# 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     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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)
```

## Data

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

```
sleep <- read.csv("Sleep_Efficiency.csv")

# Remove ID, bedtime, and wakeup time
sleep <- sleep[, -c(1, 4, 5)]
View(sleep)
```

```
colnames(sleep)
```

```
##  [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

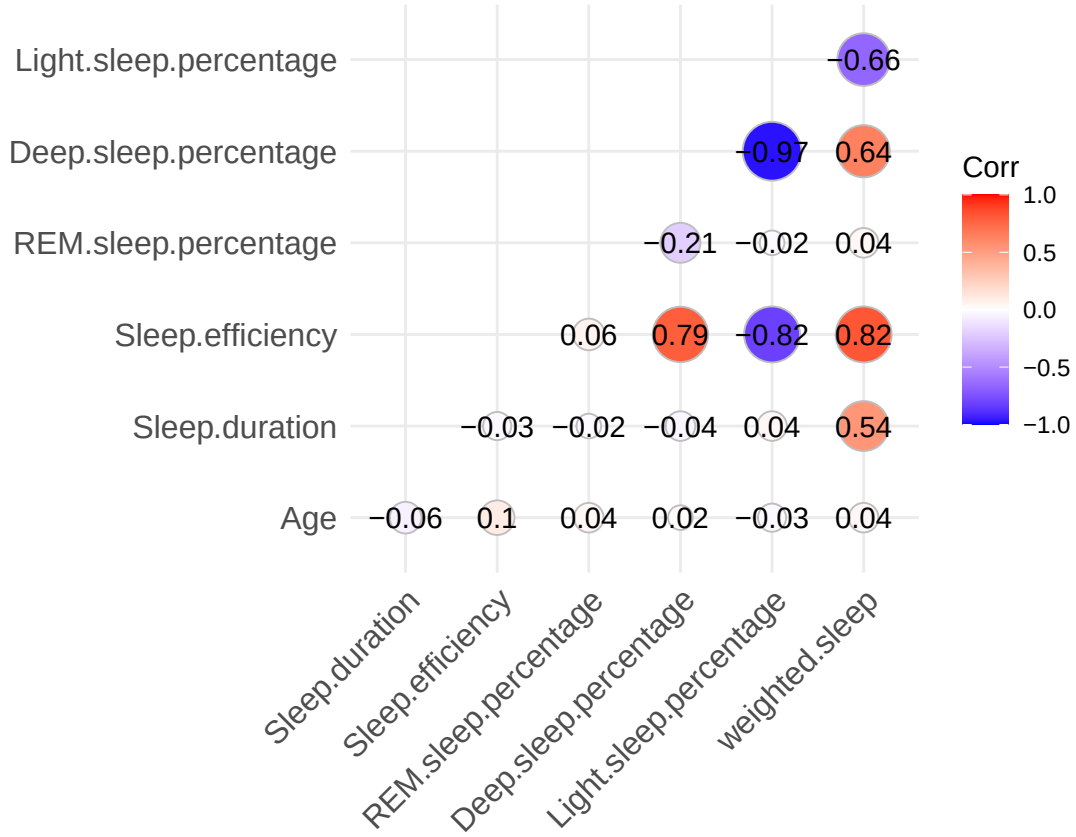
```r
# Turn percentages to proportions
sleep <- sleep %>%
  mutate(REM.sleep.percentage = REM.sleep.percentage / 100,
         Deep.sleep.percentage = Deep.sleep.percentage / 100,
         Light.sleep.percentage = Light.sleep.percentage / 100)

# 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 ...
```

## Correlation Matrix

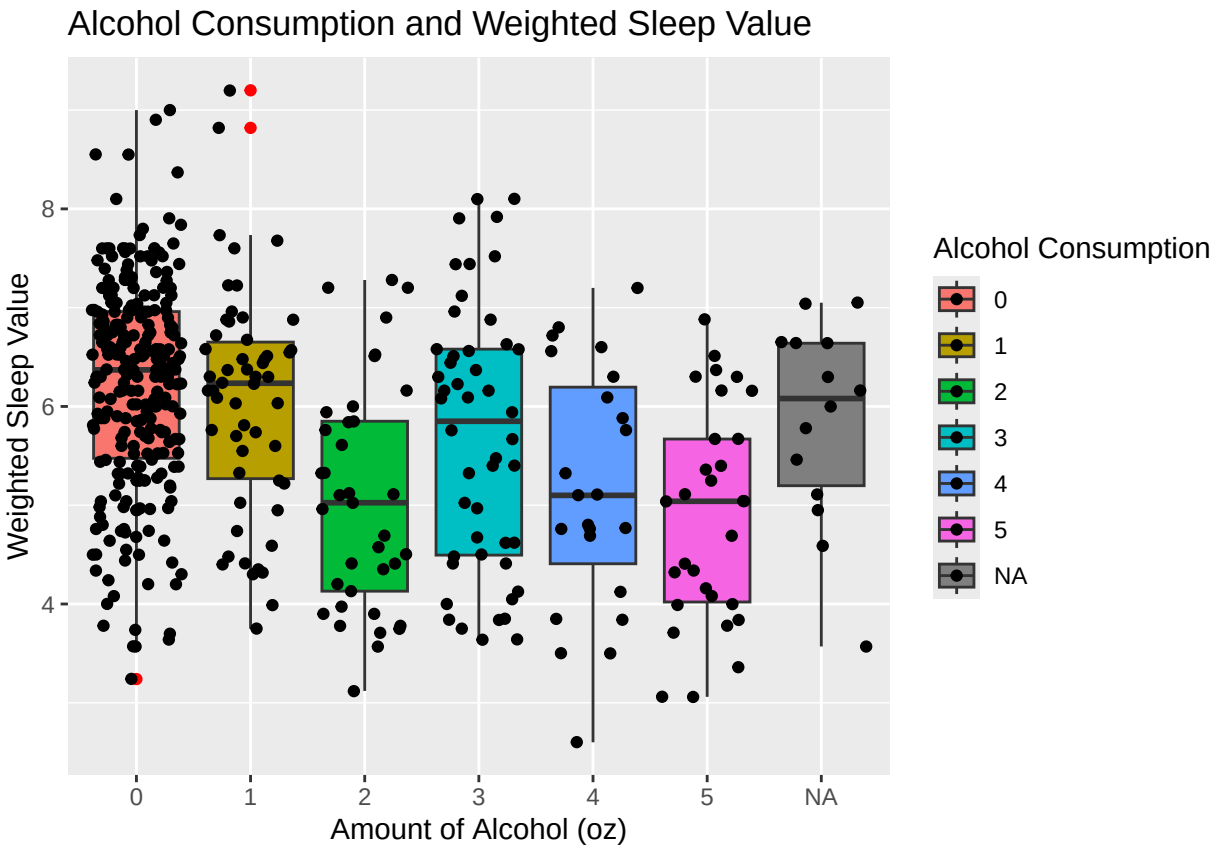
```r
sleep.num <- sleep[-c(2, 11)]
cor_mat<-round(cor(sleep.num),2)
ggcorrplot(cor_mat,lab=TRUE, type="lower", method="circle")
```




## Boxplots

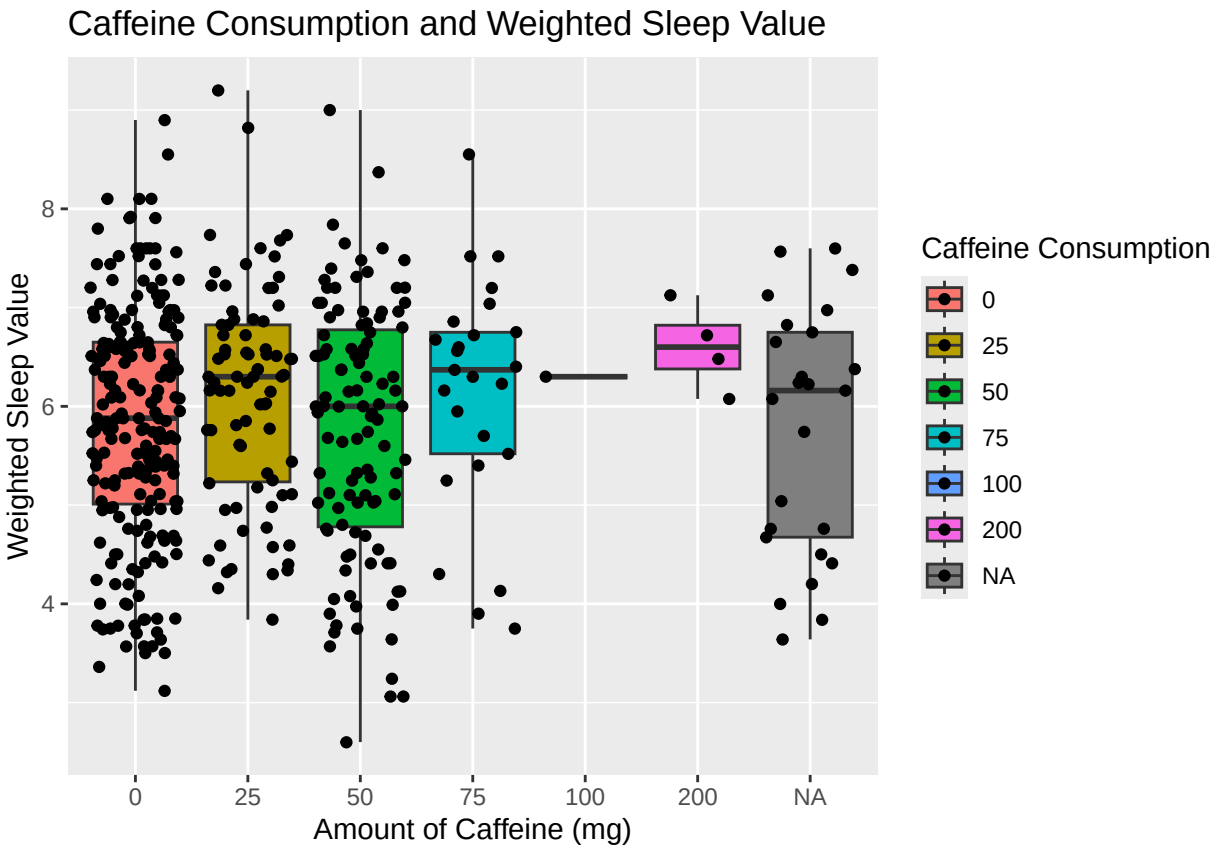
### Alcohol Consumption

```
ggplot(sleep, aes(x=as.factor(Alcohol.consumption), y=weighted.sleep,
                  fill=as.factor(Alcohol.consumption)))+
  geom_boxplot(outlier.color="red")+geom_jitter()+
  labs(title = "Alcohol Consumption and Weighted Sleep Value",
       x="Amount of Alcohol (oz)", y = "Weighted Sleep Value",
       fill="Alcohol Consumption")
```



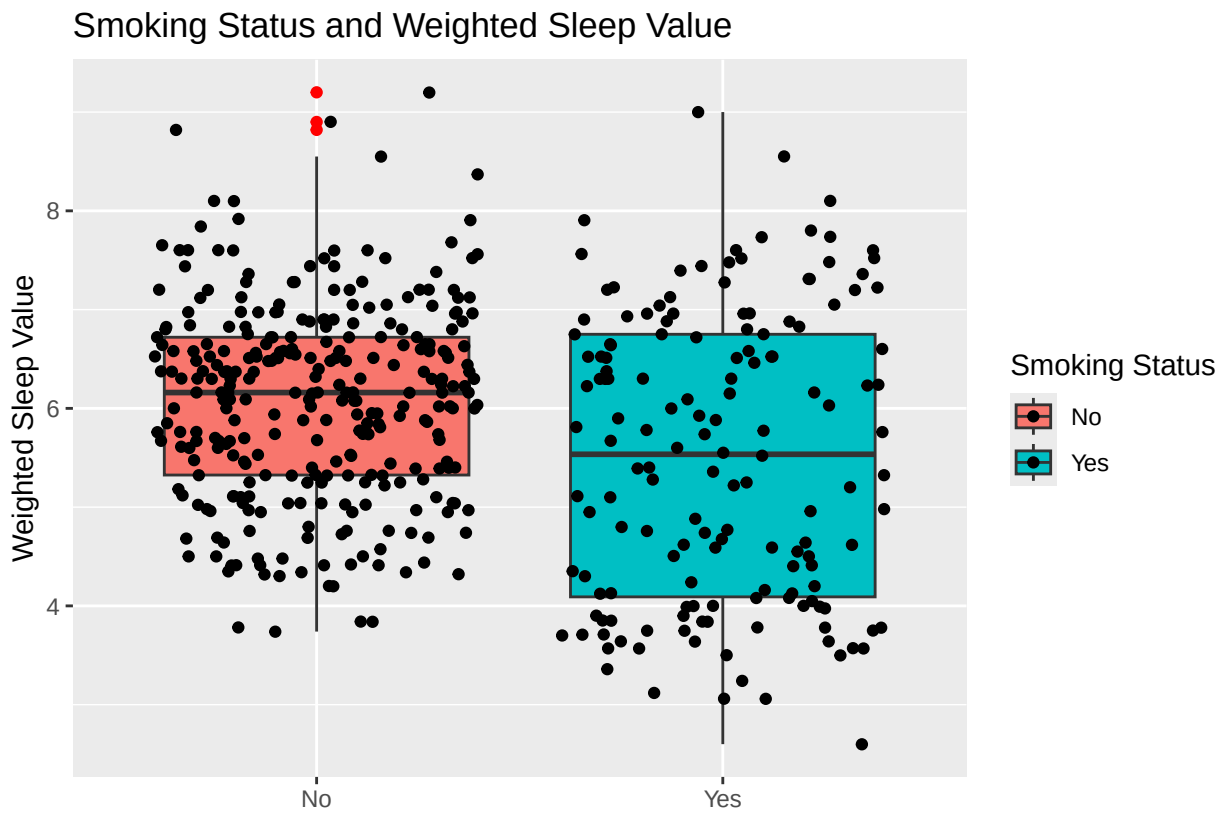

**Caffeine Consumption**

```
ggplot(sleep, aes(x=as.factor(Caffeine.consumption), y=weighted.sleep,
                  fill=as.factor(Caffeine.consumption)))+
  geom_boxplot(outlier.color="red")+geom_jitter()+
  labs(title = "Caffeine Consumption and Weighted Sleep Value",
       x="Amount of Caffeine (mg)", y = "Weighted Sleep Value",
       fill="Caffeine Consumption")
```



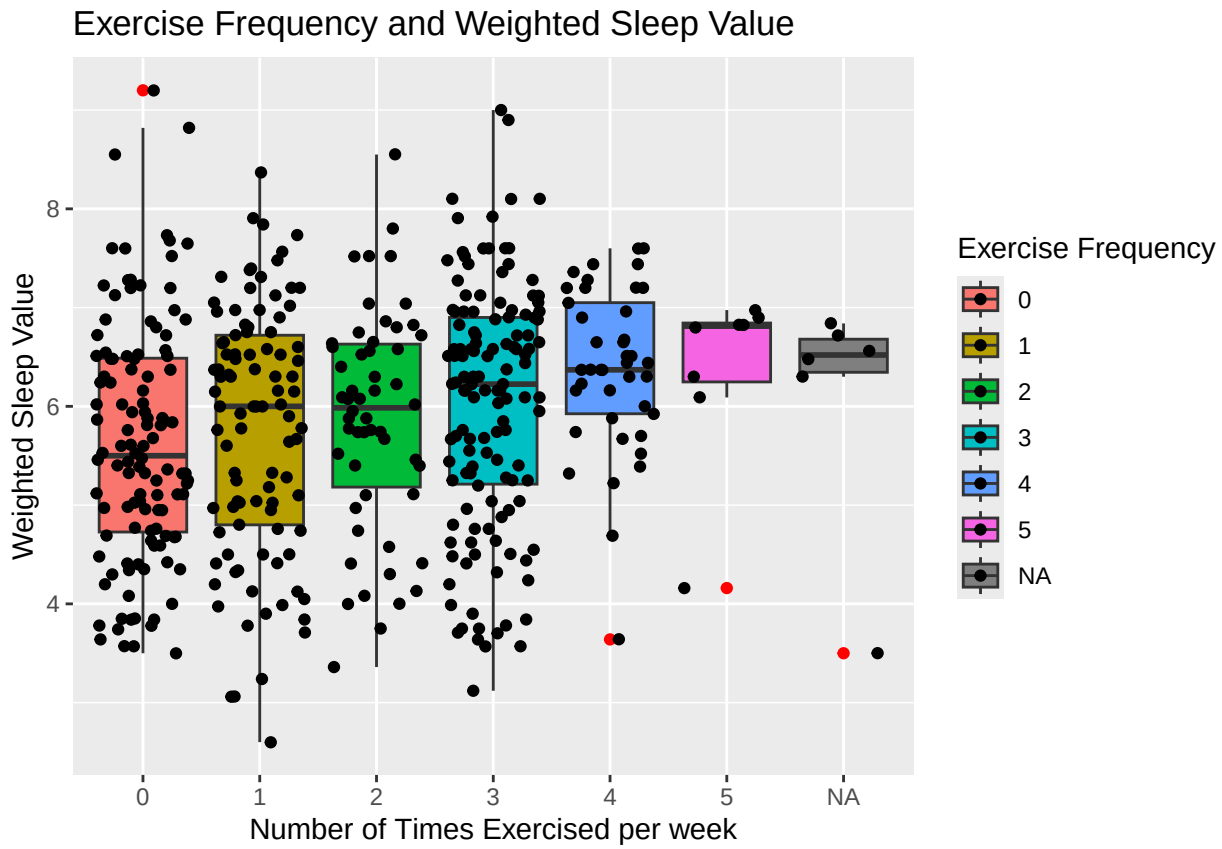

**Smoking Status**

```
ggplot(sleep, aes(x=as.factor(Smoking.status), y=weighted.sleep,
                  fill=as.factor(Smoking.status)))+
  geom_boxplot(outlier.color="red")+geom_jitter()+
  labs(title = "Smoking Status and Weighted Sleep Value",
       x="", y = "Weighted Sleep Value", fill="Smoking Status")
```



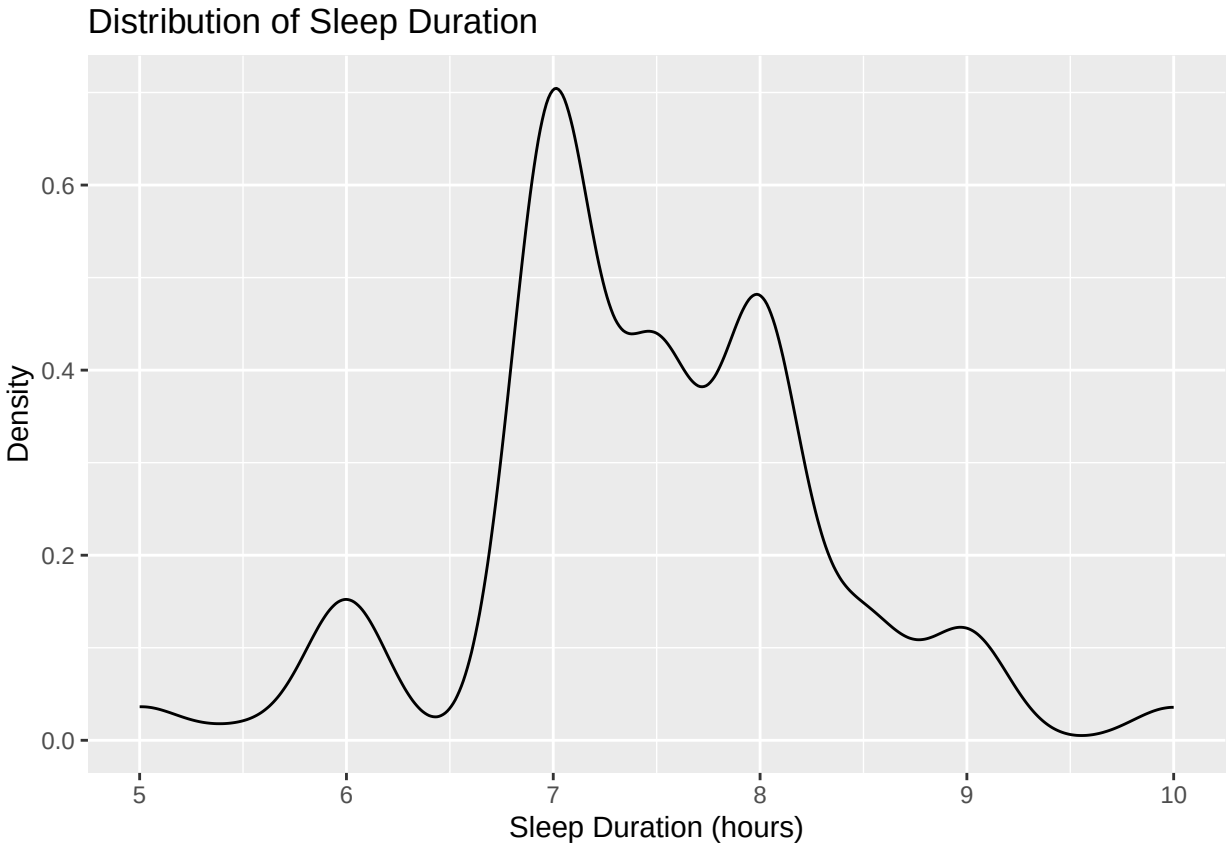

**Exercise Frequency**

```
ggplot(sleep, aes(x=as.factor(Exercise.frequency), y=weighted.sleep,
                  fill=as.factor(Exercise.frequency)))+
  geom_boxplot(outlier.color="red")+geom_jitter()+
  labs(title = "Exercise Frequency and Weighted Sleep Value",
       x="Number of Times Exercised per week",
       y = "Weighted Sleep Value", fill="Exercise Frequency")
```

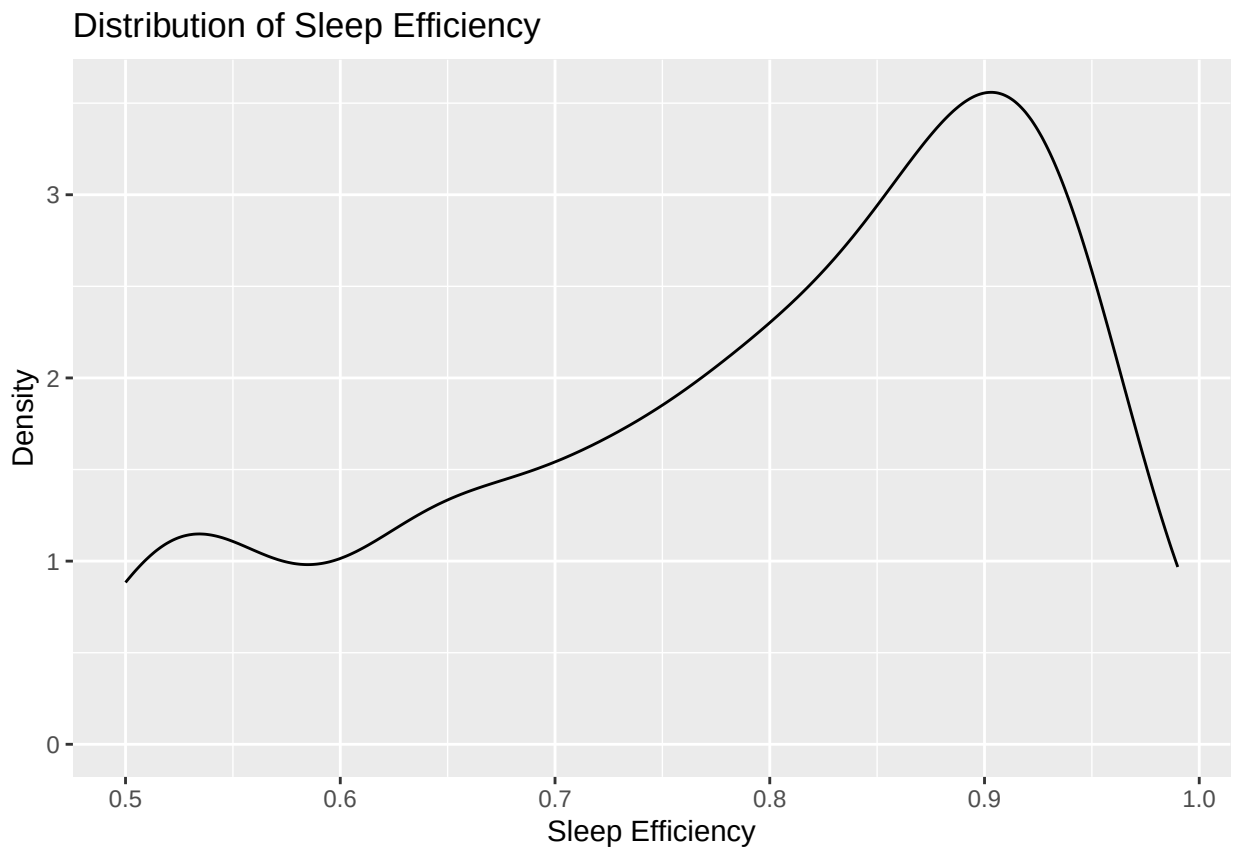



## Distributions

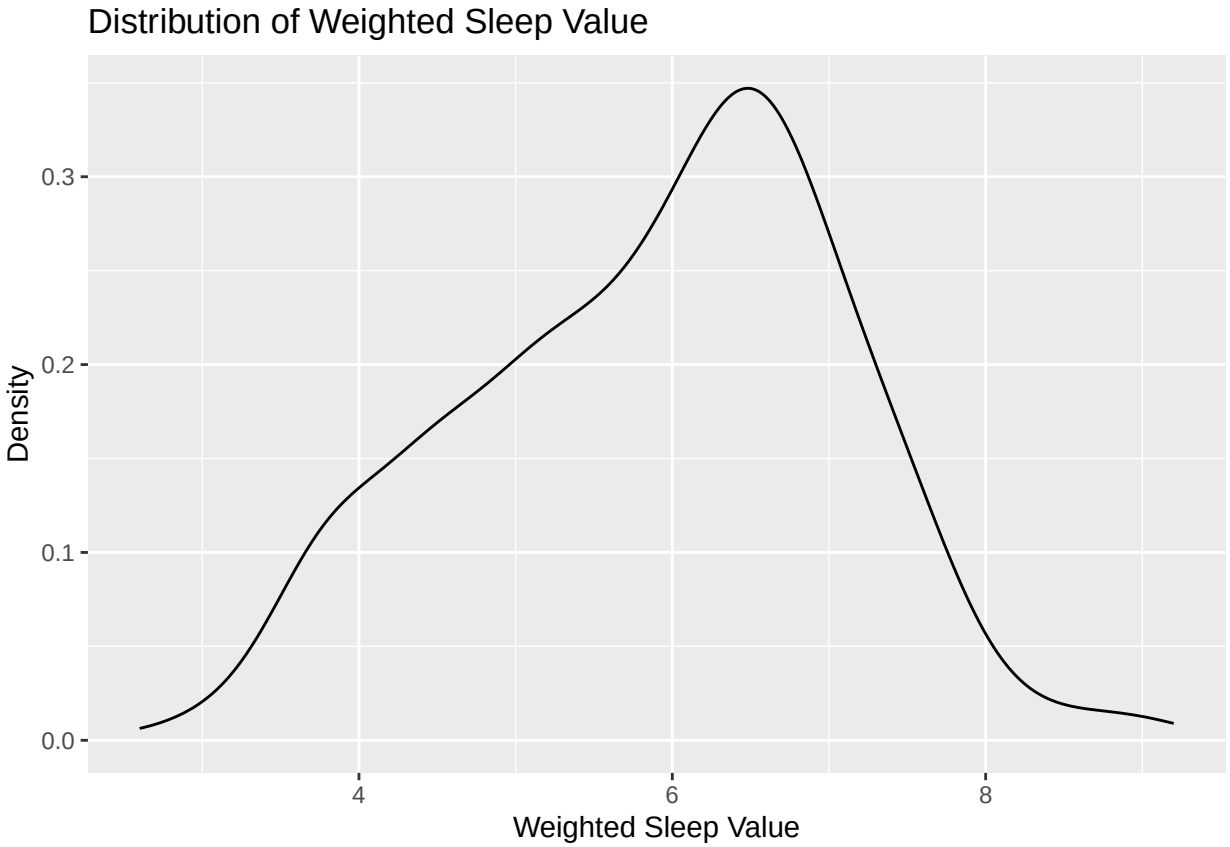
```
ggplot(sleep,aes(x=Sleep.duration))+
  geom_density()+
  labs(title = "Distribution of Sleep Duration",
       x = "Sleep Duration (hours)",
       y = "Density")
```




```
ggplot(sleep,aes(x=Sleep.efficiency))+
  geom_density()+
  labs(title = "Distribution of Sleep Efficiency",
       x = "Sleep Efficiency",
       y = "Density")
```




```
ggplot(sleep,aes(x=weighted.sleep))+
  geom_density()+
  labs(title = "Distribution of Weighted Sleep Value",
       x = "Weighted Sleep Value",
       y = "Density")
```


Distribution of Weighted Sleep Value
Density
0.0
0.1
0.2
0.3
4
6
8
Weighted Sleep Value