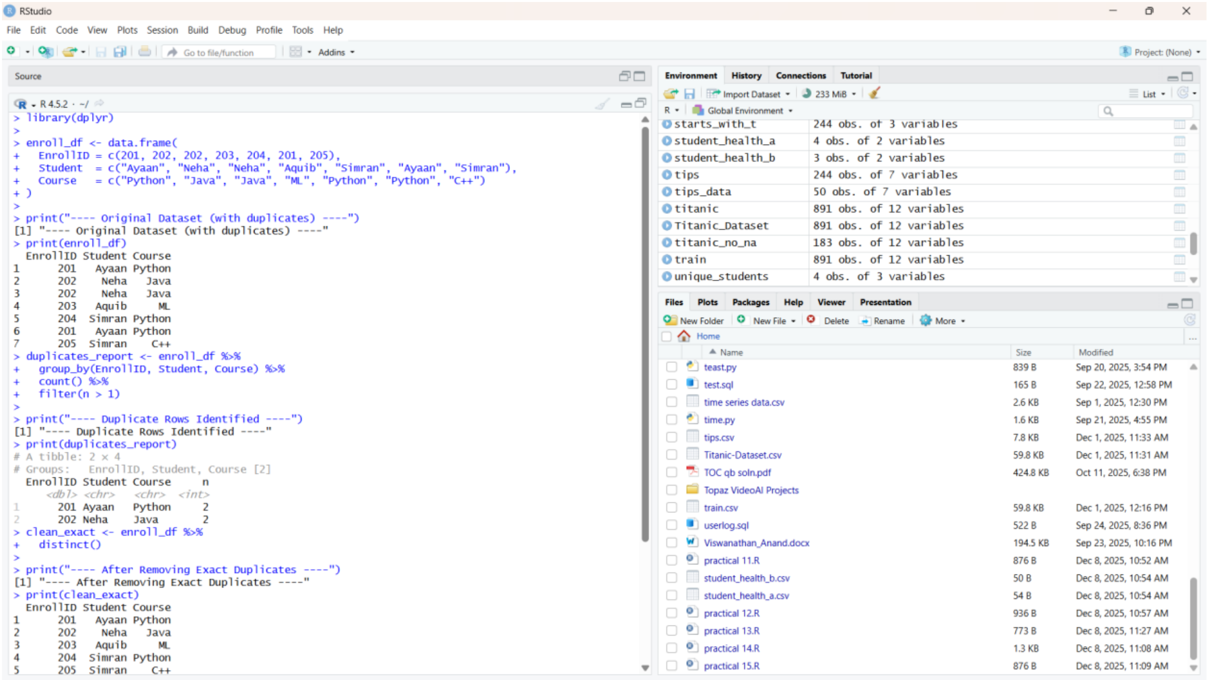
PRACTICAL NO 13

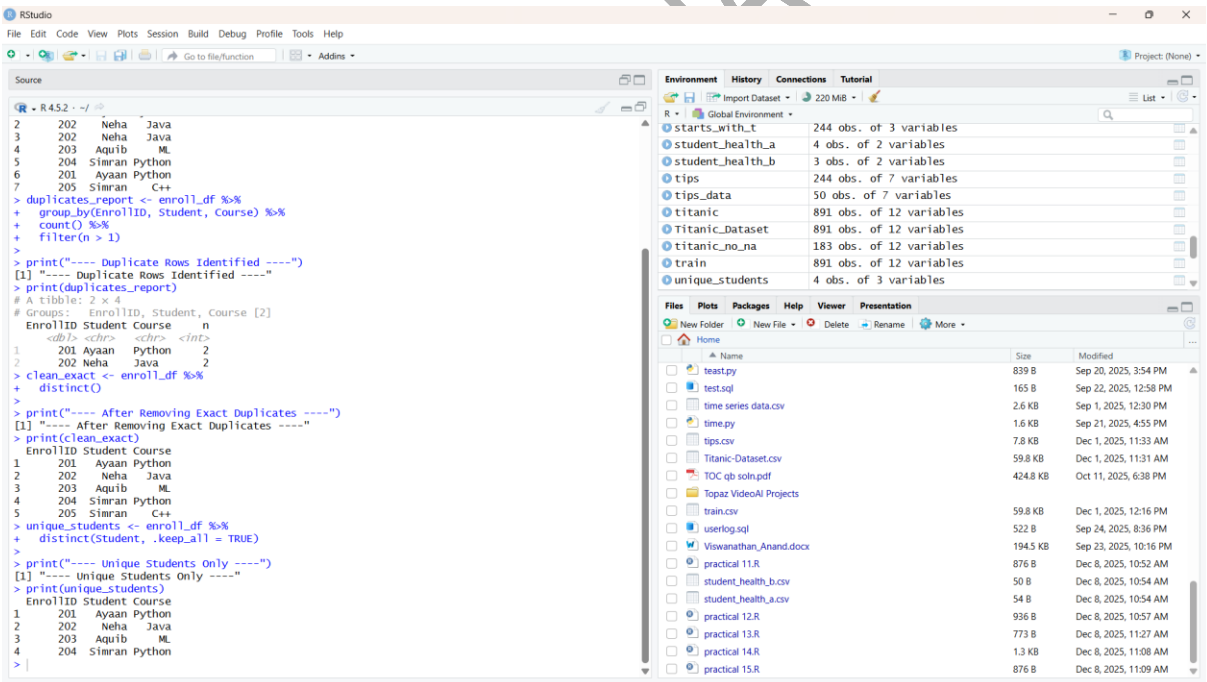
AIM: Identifying and handling duplicates using distinct() (R).


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.5.2 · ~/
Environment History Connections Tutorial
Import Dataset
233 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
New Folder
New File
Delete
Rename
More
Home

```
> library(dplyr)
>
> enroll_df <- data.frame(
+   EnrollID = c(201, 202, 202, 203, 204, 201, 205),
+   Student  = c("Ayaan", "Neha", "Neha", "Aquib", "Simran", "Ayaan", "Simran"),
+   Course   = c("Python", "Java", "Java", "ML", "Python", "Python", "C++")
+ )
>
> print("---- Original Dataset (with duplicates) ----")
[1] "---- Original Dataset (with duplicates) ----"
> print(enroll_df)
  EnrollID Student Course
1      201   Ayaan Python
2      202    Neha   Java
3      202    Neha   Java
4      203   Aquib     ML
5      204  Simran Python
6      201   Ayaan Python
7      205  Simran    C++
> duplicates_report <- enroll_df %>%
+   group_by(EnrollID, Student, Course) %>%
+   count() %>%
+   filter(n > 1)
>
> print("---- Duplicate Rows Identified ----")
[1] "---- Duplicate Rows Identified ----"
> print(duplicates_report)
# A tibble: 2 × 4
# Groups:   EnrollID, Student, Course [2]
  EnrollID Student Course     n
     <dbl> <chr>   <chr>  <int>
1      201 Ayaan   Python     2
2      202 Neha    Java       2
> clean_exact <- enroll_df %>%
+   distinct()
>
> print("---- After Removing Exact Duplicates ----")
[1] "---- After Removing Exact Duplicates ----"
> print(clean_exact)
  EnrollID Student Course
1      201   Ayaan Python
2      202    Neha   Java
3      203   Aquib     ML
4      204  Simran Python
5      205  Simran    C++
```

starts_with_t	244 obs. of 3 variables
student_health_a	4 obs. of 2 variables
student_health_b	3 obs. of 2 variables
tips	244 obs. of 7 variables
tips_data	50 obs. of 7 variables
titanic	891 obs. of 12 variables
Titanic_Dataset	891 obs. of 12 variables
titanic_no_na	183 obs. of 12 variables
train	891 obs. of 12 variables
unique_students	4 obs. of 3 variables

Name	Size	Modified
teast.py	839 B	Sep 20, 2025, 3:54 PM
test.sql	165 B	Sep 22, 2025, 12:58 PM
time series data.csv	2.6 KB	Sep 1, 2025, 12:30 PM
time.py	1.6 KB	Sep 21, 2025, 4:55 PM
tips.csv	7.8 KB	Dec 1, 2025, 11:33 AM
Titanic-Dataset.csv	59.8 KB	Dec 1, 2025, 11:31 AM
TOC qb soln.pdf	424.8 KB	Oct 11, 2025, 6:38 PM
Topaz VideoAI Projects		
train.csv	59.8 KB	Dec 1, 2025, 12:16 PM
userlog.sql	522 B	Sep 24, 2025, 8:36 PM
Viswanathan_Anand.docx	194.5 KB	Sep 23, 2025, 10:16 PM
practical 11.R	876 B	Dec 8, 2025, 10:52 AM
student_health_b.csv	50 B	Dec 8, 2025, 10:54 AM
student_health_a.csv	54 B	Dec 8, 2025, 10:54 AM
practical 12.R	936 B	Dec 8, 2025, 10:57 AM
practical 13.R	773 B	Dec 8, 2025, 11:27 AM
practical 14.R	1.3 KB	Dec 8, 2025, 11:08 AM
practical 15.R	876 B	Dec 8, 2025, 11:09 AM


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
R 4.5.2 · ~/
Environment History Connections Tutorial
Import Dataset
220 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
New Folder
New File
Delete
Rename
More
Home

```
2      202    Neha   Java
3      202    Neha   Java
4      203   Aquib     ML
5      204  Simran Python
6      201   Ayaan Python
7      205  Simran    C++
> duplicates_report <- enroll_df %>%
+   group_by(EnrollID, Student, Course) %>%
+   count() %>%
+   filter(n > 1)
>
> print("---- Duplicate Rows Identified ----")
[1] "---- Duplicate Rows Identified ----"
> print(duplicates_report)
# A tibble: 2 × 4
# Groups:   EnrollID, Student, Course [2]
  EnrollID Student Course     n
     <dbl> <chr>   <chr>  <int>
1      201 Ayaan   Python     2
2      202 Neha    Java       2
> clean_exact <- enroll_df %>%
+   distinct()
>
> print("---- After Removing Exact Duplicates ----")
[1] "---- After Removing Exact Duplicates ----"
> print(clean_exact)
  EnrollID Student Course
1      201   Ayaan Python
2      202    Neha   Java
3      203   Aquib     ML
4      204  Simran Python
5      205  Simran    C++
> unique_students <- enroll_df %>%
+   distinct(Student, .keep_all = TRUE)
>
> print("---- Unique Students Only ----")
[1] "---- Unique Students Only ----"
> print(unique_students)
  EnrollID Student Course
1      201   Ayaan Python
2      202    Neha   Java
3      203   Aquib     ML
4      204  Simran Python
> 
```

starts_with_t	244 obs. of 3 variables
student_health_a	4 obs. of 2 variables
student_health_b	3 obs. of 2 variables
tips	244 obs. of 7 variables
tips_data	50 obs. of 7 variables
titanic	891 obs. of 12 variables
Titanic_Dataset	891 obs. of 12 variables
titanic_no_na	183 obs. of 12 variables
train	891 obs. of 12 variables
unique_students	4 obs. of 3 variables

Name	Size	Modified
teast.py	839 B	Sep 20, 2025, 3:54 PM
test.sql	165 B	Sep 22, 2025, 12:58 PM
time series data.csv	2.6 KB	Sep 1, 2025, 12:30 PM
time.py	1.6 KB	Sep 21, 2025, 4:55 PM
tips.csv	7.8 KB	Dec 1, 2025, 11:33 AM
Titanic-Dataset.csv	59.8 KB	Dec 1, 2025, 11:31 AM
TOC qb soln.pdf	424.8 KB	Oct 11, 2025, 6:38 PM
Topaz VideoAI Projects		
train.csv	59.8 KB	Dec 1, 2025, 12:16 PM
userlog.sql	522 B	Sep 24, 2025, 8:36 PM
Viswanathan_Anand.docx	194.5 KB	Sep 23, 2025, 10:16 PM
practical 11.R	876 B	Dec 8, 2025, 10:52 AM
student_health_b.csv	50 B	Dec 8, 2025, 10:54 AM
student_health_a.csv	54 B	Dec 8, 2025, 10:54 AM
practical 12.R	936 B	Dec 8, 2025, 10:57 AM
practical 13.R	773 B	Dec 8, 2025, 11:27 AM
practical 14.R	1.3 KB	Dec 8, 2025, 11:08 AM
practical 15.R	876 B	Dec 8, 2025, 11:09 AM

AQUIB HAFIZ SHAIKH
S112