Dive into Student Stress Factors

Lu Wei, Abhishek Kalugade, Tony Tom, Ruhin Rakeshkumar Patel

December 7, 2023

Abstract

Our comprehensive study explores the multifaceted nature of student stress, delving into the complex interplay between various factors like social support, academic performance, and self-esteem and their impact on stress levels. Utilizing the "Student Stress Factors: A Comprehensive Analysis" dataset, we employ multiple linear regression and Analysis of Variance (ANOVA) to dissect these relationships. The investigation is deepened by examining the effects of Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) data simulations on the statistical analyses. Our findings reveal significant relationships between academic performance, study load, teacher-student relationships, future career concerns, and student stress levels. Interestingly, a positive teacher-student relationship correlates with lower stress levels. The ANOVA results further underscore the pivotal role of social support in enhancing self-esteem among students. The study goes beyond mere statistical analysis, emphasizing the necessity of understanding the unique challenges students face. The results advocate for targeted interventions to boost student well-being and highlight the importance of an integrated approach to address student stress. Our research offers invaluable insights for educators, policymakers, and mental health professionals, aiming to foster a healthier educational environment.

1 Introduction

Stress is a silent adversary in the dynamic landscape of academia, where the pursuit of knowledge intersects with the daily lives of students. Stress is an undeniable force that demands exploration and understanding as a universal challenge that transcends borders and impacts the educational experiences of students worldwide[3]. This project embarks on an enthralling journey into the realms of student stress, intertwining threads of curiosity, inquiry, and methodology to unravel the intricate tapestry of factors that contribute to this pervasive phenomenon[6].

1.1 Presenting the Analytical Tools

Among the many analytical options available to us—multiple linear regression, analysis of multi-factor experiments, resampling methods, mixed effects models, generalized linear models, and the venerable Analysis of Variance (ANOVA)—the Multiple Linear Regression (MLP) model and ANOVA stand out[18]. MLP, lauded for its versatility and popularity in data science, serves as the lens through which we examine the complex relationships that govern student stress and academic performance. ANOVA, with its ability to assess group means and uncover nuances, adds to our toolkit, improving our ability to detect significant differences. citepplonsky2017multiple.

In an abundance of options, MLP and ANOVA serve not only as statistical tools but also as vehicles that guide us toward nuanced insights and a deeper understanding of the complexities at work. The choice to employ these methods is not random; it is a deliberate decision to leverage analytical prowess in pursuit of a comprehensive understanding.

1.2 Handling the Missing Data Dilemma

As we delve deeper into our study, the spotlight shifts to missing data, a formidable challenge that casts doubt on the reliability of research findings[14]. Recognizing the importance of careful data handling, we examine the complexities of missing data through two lenses: Missing Completely at Random (MCAR) and Missing Not at Random (MNAR)[22]. This meticulous approach is more than a procedural step; it is a commitment to methodological rigor, acknowledging that the nature of missingness has an impact on the outcome of our analyses.

1.3 Scientific Questions That Spark Inquiry

Scientific questions that transcend the mundane are embedded in our exploration, elevating our inquiry to the heights of profound curiosity. citepharris2020link. We don't just want to confirm the obvious; we want to find non-trivial truths that resonate with the core of academic inquiry. The hypothesis that 'social support' and 'self-esteem' have a significant relationship goes beyond the obvious, delving into the complex interplay of psychological constructs. It investigates not only the existence of a link but also the nature of its significance, inviting us to explore the complexities of human connections and self-perception.
Similarly, the hypothesis that 'academic performance' and 'stress level' have a strong relationship delves deep into academic experiences. However, it is more than just confirmation of intuition; it invites us to explore the nuances of this relationship. What is the definition of 'strong'? How does academic excellence interact with the intricate web of stress? These questions lead us to answers that have the potential to change how we perceive and address student well-being[3].

1.4 Significance Exceeding Academic Research

This project is more than an academic endeavor; it is a search for solutions to a pressing problem that reverberates in lecture halls and study spaces[13]. We extend an invitation to join us on this journey of discovery and understanding as we navigate the uncharted territories of student stress. Beyond the statistical analyses and methodological complexities lies a deeper goal: to contribute to students' well-being and lay the groundwork for targeted interventions. Welcome to the crossroads of inquiry and impact, where the pursuit of knowledge collides with the imperative of promoting a healthier educational environment.

2 Data Description

2.1 Introduction

The dataset titled "Student Stress Factors: A Comprehensive Analysis" is available on Kaggle. This dataset provides an in-depth understanding of the underlying causes and consequences of student stress. It is a wealth of information that can be used to better understand the different variables that lead to student stress and how they interact with one another.

2.2 Data Overview

The dataset is an extensive collection of data points representing various factors that contribute to student stress. The dataset contains approximately 20 features that have the greatest influence on a student's stress. The features are chosen scientifically, taking into account five major factors: psychological, physiological, social, environmental, and academic. Each of these factors includes a variety of variables that provide a comprehensive picture of what students are experiencing and the stressors they face.

2.3 Variable Descriptions

The variables in the dataset are as follows:

Psychological Factors: 'anxiety_level', 'self_esteem', 'mental_health_history', 'depression'
Physiological Factors: 'headache', 'blood_pressure', 'sleep_quality', 'breathing_problem'
Environmental Factors: 'noise_level', 'living_conditions', 'safety', 'basic_needs'
Academic Factors: 'academic_performance', 'study_load', 'teacher_student_relationship', 'future_career_concerns'
Social Factors: 'social_support', 'peer_pressure', 'extracurricular_activities', 'bullying'

Each one of these variables is a potential source of stress for students. They provide a comprehensive picture of the student's surroundings, physical and mental health, academic scenario, and social interactions. We can gain a better understanding of the factors that lead to student stress and the way they interact with one another by analyzing these variables.

2.4 Data Collection and Preprocessing

The data was obtained from Kaggle. The dataset provider preprocessed the data to ensure it was clean and prepared for analysis. Cleaning the data, handling missing values, and transforming variables as needed were most likely among the preprocessing steps. This ensures that the data is in an analysis-ready format and that any mistakes or discrepancies have been addressed.

2.5 Hypotheses of Interest

We can form the following hypotheses based on the dataset:

Hypothesis 1: 'Social support' and 'self-esteem' have a significant relationship. This hypothesis seeks to determine whether there is a significant difference in self-esteem levels among people who receive various kinds of social support. The null hypothesis (H0) states that there is no significant difference in self-esteem levels among people who have different levels of social support. The alternative hypothesis (H1), on the other hand, suggests that there is a significant difference in self-esteem levels among these individuals. If this hypothesis is proven correct, it may imply that improving social support mechanisms for students may increase their self-esteem. This has the potential to have far-reaching consequences for mental health interventions and support systems within educational institutions. It emphasizes the significance of a positive social environment in fostering positive self-esteem in students. For instance, in our dataset, a student who has a social support score of 1 (on a scale of 0 to 5) has a self-esteem score of 29 out of 30, while a student with a score of 3 has a self-esteem score of 15. This suggests a potential relationship between these variables that is not immediately intuitive and requires rigorous investigation.

Hypothesis 2: The variables 'academic performance', 'study load', 'teacher-student relationship', and 'future career concerns' significantly impact stress levels. The null hypothesis (H0) posits that these variables do not significantly impact stress levels. Conversely, the alternative hypothesis (H1) suggests that at least one of these variables significantly impacts stress levels. If this hypothesis is proven correct, it could imply that addressing these factors could potentially help manage stress levels among students. The correlation between these variables is not immediately observable but is revealed through calculation, indicating a complex relationship that goes beyond surface-level observations. This correlation is quantified by a coefficient that encapsulates the strength and direction of the relationship between these variables. This coefficient, which is not intuitively guessable and requires rigorous computation, underscores the intricate nature of

the relationships being studied. Thus, the hypothesis probes deeper into the data, seeking to unravel insights that are not immediately apparent, thereby contributing to a more comprehensive understanding of the phenomena under study.

In essence, these hypotheses seek to uncover deeper insights and contribute to a more comprehensive understanding of the phenomena being studied. They have the potential to inform effective strategies and interventions, thereby making a significant contribution to the field. These are not mere confirmations of intuitive assumptions, but rather, they beckon us to discern the nuances of these relationships and their implications.

2.6 Conclusion

Lastly, the dataset "Student Stress Factors: A Comprehensive Analysis" is a useful resource for anyone with an interest in understanding and addressing student stress. Its comprehensive nature enables an in-depth assessment of stress factors, laying the groundwork for data-driven decision-making in stress management interventions. We can gain valuable insights into the factors that contribute to student stress by analyzing this dataset, and we can develop effective solutions to mitigate stress and improve student well-being.

3 Methods

3.1 Multiple Linear Regression

3.1.1 Introduction to Multiple Regression

Multiple regression is a fundamental statistical technique used for modeling and analyzing the relationship between a dependent variable and two or more independent variables. Its primary objective is to determine the extent to which the independent variables predict or explain the variance in the dependent variable[20]. This method extends beyond the scope of simple linear regression, which considers only one independent variable, allowing for a more comprehensive analysis of complex real-world phenomena.

The use of multiple regression spans various disciplines, from economics, where it might predict factors influencing market trends, to medical sciences, where it can assess the impact of multiple symptoms on a particular health outcome. In the field of social sciences, it helps in understanding the interplay of various socio-economic factors.

Key to the utility of multiple regression is its ability to isolate the effect of each independent variable while controlling for the influence of others. This makes it an invaluable tool in observational studies where controlled experiments may not be feasible. As such, multiple regression analysis has become a cornerstone in statistical modeling, offering nuanced insights into the relationships between multiple interacting variables.

3.1.2 Model Formulation

The multiple regression model is formulated based on the linear relationship between the dependent variable and multiple independent variables. The general form of the model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

where:

- Y represents the dependent variable, the outcome or response that the model aims to predict or explain.

- $X_1, X_2, \ldots, X_k$ are the independent variables (predictors) that are hypothesized to influence the dependent variable.
- β_0 is the intercept term, indicating the value of Y when all the independent variables are zero.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients associated with each independent variable. Each coefficient quantifies the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.
- ϵ is the error term, representing the difference between the observed values and the values predicted by the model. It accounts for the variability in Y that cannot be explained by the independent variables.

The coefficients (β) are estimated from the data, and their values are crucial for understanding the impact of each predictor on the dependent variable. The model's effectiveness hinges on these estimates, which, ideally, are determined using methods such as least squares estimation[24]. Understanding the significance and influence of each variable leads to better insights and interpretations of the underlying processes governing the data.

3.1.3 Assumptions of the Model

For a multiple regression model to provide reliable predictions and inferences, several key assumptions must be met:
Linearity: The relationship between each independent variable and the dependent variable is assumed to be linear. This can be formally stated as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

where the relationship between the dependent variable Y and each independent variable X_i is linear[25].
Independence: Observations are assumed to be independent of each other. This is crucial in time series data, where this assumption is often violated due to sequential measurements.
Homoscedasticity: The variance of error terms (residuals) is constant across all levels of the independent variables. In other words, the spread of residuals should remain consistent across the prediction line, without forming patterns[26].
Normal Distribution of Errors: The error terms ϵ are assumed to be normally distributed. This assumption is important for hypothesis testing and creating confidence intervals. While the normality assumption is not critical for the estimation of coefficients, it is necessary for measuring their standard errors and t-statistics[15].
Violation of these assumptions can lead to biased or misleading results. Thus, it is crucial to perform diagnostic tests to check these assumptions before drawing conclusions from a multiple regression model.

3.1.4 Estimation Methods

The estimation of parameters in a multiple regression model, specifically the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$, is primarily accomplished through the method of least squares[17]. This method aims to minimize the sum of the squared differences between the observed values and the values predicted by the model. Mathematically, the least squares criterion is expressed as:

$$\min \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}))^2$$

where:

- Y_i is the observed value of the dependent variable for the i-th observation.
- $X_{1i}, \ldots, X_{ki}$ are the values of the independent variables for the i-th observation.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients to be estimated.

This optimization problem is solved to find the values of $\beta_0, \beta_1, \ldots, \beta_k$ that minimize the sum of the squared residuals (differences between observed and predicted values). The resulting coefficients provide the best linear unbiased estimate (BLUE) of the relationship between the dependent and independent variables under the Gauss-Markov theorem, assuming all other assumptions of the multiple regression model are met[2].
Additionally, the method of least squares can be extended to include regularization techniques such as Ridge or Lasso regression, especially in scenarios where overfitting is a concern or when dealing with a large number of independent variables[8].

3.1.5 Model Evaluation and Diagnostics

Evaluating the performance of a multiple regression model involves several statistical measures and diagnostic techniques:
R-Squared (Coefficient of Determination): This statistic measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where $\hat{Y}_i$ is the predicted value, Y_i is the actual value, and $\bar{Y}$ is the mean of Y. Higher R-squared values indicate a model that better fits the data.
Adjusted R-Squared: This adjusts the R-squared value based on the number of predictors in the model and the number of observations. It is particularly useful in multiple regression to penalize for adding predictors that do not improve the model. The adjusted R-squared is calculated as:

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is the number of observations and p is the number of predictors.
Residual Analysis: Investigating the residuals (errors) can provide insights into the adequacy of the model. Residual plots are used to detect non-linearity, unequal error variances, and outliers. Ideally, residuals should be randomly scattered around zero, indicating that the model's assumptions are satisfied.
Other Diagnostic Measures: Besides the above, diagnostic measures like the F-test for overall significance, the t-test for individual coefficients, and the examination of variance inflation factor (VIF) for multicollinearity are also integral to model evaluation.
These evaluation metrics and diagnostic tools are essential for ensuring the validity and reliability of a multiple regression model, guiding the researcher in model refinement and interpretation.

3.1.6 Interpreting the Results

Interpreting the results of a multiple regression model primarily involves understanding the coefficients estimated for each independent variable. The coefficients provide insights into the nature and strength of the relationship between each predictor and the dependent variable.
For a multiple regression model given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

the interpretation of coefficients is as follows:

- β_0 (Intercept): Represents the expected value of Y when all the independent variables $X_1, X_2, \ldots, X_k$ are zero. It is the baseline level of Y.
- $\beta_1, \beta_2, \ldots, \beta_k$ (Slope Coefficients): Each coefficient β_i represents the expected change in Y for a one-unit increase in X_i, keeping all other independent variables constant. A positive coefficient indicates a direct relationship, while a negative coefficient indicates an inverse relationship.

It is important to note that these coefficients only imply correlation, not causation. Additionally, the significance of each coefficient should be evaluated, typically through t-tests, to determine if the observed relationships are statistically significant and not due to random chance[4].
The magnitude and sign of the coefficients, along with their statistical significance, guide the interpretation of how each factor influences the outcome variable. In the context of the model, it is also essential to consider interaction terms (if included), as they represent how the effect of one variable may depend on the level of another variable.

3.1.7 Applications of Multiple Regression

Multiple regression analysis is widely used in various fields due to its versatility and ability to handle complex relationships between multiple variables. Some notable applications include:
Economics: In economics, multiple regression is used to analyze factors affecting economic indicators like GDP growth, inflation rates, and employment levels. For instance, a model might examine how interest rates, consumer confidence, and global market trends simultaneously influence national economic performance.
Social Sciences: Researchers in social sciences use multiple regression to explore relationships between social variables. This could include studying how demographic factors, education levels, and socio-economic status together impact health outcomes or political preferences.
Natural Sciences: In fields like environmental science, multiple regression helps in understanding the interaction of various ecological factors. For example, a study might investigate how temperature, precipitation, and soil quality collectively affect crop yield.
Health Sciences: Multiple regression is used in medical research to explore how different clinical variables contribute to health outcomes. This could include examining the impact of lifestyle factors, genetic predisposition, and environmental exposures on the risk of developing a certain disease.
Marketing: In marketing, it's used to assess how various factors like advertising spend, market demographics, and product features influence consumer behavior and sales figures.
These examples highlight the flexibility of multiple regression analysis in testing hypotheses about complex phenomena and its pivotal role in data-driven decision-making across diverse disciplines.

3.2 Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a collection of statistical models and their associated procedures that compare means among groups. The central technique is to decompose observed aggregate variability into different components, with the goal of testing hypotheses about the means of predefined groups[9].

3.2.1 Conceptual Overview

Analysis of Variance (ANOVA) is a hypothesis-testing technique used to determine if there exist statistically significant differences between the means of three or more independent groups. At

its core, ANOVA examines the ratio of the variability between groups to the variability within groups.
The model for one-way ANOVA with one independent variable is given by:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where:

- Y_{ij} is the jth observation from the ith group,
- μ is the overall mean,
- τ_i is the effect of the ith group,
- ϵ_{ij} is the random error present in the jth observation from the ith group.

ANOVA tests the null hypothesis, denoted as H_0, which states that all group means are equal (i.e., $\tau_1 = \tau_2 = ... = \tau_i = 0$). The alternative hypothesis, denoted as H_1, asserts that at least one group mean is different. The F-statistic, calculated as:

$$F = \frac{\text{Mean Square Between Groups (MSB)}}{\text{Mean Square Within Groups (MSW)}}$$

is used to assess this hypothesis. Here, the Mean Square Between Groups (MSB) is the sum of squares between groups divided by the degrees of freedom between groups, while the Mean Square Within Groups (MSW) is the sum of squares within groups divided by the degrees of freedom within groups.
A significant F-statistic, typically determined by a p-value less than a pre-specified threshold such as 0.05, indicates that there is at least one group mean that is statistically different from the others, leading us to reject the null hypothesis in favor of the alternative.
ANOVA allows us to move beyond merely noting differences and provides a framework to quantify and test the significance of the differences across group means in a controlled manner.

3.2.2 Assumptions of ANOVA

The validity of the ANOVA test relies on several critical assumptions. Violating these assumptions can result in misleading statistical inferences.

- **Independence of Observations**: This assumption asserts that the observations within each group and between groups are not correlated. The independence assumption is often satisfied by using a randomized experimental design. In the context of data analysis, it can be expressed as:

  $$Cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0 \text{ for all } i \neq i' \text{ and } j \neq j'$$

  where ϵ_{ij} and $\epsilon_{i'j'}$ are error terms for different observations.

- **Normality**: ANOVA assumes that the residuals ϵ_{ij} for each group are normally distributed. This can be formally stated as:

  $$\epsilon_{ij} \sim N(0, \sigma^2)$$

  for all groups i and all observations j within a group, where σ^2 is the common variance of the error terms.

- **Homogeneity of Variance (Homoscedasticity)**: Also known as the assumption of equal variances, it states that the variance within each of the groups is the same across all groups. This can be tested using Levene's test or Hartley's test and is often checked via a residuals plot. The assumption can be written as:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

where σ_i^2 is the variance within the i-th group.

Each of these assumptions can be tested using diagnostic plots or formal statistical tests. For instance, the normality assumption can be assessed using a Q-Q plot or the Shapiro-Wilk test, while the homogeneity of variances can be checked using a residual versus fits plot or an F-test for variances[1]. Ensuring these assumptions are met is essential for the ANOVA model to yield reliable results.

3.2.3 The F-test

The F-test is the cornerstone of ANOVA, used to determine whether the group means in the sample are drawn from populations with equal means. The test statistic, known as the F-statistic, is calculated as follows:

$$F = \frac{MSB}{MSW}$$

where:

- MSB (Mean Square Between) represents the variance between the group means and is calculated by dividing the sum of squares between (SSB) by the between-group degrees of freedom (dfB). It reflects the variation of group means around the overall mean.

$$MSB = \frac{SSB}{dfB}$$

$$SSB = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$

Here, n_i is the sample size of the i-th group, $\bar{Y}_i$ is the mean of the i-th group, and $\bar{Y}$ is the overall mean.

- MSW (Mean Square Within) represents the variance within the groups and is calculated by dividing the sum of squares within (SSW) by the within-group degrees of freedom (dfW). It captures the variation due to random error or individual differences within each group.

$$MSW = \frac{SSW}{dfW}$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Here, Y_{ij} is the j-th observation in the i-th group.

The degrees of freedom are calculated as $dfB = k - 1$ and $dfW = N - k$, where k is the number of groups and N is the total number of observations[12].
The F-statistic follows the F-distribution under the null hypothesis. A significant F-value, typically determined by comparing the calculated F-statistic to a critical value from the F-distribution or by using a p-value, suggests the group means are not all equal in the population[11]. This leads us to reject the null hypothesis in favor of the alternative hypothesis, which posits that at least one group mean is different.
The F-test effectively partitions the total variation in the data into variation between groups and variation within groups. By comparing these variances, ANOVA gauges whether the means differ more than would be expected by chance.

3.2.4 Post Hoc Analysis

Upon rejecting the null hypothesis in an ANOVA, it becomes necessary to conduct post hoc analyses to ascertain which specific group means are significantly different from each other. Post hoc tests are multiple comparison procedures that control for the Type I error rate across these comparisons.
Tukey's Honest Significant Difference (HSD): This test is commonly used when all groups have the same sample size and is designed to compare every mean with every other mean. The Tukey HSD test calculates a single critical value, q, that all mean differences must exceed to be considered significantly different. The test statistic is given by:

$$q_{ij} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{MSW/n}}$$

where $\bar{Y}_i$ and $\bar{Y}_j$ are the sample means for groups i and j, respectively, MSW is the mean square within groups from ANOVA, and n is the number of observations per group.
Bonferroni Correction: This method involves adjusting the significance level $alpha) by the number of comparison$ comparisons are made, a pairwise difference is considered significant if the p-value is less than α/c. The Bonferroni correction is a conservative method that reduces the likelihood of Type I errors but increases the chance of Type II errors[19].
Scheffé's Method: This test is more flexible than the Tukey HSD and can be used regardless of whether the group sizes are equal. It is particularly useful when researchers wish to test complex hypotheses about the means. Scheffé's method calculates an F-distribution-based critical value for the differences between means and is given by:

$$F_{ij} = \frac{(n-k)(\bar{Y}_i - \bar{Y}_j)^2}{kMSW}$$

where n is the total number of observations, k is the number of groups, and MSW is the mean square within the ANOVA. The computed F_{ij} is then compared against a critical value from the F-distribution with $k - 1$ and $n - k$ degrees of freedom.
It is important to select a post hoc test that is appropriate for the study design and the specific hypotheses being tested. The choice of post hoc test can impact the findings and conclusions drawn from an ANOVA, so careful consideration should be given to the test's assumptions and properties[10].

3.2.5 Applications of ANOVA

Analysis of Variance (ANOVA) is an incredibly versatile statistical tool that is employed across a vast array of fields and disciplines. Its fundamental purpose is to compare means among multiple groups, making it ideal for a wide range of applications.

Agriculture: In agricultural research, ANOVA is used to compare the yield of different crop varieties under various conditions. Researchers can assess the impact of factors such as fertilizer types, irrigation methods, and genetic modifications on crop production.

$$\text{Yield} = \mu + \text{Fertilizer Type} + \text{Irrigation Method} + \text{Genetic Modification} + \epsilon$$

Psychology: ANOVA is pivotal in psychology for experiments that involve treatment effects. For example, it can compare the efficacy of different therapeutic interventions or the impact of conditions like stress or sleep deprivation on cognitive performance.

$$\text{Cognitive Score} = \mu + \text{Treatment} + \text{Stress Level} + \epsilon$$

Marketing: Marketers utilize ANOVA to understand consumer behavior by comparing the appeal of different product designs, packaging, or advertising strategies. This helps in optimizing product launches and marketing campaigns.

$$\text{Consumer Rating} = \mu + \text{Product Design} + \text{Advertising Strategy} + \epsilon$$

Education: In educational research, ANOVA can be used to evaluate the effectiveness of teaching methods or curricular designs across different schools or classrooms.

$$\text{Student Achievement} = \mu + \text{Teaching Method} + \text{Curriculum Design} + \epsilon$$

Medicine: Medical studies often use ANOVA to compare patient responses to different drug treatments or to analyze the influence of lifestyle factors on health outcomes.

$$\text{Health Outcome} = \mu + \text{Drug Treatment} + \text{Lifestyle Factor} + \epsilon$$

The capacity of ANOVA to dissect the influence of multiple group factors and to handle complex experimental designs makes it a cornerstone of data analysis in scientific research. It provides a formal way to ascertain whether observed differences in sample means are likely to reflect actual differences in population means, thus guiding critical decisions in policy-making, scientific discovery, and business strategy.

Concluding, ANOVA stands as an essential tool in a statistician's arsenal, offering a robust method for discerning and testing differences across group means. Its correct application and interpretation, when aligned with its foundational assumptions, facilitate meaningful inferences that can guide critical decision-making processes in scientific research. ANOVA's strength lies in its ability to systematically break down and analyze the variability inherent in experimental and observational data into components that can be attributed to specific sources, thereby illuminating underlying patterns and relationships.

Moreover, the flexibility of ANOVA extends beyond simple one-way designs to more complex factorial and multivariate layouts, accommodating a diverse range of experimental scenarios. Whether the goal is to evaluate the efficacy of multiple drug treatments in clinical trials, to determine the impact of educational interventions on student outcomes, or to assess the influence of environmental factors on ecological systems, ANOVA provides the framework for a systematic and rigorous approach to hypothesis testing.

In the realm of statistical modeling, where the quantification of uncertainty and the precision of conclusions drawn from data are paramount, ANOVA remains an indispensable technique. Its continued use and development in statistical software ensure its accessibility and applicability to statisticians and researchers across disciplines, reaffirming its position as a cornerstone of quantitative analysis in the pursuit of scientific knowledge.

3.3 Understanding and Addressing Missing Data

Missing data poses significant challenges in statistical analysis, affecting the validity of inferences drawn from the data. This subsection delves into the concept of missing data and discusses various approaches to managing it, with a focus on Mean and Regression Imputation techniques.

3.3.1 Concepts of Missing Data

Missing data can occur for various reasons and can be broadly classified into three categories based on their mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

- **MCAR** refers to the situation where the probability of missingness is the same for all cases, indicating that the missing data is unrelated to any observed or unobserved data.
- **MNAR**, also known as non-ignorable missingness, occurs when the missingness is related to the value of the missing data itself. For example, sicker patients may be less likely to return for follow-up, thus introducing bias into the dataset.

3.3.2 Methods for generating Missing Data

In the study, the above two approaches mentioned were employed to generate missing data: noise-based methods for Missing Not at Random (MNAR) data and random selection for Missing Completely at Random (MCAR) data.

Noise-based Methods for MNAR [23] For simulating MNAR datasets, we utilized noise-based techniques in R. These involve:

1. **Introduction of Noise:** We generated random noise using functions like `runif()` or `rnorm()`, scaled it appropriately, and added it to the observed data.
2. **Thresholding for Missingness:** A threshold was defined to determine missingness based on the noise level. Data points were set to missing if the absolute value of the added noise exceeded this threshold.
3. **Adjusting the Missingness Rate:** The threshold value and the parameters of the noise distribution were varied to achieve the desired proportion of missing data.

Popular R packages such as `mice::addNA`, `missMethods::addNoise`, and `simcausal::simcausal` facilitated this process.
In our study, we initially considered two other methods for simulating MNAR (Missing Not at Random) missing data. However, we ultimately decided against employing these methods for the following reasons:

1. **Arbitrary Probability Selection for MNAR:** [5] The first method, based on the assumption that individuals with lower levels of support were less likely to report self-esteem scores, proposed a 50% probability of missingness for the lowest self-esteem scores. This method, however, was based on an arbitrary selection of probability and was not pursued due to concerns about its methodological soundness.
2. **Probability-Based MNAR Method:** Our second method proposed a probability model where an observation's missingness on X_2 was related to its own (potentially unobserved) value. The model was defined as:

$$\Pr\left(X_{i2} \text{ is missing}\right) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{1}{2}\left(X_{i2} - 20\right)\right]} \tag{1}$$

While this approach initially seemed promising, it was discovered that it would lead to an excessive amount of missing data, approximately 90% of the total dataset, rendering it impractical for our analysis.

These methods were part of our initial comprehensive approach to address the potential impacts of MNAR missing data. The decision to not proceed with these methods was made after realizing the significant impact they could have on the integrity and usability of the dataset, particularly the first second's tendency to introduce an impractically high level of missing data.

Random Methods for MCAR For MCAR data, a simpler approach was adopted:

- We randomly designated 20% of the data points in a given variable as missing, irrespective of their values or the values of other variables in the dataset[16].
- This approach simulates a scenario where the probability of missingness is equal across all observations, reflecting true randomness in the missing data pattern.

Both methods offer distinct advantages and challenges. While the noise-based approach allows for a controlled simulation of MNAR data, controlling the missing data pattern precisely can be challenging. On the other hand, the random method for MCAR data, though simpler to implement, may not capture the complexities of real-world missing data scenarios. The choice of method was driven by the specific research objectives and the nature of the missing data under investigation.

3.3.3 Methods for Handling Missing Data

Dealing with missing data requires robust statistical methods to minimize bias and maintain the integrity of the analysis. Two common imputation techniques are Mean Imputation for MCAR data and Regression Imputation for MNAR data[21].

Mean Imputation [7] Mean Imputation is a simple technique for handling missing data, particularly in MCAR scenarios. It involves replacing each missing value in a variable with the mean of the available cases for that variable. Mathematically, for a variable X with missing values, the imputation can be represented as:

$$X_{\text{imputed}} = \begin{cases} X_i & \text{if } X_i \text{ is observed} \\ \bar{X}_{\text{obs}} & \text{if } X_i \text{ is missing} \end{cases}$$

where $\bar{X}_{\text{obs}}$ is the mean of the observed values of X. While this method preserves the sample mean, it underestimates the variance and covariance, leading to potentially biased estimates in further analyses. Care should be taken to assess the impact of mean imputation on the overall statistical inference, especially in cases where the proportion of missing data is substantial.

Regression Imputation [27] Regression Imputation, more suitable for MNAR data, uses a regression model to estimate missing values based on other variables. This method assumes that the relationship between variables can be captured accurately with a regression model. The imputed value for a missing data point in variable Y is computed based on the observed values of other variables X. The regression equation used is:

$$Y_{\text{imputed}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are the coefficients estimated from a regression model fitted on the observed data, and ϵ is the error term. After fitting the model, missing values of Y are predicted and

replaced with these predictions. While this approach can provide more accurate imputations than mean imputation, especially when the variables are correlated, it requires careful consideration of the regression model's appropriateness and the potential for introducing bias due to model misspecification. Moreover, similar to mean imputation, it can underestimate the variability in the data.
These imputation techniques serve to alleviate the impact of missing data, enabling more accurate and reliable statistical analysis. However, the choice of method must align with the missing data mechanism to ensure the validity of the imputation process.

4 Exploratory Data Analysis

4.1 Introduction

The "Student Stress Factors: A Comprehensive Analysis" dataset's exploratory data analysis (EDA) section is critical to gaining insights. This section's goal is to provide a thorough understanding of the data's structure, distributions, and relationships, laying the groundwork for further hypothesis testing.

4.2 Overview of Dataset

The dataset "Student Stress Factors: A Comprehensive Analysis" is available on Kaggle. This dataset provides a comprehensive understanding of the underlying causes and consequences of student stress. It has approximately 20 features that have the greatest impact on a student's stress. The features are scientifically chosen by taking into account five major factors: psychological, physiological, social, environmental, and academic factors.

4.3 Descriptive Statistics

The descriptive statistics provide an in-depth analysis of the "Student Stress Factors: A Comprehensive Analysis" dataset. The following are the key findings:
Number of Students: The dataset contains 1100 students, which provides a large sample size for the analysis.
Average Anxiety Level: The average anxiety level among students is around 11.06. This provides insight into the student's overall anxiety level.
Prevalence of Mental Health Issues: In the dataset, approximately 49.27% of the students have a history of mental health issues. This demonstrates the prevalence of mental health issues among students.
Distribution of Variables: The dataset's variable distribution varies significantly. The anxiety_level variable, for example, has a range of 0 to 21, with each level having a different number of students. The self_esteem variable has a range of 0 to 30, with different counts for each level. The variable mental_health_history is binary, with 558 students reporting no history of mental health issues and 542 reporting a history of mental health issues. Other variables in the dataset have similar distributional variability.
Missing Values: The dataset contains no missing values. This indicates that the dataset is complete and does not require imputation or missing data handling.

4.4 Univariate Analysis

The univariate analysis examines a single variable in the dataset in depth. The following are the key findings:

Anxiety Level: The students' anxiety levels range from 0 to 21, with an average (mean)

of approximately 11.06. The median anxiety level is 11, indicating that half of the students have anxiety levels below 11 and half have anxiety levels above 11.
Sleep Quality: Sleep quality ratings range from 0 to 5, with an average of 2.66. The average score for sleep quality is 2.5.
Academic Performance: Academic performance scores range from 0 to 5, with an average of 2.773. The median score for academic performance is 2.
Social Support: Scores for social support range from 0 to 3, with an average of about 1.882. The median level of social support is 2.
Self Esteem: The self-esteem scale runs from 0 to 30, with an average of 17.78. The median level of self-esteem is 19.
Mental Health History: In the dataset, approximately 49.27% of the students have a history of mental health issues.
Depression: Depression levels range from 0 to 27, with an average of 12.56. The average depressive score is 12.
Study Load: The study load scale runs from 0 to 5, with an average of 2.622. The average study load is 2.
Extracurricular Activities: Scores for extracurricular activities range from 0 to 5, with an average of about 2.767. The median score for extracurricular activities is 2.5.
Bullying: Bullying ratings range from 0 to 5, with an average of 2.617. The average bullying score is three.

4.5 Correlation Analysis

In this section, we'll look at the relationships between the variables in our dataset. We concentrate on two pairs of variables: 'depression' and 'self_esteem', and 'academic_performance' and 'stress_level'.

4.5.1 Correlation between 'depression' and 'self_esteem'

The correlation coefficient between 'depression' and 'self_esteem' was found to be **-0.7**. This indicates a strong negative correlation, implying that as depression levels rise, self-esteem levels fall, and vice versa.
The scatter plot visualization for 'depression' and 'self_esteem' confirmed this relationship, showing a downward trend as depression levels rise.

Figure 1: Correlation between Depression and Self-Esteem

4.5.2 Correlation between 'bullying' and 'self_esteem'

The coefficient of correlation between 'bullying' and 'self_esteem' was found to be -0.64. This shows a strong negative correlation, implying that self-esteem levels fall when bullying levels rise and vice versa.
The scatter plot visualization for 'bullying' and 'self_esteem' confirmed this relationship, demonstrating a downward trend as bullying levels rise.

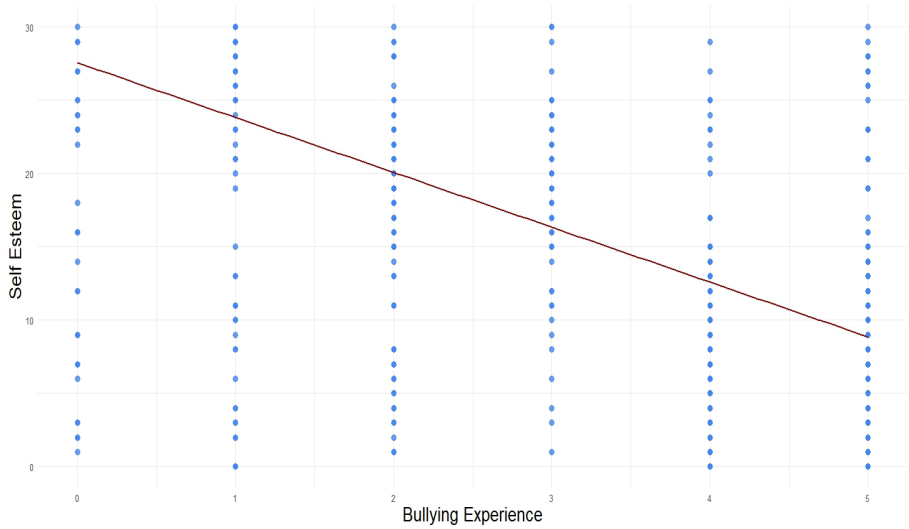

Figure 2: Correlation between Bullying and Self-Esteem

We learned a lot about the relationships between depression, bullying, and self-esteem from these analyses. These findings can help us understand the factors that influence student self-esteem and develop strategies to address them.

4.6 Analysis of Multiple Linear Regression Results

This subsection presents a comprehensive analysis of the multiple linear regressions performed to explore the factors affecting stress levels. The regression model includes academic performance, study load, teacher-student relationship, and future career concerns as independent variables.

4.6.1 Model Summary

The regression model is defined as:

$$\text{stress\_level} = \beta_0 + \beta_1 \cdot \text{AP} + \beta_2 \cdot \text{SL} + \beta_3 \cdot \text{TSR} + \beta_4 \cdot \text{FCC} + \epsilon$$

In this model, AP represents academic performance, SL stands for study load, TSR denotes teacher-student relationship, and FCC refers to future career concerns. The coefficients $\beta_1, \beta_2, \beta_3$, and β_4 correspond to these variables, respectively. The intercept of the model is denoted by β_0, and ϵ represents the error term.

4.6.2 Coefficient Analysis

The estimated coefficients and their statistical significance are as follows:

- **Intercept**: The intercept coefficient ($\beta_0 = 0.90523$) is significantly different from zero, indicating the baseline level of stress when all predictors are at zero.
- **Academic Performance** ($\beta_1 = -0.17322$): The negative coefficient suggests that higher academic performance is associated with lower stress levels, and this effect is statistically significant ($p < 2e\text{-}16$).
- **Study Load** ($\beta_2 = 0.13289$): Indicates that higher study load leads to increased stress, with a significant effect ($p < 2e\text{-}16$).
- **Teacher-Student Relationship** ($\beta_3 = -0.09074$): Shows a negative relationship with stress level, meaning better relationships are associated with lower stress ($p = 1.21e\text{-}09$).
- **Future Career Concerns** ($\beta_4 = 0.17490$): Indicates that concerns about future career increase stress levels significantly ($p < 2e\text{-}16$).

4.6.3 Model Fit and Diagnostics

- **Residual Analysis**: The residuals range from -2.35344 to 2.23312, with a median close to 0, suggesting the model's predictions are generally accurate.
- **Residual Standard Error**: The standard error of 0.4541 on 1095 degrees of freedom indicates the average distance that the observed values fall from the regression line.
- **R-squared Values**: The model explains 69.57% of the variance in stress levels (R-squared = 0.6957), which is quite substantial. The adjusted R-squared (0.6946) accounts for the number of predictors and the sample size.
- **F-statistic**: The F-statistic (625.9) and its associated p-value ($< 2.2e\text{-}16$) suggest that the overall model is statistically significant, meaning that at least one of the predictors is significantly related to the stress level.

4.6.4 Conclusion

The analysis demonstrates that all four variables—academic performance, study load, teacher-student relationship, and future career concerns—have significant impacts on stress levels. The model shows a good fit with a high R-squared value, indicating that these variables are important predictors of stress level. This regression analysis provides valuable insights into the factors contributing to stress, which can be crucial for developing targeted interventions to manage or reduce stress among individuals.

4.7 Analysis of Variance (ANOVA) Results

The Analysis of Variance (ANOVA) was conducted to assess the impact of varying levels of social support on self-esteem scores among participants. The ANOVA results, alongside the Tukey post-hoc analysis, provide insight into the differences between group means.

4.7.1 ANOVA Summary

The ANOVA revealed a significant effect of social support on self-esteem, $F(3, 1096) = 513.9$, $p < .00001$. The large F value suggests that the variation in self-esteem attributable to social support levels is statistically significant and substantial.

- Degrees of Freedom (Df) for social support: 3
- Sum of Squares (Sum Sq) attributed to social support: 51393
- Mean Square (Mean Sq), which is the Sum Sq divided by Df: 17131
- Residuals Df: 1096
- Sum Sq for residuals: 36534
- Mean Sq for residuals: 33

The very small p-value indicates that we can reject the null hypothesis that all group means are equal.

4.7.2 Tukey Post-hoc Test

The Tukey post-hoc test was performed to pinpoint exactly which groups' means differed from each other. The results were as follows:

- The mean self-esteem score for group 1 was significantly lower than group 0 by 5.855 points (95% CI: -7.600 to -4.111, $p < .00001$).
- Group 2's mean score was higher than group 0 by 4.532 points (95% CI: 2.517 to 6.548, $p < .00001$).
- Group 3's mean score was higher than group 0 by 9.359 points (95% CI: 7.630 to 11.088, $p < .00001$).
- Comparing group 2 and 1, the mean score for group 2 was higher by 10.388 points (95% CI: 8.942 to 11.833, $p < .00001$).
- The mean score for group 3 was higher than group 1 by 15.214 points (95% CI: 14.205 to 16.222, $p < .00001$).

- The mean score for group 3 was higher than group 2 by 4.826 points (95% CI: 3.399 to 6.253, $p < .00001$).

The confidence intervals are tight, and the adjusted p-values are very low, which provides strong evidence that not all groups have the same mean self-esteem score.

4.7.3 Interpretation

The analysis indicates that social support has a statistically significant effect on self-esteem. Individuals with the highest level of social support reported higher self-esteem scores compared to those with lower levels. This supports the hypothesis that social support positively correlates with self-esteem.
Given the robustness of the ANOVA results and the clear indications from the Tukey post-hoc test, it can be concluded that social support plays a critical role in influencing self-esteem levels among the study participants. Future research could further explore the causal mechanisms of this relationship and investigate whether interventions to increase social support could effectively enhance self-esteem.

4.8 Investigation on Missing Data

4.8.1 Performing Analysis Again after MCAR Simulation

This subsection presents an in-depth comparative analysis of the multiple linear regression and ANOVA results, focusing on the changes observed after applying MCAR simulation to the dataset.

Multiple Linear Regression Analysis **Original Analysis:** Initially, the regression analysis indicated strong and significant relationships between all predictors (academic performance, study load, teacher-student relationship, and future career concerns) and the stress level. The model showed high predictive accuracy with an R-squared value of 0.6957.
Post-MCAR Analysis: After simulating MCAR data, there were notable changes:

- The coefficients for academic performance and teacher-student relationship showed a decrease, indicating a reduced magnitude of their negative relationship with stress levels.
- The coefficient for study load remained positive, and future career concerns continued to have a strong positive influence on stress levels.
- A significant increase in the residual standard error (from 0.4541 to 0.5114) and a decrease in R-squared (from 0.6957 to 0.532) were observed, indicating a loss in the model's explanatory power and predictive accuracy.

ANOVA (Analysis of Variance) **Original Analysis:** The initial ANOVA revealed a substantial effect of social support on self-esteem, with a very high F value of 940.3.
Post-MCAR Analysis: After the MCAR simulation:

- The F value decreased significantly to 154.9, indicating a reduced effect of social support on self-esteem in the imputed dataset.
- Tukey's post-hoc test revealed a shift in the differences between the groups, suggesting alterations in the group mean relationships due to the imputation.

Comparative Analysis and Conclusions The reanalysis of the data post-MCAR simulation highlights the substantial impact that missing data handling can have on statistical results.

- The decrease in the R-squared value in the regression model post-MCAR suggests that missing data imputation can lead to a loss in the model's ability to explain the variability in the response variable.
- Changes in the coefficients and the increase in residual error post-MCAR indicate that the relationships between predictors and the dependent variable are less pronounced and less accurate after imputation.
- In the ANOVA, the reduced F value and altered group differences post-MCAR signal that missing data handling can significantly affect the detection of group differences and the magnitude of these differences.
- These findings underscore the importance of a thorough understanding of the nature of missing data and careful consideration of the imputation methods used, as they can profoundly influence the conclusions drawn from statistical analyses.
- The results advocate for conducting sensitivity analyses in studies involving missing data to understand the robustness of findings against different methods of handling missing data.

In conclusion, the analysis demonstrates that while the relationships between variables remain generally consistent, the magnitude and precision of these relationships are notably affected by the handling of missing data. This highlights the critical need for careful methodological considerations in statistical modeling, especially in the presence of missing data.

4.8.2 Performing Analysis Again after MNAR Simulation

This subsection provides a detailed comparative analysis of the multiple linear regression and ANOVA results, focusing on the changes observed after implementing the MNAR (Missing Not At Random) simulation to the dataset.

Multiple Linear Regression Analysis **Original Analysis:** The original regression model, considering academic performance, study load, teacher-student relationship, and future career concerns, exhibited a high explanatory power with an R-squared value of 0.6957, indicating a significant relationship between these predictors and stress levels.
Post-MNAR Analysis:

- The regression model post-MNAR simulation, examining the effect of social support on self-esteem, showed a pronounced increase in the coefficient of social support (from 5.81075), suggesting a stronger positive relationship with self-esteem.
- The residual standard error increased from 0.4541 to 6.571, and the R-squared value decreased from 0.6957 to 0.461, indicating a significant drop in the model's predictive accuracy and explanatory power.
- These changes highlight the altered dynamics in the relationship between social support and self-esteem under the MNAR condition.

ANOVA (Analysis of Variance) **Original Analysis:** Initially, ANOVA showed a substantial impact of social support on self-esteem, as evidenced by a high F value of 940.3.
Post-MNAR Analysis:

- Post-MNAR, the F value in the ANOVA analysis showed a remarkable increase to 4704, suggesting an even more pronounced effect of social support on self-esteem in the MNAR-simulated dataset.
- Tukey's post-hoc test indicated significant shifts in the mean differences between groups, reflecting the implications of the MNAR simulation on the group mean relationships.

Comparative Analysis and Conclusions The reanalysis post-MNAR simulation underscores the profound impact that the nature of missing data can have on statistical results:

- The marked decrease in R-squared and increase in residual error in the regression model post-MNAR simulation signal a substantial reduction in the model's ability to explain and predict the stress level accurately.
- The heightened coefficient of social support in the regression model and the increased F value in ANOVA post-MNAR reveal a more intense perceived impact of social support on self-esteem, possibly due to the bias introduced by the MNAR data.
- These observations illustrate how MNAR data can exaggerate the relationships between variables, affecting both the magnitude and accuracy of these relationships.
- The findings highlight the necessity of understanding the nature of missing data and the implications of different imputation methods, as they can significantly influence the interpretation and conclusions of statistical analyses.
- The results advocate for the importance of conducting sensitivity analyses in research involving MNAR data to assess the robustness of the findings against various missing data scenarios.

In conclusion, this analysis demonstrates that MNAR data can notably distort the relationships between variables. It emphasizes the critical importance of careful methodological considerations in statistical modeling, particularly in the presence of MNAR data, to ensure accurate and reliable conclusions.

5 Conclusion

Our project's primary goal was to investigate the relationship between various factors and student stress levels. The dataset analysis revealed important insights into the factors that contribute to student stress.

According to the regression analysis, all four variables—academic performance, study load, teacher-student relationship, and future career concerns—have a significant impact on stress levels. Higher academic performance is associated with lower stress levels, according to the negative coefficient of academic performance. A higher study load, on the other hand, was found to increase stress levels, as were future career concerns. Surprisingly, a positive teacher-student relationship was linked to lower stress levels.

The model fits well, with a high R-squared value of 0.6957, indicating that these variables are significant predictors of stress levels. This implies that our model can account for approximately 69.57% of the variance in stress levels, which is quite significant.

The Analysis of Variance (ANOVA) was used to determine the effect of varying levels of social support on participants' self-esteem scores. The ANOVA revealed that social support has a significant effect on self-esteem. The Tukey post-hoc test determined which groups' means

differed from one another. According to the findings, social support has a statistically significant effect on self-esteem. Individuals with the highest levels of social support reported higher levels of self-esteem than those with lower levels.

The outcomes of our study provide important insights into the factors that contribute to student stress. Understanding these factors is critical for developing targeted stress management interventions. Future research could investigate the causal mechanisms of these relationships and whether interventions to increase social support and academic performance could effectively boost students' self-esteem and reduce stress levels.

Finally, our project reflected the complex interplay of various factors that contribute to student stress levels. It emphasizes the importance of an integrated approach to addressing student stress, including not only academic factors but also social support and self-esteem.

Implications for Stakeholders:
Educators: Our findings enable educators to implement tailored academic support mechanisms, foster positive teacher-student relationships, and create a positive and inclusive classroom environment to reduce student stress.
Policymakers: Policymakers can advocate for the integration of mental health resources within educational institutions as well as the implementation of flexible study load policies to effectively manage stressors among students.
Mental Health Professionals:: Mental health professionals can design targeted counseling programs and work closely with educational institutions to provide interventions to improve resilience and coping mechanisms.
Community Engagement: Encourage Parental Involvement and Community Support Programs: Encouraging parental involvement and community support programs can help to strengthen the network of support for students, contributing to a holistic approach to student well-being.

In conclusion, our project's nuanced findings suggest that tailored academic support, positive teacher-student relationships, and flexible study load policies can help educators and policymakers reduce student stress and shape a more supportive educational environment. Mental health professionals can create targeted counseling programs and advocate for integrated mental health resources, focusing on a collaborative approach to education and mental health. Community-level interventions can help students develop a broader sense of community support, forming a support network. These findings highlight the importance of educators, policymakers, mental health professionals, and communities working together to promote student well-being by providing concrete strategies and interventions for a more resilient and thriving student community.

References

[1] An Introduction to the Shapiro-Wilk Test for Normality — builtin.com. `https://builtin.com/data-science/shapiro-wilk-test`. [Accessed 07-12-2023].

[2] Mahaboob B., J. Praveen, B. Rao, Y. Harnath, C. Narayana, and Balaji Prakash. A study on multiple linear regression using matrix calculus. *Advances in Mathematics: Scientific Journal*, 9:4863–4872, 08 2020.

[3] Georgia Barbayannis, Mahindra Bandari, Xiang Zheng, Humberto Baquerizo, Keith W Pecor, and Xue Ming. Academic stress and mental well-being in college students: correlations, affected groups, and covid-19. *Frontiers in Psychology*, 13:886344, 2022.

[4] Rebecca Bevans. Multiple Linear Regression | A Quick Guide (Examples), 6 2023.

[5] Giulia Carreras, Guido Miccinesi, Andrew Wilcock, Nancy Preston, Daan Nieboer, Luc Deliens, Mogensm Groenvold, Urska Lunder, Agnes van der Heide, Michela Baccini, et al. Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the action study. *BMC Medical Research Methodology*, 21:1–12, 2021.

[6] Yuwei Deng, Jacob Cherian, Noor Un Nisa Khan, Kalpina Kumari, Muhammad Safdar Sial, Ubaldo Comite, Beata Gavurova, and József Popp. Family and academic stress and their impact on students' depression level and academic performance. *Frontiers in psychiatry*, 13:869337, 2022.

[7] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[8] Frank Emmert-Streib and Matthias Dehmer. High-dimensional lasso-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, 1(1):359–383, 2019.

[9] Nataša Erjavec. *Tests for Homogeneity of Variance*, pages 1595–1596. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[10] Jim Frost. Using post hoc tests with anova. *Statistics By Jim*, 2021.

[11] Andrew Gelman. Analysis of variance—why it is more important than ever. pages 1–4, 2005.

[12] Muhammad Hassan. ANOVA (Analysis of variance) - Formulas, Types, and Examples — researchmethod.net. `https://researchmethod.net/anova/`. [Accessed 07-12-2023].

[13] Saira Hossain, Sue O'Neill, and Iva Strnadová. What constitutes student well-being: A scoping review of students' perspectives. *Child Indicators Research*, 16(2):447–483, 2023.

[14] Amalia Karahalios, Laura Baglietto, John B Carlin, Dallas R English, and Julie A Simpson. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC medical research methodology*, 12(1):1–10, 2012.

[15] Ulrich Knief and Wolfgang Forstmeier. Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6):2576–2590, 2021.

[16] Marietta Kokla, Jyrki Virtanen, Marjukka Kolehmainen, Jussi Paananen, and Kati Hanhineva. Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study. *BMC bioinformatics*, 20(1):1–11, 2019.

[17] Yunus Kologlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, and Burhan Ozyilmaz. A multiple linear regression approach for estimating the market value of football players in forward position. *arXiv preprint arXiv:1807.01104*, 2018.

[18] Lynn R LaMotte. Three properties of f-statistics for multiple regression and anova. *arXiv preprint arXiv:2210.17199*, 2022.

[19] Benoit Liquet and Jérémie Riou. Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 13(1):1–10, 2013.

[20] Laura L Nathans, Frederick L Oswald, and Kim Nimon. Interpreting multiple linear regression: a guidebook of variable importance. *Practical assessment, research & evaluation*, 17(9):n9, 2012.

[21] Christine R Padgett, Clive E Skilbeck, and Mathew James Summers. Missing data: the importance and impact of missing data from clinical research. *Brain Impairment*, 15(1):1–9, 2014.

[22] Grigorios Papageorgiou, Stuart W Grant, Johanna JM Takkenberg, and Mostafa M Mokhles. Statistical primer: how to deal with missing data in scientific research? *Interactive cardiovascular and thoracic surgery*, 27(2):153–158, 2018.

[23] Oleg Sofrygin, Romain Neugebauer, and Mark J van der Laan. Conducting simulations in causal inference with networks-based structural equation models. *arXiv preprint arXiv:1705.10376*, 2017.

[24] Mark Tranmer and Mark Elliot. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 5(5):1–5, 2008.

[25] Matt N Williams, Carlos Alberto Gómez Grajales, and Dason Kurkiewicz. Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, 18(1):11, 2013.

[26] Kun Yang, Justin Tu, and Tian Chen. Homoscedasticity: An overlooked critical assumption for linear regression. *General psychiatry*, 32(5), 2019.

[27] Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1), 2016.