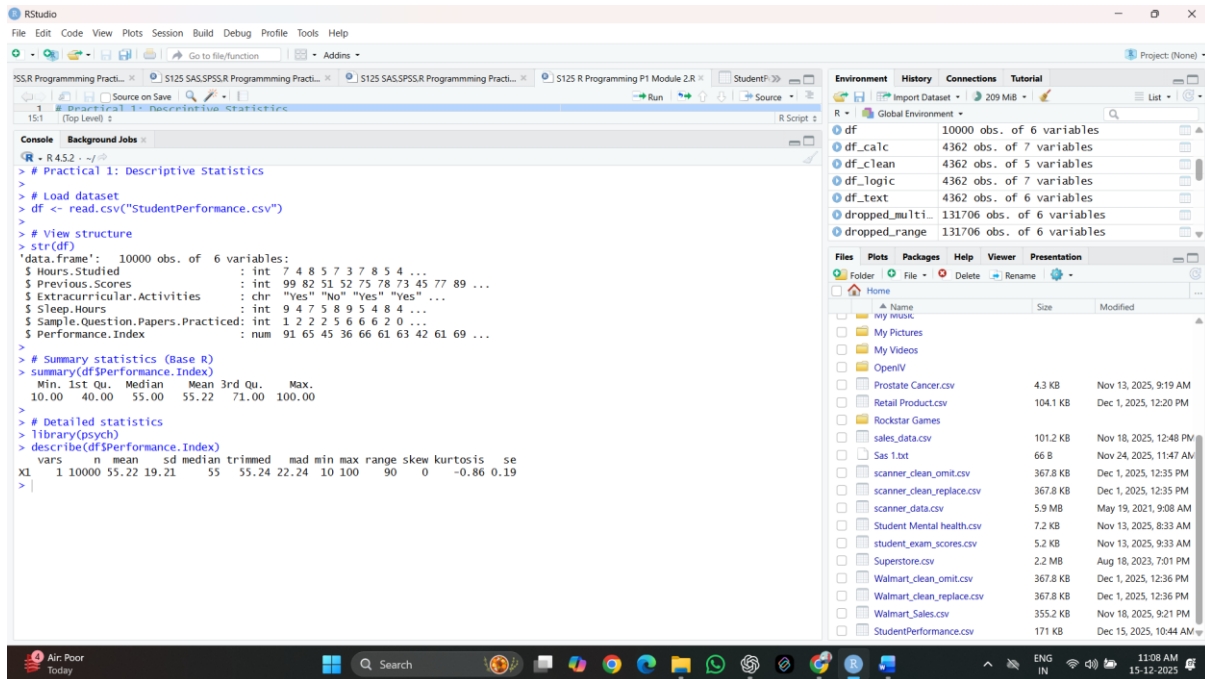


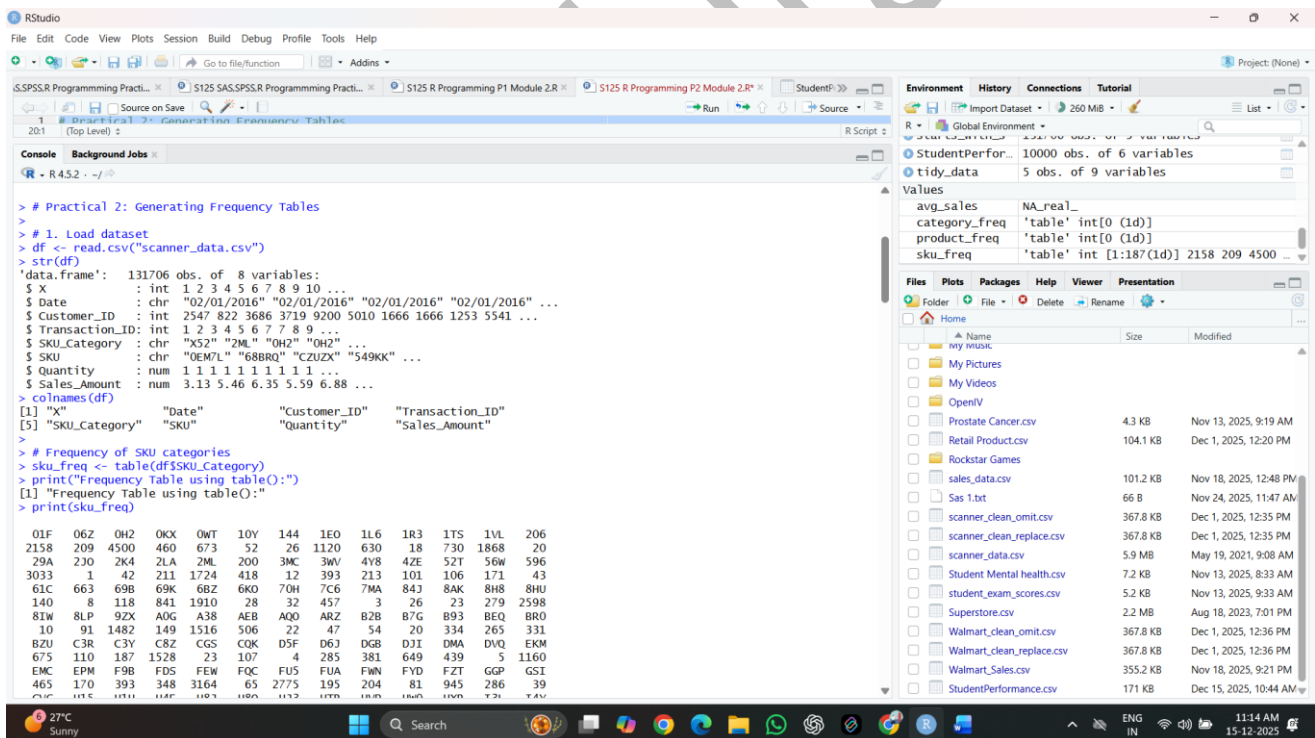
Sheth L.U.J & Sir M.V College
SAS?SPSS?R Programming
Practical No:1 to 6 Module 2

Aim 1: Generating descriptive statistics using summary() or describe() (R).



```
R - R4.5.2 - ~/R
> # Practical 1: Descriptive Statistics
>
> # Load dataset
> df <- read.csv("StudentPerformance.csv")
>
> # View structure
> str(df)
'data.frame': 10000 obs. of 6 variables:
 $ Hours.Studied : int 7 4 8 5 7 3 7 8 5 4 ...
 $ Previous.Scores : int 99 82 51 52 75 78 73 45 77 89 ...
 $ Extracurricular.Activities : chr "Yes" "No" "Yes" "Yes" ...
 $ Sleep.Hours : int 9 4 7 5 8 9 5 4 8 4 ...
 $ Sample.Question.Papers.Practiced : int 1 2 2 2 5 6 6 6 2 0 ...
 $ Performance.Index : num 91 65 45 36 66 61 63 42 61 69 ...
>
> # Summary statistics (Base R)
> summary(df$Performance.Index)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 10.00  40.00   55.00   55.22   71.00  100.00
>
> # Detailed statistics
> library(psych)
> describe(df$Performance.Index)
   vars      n mean  sd median trimmed  mad min max range skew kurtosis  se
1 1 10000 55.22 19.21   55   55.24 22.24 10 100   90   0  -0.86 0.19
>
```

Aim 2: Generating frequency tables using table() or count() (R).



```
R - R4.5.2 - ~/R
> # Practical 2: Generating Frequency Tables
>
> # 1. Load dataset
> df <- read.csv("scanner_data.csv")
>
> # Structure of the data frame
> str(df)
'data.frame': 131706 obs. of 8 variables:
 $ X : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Date : chr "02/01/2016" "02/01/2016" "02/01/2016" "02/01/2016" ...
 $ Customer_ID : int 2547 822 3686 3719 9200 5010 1666 1666 1253 5541 ...
 $ Transaction_ID : int 1 2 3 4 5 6 7 8 9 ...
 $ SKU_Category : chr "X52" "2M" "0H2" "0H2" ...
 $ SKU : chr "0E07L" "68BRQ" "CZ0ZX" "549KK" ...
 $ Quantity : num 1 1 1 1 1 1 1 1 1 ...
 $ Sales_Amount : num 3.13 5.46 6.35 5.59 6.88 ...
>
> # Frequency of SKU categories
> sku_freq <- table(df$SKU_Category)
> print("Frequency Table using table():")
[1] "Frequency Table using table():"
> print(sku_freq)
01F 06Z 0H2 0KX 0WT 10Y 144 1E0 1L6 1R3 1TS 1VL 206
2158 209 4500 460 673 52 26 1120 630 18 730 1868 20
29A 230 2K4 2LA 2ML 200 3MC 3WV 4Y8 4ZE 52T 56W 596
3033 1 42 211 1724 418 12 393 213 101 106 171 43
61C 663 698 69K 6B2 6KO 70H 7C6 7MA 843 8AK 8H8 8HU
140 8 118 841 1910 28 32 457 3 26 23 279 2598
81W 8LP 9ZX A0G A38 AEB AQO ARZ B2B B7G B93 BEQ BR0
10 91 1482 149 1516 506 22 47 54 20 334 265 331
B2U C3R C3V C8Z CGS CQK D5F D6J DGB DJI DMA DVO EKM
675 110 187 1528 23 107 4 285 381 649 439 5 1160
EMC EPM F9B FDS FEW FQC FUS FUA FWN FYD FZT GGP GSI
465 170 393 348 3164 65 2775 195 204 81 945 286 39
>
```

Sheth L.U.J & Sir M.V College

SAS?SPSS?R Programming

Practcal No:1 to 6 Module 2

The screenshot shows the RStudio interface with the following components:

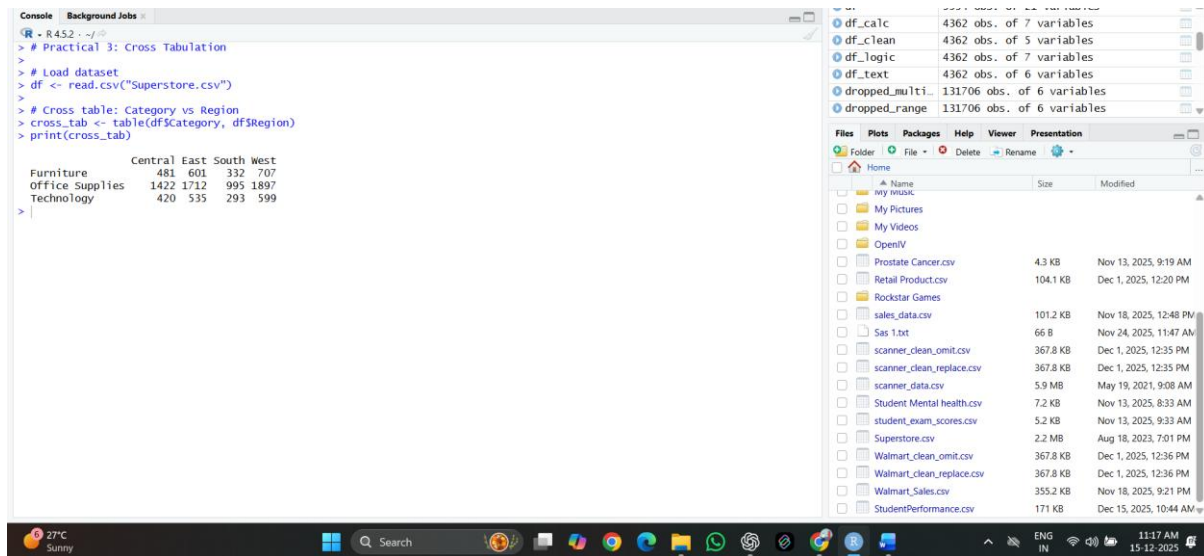
- Console:** Displays the R script execution output. It shows the loading of the `dplyr` library and the generation of a frequency table using `count()`. The output is a table with two columns: `SKU_Category` and `n`.
- Environment:** Shows the loaded objects: `StudentPerfor...` (10000 obs. of 6 variables) and `tidy_data` (5 obs. of 9 variables).
- Files:** Lists the files in the current directory, including `Prostate Cancer.csv`, `Retail Product.csv`, `Rockstar Games`, `sales_data.csv`, `Sas 1.txt`, `scanner_clean_omit.csv`, `scanner_clean_replace.csv`, `scanner_data.csv`, `Student Mental health.csv`, `student_exam_scores.csv`, `Superstore.csv`, `Walmart_clean_omit.csv`, `Walmart_clean_replace.csv`, `Walmart_Sales.csv`, and `StudentPerformance.csv`.

The screenshot shows the RStudio interface with the following components:

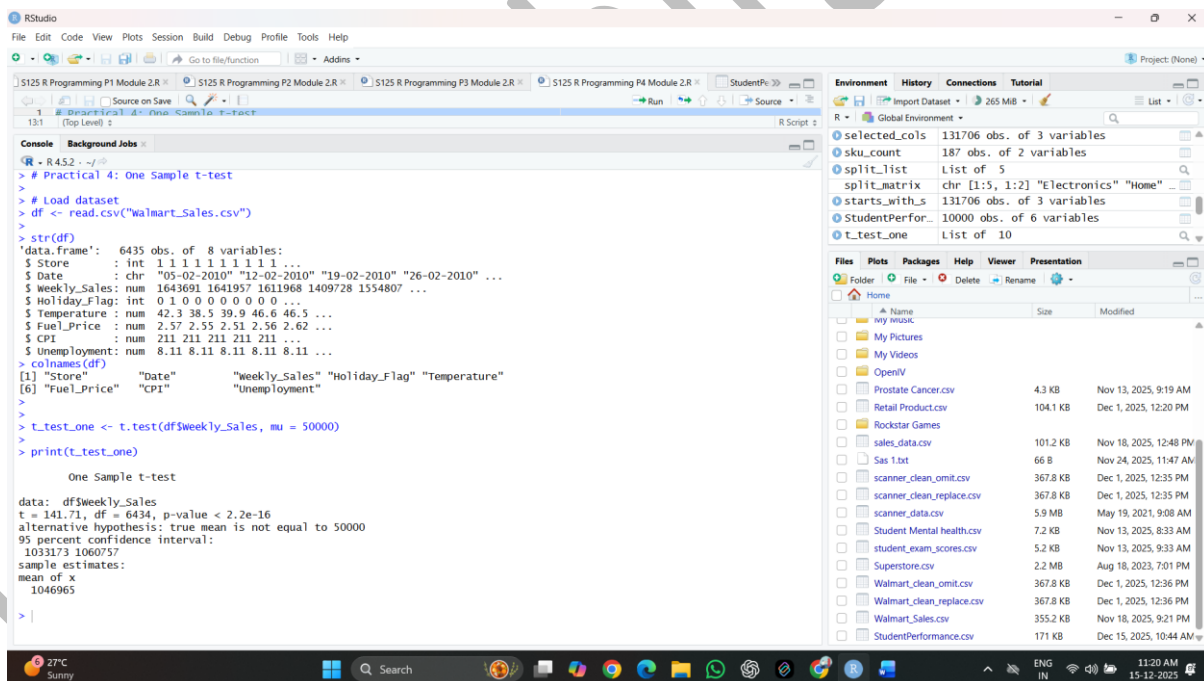
- Console:** Displays the R script execution output. It shows the loading of the `dplyr` library and the generation of a frequency table using `count()`. The output is a table with two columns: `SKU_Category` and `n`.
- Environment:** Shows the loaded objects: `StudentPerfor...` (10000 obs. of 6 variables) and `tidy_data` (5 obs. of 9 variables).
- Files:** Lists the files in the current directory, including `Prostate Cancer.csv`, `Retail Product.csv`, `Rockstar Games`, `sales_data.csv`, `Sas 1.txt`, `scanner_clean_omit.csv`, `scanner_clean_replace.csv`, `scanner_data.csv`, `Student Mental health.csv`, `student_exam_scores.csv`, `Superstore.csv`, `Walmart_clean_omit.csv`, `Walmart_clean_replace.csv`, `Walmart_Sales.csv`, and `StudentPerformance.csv`.

Sheth L.U.J & Sir M.V College
SAS?SPSS?R Programming
Practical No:1 to 6 Module 2

Aim 3: Creating cross-tabulations and two-way tables using table() (R).



Aim 4: Performing one-sample t-tests using t.test() (R).



Sheth L.U.J & Sir M.V College
SAS?SPSS?R Programming
Practical No:1 to 6 Module 2

Aim 5: Performing independent two-sample t-tests using t.test() with grouping (R).

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

S125 R Programming P3 Module 2.R S125 R Programming P4 Module 2.R S125 R Programming P5 Module 2.R
1 Practical 5: Independent Two Sample T-test
13:1 (Top Level)

Console Background Jobs
R - R452 - ~/...
> # Practical 5: Independent Two Sample t-test
> # Load dataset
> df <- read.csv("Superstore.csv")
> str(df)
'data.frame': 9994 obs. of 21 variables:
 $ Row.ID : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Order.ID : chr "CA-2013-152156" "CA-2013-152156" "CA-2013-138688" "US-2012-108966" ...
 $ Order.Date : chr "09-11-2013" "09-11-2013" "13-06-2013" "11-10-2012" ...
 $ Ship.Date : chr "12-11-2013" "12-11-2013" "17-06-2013" "18-10-2012" ...
 $ Ship.Mode : chr "Second Class" "Second Class" "Second Class" "Standard Class" ...
 $ Customer.ID : chr "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer.Name : chr "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean O'Donnell" ...
 $ Segment : chr "Consumer" "Consumer" "Corporate" "Consumer" ...
 $ Country : chr "United States" "United States" "United States" "United States" ...
 $ City : chr "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
 $ State : chr "Kentucky" "Kentucky" "California" "Florida" ...
 $ Postal.Code : int 42420 42420 90036 33311 33311 90032 90032 90032 90032 ...
 $ Region : chr "South" "South" "West" "South" ...
 $ Product.ID : chr "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-10000240" "FUR-TA-10000577" ...
 $ Category : chr "Furniture" "Furniture" "Office Supplies" "Furniture" ...
 $ Sub.Category : chr "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product.Name : chr "Bush Somerset Collection Bookcase" "Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back" "Sel
f-Adhesive Address Labels for Typewriters by Universal" "Bretford CR4500 Series Slim Rectangular Table" ...
 $ Sales : num 262 731.9 14.6 957.6 22.4 ...
 $ Quantity : int 2 3 2 5 2 7 4 6 3 5 ...
 $ Discount : num 0 0 0 0.45 0.2 0 0 0.2 0.2 0 ...
 $ Profit : num 41.91 219.58 6.87 -383.03 2.52 ...
> unique(df$Segment)
[1] "Consumer" "Corporate" "Home Office"
>
> df_two <- subset(df, Segment %in% c("Consumer", "Corporate"))
> t_test_two <- t.test(Sales ~ Segment, data = df_two)
> print(t_test_two)

Welch Two Sample t-test
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

S125 R Programming P3 Module 2.R S125 R Programming P4 Module 2.R S125 R Programming P5 Module 2.R
1 Practical 5: Independent Two Sample T-test
13:1 (Top Level)

Console Background Jobs
R - R452 - ~/...
> # Practical 5: Independent Two Sample t-test
> # Load dataset
> df <- read.csv("Superstore.csv")
> str(df)
'data.frame': 9994 obs. of 21 variables:
 $ Row.ID : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Order.ID : chr "CA-2013-152156" "CA-2013-152156" "CA-2013-138688" "US-2012-108966" ...
 $ Order.Date : chr "09-11-2013" "09-11-2013" "13-06-2013" "11-10-2012" ...
 $ Ship.Date : chr "12-11-2013" "12-11-2013" "17-06-2013" "18-10-2012" ...
 $ Ship.Mode : chr "Second Class" "Second Class" "Second Class" "Standard Class" ...
 $ Customer.ID : chr "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
 $ Customer.Name : chr "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean O'Donnell" ...
 $ Segment : chr "Consumer" "Consumer" "Corporate" "Consumer" ...
 $ Country : chr "United States" "United States" "United States" "United States" ...
 $ City : chr "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
 $ State : chr "Kentucky" "Kentucky" "California" "Florida" ...
 $ Postal.Code : int 42420 42420 90036 33311 33311 90032 90032 90032 90032 ...
 $ Region : chr "South" "South" "West" "South" ...
 $ Product.ID : chr "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-10000240" "FUR-TA-10000577" ...
 $ Category : chr "Furniture" "Furniture" "Office Supplies" "Furniture" ...
 $ Sub.Category : chr "Bookcases" "Chairs" "Labels" "Tables" ...
 $ Product.Name : chr "Bush Somerset Collection Bookcase" "Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back" "Sel
f-Adhesive Address Labels for Typewriters by Universal" "Bretford CR4500 Series Slim Rectangular Table" ...
 $ Sales : num 262 731.9 14.6 957.6 22.4 ...
 $ Quantity : int 2 3 2 5 2 7 4 6 3 5 ...
 $ Discount : num 0 0 0 0.45 0.2 0 0 0.2 0.2 0 ...
 $ Profit : num 41.91 219.58 6.87 -383.03 2.52 ...
> unique(df$Segment)
[1] "Consumer" "Corporate" "Home Office"
>
> df_two <- subset(df, Segment %in% c("Consumer", "Corporate"))
> t_test_two <- t.test(Sales ~ Segment, data = df_two)
> print(t_test_two)

Welch Two Sample t-test

data: Sales by Segment
t = -0.74178, df = 6191.3, p-value = 0.4583
alternative hypothesis: true difference in means between group Consumer and group Corporate is not equal to 0
95 percent confidence interval:
 -36.75428 16.57497
sample estimates:
mean in group Consumer mean in group Corporate
223.7336 233.8233
```

Sheth L.U.J & Sir M.V College
SAS?SPSS?R Programming
Practcal No:1 to 6 Module 2

Aim 6: Performing paired t-tests using t.test(paired=TRUE) (R).

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

```
# Practical 6: Paired t-test
>
> # Load dataset
> df <- read.csv("Prostate Cancer.csv")
> str(df)
'data.frame': 100 obs. of 10 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ diagnosis_result : chr  "M" "M" "M" "M" ...
 $ radius   : int  23 9 21 14 9 25 16 15 19 25 ...
 $ texture  : int  12 13 27 16 19 25 26 18 24 11 ...
 $ perimeter: int  151 133 130 78 135 83 120 90 88 84 ...
 $ area     : int  954 1326 1203 386 1297 477 1040 578 520 476 ...
 $ smoothness: num  0.143 0.143 0.125 0.07 0.141 0.128 0.095 0.119 0.127 0.119 ...
 $ compactness : num  0.278 0.079 0.16 0.284 0.133 0.17 0.109 0.165 0.193 0.24 ...
 $ symmetry   : num  0.242 0.181 0.207 0.26 0.181 0.209 0.179 0.22 0.235 0.203 ...
 $ fractal_dimension: num  0.079 0.057 0.06 0.097 0.059 0.076 0.057 0.075 0.074 0.082 ...
> colnames(df)
[1] "id"           "diagnosis_result" "radius"
[4] "texture"      "perimeter"       "area"
[7] "smoothness"   "compactness"     "symmetry"
[10] "fractal_dimension"
>
> df$area_before <- df$area
> df$area_after <- df$area + rnorm(nrow(df), mean = -50, sd = 20)
>
> t_testpaired <- t.test(df$area_before,
+                        df$area_after,
+                        paired = TRUE)
> print(t_testpaired)
```

The console output shows the results of the paired t-test:

```
Paired t-test
data: df$area_before and df$area_after
t = 31.39, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 50.09707 56.85787
sample estimates:
mean difference
 53.47747
```

```
> print(t_testpaired)

Paired t-test

data: df$area_before and df$area_after
t = 31.39, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 50.09707 56.85787
sample estimates:
mean difference
 53.47747

> |
```