

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
R - R452 - ~/
C:\Users\info\AppData\Local\Temp\RtmpYPWihp\downloaded_packages
> library(tidy)
> # 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(tidy)
> # 1. CREATE AND IMPORT DATASET
> # Read dataset
> taxi_df <- read_csv("C:\\Users\\info\\downloads\\taxi_tripdata.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              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 ehail_fee improvement_surcharge total_amount payment_type trip_type
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
R - R452 - ~/
> 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      0      0      0      0      0
  DOLocationID passenger_count trip_distance fare_amount extra
0      32518      0      0      0
  tip_amount tolls_amount ehail_fee improvement_surcharge total_amount payment_type
0      0      0      0      0      0
  trip_type congestion_surcharge
32518      32518

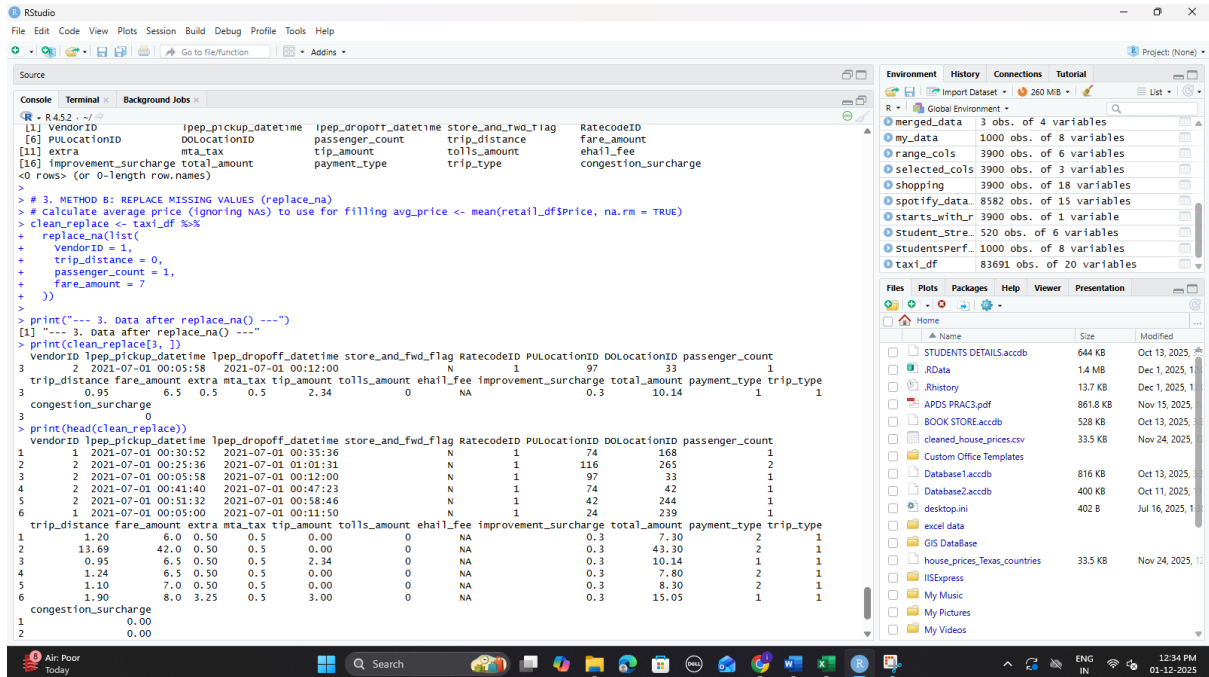
> # 2. METHOD A: REMOVE MISSING VALUES (na.omit)
> clean_omit <- na.omit(taxi_df)
>
> print("---- 2. Data after na.omit() ----")
[1] "---- 2. Data after na.omit() ----"
> print(paste("Original rows:", nrow(taxi_df)))
[1] "Original rows: 83691"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 0"
> print(head(clean_omit))
[1] vendorID lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID
[6] PULocationID DOLocationID passenger_count trip_distance fare_amount extra
[11] extra mta_tax tip_amount tolls_amount ehail_fee improvement_surcharge total_amount payment_type trip_type
[16] improvement_surcharge total_amount payment_type trip_type congestion_surcharge
<0 rows> (or 0-length row.names)
```

SUMEET JITENDRA YADAV

S124

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

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

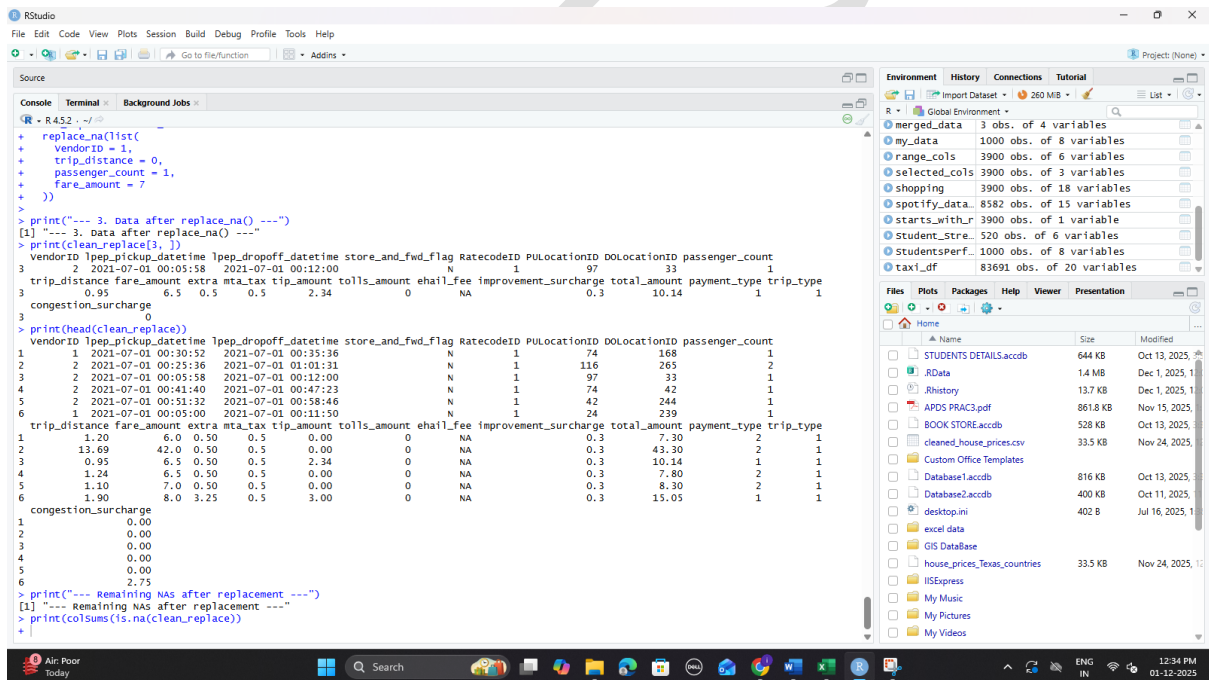


The screenshot shows the RStudio interface. The console displays the following R code and its output:

```
R - R452 - ~/ -  
[1] VendorID      lpep_pickup_datetime lpep_dropoff_datetime store_and_fwd_flag RatecodeID  
[6] PULocationID DOLocationID mta_tax trip_distance Fare_amount ehail_fee  
[11] extra          tip_amount tolls_amount improvement_surcharge total_amount payment_type trip_type  
[16] congestion_surcharge  
<0 rows> (or 0-length row.names)  
>  
> # 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() ----"  
> 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 ehail_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 ehail_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
```

The file explorer on the right shows the following files:

- 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
- taxi_df 83691 obs. of 20 variables



The screenshot shows the RStudio interface. The console displays the following R code and its output:

```
R - R452 - ~/ -  
+   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 ehail_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 ehail_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(colSums(is.na(clean_replace)))  
+ 
```

The file explorer on the right shows the following files:

- 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
- taxi_df 83691 obs. of 20 variables