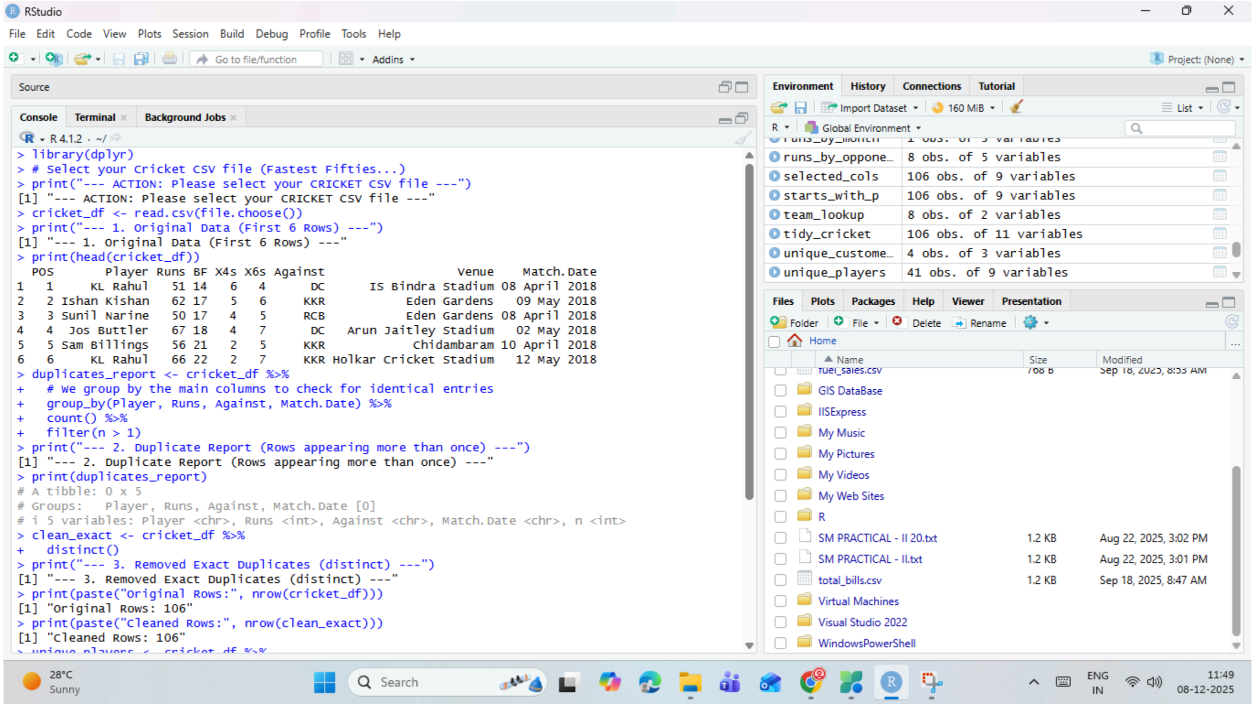
MVLU
R COLLEGE 13

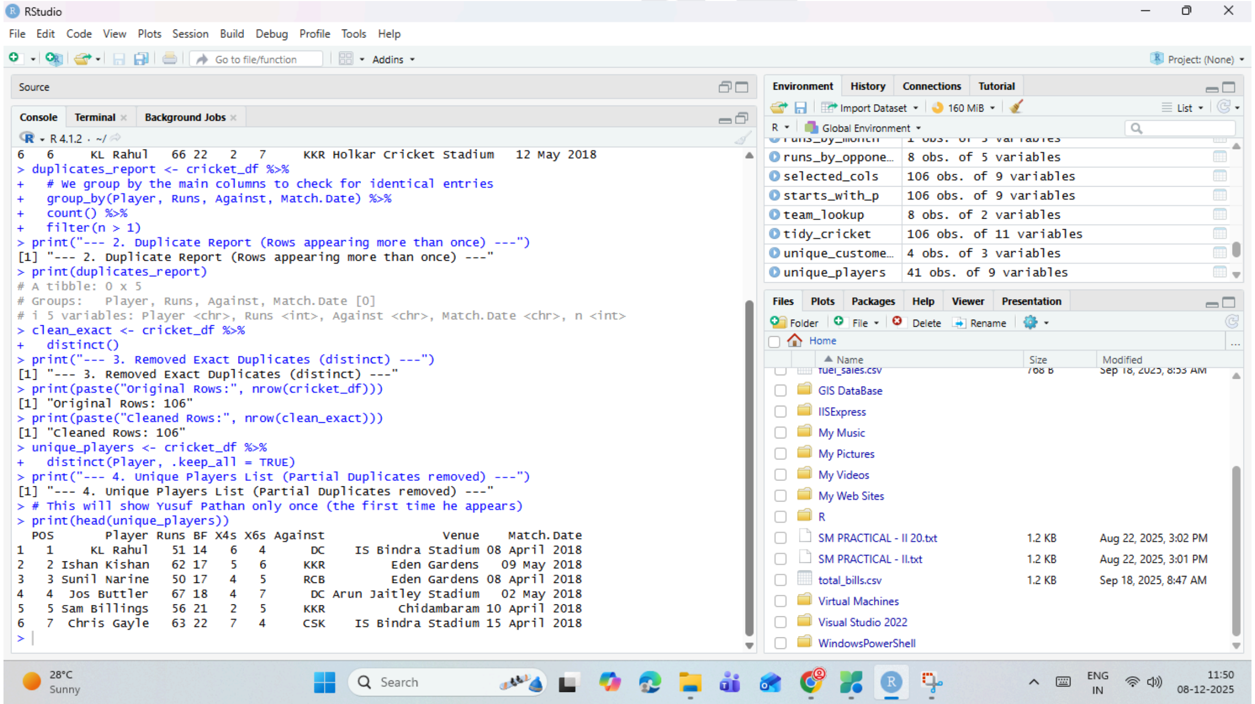
Aim:Identifying and handling duplicates using distinct() (R studio).
Output:



```
> library(dplyr)
> # Select your Cricket CSV file (Fastest Fifties...)
> print("--- ACTION: Please select your CRICKET CSV file ---")
[1] "--- ACTION: Please select your CRICKET CSV file ---"
> cricket_df <- read.csv(file.choose())
> print("--- 1. Original Data (First 6 Rows) ---")
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(cricket_df))
  POS       Player Runs BF X4s X6s Against                  Venue    Match.Date
1   1     KL Rahul   51 14   6   4      DC     IS Bindra Stadium 08 April 2018
2   2 Ishan Kishan   62 17   5   6     KKR          Eden Gardens   09 May 2018
3   3 Sunil Narine   50 17   4   5     RCB          Eden Gardens 08 April 2018
4   4  Jos Buttler   67 18   4   7      DC   Arun Jaitley Stadium   02 May 2018
5   5 Sam Billings   56 21   2   5     KKR            Chidambaram 10 April 2018
6   6     KL Rahul   66 22   2   7     KKR Holkar Cricket Stadium   12 May 2018
> duplicates_report <- cricket_df %>%
+   # We group by the main columns to check for identical entries
+   group_by(Player, Runs, Against, Match.Date) %>%
+   count() %>%
+   filter(n > 1)
> print("--- 2. Duplicate Report (Rows appearing more than once) ---")
[1] "--- 2. Duplicate Report (Rows appearing more than once) ---"
> print(duplicates_report)
# A tibble: 0 x 5
# Groups:   Player, Runs, Against, Match.Date [0]
# i 5 variables: Player <chr>, Runs <int>, Against <chr>, Match.Date <chr>, n <int>
> clean_exact <- cricket_df %>%
+   distinct()
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
> print(paste("Original Rows:", nrow(cricket_df)))
[1] "Original Rows: 106"
> print(paste("Cleaned Rows:", nrow(clean_exact)))
[1] "Cleaned Rows: 106"
```

```
runs_by_oppone…   8 obs. of 5 variables
selected_cols     106 obs. of 9 variables
starts_with_p     106 obs. of 9 variables
team_lookup       8 obs. of 2 variables
tidy_cricket      106 obs. of 11 variables
unique_custome…   4 obs. of 3 variables
unique_players    41 obs. of 9 variables
```

```
GIS DataBase
IISExpress
My Music
My Pictures
My Videos
My Web Sites
R
SM PRACTICAL - II 20.txt   1.2 KB   Aug 22, 2025, 3:02 PM
SM PRACTICAL - II.txt      1.2 KB   Aug 22, 2025, 3:01 PM
total_bills.csv            1.2 KB   Sep 18, 2025, 8:47 AM
Virtual Machines
Visual Studio 2022
WindowsPowerShell
```



```
6   6     KL Rahul   66 22   2   7     KKR Holkar Cricket Stadium   12 May 2018
> duplicates_report <- cricket_df %>%
+   # We group by the main columns to check for identical entries
+   group_by(Player, Runs, Against, Match.Date) %>%
+   count() %>%
+   filter(n > 1)
> print("--- 2. Duplicate Report (Rows appearing more than once) ---")
[1] "--- 2. Duplicate Report (Rows appearing more than once) ---"
> print(duplicates_report)
# A tibble: 0 x 5
# Groups:   Player, Runs, Against, Match.Date [0]
# i 5 variables: Player <chr>, Runs <int>, Against <chr>, Match.Date <chr>, n <int>
> clean_exact <- cricket_df %>%
+   distinct()
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
> print(paste("Original Rows:", nrow(cricket_df)))
[1] "Original Rows: 106"
> print(paste("Cleaned Rows:", nrow(clean_exact)))
[1] "Cleaned Rows: 106"
> unique_players <- cricket_df %>%
+   distinct(Player, .keep_all = TRUE)
> print("--- 4. Unique Players List (Partial Duplicates removed) ---")
[1] "--- 4. Unique Players List (Partial Duplicates removed) ---"
> # This will show Yusuf Pathan only once (the first time he appears)
> print(head(unique_players))
  POS       Player Runs BF X4s X6s Against                 Venue    Match.Date
1   1     KL Rahul   51 14   6   4      DC    IS Bindra Stadium 08 April 2018
2   2 Ishan Kishan   62 17   5   6     KKR         Eden Gardens   09 May 2018
3   3 Sunil Narine   50 17   4   5     RCB         Eden Gardens 08 April 2018
4   4  Jos Buttler   67 18   4   7      DC Arun Jaitley Stadium   02 May 2018
5   5 Sam Billings   56 21   2   5     KKR          Chidambaram 10 April 2018
6   7  Chris Gayle   63 22   7   4     CSK    IS Bindra Stadium 15 April 2018
> |
```

```
runs_by_oppone…   8 obs. of 5 variables
selected_cols     106 obs. of 9 variables
starts_with_p     106 obs. of 9 variables
team_lookup       8 obs. of 2 variables
tidy_cricket      106 obs. of 11 variables
unique_custome…   4 obs. of 3 variables
unique_players    41 obs. of 9 variables
```

```
GIS DataBase
IISExpress
My Music
My Pictures
My Videos
My Web Sites
R
SM PRACTICAL - II 20.txt   1.2 KB   Aug 22, 2025, 3:02 PM
SM PRACTICAL - II.txt      1.2 KB   Aug 22, 2025, 3:01 PM
total_bills.csv            1.2 KB   Sep 18, 2025, 8:47 AM
Virtual Machines
Visual Studio 2022
WindowsPowerShell
```

Hariprasad Vishwakarma
S126
SYCS