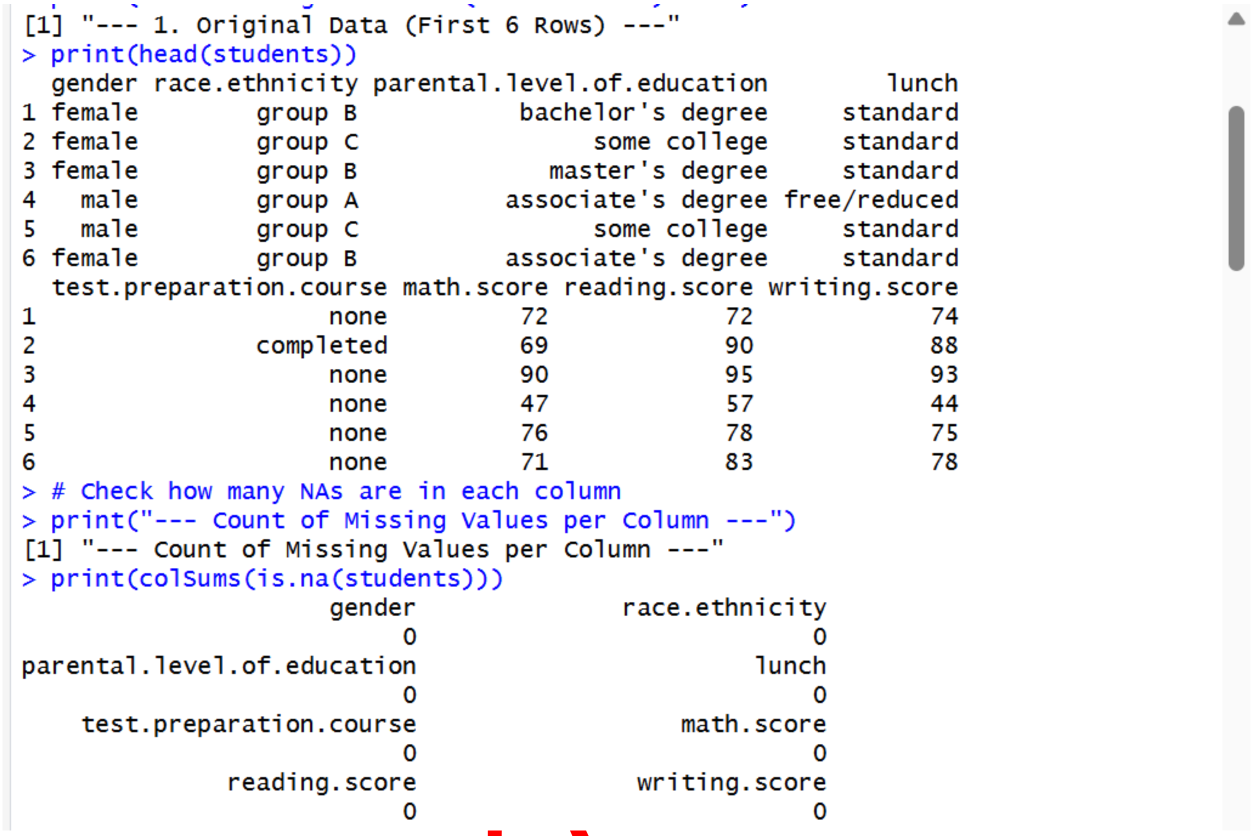
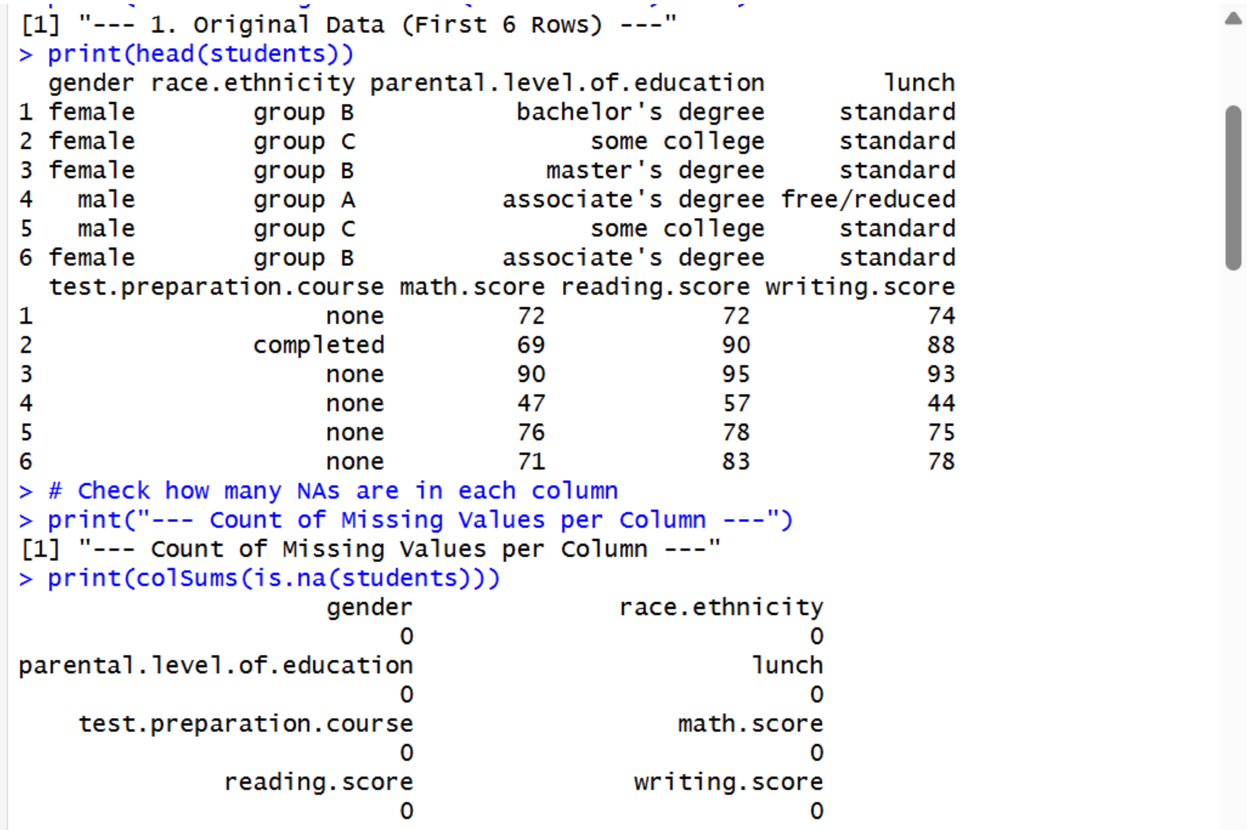
SHETH L.U.J. & SIR M.V. COLLEGE OF SCIENCE
SUBJECT - Data Analysis with SAS / SPSS / R

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

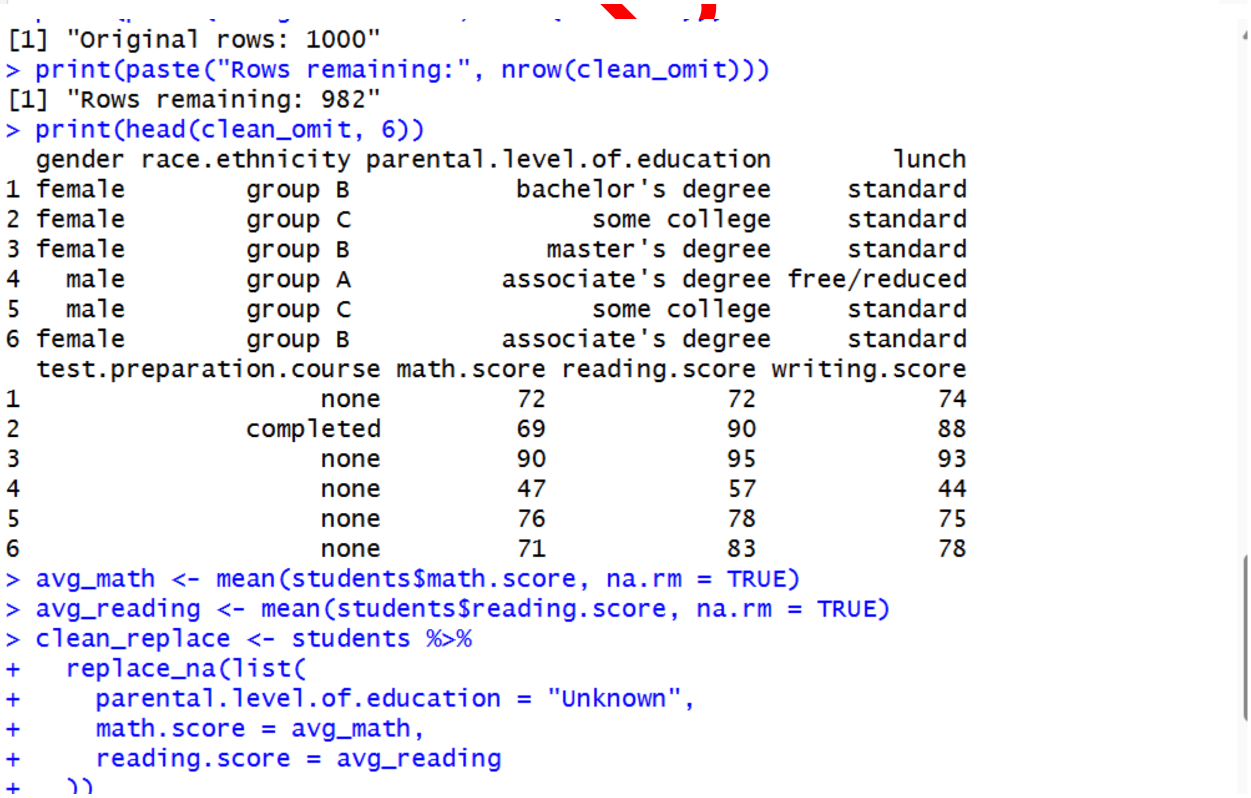


```
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(students))
  gender race.ethnicity parental.level.of.education        lunch
1 female        group B           bachelor's degree     standard
2 female        group C                some college     standard
3 female        group B             master's degree     standard
4   male        group A          associate's degree free/reduced
5   male        group C                some college     standard
6 female        group B          associate's degree     standard
  test.preparation.course math.score reading.score writing.score
1                    none         72            72            74
2               completed         69            90            88
3                    none         90            95            93
4                    none         47            57            44
5                    none         76            78            75
6                    none         71            83            78
> # Check how many NAs are in each column
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> print(colSums(is.na(students)))
                     gender              race.ethnicity 
                          0                           0 
parental.level.of.education                       lunch 
                          0                           0 
    test.preparation.course                  math.score 
                          0                           0 
              reading.score               writing.score 
                          0                           0 
```

Name - Mithil Kadam
Roll No - S083



```
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(students))
  gender race.ethnicity parental.level.of.education        lunch
1 female        group B           bachelor's degree     standard
2 female        group C                some college     standard
3 female        group B             master's degree     standard
4   male        group A          associate's degree free/reduced
5   male        group C                some college     standard
6 female        group B          associate's degree     standard
  test.preparation.course math.score reading.score writing.score
1                    none         72            72            74
2               completed         69            90            88
3                    none         90            95            93
4                    none         47            57            44
5                    none         76            78            75
6                    none         71            83            78
> # Check how many NAs are in each column
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> print(colSums(is.na(students)))
                     gender              race.ethnicity 
                          0                           0 
parental.level.of.education                       lunch 
                          0                           0 
    test.preparation.course                  math.score 
                          0                           0 
              reading.score               writing.score 
                          0                           0 
```



```
[1] "Original rows: 1000"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 982"
> print(head(clean_omit, 6))
  gender race.ethnicity parental.level.of.education        lunch
1 female        group B           bachelor's degree     standard
2 female        group C                some college     standard
3 female        group B             master's degree     standard
4   male        group A          associate's degree free/reduced
5   male        group C                some college     standard
6 female        group B          associate's degree     standard
  test.preparation.course math.score reading.score writing.score
1                    none         72            72            74
2               completed         69            90            88
3                    none         90            95            93
4                    none         47            57            44
5                    none         76            78            75
6                    none         71            83            78
> avg_math <- mean(students$math.score, na.rm = TRUE)
> avg_reading <- mean(students$reading.score, na.rm = TRUE)
> clean_replace <- students %>%
+   replace_na(list(
+     parental.level.of.education = "Unknown",
+     math.score = avg_math,
+     reading.score = avg_reading
+   ))
```