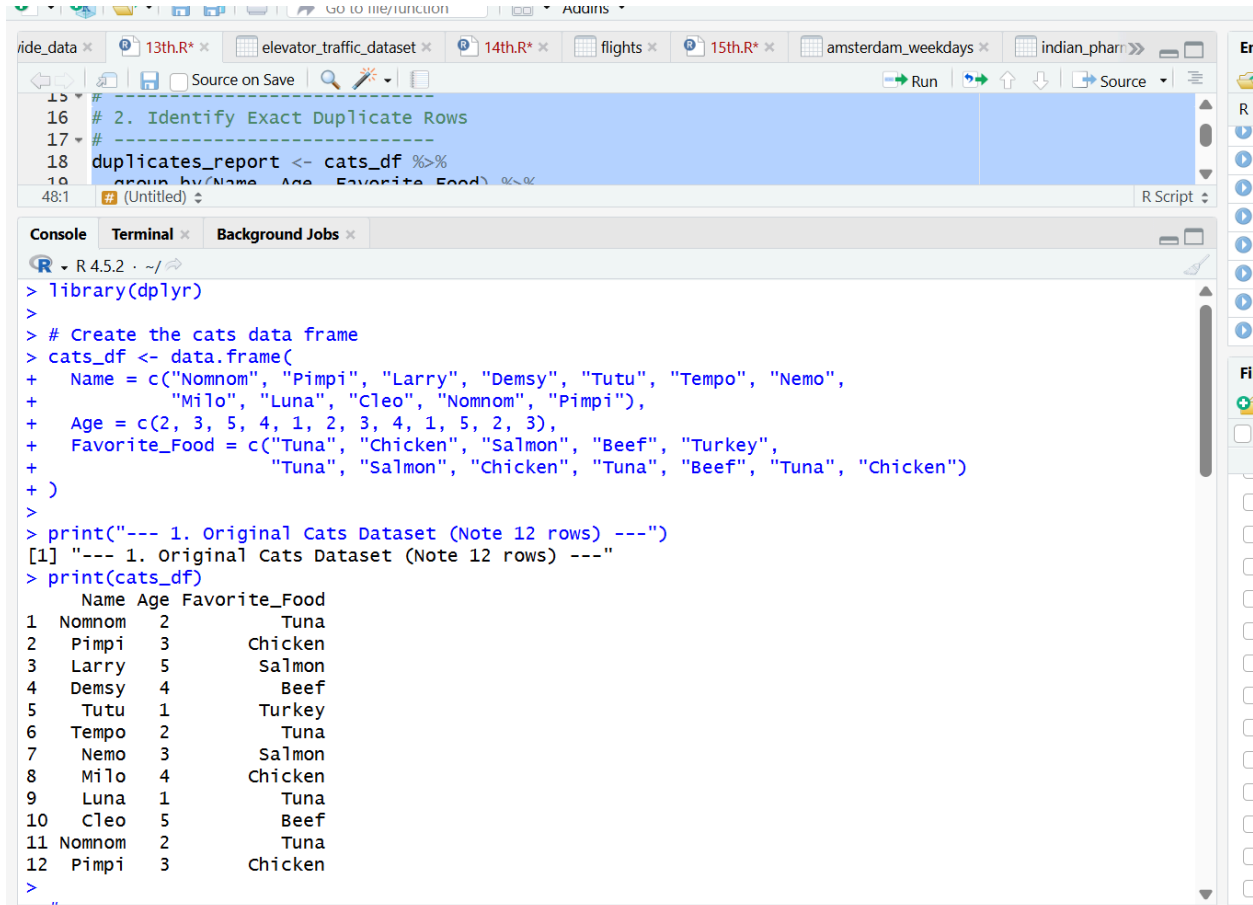


# Sheth I.u.j. And sir m.v. college of arts science and commerce

## Practical no.13th

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

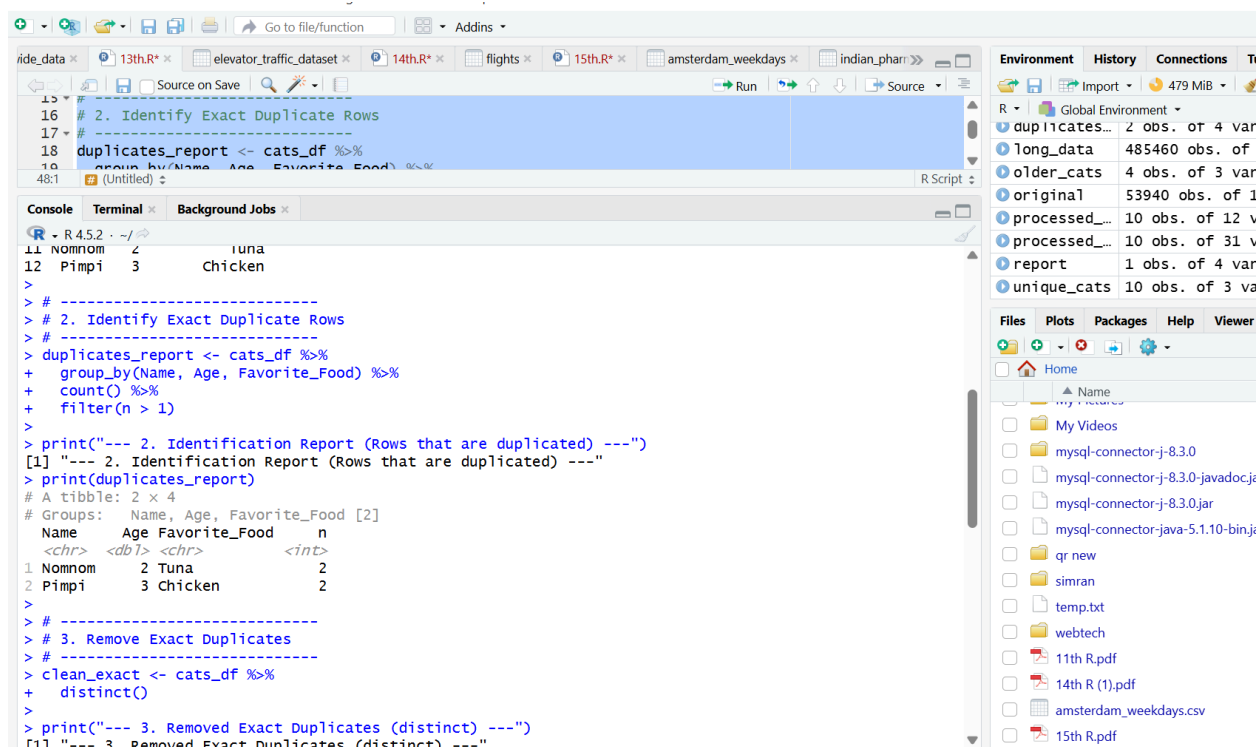


```
15 # 2. Identify Exact Duplicate Rows
16 #
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
20   distinct()

48:1 (Untitled) R Script
```

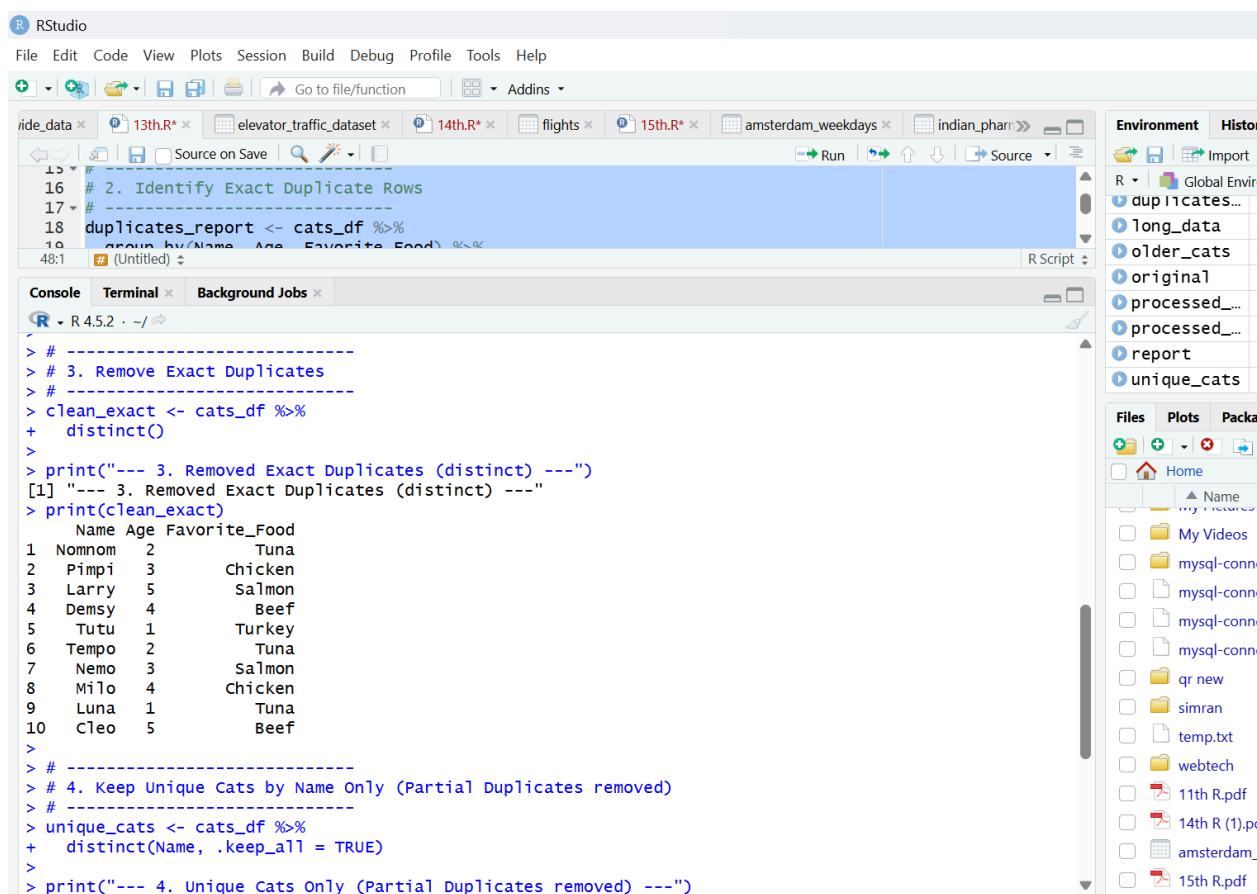
```
> library(dplyr)
>
> # Create the cats data frame
> cats_df <- data.frame(
+   Name = c("Nomnom", "Pimpi", "Larry", "Demsy", "Tutu", "Tempo", "Nemo",
+   "Milo", "Luna", "Cleo", "Nomnom", "Pimpi"),
+   Age = c(2, 3, 5, 4, 1, 2, 3, 4, 1, 5, 2, 3),
+   Favorite_Food = c("Tuna", "Chicken", "Salmon", "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 Demsy  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
>
```

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



```
15 # 2. Identify Exact Duplicate Rows
16 #
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
48:1   count() %>%
      filter(n > 1)

> # -----
> # 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 x 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) ---"
```



```
15 # 2. Identify Exact Duplicate Rows
16 #
17 #
18 duplicates_report <- cats_df %>%
19   group_by(Name, Age, Favorite_Food) %>%
48:1   count() %>%
      filter(n > 1)

> # -----
> # 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  Demsy  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 environment with several open files in the top pane. The active file is '13th.R', which contains the following R code:

```
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 # 3. Print the report
23 print(duplicates_report)
```

The bottom pane shows the R console output for the code executed in the previous file. The output displays the unique cats after removing partial duplicates:

```
> # -----
> # 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) ---")
[1] "--- 4. Unique Cats Only (Partial Duplicates removed) ---"
> print(unique_cats)
  Name Age Favorite_Food
1 Nomnom 2      Tuna
2 Pimpi 3      Chicken
3 Larry 5      Salmon
4 Damsy 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
>
> # -----
> # 5. Example print for a specific cat
> # -----
> print("Nomnom Tuna Age 2")
[1] "Nomnom Tuna Age 2"
>
```