

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
Go to file/function Addins

Source
Console Background Jobs
R - R4.5.2 - ~/
> library(dplyr)
> student_df <- data.frame(
+   studentID = c(1, 2, 2, 3, 4, 1),
+   Name = c("Nishita", "Simran", "Siya", "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 Siya   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 x 4
# Groups:   studentID, Name, Course [0]
# 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)
  studentID Name Course
1         1 Nishita Python
2         2 Simran  Java
3         2 Siya   Java
4         3 Riya   C++
5         4 Nandini Data Science
6         1 Yukta   Python
>
> # 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 Siya   Java
4         3 Riya   C++
5         4 Nandini Data Science
6         1 Yukta   Python
>
> print("Yukta Sonawane S120")
[1] "Yukta Sonawane S120"
>
```