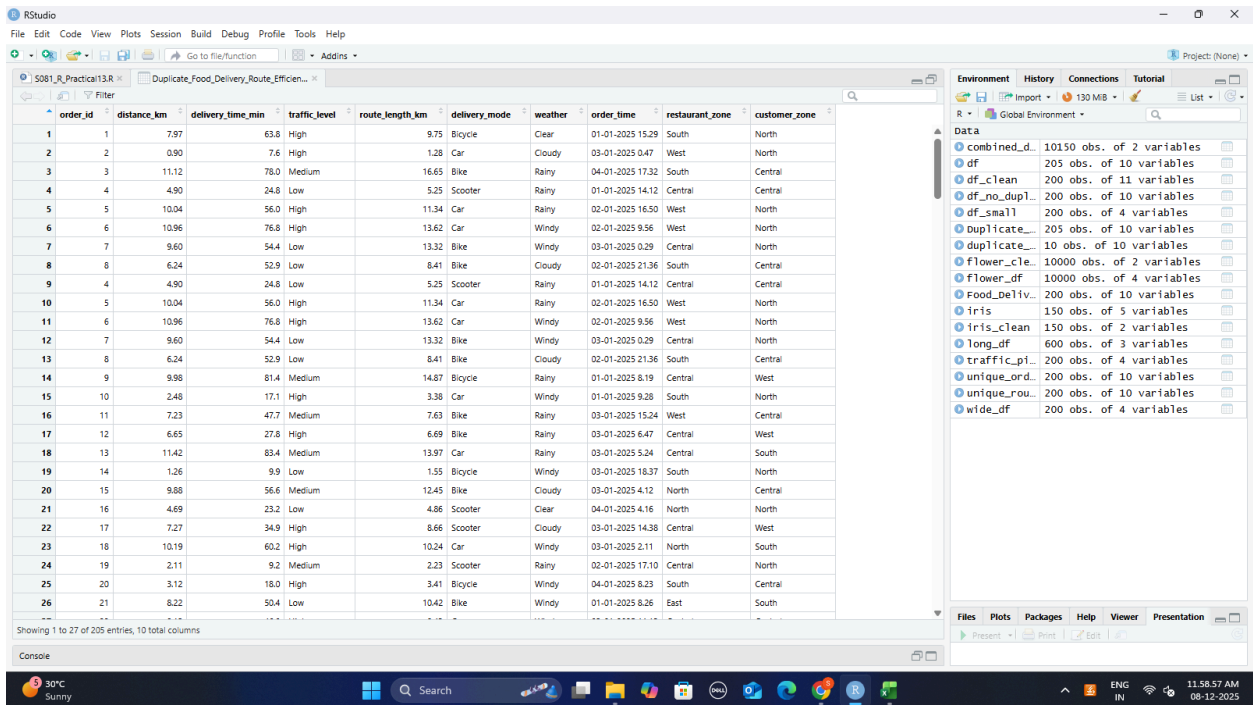
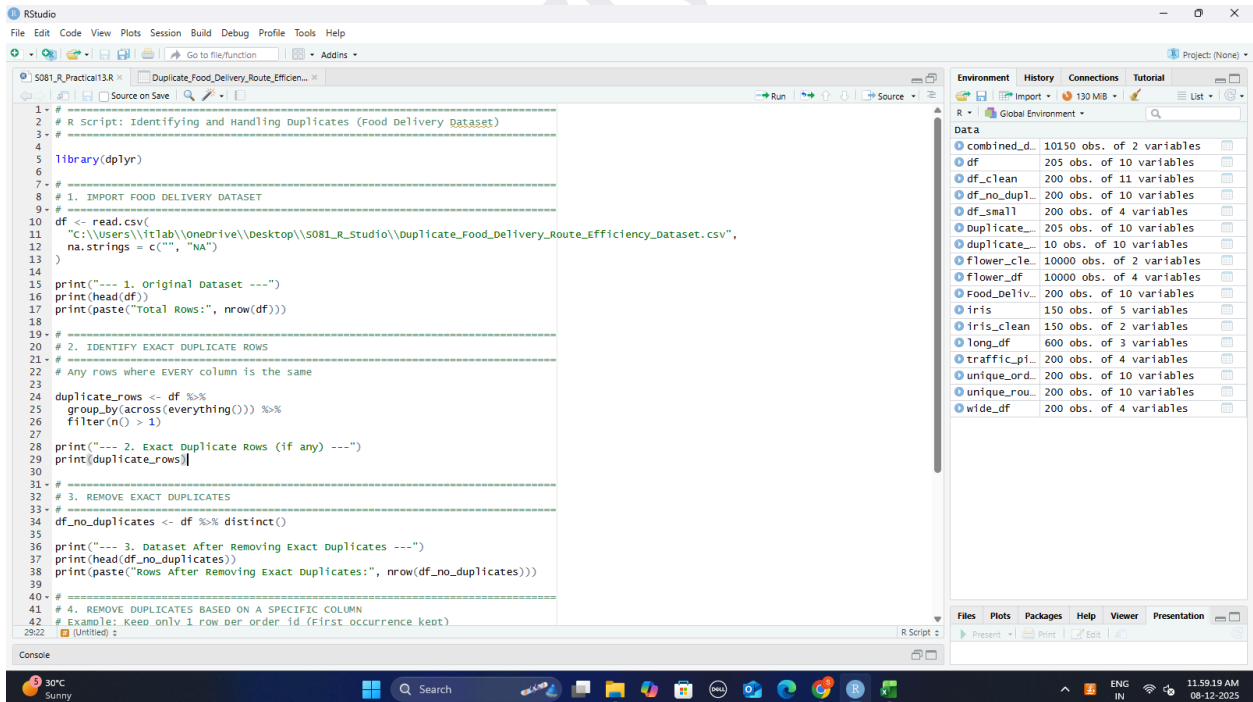


# Sheth L.U.J. & Sir M.V. College

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



order_id	distance_km	delivery_time_min	traffic_level	route_length_km	delivery_mode	weather	order_time	restaurant_zone	customer_zone
1	7.97	63.8	High	9.75	Bicycle	Clear	01-01-2025 15:29	South	North
2	0.90	7.6	High	1.28	Car	Cloudy	03-01-2025 0:47	West	North
3	11.12	78.0	Medium	16.65	Bike	Rainy	04-01-2025 17:32	South	Central
4	4.90	24.8	Low	5.25	Scooter	Rainy	01-01-2025 14:12	Central	Central
5	10.04	56.0	High	11.34	Car	Rainy	02-01-2025 16:50	West	North
6	10.96	76.8	High	13.62	Car	Windy	02-01-2025 9:56	West	North
7	9.60	54.4	Low	13.32	Bike	Windy	03-01-2025 0:29	Central	North
8	6.24	52.9	Low	8.41	Bike	Cloudy	02-01-2025 21:36	South	Central
9	4.90	24.8	Low	5.25	Scooter	Rainy	01-01-2025 14:12	Central	Central
10	10.04	56.0	High	11.34	Car	Rainy	02-01-2025 16:50	West	North
11	10.96	76.8	High	13.62	Car	Windy	02-01-2025 9:56	West	North
12	9.60	54.4	Low	13.32	Bike	Windy	03-01-2025 0:29	Central	North
13	6.24	52.9	Low	8.41	Bike	Cloudy	02-01-2025 21:36	South	Central
14	9.98	61.4	Medium	14.67	Bicycle	Rainy	01-01-2025 8:19	Central	West
15	2.48	17.1	High	3.38	Car	Windy	01-01-2025 9:28	South	North
16	7.23	47.7	Medium	7.63	Bike	Rainy	03-01-2025 15:24	West	Central
17	6.65	27.8	High	6.69	Bike	Rainy	03-01-2025 6:47	Central	West
18	11.42	83.4	Medium	13.97	Car	Rainy	03-01-2025 5:24	Central	South
19	1.26	9.9	Low	1.55	Bicycle	Windy	03-01-2025 18:37	South	North
20	9.88	56.6	Medium	12.45	Bike	Cloudy	03-01-2025 4:12	North	Central
21	4.69	23.2	Low	4.66	Scooter	Clear	04-01-2025 4:16	North	North
22	7.27	34.9	High	8.66	Scooter	Cloudy	03-01-2025 14:38	Central	West
23	10.19	60.2	High	10.24	Car	Windy	03-01-2025 2:11	North	South
24	2.11	9.2	Medium	2.23	Scooter	Rainy	02-01-2025 17:10	Central	North
25	3.12	18.0	High	3.41	Bicycle	Windy	04-01-2025 8:23	South	Central
26	8.22	50.4	Low	10.42	Bike	Windy	01-01-2025 8:26	East	South



```
1 #
2 # R script: Identifying and Handling Duplicates (Food Delivery Dataset)
3 #
4
5 library(dplyr)
6
7 #
8 # 1. IMPORT FOOD DELIVERY DATASET
9 #
10 df <- read.csv(
11   "C:\\Users\\itlab\\OneDrive\\Desktop\\S081_R_Studio\\Duplicate_Food_Delivery_Route_Efficiency_Dataset.csv",
12   na.strings = c("", "NA")
13 )
14
15 print("--- 1. Original Dataset ---")
16 print(head(df))
17 print(paste("Total Rows:", nrow(df)))
18
19 #
20 # 2. IDENTIFY EXACT DUPLICATE ROWS
21 #
22 # Any rows where EVERY column is the same
23
24 duplicate_rows <- df %>%
25   group_by(across(everything())) %>%
26   filter(n() > 1)
27
28 print("--- 2. Exact Duplicate Rows (if any) ---")
29 print(duplicate_rows)
30
31 #
32 # 3. REMOVE EXACT DUPLICATES
33 #
34 df_no_duplicates <- df %>% distinct()
35
36 print("--- 3. Dataset After Removing Exact Duplicates ---")
37 print(head(df_no_duplicates))
38 print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
39
40 #
41 # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
42 # Example: Keep only 1 row per order_id (First occurrence kept)
43 df_no_duplicates <- df %>%
44   group_by(order_id) %>%
45   filter(row_number() == 1)
46
47 print---
```

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 on Save Go to file/function Addins
S081_R_Practical13.R Duplicate_Food_Delivery_Route_Efficien...
17 print(paste("Total Rows:", nrow(df)))
18
19 #
20 # 2. IDENTIFY EXACT DUPLICATE ROWS
21 #
22 # Any rows where EVERY column is the same
23
24 duplicate_rows <- df %>%
25   group_by(across(everything())) %>%
26   filter(n() > 1)
27
28 print("---- 2. Exact Duplicate Rows (if any) ----")
29 print(duplicate_rows)
30
31 #
32 # 3. REMOVE EXACT DUPLICATES
33 #
34 df_no_duplicates <- df %>% distinct()
35
36 print("---- 3. Dataset After Removing Exact Duplicates ----")
37 print(head(df_no_duplicates))
38 print(paste("Rows After Removing Exact Duplicates:", nrow(df_no_duplicates)))
39
40 #
41 # 4. REMOVE DUPLICATES BASED ON A SPECIFIC COLUMN
42 # Example: Keep only 1 row per order_id (First occurrence kept)
43 #
44 unique_orders <- df %>% distinct(order_id, .keep_all = TRUE)
45
46 print("---- 4. Unique Orders Only (Duplicate order_id removed) ----")
47 print(head(unique_orders))
48 print(paste("Unique Orders Count:", nrow(unique_orders)))
49
50 #
51 # 5. REMOVE DUPLICATES BASED ON MULTIPLE COLUMNS
52 # Example: order_id + distance_km
53 #
54 unique_routes <- df %>% distinct(order_id, distance_km, .keep_all = TRUE)
55
56 print("---- 5. Unique Combination: order_id + distance_km ----")
57 print(head(unique_routes))
58
2922 [Untitled] R Script
Console
30°C Sunny
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Background Jobs
R • R 4.52.2 ~/\
> # R script: Identifying and handling Duplicates (Food Delivery Dataset)
>
> library(dplyr)
>
> # 1. IMPORT FOOD DELIVERY DATASET
>
> df <- read.csv(
+   "C:\\Users\\vitlab\\OneDrive\\Desktop\\S081_R_Studio\\Duplicate_Food_Delivery_Route_Efficiency_Dataset.csv",
+   na.strings = c("", "NA"))
>
> print("---- 1. Original dataset ----")
[1] "---- 1. Original Dataset ----"
> print(head(df))
  order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
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("Total Rows:", nrow(df)))
[1] "Total Rows: 205"
>
> # 2. IDENTIFY EXACT DUPLICATE ROWS
> #
> # Any rows where EVERY column is the same
>
> duplicate_rows <- df %>%
+   group_by(across(everything())) %>%
+   filter(n() > 1)
>
> print("---- 2. Exact Duplicate Rows (if any) ----")
[1] "---- 2. Exact Duplicate Rows (if any) ----"
> print(duplicate_rows)
# A tibble: 10 x 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
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
7         1         7.97             63.8          High           9.75      Bicycle   Clear 01-01-2025 15.29          South          North
8         2         0.90              7.6          High           1.28         Car    Cloudy 03-01-2025 0.47          West          North
9         3        11.12             78.0        Medium          16.65         Bike    Rainy 04-01-2025 17.32          South          Central
10        4         4.90             24.8          Low           5.25     Scooter   Rainy 01-01-2025 14.12          Central         Central
>
> # 3. 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))
  order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
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: 205"
>
> # 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))
  order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
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: 205"
>
> # 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))
  order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
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
>
2922 [Untitled] R Script
Console
30°C Sunny
```

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
Console Background Jobs
R - R452 - ~/...
# 2. Create duplicate rows (if any) ---
> print(duplicate_rows)
# A tibble: 10 x 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
#   <int>      <dbl>      <dbl>      <chr>      <dbl>      <chr>      <chr>      <chr>      <chr>      <chr>
1     4         4.9         24.8    Low         5.25    Scooter    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        13.62    Car        Windy    02-01-2025 9.56     West      North
4     7         9.6         54.4    Low         13.3    Bike       Windy    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        11.34    Car        Rainy    02-01-2025 16.50    West      North
8     6        11.0         76.8    High        13.62    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

> # 3. 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))
#   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#   <int>      <dbl>      <dbl>      <chr>      <dbl>      <chr>      <chr>      <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) ---")

Environment History Connections Tutorial
R - 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_pt... 200 obs. of 4 variables
unique_ord... 200 obs. of 10 variables
unique_row... 200 obs. of 10 variables
wide_df       200 obs. of 4 variables

Files Plots Packages Help Viewer Presentation
Present < Print < Edit <
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

Source
Console Background Jobs
R - R452 - ~/...
# 4. Unique orders only (Duplicate order_id removed) ---"
> print(head(unique_orders))
#   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#   <int>      <dbl>      <dbl>      <chr>      <dbl>      <chr>      <chr>      <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))
#   order_id distance_km delivery_time_min traffic_level route_length_km delivery_mode weather order_time restaurant_zone customer_zone
#   <int>      <dbl>      <dbl>      <chr>      <dbl>      <chr>      <chr>      <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
R - 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_pt... 200 obs. of 4 variables
unique_ord... 200 obs. of 10 variables
unique_row... 200 obs. of 10 variables
wide_df       200 obs. of 4 variables

Files Plots Packages Help Viewer Presentation
Present < Print < Edit <
```

Name :- Priya Gupta

Roll no. :- S081

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

S081 Priya Gupta

**Name :- Priya Gupta**

**Roll no. :- S081**