

SHETH L.U.J AND M.V COLLEGE  
PRACTICAL NO .13  
SUBJECT - DATA ANALYSIS

AIM- 13 Identifying and handling duplicates using distinct() (R).

INPUT-

```
library(dplyr)
```

```
orders_df <- data.frame(  
  OrderID = c(101, 102, 102, 103, 104, 101, 104),  
  Customer = c("Unnati", "Kirti", "Divya", "Geet", "Bhumika", "Aishwarya", "Parvati"),  
  Product = c("Laptop", "Phone", "Phone", "Tablet", "Monitor", "Laptop", "Mouse")  
)
```

```
print("--- 1. Original Dataset (Note 7 rows) ---")  
print(orders_df)
```

```
duplicates_report <- orders_df %>%  
  group_by(OrderID, Customer, Product) %>%  
  count() %>%  
  filter(n > 1)
```

```
print("--- 2. Identification Report (Rows that are duplicated) ---")  
print(duplicates_report)
```

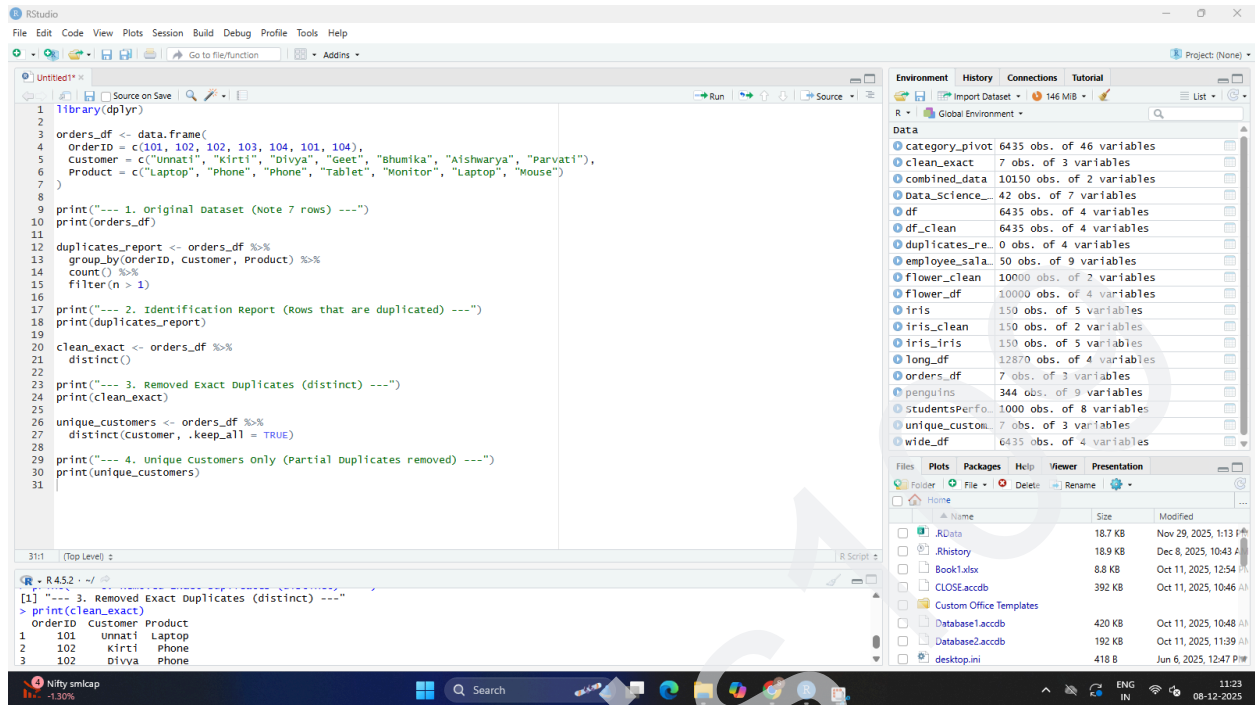
```
clean_exact <- orders_df %>%  
  distinct()
```

```
print("--- 3. Removed Exact Duplicates (distinct) ---")  
print(clean_exact)
```

```
unique_customers <- orders_df %>%  
  distinct(Customer, .keep_all = TRUE)
```

```
print("--- 4. Unique Customers Only (Partial Duplicates removed) ---")  
print(unique_customers)
```

SHETH L.U.J AND M.V COLLEGE  
PRACTICAL NO .13  
SUBJECT - DATA ANALYSIS



```
1 library(dplyr)
2
3 orders_df <- data.frame(
4   orderID = c(101, 102, 102, 103, 104, 101, 104),
5   customer = c("Unnati", "Kirti", "Divya", "Geet", "Bhumika", "Aishwarya", "Parvati"),
6   Product = c("Laptop", "Phone", "Phone", "Tablet", "Monitor", "Laptop", "Mouse")
7 )
8
9 print("--- 1. Original Dataset (Note 7 rows) ---")
10 print(orders_df)
11
12 duplicates_report <- orders_df %>%
13   group_by(OrderID, Customer, Product) %>%
14   count() %>%
15   filter(n > 1)
16
17 print("--- 2. Identification Report (Rows that are duplicated) ---")
18 print(duplicates_report)
19
20 clean_exact <- orders_df %>%
21   distinct()
22
23 print("--- 3. Removed Exact Duplicates (distinct) ---")
24 print(clean_exact)
25
26 unique_customers <- orders_df %>%
27   distinct(Customer, .keep_all = TRUE)
28
29 print("--- 4. Unique Customers Only (Partial Duplicates removed) ---")
30 print(unique_customers)
31
```

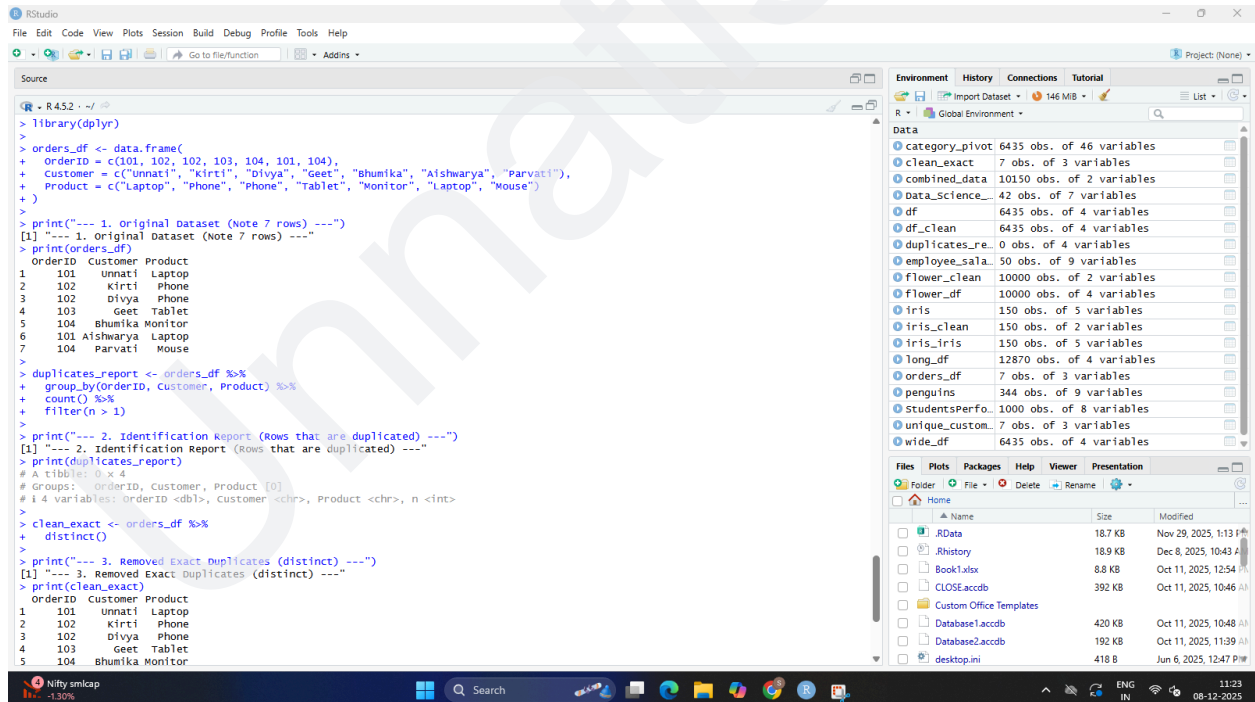
Environment: Global Environment

Object	Variables
category_pivot	6435 obs. of 46 variables
clean_exact	7 obs. of 3 variables
combined_data	10150 obs. of 2 variables
data_science	42 obs. of 7 variables
df	6435 obs. of 4 variables
df_clean	6435 obs. of 4 variables
duplicates_re	0 obs. of 4 variables
employee_sala	50 obs. of 9 variables
flower_clean	10000 obs. of 2 variables
flower_df	10000 obs. of 4 variables
iris	150 obs. of 5 variables
iris_clean	150 obs. of 2 variables
iris_iris	150 obs. of 2 variables
long_df	12870 obs. of 4 variables
orders_df	7 obs. of 3 variables
penguins	344 obs. of 9 variables
StudentsPerfo	1000 obs. of 8 variables
unique_custom	7 obs. of 3 variables
wide_df	6435 obs. of 4 variables

Source: R 4.5.2

```
> print(clean_exact)
  orderID customer Product
1    101   Unnati  Laptop
2    102    Kirti   Phone
3    102   Divya   Phone
```

OUTPUT -



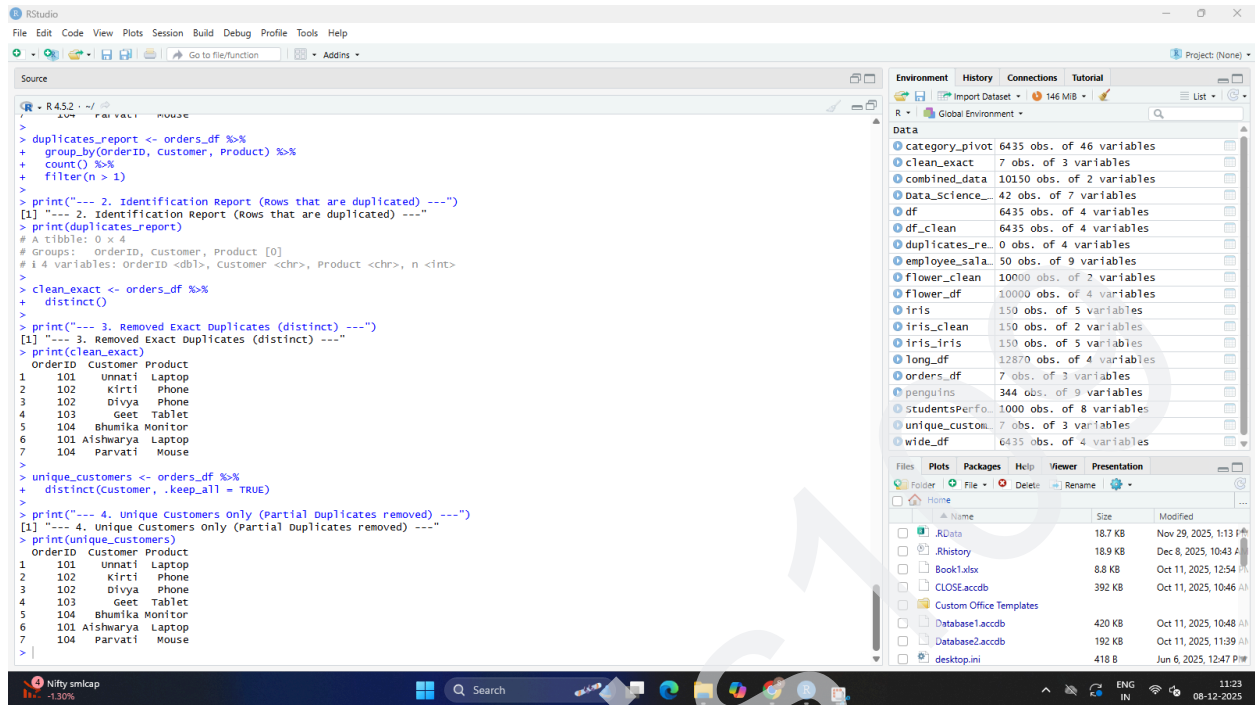
```
> library(dplyr)
>
> orders_df <- data.frame(
+   orderID = c(101, 102, 102, 103, 104, 101, 104),
+   customer = c("Unnati", "Kirti", "Divya", "Geet", "Bhumika", "Aishwarya", "Parvati"),
+   Product = c("Laptop", "Phone", "Phone", "Tablet", "Monitor", "Laptop", "Mouse")
+ )
>
> print("--- 1. Original Dataset (Note 7 rows) ---")
[1] "--- 1. Original Dataset (Note 7 rows) ---"
> print(orders_df)
  orderID customer Product
1    101   Unnati  Laptop
2    102    Kirti   Phone
3    102   Divya   Phone
4    103     Geet  Tablet
5    104   Bhumika Monitor
6    101 Aishwarya Laptop
7    104   Parvati  Mouse
>
> duplicates_report <- orders_df %>%
+   group_by(OrderID, Customer, Product) %>%
+   count() %>%
+   filter(n > 1)
>
> print("--- 2. Identification Report (Rows that are duplicated) ---")
[1] "--- 2. Identification Report (Rows that are duplicated) ---"
> print(duplicates_report)
# A tibble: 0 x 4
# Groups:   OrderID, Customer, Product [0]
# 4 variables: OrderID <dbl>, Customer <chr>, Product <chr>, n <int>
>
> clean_exact <- orders_df %>%
+   distinct()
>
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
> print(clean_exact)
  orderID customer Product
1    101   Unnati  Laptop
2    102    Kirti   Phone
3    102   Divya   Phone
4    103     Geet  Tablet
5    104   Bhumika Monitor
```

Environment: Global Environment

Object	Variables
category_pivot	6435 obs. of 46 variables
clean_exact	7 obs. of 3 variables
combined_data	10150 obs. of 2 variables
data_science	42 obs. of 7 variables
df	6435 obs. of 4 variables
df_clean	6435 obs. of 4 variables
duplicates_re	0 obs. of 4 variables
employee_sala	50 obs. of 9 variables
flower_clean	10000 obs. of 2 variables
flower_df	10000 obs. of 4 variables
iris	150 obs. of 5 variables
iris_clean	150 obs. of 2 variables
iris_iris	150 obs. of 2 variables
long_df	12870 obs. of 4 variables
orders_df	7 obs. of 3 variables
penguins	344 obs. of 9 variables
StudentsPerfo	1000 obs. of 8 variables
unique_custom	7 obs. of 3 variables
wide_df	6435 obs. of 4 variables

NAME - UNNATI RATHOD  
ROLL NO S109

SHETH L.U.J AND M.V COLLEGE  
PRACTICAL NO .13  
SUBJECT - DATA ANALYSIS



The screenshot shows the RStudio interface with the following components:

- Source Pane:** Contains R code for data cleaning. The code identifies duplicates, removes exact duplicates, and keeps unique customers. It includes print statements for each step.
- Environment Pane:** Lists the objects created in the R session, including data frames like `category_pivot`, `clean_exact`, `combined_data`, `Data_Science`, `df`, `df_clean`, `duplicates_re`, `employee_sala`, `flower_clean`, `flower_df`, `iris`, `iris_clean`, `iris_iris`, `long_df`, `orders_df`, `penguins`, `studentsPerfo`, `unique_custom`, and `wide_df`.
- Files Pane:** Shows a list of files in the current directory, including `.RData`, `.Rhistory`, `Book1.xlsx`, `CLOSE.accd`, `Custom Office Templates`, `Database1.accd`, `Database2.accd`, and `desktop.ini`.

```
> duplicates_report <- orders_df %>%
+ group_by(orderID, customer, Product) %>%
+ count() %>%
+ filter(n > 1)
>
> print("--- 2. Identification Report (Rows that are duplicated) ---")
[1] "--- 2. Identification Report (Rows that are duplicated) ---"
> print(duplicates_report)
# A tibble: 0 x 4
# Groups:   orderID, customer, Product [0]
# 1 4 variables: orderID <dbl>, customer <chr>, Product <chr>, n <int>
>
> clean_exact <- orders_df %>%
+ distinct()
>
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
> print(clean_exact)
  orderID customer Product
1     101  Unnati  Laptop
2     102   Kirti   Phone
3     102   Divya   Phone
4     103    Geet   Tablet
5     104  bhumika Monitor
6     101 Aishwarya Laptop
7     104  Parvati   Mouse
>
> unique_customers <- orders_df %>%
+ distinct(customer, .keep_all = TRUE)
>
> print("--- 4. Unique Customers only (Partial Duplicates removed) ---")
[1] "--- 4. Unique Customers only (Partial Duplicates removed) ---"
> print(unique_customers)
  orderID customer Product
1     101  Unnati  Laptop
2     102   Kirti   Phone
3     102   Divya   Phone
4     103    Geet   Tablet
5     104  bhumika Monitor
6     101 Aishwarya Laptop
7     104  Parvati   Mouse
>
```

NAME - UNNATI RATHOD  
ROLL NO S109