STAT 6021: Final Project EDA

Group 1

Libraries

```
library(tidyverse)
```

```
## -- 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 become errors
```

```
library(ggcorrplot)
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

Data

Looking at sleep efficiency, with hours of sleep as the response variable

```
##  [1] "Age"                    "Gender"                 "Sleep.duration"
##  [4] "Sleep.efficiency"       "REM.sleep.percentage"   "Deep.sleep.percentage"
##  [7] "Light.sleep.percentage" "Awakenings"             "Caffeine.consumption"
## [10] "Alcohol.consumption"    "Smoking.status"         "Exercise.frequency"
```

Cleaning

```
# Turn percentages to proportions
sleep <- sleep %>%
  mutate(REM.sleep.percentage = REM.sleep.percentage / 100,
         Deep.sleep.percentage = Deep.sleep.percentage / 100,
         Light.sleep.percentage = Light.sleep.percentage / 100)

# Weighted sleep (Sleep.duration * Sleep.efficiency)
sleep <- sleep %>%
  mutate(weighted.sleep = Sleep.duration * Sleep.efficiency)
```

```r
str(sleep)
```

```
## 'data.frame':    452 obs. of  13 variables:
##  $ Age                   : int  65 69 40 40 57 36 27 53 41 11 ...
##  $ Gender                : chr  "Female" "Male" "Female" "Female" ...
##  $ Sleep.duration        : num  6 7 8 6 8 7.5 6 10 6 9 ...
##  $ Sleep.efficiency      : num  0.88 0.66 0.89 0.51 0.76 0.9 0.54 0.9 0.79 0.55 ...
##  $ REM.sleep.percentage  : num  0.18 0.19 0.2 0.23 0.27 0.23 0.28 0.28 0.28 0.18 ...
##  $ Deep.sleep.percentage : num  0.7 0.28 0.7 0.25 0.55 0.6 0.25 0.52 0.55 0.37 ...
##  $ Light.sleep.percentage: num  0.12 0.53 0.1 0.52 0.18 0.17 0.47 0.2 0.17 0.45 ...
##  $ Awakenings            : num  0 3 1 3 3 0 2 0 3 4 ...
##  $ Caffeine.consumption  : num  0 0 0 50 0 NA 50 50 50 0 ...
##  $ Alcohol.consumption   : num  0 3 0 5 3 0 0 0 0 0 ...
##  $ Smoking.status        : chr  "Yes" "Yes" "No" "Yes" ...
##  $ Exercise.frequency    : num  3 3 3 1 3 1 1 3 1 0 ...
##  $ weighted.sleep        : num  5.28 4.62 7.12 3.06 6.08 6.75 3.24 9 4.74 4.95 ...
```

Summary Statistics

```r
#sleep duration
summary(sleep$Sleep.duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.000   7.500   7.466   8.000  10.000
```

```r
#sleep efficiency
summary(sleep$Sleep.efficiency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5000  0.6975  0.8200  0.7889  0.9000  0.9900
```

```r
#weighted sleep
summary(sleep$weighted.sleep)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.600   5.025   6.075   5.887   6.728   9.200
```

```r
# Model for sleep duration with consumption predictors
lm.duration <- lm(Sleep.duration ~ Caffeine.consumption + Alcohol.consumption + Smoking.status + Exercis
summary(lm.duration)
```

```
##
## Call:
## lm(formula = Sleep.duration ~ Caffeine.consumption + Alcohol.consumption +
##     Smoking.status + Exercise.frequency, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.54074 -0.47209  0.01386  0.52791  2.58649
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.575074   0.092103  82.246   <2e-16 ***
## Caffeine.consumption -0.001280   0.001525  -0.839    0.402
## Alcohol.consumption  -0.019320   0.027465  -0.703    0.482
## Smoking.statusYes     0.014050   0.092724   0.152    0.880
## Exercise.frequency   -0.034329   0.030372  -1.130    0.259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8824 on 402 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.005622,   Adjusted R-squared:  -0.004272
## F-statistic: 0.5683 on 4 and 402 DF,  p-value: 0.6858
```

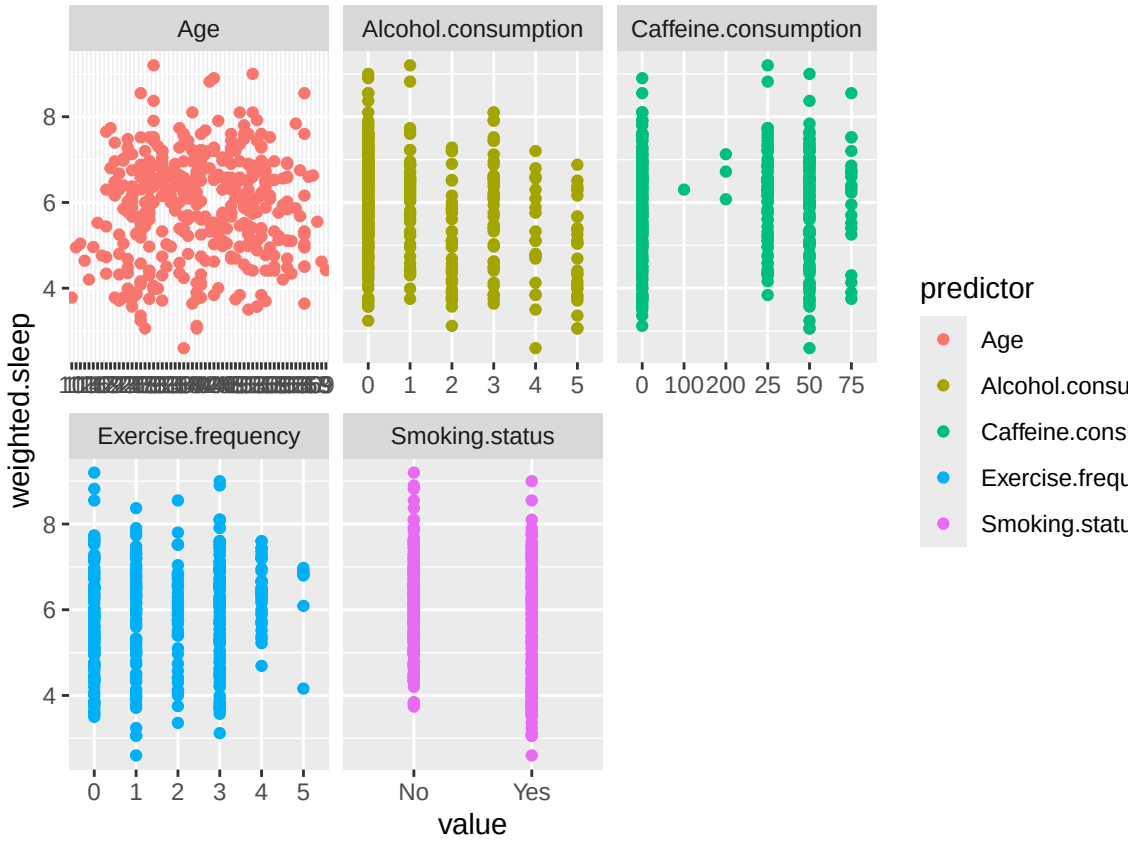
```
# Model for weighted sleep with consumption predictors
lm.weightsleep <- lm(weighted.sleep ~ Caffeine.consumption + Alcohol.consumption + Smoking.status + Exe
summary(lm.weightsleep)
```

```
##
## Call:
## lm(formula = weighted.sleep ~ Caffeine.consumption + Alcohol.consumption +
##     Smoking.status + Exercise.frequency, data = sleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4926 -0.7818 -0.0038  0.7429  3.3723
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.0456482  0.1145083  52.797  < 2e-16 ***
## Caffeine.consumption  0.0007615  0.0018962   0.402    0.688
## Alcohol.consumption  -0.2370237  0.0341465  -6.941 1.57e-11 ***
## Smoking.statusYes    -0.5071377  0.1152799  -4.399 1.39e-05 ***
## Exercise.frequency    0.1560424  0.0377602   4.132 4.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 402 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.1842, Adjusted R-squared:  0.176
## F-statistic: 22.68 on 4 and 402 DF,  p-value: < 2.2e-16
```

Model Assumptions

```
# Only predictors
sleep2 <- sleep[,c(13,1,9,10,11,12)]

# Drop missing predictors
sleep2 <- na.omit(sleep2)
```

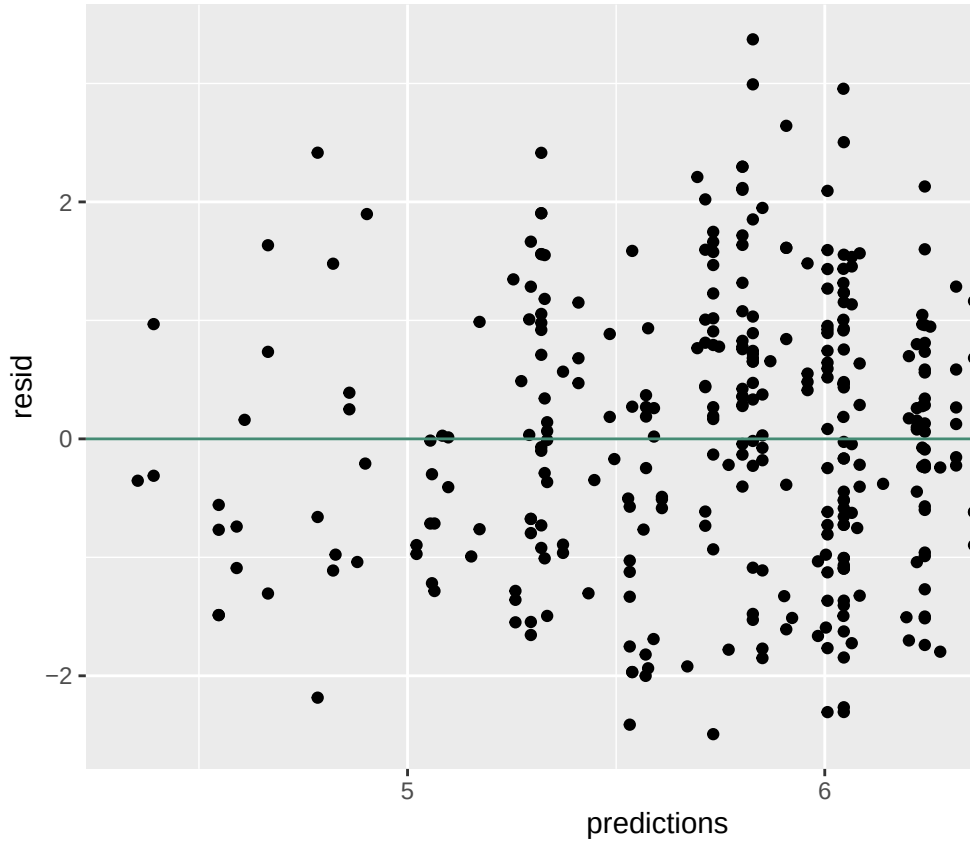
```
long <- gather(sleep2, key="predictor", value="value",
               Age, Caffeine.consumption, Alcohol.consumption,
               Smoking.status, Exercise.frequency)

ggplot(long, aes(x=value, y=weighted.sleep, color=predictor)) +
  geom_point() +
  facet_wrap(~predictor, scale="free_x")
```



Linearity assumption

The linearity assumption isn't totally met, due to the nature of our predictors. Besides age, which doesn't seem to exhibit any sort of linear relationship, our data is broken up into distinct categories so our x-axis wouldn't be continuous. As a result, a linear relationship isn't super clear.

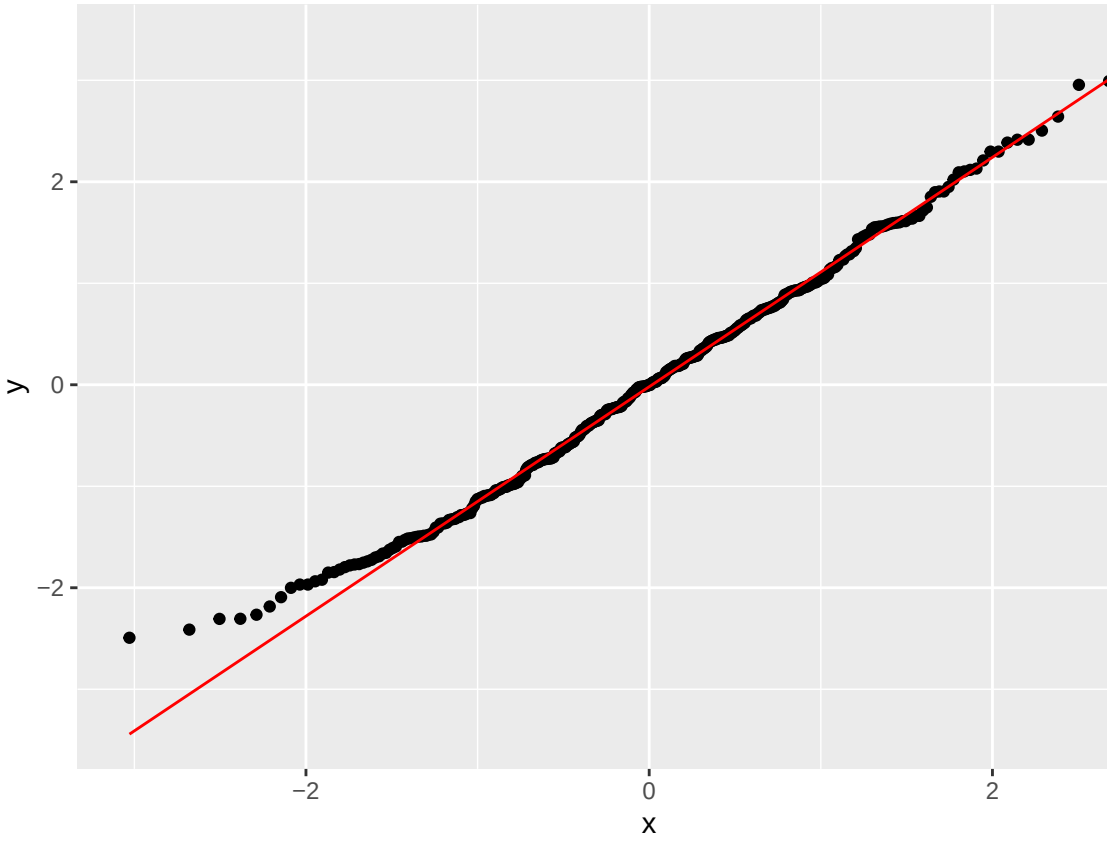
```
model_pred <- mutate(sleep2, predictions=fitted(lm.weightsleep),
                     resid=residuals(lm.weightsleep))

ggplot(model_pred, aes(x=predictions, y=resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color="aquamarine4")
```



Equal variance and Independence

The equal variance and independence assumptions appear to be met as the residuals appear to be scattered around 0 and there are no apparent clusters

```
ggplot(model_pred, aes(sample=resid)) +
  stat_qq() + stat_qq_line(color="red")
```



Normality assumption

Our model fits the normality assumption as it very closely follows the line of normality for the Q-Q plot. This means that our residuals follow a normal distribution and with it meeting all of the assumptions except for potentially the linear assumption. Therefore, we deem this model to be acceptable

Case A: Unhealthy

```
caseA <- data.frame(Age=23, Caffeine.consumption=200, Alcohol.consumption=5,
                    Smoking.status="Yes", Exercise.frequency=0)

predict(lm.weightsleep, caseA, interval="prediction")
```

```
##        fit      lwr      upr
## 1 4.505685 2.221789 6.789581
```

Case B: Health nut

```
caseB <- data.frame(Age=50, Caffeine.consumption=0, Alcohol.consumption=0,
                    Smoking.status="No", Exercise.frequency=5)

predict(lm.weightsleep, caseB, interval="prediction")
```

```
##       fit      lwr      upr
## 1 6.82586 4.649688 9.002032
```

Case C: Average College Student on weekend

```
caseC <- data.frame(Age=22, Caffeine.consumption=100, Alcohol.consumption=3,
                    Smoking.status="No", Exercise.frequency=3)

predict(lm.weightsleep, caseC, interval="confidence")
```

```
##        fit      lwr      upr
## 1 5.878851 5.509233 6.248468
```