2704 Final Project Presentation

Jackson McIntyre, Nicolas Serrano and Taryn Cail

Background

Students' academic success (GPA) is linked to various features such as the amount of studying they do, their sleep schedules, eating habits and exercise patterns. But what feature affects their performance the most? We believe it is studying habits.

Hypothesis

Null Hypothesis: There is no correlation between students' amount of studying and their academic performance.

Alternative Hypothesis: The higher the amount of studying, the higher the student's academic performance will be.

We wish to prove the Alternative Hypothesis, that the more studying a student performs the better they will achieve academically.

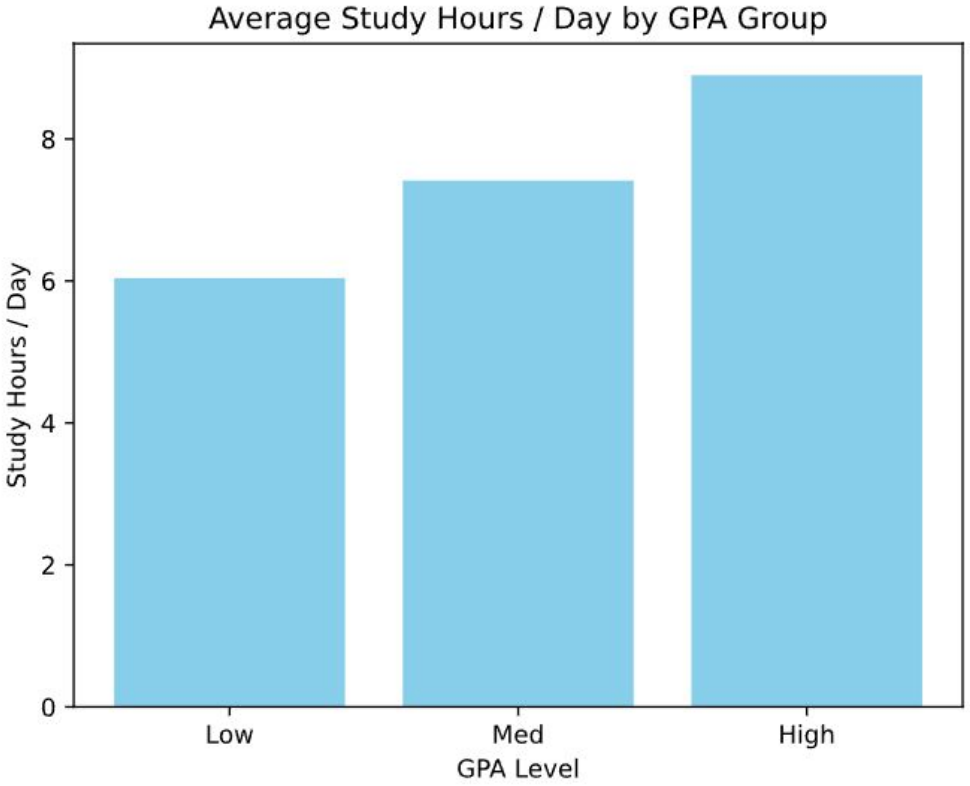
The Dataset

The dataset we will be using was found on Kaggle, an open-source data website.

The dataset contains 2000 rows and 8 columns. It includes various data including student ID's, study hours, extracurricular hours, sleep hours, social hours, physical activity hours per day as well as GPA levels and stress levels.

Student_ID	Study_Hours_Per_Day	Extracurricular_Hours_Per_Day	Sleep_Hours_Per_Day	Social_Hours_Per_Day	Physical_Activity_Hours_Per_Day	GPA	Stress_Level
1	6.9	3.8	8.7	2.8	1.8	2.99	Moderate
2	5.3	3.5	8	4.2	3	2.75	Low
3	5.1	3.9	9.2	1.2	4.6	2.67	Low
4	6.5	2.1	7.2	1.7	6.5	2.88	Moderate
5	8.1	0.6	6.5	2.2	6.6	3.51	High
6	6	2.1	8	0.3	7.6	2.85	Moderate
7	8	0.7	5.3	5.7	4.3	3.08	High
8	8.4	1.8	5.6	3	5.2	3.2	High
9	5.2	3.6	6.3	4	4.9	2.82	Low

Bar Graph

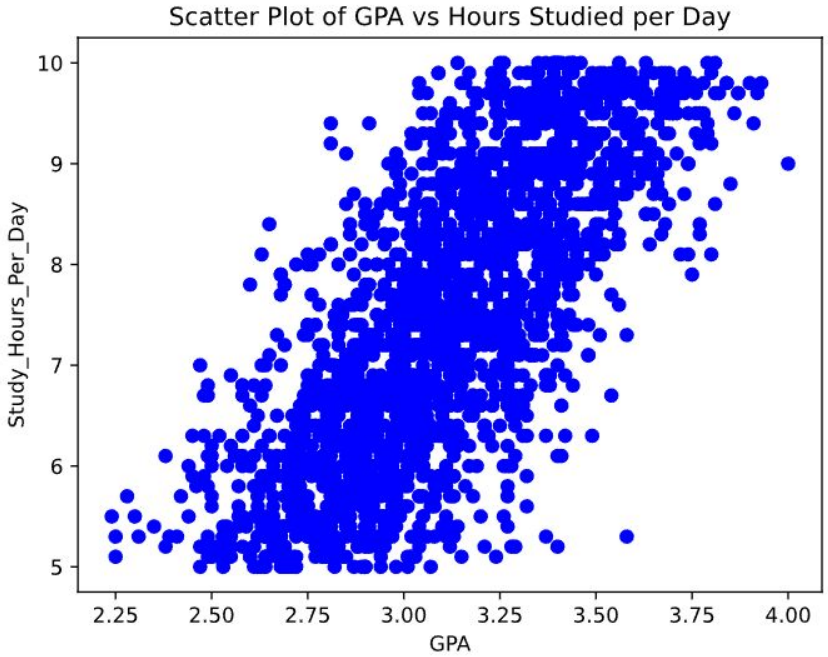


This bar graph further illustrates the relationship between a higher GPA and the amount of hours studied per day. This is evident from the larger bar in the ‘High’ GPA section compared to the ‘Low’ GPA section, indicating that students with a higher GPA score tend to spend more time studying.

Code Behind the Bar Graph

```
#Visualize - bar graph
plt.clf()
plt.bar(gpa_by_studyH['GPA Level'], gpa_by_studyH['Study_Hours_Per_Day'], color='skyblue')
plt.xlabel("GPA Level")
plt.ylabel("Study Hours / Day")
plt.title("Average Study Hours / Day by GPA Group")
plt.savefig("bar_graph.svg", format="svg", bbox_inches="tight")
plt.show()
```

Scatter Plot

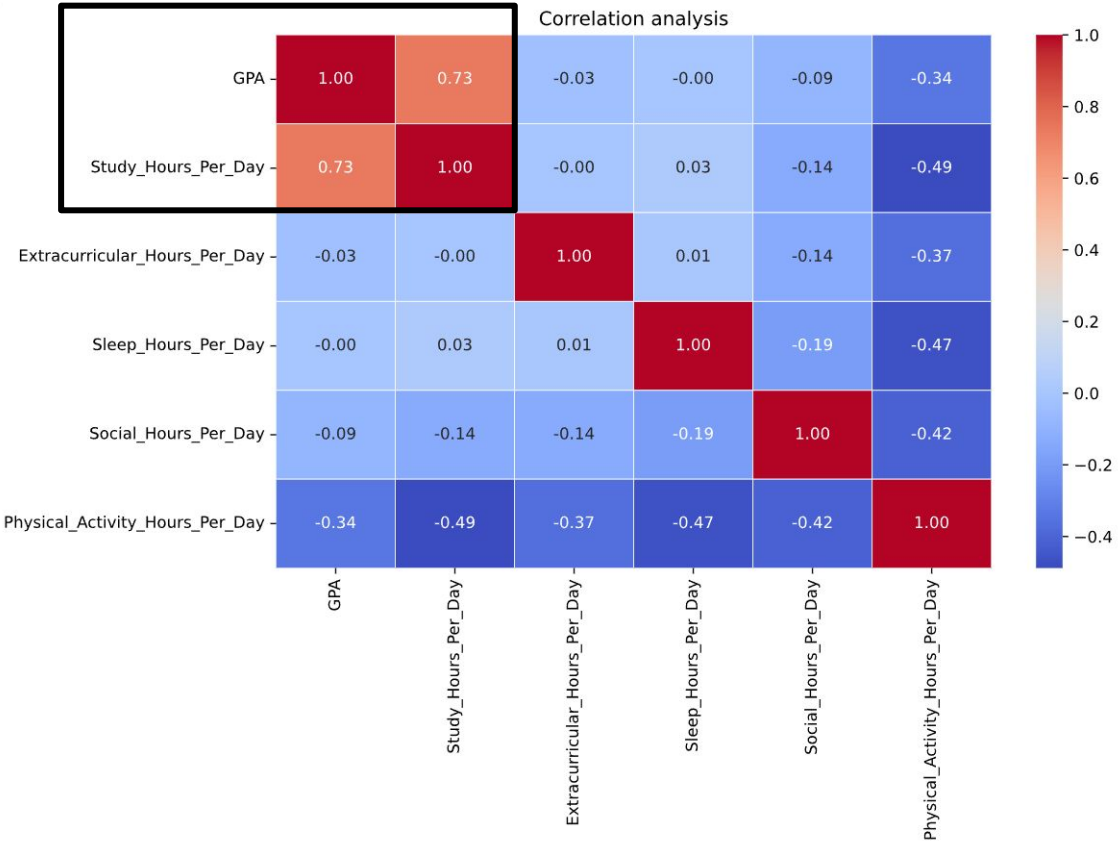


As the Scatter Plot demonstrates, there is a clear trend where higher GPA levels are associated with more study hours and vice versa. This indicates that students who spend time studying may see better academic performance compared to their counterparts.

Code Behind the Scatter Plot

```
#Visualize - scatter plot
plt.clf()
plt.scatter(df['GPA'], df['Study_Hours_Per_Day'], color='blue', marker='o')
plt.xlabel('GPA')
plt.ylabel('Study_Hours_Per_Day')
plt.title('Scatter Plot of GPA vs Hours Studied per Day')
plt.savefig("scatter_plot.svg", format="svg", bbox_inches="tight")
plt.plot()
```

Heatmap



The heatmap illustrates that the closer the correlation value is to +1, the stronger the relationship. For instance, the correlation between GPA levels and Study_Hours_Per_Day is strong, with a value of 0.73. This suggests a positive relationship where an increase to one variable leads to an increase in the other. In contrast, the other variables do not exhibit a strong positive correlation or a negative one.

Code Behind the Heatmap

```
#Visualize - heatmap
correltaion_data = df[['GPA','Study_Hours_Per_Day', 'Extracurricular_Hours_Per_Day', 'Sleep_Hours_Per_Day', 'Social_Hours
corr_matrix = correltaion_data.corr()
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Correlation analysis")
plt.savefig("heatmap.svg", format="svg", bbox_inches="tight")
plt.show()
```

P Value Test and Pearson Correlation Test

We performed two tests to decide whether we are able to accept or reject our hypothesis, these were the P-Value test and the Pearson Correlation Test.

For the P-Value Test, a p-value less than 0.05 indicates that the correlation is statistically significant.

For the Pearson Correlation Test, the coefficient ‘r’ indicates a positive or negative relationship between the two variables. Specifically an ‘r’ value greater than 0.7 suggests a strong positive linear relationship.

Based on our results:

P-Value = ~0

Pearson Correlation Coefficient = 0.7345

We can confidently reject the null hypothesis and conclude that an increase in study hours is associated with a higher GPA.

Code Behind the P Value Test and Pearson Correlation Test

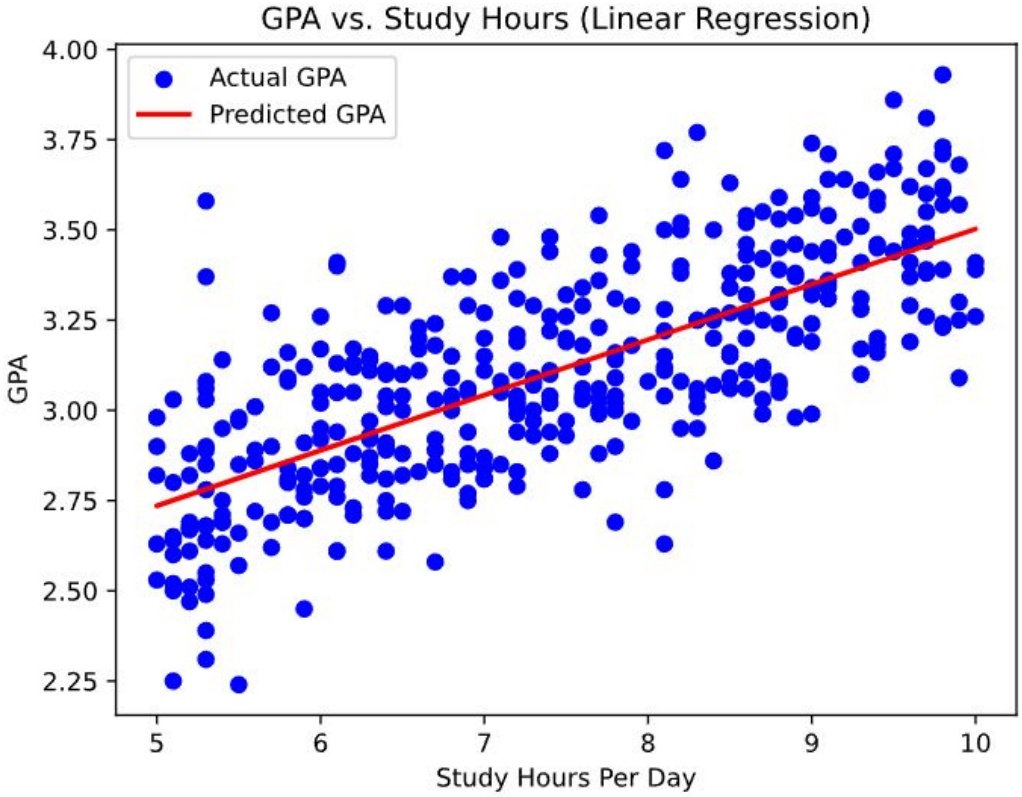
```
# Perform Pearson correlation test
r, p_value = stats.pearsonr(df['Study_Hours_Per_Day'], df['GPA'])

# Print Results
print(f"Pearson correlation coefficient: {r:.4f}")
print(f"P-value: {p_value:.4f}")
```

Predictive Analytics

Response Variable: GPA levels (outcome we want to predict)

Dependent Variable: Study Hours per day (input that affects the outcome)

Simple Linear Regression Model



Testing our Linear Regression Model with our Own Values

```
Predicted GPA for 0 hours/day: 1.97
Predicted GPA for 2 hours/day: 2.28
Predicted GPA for 4 hours/day: 2.58
Predicted GPA for 6 hours/day: 2.89
Predicted GPA for 8 hours/day: 3.20
Predicted GPA for 10 hours/day: 3.50
```

```
Mean Absolute Error : 0.58
```

Linear Discriminant Analysis

Accuracy: 0.67

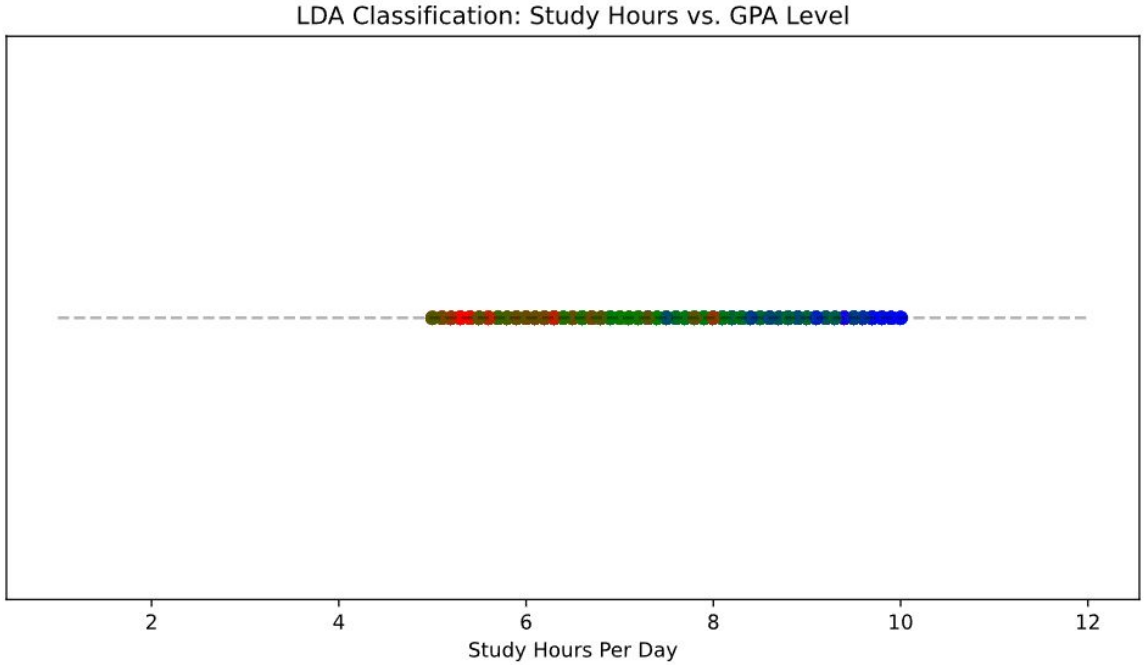
Classification Report:

	precision	recall	f1-score	support
High	0.66	0.59	0.62	94
Low	0.65	0.51	0.57	88
Med	0.68	0.77	0.72	218
accuracy			0.67	400
macro avg	0.66	0.62	0.64	400
weighted avg	0.67	0.67	0.67	400

Testing our Linear Discriminant Analysis Model with our Own Values

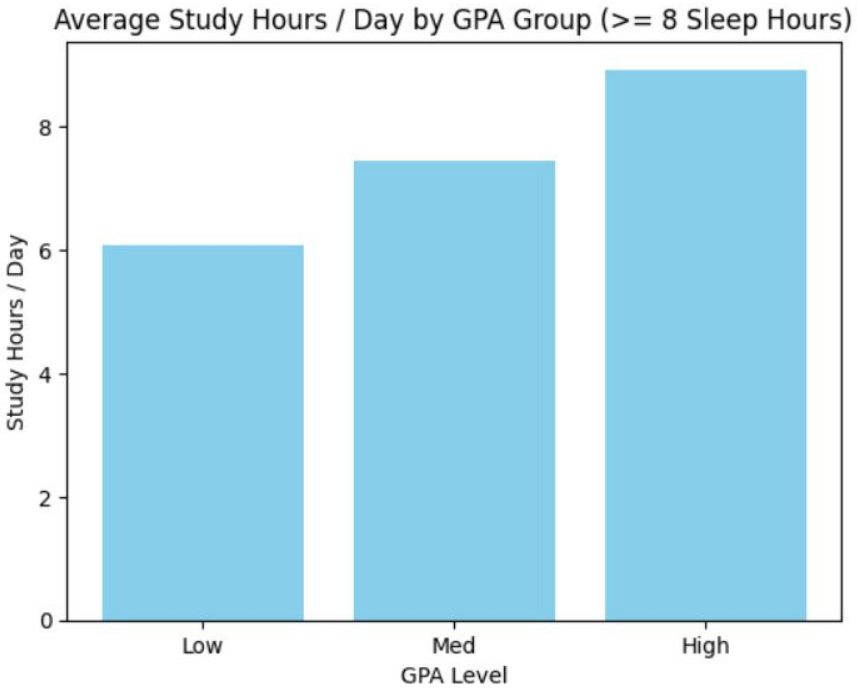
```
For 0 study hours/day:
Predicted GPA Level: Low

For 2 study hours/day:
Predicted GPA Level: Low

For 4 study hours/day:
Predicted GPA Level: Low

For 6 study hours/day:
Predicted GPA Level: Med

For 8 study hours/day:
Predicted GPA Level: Med

For 10 study hours/day:
Predicted GPA Level: High
```

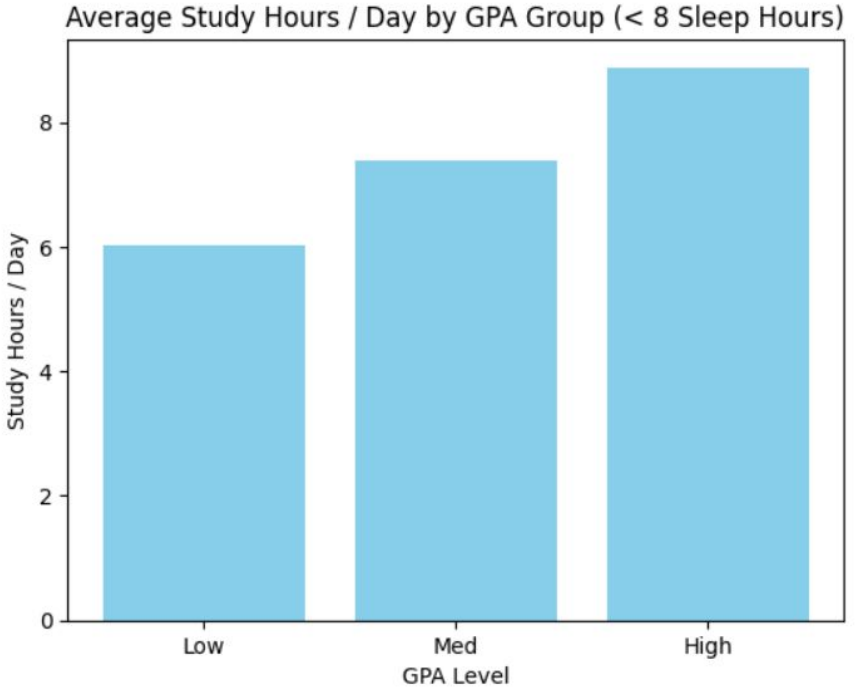
Visualizing the Data

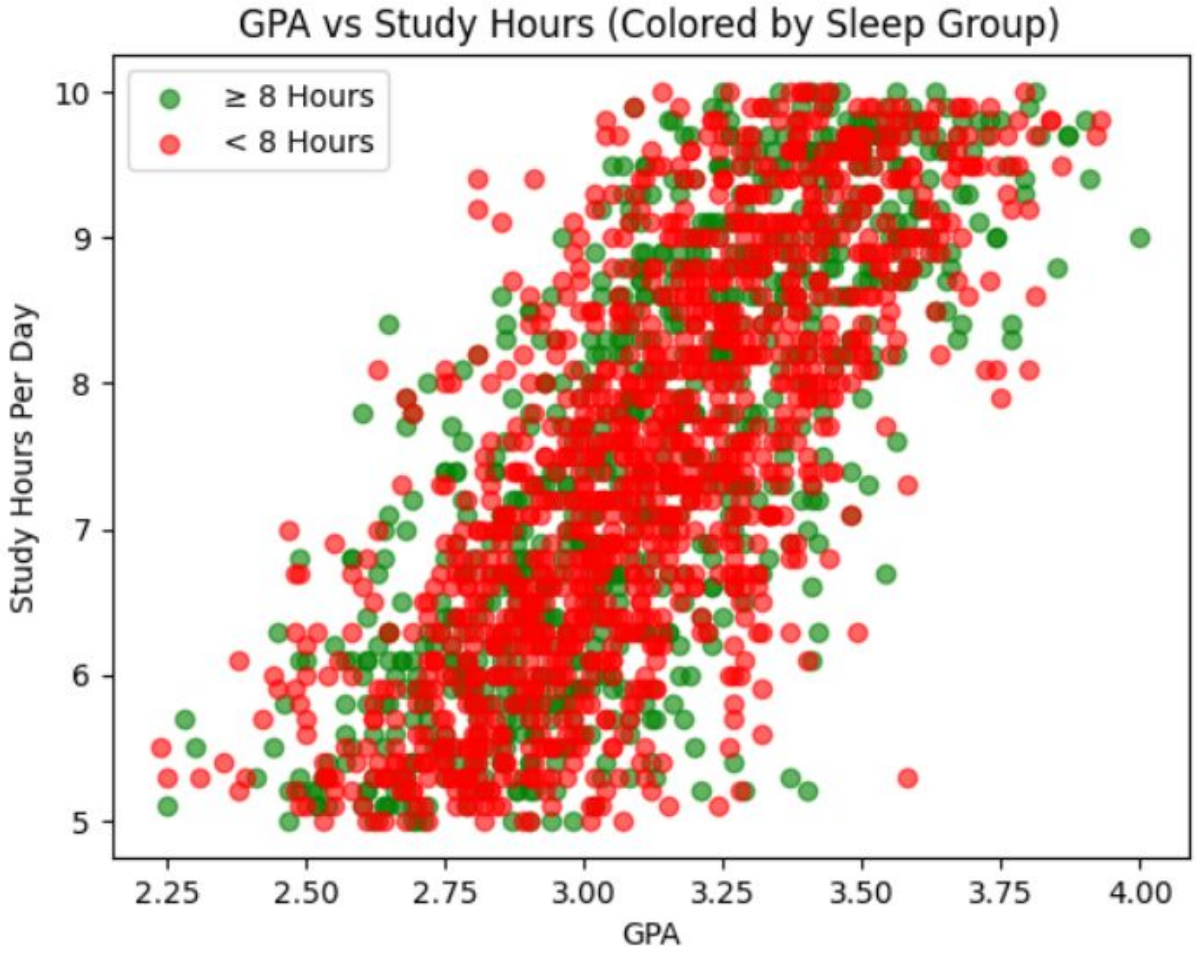


The heavy overlap shows that study hours per day may not be the sole determinant for GPA levels.

Testing whether Sleep Affects GPA and Study Hours






GPA vs Study Hours (Colored by Sleep Group)
≥ 8 Hours
< 8 Hours
Study Hours Per Day
10
9
8
7
6
5
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
GPA

Discussion

The use of graphs was effective in visualizing trends, while the heatmap provided valuable insights into correlations between variables. Additionally, the linear regression model helped illustrate patterns of change and growth, making the relationship between study hours and GPA more interpretable.

One major challenge was interpreting the p-value, as it produced unexpected results and ultimately equaled zero. This suggests that one or more variables may have had minimal variation and a low standard deviation, potentially affecting the reliability of the statistical analysis.

Further Research

To enhance the study's accuracy and reliability, future work should focus on expanding the dataset by collecting more diverse and extensive data. This would help improve statistical robustness and provide a more comprehensive understanding of the relationship between study hours and GPA.

Conclusion

A correlation coefficient (r) greater than 0.7 indicates a strong positive relationship between study hours per day and GPA. Additionally, a p-value below 0.05 signifies that the correlation is statistically significant.

Our analysis yielded an (r) value equal to 0.7345 and our p-value near 0.

Therefore, we can confidently reject the null hypothesis and accept the alternative hypothesis, concluding that there is a statistically significant positive correlation between study hours per day and GPA level.