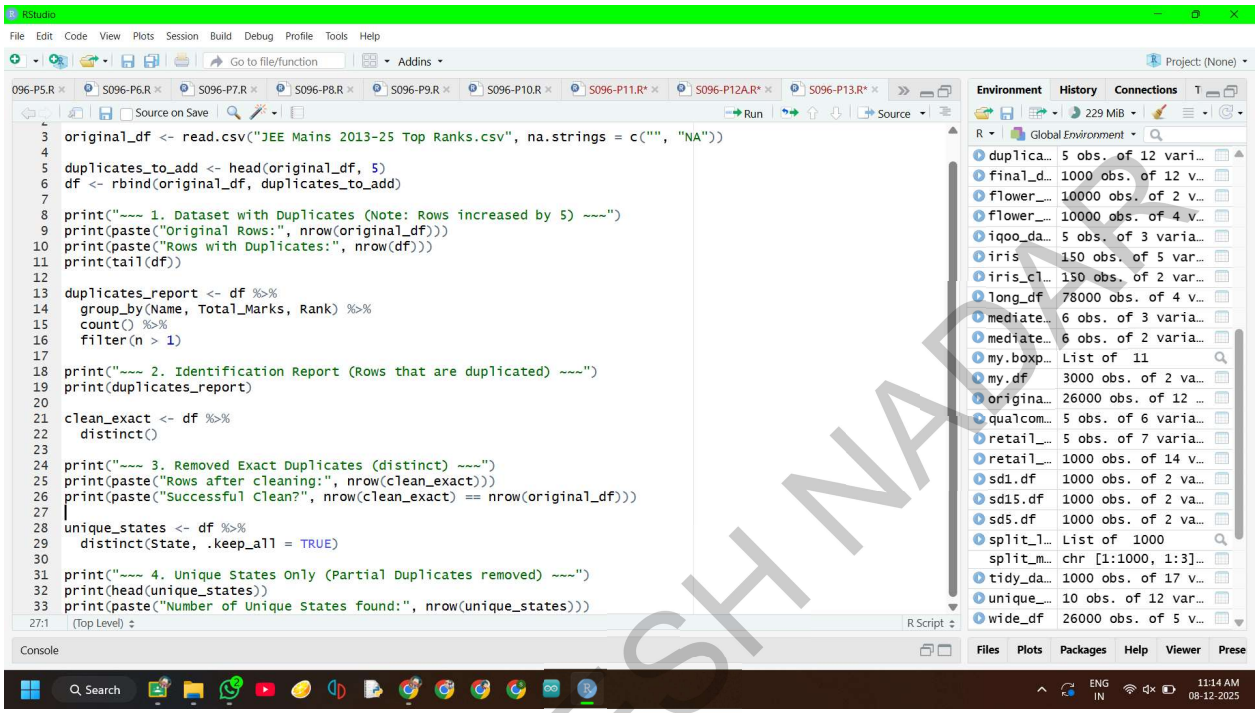
SHETH L.U.J. AND SIR M.V. COLLEGE
DATA ANALYSIS WITH SAS/SPSS/R

PRACTICAL NO: 13

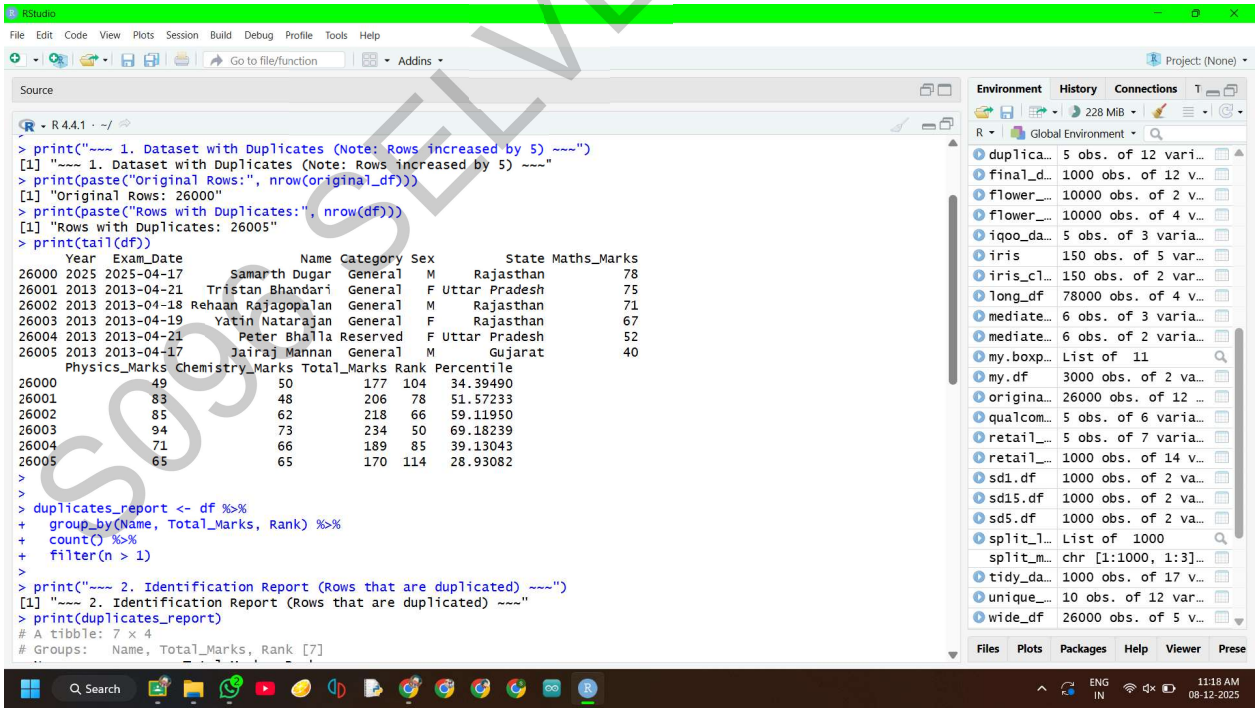
AIM: Identifying and handling duplicates using distinct() (R).

CODE:

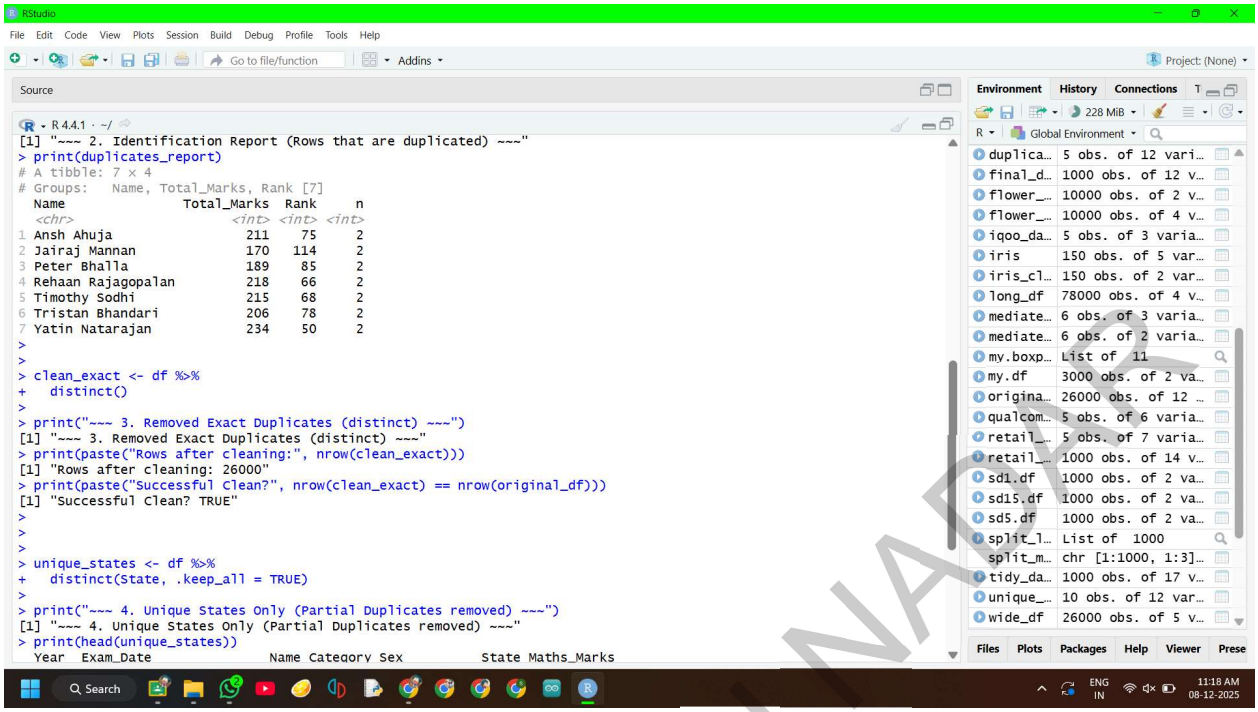


```
original_df <- read.csv("JEE Mains 2013-25 Top Ranks.csv", na.strings = c("", "NA"))

duplicates_to_add <- head(original_df, 5)
df <- rbind(original_df, duplicates_to_add)

print("~~~ 1. Dataset with Duplicates (Note: Rows increased by 5) ~~~")
print(paste("Original Rows:", nrow(original_df)))
print(paste("Rows with Duplicates:", nrow(df)))
print(tail(df))

duplicates_report <- df %>%
  group_by(Name, Total_Marks, Rank) %>%
  count() %>%
  filter(n > 1)

print("~~~ 2. Identification Report (Rows that are duplicated) ~~~")
print(duplicates_report)

clean_exact <- df %>%
  distinct()

print("~~~ 3. Removed Exact Duplicates (distinct) ~~~")
print(paste("Rows after cleaning:", nrow(clean_exact)))
print(paste("Successful Clean?", nrow(clean_exact) == nrow(original_df)))

unique_states <- df %>%
  distinct(State, .keep_all = TRUE)

print("~~~ 4. Unique States Only (Partial Duplicates removed) ~~~")
print(head(unique_states))
print(paste("Number of Unique States found:", nrow(unique_states)))
```

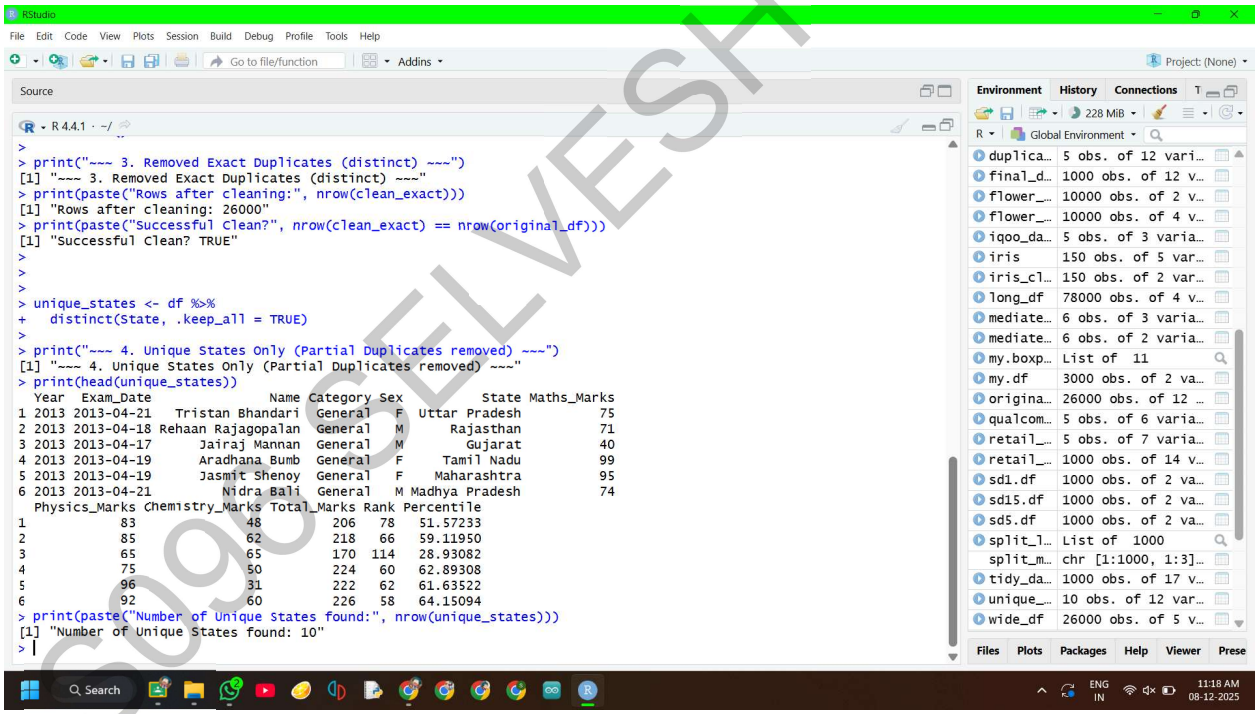
OUTPUT:



```
> print("~~~ 1. Dataset with Duplicates (Note: Rows increased by 5) ~~~")
[1] "~~~ 1. Dataset with Duplicates (Note: Rows increased by 5) ~~~"
> print(paste("Original Rows:", nrow(original_df)))
[1] "Original Rows: 26000"
> print(paste("Rows with Duplicates:", nrow(df)))
[1] "Rows with Duplicates: 26005"
> print(tail(df))
      Year  Exam_Date              Name Category Sex         State Maths_Marks
26000 2025 2025-04-17      Samarth Dugar  General   M     Rajasthan          78
26001 2013 2013-04-21   Tristan Bhandari  General   F Uttar Pradesh          75
26002 2013 2013-04-18 Rehaan Rajagopalan  General   M     Rajasthan          71
26003 2013 2013-04-19     Yatin Natarajan  General   F     Rajasthan          67
26004 2013 2013-04-21       Peter Bhalla Reserved   F Uttar Pradesh          52
26005 2013 2013-04-17      Jairaj Mannan  General   M       Gujarat          40
      Physics_Marks Chemistry_Marks Total_Marks Rank Percentile
26000            49              50         177  104   34.39490
26001            83              48         206   78   51.57233
26002            85              62         218   66   59.11950
26003            94              73         234   50   69.18239
26004            71              66         189   85   39.13043
26005            65              65         170  114   28.93082
>
>
> duplicates_report <- df %>%
+   group_by(Name, Total_Marks, Rank) %>%
+   count() %>%
+   filter(n > 1)
>
> print("~~~ 2. Identification Report (Rows that are duplicated) ~~~")
[1] "~~~ 2. Identification Report (Rows that are duplicated) ~~~"
> print(duplicates_report)
# A tibble: 7 x 4
# Groups:   Name, Total_Marks, Rank [7]
```



```
[1] "~~~ 2. Identification Report (Rows that are duplicated) ~~~"
> print(duplicates_report)
# A tibble: 7 x 4
# Groups:   Name, Total_Marks, Rank [7]
  Name              Total_Marks  Rank     n
  <chr>                   <int> <int> <int>
1 Ansh Ahuja                211    75     2
2 Jairaj Mannan             170   114     2
3 Peter Bhalla              189    85     2
4 Rehaan Rajagopalan        218    66     2
5 Timothy Sodhi             215    68     2
6 Tristan Bhandari          206    78     2
7 Yatin Natarajan           234    50     2
>
>
> clean_exact <- df %>%
+   distinct()
>
> print("~~~ 3. Removed Exact Duplicates (distinct) ~~~")
[1] "~~~ 3. Removed Exact Duplicates (distinct) ~~~"
> print(paste("Rows after cleaning:", nrow(clean_exact)))
[1] "Rows after cleaning: 26000"
> print(paste("Successful Clean?", nrow(clean_exact) == nrow(original_df)))
[1] "Successful Clean? TRUE"
>
>
>
> unique_states <- df %>%
+   distinct(State, .keep_all = TRUE)
>
> print("~~~ 4. Unique States Only (Partial Duplicates removed) ~~~")
[1] "~~~ 4. Unique States Only (Partial Duplicates removed) ~~~"
> print(head(unique_states))
  Year  Exam_Date             Name Category Sex          State Maths_Marks
```



```
>
> print("~~~ 3. Removed Exact Duplicates (distinct) ~~~")
[1] "~~~ 3. Removed Exact Duplicates (distinct) ~~~"
> print(paste("Rows after cleaning:", nrow(clean_exact)))
[1] "Rows after cleaning: 26000"
> print(paste("Successful Clean?", nrow(clean_exact) == nrow(original_df)))
[1] "Successful Clean? TRUE"
>
>
>
> unique_states <- df %>%
+   distinct(State, .keep_all = TRUE)
>
> print("~~~ 4. Unique States Only (Partial Duplicates removed) ~~~")
[1] "~~~ 4. Unique States Only (Partial Duplicates removed) ~~~"
> print(head(unique_states))
  Year  Exam_Date             Name Category Sex          State Maths_Marks
1 2013 2013-04-21   Tristan Bhandari  General   F  Uttar Pradesh          75
2 2013 2013-04-18 Rehaan Rajagopalan  General   M      Rajasthan          71
3 2013 2013-04-17      Jairaj Mannan  General   M        Gujarat          40
4 2013 2013-04-19      Aradhana Bumb  General   F     Tamil Nadu          99
5 2013 2013-04-19      Jasmit Shenoy  General   F    Maharashtra          95
6 2013 2013-04-21         Nidra Bali  General   M Madhya Pradesh          74
  Physics_Marks Chemistry_Marks Total_Marks Rank Percentile
1            83             48         206   78   51.57233
2            85             62         218   66   59.11950
3            65             65         170  114   28.93082
4            75             50         224   60   62.89308
5            96             31         222   62   61.63522
6            92             60         226   58   64.15094
> print(paste("Number of Unique States found:", nrow(unique_states)))
[1] "Number of Unique States found: 10"
>
```