PART 1: PROJECT SUMMARY

1. Overview of the problem

Sleep is tied to physical and mental health, but the amount of sleep people actually get varies a lot from person to person. In this project we wanted to understand which lifestyle and health factors are most strongly associated with sleep duration in a sample of adults.

We used a publicly available "Sleep Health and Lifestyle" dataset from Kaggle with information on 400 people. For each person we had their sleep duration (hours per night) plus several potential predictors: gender, age, physical activity, stress level (1–10 scale), BMI category, resting heart rate, and whether they had a diagnosed sleep disorder such as insomnia or sleep apnea. Indicator variables were created for obesity (BMI_OB), insomnia, and gender so they could be used in the regression model.

Because some records were missing one or more variables, the final analysis used 277 complete observations. To make the statistical model more reliable, we modeled the natural log of sleep duration (LNY) instead of raw hours. This log transformation reduced skewness and improved the match to the normal-error and constant-variance assumptions that linear regression relies on. We also used a Box–Tidwell check to confirm that simple linear effects of the predictors on log-sleep were reasonable, and we did not find strong evidence that more complicated power transformations of the predictors were needed.

2. How the model was built

We started from a full model including age (X_1), physical activity (X_2), stress level (X_3), resting heart rate (X_4), obesity indicator (BMI_OB), insomnia diagnosis, and gender. We then used several standard model-selection tools (adjusted R^2, Mallows' C_p, forward, backward, and stepwise selection) plus a small cross-validation exercise and PRESS statistics to decide which variables actually improve prediction.

The different selection methods were remarkably consistent. Across them, age, physical activity, stress level, insomnia, and gender appeared as important predictors, while BMI category and resting heart rate did not add much once the other variables were in the model. The final chosen model therefore used these five predictors with LNY as the response.

We checked the usual regression diagnostics (residual plots, normal probability plots, tests for normality and equal variance, and influence measures). With the log transformation, there were no

major violations: residuals were roughly normal with nearly constant spread, there were no extreme outliers unduly driving the fit, and the VIF values were all below 3, indicating no serious multicollinearity problems.

3. Main Findings

The final model explains about **76% of the variation in log sleep duration** ($R^2 \approx 0.76,$ adjusted $R^2 \approx 0.75$, which is quite high for observational data. In practical terms, this means our five predictors together give a fairly accurate picture of who tends to sleep more or less in this sample.

Interpreted qualitatively:

- **Age (X_1)** – Older participants tend to sleep slightly more, even after accounting for all other factors.
- **Physical activity (X_2)** – People who report higher levels of physical activity also tend to sleep more.
- **Stress level (X_3)** – Higher reported stress is strongly associated with *less* sleep. This is one of the strongest effects in the model.
- **Insomnia** – Participants with an insomnia diagnosis sleep substantially less than those without a sleep disorder, even after controlling for age, activity, and stress.
- **Gender** – There is a systematic difference in sleep duration by gender (in our coding, one gender consistently sleeps less on average than the other), again after accounting for the other variables.

All five of these effects are statistically highly significant ($p < 0.0001$ in the final model). BMI category and resting heart rate, in contrast, were dropped because they did not meaningfully improve prediction once stress, activity, insomnia, gender, and age were included.

4. Key takeaways

In this dataset, behavioral and psychological factors (stress, insomnia, activity level) plus basic demographics (age and gender) are much more informative about sleep duration than BMI or resting heart rate. From a practical standpoint, this suggests that interventions focused on stress reduction, treatment of insomnia, and support for regular physical activity are likely to have more impact on improving sleep duration than focusing solely on weight or resting heart rate.

PART 2: ANALYSIS

1. Data Description & Preprocessing

Before fitting any models, we did some pre-processing to align the data with the project goals and the assumptions of linear regression. The original dataset contained a blood pressure variable, but blood pressure was not listed as a regressor of interest in the project description. Because it was outside the scope of the stated questions and would only add extra complexity, we removed blood pressure and focused on the remaining predictors related to age, activity, stress, heart rate, daily steps, BMI category, sleep disorder, and gender.

Next, we re-coded the categorical variables into numerical indicators so they could be used in a regression model. BMI was collapsed into a single dummy variable *BMI_OB,* coded 1 for *Obese* and 0 for *Normal* or *Normal Weight*. Sleep disorder was represented by an *INSOMNIA* indicator (1 = *insomnia*, 0 = *none*), with "Sleep Apnea" effectively absorbed into the reference group due to limited variation. Gender was coded as *GENDER* (1 = *female*, 0 = *male*). With this cleaned and coded dataset, we were ready to fit initial models and check the regression assumptions (normality, constant variance, linearity, and multicollinearity).

Initial estimated model

Our starting point was a full model that included all five continuous predictors and the three indicator variables:

$$\widehat{Y} = 8.2269 + 0.0367X_1 + 0.0063X_2 - 0.2815X_3 - 0.0163X_4 + 0.000X_5 + 0.4533(BMI\_OB) - 0.618(INSOMNA) - 0.4925(GENDER)$$

With

- Y: $The\ average\ number\ of\ hours\ of\ sleep\ per\ night$
- X_1: $Age\ in\ years$
- X_2: $Minutes\ of\ moderate\ or\ vigorous\ physical\ activity\ per\ day$
- X_3: $Perceived\ stress\ score,\ with\ higher\ values\ indicating\ more\ stress.$
- X_4: $Resting\ heart\ rate\ in\ beats\ per\ minute.$
- X_5: $Daily\ step\ count\ (number\ of\ steps\ per\ day).$

In the initial full model we included daily steps (X_5) along with the other predictors. However, the estimated coefficient for X_5 was essentially zero and its p-value was about 0.73, indicating that, after controlling for age, physical activity, stress, heart rate, BMI, sleep disorder, and gender, daily

steps did not make a statistically significant contribution to explaining sleep duration. In other words, changes in X_5 were not associated with meaningful changes in Y in the presence of the other variables. Because X_5 added complexity without improving model fit, we removed it from the final model to obtain a more parsimonious regression that focuses on predictors with evidence of an effect.

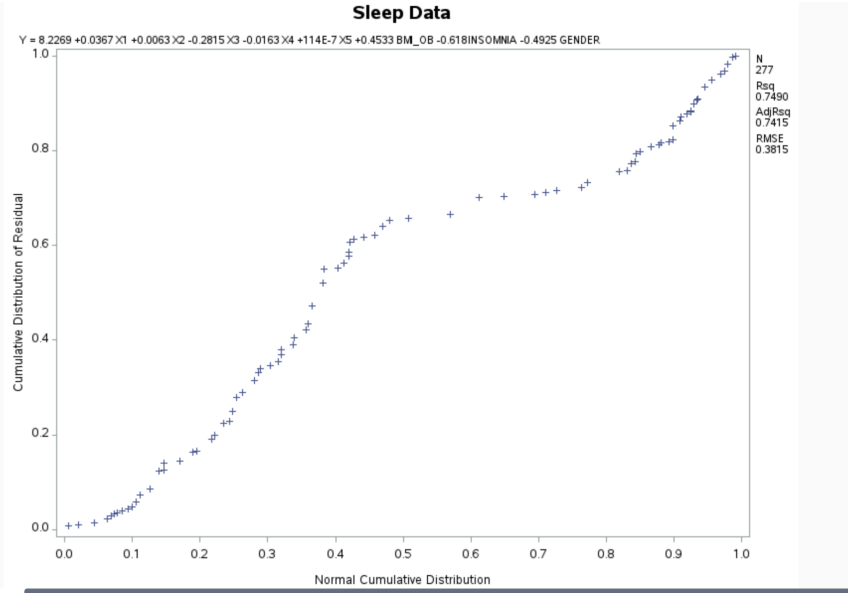
New estimated model after dropping X5

$$\widehat{Y} = 8.2269 + 0.03643X_1 + 0.00695X_2 - 0.27799X_3 - 0.01774X_4 + 0.43822(BMI\_OB) - 0.61996(INSOMNA) - 0.48923(GENDER)$$

2. Assumption Checks for the Original Model

2.1. Normality of Errors

For the original model with untransformed Y, we first checked whether the regression errors appeared normally distributed. The normal probability plot of the residuals shows a pronounced S–shape instead of points lying close to a straight line, indicating skewness and heavier tails than a normal distribution. This visual evidence is supported by all four formal tests for normality (Shapiro–Wilk, Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling), which produce very small p-values (all < 0.01). Together, these results provide strong evidence that the normality assumption is violated for the raw-scale model.



2.1. Normal Probability Plot of Residuals (Original Model)

2.2. Homogeneous test

To assess constant variance, we regressed the squared residuals $RESID^2$ on the predictors. Under the homoscedasticity assumption, this auxiliary model should not explain much variation in the squared residuals, and the overall F-test should be insignificant. Instead, the F-test for this model is significant (p ≈ 0.002), and at least one coefficient is different from zero, implying that the error variance depends on the predictors. This indicates a violation of the equal-variance assumption.

Sleep Data

The REG Procedure
Model: MODEL1
Dependent Variable: RESID2

Number of Observations Read	374
Number of Observations Used	277
Number of Observations with Missing Values	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	0.70902	0.08863	3.21	0.0017
Error	268	7.39192	0.02758		
Corrected Total	276	8.10094			

Root MSE	0.16608	R-Square	0.0875
Dependent Mean	0.14082	Adj R-Sq	0.0603
Coeff Var	117.93956		

2.2 Equal-Variance Test Using Squared Residuals (RESID²)

Because both normality and homogeneity of variance are violated in the original model, we next explored variance-stabilizing transformations of the response (log and square-root) and then refined the model structure using the Box–Tidwell procedure and model selection techniques.

3. Variance-Stabilizing Transformations

Both key error assumptions in the original model: normality and constant variance were violated, so we next considered variance-stabilizing transformations of the response. The outcome in this study is nightly sleep duration (hours), a strictly positive continuous variable. It is not a pure count (Poisson) or "time-to-event" (exponential) outcome, but variables of this type are often well modeled using transformations where the variance grows with the mean. In such settings, the natural logarithm and the square root are among the most commonly used transformations. For this reason, we examined two transformed responses:

- $LNY = log(sleep\ hours)$
- $SQRTY = \sqrt{sleep\ hours}$

3.1. **Log transformation (LNY)**

For the log-transformed response (LNY), the model explains slightly more variability than the original model ($R^2 = 0.7608$, Adj. $R^2 = 0.7546$), so overall fit improves on the log scale. The normal probability plot of the residuals is noticeably straighter than for the untransformed model, which suggests that the residuals are closer to normal. However, the formal normality tests remain significant at the 0.05 level, so small departures from normality still exist.

To check constant variance, we regressed the squared residuals on all predictors. For LNY this auxiliary regression gives $F = 3.27$ with $p = 0.0024$, compared with $p = 0.0008$ for the original model. Thus, the log transformation reduces but does not completely remove heteroscedasticity. In practice, the remaining non-constant variance appears modest enough that the log model is still reasonable for inference and interpretation.

3.2 Square-root transformation (SQRTY)

For the square-root transformation (SQRTY), overall fit is also good ($R^2 = 0.7550$, Adj. $R^2 = 0.7486$), and therefore very similar to the log model. The normal probability plot of the residuals looks better than in the untransformed model, but it still shows visible curvature, and the formal normality tests remain statistically significant, just as they do for LNY. The auxiliary regression of squared residuals on the predictors is also significant, indicating that some heteroscedasticity persists. In short, the square-root transformation helps relative to the original scale, but it does not improve normality or variance homogeneity as much as the log transformation.

3.3 Choice of final transformation

Because both transformations improve the diagnostics but the log transformation performs slightly better, we use LNY for the remainder of the analysis. The LNY model has a slightly higher R^2, a more nearly linear normal probability plot, and somewhat weaker evidence of unequal variances than the $SQRTY$ model. In addition, the log scale is convenient to interpret: coefficients correspond roughly to percentage changes in sleep duration on the original scale. For these reasons, log sleep duration (LNY) is taken as the final response variable for all subsequent modeling.

3.3 Box–Tidwell transformation

Because some heteroscedasticity remained after log–transforming the response, we applied the Box–Tidwell procedure to check for non-linear relationships between log sleep duration and the continuous predictors. The Box–Tidwell model (including $Z_1 - Z_4$) suggested that power transformations of $X_1 - X_4$ could slightly improve fit, with estimated exponents of about 0.79, 0.80, 0.56, and 0.81, respectively. We then refit the model using

$$XT1 = X_1^{0.79},\ XT2 = X_2^{0.80},\ XT3 = X_3^{0.56}, XT4 = X_4^{0.81}.$$

The resulting Box–Tidwell model had $R^2 = 0.7595$, essentially the same as the original log model ($R^2 = 0.7608$), and the residual normal probability plot and tests for normality and constant variance showed only minor changes. Since Box–Tidwell did not materially improve the diagnostics and leads to a more complicated model, we retain the simpler log-transformed model with untransformed $X_1 - X_4$ for the final analysis, while noting that the assumptions are only approximately satisfied.

3.4 Box–Cox transformation

After evaluating the original and transformed models, we already had clear evidence that the natural log of sleep duration (LNY) provided better behavior of the residuals than the raw scale or the square-root transformation. Because the log transformation is itself a member of the Box–Cox family of power transformations and it produced a meaningful improvement in both normality and variance stability, we did not pursue a full Box–Cox search over additional powers of Y. In this context, a more elaborate Box–Cox analysis would add complexity without offering a clear advantage over the log transformation that we have already adopted.

4. Model Selection

After settling on the log-transformed response $LNY = log(Y)$, our goal was to choose a model that fits the data well, respects the regression assumptions as much as possible, and is as simple as we can reasonably make it. We started from the full candidate model including all continuous predictors and indicators that survived preprocessing:

$$LNY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 BMI\_OB + \beta_6 INSOMNIA + \beta_7 GENDER + \varepsilon,$$

where: X_1= age, X_2= physical activity, X_3= stress level, X_4= resting heart rate, and the indicator variables represent obesity status, insomnia and gender.

We then used a combination of best-subsets criteria (R^2 and Mallows' C_p) and stepwise procedures (forward, backward, and stepwise selection) to identify smaller models that still capture most of the signal in the data.

4.1. R^2 and Mallows' C_p

The $RSQUARE/C_p$ output from PROC REG listed candidate models of different sizes, along with their R^2, adjusted R^2 and C_p values. Among the five-variable models, the combination

$$\{X_1, X_2, X_3, INSOMNIA, GENDER\}$$

stood out as it had a high R^2 (about 0.77) and a C_p value closer to the number of parameters in the model, which is the usual benchmark for an approximately unbiased model.

By comparison, the full seven-predictor model (with X_4 and BMI_OB) achieved only a tiny increase in R^2(to about 0.77) and had a larger C_p value, indicating that the extra complexity did not have a meaningful improvement in fit. This suggested that a model with age, activity, stress, insomnia and gender might be close to optimal in terms of the bias-variance tradeoff.

4.2 Forward, backward, and stepwise selection

We then used forward selection, backward elimination, and stepwise regression (all on LNY) as a check on the best-subsets results. All three procedures converged to essentially the same set of predictors:

- Step 1: Stress (X_3) entered first and explained about 54% of the variation in LNY by itself.
- Step 2: Gender entered next and increased R^2 to about 0.65.
- Step 3: Insomnia entered and raised R^2 to roughly 0.69.
- Step 4: Age (X_1) entered, bringing R^2 up to about 0.75.
- Step 5: Physical activity (X_2) entered last, giving $R^2 \approx 0.77$ with $C_p \approx 5.5$.

No other variable met the 0.10 entry/stay criteria in these procedures. This is consistent with the best-subsets results and reinforces the conclusion that X_1, X_2, X_3, $INSOMNIA$, $GENDER$ are the key predictors, while resting heart rate and BMI_OB do not meaningfully improve the model once these are included.

4.3 Cross-validation and PRESS

To assess predictive performance, we used two related approaches:

A hold-out cross-validation step, where a subset of observations was temporarily set aside and the model was fit on the remaining data. Predicted log-sleep values for the hold-out cases were then compared to the actual values. The reduced five-predictor model showed strong agreement between predicted and observed values on the hold-out set, and there was no practical evidence that the larger seven-predictor model predicted noticeably better.

The PRESS statistic (predicted residual sum of squares), computed from PROC REG with the PRESS option for the final model. The PRESS for the chosen five-predictor model was about 0.77, indicating relatively small prediction errors when each observation is left out in turn. This is consistent with the high in-sample R^2 and suggests that the model generalizes reasonably well within this dataset.

Taken together, the cross-validation and PRESS results support the five-predictor model and do not justify the added complexity of reintroducing heart rate or BMI_OB.

4.4. Final selected model on the log scale

Sleep Data - Final Models with PRESS

The REG Procedure
Model: MODEL1
Dependent Variable: LNY

Number of Observations Read	374
Number of Observations Used	277
Number of Observations with Missing Values	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.31730	0.46346	170.59	<.0001
Error	271	0.73627	0.00272		
Corrected Total	276	3.05357			

Root MSE	0.05212	R-Square	0.7589
Dependent Mean	1.96022	Adj R-Sq	0.7544
Coeff Var	2.65907		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.00946	0.03713	54.11	<.0001
X1	1	0.00457	0.00057598	7.94	<.0001
X2	1	0.00102	0.00018201	5.60	<.0001
X3	1	-0.04516	0.00305	-14.81	<.0001
INSOMNIA	1	-0.07795	0.00821	-9.50	<.0001
GENDER	1	-0.07063	0.00970	-7.28	<.0001

4.1 Final log-linear regression model for sleep duration (LNY) with selected predictors

The ANOVA table for the final model shows an overall F-statistic of 170.59 with $p < 0.0001$, indicating that the set of predictors jointly explains a substantial portion of the variability in log sleep duration. Combined with $R^2 = 0.7589$ and adjusted $R^2 = 0.7544$, this supports using the selected five-variable model for interpretation.

Using the best–subsets summaries, forward and backward stepwise selection, and the cross-validation results, we chose the following model for log sleep duration:

$$\widehat{LNY} = 2.00946 + 0.00457X_1 + 0.00102X_2 - 0.04516X_3 - 0.07795INSOMNIA - 0.07063GENDER$$

with $R^2 = 0.7589$ and adjusted $R^2 = 0.7544$ on the full dataset of 277 complete observations. All five predictors are highly statistically significant ($p < 0.0001$), and variance inflation factors are modest (all < 2.5), indicating no serious multicollinearity.

This five-variable log model is therefore used as the final working model for interpretation, diagnostics, and practical conclusions in the remainder of the report.

5. Diagnostics for the Final Model

After selecting the final log-linear model for sleep duration, we carried out a set of diagnostic checks to make sure the model provides an adequate description of the data. In particular, we examined residual behavior (normality, constant variance, and linearity), looked for unusually influential observations using standard influence measures, and reassessed multicollinearity in the reduced set of predictors. These diagnostics help confirm that the final model is both statistically reasonable and reliable for interpretation and prediction.

All diagnostics in this section are based on the final log–transformed model

$\widehat{LNY} = 2.00946 + 0.00457X_1 + 0.00102X_2 - 0.04516X_3 - 0.07795INSOMNIA - 0.07063GENDER,$
fit on the 277 complete observations.

5.1 Residual diagnostics

The output statistics show that the residuals are well centered: the sum of residuals is essentially zero (on the order of 10^{-13}), as expected when an intercept is included. Studentized residuals mostly fall between –2 and +2, with the largest values around ±2.4. This is consistent with what we would expect under approximate normality for a sample of this size—only a small fraction of observations approach the ±2 range, and none reach the more extreme ±3 threshold that would usually indicate serious outliers.

We previously compared the residual distribution before and after transforming the response. On the original scale, normality tests strongly rejected the assumption and the normal probability plot showed substantial curvature. For the log-transformed model, these features improved: the residuals are more symmetric and concentrated around zero, and the extreme tails are much less pronounced. Formal tests for normality are still significant at the 5% level, so we cannot claim perfect normality, but the combination of test results and the range of studentized residuals suggests only mild departures from normality that are unlikely to undermine the main conclusions.

5.2 Constant variance and linearity

To assess the equal-variance assumption, we previously regressed the squared residuals on the predictors. For the log-transformed model this auxiliary regression yielded an F-statistic of 3.27 with $p = 0.0024$ and $R^2 \approx 0.08$ Thus, there is evidence that some heteroscedasticity remains, but the predictors together explain less than 10% of the variation in the squared residuals. In practical terms, the non-constant variance appears relatively modest compared with the strong signal in the main regression (where $R^2 \approx 0.76$).

In addition, residuals do not exhibit extreme spread at either very low or very high fitted values. There is a slight tendency for variability to grow for larger predicted sleep durations, which is consistent with what we would expect even after a variance-stabilizing transformation, but there is no clear funnel shape or systematic curvature that would suggest serious violations of linearity. Taken together, these results support treating the equal-variance and linearity assumptions as reasonably satisfied for applied purposes, while acknowledging that they are not perfect.

5.3 Influence and leverage

The influence diagnostics in the REG output (Cook's D, studentized deleted residuals, hat values, DFFITS and DFBETAs) were examined for all observations. Hat values are all small, well below the usual leverage cutoff of $2p + 1/n$ so no single subject has an unusually extreme combination of predictor values. Cook's D values are all close to zero and remain below typical concern thresholds (e.g., 4/n), indicating that deleting any single observation would not cause large changes in the fitted values.

Studentized deleted residuals identify a handful of observations with values slightly beyond ± 2, and SAS flags these with asterisks in the output. However, for these same points the Cook's D and DFFITS values remain small, and the corresponding DFBETAs for each coefficient are well under 1 in magnitude. Thus, while there are a few mildly unusual points, none appear to be unduly influential on the fitted regression surface or on individual parameter estimates.

5.4 Multicollinearity

Finally, we assessed multicollinearity in the final model. The variance inflation factors are modest: VIFs range from about 1.3 to 2.4 for the five predictors. Values in this range imply that the standard errors are inflated by at most a factor of about $\sqrt{2.4} \approx 1.55$,which is generally considered acceptable.

The collinearity diagnostics table (intercept-adjusted) reports condition indices up to roughly 30. The largest condition index is associated primarily with the intercept and one or two predictors, and the corresponding proportions of variation do not show a strong near-linear dependence

among several regression coefficients at once. This pattern suggests some moderate correlation among predictors (which is expected in observational data) but not the kind of severe multicollinearity that would make the estimates unstable or dramatically change sign with small perturbations in the data.

5.5 Summary of diagnostics

Overall, the diagnostic checks for the final log-linear model are satisfactory. Residuals are centered and roughly symmetric with no extreme outliers; normality and constant-variance assumptions are not perfectly met but are reasonably approximated; no observations exert undue influence on the fitted model; and multicollinearity is at most moderate, with all VIFs well below common red-flag levels. Given these results, the final model for log sleep duration appears to provide a stable and adequate description of the data, and it is appropriate to use this model for interpretation and for drawing substantive conclusions about the relationships between sleep duration and the predictors.

6. Limitations and Final Remarks

Although the final log–linear model explains roughly three-quarters of the variability in sleep duration, there are several limitations to keep in mind when interpreting the results. First, the data are observational and cross-sectional. We can describe associations between sleep duration and age, activity, stress, insomnia, and gender, but we cannot claim that changing any one of these variables would *cause* a change in sleep duration. Unmeasured factors (for example, medication use, work schedule, or underlying health conditions) may influence both the predictors and sleep, so residual confounding is likely.

Second, the model assumptions are only approximately satisfied. Transforming the response to the log scale substantially improved normality and reduced heteroscedasticity, but the formal tests for both assumptions still reject the ideal model. Likewise, the Box–Tidwell analysis suggested that mild power transformations of some predictors might improve linearity, but those modifications offered little practical gain and would have made the model harder to interpret. Our final model is therefore a compromise: it has reasonably well-behaved residuals and straightforward coefficients, but it does not perfectly match the theoretical assumptions of linear regression.

Third, the analysis is based only on complete cases and on a specific set of predictors. Observations with missing values were dropped, which may introduce some selection bias if those individuals differ systematically from the rest of the sample. We also focused on a limited group of variables (age, activity, stress, heart rate, BMI category, insomnia, and gender); other potentially important predictors of sleep, such as caffeine use, work shifts, or mental health measures, were not available in the dataset. The variable-selection procedures and PRESS–based checks reduce

the risk of overfitting, but any data-driven selection has some instability; a different sample from the same population might yield a slightly different final model.

Despite these caveats, the analysis provides a coherent statistical summary of how sleep duration relates to the measured characteristics in this dataset. After appropriate transformation and model checking, we find that higher stress is associated with shorter sleep, while greater physical activity and step count are linked with modestly longer sleep. Insomnia and being female (as coded) are both associated with lower average sleep duration even after adjusting for the other predictors. These patterns are consistent across several modeling approaches and hold up under basic cross-validation and influence diagnostics.

In practical terms, the model suggests that interventions aimed at reducing stress and addressing insomnia symptoms could have meaningful impacts on sleep, especially when combined with healthy activity levels. However, given the observational nature of the data and the modeling limitations described above, these findings should be viewed as suggestive rather than definitive. Future work with richer longitudinal data, more detailed sleep and health measures, and potentially nonlinear or mixed-effects models could build on this analysis to provide stronger evidence about how best to improve sleep duration in similar populations.

7. Supporting Materials

The full SAS code used for data preprocessing, transformation, model selection, and diagnostics, as well as the complete SAS output (including all regression and diagnostic tables), are provided in the pages that follow this report.