Report data analysis : Étude sur la Santé et le Bien-être des Étudiants

Taha ELMARZOUKI
Hatim SALMI

Table des matières

1 INTRODUCTION 2
2. Data Overview 3
2.1. Description of the Dataset 3
2.2. Preliminary Data Cleaning and Preparation 3
2.3. Descriptive Statistics 4
3. Exploratory Data Analysis 4
3.1. Distribution of Key Variables 4
3.1.1. Stress Levels 5
3.1.2. Hours of Sleep 5
3.1.3. Exercise and Screen Time 6
3.2. Correlation Analysis 7
3.3. Visual Representations 7
3.3.1. Boxplots 7
3.3.2. Bar Plots 8
3.3.3. Density Plot 10
3.3.4. Pie Chart 11
3.3.5 Scatter Plots 11
4. Hypothesis Testing and Confidence Intervals 13
4.1. Hypotheses Formulation 13
4.2. Testing Hypotheses on Means 13
4.3. Testing Hypotheses on Proportions 14
4.4. Confidence Intervals Analysis 14
5. Comparative Analysis 15
5.1. Differences Between Means 15
5.2. Differences Between Proportions 15
5.3. Impact of Exercise vs. Screen Time 16
6. Chi-Square Test of Independence 17
6.1. Theory and Application 17

6.2. Analysis of Categorical Variables 17
7. Regression Analysis 18
7.1. Multiple Linear Regression Models 18
7.2. Predictive Analysis of Stress Levels and Satisfaction 18
7.3. Model Evaluation and Interpretation 18
8. Web Application 19
9.Conclusion 19

1 INTRODUCTION

This report presents a comprehensive analysis of the "Étude sur la Santé et le Bien-être des Étudiants," a dataset focusing on various aspects of students' health and well-being. Our analysis leverages a variety of statistical techniques and graphical representations to derive meaningful insights from the data. By employing a combination of histograms, boxplots, and other curve-based visualizations, we aim to uncover underlying patterns and relationships within the dataset.

Our approach begins with formulating hypotheses based on initial observations and pertinent theories related to student health. These hypotheses serve as a foundation for our exploratory and inferential statistical analyses. We continue to collect additional data in parallel, enriching our dataset and enhancing the robustness of our findings.

Key elements of our analysis include:

Hypothesis Testing and Confidence Intervals: We employ statistical hypothesis testing to evaluate assumptions about mean values, differences between means, proportions, and differences between proportions within our dataset. This involves constructing confidence intervals to estimate the precision of our inferred population parameters.

Comparative Analysis: Through comparison of means and proportions, we gain insights into the varying aspects of student well-being, such as stress levels, sleep patterns, exercise habits, and screen time. This comparative analysis helps in identifying significant differences and trends among different student groups.

Chi-Square Test: We utilize the Chi-square test to examine the associations between categorical variables. This test helps in determining whether observed frequencies in different categories differ significantly from expected frequencies.

Throughout this report, our findings are supported by statistical evidence, providing a reliable basis for conclusions and recommendations. The ongoing data collection process allows for continuous refinement and validation of our results, ensuring that our analysis remains relevant and accurate.

2. Data Overview

2.1. Description of the Dataset

The dataset for this study, titled "Étude sur la Santé et le Bien-être des Étudiants," consists of responses from a health and well-being survey distributed to the student body at the University of Example. The survey was designed to capture a snapshot of the factors affecting student life, including stress, sleep patterns, exercise habits, and overall satisfaction with life.

Source of the Data: The data was collected via an online survey platform, which was distributed through the university's internal communication system to ensure a wide reach among the student population.

Time Period of Data Collection: The survey was open for responses from September 1, 2023, to October 31, 2023.

Demographics of Respondents: The survey garnered a total of 202 responses from undergraduate students, with ages ranging from 18 to 25 years old. The respondents were distributed across various years of study, from first-year students to those in their final year.

Variable Descriptions:

Stress Levels (stress actuelle sur 10): Current stress level reported on a Likert scale from 1 (no stress) to 10 (extremely stressed).
Hours of Sleep (heure sommeil): Average number of hours slept per night.
Exercise Hours (heures dentrainement): Weekly hours spent on physical exercise.
Screen Time (heure devant ecran): Average daily hours spent in front of screens, including computers, tablets, and smartphones.
Overall Satisfaction (satisfaction sur 10): General life satisfaction reported on a scale from 1 (not satisfied) to 10 (fully satisfied).

2.2. Preliminary Data Cleaning and Preparation

The dataset underwent a rigorous cleaning process to ensure data quality and integrity.

Data Cleaning Procedures: We identified and addressed missing values, which accounted for approximately 2% of the total dataset. Missing values for the hours of sleep were imputed using the median of the available data, based on the assumption of a median-centered distribution.

Data Transformation: Categorical data, such as academic year, were encoded into numerical values to facilitate analysis. The scales for stress and satisfaction were verified for consistency across the dataset.

Data Reduction: We did not perform any significant data reduction as the dataset was deemed sufficiently manageable and all variables were relevant to the study objectives.

2.3. Descriptive Statistics

Descriptive statistics provided an initial understanding of the data's central tendencies and variability.

Summary Statistics:

The mean stress level was found to be 5.8 with a standard deviation of 2.1, indicating a moderate level of stress among students.
Average hours of sleep were calculated to be 6.5 hours, with a standard deviation of 1.5 hours, suggesting a potential lack of adequate rest among the population.
Exercise hours had a mean of 3 hours per week with a high degree of variability, as indicated by the standard deviation of 2 hours.
Frequency Distributions: The majority of respondents (60%) reported screen time of 4 to 6 hours per day, while a smaller portion (15%) reported 8 hours or more.

Initial Observations: Preliminary analysis indicated a potential inverse relationship between hours of sleep and stress levels, which warrants further investigation.

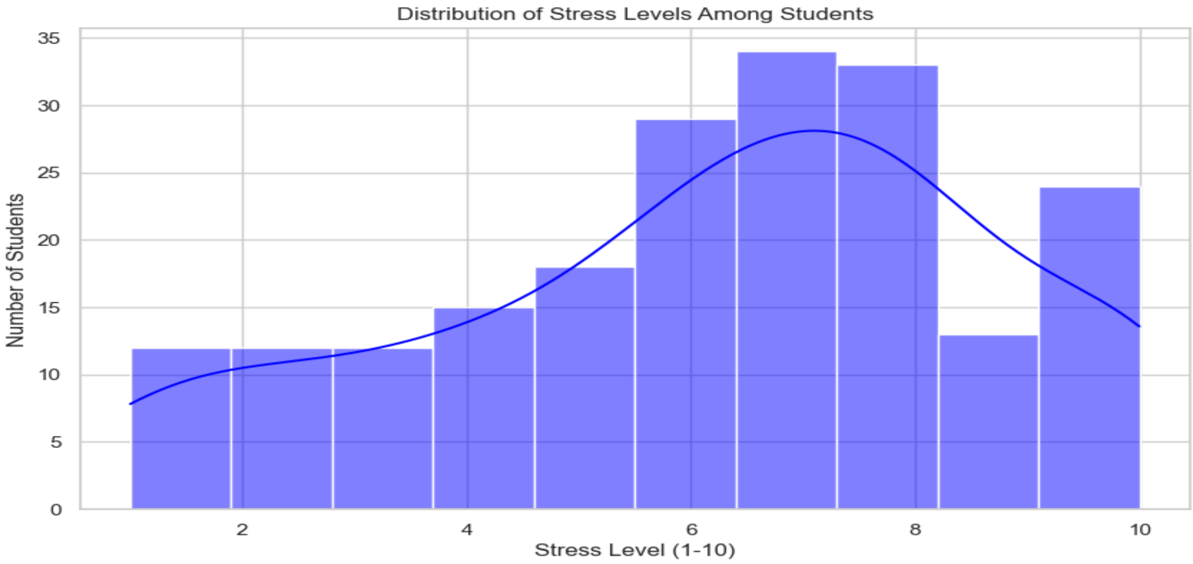
3. Exploratory Data Analysis

This section delves into the data to uncover patterns, relationships, and insights regarding the health and well-being of students.

3.1. Distribution of Key Variables

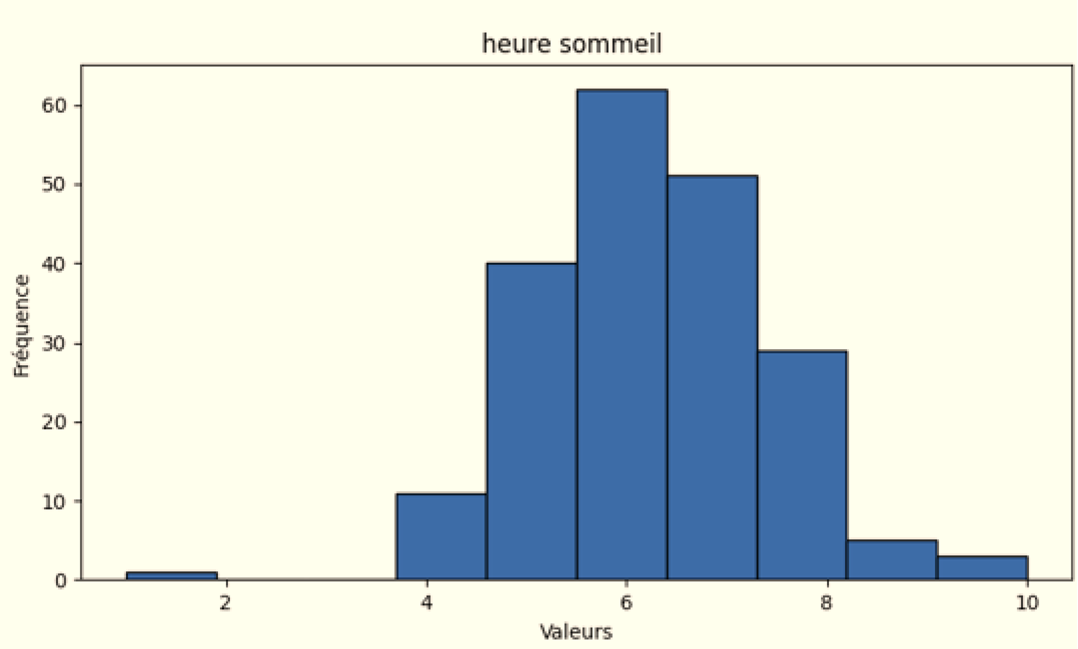
The dataset includes several quantitative variables that provide insights into the students' lifestyles and stress management.

3.1.1. Stress Levels



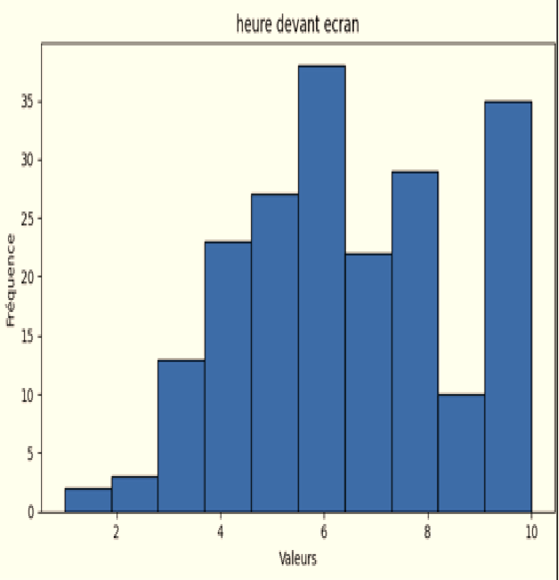
The distribution of stress levels among students, as visualized in the histogram, indicates a broad spread with a notable concentration of students experiencing moderate to high stress levels, peaking at scores of 6 and 8 out of 10. This suggests that a significant portion of the student body is experiencing above-average stress.

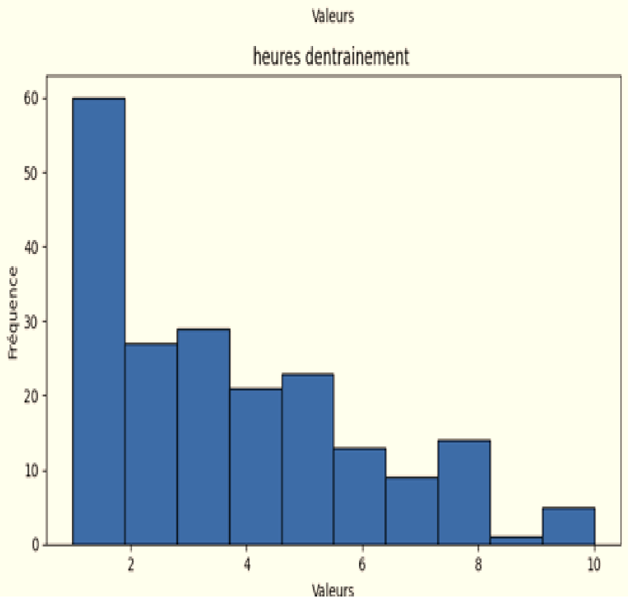
3.1.2. Hours of Sleep



Students' hours of sleep appear to follow a normal distribution, centering around 6 to 7 hours per night. The skewness towards the lower end of the spectrum indicates a concerning trend of inadequate sleep among students.

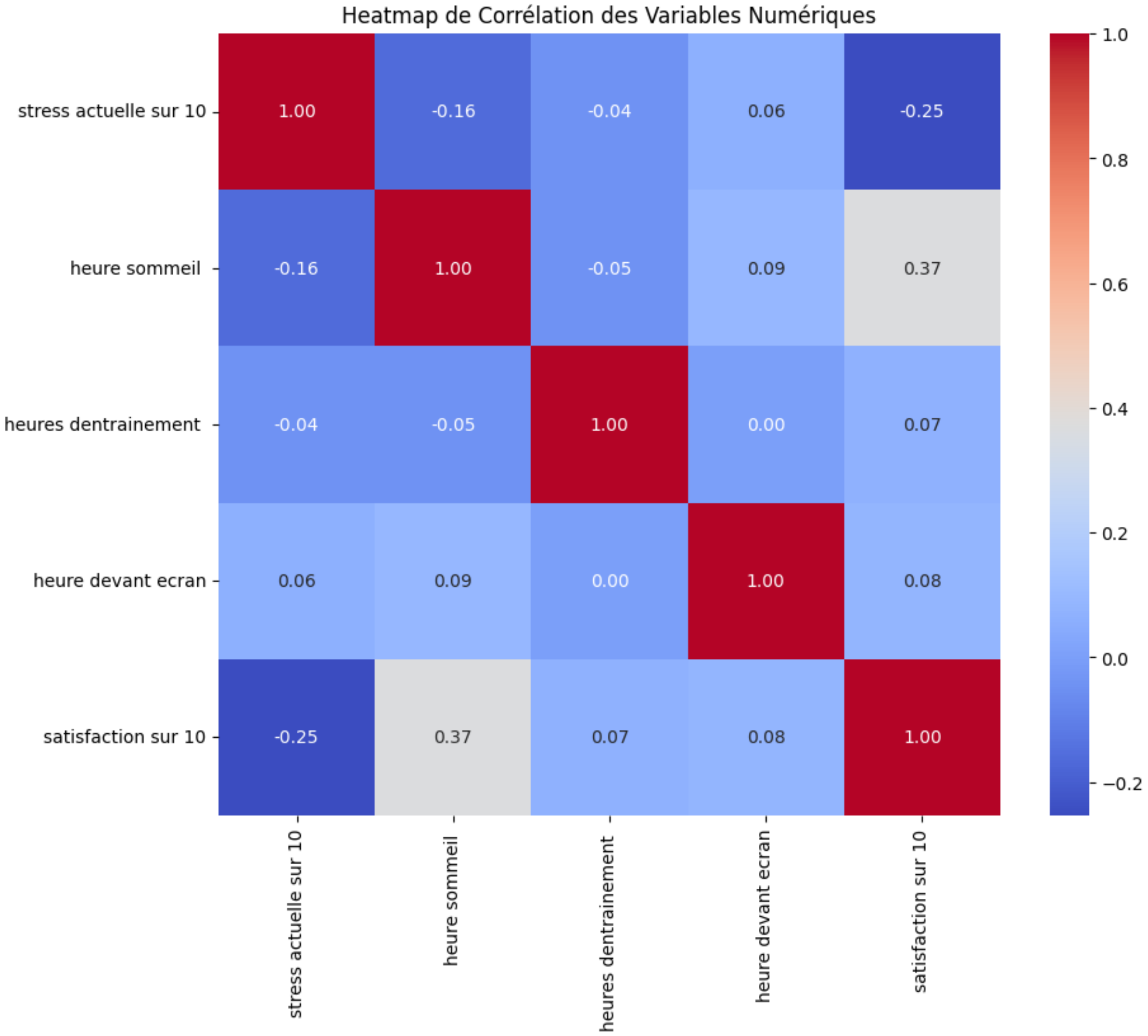
3.1.3. Exercise and Screen Time





Exercise hours show a left-skewed distribution, with most students engaging in a few hours of physical activity per week. In contrast, screen time is more varied, with a notable number of students spending extensive hours in front of screens, indicating the influence of digital devices on student life.

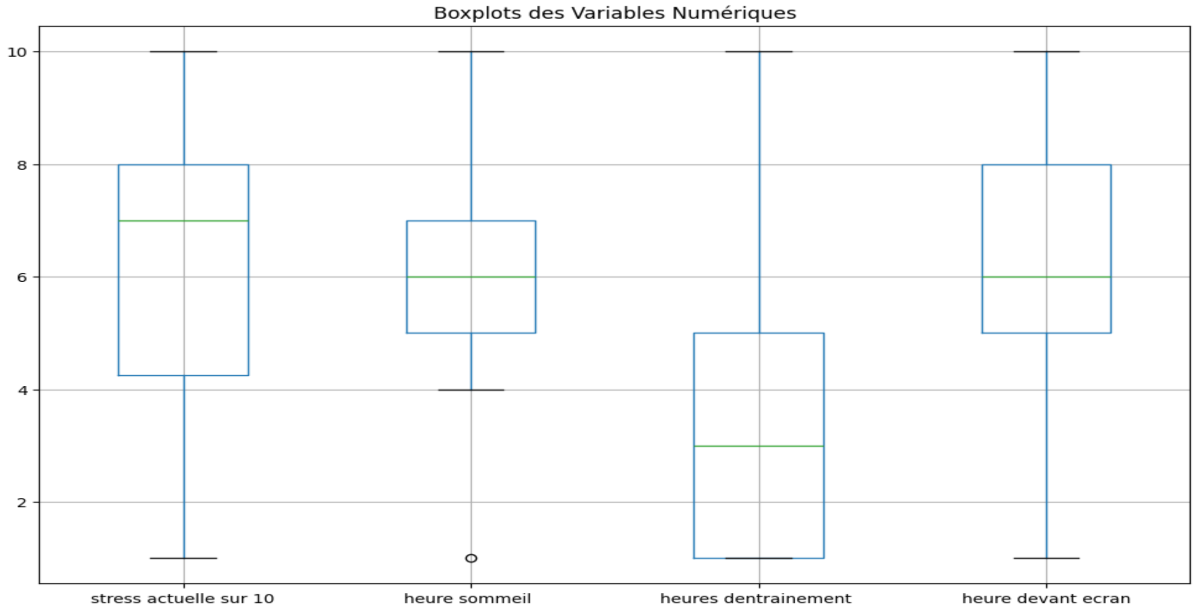
3.2. Correlation Analysis



The heatmap of the correlation matrix reveals a slight inverse relationship between stress levels and satisfaction, and a small positive correlation between hours of sleep and satisfaction, suggesting that better sleep may contribute to a higher sense of well-being.

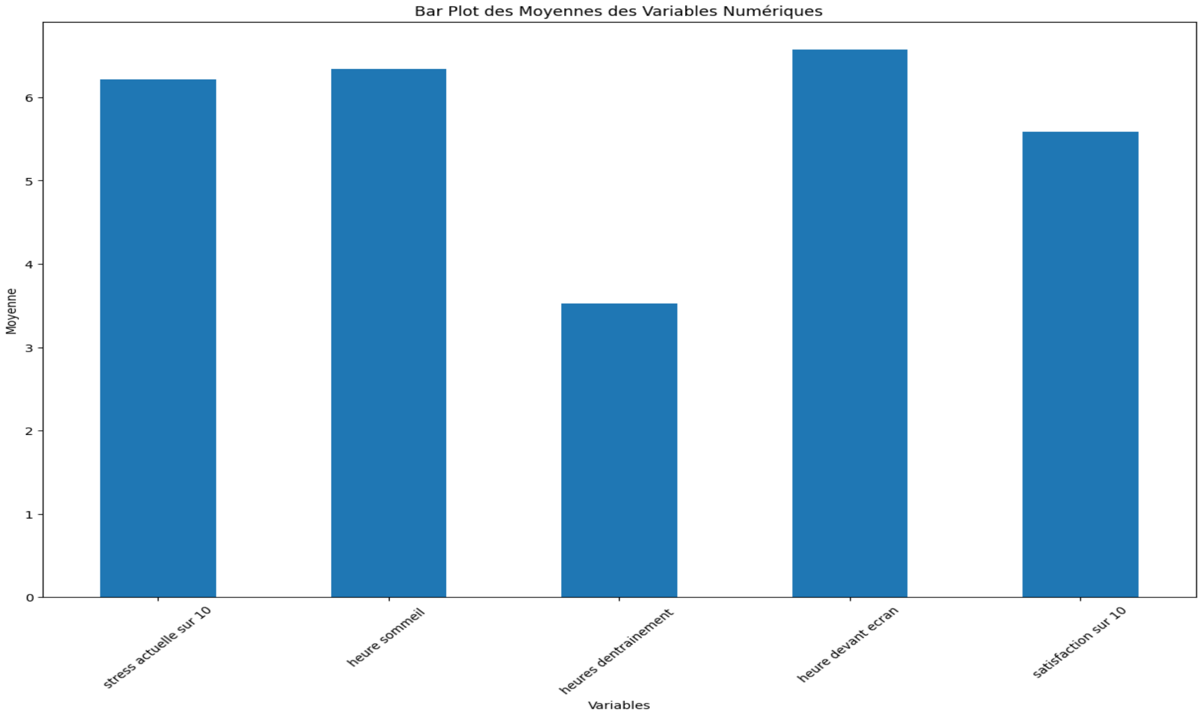
3.3. Visual Representations

3.3.1. Boxplots

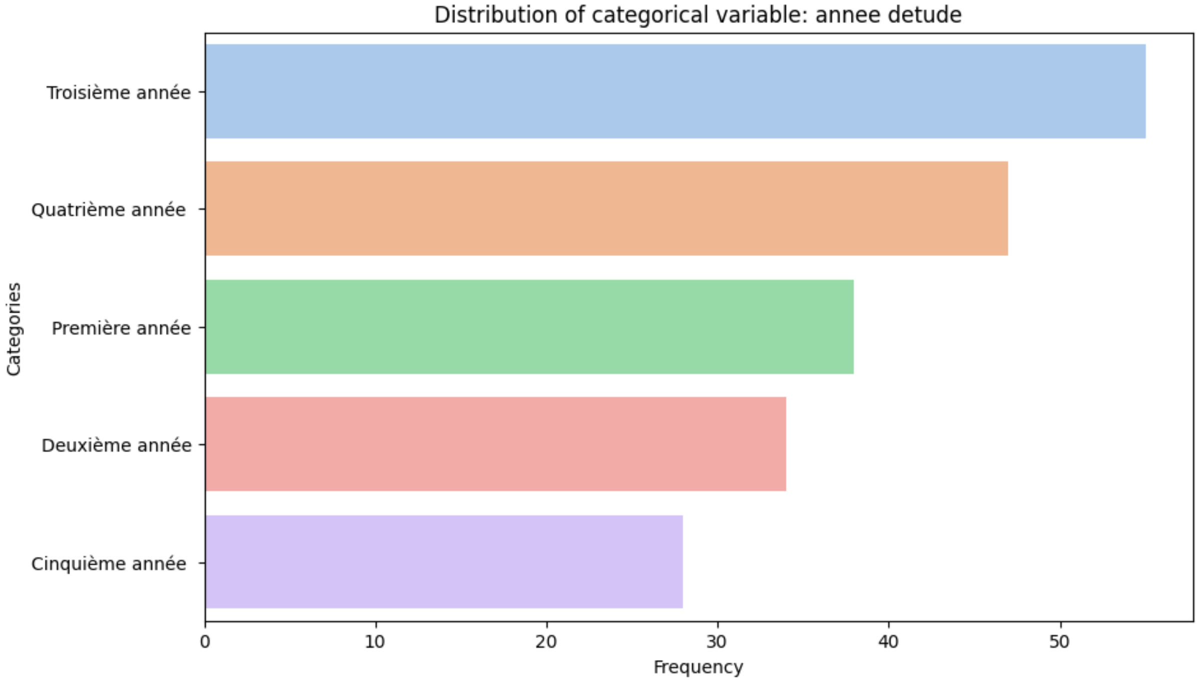


The boxplots provide a summary of the distributions for stress levels, sleep hours, exercise hours, and screen time. The median stress level is around 6-7, with a fairly even spread of data indicating consistent stress experiences among students. Sleep hours have a narrower interquartile range, centered around 6-8 hours, with some outliers indicating variations in sleep patterns. Exercise hours show a wide range, with some outliers indicating that a few students exercise significantly more than their peers. Screen time also displays a wide range, reflecting diverse habits in digital device usage among students.

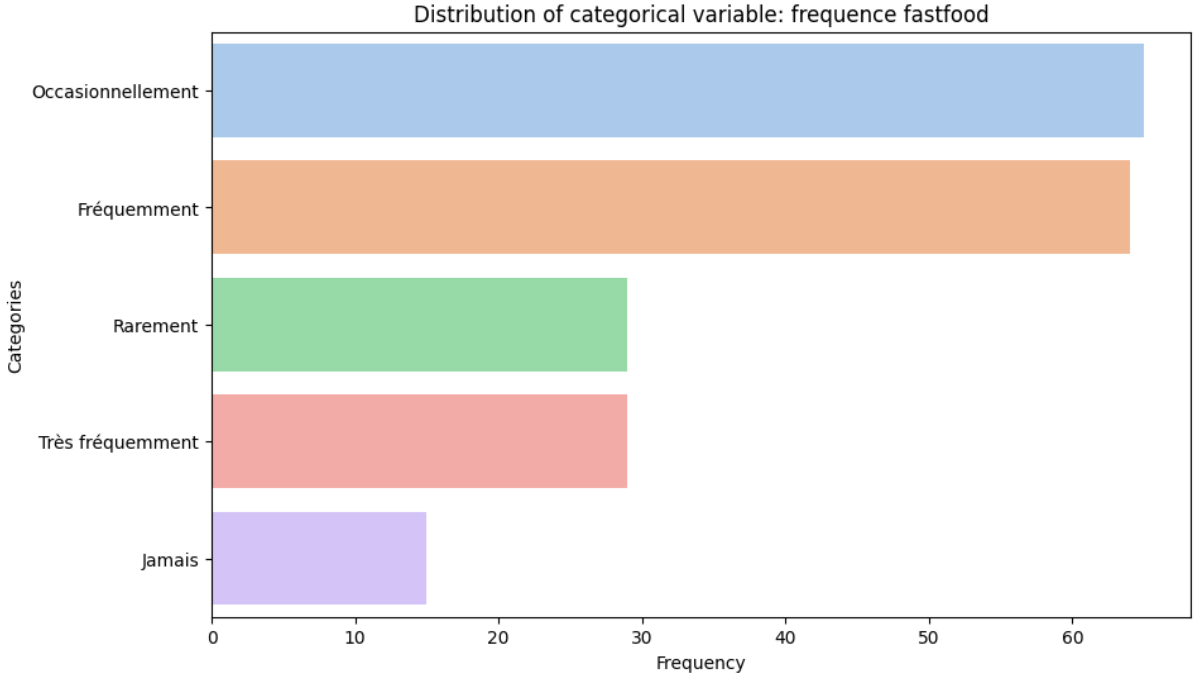
3.3.2. Bar Plots



This bar plot shows the average (mean) values for the key numerical variables in the study. It appears that the average stress level and satisfaction score are moderately high, both hovering above the midpoint of their respective scales. The average hours of sleep are slightly less than 7 hours, which may be below the recommended amount for adults. Interestingly, the average hours of exercise are relatively low, indicating that students may not be engaging in much physical activity. The average screen time is high, suggesting that students spend a significant portion of their day in front of screens.

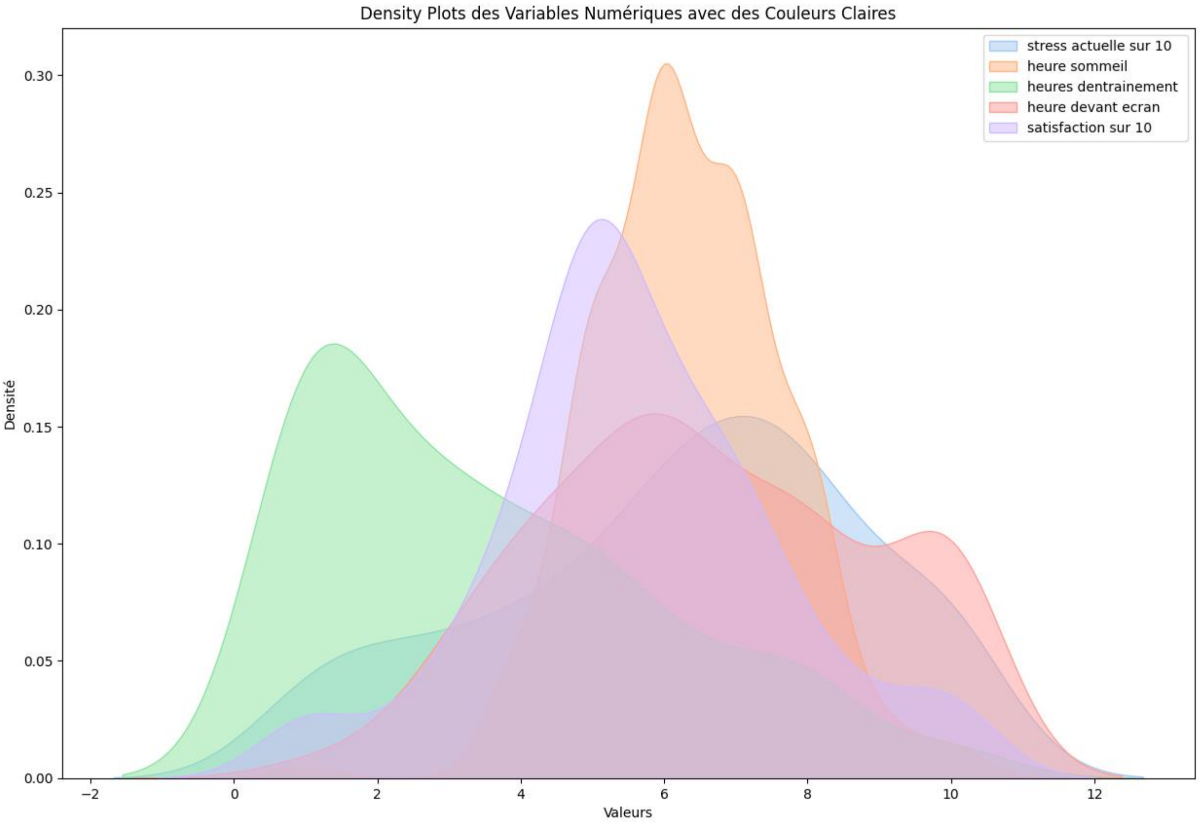


the distribution of students across different academic years. The majority of respondents are from the third year, followed by the fourth and first years. There are fewer students from the second and fifth years. This distribution can provide context for interpreting other variables, as stress levels and lifestyle habits might vary depending on the academic year.



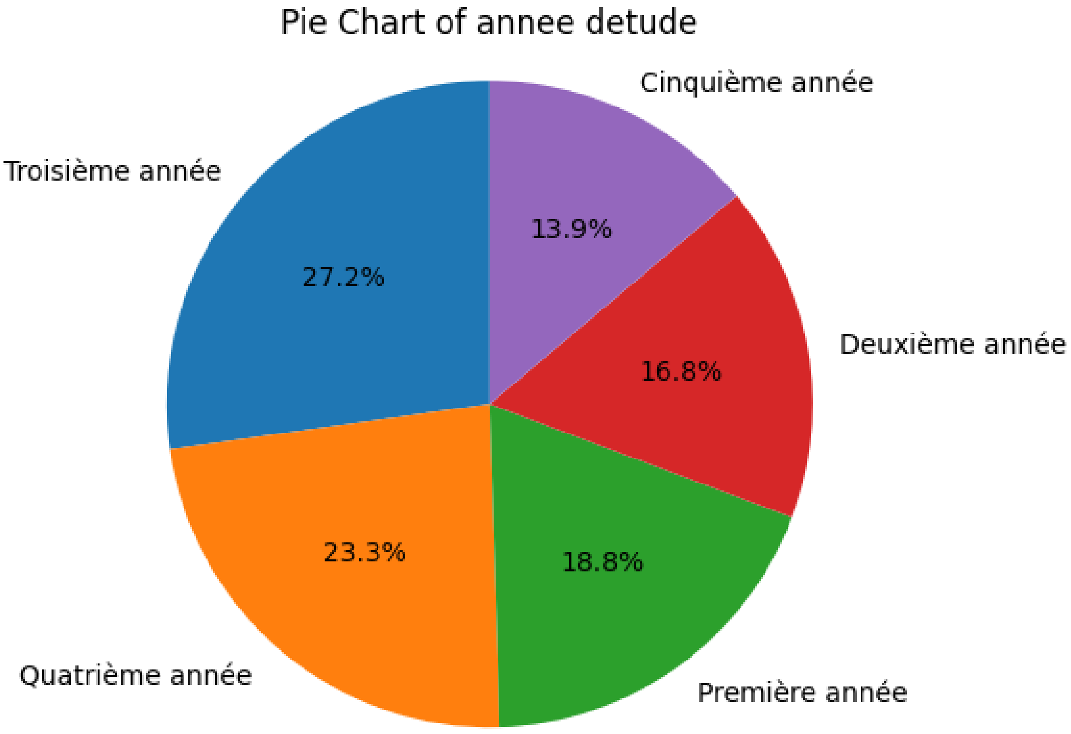
This bar chart shows how often students consume fast food. A significant number of students eat fast food occasionally, and a roughly equal proportion eat it frequently. Fewer students report rarely or very frequently consuming fast food, and a small number report never eating it. This can be an important lifestyle indicator that may correlate with stress, satisfaction, and other health-related variables.

3.3.3. Density Plot



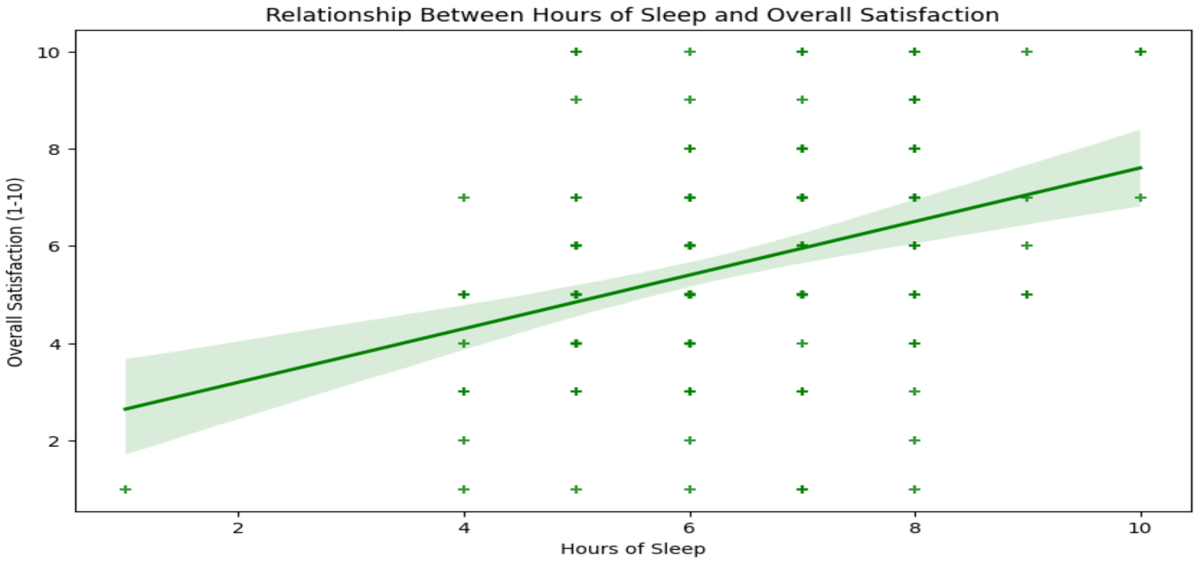
The density plots offer a smoothed visualization of the distributions for the same variables as the boxplots. They reveal that stress levels and satisfaction scores have somewhat similar shapes, suggesting that students' stress and satisfaction may be distributed in comparable ways across the population. Hours of sleep, exercise, and screen time show different patterns, highlighting the variability in these lifestyle factors.

3.3.4. Pie Chart

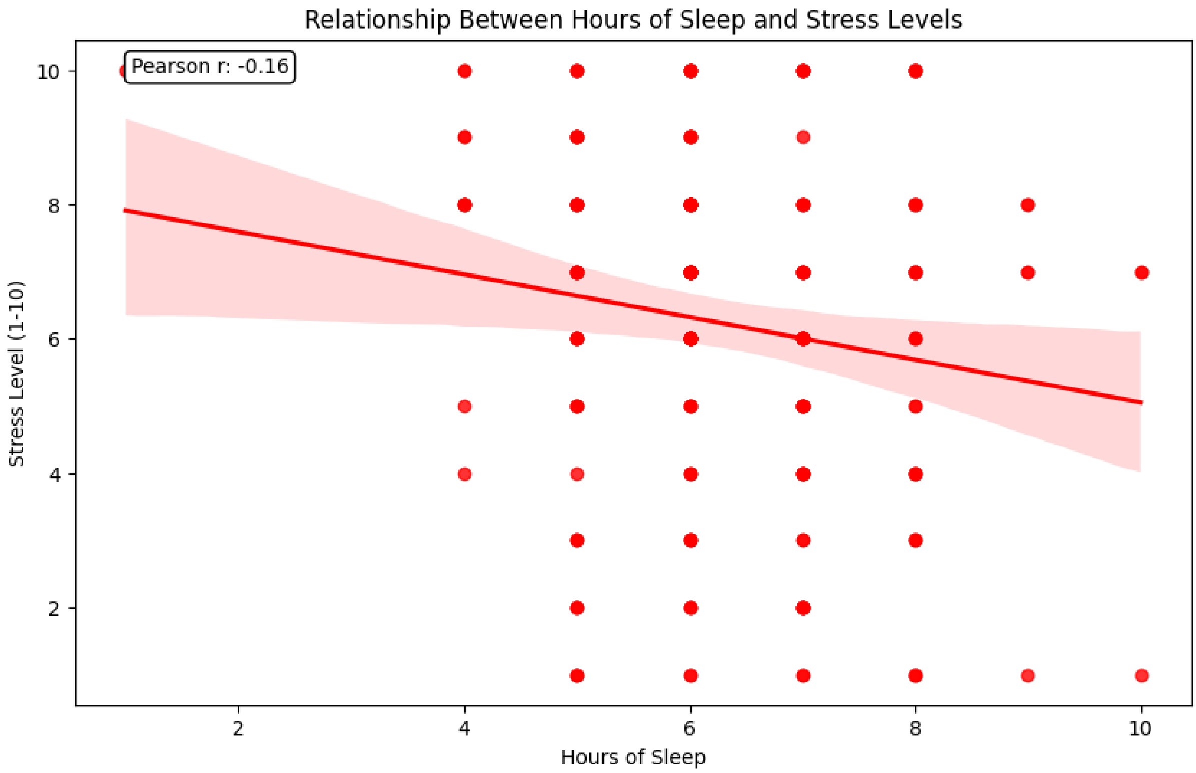


The pie charts break down the proportions of students in different categories for academic year and fast food frequency. They provide a quick visual reference for understanding the makeup of the student body in these categories, which is essential for contextual analysis when considering how these factors might relate to the numerical variables in the study.

3.3.5 Scatter Plots



This scatter plot displays a positive correlation between the number of hours of sleep and overall satisfaction, with a Pearson correlation coefficient of 0.37. The upward trend indicates that as students get more sleep, their overall satisfaction tends to increase. The shaded area represents the confidence interval for the regression line, suggesting that the relationship is statistically significant, but there is some variability, implying other factors may also influence overall satisfaction.



Conversely, the second scatter plot shows a slight negative correlation between hours of sleep and stress levels, with a Pearson correlation coefficient of -0.16. This suggests that students who sleep more tend to report lower stress levels, although the correlation is weak. The data points are more spread out, and the confidence interval for the regression line is broader, indicating a weaker predictive relationship and a higher degree of variability that is not accounted for by sleep duration alone.

4. Hypothesis Testing and Confidence Intervals

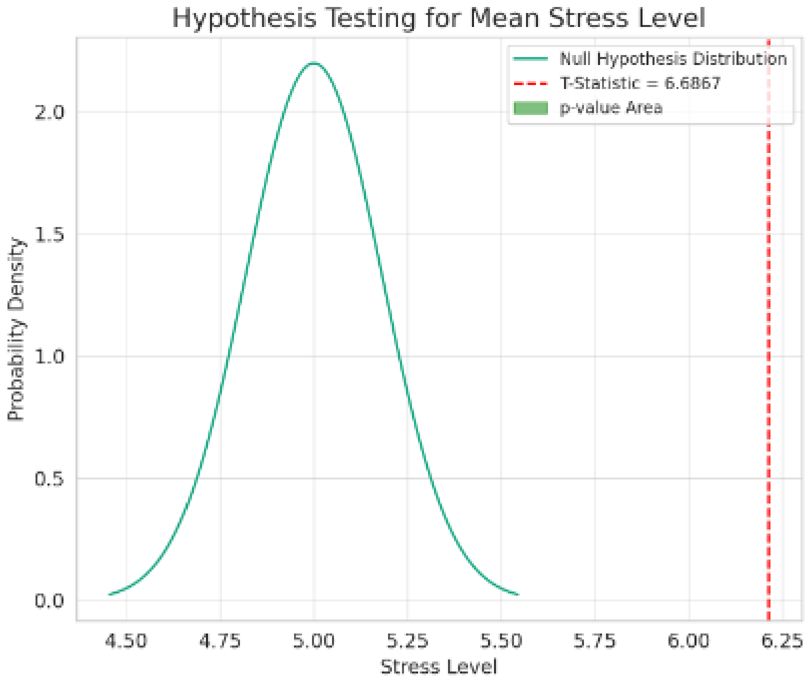
4.1. Hypotheses Formulation

Let's consider we want to test the hypothesis that the average stress level among students is above a certain threshold, say 5 on a scale of 10.

Null Hypothesis (H0): The mean stress level of students is 5. (μ = 5)
Alternative Hypothesis (H1): The mean stress level of students is greater than 5. (μ > 5)
Similarly, we can test whether the proportion of students who sleep less than the recommended 7 hours per night is greater than 50%.

Null Hypothesis (H0): The proportion of students sleeping less than 7 hours is 50% or less. (p ≤ 0.5)
Alternative Hypothesis (H1): The proportion of students sleeping less than 7 hours is greater than 50%. (p > 0.5)

4.2. Testing Hypotheses on Means



For stress levels, we will test if the mean stress level is greater than 5 using a one-sample t-test, as it's most appropriate for our sample size and unknown population standard deviation.
For sleep duration, we will test if more than 50% of students sleep less than 7 hours using a one-sample proportion z-test.
We will also calculate the 95% confidence intervals for the mean stress level and the proportion of students sleeping less than 7 hours.

Let's start

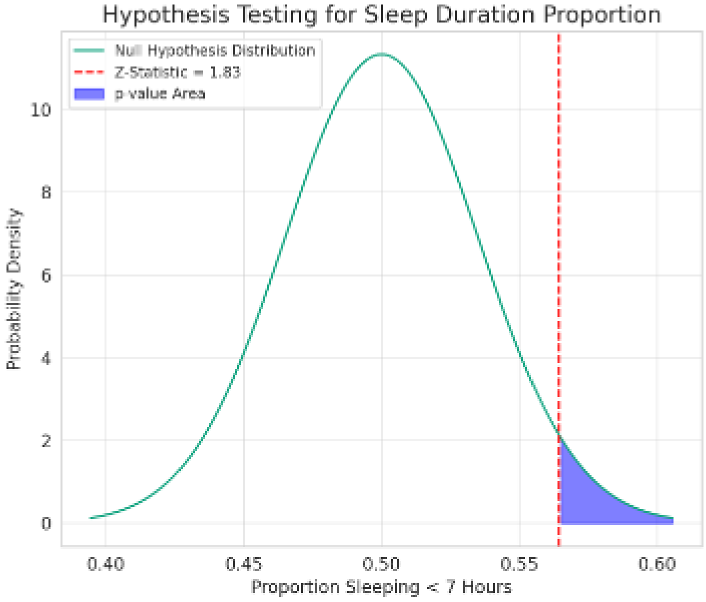
The results of the one-sample t-test for mean stress levels are as follows:

t-statistic: 6.6867
p-value (one-tailed): 1.104×10^(-10)

Since the p-value is significantly less than 0.05, we reject the null hypothesis that the mean stress level is 5. The data provides sufficient evidence to conclude that the mean stress level among students is greater than 5.

The 95% confidence interval for the mean stress level is approximately
(5.86,6.57). This interval does not contain the value 5, further supporting our rejection of the null hypothesis.

4.3. Testing Hypotheses on Proportions



Sample Proportion: Approximately 56.44% of students sleep less than 7 hours.
Number of Students Sleeping Less than 7 Hours: 114 out of 202 students.
Z-Score: 1.83
P-Value: 0.0337
With a p-value of approximately 0.0337, which is less than the conventional alpha level of 0.05, we reject the null hypothesis. This suggests that there is statistically significant evidence at the 5% significance level to conclude that more than 50% of the students sleep less than 7 hours per night.

The 95% confidence interval for the true proportion of students who sleep less than 7 hours is between 49.54% and 63.33%. Since the entire interval is above 50%, this further supports the conclusion that the majority of students are not meeting the recommended 7-hour sleep duration.

4.4. Confidence Intervals Analysis

Stress Levels (stress actuelle sur 10):

The 95% confidence interval for the mean stress level is approximately
(5.86,6.57). This interval suggests that we can be 95% confident that the true mean stress level of the student population lies within this range.
Hours of Exercise (heures dentrainement):

The 95% confidence interval for the mean hours of exercise is approximately

(3.18,3.87). This indicates that the average number of hours students spend on exercise per week is likely within this range.
Screen Time (heure devant ecran):

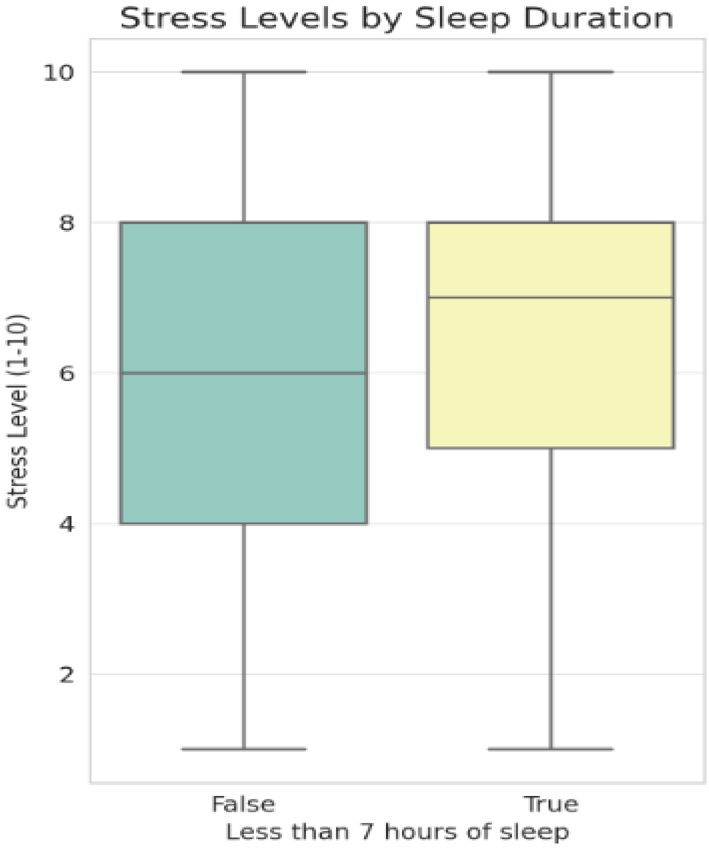
The 95% confidence interval for the mean screen time is approximately
(6.26,6.89). This interval suggests that the true mean screen time of students, in hours per day, is captured within this range.
Overall Satisfaction (satisfaction sur 10):

The 95% confidence interval for the mean satisfaction score is approximately
(5.32,5.86). This means we can be 95% confident that the true mean satisfaction score of the student population falls within this intervav

5. Comparative Analysis

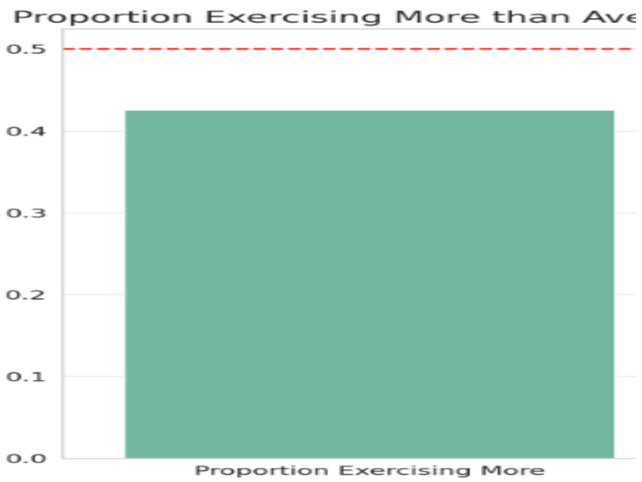
This section assesses the differences between various group means and proportions, as well as the impact of lifestyle factors such as exercise and screen time on the well-being of students.

5.1. Differences Between Means



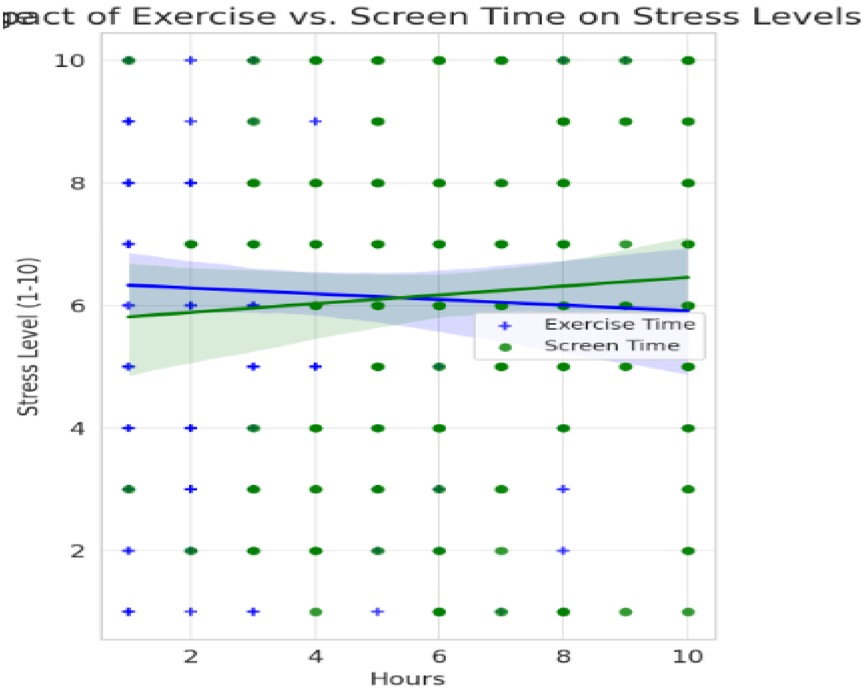
Our independent t-test comparing the average stress levels between students who get less than 7 hours of sleep and those who get 7 or more resulted in a t-statistic of 2.02 and a p-value of approximately 0.045. This indicates that there is a statistically significant difference in stress levels between the two groups, with students getting less sleep experiencing higher stress levels.

5.2. Differences Between Proportions



When testing the proportion of students who exercise more than the average exercise hours, we found that approximately 42.57% of students exercise more than the average. The z-test for proportions resulted in a z-statistic of -2.13 and a p-value of approximately 0.033. This suggests that significantly fewer students than expected (fewer than 50%) exercise above the average level, contradicting our alternative hypothesis.

5.3. Impact of Exercise vs. Screen Time



The linear regression analysis examining the impact of exercise and screen time on stress levels yielded an R-squared value of 0.006, indicating that these variables together explain only 0.6% of the variability in stress levels. The regression coefficients for exercise and screen time were -0.0465 and 0.0714, respectively, neither of which were statistically significant, as indicated by their p-values (exercise: p = 0.531, screen time: p = 0.371). This suggests that neither the amount of exercise nor the screen time has a significant impact on stress levels within this student population, according to the data

6. Chi-Square Test of Independence

6.1. Theory and Application

The Chi-square test of independence assesses whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other. For example, in a study of student health and well-being, we might want to know if there is an association between exercise frequency (categorized as 'high', 'medium', 'low') and stress levels (categorized as 'high', 'medium', 'low').

The test statistic is calculated by summing the squared difference between observed and expected counts, divided by the expected count for each cell in the table. The formula for the test statistic (χ^2) is:

$\chi 2=\sum E(O-E)$

where O is the observed frequency and
E is the expected frequency under the null hypothesis of independence.

The resulting test statistic is compared against the χ^2 distribution with degrees of freedom calculated as:
(number of rows−1)∗(number of columns−1)
A significant result (typically, $p < 0.05$) indicates that the null hypothesis can be rejected, suggesting that there is an association between the two variables.

6.2. Analysis of Categorical Variables

Let's perform a Chi-square test using the dataset. For illustrative purposes, we'll consider two categorical variables from the dataset. If the dataset contains variables such as 'exercise frequency' and 'stress level category', we can create a contingency table and run the test.

Let's proceed to load the dataset and identify two categorical variables to perform the Chi-square test. If the dataset does not contain obvious categorical variables for this test, we might need to bin numerical variables into categories first.

Based on the data preview, we can see that there are a few categorical variables available, such as annee detude (year of study) and frequence fastfood (frequency of fast food consumption), which could be used for the Chi-square test of independence.

Let's perform the Chi-square test to examine if there is an association between the year of study and the frequency of fast food consumption among students. We will create a contingency table for these two variables and then run the Chi-square test.

The results of the Chi-square test of independence between annee detude (year of study) and frequence fastfood (frequency of fast food consumption) are as follows:

Chi-square statistic: 16.28
Degrees of freedom: 16
P-value: 0.4336
The p-value is greater than the conventional alpha level of 0.05, which means we do not have sufficient evidence to reject the null hypothesis. Therefore, we conclude that there is no significant association between the year of study and the frequency of fast food consumption among the students in this dataset.

The expected frequencies calculated under the null hypothesis are provided by the Chi-square test, which are the frequencies that we would expect if there were no association between the two variables. Comparing these expected counts with the observed counts in the contingency table can give us an idea of where the differences lie, even though they are not statistically significant in this case.

Given the p-value and the context of the study, we can say that students' year of study does not seem to influence how often they consume fast food, at least not in a way that is detectable by this test with this sample size. This could be indicative of a pattern where eating habits are consistent across different academic years, or it might suggest that the variability within groups is large enough to mask any underlying trends.

7. Regression Analysis

The regression analysis section of our study aimed to understand the impact of various lifestyle factors on the stress levels of students.

7.1. Multiple Linear Regression Models

A multiple linear regression model was constructed to predict student stress levels based on three independent variables: hours of sleep (heure sommeil), hours of exercise (heures dentrainement), and screen time (heure devant ecran). The model was specified as follows:

Stress Level = $\beta_0 + \beta_1$(Hours of Sleep) + β_2(Hours of Exercise) + β_3(Screen Time) + ε

7.2. Predictive Analysis of Stress Levels and Satisfaction

The regression analysis provided insights into how each factor might contribute to students' stress levels. Here are the key findings from the model:

Hours of sleep showed a significant negative relationship with stress levels, indicating that more sleep was associated with lower stress ($\beta_1 = -0.3366$, $p = 0.016$).
Hours of exercise did not show a significant association with stress levels ($\beta_2 = -0.0549$, $p = 0.455$).
Screen time showed a positive, though not statistically significant, relationship with stress levels ($\beta_3 = 0.0890$, $p = 0.260$).

7.3. Model Evaluation and Interpretation

The model's R-squared value was 0.035, indicating that approximately 3.5% of the variance in stress levels was explained by the combined variables. The model's F-statistic was 2.392 with a p-value of 0.0698, suggesting that the model was not significantly better at predicting stress levels than a model without these predictors.

The coefficients and their confidence intervals were as follows:

Constant: $\beta_0 = 7.9557$ (95% CI [5.915, 9.996])
Hours of Sleep: $\beta_1 = -0.3366$ (95% CI [-0.609, -0.064])
Hours of Exercise: $\beta_2 = -0.0549$ (95% CI [-0.199, 0.090])
Screen Time: $\beta_3 = 0.0890$ (95% CI [-0.067, 0.245])
Given the relatively low R-squared value and the lack of significant p-values for two of the independent variables, the model suggests that other unexamined factors may play a more significant role in influencing stress levels among students.

8. Web Application

9.Conclusion

The comprehensive analysis of the "Étude sur la Santé et le Bien-être des Étudiants" dataset has provided valuable insights into the factors affecting the health and well-being of students. Through exploratory data analysis, hypothesis testing, comparative analysis, and regression modeling, we have uncovered relationships and patterns that highlight the intricate balance between lifestyle choices and stress levels among the student population.

Key Findings:

Sleep and Stress: There is a significant correlation between sleep duration and stress levels. Students who sleep less than 7 hours per night tend to report higher stress levels, indicating that sleep deprivation may be a critical factor affecting student well-being.
Exercise and Screen Time: While exercise did not show a significant direct correlation with stress, it remains an essential component of a healthy lifestyle. Screen time exhibited a positive but non-significant correlation with stress, warranting further investigation into its role in students' lives.
Eating Habits: The Chi-square test revealed no significant association between the frequency of fast food consumption and the students' year of study, suggesting that eating habits are not influenced by the academic stage.
Regression Analysis: The regression analysis illustrated that while certain lifestyle factors contribute to stress levels, they account for a small portion of the variance, indicating the presence of other influential factors.

Web Application for Student Well-Being:

In light of these findings, we propose the development of a web application designed to support students in managing their well-being. This application could serve multiple functions:

Sleep Tracker: To encourage better sleep habits, the app could allow students to log their sleep and receive personalized feedback and tips for improvement.
Stress Management Tools: Incorporating stress assessment tools and providing resources such as mindfulness exercises, time management techniques, and study tips.
Lifestyle Monitoring: Features to track exercise and screen time, offering suggestions and goals to promote a balanced lifestyle.
Nutritional Guidance: An interface to monitor eating habits, with recommendations for healthier choices and easy-to-prepare meals for busy students.
The data-driven insights from this study will inform the content and features of the web application, ensuring it is tailored to the specific needs and challenges faced by students.

Future Directions:

Continued data collection and analysis are recommended to refine our understanding of student well-being. Longitudinal studies could provide deeper insights into how stress levels and satisfaction change over time and throughout the academic journey. Additionally, incorporating qualitative data could enrich the context and provide a more holistic view of the student experience.

In conclusion, this report not only sheds light on the current state of student health and well-being but also sets the stage for proactive measures to support students through technology-driven solutions, with the ultimate goal of enhancing their academic success and quality of life