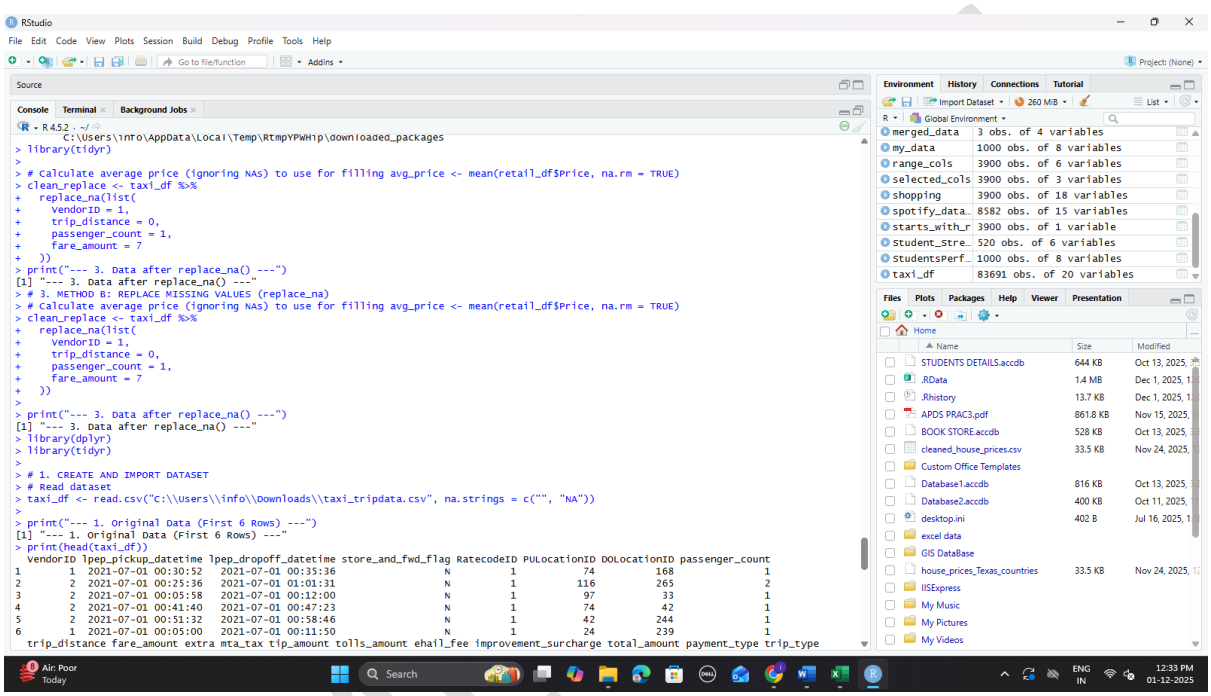
SHETH L.U.J AND SIR M.V. COLLEGE

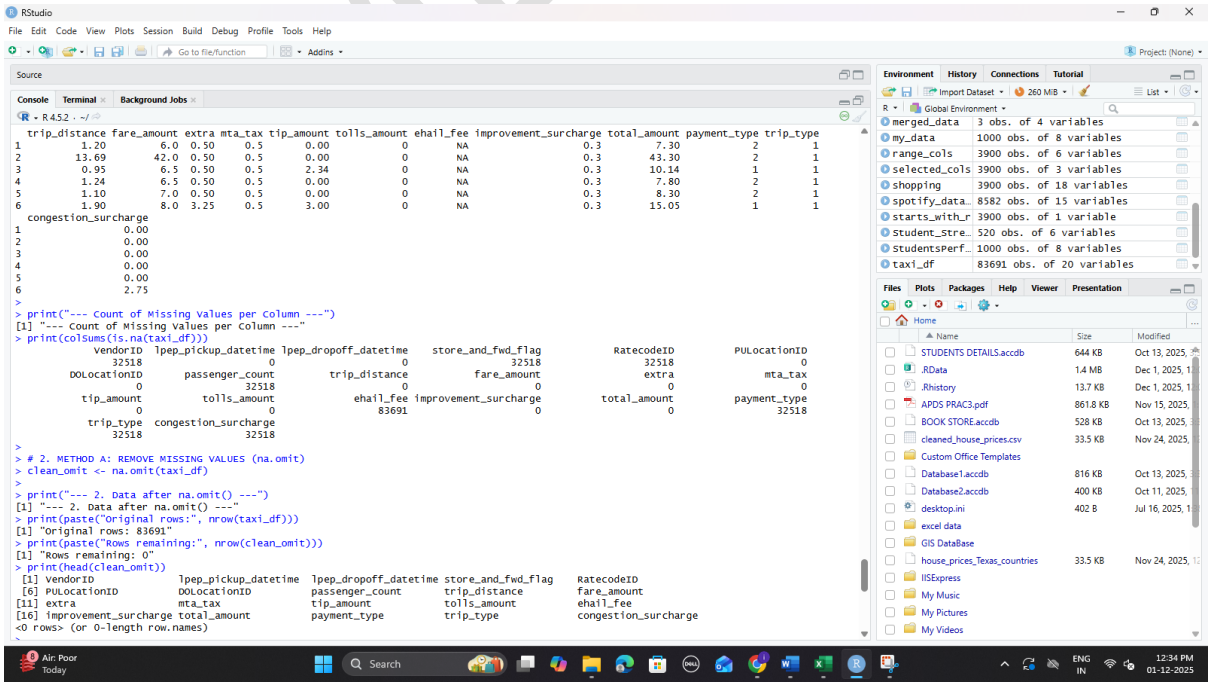
SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

PRACTICAL – 8

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

OUTPUT:-

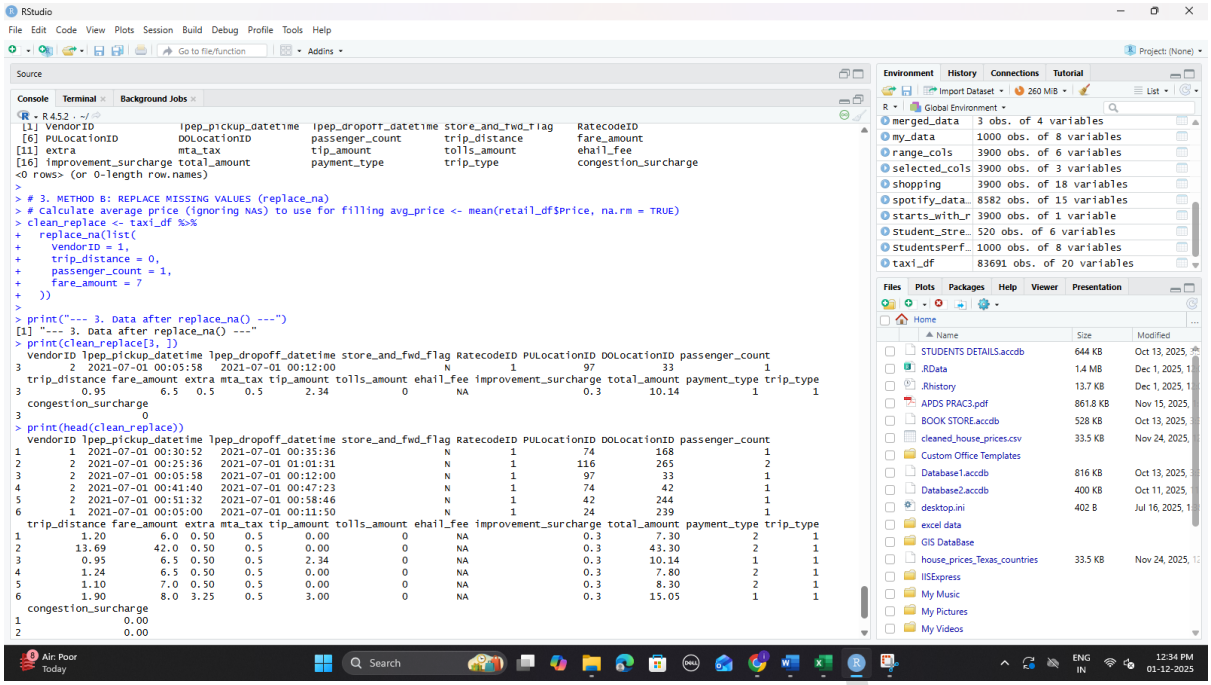




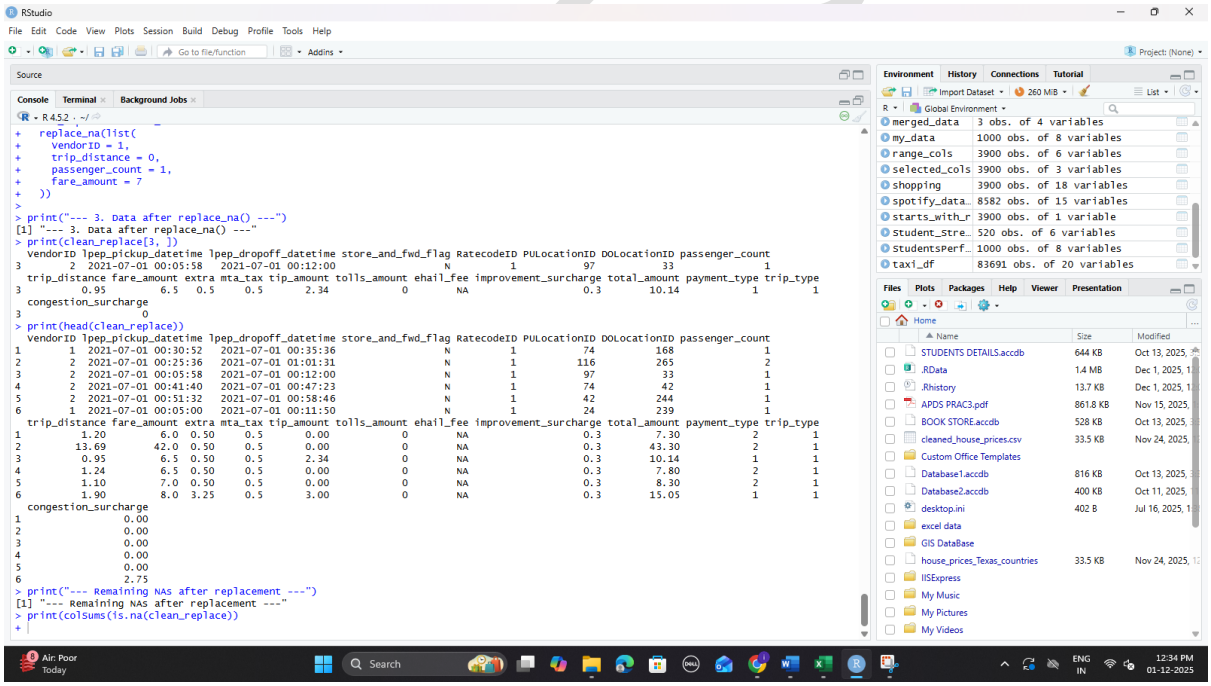
SUMEET JITENDRA YADAV
S124

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

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R



```
 [1] VendorID              lpep_pickup_datetime  lpep_dropoff_datetime store_and_fwd_flag   RatecodeID
 [6] PULocationID          DOLocationID          passenger_count       trip_distance       fare_amount
[11] extra                 mta_tax               tip_amount            tolls_amount        ehail_fee
[16] improvement_surcharge total_amount          payment_type          trip_type           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
```



```
+   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))
+
```


SUMEET JITENDRA YADAV

S124