Table des matières

1 INTRODUCTION 3
2. Data Overview 4
2.1. Description of the Dataset 4
2.2 Descriptive Statistics 5
3. Exploratory Data Analysis 5
3.1. Distribution of Key Variables 5
3.1.1. Stress Levels 6
3.1.2. Hours of Sleep 6
3.1.3. Exercise and Screen Time 7
3.2. Correlation Analysis 8
3.3. Visual Representations 9
3.3.1. Boxplots 9
3.3.2. Bar Plots 9
3.3.3. Density Plot 12
3.3.4. Pie Chart 13
3.3.5 Scatter Plots 14
4. Hypothesis Testing and Confidence Intervals 15
4.1. Hypotheses Formulation 15
4.2. Testing Hypotheses on Means 16
4.3. Testing Hypotheses on Proportions 17
4.4. Confidence Intervals Analysis 17
5. Comparative Analysis 18
5.1. Differences Between Means 18
5.2. Differences Between Proportions 19
5.3. Impact of Exercise vs. Screen Time 20
6. Chi-Square Test of Independence 20
6.1. Theory and Application 20
6.2. Analysis of Categorical Variables 21
7. Regression Analysis 21
7.1. Multiple Linear Regression Models 22
7.2. Predictive Analysis of Stress Levels and Satisfaction 22
7.3. Model Evaluation and Interpretation 22
8. Web Application 22
9.Conclusion 34

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).

University Year: Categorical variable representing the academic year of the student.

Stress Source: Categorical variable indicating the primary source of stress for the student.

Fast-Food Frequency: Categorical variable representing how often a student consumes fast food (e.g., 'Rarely', 'Occasionally', 'Frequently').

Stress Management: Categorical variable describing the methods used by students to manage stress (e.g., 'Exercise', 'Meditation', 'Socializing').

2.2. Preliminary Data Cleaning and Preparation

In the preliminary phase of data cleaning and preparation, the dataset, derived from a comprehensive survey of 202 respondents, underwent several crucial transformations. To enhance clarity and conciseness, column names were revised, and extraneous variables

such as "Start time," "End time," "Email," "Address," "Name," and "Last modified time" were omitted, as they were deemed irrelevant to the current analysis. Additionally, object columns containing string values, notably "Stress Source," "Fast-Food Frequency," and "Stress Management," were refined to ensure brevity and improved descriptive accuracy.

Moreover, the arrangement of the ID column was modified to start from 1 instead of 2, aligning the IDs with a more conventional indexing structure where the first line corresponds to the headers. This adjustment facilitates a seamless understanding of the dataset and ensures consistency in subsequent analyses. This preliminary cleaning step lays the foundation for a more structured and comprehensible dataset, setting the stage for subsequent exploratory data analysis and modeling.

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.

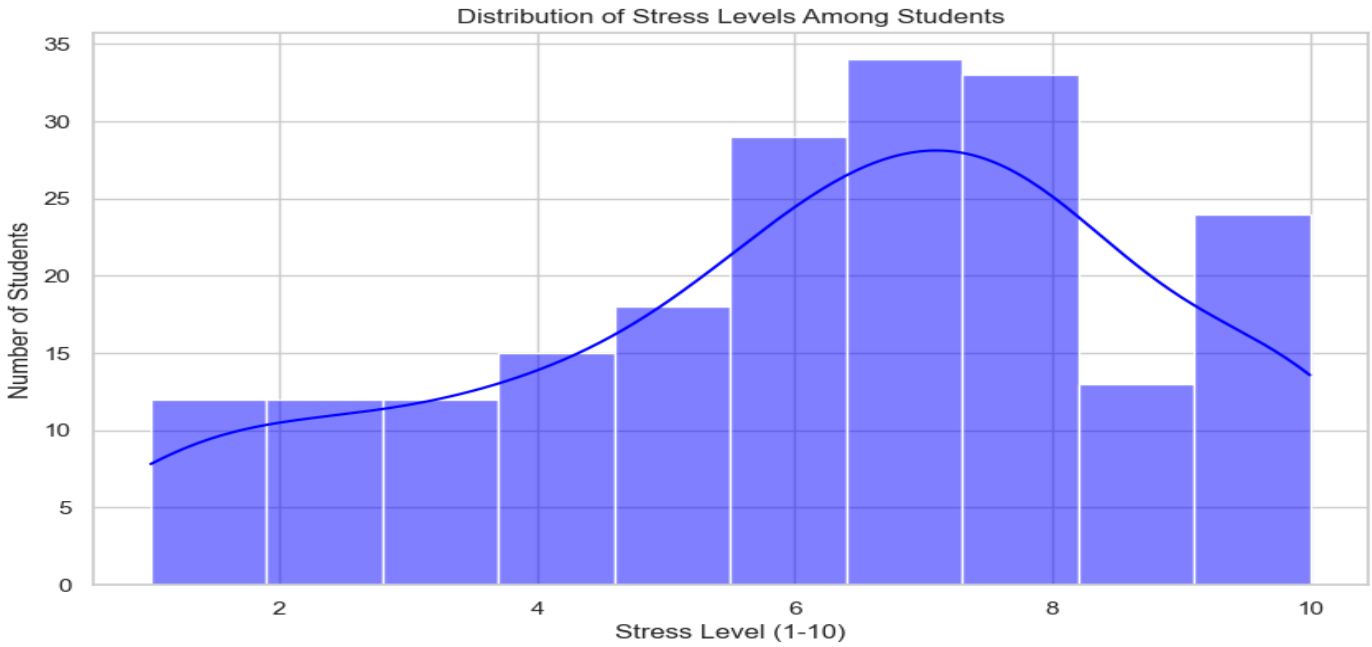
3. Exploratory Data Analysis

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

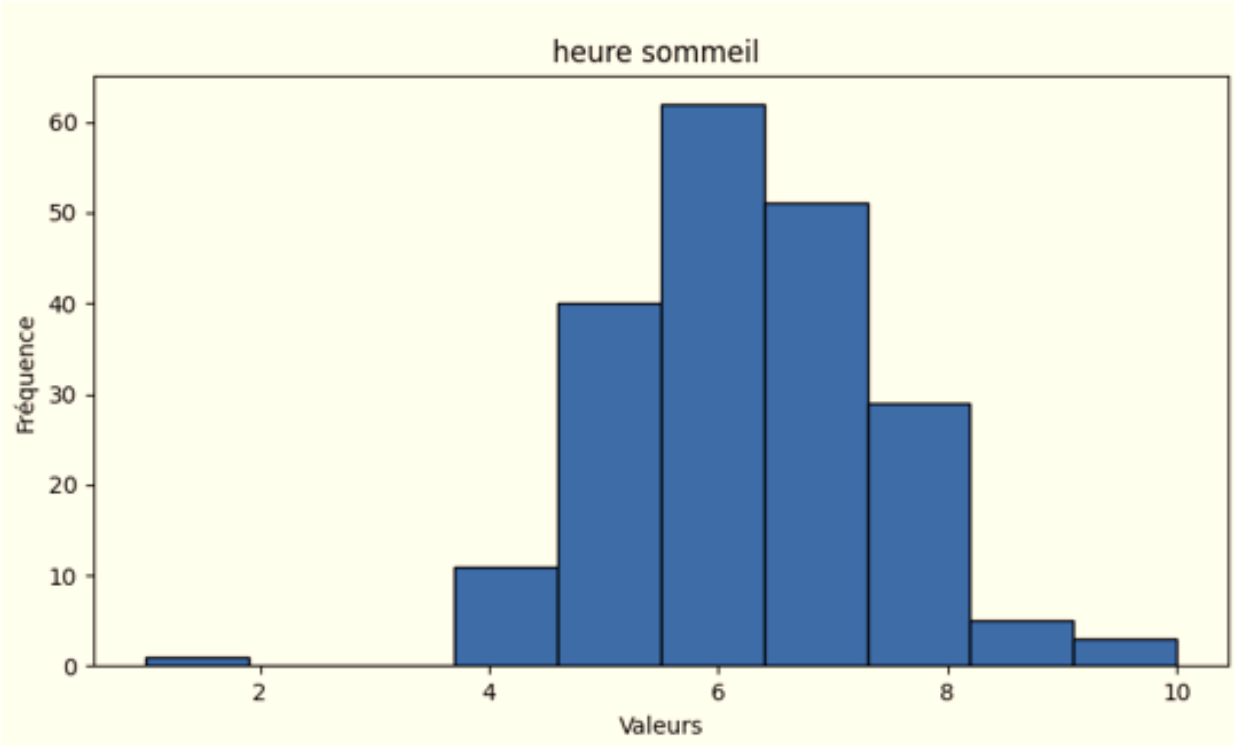
The dataset includes several quantitative variables that provide insights into the students' lifestyles and stress management.

3.1.1. Stress Levels



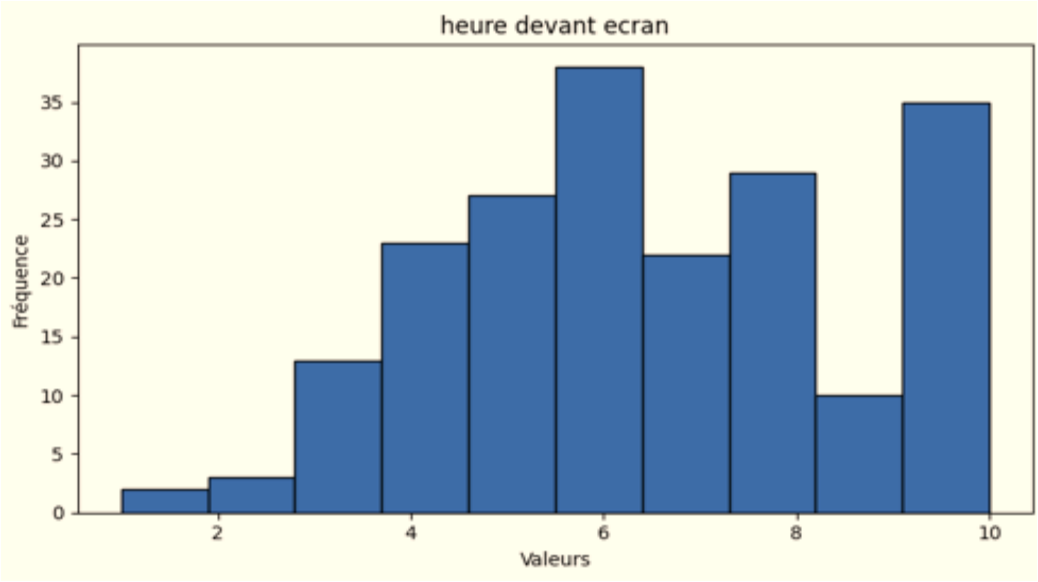
The distribution of stress levels among students, as visualized in the histogram, indicates a broad spread with a notable concentration of students experiencing moderate to high stress levels, peaking at scores of 6 and 8 out of 10. This suggests that a significant portion of the student body is experiencing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

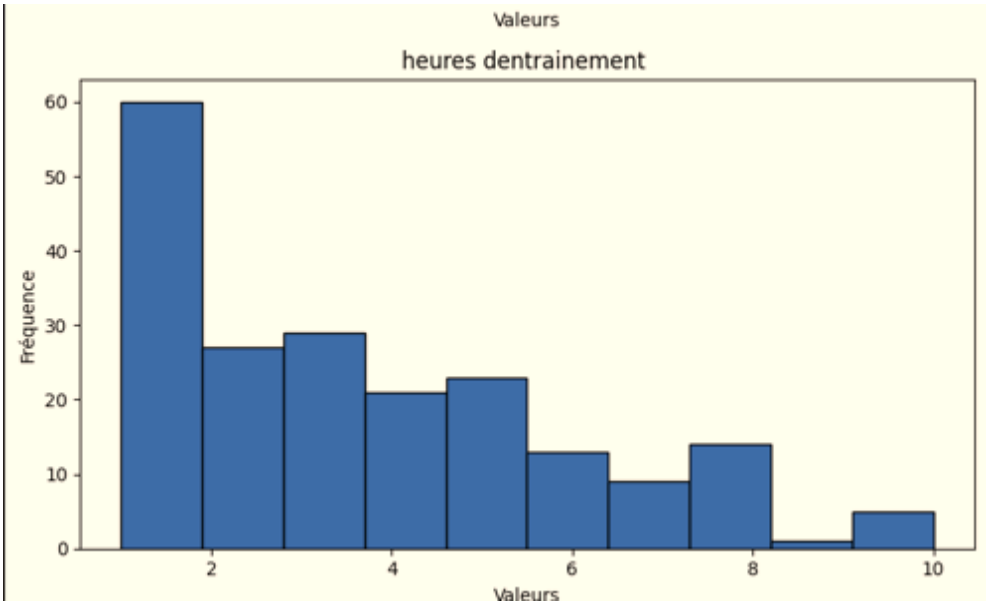
3.1.2. Hours of Sleep



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.

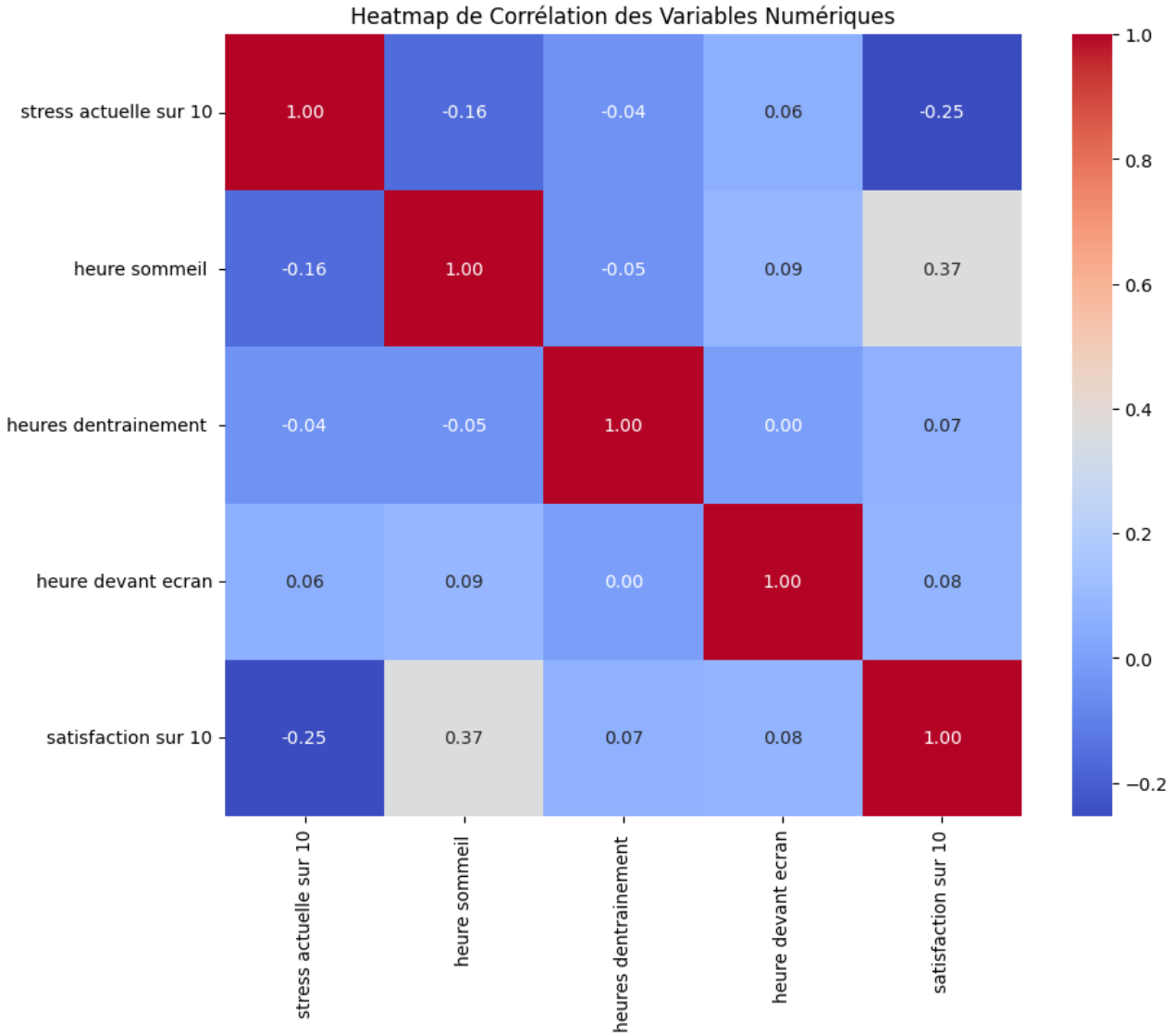
3.1.3. Exercise and Screen Time





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.

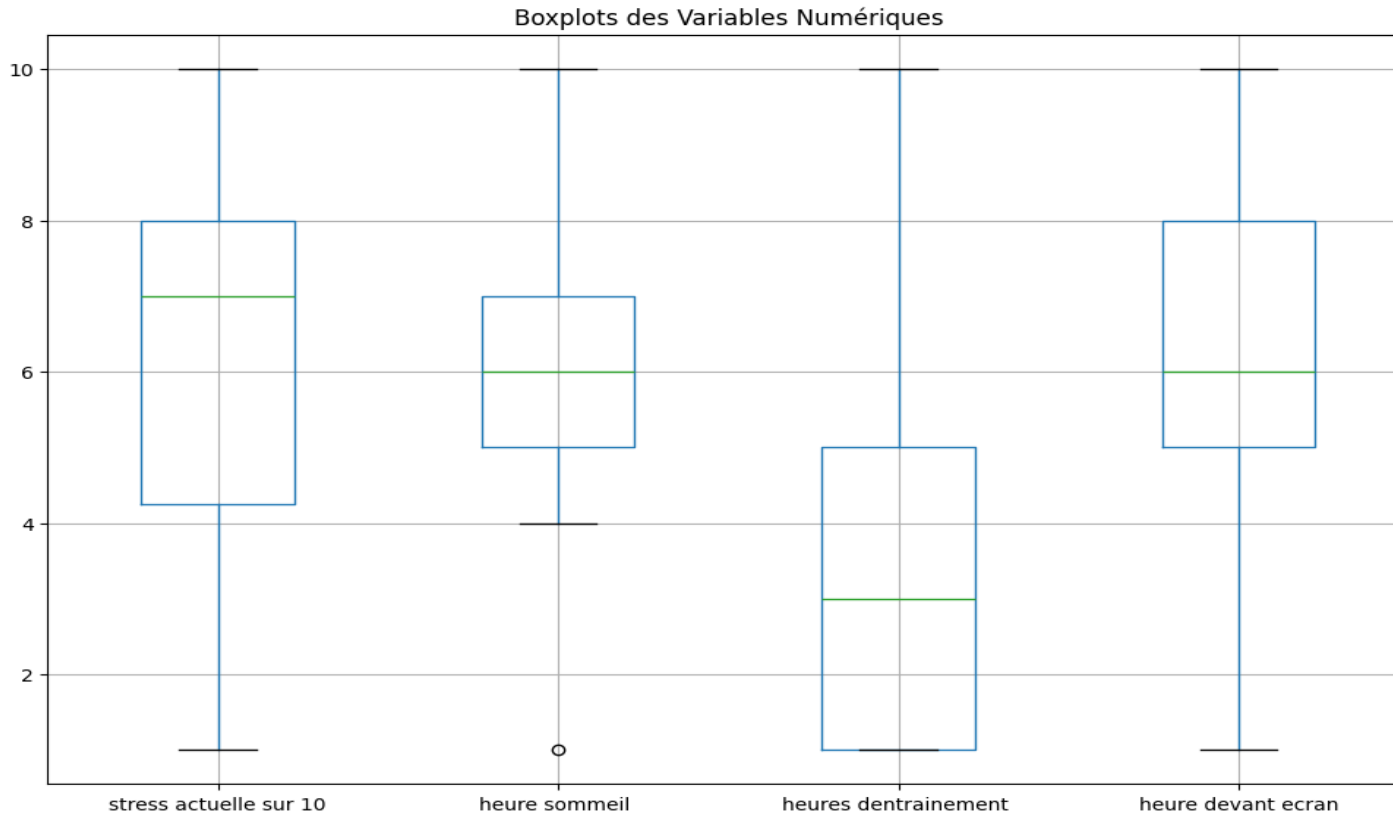
3.2. Correlation Analysis



The heatmap of the correlation matrix reveals a slight inverse relationship between stress levels and satisfaction 0.37, and a small positive correlation between hours of sleep and satisfaction, suggesting that better sleep may contribute to a higher sense of well-being.

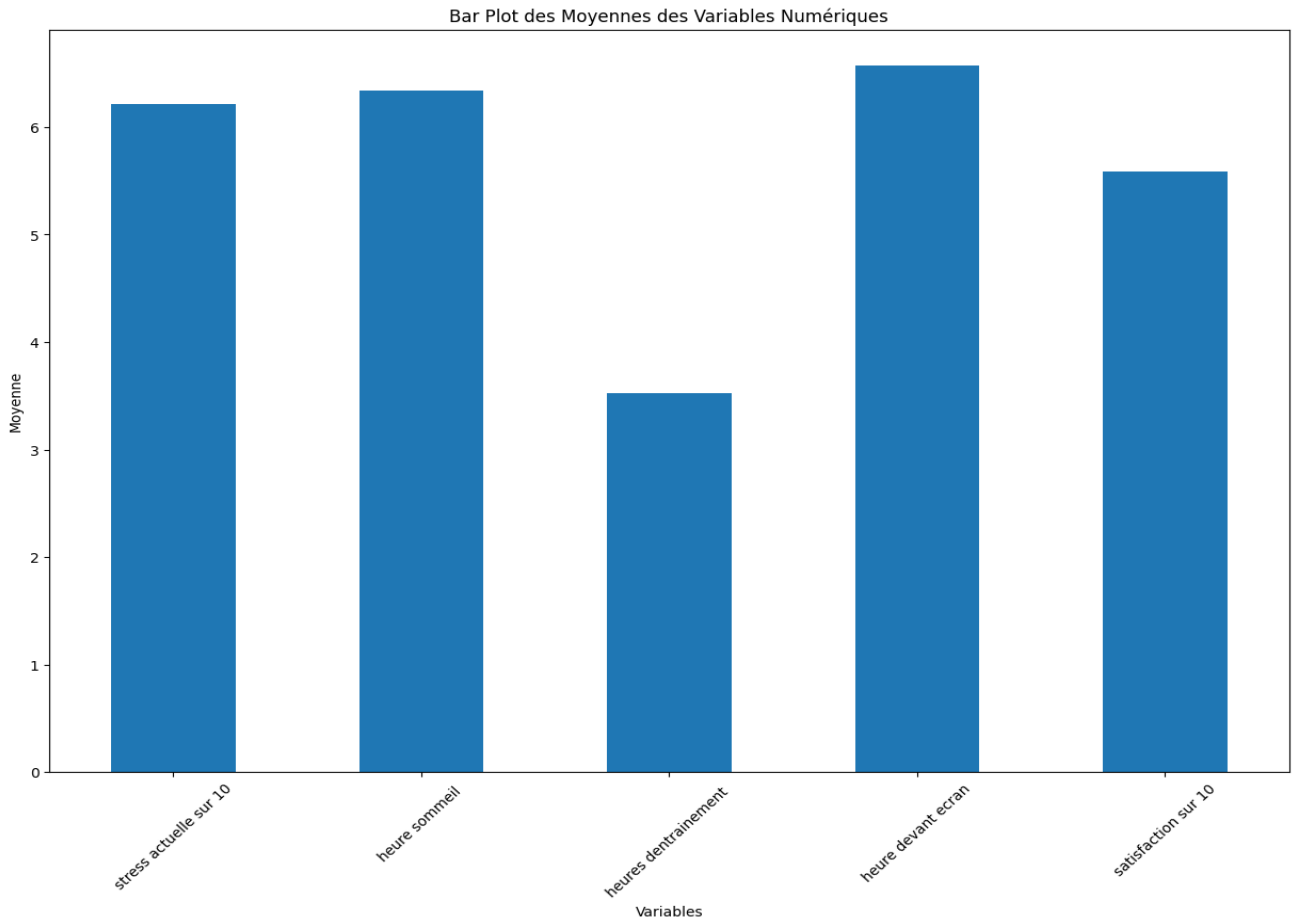
3.3. Visual Representations

3.3.1. Boxplots

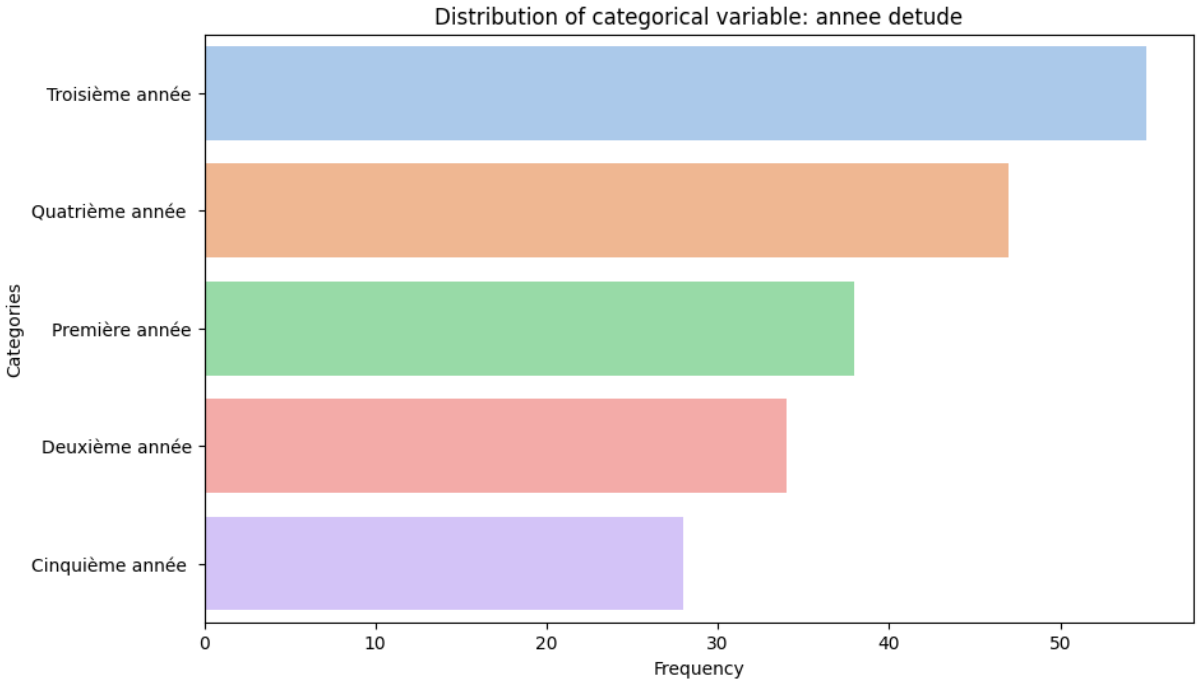


The boxplots provide a summary of the distributions for stress levels, sleep hours, exercise hours, and screen time. The median stress level is around 6-7, with a fairly even spread of data indicating consistent stress experiences among students. Sleep hours have a narrower interquartile range, centered around 5-7 hours, with some outliers indicating variations in sleep patterns. Exercise hours show a wide range, 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.

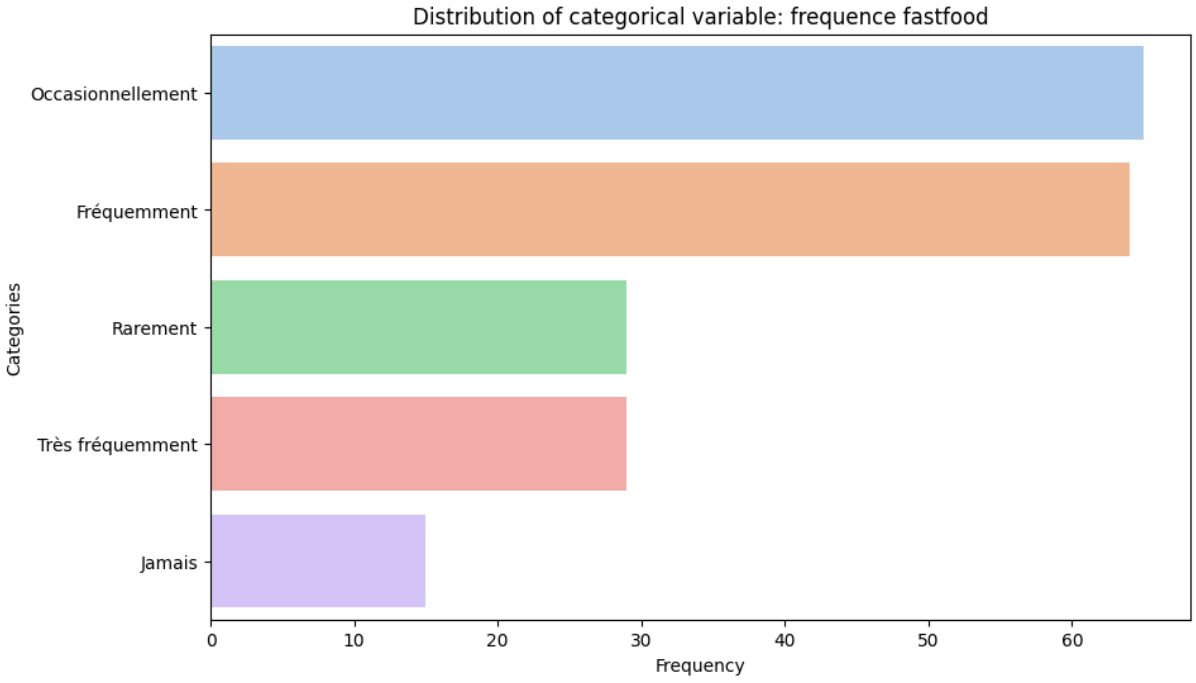
3.3.2. Bar Plots



This bar plot shows the average (mean) values for the key numerical variables in the study. It appears that the average stress level are moderately high and satisfaction score is moderate or low, 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.

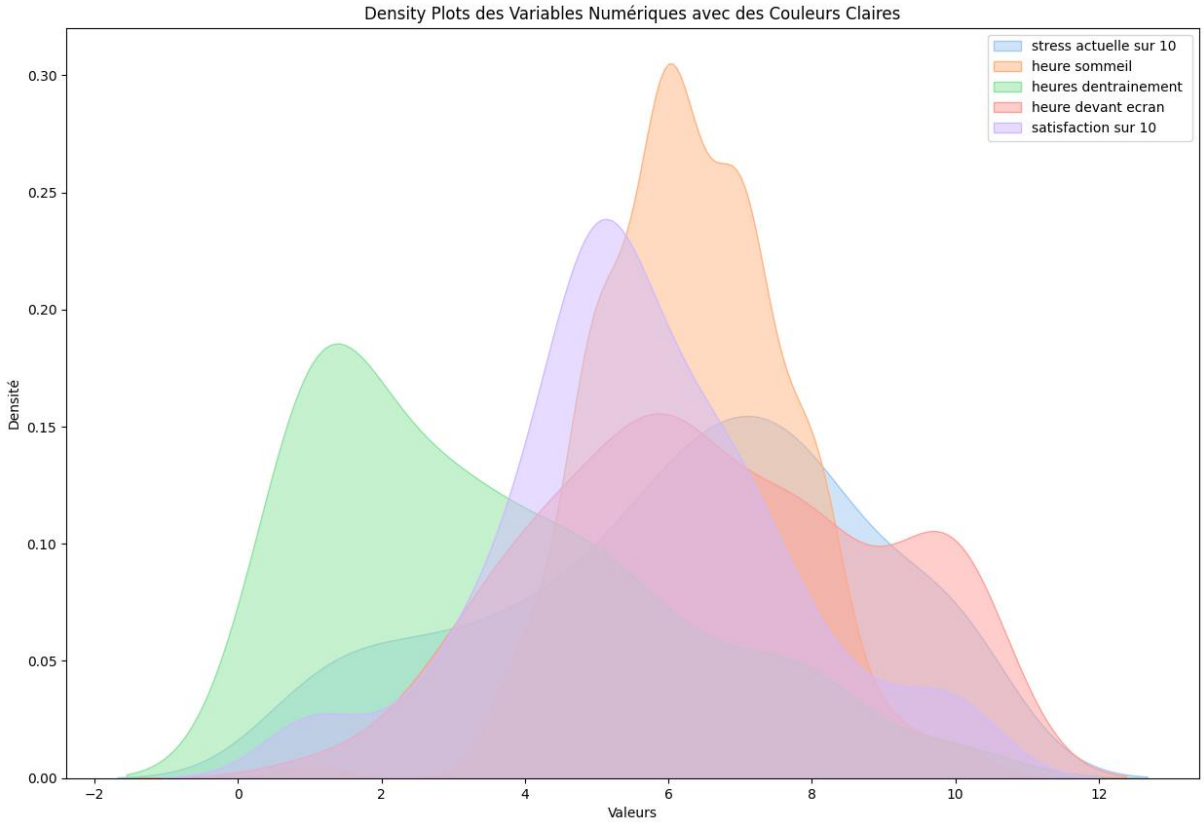


the distribution of students across different academic years. The majority of respondents are from the third year, followed by the fourth and first years. There are fewer students from the second and fifth years. This distribution can provide context for interpreting other variables, as stress levels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



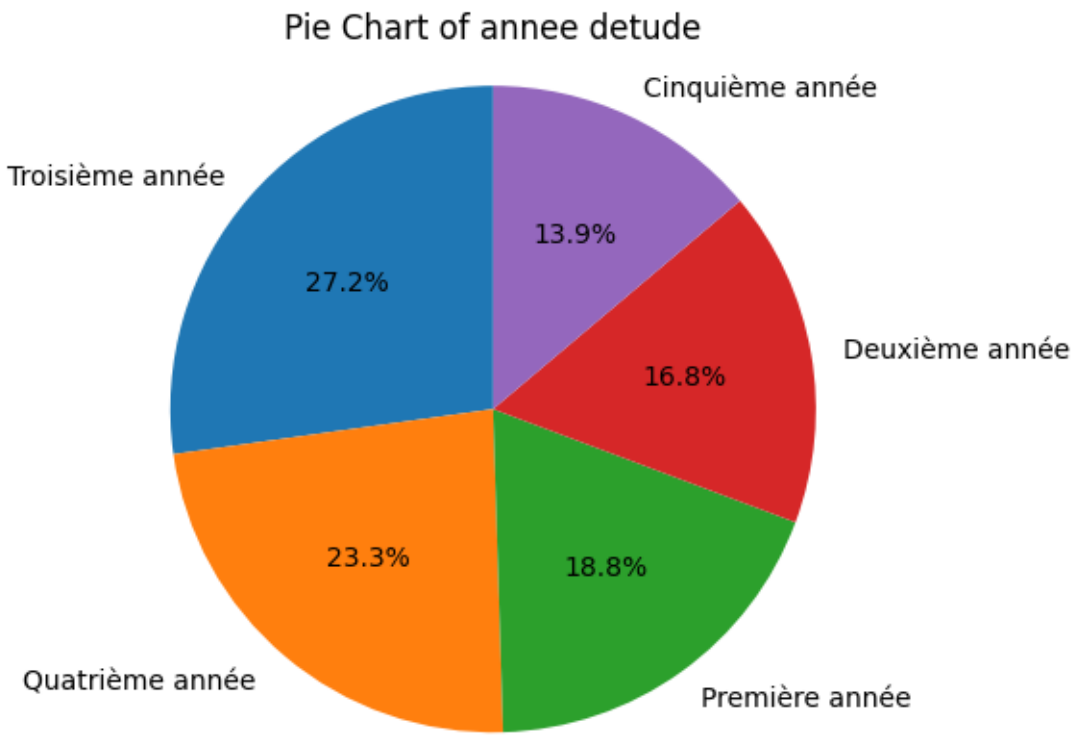
This bar chart shows how often students consume fast food. A significant number of students eat fast food occasionally, and a roughly equal proportion eat it frequently. Fewer students report rarely or very frequently consuming fast food, and a small number report never eating it. This can be an important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

3.3.3. Density Plot



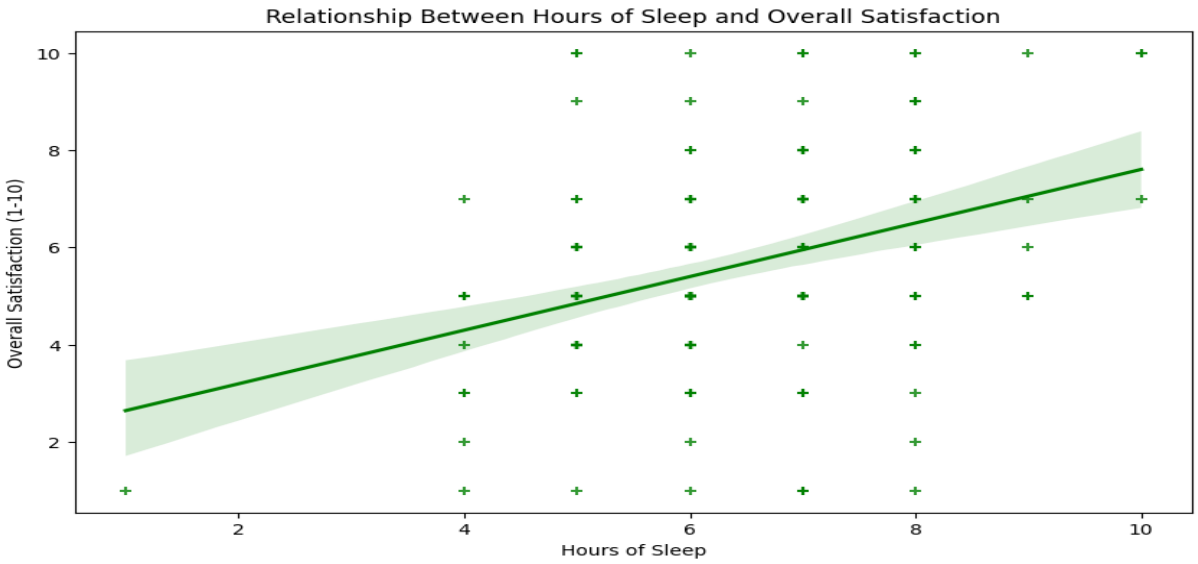
The density plots offer a smoothed visualization of the distributions for the same variables as the boxplots. They reveal that stress levels and satisfaction scores have somewhat similar shapes, suggesting that students' stress and satisfaction may be distributed in comparable ways across the population.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

3.3.4. Pie Chart



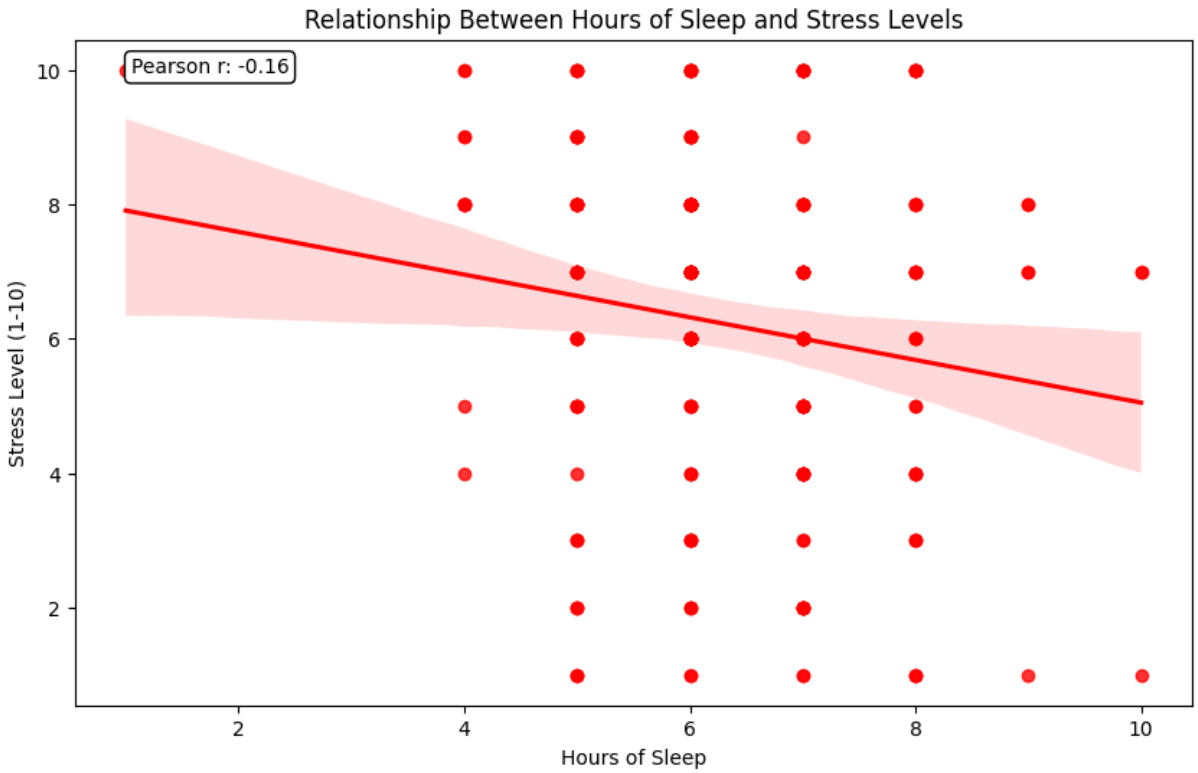
The pie charts break down the proportions of students in different categories for academic year and fast food frequency. They provide a quick visual reference for understanding the makeup of the student body in these categories, which is essential for contextual analysis when considering how these factors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

3.3.5 Scatter Plots



In the context of a scatter plot with a positive correlation (Pearson correlation coefficient of 0.37) between the number of hours of sleep and overall satisfaction, the upward trend indicates that as students get more sleep, their overall satisfaction tends to increase., the regression line represents the best-fit line that minimizes the overall distance between the observed data points and the predicted values based on the linear relationship. Essentially, it's a line that estimates the average trend in the data.

The shaded area around the regression line represents the confidence interval. In simple terms, this interval provides a range within which we are reasonably confident that the true regression line lies. It accounts for the uncertainty associated with estimating the relationship between sleep hours and overall satisfaction from a sample of data. A wider shaded area indicates higher uncertainty, while a narrower one suggests greater confidence in the accuracy of the estimated relationship. It's important to note that the regression line itself represents the average trend, while the confidence interval reflects the range of likely positions for the true relationship in the population



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

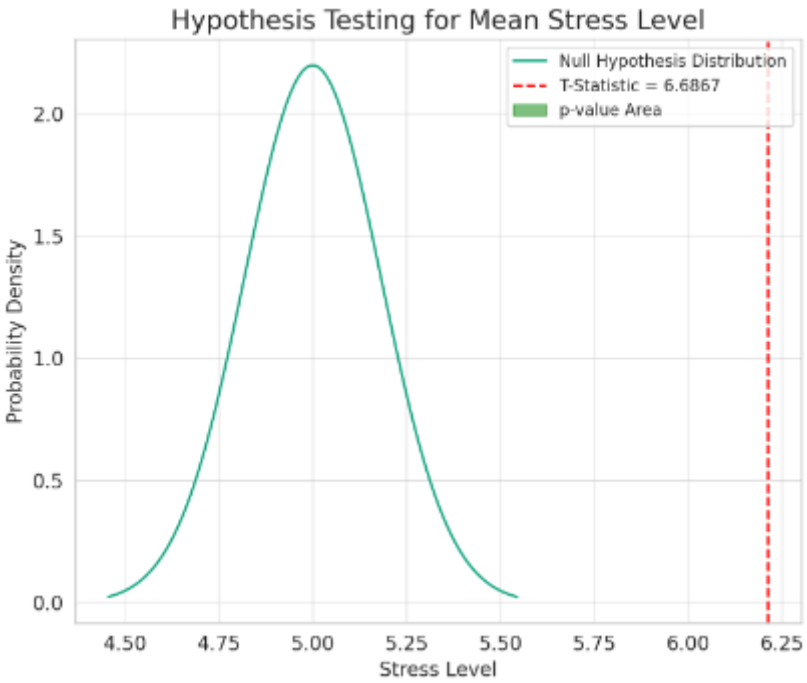
Let's consider we want to test the hypothesis that the average stress level among students is above a certain threshold, say 5 on a scale of 10.

Null Hypothesis (H0): The mean stress level of students is 5. ($\mu = 5$)
Alternative Hypothesis (H1): The mean stress level of students is greater than 5. ($\mu > 5$)

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. ($p \leq 0.5$)
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:

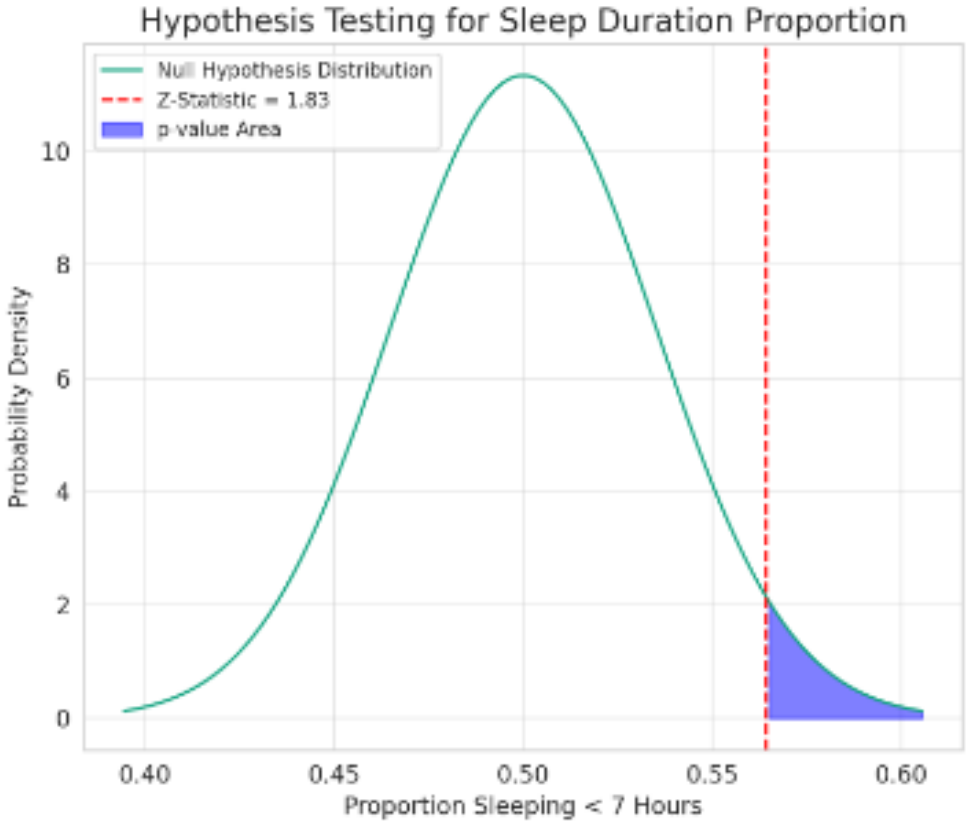
t-statistic: 6.6867

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.

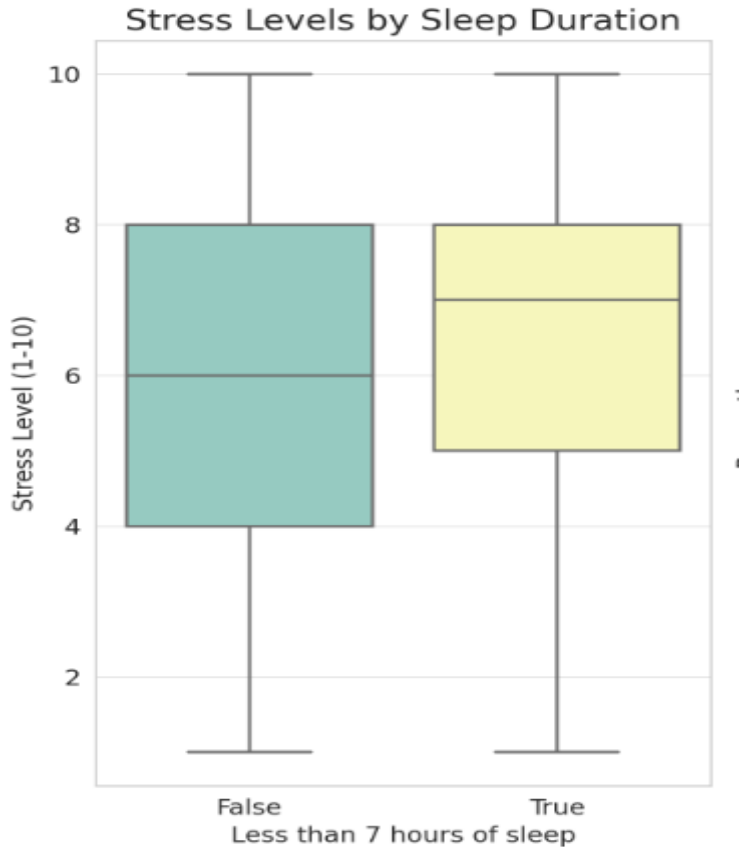
Overall Satisfaction (satisfaction sur 10):

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

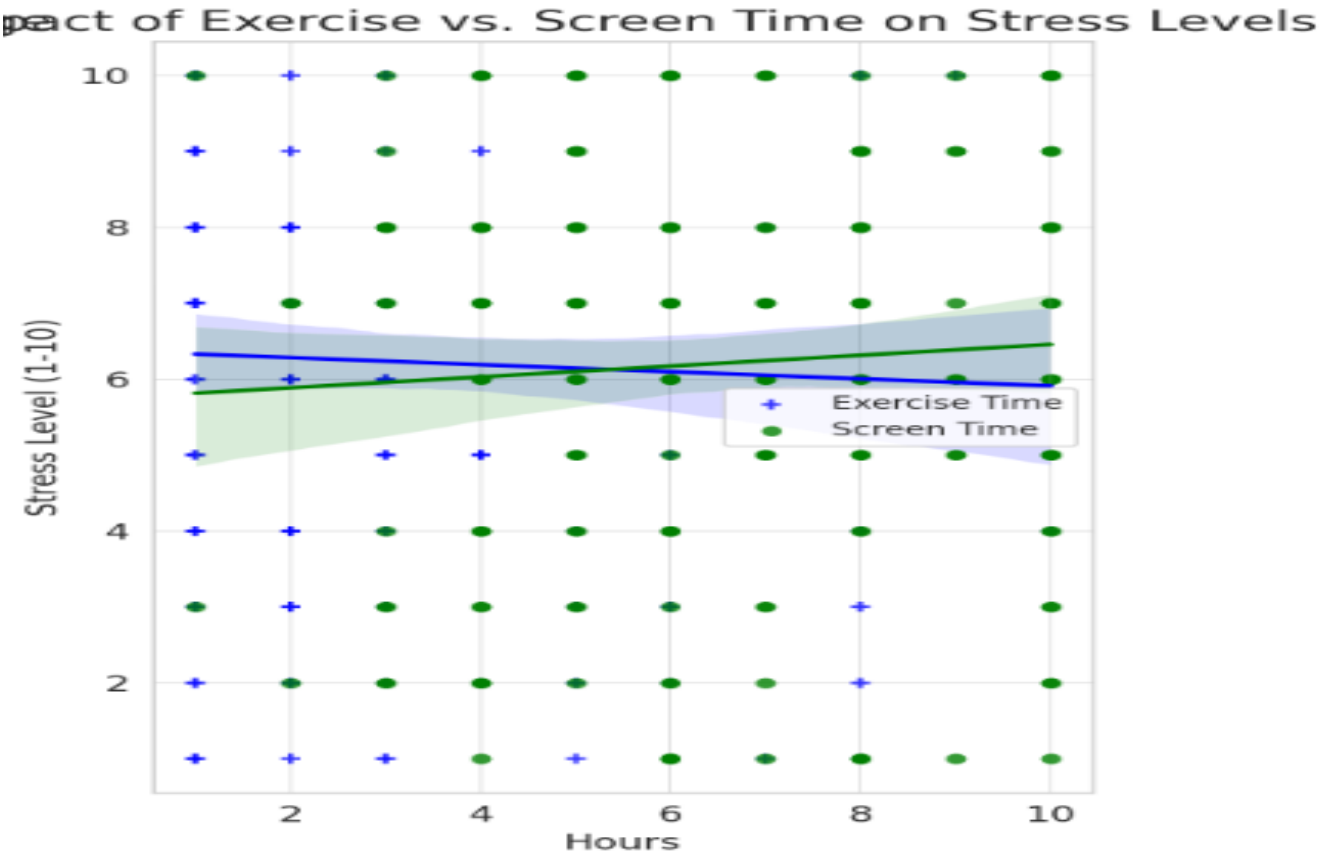


Our independent t-test comparing the average stress levels between students who get less than 7 hours of sleep and those who get 7 or more resulted in a t-statistic of 2.02 and a p-value of approximately 0.045. This indicates that there is a statistically significant difference in stress levels between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

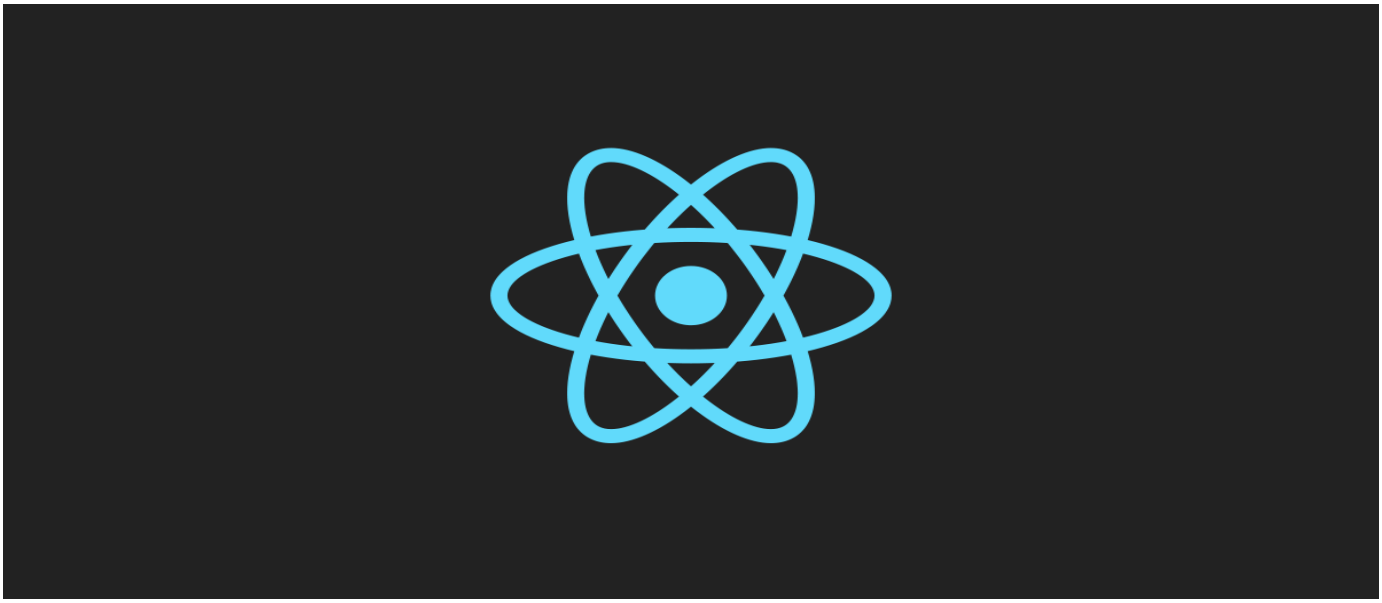
The coefficients and their confidence intervals were as follows:

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

Technologies used : React and nivo.rocks

React: React is a JavaScript library for building user interfaces, developed and maintained by Facebook. It allows developers to create reusable UI components and efficiently update the view when the underlying data changes. React follows a component-based architecture, making it easier to manage and scale complex applications. It also supports a virtual DOM, optimizing rendering performance by updating only the necessary parts of the actual DOM. React is widely used for building modern, interactive, and responsive web applications.



Nivo: Nivo is a JavaScript library for building data visualization components, particularly well-suited for React applications. It provides a set of customizable and responsive chart components that make it easy to create visually appealing and interactive data visualizations. Nivo supports various chart types, including bar charts, line charts, pie charts, scatter plots, and more. It is built with D3 and React under the hood, allowing developers to leverage the power of D3 for data manipulation while enjoying the simplicity of React for building UI components. Nivo simplifies the process of integrating complex data visualizations into React applications with its modular and customizable approach.

Architecture of the app :

The app architecture consists of a straightforward directory structure in the "src" folder:

components: This directory holds reusable React components that are used across different scenes. Examples include various chart components (BarChart, BoxPlot, LineChart, PieChart, ScatterPlot, etc.), a component for downloading PDFs (DownloadPdf), a header component (Header), a progress circle component (ProgressCircle), a scatter plot for predictions (ScatterplotPrediction), and a statistical box component (StatBox).

data: This directory contains a file (mockData.js) that likely holds mock data for testing and development purposes.

scenes: Scenes represent different sections or pages of your app. Each scene has its own directory, and within each scene directory, there are components or containers related to that specific scene. The scenes include:

bar: Scene for bar chart-related components.
boxplot: Scene for box plot-related components.
dashboard: Main scene for the overall dashboard of the app.
form: Scene related to form components.
global: Contains global components like the sidebar (Sidebar) and top bar (Topbar).
line: Scene for line chart-related components.
pie: Scene for pie chart-related components.
prediction: Scene related to prediction components.
scatter: Scene for scatter plot-related components.
App.js: The main entry point of the React application, where components and scenes are likely imported and composed.

index.css and index.js: Basic styling and the main entry point for the React app, respectively.

theme.js: A file that likely contains the theme configurations for styling the app.

Overall, this structure separates components based on their functionalities and scenes based on the major sections of the app, promoting modularity and maintainability.

User Interface (dark and light mode):

Dashboard

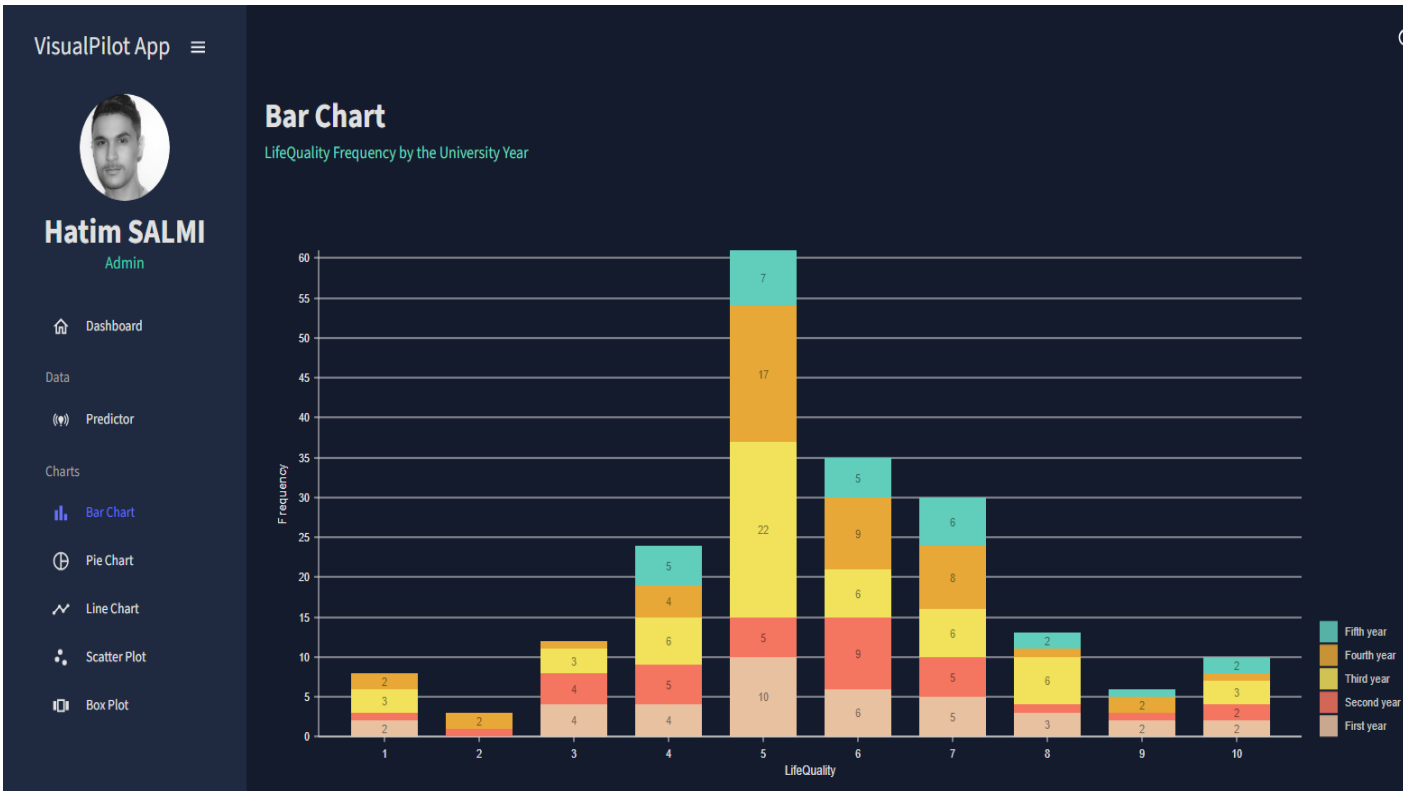
As you can see here, in this dashboard we have a sidebar which the different tabs that include charts and plots…, in the top bar there are some statistical analyses that we wanted to created visually, it showcases things like How many students surveyed and the pourcentage of those who have a good life quality, mediocre or bad life quality … We assumed here that any student that has 7 or more in their life quality rating has a good life quality , anything less than that is mediocre or bad life quality

Below that we have on overview of the different charts and plots that we can see in more details when clicking on their specific tabs.

Light mode:

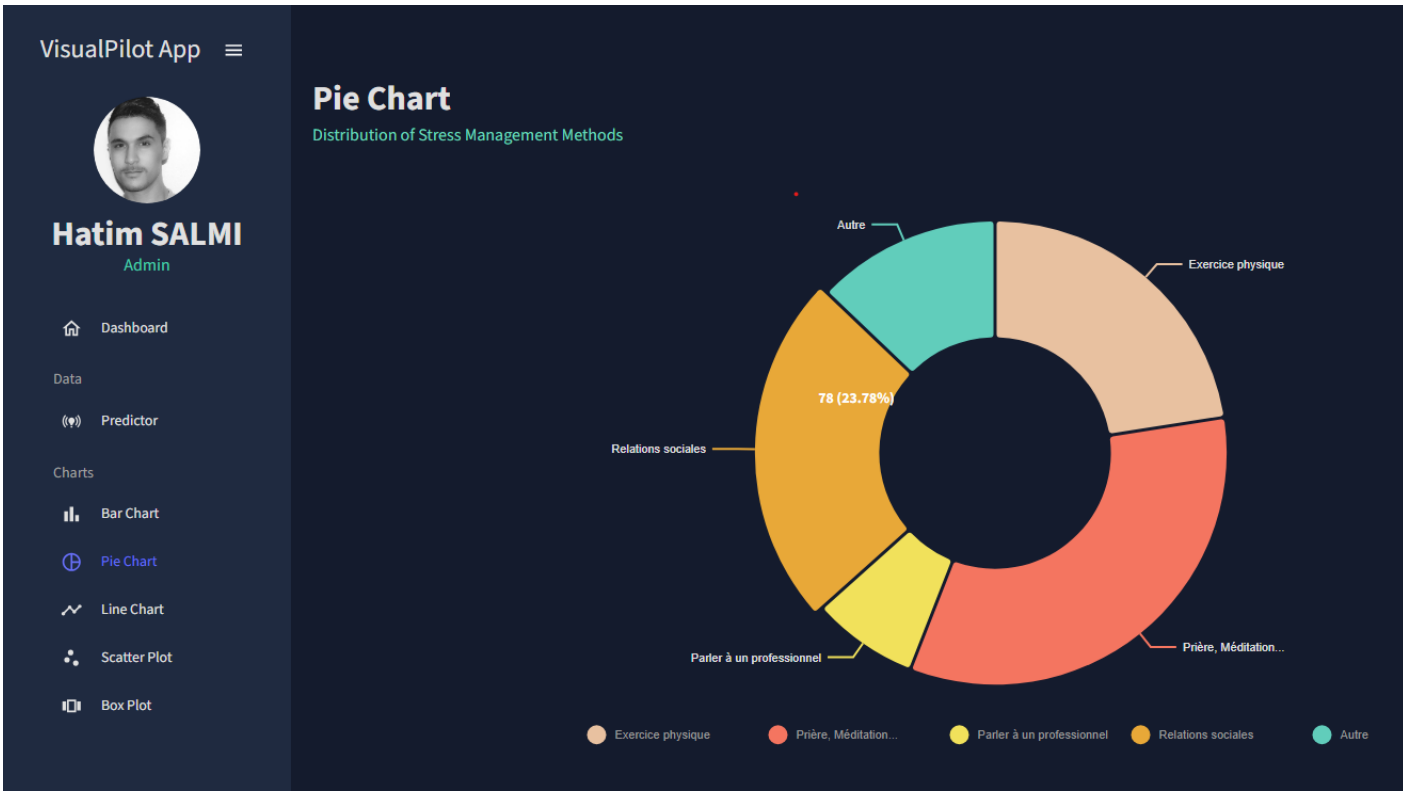
Charts :

Barchart :



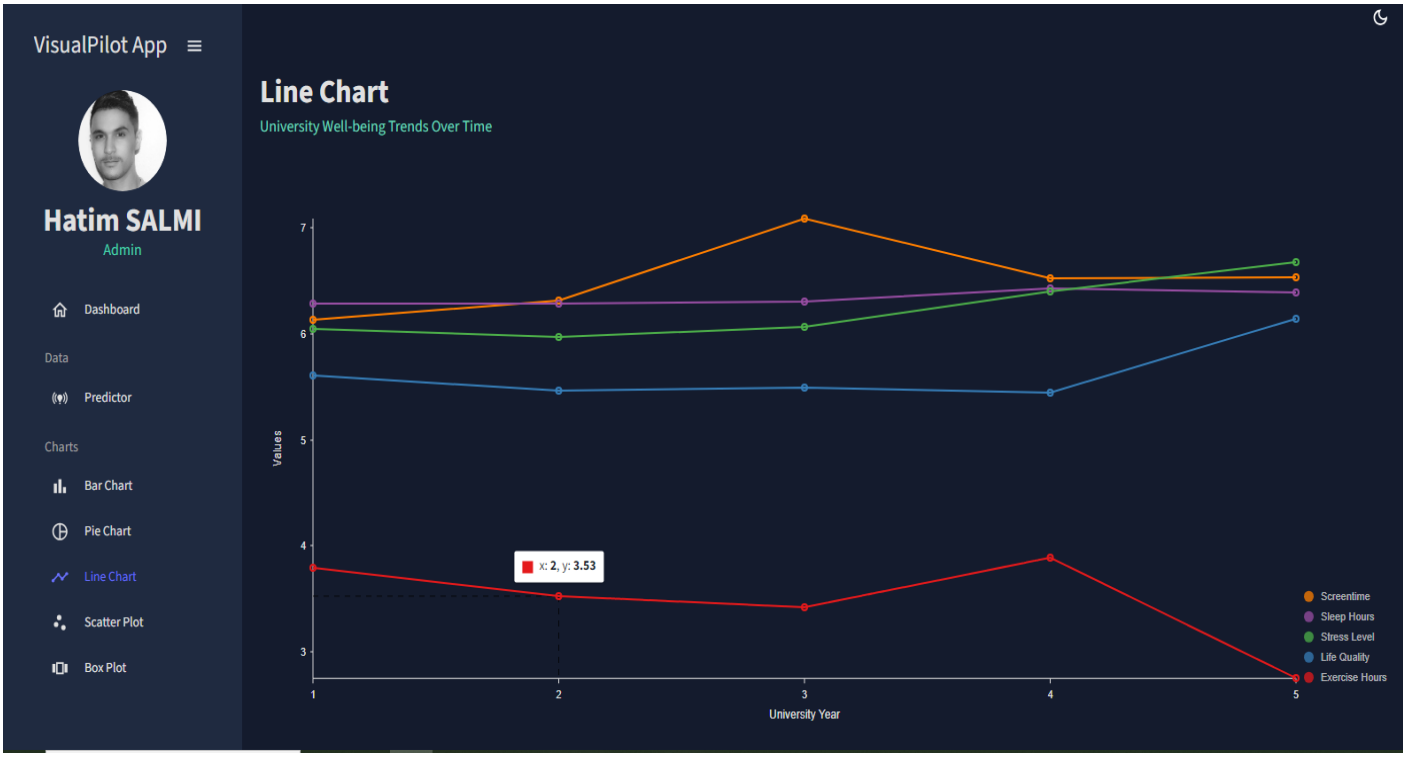
The bar chart here showcases the frequency of the students of each university year depending on their Life Quality rating that they gave , we can clearly see that the majority fall into 4-7 range which assumes that most of the students think they have a mediocre life quality

PieChart :



The pie Chart shows the pourcentage of the students that use each of the stress management methods above , for example around 24% use social interactions to manage their stress , the majority use that and also prayer , mediation and physical exercice , rarely do people visit a mental health professional or use other management methods.

Line chart :



Exercise Hours Per Week:

The average exercise hours per week vary across university years, with the highest average in the fourth year.
Analyzing the trend, it appears that there might be a positive correlation between exercise hours and life quality.

Stress Level:

Stress levels show some fluctuations across university years, with a peak in the fifth year.
There might be a potential negative correlation between stress levels and life quality.

Sleep Hours:

The average sleep hours per day remain relatively consistent across university years.
There might not be a strong correlation between sleep hours and life quality based on the provided data.

Screentime Hours Per Day:

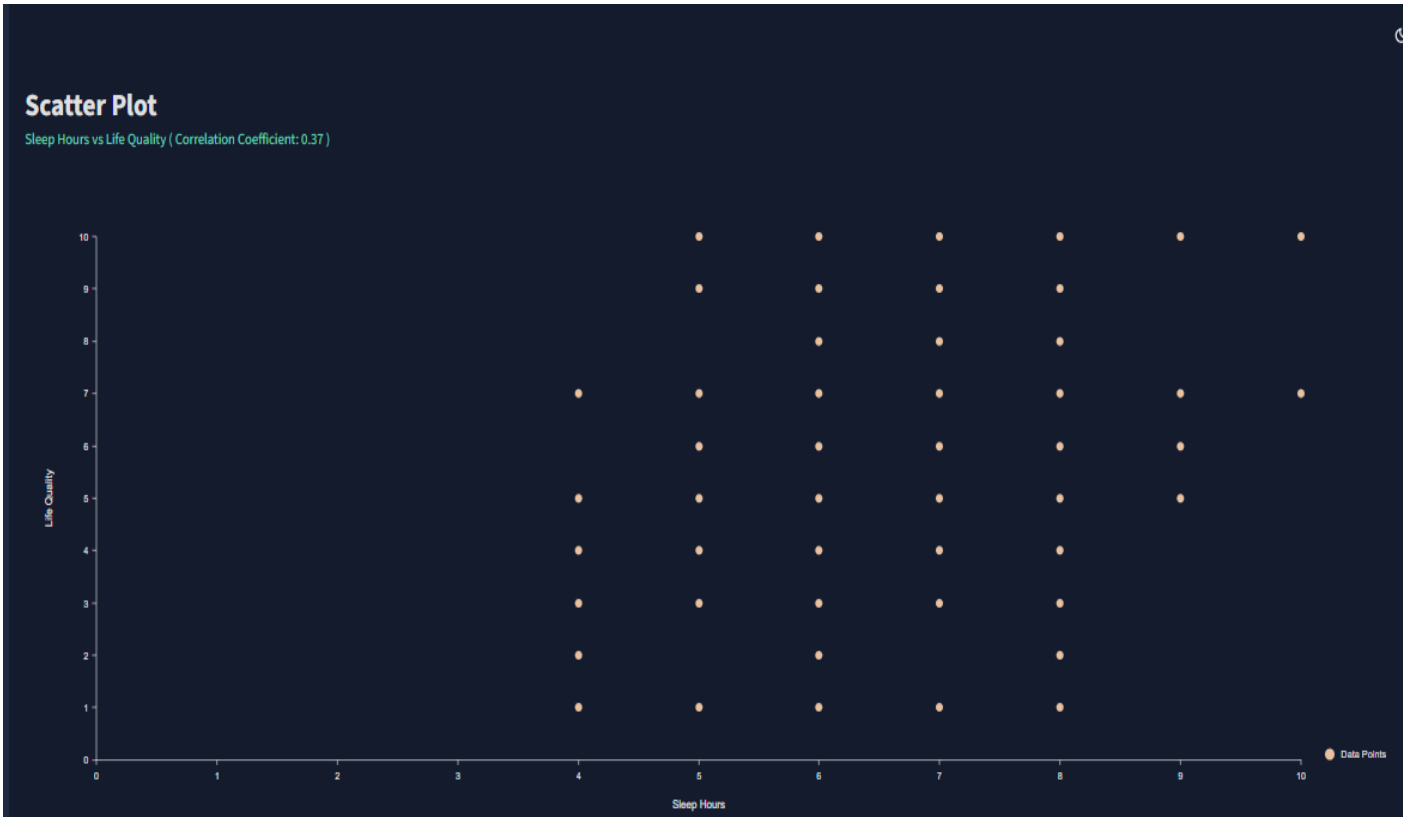
Screentime hours per day exhibit some variation, with a notable increase in the fourth and fifth years.
Increased screentime might be associated with changes in life quality, potentially indicating a negative correlation.

Life Quality:

Life quality scores show variations across university years, with the highest average in the fifth year.
The overall trend suggests a potential improvement in life quality as students progress through their university years.

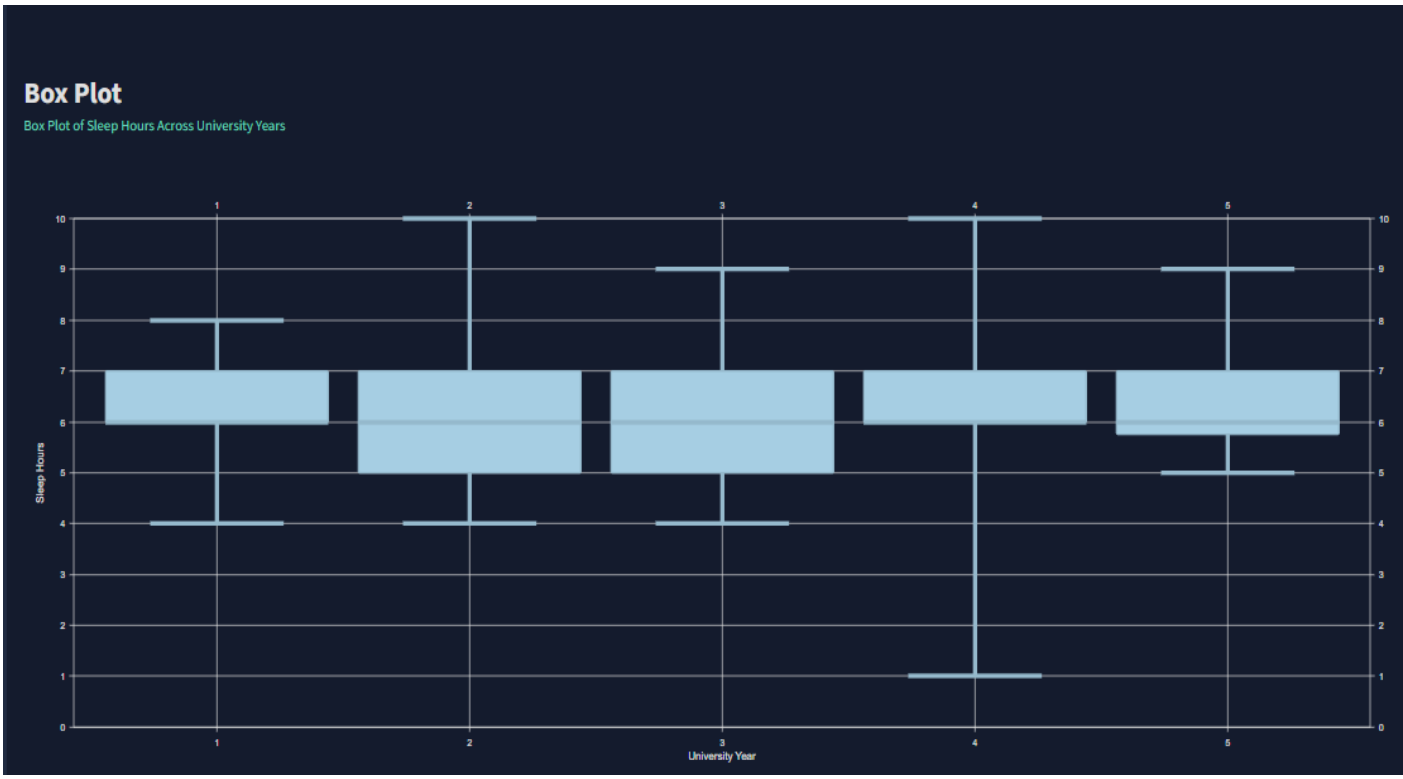
Overall, these observations are based on the provided average values, and individual experiences may vary. Correlation does not imply causation, and other factors not included in the dataset could influence these relationships. Further statistical analysis, such as correlation coefficients, regression analysis, or machine learning models, could provide more insights into the strength and significance of these relationships.

Scatterplot :



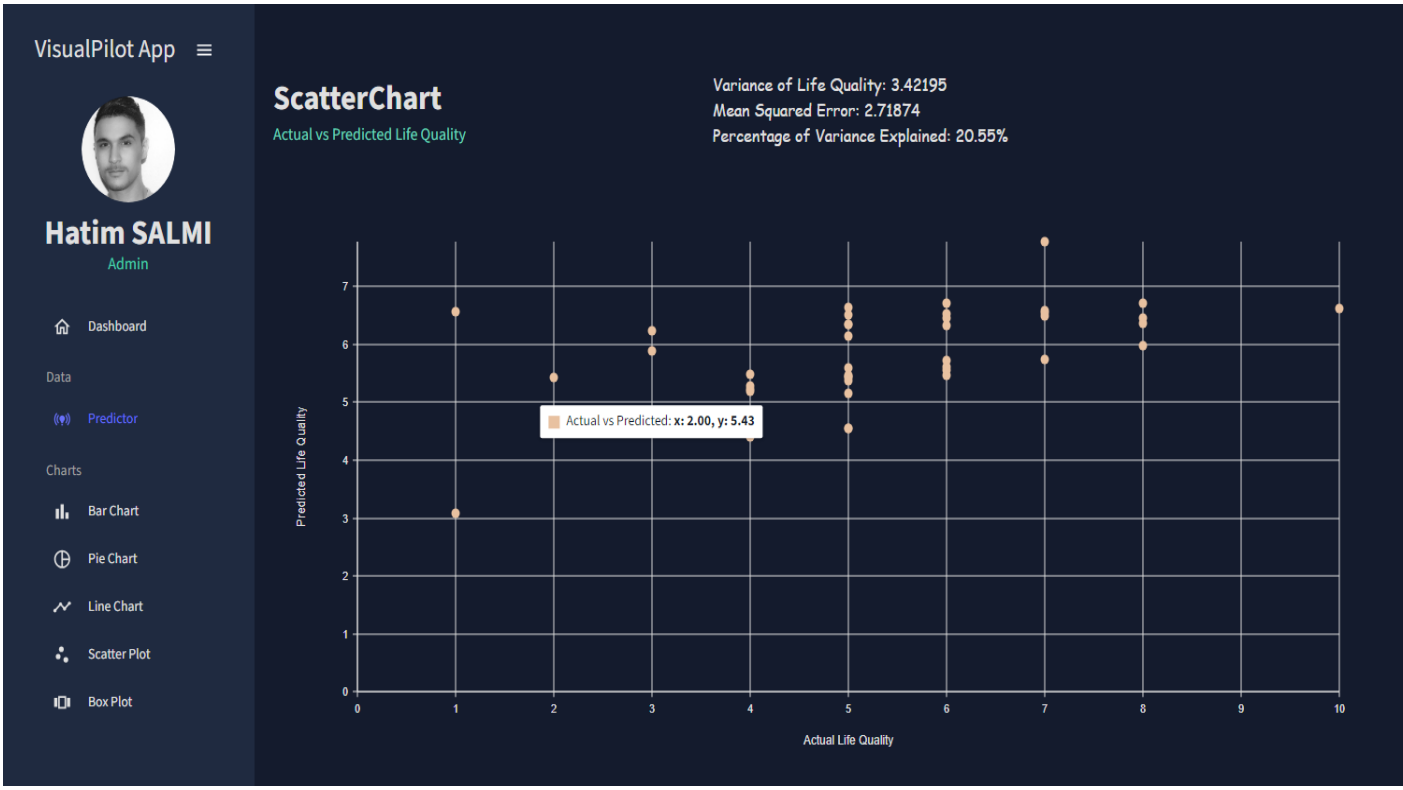
A coefficient of 0.37 showcases a positive correlation between the number of hours of sleep and overall satisfaction , we went into depth in the chapters before

BoxPlot :



These 5 box plots showcase the sleep hours distribution across the university years.

Predictor :



Variance of Life Quality (3.42195):This value represents the spread or variability in the actual Life Quality scores. In other words, it indicates how much the actual Life Quality values deviate from their mean. A higher variance suggests a wider range of Life Quality scores.

Mean Squared Error (MSE - 2.71874): The MSE is a measure of how well the predicted values align with the actual values. It calculates the average squared difference between each predicted and actual Life Quality score. A lower MSE indicates that the model's predictions are closer to the actual values.

Percentage of Variance Explained (20.55%): This value represents the proportion of the total variance in the actual Life Quality scores that is explained by the model. In your case, the model explains approximately 20.55% of the variability in Life Quality based on the provided features. A higher percentage suggests a better ability of the model to capture the patterns in the data.

Interpretation: The model is capturing a portion of the variance in Life Quality, as indicated by the percentage of variance explained. However, it's important to note that there is still a substantial amount of variability that the model has not accounted for (79.45%). The MSE is relatively low (2.71874), indicating that, on average, the model's predictions are reasonably close to the actual Life Quality scores. This is a positive aspect, but the percentage of variance explained suggests room for improvement.

In summary, while the model is making reasonably accurate predictions, there is still room for improvement to better capture the complexity of factors influencing Life Quality.

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.

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