Personal Sleep Analysis

Nathan Vernon

December 17, 2025

Project Overview

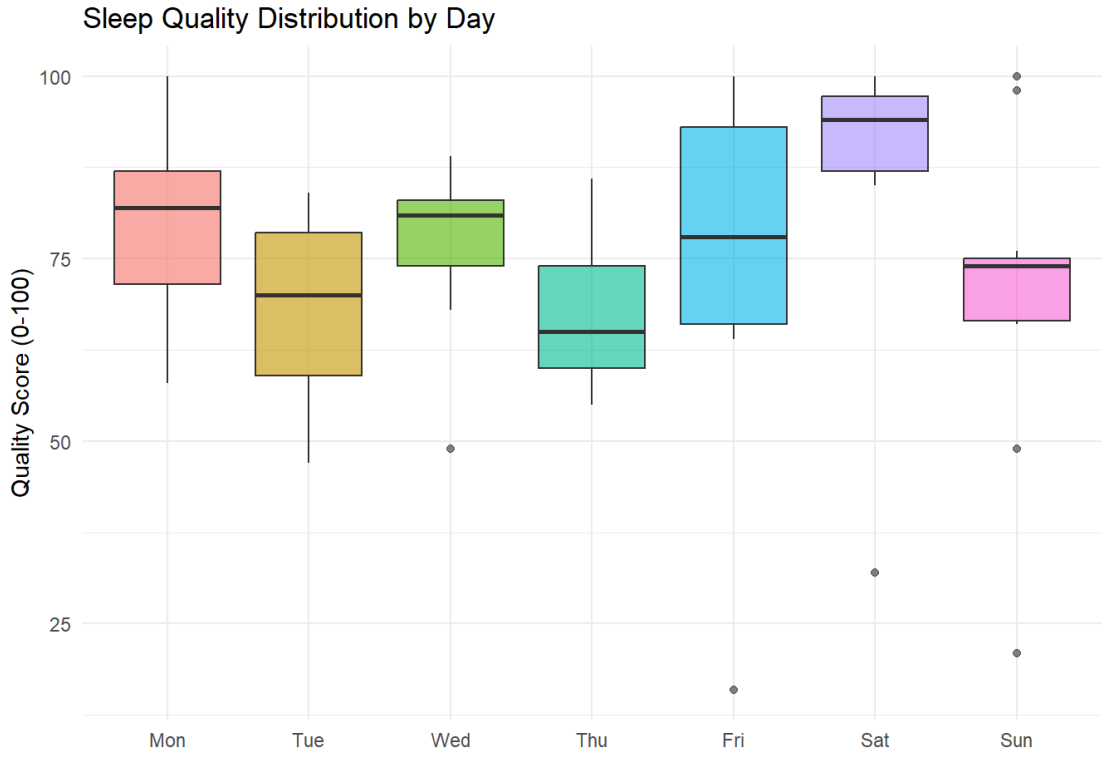
I have tracked my sleep data over approximately 70 nights using an iPhone App. The goal of this analysis is to identify factors that contribute to higher sleep quality and deeper rest and also attempt to reverse engineer the formula the app “SleepCycle” uses to determine their Quality Score

1. Exploratory Analysis: The “Weekend Effect”

First, I wanted to see if my sleep quality varies significantly by the day of the week. The day represents the night of sleep. For example, going to bed Monday and waking up Tuesday is considered Monday. I accounted for going to bed past midnight by subtracting 6 hours from every night so that any bedtime before 6 AM was listed as the correct night.

```
# Reorder days so they start on Monday instead of alphabetical
sleep$day_of_week <- factor(sleep$day_of_week,
                            levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))

ggplot(sleep, aes(x = day_of_week, y = sleep_quality, fill = day_of_week)) +
  geom_boxplot(alpha = 0.6) +
  theme_minimal() +
  labs(title = "Sleep Quality Distribution by Day",
       y = "Quality Score (0-100)",
       x = "") +
  theme(legend.position = "none")
```

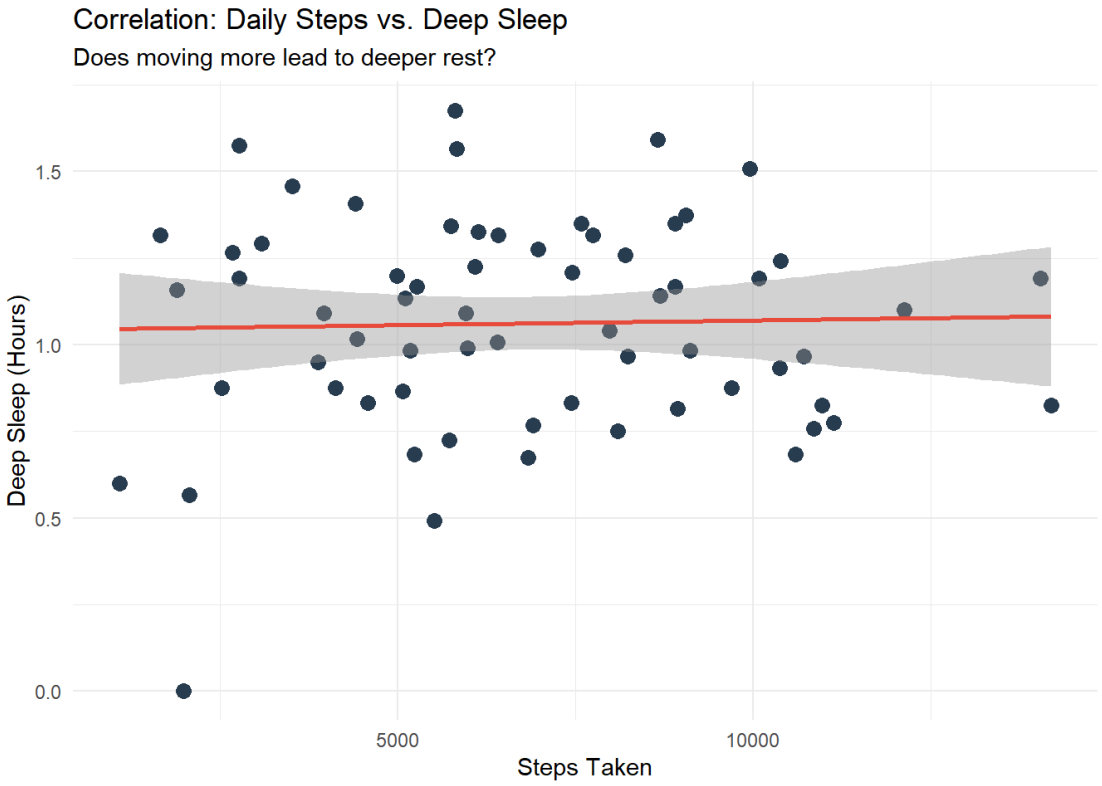


Observation: This makes sense and follows a normal students sleeping habits. During the week, I get less hours of sleep having to wake up for classes. But on the weekends, I have more flexibility to sleep later and longer.

2. Hypothesis Testing: Does Walking More Improve Deep Sleep?

My hypothesis is that physical activity (Steps) correlates with more recovery (Deep Sleep).

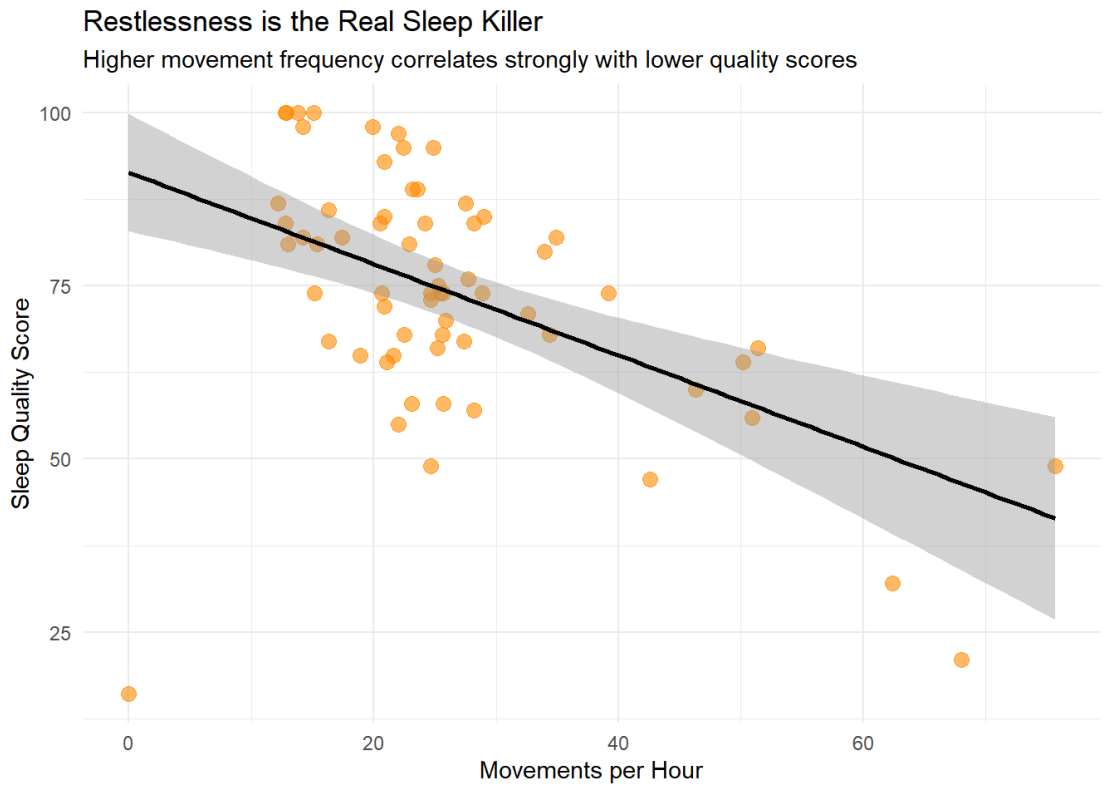
```
ggplot(sleep, aes(x = steps, y = deep_sleep_hours)) +
  geom_point(color = "#2c3e50", size = 3) +
  geom_smooth(method = "lm", color = "#e74c3c") +
  theme_minimal() +
  labs(title = "Correlation: Daily Steps vs. Deep Sleep",
       subtitle = "Does moving more lead to deeper rest?",
       x = "Steps Taken",
       y = "Deep Sleep (Hours)")
```



Observation: There seems to be no relation between steps taken and the hours of deep sleep I get. Steps taken may not be the best way to track movement and perhaps using other activity metrics, there could a correlation.

3.Hypothesis Testing: Does “Restlessness” really mean Restless?

```
ggplot(sleep, aes(x = movements_per_hour, y = sleep_quality)) +
  geom_point(color = "darkorange", alpha = 0.6, size = 3) +
  geom_smooth(method = "lm", color = "black") +
  theme_minimal() +
  labs(title = "Restlessness is the Real Sleep Killer",
       subtitle = "Higher movement frequency correlates strongly with lower quality scores",
       x = "Movements per Hour",
       y = "Sleep Quality Score")
```



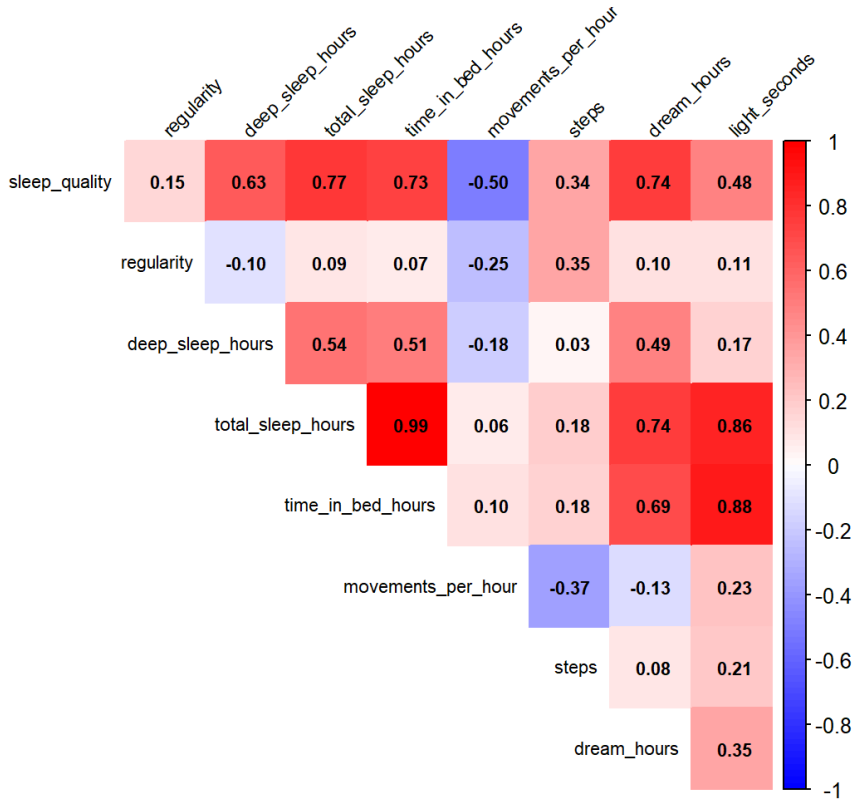
Observation: Movements per hour show a medium negative correlation to sleep quality. This makes sense as tossing and turning often indicates poorer and lighter sleep.

4. Correlation Heatmap

This Heatmap will help me determine what variables I should use to determine sleep quality.

```
# Select only the numeric columns to analyze
cor_data <- sleep |>
  select(sleep_quality, regularity, deep_sleep_hours, total_sleep_hours, time_in_bed_hours,movements_per_hour, steps,
dream_hours, light_seconds) |>
  cor() # Calculate the correlation matrix

corrplot(cor_data,
         method = "color",
         type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(600),
         addCoef.col = "black", # Add the numbers
         tl.col = "black",      # Text label color
         # Make plot readable
         number.cex = .7,
         tl.cex = .7,
         tl.srt = 45,
         diag = FALSE)
```



Observation This plot shows that deep sleep hours, total sleep hours time in bed, and dream hours all strongly positively correlate to sleep quality. Steps has a medium negative correlation as shown earlier. However, using the heatmap, we can see that total sleep hours also highly correlate to time in bed and dream hours. This leads me to believe that these are multicollinear.

5. Checking Multicollinearity

Based on the heatmap, I want to check to make sure that these variable are multicollinear before I leave them out of the model.

```
#build a linear model with all strongly correlated variables
model <- lm(sleep_quality ~ total_sleep_hours + deep_sleep_hours + movements_per_hour + dream_hours + time_in_bed_hou
rs,
            data = sleep)

#Check for Multicollinearity
vif(model)
```

```
##  total_sleep_hours   deep_sleep_hours movements_per_hour        dream_hours
##          102.804712           1.552526           1.169918           3.192185
##  time_in_bed_hours
##           87.371569
```

Observation By looking at the VIF scores for each variable, total sleep hours and time in bed hours are very multicollinear. Because of the nature of the exploration, I will go forward using total sleep hours.

6. Statistical Model

I am running a Linear Regression to reverse engineer the formula the app uses to determine the Quality Score. I am using the variables that gave strong correlations to Sleep Quality in the above Correlation Matrix and not multicollinear.

```
quality_model <- lm(sleep_quality ~ total_sleep_hours + deep_sleep_hours + movements_per_hour + dream_hours,
                    data = sleep)

summary(quality_model)
```

```
##
## Call:
## lm(formula = sleep_quality ~ total_sleep_hours + deep_sleep_hours +
##     movements_per_hour + dream_hours, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3750  -2.5375   0.4759   2.7773  10.9976
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        19.03263    3.49617   5.444 1.02e-06 ***
## total_sleep_hours   8.07407    0.77982  10.354 5.64e-15 ***
## deep_sleep_hours    8.26799    2.53281   3.264  0.00181 **
## movements_per_hour -0.65104    0.04983 -13.064  < 2e-16 ***
## dream_hours         5.40145    1.87059   2.888  0.00539 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.016 on 60 degrees of freedom
## Multiple R-squared:  0.9242, Adjusted R-squared:  0.9192
## F-statistic:   183 on 4 and 60 DF,  p-value: < 2.2e-16
```

7. Checking for Multicollinearity

```
vif(quality_model)
```

```
##  total_sleep_hours   deep_sleep_hours movements_per_hour        dream_hours
##           2.706532           1.515091           1.136676           2.440659
```

Observation There is no multicollinearity in the model

Conclusion

Based on the above model, total sleep hours, deep sleep hours, movements per hours, and dream hours account for 92% of the variation in the sleep quality that SleepCycle gives. Given this, we can create a formula that can predict the quality score based off of those 4 variables and the intercept.the formula is as follows:

$$\text{Quality} = 19.03 + 8.07(\text{total\_sleep\_hours}) + 8.27(\text{deep\_sleep\_hours}) - 0.65(\text{movements\_per\_hour}) + 5.4(\text{dream\_hours})$$

This provides insights into what the app values and how it scores. To begin, everyone starts with a score of around 19 just for falling asleep. Then for ever hours you get, you add on 8 points. Deep sleep is worth double so every hour of deep sleep is worth a total of 16 points.Another bonus is added for dreaming which is worth 5.4 points per hour. Then for every movement you lose around half a point. With these 4 variables, we can explain 92% of the variance in the app given Sleep Quality Score.