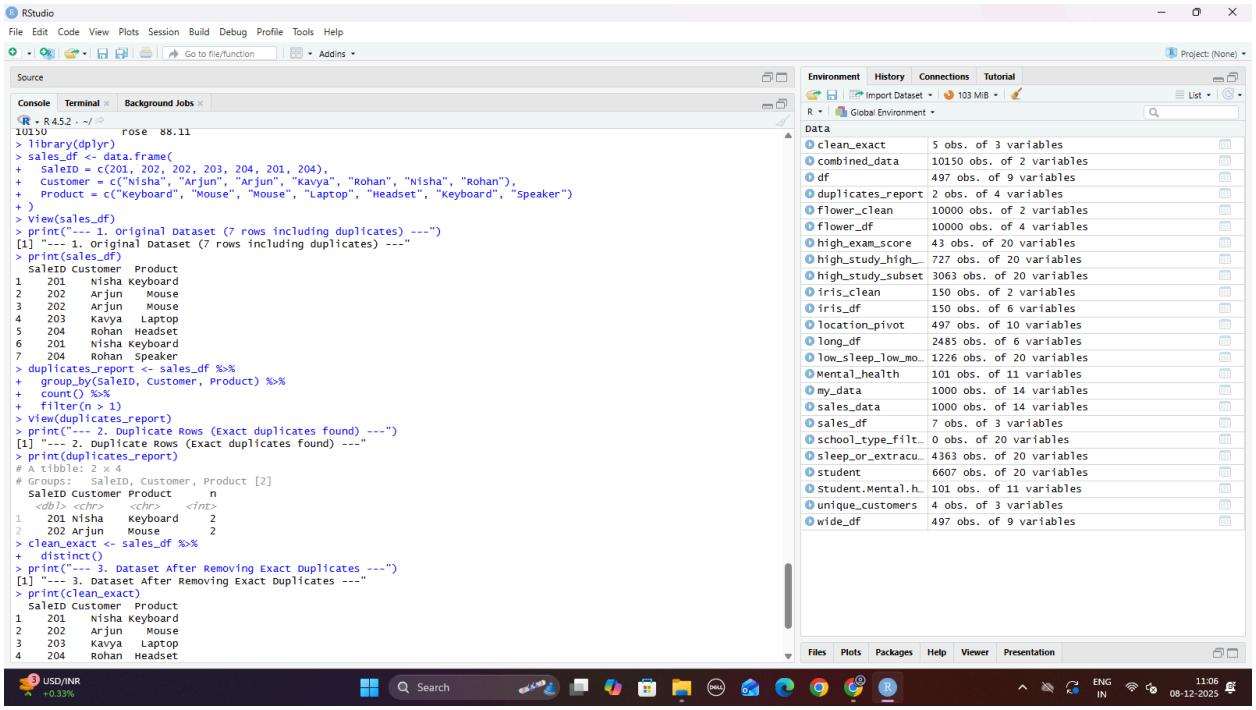
MVLU COLLEGE.

PRACTICAL NO :- 13
AIM :- Identifying and handling duplicates using distinct() (R studio).

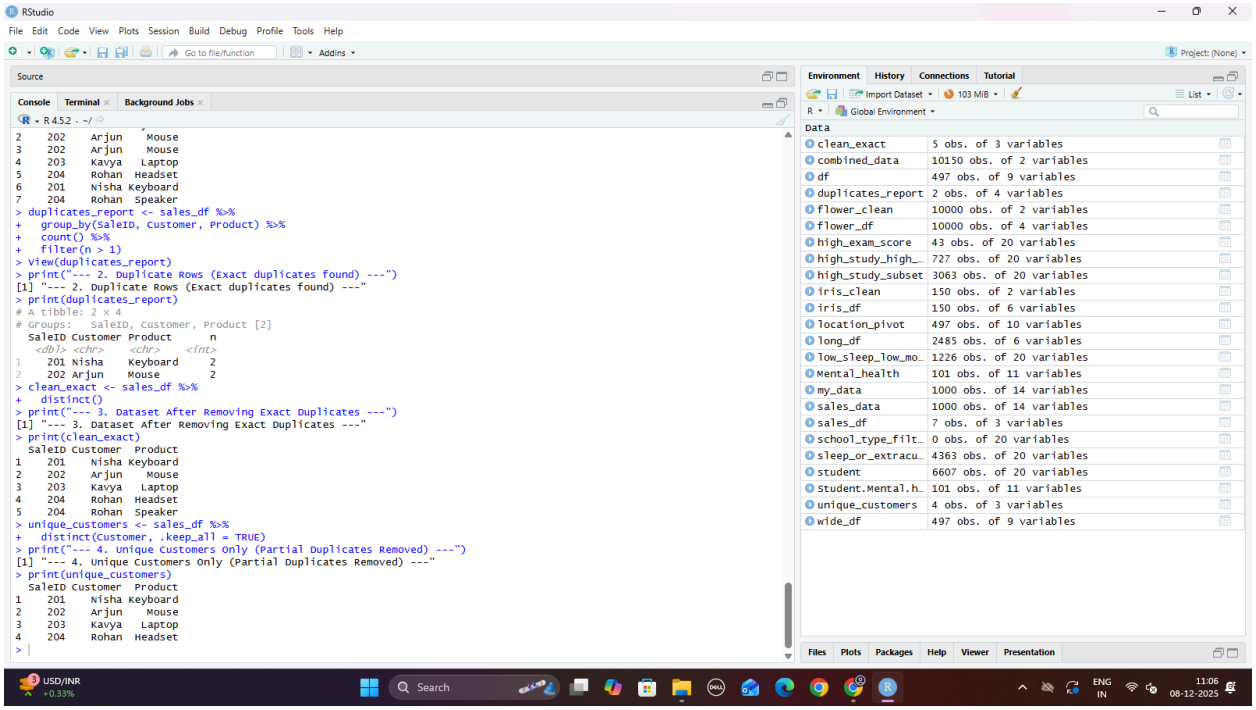
CODE :-

```
library(dplyr)

sales_df <- data.frame(
  SaleID = c(201, 202, 202, 203, 204, 201, 204),
  Customer = c("Nisha", "Arjun", "Arjun", "Kavya", "Rohan", "Nisha", "Rohan"),
  Product = c("Keyboard", "Mouse", "Mouse", "Laptop", "Headset", "Keyboard", "Speaker")
)

print("--- 1. Original Dataset (7 rows including duplicates) ---")
print(sales_df)

# 2. IDENTIFY DUPLICATES (before removing)

duplicates_report <- sales_df %>%
  group_by(SaleID, Customer, Product) %>%
  count() %>%
  filter(n > 1)

print("--- 2. Duplicate Rows (Exact duplicates found) ---")
print(duplicates_report)

# 3. REMOVE EXACT DUPLICATES

clean_exact <- sales_df %>%
  distinct()

print("--- 3. Dataset After Removing Exact Duplicates ---")
print(clean_exact)


# 4. REMOVE DUPLICATES BASED ON SPECIFIC COLUMN

unique_customers <- sales_df %>%
  distinct(Customer, .keep_all = TRUE)

print("--- 4. Unique Customers Only (Partial Duplicates Removed) ---")
print(unique_customers)
```

MVLU COLLEGE.



```
10150          rose  88.11
> library(dplyr)
> sales_df <- data.frame(
+   SaleID = c(201, 202, 202, 203, 204, 201, 204),
+   Customer = c("Nisha", "Arjun", "Arjun", "Kavya", "Rohan", "Nisha", "Rohan"),
+   Product = c("Keyboard", "Mouse", "Mouse", "Laptop", "Headset", "Keyboard", "Speaker")
+ )
> View(sales_df)
> print("--- 1. Original Dataset (7 rows including duplicates) ---")
[1] "--- 1. Original Dataset (7 rows including duplicates) ---"
> print(sales_df)
  SaleID Customer  Product
1    201    Nisha Keyboard
2    202    Arjun    Mouse
3    202    Arjun    Mouse
4    203    Kavya   Laptop
5    204    Rohan  Headset
6    201    Nisha Keyboard
7    204    Rohan  Speaker
> duplicates_report <- sales_df %>%
+   group_by(SaleID, Customer, Product) %>%
+   count() %>%
+   filter(n > 1)
> View(duplicates_report)
> print("--- 2. Duplicate Rows (Exact duplicates found) ---")
[1] "--- 2. Duplicate Rows (Exact duplicates found) ---"
> print(duplicates_report)
# A tibble: 2 x 4
# Groups:   SaleID, Customer, Product [2]
  SaleID Customer Product      n
   <dbl> <chr>    <chr>    <int>
1    201 Nisha    Keyboard     2
2    202 Arjun    Mouse        2
> clean_exact <- sales_df %>%
+   distinct()
> print("--- 3. Dataset After Removing Exact Duplicates ---")
[1] "--- 3. Dataset After Removing Exact Duplicates ---"
> print(clean_exact)
  SaleID Customer  Product
1    201    Nisha Keyboard
2    202    Arjun    Mouse
3    203    Kavya   Laptop
4    204    Rohan  Headset
```



```
2    202    Arjun    Mouse
3    202    Arjun    Mouse
4    203    Kavya   Laptop
5    204    Rohan  Headset
6    201    Nisha Keyboard
7    204    Rohan  Speaker
> duplicates_report <- sales_df %>%
+   group_by(SaleID, Customer, Product) %>%
+   count() %>%
+   filter(n > 1)
> View(duplicates_report)
> print("--- 2. Duplicate Rows (Exact duplicates found) ---")
[1] "--- 2. Duplicate Rows (Exact duplicates found) ---"
> print(duplicates_report)
# A tibble: 2 x 4
# Groups:   SaleID, Customer, Product [2]
  SaleID Customer Product      n
   <dbl> <chr>    <chr>    <int>
1    201 Nisha    Keyboard     2
2    202 Arjun    Mouse        2
> clean_exact <- sales_df %>%
+   distinct()
> print("--- 3. Dataset After Removing Exact Duplicates ---")
[1] "--- 3. Dataset After Removing Exact Duplicates ---"
> print(clean_exact)
  SaleID Customer  Product
1    201    Nisha Keyboard
2    202    Arjun    Mouse
3    203    Kavya   Laptop
4    204    Rohan  Headset
5    204    Rohan  Speaker
> unique_customers <- sales_df %>%
+   distinct(Customer, .keep_all = TRUE)
> print("--- 4. Unique Customers Only (Partial Duplicates Removed) ---")
[1] "--- 4. Unique Customers Only (Partial Duplicates Removed) ---"
> print(unique_customers)
  SaleID Customer  Product
1    201    Nisha Keyboard
2    202    Arjun    Mouse
3    203    Kavya   Laptop
4    204    Rohan  Headset
>
```

NANDINI PANDIT S100
DATA ANALYSIS WITH SAS/SPSS/R.