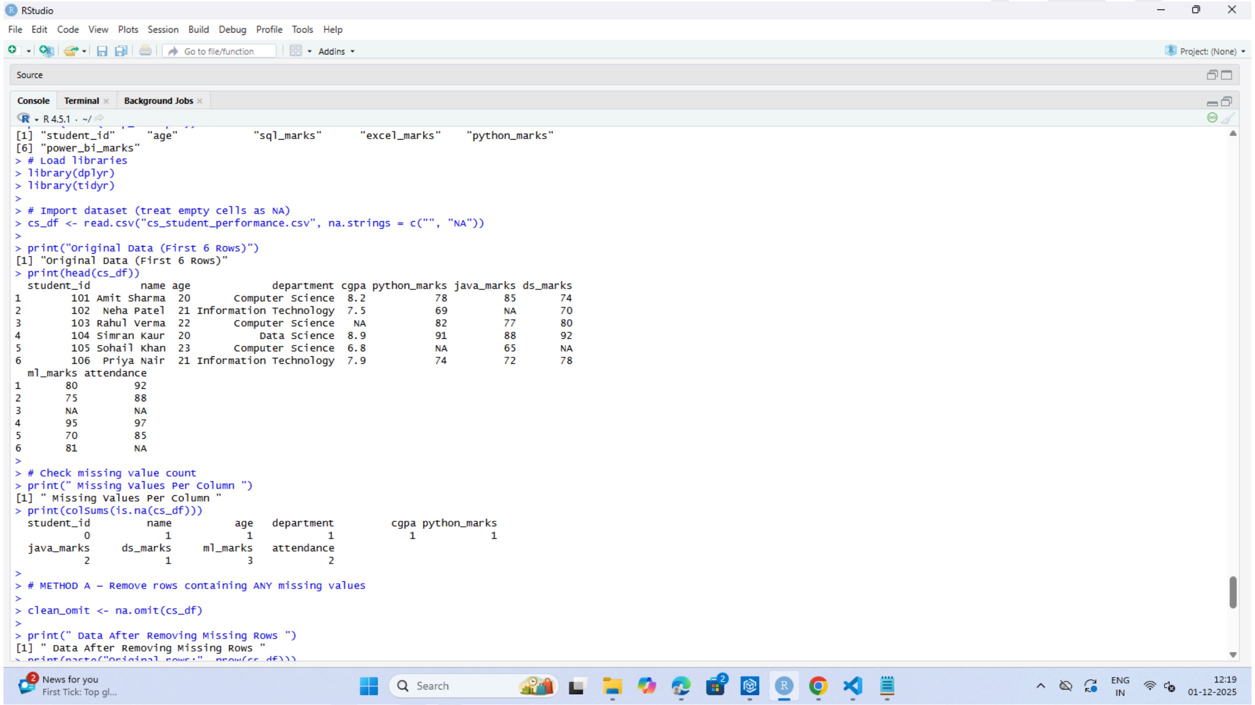
SHETH L.U.J. AND SIR M.V. COLLEGE OF ARTS SCIENCE AND COMMERCE

SUBJECT : R Programming

Practical No 8

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

Output :

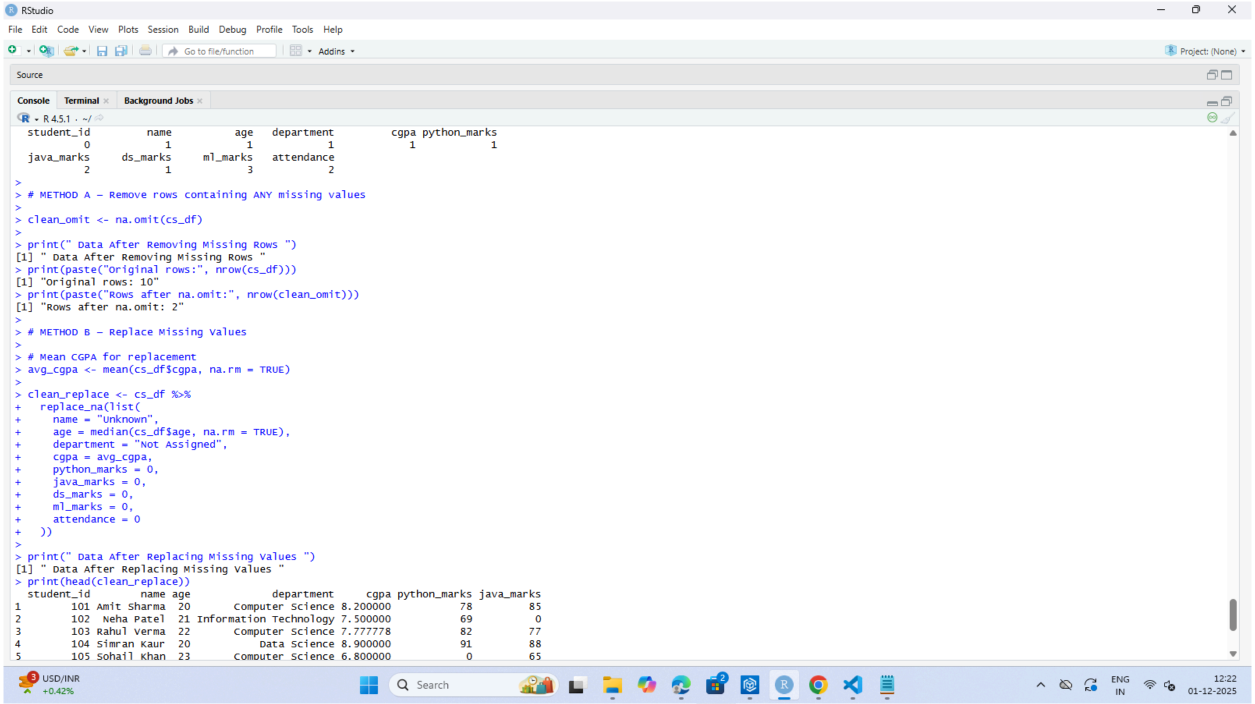


```
[1] "student_id"     "age"            "sql_marks"      "excel_marks"    "python_marks"
[6] "power_bi_marks"
> # Load libraries
> library(dplyr)
> library(tidyr)
>
> # Import dataset (treat empty cells as NA)
> cs_df <- read.csv("cs_student_performance.csv", na.strings = c("", "NA"))
>
> print("Original Data (First 6 Rows)")
[1] "Original Data (First 6 Rows)"
> print(head(cs_df))
  student_id         name age             department cgpa python_marks java_marks ds_marks
1        101 Amit Sharma  20       Computer Science  8.2           78         85       74
2        102  Neha Patel  21 Information Technology  7.5           69         NA       70
3        103 Rahul Verma  22       Computer Science   NA           82         77       80
4        104 Simran Kaur  20           Data Science  8.9           91         88       92
5        105 Sohail Khan  23       Computer Science  6.8           NA         65       NA
6        106  Priya Nair  21 Information Technology  7.9           74         72       78
  ml_marks attendance
1       80         92
2       75         88
3       NA         NA
4       95         97
5       70         85
6       81         NA
>
> # Check missing value count
> print(" Missing Values Per Column ")
[1] " Missing Values Per Column "
> print(colSums(is.na(cs_df)))
  student_id         name          age   department         cgpa python_marks
           0            1            1            1            1            1
  java_marks     ds_marks     ml_marks   attendance
           2            1            3            2
>
> # METHOD A - Remove rows containing ANY missing values
>
> clean_omit <- na.omit(cs_df)
>
> print(" Data After Removing Missing Rows ")
[1] " Data After Removing Missing Rows "
> print(paste("Original rows:", nrow(cs_df)))
```

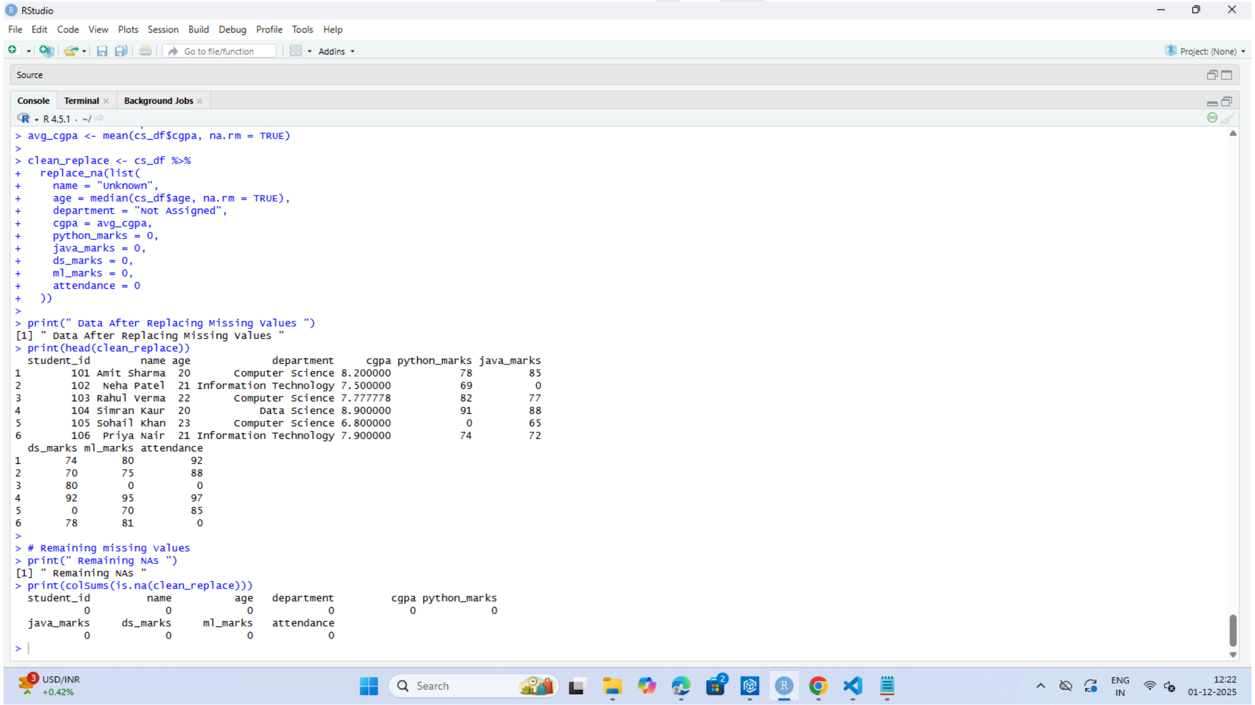
NAME : SHUBHAM SANJAY KARAPE
ROLL NO : S085

SHETH L.U.J. AND SIR M.V. COLLEGE OF ARTS SCIENCE AND COMMERCE

SUBJECT : R Programming



```
  student_id         name          age   department          cgpa python_marks 
           0            1            1            1             1            1 
  java_marks     ds_marks     ml_marks   attendance 
           2            1            3            2 
> 
> # METHOD A – Remove rows containing ANY missing values
> 
> clean_omit <- na.omit(cs_df)
> 
> print(" Data After Removing Missing Rows ")
[1] " Data After Removing Missing Rows "
> print(paste("Original rows:", nrow(cs_df)))
[1] "Original rows: 10"
> print(paste("Rows after na.omit:", nrow(clean_omit)))
[1] "Rows after na.omit: 2"
> 
> # METHOD B – Replace Missing Values
> 
> # Mean CGPA for replacement
> avg_cgpa <- mean(cs_df$cgpa, na.rm = TRUE)
> 
> clean_replace <- cs_df %>%
+   replace_na(list(
+     name = "Unknown",
+     age = median(cs_df$age, na.rm = TRUE),
+     department = "Not Assigned",
+     cgpa = avg_cgpa,
+     python_marks = 0,
+     java_marks = 0,
+     ds_marks = 0,
+     ml_marks = 0,
+     attendance = 0
+   ))
> 
> print(" Data After Replacing Missing Values ")
[1] " Data After Replacing Missing Values "
> print(head(clean_replace))
  student_id         name age             department     cgpa python_marks java_marks
1        101 Amit Sharma  20       Computer Science 8.200000           78         85
2        102  Neha Patel  21 Information Technology 7.500000           69          0
3        103 Rahul Verma  22       Computer Science 7.777778           82         77
4        104 Simran Kaur  20           Data Science 8.900000           91         88
5        105 Sohail Khan  23       Computer Science 6.800000            0         65
```



```
> avg_cgpa <- mean(cs_df$cgpa, na.rm = TRUE)
> 
> clean_replace <- cs_df %>%
+   replace_na(list(
+     name = "Unknown",
+     age = median(cs_df$age, na.rm = TRUE),
+     department = "Not Assigned",
+     cgpa = avg_cgpa,
+     python_marks = 0,
+     java_marks = 0,
+     ds_marks = 0,
+     ml_marks = 0,
+     attendance = 0
+   ))
> 
> print(" Data After Replacing Missing Values ")
[1] " Data After Replacing Missing Values "
> print(head(clean_replace))
  student_id         name age             department     cgpa python_marks java_marks
1        101 Amit Sharma  20       Computer Science 8.200000           78         85
2        102  Neha Patel  21 Information Technology 7.500000           69          0
3        103 Rahul Verma  22       Computer Science 7.777778           82         77
4        104 Simran Kaur  20           Data Science 8.900000           91         88
5        105 Sohail Khan  23       Computer Science 6.800000            0         65
6        106  Priya Nair  21 Information Technology 7.900000           74         72
  ds_marks ml_marks attendance
1       74       80         92
2       70       75         88
3       80        0          0
4       92       95         97
5        0       70         85
6       78       81          0
> 
> # Remaining missing values
> print(" Remaining NAs ")
[1] " Remaining NAs "
> print(colSums(is.na(clean_replace)))
  student_id         name          age   department          cgpa python_marks 
           0            0            0            0             0            0 
  java_marks     ds_marks     ml_marks   attendance 
           0            0            0            0 
> |
```

NAME : SHUBHAM SANJAY KARAPE

ROLL NO : S085