

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

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

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Background Jobs
R > R 4.5.2 - ~/
> library(dplyr)
>
> student_df <- data.frame(
+ StudentID = c(1, 2, 2, 3, 4, 1),
+ Name = c("Nishita", "Simran", "Siyta", "Riya", "Nandini", "Yukta"),
+ Course = c("python", "Java", "Java", "C++", "Data Science", "Python")
)
>
> print("--- 1. Original Dataset (Note 6 rows) ---")
[1] "--- 1. Original dataset (Note 6 rows) ---"
> print(student_df)
StudentID Name Course
1 1 Nishita Python
2 2 Simran Java
3 2 Siyta Java
4 3 Riya C++
5 4 Nandini Data Science
6 1 Yukta python
>
> #-----
> # IDENTIFYING DUPLICATES
> #-----
>
> duplicates_report <- student_df %>%
+ group_by(StudentID, Name, Course) %>%
+ 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: 0 × 4
Groups: StudentID, Name, Course [0]
i 4 variables: StudentID <dbl>, Name <chr>, Course <chr>, n <int>
>
> #-----
> # REMOVING EXACT DUPLICATES
> #-----
>
> clean_exact <- student_df %>%
+ distinct()
>
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] --- 3. Removed Exact Duplicates (distinct) ---
#> #-----
#> # UNIQUE STUDENTS BASED ON NAME
#> #-----
#>
#> unique_students <- student_df %>%
+ distinct(Name, .keep_all = TRUE)
>
> print("--- 4. Unique Students Only (Partial Duplicates removed) ---")
[1] --- 4. Unique Students only (Partial Duplicates removed) ---
> print(unique_students)
StudentID Name Course
1 1 Nishita Python
2 2 Simran Java
3 2 Siyta Java
4 3 Riya C++
5 4 Nandini Data Science
6 1 Yukta Python
>
#> #-----
#> # YUKTA SONAWANE S120
#> #-----
#>

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Background Jobs
R > R 4.5.2 - ~/
>
> # Groups: StudentID, Name, Course [0]
i 4 variables: StudentID <dbl>, Name <chr>, course <chr>, n <int>
>
> #-----
> # REMOVING EXACT DUPLICATES
> #-----
>
> clean_exact <- student_df %>%
+ distinct()
>
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] --- 3. Removed Exact Duplicates (distinct) ---
#> #-----
#> # Print(clean_exact)
#> #-----
#> #-----
#> #-----
#> unique_students <- student_df %>%
+ distinct(Name, .keep_all = TRUE)
>
> print("--- 4. Unique Students Only (Partial Duplicates removed) ---")
[1] --- 4. Unique Students only (Partial Duplicates removed) ---
> print(unique_students)
StudentID Name Course
1 1 Nishita Python
2 2 Simran Java
3 2 Siyta Java
4 3 Riya C++
5 4 Nandini Data Science
6 1 Yukta Python
>
#> print("Yukta Sonawane S120")
[1] "Yukta Sonawane S120"
> |