check_in_1

Group 26

Academic Grasshoppers

Yinjie Wu

Yuan Tian

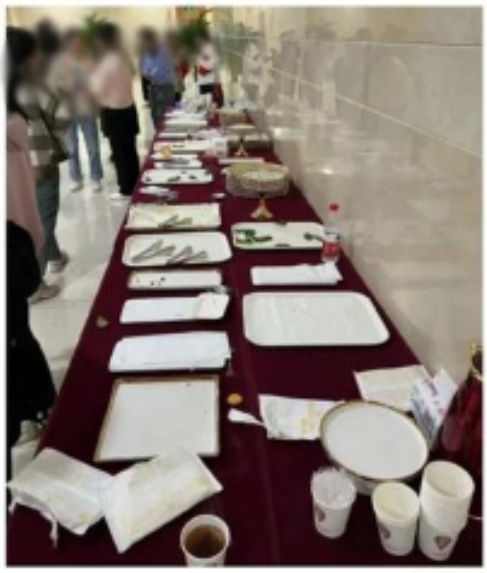
Xinran Yu

No member of this group is using these data or same/similar questions in any other course or course project, at HSPH.

Academic Grasshoppers

Only interested in the snacks at colloquiums, leaving no food behind!

Before



After

Q1 General area/domain/subject area

Sleep disorders have become a common issue for many people today, severely impacting the mental functioning and quality of life. This project explores the potential biological and behavioral factors associated with sleep disorders.

Q2 Dataset and source (describe clearly in several sentences)

Sample size: There are 15560 individuals in the dataset, 8965 individuals with selected covariates measured, and 5394 individuals with no missing values for all covariates and outcomes. The sample size is not final as we may introduce or remove covariates.

We are also considering using past year's data to see if the association changes throughout the years. For example, are weights for each feature had a similar pattern 10 years ago or not, where the sample size could vary slightly depending on the year.

Number of covariates: We are currently including 11 potential covariates, we could introduce more out of interest, or remove irrelevant covariates. Majority of the covariates follow iid.

Source: We are using NHANES 2017-March 2020 Pre-pandemicLinks to an external site. (The National Health and Nutrition Examination Survey program from CDCLinks to an external site.), which assesses the health and nutritional status of adults and children in the United States from 2017 to March 2020. This project has taken a number of health and lifestyle measurements on a specific group of people. Each set of measurements are included in separate files, but all files contain the subject's ID. The ID of each person in this project is unique and constant across all measurements. Therefore, we can merge the data in the measurements we are interested in based on these IDs. If time permits, a similar dataset from the past year or the most recent dataset may be used for comparison. The 2023 version updated in September 2024 contains much less information.

Q3 Primary questions

The primary question of this project is to investigate the association between sleep disorders/duration and a range of biological and behavioral factors. Some factors we are going to explore include BMI, blood pressure, alcohol use, physical activity, and several biomarkers. We are going to examine whether these variables are associated with the occurrence of sleep disorders or sleep duration. If so, we want to determine which factors are most strongly associated with sleep disorders and/or sleep duration.

Q4 Secondary questions

How is the concentration of ions associated with sleep disorders or sleep duration, such as sodium, calcium, and potassium? How about other compounds or measurements such as blood pressure? Do coefficients of covariates in our models hold compared to the dataset from 10 years ago or after the pandemic? Can we predict whether a person has a sleep disorder or his/her sleep duration based on his/her measurements? Other questions we may ask are: Is sleep time associated with those factors? Is the difference between workday and weekend sleep time associated with any of those factors? Are specific sleeping conditions associated with any of those factors? What about the association between sleep disorder and other diseases?

Q5 Outcome(s)/endpoint(s)

The primary outcome includes self-reported ever told doctors had trouble sleeping (questionnaire), which is a binary outcome with 1(Yes) and 2(No). We will also examine closely related measures such as sleep duration and sleep time and how they are associated with the predictor variables. These measures are continuous as reported by the study participants.

Depending on the time, we may look into categorical variables for specific sleep disorder conditions, such as snore or snort.

Q6 Draft Statistical Analysis Plan

Our project is divided into the following steps for data cleaning and subsequent analysis and interpretation.

Data Cleaning and Preprocessing

NHNAES data are classified into several categories based on the type of data collected from study participants, including demographics data, dietary data, examination data, laboratory data, and questionnaire data. Each category is further divided into multiple sections focusing on different measurements.

We imported the NHANES raw data relevant to sleep disorders and potentially associated risk factors, including sleep disorders and health measurements, standard biochemistry profile, physical activity data, total nutrient intake profile, demographic data, and other laboratory data. NHANES raw data from the selected sections will be combined based on the individual-specific ID to generate a profile containing all the information needed for subsequent analysis. Generate summaries of raw data to detect possible missing values, wrong data types. Then we will remove all rows with NAs and change date types according to the summaries. We are

considering imputing for some miss values and unrealistic errors, through averaging, regressing, or KNN.

Exploratory Data Analysis

Generate summary statistics again to gain an initial understanding of the distribution of outcomes and predictor variables.

Visualization of outcome variables

Create a barplot to visualize the distribution of the categorical outcome variable (sleep disorder) and a histogram to visualize the distribution of the continuous outcome variable (sleep duration).

Visualization of predictor variables

Create histograms to examine the distribution of continuous predictor variables (BMI, blood pressure, concentration of electrolytes in blood, sedentary activity, etc) and barplots to examine the distribution of categorical predictor variables (alcohol use, etc). Identify possible skewness.

Create boxplots to examine the initial patterns in sleep quality measurements for each covariate. Create a matrix of scatterplots and scatter plots with a loess curve for each covariates to check multicollinearity and possible important covariates.

Model Development and Selection

Assess multicollinearity among covariates using VIF

Identify regression models

Logistic regression model: We will model the diagnosis of sleep disorders against identified covariates

Linear regression model: We will model sleep quality metrics against identified covariates ### Examine potential confounding and effect modification

Confounders: we will check the classical and operational definition for several potential confounders, such as age, sex, etc.

Effect modifiers: we will test possible interactions that exist by comparing the statistical results produced with and without interaction terms and the significance of interaction terms.

Model selection

For association tasks:

We will use Elastic Net for variable selection coefficient estimation for our logistic model and linear regression model. This process allows us to determine which variables to be kept and which statistically insignificant variables can be removed from our models. We will experiment on hyperparameter alpha to search for lower AIC values.

We will also search for lower AIC values with stepwise, forward, backward selection for model comparison.

Automatic methods like GAM and automatic spline function will be performed as a reference of AIC/BIC values for comparison.

Model complexity in terms of interpretability will also be determined in a subjective manner.

Model diagnostics

We will conduct analyses (e.g. residual analysis for linear models) to check whether the underlying assumptions of logistic and linear models are satisfied.

We will also perform sensitivity analysis (e.g. Cook's distance, DFFITS, DFBETAS) to closely examine the leverage and outlying-ness of highly influential data points. Comparison of models with and without these influential points enable us to decide whether or not we can remove these data points.

Model Interpretation

Using the statistically significant covariates identified through the methods discussed above, we will explain how these variables are associated with sleep disorders and sleep durations based on our results.

Investigation of Secondary Questions

We will apply the final form of our regression models, using the same covariates, to analyze datasets from previous years (e.g., NHANES data from 10 years ago). We will compare the coefficients of the two models and interpret the differences. Then we may also explore factors that are less significant now but could be more impactful ten years ago.

Based on the full model, we may be able to explain the effect of subset of covariates such as concentrations of ions on sleep duration and sleep disorder. We may consider doing further explanation using other methods.

We will also assess how effective our interpretable model can predict whether a person has sleep disorders based on the set of measurements included as covariates, by comparing to full, complicated regression models, and compared to non regression models. Performance metrics may include accuracy, R^2, AUC, and/or etc.. for predictive tasks.

Discussion of Model Results and Key Findings

We will incorporate scientific knowledge (based on advice from experts in epidemiology and biomedical science) to assess our final fitted models within the context of sleep health.

We will also discuss the limitations to our current models and directions for future works

Q7 Biggest challenges foreseen

Since sleep disorders are a group of mental disorders characterized by a diverse range of symptoms and abnormalities, it might be hard for us to reasonably explain some covariates.

Since there are many covariates, the confounding effects and interaction effects between them will be complex, which will be a challenge.

Each covariate and outcome is from a different file, and individuals may have incomplete tests. Adding covariates reduces sample size unless missing/error values are imputed.

Some variables come from questionnaires, like self-reported trouble sleeping, which often include unmeasurable errors.

Q8 Domain expertise sought (who?)

We might contact Dr. Murray A. Mittleman and seek help on theoretical concepts about epidemiological study design and analysis. We might also consult with Dr. Immaculata De Vivo for assistance on the analysis of biological variables and the possible physiological mechanisms underlying their effects on sleep disorder and duration.

Q9 What software package(s) will you use to complete this project?

Currently, we all plan to use R.

Q10 Initial round of exploratory analyses

```
install.packages(c("tidyverse", "NHANES", "foreign", "haven", "survey", "mice"))
install.packages("Rnhanesdata")
```

Import data

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
[1] "/Users/xinranyu/Documents/GitHub/BST_210_project"
```

```
Joining with `by = join_by(SEQN)`
```

scanning raw data

```
tibble [10,195 x 22] (S3: tbl_df/tbl/data.frame)
 $ SEQN    : num [1:10195] 109266 109267 109268 109271 109273 ...
  ..- attr(*, "label")= chr "Respondent sequence number"
 $ SLQ300  : chr [1:10195] "22:00" "00:00" "22:00" "23:00" ...
  ..- attr(*, "label")= chr "Usual sleep time on weekdays or workdays"
 $ SLQ310  : chr [1:10195] "05:30" "08:00" "06:30" "09:00" ...
  ..- attr(*, "label")= chr "Usual wake time on weekdays or workdays"
 $ SLD012  : num [1:10195] 7.5 8 8.5 10 6.5 9.5 9 7 9.5 8 ...
  ..- attr(*, "label")= chr "Sleep hours - weekdays or workdays"
 $ SLQ320  : chr [1:10195] "23:00" "03:00" "23:00" "23:00" ...
  ..- attr(*, "label")= chr "Usual sleep time on weekends"
 $ SLQ330  : chr [1:10195] "07:00" "11:00" "07:00" "12:00" ...
  ..- attr(*, "label")= chr "Usual wake time on weekends"
 $ SLD013  : num [1:10195] 8 8 8 13 8 9.5 10 8 8 10 ...
  ..- attr(*, "label")= chr "Sleep hours - weekends"
 $ SLQ030  : num [1:10195] 1 0 0 0 0 1 3 3 0 0 ...
```

```
 ..- attr(*, "label")= chr "How often do you snore?"
$ SLQ040  : num [1:10195] 0 0 0 0 0 0 1 0 0 0 ...
 ..- attr(*, "label")= chr "How often do you snort or stop breathing"
$ SLQ050  : num [1:10195] 2 2 2 1 1 2 1 1 2 2 ...
 ..- attr(*, "label")= chr "Ever told doctor had trouble sleeping?"
$ SLQ120  : num [1:10195] 0 2 1 3 2 0 3 3 2 0 ...
 ..- attr(*, "label")= chr "How often feel overly sleepy during day?"
$ BMXBMI  : num [1:10195] 37.8 NA NA 29.7 21.9 30.2 NA 26.6 NA 39.1 ...
 ..- attr(*, "label")= chr "Body Mass Index (kg/m**2)"
$ ALQ121  : num [1:10195] 10 NA NA 0 0 4 NA 0 NA NA ...
 ..- attr(*, "label")= chr "Past 12 mo how often drink alcoholic bev"
$ RIAGENDR: num [1:10195] 2 NA NA 1 1 1 NA 1 NA 2 ...
 ..- attr(*, "label")= chr "Gender"
$ RIDRETH1: num [1:10195] 5 NA NA 3 3 5 NA 3 NA 1 ...
 ..- attr(*, "label")= chr "Race/Hispanic origin"
$ RIDAGEYR: num [1:10195] 29 NA NA 49 36 68 NA 76 NA 44 ...
 ..- attr(*, "label")= chr "Age in years at screening"
$ INDFMPIR: num [1:10195] 5 NA NA NA 0.83 1.2 NA 3.61 NA NA ...
 ..- attr(*, "label")= chr "Ratio of family income to poverty"
$ LBXSNASI: num [1:10195] 140 NA NA 141 139 138 NA 145 NA 141 ...
 ..- attr(*, "label")= chr "Sodium (mmol/L)"
$ LBXSKSI : num [1:10195] 3.6 NA NA 4.3 4.3 3.9 NA 4.5 NA 4.1 ...
 ..- attr(*, "label")= chr "Potassium (mmol/L)"
$ LBDSCASI: num [1:10195] 2.2 NA NA 2.23 2.42 ...
 ..- attr(*, "label")= chr "Total Calcium (mmol/L)"
$ LBXSCLSI: num [1:10195] 99 NA NA 101 100 99 NA 105 NA 103 ...
 ..- attr(*, "label")= chr "Chloride (mmol/L)"
$ PAD680  : num [1:10195] 480 NA NA 60 180 300 NA 900 NA 360 ...
 ..- attr(*, "label")= chr "Minutes sedentary activity"
- attr(*, "label")= chr "Sleep Disorders"
```

```
     SEQN           SLQ300             SLQ310             SLD012
Min.   :109266   Length:10195       Length:10195       Min.   : 2.000
1st Qu.:113208   Class :character   Class :character   1st Qu.: 7.000
Median :117039   Mode  :character   Mode  :character   Median : 7.500
Mean   :117076                                        Mean   : 7.641
3rd Qu.:120974                                        3rd Qu.: 8.500
Max.   :124822                                        Max.   :14.000
                                                      NA's   :90
   SLQ320             SLQ330              SLD013           SLQ030
Length:10195       Length:10195       Min.   : 2.000   Min.   :0.000
Class :character   Class :character   1st Qu.: 7.000   1st Qu.:0.000
```

```
Mode  :character   Mode  :character   Median : 8.000   Median :1.000
                                       Mean   : 8.362   Mean   :1.956
                                       3rd Qu.: 9.500   3rd Qu.:3.000
                                       Max.   :14.000   Max.   :9.000
                                       NA's   :96
    SLQ040           SLQ050          SLQ120          BMXBMI
Min.   :0.0000   Min.   :1.000   Min.   :0.000   Min.   :14.20
1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:24.70
Median :0.0000   Median :2.000   Median :2.000   Median :28.60
Mean   :0.8382   Mean   :1.737   Mean   :1.767   Mean   :29.88
3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:33.60
Max.   :9.0000   Max.   :9.000   Max.   :9.000   Max.   :92.30
                                                 NA's   :1405
    ALQ121          RIAGENDR        RIDRETH1        RIDAGEYR
Min.   : 0.000   Min.   :1.000   Min.   :1.000   Min.   :18.00
1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:33.00
Median : 5.000   Median :2.000   Median :3.000   Median :50.00
Mean   : 4.946   Mean   :1.514   Mean   :3.265   Mean   :49.47
3rd Qu.: 8.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:64.00
Max.   :99.000   Max.   :2.000   Max.   :5.000   Max.   :80.00
NA's   :2692     NA's   :1230    NA's   :1230    NA's   :1230
   INDFMPIR        LBXSNASI        LBXSKSI         LBDSCASI
Min.   :0.000   Min.   :121.0   Min.   :2.600   Min.   :1.600
1st Qu.:1.170   1st Qu.:139.0   1st Qu.:3.900   1st Qu.:2.250
Median :2.160   Median :141.0   Median :4.100   Median :2.325
Mean   :2.559   Mean   :140.6   Mean   :4.089   Mean   :2.318
3rd Qu.:4.140   3rd Qu.:142.0   3rd Qu.:4.300   3rd Qu.:2.375
Max.   :5.000   Max.   :151.0   Max.   :7.100   Max.   :3.075
NA's   :2490    NA's   :1937    NA's   :1945    NA's   :1940
   LBXSCLSI        PAD680
Min.   : 84.0   Min.   :   0.0
1st Qu.:100.0   1st Qu.: 180.0
Median :101.0   Median : 300.0
Mean   :101.4   Mean   : 396.3
3rd Qu.:103.0   3rd Qu.: 480.0
Max.   :117.0   Max.   :9999.0
NA's   :1937    NA's   :1247
```

Data cleaning

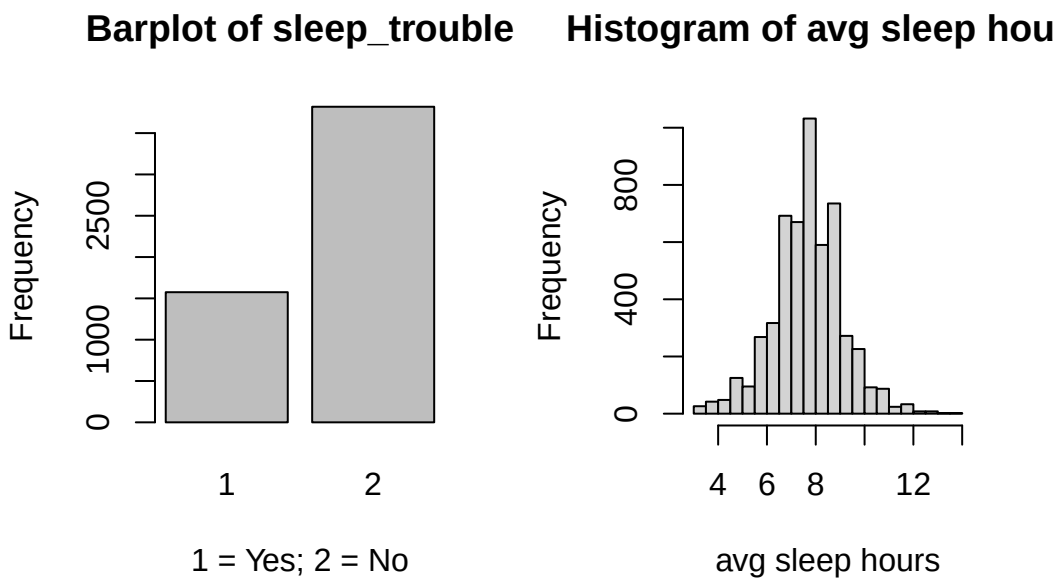
scanning cleaned data

```
tibble [5,394 x 23] (S3: tbl_df/tbl/data.frame)
 $ ID                           : Factor w/ 5394 levels "109266","109273",..: 1 2 3 4 5 6 7
 $ sleep_time_weekdays          :Formal class 'Period' [package "lubridate"] with 6 slots
  .. ..@ .Data : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ year  : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ month : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ day   : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ hour  : num [1:5394] 22 8 21 23 21 22 23 22 9 23 ...
  .. ..@ minute: num [1:5394] 0 0 30 0 0 0 0 0 0 0 ...
 $ wake_time_weekdays           :Formal class 'Period' [package "lubridate"] with 6 slots
  .. ..@ .Data : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ year  : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ month : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ day   : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ hour  : num [1:5394] 5 14 7 6 1 5 6 6 15 6 ...
  .. ..@ minute: num [1:5394] 30 35 0 0 30 30 30 30 0 0 ...
 $ sleep_hours_weekdays         : num [1:5394] 7.5 6.5 9.5 7 4.5 7.5 7.5 8.5 6 7 ...
  ..- attr(*, "label")= chr "Sleep hours - weekdays or workdays"
 $ sleep_time_weekends          :Formal class 'Period' [package "lubridate"] with 6 slots
  .. ..@ .Data : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ year  : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ month : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ day   : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ hour  : num [1:5394] 23 21 21 23 21 23 23 23 1 23 ...
  .. ..@ minute: num [1:5394] 0 0 30 0 0 0 0 0 0 0 ...
 $ wake_time_weekends           :Formal class 'Period' [package "lubridate"] with 6 slots
  .. ..@ .Data : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ year  : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ month : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ day   : num [1:5394] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ hour  : num [1:5394] 7 5 7 7 1 5 10 6 10 6 ...
  .. ..@ minute: num [1:5394] 0 0 0 0 30 30 0 30 0 0 ...
 $ sleep_hours_weekends         : num [1:5394] 8 8 9.5 8 4.5 6.5 11 7.5 9 7 ...
  ..- attr(*, "label")= chr "Sleep hours - weekends"
 $ frq_snore                    : Factor w/ 4 levels "0","1","2","3": 2 1 2 4 1 3 1 2 2 4 .
 $ frq_snort_or_stop_breathing  : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 3 1 1 1 1 .
 $ sleep_trouble                : Factor w/ 2 levels "1","2": 2 1 2 1 1 2 2 2 2 1 ...
 $ overly_sleepy                : Factor w/ 5 levels "0","1","2","3",..: 1 3 1 4 3 4 1 2 2 4
```

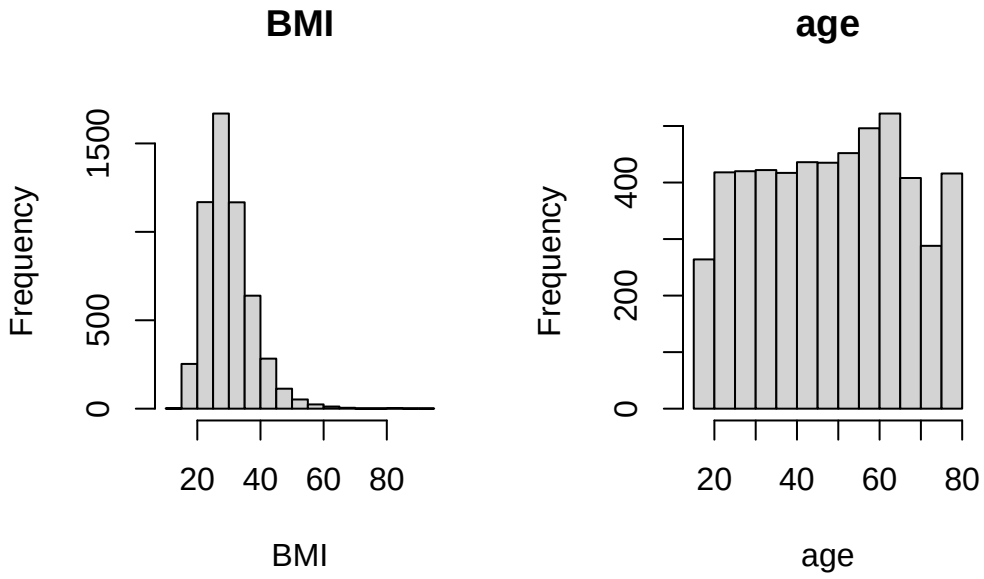
```
$ BMI                          : num [1:5394] 37.8 21.9 30.2 26.6 30.5 30.1 24.6 23.9 30.5
 ..- attr(*, "label")= chr "Body Mass Index (kg/m**2)"
$ alcohol                      : Factor w/ 11 levels "0","1","2","3",..: 11 1 5 1 5 1 10 1
$ gender                       : Factor w/ 2 levels "1","2": 2 1 1 1 1 1 1 2 1 2 ...
$ Race_Hispanic_origin         : Factor w/ 5 levels "1","2","3","4",..: 5 3 5 3 2 3 5 3 4 4
$ age                          : num [1:5394] 29 36 68 76 58 44 47 48 22 19 ...
 ..- attr(*, "label")= chr "Age in years at screening"
$ Family_income_to_poverty_ratio: num [1:5394] 5 0.83 1.2 3.61 1.6 0.02 1.38 5 4.93 0.06 ..
 ..- attr(*, "label")= chr "Ratio of family income to poverty"
$ Na                           : num [1:5394] 140 139 138 145 140 138 142 139 138 140 ...
 ..- attr(*, "label")= chr "Sodium (mmol/L)"
$ K                            : num [1:5394] 3.6 4.3 3.9 4.5 4.8 3.7 3.8 3.9 3.7 4.2 ...
 ..- attr(*, "label")= chr "Potassium (mmol/L)"
$ Ca                           : num [1:5394] 2.2 2.42 2.27 2.27 2.33 ...
 ..- attr(*, "label")= chr "Total Calcium (mmol/L)"
$ Cl                           : num [1:5394] 99 100 99 105 102 98 104 100 99 105 ...
 ..- attr(*, "label")= chr "Chloride (mmol/L)"
$ sedentary_activity           : num [1:5394] 480 180 300 900 600 360 120 180 600 360 ...
 ..- attr(*, "label")= chr "Minutes sedentary activity"
$ sleep_hours_avg              : num [1:5394] 7.64 6.93 9.5 7.29 4.5 ...
 ..- attr(*, "label")= chr "Sleep hours - weekends"
- attr(*, "label")= chr "Sleep Disorders"


      ID       sleep_time_weekdays
109266 :   1   Min.   :0S
109273 :   1   1st Qu.:3H 0M 0S
109274 :   1   Median :22H 0M 0S
109282 :   1   Mean   :16H 7M 51.134593993309S
109292 :   1   3rd Qu.:23H 0M 0S
109293 :   1   Max.   :23H 45M 0S
(Other):5388
wake_time_weekdays                sleep_hours_weekdays
Min.   :0S                        Min.   : 3.000
1st Qu.:5H 30M 0S                 1st Qu.: 6.500
Median :6H 30M 0S                 Median : 7.500
Mean   :6H 43M 20.2780867630718S  Mean   : 7.569
3rd Qu.:7H 30M 0S                 3rd Qu.: 8.500
Max.   :23H 30M 0S                Max.   :14.000

sleep_time_weekends                wake_time_weekends
Min.   :0S                         Min.   :0S
1st Qu.:1H 0M 0S                   1st Qu.:6H 30M 0S
```

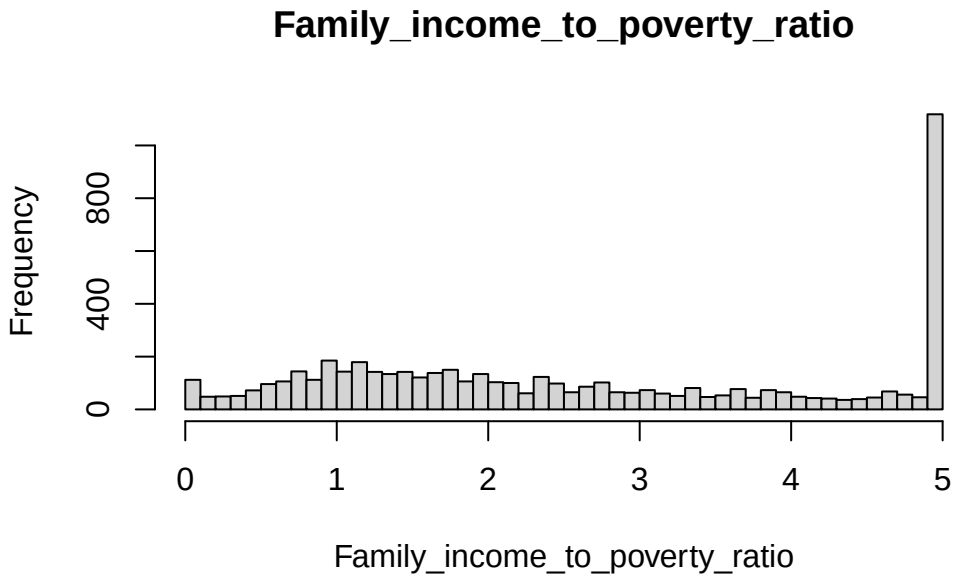
```
Median :21H 0M 0S                    Median :8H 0M 0S
Mean   :12H 55M 52.2803114571652S    Mean   :7H 49M 18.2313681868654S
3rd Qu.:23H 0M 0S                    3rd Qu.:9H 0M 0S
Max.   :23H 50M 0S                   Max.   :23H 0M 0S

sleep_hours_weekends frq_snore frq_snort_or_stop_breathing sleep_trouble
Min.   : 3.000       0:1393    0:3981                      1:1575
1st Qu.: 7.000       1:1433    1: 748                      2:3819
Median : 8.000       2:1058    2: 346
Mean   : 8.265       3:1510    3: 319
3rd Qu.: 9.000
Max.   :14.000

overly_sleepy      BMI           alcohol     gender   Race_Hispanic_origin
0: 803        Min.   :14.60   0      :1098   1:2737   1: 653
1:1291        1st Qu.:24.80   6      : 753   2:2657   2: 526
2:1882        Median :28.80   9      : 575            3:2107
3:1000        Mean   :30.02   10     : 568            4:1320
4: 418        3rd Qu.:33.80   4      : 462            5: 788
              Max.   :92.30   7      : 439
                              (Other):1499
     age        Family_income_to_poverty_ratio       Na              K
Min.   :18.00   Min.   :0.00                  Min.   :121.0   Min.   :2.60
1st Qu.:33.00   1st Qu.:1.23                  1st Qu.:139.0   1st Qu.:3.90
Median :49.00   Median :2.37                  Median :141.0   Median :4.10
Mean   :48.78   Mean   :2.68                  Mean   :140.5   Mean   :4.09
3rd Qu.:63.00   3rd Qu.:4.44                  3rd Qu.:142.0   3rd Qu.:4.30
Max.   :80.00   Max.   :5.00                  Max.   :150.0   Max.   :7.10

      Ca              Cl        sedentary_activity sleep_hours_avg
Min.   :1.600   Min.   : 84.0   Min.   :   0.0     Min.   : 3.000
1st Qu.:2.250   1st Qu.:100.0   1st Qu.: 180.0     1st Qu.: 7.000
Median :2.325   Median :101.0   Median : 300.0     Median : 7.857
Mean   :2.320   Mean   :101.3   Mean   : 368.6     Mean   : 7.768
3rd Qu.:2.375   3rd Qu.:103.0   3rd Qu.: 480.0     3rd Qu.: 8.643
Max.   :3.075   Max.   :117.0   Max.   :9999.0     Max.   :13.714
```

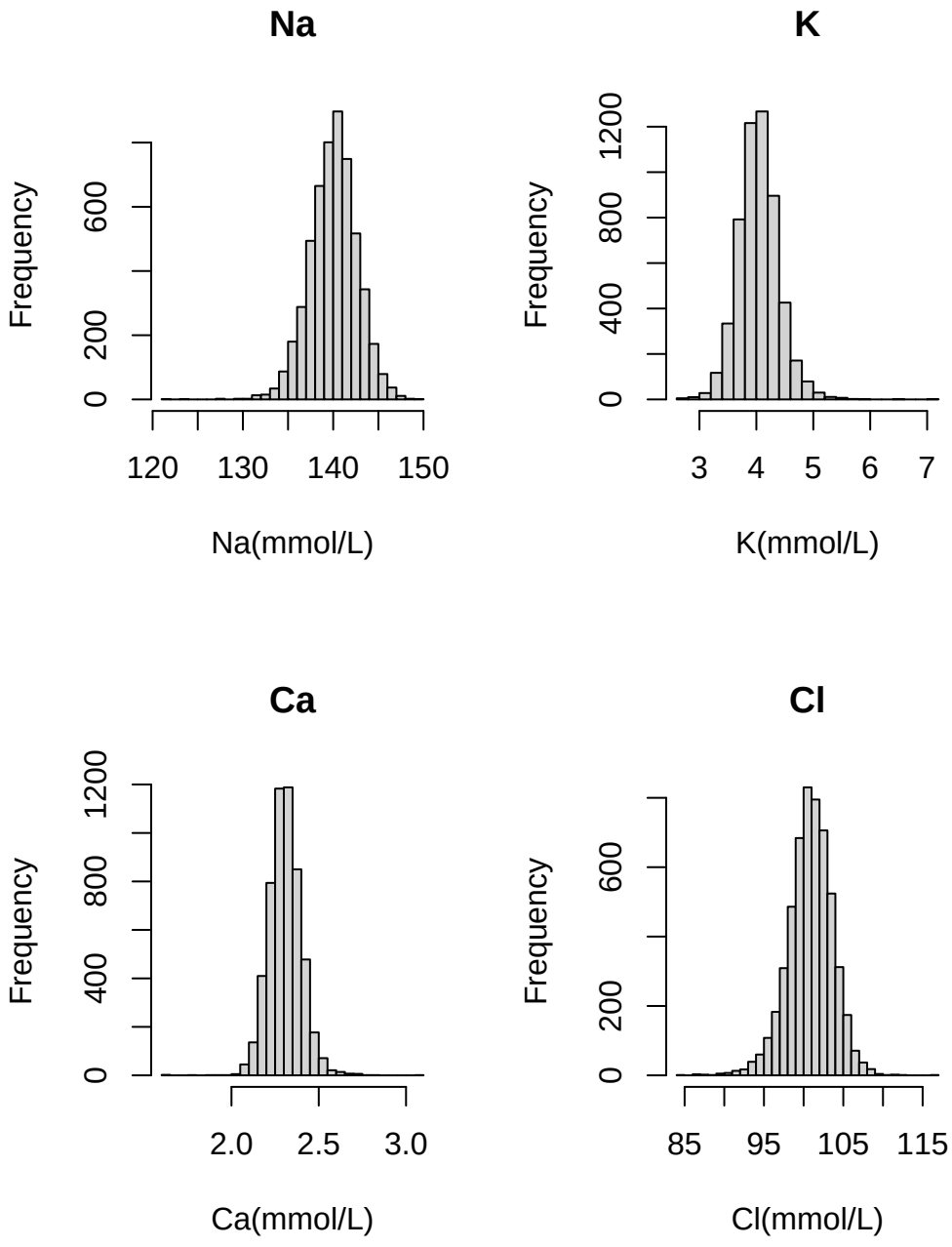
Histogram and barplot of outcome

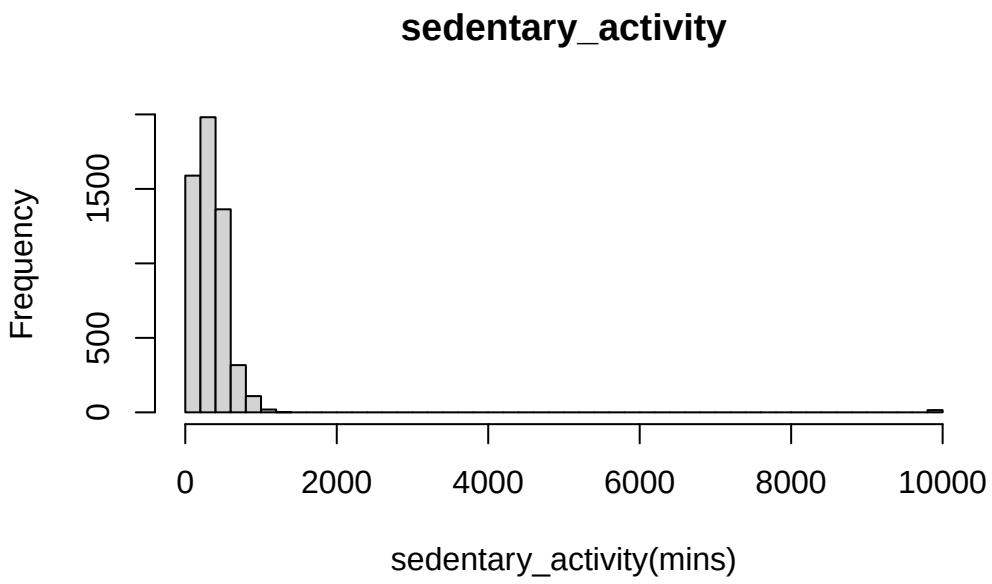


Histogram of continuous variables X

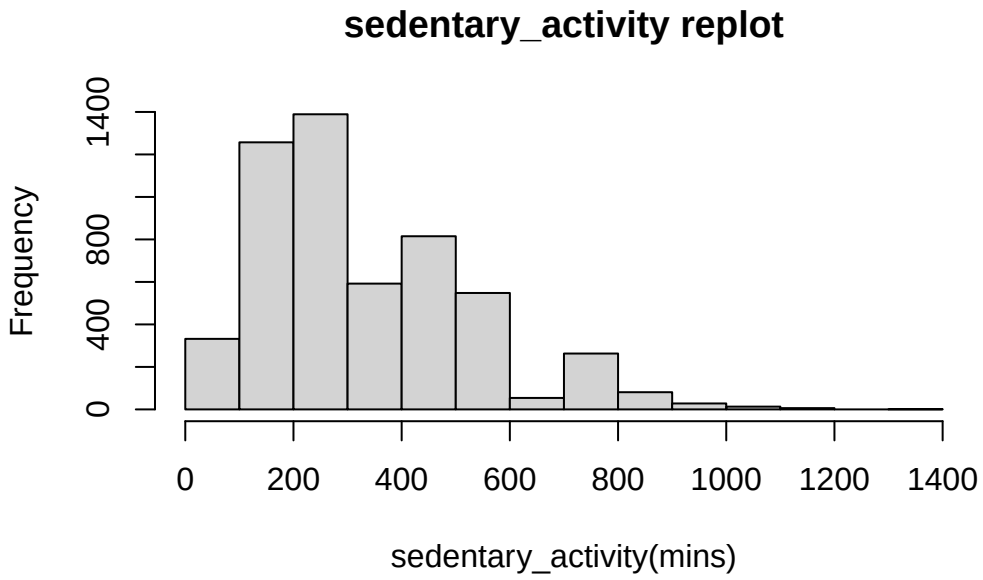




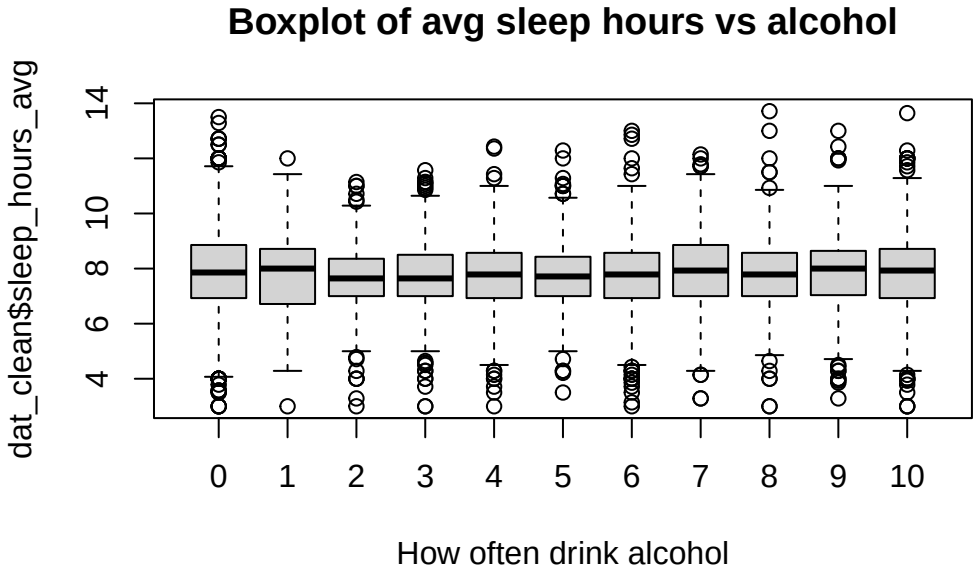

Na
Frequency
0
200
600
120
130
140
150
Na(mmol/L)
K
Frequency
0
400
800
1200
3
4
5
6
7
K(mmol/L)
Ca
Frequency
0
400
800
1200
2.0
2.5
3.0
Ca(mmol/L)
Cl
Frequency
0
200
600
85
95
105
115
Cl(mmol/L)

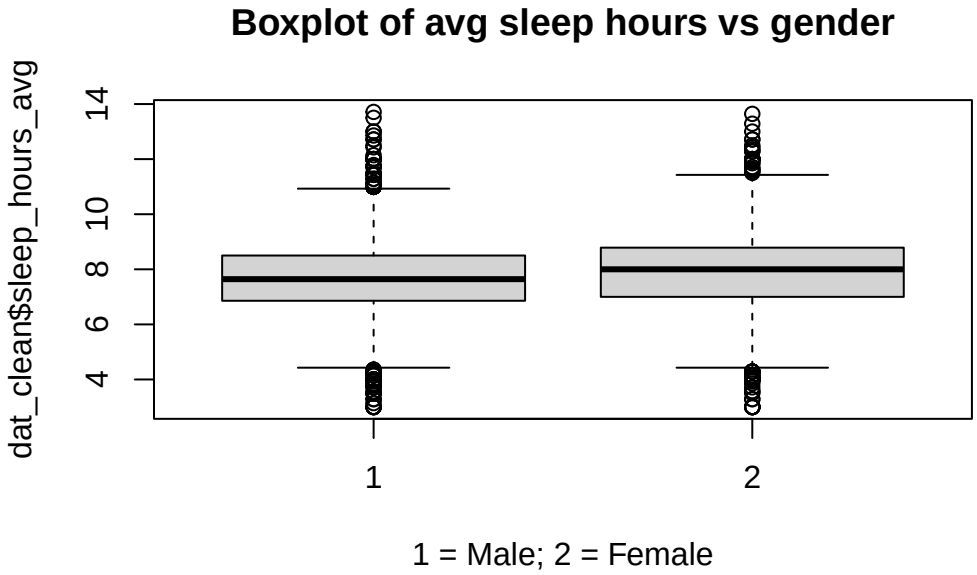


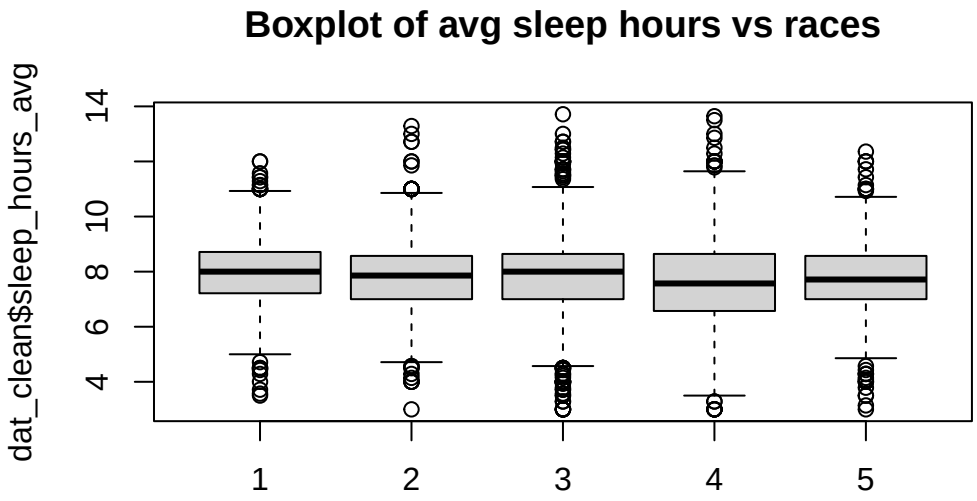
Obviously, 9000 minutes is more than the possible range of a day (1440 minutes). Thus these high leverage points are errors. It's better to remove them.



Boxplot

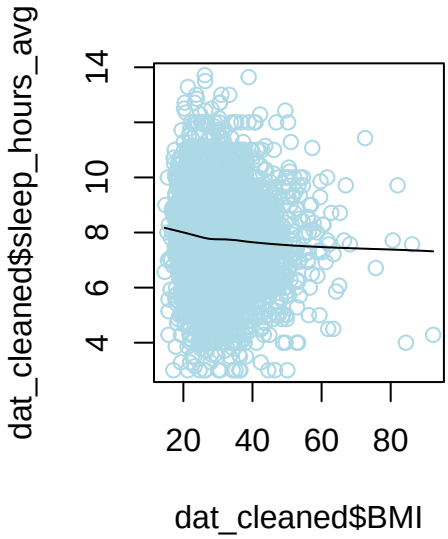

Boxplot of avg sleep hours vs alcohol
dat_clean$sleep_hours_avg
4
6
8
10
14
0
1
2
3
4
5
6
7
8
9
10
How often drink alcohol

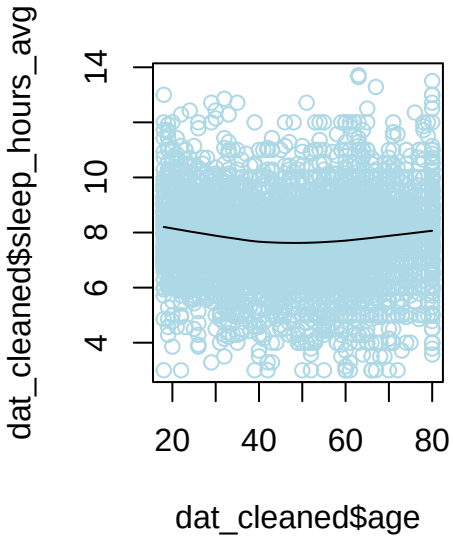

Boxplot of avg sleep hours vs gender
dat_clean$sleep_hours_avg
4
6
8
10
14
1
2
1 = Male; 2 = Female

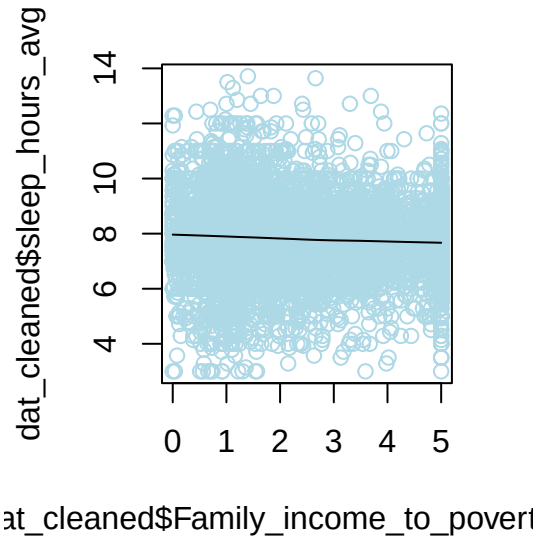


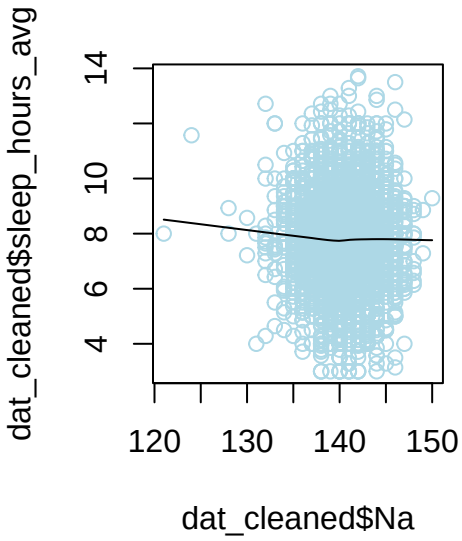
1 = Mexican American; 2 = Other Hispanic; 3 = White; 4 = Black; 5 = Other

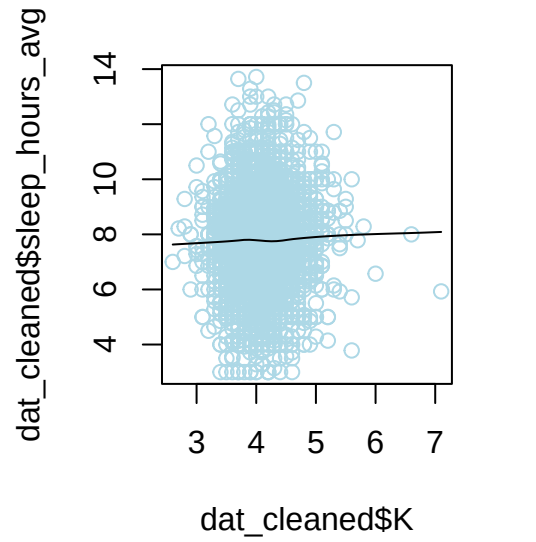
Correlations

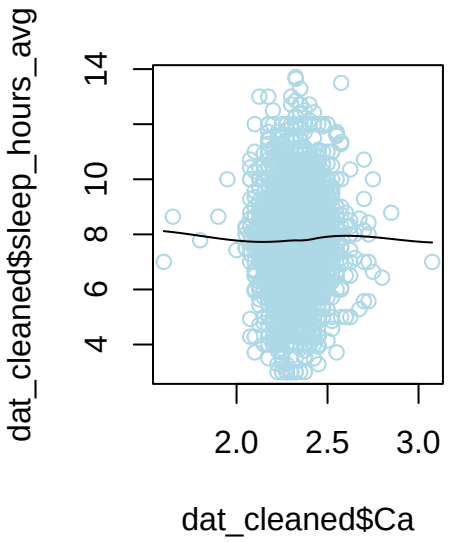


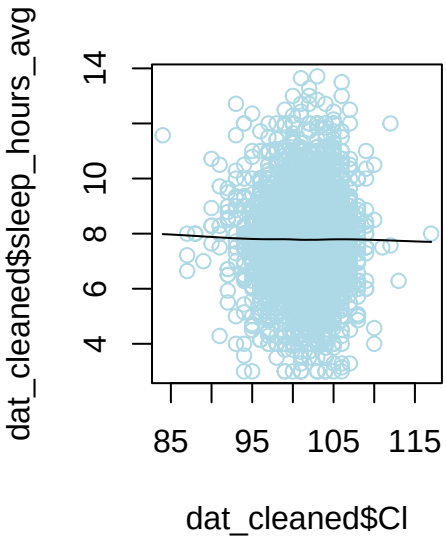


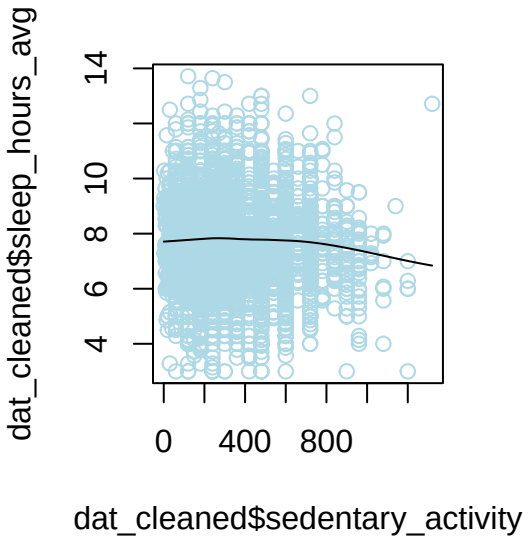

dat_cleaned$sleep_hours_avg
4
6
8
10
14
0
1
2
3
4
5
at_cleaned$Family_income_to_poverl

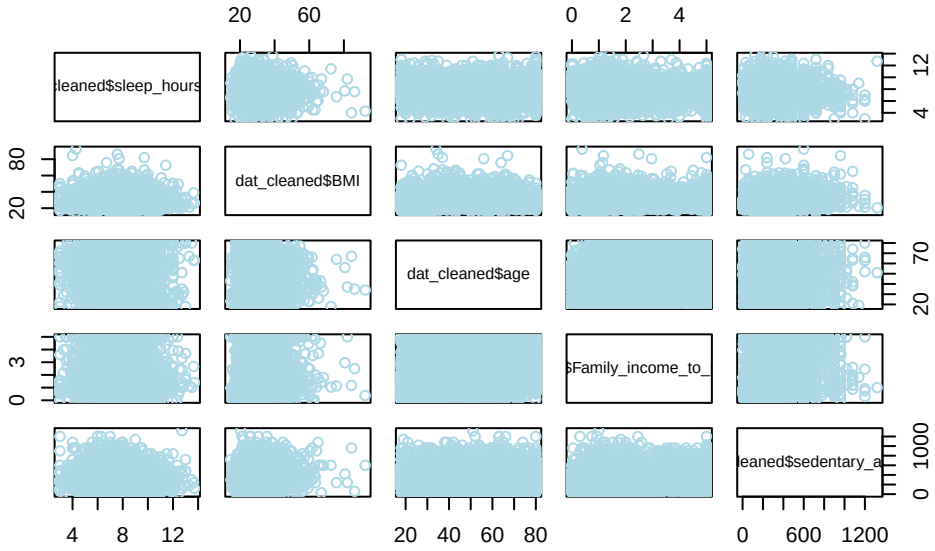

dat_cleaned$sleep_hours_avg
4
6
8
10
14
120
130
140
150
dat_cleaned$Na

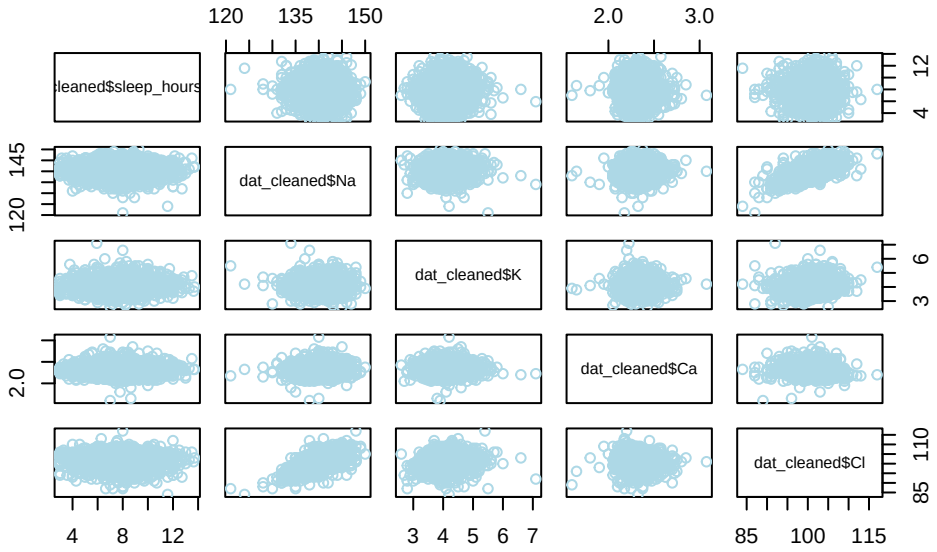

dat_cleaned$sleep_hours_avg
4
6
8
10
14
3
4
5
6
7
dat_cleaned$K


dat_cleaned$sleep_hours_avg
4
6
8
10
14
2.0
2.5
3.0
dat_cleaned$Ca


dat_cleaned$sleep_hours_avg
4
6
8
10
14
85
95
105
115
dat_cleaned$Cl


dat_cleaned$sleep_hours_avg
4
6
8
10
14
0
400
800
dat_cleaned$sedentary_activity


20
60
0
2
4
leaned$sleep_hours
4
12
dat_cleaned$BMI
20
80
dat_cleaned$age
20
70
$Family_income_to_
0
3
leaned$sedentary_a
0
1000
4
8
12
20
40
60
80
0
600
1200



Q11

Group 26

Academic Grasshoppers

Yinjie Wu

Yuan Tian

Xinran Yu

Project Attestation: No member of this group is using these data or same/similar questions in any other course or course project, at HSPH. By listing your name as a group member on your project, and submitting this assignment, you are attesting to this statement above. Groups must include this attestation here under Question 11 in order to receive credit for HW4!

Code Appendix

```r
library(tidyverse)
library(haven)
library(NHANES)
library(dplyr)
library(purrr)
library(lubridate)

getwd()
setwd("./data/2020")

ALQ <- read_xpt("P_ALQ.XPT")
BIOPRO <- read_xpt("P_BIOPRO.XPT")
BMX <- read_xpt("P_BMX.XPT")
BPXO <- read_xpt("P_BPXO.XPT")
DEMO <- read_xpt("P_DEMO.XPT")
# FOLATE <- read_xpt("P_FOLATE.XPT")
PAQ <- read_xpt("P_PAQ.XPT")
SLQ <- read_xpt("P_SLQ.XPT")
TCHOL <- read_xpt("P_TCHOL.XPT")

datasets <- list(ALQ, BIOPRO, BMX, BPXO, DEMO,  PAQ, SLQ, TCHOL)
test <- reduce(datasets, inner_join, by = "SEQN")

temp <- select(test, SEQN , BMXBMI, ALQ121, RIAGENDR, RIDRETH1, RIDAGEYR, INDFMPIR, LBXSNASI
dat_raw <- left_join(SLQ, temp)

# DMDEDUC2(education level 20+)
# DMDMARTZ - Marital status
str(dat_raw)
summary(dat_raw)
# rename cols
new_colnames <-  c(
  "ID",      # SEQN;Respondent_sequence_number
  "sleep_time_weekdays",       # SLQ300;Usual_sleep_time_weekdays
  "wake_time_weekdays",        # SLQ310;Usual_wake_time_weekdays
  "sleep_hours_weekdays",            # SLD012;Sleep_hours_weekdays
  "sleep_time_weekends",       # SLQ320;Usual_sleep_time_weekends
  "wake_time_weekends",        # SLQ330;Usual_wake_time_weekends
  "sleep_hours_weekends",            # SLD013;Sleep_hours_weekends
  "frq_snore",                 # SLQ030;How_often_snore
  "frq_snort_or_stop_breathing",# SLQ040;How_often_snort_or_stop_breathing
  "sleep_trouble",  # SLQ050;Ever_told_doctor_sleep_trouble
```

```
  "overly_sleepy",          # SLQ120;Feel_overly_sleepy_day
  "BMI",                  # BMXBMI;Body_Mass_Index
  "alcohol",          # ALQ121;How_often_drink_alcohol
  "gender",                           # RIAGENDR;Gender
  "Race_Hispanic_origin",             # RIDRETH1;Race_Hispanic_origin
  "age",                 # RIDAGEYR;Age_at_screening
  "Family_income_to_poverty_ratio",  # INDFMPIR;Family_income_to_poverty_ratio
  "Na",                    # LBXSNASI;Sodium_mmol_L
  "K",                 # LBXSLSI;Potassium_mmol_L
  "Ca",             # LBDSCASI;Total_Calcium_mmol_L
  "Cl",                  # LBXSCLSI;Chloride_mmol_L
  "sedentary_activity"       # PAD680;Minutes_sedentary_activity
)
colnames(dat_raw) <- new_colnames
# remove all lines with NA
dat_raw[dat_raw == ""] <- NA
dat_clean <- dat_raw |>
  drop_na() |>
   filter(alcohol != 77 & frq_snore != 7 & frq_snore != 9 & frq_snort_or_stop_breathing != 7
  mutate(across(c(sleep_time_weekdays,wake_time_weekdays,sleep_time_weekends,wake_time_weeke
  mutate(across(c(ID,frq_snore,frq_snort_or_stop_breathing,sleep_trouble,overly_sleepy,alcoh
  mutate(across(c(,),as.numeric)) |>
  mutate(sleep_hours_avg = 2/7*sleep_hours_weekends+5/7*sleep_hours_weekdays)

str(dat_clean)
summary(dat_clean)
par(mfrow = c(1,2))
barplot(table(dat_clean$sleep_trouble), xlab = "1 = Yes; 2 = No", main = "Barplot of sleep_t
hist(dat_clean$sleep_hours_avg, main = "Histogram of avg sleep hours",breaks = 30, xlab = "a
par(mfrow = c(1,2))
hist(dat_clean$BMI,main = "BMI",xlab = "BMI")
hist(dat_clean$age,main = "age", xlab = "age")
hist(dat_clean$Family_income_to_poverty_ratio,main = "Family_income_to_poverty_ratio", xlab =
par(mfrow = c(1,2))
hist(dat_clean$Na,main = "Na",xlab = "Na(mmol/L)",breaks = 30)
hist(dat_clean$K,main = "K", xlab = "K(mmol/L)",breaks = 30)
par(mfrow = c(1,2))
hist(dat_clean$Ca,main = "Ca",xlab = "Ca(mmol/L)",breaks = 30)
hist(dat_clean$Cl,main = "Cl", xlab = "Cl(mmol/L)",,breaks = 30)
hist(dat_clean$sedentary_activity,main = "sedentary_activity",xlab = "sedentary_activity(min
dat_cleaned <- dat_clean |>
  filter(sedentary_activity < 9000)
```

```r
hist(dat_cleaned$sedentary_activity,main = "sedentary_activity replot",xlab = "sedentary_acti
boxplot(dat_clean$sleep_hours_avg~dat_clean$alcohol, main = "Boxplot of avg sleep hours vs al
boxplot(dat_clean$sleep_hours_avg~dat_clean$gender,main = "Boxplot of avg sleep hours vs gend
boxplot(dat_clean$sleep_hours_avg~dat_clean$Race_Hispanic_origin,main = "Boxplot of avg sleep
par(mfrow = c(1,2))
scatter.smooth(dat_cleaned$BMI,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$age,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$Family_income_to_poverty_ratio,dat_cleaned$sleep_hours_avg,col = "
scatter.smooth(dat_cleaned$Na,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$K,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$Ca,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$Cl,dat_cleaned$sleep_hours_avg,col = "light blue")
scatter.smooth(dat_cleaned$sedentary_activity,dat_cleaned$sleep_hours_avg,col = "light blue")
pairs(dat_cleaned$sleep_hours_avg ~ dat_cleaned$BMI + dat_cleaned$age + dat_cleaned$Family_in
pairs(dat_cleaned$sleep_hours_avg ~ dat_cleaned$Na + dat_cleaned$K + dat_cleaned$Ca + dat_cle
```