

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**

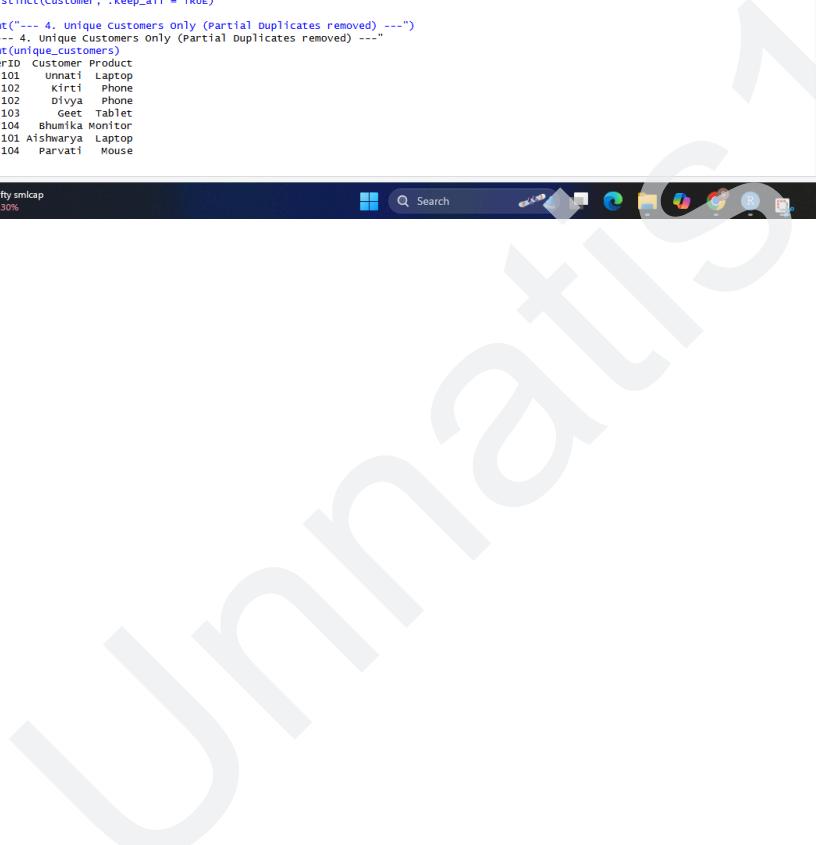
The screenshot shows the RStudio interface with the following components:

- Script Editor:** Displays an R script titled "Untitled1.R". The code performs several operations on a dataset:
 - Imports dplyr.
 - Creates a data frame "orders_df" with columns OrderID, Customer, and Product.
 - Prints the original dataset.
 - Finds and prints duplicate rows.
 - Creates a clean dataset by removing duplicates.
 - Prints unique customers.
- Environment Browser:** Shows a list of available datasets and objects, including "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", "penguins", "studentsPerfo...", "unique_custom...", and "wide_df".
- File Browser:** Shows the project's directory structure with files like .RData, .History, Book1.xlsx, CLOSE.accdb, CustomOffice Templates, Database1.accdb, Database2.accdb, and desktop.ini.
- Status Bar:** Shows the R version (R 4.5.2), the current file path (~\Untitled1.R), and the date (Nov 29, 2023).
- System Tray:** Shows icons for battery (13%), signal strength, and system status.

OUTPUT -

NAME - UNNATI RATHOD
ROLL NO S109

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



R Studio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

```
R + R4.5.2 - /Users/unnati/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 × 4
# Groups:   OrderID, Customer, Product [0]
# i 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)
# # A tibble: 7 × 3
# Groups:   OrderID, Customer, Product [7]
# i OrderID Customer Product
# 1 101     Unnati Laptop
# 2 102     Kirti Phone
# 3 102     Divya Phone
# 4 103     Geet Tablet
# 5 104     Bhumiika 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] "... 2 Unique Customers Only (Partial duplicates removed) ---"
> print(unique_customers)
# # A tibble: 7 × 3
# Groups:   OrderID, Customer, Product [7]
# i OrderID Customer Product
# 1 101     Unnati Laptop
# 2 102     Kirti Phone
# 3 102     Divya Phone
# 4 103     Geet Tablet
# 5 104     Bhumiika Monitor
# 6 101     Aishwarya Laptop
# 7 104     Parvati Mouse
> |
```

Environment History Connections Tutorial

Import Dataset 146 MB

Global Environment

Data

- 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 3 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

Files Plots Packages Help Viewer Presentation

Folder File Delete Rename

Name	Size	Modified
RData	18.7 KB	Nov 29, 2025, 1:13 PM
Rhistory	18.9 KB	Dec 8, 2025, 10:43 AM
Book1.xlsx	8.8 KB	Oct 11, 2025, 12:54 PM
CLOSEaccdb	392 KB	Oct 11, 2025, 10:46 AM
Custom Office Templates		
Database1.accdb	420 KB	Oct 11, 2025, 10:48 AM
Database2.accdb	192 KB	Oct 11, 2025, 11:39 AM
desktop.ini	418 B	Jun 6, 2025, 12:47 PM

11:23 ENG IN 08-12-2025