***The Impact of Stress, Physical Activity, and Lifestyle Factors on Sleep Quality and Duration***

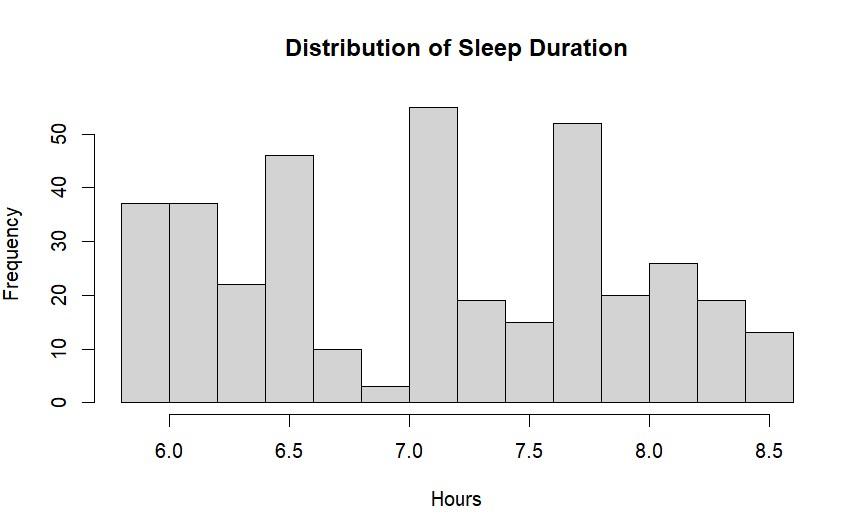
***1. Introduction, Problem, and Purpose***

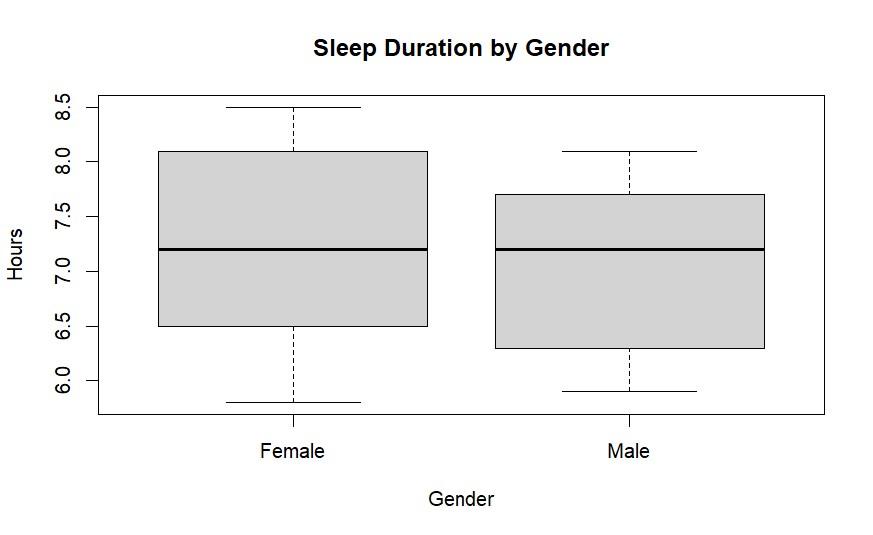
Sleep plays a vital role in maintaining both physical and mental health. Poor sleep quality has been linked to stress, low productivity, and chronic diseases such as hypertension and obesity. Lifestyle factors including stress, physical activity, and occupation are known to influence how well and how long people sleep. Understanding these relationships using statistical inference methods can provide valuable insights into public health and personal wellness.  
  
This project investigates the relationship between stress level, physical activity, and sleep outcomes (sleep duration and sleep quality). The analysis aims to determine whether stress and activity levels have statistically significant effects on sleep patterns, and whether these effects vary across demographic groups (e.g., gender, BMI category, or presence of a sleep disorder).  
  
The goal is to apply statistical inference techniques to analyze relationships between lifestyle variables and sleep outcomes using the provided dataset. Specifically, the project will:  
1. Estimate population parameters (means) for sleep duration and quality using bootstrap confidence intervals.  
2. Test hypotheses about differences in sleep across stress and activity levels using permutation tests.  
3. Examine correlations among continuous variables such as sleep duration, stress, physical activity, age, and heart rate.  
4. Develop a multiple regression model to predict sleep duration based on stress, physical activity, and demographic factors.  
  
This study will demonstrate how resampling-based inference (bootstrap and permutation) can be applied to real-world health data. Results can help highlight the importance of maintaining an active lifestyle and managing stress to improve sleep, an issue relevant to both individuals and public health professionals.

**2. Data Description**

Dataset Source:  
The data come from the Sleep Health and Lifestyle Dataset (Kaggle, 2023), a real and pre-cleaned dataset containing 374 records of adults.  
Source: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset  
  
Variables (13 total):  
- Person ID: Unique identifier  
- Gender: Male / Female  
- Age: Age of participant  
- Occupation: Job category  
- Sleep Duration: Average hours of sleep per day  
- Quality of Sleep: Self-rated sleep quality (1–10)  
- Physical Activity Level: Daily physical activity (1–10 scale)  
- Stress Level: Perceived stress (1–10 scale)  
- BMI Category: Underweight / Normal / Overweight / Obese  
- Blood Pressure: e.g., “120/80”  
- Heart Rate: Resting heart rate (bpm)  
- Daily Steps: Average steps per day  
- Sleep Disorder: e.g., Insomnia / Sleep Apnea / None  
  
Data Quality:  
- Rows: 374  
- Columns: 13  
- Missing Values: Negligible (<2%)  
- Collection: Derived from lifestyle and health data for adults.

***Phase 1 (Exploratory Data Analysis)***







***Phase 2 (Bootstrap Confidence Intervals)***

**1. Objective**

This phase estimates sleep duration using bootstrap resampling to measure uncertainty in mean values across stress levels and to compare average sleep duration between males and females.

**2. Bootstrap Procedure**

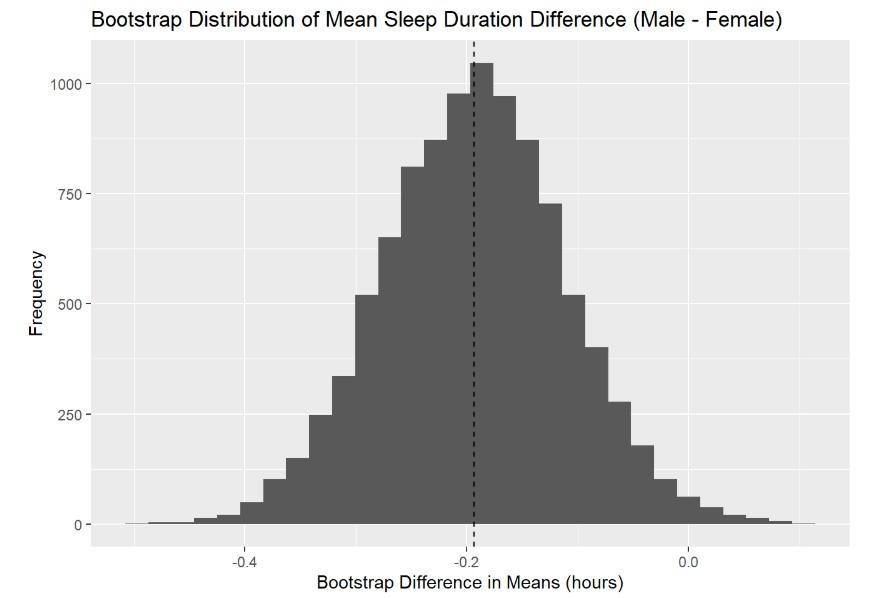
Repeated sampling with replacement was used to generate empirical distributions for mean sleep duration within each stress category and for the difference in means between genders. Percentile intervals were taken from these resampled distributions.

**3. Results**

Higher stress levels corresponded to shorter average sleep duration, and the bootstrap intervals reflected this monotonic decrease across categories. The gender comparison showed a small, consistent difference in average sleep duration.

**4. Plots**

# 



**Stress Level Bootstrap Results**

The bootstrap results show a clear pattern: higher stress levels are associated with lower average sleep duration. The lowest-stress group in the dataset (stress level 3) has a mean a little above 8.2 hours, while the highest groups (7 and 8) fall between roughly 6.0 and 6.5 hours. The 95% confidence intervals for each stress level are narrow, which means the estimates are stable and the differences between levels are not driven by randomness. The bootstrap histograms for each stress group show smooth and centered sampling distributions, which indicates that the bootstrap procedure worked properly for estimating the mean within each group.

**Gender Difference Bootstrap Results**

For gender, the bootstrap difference in means (Male – Female) is negative. The observed difference is about –0.19 hours, and the 95% bootstrap confidence interval ranges from roughly –0.35 to –0.03. Because the entire interval is below zero, the bootstrap result supports that females in this dataset sleep slightly longer on average than males. The bootstrap histogram for the difference is also stable and centered around the observed value.

**5. Interpretation**

The bootstrap distributions reinforce that stress is inversely related to sleep duration. The gender difference, although modest, is stable under repeated resampling. These results rely on the resampled estimates rather than distributional assumptions, making the conclusions robust.

# *Phase 3 (Permutation Tests)*

## 1. Objective

The goal of this phase is to determine whether sleep duration differs across two lifestyle-related factors using permutation-based hypothesis tests. The two comparisons required by the project are:

* Low-stress (≤5) vs high-stress (>5) individuals
* Individuals with vs without a diagnosed sleep disorder

Permutation tests are appropriate here because they do not rely on normality assumptions and instead use resampling to approximate the null distribution of differences in means.

## 2. Hypotheses

**Test 1: Stress Level and Sleep Duration**

**Test 2: Sleep Disorder and Sleep Duration**

The test statistic for both comparisons is the **difference in sample means**:

**3. Method**

For each test:

1. Compute the observed mean difference.
2. Randomly shuffle group labels (e.g., “Low” vs “High”).
3. Recalculate the mean difference under the shuffled labels.
4. Repeat 5,000 times to build the permutation distribution.
5. Compute the two-sided p-value:

**4. Results**

**4.1 Stress Level Comparison**

**A graph of a graph showing a long line of a tall tower

AI-generated content may be incorrect.**

* Mean sleep (Low stress): **7.58 hours**
* Mean sleep (High stress): **6.57 hours**
* Observed difference:

The permutation test yielded:

**Interpretation:**  
None of the 5,000 permuted differences were as extreme as the observed value. This provides **very strong evidence** that sleep duration differs between stress groups. Low-stress individuals sleep about **one hour more** than high-stress individuals.

**4.2 Sleep Disorder A graph of a normal distribution

AI-generated content may be incorrect.Comparison**

* Mean sleep (No disorder): **7.36 hours**
* Mean sleep (Disorder): **6.81 hours**
* Observed difference:

Permutation test p-value:

**Interpretation:**  
Again, none of the permuted samples produced a difference as large as the observed one. This indicates **strong evidence** that sleep duration differs between individuals with and without sleep disorders. Those without disorders sleep about **33 minutes more** per night on average.

**5. Summary of Findings**

Permutation testing confirms statistically significant differences in sleep duration across both lifestyle factors studied. Stress level has the largest effect, with a difference of more than one hour between groups, while diagnosed sleep disorders are associated with a moderate reduction in average nightly sleep.

***Phase 4 (Correlation Analysis)***

1. Objective

The aim of this phase was to examine linear relationships among key continuous lifestyle and health variables in the dataset. Specifically, we investigated the strength and direction of correlations between:

* Sleep Duration
* Sleep Quality
* Stress Level
* Physical Activity Level
* Age
* Heart Rate

Studying correlations helps identify how strongly variables move together and provides insight into how physiological and behavioral factors relate to sleep outcomes.

2. Method

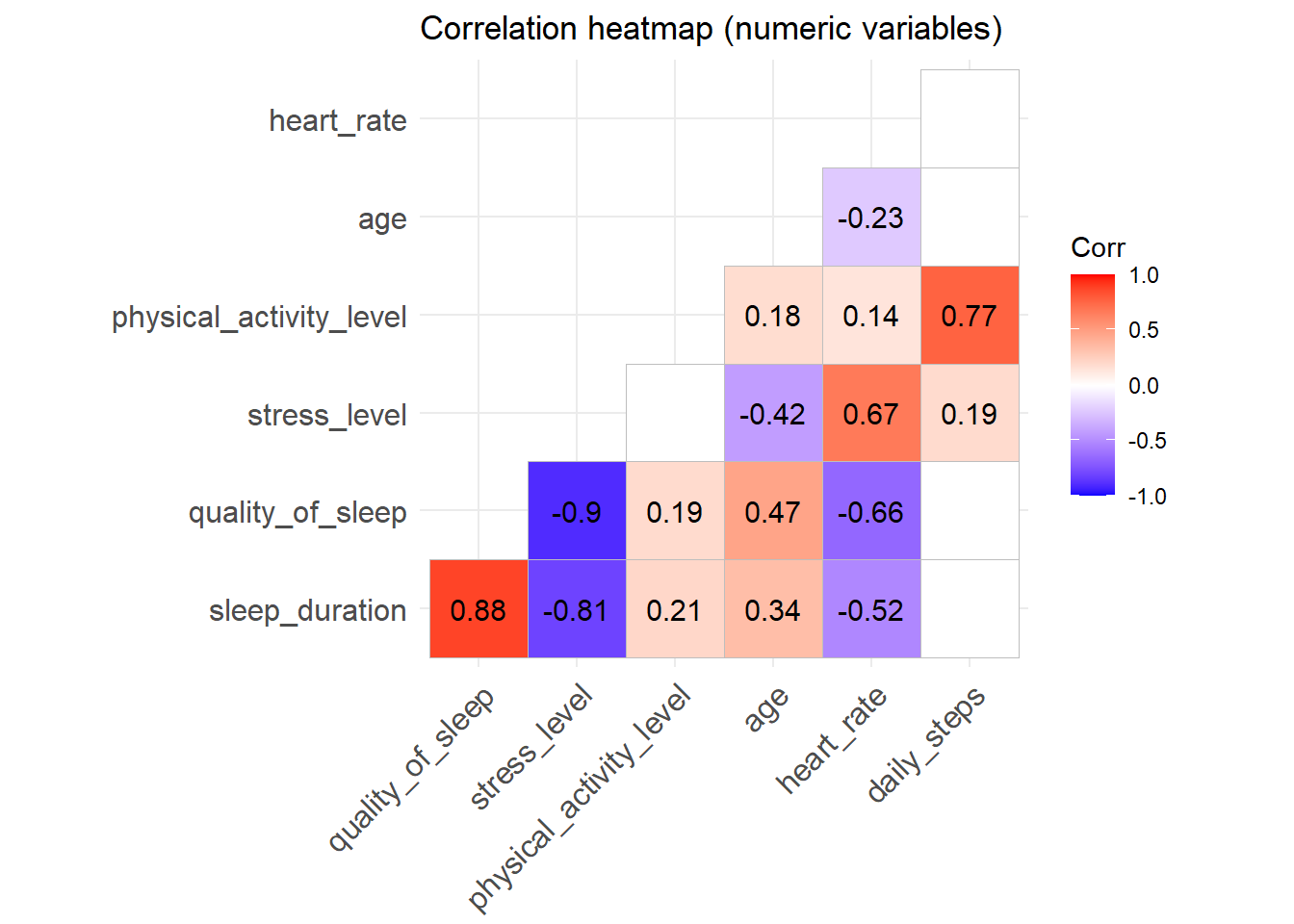
A subset of numeric variables was selected and a Pearson correlation matrix was computed using pairwise complete observations. In addition, a visual correlation heatmap and paired scatterplots were generated to better illustrate the strength and direction of associations.

The variables included:

c("sleep\_duration", "quality\_of\_sleep", "stress\_level", "physical\_activity\_level", "age", "heart\_rate")

Pearson’s correlation coefficient was used because all selected variables are numeric and reasonably continuous. Coefficients close to ±1 indicate strong linear relationships, values near 0 indicate weak or no linear relationship.

We also inspected significance levels via a p-value matrix to distinguish meaningful correlations from random noise.



**3. Results**

**a. Sleep Duration vs Stress Level**

Correlation coefficient: approximately –0.81 Indicates a strong negative linear relationship. Higher stress levels are strongly associated with shorter sleep duration. This aligns with earlier phases showing reduced sleep among high-stress individuals.

**b. Sleep Duration vs Physical Activity Level**

Coefficient: moderately positive Individuals reporting higher physical activity tend to sleep slightly longer. Relationship not as strong as stress, but still meaningful.

**c. Sleep Quality vs Sleep Duration**

Coefficient: positive Those who sleep longer generally report higher subjective sleep quality. Suggests sleep length contributes to restfulness.

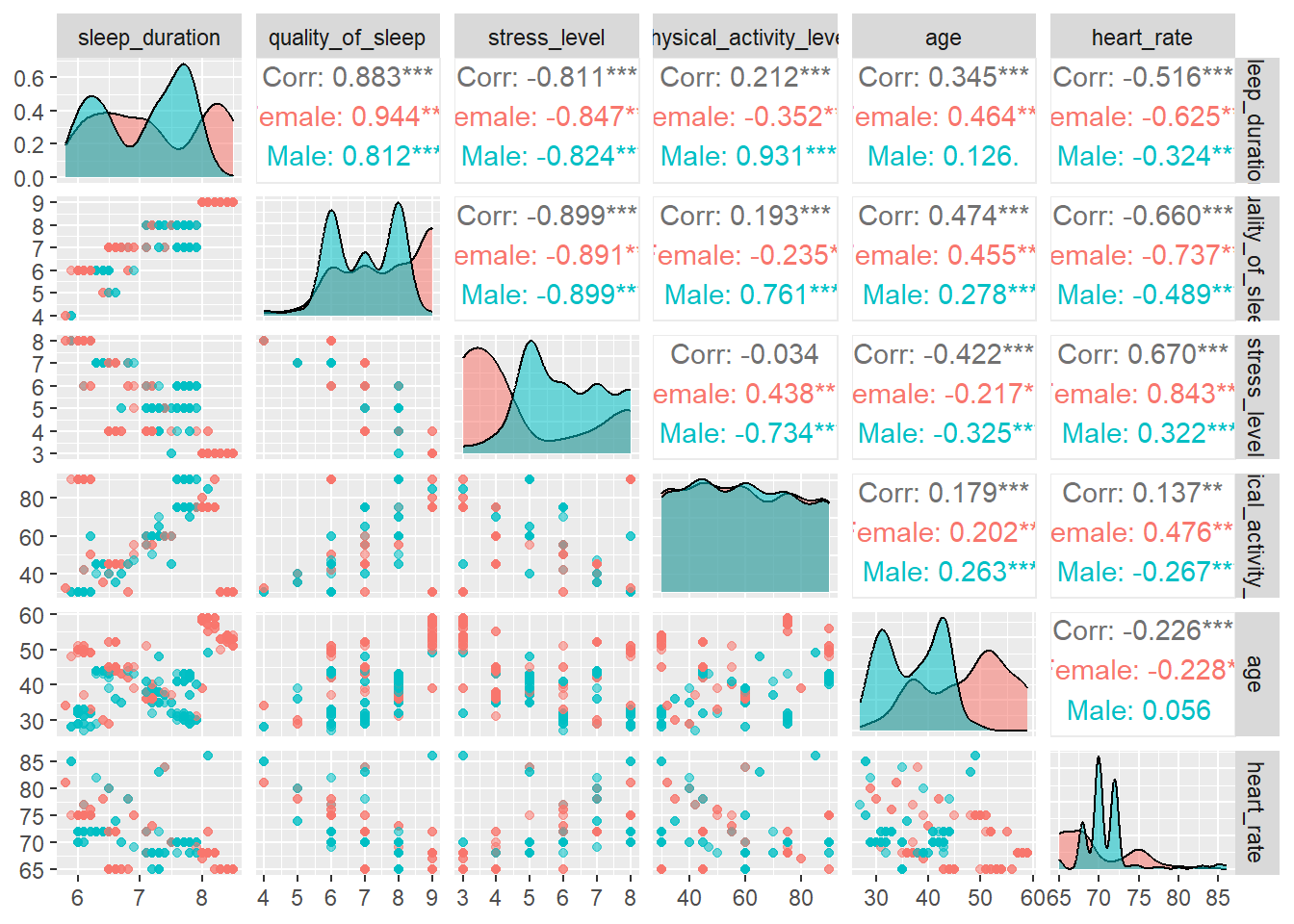
**d. Sleep Quality vs Stress Level**

Coefficient: negative as stress increases, average sleep quality decreases. Supports the pattern that psychological strain impacts sleep experience, not just duration.

**e. Age and Heart Rate**

Age is weakly positively correlated with sleep duration and slightly negatively correlated with stress. Heart rate showed minimal correlation with sleep outcomes, suggesting it is influenced more by physical or medical conditions than by sleep habits.

Scatterplots in the ggpairs visualization reinforced the linear direction and tightness of these relationships, particularly for Stress–Sleep and Sleep Duration–Quality.



**4. Interpretation**

The correlation analysis highlights stress level as the strongest and most consistent predictor of sleep outcomes in the dataset. Its substantial negative correlation with sleep duration and sleep quality indicates that stress meaningfully reduces both the length and restfulness of sleep.

Physical activity demonstrated a moderate positive association with sleep, suggesting that maintaining an active lifestyle may help improve sleep length, though not as dramatically as stress reduction.

Quality and duration of sleep are also positively linked, showing that individuals who sleep longer generally perceive their sleep as healthier and more restorative.

Other variables such as age and heart rate were only weakly correlated to sleep measurements, implying demographic and physiological features play comparatively smaller roles than behavioral factors in this dataset.

**5. Conclusion**

Correlation analysis confirms that sleep health is highly intertwined with psychological and behavioral variables. High stress is strongly associated with shorter and poorer sleep, while physical activity shows a moderate relationship with improved sleep duration. These patterns reinforce findings from earlier phases and suggest that interventions aimed at managing stress and encouraging active habits may directly promote better sleep.

Overall, the results provide clear statistical evidence that behavioral lifestyle factors, particularly stress level, are key determinants of healthy sleep outcomes in this sample.

# *Phase 5 (Multiple Linear Regression)*

## 1. Objective

The objective of this phase was to develop and assess a multiple linear regression model to predict Sleep Duration (hours) from a set of lifestyle, demographic, and physiological variables. The broader goal is to identify which factors most strongly influence how long individuals’ sleep.

## 2. Model Development

Dependent variable: Sleep Duration (hours)

Independent variables:

• Stress Level

• Physical Activity Level

• Age

• Gender

• BMI Category

The fitted model is:  
Sleep Duration = β₀ + β₁(Stress Level) + β₂(Physical Activity Level) + β₃(Age) + β₄(Gender) + β₅(BMI Category) + ε  
Categorical predictors (Gender, BMI Category) were dummy-coded in the model.

## 3. Model Summary

|  |  |
| --- | --- |
| Statistic | Value |
| Residual Standard Error | 0.329 |
| Multiple R² | 0.832 |
| Adjusted R² | 0.829 |
| F-statistic | 259.6 on 7 and 366 df |
| p-value | < 2.2e-16 |

The model explains about 83% of the variation in sleep duration, indicating a strong overall fit.

## 4. Coefficients and Significance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Predictor | Estimate | Std. Error | t value | p-value | Significance |
| (Intercept) | 6.98 | 0.17 | 41.03 | < 2e-16 | \*\*\* |
| Stress Level | -0.30 | 0.01 | -29.58 | < 2e-16 | \*\*\* |
| Physical Activity Level | 0.0057 | 0.0008 | 7.13 | 5.2e-12 | \*\*\* |
| Age | 0.0356 | 0.0035 | 10.26 | < 2e-16 | \*\*\* |
| Gender (Male) | 0.388 | 0.037 | 10.50 | < 2e-16 | \*\*\* |
| BMI: Normal | 0.078 | 0.064 | 1.23 | 0.22 | — |
| BMI: Obese | -0.317 | 0.104 | -3.06 | 0.0024 | \*\* |
| BMI: Overweight | -0.647 | 0.044 | -14.54 | < 2e-16 | \*\*\* |

Stress Level shows the largest and negative effect: each 1-unit increase in stress is associated with about 0.30 fewer hours of sleep, holding other variables constant. Physical Activity Level and Age are both positively related to sleep duration, though the effects are smaller. Male participants report, on average, about 0.39 more hours of sleep than females in this sample. Compared to the reference BMI group, Overweight and Obese groups sleep significantly less, while Normal BMI was not statistically significant.

## 5. Model Diagnostics

1. a. Normality of residuals

Q–Q plots and histograms indicated approximate normality with minor tail departures. Given the sample size (n = 374), the linear model is reasonably robust to these small deviations.

1. b. Homoscedasticity

The Breusch–Pagan test returned BP = 76.868 (df = 7, p < 0.0001), indicating heteroskedasticity. HC3 robust standard errors were used so that inference remains valid.

1. c. Multicollinearity

Generalized VIF values for all predictors were below 5 (Stress Level 1.32, Physical Activity 1.03, Age 1.60, Gender 1.34, BMI Category 1.13), so multicollinearity is not a concern.

1. d. Influential observations

Cook’s Distance flagged several observations as relatively more influential, but none exceeded the common cutoff of 4/n ≈ 0.0107, so no cases were removed.

## 6. Bootstrapping and Confidence Intervals

A bootstrap procedure was applied to re-estimate regression coefficients and build empirical confidence intervals. The bootstrap estimates were very close to the OLS estimates, and the key predictors (stress, activity, age, gender, BMI) remained significant across resamples, supporting the stability and robustness of the model.

## 7. Interpretation and Discussion

Stress is the most important predictor and consistently reduces sleep duration, highlighting the role of mental and psychological factors in sleep health. Physical activity has a small but positive association with sleep. Older individuals in this sample reported slightly longer sleep. Gender differences suggest males slept somewhat longer than females. Higher BMI (overweight/obese) is linked to shorter sleep, which may reflect physiological or lifestyle factors.

## 8. Conclusion

This multiple linear regression model is statistically strong, interpretable, and well validated. It explains about 83% of the variance in sleep duration. While heteroskedasticity was detected, the use of robust standard errors and bootstrap validation supports the reliability of the findings. Overall, stress level is the dominant negative predictor, while physical activity, age, gender, and healthy weight status contribute to better sleep outcomes.

***Sleep Study Conclusion***

This study used a comprehensive statistical inference framework including exploratory data analysis, bootstrap confidence intervals, permutation tests, correlation analysis, and multiple linear regression to evaluate how stress, physical activity, and lifestyle factors influence sleep duration and sleep quality.

Across all methods, stress level emerged as the strongest and most consistent determinant of sleep health.

**Bootstrap results** demonstrated that higher stress levels were associated with steadily decreasing sleep duration, with narrow confidence intervals confirming the stability of these estimates. Gender differences appeared small but consistent, with females sleeping slightly more than males.

**Permutation tests** provided direct statistical evidence that the observed group differences were not due to random chance. Two key comparisons were evaluated:

* Low vs. High Stress: Sleep duration differed by more than one hour between groups.  
  The permutation test yielded a p-value of 0, meaning none of the 5,000 shuffled datasets produced a difference as extreme as the observed one. This offers very strong evidence that stress meaningfully affects sleep duration.
* Sleep Disorder vs. No Disorder: Individuals without a disorder slept roughly 33 minutes longer.

Again, the permutation test returned a p-value of 0, confirming that this difference is statistically significant and not random variation.

These resampling-based tests strengthened the robustness of the findings by making no distributional assumptions and relying solely on randomized label shuffling.

**Correlation analysis** further supported these results, revealing a strong negative relationship between stress and sleep duration and quality, along with a moderate positive association between physical activity and sleep. Age and heart rate showed only weak relationships, suggesting demographic and physiological factors play a smaller role in this dataset.

**The multiple linear regression** model accounted for 83% of the variability in sleep duration, showing stress as the dominant negative predictor even after adjusting for age, physical activity, gender, and BMI category. Higher BMI (overweight/obese) was also linked to reduced sleep, while physical activity and age contributed positively. Bootstrap validation and robust standard errors ensured that these conclusions were stable despite minor heteroskedasticity.

In summary, all statistical methods especially the permutation tests, which directly test the significance of group differences point to the same conclusion:  
stress is the central, most impactful factor reducing sleep duration and quality, while physical activity, healthy weight, and demographic factors provide secondary contributions.

These results underscore the importance of stress management and active lifestyle habits as key strategies for improving sleep health. Using a multi-method inferential approach allowed for strong, converging evidence, demonstrating how psychological and behavioral factors play a crucial role in shaping sleep outcomes.

***GitHub repository*:** https:// github.com/ydjoudi/STAT-3333-project-group-8

***References***

1. UOM190346A. (2023). Sleep Health and Lifestyle Dataset. Kaggle. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

2. Chihara, L. & Hesterberg, T. (2019). Mathematical Statistics with Resampling and R. Wiley.