Final Project Report: Analysis of Coffee Consumption and Health Metrics

Inferential Statistics 2025/2026

Andrea Recchiuti and Rafaelle Richel Pearl

2025-12-18

Contents

1 Abstract

(say from 3-8 lines) describing the problem and what it has been done

2 Introduction

motivation about the problem and possibly pointers to the literature

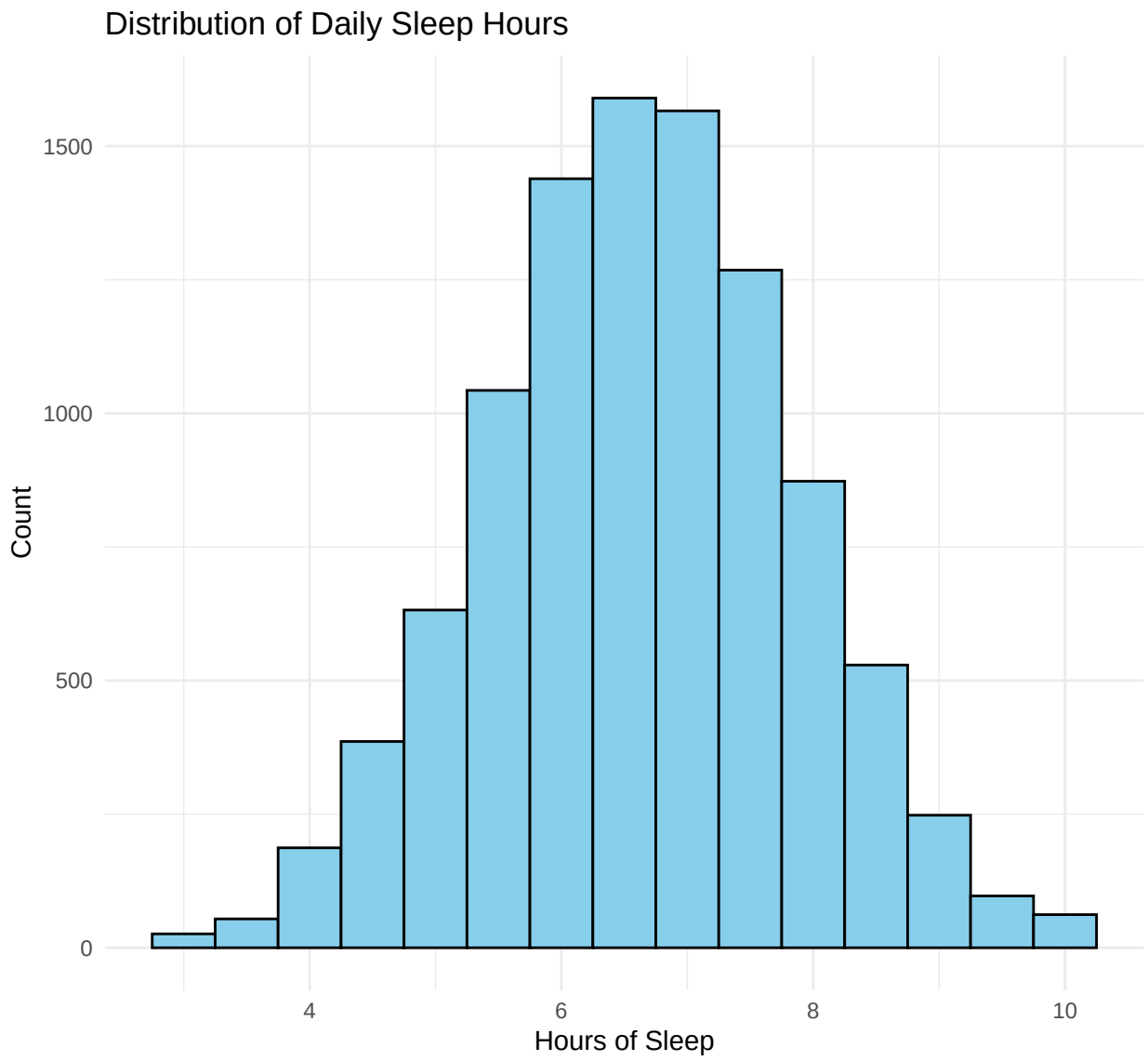
3 Data

Data description and exploration: describes where the data comes from and performs some quick exploratory analysis by means of graphs and summaries

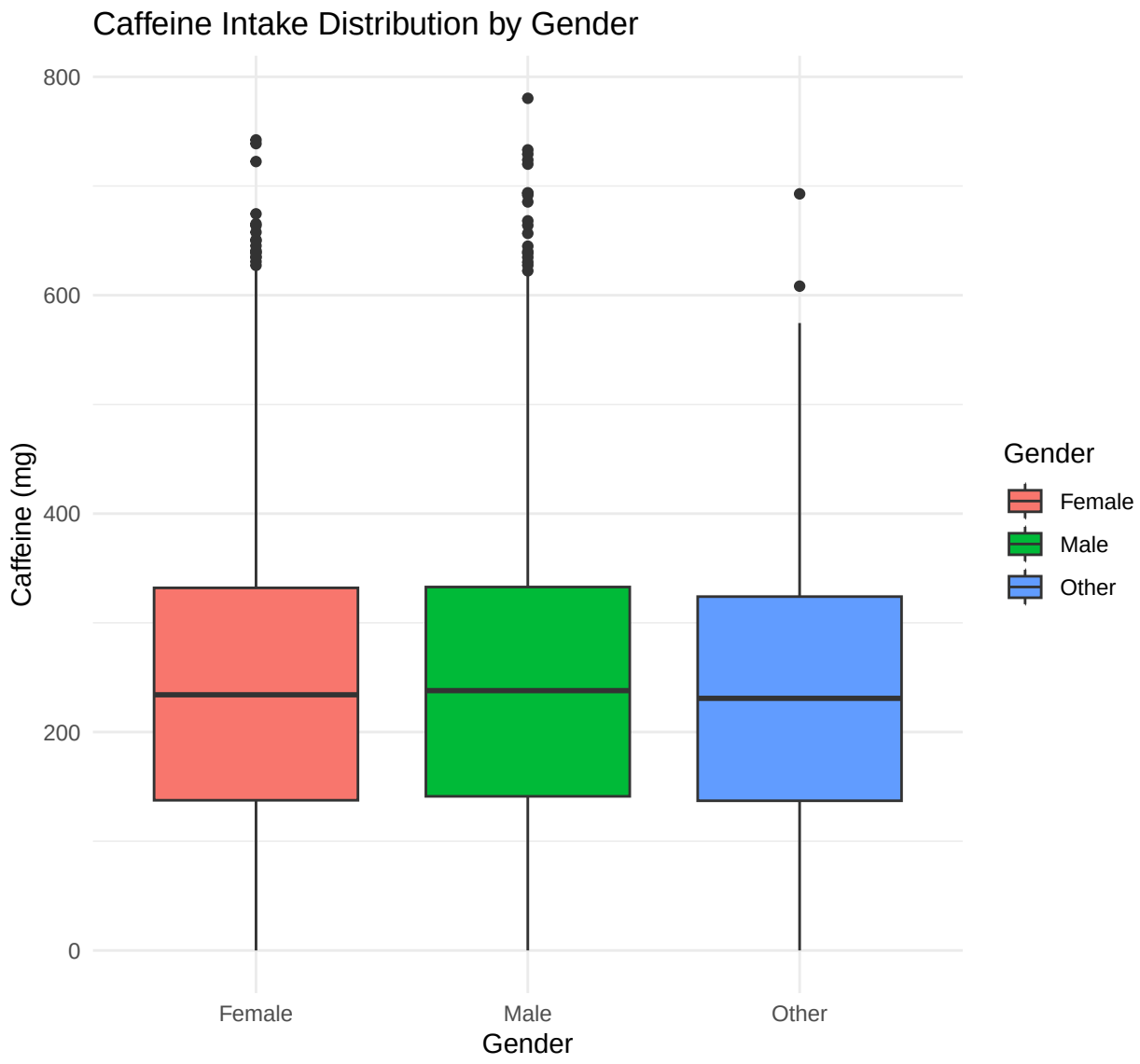
```
library(ggplot2)
coffee_data <- read.csv("synthetic_coffee_health_10000.csv", sep = ",")
summary(coffee_data)
```

```
##        ID             Age           Gender            Country
##  Min.   :    1   Min.   :18.00   Length:10000       Length:10000
##  1st Qu.: 2501   1st Qu.:26.00   Class :character   Class :character
##  Median : 5000   Median :34.00   Mode  :character   Mode  :character
##  Mean   : 5000   Mean   :34.95
##  3rd Qu.: 7500   3rd Qu.:43.00
##  Max.   :10000   Max.   :80.00
##  Coffee_Intake    Caffeine_mg     Sleep_Hours     Sleep_Quality
##  Min.   :0.000   Min.   :  0.0   Min.   : 3.000   Length:10000
##  1st Qu.:1.500   1st Qu.:138.8   1st Qu.: 5.800   Class :character
##  Median :2.500   Median :235.4   Median : 6.600   Mode  :character
##  Mean   :2.509   Mean   :238.4   Mean   : 6.636
##  3rd Qu.:3.500   3rd Qu.:332.0   3rd Qu.: 7.500
##  Max.   :8.200   Max.   :780.3   Max.   :10.000
##       BMI          Heart_Rate     Stress_Level       Physical_Activity_Hours
##  Min.   :15.00   Min.   : 50.00   Length:10000       Min.   : 0.000
##  1st Qu.:21.30   1st Qu.: 64.00   Class :character   1st Qu.: 3.700
##  Median :24.00   Median : 71.00   Mode  :character   Median : 7.500
##  Mean   :23.99   Mean   : 70.62                      Mean   : 7.487
##  3rd Qu.:26.60   3rd Qu.: 77.00                      3rd Qu.:11.200
##  Max.   :38.20   Max.   :109.00                      Max.   :15.000
##  Health_Issues       Occupation           Smoking       Alcohol_Consumption
##  Length:10000       Length:10000       Min.   :0.0000   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :0.0000   Median :0.0000
##                                        Mean   :0.2004   Mean   :0.3007
##                                        3rd Qu.:0.0000   3rd Qu.:1.0000
##                                        Max.   :1.0000   Max.   :1.0000
```

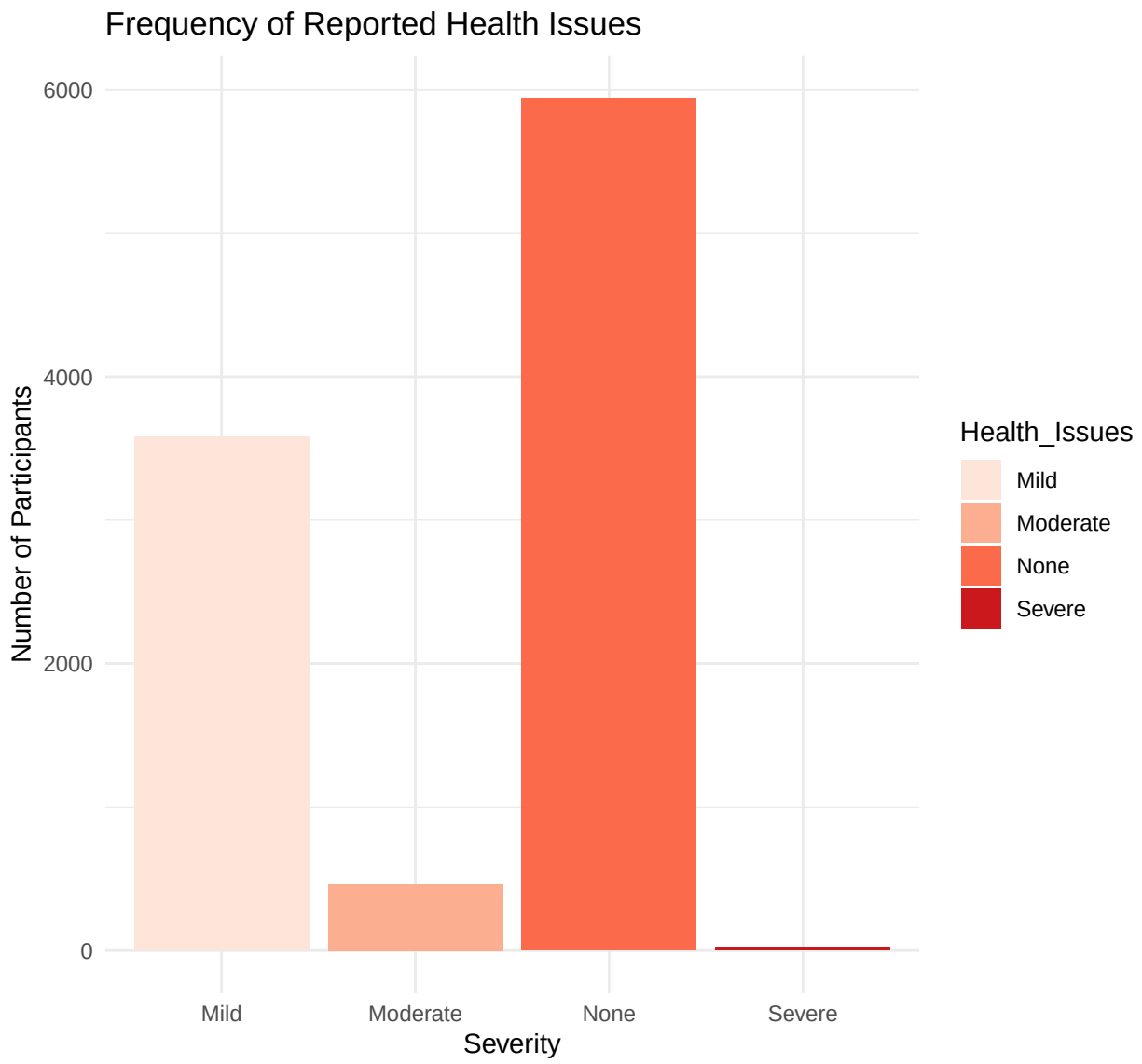
```r
# hist for sleep hours
ggplot(coffee_data, aes(x = Sleep_Hours)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Daily Sleep Hours", x = "Hours of Sleep", y = "Count")
```



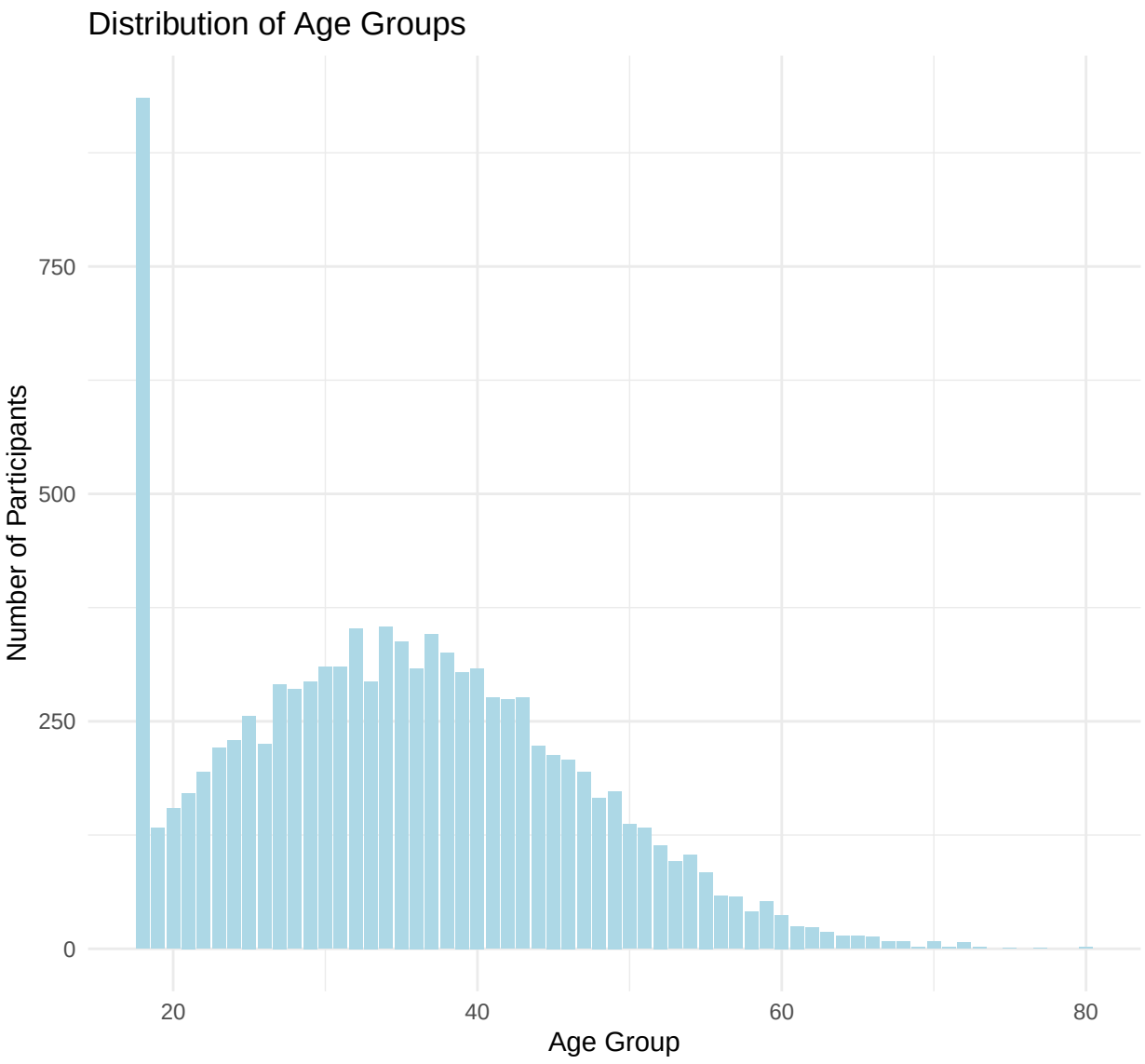

```r
# boxplot of caffeine intake per gender
ggplot(coffee_data, aes(x = Gender, y = Caffeine_mg, fill = Gender)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Caffeine Intake Distribution by Gender", y = "Caffeine (mg)")
```




```
# barchart health issues
ggplot(coffee_data, aes(x = Health_Issues, fill = Health_Issues)) +
  geom_bar() +
  scale_fill_brewer(palette = "Reds") +
  theme_minimal() +
  labs(title = "Frequency of Reported Health Issues", x = "Severity", y = "Number of Participants")
```




```r
# age group histogram
ggplot(coffee_data, aes(x = Age)) +
  geom_bar(fill = "lightblue") +
  theme_minimal() +
  labs(title = "Distribution of Age Groups", x = "Age Group", y = "Number of Participants")
```



4 Statistical Methods

statistical methods applied and the results obtained

4.0.1 Linear Regression

```
model <- lm(Heart_Rate ~ Caffeine_mg + Sleep_Hours + BMI + Physical_Activity_Hours, data = coffee_data)

summary(model)
```

```
##
## Call:
## lm(formula = Heart_Rate ~ Caffeine_mg + Sleep_Hours + BMI + Physical_Activity_Hours,
```

```
##     data = coffee_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.601  -6.808  -0.074   6.634  36.885
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              71.633281   0.871232  82.221  < 2e-16 ***
## Caffeine_mg               0.003927   0.000725   5.417 6.21e-08 ***
## Sleep_Hours              -0.206524   0.081729  -2.527   0.0115 *
## BMI                      -0.021770   0.025099  -0.867   0.3858
## Physical_Activity_Hours -0.007891   0.022722  -0.347   0.7284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.804 on 9995 degrees of freedom
## Multiple R-squared:  0.004328,   Adjusted R-squared:  0.003929
## F-statistic: 10.86 on 4 and 9995 DF,  p-value: 8.738e-09
```

Caffeine / day -> categorical variable (>400, <400)

5 Results

Conclusions