part-3

Madhav Kanna Thenappan

2023-12-11

Installing libraries

```
# install.packages("NHANES") # install
# install.packages("dplyr")
# install.packages("car")
# install.packages("MASS")
```

Getting data

```
library(NHANES) # Load package
```

```
## Warning: package 'NHANES' was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
write.csv(NHANES, "..\\dataset.csv")
raw_data <- NHANES # Load data
required_columns <- raw_data %>% select(Gender, DirectChol, SleepHrsNight, PhysActiveDays, Age, Alcohol
cleaned_data <- na.omit(required_columns)
```

Creating Preliminary Linear Regression Model

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.3.2
```

```
##
## Attaching package: 'MASS'
```

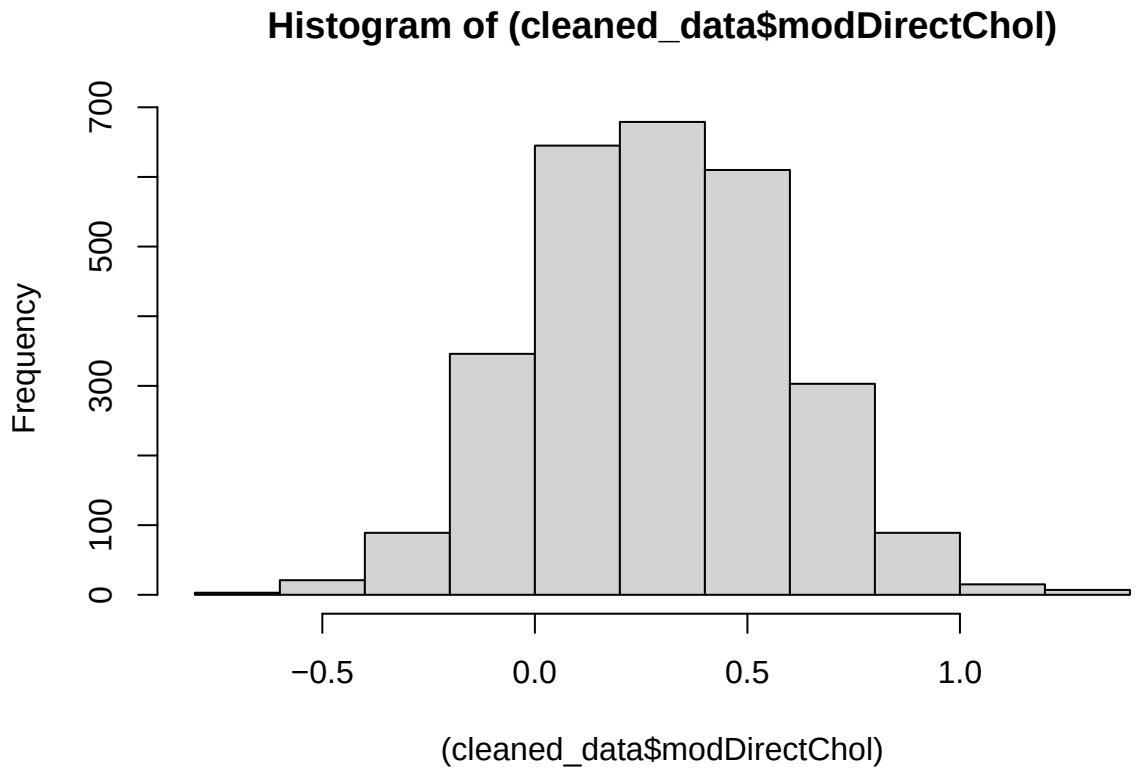
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
cleaned_data$modAlcoholYear <- log(cleaned_data$AlcoholYear + 1)
cleaned_data$modSleepHrsNight <- cleaned_data$SleepHrsNight^(2)
cleaned_data$modDirectChol <- log(cleaned_data$DirectChol)
cleaned_data$modBMI <- log(cleaned_data$BMI)
cleaned_data$modBPSysAve <- log(cleaned_data$BPSysAve)
fit <- lm(modSleepHrsNight ~ Gender + modDirectChol  + PhysActiveDays + Age  + modBMI + modBPSysAve + BI
summary(fit)
```

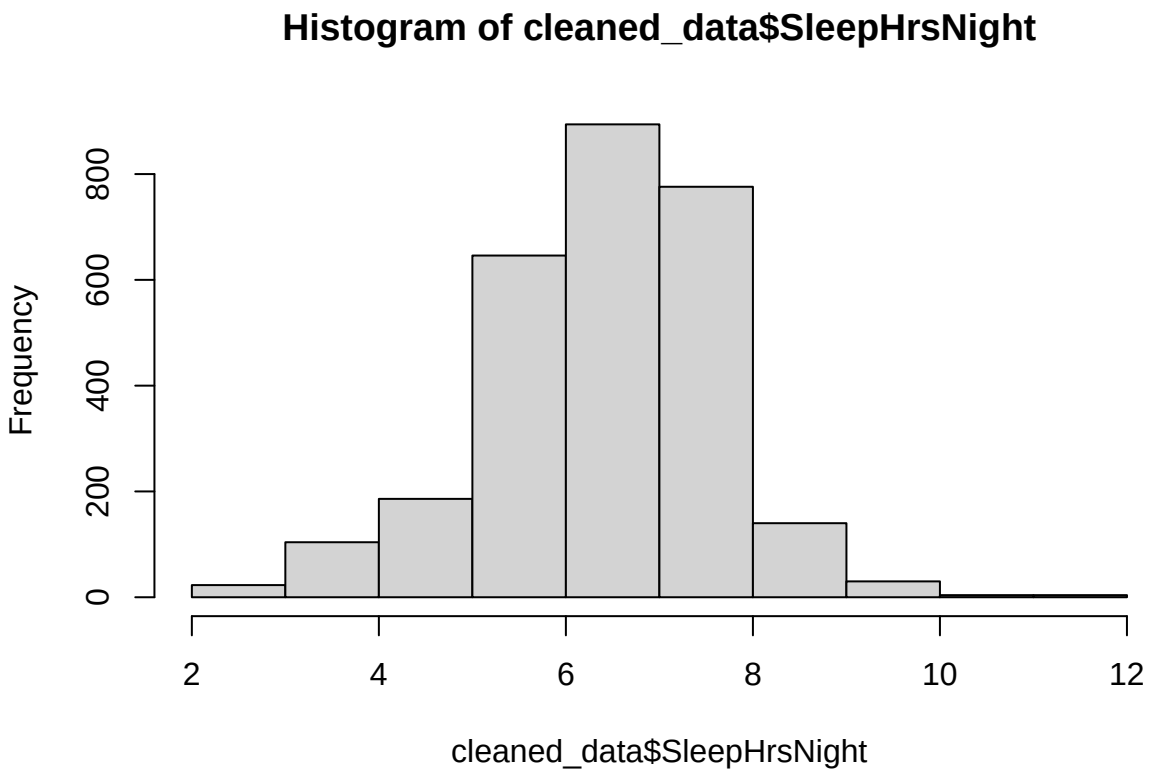
```
##
## Call:
## lm(formula = modSleepHrsNight ~ Gender + modDirectChol + PhysActiveDays +
##     Age + modBMI + modBPSysAve + BPDiaAve + modAlcoholYear, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.621 -12.083  -0.607  13.136  94.305
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    92.31715   13.57672   6.800 1.28e-11 ***
## Gendermale     -3.01004    0.71974  -4.182 2.98e-05 ***
## modDirectChol  -1.93583    1.35311  -1.431 0.152643
## PhysActiveDays  0.14383    0.17721   0.812 0.417081
```

```
## Age             0.07468    0.02201   3.393 0.000701 ***
## modBMI         -4.85520    1.76553  -2.750 0.005998 **
## modBPSysAve    -5.84422    2.93700  -1.990 0.046704 *
## BPDiaAve       -0.03695    0.02919  -1.266 0.205642
## modAlcoholYear  0.61649    0.17850   3.454 0.000561 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.95 on 2798 degrees of freedom
## Multiple R-squared:  0.02122,    Adjusted R-squared:  0.01842
## F-statistic: 7.583 on 8 and 2798 DF,  p-value: 4.503e-10
```

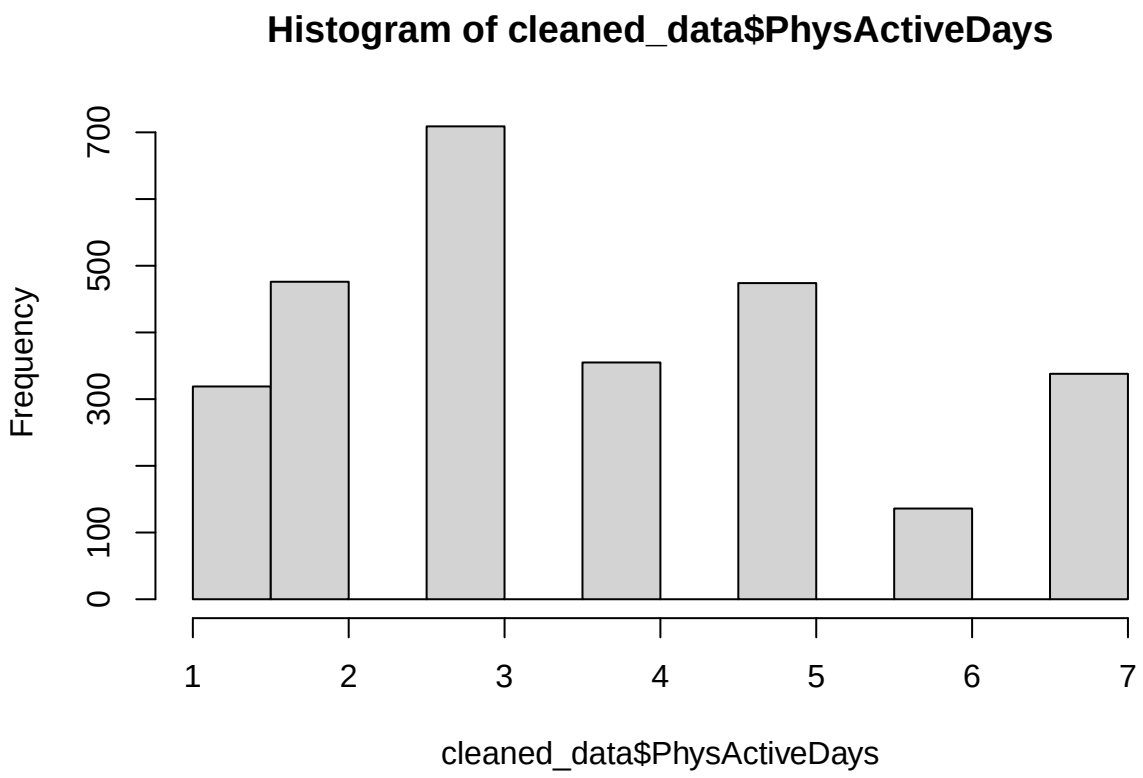
```
hist((cleaned_data$modDirectChol))
```



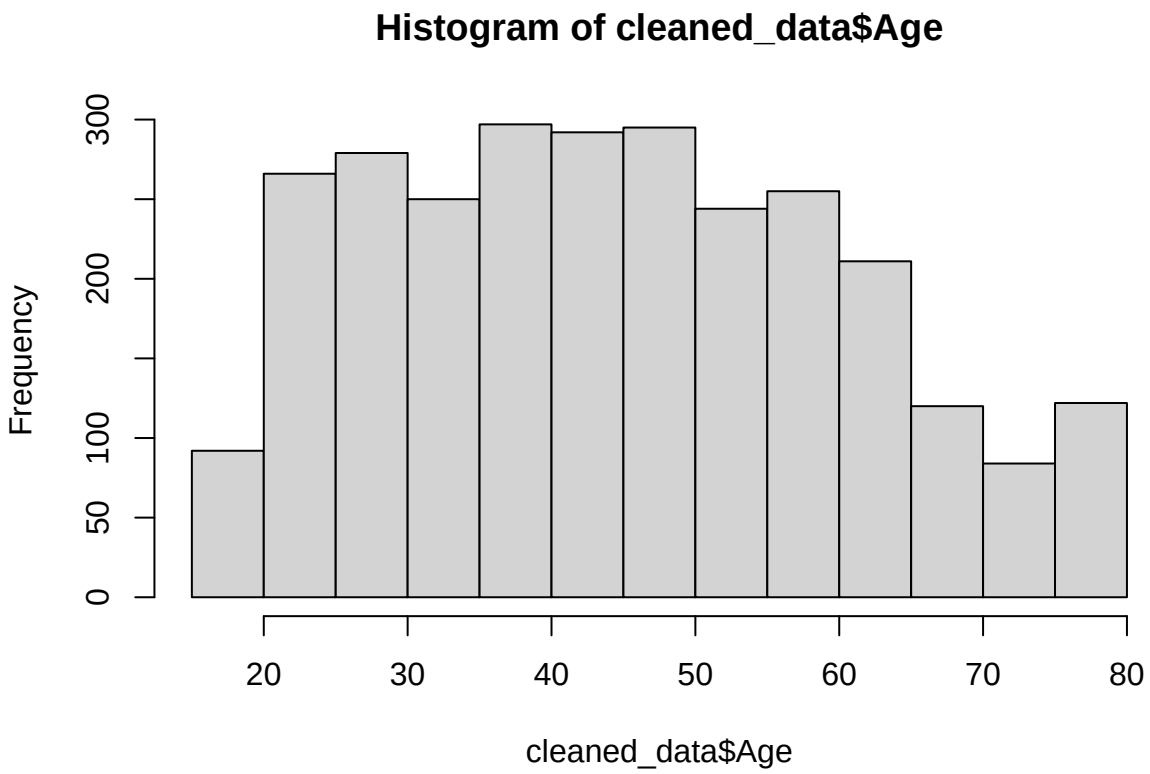

```
hist(cleaned_data$SleepHrsNight)
```



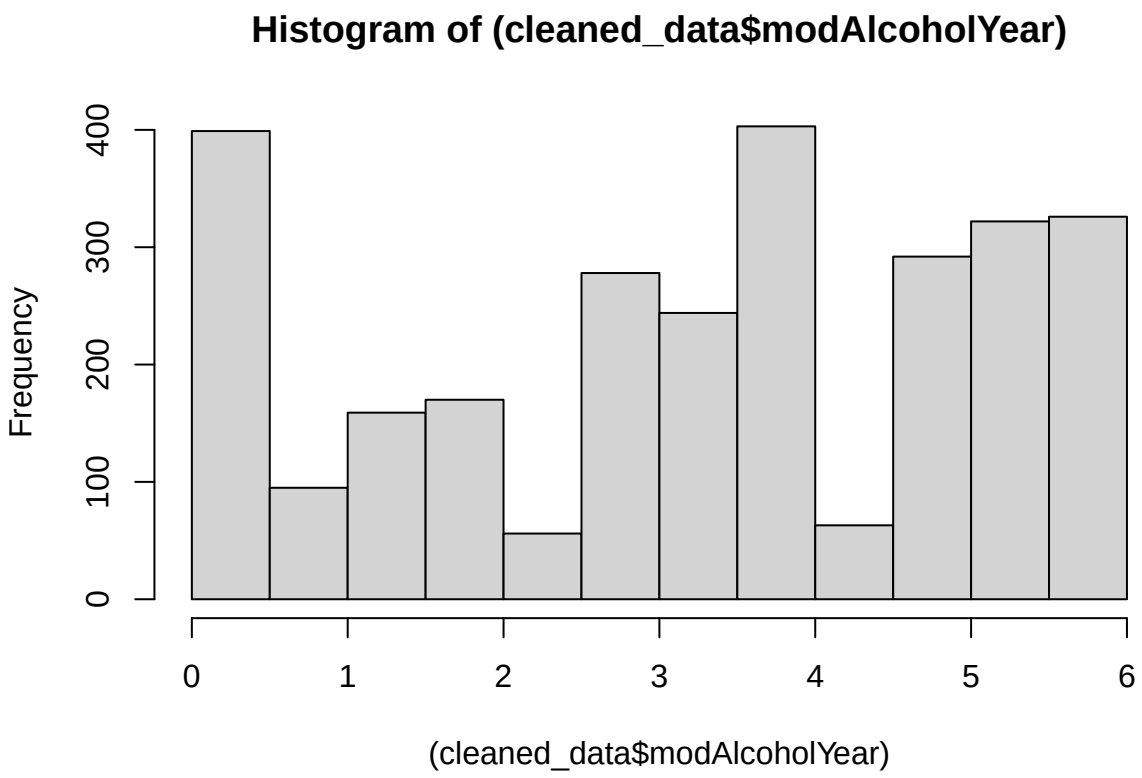

```
hist(cleaned_data$PhysActiveDays)
```



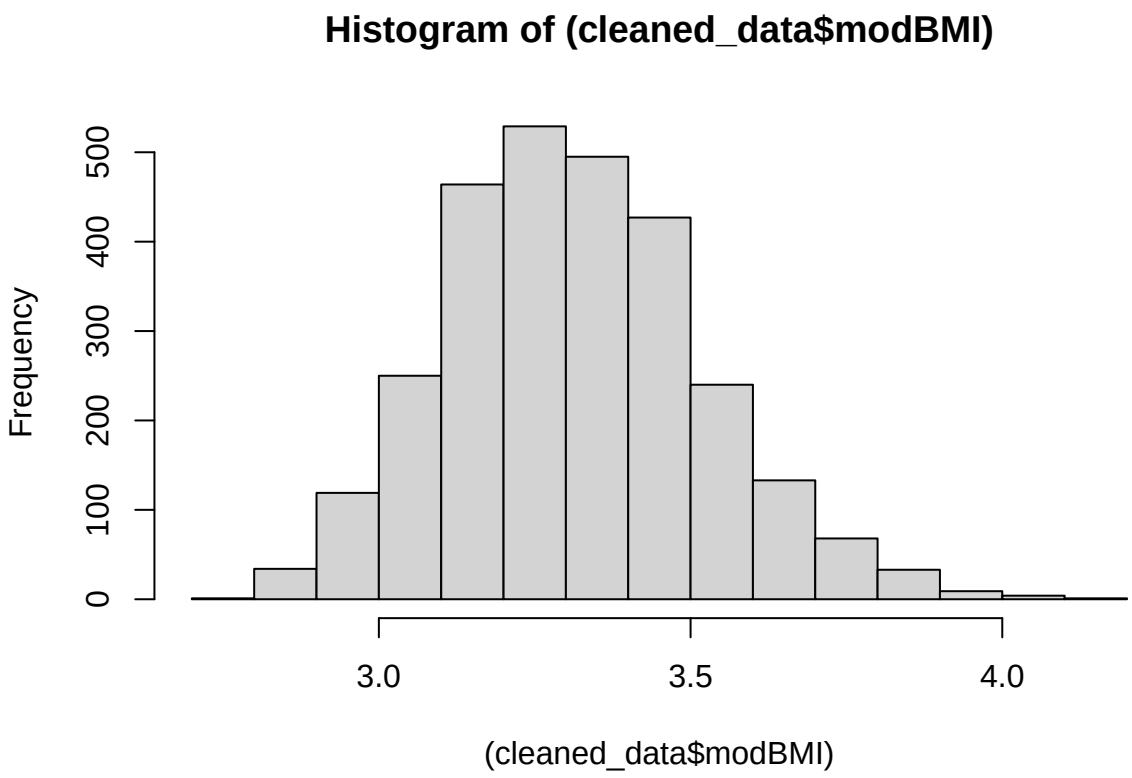

```r
hist(cleaned_data$Age)
```



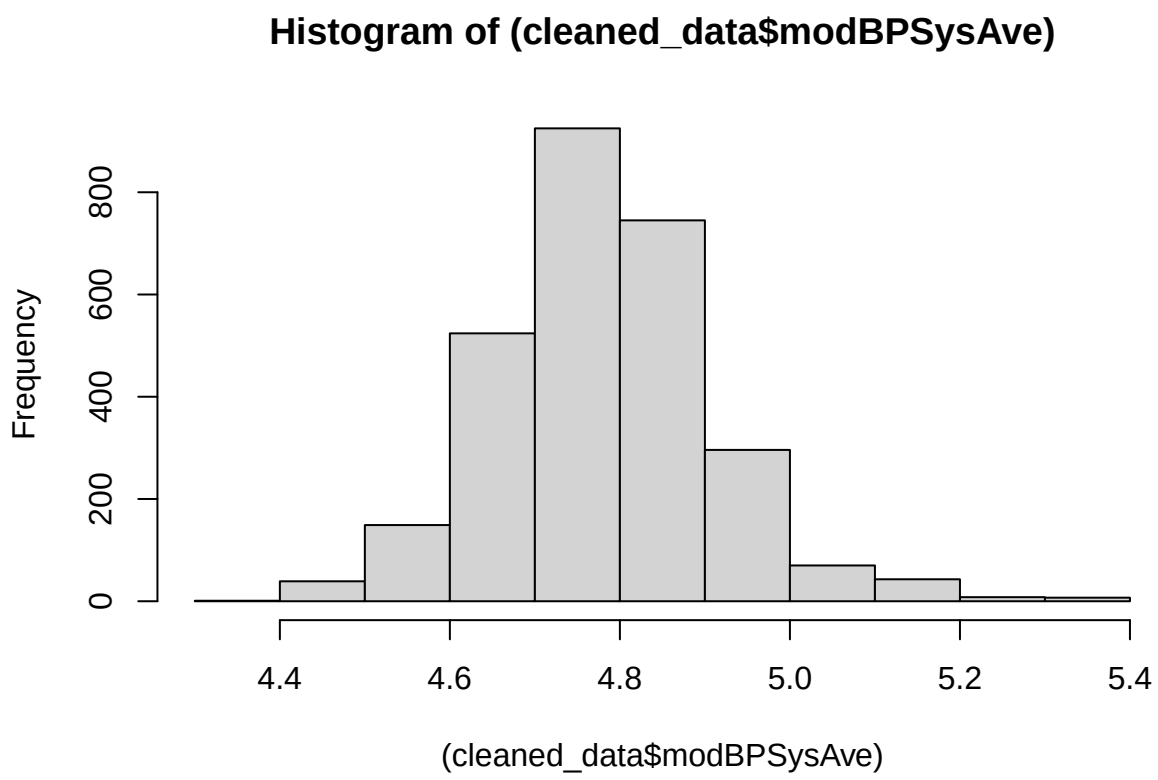

```r
hist((cleaned_data$modAlcoholYear))
```



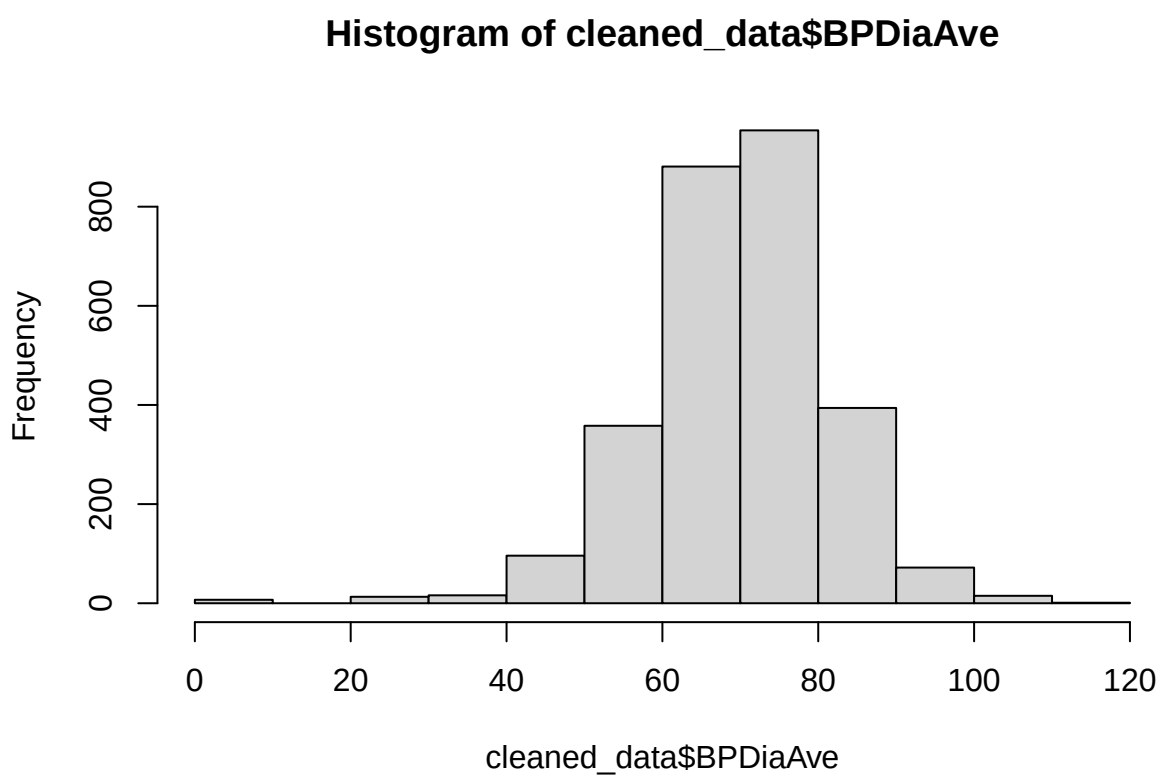

```
hist((cleaned_data$modBMI))
```



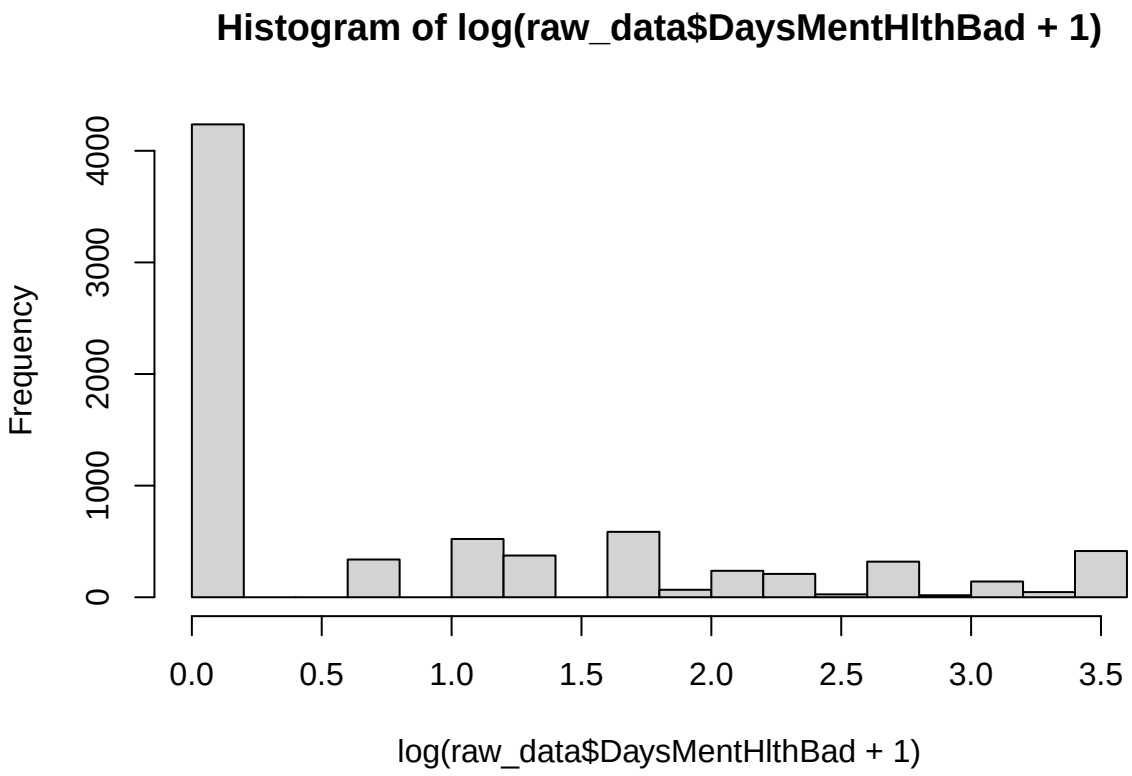

```
hist((cleaned_data$modBPSysAve))
```




```
hist(cleaned_data$BPDiaAve)
```

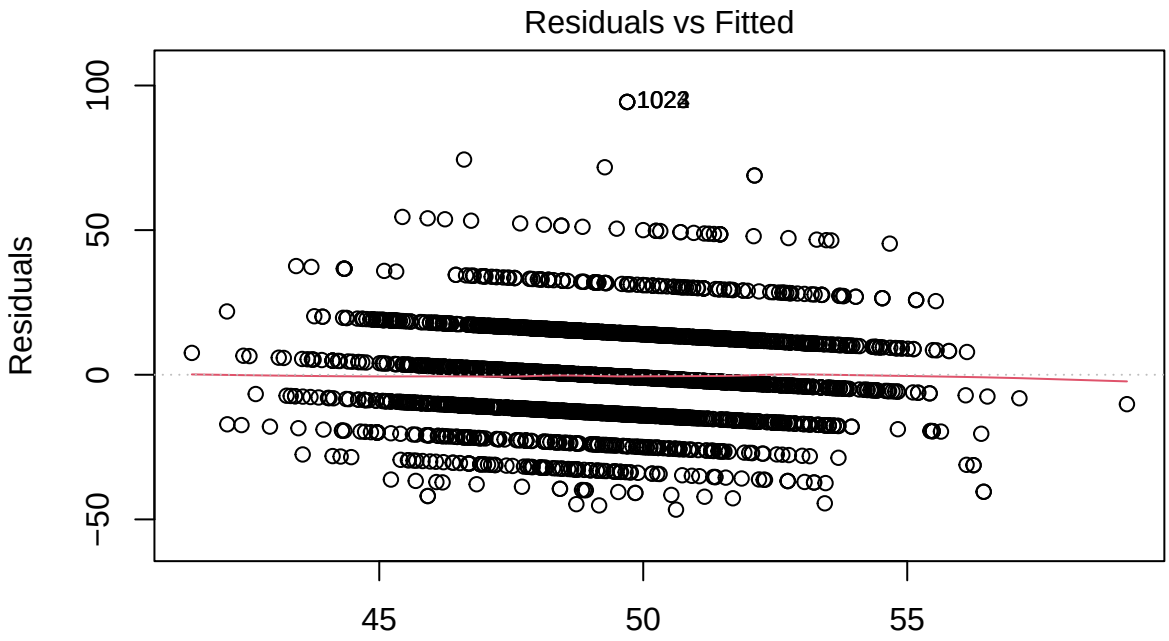



```
hist(log(raw_data$DaysMentHlthBad + 1))
```

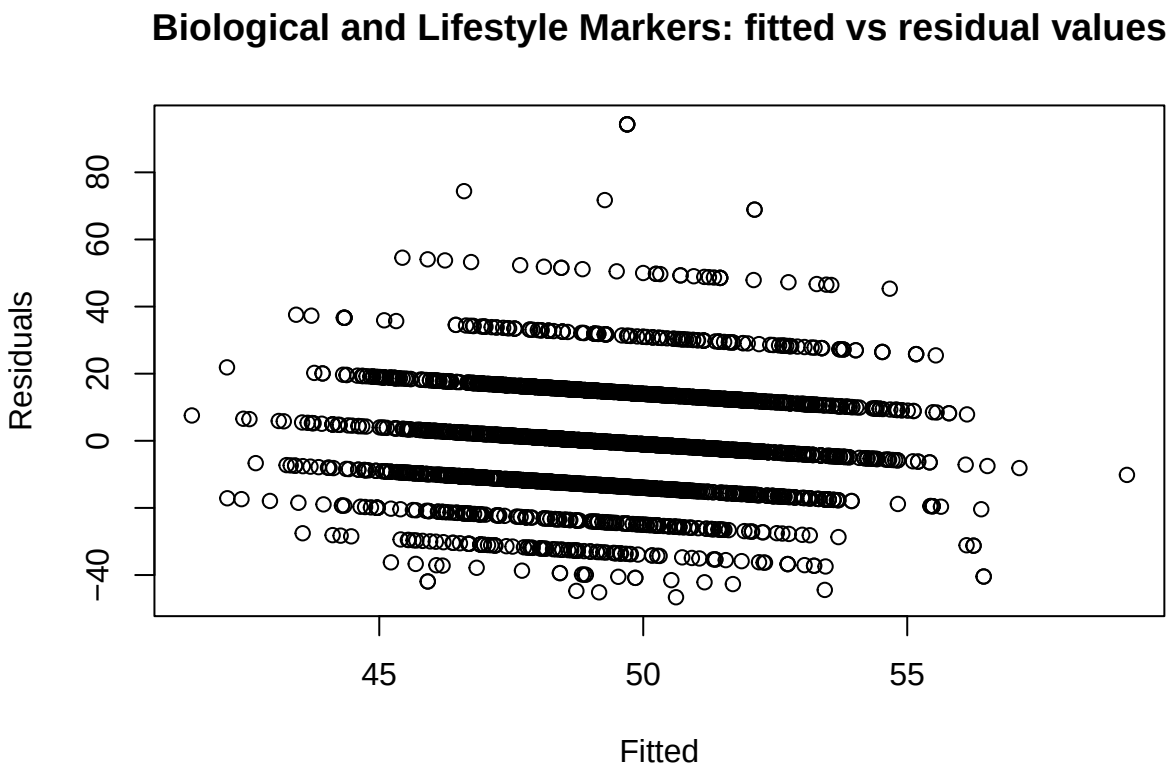


Checking Assumptions of Preliminary Linear Regression Model

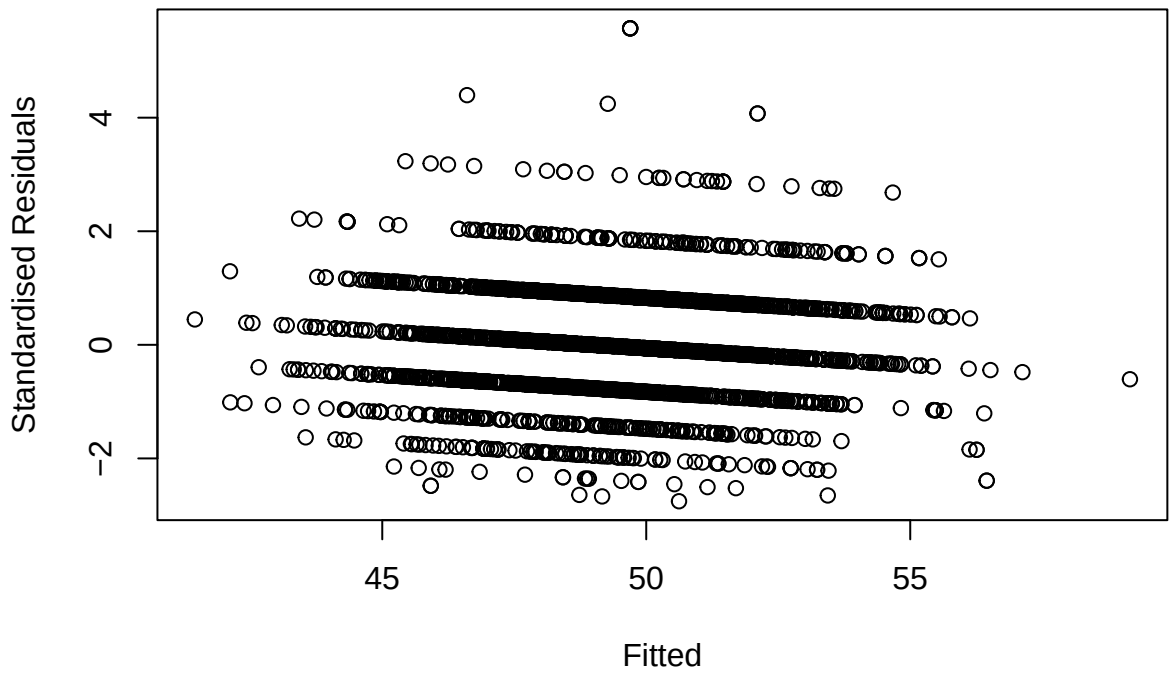
```
plot(fit, which = 1)
```



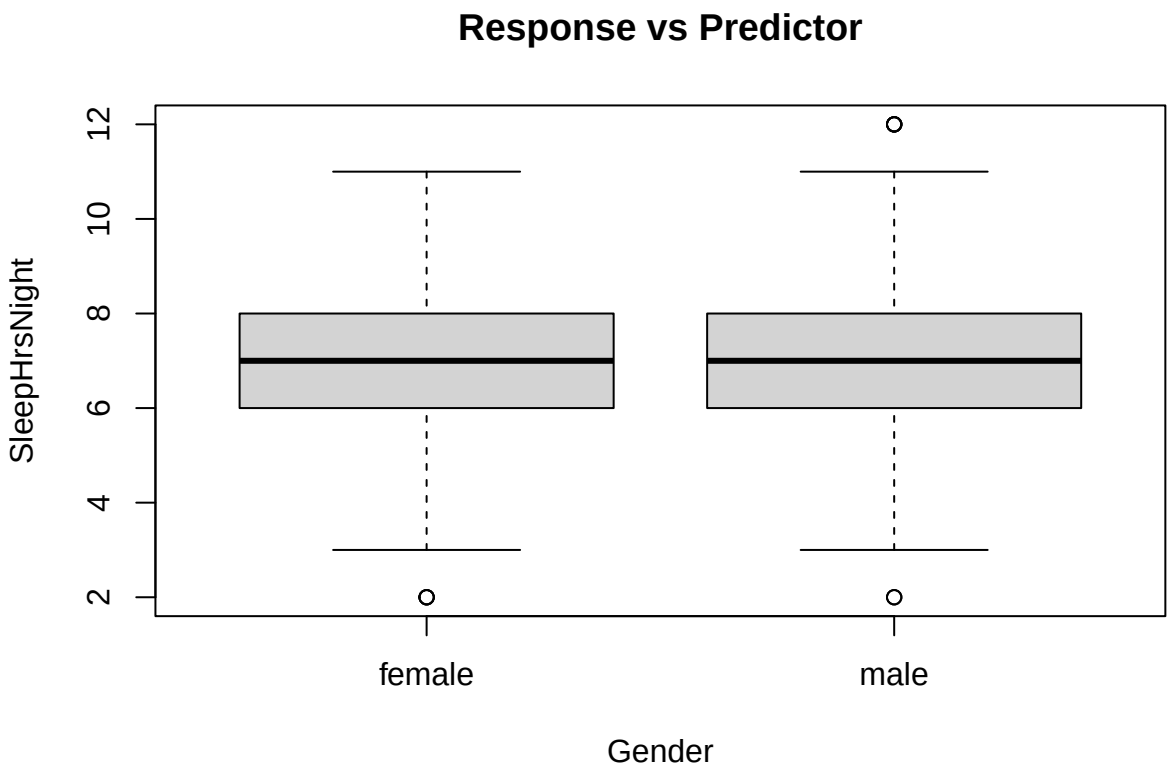

```
fitted_values = fitted(fit)
residual_values = resid(fit)
standardised_residual_values = rstandard(fit)
# Residuals against fitted values
plot(fitted_values, residual_values, main = "Biological and Lifestyle Markers: fitted vs residual values
```



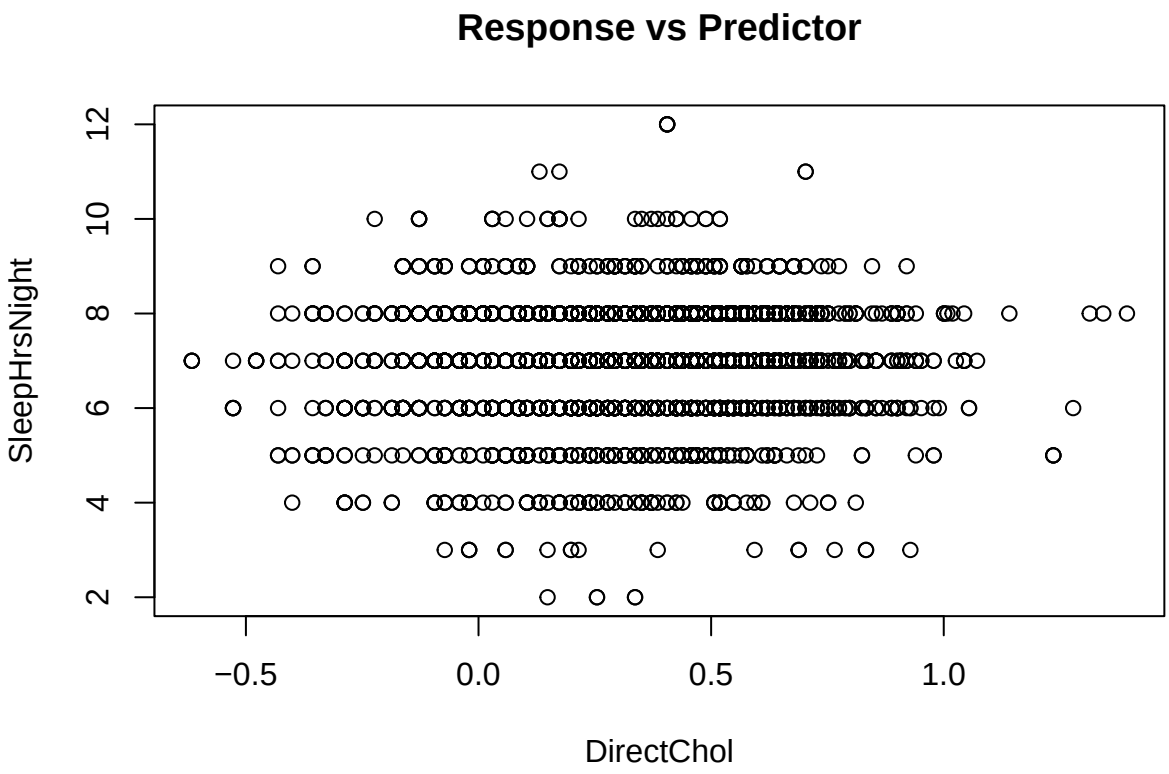

```
# Standardized residuals against fitted values
plot(fitted_values, standardised_residual_values, main = "Biological and Lifestyle Markers: fitted vs st
```



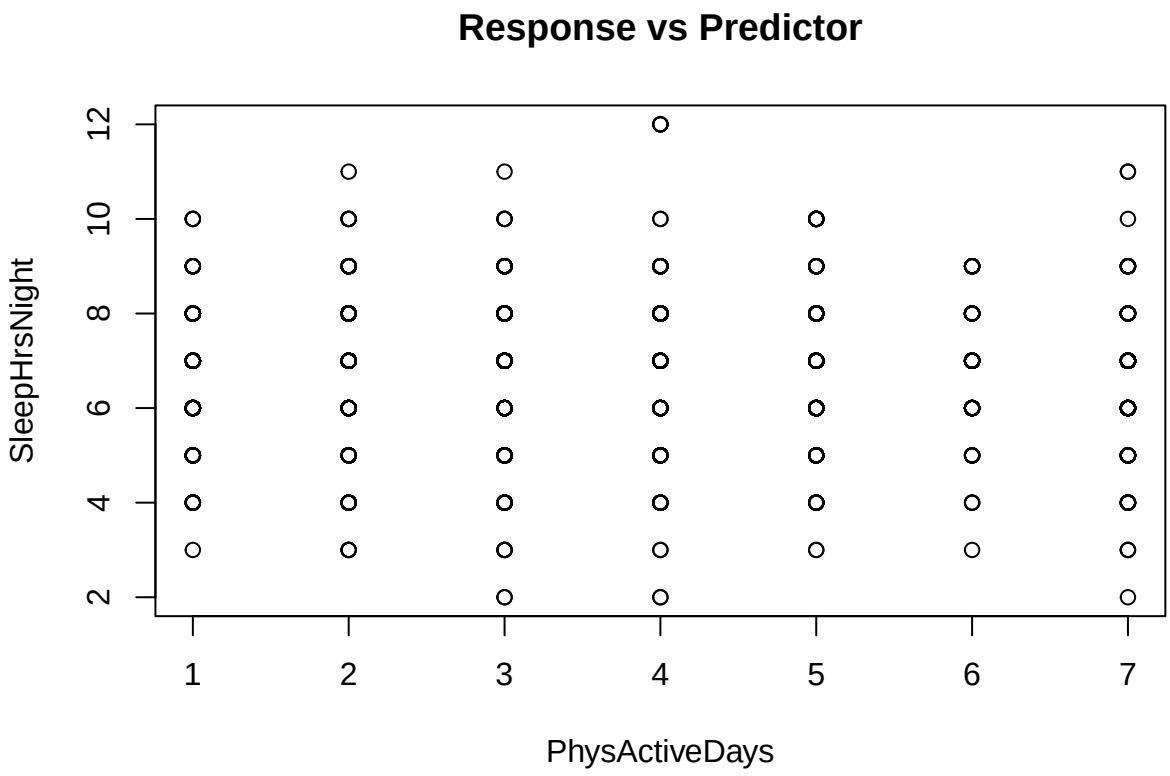

```
# Response vs predictor
plot(cleaned_data$Gender, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "Gender", yla
```



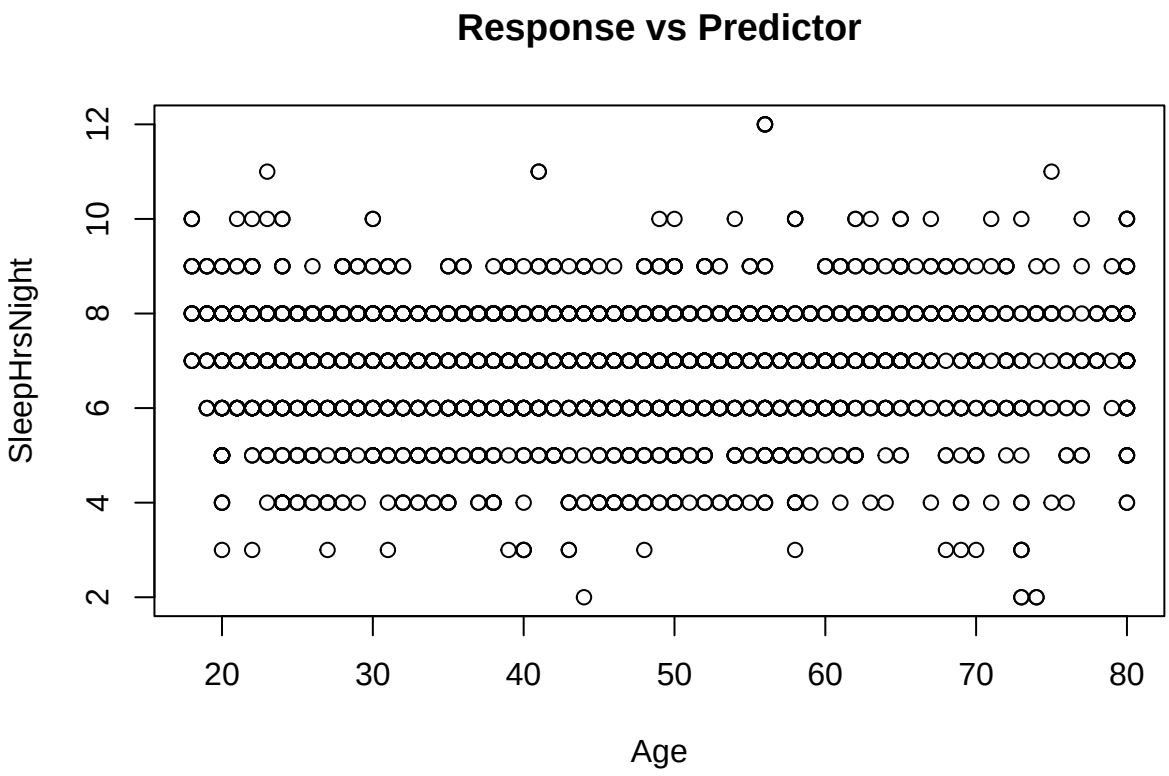

```
plot(cleaned_data$modDirectChol, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "Dire
```



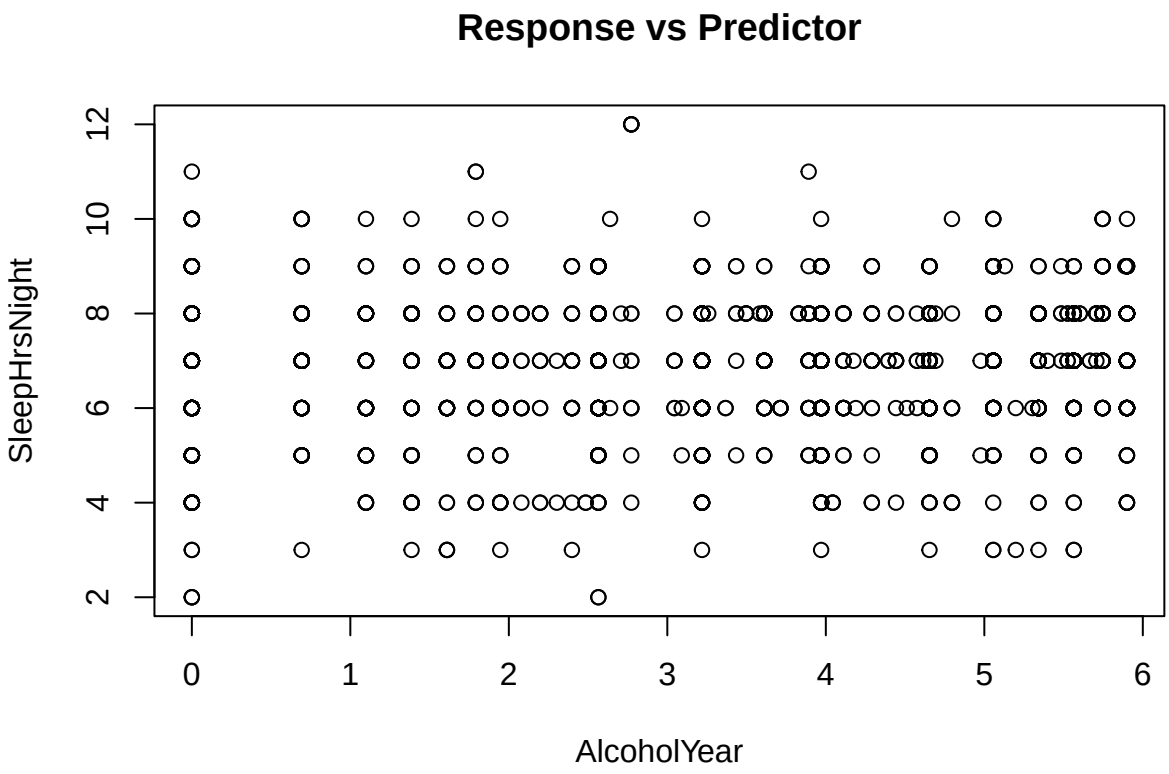

```
plot(cleaned_data$PhysActiveDays, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "Phys
```



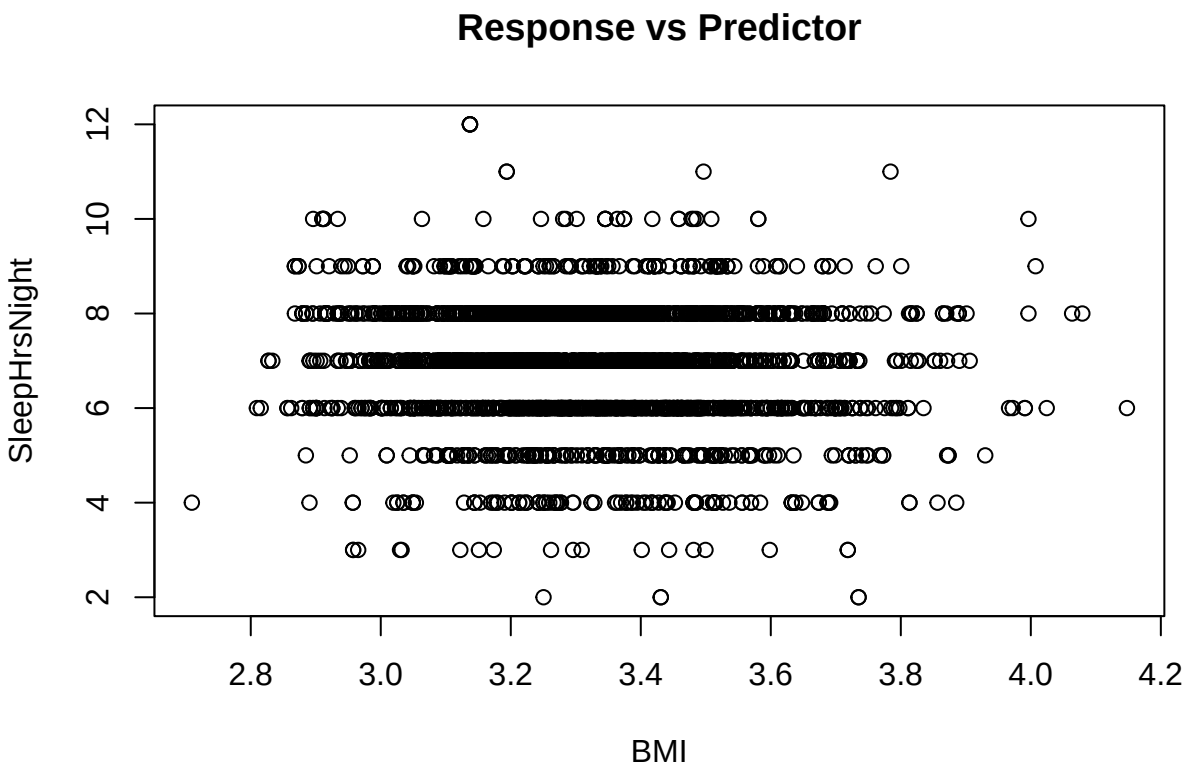

```
plot(cleaned_data$Age, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "Age", ylab = "S
```



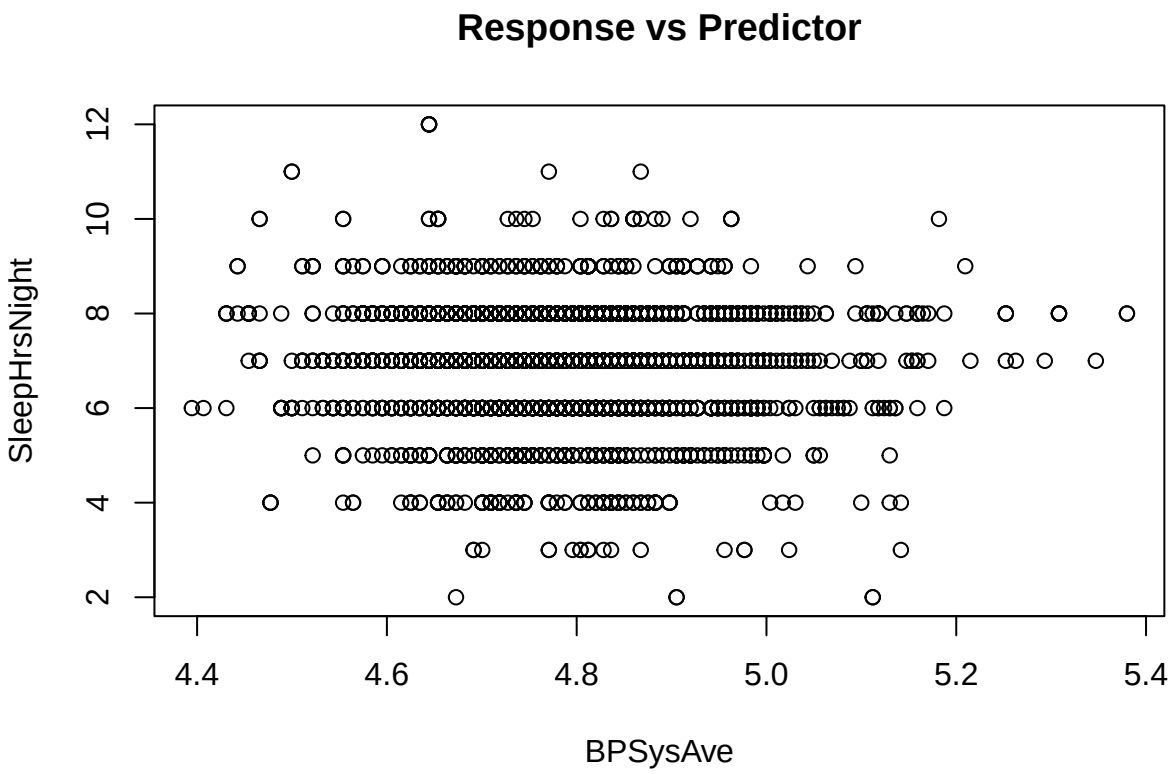

```
plot(cleaned_data$modAlcoholYear, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "Alc
```



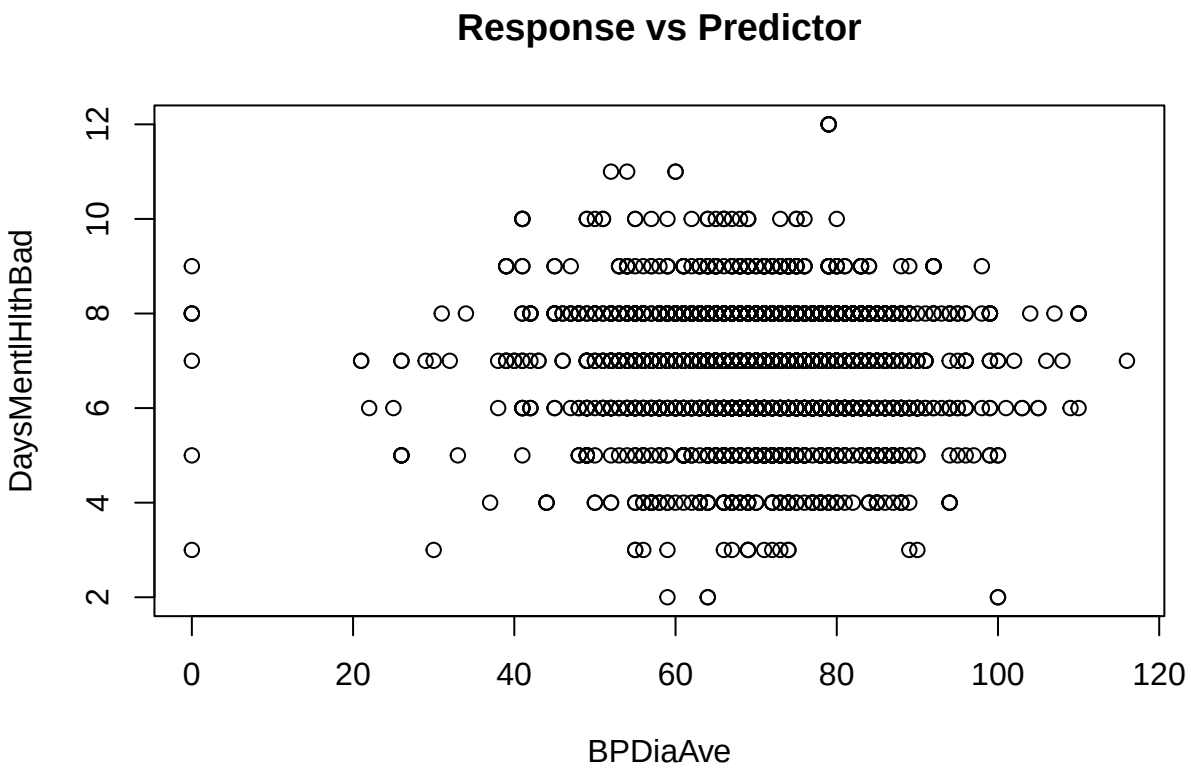

```
plot(cleaned_data$modBMI, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "BMI", ylab =
```



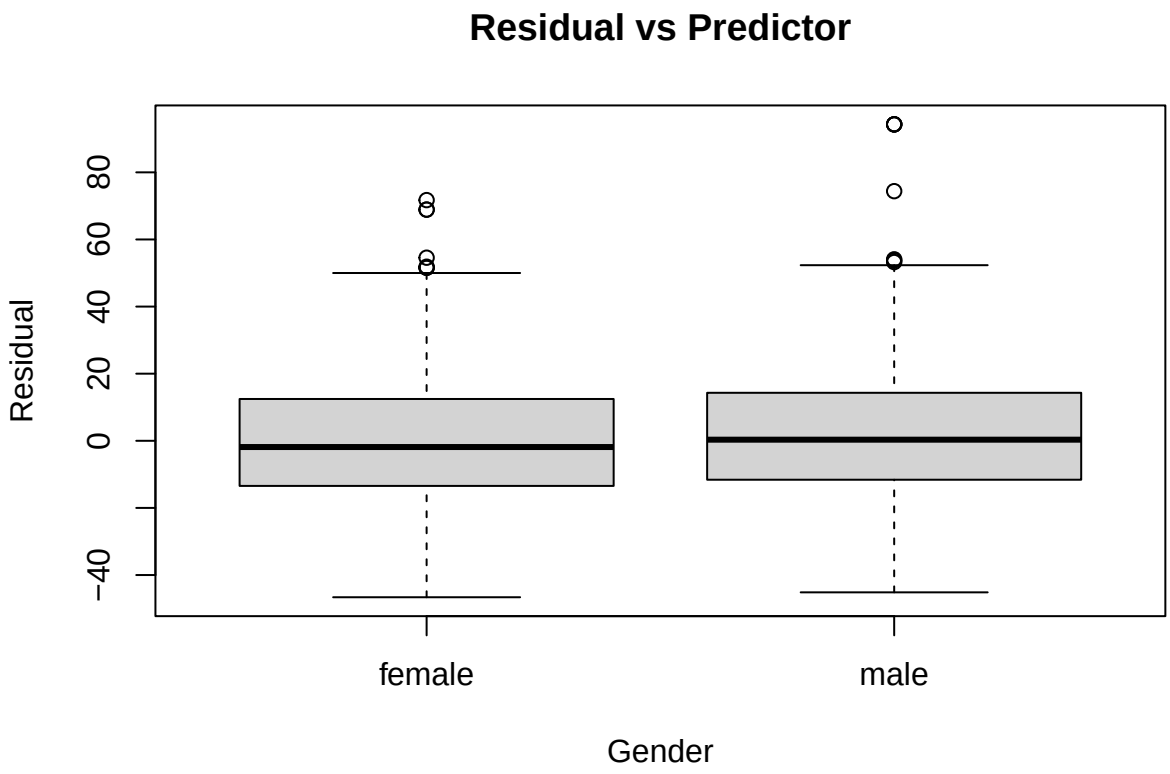

```
plot(cleaned_data$modBPSysAve, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "BPSysA
```



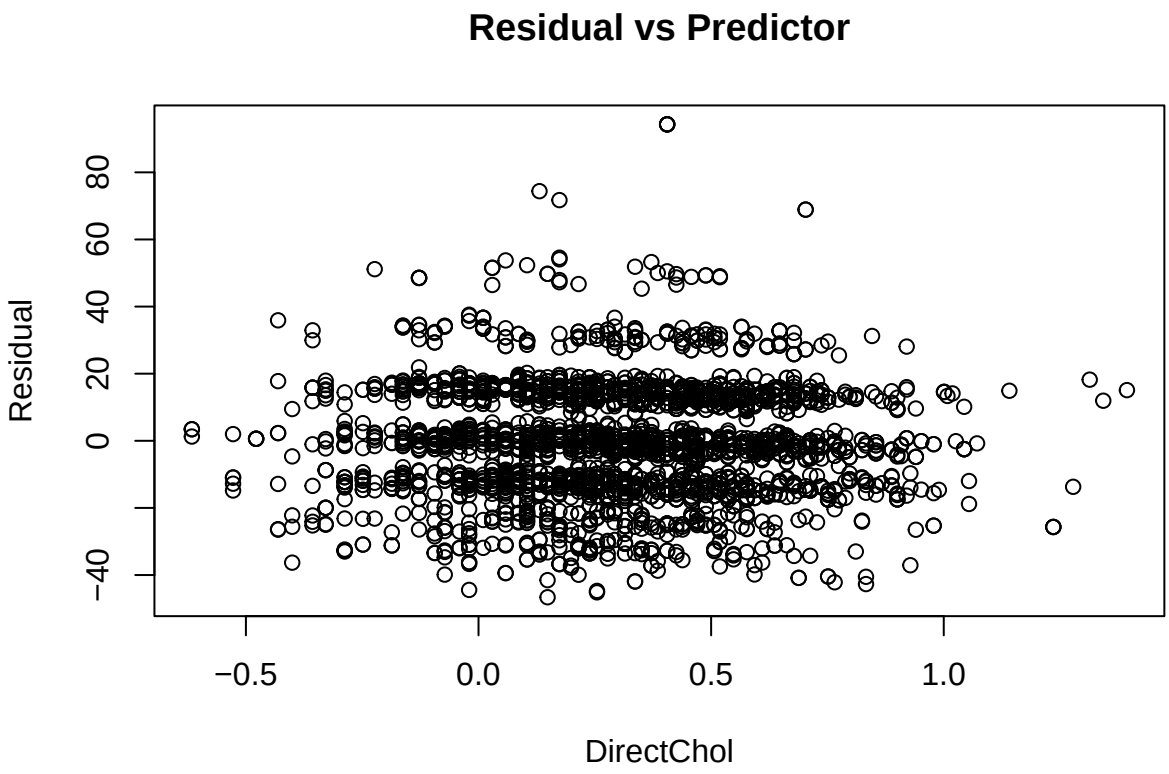

```
plot(cleaned_data$BPDiaAve, cleaned_data$SleepHrsNight, main="Response vs Predictor", xlab = "BPDiaAve"
```



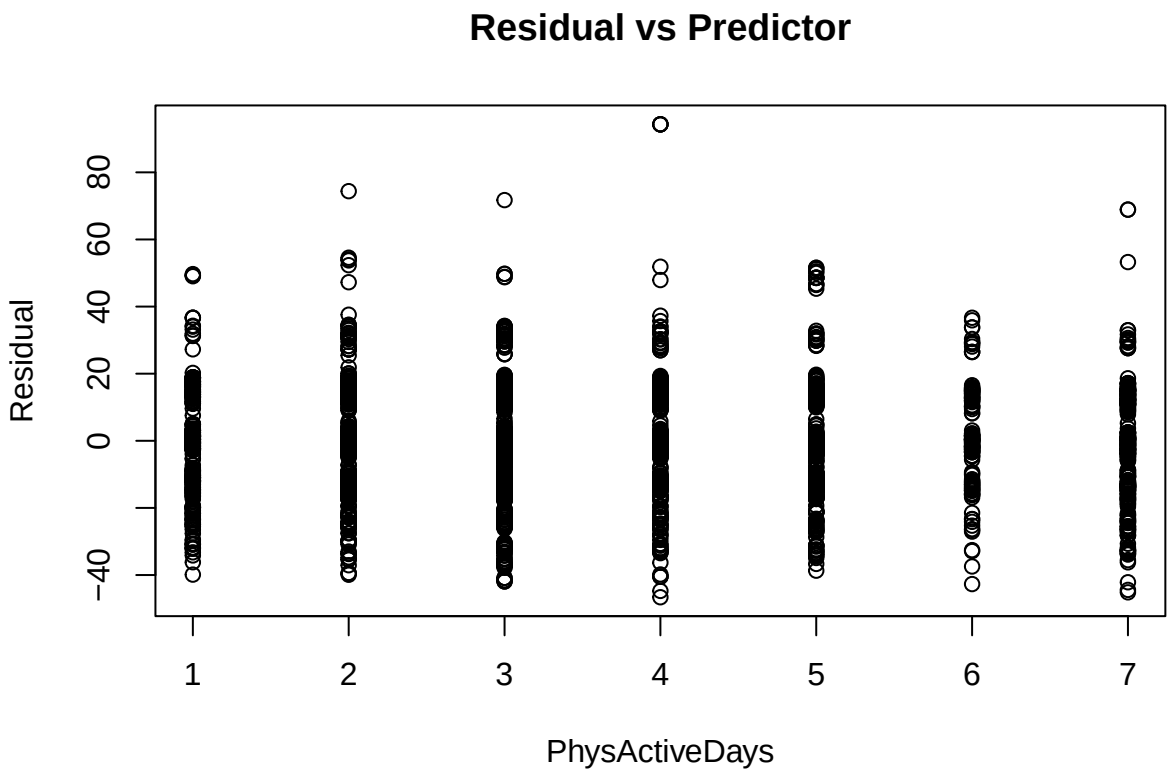

```
# Residual vs predictor
plot(cleaned_data$Gender, residual_values, main="Residual vs Predictor", xlab = "Gender", ylab = "Residu
```



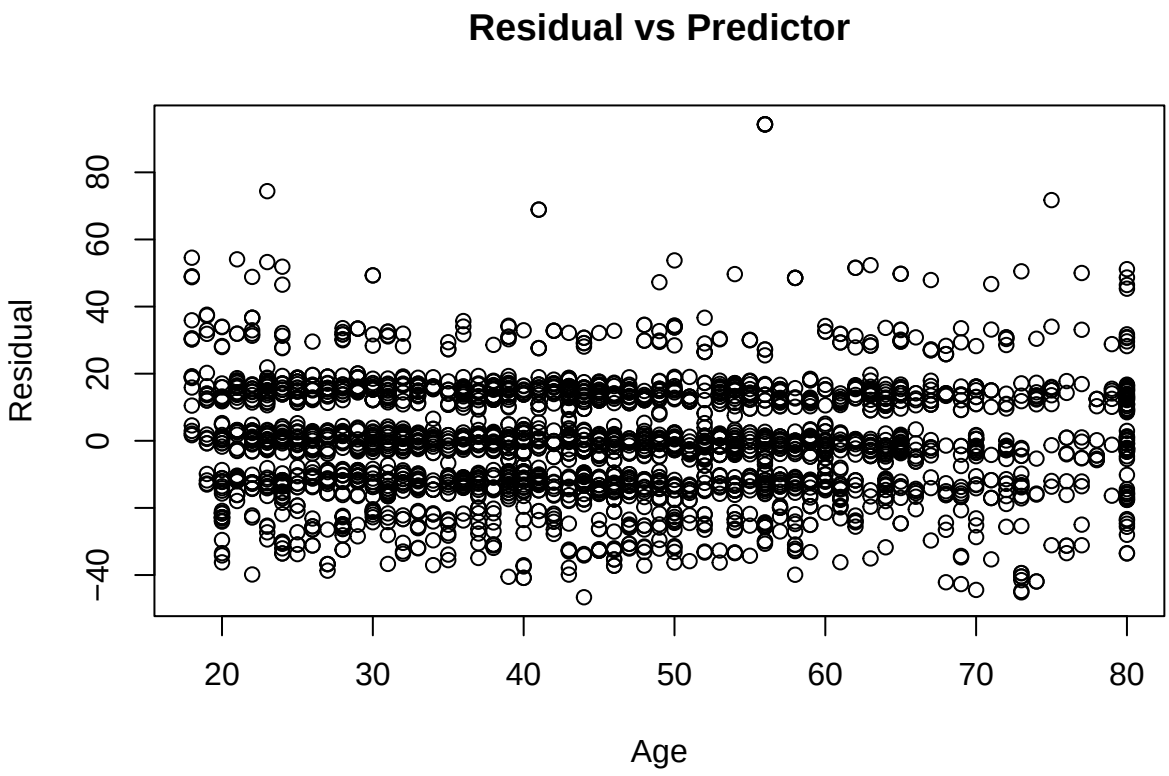

```
plot(cleaned_data$modDirectChol, residual_values, main="Residual vs Predictor", xlab = "DirectChol", yla
```



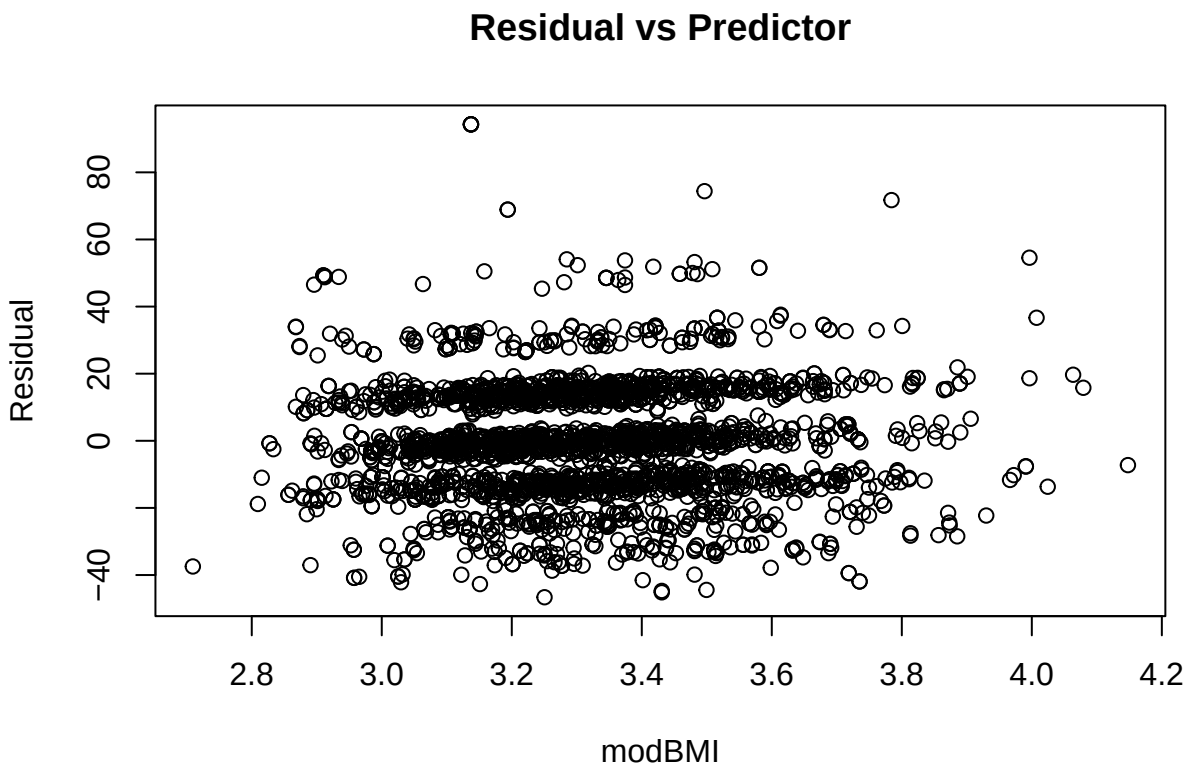

```
plot(cleaned_data$PhysActiveDays, residual_values, main="Residual vs Predictor", xlab = "PhysActiveDays
```



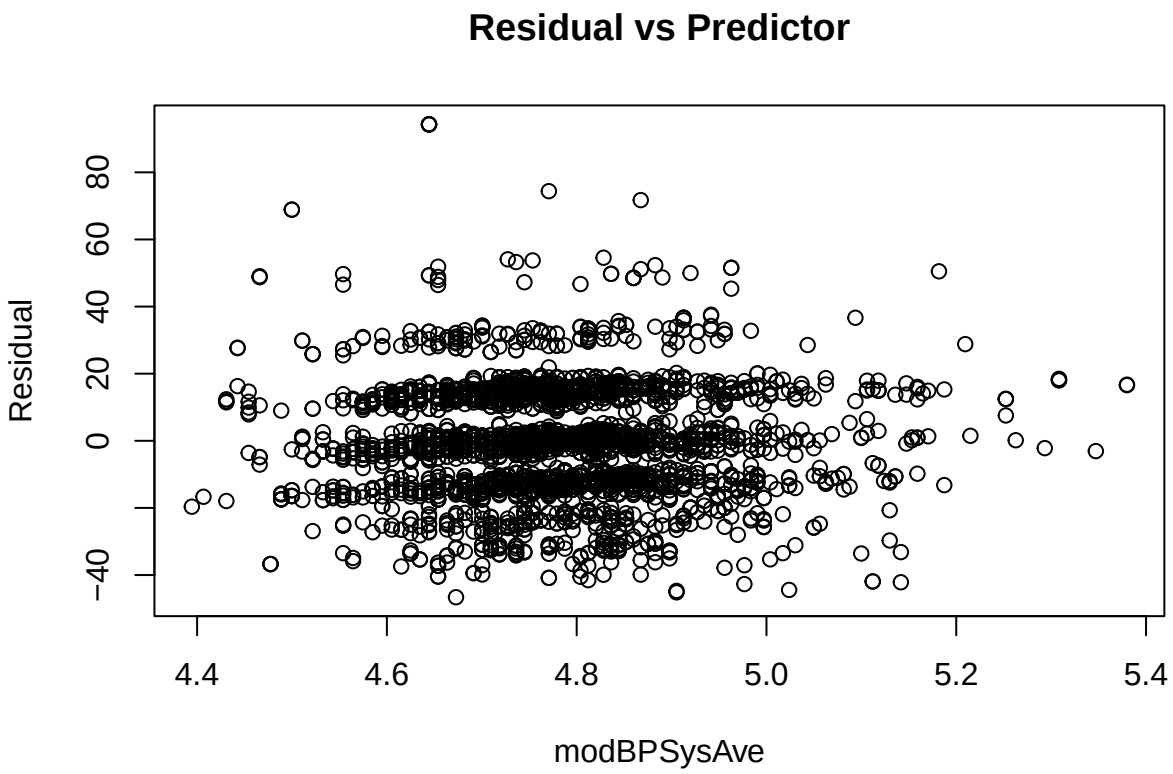

```
plot(cleaned_data$Age, residual_values, main="Residual vs Predictor", xlab = "Age", ylab = "Residual")
```



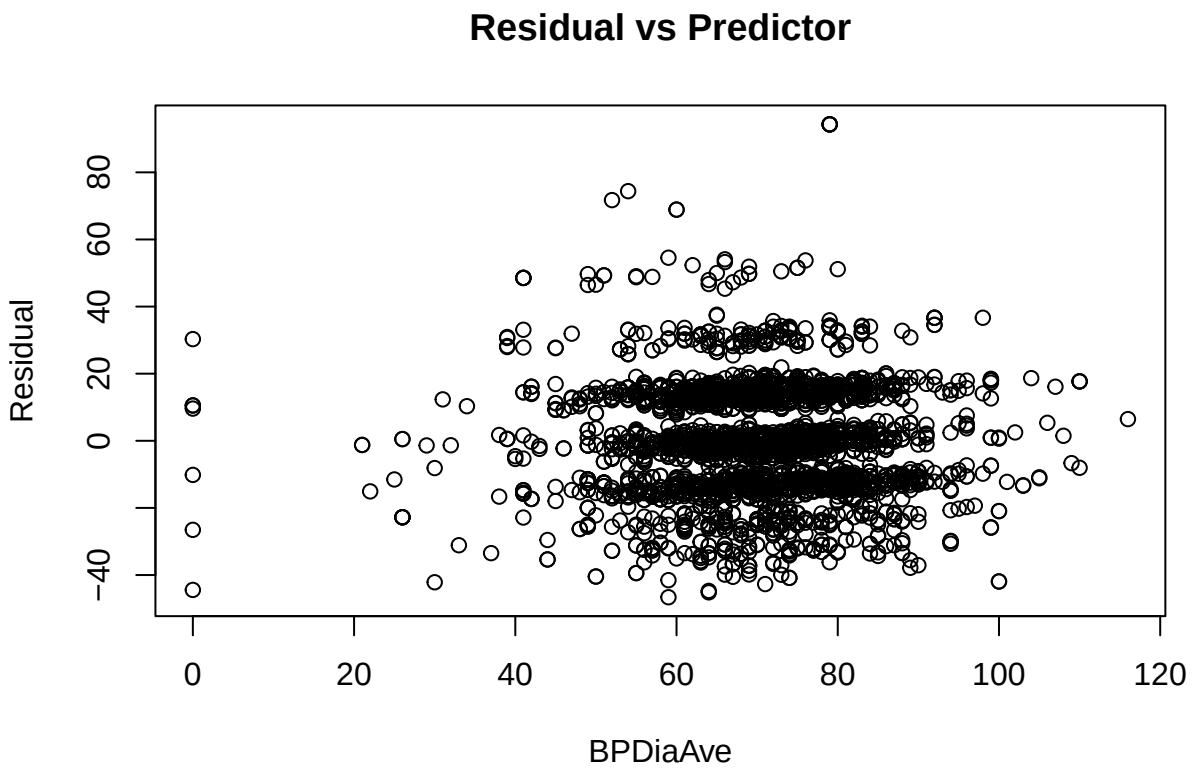

```
# plot(cleaned_data$modAlcoholYear, residual_values, main="Response vs Predictor", xlab = "modAlcoholYe
plot(cleaned_data$modBMI, residual_values, main="Residual vs Predictor", xlab = "modBMI", ylab = "Residu
```



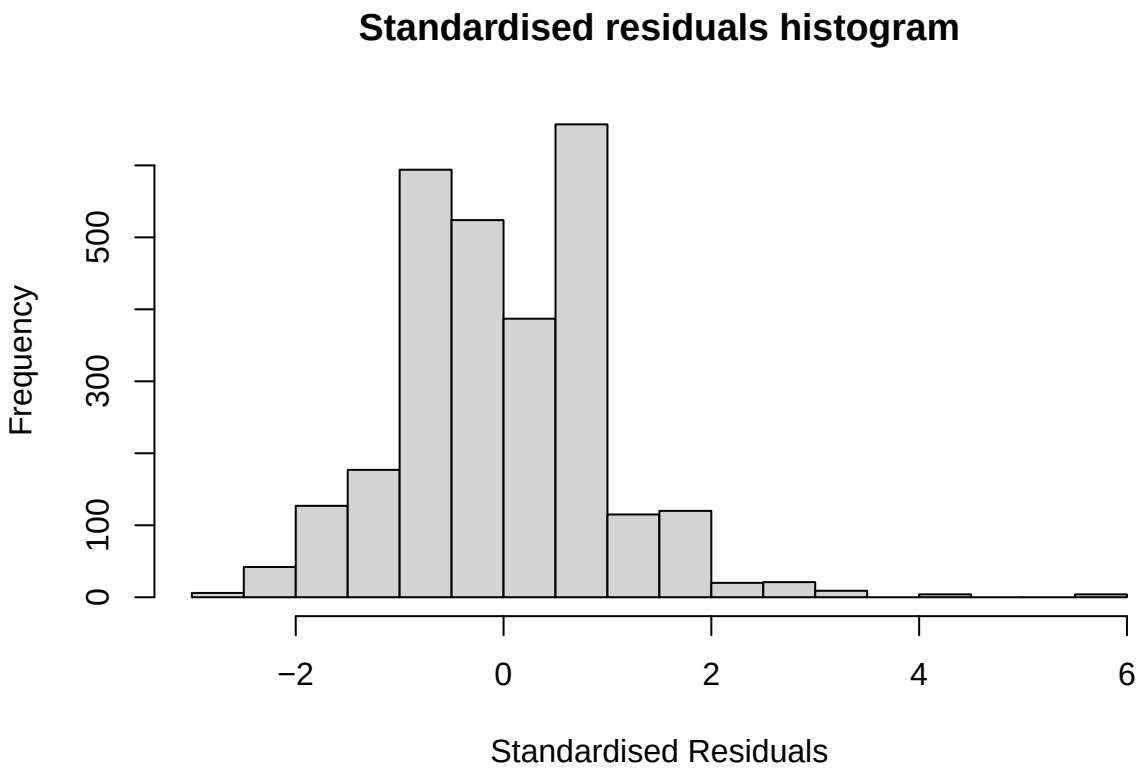

```
plot(cleaned_data$modBPSysAve, residual_values, main="Residual vs Predictor", xlab = "modBPSysAve", ylab
```



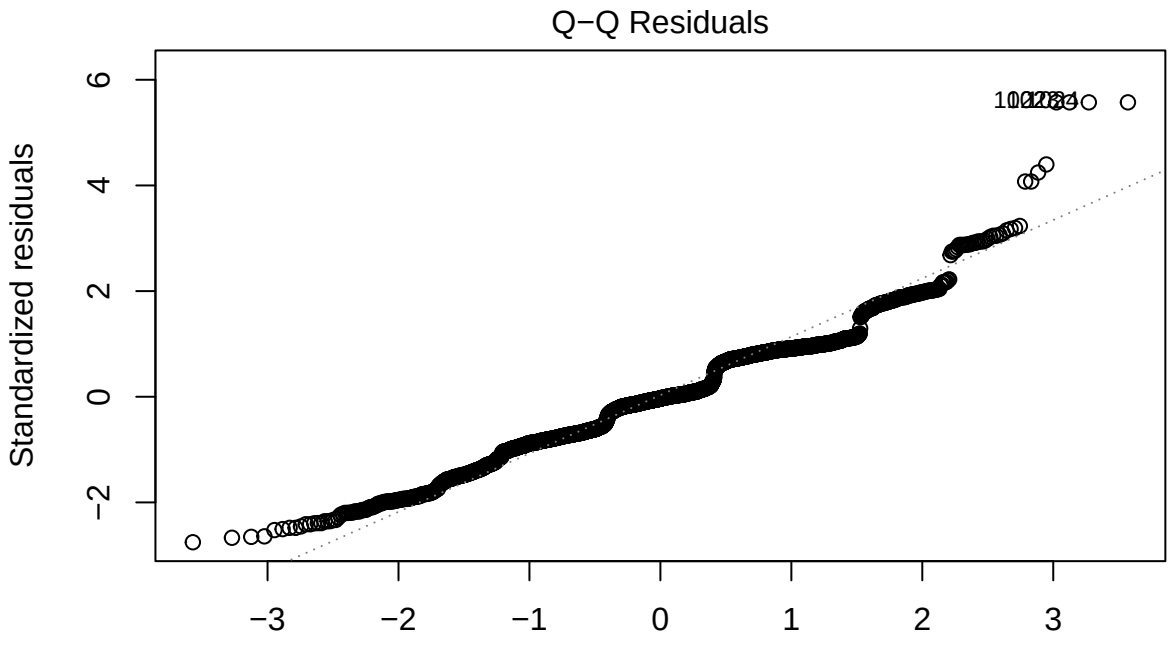

```
plot(cleaned_data$BPDiaAve, residual_values, main="Residual vs Predictor", xlab = "BPDiaAve", ylab = "R
```




```
# Residual histogram
hist(standardised_residual_values, xlab = "Standardised Residuals", main = "Standardised residuals hist
```




```
# QQ Plot
plot(fit, which = 2)
```




```
final_model <- lm(modSleepHrsNight ~ Gender+ Age+ modBMI + modBPSysAve + modAlcoholYear, data = cleaned_
summary(final_model)
```

```
##
## Call:
## lm(formula = modSleepHrsNight ~ Gender + Age + modBMI + modBPSysAve +
##     modAlcoholYear, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.050 -12.099  -0.566  13.217  93.526
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    94.58151   12.82982   7.372 2.20e-13 ***
## Gendermale     -2.63341    0.66083  -3.985 6.92e-05 ***
## Age             0.07424    0.02139   3.470 0.000528 ***
## modBMI         -4.14704    1.60195  -2.589 0.009683 **
## modBPSysAve    -7.34425    2.72175  -2.698 0.007010 **
## modAlcoholYear  0.53673    0.17241   3.113 0.001871 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.95 on 2801 degrees of freedom
## Multiple R-squared:  0.0197, Adjusted R-squared:  0.01795
## F-statistic: 11.26 on 5 and 2801 DF,  p-value: 9.081e-11
```