STATISTICAL ANALYSIS OF RISK FACTORS ASSOCIATED WITH SLEEP DISORDER

Vidya Sai Rashmitha Reddy Yellareddy

2023-11-02

Project Objective:

The primary objective of this study is to investigate the relationships between key lifestyle factors and the occurrence of sleep disorders. Additionally, the study aims to explore how these lifestyle factors are associated with sleep duration, with a particular focus on variables such as physical activity and stress level.

Hypotheses:

Null Hypothesis (H0): Various factors, including age, gender, occupation, physical activity, and stress level, do not have a statistically significant impact on the presence of sleep disorders.

Alternative Hypothesis (HA): Various factors, including age, gender, occupation, physical activity, and stress levels, have a statistically significant impact on the presence of sleep disorders.

Methods:

Data :
The data for the analysis is obtained from Kaggle's sleep health and lifestyle dataset. The data contains of 374 participants data on their age, gender, occupation, sleep duration, quality of sleep, blood pressure, heart rate, stress level, BMI category, physical activity, daily steps and their sleep disorder status.

```
sleep_data <- read.csv("~/Desktop/Sleep_health_and_lifestyle_dataset.csv",header = TRUE, row.names = 1)
```

loading libraries:

```
##
## 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
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Preprocessing Data:

```
sleep_data_mod <- sleep_data
sleep_data_mod$Gender <- ifelse(sleep_data_mod$Gender == 'Male', 1, 0)
mapping_vector <- c("None" = 0, "Sleep Apnea" = 1, "Insomnia" = 2)
sleep_data_mod$Sleep.Disorder <- mapping_vector[sleep_data_mod$Sleep.Disorder]

data = sleep_data_mod[,c(1,2,4,5,6,7,10,12)]
```

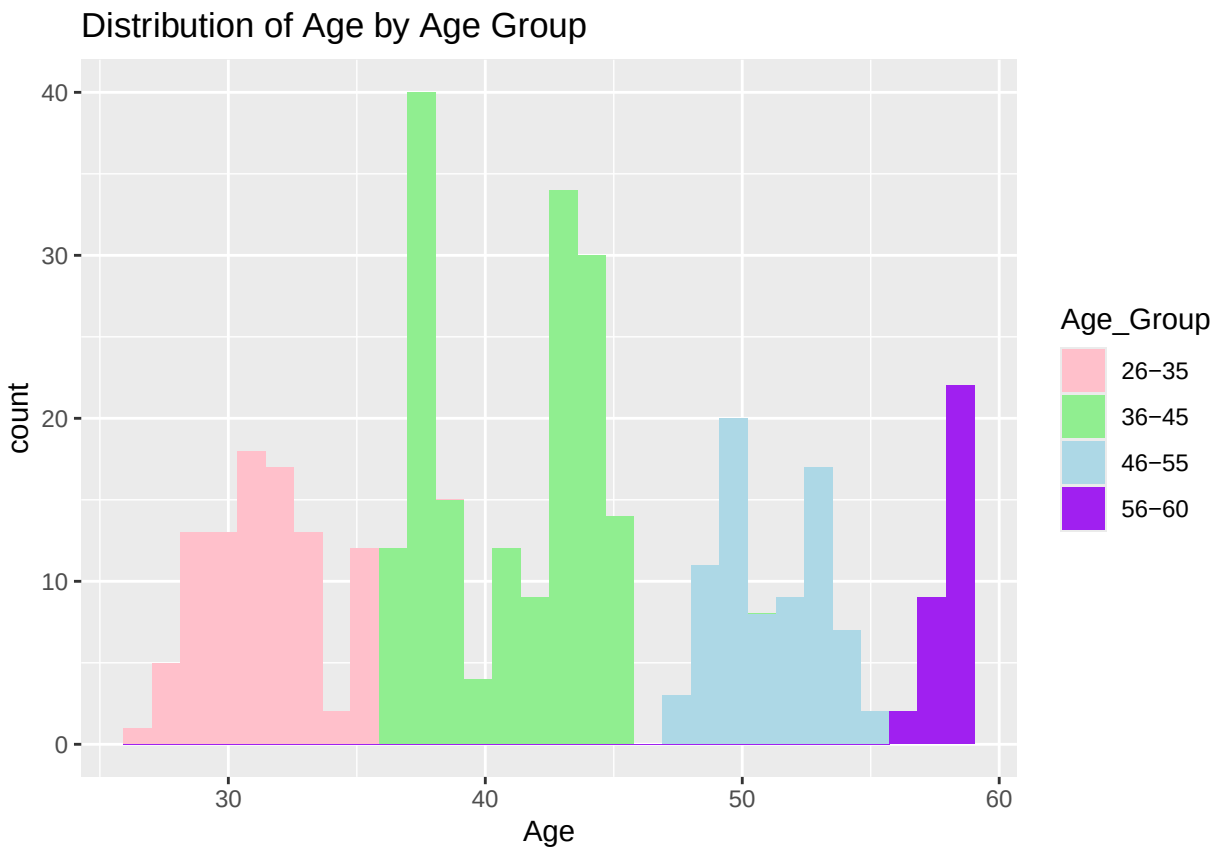
Descriptive Analysis:

```
# Calculate descriptive statistics for numerical variables
sleep_data %>%
  summarize(
    mean = mean(Age),
    median = median(Age),
    sd = sd(Age),
    iqr = IQR(Age)
  ) %>%
  print()
```

```
##       mean median       sd   iqr
## 1 42.18449     43 8.673133 14.75
```

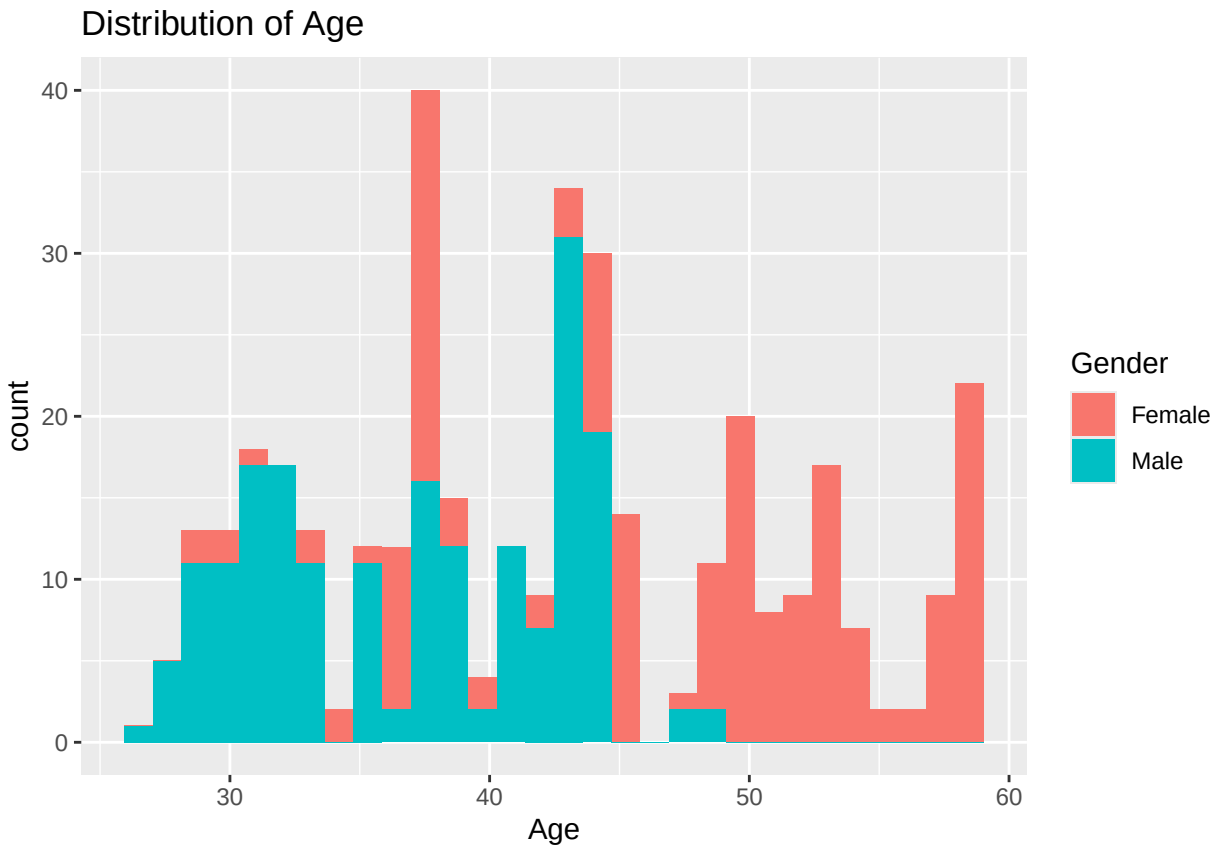
```
# Define age ranges and corresponding colors
age_ranges <- c( "26-35", "36-45", "46-55", "56-60")
age_colors <- c("pink", "lightgreen", "lightblue", "purple")

# Create a new variable that represents age ranges
sleep_data$Age_Group <- cut(sleep_data$Age, breaks = c(26, 35, 45, 55, max(sleep_data$Age)), labels = ag

# Create a histogram with custom colors for different age groups
ggplot(sleep_data, aes(x = Age, fill = Age_Group)) +
  geom_histogram() +
  labs(title = "Distribution of Age by Age Group") +
  scale_fill_manual(values = age_colors)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

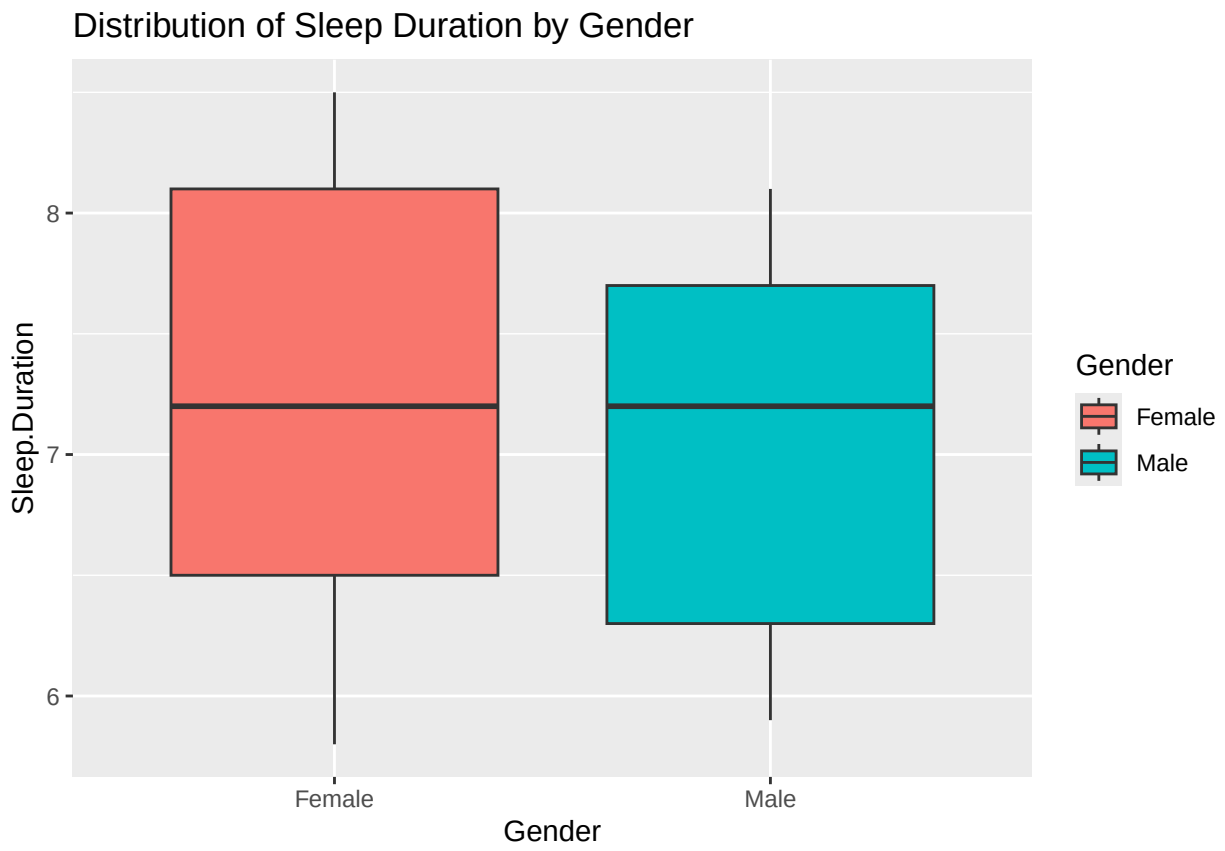



```
# Create a histogram for Age with colors
ggplot(sleep_data, aes(x = Age, fill = Gender)) +
  geom_histogram() +
  labs(title = "Distribution of Age")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



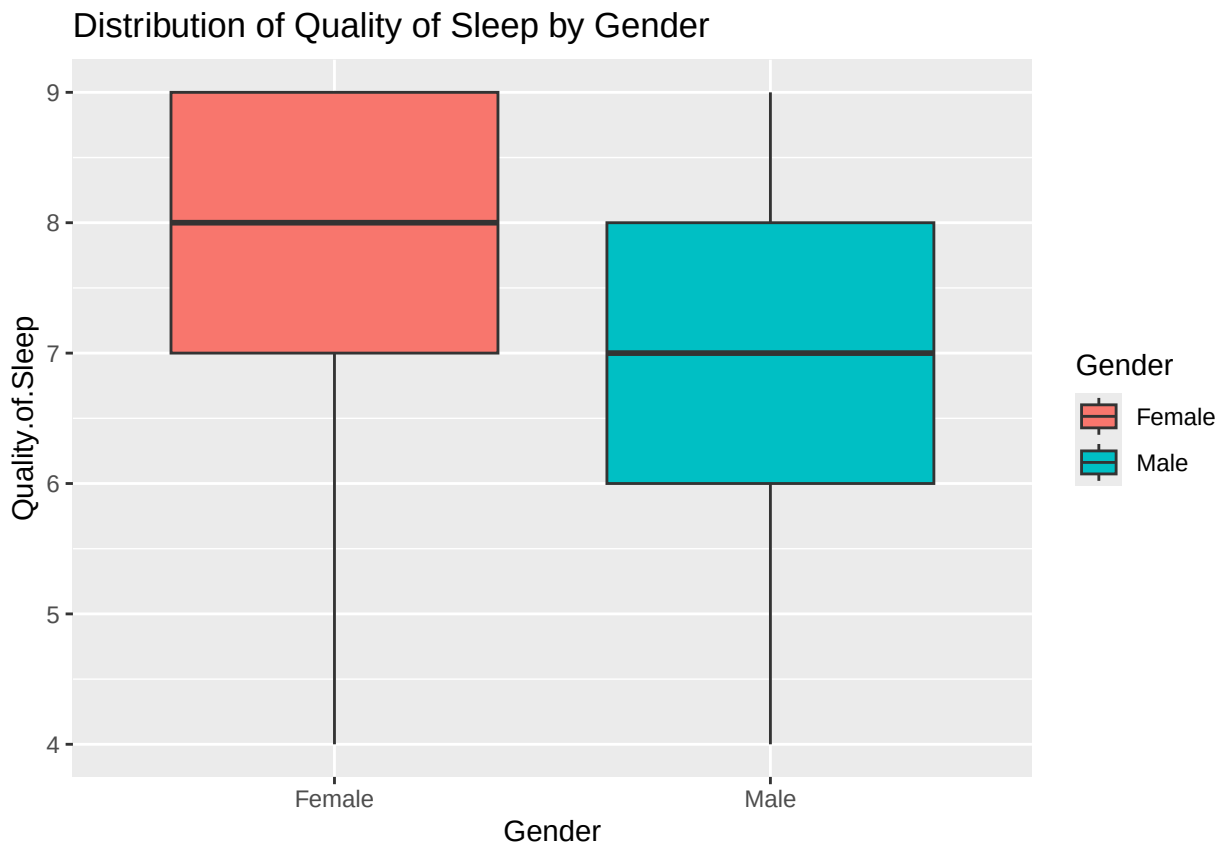

```
# Create boxplots for numerical variables
ggplot(sleep_data, aes(x = Gender, y = Sleep.Duration, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Distribution of Sleep Duration by Gender")
```



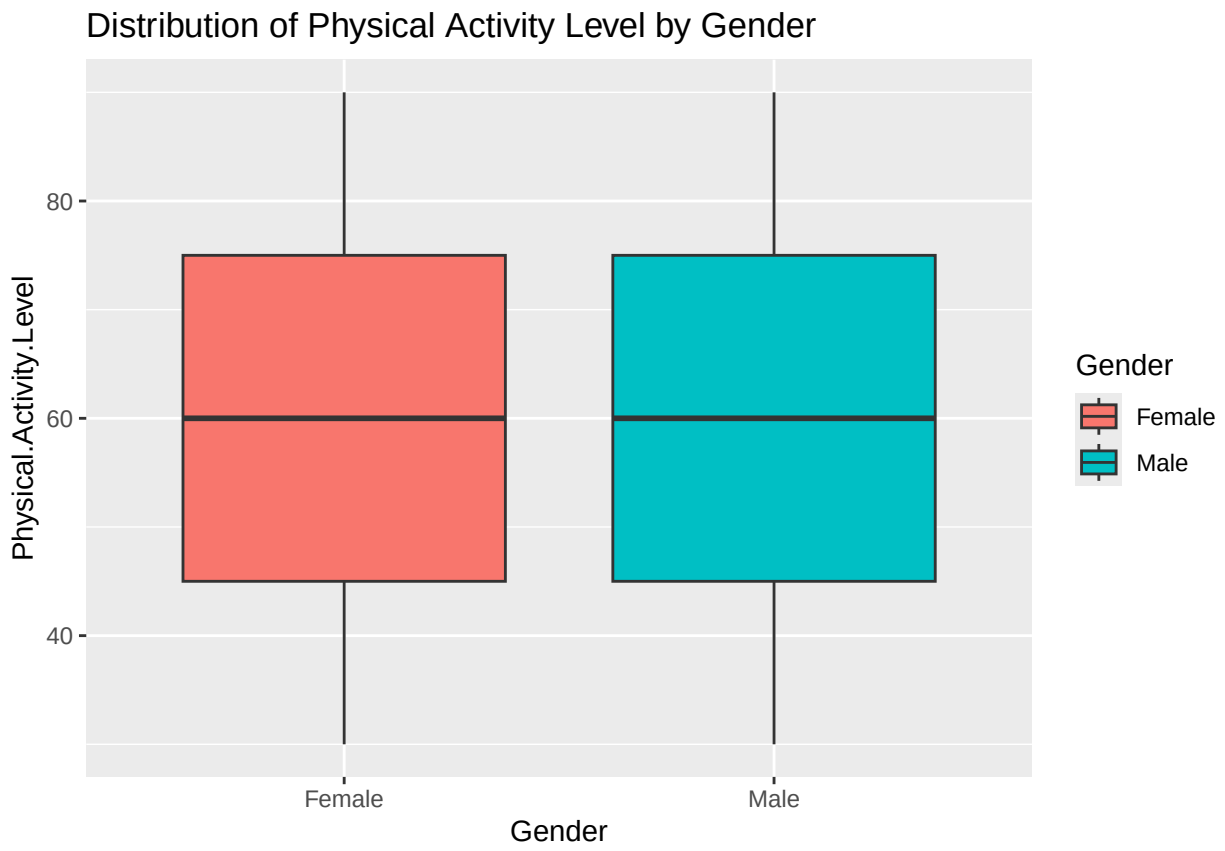

```
# Create frequency tables for categorical variables
table(sleep_data$Sleep.Disorder)
```

```
##
##    Insomnia        None Sleep Apnea
##          77         219          78
```

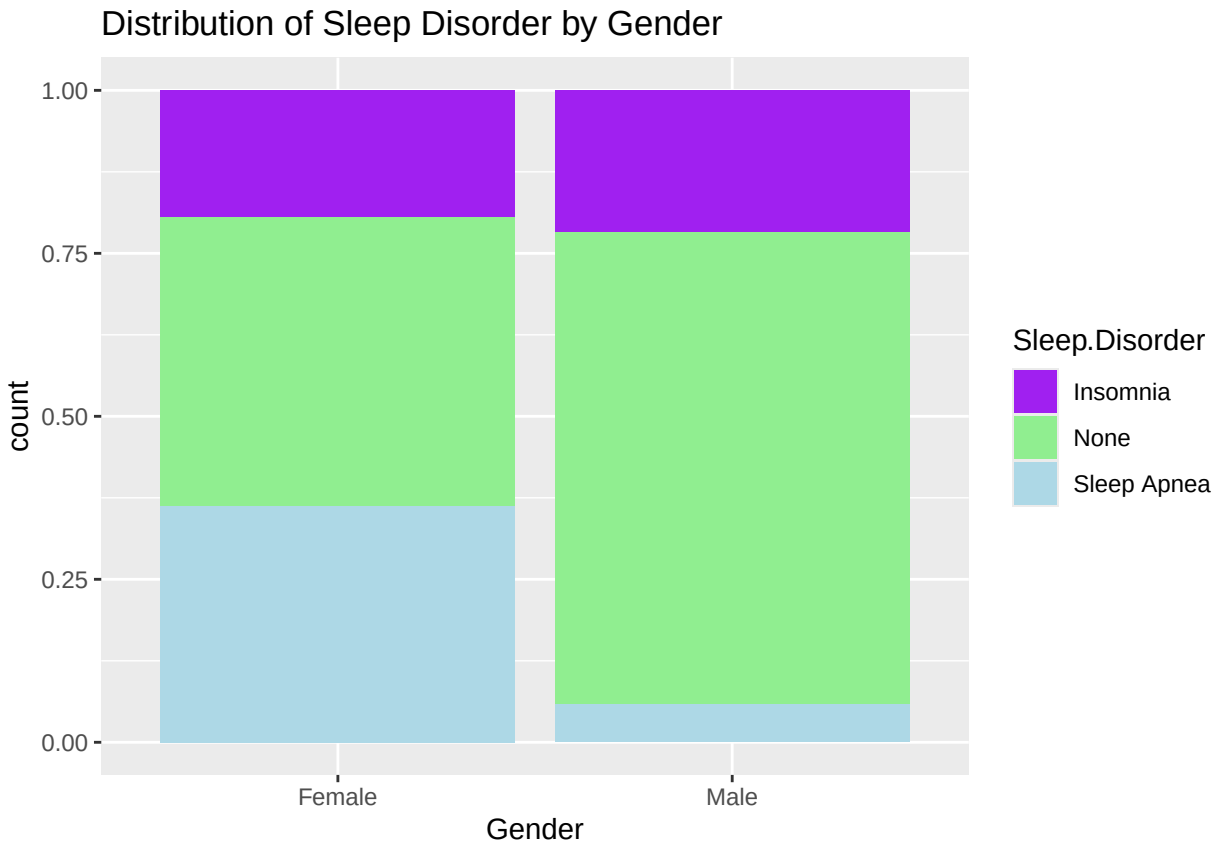
```
# Create a boxplot for Quality of Sleep by Gender
ggplot(sleep_data, aes(x = Gender, y = Quality.of.Sleep, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Distribution of Quality of Sleep by Gender")
```




```
# Create a boxplot for Physical Activity Level by Gender
ggplot(sleep_data, aes(x = Gender, y = Physical.Activity.Level, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Distribution of Physical Activity Level by Gender")
```




```r
# Create a bar plot to visualize the distribution of Sleep Disorder by Gender
ggplot(sleep_data, aes(x = Gender, fill = Sleep.Disorder)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Sleep Disorder by Gender") +
  scale_fill_manual(values = c("None" = "lightgreen", "Insomnia" = "purple", "Sleep Apnea" = "lightblue
```

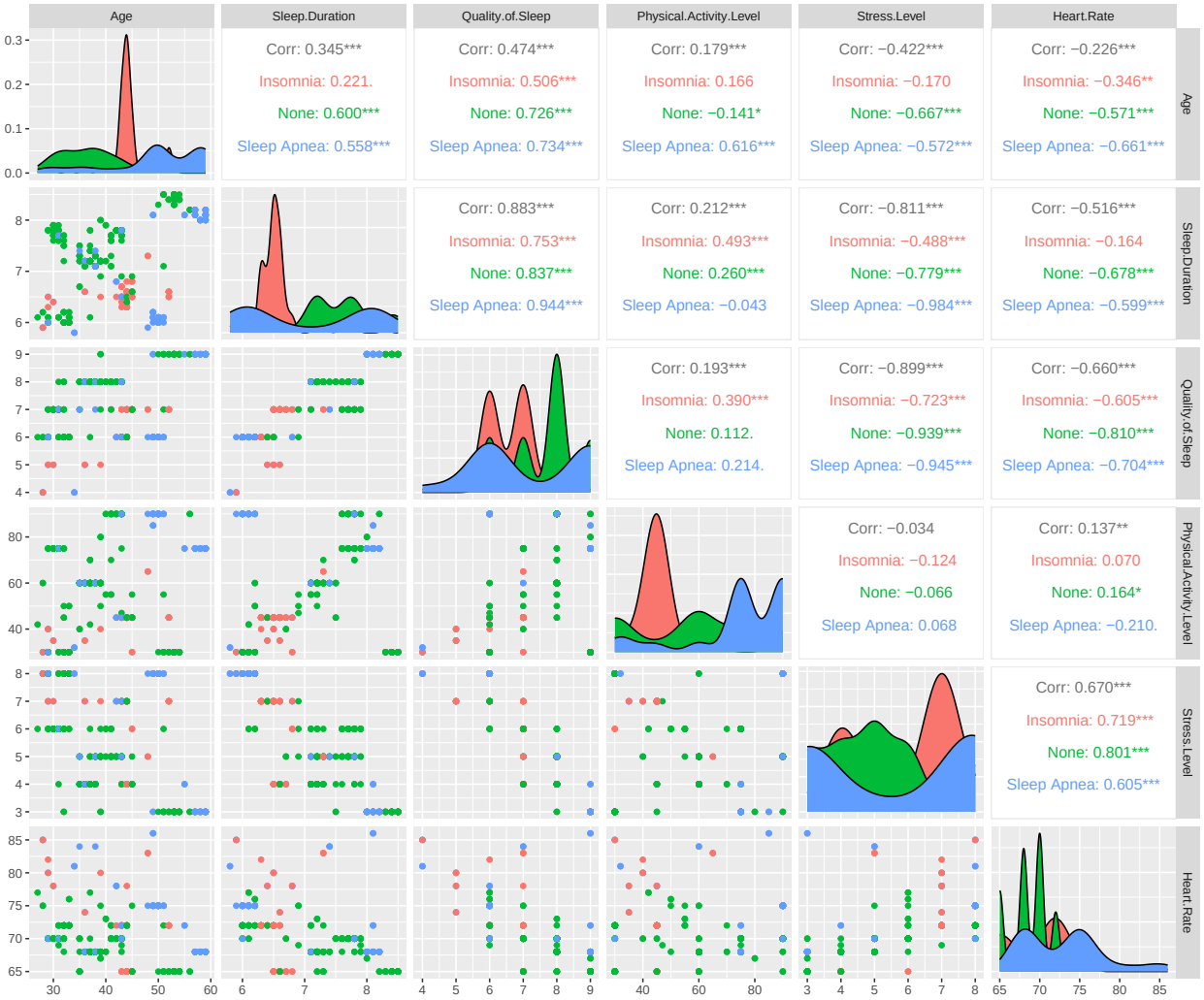



```
# Create a contingency table of "Sleep Disorder" and "Gender"
contingency_table <- table(sleep_data$Sleep.Disorder, sleep_data$Gender)
contingency_table
```

```
##
##               Female Male
##   Insomnia        36   41
##   None            82  137
##   Sleep Apnea     67   11
```

Correlation Analysis:

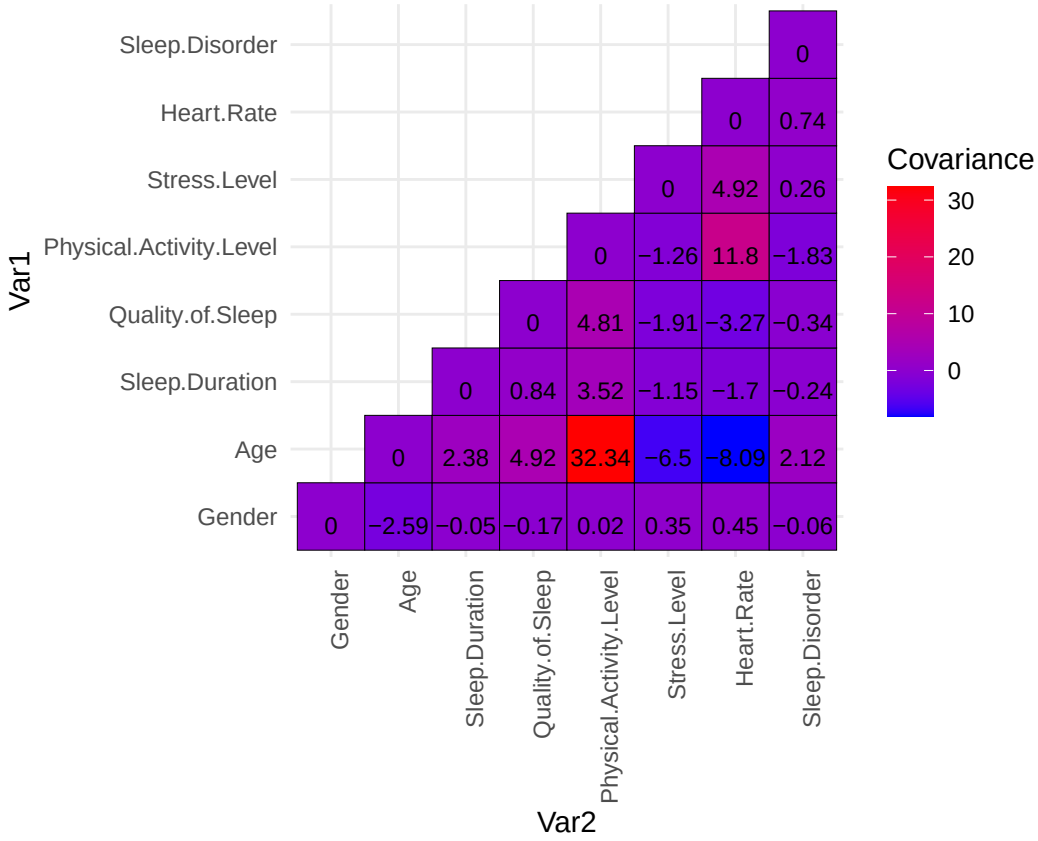
```
correlation <- cor(data)
ggpairs(sleep_data,columns = c(2,4:7,10), aes(color=Sleep.Disorder))
```



Covariance Analysis:

```
covariance <- cov(data)

diag(covariance) <- 0

# Install and load the ggplot2 package if you haven't already
library(ggplot2)
library(reshape2)
library(dplyr)

# Get upper triangle of the correlation matrix
get_upper_tri_no_diag <- function(cormat) {
  # Set the lower triangular part (including the diagonal) to NA
  cormat[lower.tri(cormat)] <- NA
  return(cormat)
}

covariance_melt <- melt(get_upper_tri_no_diag(covariance))

labels <- covariance_melt %>%
```

```
  filter(!is.na(value)) %>%
  mutate(value = round(value, 2))

# Heatmap
ggplot(data = covariance_melt, aes(Var2, Var1)) +
  geom_tile(data = subset(covariance_melt, !is.na(value)), aes(fill = value), color = "black") +
  geom_text(data = labels, aes(label = value), vjust = 1, size = 3) +
  scale_fill_gradientn(colors = c("blue", "red"), na.value = "white", name = "Covariance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  coord_fixed()
```



ANOVA:

```
summary(aov(Sleep.Disorder ~ Gender + Age + Stress.Level + Physical.Activity.Level,sleep_data_mod))
```

```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## Gender                      1   6.29   6.285   12.65 0.000424 ***
## Age                         1  16.23  16.234   32.68 2.25e-08 ***
## Stress.Level                1  28.56  28.560   57.49 2.78e-13 ***
## Physical.Activity.Level     1   7.69   7.692   15.48 9.95e-05 ***
## Residuals                 369 183.31   0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise T-test

```r
gender <- t.test(sleep_data_mod$Gender, sleep_data_mod$Sleep.Disorder, paired = TRUE)

age <- t.test(sleep_data_mod$Age, sleep_data_mod$Sleep.Disorder, paired = TRUE)

physical_activity <- t.test(sleep_data_mod$Physical.Activity.Level, sleep_data_mod$Sleep.Disorder, pair

sleep_duration <- t.test(sleep_data_mod$Sleep.Duration, sleep_data_mod$Sleep.Disorder, paired = TRUE)

gender
```

```
##
##  Paired t-test
##
## data:  sleep_data_mod$Gender and sleep_data_mod$Sleep.Disorder
## t = -2.1912, df = 373, p-value = 0.02905
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.21814642 -0.01180011
## sample estimates:
## mean difference
##      -0.1149733
```

```r
age
```

```
##
##  Paired t-test
##
## data:  sleep_data_mod$Age and sleep_data_mod$Sleep.Disorder
## t = 94.978, df = 373, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  40.70366 42.42468
## sample estimates:
## mean difference
##        41.56417
```

```r
physical_activity
```

```
##
##  Paired t-test
##
## data:  sleep_data_mod$Physical.Activity.Level and sleep_data_mod$Sleep.Disorder
## t = 54.09, df = 373, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  56.42228 60.67933
## sample estimates:
## mean difference
##         58.5508
```

```
sleep_duration
```

```
##
##  Paired t-test
##
## data:  sleep_data_mod$Sleep.Duration and sleep_data_mod$Sleep.Disorder
## t = 94.606, df = 373, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  6.376420 6.647109
## sample estimates:
## mean difference
##        6.511765
```

Chi-square Test of independence

```
gender_table <- table(sleep_data_mod$Gender, sleep_data_mod$Sleep.Disorder)
chisq.test(gender_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gender_table
## X-squared = 54.306, df = 2, p-value = 1.613e-12
```

```
# Linear Regression Model
linear_model <- lm(Sleep.Disorder ~ Gender + Age + Stress.Level + Physical.Activity.Level, data = sleep

# Summary of the Linear Regression Model
summary(linear_model)
```

```
##
## Call:
## lm(formula = Sleep.Disorder ~ Gender + Age + Stress.Level + Physical.Activity.Level,
##     data = sleep_data_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2398 -0.3821 -0.1552  0.2244  1.8900
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.801647   0.310631  -5.800 1.43e-08 ***
## Gender                  -0.040884   0.093447  -0.438    0.662
## Age                      0.045191   0.005556   8.134 6.41e-15 ***
## Stress.Level             0.177325   0.023152   7.659 1.66e-13 ***
## Physical.Activity.Level -0.007075   0.001798  -3.935 9.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 369 degrees of freedom
## Multiple R-squared:  0.2428, Adjusted R-squared:  0.2346
## F-statistic: 29.58 on 4 and 369 DF,  p-value: < 2.2e-16
```