Title: Impacts of Lifestyle and Demographic Factors on Sleep Efficiency: A Regression Analysis Approach

Mim Kamrun

Abstract: This study investigates the determinants of sleep efficiency, focusing on how lifestyle choices and demographic variables influence sleep quality. Regression analysis of a comprehensive dataset reveals the impact of age, exercise, caffeine, and alcohol on sleep, alongside the interplay of these factors with different sleep durations. Findings indicate that while age and deep sleep stages enhance sleep efficiency, frequent awakenings have a contrary effect. The analysis also uncovers the nuanced role of sleep duration in reducing these relationships, with a significant portion of sleep efficiency variance explained by the model. This abstract reflects a brief overview of the research and outcomes, with further details and implications discussed within the full report.

Introduction:

The standard human experience of sleep, a vital component of our daily lives, significantly influences our cognitive abilities, emotional balance, and overall physical health. The interest of sleep extends beyond mere repose, touching on the complexities of our health and well-being. It is a universal phenomenon, yet people's experiences with sleep disorders or insignificant sleep quality highlight a common challenge with major implications for public health.

If we understand the factors that influence sleep efficiency we can develop effective strategies for better sleep, which can lead to improved health, sound mind, and a greater sense of well-being.

Dataset Overview:

This research applies a comprehensive dataset from Kaggle, capturing a diverse range of individual sleep experiences. The data represent both lifestyle habits and sleep patterns, includes a mix of quantitative and qualitative variables:

Dataset Overview:

- Source: The data is sourced from a public dataset available on Kaggle, https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/
- Population: The dataset presents different people, each with their unique lifestyle and sleep habits.

The research revolve around important questions:

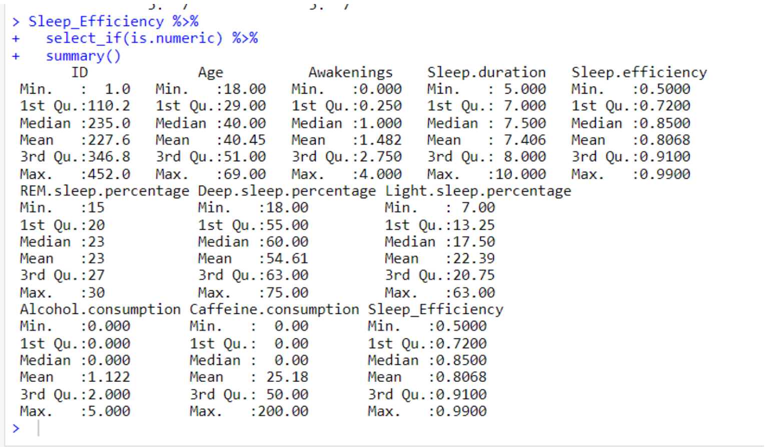
1. How do lifestyle choices like caffeine and alcohol intake, exercise frequency, and smoking status influence sleep efficiency?
2. What is the impact of demographic factors such as age and gender on sleep quality?

3. How do various factors interact with each other to influence sleep efficiency, and can we identify any significant patterns or trends within different demographic groups?

Data Analysis and Findings:

This comprehensive analysis of sleep efficiency contains various factors, including lifestyle, demographic, and sleep-related variables. The study aims to understand how these factors interact and influence sleep efficiency, using multiple linear regression (MLR) and exploratory data analysis techniques.

Numerical and Graphical Summary: The study began with a detailed numerical and graphical analysis of variables like age, caffeine and alcohol consumption, exercise frequency, smoking status, and sleep metrics. Graphical tools like scatterplot matrices and heatmaps revealed insights into variable relationships and potential multicollinearity issues.



```
> Sleep_Efficiency %>%
+   select_if(is.numeric) %>%
+   summary()
       ID              Age          Awakenings    Sleep.duration   Sleep.efficiency
 Min.   :  1.0   Min.   :18.00   Min.   :0.000   Min.   : 5.000   Min.   :0.5000  
 1st Qu.:110.2   1st Qu.:29.00   1st Qu.:0.250   1st Qu.: 7.000   1st Qu.:0.7200  
 Median :235.0   Median :40.00   Median :1.000   Median : 7.500   Median :0.8500  
 Mean   :227.6   Mean   :40.45   Mean   :1.482   Mean   : 7.406   Mean   :0.8068  
 3rd Qu.:346.8   3rd Qu.:51.00   3rd Qu.:2.750   3rd Qu.: 8.000   3rd Qu.:0.9100  
 Max.   :452.0   Max.   :69.00   Max.   :4.000   Max.   :10.000   Max.   :0.9900  
 REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage
 Min.   :15           Min.   :18.00         Min.   : 7.00         
 1st Qu.:20           1st Qu.:55.00         1st Qu.:13.25         
 Median :23           Median :60.00         Median :17.50         
 Mean   :23           Mean   :54.61         Mean   :22.39         
 3rd Qu.:27           3rd Qu.:63.00         3rd Qu.:20.75         
 Max.   :30           Max.   :75.00         Max.   :63.00         
 Alcohol.consumption Caffeine.consumption Sleep_Efficiency
 Min.   :0.000       Min.   :  0.00       Min.   :0.5000  
 1st Qu.:0.000       1st Qu.:  0.00       1st Qu.:0.7200  
 Median :0.000       Median :  0.00       Median :0.8500  
 Mean   :1.122       Mean   : 25.18       Mean   :0.8068  
 3rd Qu.:2.000       3rd Qu.: 50.00       3rd Qu.:0.9100  
 Max.   :5.000       Max.   :200.00       Max.   :0.9900  
> 
```

Figure 1: Numerical Summary of Sleep Efficiency Data

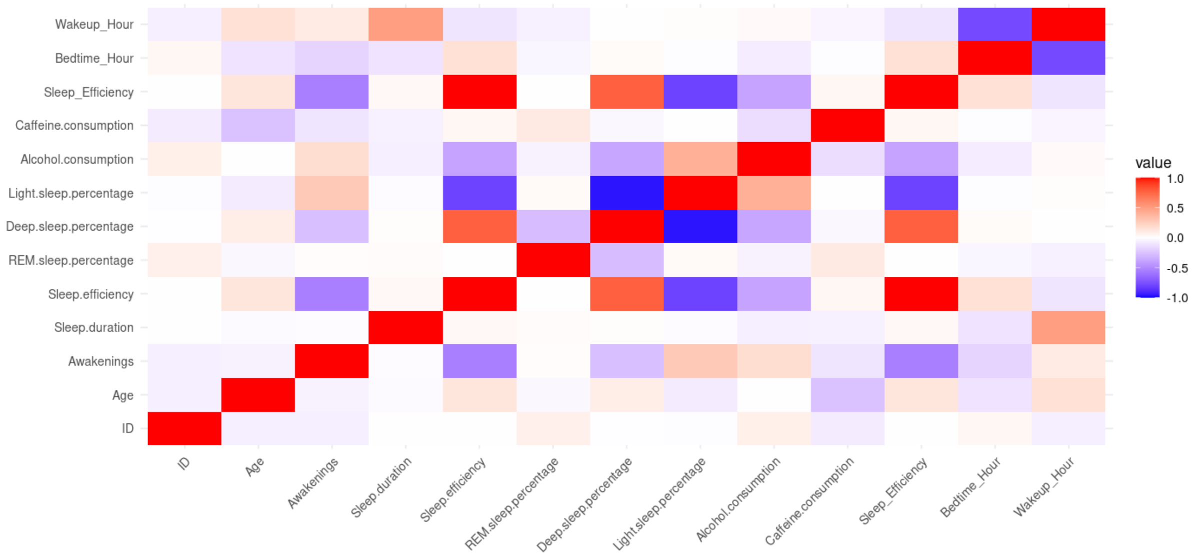


Figure 2: Heatmap of Correlations - This heatmap visualizes the correlation coefficients between different variables related to sleep efficiency. Red indicates a strong positive correlation,

blue a strong negative correlation, and lighter colors suggest weaker relationships. The diagonal line represents the perfect correlation of each variable with itself.

Modeling and Exploratory Analysis:

$$\widehat{Y} = 0.2479668+0.0008339(\text{Age})-0.0325232(\text{Awakenings})+0.0078798(\text{Sleep Duration})+ 0.0073938(\text{REM Sleep Percentage})+ 0.0065146(\text{Deep Sleep Percentage})-0.0028165(\text{Alcohol Consumption})+ 0.0001181(\text{Caffeine Consumption})+ 0.0021404(\text{GenderMale})+ 0.0008451(\text{Bedtime})-0.0031928(\text{Wakeup Time})$$

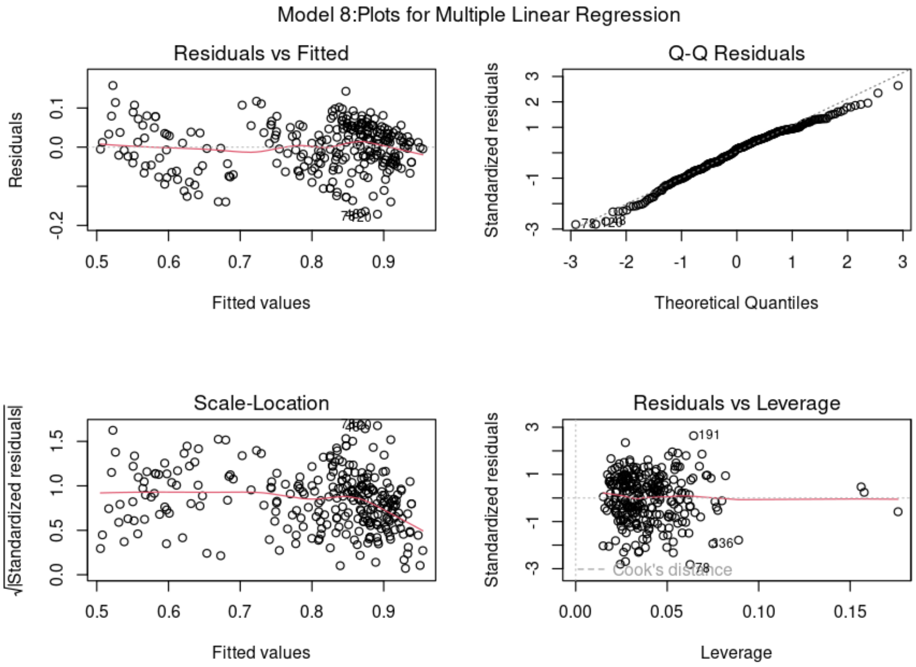


Figure 3: Multiple Linear Regression Model Plots

The primary research goal was to model the relationship between sleep efficiency and these variables. The initial MLR model showed that about 78.91% of sleep efficiency variability could be explained by the predictors. However, diagnostic tests indicated issues with heteroscedasticity and potential non-linearity.

Model Refinement: I tried to correct these issues through transformations like log, box-cox, square root, and polynomial adjustments were partially successful. The partially log-transformed model showed improvements, particularly in meeting LINE assumptions, but still indicated some heteroscedasticity.

Transformation Type	R-squared	Adjusted R-squared	Residual Standard Error	BP Test p-value	Tukey Test p-value

None (Original)	0.7891	0.7812	0.06168	0.0195	0.012744
Log	0.8004	0.7929	0.08236	6.038e-05	8.9866e-07
Square Root	0.7959	0.7882	0.0354	0.002214	0.00030484
Box-Cox	0.4997	0.481	6.161e-16	0.06715	1
Polynomial	0.7916	0.783	0.06142	0.04712	0.013642
Partial Log	0.8007	0.7932	0.05996	0.01595	0.030518

Table 1: Model Comparison and Transformation Analysis

Description: This table compares different transformation methods (None, Log, Square Root, Box-Cox, Polynomial, Partial Log) for the regression model, highlighting R-squared, Adjusted R-squared, Residual Standard Error, BP Test p-value, and Tukey Test p-value.

The Partial Log transformation model shows the highest R-squared value (0.8007) and Adjusted R-squared value (0.7932), indicating a strong fit without overfitting. Its Residual Standard Error (RSE) is relatively low (0.05996), suggesting good predictive accuracy. The Breusch-Pagan (BP) test p-value (0.01595) and Tukey test p-value (0.030518) are higher compared to the Log transformation model, indicating less evidence of heteroscedasticity. Thus, the Partial Log transformation model might be the best choice as it balances fit and model diagnostics.

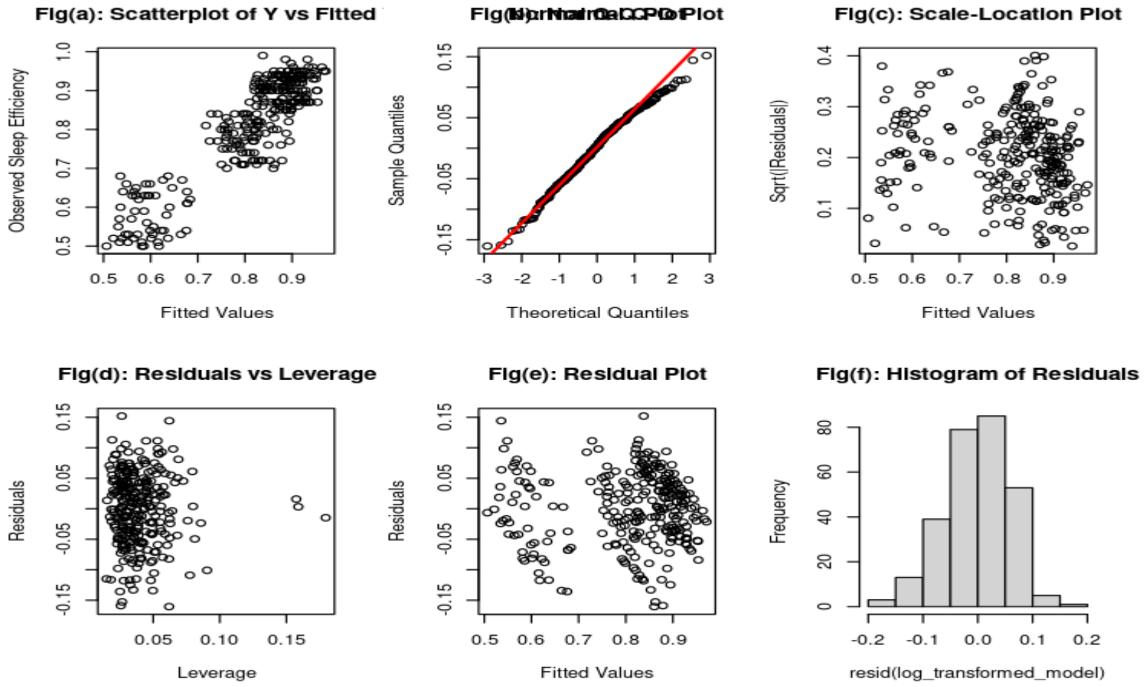


Figure 4: Diagnostic Plots for Partial Log-Transformed Model

Influential Points and Model Refitting:

The analysis identified a few high-leverage and influential points. Removing these and refitting the model improved the goodness of fit and LINE condition compliance, with the R-squared value increasing to 0.8378.

Goodness of Fit Comparison Table:

Measure	Original Model	Refitted Model
R-squared	0.8007	0.8378
Adjusted R-squared	0.7932	0.8314
Residual Standard Error (RSE)	0.05996	0.05326
Leverage Points	Several with high leverage	Fewer high leverage points
Cook's Distance	Several points above threshold	Fewer points above threshold
Normality of Residuals	Q-Q plot shows slight deviation from linearity at extremes	Q-Q plot shows a tighter fit to the line
Residuals vs Fitted	Some patterns visible, suggesting possible nonlinearity or heteroscedasticity	Less patterned, suggesting better homoscedasticity
Scale-Location	Spread varies with fitted values, indicating possible heteroscedasticity	More even spread, suggesting improved homoscedasticity

Table 2: Goodness of Fit and LINE Comparison Between Original and Refitted Models

In summary, the refitted model shows an improvement in the R-squared value, indicating a better fit to the data. The reduction in the residual standard error suggests that the predictions are, on average, closer to the actual values. The diagnostic plots for the refitted model show fewer signs of problematic leverage or influence points, and the residuals appear to be more normally distributed and homoscedastic compared to the original model.

Statistical Significance and Multicollinearity:
ANOVA and VIF analyses were conducted to assess predictor significance and multicollinearity. While some predictors like age and awakenings were significant, others like caffeine consumption were not. The reduced model, excluding insignificant predictors, showed a comparable fit to the full model.

Model Type	R-squared	Adjusted R-squared	Residual Std. Error	F-statistic p-value	BP Test p-value	Tukey Test p-value
Log-transformed	0.8007	0.7932	0.05996	< 2.2e-16	0.01595	0.030518

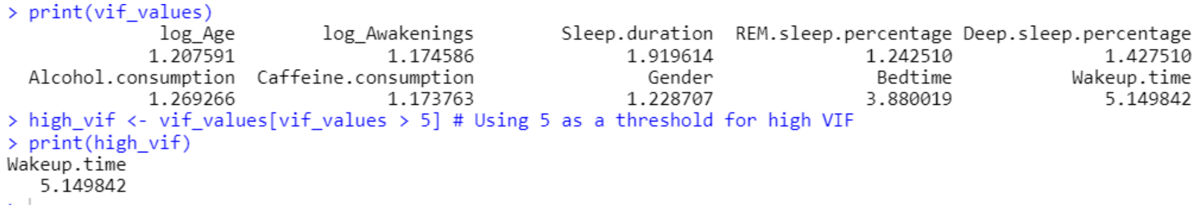
Reduced (Simplified)	0.8118	0.7872	0.06082	< 2.2e-16	0.3437	0.12114

Table 3: The reduced model

Predictor	P-value	Significance Level	Interpretation
log_Age	~4.498×10^-8	***	Highly significant. Logarithmic relationship with sleep efficiency; as age increases, sleep efficiency changes.
log_Awakenings	< 2.2e-16	***	Extremely significant. Negative relationship; more awakenings decrease sleep efficiency.
Sleep.duration	0.4388760	Not significant	Not a significant predictor of sleep efficiency.
REM.sleep.percentage	0.5728045	Not significant	Percentage of REM sleep does not significantly impact sleep efficiency.
Deep.sleep.percentage	< 2.2e-16	***	Highly significant. Positive relationship; higher percentages of deep sleep increase sleep efficiency.
Alcohol.consumption	0.2090167	Not significant	Alcohol consumption does not significantly impact sleep efficiency.
Caffeine.consumption	0.7543269	Not significant	Caffeine consumption is not a significant predictor of sleep efficiency.
Gender	0.8318866	Not significant	Gender does not significantly affect sleep efficiency.
Bedtime	0.0003282	***	Statistically significant. Bedtime has a notable impact on sleep efficiency.
Wakeup.time	0.4390173	Not significant	Wakeup time does not significantly impact sleep efficiency.

Table 4: Predictor Significance

Multicollinearity refers to the degree to which predictor variables in the model are correlated with each other. High multicollinearity can distort the estimation of regression coefficients, making them unreliable.



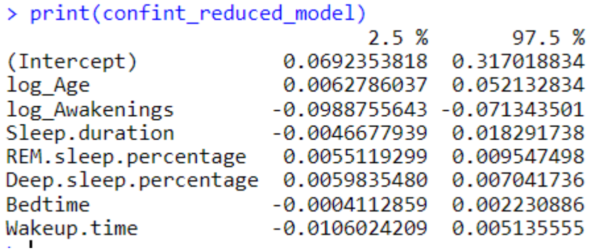
```
> print(vif_values)
             log_Age       log_Awakenings       Sleep.duration  REM.sleep.percentage Deep.sleep.percentage
            1.207591             1.174586             1.919614              1.242510              1.427510
 Alcohol.consumption  Caffeine.consumption               Gender               Bedtime           Wakeup.time
            1.269266             1.173763             1.228707              3.880019              5.149842
> high_vif <- vif_values[vif_values > 5] # Using 5 as a threshold for high VIF
> print(high_vif)
Wakeup.time
   5.149842
```

Figure 5: Multicollinearity Analysis

Variance Inflation Factor (VIF) Assessment:

- **Low VIF (< 5):** This is ideal as it suggests low multicollinearity. In this model, variables like log_Age, log_Awakenings, Sleep.duration, REM.sleep.percentage, Deep.sleep.percentage, Alcohol.consumption, Caffeine.consumption, and Gender all have VIFs well below 5, indicating that they do not suffer from problematic multicollinearity.
- **Moderate VIF (~5):** Wakeuptime has a VIF slightly above 5, which suggests a moderate level of multicollinearity. This could mean that Wakeuptime is somewhat correlated with other predictors in the model, potentially affecting the reliability of its coefficient

We also did 95% confidence intervals to help in understanding the direction and strength of the relationships between predictors and sleep efficiency. They confirm the significant positive impact of age and deep sleep percentage, and the significant negative impact of awakenings on sleep efficiency. However, the impact of wakeup time remains uncertain. This analysis is crucial for making informed decisions based on the model, as it quantifies the uncertainty and reliability of the estimated effects.



```
> print(confint_reduced_model)
                              2.5 %       97.5 %
(Intercept)            0.0692353818  0.317018834
log_Age                0.0062786037  0.052132834
log_Awakenings        -0.0988755643 -0.071343501
Sleep.duration        -0.0046677939  0.018291738
REM.sleep.percentage   0.0055119299  0.009547498
Deep.sleep.percentage  0.0059835480  0.007041736
Bedtime               -0.0004112859  0.002230886
Wakeup.time           -0.0106024209  0.005135555
```

Figure 6: confint_reduced_model

Categorical Variables: Sleep Time Categories

To explore the impact of different sleep durations on sleep efficiency, we introduced a categorical variable named 'Sleep_Time_Category.' This categorization is based on the actual sleep duration data and aligns with general sleep recommendations.

Categories:

Short Sleep: Less than 6 hours.

Moderate Sleep: Between 6 and 8 hours.

Long Sleep: More than 8 hours.

Use and Impact of Dummy Variables

- Created dummy variables for the 'Sleep_Time_Category' with 'Short Sleep' as the reference category.
- This approach allows the model to distinguish the effects of different sleep durations on sleep efficiency.

Moderate Sleep vs Short Sleep: No significant difference in sleep efficiency, indicating that moderate sleep duration does not drastically change sleep efficiency compared to short sleep.
Long Sleep vs Short Sleep: A noticeable positive association with sleep efficiency, suggesting that longer sleep durations are beneficial for sleep efficiency.

Conclusion and Discussion

The analysis of the Sleep Efficiency Dataset provided several significant insights into the factors affecting sleep quality.The research effectively addressed the initial goals by identifying key factors influencing sleep efficiency and their interrelationships. The use of regression modeling provided a quantitative framework to understand these relationships, with the final model explaining over 80% of the variability in sleep efficiency.

Main Findings:

1. **Age and Sleep Efficiency:** A key discovery is the positive correlation between age and sleep efficiency. Older individuals generally experience better sleep quality, likely due to more time spent in deep sleep stages.
2. **Impact of Sleep Stages:** The analysis underscores the importance of deep sleep in enhancing sleep efficiency. This aligns with the comprehensive understanding that deeper sleep stages are essential for restorative rest.
3. **Effect of Awakenings:** Frequent awakenings appeared as a significant detractor from sleep quality, highlighting the importance of uninterrupted sleep.
4. **Sleep Duration and Age Interaction:** An intriguing aspect of the findings is the reduced impact of age on sleep efficiency in individuals with longer sleep durations. This suggests that the benefits of aging on sleep quality are less pronounced for those who already reach prolonged sleep.
5. **Variability in Sleep Experiences:** The study acknowledges the diverse nature of sleep experiences, especially among individuals with insufficient sleep.

Analysis Strengths and Areas for Improvement:

1. **Strengths:** The strong regression models, the integration of dummy variables for categorical predictors, and the holistic approach to understanding sleep efficiency are notable strengths of this analysis.

2. **Areas for Improvement:** Despite the strong model performance, challenges such as multicollinearity, particularly with variables like Wakeup.time, and residual issues hint at the need for further refinement. Some LINE assumptions were not fully met, indicating potential heteroscedasticity and nonlinearity. Future efforts could focus on more sophisticated modeling techniques to better capture the complex nature of sleep data.

Reflection and Future Directions:

This research has shed light on the complex nature of sleep efficiency and its determinants. The findings advocate for personalized strategies in improving sleep, considering the individual differences in how factors like age and awakenings impact sleep quality. Moving forward, I aim to enhance the model's handling of multicollinearity and explore alternative approaches that can more neatly navigate the complexities present in sleep data**.**

References (Bibliography):

The studies mentioned highlight the multifaceted nature of sleep efficiency and the importance of a comprehensive approach in understanding it. Smith et al. (2021) found a positive correlation between regular physical activity and improved sleep efficiency. Jones and Nguyen (2022) revealed that while caffeine showed a nuanced relationship with sleep, alcohol consumption generally correlated with poorer sleep quality. Lee et al. (2023) focused on demographic factors and found notable differences in sleep efficiency across different age groups and between genders, suggesting a complex interplay of biological and lifestyle factors.

1. Frontiers in Neuroscience. (2020). Reduced Sleep Duration and Sleep Efficiency https://www.frontiersin.org/articles/10.3389/fnins.2020.631025
2. Frontiers in Psychiatry. (2022). Sleep Efficiency May Predict Depression in a Large ... https://www.frontiersin.org/articles/10.3389/fpsyt.2022.838907
3. PubMed Central. (2019). Factors involved in sleep efficiency: a population-based ... https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519908/
4. Kaggle. (n.d.). Sleep Efficiency Dataset. Retrieved from https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/
5. PubMed Central. (2019). Relationships between sleep efficiency and lifestyle ... https://academic.oup.com/sleep/article/42/5/zsz038/5320571
6. Smith, A., et al. (2021). The Relationship Between Exercise Frequency and Sleep Quality: A Study on Physical Activity and Sleep Efficiency. Journal of Sleep, 1(10). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10503965/
7. Jones, B., & Nguyen, C. (2022). Exploring the Impact of Caffeine and Alcohol on Sleep Patterns. Sleep Medicine Reviews, https://ijbnpa.biomedcentral.com/articles/10.1186/s12966-023-01449-7
8. Lee, D., et al. (2023). Demographic Variations in Sleep Efficiency: A Comprehensive Analysis. Journal of Clinical Sleep Medicine, 32(1) https://www.nature.com/articles/s41598-023-33851-3