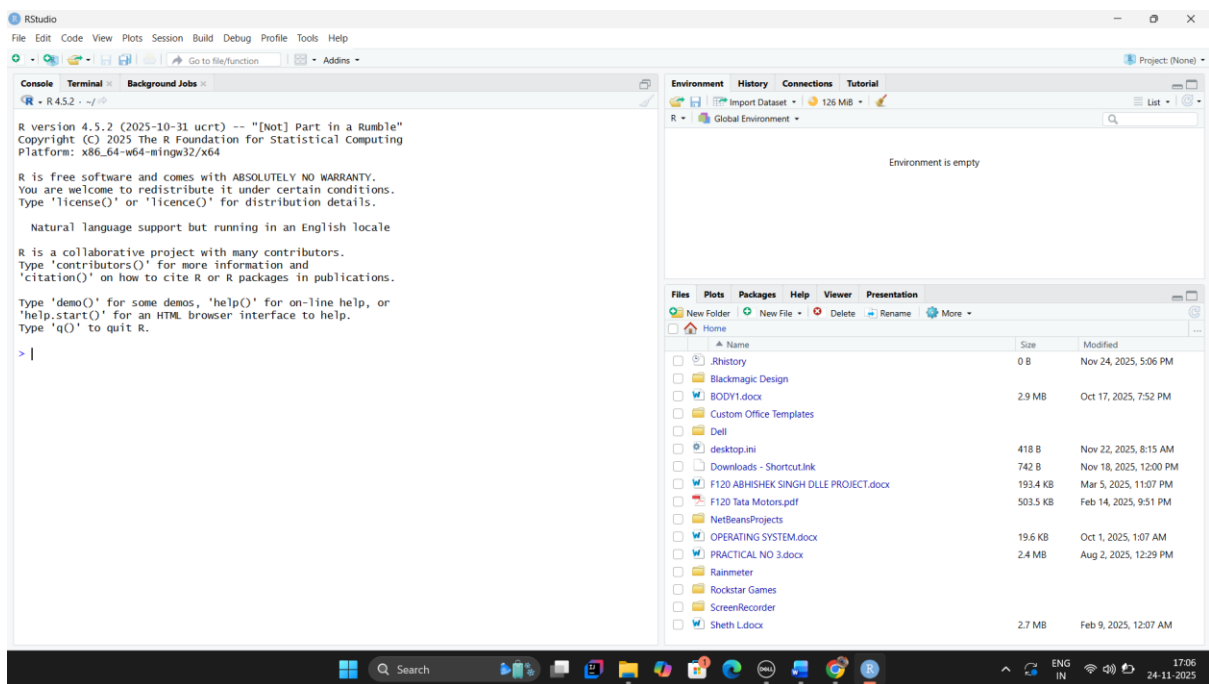


## PRACTICAL NO 2

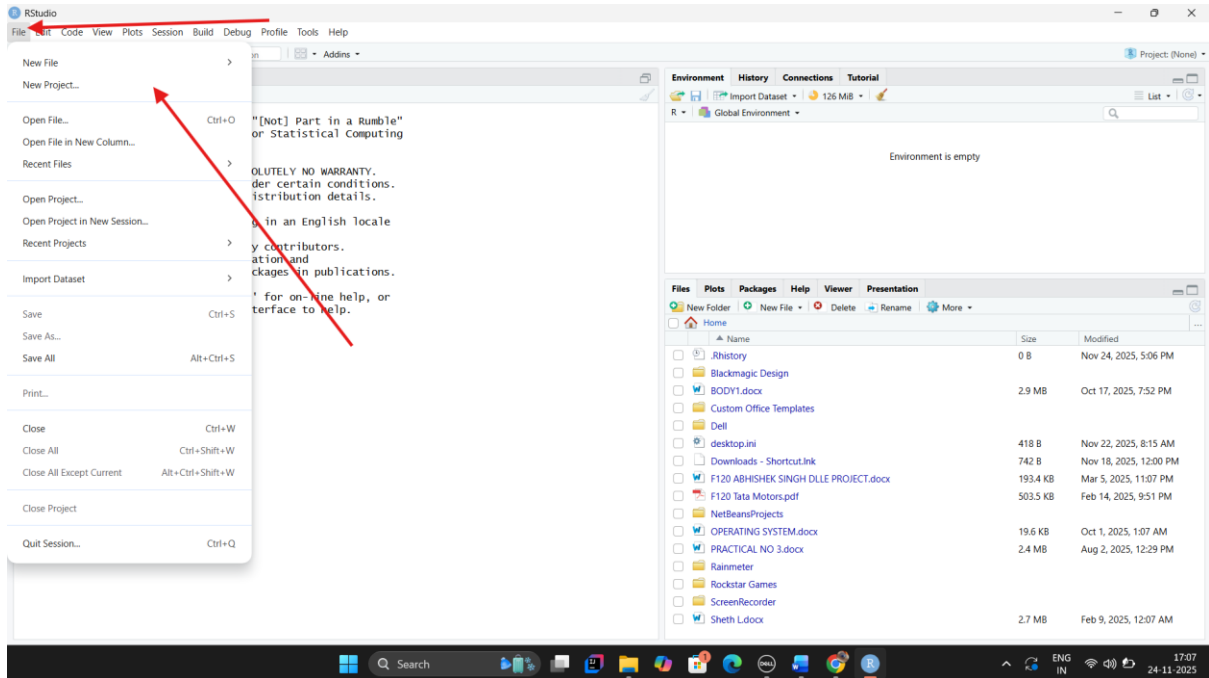
**AIM:-**Creating datasets from raw data (text files, CSV files, Excel sheets) and importing data into SAS/SPSS/R.

### STEP 1:-

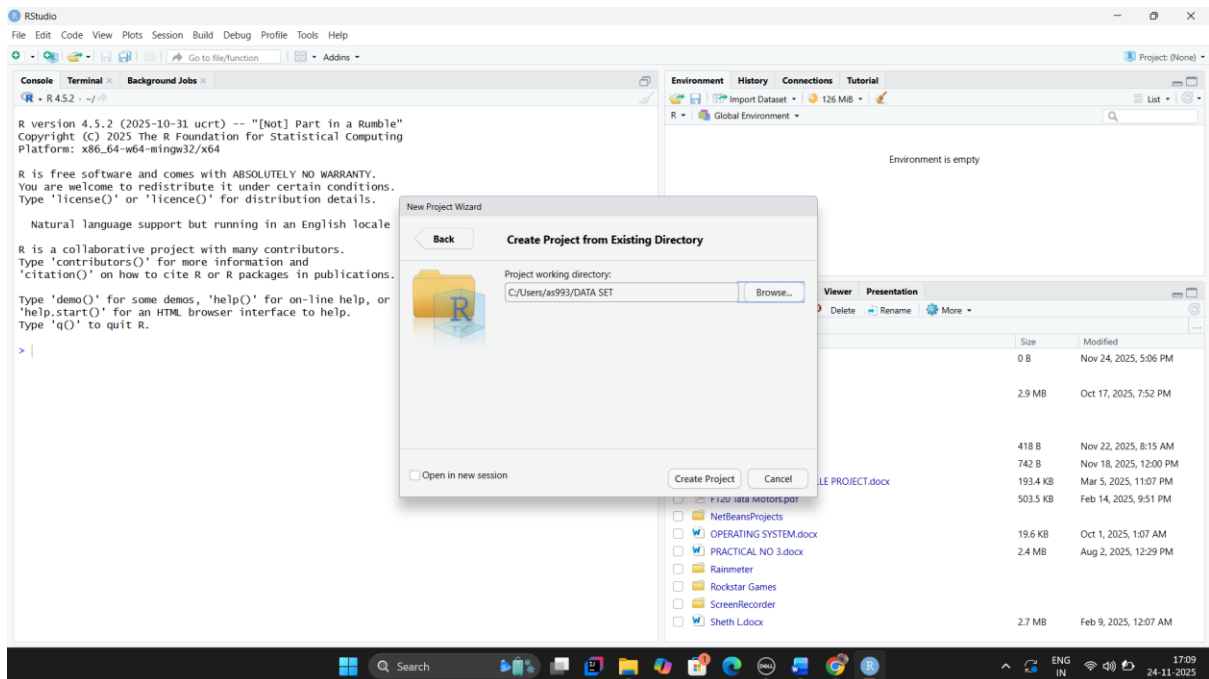
### OPEN R STUDIO



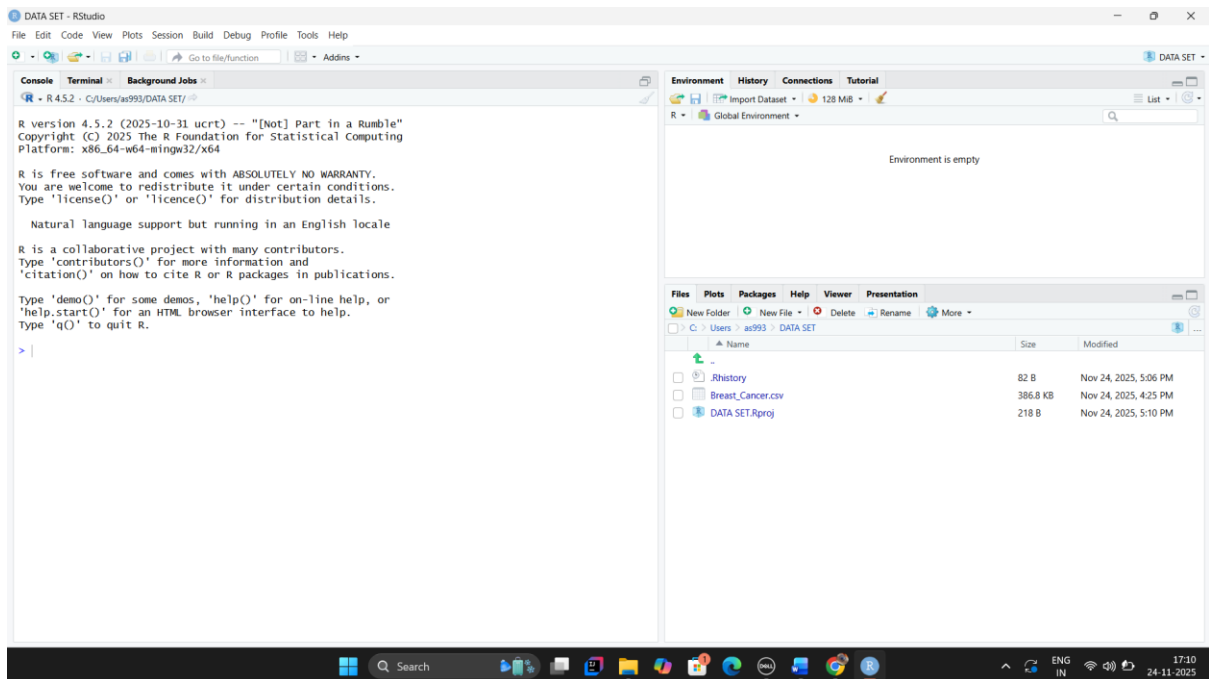
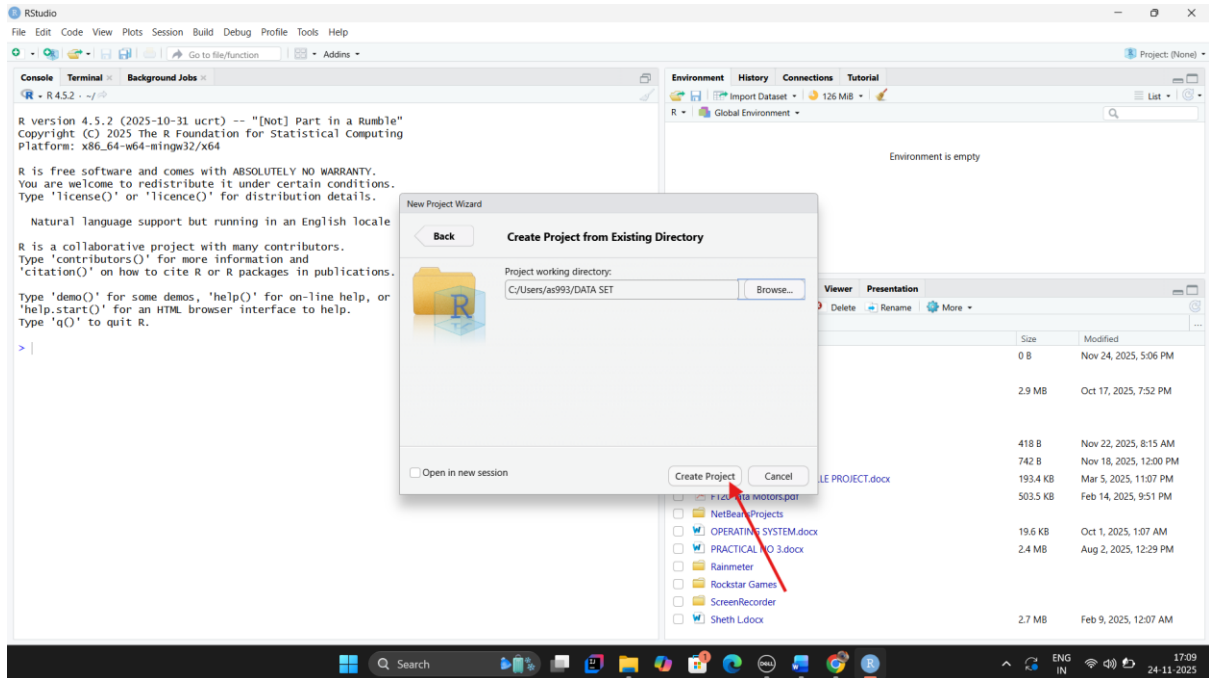
## STEP 2 CREATING A NEW PROJECT



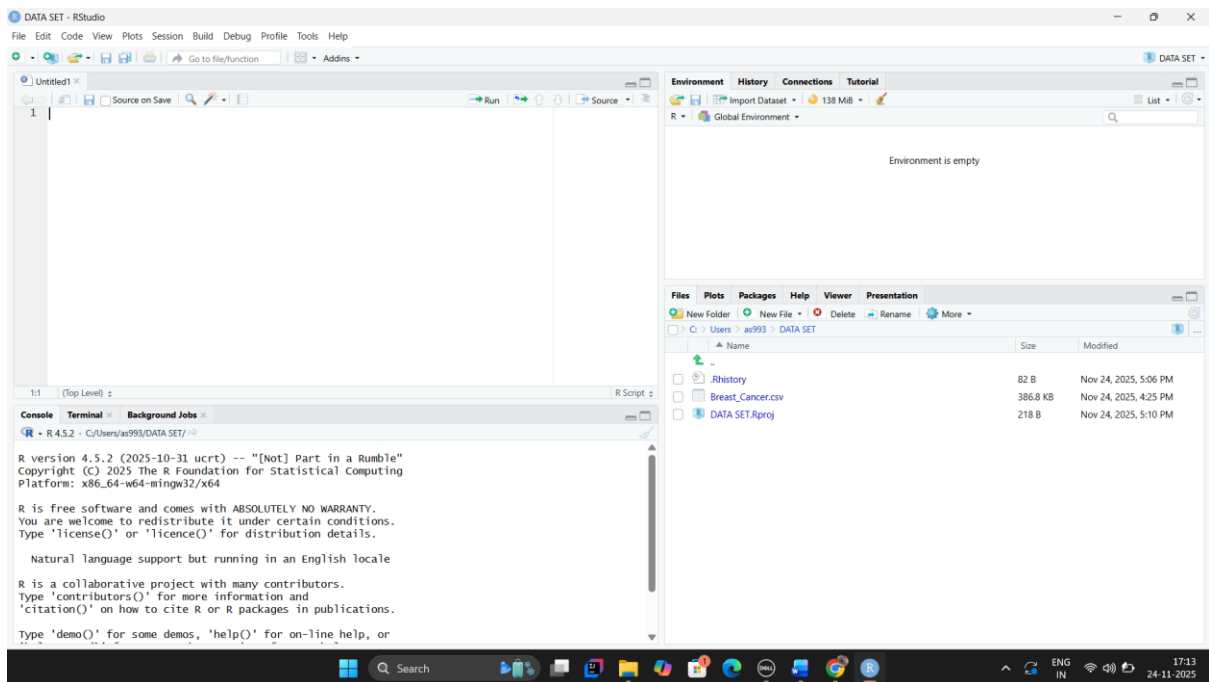
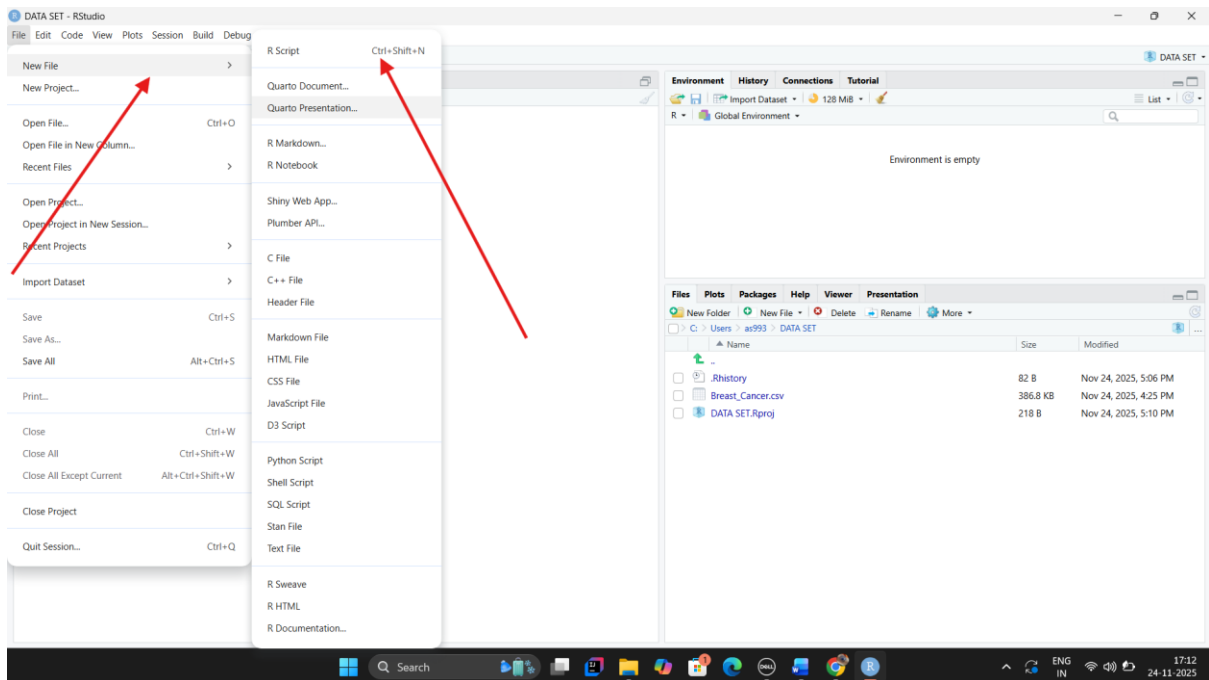
## STEP 3 CHOOSE A DIRECTORY FOR PROJECT



## STEP 4 CLICK ON CREAT PROJECT



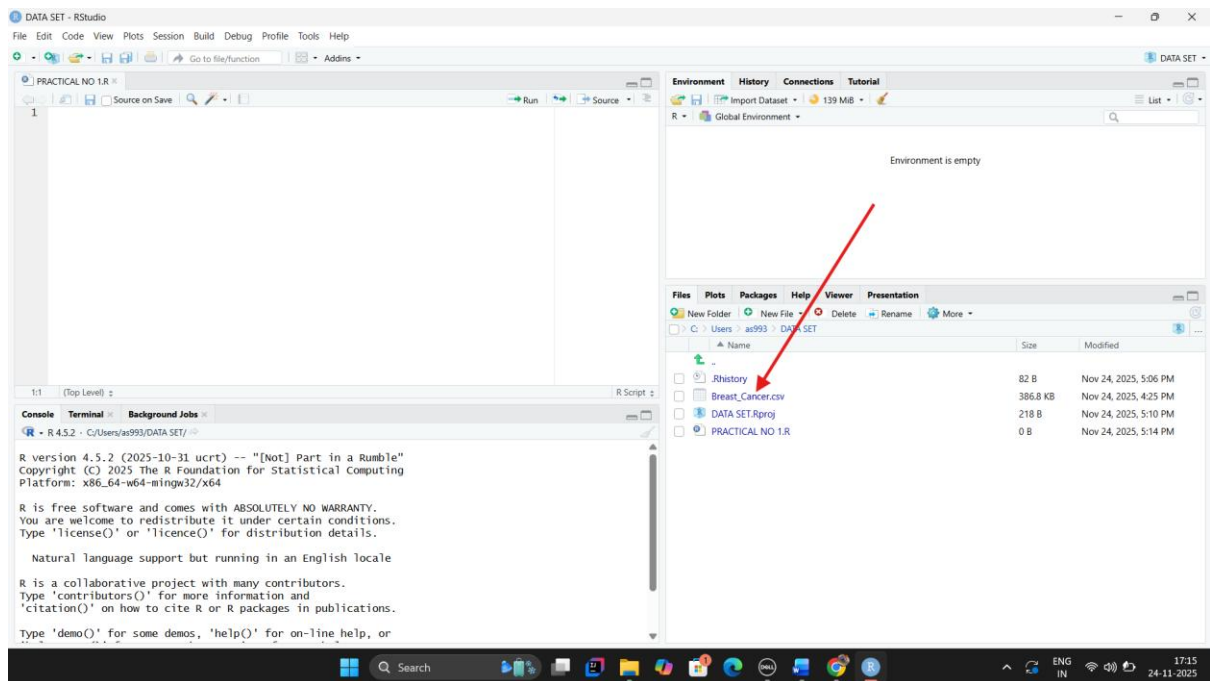
## STEP 5 CREATING A NEW FILE



## STEP 6 SAVE THE FILE AND WRITE THIS CODE ALONG NAME WITH YOUR DATA SET

EX:- `data <- read.csv("Breast_Cancer.csv")`

`View(data)`



# Sheth L.U.J. College of Arts And Sir M.V. College of Science and Commerce

## Data Analysis with SAS / SPSS / R

DATA SET - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Environment History Connections Tutorial

Import Text Data

File/URL: C:/Users/as993/DATA SET/Breast\_Cancer.csv Update

Data Preview:

Age (double)	Race (character)	Marital Status (character)	T Stage (character)	N Stage (character)	6th Stage (character)	differentiate (character)	Grade (double)	A Stage (character)	Tumor Size (double)	Estrogen Status (character)	Progesterone Status (character)
68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive
50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive
58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive
58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	18	Positive	Positive
47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	41	Positive	Positive
51	White	Single	T1	N1	IIA	Moderately differentiated	2	Regional	20	Positive	Positive
51	White	Married	T1	N1	IIA	Well differentiated	1	Regional	8	Positive	Positive
40	White	Married	T2	N1	IIB	Moderately differentiated	2	Regional	30	Positive	Positive
40	White	Divorced	T4	N3	IIIC	Poorly differentiated	3	Regional	103	Positive	Positive
69	White	Married	T4	N3	IIIC	Well differentiated	1	Distant	32	Positive	Positive
68	White	Widowed	T1	N1	IIA	Moderately differentiated	2	Regional	13	Positive	Positive

Import Options:

Name: Breast\_Cancer Skip: 0

☒ First Row as Names ☒ Trim Spaces ☒ Open Data Viewer

Delimiter: Comma Escape: None Quotes: Default Comment: Default

Locale: Configure... NA: Default

Code Preview:

```
library(readr)
Breast_Cancer <- read_csv("Breast_Cancer.csv")
View(Breast_Cancer)
```

Import Cancel

DONE WE IMPORTED COMMA SEPERTAED VALUE

DATA SET - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Environment History Connections Tutorial

Import Dataset 156 MB

Data

Breast\_Cancer 4024 obs. of 16 variables

data 4024 obs. of 16 variables

Files Plots Packages Help Viewer Presentation

New Folder New File Delete Rename More

C:/Users/as993/DATA SET

Name	Size	Modified
Rhistory	82 B	Nov 24, 2025, 5:06 PM
Breast_Cancer.csv	386.8 KB	Nov 24, 2025, 4:25 PM
DATA SET.Rproj	218 B	Nov 24, 2025, 5:10 PM
PRACTICAL NO 1.R	51 B	Nov 24, 2025, 5:17 PM

Console

```
> library(readr)
> Breast_Cancer <- read_csv("Breast_Cancer.csv")
Rows: 4024 Columns: 16
Column specification
Delimiter: ","
chr (11): Race, Marital Status, T Stage, N Stage, 6th Stage, differentiate, Grade, A Stage,...
dbl (5): Age, Tumor Size, Regional Node Examined, Regional Node Positive, Survival Months
i Use 'spec()' to retrieve the full column specification for this data.
i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
> View(Breast_Cancer)
> data <- read_csv("Breast_Cancer.csv")
> View(data)
>
```

ABHISHEK DINESH SINGH  
S116

# Sheth L.U.J. College of Arts And Sir M.V. College of Science and Commerce

## Data Analysis with SAS / SPSS / R

### TXT FILE

The screenshot shows the RStudio interface. The script editor on the left contains the following R code:

```
1. Definition
2. Breast cancer is a disease in which cells in the breast grow uncontrollably.
3. It usually begins in the milk ducts or lobules.
4. It is the most common cancer among women worldwide.
5.
6. 2. Causes
7. Genetic factors (BRCA1, BRCA2 mutations)
8. Increasing age, especially after 40
9. Hormonal factors (early menstruation, late menopause)
10. Lifestyle factors (obesity, alcohol, lack of exercise)
11. Radiation exposure
12. 3. Symptoms
13. Lump or thickening in the breast
14. Change in breast shape or size
15. Nipple pain or unusual discharge
16. Dimpling of skin (orange-peel appearance)
17. Swelling under the arm (lymph nodes)
```

The console on the bottom left shows the execution of the following code:

```
data <- read.table("BC.txt",
  header = TRUE,
  sep = "\t",
  stringsAsFactors = FALSE)
```

The Environment pane on the right shows two data objects:

- Breast\_Cancer: 4024 obs. of 16 variables
- data: 52 obs. of 1 variable

The Files pane on the bottom right shows the file structure of the project, including files like .Rhistory, BC.txt, Breast\_Cancer.csv, DATA SET.Rproj, and PRACTICAL NO 1.R.

### XLSX FILE

The screenshot shows the RStudio interface with the 'Import Excel Data' dialog box open. The 'File/URL' field contains the path 'C:/Users/as993/DATA SET/SALARY.xlsx'. The 'Data Preview' section shows a table with the following data:

ID	Name	Salary (INR)
1	Aarav Sharma	41910
2	Vivaan Patel	28493
3	Aditya Singh	66230
4	Krishna Mehta	110054
5	Ananya Sharma	91294
6	Ishaan Verma	116024
7	Riya Gupta	114099
8	Saanvi Reddy	73738
9	Arjun Yadav	116568
10	Kabir Nair	36061
11	Neha Kaur	47748
12	Tanisha Das	16586

The 'Import Options' section shows the following settings:

- Name: SALARY
- Sheet: Default
- Range: A1:D10
- Max Rows: 0
- First Row as Names: ☒
- Open Data Viewer: ☒

The 'Code Preview' section shows the following R code:

```
library(readxl)
SALARY <- read_excel("SALARY.xlsx")
View(SALARY)
```

The 'Import' button is highlighted at the bottom right of the dialog box.

ABHISHEK DINESH SINGH  
S116

# Sheth L.U.J. College of Arts And Sir M.V. College of Science and Commerce

## Data Analysis with SAS / SPSS / R

The screenshot displays the RStudio environment with the following components:

- Environment Panel:** Shows the loaded data frames: `Breast_Cancer` (4024 obs. of 16 variables), `data` (52 obs. of 1 variable), and `SALARY` (20 obs. of 3 variables).
- Files Panel:** Lists files in the project directory, including `.Rhistory`, `BC.txt`, `Breast_Cancer.csv`, `DATA SET.Rproj`, `PRACTICAL NO 1.R`, and `SALARY.xlsx`.
- Console:** Contains the following R code:

```
R - R 4.5.2 - C:/Users/as993/DATA SET/ > stringsAsFactors = FALSE) > > > > library(readxl) > SALARY <- read_excel("SALARY.xlsx") > View(SALARY) >
```
- Data Frame:** A table with 18 rows and 3 columns (ID, Name, Salary (INR)).

ID	Name	Salary (INR)
1	Aarav Sharma	41910
2	Vivaan Patel	28493
3	Aditya Singh	66230
4	Krishna Mehta	110054
5	Ananya Sharma	91294
6	Ishaan Verma	116024
7	Riya Gupta	114089
8	Saanvi Reddy	73738
9	Ajgun Yadav	116568
10	Kabir Nair	36061
11	Neha Kaur	47748
12	Tanisha Das	16586
13	Rohan Pillai	79748
14	Prisha Iyer	46755
15	Lakshay Soni	29495
16	Harshit Rana	89503
17	Meera Bansal	108951
18	Rahul Malhotra	74774