Stat107 Final Project Data Cleaning

Group 15

11-5-25

```
sleep_data <- read.csv("sleep_deprivation_dataset_detailed.csv")
head(sleep_data)
```

```
##   Participant_ID Sleep_Hours Sleep_Quality_Score Daytime_Sleepiness
## 1             P1        5.25                  15                 12
## 2             P2        8.70                  12                 14
## 3             P3        7.39                  17                 10
## 4             P4        6.59                  14                  3
## 5             P5        3.94                  20                 12
## 6             P6        3.94                  12                  6
##   Stroop_Task_Reaction_Time N_Back_Accuracy Emotion_Regulation_Score
## 1                      1.60           64.20                       12
## 2                      2.54           65.27                       21
## 3                      3.40           74.28                       35
## 4                      3.54           72.42                       25
## 5                      3.09           99.72                       60
## 6                      2.84           58.80                       46
##   PVT_Reaction_Time Age Gender   BMI Caffeine_Intake Physical_Activity_Level
## 1            365.85  35 Female 30.53               2                      1
## 2            288.95  20   Male 27.28               3                      8
## 3            325.93  18   Male 30.00               1                      2
## 4            276.86  18   Male 34.47               5                      0
## 5            383.45  36   Male 29.70               3                      4
## 6            224.48  28   Male 32.23               3                      6
##   Stress_Level
## 1           33
## 2           37
## 3           32
## 4           23
## 5           14
## 6           29
```

In order to ensure the data is ready for use in our analysis, we must first clean it such that each variable is formatted in a predicatble manner and there are not missing data values. To do this we define the below functions:

```
#########################################################
##These functions are taken from DA6
#########################################################
remove_na <- function(df) { ## Automatize the removal of missing observations.
  n_obs <- nrow(df)
```

```
  missing <- rep(FALSE, n_obs)
  for (obs_ind in 1:n_obs) {
    obs <- df[obs_ind, ]
    n_missing <- sum(is.na(obs))
    if(n_missing > 0) {
      missing[obs_ind] <- TRUE
    }
  }
  df_red <- df[!missing, ]

  return(df_red)
}

############################################################
############################################################
to_factors <- function(df) { ## Turn all character variables to factors.
  n_vars <- ncol(df)
  for (var_ind in 1:n_vars) {
    var <- df[, var_ind]
    if (class(var) == "character") {
      df[, var_ind] <- factor(var)
    }
  }
  return(df)
}
```

Apply the functions to the data set:

```
sleep_df <- to_factors(remove_na(sleep_data))
head(sleep_df)
```

```
##   Participant_ID Sleep_Hours Sleep_Quality_Score Daytime_Sleepiness
## 1             P1        5.25                  15                 12
## 2             P2        8.70                  12                 14
## 3             P3        7.39                  17                 10
## 4             P4        6.59                  14                  3
## 5             P5        3.94                  20                 12
## 6             P6        3.94                  12                  6
##   Stroop_Task_Reaction_Time N_Back_Accuracy Emotion_Regulation_Score
## 1                      1.60           64.20                       12
## 2                      2.54           65.27                       21
## 3                      3.40           74.28                       35
## 4                      3.54           72.42                       25
## 5                      3.09           99.72                       60
## 6                      2.84           58.80                       46
##   PVT_Reaction_Time Age Gender   BMI Caffeine_Intake Physical_Activity_Level
## 1            365.85  35 Female 30.53               2                      1
## 2            288.95  20   Male 27.28               3                      8
## 3            325.93  18   Male 30.00               1                      2
## 4            276.86  18   Male 34.47               5                      0
## 5            383.45  36   Male 29.70               3                      4
## 6            224.48  28   Male 32.23               3                      6
##   Stress_Level
```

```
## 1           33
## 2           37
## 3           32
## 4           23
## 5           14
## 6           29
```

Remove outliers?

Export the now cleaned data:

```
save(sleep_df, file = "sleep_df.RData")
```