

# SHETH L.U.J AND SIR M.V. COLLEGE

## SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

### PRACTICAL – 8

**AIM:-** Applying basic data cleaning functions: handling missing values using `na.omit()`/`replace_na()` in R. import dataset.

### OUTPUT:-

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Terminal Background Jobs
R - R 4.5.2 - ~/...
C:/Users/lnro/AppData/Local/Temp/RtmpYpWHP1/downloaded_packages
> library(tidyverse)
>
> # calculate average price (ignoring NAs) to use for filling avg_price <- mean(retail_df$price, na.rm = TRUE)
> clean_replace <- taxi_df %>%
+   replace_na(list(
+     vendorID = 1,
+     trip_distance = 0,
+     passenger_count = 1,
+     fare_amount = 7
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
> # calculate average price (ignoring NAs) to use for filling avg_price <- mean(retail_df$price, na.rm = TRUE)
> clean_replace <- taxi_df %>%
+   replace_na(list(
+     vendorID = 1,
+     trip_distance = 0,
+     passenger_count = 1,
+     fare_amount = 7
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> library(dplyr)
> library(tidyverse)
>
> # 1. CREATE AND IMPORT DATASET
> # Read dataset
> taxi_df <- read.csv("C:/Users/lnro/Downloads/taxi_tripsdata.csv", na.strings = c("", "NA"))
> print("--- 1. Original Data (First 6 Rows) ---")
[1] "1. Original Data (First 6 Rows) ---"
> print(head(taxi_df))
#> #> #> #> #> #>
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
1 1 2021-07-01 00:30:52 2021-07-01 00:35:36 NA 1 74 168 1
2 2 2021-07-01 00:25:36 2021-07-01 01:01:31 NA 1 116 265 2
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 NA 1 97 33 1
4 2 2021-07-01 00:41:40 2021-07-01 00:47:23 NA 1 74 42 1
5 2 2021-07-01 00:51:32 2021-07-01 00:58:46 NA 1 42 244 1
6 1 2021-07-01 00:05:00 2021-07-01 00:11:50 NA 1 24 239 1
#> #> #> #> #> #>
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehtail_fee improvement_surcharge total_amount payment_type trip_type
1 1.20 6.0 0.5 0.5 0.0 0 NA 0.3 43.30 2 1
2 1.69 42.0 0.50 0.5 0.0 0 NA 0.3 0.00 2 1
3 0.95 6.5 0.50 0.5 2.34 0 NA 0.3 10.14 1 1
4 1.24 6.5 0.50 0.5 0.00 0 NA 0.3 7.80 2 1
5 1.10 7.0 0.50 0.5 0.00 0 NA 0.3 8.30 2 1
6 1.90 8.0 3.25 0.5 3.00 0 NA 0.3 15.05 1 1
congestion_surcharge
1 0.00
2 0.00
3 0.00
4 0.00
5 0.00
6 2.75
> print("--- Count of Missing Values per Column ---")
[1] "Count of Missing Values per Column ---"
> print(colSums(is.na(taxi_df)))
#> #> #> #> #> #>
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
32518 0 0 0 32518 32518 0 0
DOLocationID passenger_count trip_distance fare_amount extra mta_tax
0 32518 0 0 0 0
tip_amount tolls_amount ehtail_fee improvement_surcharge total_amount payment_type
0 0 0 0 0 32518
trip_type congestion_surcharge
32518 32518
#> #> #> #> #> #>
> # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
> clean_omit <- na.omit(taxi_df)
>
> print("--- 2. Data after na.omit() ---")
[1] "2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(taxi_df)))
[1] "Original rows: 83691"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 0"
> print(head(clean_omit))
#> #> #> #> #> #>
lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID fare_amount
[1] 0 0 0 0 0
[2] 0 0 0 0 0
[3] extra mta_tax tip_amount tolls_amount ehtail_fee
[4] 0 0 0 0 0
[5] 0 0 0 0 0
[6] 0 0 0 0 0
improvement_surcharge total_amount payment_type trip_type
[7] 0 0 0 0 0
[8] 0 0 0 0 0
[9] 0 0 0 0 0
[10] 0 0 0 0 0
[11] 0 0 0 0 0
[12] 0 0 0 0 0
[13] 0 0 0 0 0
[14] 0 0 0 0 0
[15] 0 0 0 0 0
[16] 0 0 0 0 0
congestion_surcharge
[17] 0 0 0 0 0
<0 rows> or 0-length row.names

```

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source Terminal Background Jobs
R - R 4.5.2 - ~/...
C:/Users/lnro/AppData/Local/Temp/RtmpYpWHP1/downloaded_packages
> library(tidyverse)
>
> # calculate average price (ignoring NAs) to use for filling avg_price <- mean(retail_df$price, na.rm = TRUE)
> clean_replace <- taxi_df %>%
+   replace_na(list(
+     vendorID = 1,
+     trip_distance = 0,
+     passenger_count = 1,
+     fare_amount = 7
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> # 3. METHOD B: REPLACE MISSING VALUES (replace_na)
> # calculate average price (ignoring NAs) to use for filling avg_price <- mean(retail_df$price, na.rm = TRUE)
> clean_replace <- taxi_df %>%
+   replace_na(list(
+     vendorID = 1,
+     trip_distance = 0,
+     passenger_count = 1,
+     fare_amount = 7
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> library(dplyr)
> library(tidyverse)
>
> # 1. CREATE AND IMPORT DATASET
> # Read dataset
> taxi_df <- read.csv("C:/Users/lnro/Downloads/taxi_tripsdata.csv", na.strings = c("", "NA"))
> print("--- 1. Original Data (First 6 Rows) ---")
[1] "1. Original Data (First 6 Rows) ---"
> print(head(taxi_df))
#> #> #> #> #> #>
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
1 1 2021-07-01 00:30:52 2021-07-01 00:35:36 NA 1 74 168 1
2 2 2021-07-01 00:25:36 2021-07-01 01:01:31 NA 1 116 265 2
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 NA 1 97 33 1
4 2 2021-07-01 00:41:40 2021-07-01 00:47:23 NA 1 74 42 1
5 2 2021-07-01 00:51:32 2021-07-01 00:58:46 NA 1 42 244 1
6 1 2021-07-01 00:05:00 2021-07-01 00:11:50 NA 1 24 239 1
#> #> #> #> #> #>
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehtail_fee improvement_surcharge total_amount payment_type trip_type
1 1.20 6.0 0.5 0.5 0.0 0 NA 0.3 43.30 2 1
2 1.69 42.0 0.50 0.5 0.0 0 NA 0.3 0.00 2 1
3 0.95 6.5 0.50 0.5 2.34 0 NA 0.3 10.14 1 1
4 1.24 6.5 0.50 0.5 0.00 0 NA 0.3 7.80 2 1
5 1.10 7.0 0.50 0.5 0.00 0 NA 0.3 8.30 2 1
6 1.90 8.0 3.25 0.5 3.00 0 NA 0.3 15.05 1 1
congestion_surcharge
1 0.00
2 0.00
3 0.00
4 0.00
5 0.00
6 2.75
> print("--- Count of Missing Values per Column ---")
[1] "Count of Missing Values per Column ---"
> print(colSums(is.na(taxi_df)))
#> #> #> #> #> #>
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
32518 0 0 0 32518 32518 0 0
DOLocationID passenger_count trip_distance fare_amount extra mta_tax
0 32518 0 0 0 0
tip_amount tolls_amount ehtail_fee improvement_surcharge total_amount payment_type
0 0 0 0 0 32518
trip_type congestion_surcharge
32518 32518
#> #> #> #> #> #>
> # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
> clean_omit <- na.omit(taxi_df)
>
> print("--- 2. Data after na.omit() ---")
[1] "2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(taxi_df)))
[1] "Original rows: 83691"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 0"
> print(head(clean_omit))
#> #> #> #> #> #>
lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID fare_amount
[1] 0 0 0 0 0
[2] 0 0 0 0 0
[3] extra mta_tax tip_amount tolls_amount ehtail_fee
[4] 0 0 0 0 0
[5] 0 0 0 0 0
[6] 0 0 0 0 0
improvement_surcharge total_amount payment_type trip_type
[7] 0 0 0 0 0
[8] 0 0 0 0 0
[9] 0 0 0 0 0
[10] 0 0 0 0 0
[11] 0 0 0 0 0
[12] 0 0 0 0 0
[13] 0 0 0 0 0
[14] 0 0 0 0 0
[15] 0 0 0 0 0
[16] 0 0 0 0 0
congestion_surcharge
[17] 0 0 0 0 0
<0 rows> or 0-length row.names

```

# SHETH L.U.J AND SIR M.V. COLLEGE

## SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

RStudio Environment pane showing various datasets loaded:

- merged\_data (3 obs. of 4 variables)
- my\_data (1000 obs. of 8 variables)
- range\_cols (3900 obs. of 6 variables)
- selected\_cols (3900 obs. of 3 variables)
- shopping (3900 obs. of 18 variables)
- spotify\_data (8582 obs. of 15 variables)
- starts\_with\_r (3900 obs. of 1 variable)
- student\_stre\_ (520 obs. of 6 variables)
- studentsperf\_ (1000 obs. of 8 variables)
- taxis\_df (83691 obs. of 20 variables)

Console output showing data cleaning steps:

```

Console Terminal Background Jobs
> R 4.2.1
> replace_na(list(
+   vendorID = 1,
+   trip_distance = 0,
+   passenger_count = 1,
+   fare_amount = 7
+ ))
> print("---- 3. Data after replace_na() ----")
[1] "---- 3. Data after replace_na() ----"
> print(clean_replace[3,])
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 N 1 97 33 1
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehaif_fee improvement_surcharge total_amount payment_type trip_type
3 0.95 6.5 0.5 0.5 2.34 0 NA 0.3 10.14 1 1
congestion_surcharge
3 0
> print(head(clean_replace))
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
1 1 2021-07-01 00:05:52 2021-07-01 00:15:31 N 1 74 168 1
2 2 2021-07-01 00:05:57 2021-07-01 00:10:21 N 1 116 265 2
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 N 1 97 33 1
4 2 2021-07-01 00:41:40 2021-07-01 00:47:23 N 1 74 42 1
5 2 2021-07-01 00:51:32 2021-07-01 00:58:46 N 1 42 244 1
6 1 2021-07-01 00:05:00 2021-07-01 00:11:50 N 1 24 239 1
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehaif_fee improvement_surcharge total_amount payment_type trip_type
1 1.20 6.0 0.50 0.5 0.00 0 NA 0.3 7.30 2 1
2 13.69 42.0 0.50 0.5 0.00 0 NA 0.3 43.30 2 1
3 0.95 6.5 0.50 0.5 2.34 0 NA 0.3 10.14 1 1
4 1.24 6.5 0.50 0.5 0.00 0 NA 0.3 7.80 2 1
5 1.10 7.0 0.50 0.5 0.00 0 NA 0.3 8.30 2 1
6 1.90 8.0 3.25 0.5 3.00 0 NA 0.3 15.05 1 1
congestion_surcharge
1 0.00
2 0.00

```

RStudio Environment pane showing various datasets loaded:

- merged\_data (3 obs. of 4 variables)
- my\_data (1000 obs. of 8 variables)
- range\_cols (3900 obs. of 6 variables)
- selected\_cols (3900 obs. of 3 variables)
- shopping (3900 obs. of 18 variables)
- spotify\_data (8582 obs. of 15 variables)
- starts\_with\_r (3900 obs. of 1 variable)
- student\_stre\_ (520 obs. of 6 variables)
- studentsperf\_ (1000 obs. of 8 variables)
- taxis\_df (83691 obs. of 20 variables)

Console output showing data cleaning steps:

```

Console Terminal Background Jobs
> R 4.2.1
> replace_na(list(
+   vendorID = 1,
+   trip_distance = 0,
+   passenger_count = 1,
+   fare_amount = 7
+ ))
> print("---- 3. Data after replace_na() ----")
[1] "---- 3. Data after replace_na() ----"
> print(clean_replace[3,])
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 N 1 97 33 1
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehaif_fee improvement_surcharge total_amount payment_type trip_type
3 0.95 6.5 0.5 0.5 2.34 0 NA 0.3 10.14 1 1
congestion_surcharge
3 0
> print(head(clean_replace))
VendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID PULocationID DOLocationID passenger_count
1 1 2021-07-01 00:30:52 2021-07-01 00:35:36 N 1 74 168 1
2 2 2021-07-01 00:25:36 2021-07-01 01:01:31 N 1 116 265 2
3 2 2021-07-01 00:05:58 2021-07-01 00:12:00 N 1 97 33 1
4 2 2021-07-01 00:41:40 2021-07-01 00:47:23 N 1 74 42 1
5 2 2021-07-01 00:51:32 2021-07-01 00:58:46 N 1 42 244 1
6 1 2021-07-01 00:05:00 2021-07-01 00:11:50 N 1 24 239 1
trip_distance fare_amount extra mta_tax tip_amount tolls_amount ehaif_fee improvement_surcharge total_amount payment_type trip_type
1 1.20 6.0 0.50 0.5 0.00 0 NA 0.3 7.30 2 1
2 13.69 42.0 0.50 0.5 0.00 0 NA 0.3 43.30 2 1
3 0.95 6.5 0.50 0.5 2.34 0 NA 0.3 10.14 1 1
4 1.24 6.5 0.50 0.5 0.00 0 NA 0.3 7.80 2 1
5 1.10 7.0 0.50 0.5 0.00 0 NA 0.3 8.30 2 1
6 1.90 8.0 3.25 0.5 3.00 0 NA 0.3 15.05 1 1
congestion_surcharge
1 0.00
2 0.00
3 0.00
4 0.00
5 0.00
6 2.75
> print("---- Remaining NAs after replacement ----")
[1] "---- Remaining NAs after replacement ----"
> print(coltsum(is.na(clean_replace)))
+ 
```