

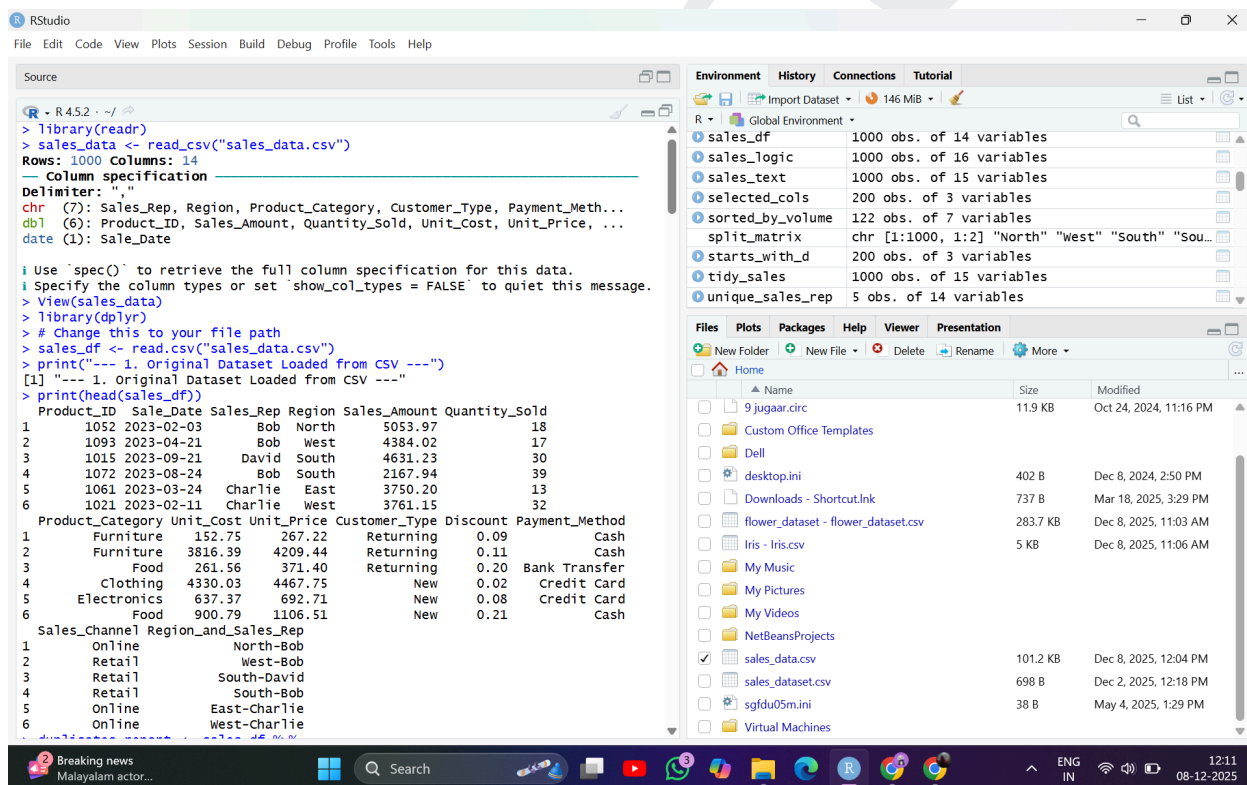
SHETH L.U.J AND SIR M.V COLLEGE

Subject: Data Analysis with SAS / SPSS /R

Practical no. 13

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

Outputs→



The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for loading a CSV file and displaying its structure and content.
- Environment:** Lists the objects created in the R session, including `sales_df`, `sales_logical`, `sales_text`, `selected_cols`, `sorted_by_volume`, `split_matrix`, `starts_with_d`, `tidy_sales`, and `unique_sales_rep`.
- Files:** Shows the file explorer with various files and folders, including `9_jugar.circ`, `Custom Office Templates`, `Dell`, `desktop.ini`, `Downloads - ShortcutLink`, `flower_dataset - flower_dataset.csv`, `Iris - Iris.csv`, `My Music`, `My Pictures`, `My Videos`, `NetBeansProjects`, `sales_data.csv`, `sales_dataset.csv`, `sgfdu05m.ini`, and `Virtual Machines`.

```
> 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_Meth...
dbl (6): Product_ID, Sales_Amount, Quantity_Sold, Unit_Cost, Unit_Price, ...
date (1): Sale_Date

# Use `spec()` to retrieve the full column specification for this data.
# 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

The screenshot shows the RStudio interface. The Source pane on the left contains R code for data cleaning. The Environment pane on the right shows the loaded datasets. The Files pane at the bottom shows the file explorer.

```
> 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,
+     Customer_Type, Discount, Payment_Method, Sales_Channel,
+     Region_and_Sales_Rep) %>%
+   count() %>%
+   filter(n > 1) # Keep only duplicates
+ print("--- 2. Duplicate Rows Detected ---")
[1] "--- 2. Duplicate Rows Detected ---"
> print(duplicates_report)
# A tibble: 0 x 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_Rep [0]
# i 15 variables: Product_ID <int>, 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_Rep <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

The screenshot shows the RStudio interface. The Source pane on the left contains R code for data cleaning. The Environment pane on the right shows the loaded datasets. The Files pane at the bottom shows the file explorer.

```
> 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,
+     Customer_Type, Discount, Payment_Method, Sales_Channel,
+     Region_and_Sales_Rep) %>%
+   count() %>%
+   filter(n > 1) # Keep only duplicates
+ print("--- 2. Duplicate Rows Detected ---")
[1] "--- 2. Duplicate Rows Detected ---"
> print(duplicates_report)
# A tibble: 0 x 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_Rep [0]
# i 15 variables: Product_ID <int>, 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_Rep <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	
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

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

Subject: Data Analysis with SAS / SPSS /R

The screenshot shows the RStudio interface. The Source pane on the left displays a data table with columns: Product_Category, Unit_Cost, Unit_Price, Customer_Type, and Discount. The Environment pane on the right shows the Global Environment with variables: 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), and unique_sales_rep (5 obs. of 14 variables). The Files pane at the bottom shows a file explorer with various files and folders.

Product_Category	Unit_Cost	Unit_Price	Customer_Type	Discount
Furniture	152.75	267.22	Returning	0.09
Furniture	3816.39	4209.44	Returning	0.11
Food	261.56	371.40	Returning	0.20
Clothing	4330.03	4467.75	New	0.02
Electronics	637.37	692.71	New	0.08
Food	900.79	1106.51	New	0.21
Furniture	2408.81	2624.09	Returning	0.14
Furniture	3702.51	3964.65	New	0.12
Furniture	738.06	1095.45	New	0.05
Clothing	2228.35	2682.34	New	0.13
Furniture	2440.11	2517.60	New	0.23
Food	1100.81	1137.44	Returning	0.08
Food	622.01	641.09	Returning	0.00
Furniture	4190.28	4270.65	New	0.20
Food	2537.20	2869.60	New	0.29
Furniture	121.19	487.65	New	0.18
Furniture	4024.76	4420.15	Returning	0.04
Electronics	4904.93	5034.35	New	0.10
Clothing	3049.33	3209.22	Returning	0.26
Clothing	2543.36	2790.10	Returning	0.07
Electronics	4398.16	4439.12	New	0.18
Electronics	4764.96	5074.42	New	0.12

The screenshot shows the RStudio interface. The Source pane on the left displays a data table with columns: Payment_Method, Sales_Channel, Region_and_Sales_Rep, and Discount. The Environment pane on the right shows the Global Environment with variables: 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), and unique_sales_rep (5 obs. of 14 variables). The Files pane at the bottom shows a file explorer with various files and folders.

Payment_Method	Sales_Channel	Region_and_Sales_Rep	Discount
Cash	Online	North-Bob	0.04
Cash	Retail	West-Bob	0.19
Bank Transfer	Retail	South-David	0.20
Credit Card	Retail	South-Bob	0.23
Credit Card	Online	East-Charlie	0.17
Cash	Online	West-Charlie	0.19
Cash	Online	West-Bob	0.00
Bank Transfer	Online	South-Eve	0.05
Bank Transfer	Online	South-David	0.30
Bank Transfer	Online	West-Charlie	0.02
Bank Transfer	Retail	North-Eve	0.30
Bank Transfer	Online	West-Bob	0.02

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

Subject: Data Analysis with SAS / SPSS /R

RStudio interface showing a data frame with 929 rows and 4 columns: Transaction Type, Channel, Region, and Sales Representative. The Environment pane on the right lists the loaded objects: 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), and unique_sales_rep (5 obs. of 14 variables).

RStudio interface showing R code for data manipulation and the resulting output in the console. The code includes: `unique_sales_rep <- sales_df %>% distinct(Sales_Rep, .keep_all = TRUE)`, `print("--- 4. Unique Sales Rep (All columns kept, duplicates removed) ---")`, and `print(unique_sales_rep)`. The console output shows the first 5 rows of the resulting data frame.

Product_ID	Sale_Date	Sales_Rep	Region	Sales_Amount	Quantity_Sold
1052	2023-02-03	Bob	North	5053.97	18
1015	2023-09-21	David	South	4631.23	30
1061	2023-03-24	Charlie	East	3750.20	13
1087	2023-01-06	Eve	South	7698.92	46

Product_Category	Unit_Cost	Unit_Price	Customer_Type	Discount	Payment_Method
Furniture	152.75	267.22	Returning	0.09	Cash
Food	261.56	371.40	Returning	0.20	Bank Transfer
Electronics	637.37	692.71	New	0.08	Credit Card
Furniture	3702.51	3964.65	New	0.12	Bank Transfer

Sales_Channel	Region_and_Sales_Rep
Online	North-Bob
Retail	South-David
Online	East-Charlie
Online	South-Eve
Online	South-Alice