Project intro: Discovering Optimal study habits

This project is about determining the correlation academic performance of a student based on their study habits and past academic history.
We would like to use this data todetermine an optimal correlation between all of these factors using Multiple Linear Regression.

The motivation for this project is dual intent. Of course, this is a study based. However, it does solve a real world, applied problem.
Finding the best study / life balances can be used to encourage certain routines and habits to enhance the performance of their peers and also predicting a students future results if they keep their habits.

The Data:

Data: For this project we are using a dataset provided by Kaggle.
(https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression?resource=download)

The data consists of 10000 rows worth of student records, with each record containing the following information:
Please note that the data is in one file with noauxiliary items.

Hours Studied: The total number of hours spent studying by each student.
Previous Scores: The scores obtained by students in previous tests.
Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
Sleep Hours: The average number of hours of sleep the student had per day.
Sample Question Papers Practiced: The number of sample question papers the student practiced.

Target Variable:
Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Projects steps:

1) Cleaning and adjusting the data
2) EDA
3) Creating the models and testing them
4) Conclusion

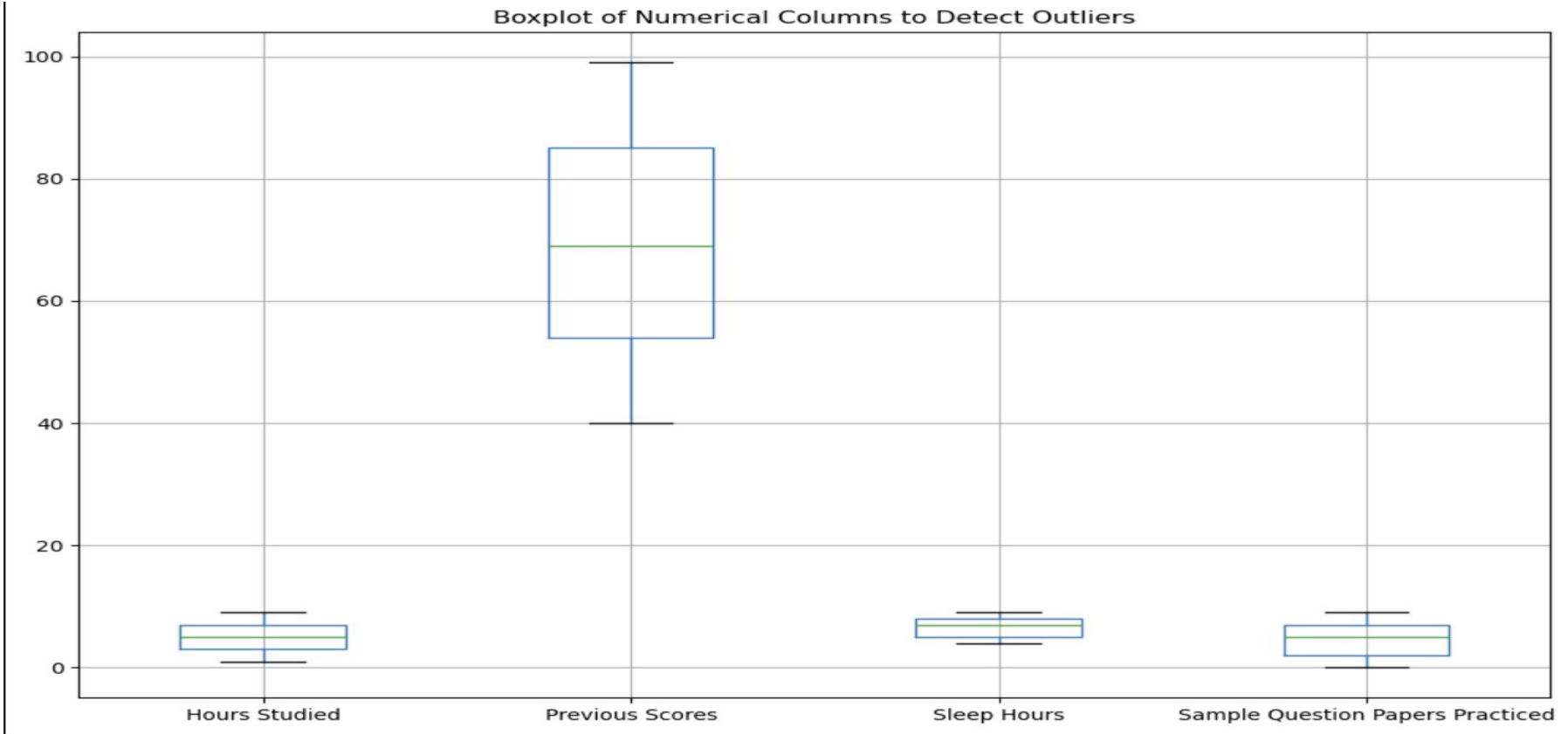
Data Cleaning and preparation:

1) Check for missing values.
2) Adjusting Data Types for better Semantics (The 'Extracurricular Activities' column's data type was explicitly set to 'category' to reflect its nature more accurately.)
3) Check for Duplicates
4) View Outliers (The Interquartile Range (IQR) method was employed to identify and remove outliers, ensuring data within reasonable limits.)
5) Summary and Visualization

Most important stats for the cleaning

1) 127 duplicate rows cleared
2) No missing values in the dataset

Boxplot of outliers



Exploratory Data Analysis

Steps performed

Summary Statistics: Calculated and displayed for all numerical features.

Target Variable Distribution: Visualized the distribution of 'Performance Index' using a histogram.

Correlation Analysis: Generated and analyzed a correlation matrix for numerical features and 'Performance Index', identifying strong correlations.

Visual Relationships: Visualized relationships using pair plots segmented by 'Extracurricular Activities'.

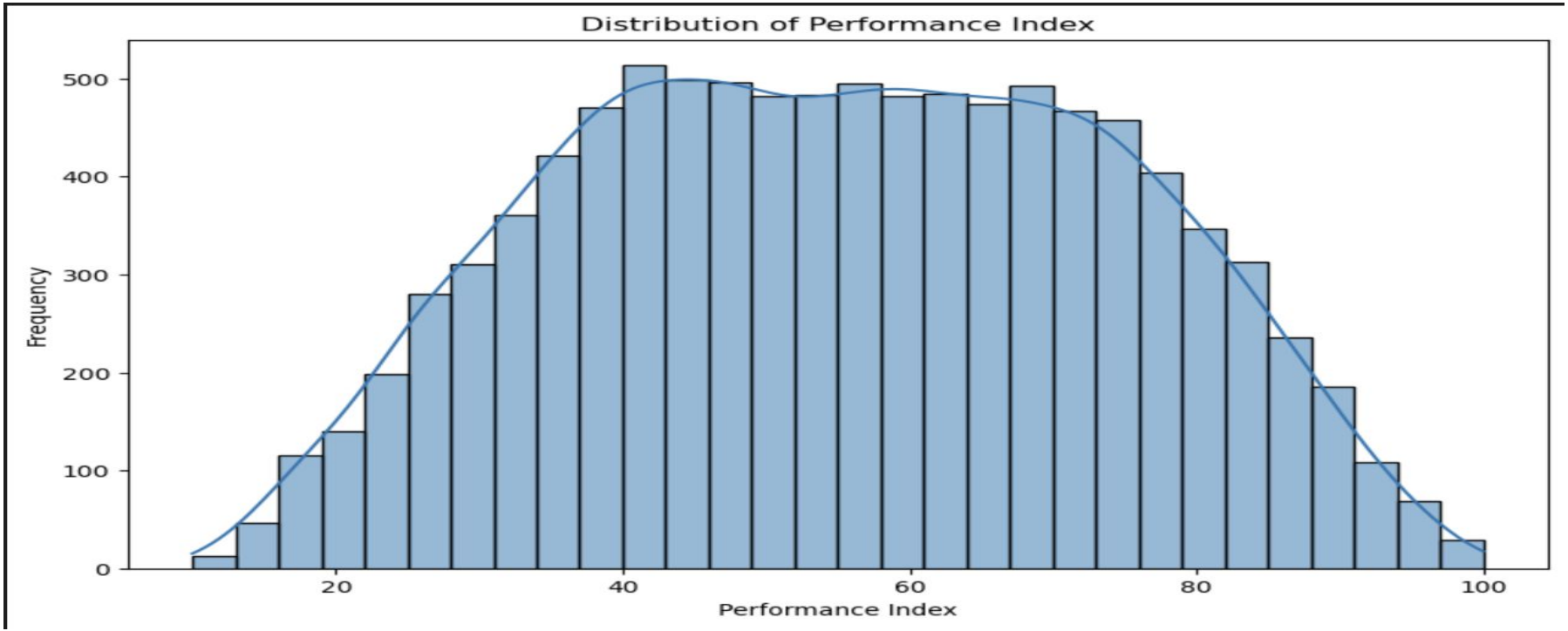
Statistical Tests:

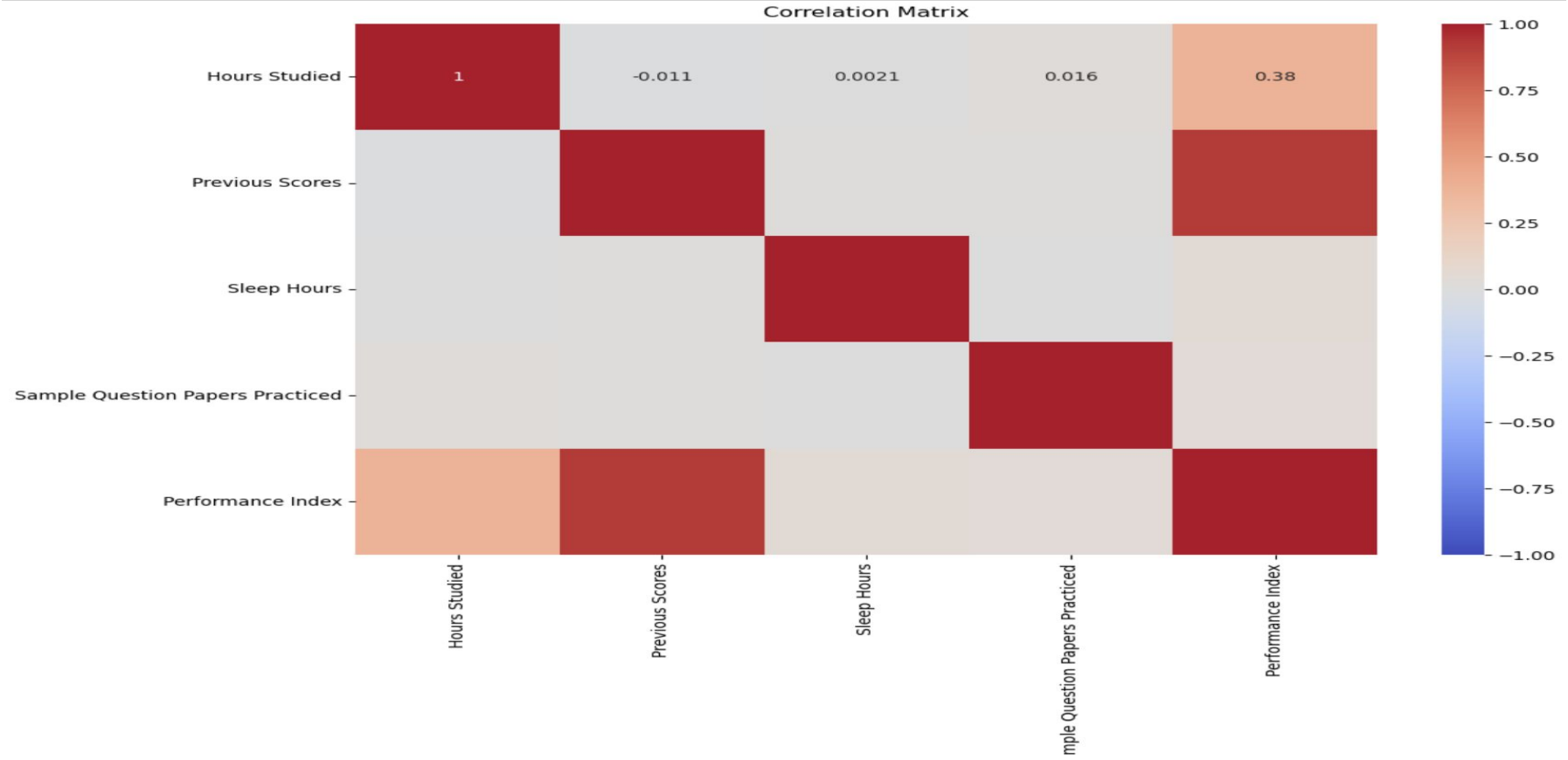
- Conducted a T-test to compare 'Hours Studied' between students involved in extracurricular activities versus those not involved.
- Performed a Chi-square test to analyze the relationship between 'Extracurricular Activities' and categorical 'Performance Index'.

Additional Visualizations:

- Visualized 'Performance Index' across categories of 'Extracurricular Activities'.
- Displayed box plots to explore 'Hours Studied' across different performance categories.

Performance index distribution




Correlation Matrix
Hours Studied
Previous Scores
Sleep Hours
Sample Question Papers Practiced
Performance Index
1
-0.011
0.0021
0.016
0.38
Hours Studied
Previous Scores
Sleep Hours
mple Question Papers Practiced
Performance Index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

EDA Conclusion

EDA Conclusion and interpretation:
let's go over the main features -

Study Hours:

Students study an average of about 5 hours per day, with a standard deviation of 2.59 hours. The range of study hours is between 1 to 9 hours.
Previous Scores:

The average previous score is approximately 69.44, with a standard deviation of 17.33. Scores range from 40 to 99, indicating a wide variation in past academic performance.
Sleep Hours:

On average, students get around 6.53 hours of sleep per day, with a standard deviation of 1.70 hours. Sleep hours range from 4 to 9 hours.
Sample Question Papers Practiced:

On average, students practice about 4.58 sample question papers, with a standard deviation of 2.87. The range is from 0 to 9 sample papers practiced.
Performance Index:

The Performance Index ranges from 10 to 100, with an average score of 55.22 and a standard deviation of 19.21. This indicates a diverse range of academic performance among the students.

Models

Feature Engineering: Categorized 'Performance Index' into 'Low', 'Medium', and 'High'.

Addressed Multicollinearity: Calculated VIF, removed 'Previous Scores' due to high VIF.

Prepared Data for Modeling: Defined target and features, encoded categorical variables.

Split Data: Divided data into training and testing sets.

Compared Models: Evaluated multiple models using cross-validation. (Linear Regression, Ridge, Lasso, and Random Forest).)

Tuned Hyperparameters: Conducted grid search for Ridge and Random Forest models.

Evaluated Final Model: Fitted and tested the best-performing model.

Model Result interpretation

Linear and Ridge regression models perform best with the lowest CV MSE, indicating a strong linear relationship between study habits and performance.
Best Ridge parameters: alpha = 10.
Best Random Forest parameters: max_depth = 10, min_samples_split = 5, n_estimators = 200.
Feature Importance:

From the Random Forest model, key features influencing performance include Hours Studied and Sample Question Papers Practiced.
Visual Insights:

Distribution of Performance Index: Shows how performance scores are spread out.
Boxplot by Extracurricular Activities: Indicates how extracurricular involvement impacts performance.
Pairplot: Visualizes relationships between study habits and performance.
Conclusion:
Focus on increasing Hours Studied and practicing more sample question papers while ensuring adequate sleep for optimal academic performance.

Conclusion

Discussion and conclusion:
Overall, i think this project is a solid demonstration of using Multilinear regression to find out which metrics are most impactful on the performance index.
It shows the applied usage of the skills learned in the course material.
For me the greatest takeaway from this experience was doing a project from scratch based on a real world scenario.

Things that I've tried but didn't include in the project. I have tried to include getting feature importance per each model used, however, i was unsuccessful as
i didn't manage to handle a python error thrown by the code. Unfortunately, it had to be omitted do to time constraints.

Overall, this project could be improved by adding in more data and better feature engineering. For example.

More Features: Collect data on other factors that might influence performance, such as nutrition, study environment, or parental involvement.
Longitudinal Data: Gather data over multiple semesters or years to track performance trends over time.

Feature Engineering:
Interaction Terms: Create new features that capture interactions between existing features, like Hours Studied * Sleep Hours.
Polynomial Features: Include polynomial terms to capture non-linear relationships.
Domain-Specific Features: Incorporate domain knowledge to engineer features that are more predictive of student performance.