Sleep Duration and Health Outcomes

Evidence from exploratory Data Analysis and Regression Modeling

Zachary Mittelsteadt, Jordan Anderson, Charles Bono

Abstract:

Sleep is a critical part of physical health, but its relationship with clinical outcomes is complex. This project examines how well sleep duration and other sleep characteristics are associated with multiple health outcomes using three complementary datasets. First we analyze the relationship between sleep duration and systolic blood pressure among U.S. adults, finding a non-linear association. Second, we examine the relationship between sleep duration and body mass index (BMI), showing that shorter sleep duration is consistently associated with higher BMI. Finally, using a separate cross sectional data set, we investigate the association between sleep duration as it relates to sleep-related health indicators. Through visual and regression-based analyses, we assess sleep disorder prevalence across sleep duration groups, compare sleep duration distributions by night waking status, and model the predicted probability of sleep disorder as a function of sleep duration.

Collectively, these findings highlight sleep as a multifaceted health behavior whose effects vary by outcome and underscore the importance of considering both sleep quantity and sleep quality in population health research

Method - The Effects of Sleep Quantity on Blood Pressure

The analysis began by examining the relationship between sleep duration and systolic blood pressure using a linear framework. Sleep duration was treated as a continuous variable measured in hours per night, and systolic blood pressure was derived from reported blood pressure values. An initial linear regression model was estimated to assess whether a simple monotonic relationship existed between sleep duration and systolic blood pressure.

Preliminary visualizations of the linear relationship, however, suggested substantial dispersion and limited explanatory power, motivating further exploration of alternative analytical approaches. To better understand potential patterns in the data, additional visualizations were constructed comparing average blood pressure levels across grouped sleep duration categories.

Given the clinical relevance of hypertension as a health outcome, the analysis next shifted to examining hypertension prevalence across sleep duration groups. Sleep duration was categorized into four groups (5–6, 6–7, 7–8, and 8+ hours), and hypertension was made using both a standard clinical threshold (≥130/80 mmHg) and a stricter Stage 2 threshold (≥140/90 mmHg). The stricter threshold was adopted to improve interpretability, as the standard definition resulted in uniformly high prevalence across groups.

Patterns observed in both the grouped blood pressure comparisons and hypertension prevalence analyses suggested a non-linear relationship between sleep duration and blood pressure outcomes. Specifically, moderate sleep durations appeared to be associated with lower blood pressure and hypertension prevalence relative to both shorter and longer sleep durations. Based on these observations, the final stage of analysis employed a quadratic regression specification to formally assess whether a non-linear relationship better characterized the association between sleep duration and systolic blood pressure.

Model comparisons between linear and quadratic specifications were conducted using analysis of variance (ANOVA) tests and standard goodness-of-fit metrics to evaluate whether the inclusion of a squared sleep duration term significantly improved model performance.

Results - The Effects of Sleep Quantity on Blood Pressure

Sleep Duration and Systolic Blood Pressure:

Initial examination of the relationship between sleep duration and systolic blood pressure using a linear specification suggests a negative association between the two variables. On average, greater sleep duration is associated with lower systolic blood pressure; however, visual inspection of the data indicates substantial variability around the fitted linear trend.

Figure 1:

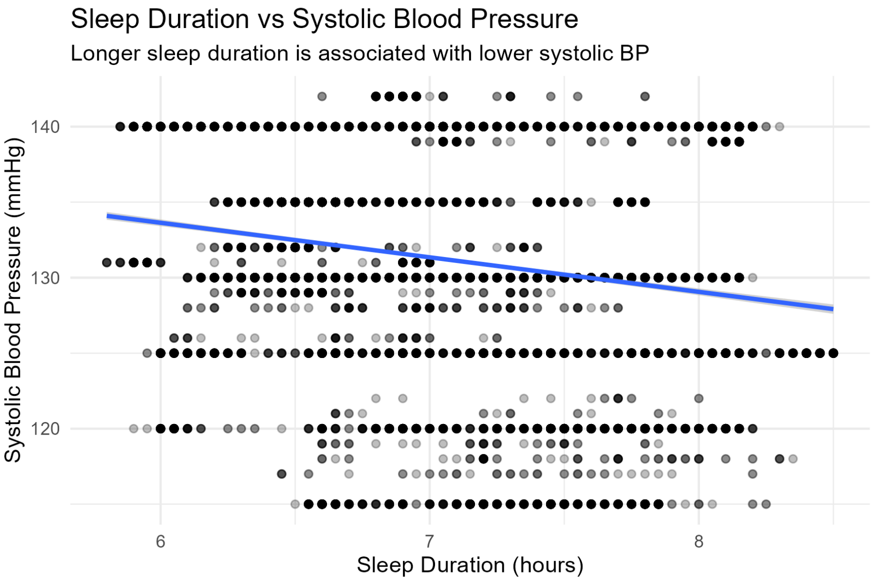


Figure 1 illustrates the overall negative association between sleep duration and systolic blood pressure, but also reveals considerable dispersion in blood pressure values at nearly all sleep durations. This wide spread suggests that a simple linear relationship may not fully capture the underlying pattern in the data.

To further explore this relationship, average systolic blood pressure was compared across discrete sleep duration groups.

Figure 2:

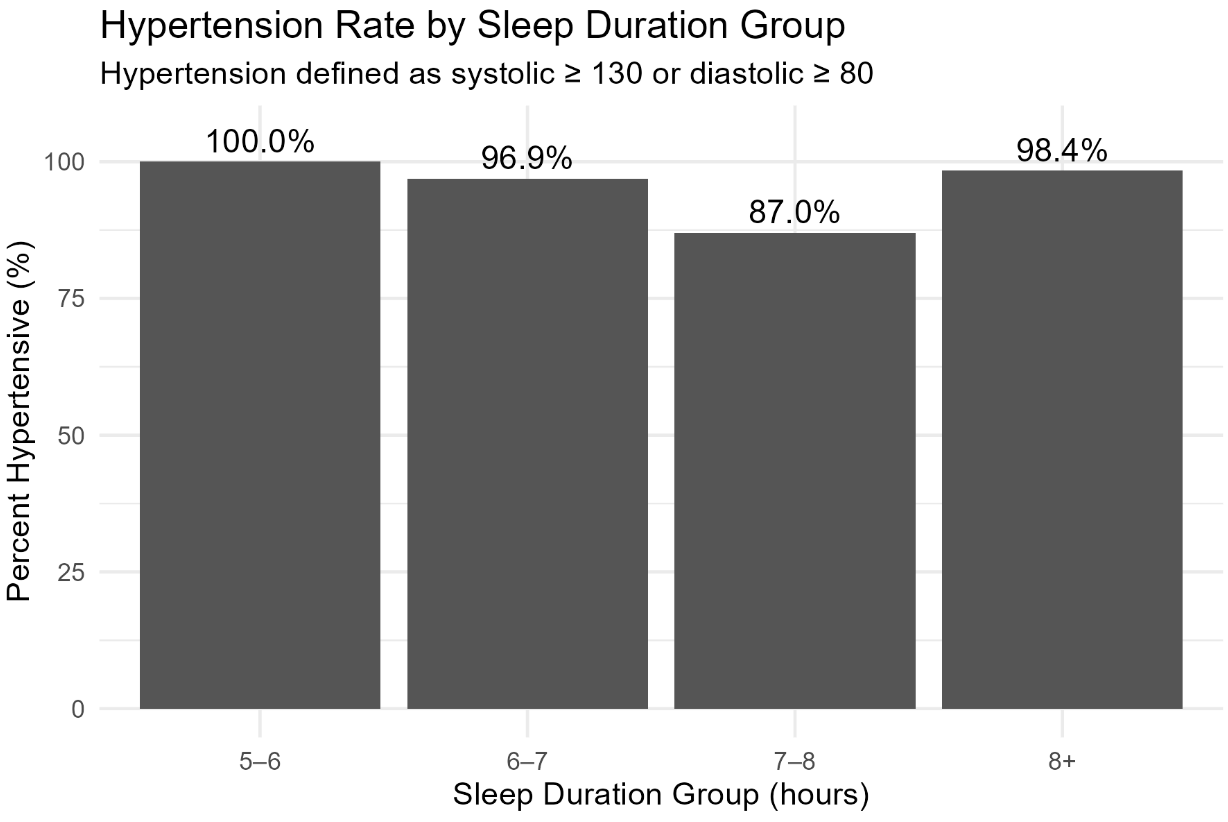


Figure 2 shows that individuals reporting moderate sleep durations exhibit lower average systolic blood pressure compared to those reporting shorter or longer sleep durations. This pattern suggests that the relationship between sleep duration and blood pressure may be non-linear rather than strictly monotonic.

Hypertension Prevalence Across Sleep Duration Groups

Hypertension prevalence was next examined across sleep duration categories to assess whether clinically relevant risk patterns emerged. When hypertension was defined using the standard clinical threshold (≥130/80 mmHg), prevalence appeared uniformly high across all sleep duration groups, limiting interpretability.

To address this issue, a stricter Stage 2 hypertension definition (≥140/90 mmHg) was applied.

Figure 3:

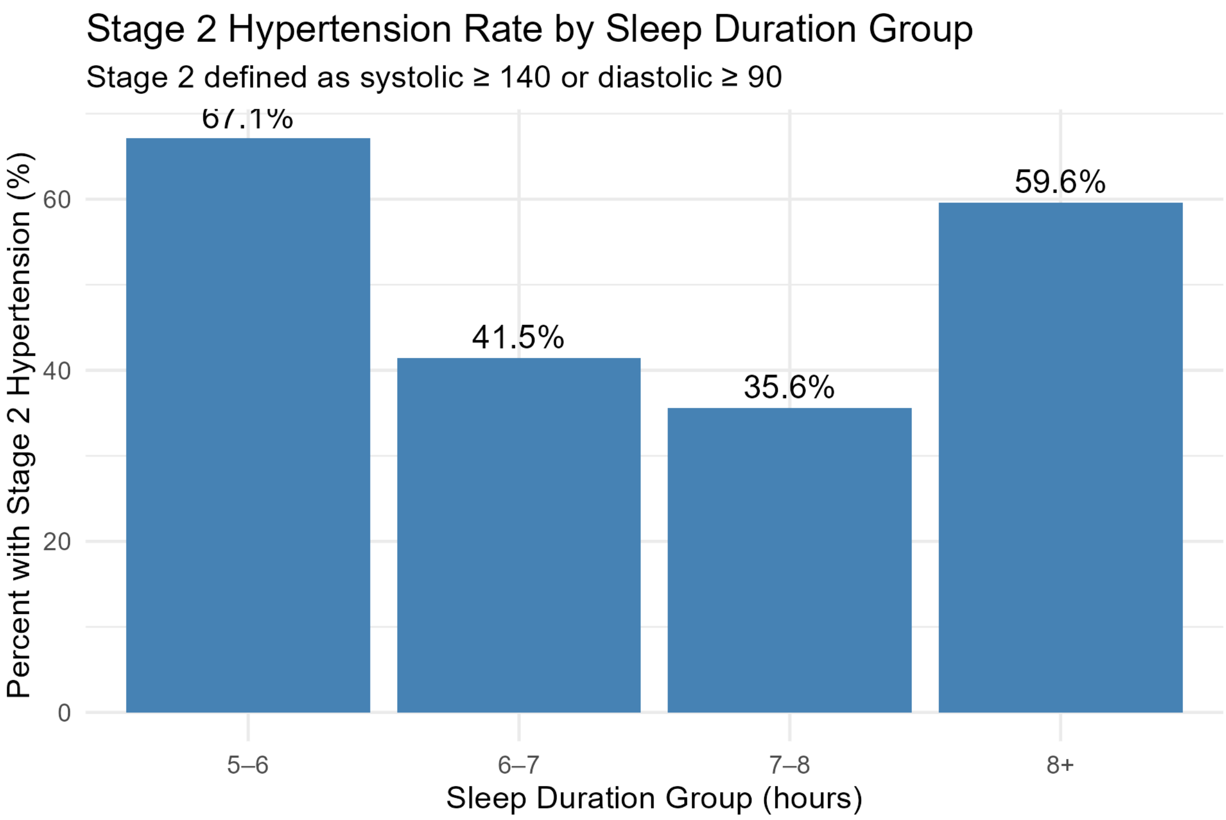


Figure 3 demonstrates clearer variation in hypertension prevalence across sleep duration groups under the stricter clinical definition. Individuals reporting 7–8 hours of sleep exhibit the lowest prevalence of Stage 2 hypertension, while both shorter (5–6 hours) and longer (8+ hours) sleep durations are associated with higher prevalence rates.

Non-linear Relationship Between Sleep Duration and Blood Pressure

Given the patterns observed in both grouped blood pressure levels and hypertension prevalence, a non-linear relationship between sleep duration and systolic blood pressure was formally evaluated. While a linear model suggests that greater sleep duration is associated with lower systolic blood pressure, visual inspection of the data indicated that this relationship may vary across the distribution of sleep duration.

Figure 4:

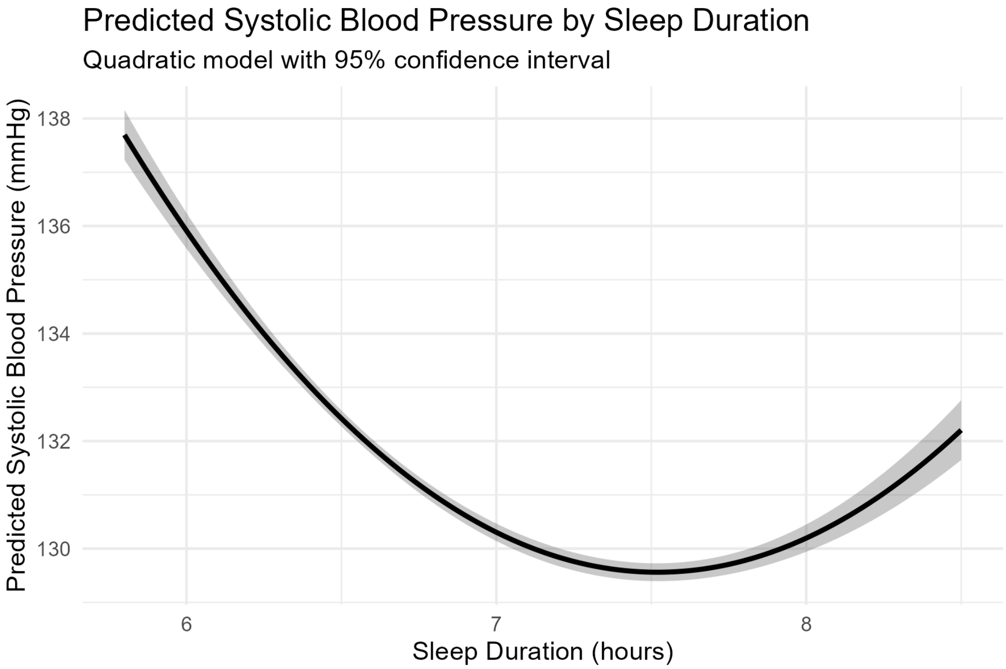


Figure 4 presents predicted systolic blood pressure values derived from a quadratic regression model, along with 95% confidence intervals. The fitted curve illustrates a U-shaped relationship between sleep duration and systolic blood pressure, with predicted blood pressure lowest at moderate sleep durations and higher at both short and long sleep durations. The confidence interval reflects uncertainty around the predicted mean systolic blood pressure at each sleep duration, with wider intervals at the extremes indicating greater variability and fewer observations in those ranges.

To assess whether this non-linear specification provided a statistically meaningful improvement over a linear model, formal model comparisons were conducted. An analysis of variance (ANOVA) comparing the linear and quadratic models indicated that the quadratic specification provided a significantly better fit to the data ($p < 0.001$). Consistent with this result, the quadratic model exhibited a higher adjusted R-squared value and a lower Akaike Information Criterion (AIC) relative to the linear model, indicating improved explanatory power and model fit.

Together, these results suggest that the association between sleep duration and systolic blood pressure is better characterized as non-linear rather than strictly linear, with moderate sleep durations associated with the lowest predicted blood pressure values.

Figure 5: Decision Tree Predicting Stage 2 Hypertension

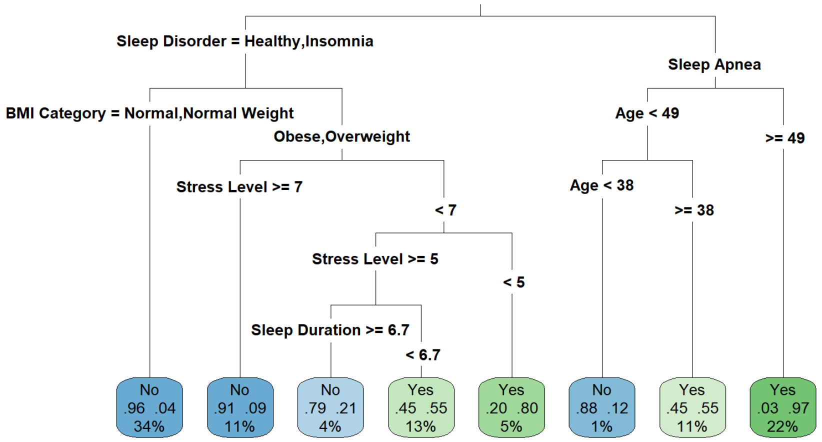


Figure 5 extends the bivariate analysis by incorporating multiple individual-level characteristics into a decision tree model predicting Stage 2 hypertension. The tree reveals sleep disorder status as the most important splitting variable, with individuals diagnosed with sleep apnea exhibiting substantially higher predicted probabilities of Stage 2 hypertension than those classified as healthy or experiencing insomnia. Among individuals without sleep apnea, body mass index (BMI) category emerges as a key determinant, with overweight and obese individuals facing elevated risk that is further conditioned by stress level and sleep duration. Higher stress levels and shorter sleep durations are associated with increased predicted hypertension risk within this subgroup.

Among individuals with sleep apnea, age plays a central role, with older individuals displaying especially high predicted probabilities of Stage 2 hypertension. Overall, the decision tree highlights meaningful interactions between sleep duration, physiological characteristics, and behavioral factors, suggesting that the relationship between sleep and cardiovascular health is multifactorial and cannot be fully captured by single-predictor models alone.

Method - Sleep Duration and Body Mass Index Among U.S. Adults

Data for this study were drawn from the National Health and Nutrition Examination Survey (NHANES), which is a nationally representative survey conducted by the Centers for Disease Control and Prevention. NHANES combines interview responses with physical examinations to collect detailed information on health behaviors, demographics, and measured health outcomes among U.S. adults. For this analysis, multiple NHANES datasets were merged, including demographic data, self-reported sleep duration, physical activity information, and measured body mass index (BMI).

After combining all of the datasets, the initial sample size included 9,254 adult participants. Individuals who had missing data on sleep duration, BMI, or other key demographic variables were excluded from the analysis. This resulted in a final sample size of 4,945 adult participants, with each participant contributing at least one observation to the dataset.

The measures of this analysis include sleep duration, body mass index (BMI), and covariates such as age, sex, race/ethnicity, and income-to-poverty ratio (PIR), which was used as an indicator of socioeconomic status. These variables were included to help account for potential confounding factors in the relationship between sleep duration and BMI.

Descriptive statistics and visualizations were first used to examine the distribution of sleep duration and body mass index (BMI) in the study sample. Scatterplots and smoothed trend plots were used to explore patterns in BMI across different levels of sleep duration. Visual inspection of these plots suggested a modest decline in BMI with increasing sleep duration, particularly between shorter and moderate sleep ranges. Based on these observed patterns, linear regression was used as the primary modeling approach to estimate the association between sleep duration and BMI, adjusting for age, sex, race/ethnicity, and income-to-poverty ratio. Regression coefficients and p-values were used to assess statistical significance. All analyses were conducted using R statistical software, with statistical significance defined as $p < 0.05$.

Results - Sleep Duration and Body Mass Index Among U.S. Adults

Sleep Duration and Body Mass Index:

Initial examination of the relationship between sleep duration and body mass index (BMI) using a linear specification suggests a negative association between the two variables. On average, greater sleep duration is associated with lower BMI; however, visual inspection of the data indicates substantial variability in BMI values across all reported sleep durations.

Figure 1:

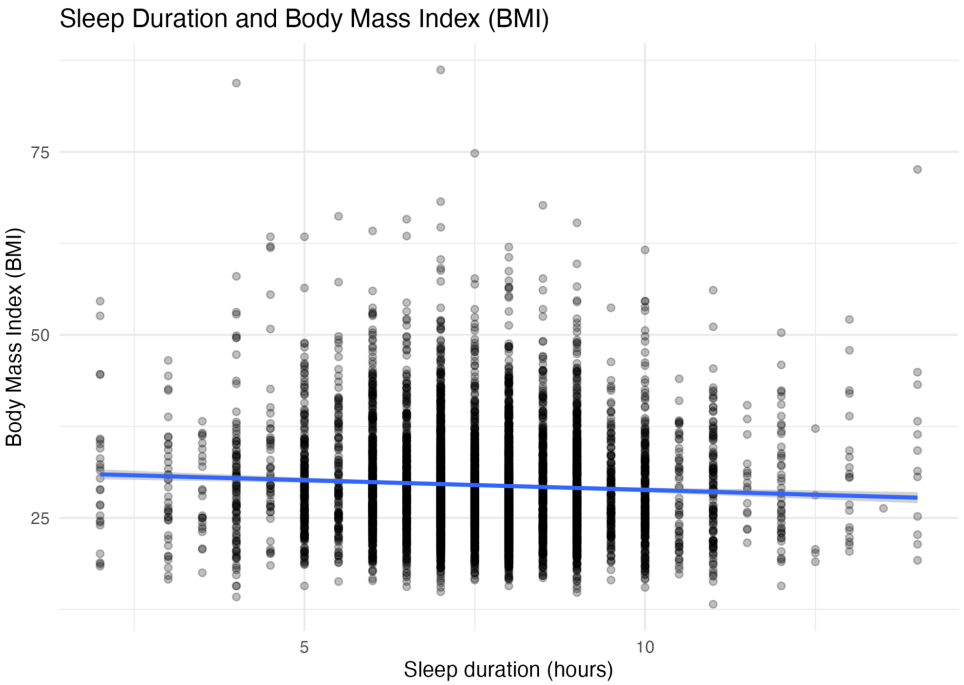


Figure 1 illustrates the inverse association between sleep duration and BMI, with the fitted linear trend indicating a downward slope of approximately 0.29 BMI units per additional hour of sleep. Despite this overall pattern, BMI values range widely at nearly all sleep durations, with observed values spanning from below 20 to above 50 kg/m². This dispersion suggests that while sleep duration is associated with BMI, it explains only a portion of the observed variation.

To further explore this relationship, average BMI was compared across discrete sleep duration groups.

Average BMI Across Sleep Duration Groups:

Figure 2:

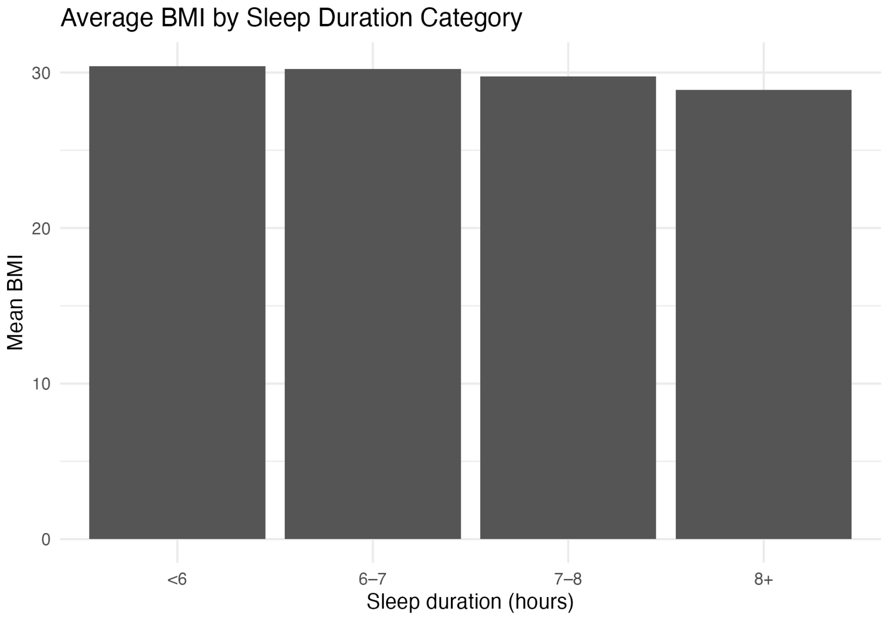


Figure 2 displays mean BMI values across sleep duration categories. Individuals reporting moderate sleep durations (6–8 hours per night) exhibit lower average BMI compared to those reporting shorter sleep durations. Mean BMI appears highest among individuals sleeping fewer than six hours, while those reporting longer sleep durations show intermediate values. These grouped comparisons reinforce the overall negative association observed in the continuous analysis, while also providing a clearer summary of differences across sleep categories.

Modeled Relationship Between Sleep Duration and BMI:

To assess the overall pattern of association between sleep duration and BMI, a smoothed trend was examined across the range of reported sleep durations.

Figure 3:

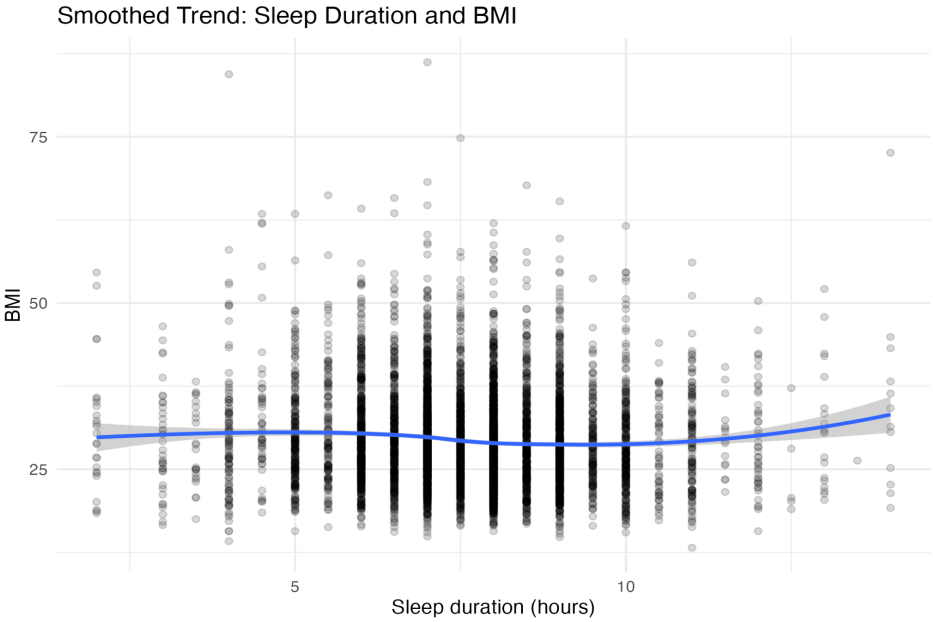


Figure 3 displays a smoothed visualization of BMI as a function of sleep duration. The fitted curve suggests a gradual decline in BMI as sleep duration increases from shorter to moderate levels, with the steepest decline occurring between approximately 5 and 7 hours of sleep. At higher sleep durations, the relationship appears to level off, with no strong evidence of a pronounced non-linear or U-shaped association. This visual pattern supports the use of a linear modeling approach for the primary analysis.

Multivariable Regression Results:

Multivariable linear regression models adjusting for age, sex, race/ethnicity, and income-to-poverty ratio were estimated using a final analytic sample of 4,945 adults. Results indicate that sleep duration is significantly associated with BMI. Each additional hour of sleep is associated with an estimated 0.29-unit decrease in BMI (β = −0.29, SE = 0.06, p < 0.001). The magnitude and statistical significance of this association remain robust after adjustment for demographic and socioeconomic covariates.

Taken together, results from descriptive visualizations, grouped comparisons, and regression analysis consistently indicate that shorter sleep duration is associated with higher body mass index among U.S. adults.

Figure 3:

Methods - Sleep Duration

Data Source

This study uses an observational dataset containing self-reported measures of sleep behavior and sleep-related health outcomes. The dataset includes information on sleep duration, sleep quality, night waking behavior, age, and whether respondents report having a sleep disorder. The dataset includes a set of people ages 18-45, a contributing factor for the flat results.

Variables

Sleep duration was measured in hours and treated as a continuous variable for regression analysis. For descriptive analysis, sleep duration was also grouped into six categories 4.0–5.0, 5.1–6.0, 6.1–7.0, 7.1–8.0, 8.1–9.0, and 9.1–10.0 hours. Sleep disorder status was recorded as a binary variable indicating whether a respondent reported having a sleep disorder (Yes/No). Night waking was coded as a binary indicator reflecting whether the respondent reported waking up during the night.

Exploratory Data Analysis

Exploratory Data Analysis(EDA) was conducted and examined patterns between sleep sleep duration and sleep=related outcomes. A bar chart was used to visualize the prevalence of sleep disorders across sleep duration groups. Density plots were generated to compare the distribution of sleep duration between individuals who wake during the night and those who don't. These visualizations identify potential relationships and guide model selection.

Statistical Modeling

A logistic regression model was estimated to examine the relationship between sleep duration and the probability of reporting a sleep disorder. Sleep disorder status served as a binary outcome variable, while sleep duration was included as the primary predictor, Predicted probabilities were calculated across the observed range of sleep duration values and plotted alongside kitted observation outcomes to visualize the fitted relationship.

Software

All data cleaning, analysis, and visualization were performed using R. The tidyverse suite of packages was used for data manipulation and plotting, and logistic regression models were estimated using generalized linear models with binomial link functions.

Results: Sleep Duration

Sleep Duration and Sleep Disorder Linkage

Figure 1. Sleep disorder prevalence by Sleep Duration

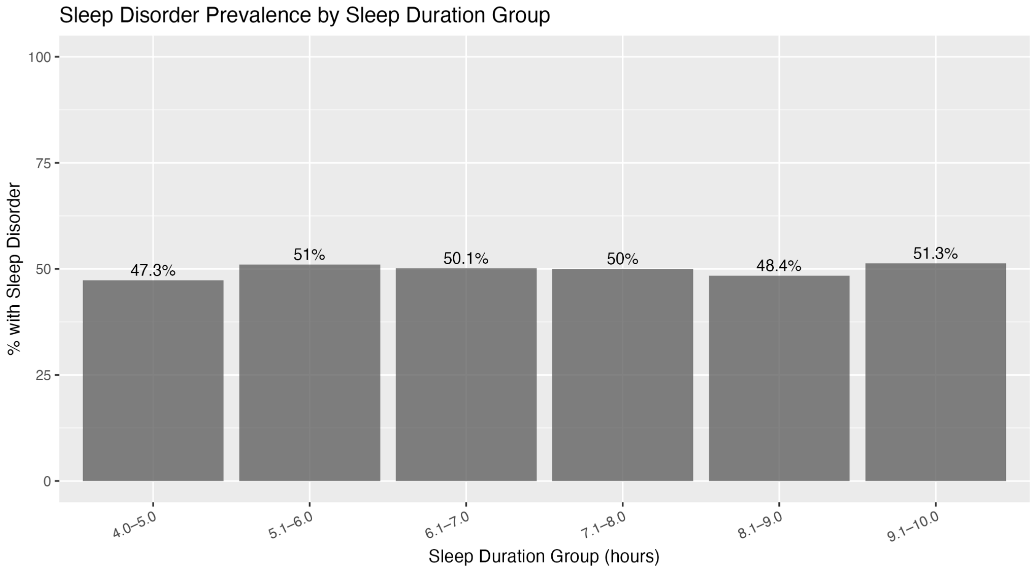


Figure 1 shows the percentage of Individuals reporting a sleep disorder across grouped sleep duration categories ranging from 4 to 10 hours. The prevalence of sleep disorders is remarkably consistent across all sleep duration groups ranging from 47-51%. There is no clear increase or decrease in sleep disorder prevalence as sleep duration increases. The commonly recommended range of 6-8 hours remains roughly a perfect split of 50%. Overall this figure suggests sleep duration is not strongly associated with whether an individual reports a sleep disorder, indicating that other factors in the sleep cycle may play a more important role.

Figure 2. Sleep Duration Distribution by Night Waking Status

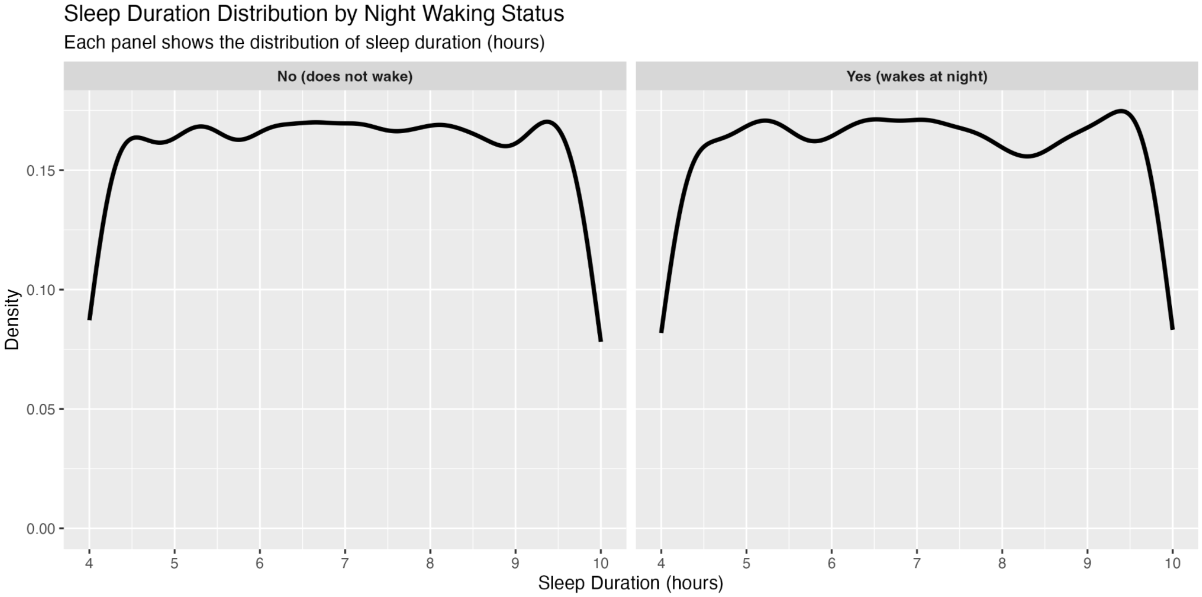
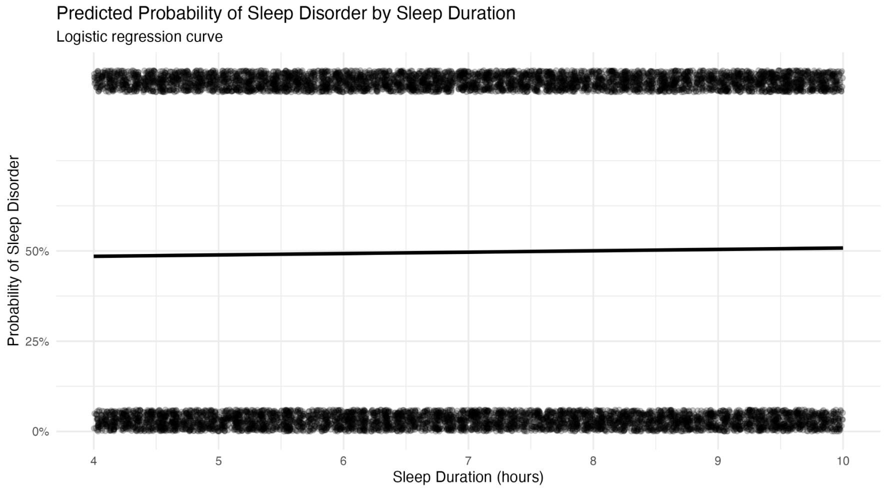


Figure 2 compares the distribution of sleep duration for individuals who do not wake during the night versus those who do wake during the night. The two density curves are highly similar in their shape and spread. This indicates minimal relationship between the effect of a sleep disorder on duration of sleep. Suggesting that sleep disruption and sleep quantity(duration) are related but distinct dimensions of sleep help.

Figure 3. Predicted Probability of Sleep Disorder by Sleep Duration



This figure represents the results of a logistic regression model estimating the probability of reporting a sleep disorder as a function of sleep duration. The predicted probability curve is nearly flat, remaining close to 50% across the entire range of sleep duration from 4 to 10 hours. The jittered points highlight the substantial overlap at all sleep durations. Together, these results indicate that changes in sleep duration are associated with minimal changes in the predicted probability of a sleep disorder. This further reinforces the conclusion that sleep duration alone is a weak predictor of sleep disorder status in this sample.

Figure 4. Predicted Probability of Sleep Disorder by sleep duration

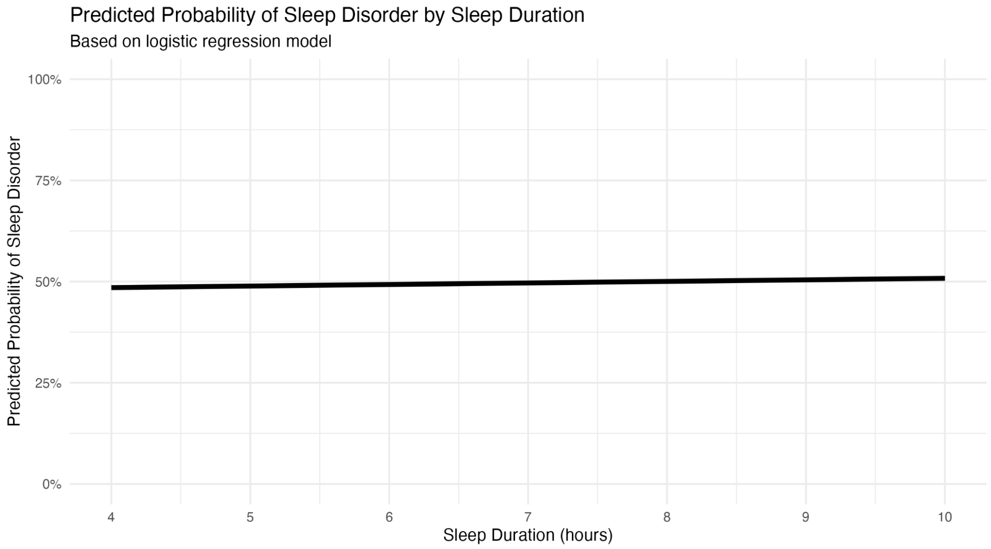
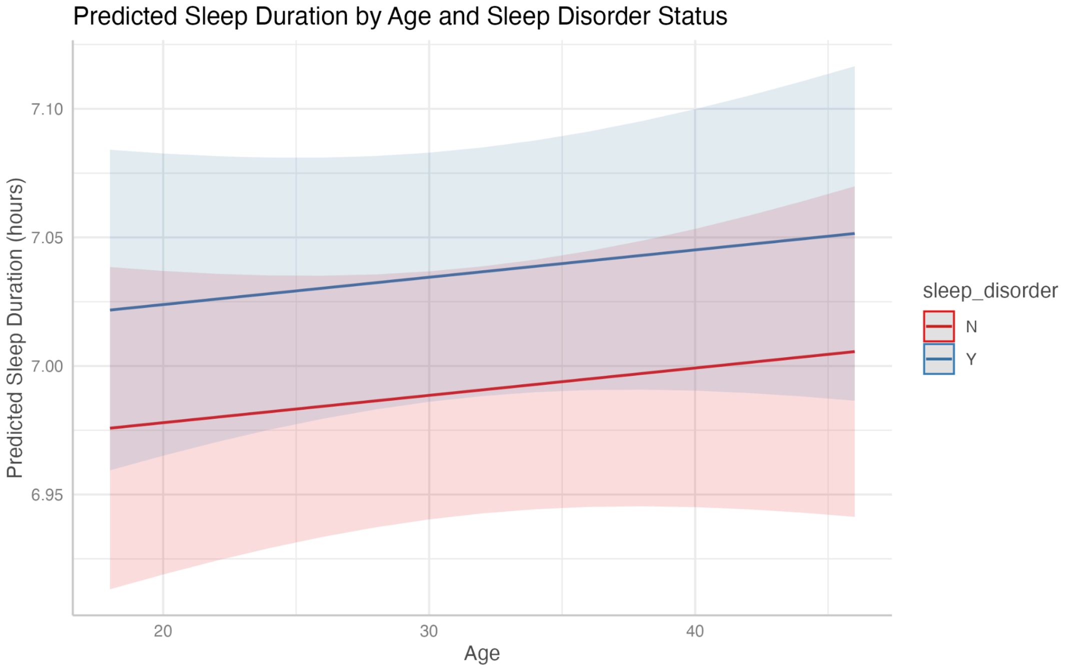


Figure 4 presents the predicted probability of reporting a sleep disorder across varying sleep duration based on a logistic regression model. Results show that predicted probability of a sleep disorder remains relatively stable across the observed range of sleep duration(roughly 4-10 hours). While there is a small upward trend as sleep duration increases, the overall change in predicted probability is minimal, hovering right around the 50% mark. This suggests that within this data set, sleep duration alone is not a strong predictor of sleep disorder status, and factors beyond total hours of sleep likely play a greater role in explaining sleep disorder prevalence.

Figure 5: Sleep Duration by Age and Sleep Disorder



Due to the age range of participants within this self-reported data set, figures 1-4 reveal limited evidence that sleep duration alone meaningfully impacts the likelihood of reporting a sleep disorder. Instead, to further explore this relationship, a multivariable regression model was estimated and used to generate predicted values of sleep duration across age. As shown in Figure 5, predicted sleep duration increases modestly with age for both individuals with and without a reported sleep disorder. This indicates a weak but positive linear association between age and sleep duration. Across the age range, individuals with a sleep disorder are consistently predicted to sleep slightly fewer hours than those without. However, the overlapping confidence intervals suggest that these differences are relatively small. This reinforces the conclusion that sleep duration alone may not be a strong distinguishing factory for sleep disorder prevalence in the sample.

Conclusion

This project demonstrates that sleep is meaningfully associated with a range of health outcomes, though the nature of these relationships differs across physiological systems. Analyses of blood pressure revealed that the association between sleep duration and systolic blood pressure is best characterized as non-linear, with moderate sleep durations with the lowest predicted blood pressure and elevated risk observed at both short and long extremes of sleep. In contrast, the relationship between sleep duration and BMI appears more consistently linear, with shorter sleep associated with higher BMI.

The analysis of sleep duration relative to sleep characteristics extends this framework by illustrating how sleep quality and disruption may influence more localized health conditions. Together the results across datasets suggest that sleep health cannot be reduced to duration alone; rather, both quantity and quality of sleep contribute to health outcomes in distinct ways. Future research could build on this work by incorporating longitudinal data, objective sleep measurements, or experimental designs to better identify causal mechanisms. Overall, this project highlights sleep as a critical component of health and demonstrates the value of combining exploratory visualizations with regression modeling to study complex behavioral risk factors.

References

Hernán M. A. (2018). The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American journal of public health*, *108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Nagra, D. (2025, February 13). *Dry eye disease*. Kaggle. https://www.kaggle.com/datasets/dakshnagra/dry-eye-disease?resource=download

Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P., & Miller, M. A. (2011). Sleep duration predicts cardiovascular outcomes: A systematic review and meta-analysis of prospective studies. *European Heart Journal, 32*(12), 1484–1492. https://doi.org/10.1093/eurheartj/ehr007

Gangwisch, J. E., Heymsfield, S. B., Boden-Albala, B., Buijs, R. M., Kreier, F., Pickering, T. G., Rundle, A. G., Zammit, G. K., & Malaspina, D. (2006). Short sleep duration as a risk factor for hypertension: Analyses of the first National Health and Nutrition Examination Survey. *Hypertension, 47*(5), 833–839. https://doi.org/10.1161/01.HYP.0000217362.34748.e0

Liu, Y., Wheaton, A. G., Chapman, D. P., Cunningham, T. J., Lu, H., & Croft, J. B. (2013). Prevalence of healthy sleep duration among adults — United States, 2014. *Morbidity and Mortality Weekly Report, 65*(6), 137–141.

St-Onge, M. P., Grandner, M. A., Brown, D., Conroy, M. B., Jean-Louis, G., Coons, M., & Bhatt, D. L. (2016). Sleep duration and quality: Impact on lifestyle behaviors and cardiometabolic health. *Circulation, 134*(18), e367–e386. https://doi.org/10.1161/CIR.0000000000000444

ImaginativeCoder. (2023). *Sleep health and lifestyle dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/imaginativecoder/sleep-health-data-sampled

Centers for Disease Control and Prevention. (2018). *National Health and Nutrition Examination Survey (NHANES), 2017–2018* [Data set]. U.S. Department of Health and Human Services. https://www.cdc.gov/nchs/nhanes/

Data: Sleep Health Data NHANES Sleep & Activity Data Sleep Duration