Predicting Academic Performance from Student Habits

Names: Calef Alesse, Sam Amorsolo

Introduction

The dataset we used was from kaggle, and included information on students' habits and behaviors vs exam scores. This data is synthetic, which means it was generated to reflect real-life patterns found in similar data, but it itself is not actual data collected from real students. Benefits of this include that it is easier to work with when generating and training models, as well as cost-effectiveness of not having to actually conduct a study to collect data. The 16 variables included in the dataset were: student id (unique id), age (age of student), gender (male/female/other), study_hours_per_day (avg. daily study time), social_media_hours (daily social media time), netflix_hours (avg. daily Netflix/binging time), part_time_job (yes/no), attendance_percentage (class attendance 0–100%), sleep_hours (avg. daily sleep), diet_quality (poor/fair/good), exercise_frequency (times per week), parental_education_level (high school/bachelor/other), internet_quality (good/average/other), mental_health_rating (scale 1–10), extracurricular_participation (yes/no), and exam_score (final exam score 0–100). This dataset included 1000 rows of unique synthetic data, making it a large enough dataset to use effectively on building, training, and testing all our models.

Question 1: What student habits are the strongest predictors of academic performance (exam score), and are these predictors consistent across genders?

Methods

For this question, we used the dataset to find the most influential predictors of exam score by using linear regression and dimensionality reduction. For preprocessing, first, the data was cleaned by dropping any missing values, dummying categorical variables into continuous predictors, splitting the data into training and testing data sets, and standardizing all continuous variables through z-scoring, so they are all on the same comparable scale.

We first built the linear regression model, which is under the assumption that a linear relationship between the habits and exam performance. This model considered all the variables to see how well they predicted final exam scores altogether. To identify the most influential predictors, we used LASSO regression, a type of dimensionality reduction called feature selection, which reduces the number of variables in the model

by making less important ones zero, essentially removing their influence from the regression.

Finally, the models were evaluated on how much of the variance of exam scores they captured and the mean average error (MAE) between the predictions and actual exam scores. This entire process was repeated for each gender (male, female, and other) to see if the results were consistent between all subgroups. To visualize the results, we used bar charts, which showed each predictor's influence on academic performance, one that showed overall and another for within each gender subgroup. Additional scatter plots and boxplots helped show the relationship between found top predictors, like study time and mental health, and student exam scores.

Results

This analysis showed us results that the linear regression model was able to capture about 88% of the variance in exam scores, and had an MAE of 4.26. After LASSO was applied, the variance increased slightly to 89%, and MAE decreased to 4.21. LASSO also helped discover the top student habits that were associated with academic success.

Looking at all the variables, the number of hours spent studying per day was determined to be the most influential predictor of exam scores, with a coefficient of 14.125. This showed that students who studied more on average tended to achieve higher exam scores, with a strong positive correlation. Another strong predictor was mental health rating; students who rated their mental well-being higher also performed better academically, with a coefficient of 5.49. On the other hand, predictors like social media time and high Netflix hours were associated with lower exam scores, with negative coefficients of -2.88 and -2.12, respectively.

When the analysis was repeated by gender, the most important predictors remained constant for all 3 categories (male, female, "other"), both positive and negative. Study hours and mental health remained top predictors for male, female, and other-gender students. However, some minor differences were observed. For example, attendance happened to be a stronger factor for female students, while sleep and exercise had slightly more influence for the male students. For students identifying as "other," the smaller sample size made patterns harder to generalize, but study hours and mental health still stood out as important influences.

These results were visualized through multiple graphs (figures 1-6) that help conceptualize the relationships. The bar charts showed the relationship between each predictor and the exam score, for the entire dataset, as well as each gender subgroup. The scatterplots visualized the top two most influential predictors, study time and mental health, where you can see a positive upward trend, with color to indicate gender.

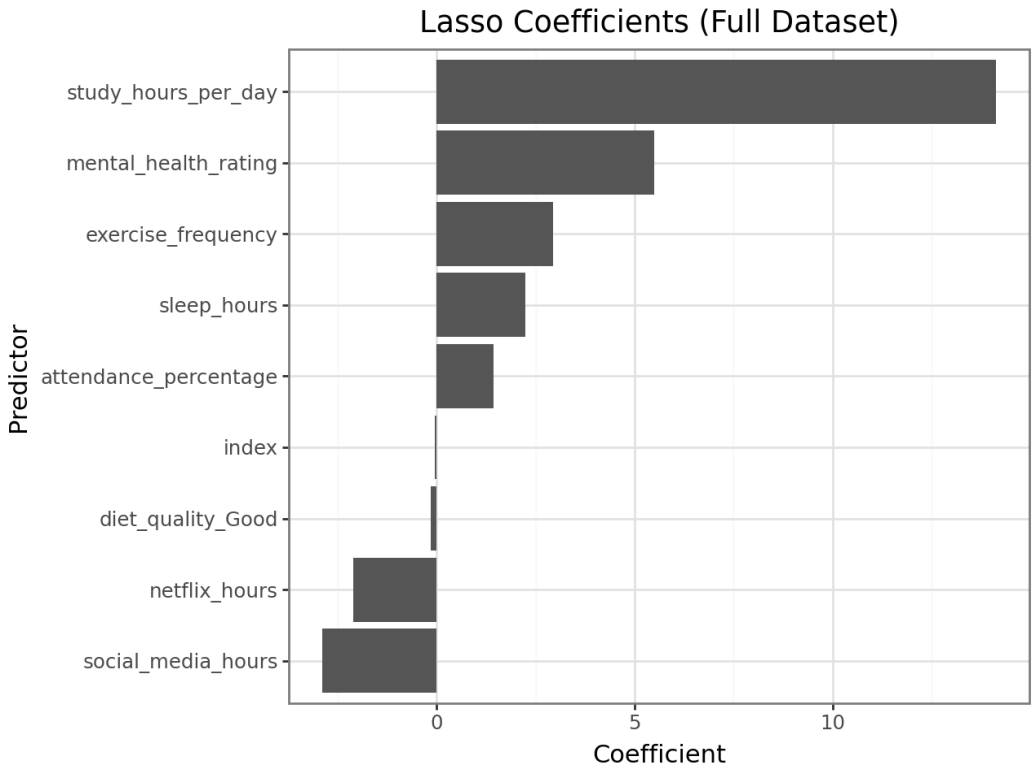

Figure 1: Bar chart of LASSO coefficients for all genders

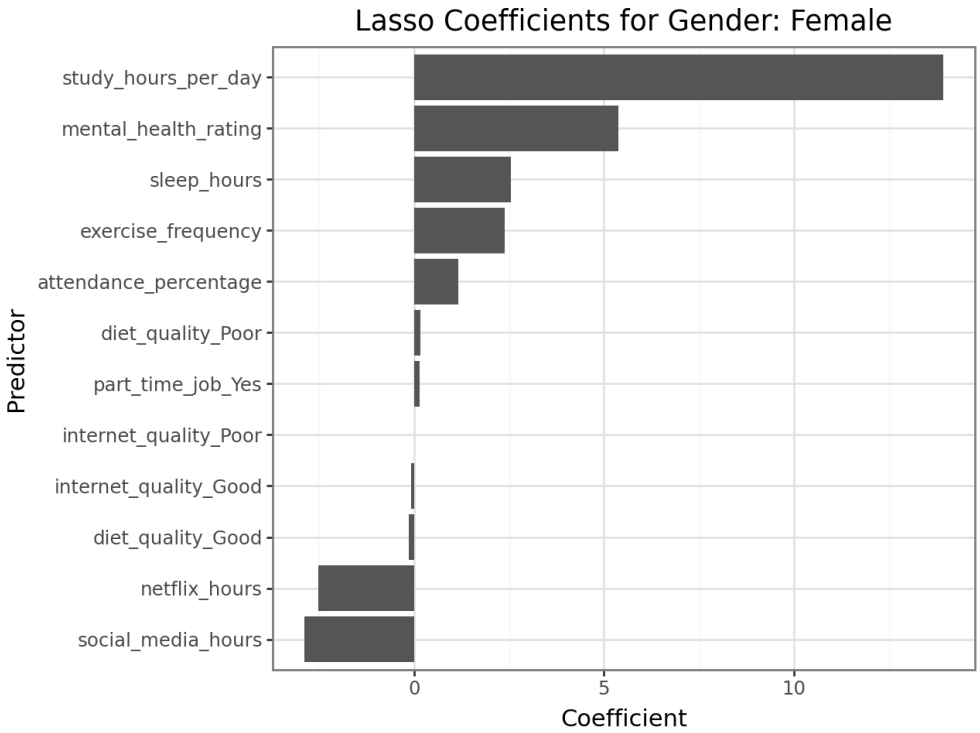

Figure 2: Bar chart of LASSO coefficients for females

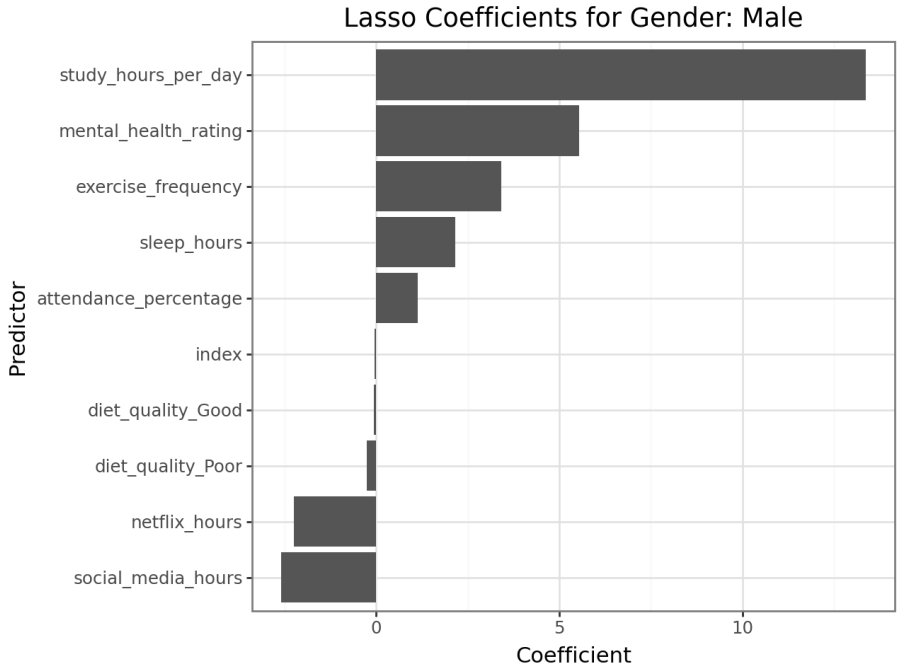

Figure 3: Bar chart of LASSO coefficients for males

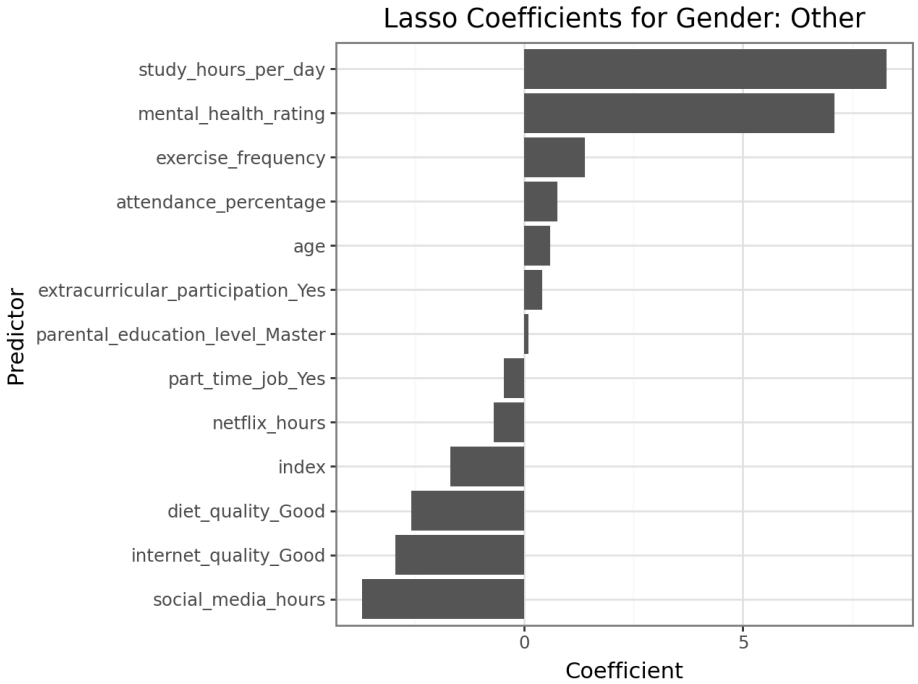

Figure 4: Bar chart of LASSO coefficients for students who listed “other”

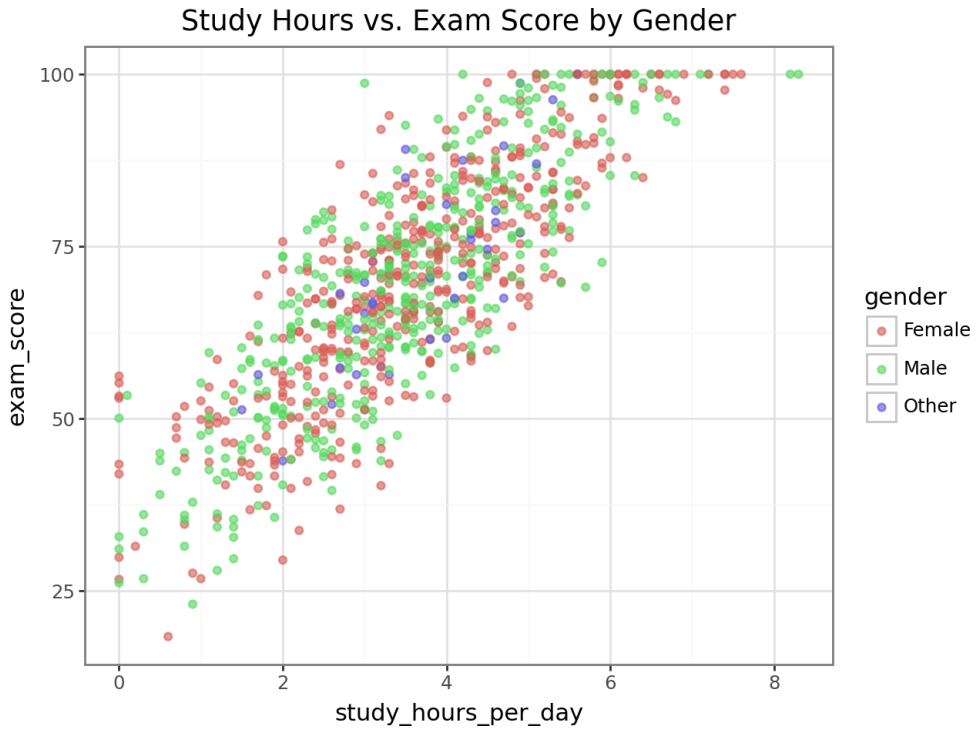

Figure 5: Scatterplot of hours studied per day and exam score, color indicates gender

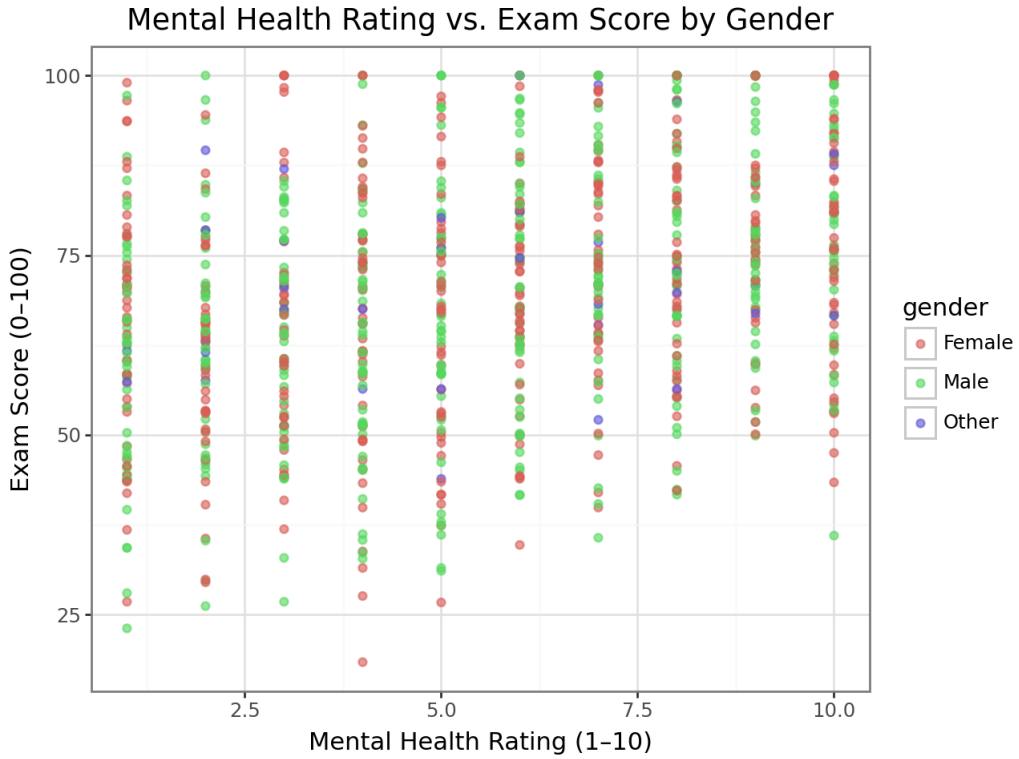

Figure 6: Scatterplot of mental health rating and exam score, color indicates gender

Table 1: Model Performance

	R^2	MAE
Linear Regression	0.889	4.2575
LASSO	0.891	4.2089

Discussion

Overall, this analysis is useful for helping to identify which student habits most significantly influence academic performance, specifically final exam scores. For example, the model indicates that consistent study habits and strong student mental health can lead to better exam scores for all students, regardless of gender. This insight is valuable for many communities, including educators, parents, and students themselves, as it highlights the behaviors and habits that are most worth spending time doing or avoiding. Additionally, the consistency of the top predictors (like study hours and mental health) across gender subgroups reinforces their importance and suggests that strategies that target these specific factors could be effective for the broader student population.

However, there are also limitations to this model and analysis. Firstly, using a linear regression that assumes all the variables have an independent linear relationship with the outcome (exam score), does not necessarily reflect real-life behavior. It is also important to note that the dataset used was made of synthetic data, which is artificially generated to mimic the structure and patterns of real-world data, but is not actually collected from real students. This means that the models' findings may show certain patterns, but they may not actually be applicable to real-world scenarios without additional validation on actual datasets.

To improve this analysis in the future, it would be interesting to apply the same modeling approach to real data collected from actual students. Also, using non-linear models could help capture more complex relationships between habits and performance that could appear from working with non-synthetic data.

Question 2: When clustering students based on their average study hours per week, sleep hours per night, and social media usage, what clusters emerge, and how do these clusters differ in terms of average academic performance?

Methods

To observe how students' habits of average study hours per day, average sleep hours per night, and daily social media usage cluster together, and how they relate to exam scores, we performed K-Means clustering. First, the data was cleaned and preprocessed by dropping any missing variables and standardizing the variables using z-scores so that they would be on the same scale, which is necessary when clustering algorithms that use distance metrics.

We chose K-Means clustering because it works well with distinct clusters and continuous data. K-Means is efficient, interpretable, and performs well when data is assumed spherical and equally distributed, which we assume to be reasonable based on the chosen variables. Also, K-Means uses hard assignment, where each student can only belong to one cluster, rather than every cluster with some probability. This is more desirable in our context, where we aim to categorize students into distinct lifestyle groups to make interpretation and comparisons easier.

To tune the hyperparameter k, the number of clusters, silhouette scores from 2-10 were graphed to determine the best k, with the highest correlating silhouette score. Once we determined the best number of clusters, K-Means was applied to the data, and assigned each student to a cluster. To evaluate among the clusters, we calculated the average exam score to explore the differences between the groups. For visualization, PCA was used to reduce the dimensions into 2, which were used to create a 2D scatter plot of the clusters. Additionally, a bar chart was created to show the average exam scores by cluster, making it easy to compare performance across different lifestyles defined by the clusters. Finally, the clusters were summarized through a table that shows the mean values of the clustering input features for every cluster.

Results

This analysis identified 6 clusters to be best to represent students based on the chosen variables of average study hours per day, sleep hours per night, and social media usage. This was chosen through the silhouette score graph having the highest peak at k=6. The goal of these clusters are to identify a variety of lifestyles, from the extremes of students who have good study habits and low social media consumption to those who have poor sleep patterns, high social media engagement and rarely study, as well as intermediate groups with a mix of habits.

When comparing these clusters' academic performance, there were clear differences of average exam scores between the groups. Students in Cluster 0 had the highest average exam scores around 85%, and was made up of those who had an average study time of 5 hours per day, 6 hours of sleep per night, and 1.5 hours of social media time. On the other hand, Cluster 2 had the lowest academic performance

of around 50%, and was made up of students who studied on average 2 hours per day, slept 7 hours a night, and spent over 3.5 hours on social media. The intermediate clusters showed mixed results, with exam scores falling between the highest and lowest groups, reflecting varying lifestyle combinations, but consistently reflecting healthier lifestyles with higher academic performance.

The clusters were not very highly separated or cohesive on the 2D scatterplot, as its silhouette score did not rise above 0.25, indicating considerable overlap between the clusters, which is evident on the graph. So, while some patterns emerged from the clusters, like correlation between healthier studying/living habits and higher exam scores, the low silhouette score cautions users to limit the confidence they can have in clearly defining behavioral types based on this clustering alone.

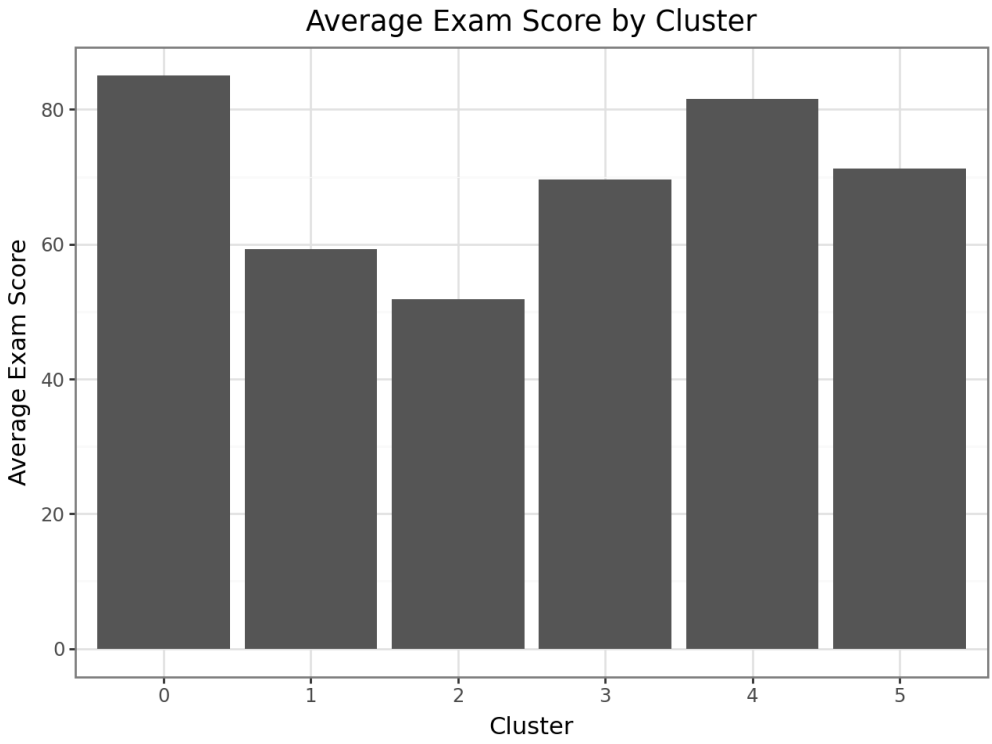

Figure 1: Bar chart of Clusters and Average Exam Score

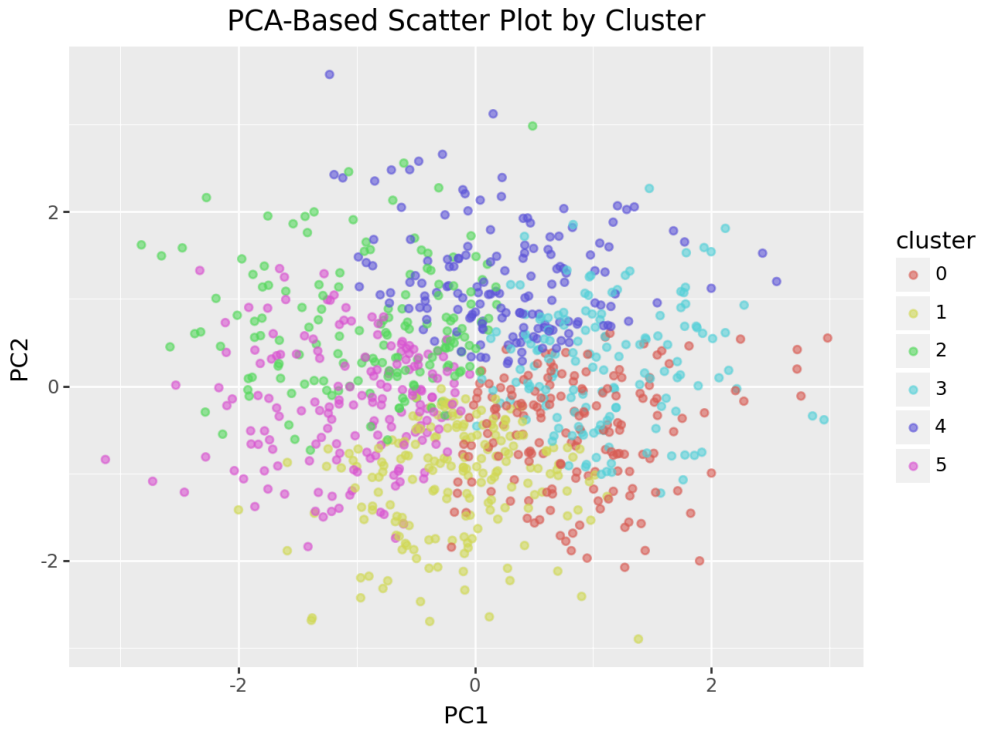

Figure 2: Scatterplot of clusters

Table 1: Summary of mean values of input features for each cluster

Cluster	Study Hours Per Day	Sleep Hours Per Night	Social Media hours
0	4.98	6.07	1.57
1	2.48	5.68	1.71
2	2.07	6.91	3.61
3	4.01	5.01	3.45
4	4.87	7.32	3.39
5	3.08	7.82	1.66

Discussion

This clustering model is good for identifying patterns in student habits and groups with similar characteristics based on the features of study hours, average sleep, and

social media use. This is useful for students, families, and schools to understand the impacts of different combinations of these features and how they relate to academic performance. For example, a cluster that showed a combination of higher study time and lower social media use, happened to correlate with higher exam scores. Insights like this could help in developing targeted programs for lower performing students, based on their levels of study hours, average sleep, and social media time.

However, the clustering results also revealed important cautions and limitations. The silhouette score for the model was below 0.25, which suggests that the clusters were not very distinct. The PCA scatter plot further confirmed this, showing that the clusters overlapped significantly with each other, rather than optimally being clearly separated groups. This could be attributed to the fact that only three randomly selected features were used in the clustering, which may have caused oversimplification in the model and not represented the complexity of student behavior well.

To improve this model in the future, additional variables could be included to widen the scope of the model, specifically features that were identified as more influential on academic performance from the test conducted in Question1. Also, exploring other clustering algorithms like GMM or DBSCAN, which are better for working with non spherical clusters, could uncover more meaningful clusters and insights. Once again, collecting real world data, over synthetic, would make the findings more realistic and reliable when considering application to academic settings.

Question 3: Which student habit variables have the strongest relationship with whether student scores above or below the average exam score?

Methods

For this question, we used a logistic regression model to predict whether a student's exam score was above or below the average exam score. First the data was cleaned by dropping any missing values. Then, the categorical variables, diet_quality and extracurricular_participation, were converted into dummy variables. Additionally, a new binary target column was added to classify each student if they scored above or below the mean exam score.

Next, we split the data into training and test sets using 80/20 split to ensure the model would be evaluated on unseen data. All the continuous variables were standardized by z-scoring. Then, we performed the Logistic Regression using the training data to model the probability of a student scoring above the average exam score based on their daily habits. The coefficients from this model influence the strength and direction of the relationship between each habit and the result. To determine which habits had the strongest relationship with the performance, we looked at the absolute values of the standardized coefficients. Looking at the absolute values allows you to

observe the magnitude and strength of the relationship, regardless of the direction. We wanted to identify which habits have the strongest impact, not solely looking at the positive relationships, and this allows us to consider both positive and negative effects equally. We then can rank the standardized coefficients based on importance even if they help or hurt the performance.

The results were visualized using two graphs (horizontal bar chart and box plot) to help interpret the model's results. First, the horizontal chart presented the absolute standardized coefficients from the Logistic Regression evaluation. This chart allowed us to observe the strengths of each predictor (daily habits) and whether or not they had a strong influence on the likelihood of scoring an above average exam score. Second, a box plot was used to show how the strongest habit predictor, study_hours_per_day, differed between students who scored above the average exam score and those who did not. This helped analyze the predictor's impact by showing the distribution of study time across the observed groups.

Results

The Logistic Regression evaluation resulted in a strong predictive performance with a test test accuracy of 81.3%. With this score it means the model was able to correctly classify whether a student scored above or below average in over 8 out of 10 cases. The recall score of 77.8% indicates that the model successfully identified a majority of the students who scored above average, highlighting its ability to detect true positives. A precision score of 86.5% means that the model was able to predict if a student would score above the average, reflecting a low false positive rate. Finally, the ROC AUC score of 0.91 indicates a great classification ability across all categories. The model suggests that it was able to successfully distinguish between above-average and below-average student scores effectively.

The most influential predictor of exam performance was study_hours_per_day, with a coefficient of 3.02, which is significantly higher than any other variable. This suggests that the more time spent studying during the student's day has a strong positive correlation of scoring above the mean average exam score. Additionally, other strong predictors included exercise_frequency and sleep_hours which had positive relationships indicating that healthy habits of including more exercise and sleep along with longer study times in the students' routine, correlated with high exam scores. Comparatively, the results revealed that high social media and Netflix screen times are associated with lower exam scores and academic performance as it can be a distraction for students taking away from their study times. It can also be noted that variables such as attendance_percentage and diet_quality showed a small difference, while extracurricular_participation had very little influence on the results.

These findings were visualized using two graphs (Figure 1- Horizontal Bar Chart & Figure 2- Boxplot). Figure 1 presents a Horizontal Bar Chart of the absolute

standardized coefficients from the Logistic Regression Model. The graph compares their relative impact and strength of each habit predictor, regardless of their direction being positive or negative. Figure 2 shows a Box Plot to compare the daily study hours between students who scored above and below the average exam score. This helps visualize the real-world difference in behavior between the students. The spread and medians in the plot highlight that students who spent more time studying had a noticeable difference in exam scores being above-average.

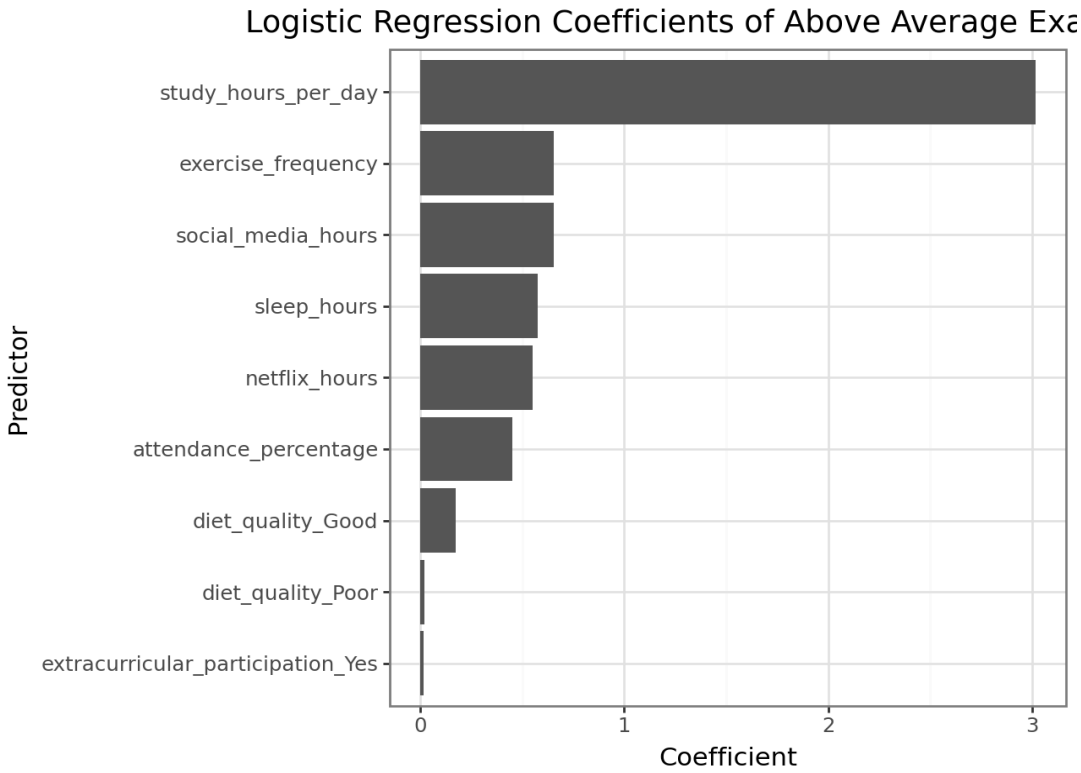

Figure 1: Bar chart of Absolute Standardized Logistic Regression Coefficients For Predicting Above-Average Exam Scores

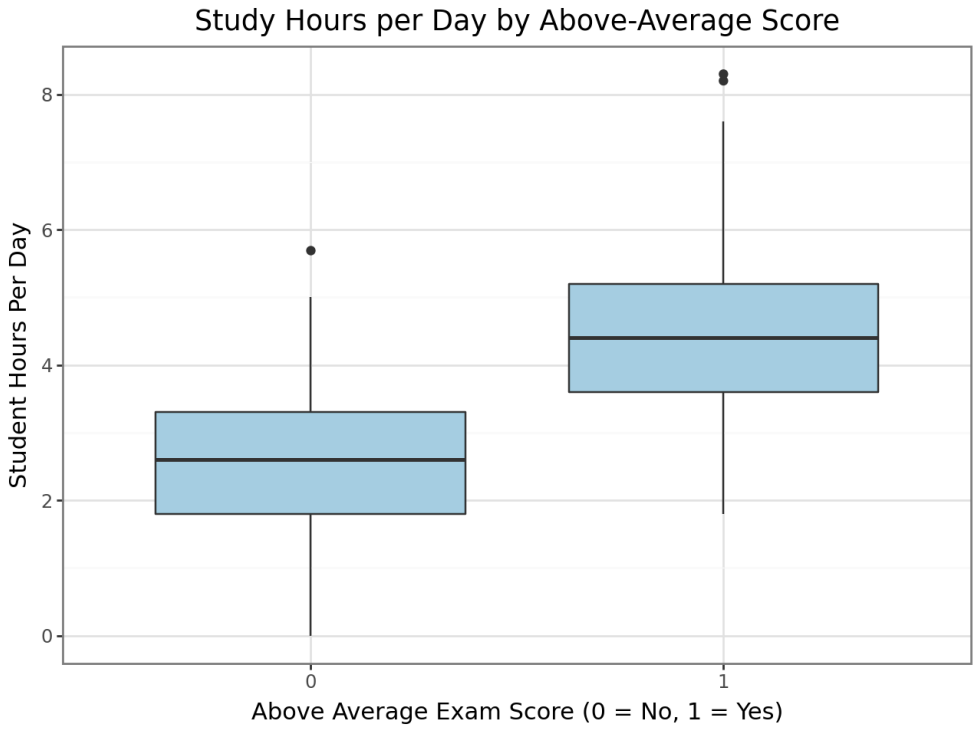

Figure 2: Boxplot of Comparing Study Hours between High and Low Performing Students

Discussion

This analysis effectively answers the question by looking closely at the data and quantifying the relationship between students' daily habits and their academic performance using a Logistic Regression evaluation. By standardizing the predictors, we were able to compare their impact revealing that study_hours_per_day was the most impactful predictor when scoring an above-average exam score. The analysis is valuable because it helps inform educators and students. Knowing this knowledge helps encourage students to prioritize healthy habits and be mindful of how their habits impact their academic performance. It allows educators to advise students that are struggling to reflect on their habits and what they could change about their routine to improve their academic performance.

One advantage of using Logistic Regression is the interpretability of the coefficient performance. We were able to analyze which coefficients have the most influence through the outcome probability. The model performed well on unseen test data indicating strong relationships. However, a limitation in using this data was that it is synthetic and not drawn from real students. Therefore our findings have the possibility of being skewed, although it also has a possibility of being true. Our data should be validated through actual survey data tested on real students.

In the future, exploring non-linear models or interactions can impact how study time combined with sleep patterns and diet can impact the model's performance. This analysis clearly shows how a student's daily habits correlate with their academic performance.

Question 4: How does having a part-time job (or not) affect a student's exam scores? Is there a relationship between the two variables, and is it linear or more complex?

Methods

We explored the relationship between students having part-time work and their academic performance. We performed two Linear Regression models to predict students' exam scores based on whether or not they have a part-time job. So we analyzed the data. There were no missing values and all the variables were continuous. Then, we standardized the variables by z-scoring.

The first model we tested, isolated the effect of having part-time employment by using only the binary variable, part_time_job, as the independent predictor of predicting the exam score. This allowed us to assess the impact of how having a part-time job affects academic performance. Then, for the second model we added additional predictors of academic and wellness variables including study_hours_per_day, sleep_hours, attendance_percentage, and mental_health_rating. This broadened the model as it was able to help determine whether the part-work affected the exam scores or whether the other potential predictors were controlled for. Both models were Linear Regressions, assuming a straight-line relationship between predictors and the exam score.

Both models were trained on an 80/20 split to be evaluated on unseen data. Then their performance was evaluated using R^2 and Mean Absolute Error (MAE) metrics for both training and test sets. We then further analyzed by adding visualizations. First visualization is a box plot to compare the exam score distributions between students with and without part-time jobs. This graph helped show a clear performance comparison among job status within the students. Second graph we created is a scatter plot which helped picture the relationship between daily study hours and exam scores, with color coding based on having a part-time job. This helped visualize whether the relationship between study time and performance differed for students who had a job and didn't. We also assessed if there were any non-linear patterns to see if it was more complex.

Results

The first Linear Regression model predicted exam scores independently on part-time job status as a predictor of exam scores. This resulted in extremely low predictive power, shown by the training set of achieving a R^2 score of 0.0004, indicating none of the variation in exam scores. On the test set the R^2 score dropped to -0.012, meaning the model actually performed worse than predicting the average exam score for all students. The MAE was approximately 13.5 in both sets, suggesting the model has higher error. These results show that students with a part-time job alone is not meaningful to predict a student's exam performance.

However, the second model included four additional continuous predictors that included study_hours_per_day, sleep_hours, attendance_percentage, and mental_health_rating. With these additional predictors the model performed significantly better. The model had an R^2 score of 0.815 on the training set and 0.809 on the test set, indicating that the model explained over 80% of the variance in exam scores. The MAE also dropped significantly to 5.7 on the test set, gaining major improvement and a more accurate model.

We created visualizations that helped support the results. The Box Plot in Figure 1 compared exam scores by the status of having employment or not, but presented minimal differences in median and spread. Reinforcing the idea that having a part-time job is not a sole indicator of predicting exam scores. Meanwhile, the Scatter Plot in Figure 2 revealed a positive linear relationship between study hours and exam scores. Particularly a stronger, steeper relationship for students without jobs although some overlap is present. This suggests that study habits and behaviors overall are more predictive of academic performance, and part-time work may not greatly impact your academic career.

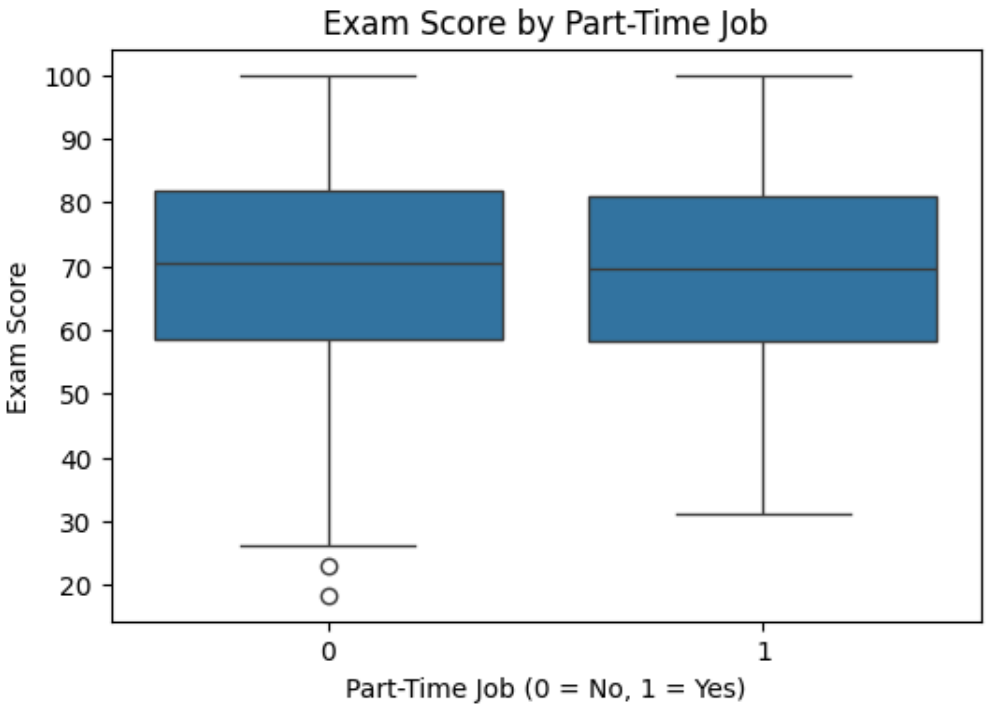

Figure 1: Box plot of Median Exam Scores for Students with and without Part-Time Jobs

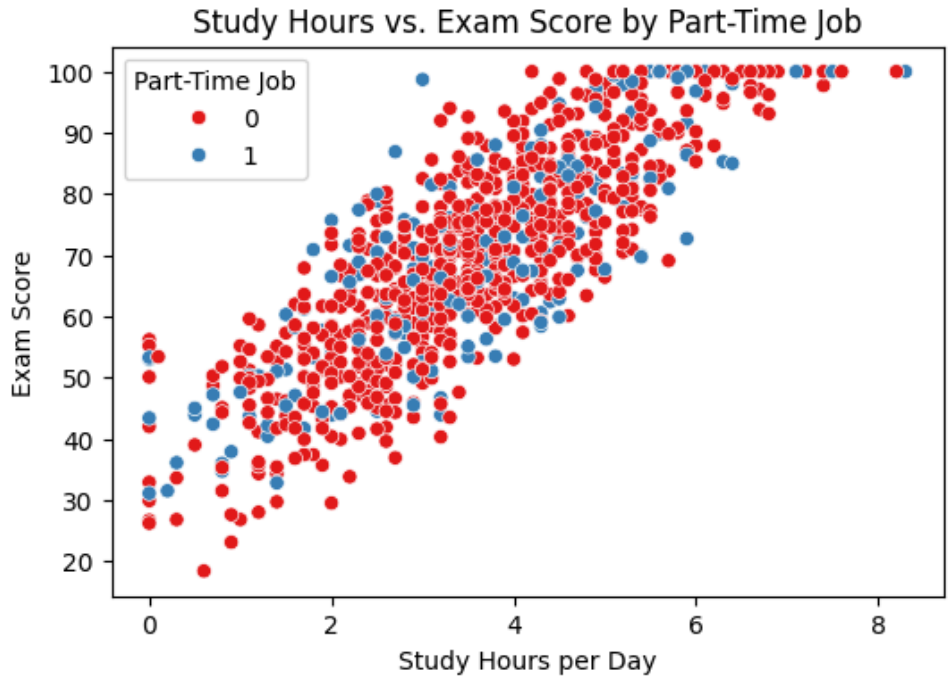

Figure 2: Scatterplot of Study Hours vs. Exam Score colored by Part-Time Job Status

Table 1: Regression Performance Summary

Model	Predictors	Train R^2	Test R^2	Train MAE	Test MAE
Model 1	Part-Time Job	0.004	-0.012	13.700	13.523
Model 2	Part-Time Job, Study Hours, Sleep, Attendance Percentage, Mental Health Rating	0.815	0.809	5.900	5.725

Discussion

This analysis focused on the effect of students having a part-time job on their academic performance while accounting for other important academic and wellness variables. We used two regression models of one solely using employment status as a predictor and another that included the other study habit predictors. Comparing the performance of the two models, we saw that having a part-time job alone is not a meaningful determinant of exam performance. When analyzed by itself, it has no significant relationship compared when study habits of sleep, attendance, and mental health are accounted for. This was supported by our model as the accuracy improved dramatically, and then the contribution of part-time jobs became minimal.

The graphs we created further reinforced these results, as Figure 1 of the box plot showed no significant performance gap between students who had a part-time job and those without. This aligned with the weak relationship of our test results. Additionally, Figure 2 the scatter plot demonstrated a strong, positive relationship between study hours and exam scores. There was no strong evidence of non-linearity, suggesting that using linear regression models was appropriate for this analysis.

Overall, our results suggest that students with part-time jobs do not negatively impact their academic performance on its own, and are not a strong predictor of whether or not they will achieve success. Instead, consistent study habits and keeping up with your mental health are strong predictors that impact the students' exam scores more significantly. This information can give educators and students insight about having a part-time job. Although it's important to keep up healthy study habits, students shouldn't feel discouraged getting a part-time job in fear of how it will affect their academics. As well as suggesting that it is possible to balance work and school assignments and assessments.