

Sheth L.U.J. & Sir M.V. College

13. Identifying and handling duplicates using distinct() (R studio).

The screenshot shows the RStudio interface. The left pane displays a data frame titled "Duplicate_Food_Delivery_Route_Efficiency" with 205 rows and 10 columns. The columns include order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, and customer_zone. The right pane shows the "Environment" tab with various objects listed, such as combined_d, df, df_clean, df_no_dupl, df_small, duplicate, duplicate_10, flower_cle, flower_df, Food_deliv, iris, iris_clean, long_df, traffic_cpl, unique_ord, unique_rou, and wide_df. The bottom status bar shows the date and time as 08-12-2025 at 11:58:47 AM.

The screenshot shows the RStudio interface with an R script titled "S081_R_Practical13.R". The script contains code for identifying and handling duplicates in a food delivery dataset. It includes reading the CSV file, printing the head, calculating total rows, identifying exact duplicate rows, removing them, and finally removing duplicates based on a specific column (order_id). The right pane shows the "Environment" tab with the same list of objects as the previous screenshot. The bottom status bar shows the date and time as 08-12-2025 at 11:58:19 AM.

```
1 # R Script: Identifying and Handling Duplicates (Food Delivery Dataset)
2 #
3 #
4 # library(dplyr)
5 #
6 # -----
7 # 1. IMPORT FOOD DELIVERY DATASET
8 #
9 # df <- read.csv(
10 #   "c:\\Users\\itlab\\OneDrive\\Desktop\\S081_R_Studio\\Duplicate_Food_Delivery_Route_Efficiency_dataset.csv",
11 #   na.strings = c("", "NA")
12 # )
13 #
14 print("--- 1. Original Dataset ---")
15 print(head(df))
16 print(paste("Total Rows:", nrow(df)))
17 #
18 #
19 # 2. IDENTIFY EXACT DUPLICATE ROWS
20 #
21 # Any rows where EVERY column is the same
22 #
23 duplicate_rows <- df %>%
24   group_by(across(everything())) %>%
25   filter(n() > 1)
26 #
27 print("--- 2. Exact Duplicate Rows (if any) ---")
28 print(duplicate_rows)
29 #
30 #
31 # 3. REMOVE EXACT DUPLICATES
32 #
33 # df_no_duplicates <- df %>% distinct()
34 #
35 print("--- 3. Dataset After Removing Exact Duplicates ---")
36 print(head(df_no_duplicates))
37 print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
38 #
39 #
40 # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
41 # Example: Keep only 1 row per order_id (First occurrence kept)
42 #
```

Name :- Priya Gupta
Roll no. :- S081

Sheth L.U.J. & Sir M.V. College

The screenshot shows the RStudio interface with a script editor and a data viewer.

Script Editor:

```
1 print(paste("Total Rows:", nrow(df)))
2 # 2. IDENTIFY EXACT DUPLICATE ROWS
3 # Any rows where EVERY column is the same
4
5 duplicate_rows <- df %>%
6 group_by(across(everything())) %>%
7 filter(n() > 1)
8
9 print("---- 2. Exact Duplicate Rows (if any) ---")
10 print(duplicate_rows)
11
12 # 3. REMOVE EXACT DUPLICATES
13
14 df_no_duplicates <- df %>% distinct()
15
16 print("---- 3. Dataset After Removing Exact Duplicates ---")
17 print(head(df_no_duplicates))
18 print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
19
20 # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
21 # Example: Keep only 1 row per order_id (First occurrence kept)
22
23 unique_orders <- df %>% distinct(order_id, .keep_all = TRUE)
24
25 print("---- 4. Unique orders only (Duplicate order_id removed) ---")
26 print(head(unique_orders))
27 print(paste("Unique orders count:", nrow(unique_orders)))
28
29 # 5. REMOVE DUPLICATES BASED ON MULTIPLE COLUMNS
30 # Example: order_id + distance_km
31
32 unique_routes <- df %>% distinct(order_id, distance_km, .keep_all = TRUE)
33
34 print("---- 5. Unique combination: order_id + distance_km ---")
35 print(head(unique_routes))
36
```

Data View:

Object	Type	Observations	Variables
combined_df	10150 obs. of 2 variables		
df	205 obs. of 10 variables		
df_clean	200 obs. of 11 variables		
df_no_dups	200 obs. of 10 variables		
df_small	200 obs. of 4 variables		
duplicate_df	205 obs. of 10 variables		
duplicate_routes	10 obs. of 10 variables		
flower_clean	10000 obs. of 2 variables		
flower_df	10000 obs. of 4 variables		
Food_delivery	200 obs. of 10 variables		
iris	150 obs. of 5 variables		
iris_clean	150 obs. of 2 variables		
long_df	600 obs. of 3 variables		
traffic_patterns	200 obs. of 4 variables		
unique_order_ids	200 obs. of 10 variables		
unique_routes	200 obs. of 10 variables		
wide_df	200 obs. of 4 variables		

Name :- Priya Gupta
Roll no. :- S081

Sheth L.U.J. & Sir M.V. College

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source Environment History Connections Tutorial
Console Background Jobs
R - R 4.5.2 ~/ ~
> # 1. REMOVE DUPLICATES (BY ANY ---)
> print(duplicate_rows)
# A tibble: 10 × 10
# Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
1     4       7.97        63.8      High     9.75      Bicycle  Rainy  01-01-2025 14.12 Central  Central
2     5       10.0         56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
3     6       11.0         76.8      High     13.6      Car     Windy  02-01-2025 9.56 West    North
4     7       9.6          54.4      Low      5.25      Scooter Rainy  03-01-2025 0.29 Central  North
5     8       6.24         52.9      Low      8.41      Bike    Cloudy 02-01-2025 21.36 South   Central
6     4       4.9          24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
7     5       10.0         56.0      High     11.3      Car     Rainy  02-01-2025 16.50 West    North
8     6       11.0         76.8      High     13.6      Car     Windy  02-01-2025 9.56 West    North
9     7       9.6          54.4      Low      13.3      Bike    Windy  03-01-2025 0.29 Central  North
10    8       6.24         52.9      Low      8.41      Bike    Cloudy 02-01-2025 21.36 South   Central
> # =====
> # 2. REMOVE EXACT DUPLICATES
> df_no_duplicates <- df %>% distinct()
> 
> print("--- 3. dataset After Removing Exact Duplicates ---")
[1] "--- 3. dataset After Removing Exact Duplicates ---"
> print(head(df_no_duplicates))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
[1] "Rows After Removing Exact Duplicates: 200"
> 
> # =====
> # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
> # Example: keep only 1 row per order_id (first occurrence kept)
> # =====
> unique_orders <- df %>% distinct(order_id, .keep_all = TRUE)
> 
> print("--- 4. unique orders only (duplicate order_id removed) ---")
[1] "--- 4. unique orders only (duplicate order_id removed) ---"
> print(head(unique_orders))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> print(paste("Unique Orders Count:", nrow(unique_orders)))
[1] "Unique Orders count: 200"
> 
> # =====
> # 5. REMOVE DUPLICATES BASED ON MULTIPLE COLUMNS
> # Example: order_id + distance_km
> # =====
> unique_routes <- df %>% distinct(order_id, distance_km, .keep_all = TRUE)
> 
> print("--- 5. unique Combination: order_id + distance_km ---")
[1] "--- 5. unique Combination: order_id + distance_km ---"
> print(head(unique_routes))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> 
```

Environment History Connections Tutorial

Import 130 MB List Global Environment

Data

- combined_d... 10150 obs. of 2 variables
- df 205 obs. of 10 variables
- df_clean 200 obs. of 11 variables
- df_no_dupl... 200 obs. of 10 variables
- df_small 200 obs. of 4 variables
- duplicate_... 205 obs. of 10 variables
- duplicate_... 10 obs. of 10 variables
- flower_cle... 10000 obs. of 2 variables
- flower_df 10000 obs. of 4 variables
- food_deliv... 200 obs. of 10 variables
- iris 150 obs. of 5 variables
- iris_clean 150 obs. of 2 variables
- long_df 600 obs. of 3 variables
- traffic_pl... 200 obs. of 4 variables
- unique_ord... 200 obs. of 10 variables
- unique_rout... 200 obs. of 10 variables
- wide_df 200 obs. of 4 variables

Files Plots Packages Help Viewer Presentation

Present Print Edit

ENG IN 11:59:55 AM 08-12-2025

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source Environment History Connections Tutorial
Console Background Jobs
R - R 4.5.2 ~/ ~
> # 1. REMOVE DUPLICATES (BY ANY ---)
> print(duplicate_rows)
# A tibble: 10 × 10
# Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
1     4       7.97        63.8      High     9.75      Bicycle  Rainy  01-01-2025 14.12 Central  Central
2     5       10.0         56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
3     6       11.0         76.8      High     13.6      Car     Windy  02-01-2025 9.56 West    North
4     7       9.6          54.4      Low      5.25      Scooter Rainy  03-01-2025 0.29 Central  North
5     8       6.24         52.9      Low      8.41      Bike    Cloudy 02-01-2025 21.36 South   Central
6     4       4.9          24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
7     5       10.0         56.0      High     11.3      Car     Rainy  02-01-2025 16.50 West    North
8     6       11.0         76.8      High     13.6      Car     Windy  02-01-2025 9.56 West    North
9     7       9.6          54.4      Low      13.3      Bike    Windy  03-01-2025 0.29 Central  North
10    8       6.24         52.9      Low      8.41      Bike    Cloudy 02-01-2025 21.36 South   Central
> # =====
> # 2. REMOVE EXACT DUPLICATES
> df_no_duplicates <- df %>% distinct()
> 
> print("--- 3. dataset After Removing Exact Duplicates ---")
[1] "--- 3. dataset After Removing Exact Duplicates ---"
> print(head(df_no_duplicates))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
[1] "Rows After Removing Exact Duplicates: 200"
> 
> # =====
> # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
> # Example: keep only 1 row per order_id (first occurrence kept)
> # =====
> unique_orders <- df %>% distinct(order_id, .keep_all = TRUE)
> 
> print("--- 4. unique orders only (duplicate order_id removed) ---")
[1] "--- 4. unique orders only (duplicate order_id removed) ---"
> print(head(unique_orders))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> print(paste("Unique Orders Count:", nrow(unique_orders)))
[1] "Unique Orders count: 200"
> 
> # =====
> # 5. REMOVE DUPLICATES BASED ON MULTIPLE COLUMNS
> # Example: order_id + distance_km
> # =====
> unique_routes <- df %>% distinct(order_id, distance_km, .keep_all = TRUE)
> 
> print("--- 5. unique Combination: order_id + distance_km ---")
[1] "--- 5. unique Combination: order_id + distance_km ---"
> print(head(unique_routes))
#> # A tibble: 10 × 10
#> # Groups: order_id, distance_km, delivery_time_min, traffic_level, route_length_km, delivery_mode, weather, order_time, restaurant_zone, customer_zone [5]
#> #   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#> #   <dbl>      <dbl>          <dbl>      <dbl>          <dbl>      <dbl>    <chr>    <date>    <chr>        <chr>
#> 1     1       7.97        63.8      High     9.75      Bicycle  Clear  01-01-2025 15.29 South   North
#> 2     2       0.90         7.6      High     1.28      Car     Cloudy 03-01-2025 0.47 West    North
#> 3     3      11.12        78.0      Medium   16.65      Bike    Rainy  04-01-2025 17.32 South   Central
#> 4     4       4.90         24.8      Low      5.25      Scooter Rainy  01-01-2025 14.12 Central  Central
#> 5     5       10.04        56.0      High     11.34      Car     Rainy  02-01-2025 16.50 West    North
#> 6     6       10.96        76.8      High     13.62      Car     Windy  02-01-2025 9.56 West    North
> 
```

Environment History Connections Tutorial

Import 130 MB List Global Environment

Data

- combined_d... 10150 obs. of 2 variables
- df 205 obs. of 10 variables
- df_clean 200 obs. of 11 variables
- df_no_dupl... 200 obs. of 10 variables
- df_small 200 obs. of 4 variables
- duplicate_... 205 obs. of 10 variables
- duplicate_... 10 obs. of 10 variables
- flower_cle... 10000 obs. of 2 variables
- flower_df 10000 obs. of 4 variables
- food_deliv... 200 obs. of 10 variables
- iris 150 obs. of 5 variables
- iris_clean 150 obs. of 2 variables
- long_df 600 obs. of 3 variables
- traffic_pl... 200 obs. of 4 variables
- unique_ord... 200 obs. of 10 variables
- unique_rout... 200 obs. of 10 variables
- wide_df 200 obs. of 4 variables

Files Plots Packages Help Viewer Presentation

Present Print Edit

ENG IN 12:00:04 PM 08-12-2025

Name :- Priya Gupta
 Roll no. :- S081

Sheth L.U.J. & Sir M.V. College

S081 Priya Gupta

Name :- Priya Gupta

Roll no. :- S081