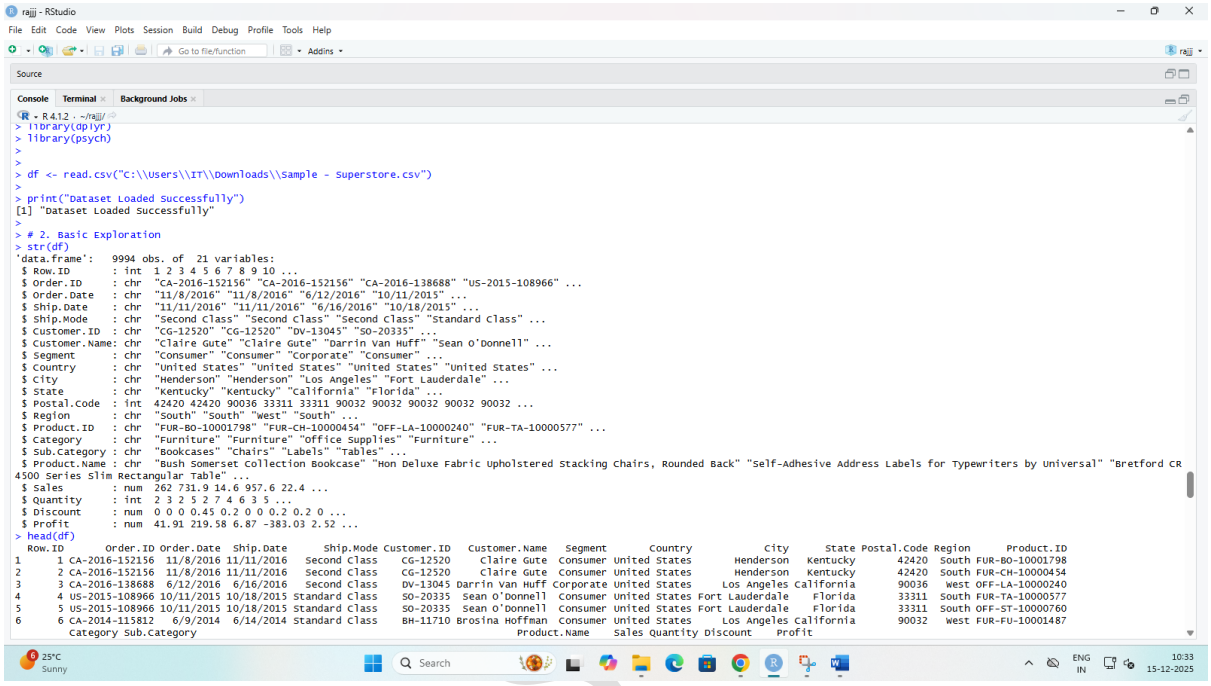
SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

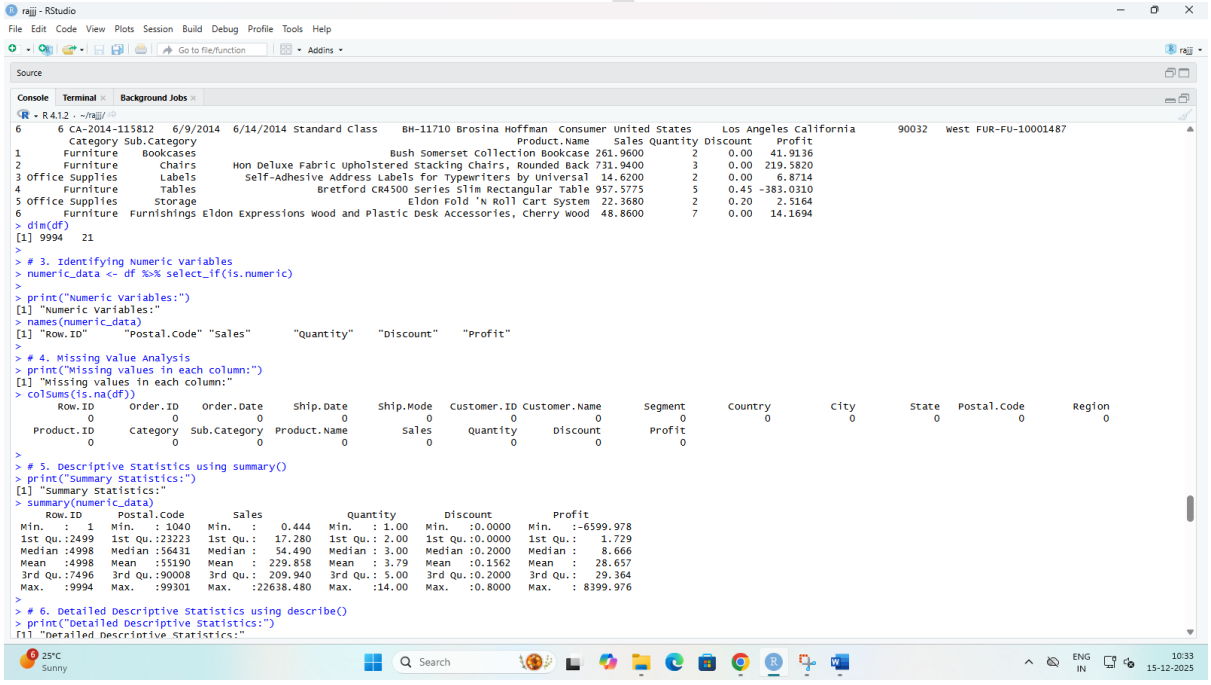
MODULE 2 - PRACTICAL – 1

AIM:- Generating descriptive statistics using summary() or describe() (R)

OUTPUT:-



```
> library(dplyr)
> library(psych)
>
>
> df <- read.csv("C:\\Users\\IT\\Downloads\\Sample - Superstore.csv")
>
> print("Dataset Loaded Successfully")
[1] "Dataset Loaded Successfully"
>
> # 2. Basic Exploration
> 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-2016-152156" "CA-2016-152156" "CA-2016-138688" "US-2015-108966" ...
 $ Order.Date   : chr  "11/8/2016" "11/8/2016" "6/12/2016" "10/11/2015" ...
 $ Ship.Date    : chr  "11/11/2016" "11/11/2016" "6/16/2016" "10/18/2015" ...
 $ 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 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" "Self-Adhesive Address Labels for Typewriters by Universal" "Bretford CR
4500 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 ...
> head(df)
  Row.ID       Order.ID Order.Date  Ship.Date      Ship.Mode Customer.ID   Customer.Name    Segment       Country           City       State Postal.Code Region      Product.ID
1      1 CA-2016-152156  11/8/2016 11/11/2016   Second Class    CG-12520     Claire Gute   Consumer United States      Henderson    Kentucky       42420  South FUR-BO-10001798
2      2 CA-2016-152156  11/8/2016 11/11/2016   Second Class    CG-12520     Claire Gute   Consumer United States      Henderson    Kentucky       42420  South FUR-CH-10000454
3      3 CA-2016-138688  6/12/2016  6/16/2016   Second Class    DV-13045 Darrin Van Huff Corporate United States    Los Angeles California       90036   West OFF-LA-10000240
4      4 US-2015-108966 10/11/2015 10/18/2015 Standard Class    SO-20335  Sean O'Donnell   Consumer United States Fort Lauderdale     Florida       33311  South FUR-TA-10000577
5      5 US-2015-108966 10/11/2015 10/18/2015 Standard Class    SO-20335  Sean O'Donnell   Consumer United States Fort Lauderdale     Florida       33311  South OFF-ST-10000760
6      6 CA-2014-115812   6/9/2014  6/14/2014 Standard Class    BH-11710 Brosina Hoffman   Consumer United States    Los Angeles California       90032   West FUR-FU-10001487
     Category Sub.Category                                                 Product.Name    Sales Quantity Discount    Profit
```



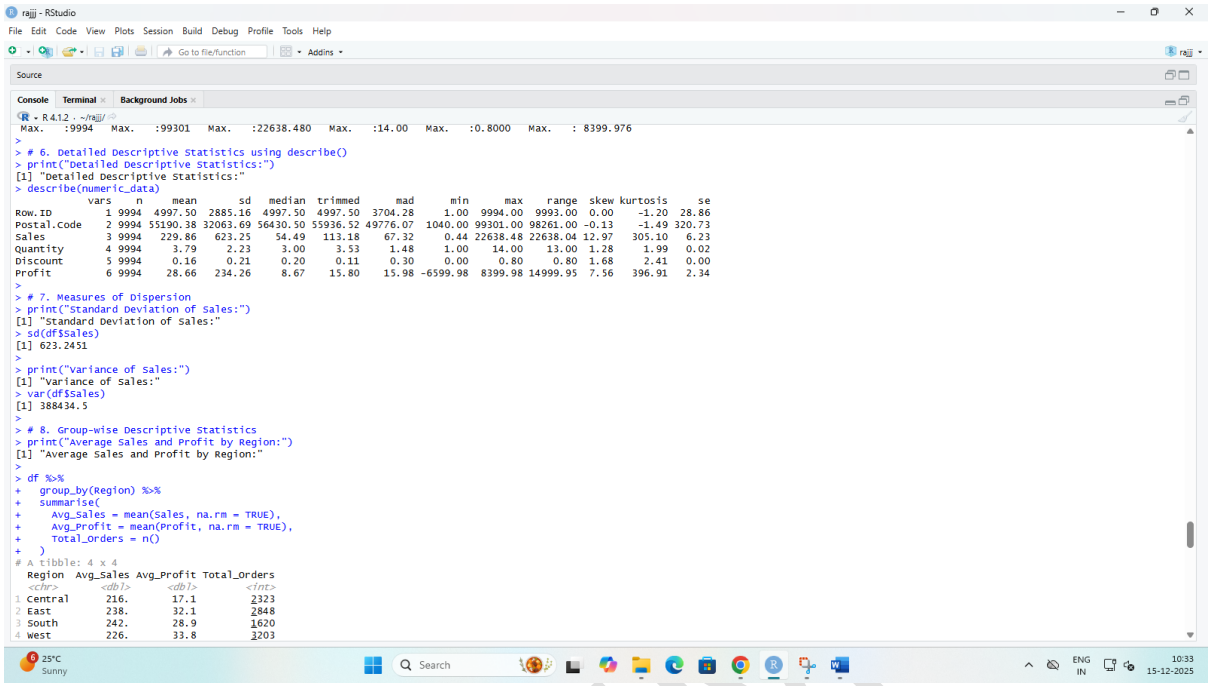
```
6      6 CA-2014-115812   6/9/2014  6/14/2014 Standard Class    BH-11710 Brosina Hoffman  Consumer United States     Los Angeles California       90032   West FUR-FU-10001487
     Category Sub.Category                                                 Product.Name    Sales Quantity Discount    Profit
1       Furniture    Bookcases                                Bush Somerset Collection Bookcase 261.9600        2     0.00   41.9136
2       Furniture       Chairs      Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.9400        3     0.00  219.5820
3 Office Supplies       Labels        Self-Adhesive Address Labels for Typewriters by Universal  14.6200        2     0.00    6.8714
4       Furniture       Tables                     Bretford CR4500 Series Slim Rectangular Table 957.5775        5     0.45 -383.0310
5 Office Supplies      Storage                                   Eldon Fold 'N Roll Cart System  22.3680        2     0.20    2.5164
6       Furniture  Furnishings Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood  48.8600        7     0.00   14.1694
> dim(df)
[1] 9994   21
>
> # 3. Identifying Numeric Variables
> numeric_data <- df %>% select_if(is.numeric)
>
> print("Numeric Variables:")
[1] "Numeric Variables:"
> names(numeric_data)
[1] "Row.ID"      "Postal.Code" "Sales"       "Quantity"    "Discount"    "Profit"
>
> # 4. Missing Value Analysis
> print("Missing values in each column:")
[1] "Missing values in each column:"
> colSums(is.na(df))
       Row.ID      Order.ID    Order.Date     Ship.Date     Ship.Mode   Customer.ID Customer.Name       Segment       Country          City         State   Postal.Code        Region
            0             0             0             0             0             0             0             0             0             0             0             0             0
   Product.ID      Category  Sub.Category  Product.Name         Sales      Quantity      Discount        Profit
            0             0             0             0             0             0             0             0
>
> # 5. Descriptive Statistics using summary()
> print("Summary Statistics:")
[1] "Summary Statistics:"
> summary(numeric_data)
     Row.ID       Postal.Code        Sales              Quantity       Discount          Profit
 Min.   :   1   Min.   : 1040   Min.   :    0.444   Min.   : 1.00   Min.   :0.0000   Min.   :-6599.978
 1st Qu.:2499   1st Qu.:23223   1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000   1st Qu.:    1.729
 Median :4998   Median :56431   Median :   54.490   Median : 3.00   Median :0.2000   Median :    8.666
 Mean   :4998   Mean   :55190   Mean   :  229.858   Mean   : 3.79   Mean   :0.1562   Mean   :   28.657
 3rd Qu.:7496   3rd Qu.:90008   3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000   3rd Qu.:   29.364
 Max.   :9994   Max.   :99301   Max.   :22638.480   Max.   :14.00   Max.   :0.8000   Max.   : 8399.976
>
> # 6. Detailed Descriptive Statistics using describe()
> print("Detailed Descriptive Statistics:")
[1] "Detailed Descriptive Statistics:"
```

SUMEET JITENDRA YADAV

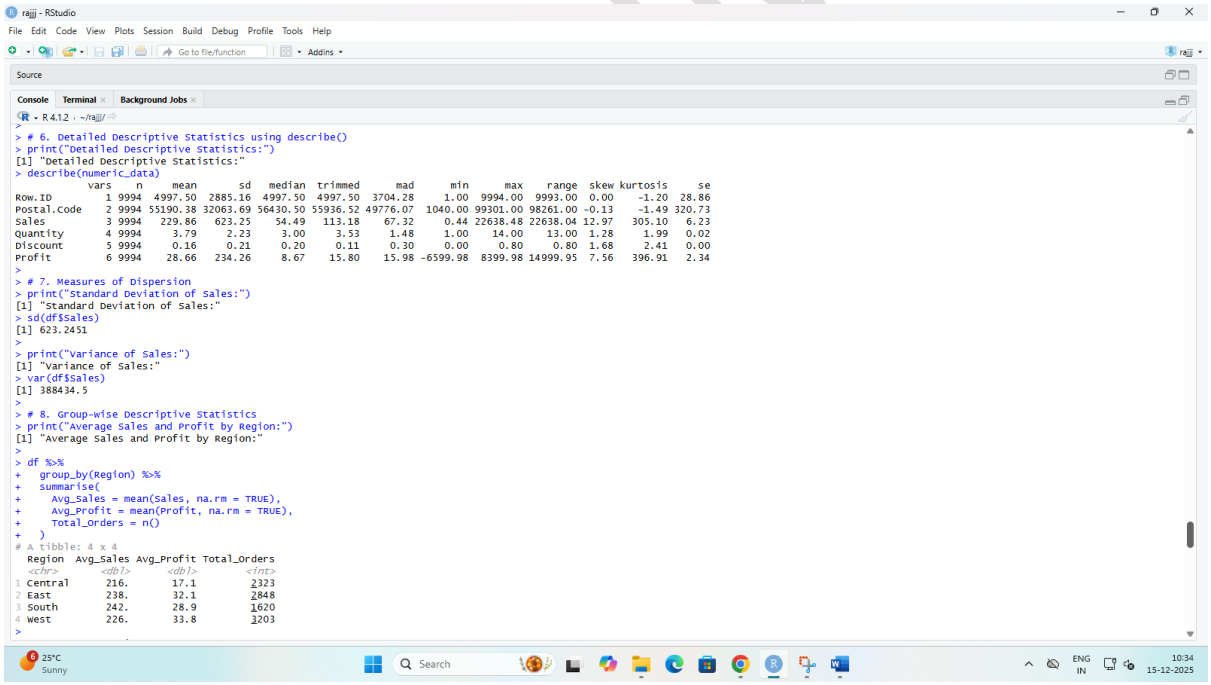
S124

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

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R



```
 Max.   :9994   Max.   :99301   Max.   :22638.480   Max.   :14.00   Max.   :0.8000   Max.   : 8399.976
>
> # 6. Detailed Descriptive Statistics using describe()
> print("Detailed Descriptive Statistics:")
[1] "Detailed Descriptive Statistics:"
> describe(numeric_data)
             vars    n     mean       sd   median  trimmed      mad      min      max    range  skew kurtosis     se
Row.ID          1 9994  4997.50  2885.16  4997.50  4997.50  3704.28     1.00  9994.00  9993.00  0.00    -1.20  28.86
Postal.Code     2 9994 55190.38 32063.69 56430.50 55936.52 49776.07  1040.00 99301.00 98261.00 -0.13    -1.49 320.73
Sales           3 9994   229.86   623.25    54.49   113.18    67.32     0.44 22638.48 22638.04 12.97   305.10   6.23
Quantity        4 9994     3.79     2.23     3.00     3.53     1.48     1.00    14.00    13.00  1.28     1.99   0.02
Discount        5 9994     0.16     0.21     0.20     0.11     0.30     0.00     0.80     0.80  1.68     2.41   0.00
Profit          6 9994    28.66   234.26     8.67    15.80    15.98 -6599.98  8399.98 14999.95  7.56   396.91   2.34
>
> # 7. Measures of Dispersion
> print("Standard Deviation of Sales:")
[1] "Standard Deviation of Sales:"
> sd(df$Sales)
[1] 623.2451
>
> print("Variance of Sales:")
[1] "Variance of Sales:"
> var(df$Sales)
[1] 388434.5
>
> # 8. Group-wise Descriptive Statistics
> print("Average Sales and Profit by Region:")
[1] "Average Sales and Profit by Region:"
>
> df %>%
+   group_by(Region) %>%
+   summarise(
+     Avg_Sales = mean(Sales, na.rm = TRUE),
+     Avg_Profit = mean(Profit, na.rm = TRUE),
+     Total_Orders = n()
+   )
# A tibble: 4 x 4
  Region  Avg_Sales Avg_Profit Total_Orders
  <chr>       <dbl>      <dbl>        <int>
1 Central      216.       17.1         2323
2 East         238.       32.1         2848
3 South        242.       28.9         1620
4 West         226.       33.8         3203
```



```
> # 6. Detailed Descriptive Statistics using describe()
> print("Detailed Descriptive Statistics:")
[1] "Detailed Descriptive Statistics:"
> describe(numeric_data)
             vars    n     mean       sd   median  trimmed      mad      min      max    range  skew kurtosis     se
Row.ID          1 9994  4997.50  2885.16  4997.50  4997.50  3704.28     1.00  9994.00  9993.00  0.00    -1.20  28.86
Postal.Code     2 9994 55190.38 32063.69 56430.50 55936.52 49776.07  1040.00 99301.00 98261.00 -0.13    -1.49 320.73
Sales           3 9994   229.86   623.25    54.49   113.18    67.32     0.44 22638.48 22638.04 12.97   305.10   6.23
Quantity        4 9994     3.79     2.23     3.00     3.53     1.48     1.00    14.00    13.00  1.28     1.99   0.02
Discount        5 9994     0.16     0.21     0.20     0.11     0.30     0.00     0.80     0.80  1.68     2.41   0.00
Profit          6 9994    28.66   234.26     8.67    15.80    15.98 -6599.98  8399.98 14999.95  7.56   396.91   2.34
>
> # 7. Measures of Dispersion
> print("Standard Deviation of Sales:")
[1] "Standard Deviation of Sales:"
> sd(df$Sales)
[1] 623.2451
>
> print("Variance of Sales:")
[1] "Variance of Sales:"
> var(df$Sales)
[1] 388434.5
>
> # 8. Group-wise Descriptive Statistics
> print("Average Sales and Profit by Region:")
[1] "Average Sales and Profit by Region:"
>
> df %>%
+   group_by(Region) %>%
+   summarise(
+     Avg_Sales = mean(Sales, na.rm = TRUE),
+     Avg_Profit = mean(Profit, na.rm = TRUE),
+     Total_Orders = n()
+   )
# A tibble: 4 x 4
  Region  Avg_Sales Avg_Profit Total_Orders
  <chr>       <dbl>      <dbl>        <int>
1 Central      216.       17.1         2323
2 East         238.       32.1         2848
3 South        242.       28.9         1620
4 West         226.       33.8         3203
>
```


SUMEET JITENDRA YADAV

S124