1.Motivation

Nutrition and sleep are two of the most fundamental components of human well-being, and they are closely interconnected. Poor dietary habits may disrupt sleep patterns, while insufficient or low-quality sleep can negatively influence eating behavior. Despite widespread assumptions about the importance of healthy eating for better sleep, empirical evidence—especially among university students—remains limited.

University students often experience irregular meal timing, high caffeine consumption, and inconsistent sleep schedules due to academic pressure and lifestyle factors. This project aims to investigate whether healthier eating habits are actually associated with improved sleep quality and sleep duration in a student population. Rather than relying on intuition, the project adopts a data-driven approach using statistical analysis and machine learning techniques to explore these relationships.

2. Data Source

The dataset used in this project consists of self-reported survey responses collected from **26 university students**. The data was obtained through questionnaires designed to capture participants' dietary habits, sleep characteristics, and selected lifestyle behaviors. The choice of a self-reported survey approach allows the collection of subjective perceptions of sleep quality and eating habits, which are particularly relevant when studying well-being-related outcomes.

The dataset includes three main types of variables. First, **diet-related variables** measure participants' perceived diet quality and related nutritional behaviors, summarized through a healthy eating score. Second, **sleep-related variables**capture both quantitative and qualitative aspects of sleep, including sleep duration and subjective sleep quality. Third, **lifestyle variables** such as caffeine intake, hydration level, exercise frequency, and screen time before bed are included to account for external factors that may influence sleep outcomes beyond diet alone.

Although the dataset is relatively small and relies on self-reported information, it reflects realistic daily habits of university students. This makes it suitable for exploratory analysis, hypothesis testing, and small-scale machine learning experiments aimed at understanding patterns rather than making population-level generalizations.

3. Data Analysis

3.1 Data Cleaning and Preparation

The first stage of the analysis involved transforming the raw survey data into a structured and analyzable format. Variables were converted to appropriate numeric types, and missing values were handled using systematic imputation strategies to preserve as much information as possible.

To facilitate analysis, two binary categorical variables were derived:

- **healthy_diet**, defined as 1 if the diet_score exceeds 6 and 0 otherwise
- **good_sleep**, defined as 1 if the sleep_score exceeds 6 and 0 otherwise

These derived variables enabled both group-based statistical comparisons and supervised machine learning classification.

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to visually inspect relationships and distributions within the dataset. A combination of scatter plots, histograms, boxplots,

and correlation heatmaps was used to examine potential associations between dietary quality, sleep outcomes, and lifestyle factors.
The EDA stage revealed substantial variability across participants in both diet scores and sleep measures. However, visual inspection did not indicate a strong or consistent trend linking healthier eating habits to improved sleep quality. In contrast, sleep duration appeared to show a clearer association with sleep quality, suggesting that sleep-related variables may play a more prominent role than diet alone.

3.3 Statistical Analysis

To formally evaluate relationships suggested by the exploratory analysis, multiple hypothesis tests were performed. Pearson correlation coefficients were calculated to assess linear relationships between continuous variables, while two-sample t-tests were used to compare sleep quality between healthy and unhealthy diet groups.
All tests were conducted using a significance level of **$\alpha = 0.05$**. The results indicated that neither diet category nor dietary quality showed a statistically significant association with sleep quality. These findings suggest that, within this dataset, dietary habits alone do not exert a strong measurable effect on sleep outcomes.

4. Machine Learning Analysis

4.1 Problem Formulation

To complement the statistical analysis, a supervised machine learning approach was employed. While statistical tests focus on identifying significant relationships at the group or variable level, machine learning aims to assess whether sleep quality can be **predicted** at the individual level based on available features.
The task was formulated as a **binary classification problem**, where the goal was to predict whether a participant has good sleep quality (good_sleep) using dietary and lifestyle variables as inputs. This framing allows for evaluating the practical usefulness of the data in making predictions, even in the absence of statistically significant associations.

4.2 Feature Selection and Preprocessing

The input features included diet score, sleep duration, caffeine intake, hydration level, exercise frequency, screen time, and calorie intake. To ensure consistency and fairness across models, a unified preprocessing pipeline was implemented.
This pipeline consisted of:

- Imputation of missing numerical values using the median
- Feature scaling for numerical variables
- Train-test splitting to evaluate generalization performance

Categorical and numerical preprocessing steps were handled systematically within a pipeline structure, following best practices in applied machine learning.

4.3 Models Used

Two supervised learning models were employed:

1. **Logistic Regression**
   Logistic Regression was selected as a baseline model due to its interpretability and suitability for binary classification tasks. It provides insight into the linear

contribution of each feature to the prediction outcome and serves as a standard benchmark for comparison.

2. **Random Forest Classifier**
   A Random Forest model was implemented as a more flexible, non-linear alternative. By combining multiple decision trees, Random Forest can capture interactions and non-linear relationships that simpler models may fail to detect.

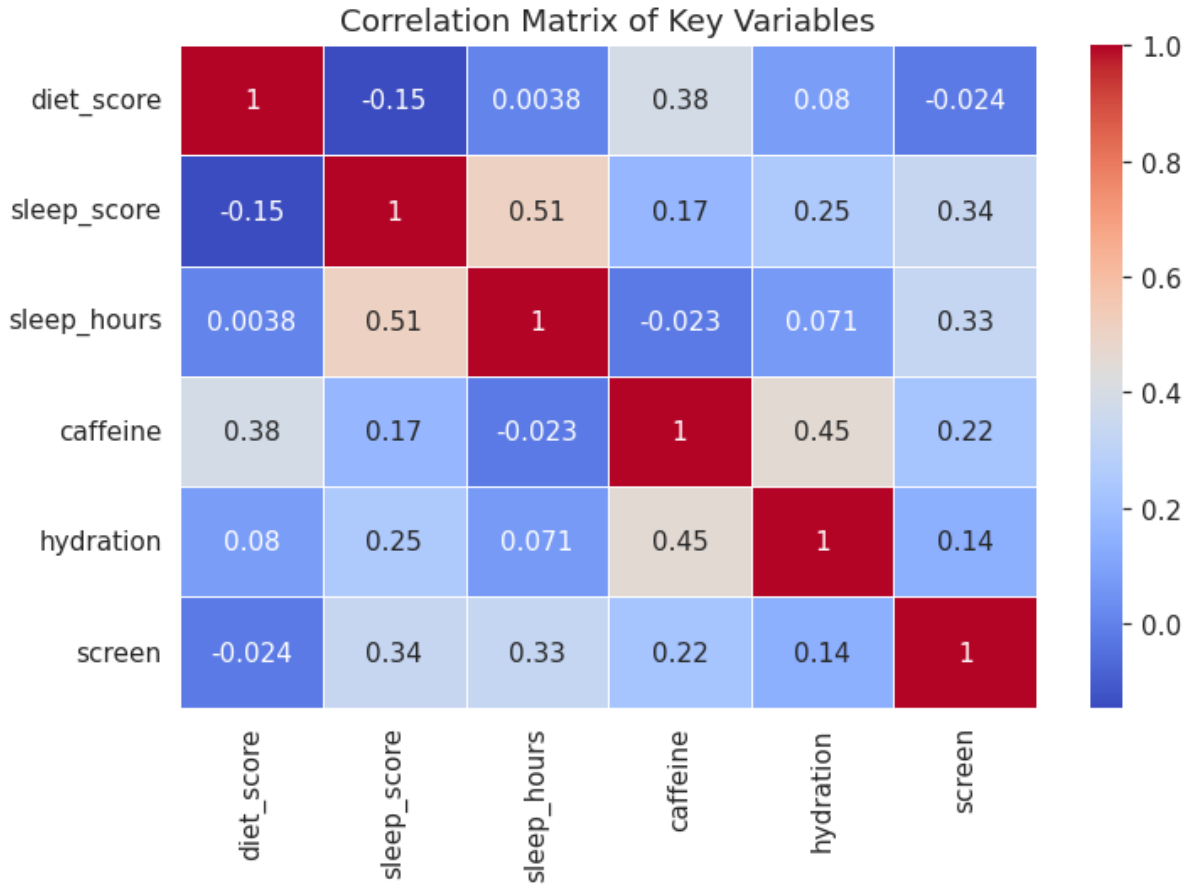
Using both models allows for evaluating whether increased model complexity leads to improved predictive performance.

4.4 Model Evaluation

Both models were evaluated using the same train-test split and performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.
The evaluation results showed that Logistic Regression and Random Forest achieved comparable performance. The Random Forest model did not provide a substantial improvement over the baseline Logistic Regression, indicating that the underlying relationships in the dataset are likely weak or approximately linear.
This outcome suggests that dietary and lifestyle variables, as measured in this dataset, contain limited predictive signal for sleep quality classification.

5. Findings

The findings from statistical and machine learning analyses converge on a consistent conclusion. Statistical tests reveal no significant relationship between diet category and sleep quality, while machine learning models demonstrate limited ability to predict sleep quality outcomes based on dietary and lifestyle features.
Notably, sleep duration emerges as a more relevant variable than diet quality when explaining variations in sleep quality. However, even this relationship does not translate into strong predictive power in the classification setting.
These findings highlight the complexity of sleep behavior and suggest that diet alone may not be a dominant factor in determining sleep quality among university students.



Correlation Matrix of Key Variables

This correlation matrix presents the Pearson correlation coefficients among key variables related to eating habits, sleep characteristics, and lifestyle factors, including diet score, sleep quality, sleep duration, caffeine intake, hydration level, and screen time.

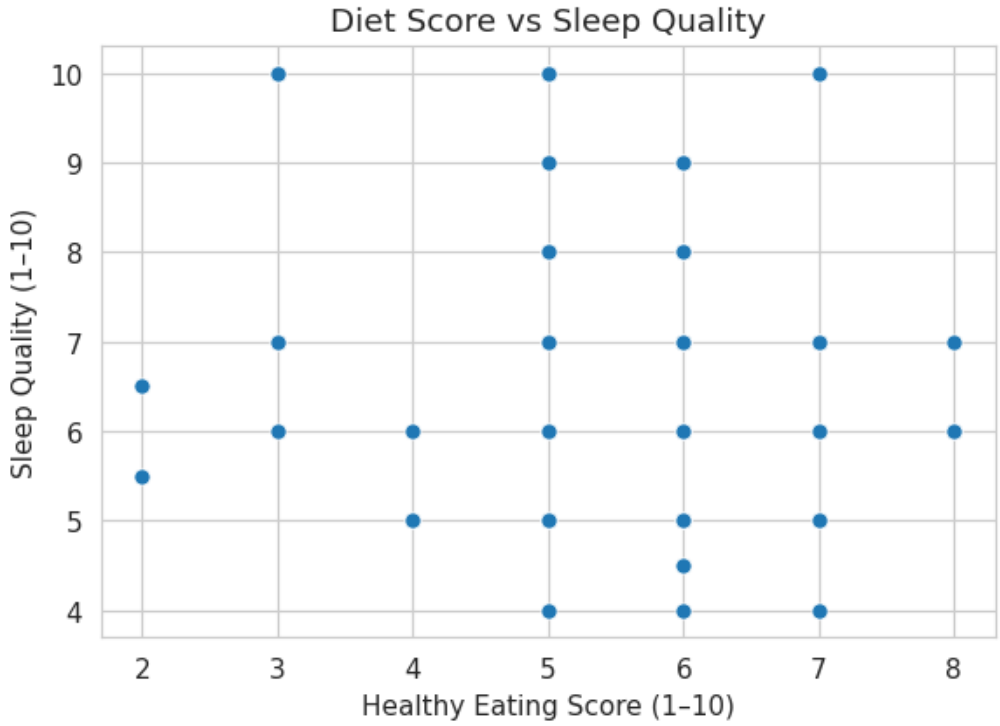
The strongest positive correlation is observed between **sleep quality and sleep duration (r = 0.51)**, indicating that participants who sleep longer hours tend to report higher sleep quality. This relationship is the most pronounced association in the matrix and aligns with general expectations about sleep behavior.

A moderate positive correlation is also found between **caffeine intake and hydration level (r = 0.45)**, suggesting that individuals who consume more caffeine also tend to drink more water. This may reflect compensatory hydration behavior among participants with higher caffeine consumption.

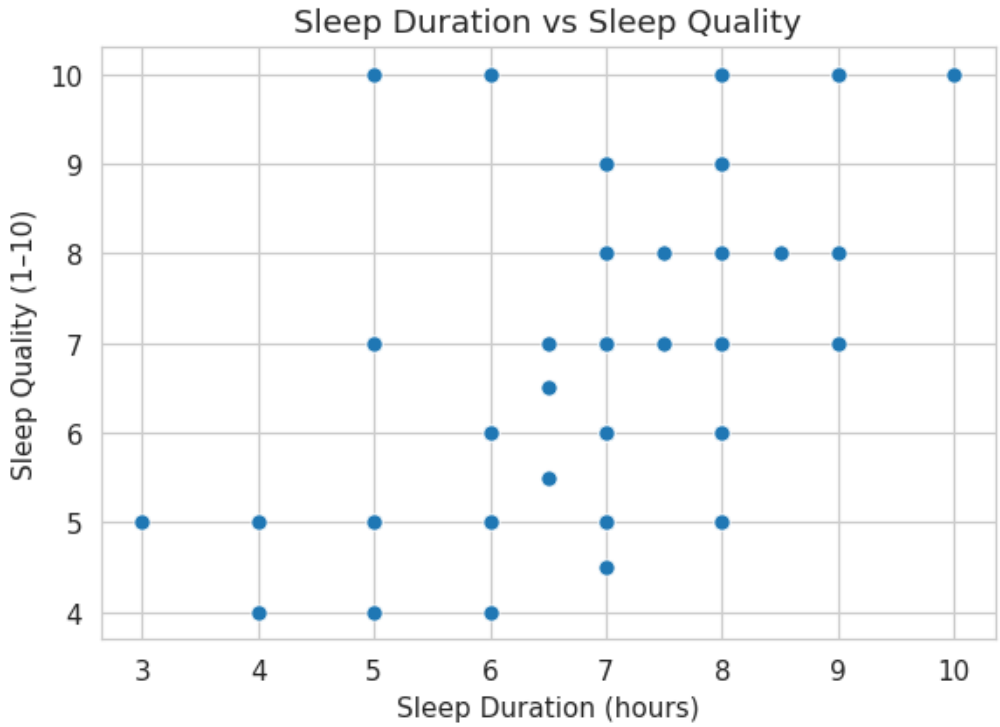
The relationship between **diet score and sleep quality is weak and slightly negative (r = −0.15)**, indicating no meaningful linear association between healthier eating habits and perceived sleep quality within this dataset. Similarly, diet score shows a near-zero correlation with sleep duration ($r \approx 0.00$), reinforcing the lack of evidence for a strong diet–sleep link.

Screen time displays modest positive correlations with **sleep quality (r = 0.34)** and **sleep duration (r = 0.33)**. While this may appear counterintuitive, these correlations are relatively small and may be influenced by confounding factors or self-reporting bias rather than indicating a direct causal relationship.

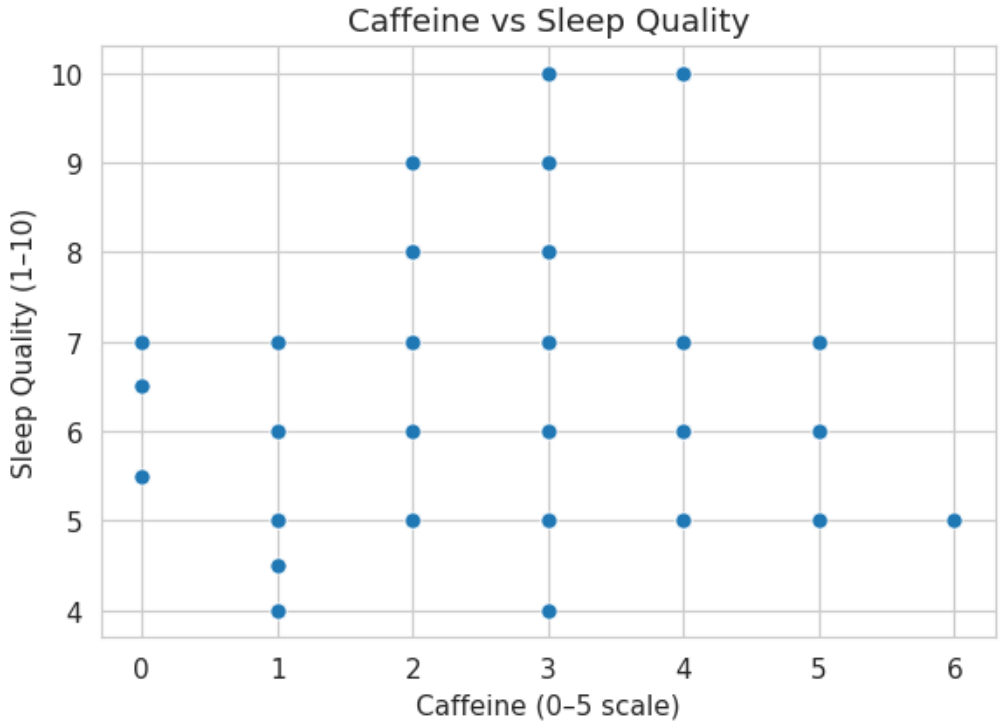
Overall, most correlations in the matrix are weak to moderate, with no strong associations between dietary quality and sleep outcomes. This visualization supports the statistical and machine learning findings of the project, which suggest limited linear relationships between eating habits and sleep quality in the analyzed sample.



Diet Score vs Sleep Quality
This scatter plot shows the relationship between healthy eating score and self-reported sleep quality. The points are widely dispersed and no clear trend is observed, indicating a weak relationship between diet score and sleep quality. This visual finding is consistent with the correlation analysis, which suggests no statistically significant association between healthier eating habits and sleep quality in this dataset.

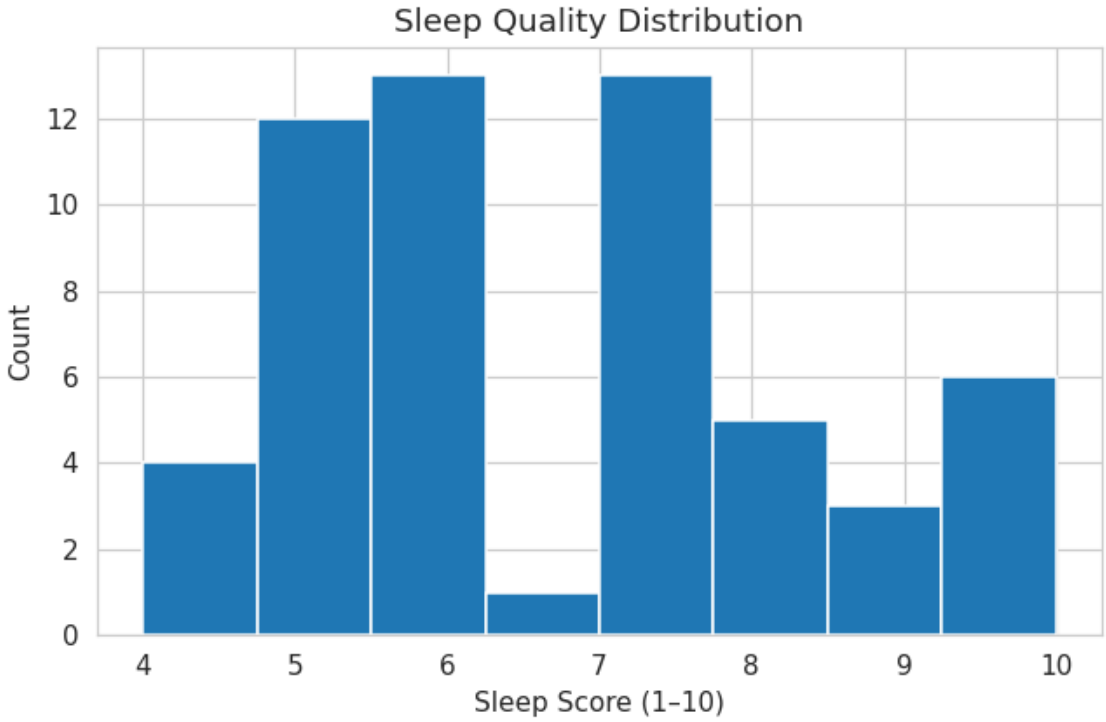


Sleep Duration vs Sleep Quality
This scatter plot illustrates the relationship between sleep duration and self-reported sleep quality. A moderate positive pattern is visible, suggesting that participants who sleep longer hours tend to report better sleep quality. This observation is consistent with the correlation analysis, which indicates a positive association between sleep duration and sleep quality.



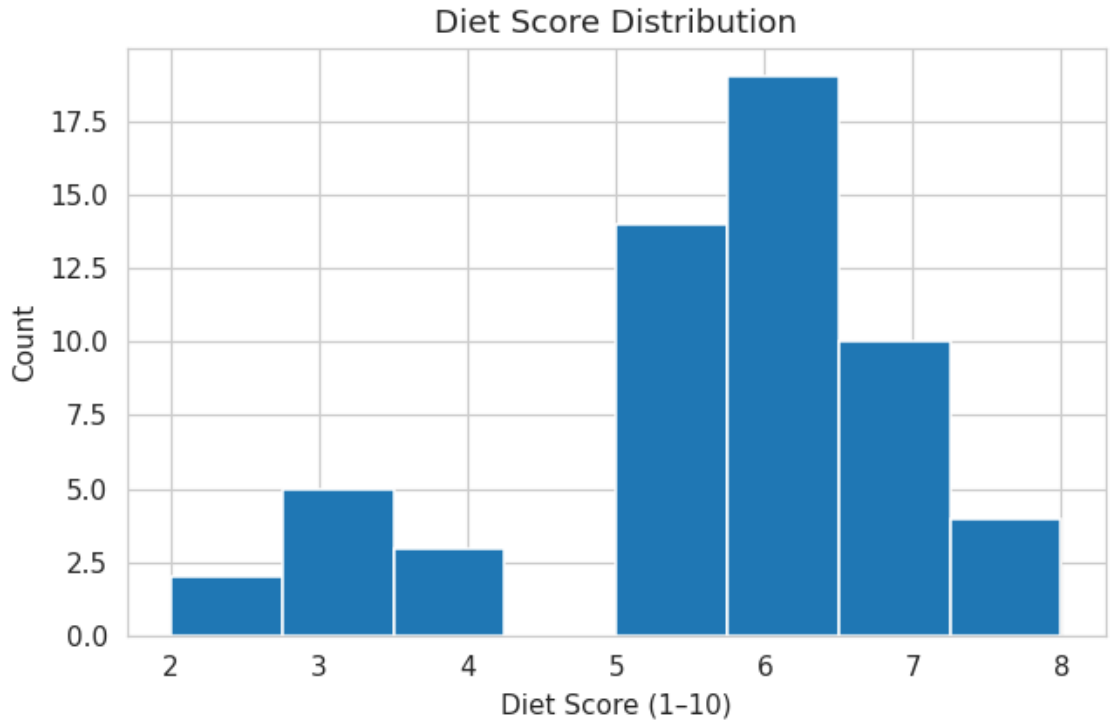
Caffeine Intake vs Sleep Quality

This scatter plot shows the relationship between caffeine intake and sleep quality. The data points are widely scattered with no clear downward trend, suggesting that caffeine intake does not show a strong measurable effect on sleep quality in this sample. This visual observation is consistent with the correlation analysis, which indicates a weak association between caffeine consumption and sleep quality.



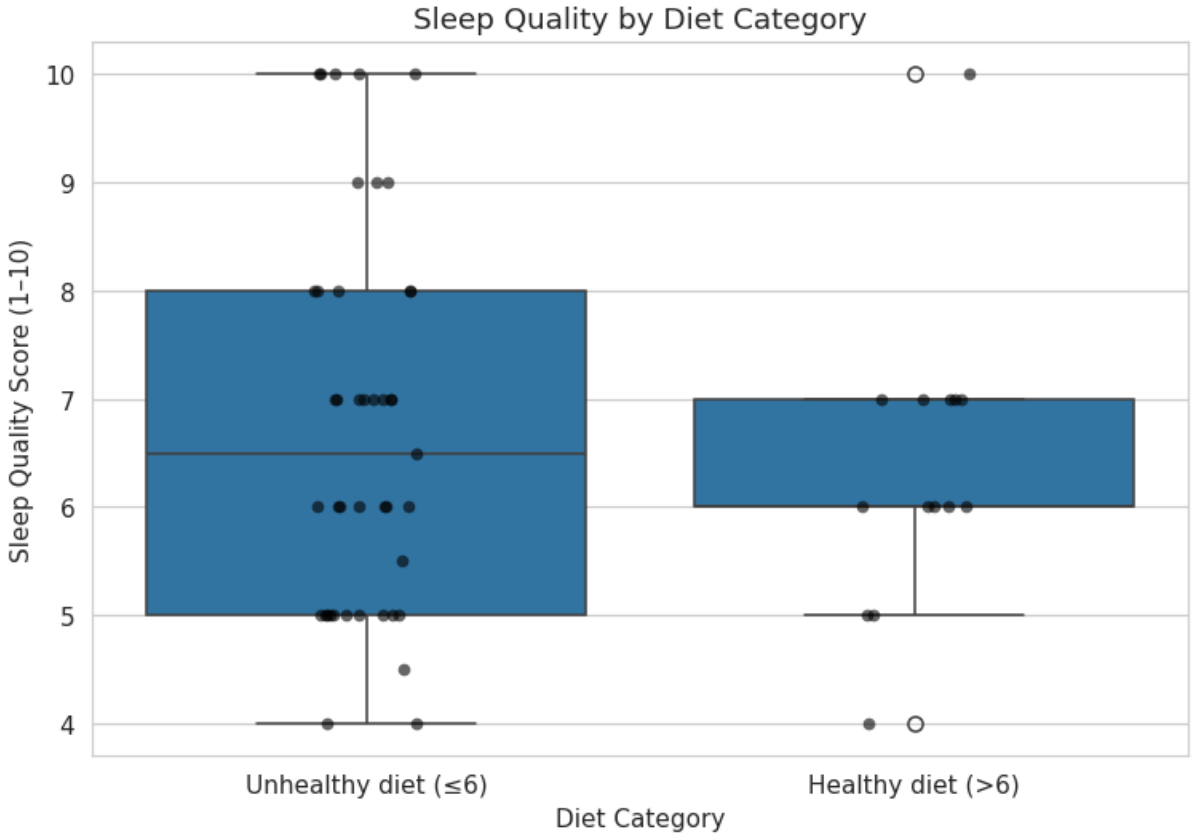
Distribution of Sleep Quality Scores

This histogram shows the distribution of self-reported sleep quality scores among participants. Most observations are concentrated in the mid-to-high range, indicating that participants generally perceive their sleep quality as moderate to good. The spread of the distribution suggests sufficient variability for comparative analysis.



Distribution of Diet Score

This histogram presents the distribution of healthy eating scores among participants. Most scores are clustered around the mid-range, indicating moderate dietary habits within the sample. The distribution shows sufficient variation to compare different diet quality groups in subsequent analyses.



Sleep Quality by Diet Category
This boxplot compares sleep quality scores between participants with healthy and unhealthy diet categories. The median sleep quality values and interquartile ranges overlap substantially between the two groups, indicating no clear difference in sleep quality. This visual observation is consistent with the two-sample t-test results, which show no statistically significant difference between the groups.

6. Limitations and Future Work

Limitations

Several limitations should be considered when interpreting the results of this project. First, the small sample size restricts statistical power and limits the generalizability of the findings. Second, the reliance on self-reported data introduces potential bias and measurement error. Third, important external factors such as stress levels, mental health, academic workload, and sleep environment were not directly measured.

Future Work

Future research could address these limitations by collecting larger and more diverse datasets, incorporating objective sleep measurements from wearable devices, and including psychological and environmental variables. Longitudinal data would allow for analyzing changes over time and causal relationships. Additionally, advanced feature engineering and alternative machine learning models may uncover more subtle patterns not captured in the current analysis.