

PRACTICAL NO : 13

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

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Addins:** A dropdown menu containing various icons for different add-ins.
- Source Editor:** Displays R code for identifying and handling duplicates. The code includes loading the dplyr package, creating a dataset with intentional duplicates, and printing the original dataset.
- Console:** Shows the output of the R code, including the creation of the orders_df data frame and its contents.
- Environment:** Shows the global environment with objects like superstore_logic, superstore_text, survey_clean, survey_df, surveys, tidy_df, and unique_customers.
- Files:** A sidebar showing the file structure, including PDFs, ZIP files, and various document formats.
- Search Bar:** A search bar at the bottom left.
- System Icons:** Icons for battery level (29%), signal strength (gg), and system status.
- Bottom Status Bar:** Shows the date (08-12-2025) and time (12:04 PM).

R Studio

File Edit Code View Plots Session Build Debug Profile Tools Help

Source

Console Terminal Background Jobs

R > R 4.5.2. ~

```
> #-----  
> # 2. IDENTIFYING DUPLICATES (Before removing them)  
> #-----  
> # we can use group_by() and count() to see which rows appear more than once.  
> duplicates_report <- orders_df %>%  
+   group_by(OrderID, Customer, Product) %>%  
+   count() %>% # counts occurrences  
+   filter(n > 1) # Keeps only rows that appear more than once  
> print("---- 2. Identification Report (Rows that are duplicated) ----")  
[1] "---- 2. Identification Report (Rows that are duplicated) ----"  
> print(duplicates_report)  
# A tibble: 0 × 3  
# Groups:   OrderID, Customer, Product [0]  
# i 4 variables: OrderID <dbl>, Customer <chr>, Product <chr>, n <int>  
> #-----  
> # 3. HANDLING DUPLICATES: Exact Matches  
> #-----  
> # Scenario: Remove rows where EVERY column is identical.  
> # Result: Alice (101) and Bob (102) duplicates are removed.  
> # Dan (104) is kept twice because his Products are different ("Monitor" vs "Mouse").  
> 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	Dheeraj Laptop
2	102	Manya Phone
3	102	Kaju Phone
4	103	Pista Tablet
5	104	Badam Monitor
6	101	kaju Laptop
7	104	Manya Mouse

Environment History Connections Tutorial

Import Dataset 233 MB

R Global Environment

superstore_logic 9994 obs. of 23 variables
superstore_text 9994 obs. of 22 variables
survey_clean 35549 obs. of 2 variables
survey_df 35549 obs. of 9 variables
surveys 35549 obs. of 9 variables
tidy_df 9994 obs. of 27 variables
unique_customers 6 obs. of 3 variables

Files Plots Packages Help Viewer Presentation

New Folder New File Delete Rename More

Home Downloads

Name	Size	Modified
TCN 1.D.pdf	209.2 KB	Nov 27, 2025, 3:28 PM
train.csv.zip	22 B	Nov 25, 2025, 12:21 PM
Unconfined 187290.crdownload	7569 MB	Dec 5, 2025, 2:02 PM
Unnati prac 1.pdf.odt	5 MB	Nov 13, 2025, 9:30 AM
Unnati prac 1.pdf.pdf	1.7 MB	Nov 13, 2025, 9:30 AM
Untitled document (1).pdf	1.7 MB	Nov 13, 2025, 9:29 AM
Untitled document.pdf	39.6 KB	Nov 12, 2025, 3:35 PM
WhatsApp Unknown 2025-11-18 at 9:41.15 AM		
WhatsApp Unknown 2025-11-18 at 9:41.15 AM.zip	1013.5 KB	Nov 18, 2025, 11:45 AM
Whols		
Whols.zip	585.1 KB	Nov 12, 2025, 2:24 PM
penguins_lter.csv	49.7 KB	Dec 8, 2025, 10:34 AM
surveys.csv	997.7 KB	Dec 8, 2025, 10:36 AM
~\$nati prac 1.pdf.odt	0 B	Jan 1, 1970, 5:30 AM

29°C 12:04

EN IN 08-12-2025

Sheth L.U.J. College of Arts & Sir M.V. College Of Science & Commerce

SUBJECT NAME: Data Analysis with SAS / SPSS/R

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Project dropdown set to '(None)'. The left sidebar has tabs for Source, Console (active), Terminal, and Background Jobs. The console window displays R code and its output, including data cleaning steps and a final dataset named 'unique_customers'. The right sidebar features tabs for Environment, History, Connections, and Tutorial, with Environment selected. It shows a global environment with various datasets like 'superstore_logic' and 'superstore_text'. Below the environment is a file browser with a 'Downloads' folder containing several files, including 'TCN 1.pdf', 'train.csv.zip', and 'Unnati prac 1.pdf.odt'. The bottom status bar shows the date '08-12-2025' and time '12:04'.

DHEERAJ SINGH
S118