Leading Question:

Which Factors Are Associated With Exam Score After Accounting for Confounding Variables?

Purpose:

The purpose of this project is to explore possible relationships between student habits and exam score performance. By diving into what factors affect exam score taking into account confounding variables, we can identify behaviors that most directly affect academic performance.

Dataset Information and Preprocessing Data:

Source: Kaggle

Author: Jayanta Nath

Title: Student Habits vs Academic Performance

Columns: Student ID, Age, Gender, Hours of Study per Day, Social Media Hours per Day, Netflix Hours per day, Part Time Job, Attendance Percentage, Sleep Hours, Diet Quality, Exercise Frequency, Parental Education Level, Internet Quality, Mental Health Rating, Extracurricular Participation, Exam Score

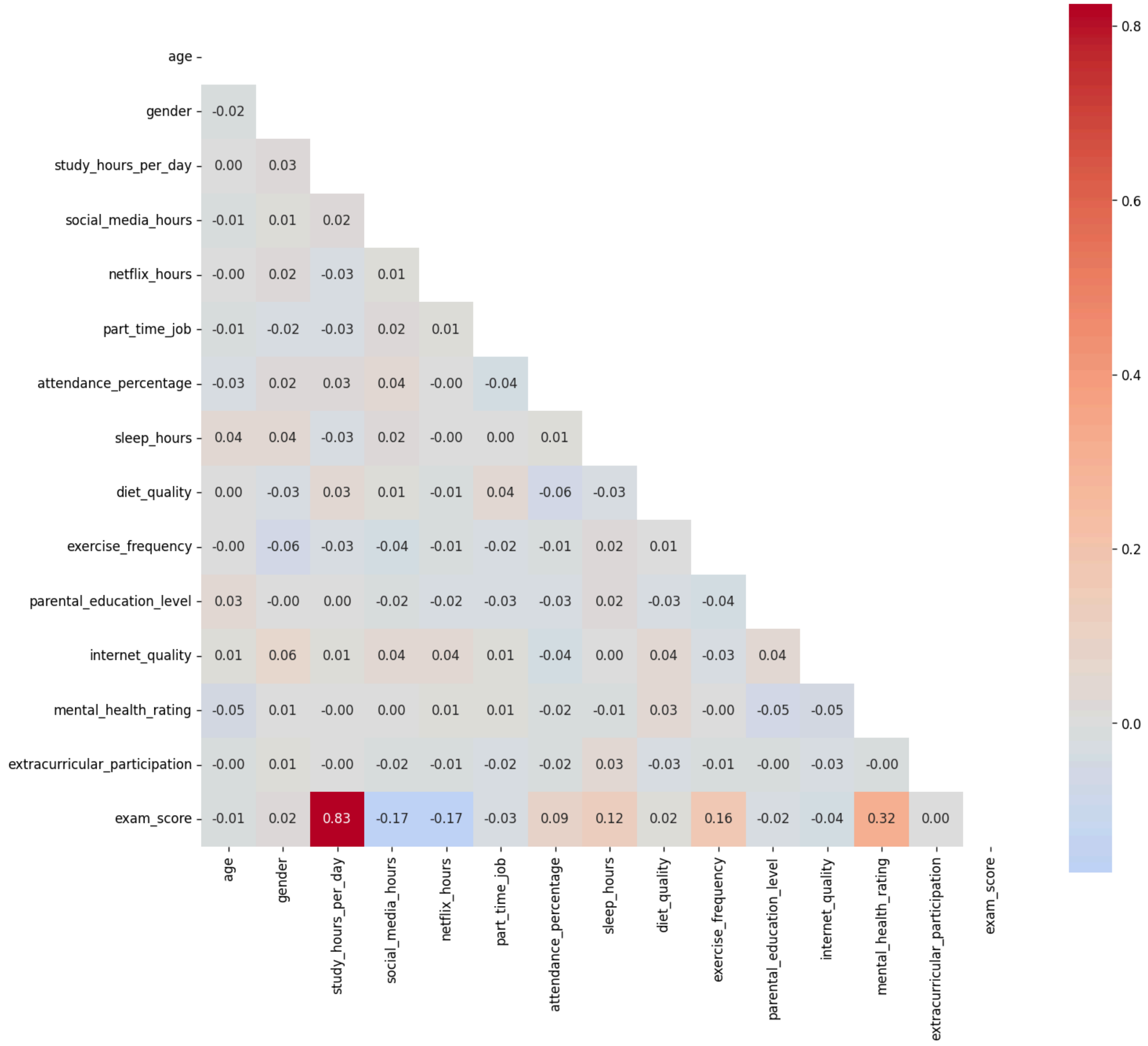
Length: 16 columns, 1000 rows

Target Feature: Exam Score

This dataset contains information on student habits as well as their exam score outcomes. There were no duplicates and the only null values were in the parental education level column. Categorical variables were all encoded numerically. Exam score was selected as the target feature.

Exploratory Analysis:

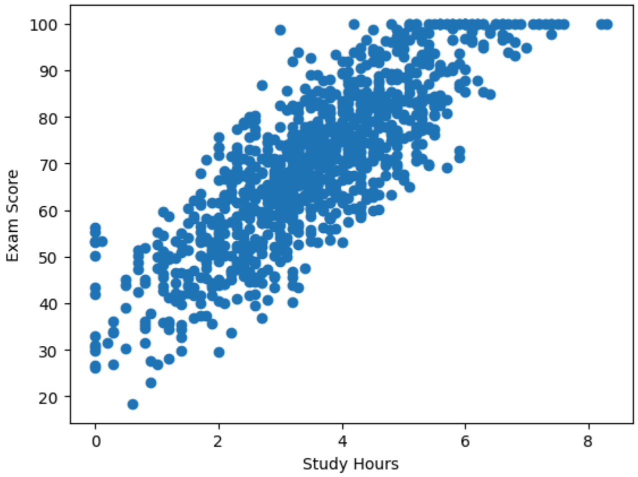
Heat Map



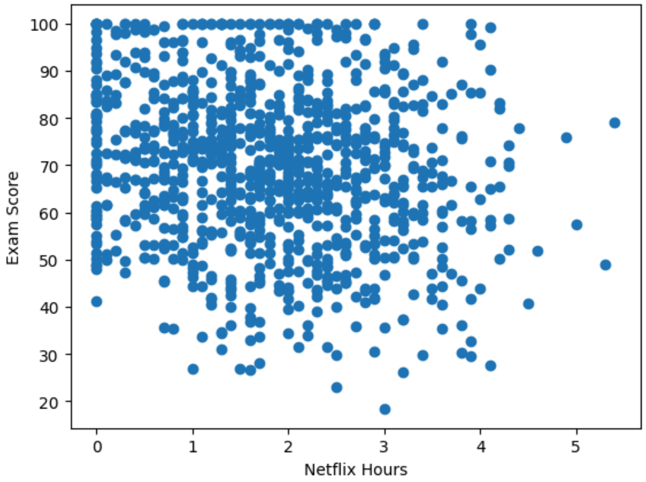
	age	gender	study_hours_per_day	social_media_hours	netflix_hours	part_time_job	attendance_percentage	sleep_hours	diet_quality	exercise_frequency	parental_education_level	internet_quality	mental_health_rating	extracurricular_participation	exam_score
age															
gender	-0.02														
study_hours_per_day	0.00	0.03													
social_media_hours	-0.01	0.01	0.02												
netflix_hours	-0.00	0.02	-0.03	0.01											
part_time_job	-0.01	-0.02	-0.03	0.02	0.01										
attendance_percentage	-0.03	0.02	0.03	0.04	-0.00	-0.04									
sleep_hours	0.04	0.04	-0.03	0.02	-0.00	0.00	0.01								
diet_quality	0.00	-0.03	0.03	0.01	-0.01	0.04	-0.06	-0.03							
exercise_frequency	-0.00	-0.06	-0.03	-0.04	-0.01	-0.02	-0.01	0.02	0.01						
parental_education_level	0.03	-0.00	0.00	-0.02	-0.02	-0.03	-0.03	0.02	-0.03	-0.04					
internet_quality	0.01	0.06	0.01	0.04	0.04	0.01	-0.04	0.00	0.04	-0.03	0.04				
mental_health_rating	-0.05	0.01	-0.00	0.00	0.01	0.01	-0.02	-0.01	0.03	-0.00	-0.05	-0.05			
extracurricular_participation	-0.00	0.01	-0.00	-0.02	-0.01	-0.02	-0.02	0.03	-0.03	-0.01	-0.00	-0.03	-0.00		
exam_score	-0.01	0.02	0.83	-0.17	-0.17	-0.03	0.09	0.12	0.02	0.16	-0.02	-0.04	0.32	0.00	

Notable Aspects:

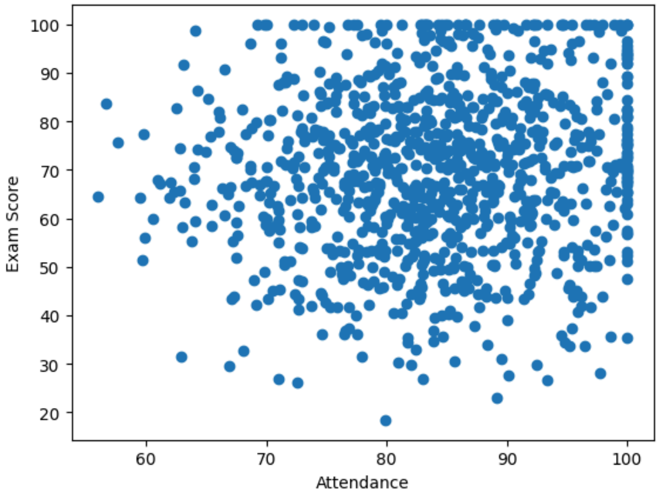
Study Hours	r = 0.83	strong positive correlation
Mental Health Rating	r = 0.32	weak positive correlation
Social Media Hours	r = -0.17	weak negative correlation
Netflix Hours	r = -0.17	weak negative correlation
Exercise Frequency	r = 0.16	weak positive correlation



Scatter Plot 1)
Exam Score vs Study Hours
Strong Positive Linear Correlation



Scatter Plot 2)
Exam Score vs Netflix Hours
Weak Negative Linear Correlation



Scatter Plot 3)
Exam Score vs Attendance
Weak Positive Linear Correlation

Modeling:
The chosen features continuing forward are study hours per day, mental health rating, netflix hours per day, and social media hours per day. The magnitude of the r value must be above or equal to 0.17 to be considered a feature for the following models.

Linear Regression
A linear regression model was trained using the selected features: study hours per day, netflix hours per day, mental health rating, and social media hours per day. The dataset was split into training and testing with 20% of the data used for testing. Feature scaling was also applied to allow for comparison of relative feature importance ignoring units. After standardization, study hours had the largest positive coefficient followed by mental health rating, while Netflix hours showed a negative coefficient. Fixed random seed.

MSE: 47.825
R^2: 0.813

Random Forest Regression

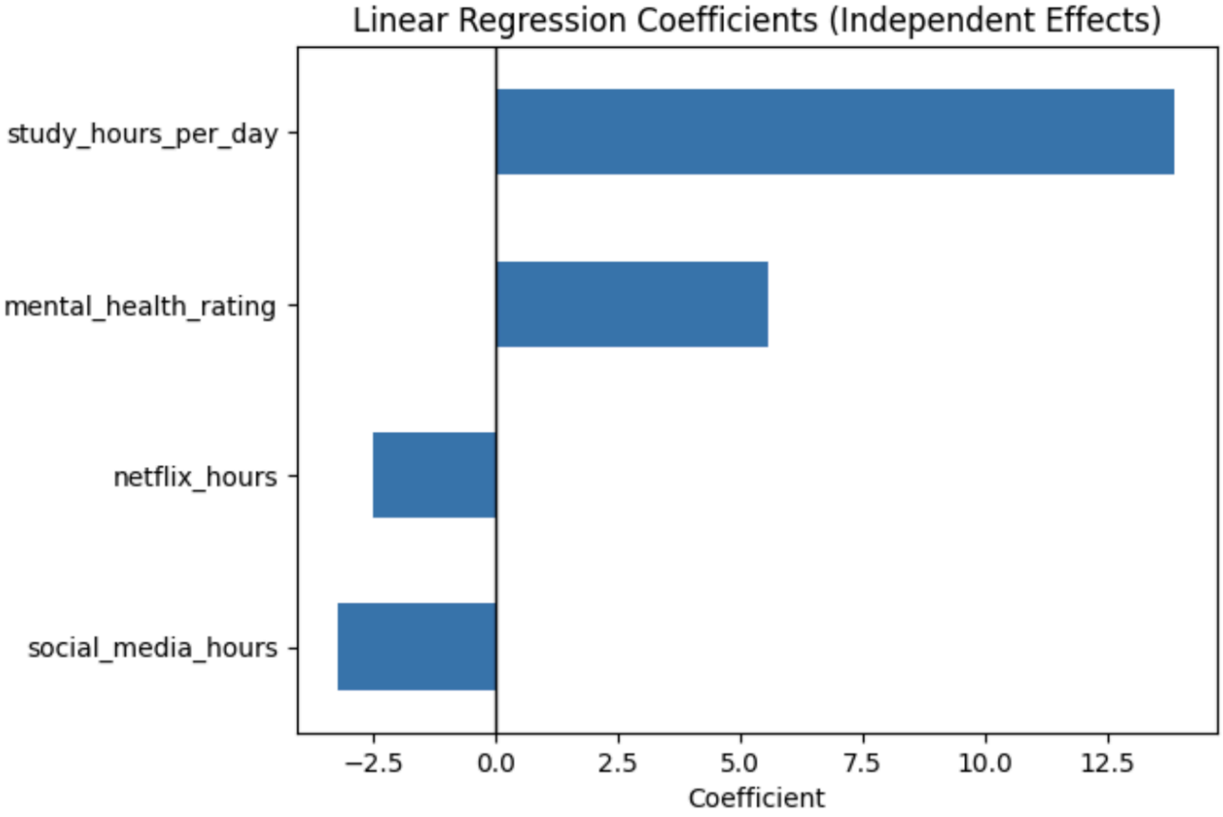
Secondly, a random forest regression was trained using the same selected features: study hours per day, netflix hours per day, mental health rating, and social media hours per day. The dataset was split into training and testing with 20% of the data used for testing. Feature scaling was applied for consistency across models, though it does not affect tree-based methods. To note, hyperparameters were not tuned.

MSE: 55.257
R^2: 0.784

These results suggest that the predominant relationships in the data are linear and increased model complexity actually lowered the accuracy of the model and did not improve overall performance. Therefore, the linear regression, although simpler, is a better fit to model this data.

TLDR: Continuing with linear regression.

Accounting for Confounding Variables Using Regression



From the standardized regression coefficients shown in the bar graph, when all other variables are held constant netflix hours and social media hours have weak negative independent effects on exam scores. Mental health demonstrates a moderate positive effect on exam score. In contrast, study hours per day demonstrates a strong independent effect on exam score after controlling for other predictors in the regression model. Although mental health, netflix hours and social media hours have weaker coefficients they are retained to control for confounding variables and capture additional trends.