

SHETH L.U.J. & SIR M.V. COLLEGE OF SCIENCE

SUBJECT - Data Analysis with SAS / SPSS / R

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

Input :

The screenshot shows the RStudio interface with a script editor on the left and the Environment pane on the right. The script editor contains the following code:

```
1 library(dplyr)
2
3 sw <- starwars %>%
4   select(name, species, gender)
5
6 sw_dup <- rbind(
7   sw[1:5, ],
8   sw[1, ],
9   sw[2, ],
10  sw[2, ],
11  sw[10, ],
12  sw[11, ]
13 )
14
15 print("--- Original Dataset with Duplicates ---")
16 print(sw_dup)
17
18 duplicates_report <- sw_dup %>%
19   group_by(name, species, gender) %>%
20   count() %>%
21   filter(n > 1)
22
23 print("--- Duplicates Found ---")
24 print(duplicates_report)
25
26 clean_exact <- sw_dup %>%
27   distinct()
28
29 print("--- Exact Duplicates Removed ---")
30 print(clean_exact)
31
32 unique_names <- sw_dup %>%
33   distinct(name, .keep_all = TRUE)
34
35 print("--- Unique Characters by Name ---")
36 print(unique_names)
37
38
```

The Environment pane on the right shows the following datasets:

Dataset	Observations	Variables
small_bills	17	7
smoker_or_big_tip	97	7
starwars	87	14
sw	87	3
sw_dup	10	3
team_records	12	19
tips	244	7
unique_names	7	3

The screenshot shows the RStudio interface after running the code. The script editor now contains the following code:

```
3 sw <- starwars %>%
4   select(name, species, gender)
5
6 sw_dup <- rbind(
7   sw[1:5, ],
8   sw[1, ],
9   sw[2, ],
10  sw[2, ],
11  sw[10, ],
12  sw[11, ]
13 )
14
15 print("--- Original Dataset with Duplicates ---")
16 print(sw_dup)
17
18 duplicates_report <- sw_dup %>%
19   group_by(name, species, gender) %>%
20   count() %>%
21   filter(n > 1)
22
23 print("--- Duplicates Found ---")
24 print(duplicates_report)
25
26 clean_exact <- sw_dup %>%
27   distinct()
28
29 print("--- Exact Duplicates Removed ---")
30 print(clean_exact)
31
32 unique_names <- sw_dup %>%
33   distinct(name, .keep_all = TRUE)
34
35 print("--- Unique Characters by Name ---")
36 print(unique_names)
37
38
```

The Environment pane on the right shows the following datasets:

Dataset	Observations	Variables
small_bills	17	7
smoker_or_big_tip	97	7
starwars	87	14
sw	87	3
sw_dup	10	3
team_records	12	19
tips	244	7
unique_names	7	3

Name - Mithil Kadam
Roll No - S083

SHETH L.U.J. & SIR M.V. COLLEGE OF SCIENCE

SUBJECT - Data Analysis with SAS / SPSS / R

Output :

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)

Source
Console Terminal Background Jobs
R - R4.5.2 - /
> library(dplyr)
> sw <- starwars %>%
+   select(name, species, gender)
> sw_dup <- rbind(
+   sw[1:5, ],
+   sw[1, ],
+   sw[2, ],
+   sw[2, ],
+   sw[10, ],
+   sw[11, ]
+ )
> print("--- Original Dataset with Duplicates ---")
[1] "--- Original Dataset with Duplicates ---"
> print(sw_dup)
# A tibble: 10 x 3
  name      species gender
  <chr>      <chr>   <chr>
1 Luke Skywalker Human    masculine
2 C-3PO      Droid    masculine
3 R2-D2      Droid    masculine
4 Darth Vader Human    masculine
5 Leia Organa Human    feminine
6 Luke Skywalker Human    masculine
7 C-3PO      Droid    masculine
8 C-3PO      Droid    masculine
9 Obi-wan Kenobi Human    masculine
10 Anakin Skywalker Human    masculine
> duplicates_report <- sw_dup %>%
+   group_by(name, species, gender) %>%
+   count() %>%
+   filter(n > 1)
> print("--- Duplicates Found ---")
[1] "--- Duplicates Found ---"
> print(duplicates_report)
# A tibble: 2 x 4
  Groups: name, species, gender [2]
  name      species gender      n
  <chr>      <chr>   <chr>   <int>
1 C-3PO      Droid    masculine    3
2 Luke Skywalker Human    masculine    2
> clean_exact <- sw_dup %>%
+   distinct()
> print("--- Exact Duplicates Removed ---")
[1] "--- Exact Duplicates Removed ---"
> print(clean_exact)
# A tibble: 7 x 3
  name      species gender
  <chr>      <chr>   <chr>
1 Luke Skywalker Human    masculine
2 C-3PO      Droid    masculine
3 R2-D2      Droid    masculine
4 Darth Vader Human    masculine
5 Leia Organa Human    feminine
6 Obi-wan Kenobi Human    masculine
7 Anakin Skywalker Human    masculine
> unique_names <- sw_dup %>%
+   distinct(name, .keep_all = TRUE)
> print("--- Unique Characters by Name ---")
[1] "--- Unique Characters by Name ---"
> print(unique_names)
# A tibble: 7 x 3
  name      species gender
  <chr>      <chr>   <chr>
1 Luke Skywalker Human    masculine
2 C-3PO      Droid    masculine
3 R2-D2      Droid    masculine
4 Darth Vader Human    masculine
5 Leia Organa Human    feminine
6 Obi-wan Kenobi Human    masculine
7 Anakin Skywalker Human    masculine
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Project: (None)

Source
Console Terminal Background Jobs
R - R4.5.2 - /
# A tibble: 2 x 4
  Groups: name, species, gender [2]
  name      species gender      n
  <chr>      <chr>   <chr>   <int>
1 C-3PO      Droid    masculine    3
2 Luke Skywalker Human    masculine    2
> clean_exact <- sw_dup %>%
+   distinct()
> print("--- Exact Duplicates Removed ---")
[1] "--- Exact Duplicates Removed ---"
> print(clean_exact)
# A tibble: 7 x 3
  name      species gender
  <chr>      <chr>   <chr>
1 Luke Skywalker Human    masculine
2 C-3PO      Droid    masculine
3 R2-D2      Droid    masculine
4 Darth Vader Human    masculine
5 Leia Organa Human    feminine
6 Obi-wan Kenobi Human    masculine
7 Anakin Skywalker Human    masculine
> unique_names <- sw_dup %>%
+   distinct(name, .keep_all = TRUE)
> print("--- Unique Characters by Name ---")
[1] "--- Unique Characters by Name ---"
> print(unique_names)
# A tibble: 7 x 3
  name      species gender
  <chr>      <chr>   <chr>
1 Luke Skywalker Human    masculine
2 C-3PO      Droid    masculine
3 R2-D2      Droid    masculine
4 Darth Vader Human    masculine
5 Leia Organa Human    feminine
6 Obi-wan Kenobi Human    masculine
7 Anakin Skywalker Human    masculine
```

Name - Mithil Kadam
Roll No - S083