

PRACTICAL NO : 13

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

```

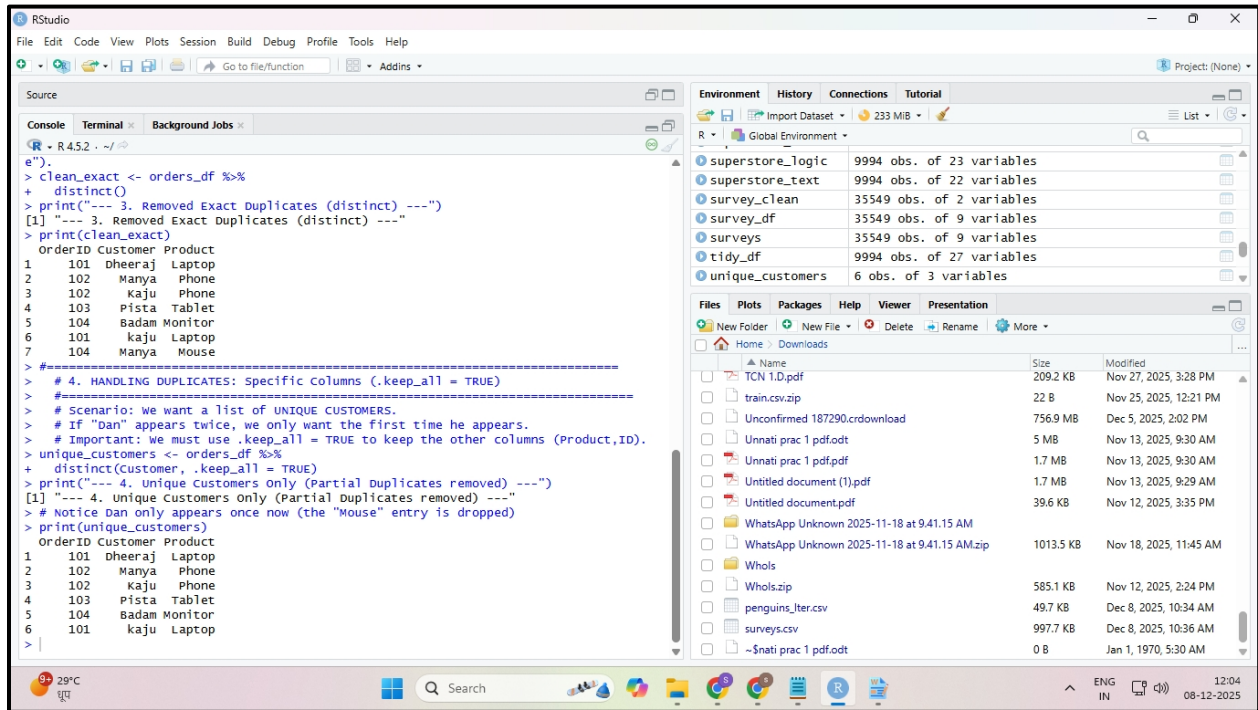
# R Script: Identifying and Handling duplicates
# Function: distinct() from the dplyr package
#
#=====
# Load the library
library(dplyr)
#=====
# 1. SETUP: Create a Dataset with Intentional Duplicates
#=====
# - Row 1 and Row 6 are EXACT duplicates (Same ID, Name, and Item)
# - Row 2 and Row 3 are EXACT duplicates
# - Row 5 and 7 share the same Name ("Dan"), but bought different items.
orders_df <- data.frame(
+   orderID = c(101, 102, 102, 103, 104, 101, 104),
+   Customer = c("Dheeraj", "Manyia", "Kaju", "Pista", "Badam", "kaju", "Manyia"),
+   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 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
#=====
# 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 x 4
# Groups:   orderID [db], 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

```

```

#=====
# 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

```



The screenshot displays the RStudio environment with the following components:

- Console:** Shows R code execution for cleaning data and handling duplicates. The output includes a list of unique customers after removing partial duplicates.
- Environment:** Lists loaded datasets with their dimensions.
- Files:** Shows the file explorer with various documents and folders.

```
e").
> 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

# 4. HANDLING DUPLICATES: Specific columns (.keep_all = TRUE)
# Scenario: We want a list of UNIQUE CUSTOMERS.
# If "Dan" appears twice, we only want the first time he appears.
# Important: We must use .keep_all = TRUE to keep the other columns (Product, ID).
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) ---"
# Notice Dan only appears once now (the "Mouse" entry is dropped)
> print(unique_customers)
  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
```

Dataset	Observations	Variables
superstore_logic	9994	23
superstore_text	9994	22
survey_clean	35549	2
survey_df	35549	9
surveys	35549	9
tidy_df	9994	27
unique_customers	6	3