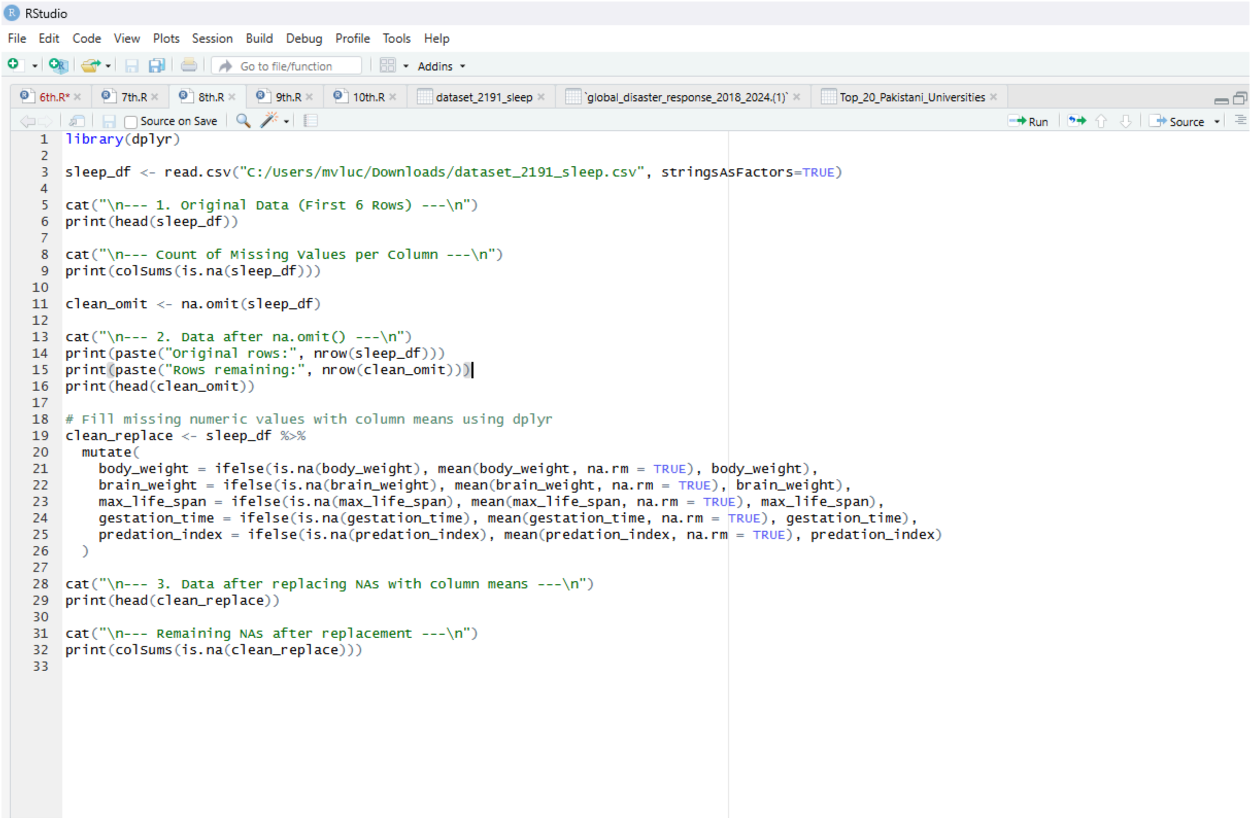
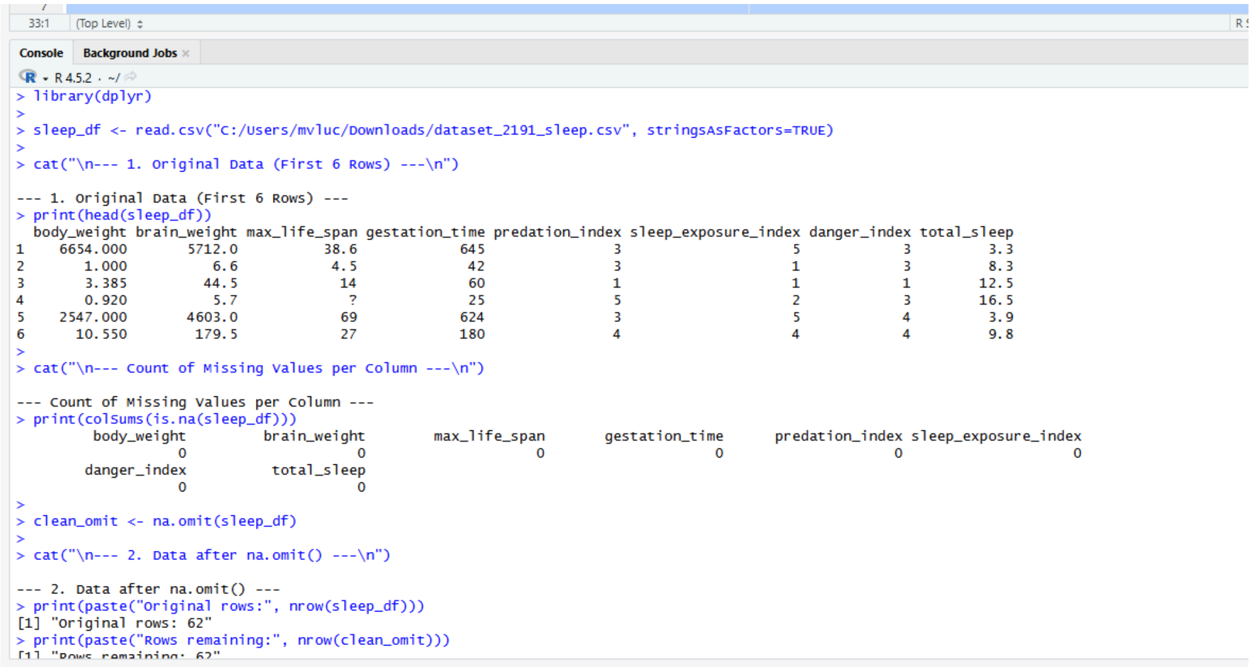
Sheth l.u.j. And sir m.v. college of arts science and commerce

Practical 8 R

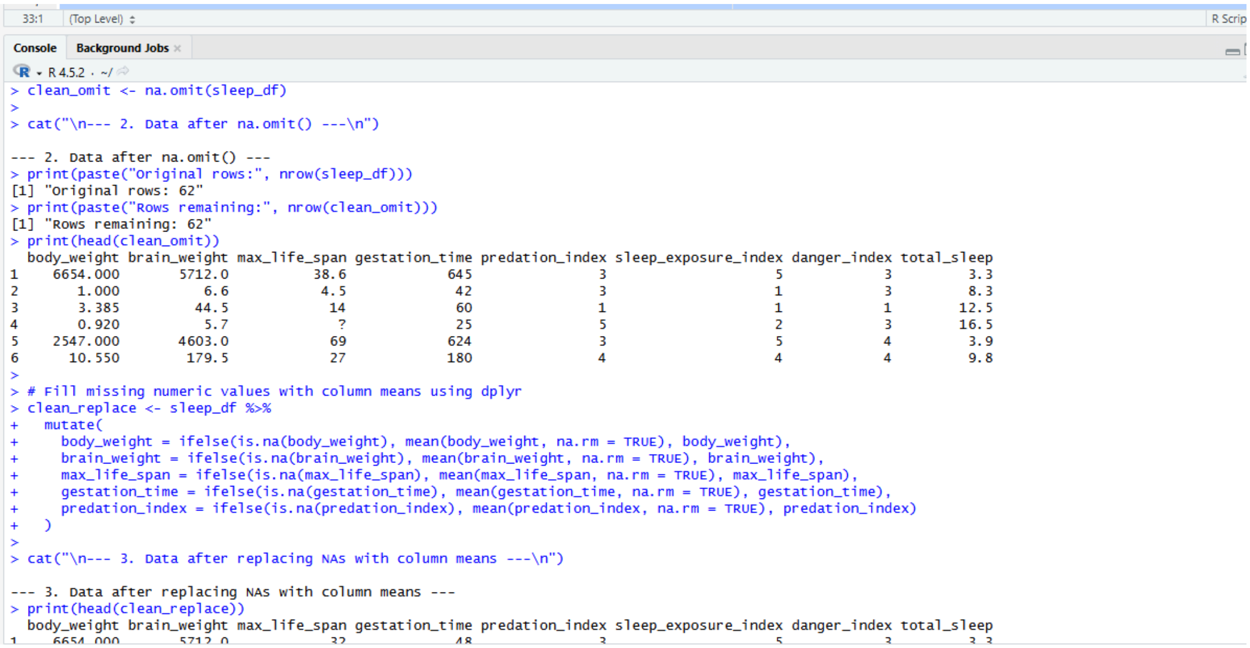
Applying basic data cleaning functions: handling missing values using na.omit()/replace_na() in R. import dataset.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
6th.R*
7th.R
8th.R
9th.R
10th.R
dataset_2191_sleep
`global_disaster_response_2018_2024.(1)`
Top_20_Pakistani_Universities
Source on Save
Run
Source

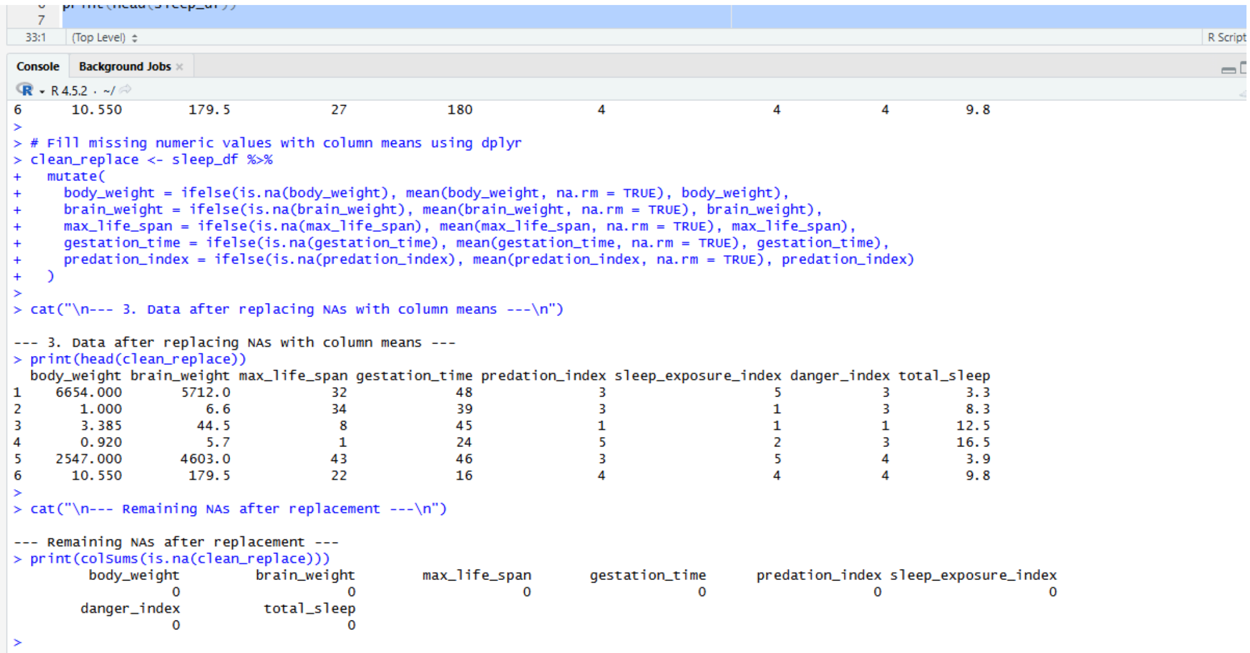
```r
library(dplyr)

sleep_df <- read.csv("C:/Users/mvluc/Downloads/dataset_2191_sleep.csv", stringsAsFactors=TRUE)

cat("\n--- 1. Original Data (First 6 Rows) ---\n")
print(head(sleep_df))

cat("\n--- Count of Missing Values per Column ---\n")
print(colSums(is.na(sleep_df)))

clean_omit <- na.omit(sleep_df)

cat("\n--- 2. Data after na.omit() ---\n")
print(paste("Original rows:", nrow(sleep_df)))
print(paste("Rows remaining:", nrow(clean_omit)))
print(head(clean_omit))

# Fill missing numeric values with column means using dplyr
clean_replace <- sleep_df %>%
  mutate(
    body_weight = ifelse(is.na(body_weight), mean(body_weight, na.rm = TRUE), body_weight),
    brain_weight = ifelse(is.na(brain_weight), mean(brain_weight, na.rm = TRUE), brain_weight),
    max_life_span = ifelse(is.na(max_life_span), mean(max_life_span, na.rm = TRUE), max_life_span),
    gestation_time = ifelse(is.na(gestation_time), mean(gestation_time, na.rm = TRUE), gestation_time),
    predation_index = ifelse(is.na(predation_index), mean(predation_index, na.rm = TRUE), predation_index)
  )

cat("\n--- 3. Data after replacing NAs with column means ---\n")
print(head(clean_replace))

cat("\n--- Remaining NAs after replacement ---\n")
print(colSums(is.na(clean_replace)))
```



```
> library(dplyr)
>
> sleep_df <- read.csv("C:/Users/mvluc/Downloads/dataset_2191_sleep.csv", stringsAsFactors=TRUE)
>
> cat("\n--- 1. Original Data (First 6 Rows) ---\n")

--- 1. Original Data (First 6 Rows) ---
> print(head(sleep_df))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index danger_index total_sleep
1    6654.000       5712.0          38.6            645               3                    5            3         3.3
2       1.000          6.6           4.5             42               3                    1            3         8.3
3       3.385         44.5            14             60               1                    1            1        12.5
4       0.920          5.7             ?             25               5                    2            3        16.5
5    2547.000       4603.0            69            624               3                    5            4         3.9
6      10.550        179.5            27            180               4                    4            4         9.8
>
> cat("\n--- Count of Missing Values per Column ---\n")

--- Count of Missing Values per Column ---
> print(colSums(is.na(sleep_df)))
         body_weight         brain_weight        max_life_span       gestation_time      predation_index sleep_exposure_index
                   0                    0                    0                    0                    0                    0
        danger_index          total_sleep
                   0                    0
>
> clean_omit <- na.omit(sleep_df)
>
> cat("\n--- 2. Data after na.omit() ---\n")

--- 2. Data after na.omit() ---
> print(paste("Original rows:", nrow(sleep_df)))
[1] "Original rows: 62"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 62"
```



```
> clean_omit <- na.omit(sleep_df)
>
> cat("\n--- 2. Data after na.omit() ---\n")

--- 2. Data after na.omit() ---
> print(paste("Original rows:", nrow(sleep_df)))
[1] "Original rows: 62"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 62"
> print(head(clean_omit))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index danger_index total_sleep
1    6654.000       5712.0          38.6            645               3                    5            3         3.3
2       1.000          6.6           4.5             42               3                    1            3         8.3
3       3.385         44.5            14             60               1                    1            1        12.5
4       0.920          5.7             ?             25               5                    2            3        16.5
5    2547.000       4603.0            69            624               3                    5            4         3.9
6      10.550        179.5            27            180               4                    4            4         9.8
>
> # Fill missing numeric values with column means using dplyr
> clean_replace <- sleep_df %>%
+   mutate(
+     body_weight = ifelse(is.na(body_weight), mean(body_weight, na.rm = TRUE), body_weight),
+     brain_weight = ifelse(is.na(brain_weight), mean(brain_weight, na.rm = TRUE), brain_weight),
+     max_life_span = ifelse(is.na(max_life_span), mean(max_life_span, na.rm = TRUE), max_life_span),
+     gestation_time = ifelse(is.na(gestation_time), mean(gestation_time, na.rm = TRUE), gestation_time),
+     predation_index = ifelse(is.na(predation_index), mean(predation_index, na.rm = TRUE), predation_index)
+   )
>
> cat("\n--- 3. Data after replacing NAs with column means ---\n")

--- 3. Data after replacing NAs with column means ---
> print(head(clean_replace))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index danger_index total_sleep
1    6654.000       5712.0            32             48               3                    5            3         3.3
```



```
    print(head(sleep_df))
```

```
6      10.550        179.5            27            180               4                    4            4         9.8
>
> # Fill missing numeric values with column means using dplyr
> clean_replace <- sleep_df %>%
+   mutate(
+     body_weight = ifelse(is.na(body_weight), mean(body_weight, na.rm = TRUE), body_weight),
+     brain_weight = ifelse(is.na(brain_weight), mean(brain_weight, na.rm = TRUE), brain_weight),
+     max_life_span = ifelse(is.na(max_life_span), mean(max_life_span, na.rm = TRUE), max_life_span),
+     gestation_time = ifelse(is.na(gestation_time), mean(gestation_time, na.rm = TRUE), gestation_time),
+     predation_index = ifelse(is.na(predation_index), mean(predation_index, na.rm = TRUE), predation_index)
+   )
>
> cat("\n--- 3. Data after replacing NAs with column means ---\n")

--- 3. Data after replacing NAs with column means ---
> print(head(clean_replace))
  body_weight brain_weight max_life_span gestation_time predation_index sleep_exposure_index danger_index total_sleep
1    6654.000       5712.0            32             48               3                    5            3         3.3
2       1.000          6.6            34             39               3                    1            3         8.3
3       3.385         44.5             8             45               1                    1            1        12.5
4       0.920          5.7             1             24               5                    2            3        16.5
5    2547.000       4603.0            43             46               3                    5            4         3.9
6      10.550        179.5            22             16               4                    4            4         9.8
>
> cat("\n--- Remaining NAs after replacement ---\n")

--- Remaining NAs after replacement ---
> print(colSums(is.na(clean_replace)))
         body_weight         brain_weight        max_life_span       gestation_time      predation_index sleep_exposure_index
                   0                    0                    0                    0                    0                    0
        danger_index          total_sleep
                   0                    0
>
```

Sheth l.u.j. And sir m.v. college of arts science and commerce

Name: Simran S113