By: Aryan Pradhan

Student Grades Capstone Project

Today's agenda

- Understand the dataset
- Analyze the variables in the dataset
- Understand relationships in the dataset
- Apply data-preprocessing
- Compare the performance of linear regression and random forest regression models on the dataset
- Limitations

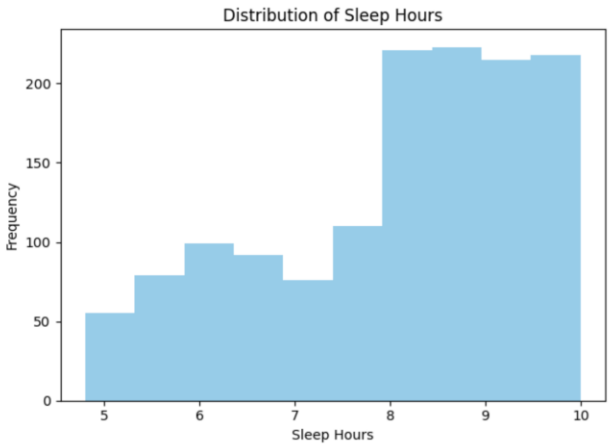
Understanding the dataset

- **Study Hours**
  - **Description:** Average daily hours spent studying.
- **Sleep Hours**
  - **Description:** Average daily hours spent sleeping.
- **Socioeconomic Score**
  - **Description:** A normalized score (0-1) indicating the student's socioeconomic background.
- **Attendance (%)**
  - **Description:** The percentage of classes attended by the student.
- **Grades** (TARGET)
  - **Description:** The final performance score of the student, derived from a combination of study hours, sleep hours, socioeconomic score, and attendance (out of 100).

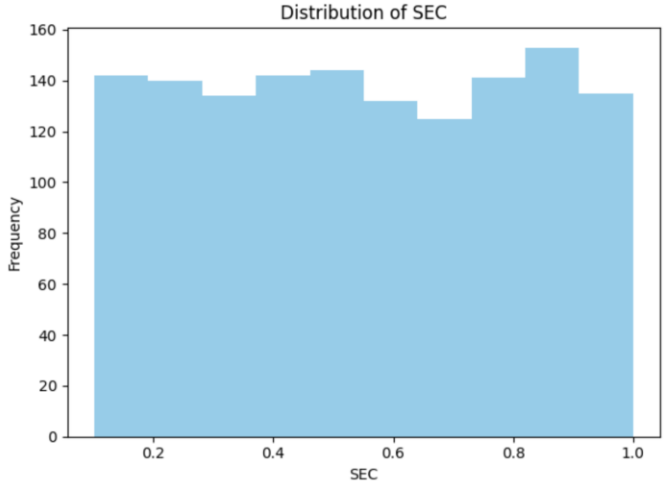
	Socioeconomic Score	Study Hours	Sleep Hours	Attendance (%)	Grades
0	0.95822	3.4	8.2	53.0	47.0
1	0.85566	3.2	5.9	55.0	35.0
2	0.68025	3.2	9.3	41.0	32.0
3	0.25936	3.2	8.2	47.0	34.0
4	0.60447	3.8	10.0	75.0	33.0

Snippet of the dataset (Candidate).

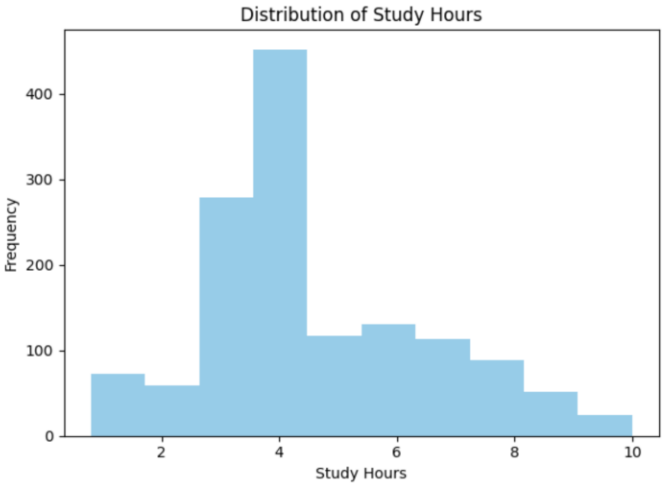
Analyzing the variables



Analyzing sleep hours spread:
•Distribution is **left-skewed**, centered around **8 to 10 hours/day**.
•Standard deviation is **low to moderate**, with most students getting sufficient sleep.
•Very few students sleep under **6 hours**, indicating overall healthy habits.
•Weak or slightly negative correlation with Grades — may suggest oversleeping correlates with underperformance or disengagement.

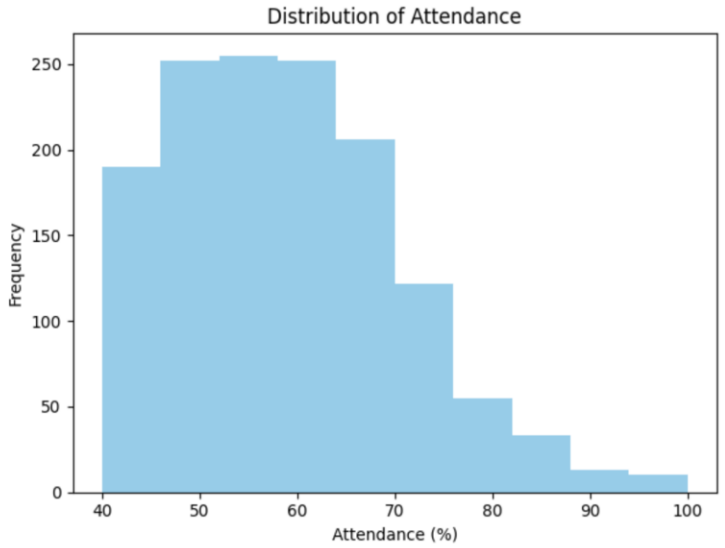


Analyzing SEC spread:
•Distribution is **nearly uniform**, ranging between 0 and 1.
•Standard deviation is likely **high** due to even spread across the full scale.
•Implies a **diverse socioeconomic background** among students.
•No clear skew; students come equally from all levels of the socioeconomic spectrum.
•Indicates that SEC won't dominate prediction just by volume — useful for balanced model learning.

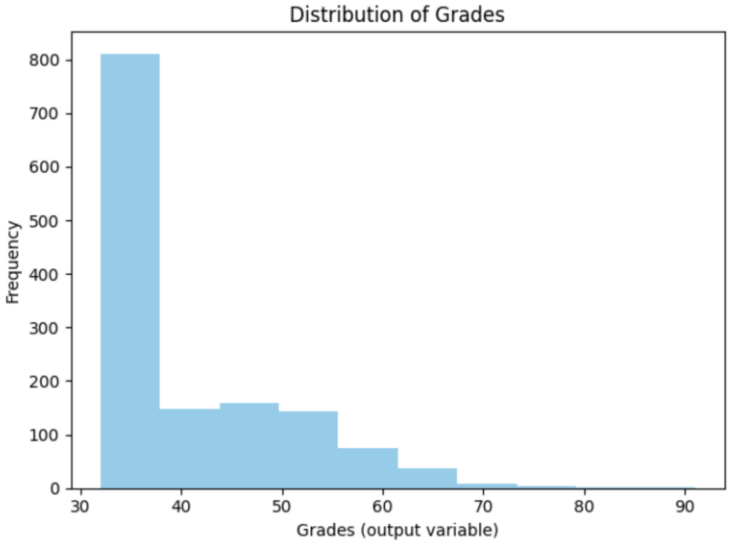


Analyzing study hours spread:
•Distribution is **right-skewed**, peaking around **4 hours/day**.
•Standard deviation is **moderate**, with some students studying up to 10 hours.
•Majority of students study between **3–5 hours**, forming a sharp mode.
•Long tail suggests a few outlier students with **very high dedication**.
•Strongest positive correlation with Grades — more study typically means better performance.

Analyzing the variables



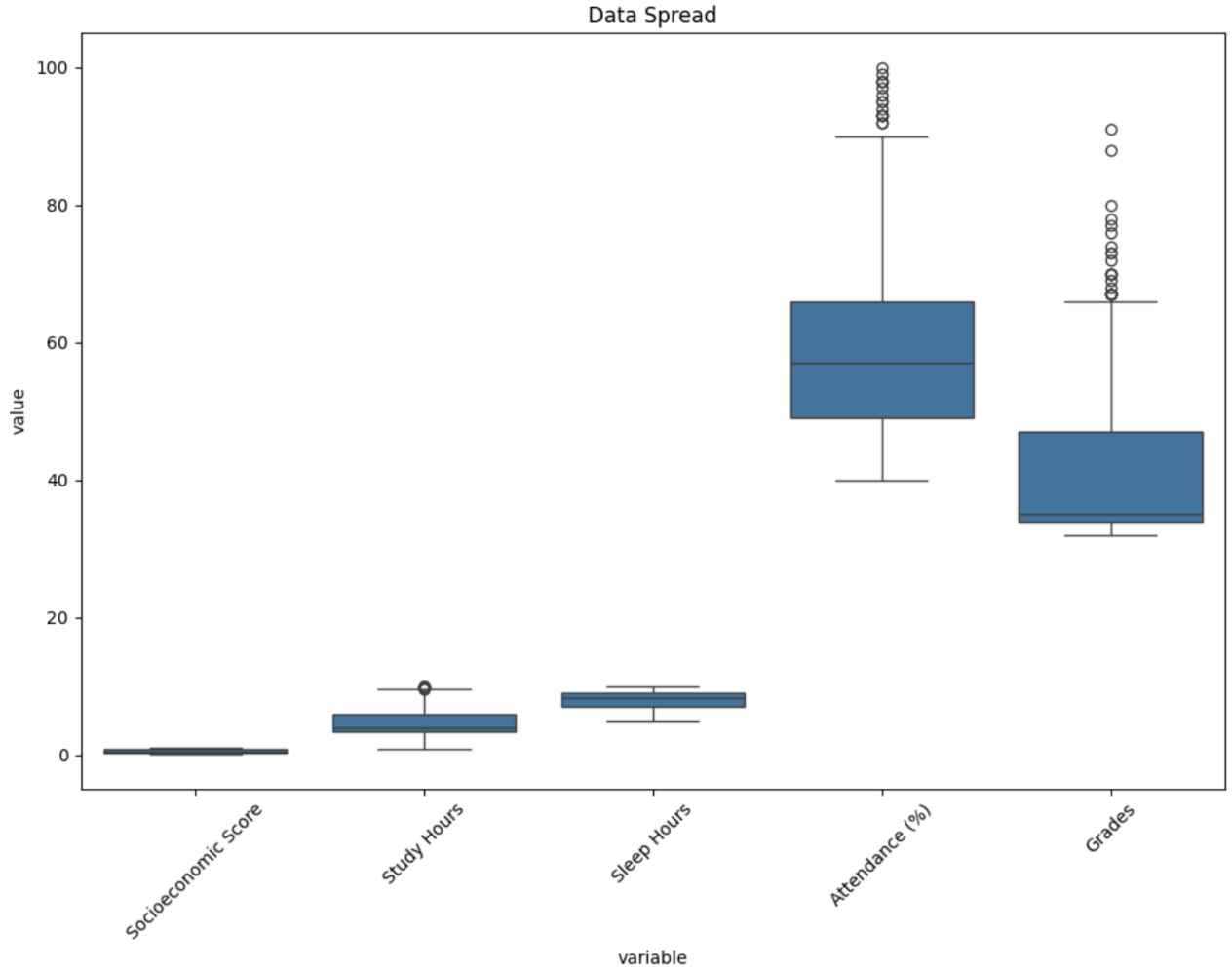
Analyzing attendance spread:
•Distribution is **left-skewed**, with a mode around **55–60%**.
•Standard deviation is **moderate to high**, with values ranging from **40% to nearly 100%**.
•Indicates a large number of students have **poor attendance**.
•Could be a key explanatory variable for low grades across the dataset.
•A small group of highly-attending students could act as **positive outliers** in performance.



Analyzing attendance spread:
•Distribution is **heavily right-skewed**, with the majority scoring **between 30–40**.
•Standard deviation is **high**, suggesting significant performance variance.
•Very few students score above **70**, indicating an overall **low-achieving population**.
•Tail toward higher grades likely contains those who studied consistently and attended regularly.
•Justifies the need for strong predictive models to identify **which inputs drive high grades**.

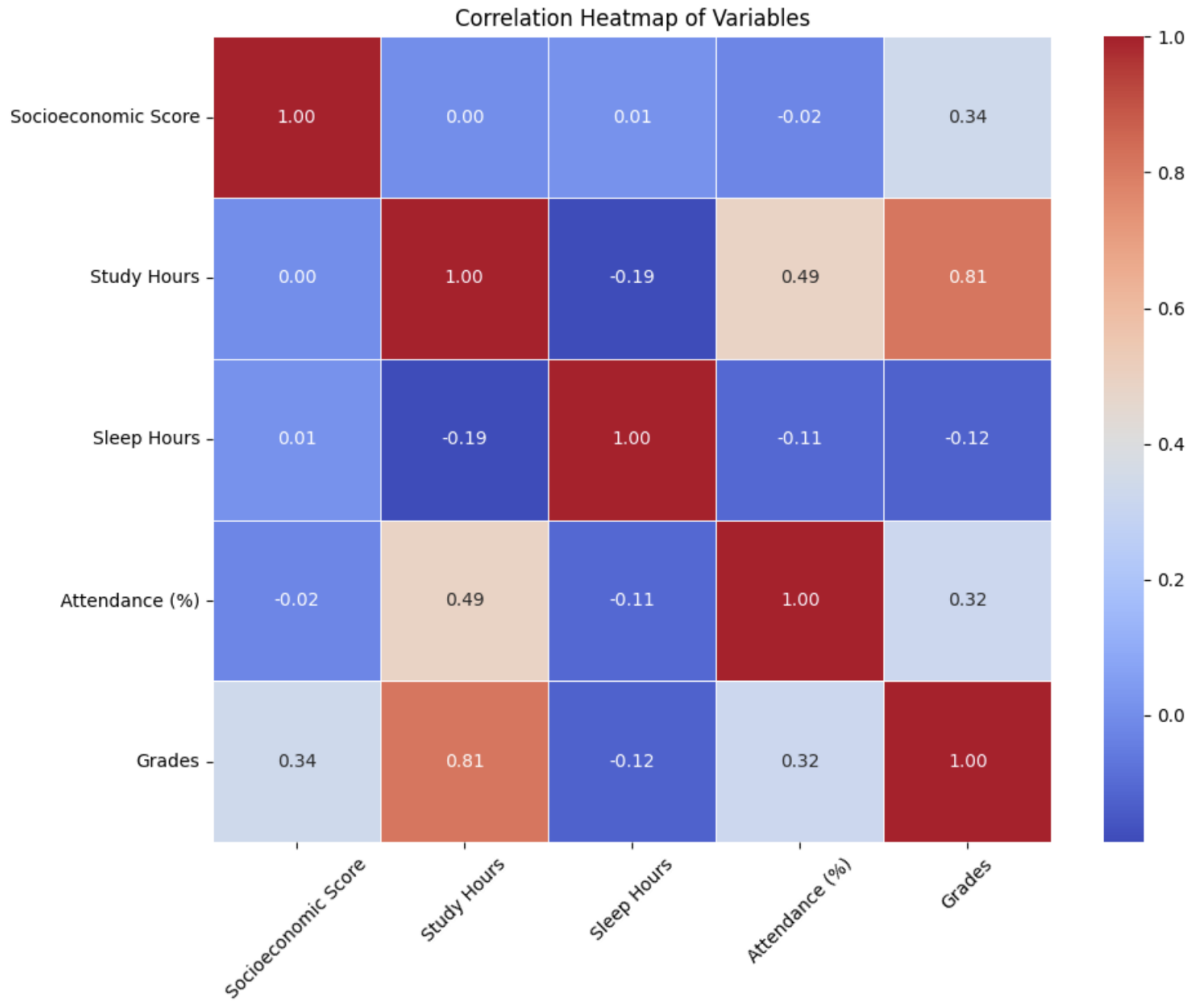
Understand relationships in the dataset

- Socioeconomic Score is tightly distributed with minimal outliers, suggesting it's fairly uniform across students and may not independently drive significant variation in grades.
- Study Hours shows a wide spread with several high-end outliers, indicating that some students invest significantly more time, and this variation could strongly influence academic performance.
- Sleep Hours is tightly clustered around 8–9 hours with low variability, implying that most students have similar sleep patterns, limiting its predictive power for distinguishing grade outcomes.
- Attendance displays a wide interquartile range and numerous high outliers near 100%, indicating variability in commitment and suggesting a strong potential relationship with grades.
- Grades themselves are right-skewed with a large concentration of students scoring between 30–40, while only a few achieve scores above 70, showing a long-tailed distribution.
- The overlapping spread and outlier patterns between Study Hours, Attendance, and Grades suggest that these input features are likely the most influential in predicting performance.
- In contrast, the low spread in Sleep Hours and Socioeconomic Score suggests these variables may only have marginal or indirect effects on academic outcomes.



Understand relationships in the dataset

- **Study Hours and Grades** show a very strong positive correlation (**0.81**), indicating that the more time students spend studying, the higher their grades tend to be.
- **Attendance and Study Hours** have a moderately strong positive correlation (**0.49**), suggesting that students who attend school more regularly also tend to study more.
- **Socioeconomic Score and Grades** show a weak positive correlation (**0.34**), implying that students from slightly higher socioeconomic backgrounds may perform better, though the effect isn't dominant.
- **Attendance and Grades** also show a moderate correlation (**0.32**), indicating a meaningful but not overwhelming impact of attendance on academic outcomes.
- **Sleep Hours** has a weak negative correlation with both **Grades (-0.12)** and **Study Hours (-0.19)**, possibly suggesting a trade-off between sleep and study time that doesn't necessarily lead to higher grades.
- **Socioeconomic Score** appears largely uncorrelated with **Study Hours (0.00)** and **Attendance (-0.02)**, indicating that it does not strongly influence how much students engage with their academic responsibilities.
- No extremely strong multicollinearity is observed except between **Study Hours and Grades**, which is expected given the context.



Apply data-preprocessing

Null detection

```
To check if there are any null values in the dataset
Socioeconomic Score    0
Study Hours            0
Sleep Hours            0
Attendance (%)         0
Grades                 0
dtype: int64

Any missing or blank values in the dataset?: False
```

I checked whether there were any null-values in the dataset by using a search code. The output represents that no null-values (meaning missing values) were reported in the dataset. The reported zeros exemplify that no missing values were reported in each variable. Also when analyzing the variables, as per the graph, there were no 0 values which makes sense.

Train-test split

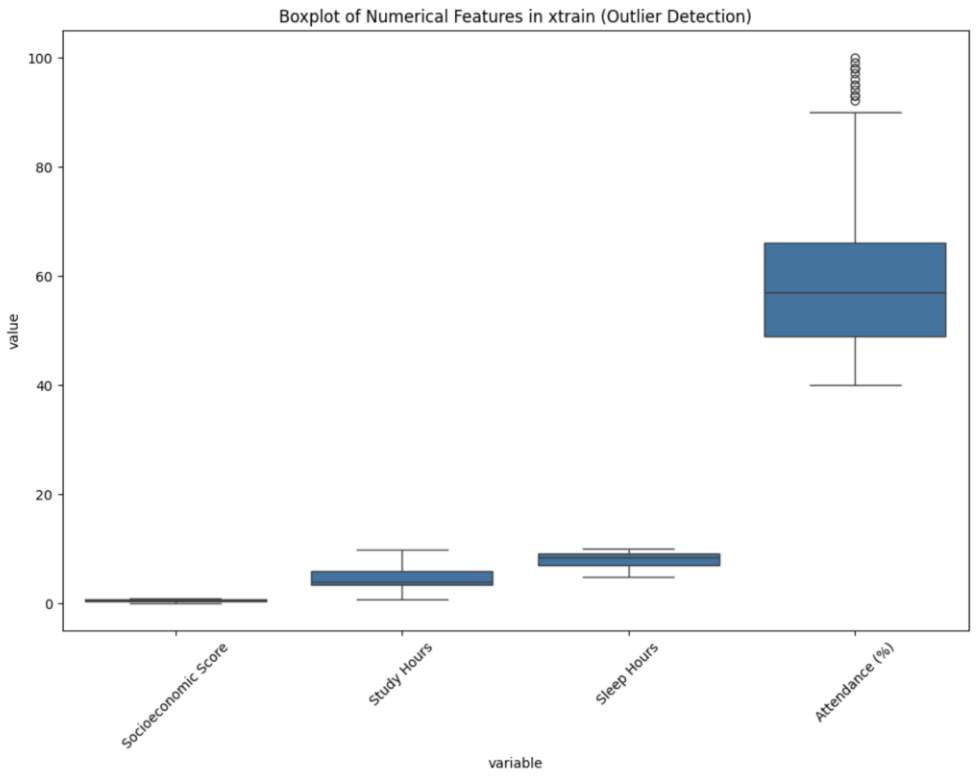
```
y = df["Grades"]
x = df.drop('Grades', axis=1)

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x , y, test_size = 0.2, random_state = 42)
```

For data preprocessing and further model implementation it is important to split the data into a train-test set. I made the test-size 20% of the dataset and 80% was the training data. I also implemented a random state of 42. Given the purpose of this capstone project, the random state value is not that important which is why I used a very common random state value of 42.

Apply data-preprocessing



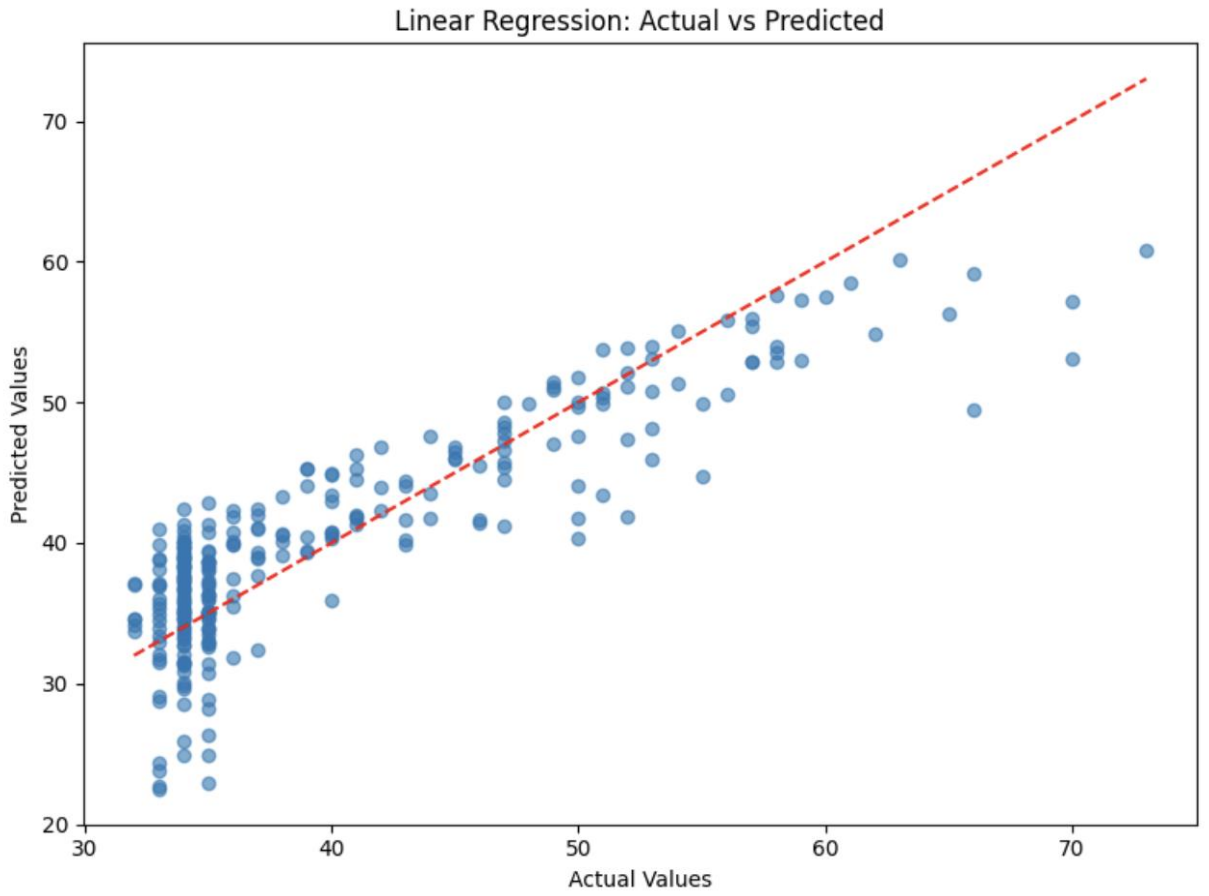

```
Q1 = x_train.quantile(0.25)
Q3 = x_train.quantile(0.75)
IQR = Q3 - Q1
mask = ~((x_train < (Q1 - 1.5 * IQR)) | (x_train > (Q3 + 1.5 * IQR))).any(axis=1)

x_train_no_outliers = x_train[mask]
y_train_no_outliers = y_train[mask]
```

As shown by the graph, there were outliers in the dataset. Thus, to optimize the model training in sequent events, I eliminated outliers from the dataset. Furthermore, to reach conclusions about the grades dataset, it is important to not consider the outliers.

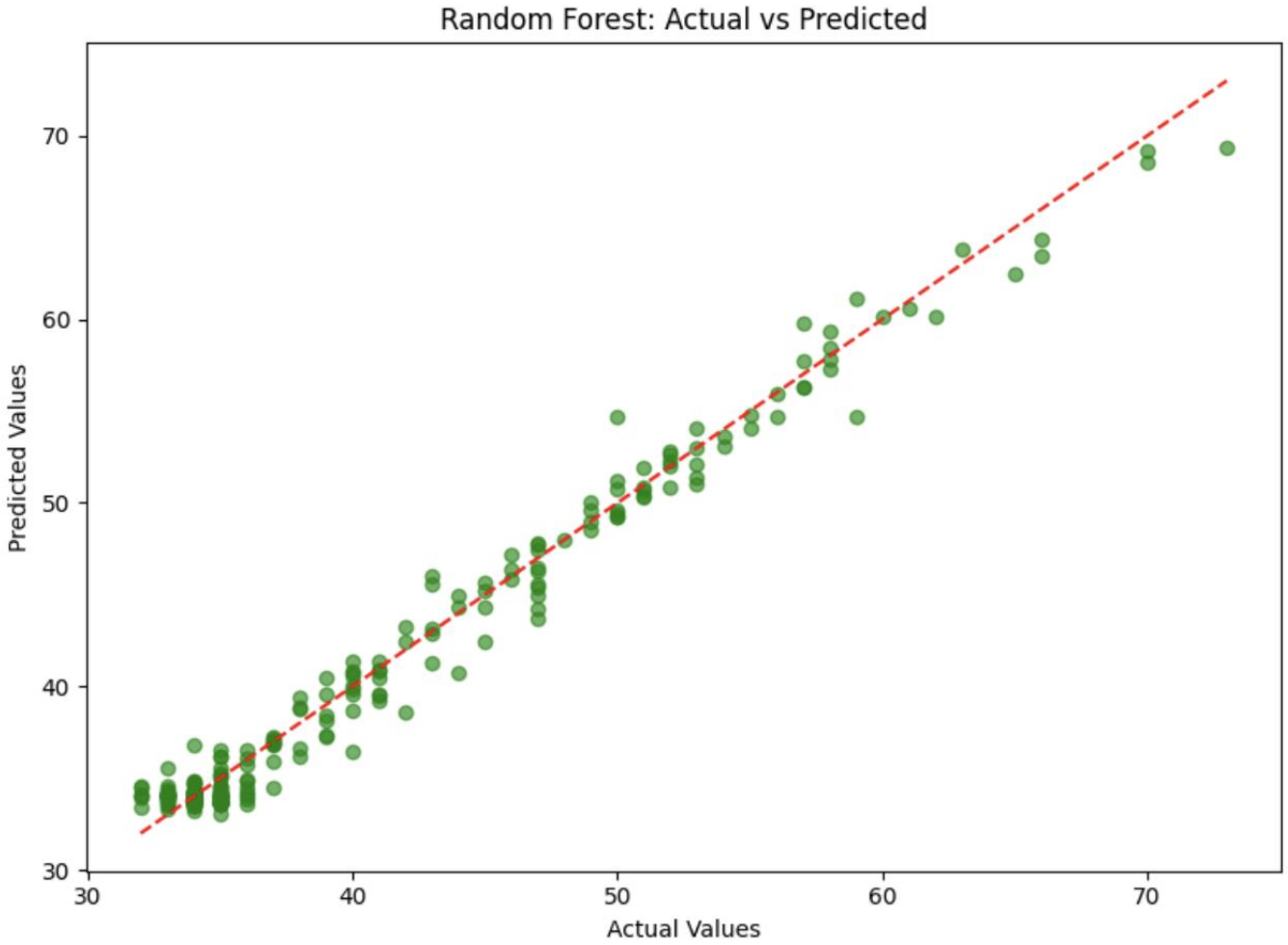
Linear Regression

I applied the masked training set for the linear regression's training data. I did not conduct any hyper-parameter tuning. The graph represents the linear relationship that the model fitted. The dispersion between the actual and predicted values are relatively spaced out. This suggests that the linear model was not able to work well enough with the dataset.

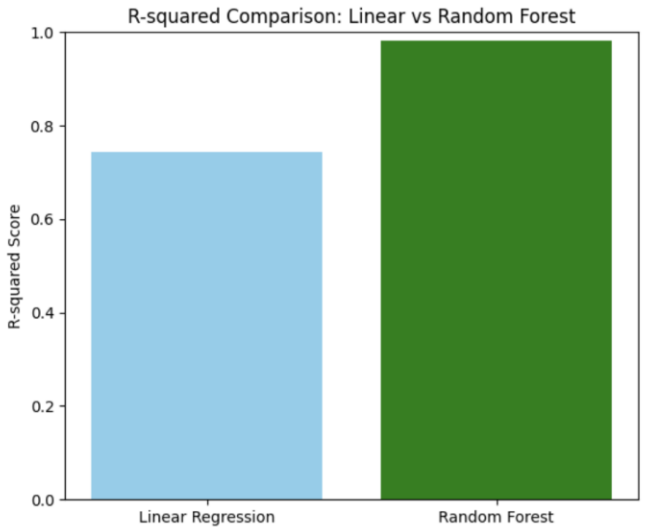


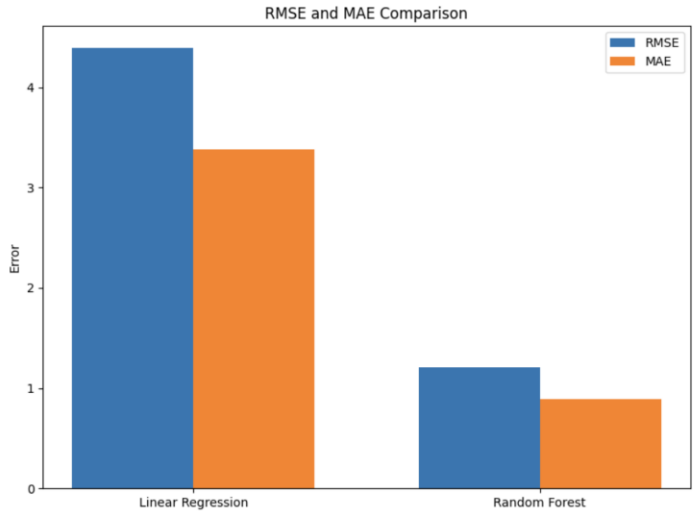
Random Forest Regressor

I applied the masked training set for the random forest regressor's training data. I did not conduct any hyper-parameter tuning. The graph represents the linear relationship that the model fitted. The dispersion between the actual and predicted values are relatively pretty close. This suggests that the random forest model was able to work well with the dataset.



Comparative results





Model Evaluation Comparison: Linear vs Random Forest

Model	RMSE	MAE	R-squared Score
Linear Regression	4.390	3.380	0.745
Random Forest	1.204	0.888	0.981

The random forest regressor had a significantly greater R-squared coefficient which illustrates how its accuracy was better compared to that of the linear model for this dataset. Moving on to the RMSE and MAE comparison, the linear model had a greater RMSE and MAE compared to random forest. This encapsulates how the linear model had a significantly greater error margin when predicting; the linear model's error margin was also captured from the actual vs predicted graph. For a numerical comparison, the random forest model had better statistical outcomes. Henceforth, to generalize, a random forest might prove to work better with student performance datasets for grade prediction or optimization in the future. However, one must consider that these results represent the performance for random forests specifically for this dataset.

Limitations

- When analyzing the variables it is apparent that the grades are highly right-skewed with a very high standard deviation. This might not be an accurate representative of the average school in almost any country. Thus, for further exploration one must select a dataset that comes from an actual school rather than a generated dataset found in Kaggle. This might be the same case when it comes to the spread of the attendance variable.
- When considering the two models, linear regression and random forest, I did not apply hyper-parameter tuning by passing test values and optimizing the models. Hence, the results for the models are based on randomized states of the two models. Further exploration can consider to optimize the models before testing them.
- There was an anomalous but very strong relationship between the input variable study hours with the output grades as compared to any other input variable. This suggests biased data and for further exploration one could apply weightages in the data-preprocessing to the variables.