

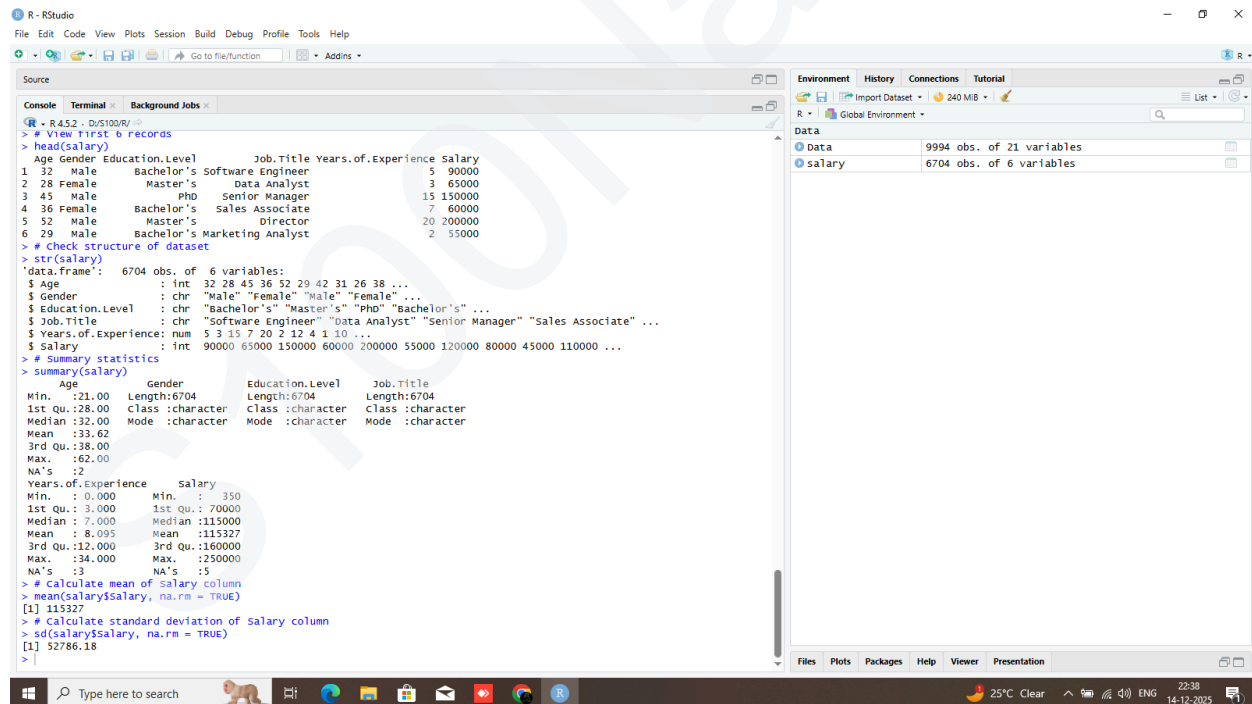
MVLU COLLEGE

Practical no :- 01 (Mod 2)

Aim :- Generating descriptive statistics using summary() or describe() ®

Code :-

```
install.packages(c("psych", "dplyr", "tidyverse"))
library(psych)
library(dplyr)
library(tidyverse)
# Load the dataset
salary <- read.csv("C:/Users/Arvind/Downloads/Salary_Data.csv")
# View first 6 records
head(salary)
# Check structure of dataset
str(salary)
# Summary statistics
summary(salary)
# Calculate mean of Salary column
mean(salary$Salary, na.rm = TRUE)
# Calculate standard deviation of Salary column
sd(salary$Salary, na.rm = TRUE)
```



The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains the R script code for loading the dataset, viewing records, checking structure, and calculating summary statistics.
- Console:** Shows the output of the R commands, including the first 6 rows of the dataset, the structure of the data frame, and the summary statistics for each variable.
- Environment:** Lists the objects in the global environment: 'Data' (9994 obs. of 21 variables) and 'salary' (6704 obs. of 6 variables).
- Terminal:** Shows the command prompt for the R session.
- Background Jobs:** A tab for monitoring background processes.
- Files:** A list of files in the current project directory.
- Plots:** A tab for viewing plots.
- Packages:** A list of installed and available packages.
- Help:** A tab for accessing R documentation.
- Viewer:** A tab for viewing plots and other visualizations.
- Presentation:** A tab for presenting slides.

The console output for the summary statistics is as follows:

```
> # Summary statistics
> summary(salary)
  Age      Gender      Education.Level      Job.Title      Years.of.Experience      Salary
Min.   :21.00   Length:6704   Length:6704   Length:6704   Length:6704   Length:6704
1st Qu.:28.00   Class :character   Class :character   Class :character   Class :character   Class :character
Median :32.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mean   :33.62
3rd Qu.:38.00
Max.   :62.00
NA's   :2
Years.of.Experience      Salary
Min.   :0.000   Min.   : 350
1st Qu.:3.000   1st Qu.: 70000
Median :7.000   Median :115000
Mean   :8.095   Mean   :115327
3rd Qu.:12.000   3rd Qu.:160000
Max.   :34.000   Max.   :250000
NA's   :3       NA's   :5
> # calculate mean of Salary column
> mean(salary$Salary, na.rm = TRUE)
[1] 115327
> # calculate standard deviation of Salary column
> sd(salary$Salary, na.rm = TRUE)
[1] 52786.18
>
```

S100 Nandini Pandit

Data Analysis with SAS/SPSS/R

MVLU COLLEGE

Practical N :- 02 (Mod 2)

Aim :- Generating frequency tables using table() or count() (R).

Code :-

```
library(psych)
library(dplyr)
library(tidyverse)
```

```
# Load dataset
```

```
jobs <- read.csv("C:/Users/Arvind/Downloads/Data_Science_Jobs_in_India.csv")
colnames(jobs)
```

```
# View structure
```

```
str(jobs)
```

```
job_table <- table(jobs$job_title)
job_table
```

```
# Proportion table
```

```
prop.table(job_table)
```

MVLU COLLEGE

```
R - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Source
Console Terminal Background Jobs
R - R 4.5.2 - D:\S100\RV
> library(tidyverse)
> # Load dataset
> jobs <- read.csv("C:/Users/Arvind/Downloads/Data_Science_Jobs_in_India.csv")
> View(jobs)
> # view structure
> str(jobs)
'data.frame': 1602 obs. of 8 variables:
 $ X : int 0 1 2 3 4 5 6 7 8 9 ...
 $ company_name : chr "TCS" "Accenture" "IBM" "Cognizant" ...
 $ job_title : chr "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
 $ min_experience : int 2 2 2 2 2 2 2 2 2 ...
 $ avg_salary : chr "7.8L" "12.8L" "13.4L" "9.8L" ...
 $ min_salary : chr "4.5L" "5.8L" "5.3L" "5.0L" ...
 $ max_salary : chr "16.0L" "23.0L" "25.0L" "18.0L" ...
 $ num_of_salaries : int 841 501 394 318 300 228 225 218 166 163 ...
> # Frequency table for job locations
> table(jobs$location)
< table of extent 0 >
> # Proportions
> prop.table(table(jobs$location))
numeric(0)
> colnames(jobs)
[1] "X" "company_name" "job_title" "min_experience"
[5] "avg_salary" "min_salary" "max_salary" "num_of_salaries"
> job_table <- table(jobs$job_title)
> # Load dataset
> jobs <- read.csv("C:/Users/Arvind/Downloads/Data_Science_Jobs_in_India.csv")
> # Display frequency table
> job_table

      Business Analyst      Data Analyst      Data Architect
      188                  187                  50
      Data Engineer      Data Scientist Machine Learning Engineer
      188                  188                  59
      Senior Business Analyst Senior Data Analyst Senior Data Engineer
      187                  187                  183
      Senior Data Scientist
      185
> # Proportion table
> prop.table(job_table)

      Business Analyst      Data Analyst      Data Architect
0.11735331      0.11672909      0.03121099

      Data Engineer      Data Scientist Machine Learning Engineer
0.11735331      0.11735331      0.03682896
      Senior Business Analyst Senior Data Analyst Senior Data Engineer
0.11672909      0.11672909      0.11423221
      Senior Data Scientist
0.11548065

Environment History Connections Tutorial
R - Global Environment
Data
Data 9994 obs. of 21 variables
jobs 1602 obs. of 8 variables
salary 6704 obs. of 6 variables
Values
company_table 'table' int [1:642(1d)] 2 1 1 5 1 3 1 10 2 2 ...
job_table 'table' int [1:10(1d)] 188 187 50 188 188 59 187...
```

```
R - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Source
Console Terminal Background Jobs
R - R 4.5.2 - D:\S100\RV
> str(jobs)
'data.frame': 1602 obs. of 8 variables:
 $ X : int 0 1 2 3 4 5 6 7 8 9 ...
 $ company_name : chr "TCS" "Accenture" "IBM" "Cognizant" ...
 $ job_title : chr "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
 $ min_experience : int 2 2 2 2 2 2 2 2 2 ...
 $ avg_salary : chr "7.8L" "12.8L" "13.4L" "9.8L" ...
 $ min_salary : chr "4.5L" "5.8L" "5.3L" "5.0L" ...
 $ max_salary : chr "16.0L" "23.0L" "25.0L" "18.0L" ...
 $ num_of_salaries : int 841 501 394 318 300 228 225 218 166 163 ...
> # Frequency table for job locations
> table(jobs$location)
< table of extent 0 >
> # Proportions
> prop.table(table(jobs$location))
numeric(0)
> colnames(jobs)
[1] "X" "company_name" "job_title" "min_experience"
[5] "avg_salary" "min_salary" "max_salary" "num_of_salaries"
> job_table <- table(jobs$job_title)
> # Load dataset
> jobs <- read.csv("C:/Users/Arvind/Downloads/Data_Science_Jobs_in_India.csv")
> # Display frequency table
> job_table

      Business Analyst      Data Analyst      Data Architect
      188                  187                  50
      Data Engineer      Data Scientist Machine Learning Engineer
      188                  188                  59
      Senior Business Analyst Senior Data Analyst Senior Data Engineer
      187                  187                  183
      Senior Data Scientist
      185
> # Proportion table
> prop.table(job_table)

      Business Analyst      Data Analyst      Data Architect
0.11735331      0.11672909      0.03121099
      Data Engineer      Data Scientist Machine Learning Engineer
0.11735331      0.11735331      0.03682896
      Senior Business Analyst Senior Data Analyst Senior Data Engineer
0.11672909      0.11672909      0.11423221
      Senior Data Scientist
0.11548065

Environment History Connections Tutorial
R - Global Environment
Data
Data 9994 obs. of 21 variables
jobs 1602 obs. of 8 variables
salary 6704 obs. of 6 variables
Values
company_table 'table' int [1:642(1d)] 2 1 1 5 1 3 1 10 2 2 ...
job_table 'table' int [1:10(1d)] 188 187 50 188 188 59 187...
```

MVLU COLLEGE

Practical No :- 03 (Mod 2)

Aim :- Creating cross-tabulations and two-way tables using table() (R).

CODE :-

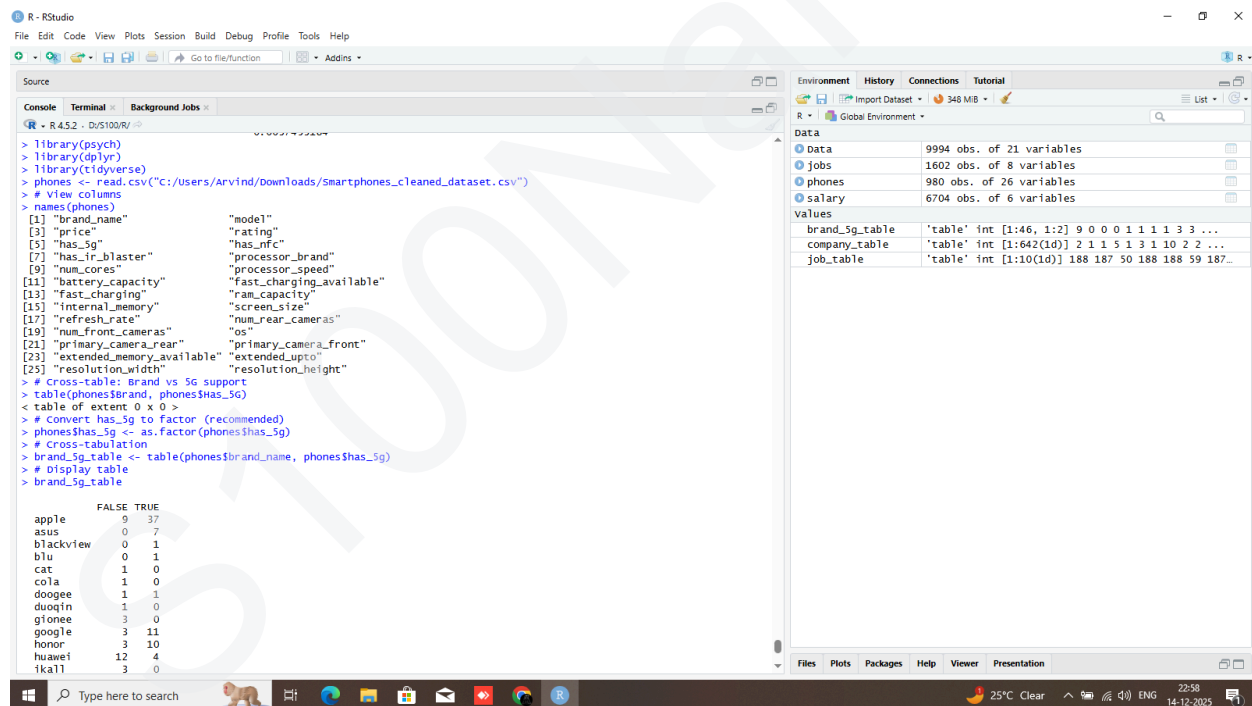
```
library(psych)
library(dplyr)
library(tidyverse)
```

```
phones <- read.csv("C:/Users/Arvind/Downloads/Smartphones_cleaned_dataset.csv")
names(phones)
```

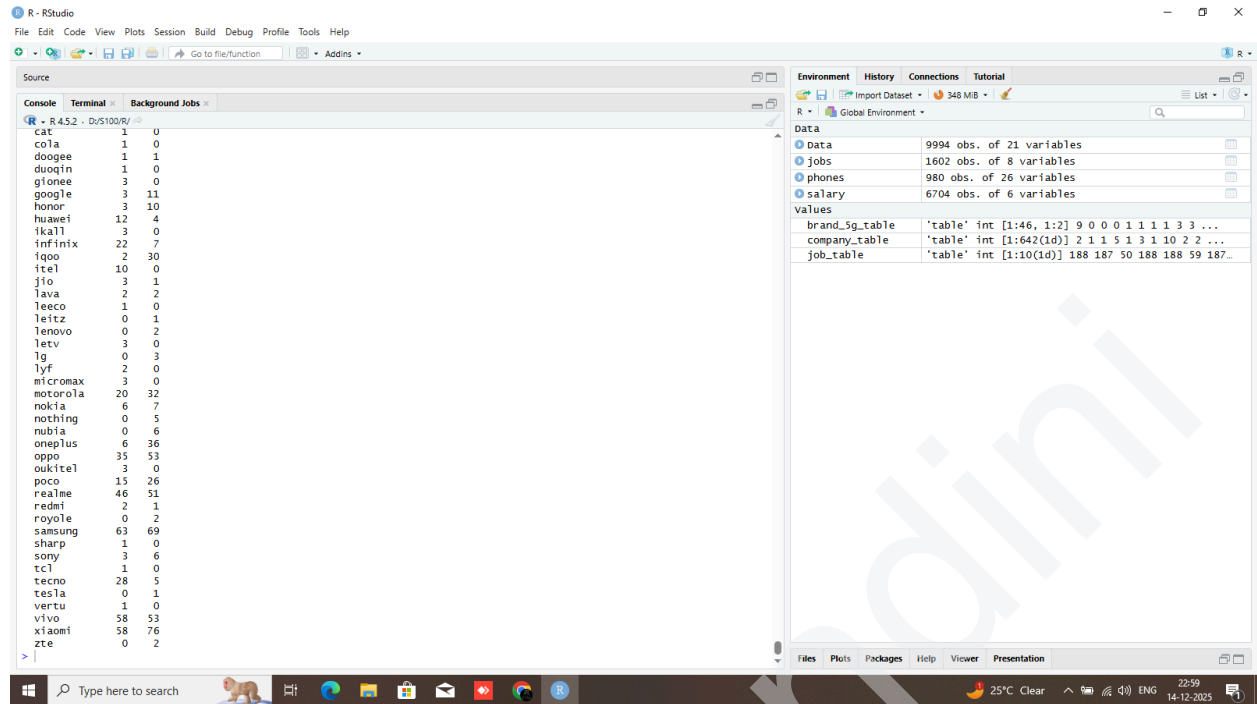
```
# Convert has_5g to factor (recommended)
phones$has_5g <- as.factor(phones$has_5g)
```

```
# Cross-tabulation
brand_5g_table <- table(phones$brand_name, phones$has_5g)
```

```
# Display table
brand_5g_table
```



MVLU COLLEGE



Practical No :- 04 (Mod 2)

Aim :- Performing one-sample t-tests using `t.test()` (R).

Code :-

```
library(psych)
library(dplyr)
library(tidyverse)
```

Load dataset

```
lang <- read.csv("C:/Users/Arvind/Downloads/programming language trend over time.csv")
names(lang)
```

View data

```
head(lang)
```

```
lang_clean <- na.omit(lang[, c("Python", "Java")])
```

Compare popularity of Python vs Java using independent t-test

```
t_test_result <- t.test(lang_clean$Python, lang_clean$Java)
```

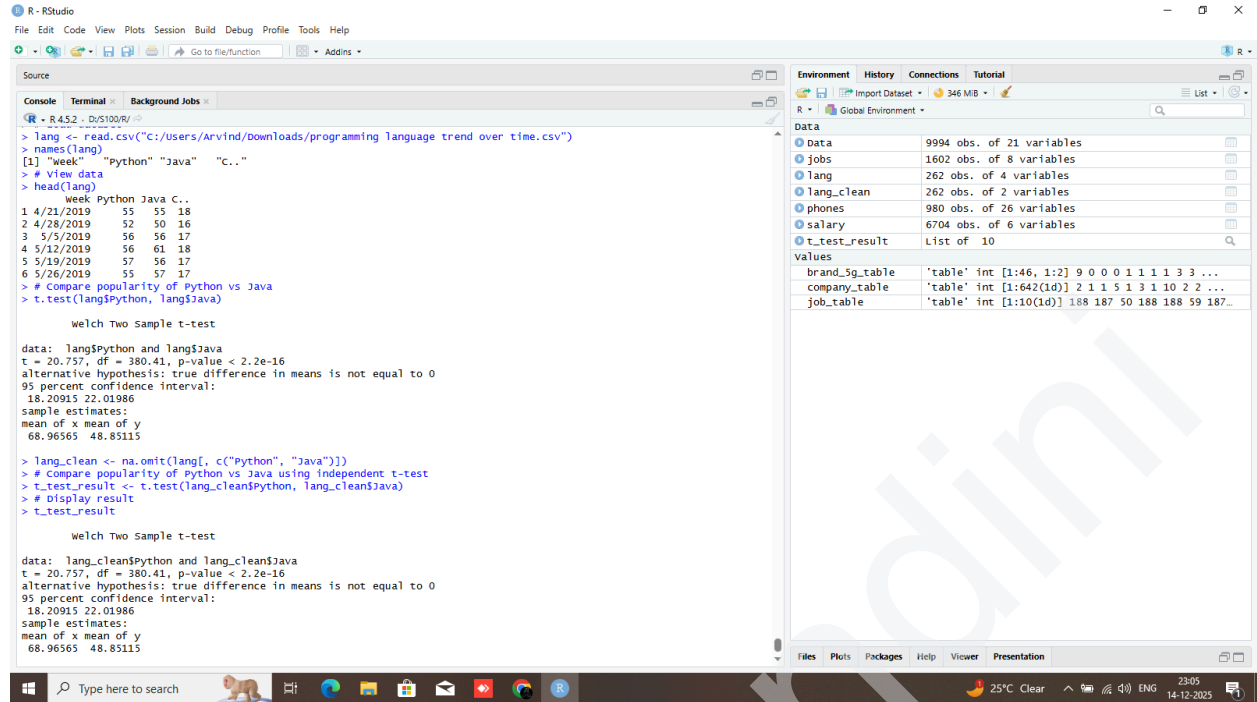
Display result

```
t_test_result
```

S100 Nandini Pandit

Data Analysis with SAS/SPSS/R

MVLU COLLEGE



The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for reading a CSV file, viewing data, and performing t-tests.
- Console:** Shows the output of the R code, including data structure, t-test results, and sample estimates.
- Environment:** Lists loaded data objects and their dimensions.
- Taskbar:** Shows the Windows taskbar with various application icons and system clock.

```
> lang <- read.csv("c:/users/arvind/downloads/programming language trend over time.csv")
> names(lang)
[1] "week" "python" "java" "c.."
> # View data
> head(lang)
  week python java c..
1 4/21/2019    55    55  18
2 4/28/2019    52    50  16
3 5/5/2019    56    56  17
4 5/12/2019    56    61  18
5 5/19/2019    57    56  17
6 5/26/2019    55    57  17
> # Compare popularity of Python vs Java
> t.test(lang$python, lang$java)

welch Two Sample t-test

data: lang$python and lang$java
t = 20.757, df = 380.41, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.20915 22.01986
sample estimates:
mean of x mean of y
 68.96565  48.85115

> lang_clean <- na.omit(lang[, c("python", "java")])
> # Compare popularity of Python vs Java using independent t-test
> t_test_result <- t.test(lang_clean$python, lang_clean$java)
> # Display result
> t_test_result

welch Two Sample t-test

data: lang_clean$python and lang_clean$java
t = 20.757, df = 380.41, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.20915 22.01986
sample estimates:
mean of x mean of y
 68.96565  48.85115
```

Environment Panel:

Object	Class	Dimensions
Data	data.frame	9994 obs. of 21 variables
jobs	data.frame	1602 obs. of 8 variables
lang	data.frame	262 obs. of 4 variables
lang_clean	data.frame	262 obs. of 2 variables
phones	data.frame	980 obs. of 26 variables
salary	data.frame	6704 obs. of 6 variables
t_test_result	list	List of 10

Taskbar: 25°C Clear, 23:05, 14-12-2025

MVLU COLLEGE

Practical No :- 05 (Mod 2)

Aim : Performing independent two-sample t-tests using t.test() with grouping (R).

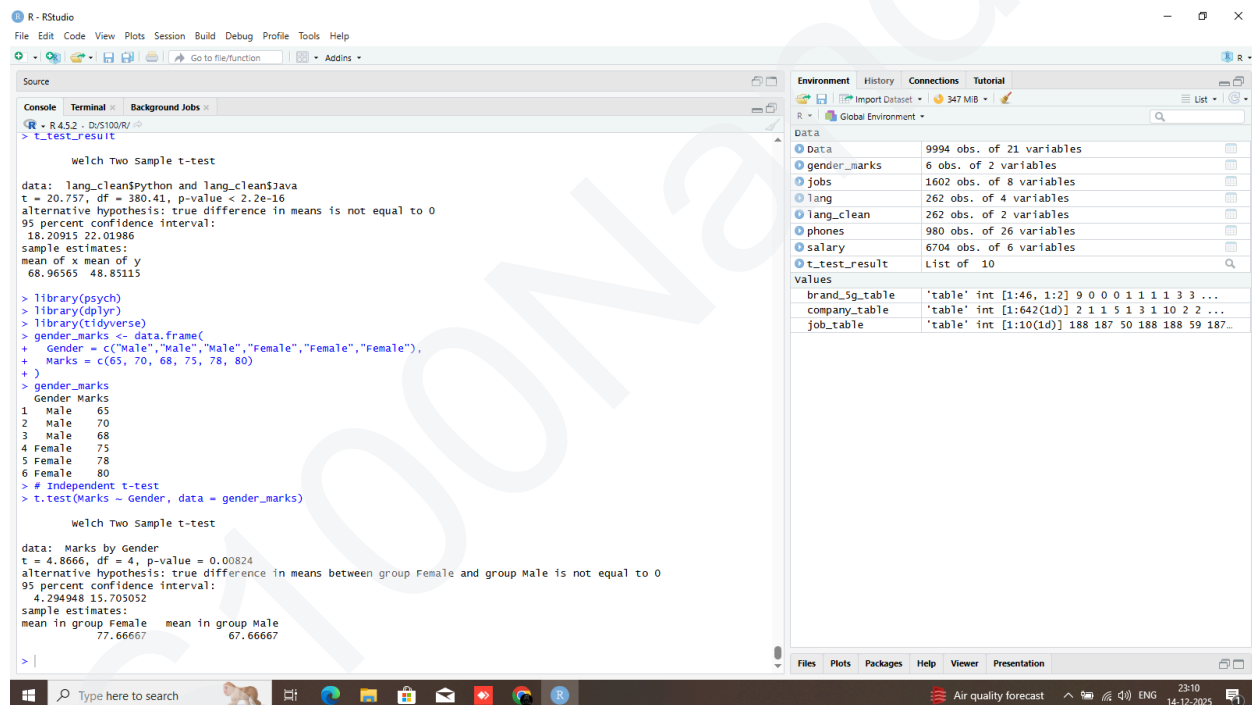
Code :-

```
library(psych)
library(dplyr)
library(tidyverse)
```

```
gender_marks <- data.frame(
  Gender = c("Male","Male","Male","Female","Female","Female"),
  Marks = c(65, 70, 68, 75, 78, 80)
)
```

Independent t-test

```
t.test(Marks ~ Gender, data = gender_marks)
```



S100 Nandini Pandit

Data Analysis with SAS/SPSS/R

MVLU COLLEGE

Practical No :- 06 (Mod 2)

Aim : Performing paired t-tests using `t.test(paired=TRUE)` (R).

```
library(psych)
```

```
library(dplyr)
```

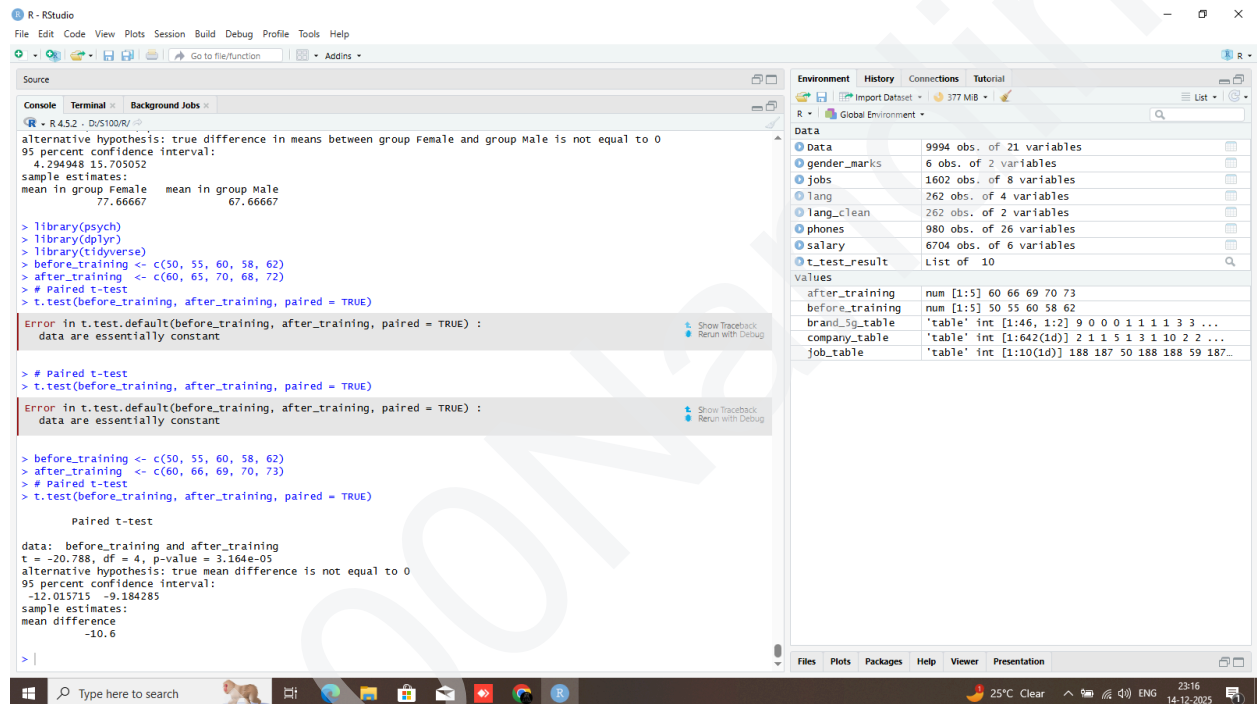
```
library(tidyverse)
```

```
before_training <- c(50, 55, 60, 58, 62)
```

```
after_training <- c(60, 66, 69, 70, 73)
```

Paired t-test

```
t.test(before_training, after_training, paired = TRUE)
```



```
R - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R 4.5.2 - D:\S100\R\
alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
95 percent confidence interval:
 4.294948 15.705052
sample estimates:
mean in group Female mean in group Male
       77.66667       67.66667

> library(psych)
> library(dplyr)
> library(tidyverse)
> before_training <- c(50, 55, 60, 58, 62)
> after_training <- c(60, 66, 69, 70, 73)
> # Paired t-test
> t.test(before_training, after_training, paired = TRUE)

Error in t.test.default(before_training, after_training, paired = TRUE) :
data are essentially constant

> # Paired t-test
> t.test(before_training, after_training, paired = TRUE)

Error in t.test.default(before_training, after_training, paired = TRUE) :
data are essentially constant

> before_training <- c(50, 55, 60, 58, 62)
> after_training <- c(60, 66, 69, 70, 73)
> # paired t-test
> t.test(before_training, after_training, paired = TRUE)

Paired t-test

data: before_training and after_training
t = -20.788, df = 4, p-value = 3.164e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -12.015715 -9.184285
sample estimates:
mean difference
      -10.6

> |

Environment History Connections Tutorial
R - Global Environment
Data
Data 9994 obs. of 21 variables
gender_marks 6 obs. of 2 variables
jobs 1602 obs. of 8 variables
lang 262 obs. of 4 variables
lang_clean 262 obs. of 2 variables
phones 980 obs. of 26 variables
salary 6704 obs. of 6 variables
t_test_result List of 10
values
after_training num [1:5] 60 66 69 70 73
before_training num [1:5] 50 55 60 58 62
brand_sg_table 'table' int [1:46, 1:2] 9 0 0 0 1 1 1 1 3 3 ...
company_table 'table' int [1:642(1d)] 2 1 1 5 1 3 1 10 2 2 ...
job_table 'table' int [1:10(1d)] 188 187 50 188 188 59 187...
```

S100 Nandini Pandit

Data Analysis with SAS/SPSS/R