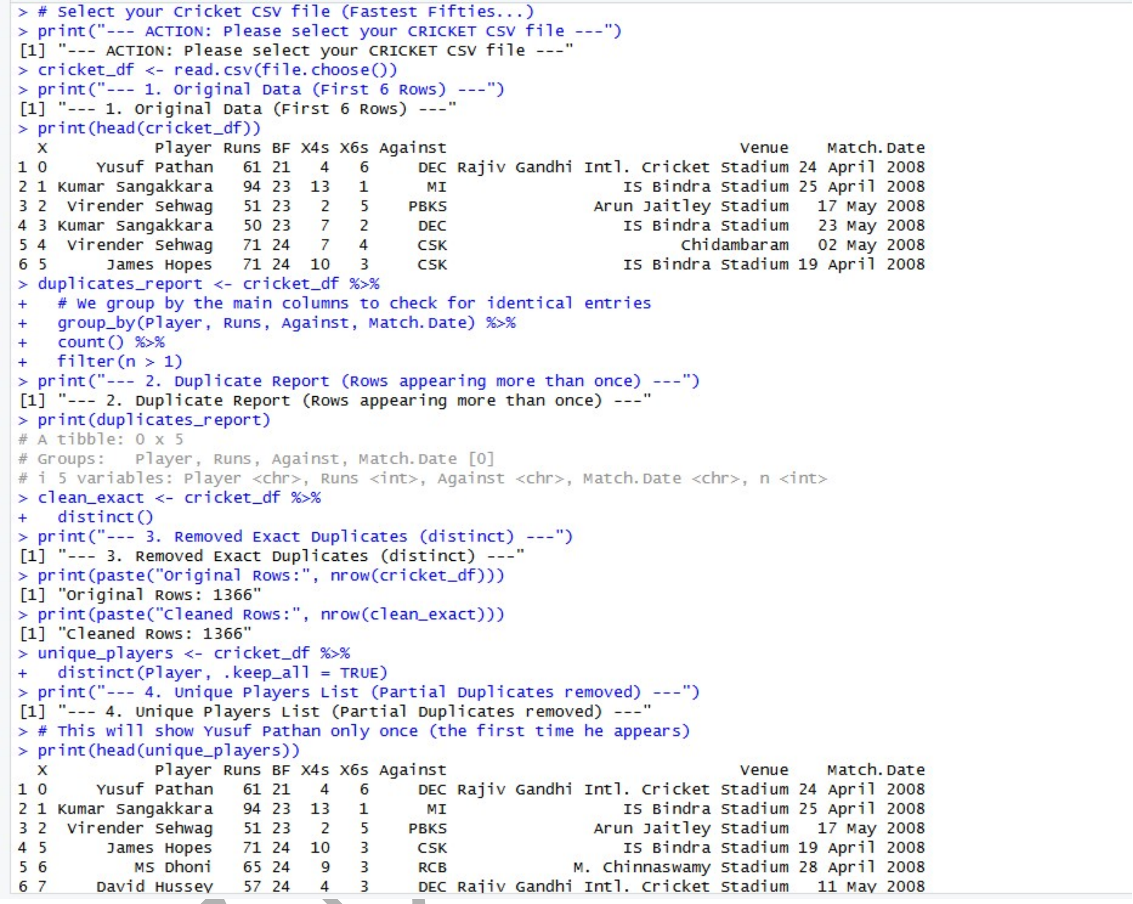
SHETH L.U.J AND SIR M.V. COLLEGE
SUBJECT NAME: **Data Analysis with SAS / SPSS /R**
Practical No. 13

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

Output:



```
> # 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))
  X            Player Runs BF X4s X6s Against                              Venue    Match.Date
1 0     Yusuf Pathan   61 21   4   6     DEC Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 1 Kumar Sangakkara   94 23  13   1      MI                  IS Bindra Stadium 25 April 2008
3 2  Virender Sehwag   51 23   2   5    PBKS               Arun Jaitley Stadium   17 May 2008
4 3 Kumar Sangakkara   50 23   7   2     DEC                  IS Bindra Stadium   23 May 2008
5 4  Virender Sehwag   71 24   7   4     CSK                        Chidambaram   02 May 2008
6 5      James Hopes   71 24  10   3     CSK                  IS Bindra Stadium 19 April 2008
> 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: 1366"
> print(paste("Cleaned Rows:", nrow(clean_exact)))
[1] "Cleaned Rows: 1366"
> 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))
  X            Player Runs BF X4s X6s Against                              Venue    Match.Date
1 0     Yusuf Pathan   61 21   4   6     DEC Rajiv Gandhi Intl. Cricket Stadium 24 April 2008
2 1 Kumar Sangakkara   94 23  13   1      MI                  IS Bindra Stadium 25 April 2008
3 2  Virender Sehwag   51 23   2   5    PBKS               Arun Jaitley Stadium   17 May 2008
4 5      James Hopes   71 24  10   3     CSK                  IS Bindra Stadium 19 April 2008
5 6         MS Dhoni   65 24   9   3     RCB             M. Chinnaswamy Stadium 28 April 2008
6 7     David Hussey   57 24   4   3     DEC Rajiv Gandhi Intl. Cricket Stadium   11 May 2008
```