

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

Subject: Data Analysis with SAS / SPSS /R

Practical no. 13

Aim:Identifying and handling duplicates using distinct() (R).

Outputs →

The screenshot shows an RStudio interface with the following components:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Source Editor:** Displays R code for reading a CSV file and printing its head.
- Environment Browser:** Shows a list of objects in the global environment, including sales_df, sales_logic, sales_text, selected_cols, sorted_by_volume, split_matrix, starts_with_d, tidy_sales, and unique_sales_rep.
- Files Browser:** Shows a file tree with files like 9_jugar.circ, Custom Office Templates, Dell, desktop.ini, Downloads - Shortcut.lnk, flower_dataset - flower_dataset.csv, Iris - Iris.csv, My Music, My Pictures, My Videos, NetBeansProjects, sales_data.csv, sales_dataset.csv, sgfd05.mini, and Virtual Machines.

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

Source
R - R 4.5.2 - ~/
> library(readr)
> sales_data <- read_csv("sales_data.csv")
Rows: 1000 Columns: 14
-- Column specification --
Delimiter: ","
chr (7): Sales_Rep, Region, Product_Category, Customer_Type, Payment_Method...
dbl (6): Product_ID, Sales_Amount, Quantity_Sold, Unit_Cost, Unit_Price, ...
date (1): Sale_Date

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(sales_data)
> library(dplyr)
> # Change this to your file path
> sales_df <- read.csv("sales_data.csv")
> print("--- 1. Original Dataset Loaded from CSV ---")
[1] "--- 1. Original Dataset Loaded from CSV ---"
> print(head(sales_df))
#> #> #> #> #> #>
Product_ID Sale_Date Sales_Rep Region Sales_Amount Quantity_Sold
1 1052 2023-02-03 Bob North 5053.97 18
2 1093 2023-04-21 Bob West 4384.02 17
3 1015 2023-09-21 David South 4631.23 30
4 1072 2023-08-24 Bob South 2167.94 39
5 1061 2023-03-24 Charlie East 3750.20 13
6 1021 2023-02-11 Charlie West 3761.15 32
#> #> #> #> #> #>
Product_Category Unit_Cost Unit_Price Customer_Type Discount Payment_Method
1 Furniture 152.75 267.22 Returning 0.09 Cash
2 Furniture 3816.39 4209.44 Returning 0.11 Cash
3 Food 261.56 371.40 Returning 0.20 Bank Transfer
4 Clothing 4330.03 4467.75 New 0.02 Credit Card
5 Electronics 637.37 692.71 New 0.08 Credit Card
6 Food 900.79 1106.51 New 0.21 Cash
#> #> #> #> #> #>
Sales_Channel Region_and_Sales_Rep
1 Online North-Bob
2 Retail West-Bob
3 Retail South-David
4 Retail South-Bob
5 Online East-charlie
6 Online West-charlie
```

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

Subject: Data Analysis with SAS / SPSS /R

RStudio - R 4.5.2 · ~

```
> duplicates_report <- sales_df %>%  
+   group_by(Product_ID, Sale_Date, Sales_Rep, Region, Sales_Amount,  
+             Quantity_Sold, Product_Category, Unit_Cost, Unit_Price,  
+             Customer_Type, Discount, Payment_Method, Sales_Channel,  
+             Region_and_Sales_Report) %>%  
+   count() %%>% # Count how many times each row appears  
+   filter(n > 1) # Keep only duplicates  
> print("--- 2. Duplicate Rows Detected ---")  
[1] "--- 2. Duplicate Rows Detected ---"  
> print(duplicates_report)  
# A tibble: 0 × 15  
# Groups: Product_ID, Sale_Date, Sales_Rep, Region, Sales_Amount,  
#         Quantity_Sold, Product_Category, Unit_Cost, Unit_Price, Customer_Type,  
#         Discount, Payment_Method, Sales_Channel, Region_and_Sales_Report [0]  
# i 15 variables: Product_ID <chr>, Sale_Date <chr>, Sales_Rep <chr>,  
#   Region <chr>, Sales_Amount <dbl>, Quantity_Sold <int>,  
#   Product_Category <chr>, Unit_Cost <dbl>, Unit_Price <dbl>,  
#   Customer_Type <chr>, Discount <dbl>, Payment_Method <chr>,  
#   Sales_Channel <chr>, Region_and_Sales_Report <chr>, n <int>  
> clean_exact <- sales_df %>%  
+   distinct() # Removes perfectly identical rows  
> print("--- 3. Dataset After Removing Exact Duplicates ---")  
[1] "--- 3. Dataset After Removing Exact Duplicates ---"  
> print(clean_exact)  
Product_ID Sale_Date Sales_Rep Region Sales_Amount Quantity_Sold  
1 1052 2023-02-03 Bob North 5053.97 18  
2 1093 2023-04-21 Bob West 4384.02 17  
3 1015 2023-09-21 David South 4631.23 30  
4 1072 2023-08-24 Bob South 2167.94 39  
5 1061 2023-03-24 Charlie East 3750.20 13  
6 1021 2023-02-11 Charlie West 3761.15 32  
7 1083 2023-04-11 Bob West 618.31 29  
8 1087 2023-01-06 Eve South 7698.92 46  
9 1075 2023-06-29 David South 4223.39 30  
10 1075 2023-10-09 Charlie West 8239.58 18  
11 1088 2023-11-16 Eve North 8518.45 13  
12 1100 2023-08-14 Bob West 2198.74 43  
13 1024 2023-11-11 Eve West 6607.80 21  
14 1003 2023-12-31 Alice South 4775.59 30
```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation

9 jugaar.circ 11.9 KB Oct 24, 2024, 11:16 PM

Home

12:11 ENG IN 08-12-2025

RStudio - R 4.5.2 · ~

```
14 1003 2023-12-31 Alice South 4775.59 30  
15 1022 2023-08-17 Charlie South 8813.55 21  
16 1053 2023-10-16 Bob North 2235.83 48  
17 1002 2023-05-30 David North 6810.35 17  
18 1088 2023-10-04 Bob East 6116.75 40  
19 1030 2023-07-17 David West 3023.48 19  
20 1038 2023-03-11 Bob South 1452.35 15  
21 1002 2023-04-22 Eve North 6551.23 9  
22 1064 2023-01-04 Eve East 7412.11 10  
23 1060 2023-12-16 Eve East 3224.71 44  
24 1021 2023-11-27 Alice South 6483.84 31  
25 1033 2023-11-14 David South 4011.80 23  
26 1076 2023-12-16 Eve East 7160.75 30  
27 1058 2023-04-05 Alice North 2072.23 33  
28 1022 2023-06-01 David East 8913.13 9  
29 1089 2023-11-07 Bob West 2945.36 47  
30 1049 2023-05-17 Alice West 3741.08 1  
31 1091 2023-09-04 Charlie South 675.11 44  
32 1059 2023-08-31 Alice North 1203.97 35  
33 1042 2023-01-31 Alice North 5207.03 11  
34 1092 2023-02-09 David West 2749.17 34  
35 1060 2023-08-17 Bob East 8371.25 16  
36 1080 2023-02-05 Charlie South 245.46 9  
37 1015 2023-08-11 Eve East 3853.03 32  
38 1062 2023-01-06 Eve East 3439.72 15  
39 1062 2023-11-18 Alice East 291.34 12  
40 1047 2023-08-08 David North 1331.25 33  
41 1062 2023-03-16 David West 4195.06 45  
42 1051 2023-01-04 Eve North 4979.36 14  
43 1055 2023-12-01 Charlie West 4102.47 8  
44 1064 2023-05-14 Bob South 5356.28 8  
45 1003 2023-04-28 Charlie South 5991.80 27  
46 1051 2023-04-04 Bob East 198.25 12  
47 1007 2023-03-03 Eve West 4694.54 1  
48 1021 2023-07-13 Bob West 9638.64 43  
49 1073 2023-12-01 David West 5238.42 40  
50 1039 2023-07-22 Alice East 6807.67 42  
51 1018 2023-01-26 Alice East 3187.45 11  
52 1004 2023-06-22 Eve South 7762.51 39
```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation

9 jugaar.circ 11.9 KB Oct 24, 2024, 11:16 PM

Home

12:11 ENG IN 08-12-2025

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

Subject: Data Analysis with SAS / SPSS /R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source

```
R - R 4.5.2 · ~/
55 1014 2023-11-22 David West 9762.54 17
56 1009 2023-04-28 Alice North 1342.95 33
57 1090 2023-11-20 Alice West 267.78 32
58 1053 2023-10-02 David East 7724.57 29
59 1002 2023-07-31 David East 8090.84 21
60 1084 2023-08-13 Alice West 1290.05 21
61 1092 2023-01-20 Charlie East 2729.27 40
62 1060 2023-04-23 Eve East 273.77 23
63 1071 2023-05-20 Eve West 3003.76 6
64 1044 2023-02-16 David South 7754.10 22
65 1008 2023-01-01 Eve North 5227.81 38
66 1047 2023-03-31 Alice East 3546.15 37
67 1035 2023-05-22 Alice South 3780.22 45
68 1078 2023-11-16 David East 113.40 8
69 1081 2023-02-07 David East 3068.03 41
70 1036 2023-10-20 David North 6499.94 49
71 1050 2023-05-19 Charlie East 9744.52 35
```

Environment History Connections Tutorial

R - Global Environment

- `sales_df` 1000 obs. of 14 variables
- `sales_logic` 1000 obs. of 16 variables
- `sales_text` 1000 obs. of 15 variables
- `selected_cols` 200 obs. of 3 variables
- `sorted_by_volume` 122 obs. of 7 variables
- `split_matrix` chr [1:1000, 1:2] "North" "West" "South" "Sou...
- `starts_with_d` 200 obs. of 3 variables
- `tidy_sales` 1000 obs. of 15 variables
- `unique_sales_rep` 5 obs. of 14 variables

Files Plots Packages Help Viewer Presentation

New Folder New File Delete Rename More

Home

Name	Size	Modified
9 jugaar.circ	11.9 KB	Oct 24, 2024, 11:16 PM
Custom Office Templates		
Dell		
desktop.ini	402 B	Dec 8, 2024, 2:50 PM
Downloads - Shortcut.lnk	737 B	Mar 18, 2025, 3:29 PM
flower_dataset - flower_dataset.csv	283.7 KB	Dec 8, 2025, 11:03 AM
Iris - Iris.csv	5 KB	Dec 8, 2025, 11:06 AM
My Music		
My Pictures		
My Videos		
NetBeansProjects		
<input checked="" type="checkbox"/> sales_data.csv	101.2 KB	Dec 8, 2025, 12:04 PM
<input checked="" type="checkbox"/> sales_dataset.csv	698 B	Dec 2, 2025, 12:18 PM
sgfdu05m.ini	38 B	May 4, 2025, 1:29 PM
Virtual Machines		

Breaking news Malayalam actor... Search ENG IN 12:11 08-12-2025

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source

```
R - R 4.5.2 · ~/
1 Furniture 152.75 267.22 Returning 0.09
2 Furniture 3816.39 4209.44 Returning 0.11
3 Food 261.56 371.40 Returning 0.20
4 Clothing 4330.03 4467.75 New 0.02
5 Electronics 637.37 692.71 New 0.08
6 Food 900.79 1106.51 New 0.21
7 Furniture 2408.81 2624.09 Returning 0.14
8 Furniture 3702.51 3964.65 New 0.12
9 Furniture 738.06 1095.45 New 0.05
10 Clothing 2228.35 2682.34 New 0.13
11 Furniture 2440.11 2517.60 New 0.23
12 Food 1100.81 1137.44 Returning 0.08
13 Food 622.01 641.09 Returning 0.00
14 Furniture 4190.28 4270.65 New 0.20
15 Food 2537.20 2869.60 New 0.29
16 Furniture 121.19 487.65 New 0.18
17 Furniture 4024.76 4420.15 Returning 0.04
18 Electronics 4904.93 5034.35 New 0.10
19 Clothing 3049.33 3209.22 Returning 0.26
20 Clothing 2543.36 2790.10 Returning 0.07
21 Electronics 4398.16 4439.12 New 0.18
22 Electronics 4764.98 5074.42 New 0.12
```

Environment History Connections Tutorial

R - Global Environment

- `sales_df` 1000 obs. of 14 variables
- `sales_logic` 1000 obs. of 16 variables
- `sales_text` 1000 obs. of 15 variables
- `selected_cols` 200 obs. of 3 variables
- `sorted_by_volume` 122 obs. of 7 variables
- `split_matrix` chr [1:1000, 1:2] "North" "West" "South" "Sou...
- `starts_with_d` 200 obs. of 3 variables
- `tidy_sales` 1000 obs. of 15 variables
- `unique_sales_rep` 5 obs. of 14 variables

Files Plots Packages Help Viewer Presentation

New Folder New File Delete Rename More

Home

Name	Size	Modified
9 jugaar.circ	11.9 KB	Oct 24, 2024, 11:16 PM
Custom Office Templates		
Dell		
desktop.ini	402 B	Dec 8, 2024, 2:50 PM
Downloads - Shortcut.lnk	737 B	Mar 18, 2025, 3:29 PM
flower_dataset - flower_dataset.csv	283.7 KB	Dec 8, 2025, 11:03 AM
Iris - Iris.csv	5 KB	Dec 8, 2025, 11:06 AM
My Music		
My Pictures		
My Videos		
NetBeansProjects		
<input checked="" type="checkbox"/> sales_data.csv	101.2 KB	Dec 8, 2025, 12:04 PM
<input checked="" type="checkbox"/> sales_dataset.csv	698 B	Dec 2, 2025, 12:18 PM
sgfdu05m.ini	38 B	May 4, 2025, 1:29 PM
Virtual Machines		

Breaking news Malayalam actor... Search ENG IN 12:11 08-12-2025

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

Subject: Data Analysis with SAS / SPSS /R

RStudio interface showing a session in R 4.5.2. The Source pane displays a large dataset with 929 rows omitted due to max.print. The Environment pane shows various global variables and datasets. The Files pane lists local files including 'sales_data.csv' and 'iris.csv'. The bottom status bar shows the date as 08-12-2025.

```
Source
R - R 4.5.2 · ~/ ~
34 Credit Card Online West-David
35 Cash Online East-Bob
36 Credit Card Retail South-Charlie
37 Cash Retail East-Eve
38 Credit Card Online East-Eve
39 Bank Transfer Retail East-Alice
40 Credit Card Online North-David
41 Credit Card Online West-David
42 Bank Transfer Retail North-Eve
43 Bank Transfer Online West-Charlie
44 Cash Online South-Bob
45 Cash Online South-Charlie
46 Cash Online East-Bob
47 Credit Card Online West-Eve
48 Cash Retail West-Bob
49 Credit Card Retail West-David
50 Credit Card Retail East-Alice
51 Cash Online East-Alice
52 Cash Online South-Eve
53 Credit Card Online North-David
54 Bank Transfer Online West-Charlie
55 Bank Transfer Online West-David
56 Credit Card Online North-Alice
57 Cash Retail West-Alice
58 Credit Card Online East-David
59 Credit Card Retail East-David
60 Cash Online West-Alice
61 Bank Transfer Retail East-Charlie
62 Credit Card Online East-Eve
63 Bank Transfer Retail West-Eve
64 Bank Transfer Retail South-David
65 Cash Online North-Eve
66 Credit Card Online East-Alice
67 Credit Card Retail South-Alice
68 Bank Transfer Online East-David
69 Bank Transfer Online East-David
70 Cash Online North-David
71 Bank Transfer Retail East-Charlie
[ reached 'max' / getoption("max.print") -- omitted 929 rows ]
```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation

Home Name Size Modified

- 9 jugaar.circ 11.9 KB Oct 24, 2024, 11:16 PM
- Custom Office Templates
- Dell
- desktop.ini 402 B Dec 8, 2024, 2:50 PM
- Downloads - Shortcut.lnk 737 B Mar 18, 2025, 3:29 PM
- flower_dataset - flower_dataset.csv 283.7 KB Dec 8, 2025, 11:03 AM
- Iris - Iris.csv 5 KB Dec 8, 2025, 11:06 AM
- My Music
- My Pictures
- My Videos
- NetBeansProjects
- sales_data.csv 101.2 KB Dec 8, 2025, 12:04 PM
- sales_dataset.csv 698 B Dec 2, 2025, 12:18 PM
- sgfdu05m.ini 38 B May 4, 2025, 1:29 PM
- Virtual Machines

ENG IN 12:11 08-12-2025

RStudio interface showing a session in R 4.5.2. The Source pane displays a dataset with 929 rows omitted due to max.print. The Environment pane shows various global variables and datasets. The Files pane lists local files including 'sales_data.csv' and 'iris.csv'. The bottom status bar shows the date as 08-12-2025.

```
Source
R - R 4.5.2 · ~/ ~
60 Cash Online West-Alice
61 Bank Transfer Retail East-Charlie
62 Credit Card Online East-Eve
63 Bank Transfer Retail West-Eve
64 Bank Transfer Retail South-David
65 Cash Online North-Eve
66 Credit Card Online East-Alice
67 Credit Card Retail South-Alice
68 Bank Transfer Online East-David
69 Bank Transfer Online East-David
70 Cash Online North-David
71 Bank Transfer Retail East-Charlie
[ reached 'max' / getoption("max.print") -- omitted 929 rows ]
> unique_sales_rep <- sales_df %>%
+ distinct(Sales_Rep, .keep_all = TRUE)
> print("--- 4. Unique Sales Rep (All columns kept, duplicates removed) ---")
[1] "--- 4. Unique Sales Rep (All columns kept, duplicates removed) ---"
> print(unique_sales_rep)
Product_ID Sale_Date Sales_Rep Region Sales_Amount Quantity_Sold
1 1052 2023-02-03 Bob North 5053.97 18
2 1015 2023-09-21 David South 4631.23 30
3 1061 2023-03-24 Charlie East 3750.20 13
4 1087 2023-01-06 Eve South 7698.92 46
5 1003 2023-12-31 Alice South 4775.59 30
Product_Category Unit_Cost Unit_Price Customer_Type Discount Payment_Method
1 Furniture 152.75 267.22 Returning 0.09 Cash
2 Food 261.56 371.40 Returning 0.20 Bank Transfer
3 Electronics 637.37 692.71 New 0.08 Credit Card
4 Furniture 3702.51 3964.65 New 0.12 Bank Transfer
5 Furniture 4190.28 4270.65 New 0.20 Cash
Sales_Channel Region_and_Sales_Rep
1 Online North-Bob
2 Retail South-David
3 Online East-Charlie
4 Online South-Eve
5 Online South-Alice
>
>
>
```

Environment History Connections Tutorial

Files Plots Packages Help Viewer Presentation

Home Name Size Modified

- 9 jugaar.circ 11.9 KB Oct 24, 2024, 11:16 PM
- Custom Office Templates
- Dell
- desktop.ini 402 B Dec 8, 2024, 2:50 PM
- Downloads - Shortcut.lnk 737 B Mar 18, 2025, 3:29 PM
- flower_dataset - flower_dataset.csv 283.7 KB Dec 8, 2025, 11:03 AM
- Iris - Iris.csv 5 KB Dec 8, 2025, 11:06 AM
- My Music
- My Pictures
- My Videos
- NetBeansProjects
- sales_data.csv 101.2 KB Dec 8, 2025, 12:04 PM
- sales_dataset.csv 698 B Dec 2, 2025, 12:18 PM
- sgfdu05m.ini 38 B May 4, 2025, 1:29 PM
- Virtual Machines

ENG IN 12:11 08-12-2025