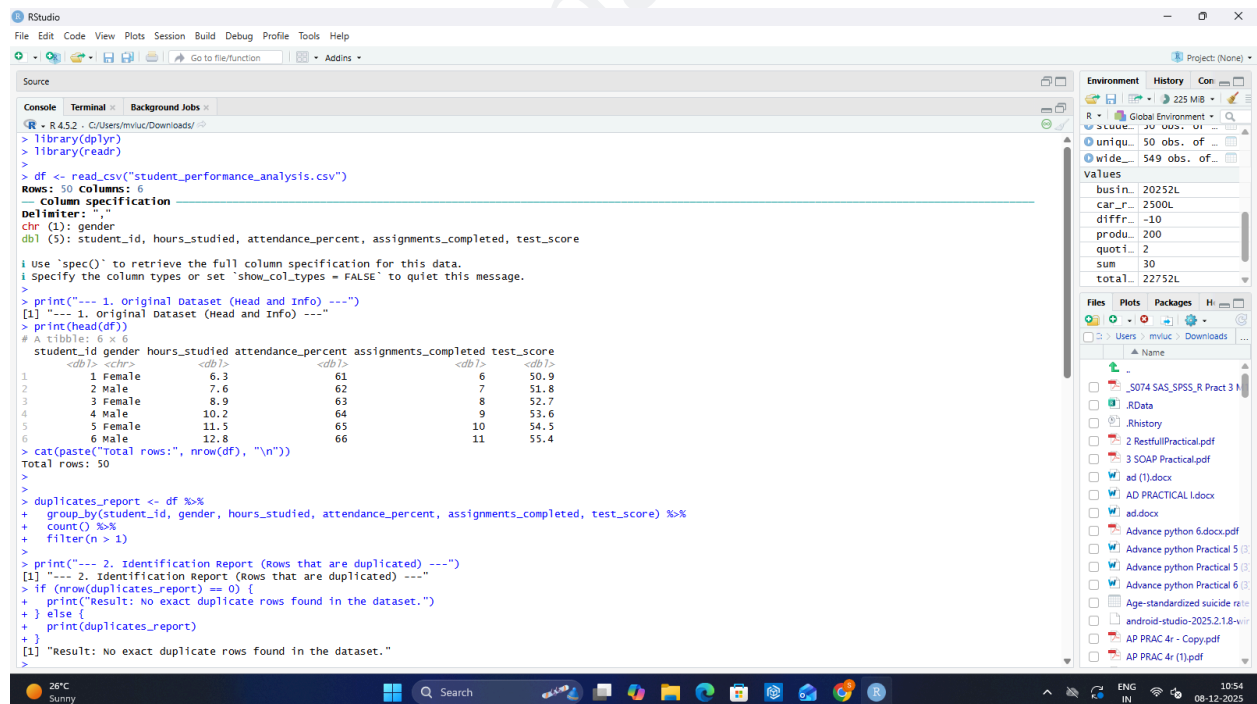


SHETH L.U.J AND SIR M.V. COLLEGE
SUBJECT NAME: DATA ANALYSIS WITH SAS/SPSS/R

Module 1 Practical 13

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

OUTPUT:



```
R - R 4.5.2 - C:/Users/mvut/Downloads/
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Terminal Background Jobs
> library(dplyr)
> library(readr)
> df <- read_csv("student_performance_analysis.csv")
Rows: 50 Columns: 6
Column specification
Delimiter: ","
chr (1): gender
dbl (5): student_id, hours_studied, attendance_percent, assignments_completed, test_score
i use 'spec()' to retrieve the full column specification for this data.
i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
> print("---- 1. Original Dataset (Head and Info) ----")
[1] "---- 1. Original Dataset (Head and Info) ----"
> print(head(df))
# A tibble: 6 x 6
  student_id gender hours_studied attendance_percent assignments_completed test_score
  <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
1 1 Female 6.3 61 6 50.9
2 2 Male 7.6 62 7 51.8
3 3 Female 8.9 63 8 52.7
4 4 Male 10.2 64 9 53.6
5 5 Female 11.5 65 10 54.5
6 6 Male 12.8 66 11 55.4
> cat(paste("Total rows:", nrow(df), "\n"))
Total rows: 50
>
> duplicates_report <- df %>%
+   group_by(student_id, gender, hours_studied, attendance_percent, assignments_completed, test_score) %>%
+   count() %>%
+   filter(n > 1)
>
> print("---- 2. Identification Report (Rows that are duplicated) ----")
[1] "---- 2. Identification Report (Rows that are duplicated) ----"
> if (nrow(duplicates_report) == 0) {
+   print("Result: No exact duplicate rows found in the dataset.")
+ } else {
+   print(duplicates_report)
+ }
[1] "Result: No exact duplicate rows found in the dataset."
```

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

SUBJECT NAME: DATA ANALYSIS WITH SAS/SPSS/R

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R452 - C:/Users/mvuc/Downloads/
> cat(paste("Total rows:", nrow(df), "\n"))
Total rows: 50
>
> duplicates_report <- df %>%
+ group_by(student_id, gender, hours_studied, attendance_percent, assignments_completed, test_score) %>%
+ count() %>%
+ filter(n > 1)
>
> print("---- 2. Identification Report (Rows that are duplicated) ----")
[1] "---- 2. Identification Report (Rows that are duplicated) ----"
> if (nrow(duplicates_report) == 0) {
+   print("Result: No exact duplicate rows found in the dataset.")
+ } else {
+   print(duplicates_report)
+ }
[1] "Result: No exact duplicate rows found in the dataset."
>
> clean_exact <- df %>%
+ distinct()
>
> print("---- 3. Removed Exact Duplicates (distinct) ----")
[1] "---- 3. Removed Exact Duplicates (distinct) ----"
> cat(paste("Original rows:", nrow(df), "\n"))
Original rows: 50
> cat(paste("Clean exact rows:", nrow(clean_exact), "\n"))
Clean exact rows: 50
> print(head(clean_exact, 10))
# A tibble: 10 x 6
  student_id gender hours_studied attendance_percent assignments_completed test_score
  <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
1 1 Female 6.3 61 6 50.9
2 2 Male 7.6 62 7 51.8
3 3 Female 8.9 63 8 52.7
4 4 Male 10.2 64 9 53.6
5 5 Female 11.5 65 10 54.5
6 6 Male 12.8 66 11 55.4
7 7 Female 14.1 67 12 56.3
8 8 Male 15.4 68 13 57.2
9 9 Female 16.7 69 14 58.1
10 10 Male 5 70 5 59
>
26°C Sunny
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Terminal Background Jobs
R - R452 - C:/Users/mvuc/Downloads/
> cat(paste("Clean exact rows:", nrow(clean_exact), "\n"))
Clean exact rows: 50
> print(head(clean_exact, 10))
# A tibble: 10 x 6
  student_id gender hours_studied attendance_percent assignments_completed test_score
  <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
1 1 Female 6.3 61 6 50.9
2 2 Male 7.6 62 7 51.8
3 3 Female 8.9 63 8 52.7
4 4 Male 10.2 64 9 53.6
5 5 Female 11.5 65 10 54.5
6 6 Male 12.8 66 11 55.4
7 7 Female 14.1 67 12 56.3
8 8 Male 15.4 68 13 57.2
9 9 Female 16.7 69 14 58.1
10 10 Male 5 70 5 59
>
> unique_students <- df %>%
+ distinct(student_id, .keep_all = TRUE)
>
> print("---- 4. Unique Students Only (Partial Duplicates removed) ----")
[1] "---- 4. Unique Students Only (Partial Duplicates removed) ----"
> cat(paste("Original rows:", nrow(df), "\n"))
Original rows: 50
> cat(paste("Unique student ID rows:", nrow(unique_students), "\n"))
Unique student ID rows: 50
> print(head(unique_students, 10))
# A tibble: 10 x 6
  student_id gender hours_studied attendance_percent assignments_completed test_score
  <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
1 1 Female 6.3 61 6 50.9
2 2 Male 7.6 62 7 51.8
3 3 Female 8.9 63 8 52.7
4 4 Male 10.2 64 9 53.6
5 5 Female 11.5 65 10 54.5
6 6 Male 12.8 66 11 55.4
7 7 Female 14.1 67 12 56.3
8 8 Male 15.4 68 13 57.2
9 9 Female 16.7 69 14 58.1
10 10 Male 5 70 5 59
>
26°C Sunny
```