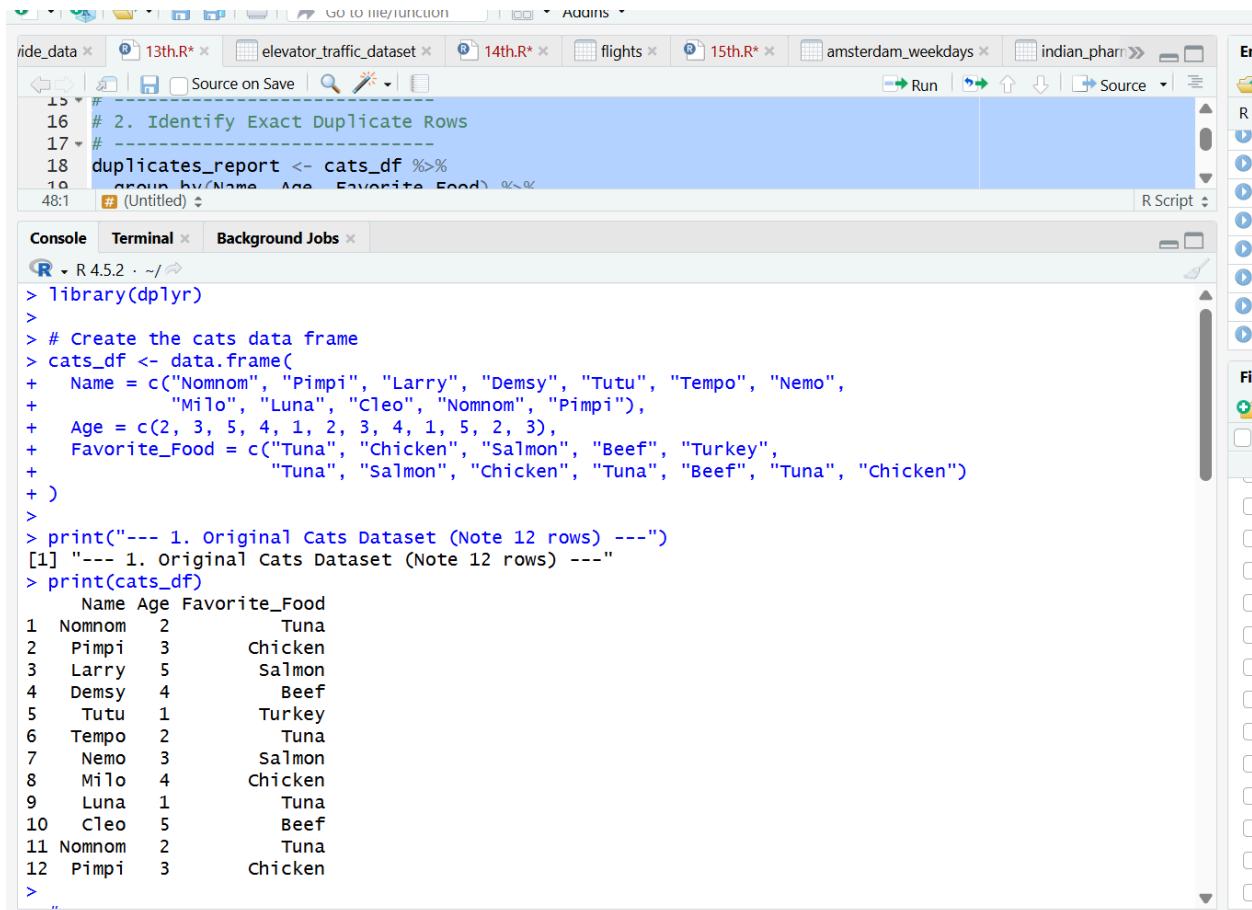


# Sheth I.u.j. And sir m.v. college of arts science and commerce

## Practical no.13th

Aim: Identifying and handling duplicates using PROC SORT NODUPKEY (SAS), Identify Duplicate Cases (SPSS), and distinct() (R).



The screenshot shows the RStudio interface with the following details:

- File Bar:** ride\_data, 13th.R\*, elevator\_traffic\_dataset, 14th.R\*, flights, 15th.R\*, amsterdam\_weekdays, indian\_pharm.
- Toolbar:** Source on Save, Run, Source.
- Code Editor:** An R script titled "Untitled" containing the following code:

```
15 # 
16 # 2. Identify Exact Duplicate Rows
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
48:1 # (Untitled)
```
- Console Tab:** Shows the R session output:

```
R 4.5.2 · ~/r
> library(dplyr)
>
> # Create the cats data frame
> cats_df <- data.frame(
+   Name = c("Nomnom", "Pimpi", "Larry", "Dempsy", "Tutu", "Tempo", "Nemo",
+           "Milo", "Luna", "Cleo", "Nomnom", "Pimpi"),
+   Age = c(2, 3, 5, 4, 1, 2, 3, 4, 1, 5, 2, 3),
+   Favorite_Food = c("Tuna", "Chicken", "Salmon", "Beef", "Turkey",
+                     "Tuna", "Salmon", "Chicken", "Tuna", "Beef", "Tuna", "Chicken")
+ )
>
> print("--- 1. Original Cats Dataset (Note 12 rows) ---")
[1] "--- 1. Original Cats Dataset (Note 12 rows) ---"
> print(cats_df)
  Name Age Favorite_Food
1 Nomnom  2        Tuna
2 Pimpi   3     Chicken
3 Larry   5      Salmon
4 Dempsy  4       Beef
5 Tutu    1      Turkey
6 Tempo   2        Tuna
7 Nemo    3      Salmon
8 Milo    4     chicken
9 Luna    1        Tuna
10 Cleo   5       Beef
11 Nomnom  2        Tuna
12 Pimpi  3     chicken
> "
```
- Environment Tab:** Shows the current environment variables.
- File Explorer:** Shows the project structure.

**Sheth I.u.j. And sir m.v. college of arts science and commerce**

The screenshot shows an RStudio interface with several tabs open at the top: 'ride\_data', '13th.R\*', 'elevator\_traffic\_dataset', '14th.R\*', 'flights', '15th.R\*', 'amsterdam\_weekdays', and 'indian\_phan'. The 'R Script' tab is active, displaying the following R code:

```
16 # 2. Identify Exact Duplicate Rows
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
48:1   (Untitled) 53940 obs. of 1

Console Terminal Background Jobs
R 4.5.2 · ~/ ◊
11 Nomnom 2 Tuna
12 Pimpi 3 Chicken
>
> # -----
> # 2. Identify Exact Duplicate Rows
> #
> duplicates_report <- cats_df %>%
+   group_by(Name, Age, Favorite_Food) %>%
+   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: 2 × 4
# Groups: Name, Age, Favorite_Food [2]
  Name     Age Favorite_Food     n
  <chr>   <dbl> <chr>       <int>
1 Nomnom    2 Tuna          2
2 Pimpi     3 Chicken       2
>
> # -----
> # 3. Remove Exact Duplicates
> #
> clean_exact <- cats_df %>%
+   distinct()
>
> print("--- 3. Removed Exact Duplicates (distinct) ---")
[1] "--- 3. Removed Exact Duplicates (distinct) ---"
```

The 'Environment' sidebar on the right lists variables and their values:

Variable	Type	Value
duplicates...	tbl_df	2 obs. of 4 var
long_data	tbl_df	485460 obs. of 1
older_cats	tbl_df	4 obs. of 3 var
original	tbl_df	53940 obs. of 1
processed...	tbl_df	10 obs. of 12 var
processed...	tbl_df	10 obs. of 31 var
report	tbl_df	1 obs. of 4 var
unique_cats	tbl_df	10 obs. of 3 var

The 'File' sidebar on the right shows the project structure:

- Home
  - Name
  - My Videos
- mysql-connector-j-8.3.0
- mysql-connector-j-8.3.0-javadoc.jar
- mysql-connector-j-8.3.0.jar
- mysql-connector-java-5.1.10-bin.jar
- qr new
- simran
- temp.txt
- webtech
- 11th R.pdf
- 14th R (1).pdf
- amsterdam\_weekdays.csv
- 15th R.pdf

The screenshot shows the RStudio interface with the following components:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Addins:** A dropdown menu containing "Addins".
- Script Editor:** Shows an R script with code for identifying and removing exact duplicates from a dataset named "cats\_df". The script includes sections for identifying exact duplicates, removing them, and keeping unique cats by name only.
- Console:** Displays the R session history, including the execution of the script and the resulting output. The output shows the original dataset and the cleaned dataset, which has 10 rows and 3 columns: Name, Age, and Favorite\_Food.
- Environment:** A pane on the right showing the global environment with objects like "duplicates...", "long\_data", "older\_cats", "original", "processed\_...", "report", and "unique\_cats".
- Files:** A pane showing the file structure, including "My Videos", "mysql-conn-", "qr new", "simran", "temp.txt", "webtech", "11th R.pdf", "14th R (1).pk", "amsterdam\_...", and "15th R.pdf".

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Addins
ide_data x 13th.R* elevator_traffic_dataset x 14th.R* flights x 15th.R* amsterdam_weekdays x indian_phar> Run Source
15 # 2. Identify Exact Duplicate Rows
16 # 2. Identify Exact Duplicate Rows
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
48:1 # (Untitled) R Script

Console Terminal x Background Jobs x
R 4.5.2 ~ /
> # -----
> # 3. Remove Exact Duplicates
> #
> clean_exact <- cats_df %>%
+   distinct()
>
> print(" --- 3. Removed Exact Duplicates (distinct) ---")
[1] " --- 3. Removed Exact Duplicates (distinct) ---"
> print(clean_exact)
  Name Age Favorite_Food
1  Nomnom  2        Tuna
2   Pimpi   3      Chicken
3   Larry   5      Salmon
4  Dempsy   4       Beef
5    Tutu   1     Turkey
6   Tempo   2        Tuna
7    Nemo   3      Salmon
8    Milo   4      Chicken
9    Luna   1        Tuna
10   Cleo   5       Beef
>
> # -----
> # 4. Keep Unique Cats by Name Only (Partial Duplicates removed)
> #
> unique_cats <- cats_df %>%
+   distinct(Name, .keep_all = TRUE)
>
> print(" --- 4. Unique Cats Only (Partial Duplicates removed) ---")
```

Name: Simran S113

# Sheth I.u.j. And sir m.v. college of arts science and commerce

The screenshot shows the RStudio interface with the following details:

- File menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and a search bar labeled "Go to file/function".
- Project Explorer:** Shows multiple R files: ride\_data.R, 13th.R\*, elevator\_traffic\_dataset.R, 14th.R\*, flights.R, 15th.R\*, amsterdam\_weekdays.R, indian\_pharm.R. The 15th.R\* file is currently selected.
- Code Editor:** Displays R code for identifying and removing duplicate rows from a dataset named "cats\_df". The code includes filtering by Name, Age, and Favorite\_Food, and then grouping by Name, Age, and Favorite\_Food to find duplicates.
- Console:** Shows the R session output. It prints a table of unique cat data and then prints "Nomnom Tuna Age 2".
- Environment:** A sidebar showing the current environment variables.
- Files:** A sidebar showing the project's file structure.

```
15 # 2. Identify Exact Duplicate Rows
16 # -----
17 # -----
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
20   summarise(n = n())
21
22 # (Untitled) 1
23
24 # 3. Remove Exact Duplicates
25 # -----
26 # 4. Keep Unique Cats by Name Only (Partial Duplicates removed)
27 # -----
28 unique_cats <- cats_df %>%
29   distinct(Name, .keep_all = TRUE)
30
31 print(" --- 4. Unique Cats Only (Partial Duplicates removed) ---")
32 [1] " --- 4. Unique Cats Only (Partial Duplicates removed) ---"
33 > print(unique_cats)
34   Name Age Favorite_Food
35 1 Nomnom  2      Tuna
36 2 Pimpi    3     Chicken
37 3 Larry    5     Salmon
38 4 Demsy    4      Beef
39 5 Tutu    1     Turkey
40 6 Tempo    2      Tuna
41 7 Nemo    3     Salmon
42 8 Milo    4     Chicken
43 9 Luna    1     Tuna
44 10 Cleo   5      Beef
45
46 # -----
47 # 5. Example print for a specific cat
48 # -----
49 print("Nomnom Tuna Age 2")
50 [1] "Nomnom Tuna Age 2"
```