Applied Statistical Data Analysis

48-Hour Exam

Group Number: 01

Mohamed Khaled Elsafty

Seyed Bahram Taghavi Araghi

Hossein Ojaghi

Mayurbhai J. Odedra

Ayesha Jabeen

1 Contents

Heads bowed, whispers tense,
Deadline looms, confusion deep,
Coffee fuels their hope

1 Introduction

Student academic performance is influenced by a variety of factors, including study habits, attendance, participation, stress levels, and even lifestyle choices such as sleep duration and social media usage. Understanding these factors can help educators, students, and policymakers develop strategies to improve learning outcomes and overall academic success. The goal of this analysis is to determine which variables have the most significant impact on students' final grades and to identify actionable insights for academic improvement.

1.1 Objectives of the Report

This report is structured to achieve the following objectives:

- Identify key determinants of student success, such as study hours, attendance, and participation.
- Assess the impact of external lifestyle factors (stress levels, sleep patterns, social media usage) on academic performance.
- Explore gender-based performance differences to determine if certain groups require targeted interventions.
- Use data-driven visualizations to highlight trends and relationships between variables.
- Provide practical recommendations to improve study habits and academic performance.

1.2 Dataset Overview

The dataset used in this analysis consists of **10,000 student records**, each containing multiple variables that describe their academic behaviors, lifestyle choices, and performance indicators. The key variables included in the dataset are displayed in Table 1:

Table 1: Overview of variables:

Variables	**Description**
Demographic Characteristics	
Student ID	Unique identifier for each student
Age	Age of the student
Gender	Gender identity of the student (Male/Female/Other)
Study Habits	
Study Hours per week	Total hours dedicated to studying each week
Preferred Learning Style	The method in which the student learns best (e.g., visual, auditory, kinesthetic)
Online Courses Completed	Number of online courses successfully completed
Participations of Discussions	Indicates whether the student actively engages in classroom or

	online discussions
Attendance Rate	Percentage of classes attended by the student
Lifestyle Factors	
Self Reported Stress Level	A measure of how much stress the student perceives in their academic life (high/medium/low)
Time Spent on Social Media	Time spent on social media platforms (in hours per week)
Sleep Hours per Night	Average hours of sleep a student gets each night
Technological Usage	
Use of Educational Tech	Whether the student utilizes digital tools for learning (e.g., e-books, learning management systems)
Academic Performance Indicators	
Assignment Completion rate	Percentage of assignments submitted on time
Exam Score	Percentage of classes attended by the student
Final Grade	The overall academic performance classification, based on the exam score (A, B, C, D)

1.3 Data Cleaning

Before conducting any analysis, it is crucial to clean and preprocess the data to ensure accuracy and reliability. The following data cleaning steps were performed:

1.3.1 Handling Missing Values

The dataset was examined for **missing values** to ensure a complete sample. No significant gaps in the data were found, so that no removal of entries was necessary. All variables contained complete records.

(Any missing data in essential fields (e.g., Exam Score, Study Hours per Week) was either imputed using the median or removed if excessive.)

1.3.2 Removing Duplicates

A check for **duplicate records** was performed in order to avoid redundant data entries. The results showed no duplicate entries were present, which confirmed that the dataset did not require deduplication.

(Checked for and eliminated any duplicate student records to ensure data integrity.)

1.3.3 Standardizing Data Formats

Reformatted categorical variables (e.g., Gender, Preferred Learning Style) for consistency.

1.3.4 Outlier Detection

Identified and addressed extreme values in numeric fields.

- **Encoding Categorical Variables** – Converted categorical variables into numerical values where necessary for analysis.
- **Identification of Connected Variables** – A key relationship in the dataset was identified between Exam Score (%) and Final Grade. Since these two variables are inherently linked, this dependency was taken into account to prevent redundancy and misleading interpretations in subsequent analyses

1.4 Analysis of Data

Table 2 provides the descriptive statistics for the continuous variables. The average age of respondents is approximately 23.5 years, with ages ranging from 18 to 29 years. On average, students dedicate about 27 hours per week to studying, although the variation is substantial (standard deviation of 13 hours), which suggests big differences in study habits. Participants completed around 10 online courses on average, with the completion rates showing relatively high variation. Assignment completion rates and exam scores averaged roughly 75% and 70%, respectively, with, again, wide variations indicating diverse academic performances among students. Attendance rates average around 75%, also displaying significant variability. The average time spent on social media per week was about 15 hours. Lastly, sleep duration averages around 7 hours nightly, with limited variability.

Table 2: Descriptive Statistics of continuous variables

Variable	Obs	Mean	Std. Dev.	Min	p25	p50	p75	Max
Age	10000	23.479	3.462	18	20	23	27	29
Study Hours per Week	10000	27.130	13.003	5	16	27	38	49
Online Courses Completed	10000	10.008	6.137	0	5	10	15	20
Assignment Completion (%)	10000	74.922	14.675	50	62	75	88	100
Exam Score (%)	10000	70.189	17.649	40	55	70	85	100
Attendance Rate (%)	10000	75.085	14.749	50	62	75	88	100
Time spent on Social Media (hours/week)	10000	14.937	9.023	0	7	15	23	30
Sleep Hours per night	10000	6.979	1.997	4	5	7	9	10

Note: Table 2 displays all continuous variables included in the analysis. Table 1 provides detailed descriptions of the variables

Table 3 displays the Spearmen`s rank correlation coefficient matrix for all variables used in the data analysis. All variables display a very small correlation coefficient (between -0.02 and 0.03), which indicates a non-existent linear relationship between the different variables. This is strengthened by the statistical insignificance almost throughout every interaction. Combined with the very high sample size, this indicates that the small associations exist due to randomness and that no meaningful linear dependencies can be taken out from the correlation analysis. The only variables which correlate at 10%-significance level ($p < 0.1$) are the interactions between the Number of Online Courses and Study Hours per week as well as Sleep Hours per night and Exam Score. However, also here the correlations remain at a very low level, which also indicates very weak linear relationships.

Table 3: Correlation matrix

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Age	1.00							
(2) Study Hours per Week	0.00	1.00						
(3) Online Courses Completed	0.00	0.03*	1.00					
(4) Assignment Completion Rate	-0.01	0.00	0.00	1.00				
(5) Exam Score (%)	0.00	0.00	0.00	0.01	1.00			
(6) Attendance Rate(%)	-0.01	0.01	0.00	0.00	0.00	1.00		
(7) Social Media Time (h/week)	0.01	0.01	0.00	0.01	0.00	-0.01	1.00	
(8) Sleep Hours per night	0.01	0.01	-0.01	0.01	-0.02*	-0.01	0.00	1.00

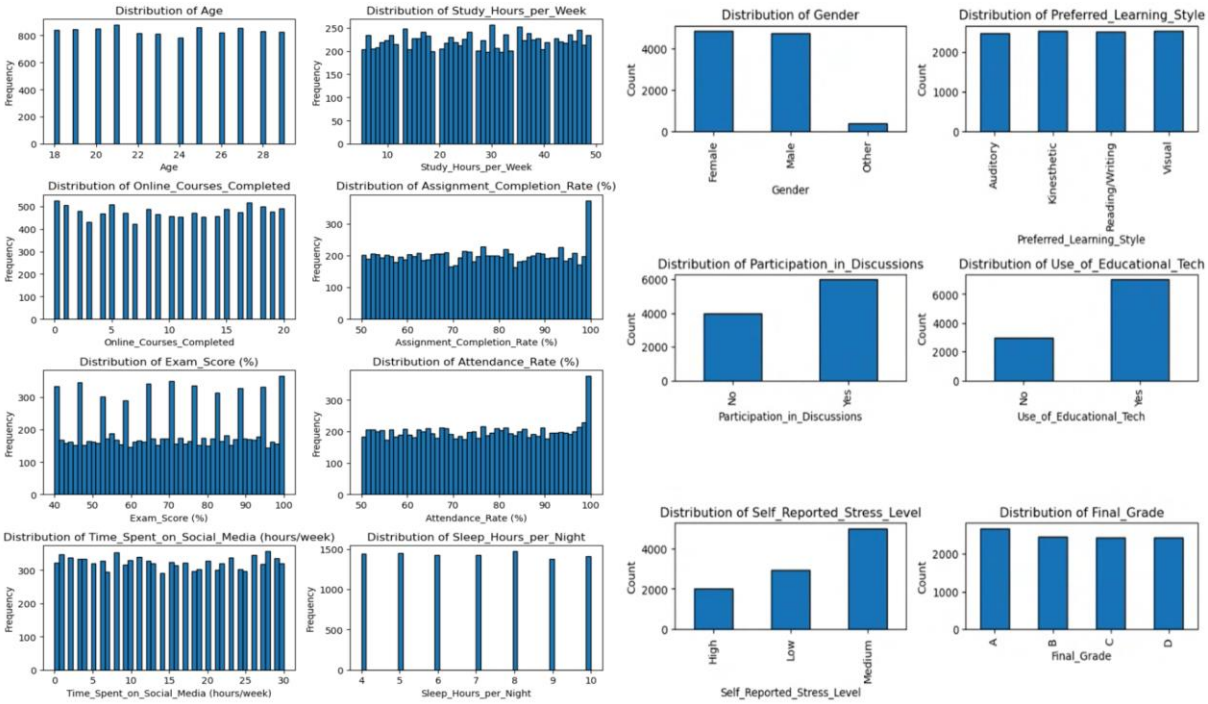
*** p<0.01, ** p<0.05, * p<0.1

Note: Table 3 represents Spearmen`s rank correlation coefficients for the variables used in the study. Table 1 provides detailed descriptions of the variables

Figure 1 displays the distributions of the continuous and categorical variables. The continuous variables exhibit evenly spread distributions with uniform-like patterns. This is investigated further by conducting normality tests. Since the dataset contains a large number of statistical units ($n = 10{,}000$), the D'Agostino-Pearson and Anderson-Darling tests were chosen for their effectiveness in assessing normality in large samples. The results show an extremely low p-value ($p < 0.01$) and high test statistics, indicating that all tested variables significantly deviate from normality. Therefore this report does not

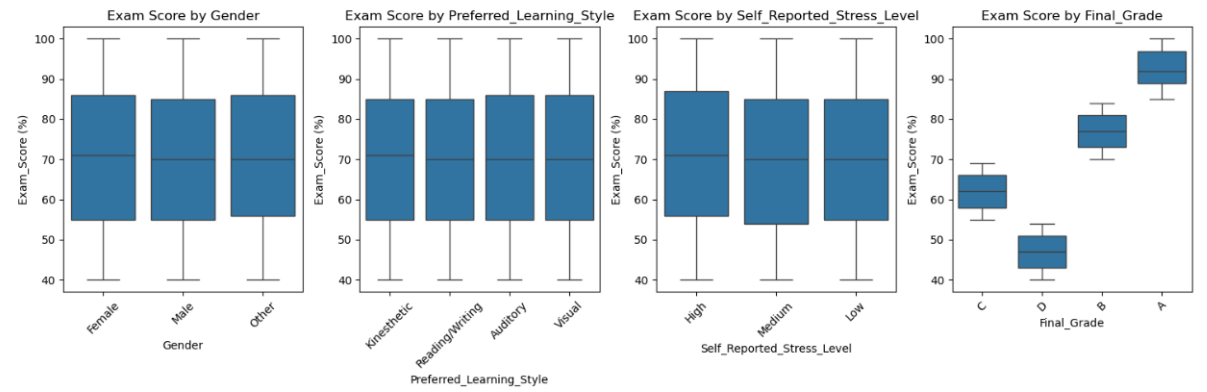
use t-tests and ANOVAS for group comparisons, but focuses on the Mann-Whitney U test and the Kruskal-Wallis test.

Figure 1: Distribution of continuous (left) and categorical variables (right)



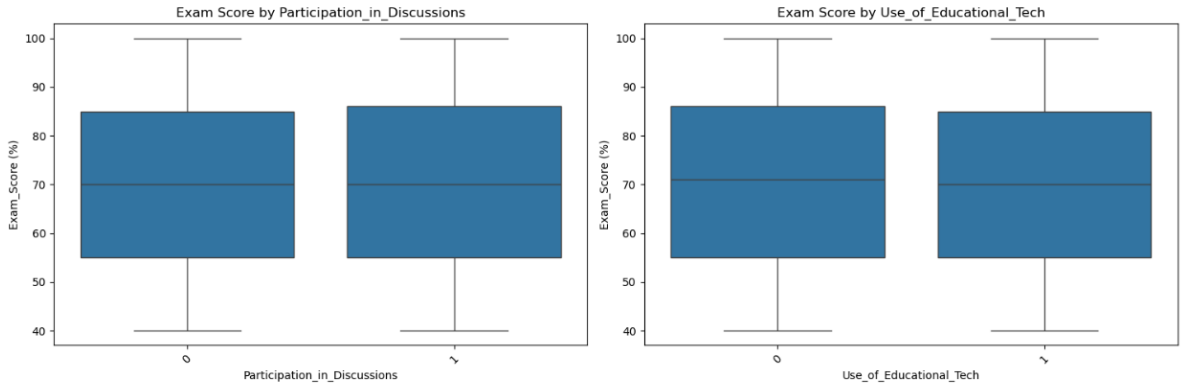
In this analysis, we aimed to compare exam scores across different categorical variables. Since the data did not meet the normality assumption, a key requirement for ANOVA and t-tests, non-parametric alternatives were applied. The Mann-Whitney U Test was applied to binary categorical variables, while the Kruskal-Wallis Test was applied to variables with three or more groups.

Figure 2: Comparison of Exam Scores Across Demographic and Academic Factors



The Mann-Whitney U Test showed no significant difference in exam scores between genders (p = 0.92128). The Kruskal-Wallis Test revealed no significant difference in exam scores based on preferred learning style (p = 0.54167). However, a significant difference was found for self-reported stress level (p = 0.02659), suggesting stress may impact performance. Finally, a highly significant difference in exam scores was observed based on final grades (p < 0.00001), indicating a strong correlation between final grades and exam performance.

Figure 3: Impact of Participation and Technology Use on Exam Performance



The Mann-Whitney U test results indicate no statistically significant difference in exam scores between students who participated in discussions and those who did not (p = 0.88242). Similarly, the use of educational technology did not show a significant impact on exam scores (p = 0.57714). The boxplots visually confirm this, as the distributions of exam scores appear nearly identical across both groups. This suggests that neither participation in discussions nor the use of educational technology had a strong influence on students' exam performance in this dataset.

There is no significant difference in exam scores based on gender (p = 0.92128), indicating that gender does not correlate with exam performance. Similarly, preferred learning style does not strongly influence exam scores (p = 0.54167), and participation in discussions shows no significant difference (p = 0.88242), suggesting no strong relationship between discussion participation and exam performance. Additionally, the use of educational technology does not show a significant difference (p = 0.57714), pointing to no clear correlation between technology use and exam scores. However, self-reported stress levels significantly impact exam performance, with higher or lower stress potentially correlating with better or worse scores (p = 0.02659).

2 Feature Engineering and Data Transformation

To extract deeper insights from the raw dataset, several new features were engineered to capture nuanced aspects of student performance. These derived metrics not only simplify the dataset but also provide more interpretable relationships between behavioral factors and academic outcomes. The key engineered features are as follows:

- **Study Efficiency**: Defined as the ratio of Exam Score (%) to Study Hours per Week (with an added constant to avoid division by zero), this metric quantifies the effectiveness of study time. It offers a normalized view of how well students convert study efforts into exam performance.
- **Participation Effectiveness:** Calculated as the product of Assignment Completion Rate (%) and the encoded measure of Participation in Discussions, this feature reflects the combined impact of homework diligence and classroom engagement on student success.
- **Social Media Impact:** Expressed as the ratio of Exam Score (%) to Time Spent on Social Media, this metric provides an indication of how distractions may affect academic performance.
- **Stress Performance:** By dividing Exam Score (%) by the encoded Self-Reported Stress Level, this metric seeks to capture the interplay between stress and academic achievement, offering insight into how stress management might relate to performance.
- **Sleep Efficiency:** Derived as Exam Score (%) divided by Sleep Hours per Night, this feature underscores the potential influence of sleep quality on exam outcomes.
- **Attendance Effectiveness:** This is the product of Exam Score (%) and Attendance Rate (%), encapsulating the idea that consistent class attendance may directly boost academic performance.

After creating these features, the original columns used for these calculations were removed, resulting in a refined dataset that focuses on the most performance-relevant variables.

2.1 Correlation Analysis

Table 4: Correlation matrix of Feature Engineering

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Age	1.00											
(2) Online Courses Completed	0.00	1.00										
(3) Assignment Completion Rate	-0.01	0.00	1.00									
(4) Exam Score (%)	0.00	0.00	0.01	1.00								
(5) Attendance Rate(%)	-0.01	0.00	0.00	0.00	1.00							
(6) Final Grade	0.00	0.00	0.01	**0.97**	0.01	1.00						
(7) Study Efficiency	0.00	-0.02	0.01	**0.33**	0.00	**0.32**	1.00					
(8) Participation Effectiveness	0.00	-0.01	**0.23**	0.00	-0.01	0.00	0.00	1.00				
(9) Social Media Impact	-0.01	0.00	0.00	**0.15**	0.02	**0.15**	0.05	0.00	1.00			
(10) Stress Performance	0.00	0.01	0.00	**0.46**	0.00	**0.45**	**0.14**	0.00	0.07	1.00		
(11) Sleep Efficiency	-0.01	0.01	0.00	**0.68**	0.01	**0.66**	**0.23**	-0.01	**0.10**	**0.32**	1.00	
(12) Attendance Effectiveness	-0.01	0.00	0.01	**0.78**	**0.62**	**0.76**	**0.26**	-0.01	**0.13**	**0.37**	**0.53**	1.00

Note: Table 3 represents Spearmen`s rank correlation coefficients for the variables used in the study. Table 1 provides detailed descriptions of the variables; Values over 0.1 are presented thick

Table 4 displays the results of the correlation coefficient matrix based on the Feature Engineering. Most of the variables still display a very weak relationship with each other. However some more positive correlation could be observed with the new variables. While the high correlation (0.97) between Final Grade and Exam Score is unsurprising due to the automatic interaction with each other, other factors also offer positive relationships with both variables. Here it is observed that Study efficiency (0.33) and Stress performance (0.46) have both a solid positive relationship to the Exam Score, while Sleep Efficiency (0.68) and Attendance Effectiveness (0.78) have a strong relationship.

2.2 Bivariate Analysis

Figure 4 displays histograms, which illustrate how each feature in the refined dataset is distributed. The distribution has changed due to the adjustments and includement of new variables. Attendance effectiveness, Sleep & Study efficiency as well as stress performance now offer a right-skewed distribution.

Figure 4: Distribution of variables in Feature Engineering

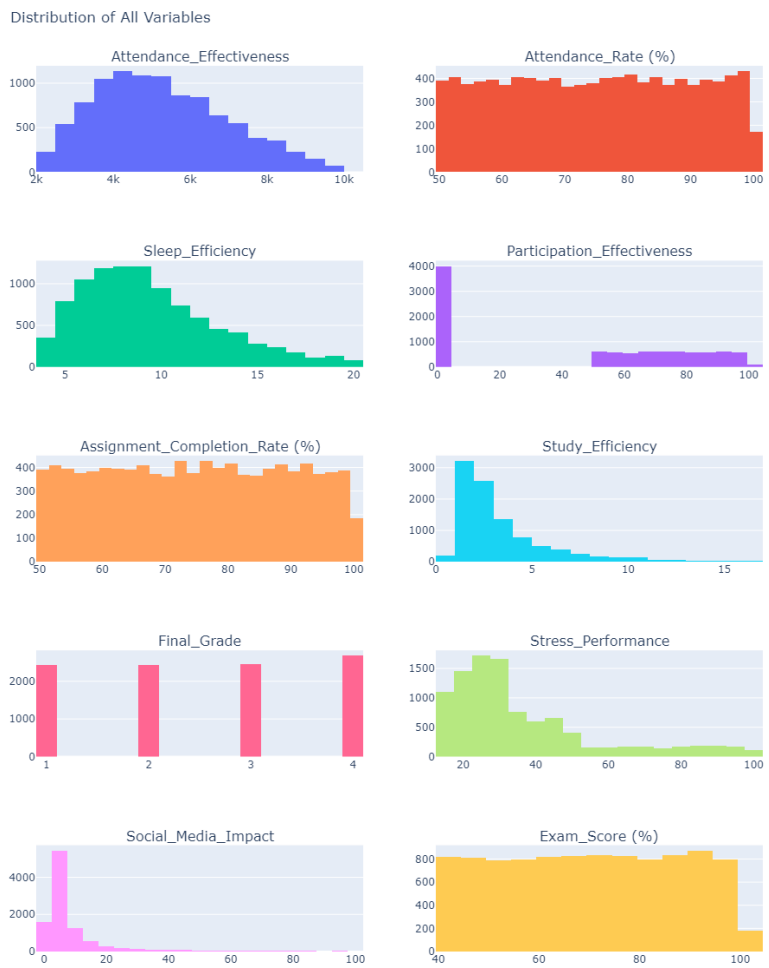


Figure 5 displays the positive linear relationship between Attendance Effectiveness and the Exam Score, indicating that a positive relationship might exist between both variables

Figure 5: Linear relationship of Attendance effectiveness and Exam Score

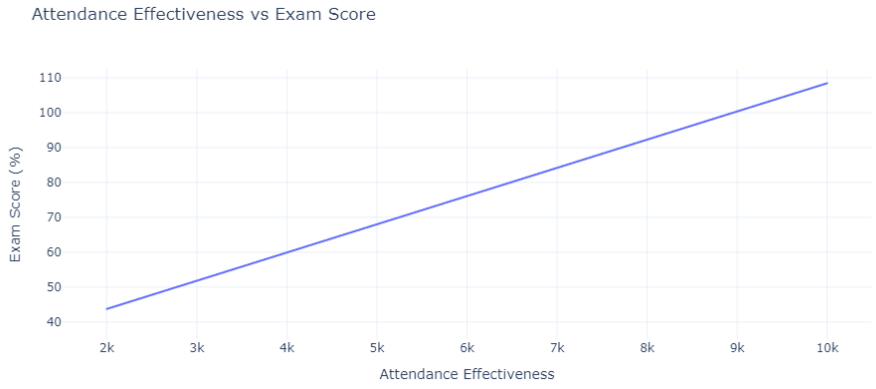
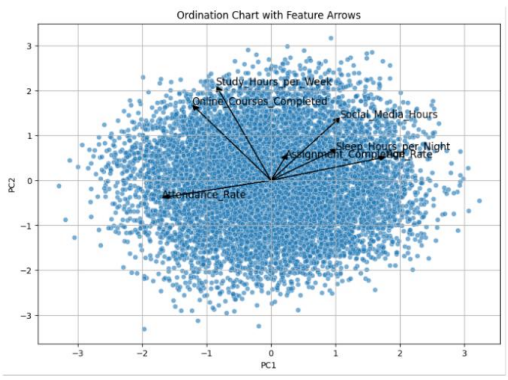
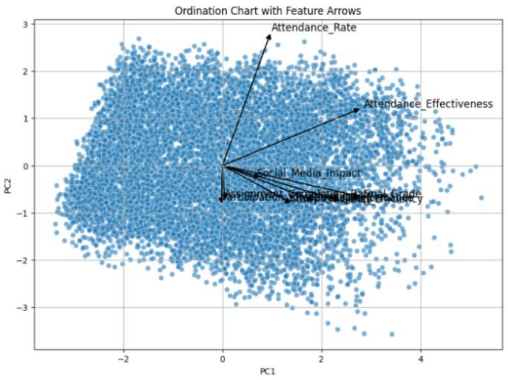


Figure 6 displays the ordination charts with normal and adjusted variables. The first ordination chart, which included the given variables indicates that Attendance, assignment completion, and good sleep habits group together, suggesting that students who attend regularly and sleep enough usually complete assignments successfully. On the other side, online courses, social media use, and study hours form another group, indicating students who prefer digital learning and flexible study methods. The right diagram focuses more on performance and participation. Attendance, participation, and attendance effectiveness strongly relate, meaning actively attending classes significantly influences student success. Social media impact and assignment proficiency form separate patterns, suggesting they influence students differently. While attending and participating strongly predict performance, assignment skills and social media habits show unique ways students might achieve success or face challenges in their academic journey.

Figure 6: Ordination charts with normal and adjusted variables





2.3 Model Performance Analysis

Several models were applied to predict exam scores, and their performance was evaluated using Mean Squared Error (MSE) and Akaike Information Criterion (AIC). Support Vector Machine (SVM) achieved the lowest MSE (3.0791), indicating the best predictive accuracy among all models. Linear Regression, OLS, and GLM (Gaussian) had identical MSE values (7.0095) and AIC (38613.16), suggesting similar performance. GLM (Poisson) - Model 4 performed the worst, with MSE (93.768) and the highest AIC (59216.46), indicating poor model fit. Refer Appendix for graphs.

Table 7: Model Performance Analysis

Models	**Mean Square Error**	**AIC**
Linear Regression	7.0095	
OLS	7.0095	38613.16
Support Vector Machine	3.0791	
GLM (Gaussian)	7.0095	38613.16
GLM (Poisson) - Model 4 (Exam Score ~ Attendance Effectiveness + Sleep Efficiency + Stress_Performance + study_efficiency +social_media_impact)	93.768	59216.46

3 Limitations to the Analysis

We attempted to identify factors which influence student performance in exams. However, the effects, which were observed, were either statistically insignificant or simply too weak to show the effects. Based on this, we come along with various limitations regarding the accuracy and extent to which the results can be interpreted.

One major challenge is that most features in this dataset have weak correlations to exam scores to yield meaningful results, which means that additional factors should be included in order to identify the influence on academic performance more effectively. For example a student's home environment and mental well-being can have a big impact on their academic journey. Students from wealthy backgrounds often have access to private tutors, who can teach them better skills and they also have access to high quality learning materials. Parents who receive a higher level of education at home can also provide more academic support. Mental health is another important factor. High levels of stress, fear, and lack of sleep can make it more difficult to concentrate and retain information. Because data records do not cover these aspects, some of the most important predictors of the test may be missing. Additionally, the dataset does not distinguish between private and public institutions, despite significant differences between these settings in terms of resources. Such differences can have a big influence on academic results. By omitting these institutional factors, the current analysis has the risk of overlooking meaningful variations that could explain differences in academic achievement.

Besides characteristic variables, this data record also does not account for any important regional and structural factors. Differences in educational infrastructure, school fundings or quality of teaching are closely linked to geographic locations and can significantly affect student performance. Moreover, cultural attitudes toward education vary greatly depending on the region, which can influence academic outcomes.

Moreover, although including the adjusted variables brought better results, it can be biased due to the fact that these variables were already influenced by the ratio to exam score. In General, a combination of different methodological approaches didn't show meaningful results, which indicates that the information given in the dataset lacks more information to conduct a data analysis with strong, meaningful results.

4 Appendix

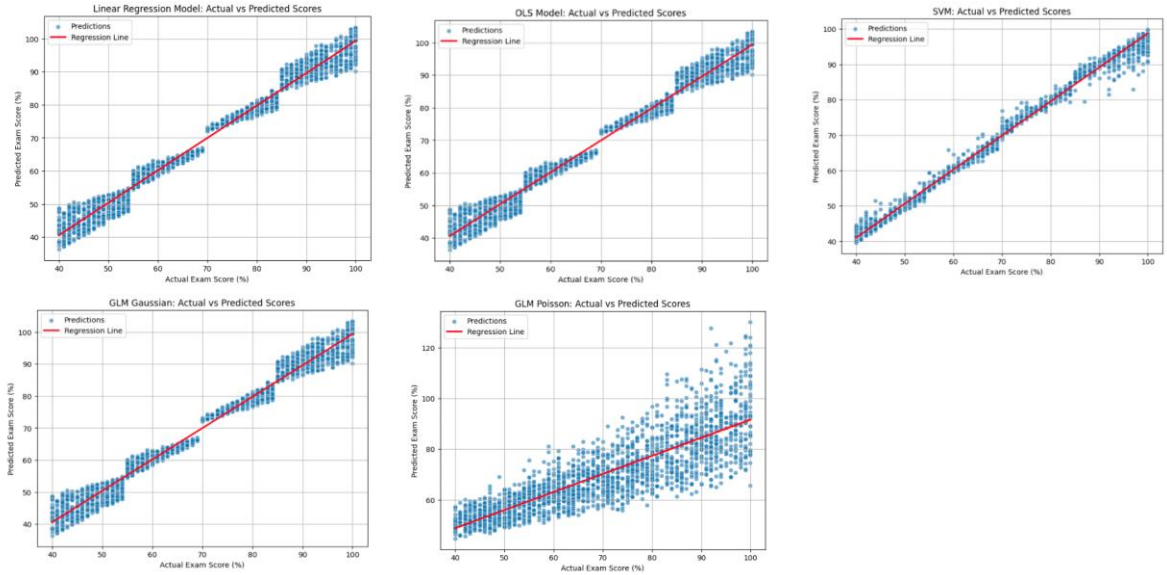
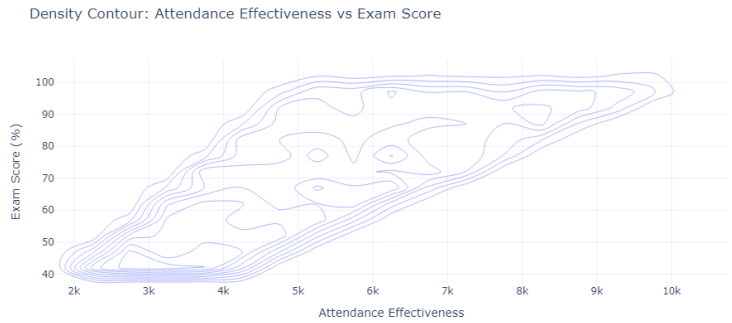
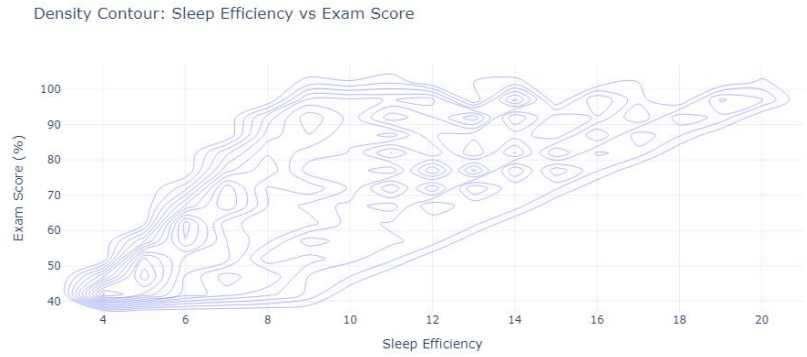


Figure: Regression graphs of the different models comparing Actual vs Predicted value

4.1 Density Contour Plots

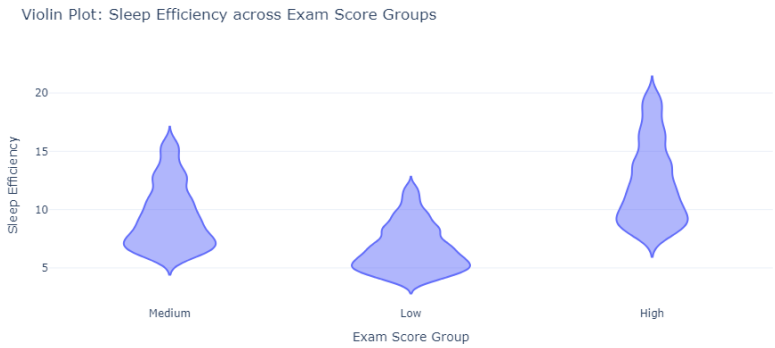


This plot displays the concentration of data points, indicating common ranges of attendance and exam performance.



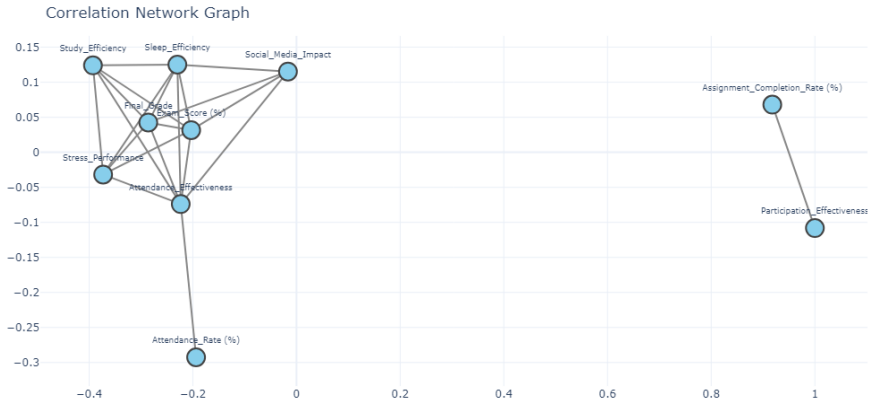
Another contour plot reveals the density of observations, providing a nuanced view of how sleep efficiency is distributed relative to exam scores.

4.2 Violin Plots by Exam Score Groups



These violins show how Sleep Efficiency is distributed among Low, Medium, and High exam score groups. Students with higher exam scores generally have higher Sleep Efficiency, suggesting a link between adequate sleep and stronger performance.

4.3 Correlation Network Graph



This graph visualizes how features interconnect based on their correlations. Each node represents a feature, and edges indicate statistically significant relationships. The cluster on the left centers around exam performance, attendance, and sleep efficiency, whereas assignment completion rate and participation effectiveness form a separate cluster on the right, highlighting distinct patterns in the data.