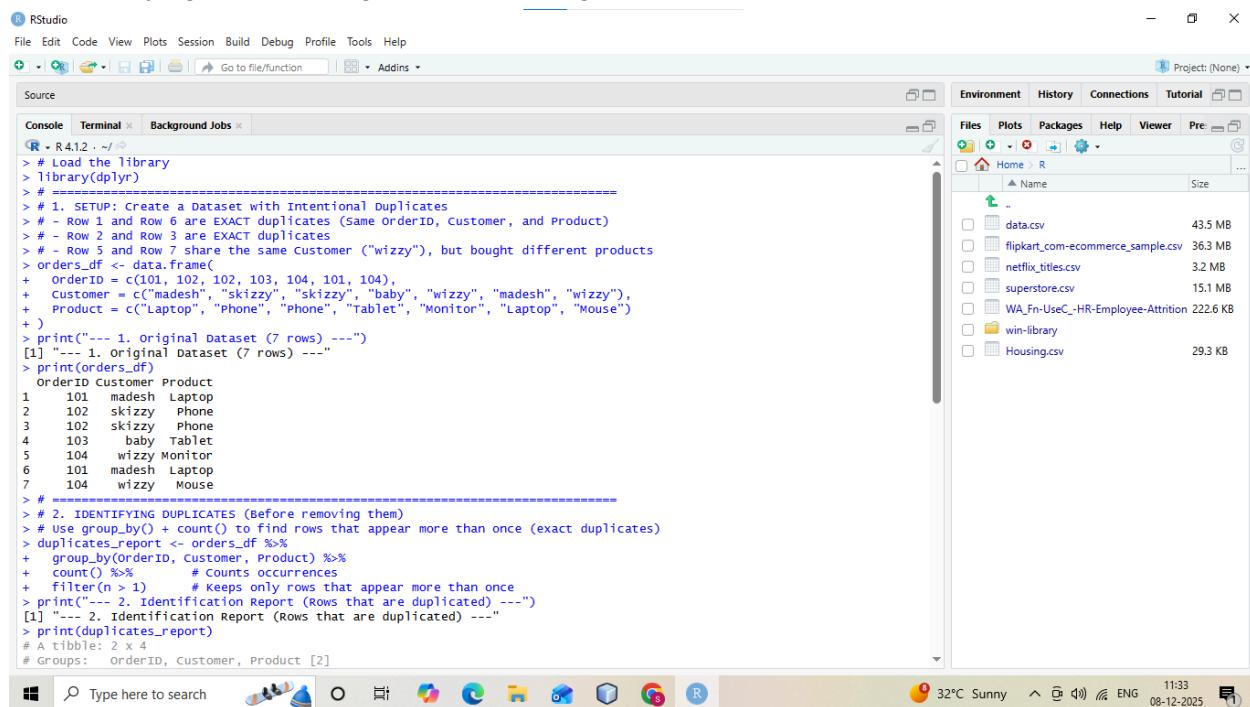


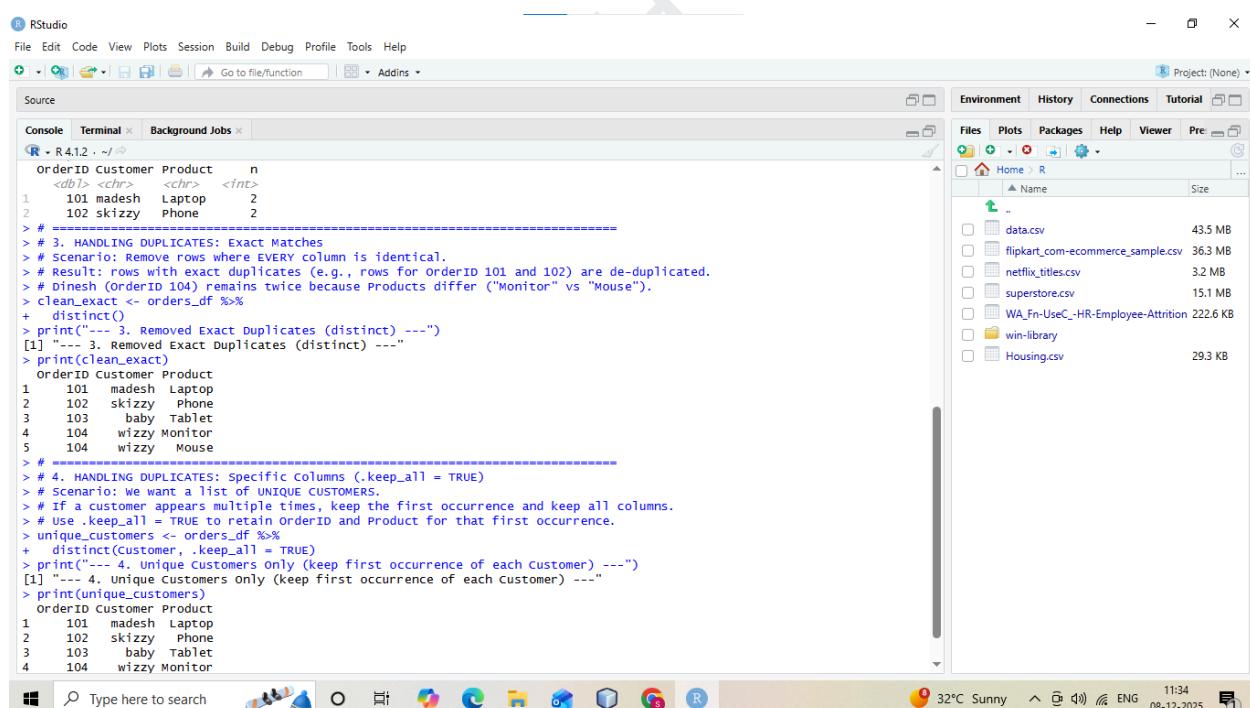
SHETH L.U.J. AND SIR M.V. COLLEGE

PRACTICAL NO 13

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



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R # Load the library
R library(dplyr)
R # =====
R # 1. SETUP: Create dataset with Intentional Duplicates
R # - Row 1 and Row 6 are EXACT duplicates (Same OrderID, Customer, and Product)
R # - Row 2 and Row 3 are EXACT duplicates
R # - Row 5 and Row 7 share the same Customer ("wizzy"), but bought different products
R 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")
+)
R print("--- 1. original dataset (7 rows) ---")
[1] "--- 1. original dataset (7 rows) ---"
R 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
R # =====
R # 2. IDENTIFYING DUPLICATES (Before removing them)
R # use group_by() + count() to find rows that appear more than once (exact duplicates)
R duplicates_report <- orders_df %>%
+ group_by(OrderID, Customer, Product) %>%
+ count() %>%
Counts occurrences
+ filter(n > 1) # keeps only rows that appear more than once
R print("--- 2. Identification Report (Rows that are duplicated) ---")
[1] "--- 2. Identification Report (Rows that are duplicated) ---"
R print(duplicates_report)
A tibble: 2 × 4
Groups: OrderID, Customer, Product [2]



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R # R 4.1.2 - ~/
R # 3. HANDLING DUPLICATES: Exact Matches
R # Scenario: Remove rows where EVERY column is identical.
R # Result: rows with exact duplicates (e.g., rows for orderID 101 and 102) are de-duplicated.
R # Dinesh (orderID 104) remains twice because Products differ ("Monitor" vs "Mouse").
R clean_exact <- orders_df %>%
+ distinct()
R print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
R 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
R # =====
R # 4. HANDLING DUPLICATES: specific columns (.keep_all = TRUE)
R # Scenario: We want a list of UNIQUE CUSTOMERS.
R # If a customer appears multiple times, keep the first occurrence and keep all columns.
R # Use .keep_all = TRUE to retain OrderID and Product for that first occurrence.
R unique_customers <- orders_df %>%
+ distinct(Customer, .keep_all = TRUE)
R print("--- 4. Unique customers only (keep first occurrence of each customer) ---")
[1] "--- 4. Unique customers only (keep first occurrence of each customer) ---"
R print(unique_customers)
OrderID Customer Product
1 101 madesh Laptop
2 102 skizzy Phone
3 103 baby Tablet
4 104 wizzy Monitor