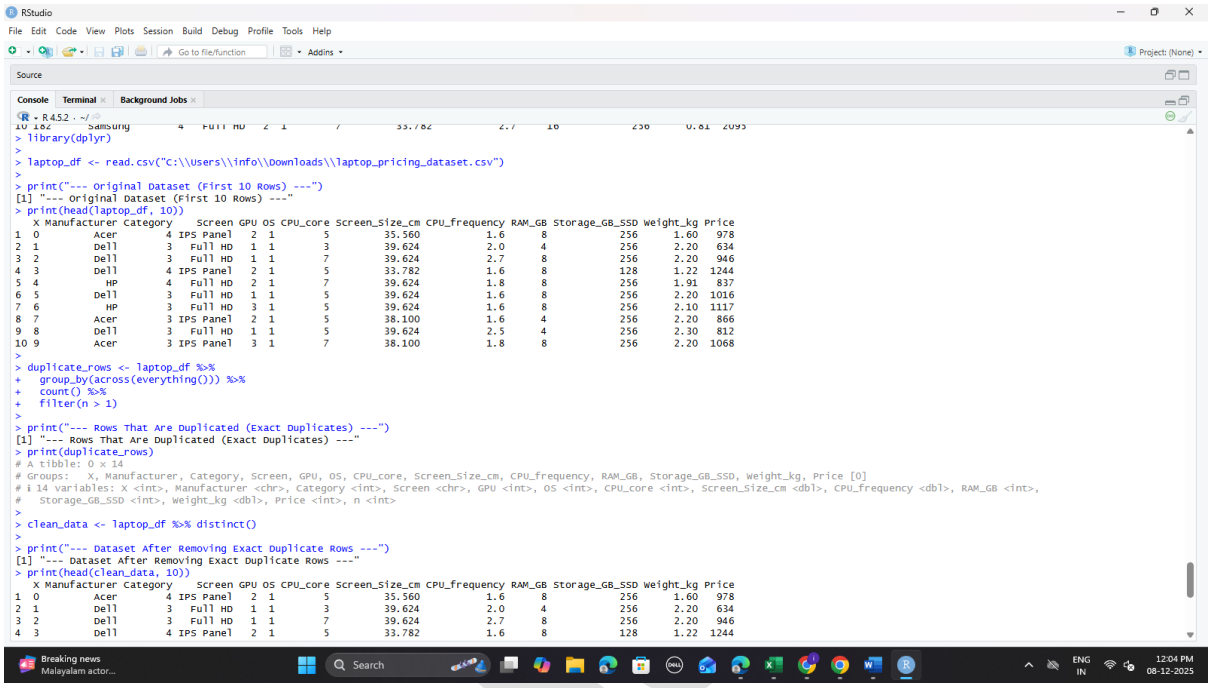
SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

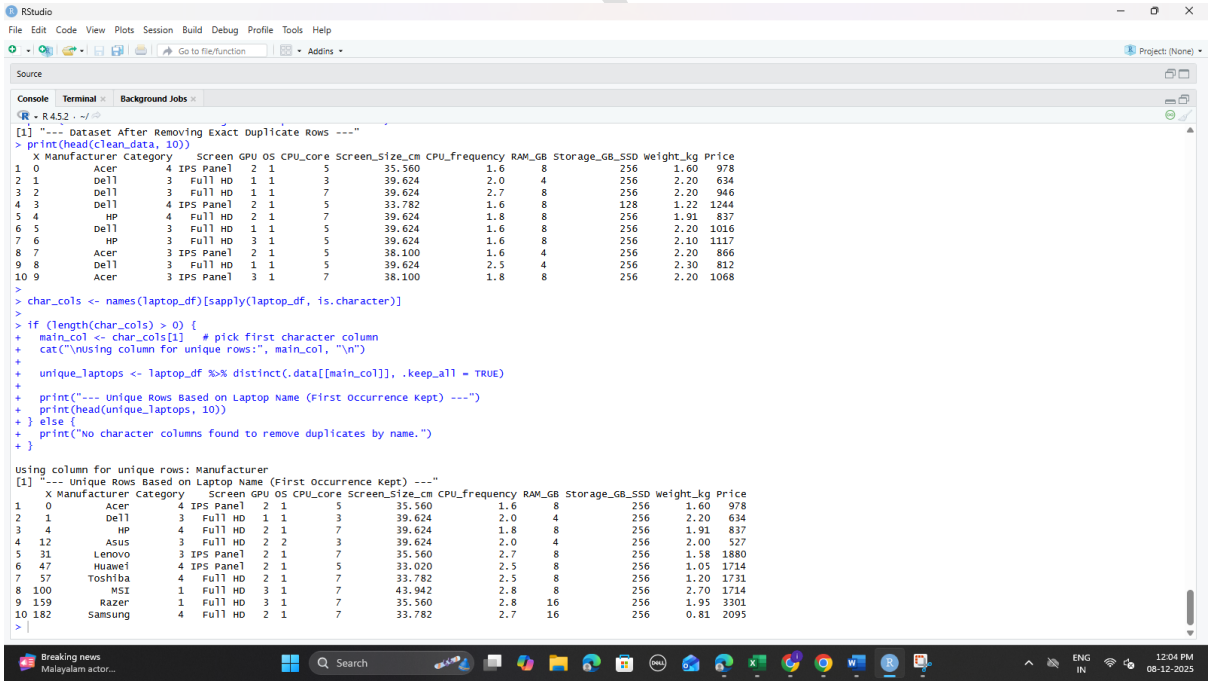
PRACTICAL – 13

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

OUTPUT:-



```
> library(dplyr)
>
> laptop_df <- read.csv("C:\\Users\\info\\Downloads\\laptop_pricing_dataset.csv")
>
> print("--- Original Dataset (First 10 Rows) ---")
[1] "--- Original Dataset (First 10 Rows) ---"
> print(head(laptop_df, 10))
   X Manufacturer Category    Screen GPU OS CPU_core Screen_Size_cm CPU_frequency RAM_GB Storage_GB_SSD Weight_kg Price
1  0         Acer        4 IPS Panel   2  1        5         35.560           1.6      8             256      1.60   978
2  1         Dell        3   Full HD   1  1        3         39.624           2.0      4             256      2.20   634
3  2         Dell        3   Full HD   1  1        7         39.624           2.7      8             256      2.20   946
4  3         Dell        4 IPS Panel   2  1        5         33.782           1.6      8             128      1.22  1244
5  4           HP        4   Full HD   2  1        7         39.624           1.8      8             256      1.91   837
6  5         Dell        3   Full HD   1  1        5         39.624           1.6      8             256      2.20  1016
7  6           HP        3   Full HD   3  1        5         39.624           1.6      8             256      2.10  1117
8  7         Acer        3 IPS Panel   2  1        5         38.100           1.6      4             256      2.20   866
9  8         Dell        3   Full HD   1  1        5         39.624           2.5      4             256      2.30   812
10 9         Acer        3 IPS Panel   3  1        7         38.100           1.8      8             256      2.20  1068
>
> duplicate_rows <- laptop_df %>%
+   group_by(across(everything())) %>%
+   count() %>%
+   filter(n > 1)
>
> print("--- Rows That Are Duplicated (Exact Duplicates) ---")
[1] "--- Rows That Are Duplicated (Exact Duplicates) ---"
> print(duplicate_rows)
# A tibble: 0 x 14
# Groups:   X, Manufacturer, Category, Screen, GPU, OS, CPU_core, Screen_Size_cm, CPU_frequency, RAM_GB, Storage_GB_SSD, Weight_kg, Price [0]
# i 14 variables: X <int>, Manufacturer <chr>, Category <int>, Screen <chr>, GPU <int>, OS <int>, CPU_core <int>, Screen_Size_cm <dbl>, CPU_frequency <dbl>, RAM_GB <int>,
#   Storage_GB_SSD <int>, Weight_kg <dbl>, Price <int>, n <int>
>
> clean_data <- laptop_df %>% distinct()
>
> print("--- Dataset After Removing Exact Duplicate Rows ---")
[1] "--- Dataset After Removing Exact Duplicate Rows ---"
> print(head(clean_data, 10))
   X Manufacturer Category    Screen GPU OS CPU_core Screen_Size_cm CPU_frequency RAM_GB Storage_GB_SSD Weight_kg Price
1  0         Acer        4 IPS Panel   2  1        5         35.560           1.6      8             256      1.60   978
2  1         Dell        3   Full HD   1  1        3         39.624           2.0      4             256      2.20   634
3  2         Dell        3   Full HD   1  1        7         39.624           2.7      8             256      2.20   946
4  3         Dell        4 IPS Panel   2  1        5         33.782           1.6      8             128      1.22  1244
```



```
[1] "--- Dataset After Removing Exact Duplicate Rows ---"
> print(head(clean_data, 10))
   X Manufacturer Category    Screen GPU OS CPU_core Screen_Size_cm CPU_frequency RAM_GB Storage_GB_SSD Weight_kg Price
1  0         Acer        4 IPS Panel   2  1        5         35.560           1.6      8             256      1.60   978
2  1         Dell        3   Full HD   1  1        3         39.624           2.0      4             256      2.20   634
3  2         Dell        3   Full HD   1  1        7         39.624           2.7      8             256      2.20   946
4  3         Dell        4 IPS Panel   2  1        5         33.782           1.6      8             128      1.22  1244
5  4           HP        4   Full HD   2  1        7         39.624           1.8      8             256      1.91   837
6  5         Dell        3   Full HD   1  1        5         39.624           1.6      8             256      2.20  1016
7  6           HP        3   Full HD   3  1        5         39.624           1.6      8             256      2.10  1117
8  7         Acer        3 IPS Panel   2  1        5         38.100           1.6      4             256      2.20   866
9  8         Dell        3   Full HD   1  1        5         39.624           2.5      4             256      2.30   812
10 9         Acer        3 IPS Panel   3  1        7         38.100           1.8      8             256      2.20  1068
>
> char_cols <- names(laptop_df)[sapply(laptop_df, is.character)]
>
> if (length(char_cols) > 0) {
+   main_col <- char_cols[1]   # pick first character column
+   cat("\nUsing column for unique rows:", main_col, "\n")
+
+   unique_laptops <- laptop_df %>% distinct(.data[[main_col]], .keep_all = TRUE)
+
+   print("--- Unique Rows Based on Laptop Name (First Occurrence Kept) ---")
+   print(head(unique_laptops, 10))
+ } else {
+   print("No character columns found to remove duplicates by name.")
+ }

Using column for unique rows: Manufacturer
[1] "--- Unique Rows Based on Laptop Name (First Occurrence Kept) ---"
     X Manufacturer Category    Screen GPU OS CPU_core Screen_Size_cm CPU_frequency RAM_GB Storage_GB_SSD Weight_kg Price
1    0         Acer        4 IPS Panel   2  1        5         35.560           1.6      8             256      1.60   978
2    1         Dell        3   Full HD   1  1        3         39.624           2.0      4             256      2.20   634
3    4           HP        4   Full HD   2  1        7         39.624           1.8      8             256      1.91   837
4   12         Asus        3   Full HD   2  2        3         39.624           2.0      4             256      2.00   527
5   31       Lenovo        3 IPS Panel   2  1        7         35.560           2.7      8             256      1.58  1880
6   47       Huawei        4 IPS Panel   2  1        5         33.020           2.5      8             256      1.05  1714
7   57      Toshiba        4   Full HD   2  1        7         33.782           2.5      8             256      1.20  1731
8  100          MSI        1   Full HD   3  1        7         43.942           2.8      8             256      2.70  1714
9  159        Razer        1   Full HD   3  1        7         35.560           2.8     16             256      1.95  3301
10 182      Samsung        4   Full HD   2  1        7         33.782           2.7     16             256      0.81  2095
>
```

SUMEET JITENDRA YADAV
S124