Data Analytics Group Project
Undergraduate Health Concern

The Johns Hopkins University Carey Business School

BU.510. 650.X1. FA24: Data Analytics

Professor: Dr. Ruxian Wang

Team members:

Jingtian Cao, Yunhong Wang, Cassie Liu, Caesar Feng, Manman Wei

December 10, 2024

Group member contributions: Jingtian Cao is responsible for Data Description, Cassie Liu and Yunhong Wang are responsible for Methodologies: Data Preparation and Generalized Linear Model. Manman Wei and Caesar Feng are responsible for Decision Trees model, Research Background description, Research Objectives and Questions.

1. Research background

In today's academic environment, college students face unprecedented levels of stress and health challenges. According to the American Psychological Association (APA), over **60% of college students met the criteria for at least one mental health issue during the 2020–2021 academic year**, highlighting the scale of the problem (American Psychological Association, 2022). Stressors such as academic pressure, financial concerns, and social obligations play a significant role in this growing crisis. Additionally, a Gallup poll conducted in 2023 revealed that **66% of college students experienced frequent stress**, and over half reported feeling worried on most days (Seltzer, 2023). These findings illustrate how stress and mental health struggles have become the potential culprits behind students' overall health.

Another key factor exacerbating health concerns among students is the prevalence of sedentary behavior and poor lifestyle habits. A recent study highlighted that **university students spend an average of 7.5–8 hours per day engaged in sedentary activities**, a trend associated with increased risks of obesity, cardiovascular issues, and other long-term health problems (University of South Carolina, 2020). In parallel, irregular sleep schedules, lack of exercise, and poor dietary choices further contribute to the deterioration of health among college students. A systematic review found that **sedentary behavior among university students is higher compared to the general population**, amplifying negative physical and mental health outcomes (Carter, Carter, & Du, 2020). Taken together, these lifestyle and behavioral patterns suggest that health is not solely determined by medical or physical attributes but is deeply influenced by a broader spectrum of factors, including mental well-being, daily routines, and activity levels.

This study aims to explore the complex and evolving nature of student health conditions and assessment through the analysis of a publicly available student health risk assessment dataset. The dataset comprises health risk assessments for **1,000 undergraduate students**, evaluated by physicians who categorized each student's health risk level as **low**, **moderate**, or **high**. The dataset includes three primary categories of information:

1. **Physical (Objective) Data**: Demographic and medical metrics, such as age, gender, blood pressure, and heart rate.

2. **Lifestyle and Mental Health Data**: Subjective assessments, including sleep quality, physical activity levels, and self-reported stress.
3. **Behavioral Data**: Daily activity patterns, such as hours spent studying, working on projects, or engaging in recreational activities.

2. Research Objectives and Questions

The primary goal of this research is to identify the key determinants of physicians' health risk evaluations. Using quantitative methods such as **regression analysis** and **decision trees**, we try to address the following questions:

- Which student health attributes most strongly influence a physician's diagnosis that a student has "high health risk"?
- Based on specific health attributes, how can students' overall health risk levels be classified?

This research also aims to provide actionable recommendations for college students to improve their health outcomes. For instance, we hypothesize that **mental health stress levels**, **sleep quality**, and **hours spent studying** are the most significant contributors to health risk. By analyzing and testing these assumptions, the findings will serve as a guide for developing practical strategies to support student well-being.

3. Data Description

a. Demographics - Age, Gender

This dataset includes typical undergraduate ages ranging from 18 to 25, with an average age of 21 (see Appendix 1 for details). The distribution of gender is nearly balanced between male and female respondents (see Appendix 2 for details).

b. Distributions

- Objective Health Factors
    - Heart Rate: Ranges from 50 to 100 bpm (see Appendix 3 for details).
    - Blood Pressure (Systolic): Ranges from 90 to 170 mmHg (see Appendix 4 for details).
    - Blood Pressure (Diastolic): Ranges from 60 to 110 mmHg (see Appendix 5 for details).
    - Stress Level (Biosensor): Ranges from 1 to 10, with a mean of 5.48 (see Appendix 6 for details)

  - Stress Level (Self-Report): Ranges from 1 to 10, with a mean of 5.36 (see Appendix 7 for details)
- Subjective Health Factors
  - Physical Activity Levels: Categorized into High, Moderate, and Low. Moderate level of activity is reported by most students (see Appendix 8 for details).
  - Sleep Quality: Categorized into Good, Moderate, and Poor, with a large portion of respondents reporting "Good" sleep (see Appendix 9 for details).
  - Mood: Categorized into Happy, Neutral, and Stressed, where the counts of Happy and Neutral take most of the portion of data (see Appendix 10 for details).
- Behavioral Health Factors
  - Study Hours: Daily hours spent on studying, ranging from 5 to 60 with a mean of 30.23 (see Appendix 11 for details).
  - Project Hours: Daily hours spent on projects, ranging from 0 to 32.72, with a mean of 14.89 (see Appendix 12 for details).
  - Workload: Daily total hours spent on studying and projects, ranging from 11.82 to 82.14, with a mean of 45.11.

c. Variable Categorization

Category	Variables	Description	Our initial Analysis & Assumptions
Objective Health Factors	Age, Gender, Heart rate, Blood pressure, Stress Level.	Basic demographic information and measurable health indices	Extremely high or low heart rate and high stress level could lead to high health risk level
Subjective Factors	Physical activity, Sleep quality, and Mood.	Lifestyle, Mental Health Evaluation metrices measuring in three levels.	Lack of physical activity, poor sleep quality and stressed moods are linked to higher health risks.

Behavioral	Study Hours, Project Hours, Workload = Study + Project Hours.	Daily hours spent on studying, projects, and the sum of them.	Higher workloads are expected to increase health risk levels.

4. Methodologies

a. Data Preparation: Generalized Linear Model

Since we have 1000 rows in total, we set the first 800 as our training dataset and the remaining 200 rows as testing dataset.

For our dependent variable “Health risk level”, it has three outcomes: High, Moderate and Low. After assigning them with 3,2,1 respectively, we set Moderate and Low health risk level as 0, High health risk level as 1, and named this variable as “Binary_Health_Risk” for our further analysis. Then the binomial family was selected to account for the distribution of this health risk level with binary characteristics.

For our remaining 12 independent variables, there are 4 variables in text format. We converted them into numerical values, including:

Independent Variable Names	Numerical Values Description
Gender	M (Male)=1, F(Female)=0
Physical Activities	Low = 1, Moderate =2, High = 3
Sleep Quality	Poor=1, Moderate=2, Good = 3
Mood	Stressed =1, Neutral = 2, Happy = 3

b. Generalized Linear Model: What factors caused high risk level?

To investigate the relationship between Health risk level and the corresponding significant independent variables, we employed a Generalized Linear Model (GLM). The GLM framework allows for modeling non-linear relationships while accommodating response variables that follow distributions other than the normal distribution.

The dependent variable, “Binary_Health_Risk”, and the other 12 columns were set as independent variables for model generation. Here is our result and we can see that **Age with 90% significance level, Stress Level biosensor (Stress Level detect by biosensor), Stress Level Self Report(Stress Level reported by students themselves), Physical activity, Sleep Quality are the significant variables with 99% significance level.**

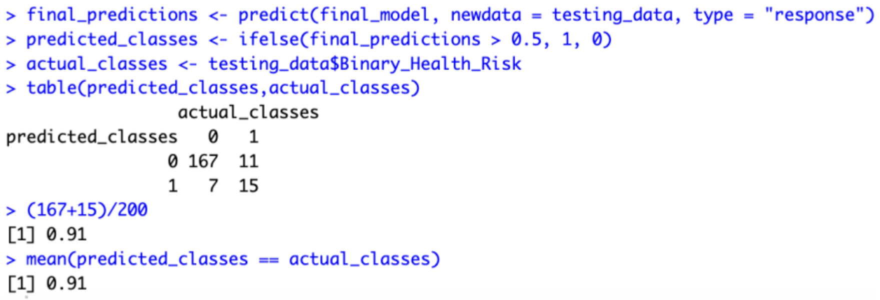
We used the above significant variables for further GLM generation and called it “final_model”. From the result we can see that since the AIC of the first model is 396.05, and the

AIC of our final model decreased to 388.35, thus the final model will be better for our further prediction.

$$logit(\mu) = -13.622 + 0.180 * Age + 0.656 * \mathit{Stress\_Level\_Biosensor} + 0.534* \mathit{Stress\_Level\_Self\_Report} + 1.325 * \mathit{Physical\_Activity} + -1.134 * \mathit{Sleep\_Quality}$$

$$\mu = 1 / (1 + exp\ (-(logit(mu) = -13.622 + 0.180 * Age + 0.656 * \mathit{Stress\_Level\_Biosensor} + 0.534 * \mathit{Stress\_Level\_Self\_Report} + 1.325 * \mathit{Physical\_Activity} + -1.134 *\mathit{Sleep\_Quality})))$$

Then we used our testing dataset to examine the accuracy of our final model. From the summary we can see that the accuracy ratio is 91%.



```
> final_predictions <- predict(final_model, newdata = testing_data, type = "response")
> predicted_classes <- ifelse(final_predictions > 0.5, 1, 0)
> actual_classes <- testing_data$Binary_Health_Risk
> table(predicted_classes,actual_classes)
                 actual_classes
predicted_classes   0   1
                0 167  11
                1   7  15
> (167+15)/200
[1] 0.91
> mean(predicted_classes == actual_classes)
[1] 0.91
```

c. Data Preparation: Decision Tree

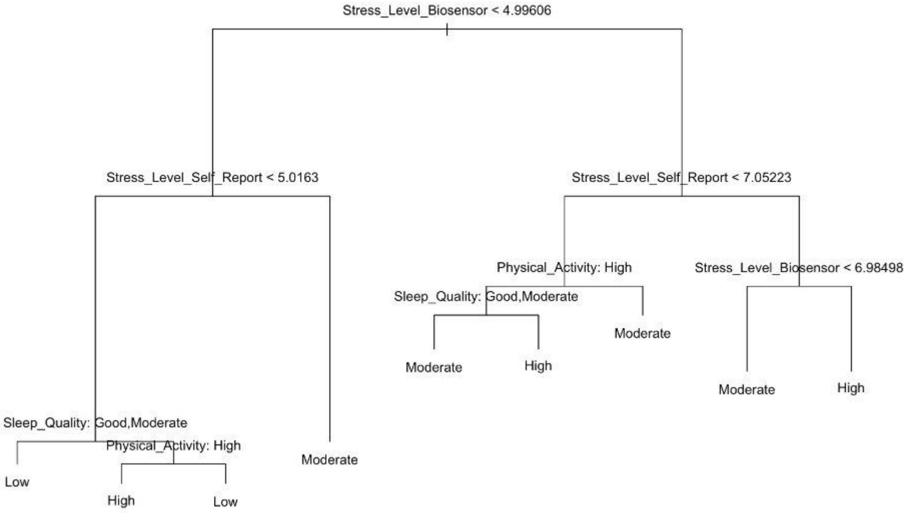
Same to the Generalized Linear Model data preparation, the dataset is split into a training set containing the first 800 rows and a testing set containing the remaining 200 rows.

The categorical variables are converted to factors to ensure proper handling during the Decision Tree Model development. The categorical data includes “Gender”, “Physical_Activity”, “Sleep_Quality”, “Mood”, and “Health_Risk_Level”, which were initially stored as character strings. They are transformed into factor variables, which allows the model to treat them appropriately as categorical predictors rather than numerical or text-based data.

d. Decision Tree Model: how can students’ overall health risk levels be classified?

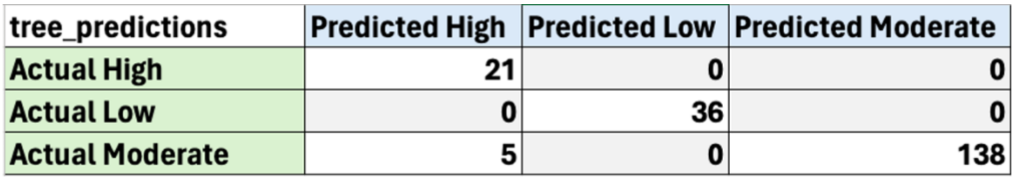
To examine the relationship between Health Risk Level and the corresponding significant independent variables, Decision Tree Model is developed. The Decision Tree Model provides a non-parametric approach that segments the data into subsets based on splitting rules from the predictors. The intuitive visualization of the relationships between variables is derived from the non-linear interactions and patterns within the data.

The Decision Tree has 9 terminal nodes. It identifies “Stress_Level_Biosensor” and “Stress_Level_Self_Report” as the most significant predictors of health risk levels, with “Sleep_Quality” as secondary protective factor that moderate risk levels and “Physical_Activity” as the conditional predictor to be either protective or risk-augmenting, depending on the context.


Stress_Level_Biosensor < 4.99606
Stress_Level_Self_Report < 5.0163
Stress_Level_Self_Report < 7.05223
Physical_Activity: High
Stress_Level_Biosensor < 6.98498
Sleep_Quality: Good,Moderate
Moderate
Moderate
High
Moderate
High
Sleep_Quality: Good,Moderate
Physical_Activity: High
Moderate
Low
High
Low

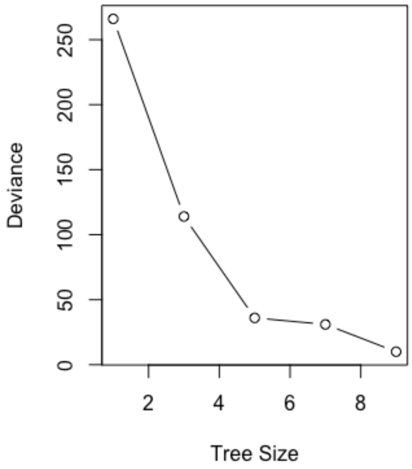
The residual mean deviance of 0.06139 reflects a strong fit to the training data. The misclassification error rate of 0.00625 indicates high accuracy in predicting health risk levels, which showcases the testing accuracy, making the model a reliable tool for health risk assessment.

The confusion matrix derived from the Decision Tree Model has correctly classified all High Risk and Low Risk with only five misclassifications in the Moderate Risk category. The minimal error rate indicates the model's effectiveness in predicting three health risk levels. The Decision Tree Model with a test accuracy of 97.5% indicates its strong performance on predicting health risk levels, which further proves it to be a reliable tool for predicting health risk levels.



tree_predictions	Predicted High	Predicted Low	Predicted Moderate
Actual High	21	0	0
Actual Low	0	36	0
Actual Moderate	5	0	138

During the pruning process, the graph is formulated as part of the pruning process to evaluate the optimal size of the decision tree based on its performance. Cross-validation is conducted to calculate the deviance for trees of varying sizes, starting from the largest unpruned tree down to a root-only tree. The optimal size is determined by identifying the size with the lowest deviance which indicates the best balance between accuracy and simplicity. The relationship between the tree size and deviance indicates the pattern that additional complexity in the unpruned tree does not harm the model's performance. Therefore, the pruned tree remains identical to the unpruned tree with nine terminal nodes. Since every split contributes to reducing deviance and improving predictions, the model remains high accuracy without requiring additional pruning.



5. Results and Conclusion

From our GLM result, we can find that the factors caused high health risk level including Age (with coefficient 0.18), Stress Level biosensor (with coefficient 0.656), Stress Level Self Report (with coefficient 0.534), Physical activity (with coefficient 1.325), Sleep Quality (with coefficient –1.134), which has already been examined by our testing dataset with 91% accuracy.

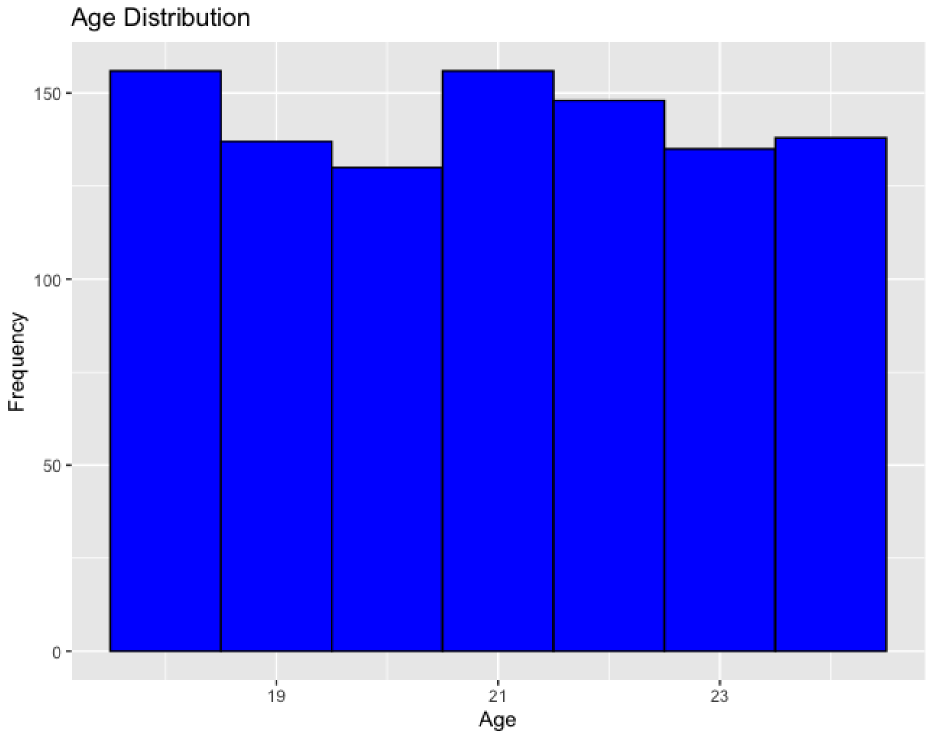
Surprisingly, we found a positive coefficient of 1.325 for physical activity, which leads to the conclusion that the more physical activity an undergraduate student has, the higher the health risk. Peter et al. (2021) explained that participation in physical activity should be promoted, but there are potential risks associated with too many hours of weekly physical activity, as they observed a U-shape relationship between the number of hours of weekly physical activity and cardiovascular disease and all-cause mortality in a U-shaped relationship, with those who participated in 2.6 to 4.5 hours of recreational physical activity per week being at the lowest risk. Therefore, we can conclude that undergraduate students with moderate levels of physical activity, lower stress levels, and better sleep quality have lower levels of health risks.

From our Decision Tree Model result, we find out that the most significant predictors are Biosensor Stress and Self-Reported Stress: 1. lower Biosensor Stress (< 4.99606) and lower Self-Reported Stress (< 5.0163) would lead to low risk. 2. higher Biosensor Stress (> 6.98498) and higher Self-Reported Stress (> 7.05223) would lead to high risk if protective predictors are not present. We also find out that the protective predictor is Sleep Quality: for students experiencing Moderate or Higher Stress Levels (Biosensor Stress and/or Self-Reported Stress), maintaining Good or Moderate Sleep Quality would reduce their Health Risk Levels to Moderate even when Stress Levels are high. We then find out that the conditional predictor is Physical Activity: despite being typically protective, High Physical Activity amplifies health risk level under conditions of High Stress Levels and Good or Moderate Sleep Quality. This finding further proved the GLM result where high physical activity could result in an elevated health risk level.

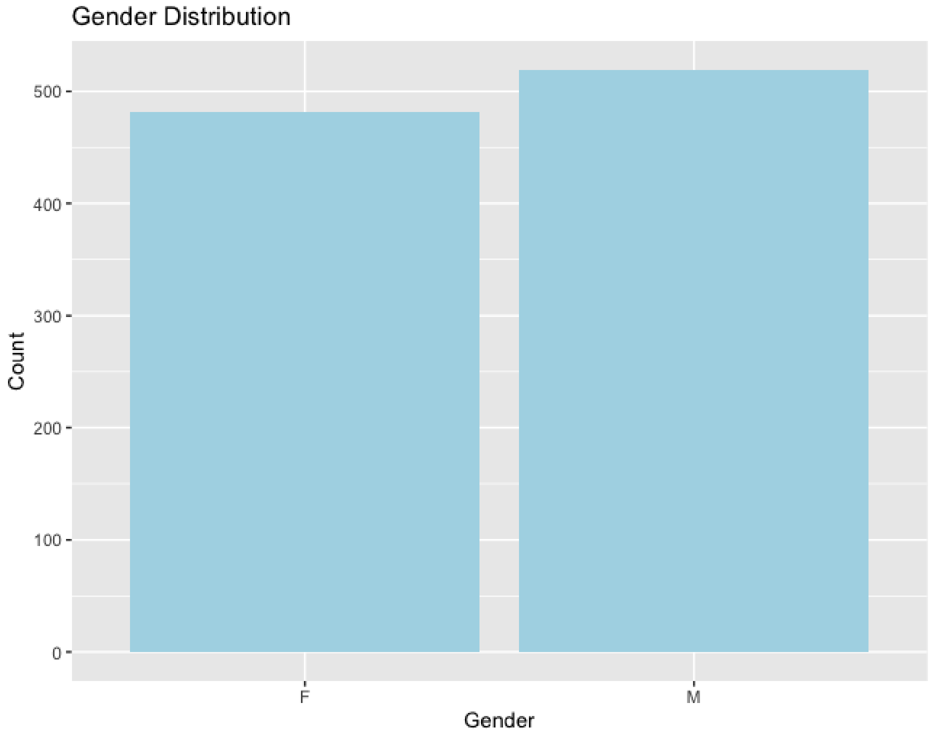
Therefore, the Decision Tree Model result also concludes that to achieve lower levels of health risks, undergraduate students can focus on maintaining moderate levels of physical activity, lower stress levels, and improving sleep quality. Implementing stress management programs on campus, such as mindfulness workshops, counseling sessions, or peer support groups, can help students effectively cope with academic and personal pressures (Journalists Resource, n.d.). Additionally, establishing systems for early detection and intervention for high stress, such as regular stress assessments or mental health screenings, can identify at-risk individuals before their health deteriorates (International Journal of Mental Health Systems, 2020). Lastly, offering lifestyle education initiatives to promote healthy habits—like balanced physical activity, sleep hygiene, and nutrition—can empower students to make informed choices that support their overall well-being (Desimone, 2013).

6. Appendices

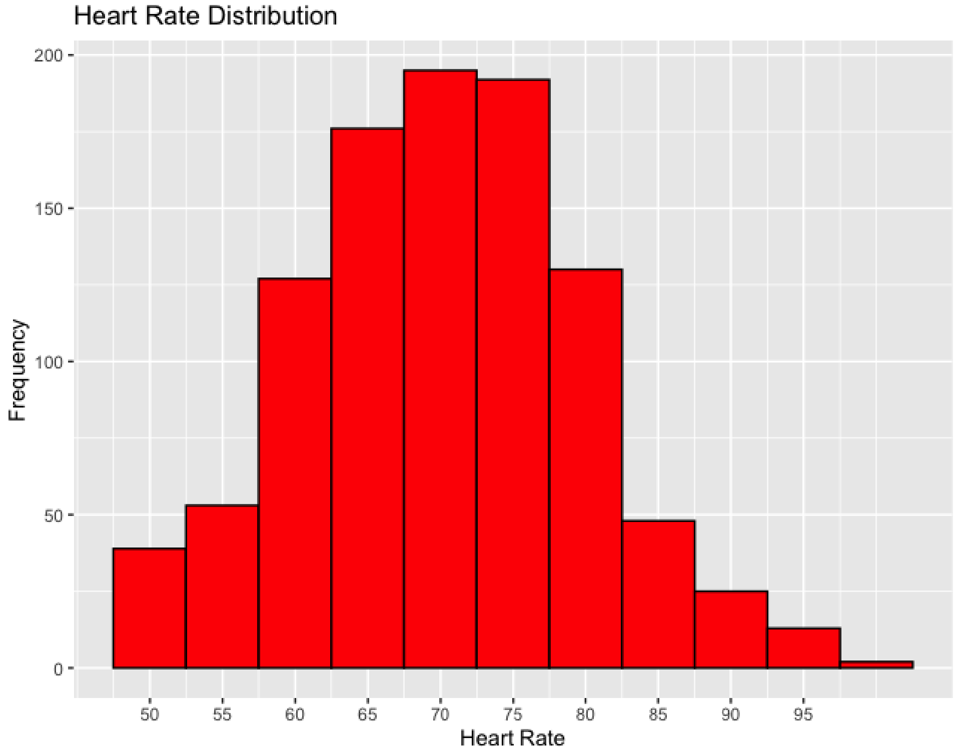
Appendix 1 Age Distribution



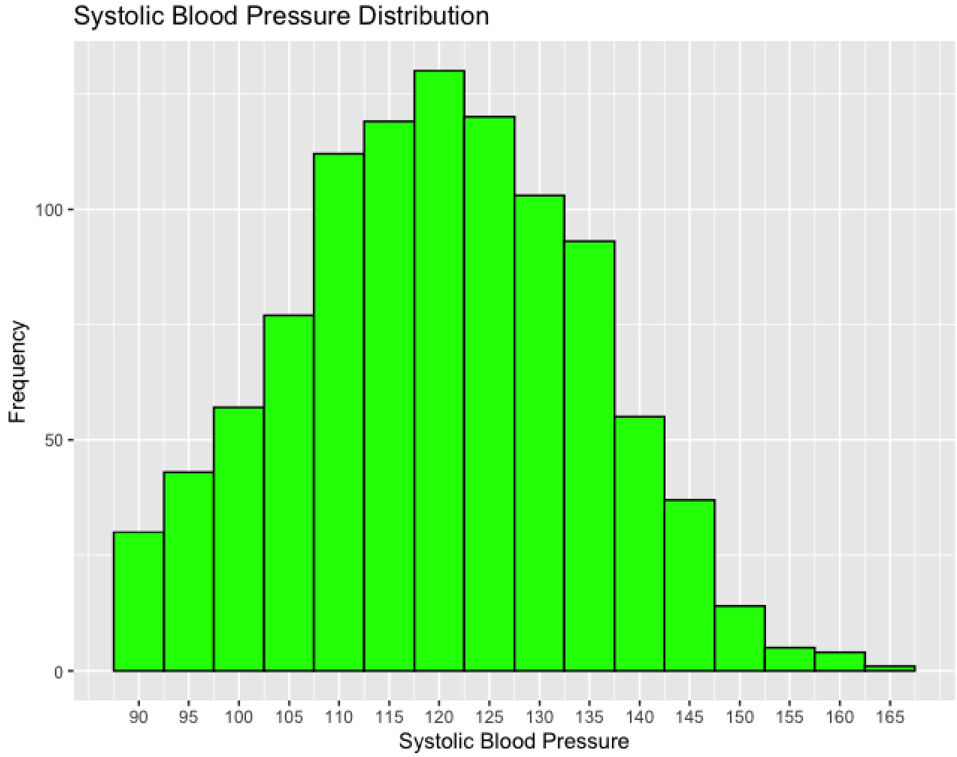
Appendix 2 Gender Distribution



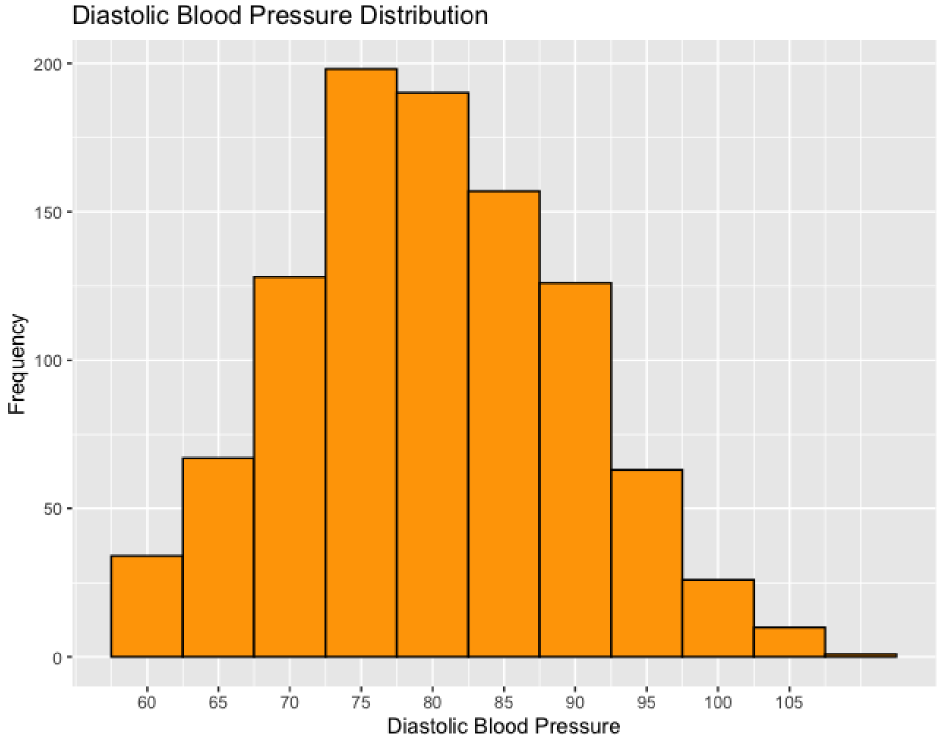
Appendix 3 Heart Rate Distribution



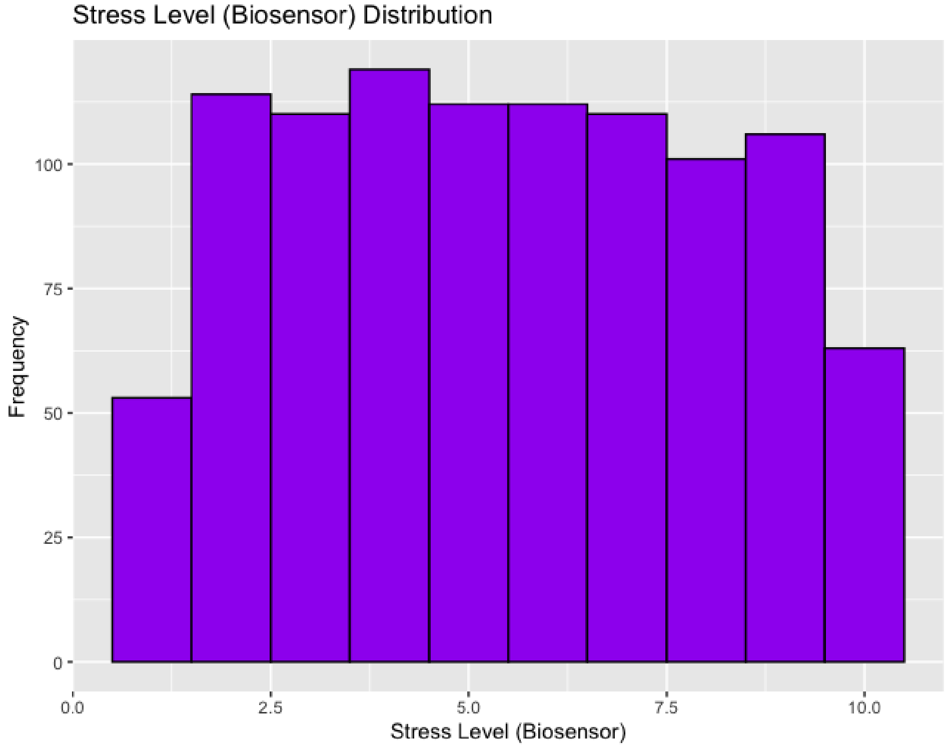
Appendix 4 Blood Pressure (Systolic) Distribution



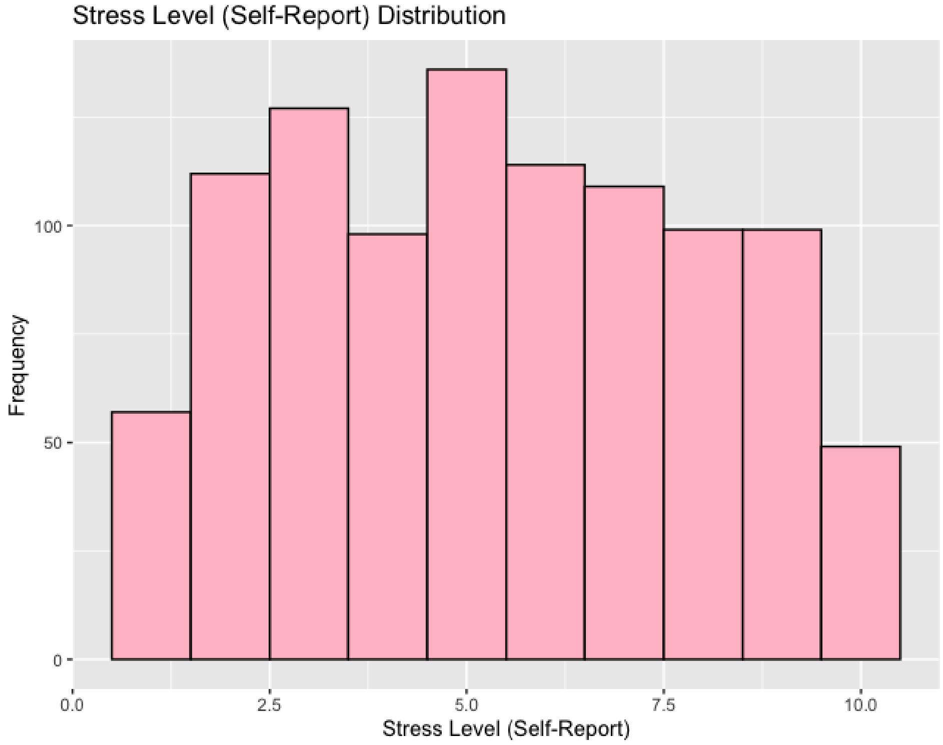
Appendix 5 Blood Pressure (Diastolic) Distribution



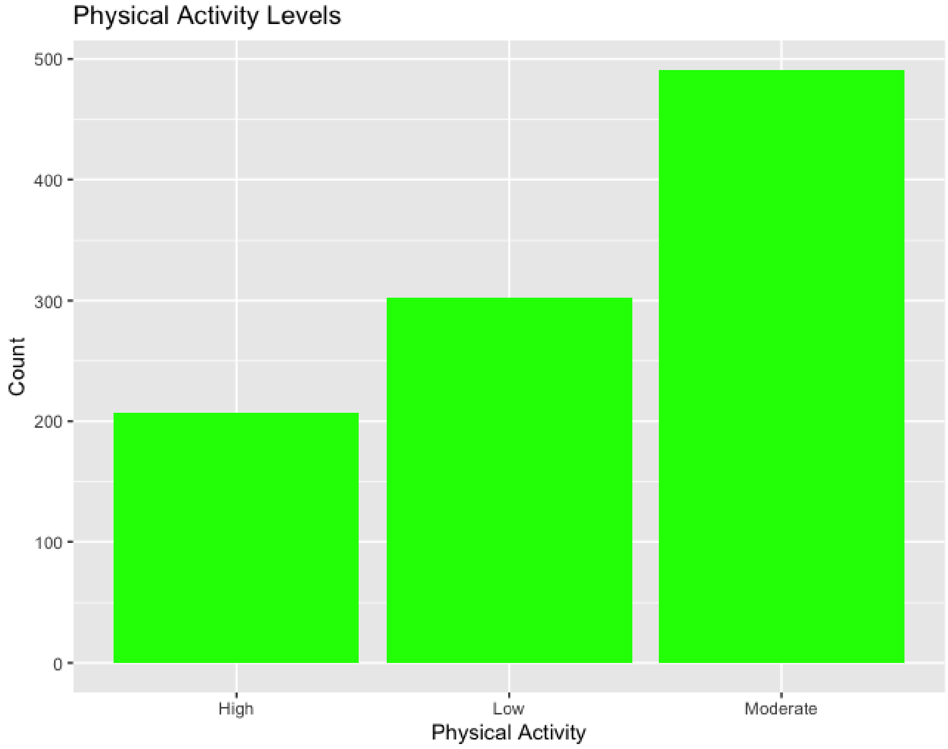
Appendix 6 Stress Level (Biosensor) Distribution



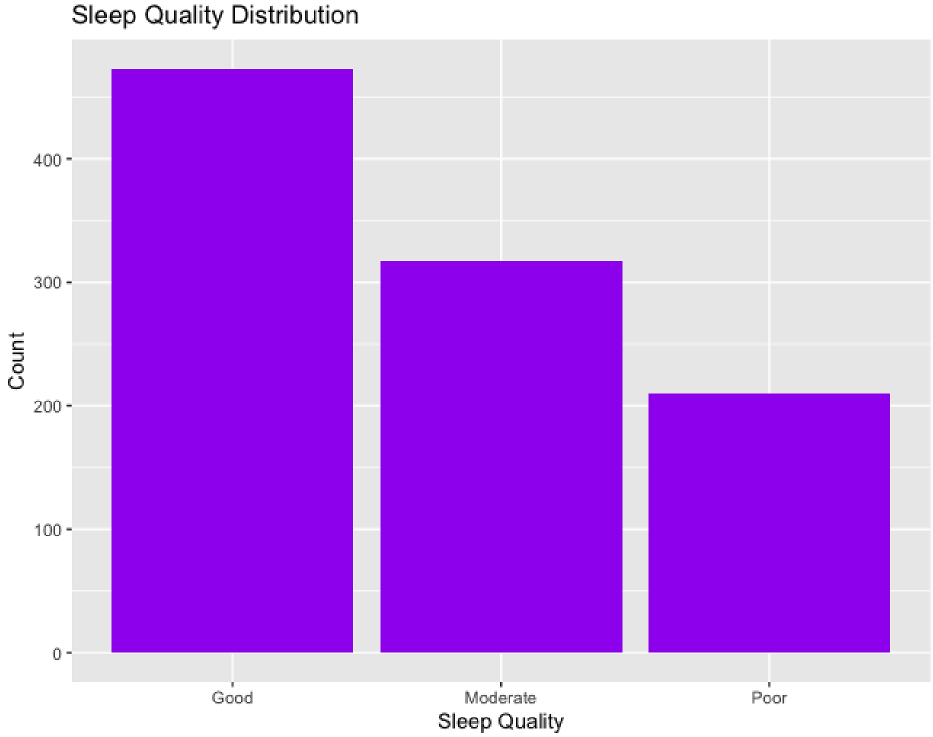
Appendix 7 Stress Level (Self-Report) Distribution



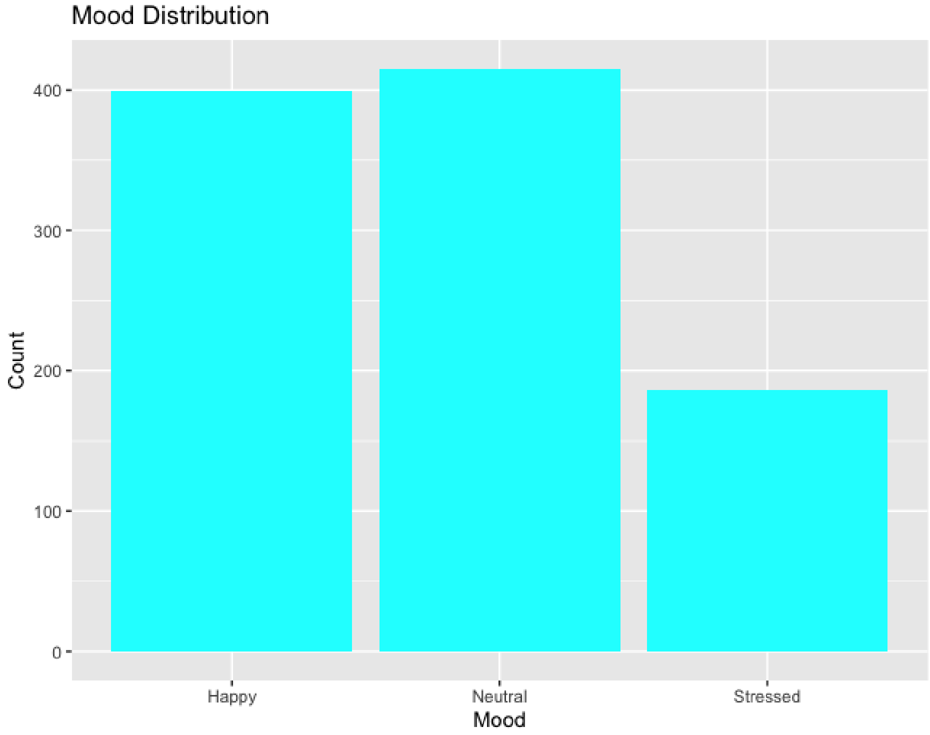
Appendix 8 Physical Activity Levels



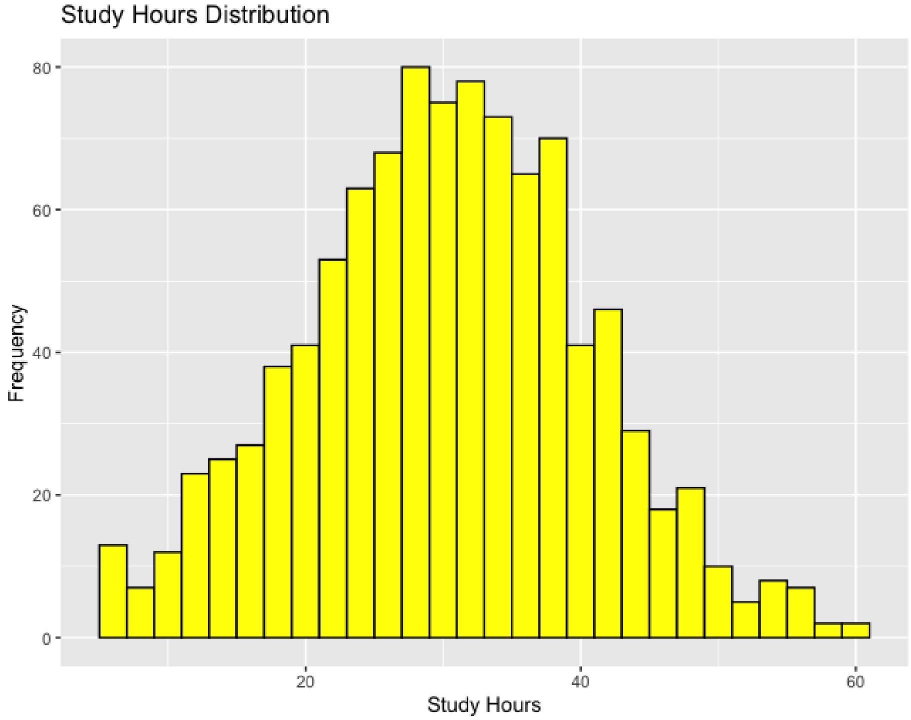
Appendix 9 Sleep Quality Distribution



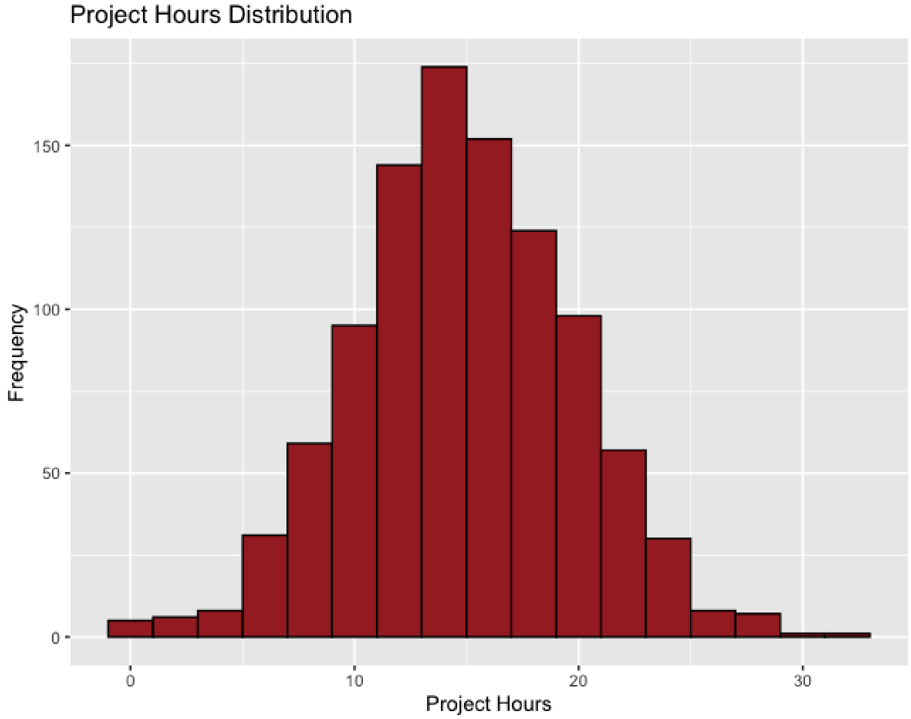
Appendix 10 Mood Distribution



Appendix 11 Study Hours Distribution



Appendix 12 Project Hours Distribution



7. References

American Psychological Association. (2022). *Mental health on campus: Experts discuss trends, obstacles, and opportunities for addressing college students' mental health.* Retrieved from https://www.apa.org/monitor/2022/10/mental-health-campus-care

Carter, B., Carter, H., & Du, H. (2020). Impact of sedentary behavior on university students: A systematic review. *Prevention Science.* Retrieved from https://link.springer.com/article/10.1007/s11121-020-01093-8

Desimone, M. (2013). A study on the effectiveness of a stress management program for college students. Retrieved from https://www.researchgate.net/profile/Martin-Desimone/publication/235419917

International Journal of Mental Health Systems. (2020). Strategies for early intervention in mental health care. Retrieved from https://ijmhs.biomedcentral.com/articles/10.1186/s13033-020-00356-9

Junn, N. (n.d.). *University students mental health.* Kaggle. Retrieved from https://www.kaggle.com/datasets/junnn0126/university-students-mental-health?resource=download

Journalists Resource. (n.d.). Stress and mental health among college students: Research roundup. Retrieved from https://journalistsresource.org/education/college-student-mental-health-research-interventions/

Seltzer, R. (2023). Stress in college students: What to know. *U.S. News & World Report.* Retrieved from https://www.usnews.com/education/best-colleges/articles/stress-in-college-students-what-to-know

Schnohr, P., O'Keefe, J. H., Lavie, C. J., Holtermann, A., Lange, P., Jensen, G. B., & Marott, J. L. (2021). U-shaped association between duration of sports activities and mortality: Copenhagen City Heart Study. *Mayo Clinic Proceedings, 96*(12), 3012–3020. https://doi.org/10.1016/j.mayocp.2021.05.028

University of South Carolina. (2020). Sedentary behaviors in university students. Retrieved from https://sc.edu/about/offices_and_divisions/research/news_and_pubs/caravel/archive/2020_fall/2020_sedentarybehaviors.php