

PRACTICAL NO 13

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

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
> # Load the library
> library(dplyr)
> # =====
> # 1. SETUP: Create a Dataset with Intentional Duplicates
> # - Row 1 and Row 6 are EXACT duplicates (Same OrderID, Customer, and Product)
> # - Row 2 and Row 3 are EXACT duplicates
> # - Row 3 and Row 7 share the same Customer ("wizzy"), but bought different products
> orders_df <- data.frame(
+   orderID = c(101, 102, 102, 103, 104, 101, 104),
+   customer = c("madesh", "skizzy", "skizzy", "baby", "wizzy", "madesh", "wizzy"),
+   product = c("Laptop", "Phone", "Phone", "Tablet", "Monitor", "Laptop", "Mouse")
+ )
> print("--- 1. Original Dataset (7 rows) ---")
[1] "--- 1. Original Dataset (7 rows) ---"
> print(orders_df)
  orderID customer product
1      101  madesh  Laptop
2      102  skizzy   Phone
3      102  skizzy   Phone
4      103   baby  Tablet
5      104  wizzy  Monitor
6      101  madesh  Laptop
7      104  wizzy   Mouse
> # =====
> # 2. IDENTIFYING DUPLICATES (Before removing them)
> # Use group_by() + count() to find rows that appear more than once (exact duplicates)
> duplicates_report <- orders_df %>%
+   group_by(orderID, customer, product) %>%
+   count() %>%
+   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: 2 x 4
# Groups:   orderID, customer, product [2]

```

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R4.1.2 - ~/
  orderID customer product      n
  <dbl> <chr> <chr> <int>
1      101  madesh  Laptop      2
2      102  skizzy   Phone      2
> # =====
> # 3. HANDLING DUPLICATES: Exact Matches
> # Scenario: Remove rows where EVERY column is identical.
> # Result: rows with exact duplicates (e.g., rows for orderID 101 and 102) are de-duplicated.
> # Dinesh (OrderID 104) remains twice because Products differ ("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  madesh  Laptop
2      102  skizzy   Phone
3      103   baby  Tablet
4      104  wizzy  Monitor
5      104  wizzy   Mouse
> # =====
> # 4. HANDLING DUPLICATES: Specific columns (.keep_all = TRUE)
> # Scenario: We want a list of UNIQUE CUSTOMERS.
> # If a customer appears multiple times, keep the first occurrence and keep all columns.
> # Use keep_all = TRUE to retain OrderID and Product for that first occurrence.
> unique_customers <- orders_df %>%
+   distinct(customer, .keep_all = TRUE)
> print("--- 4. Unique Customers Only (keep first occurrence of each Customer) ---")
[1] "--- 4. Unique Customers Only (keep first occurrence of each Customer) ---"
> print(unique_customers)
  orderID customer product
1      101  madesh  Laptop
2      102  skizzy   Phone
3      103   baby  Tablet
4      104  wizzy  Monitor

```