**CY2004**

**Probability & Statistics**

**Final Project**

**Lifestyle Factors & their Impact on Grades (GPA)**

**Submitted by:** Abdul Ahad, M. Faizan, Waleed Ahmed Zai

**Roll number:** 23i-2014, 23i-2074, 23i-2128

**Date:** 04/05/2025

Table of Contents

[Introduction: 5](#_Toc197256449)

[Select Secondary Dataset: 5](#_Toc197256450)

[Descriptive Statistics: 5](#_Toc197256451)

[Clean Data Set Summary: 6](#_Toc197256452)

[Unclean Data Set Summary: 6](#_Toc197256453)

[Tabular Analysis (Clean and Unclean dataset): 8](#_Toc197256454)

[Descriptive Statistical Analysis: Clean vs. Unclean Dataset 10](#_Toc197256455)

[1. Student\_ID 11](#_Toc197256456)

[2. Study\_Hours\_Per\_Day 11](#_Toc197256457)

[3. Extracurricular\_Hours\_Per\_Day 11](#_Toc197256458)

[4. Sleep\_Hours\_Per\_Day 11](#_Toc197256459)

[5. Social\_Hours\_Per\_Day 11](#_Toc197256460)

[6. Physical\_Activity\_Hours\_Per\_Day 12](#_Toc197256461)

[7. Stress\_Level 12](#_Toc197256462)

[8. Gender 12](#_Toc197256463)

[9. Grades 12](#_Toc197256464)

[Final Remarks: 12](#_Toc197256465)

[Data Preprocessing: 13](#_Toc197256466)

[Step 1: Fixing Wrong Data Types 13](#_Toc197256467)

[Step 2: Handling Missing Values 13](#_Toc197256468)

[Step 3: Removing Duplicates 14](#_Toc197256469)

[Step 4: Fixing Categorical Mess 14](#_Toc197256470)

[Step 5: Outliers and Invalids 15](#_Toc197256471)

[Step 6: Saved the Clean Data 16](#_Toc197256472)

[Step 7: Feature Scaling: 16](#_Toc197256473)

[Final Check with Summary 17](#_Toc197256474)

[Exploratory Data Analysis (EDA): 17](#_Toc197256475)

[1. Grade Distribution x and y: 17](#_Toc197256476)

[Inference: 17](#_Toc197256477)

[2. Boxplot: Grades by Gender 17](#_Toc197256478)

[3. Boxplot: Grades by Stress Level 18](#_Toc197256479)

[4. Study Hours vs Grades: 19](#_Toc197256480)

[Inference: 19](#_Toc197256481)

[5. Sleep Hours vs Grades: 19](#_Toc197256482)

[Inference: 20](#_Toc197256483)

[6. Social Hours vs Grades: 20](#_Toc197256484)

[Inference: 20](#_Toc197256485)

[7. Physical Activity vs Grades: 20](#_Toc197256486)

[Inference: 21](#_Toc197256487)

[8. Extracurricular Hours vs Grades: 21](#_Toc197256488)

[9. Stress Level by Gender: 22](#_Toc197256489)

[10. Study Hours by Stress Level: 22](#_Toc197256490)

[Inference: 22](#_Toc197256491)

[11. Sleep Hours by Stress Level: 23](#_Toc197256492)

[Inference: 23](#_Toc197256493)

[12. Physical Activity by Gender: 23](#_Toc197256494)

[Inference: 24](#_Toc197256495)

[13. Social Hours by Gender: 24](#_Toc197256496)

[14. Combined Activity Time vs Grades: 25](#_Toc197256497)

[Inference: 25](#_Toc197256498)

[15. Density Plots: 25](#_Toc197256499)

[Top-Left: Study Hours Per Day 26](#_Toc197256500)

[Top-Right: Sleep Hours Per Day 26](#_Toc197256501)

[Bottom-Left: Social Hours Per Day 26](#_Toc197256502)

[Bottom-Right: Extracurricular Hours Per Day 26](#_Toc197256503)

[Overall Insight: 26](#_Toc197256504)

[16. Violin Plot for Gender vs Grades: 26](#_Toc197256505)

[Summary: 27](#_Toc197256506)

[17. Gender Distribution Pie chart: 27](#_Toc197256507)

[Summary: 27](#_Toc197256508)

[18. Scatterplot matrix: 28](#_Toc197256509)

[Summary: 28](#_Toc197256510)

[Inference: 29](#_Toc197256511)

[Interactive Visualization: 29](#_Toc197256512)

[Statistical Modeling and Inference: 29](#_Toc197256513)

[1. Correlation Analysis Heat Map: 30](#_Toc197256514)

[Summary: 30](#_Toc197256515)

[Inference: 30](#_Toc197256516)

[SLR (Simple Linear Regression): 30](#_Toc197256517)

[1. Grades - Study Hours Per Day: 30](#_Toc197256518)

[2. Grades – Sleep Hours Per Day: 31](#_Toc197256519)

[3. Grades – Social Hours Per Day: 32](#_Toc197256520)

[4. Grades - Physical Activity Hours Per Day: 33](#_Toc197256521)

[5. Grades – Extracurricular Hours Per Day: 34](#_Toc197256522)

[6. Grades - Total Activity: 35](#_Toc197256523)

[Comparison Table: SLR Models Predicting Grades 36](#_Toc197256524)

[MLR (Multi Linear Regression): 37](#_Toc197256525)

[Model 1: Study Hours + Sleep Hours 37](#_Toc197256526)

[Model 2: Study Hours + Sleep Hours + Social Hours 38](#_Toc197256527)

[Model 3: Study Hours + Sleep Hours + Social Hours + Extracurricular Time 40](#_Toc197256528)

[Model 4: Full Model (Add Physical Activity) 41](#_Toc197256529)

[Model 5: Full Model + Gender + Stress Level 42](#_Toc197256530)

[Comparison Table: MLR Models 1–5 43](#_Toc197256531)

[Key Takeaways: 44](#_Toc197256532)

[Conclusion: 44](#_Toc197256533)

[Summary of Our Study: 44](#_Toc197256534)

[Key Findings: 44](#_Toc197256535)

[Limitations of Our Study: 45](#_Toc197256536)

[Final Thoughts: 45](#_Toc197256537)

# Introduction:

For this project, we picked a topic that most students in Pakistan can relate to “how our **daily lifestyle habits affect our grades**.” You know, things like how many hours we study, sleep, hang out with friends, or play sports.

We were curious to find out if there’s any actual link between these habits and academic performance. Like, does more sleep really help? Or is it better to study long hours? Maybe too much socializing affects grades? So, we decided to dig into this using real data.

The goal of our study is simple:  
**To understand how different lifestyle factors impact students’ GPA or grades.**

We used a dataset with details like study hours, sleep time, physical activity, social life, etc. The idea was to see if there’s any pattern in which habits help, and which might hurt your academic performance.

# Select Secondary Dataset:

We got our dataset from **Kaggle**, which is a trusted platform for public datasets kind of like a big data library for students and researchers. The dataset was originally clean and well-structured.

But since our project required **data preprocessing**, we decided to mess it up a bit (on purpose, of course). We wanted to practice real-world data cleaning, the way it's done in actual projects.

So, here’s what we did:

* We **removed some values** randomly from a few columns (just like missing data happens in real life).
* We **added outliers** in over 200 rows — like making someone study 40 hours a day (which is totally impossible, you know).
* In total, we worked with **2010+ rows and 8 columns**, covering different lifestyle habits and the students’ grades.

This gave us a mix of realistic and problematic data, perfect for cleaning and analysis.

# Descriptive Statistics:

We performed both the description and summary techniques along with evaluating the mode of the data values. We have also added snapshots along with the tabular and descriptive analysis of our findings.

## Clean Data Set Summary:

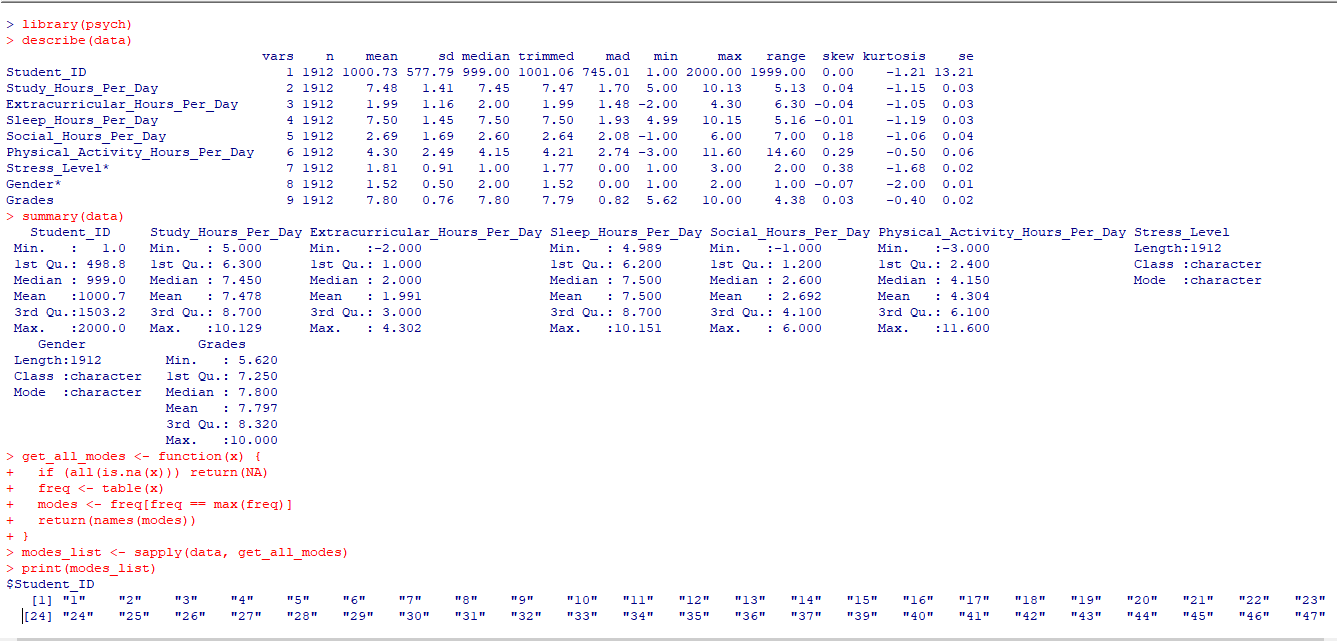


Figure (Clean Data Set Summary)

## Unclean Data Set Summary:

A screen shot of a computer

AI-generated content may be incorrect.

Figure (Unclean Data Set Summary)

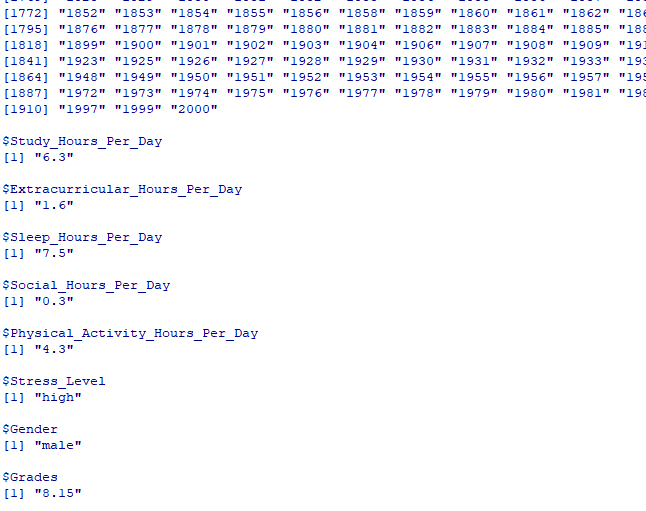


Figure (MODE of clean data set)

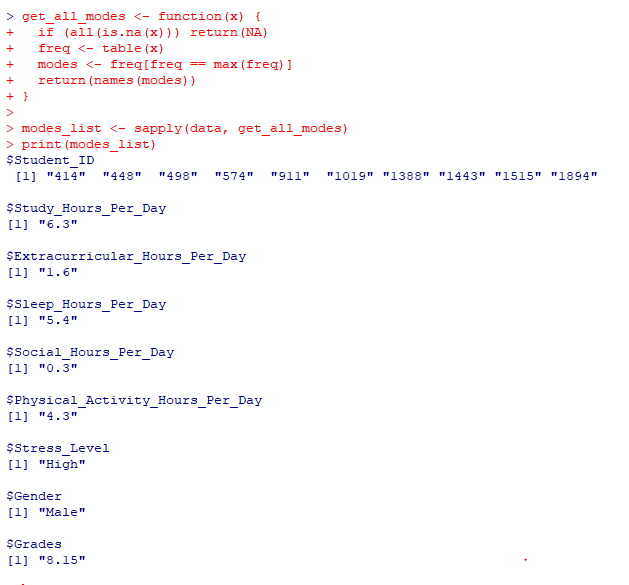


Figure (Mode Unclean Data Set)

Below is the tabulated analysis of our findings.

## Tabular Analysis (Clean and Unclean dataset):

|  |  |  |
| --- | --- | --- |
| Variable | Clean Data Summary | Unclean Data Summary |
| Student\_ID | Mean: 1000.73  Median: 999  Min: 1  Max: 2000  Mode: All unique (no single mode) | Mean: 1000.55  Median: 1000.5  Min: 1  Max: 2000  Mode: 10 IDs occurred more than once |
| Study\_Hours\_Per\_Day | Mean: 7.48  Median: 7.45  Min: 5  Max: 10.13  SD: 1.41  Mode: 6.3 | Mean: 56.87 (erroneous scale)  Median: 56  Min: 1  Max: 108  SD: 29.00  Mode: 6.3 |
| Extracurricular\_Hours\_Per\_Day | Mean: 1.99  Median: 2.00  Min: -2 (outlier)  Max: 4.30  SD: 1.16  Mode: 1.6 | Mean: 2.08  Median: 2.00  Min: -20 (extreme outlier)  Max: 29  SD: 1.73  Mode: 1.6 |
| Sleep\_Hours\_Per\_Day | Mean: 7.50  Median: 7.50  Min: 4.99  Max: 10.15  SD: 1.45  Mode: 7.5 | Mean: 62.57 (erroneous scale)  Median: 66  Min: 1  Max: 109  SD: 31.19  Mode: 5.4 |
| Social\_Hours\_Per\_Day | Mean: 2.69  Median: 2.6  Min: -1 (slight outlier) Max: 6  SD: 1.69  Mode: 0.3 | Mean: 50.23 (erroneous scale)  Median: 45  Min: 1  Max: 121  SD: 31.66  Mode: 0.3 |
| Physical\_Activity\_Hours\_Per\_Day | Mean: 4.30  Median: 4.15  Min: -3  Max: 11.60  SD: 2.49  Mode: 4.3 | Mean: 4.35  Median: 4.1  Min: -3  Max: 23  SD: 2.58  Mode: 4.3 |
| Stress\_Level | Mean: 1.81  Median: 1  Min: 1  Max: 3  SD: 0.91  Mode: "high" | Mean: 8.21 (likely erroneous encoding) Median: 5  Min: 1  Max: 14  SD: 3.69  Mode: "High" |
| Gender | Mean: 1.52  Median: 2  Mode: "male" | Mean: 6.47 (incorrect encoding)  Median: 7.5  Mode: "Male" |
| Grades | Mean: 7.80  Median: 7.8  Min: 5.62  Max: 10  SD: 0.76  Mode: 8.15 | Mean: 103.48 (erroneous scale)  Median: 104  Min: 1  Max: 206  SD: 39.90  Mode: 8.15 |

## Descriptive Statistical Analysis: Clean vs. Unclean Dataset

This section presents a comparative statistical analysis of both the **cleaned** and **uncleaned** versions of the dataset, focusing on summary statistics such as mean, median, standard deviation, and mode. These insights help assess the central tendency, variability, and overall data quality. Identifying issues in the unclean dataset reinforces the importance of preprocessing steps in data analysis.

### 1. Student\_ID

* **Clean Data**: The Student\_ID variable ranges from 1 to 2000, with a mean of 1000.73 and median of 999. This indicates an even distribution of IDs. The mode reveals all values are unique, so no repeated IDs exist.
* **Unclean Data**: Similar range and mean; however, 10 specific IDs (e.g., 414, 498, 911) appeared more than once, suggesting **duplicate entries** or identity conflicts.
* **Inference**: Duplicate entries in unclean data may compromise individual-level analysis and violate assumptions of independent observations.

### 2. Study\_Hours\_Per\_Day

* **Clean Data**: Mean of 7.48 hours/day with a narrow standard deviation (SD) of 1.41 suggests consistent study behavior. The mode is 6.3 hours, close to both mean and median.
* **Unclean Data**: Mean of **56.87** and SD of **29.00**, with values extending up to 108 hours/day, which is **physically impossible**. Mode still appears as 6.3, but this is masked by erroneous high values.
* **Inference**: Clear indication of **data entry or unit errors** in unclean data. These extreme values inflate statistical metrics and misrepresent study behaviors.

### 3. Extracurricular\_Hours\_Per\_Day

* **Clean Data**: Mean of 1.99 hours/day and SD of 1.16 indicates moderate, consistent extracurricular involvement. The mode is 1.6 hours.
* **Unclean Data**: Slightly higher mean (2.08), but the presence of outliers (min = -20, max = 29) skews the data. SD increases to 1.73, and the mode remains 1.6.
* **Inference**: Negative values like -20 are **not logically valid** and indicate data corruption. Outliers distort distribution and central tendency.

### 4. Sleep\_Hours\_Per\_Day

* **Clean Data**: Average of 7.50 hours/day with tight distribution. Mode is 7.5, aligning well with mean and median.
* **Unclean Data**: Drastically inflated mean (62.57), SD (31.19), and max value of 109 suggest unit issues (e.g., minutes wrongly labeled as hours). Mode shifts to 5.4.
* **Inference**: Sleep data is severely compromised. Correcting this is critical as sleep duration is a key variable linked to academic performance and stress.

### 5. Social\_Hours\_Per\_Day

* **Clean Data**: Mean of 2.69 hours with SD of 1.69 indicates generally low social activity. A slight outlier at -1 exists but is minimal. Mode is 0.3.
* **Unclean Data**: Mean jumps to 50.23 with an SD of 31.66, which is **unrealistic**. Mode remains 0.3.
* **Inference**: Likely unit conversion or scaling issues; needs correction to avoid misinterpretation in behavior modeling.

### 6. Physical\_Activity\_Hours\_Per\_Day

* **Clean Data**: Mean of 4.30 hours/day, SD of 2.49, and mode of 4.3 suggests healthy, consistent physical activity. Some mild negative outliers exist.
* **Unclean Data**: Slightly higher max (23 hours), but the core distribution remains similar. SD is 2.58.
* **Inference**: This variable is relatively stable across both datasets, although negative or high-end values still require review.

### 7. Stress\_Level

* **Clean Data**: Mean of 1.81 with SD of 0.91 on a 1–3 scale. Mode is "high", indicating high stress is most common.
* **Unclean Data**: Mean spikes to **8.21**, with max of 14, indicating likely **categorical encoding errors** or wrong scaling. Mode is "High" (capitalized).
* **Inference**: This is a major data integrity issue. Incorrect encodings distort statistical summaries and downstream analytics such as classification.

### 8. Gender

* **Clean Data**: Binary coded (1 = Male, 2 = Female). Mean of 1.52 indicates slightly more females. Mode: "male".
* **Unclean Data**: Mean of 6.47 and range from 1 to 11 suggest **improper or multi-category encodings**, possibly due to survey input errors. Mode is "Male".
* **Inference**: Gender encoding must be **normalized and validated** to ensure meaningful categorical analysis.

### 9. Grades

* **Clean Data**: Mean of 7.80, SD of 0.76, and mode of 8.15 reflect consistent academic performance.
* **Unclean Data**: Mean of **103.48**, max of 206, and SD of 39.90 suggest **unit mismatches or percentage values without normalization**. Mode still 8.15.
* **Inference**: These inconsistencies can lead to incorrect conclusions about academic success unless corrected.

### Final Remarks:

The analysis demonstrates the importance of data cleaning as a foundational step in data science. The **unclean dataset** contains:

* Extreme outliers (e.g., -20 or 108 hours)
* Encoding issues (e.g., Stress\_Level, Gender)
* Unit inconsistencies (e.g., minutes vs. hours)
* Redundant or duplicate entries (Student\_IDs)

By contrast, the **clean dataset** exhibits:

* Statistically plausible values
* Improved alignment between mean, median, and mode
* Enhanced reliability for downstream modeling or visualization

This comparison illustrates that without rigorous data preprocessing, any analysis can become **misleading** or even **invalid**. Hence, **data cleaning is not optional, essential**.

# Data Preprocessing:

## Step 1: Fixing Wrong Data Types

When we loaded the dataset, two numeric columns (Study\_Hours\_Per\_Day and Grades) were wrongly read as **character type**. We fixed them using as.numeric().

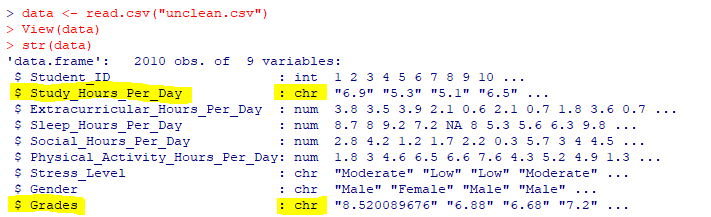


Figure (Data Types Issues)

A screenshot of a computer

AI-generated content may be incorrect.

Figure (Data Type Fix)

That gave us a warning that some entries were invalid and would be converted to NA — no worries, we dealt with that in the next step.

## Step 2: Handling Missing Values

Since NA values popped up, we used **mean imputation** to fill them in for all numeric columns:

* Extracurricular hours
* Sleep hours
* Social hours
* Physical activity

A screen shot of a computer code

AI-generated content may be incorrect.

Figure (NA issue due to conversion)

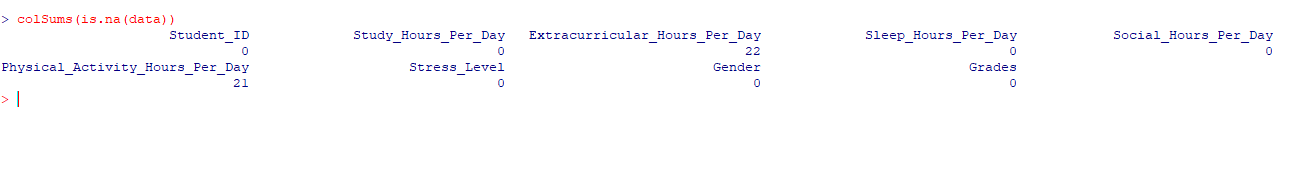


Figure (NA count in Data Set)

Just took the column mean and filled the blanks, done and dusted.

A close-up of a document

AI-generated content may be incorrect.

Figure (Fixed All NA's)

## Step 3: Removing Duplicates

Since there were no duplicates so it didn’t make a special change over the data set.

data <- data[!duplicated(data), ]

## Step 4: Fixing Categorical Mess

The Gender and Stress\_Level columns were full of **typos, extra spaces, and weird formats**.

A screen shot of a number

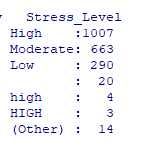
AI-generated content may be incorrect. A computer screen shot of a data

AI-generated content may be incorrect.

Figure (Gender Issue) Figure (Gender Issue Fixed)

We had values like "Femaale", "Mod", "Hig" — typical dirty data stuff. Here's what we did:

* Converted all values to **lowercase** for consistency.
* Cleaned up common typos.
* Replaced anything weird or unknown with the **most frequent** value (just to keep things smooth).

 A computer screen shot of a data

AI-generated content may be incorrect.

Figure (Stress Level Issue) Figure (Stress Issue Fixed)

Now both columns look clean and proper. You can also have a look at them.

## Step 5: Outliers and Invalids

Some extreme values were messing up the scale — like someone with 35 hours of physical activity a day. So, we dealt with outliers and trimmed those rows.

A screen shot of a computer code

AI-generated content may be incorrect.

Figure (Removed Rows with outliers)

## Step 6: Saved the Clean Data

Once everything was clean, we saved the final dataset using:

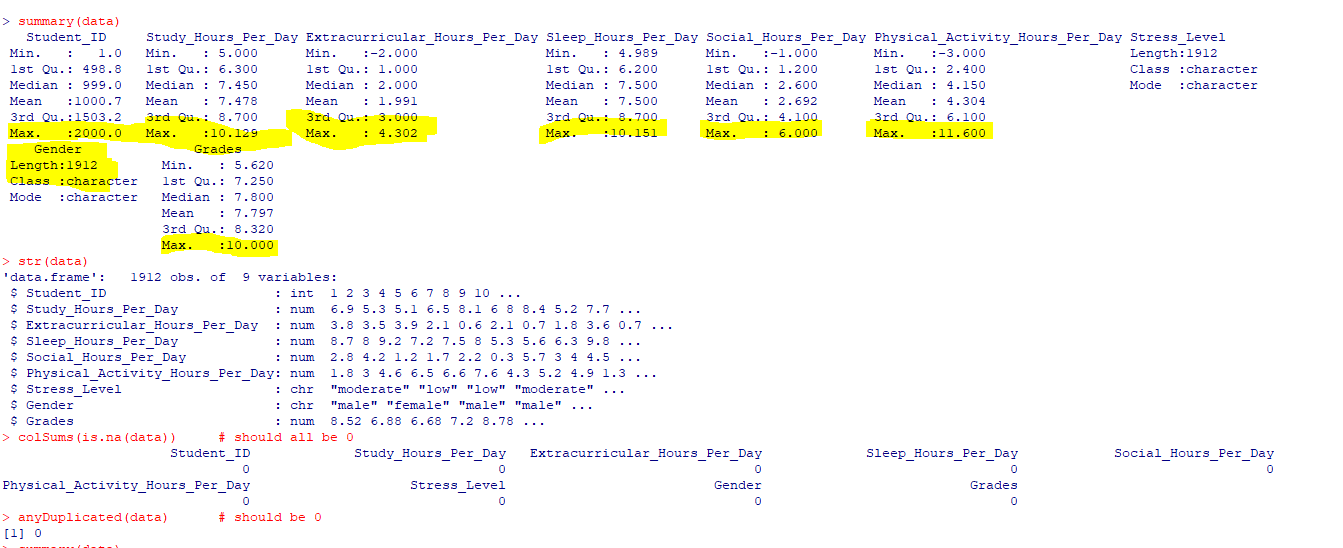


Figure (Cleaned Data Set)

write.csv(data, "cleaned\_student\_data.csv", row.names = FALSE)

That way we kept both the **dirty** and **clean** versions for future comparisons.

## Step 7: Feature Scaling:

We also applied **feature scaling** to bring all numeric values onto the same level. This means converting them to a common scale usually with a mean of 0. We used the scale() function in R, which helped us make sure none of the variables dominate the others just because of their range. **It’s not always necessary for regression, but it makes the data neater and ready for future modeling too.**

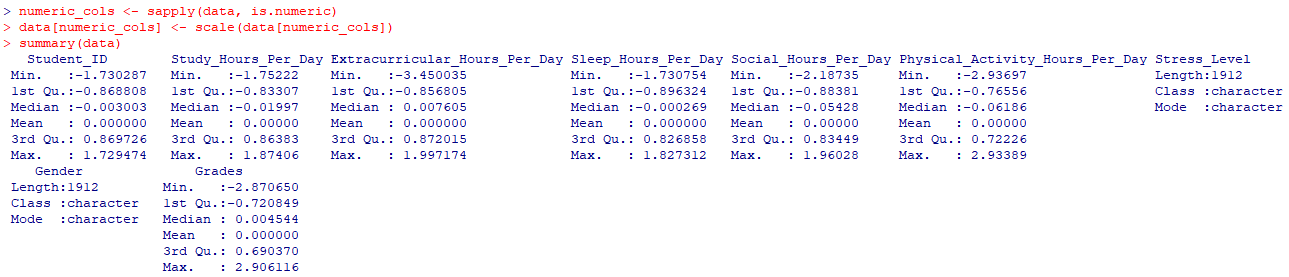


Figure (Feature Scaling on Data set, NOT specifically used)

## Final Check with Summary

We ran summary(data) again just to make sure everything looked fine — and yes, it finally did. Also saved the summary as a text file using **sink()** just in case we need to show proof during the viva.

# Exploratory Data Analysis (EDA):

## 1. Grade Distribution x and y:

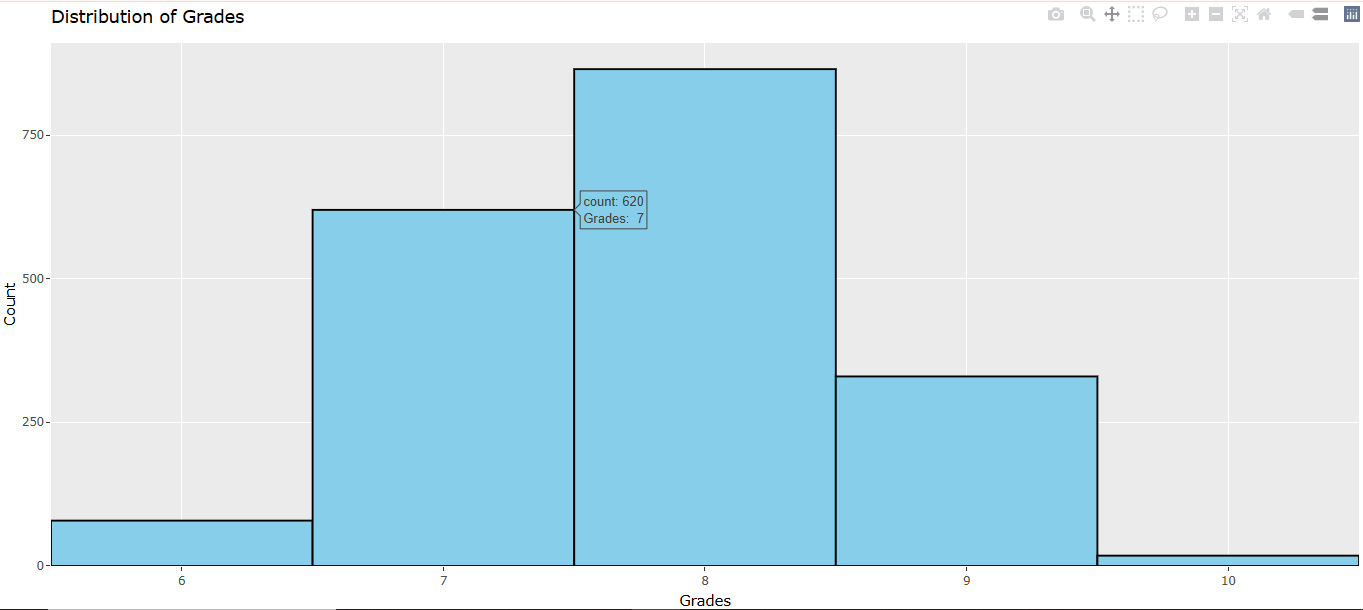


Figure (grade distribution x and y)

### Inference:

Most students have grades concentrated around 7.5–8.5. The distribution is slightly skewed to the left.

## 2. Boxplot: Grades by Gender

**Both genders perform similarly, but females have slightly higher median grades.**

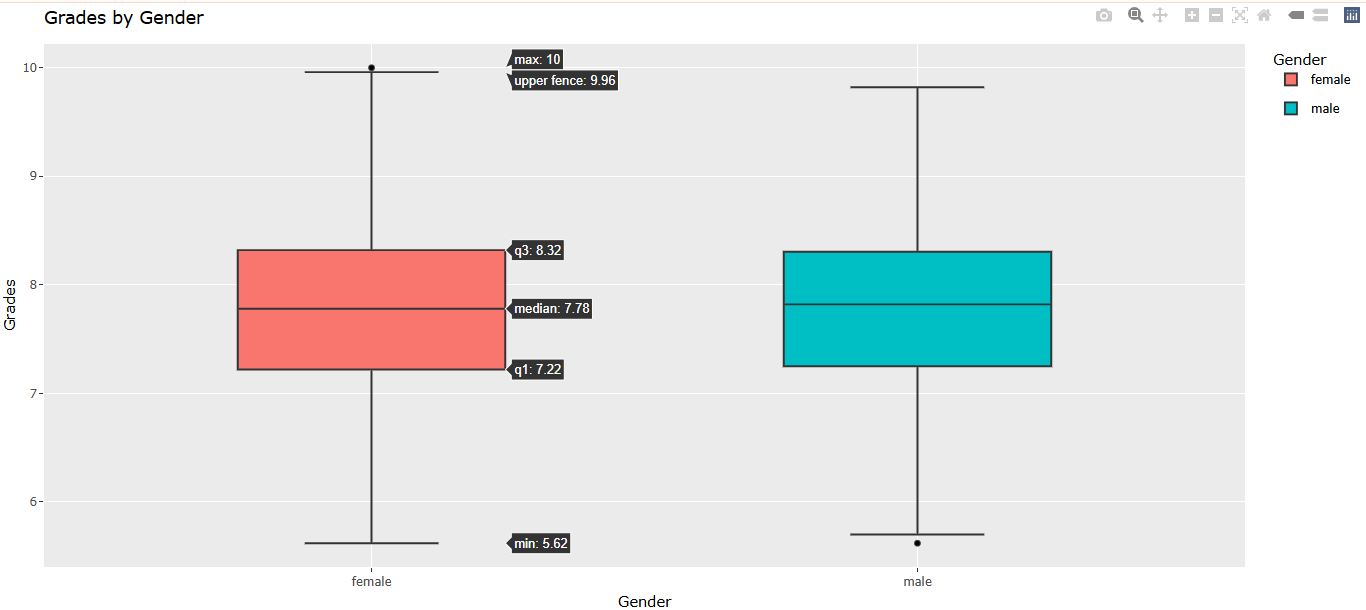


Figure (Boxplot: Grades by Gender)

## 3. Boxplot: Grades by Stress Level

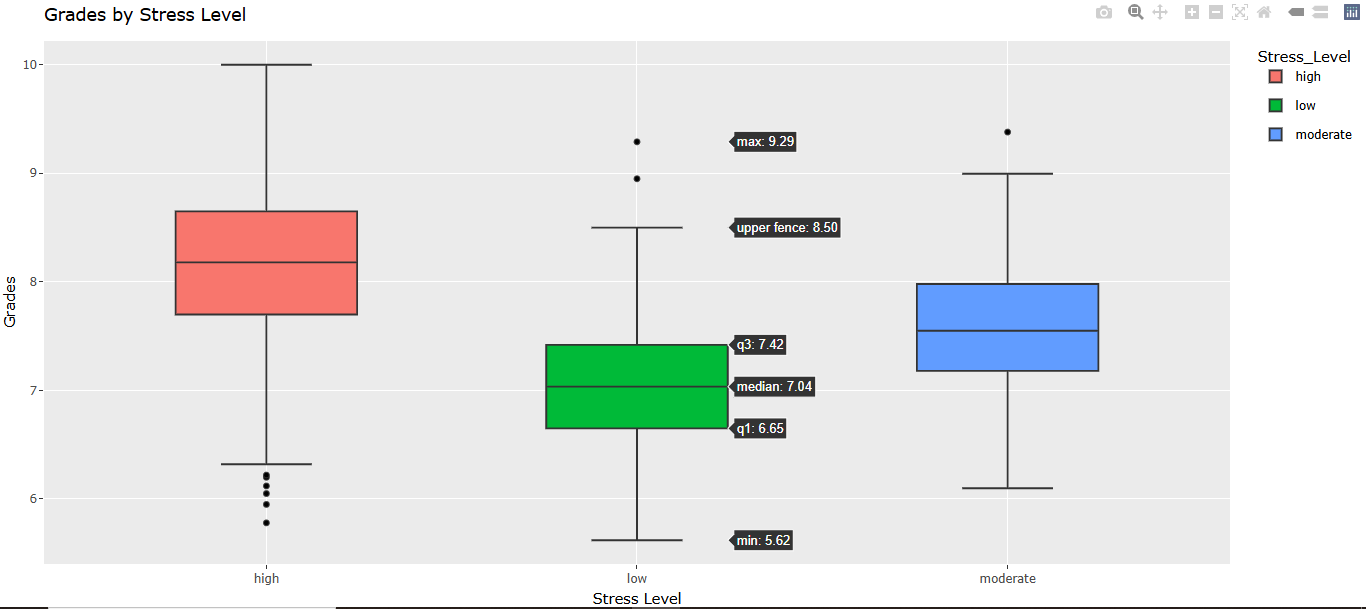


Figure (Boxplot: Grades by Stress Level)

**Overall Insights:**

* A moderate level of stress may be optimal for maintaining decent academic performance.
* High stress can lead to both high achievers and struggling students, showing a polarized impact.
* Low stress might correlate with lower average performance, possibly due to lack of pressure or urgency.

## 4. Study Hours vs Grades:

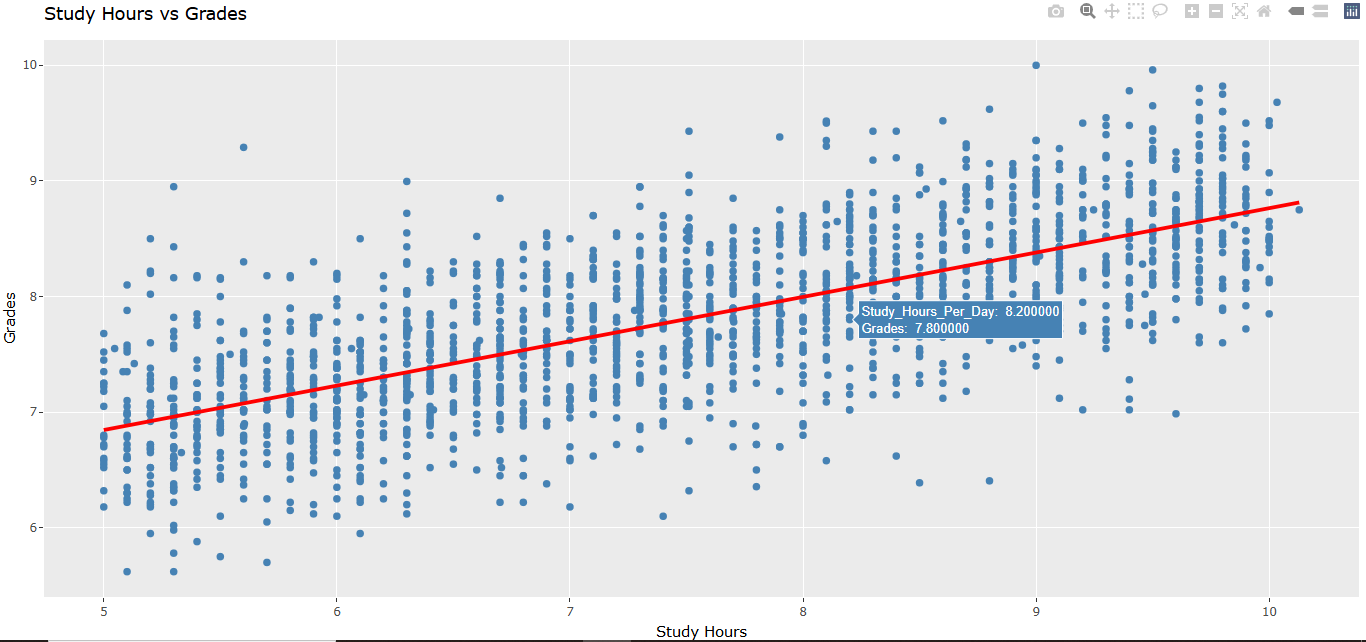


Figure (Study Hours vs Grades)

### Inference:

* Studying more hours per day is generally associated with higher grades, supporting the idea that increased study time positively impacts academic performance.
* However, the scatter and outliers indicate that studying hours alone does not determine grades efficiency, comprehension, and individual learning styles can also play important roles.

## 5. Sleep Hours vs Grades:

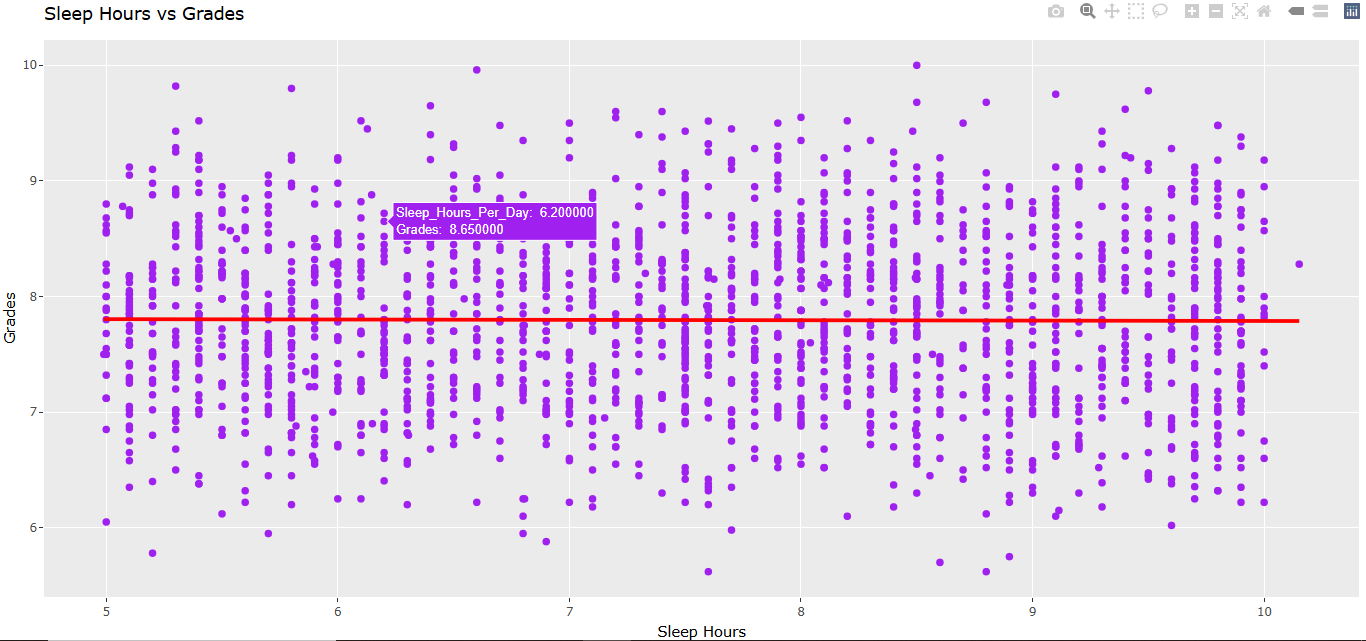


Figure (Sleep Hours vs Grades)

### Inference:

* Weak or No Correlation: The regression line is almost flat, indicating little to no linear correlation between sleep hours and grades. This means increasing or decreasing sleep doesn't strongly predict changes in grades.
* Wide Variability: At nearly every level of sleep (from ~5 to 10 hours), grades vary widely from as low as ~6 to nearly 10. This suggests that many other factors besides sleep likely influence grades.
* Clustered Data: Most data points seem to cluster between 6 and 9 hours of sleep and grades between 7 and 9.

## 6. Social Hours vs Grades:

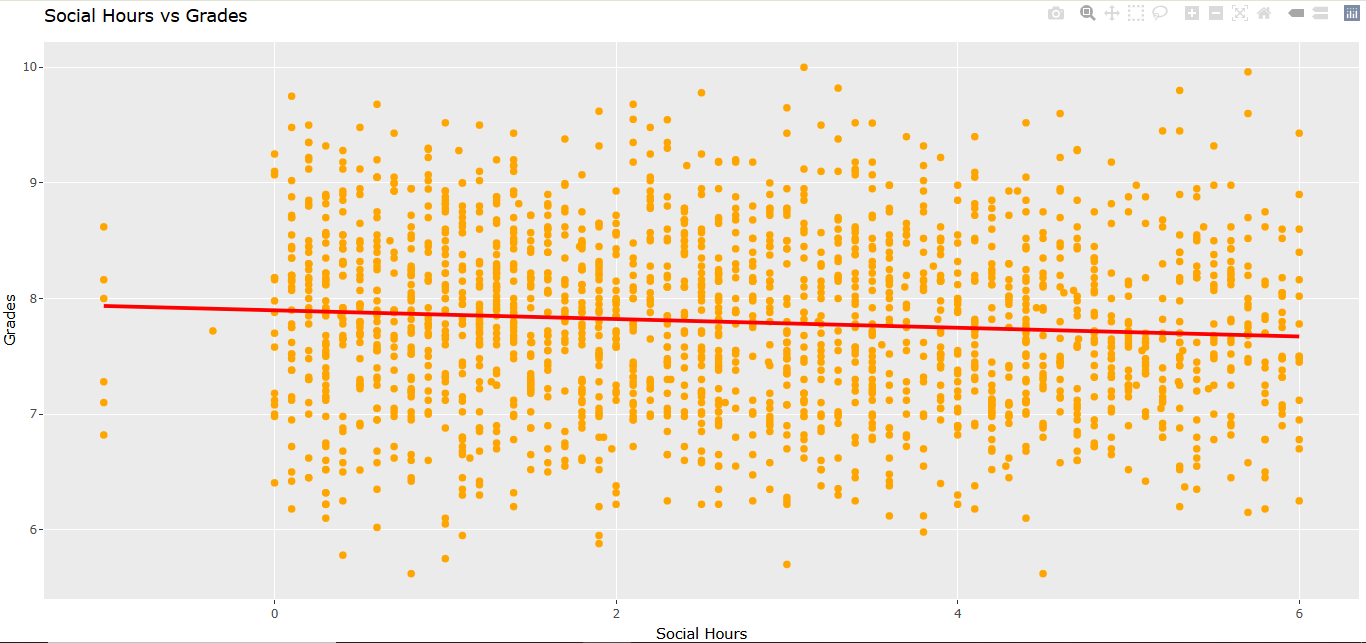


Figure (Social Hours vs Grades)

### Inference:

* Slight Negative Correlation: The red regression line has a gentle downward slope, suggesting a small negative correlation between social hours and grades. As social hours increase, grades tend to decrease slightly.
* Still High Variability: Despite the slight trend, students with a wide range of grades are found at almost all levels of social hours (0 to ~6). This indicates that social time is not a strong predictor of its own.
* Concentration of Data: Most students appear to spend between 0 and 4 hours on social activities, and most grades still fall between 7 and 9, just like in the sleep analysis.

## 7. Physical Activity vs Grades:

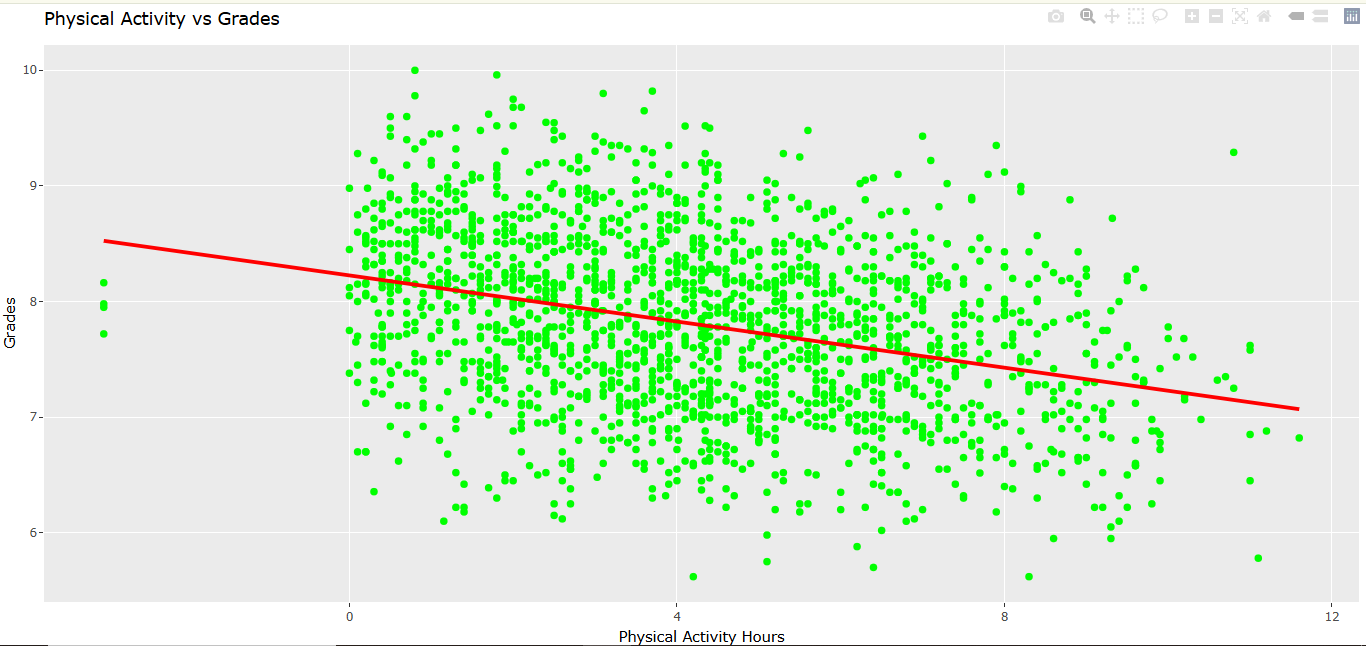


Figure (Physical Activity vs Grades)

### Inference:

* Moderate Negative Correlation: The red regression line has a clear downward slope, indicating a stronger negative correlation than in the previous plots. As physical activity hours increase, grades tend to decrease more noticeably.
* Pattern Observed: Students who spend fewer hours (0–3 hours) on physical activity generally have higher grades, while those spending more time (6+ hours) tend to have lower grades, on average.
* Still Considerable Variation: Although the trend is clearer here, there is still a wide range of grades at most levels of physical activity. So, while there's a trend, it is not deterministic.

## 8. Extracurricular Hours vs Grades:

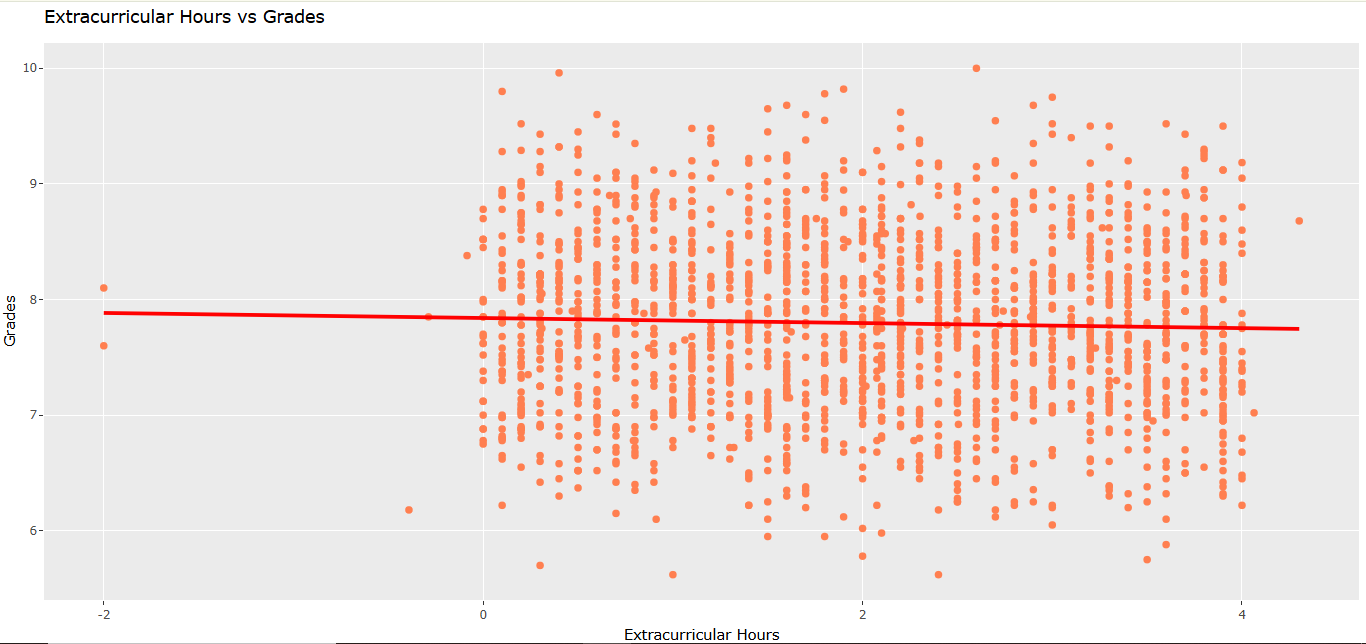


Figure (Extracurricular Hours vs Grades)

From the scatter plot we can draw the following inferences:

* **Trend Line (Red Line):** The red regression line is nearly flat with a very slight negative slope, indicating a very weak or negligible negative correlation between extracurricular hours and grades.
* **Correlation:** There appears to be no strong relationship between the number of extracurricular hours and students' grades. As extracurricular hours increase or decrease, grades remain mostly constant, scattered widely.
* **Data Spread:**
  + Grades are generally concentrated between 6 and 9.
  + Extracurricular hours mostly fall between 0 and 4.
  + There's some sparse data outside the main range (e.g., extracurricular hours below 0 and grades above 9.5 or below 6), which may be outliers.
* **Conclusion:** Participating in extracurricular activities, based on this data, does not have a significant impact (positive or negative) on academic grades.

## 9. Stress Level by Gender:

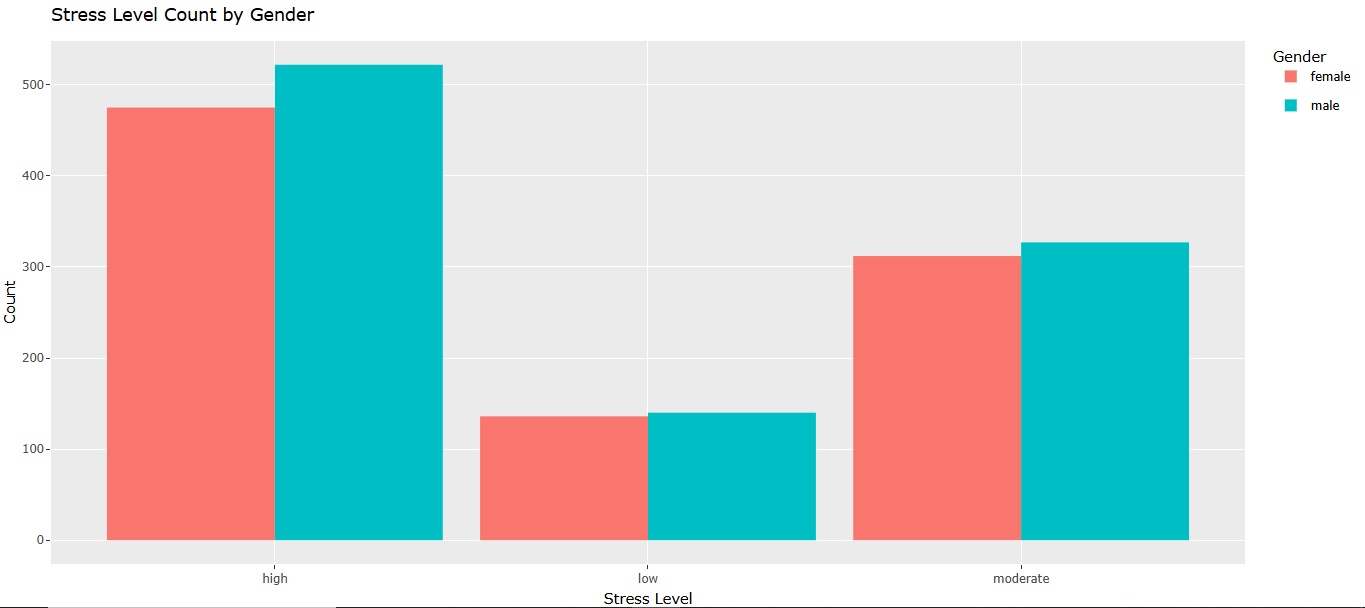


Figure (Stress Level by Gender:)

**Inference:** Male experience noticeably more stress but there is a little or almost no difference in all the three stress categories. This might affect performance or well-being.

## 10. Study Hours by Stress Level:

### Inference:

Students with high stress tend to study the most, with a relatively widespread. Moderately stressed students study less than highly stressed ones but more than low-stress students. Spread is narrower. Students with low stress study the least, with the tightest distribution and lowest hours overall.

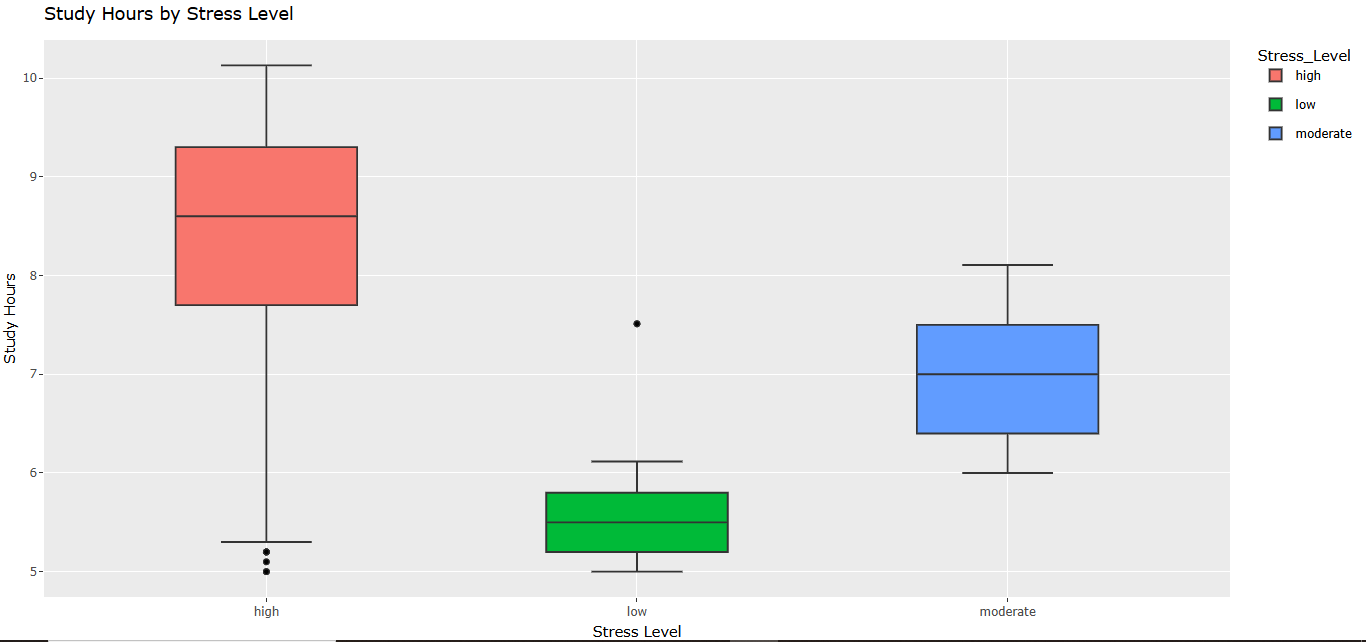


Figure (Study Hours by Stress Level)

## 11. Sleep Hours by Stress Level:

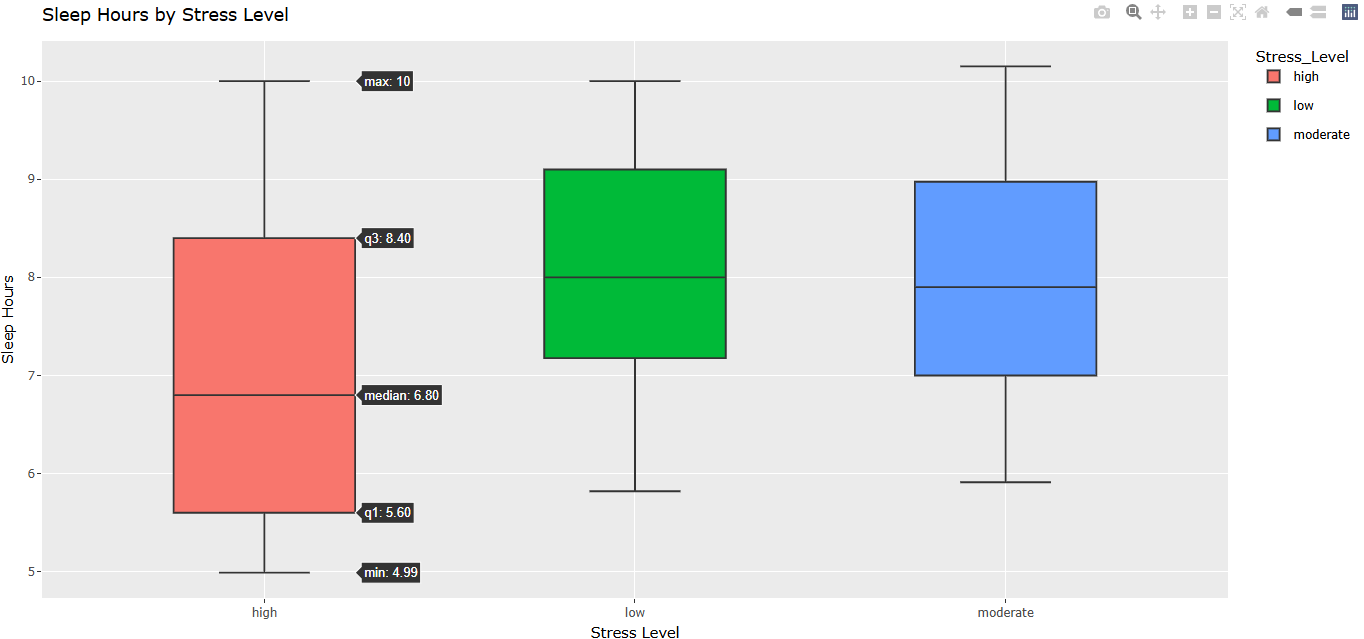


Figure (Sleep Hours by Stress Level)

### Inference:

This boxplot shows that individuals with higher stress levels tend to sleep less and have more variability in sleep duration. Those with low stress levels generally sleep more, with consistent sleep patterns. The median sleep hours increase as stress levels decrease, highlighting a negative correlation between stress and sleep duration.

## 12. Physical Activity by Gender:

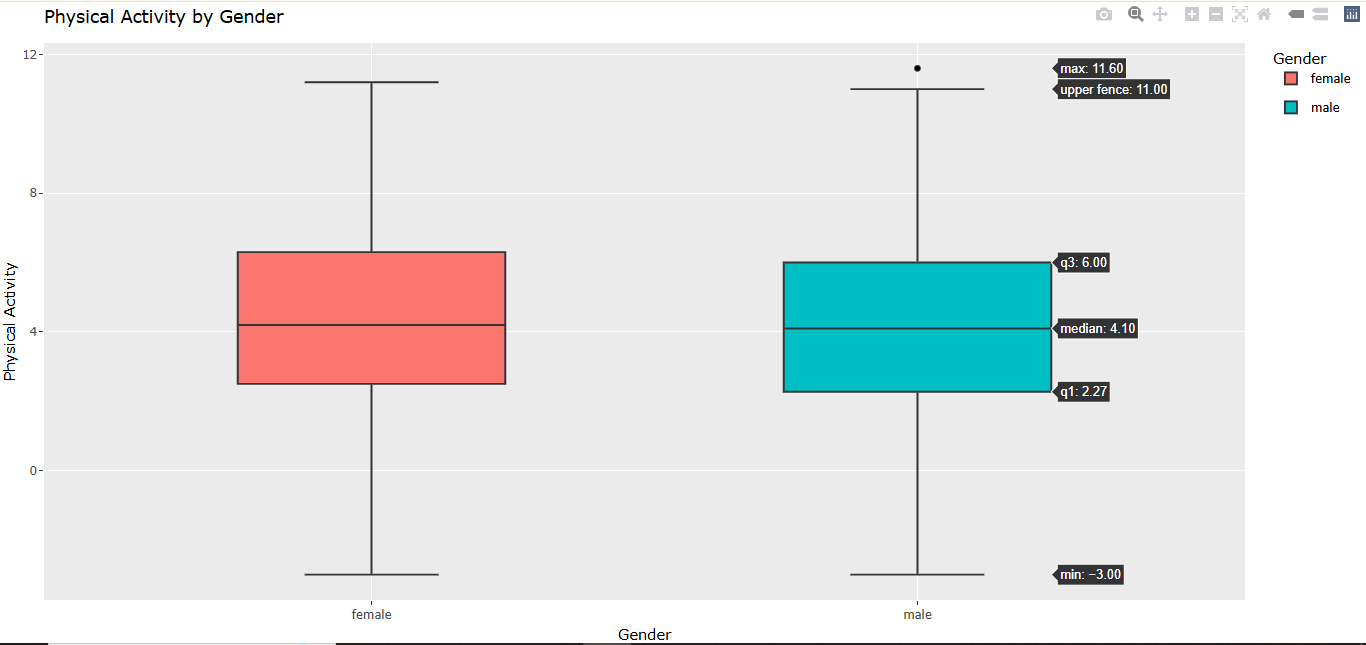


Figure (Physical Activity by Gender)

### Inference:

Males tend to be more physically active than the other gender. As majority of them fall above the median while the female’s majority falls below the median.

## 13. Social Hours by Gender:

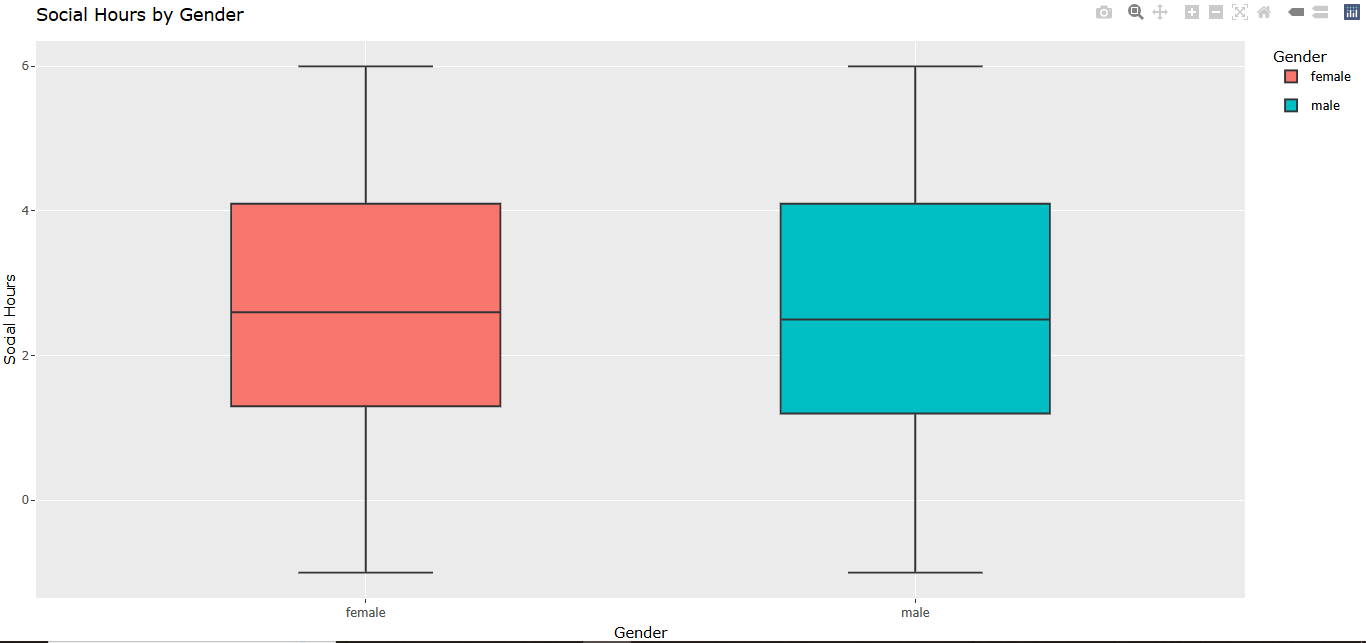


Figure (Social Hours by Gender)

**Key Observations:**

**1. Median Social Hours:**

* Both males and females have a very similar median social time (~2.5 hours).
* This suggests no meaningful difference in average social activity based on gender.

**2. Interquartile Range (IQR):**

* Both groups have similar IQRs, indicating comparable variability in social hours.
* For both, the IQR spans roughly from 1 to 4 hours.

**3. Range:**

* The **minimum** and **maximum** values are roughly the same across genders (~0 to 6 hours).
* Both genders have a few individuals with very low (even slightly negative) or very high social hours.

**4. Outliers:** There are minor outliers, but not enough to indicate major anomalies or group-specific trends.

**Conclusion:** There is no significant gender-based difference in social hours among the students. Both distributions are nearly identical in shape, spread, and central tendency.

## 14. Combined Activity Time vs Grades:

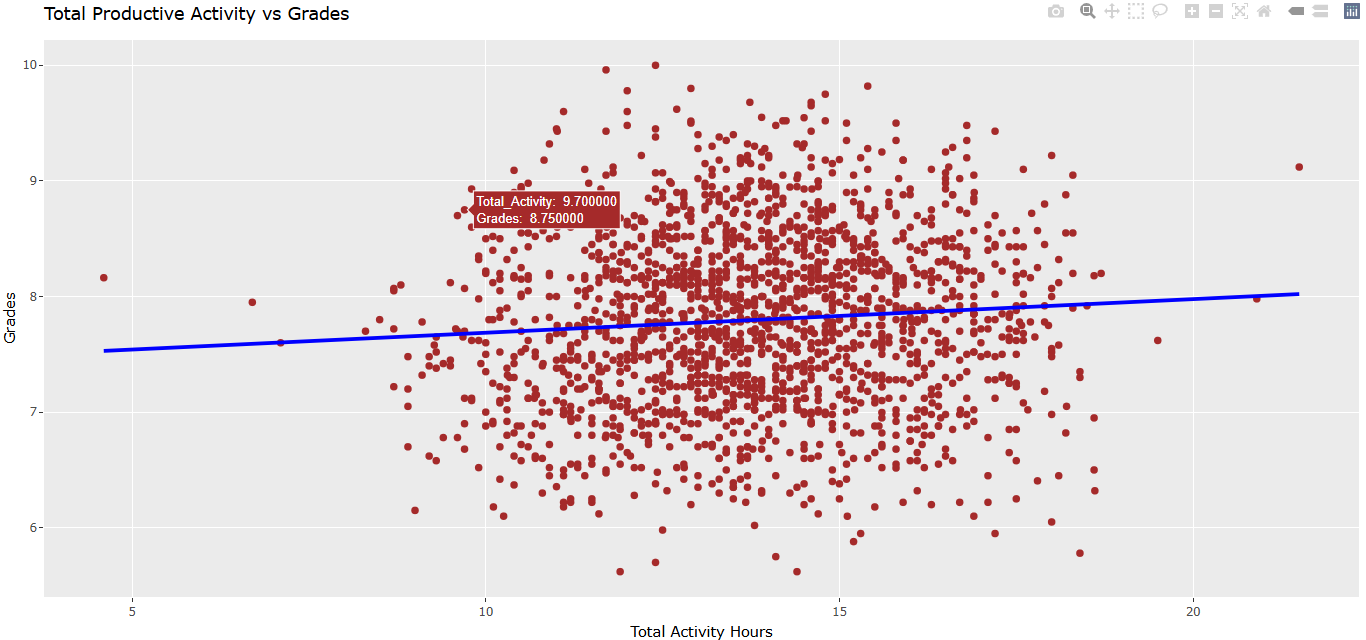


Figure (Combined Activity Time vs Grades)

### Inference:

The graph shows a weak positive correlation between productive activity hours and grades, suggesting that while more productive time may slightly boost performance, it's not a strong predictor of academic success. Grades vary widely across all activity levels, highlighting the importance of other factors like efficiency and individual learning styles.

## 15. Density Plots:

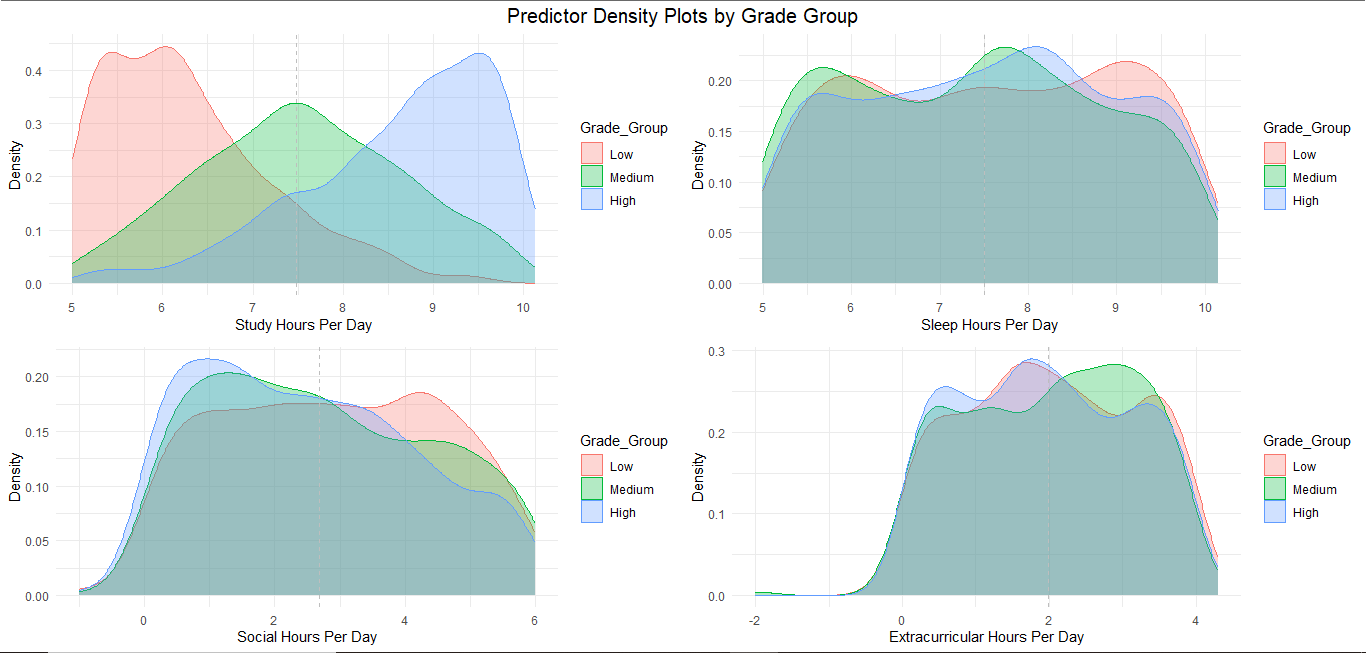


Figure (Density Plots)

### Top-Left: Study Hours Per Day

High-performing students (blue) tend to study more, clustering around 8–9 hours. Low-performing students (red) mostly study less, around 5–6 hours.

Clear pattern: more study time is associated with higher grades.

### Top-Right: Sleep Hours Per Day

All grade groups show a similar distribution, mostly between 7–9 hours. Minor shift: high performers may get slightly more sleep, but differences are not substantial.

Implication: Sleep alone may not strongly differentiate grade groups.

### Bottom-Left: Social Hours Per Day

High-grade students (blue) tend to spend fewer hours on social activities (peak near 1–2 hours). Low-grade students show a broader range, and more time spent socially. Suggests excessive social time may negatively affect grades.

### Bottom-Right: Extracurricular Hours Per Day

All groups are similar, but medium-grade students (green) show a slight peak around 2 hours. No strong trend connecting extracurricular time directly to grades. Implies extracurriculars may support balance but are not a key differentiator.

### Overall Insight:

Study time is the most clearly associated with higher grades. Social time might inversely affect performance. Sleep and extracurriculars appear less influential on grade group differences.

## 16. Violin Plot for Gender vs Grades:

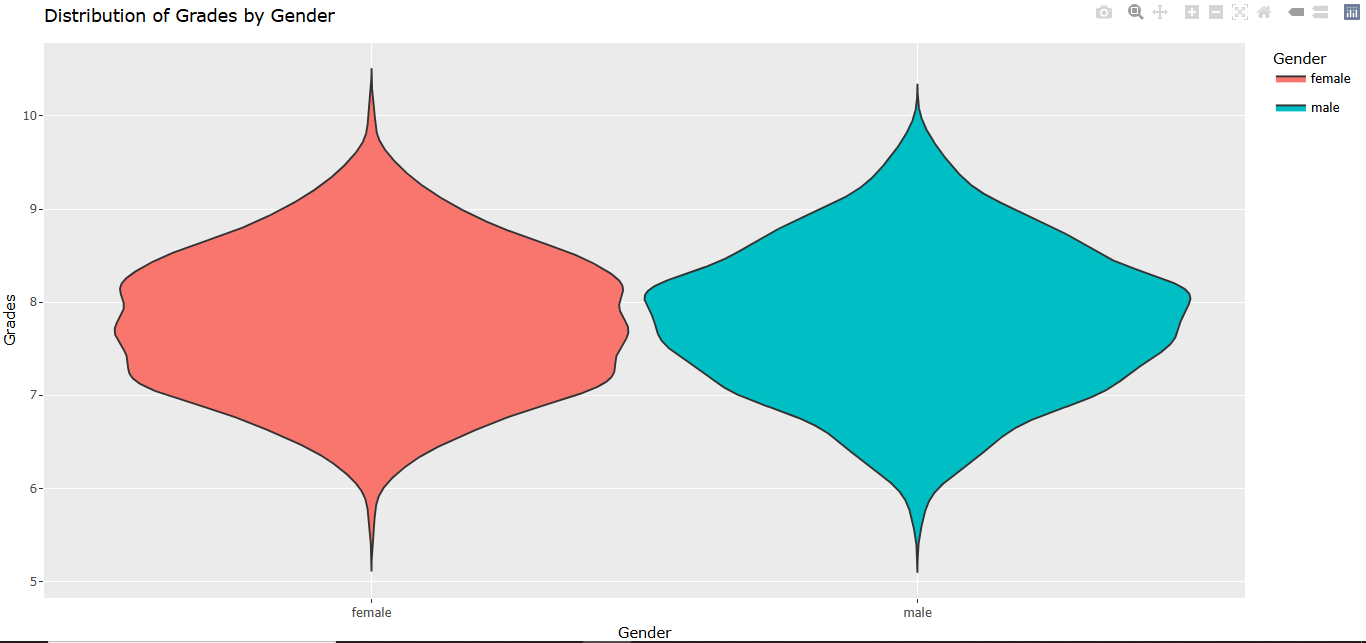


Figure (Violin Plot for Gender vs Grades)

### Summary:

The graph shows a violin plot comparing grade distributions between female and male students. Female students (shown in red on the left) have a distribution centered roughly around 8, with a notable peak near grade 10 and a spread primarily between grades 7 and 9. Male students (shown in teal on the right) have a slightly lower central tendency, with their distribution more concentrated around grades 7-8, though they also show a peak near grade 10. Both distributions have similar overall ranges from approximately grade 5 to just above 10, but the female distribution appears somewhat more concentrated toward higher grades while the male distribution shows greater density in the middle range.

## 17. Gender Distribution Pie chart:

### Summary:

The pie chart displays the gender distribution of what appears to be a population or sample group. Males (represented by the blue segment) make up a slight majority at 51.7% of the total, while females (represented by the orange segment) constitute 48.3%. The distribution is nearly balanced with only a 3.4 percentage point difference between the genders, indicating an even representation in whatever group or dataset is being analyzed.

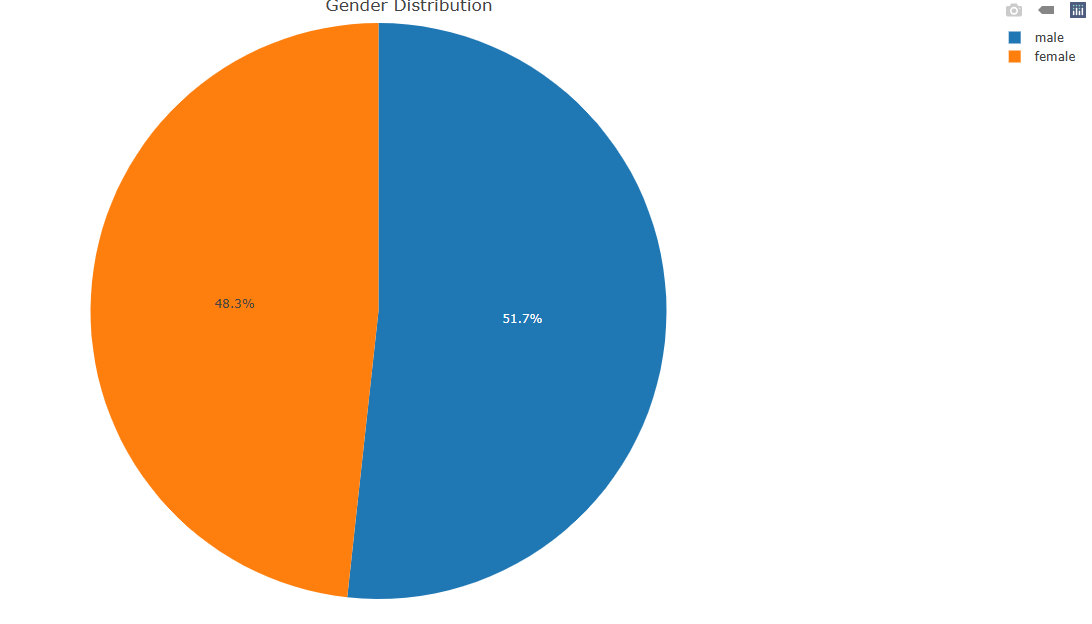


Figure (Gender Distribution Pie chart)

## 18. Scatterplot matrix:

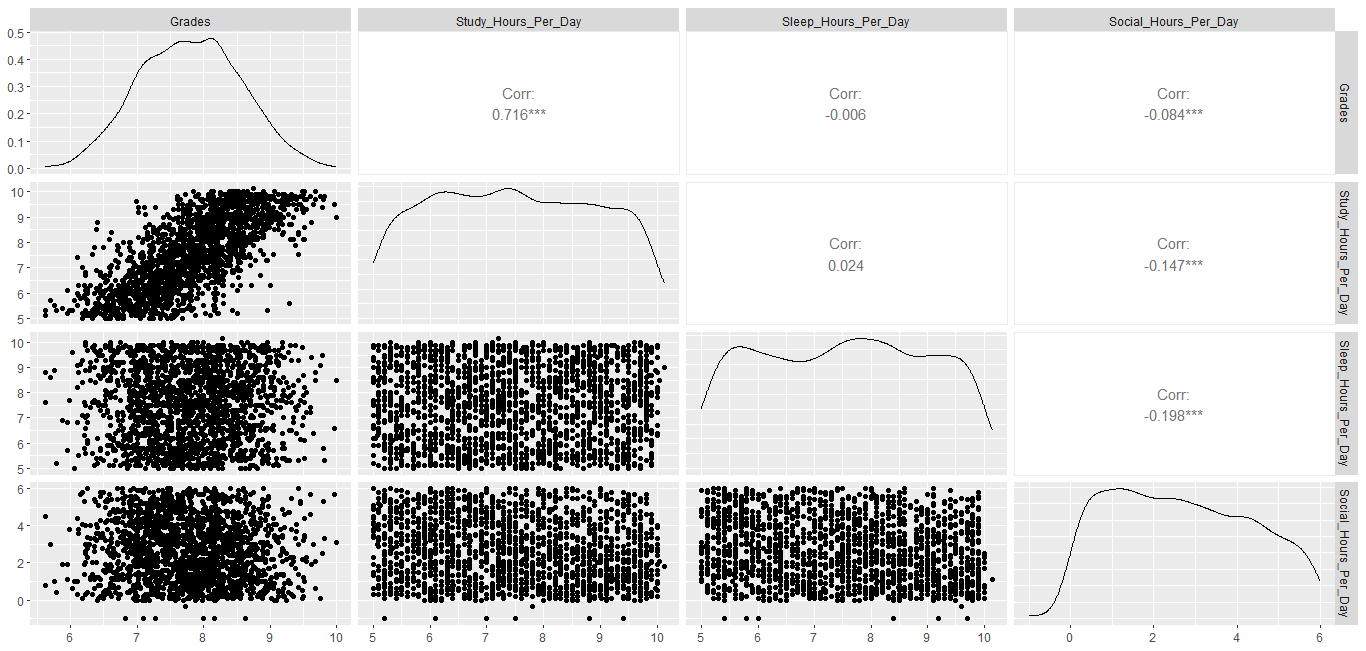


Figure (Scatterplot matrix)

### Summary:

This image shows a pairs plot (scatterplot matrix) displaying relationships between four variables: Grades, Study\_Hours\_Per\_Day, Sleep\_Hours\_Per\_Day, and Social\_Hours\_Per\_Day. The diagonal shows distribution histograms for each variable, while the off-diagonal cells show scatterplots with correlation coefficients. Key findings include a strong positive correlation (0.716\*\*\*) between study hours and grades, indicating that more study time is associated with better academic performance. Sleep hours show no significant correlation with grades (-0.006). Both social hours and sleep hours have significant negative correlations with study hours (-0.147\*\*\* and -0.198\*\*\* respectively), suggesting that time spent socializing or sleeping reduces study time. Social hours also show a slight negative correlation with grades (-0.084\*\*\*), though this effect is weaker than the positive impact of studying. The asterisks (\*\*\*) indicate statistical significance of the correlations.

### Inference:

The primary inference is that study time is the most critical factor for academic success, while maintaining balance between studying, sleeping, and socializing requires careful time management since these activities compete for students' limited time.

The strongest factor influencing academic performance is study time, with a substantial positive correlation (0.716\*\*\*) between study hours and grades. Social activities have a slight negative impact on grades (-0.084\*\*\*), likely because they reduce available study time (-0.147\*\*\* correlation between social hours and study hours). Sleep duration itself doesn't directly correlate with grades, but it does negatively correlate with study time (-0.198\*\*\*), suggesting a time allocation trade-off between different activities.

# Interactive Visualization:

Although all our graphs are interactive, but this one is special “Interactive Bubble Graph.”

It gives the relation between Grades and the study hours (the variable that matters the most as per our study with this data set). Below is the Bubble Plot.

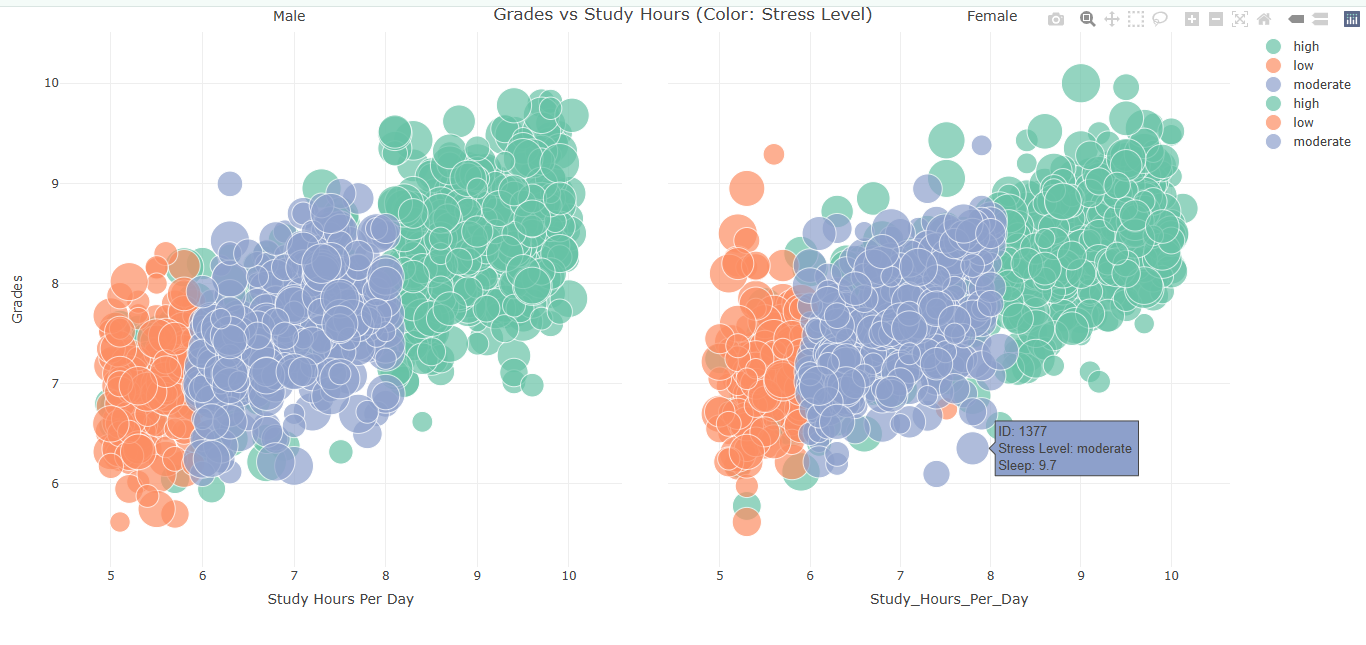


Figure (Interactive Bubble Plot)

# Statistical Modeling and Inference:

## 1. Correlation Analysis Heat Map:

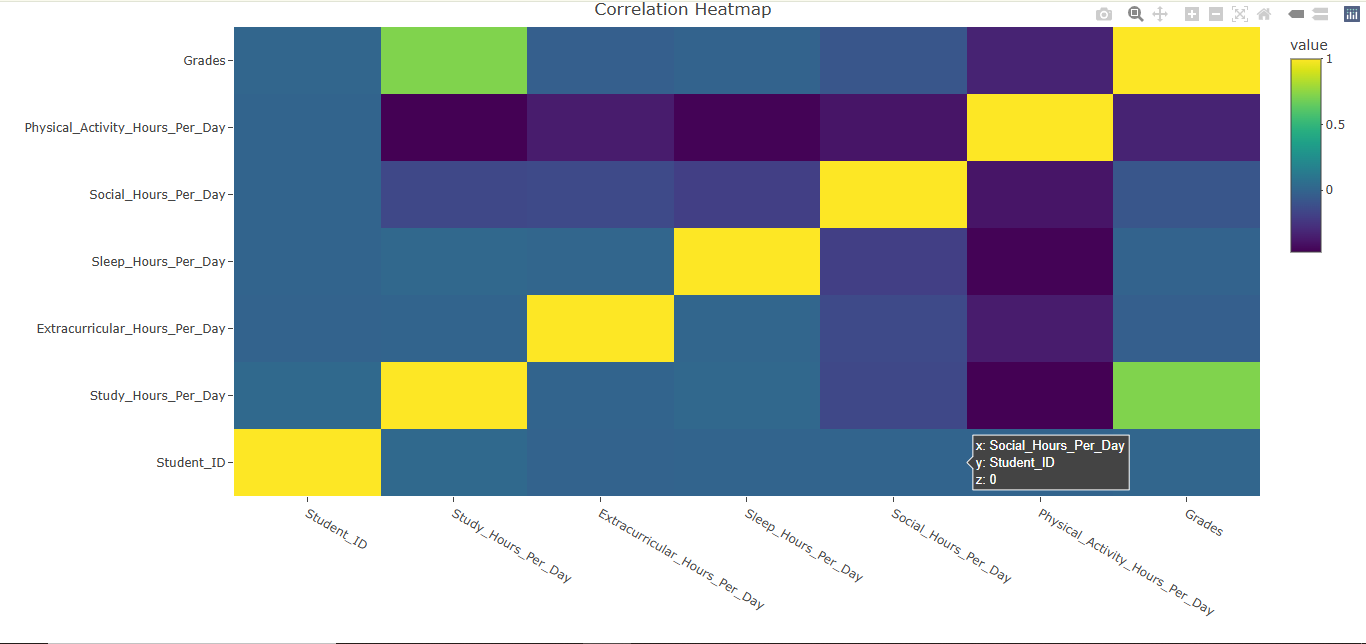


Figure (Correlation Heat Map)

### Summary:

The correlation heatmap displays relationships between various student activities and academic performance. The color scale ranges from dark purple (strong negative correlation) to bright yellow (strong positive correlation), with blue representing neutral or weak correlations.

### Inference:

The primary inference from this heatmap is that time allocation represents a zero-sum game for students, where time spent on one activity necessarily reduces time available for others. Study time maintains its position as the most important factor for academic success, while physical activity appears to have the strongest negative relationship with study time. This suggests students may need guidance on balancing academic commitments with physical activity to maintain both health and academic performance. The data reinforces the need for effective time management skills among students to balance competing priorities.

## SLR (Simple Linear Regression):

### 1. Grades - Study Hours Per Day:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Model Summary:

* **Slope**: 0.384 → Each additional hour of studying per day increases grades by 0.384 points.
* **R²**: 0.5125 → Explains 51.25% of the variance in grades.
* **Highly significant**, best predictor.

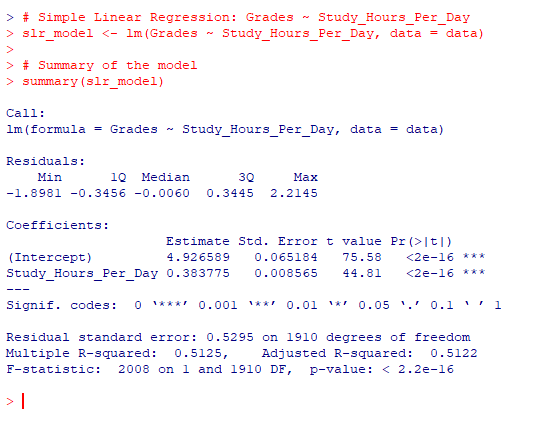


Figure (Summary SLR MODEL 1)

#### Plot:

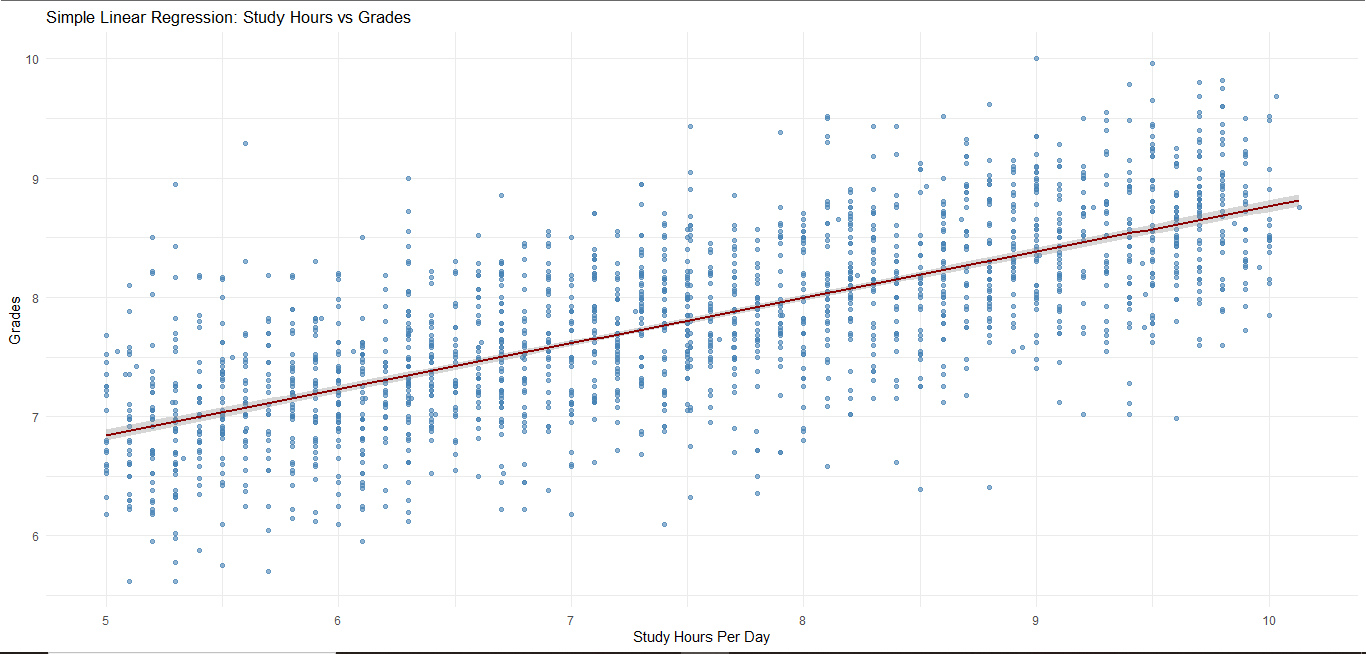


Figure (Plot SLR MODEL 1 GRADES vs STUDY HRS)

### 2. Grades – Sleep Hours Per Day:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **Slope**: -0.00315 → Essentially no effect.
* **R²**: ~0 → Sleep has no predictive power on grades.
* **Not significant**.

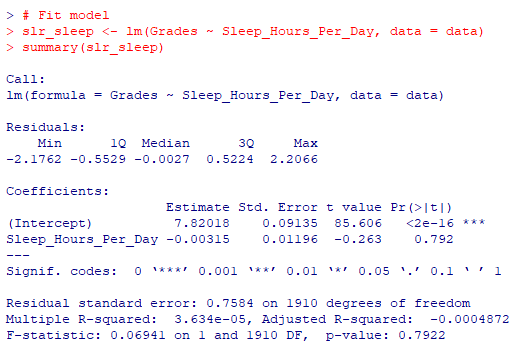


Figure (Summary SLR MODEL 2 GRADE vs SLEEP)

#### Plot:

A screen shot of a graph

AI-generated content may be incorrect.

Figure (Plot SLR MODEL 2 SLEEP vs GRADE)

### 3. Grades – Social Hours Per Day:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **Slope**: -0.0377 → Each additional hour socializing slightly lower grades.
* **R²**: 0.007 → Very small effect.
* **Significant**, but weak.

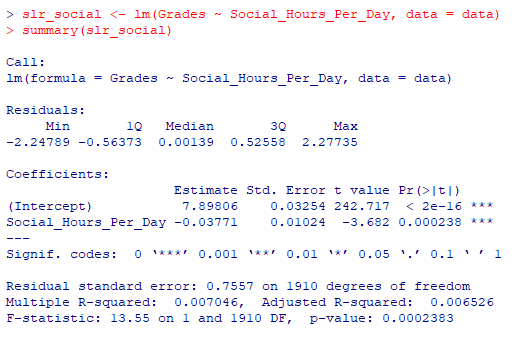


Figure (Summary SLR MODEL 3 GRADE vs SOCIAL HRS)

#### Plot:

A graph with a red line

AI-generated content may be incorrect.

Figure (Plot SLR MODEL 3 GRADE vs SOCIAL HRS)

### 4. Grades - Physical Activity Hours Per Day:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **Slope**: -0.0998 → Higher physical activity slightly reduces grades.
* **R²**: 0.1071 → Explains ~10.7% of grade variance.
* **Highly significant**.

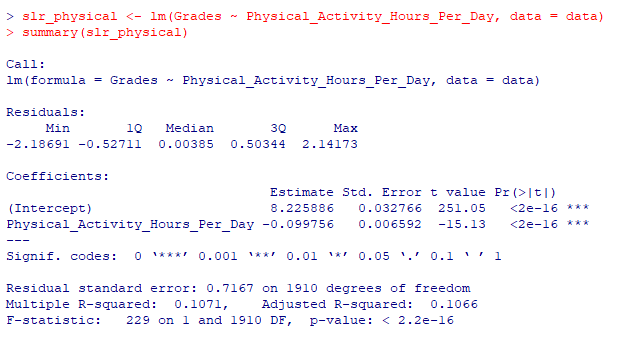


Figure (Summary SLR MODEL 4 GRADES vs PA / HRS)

#### Plot:

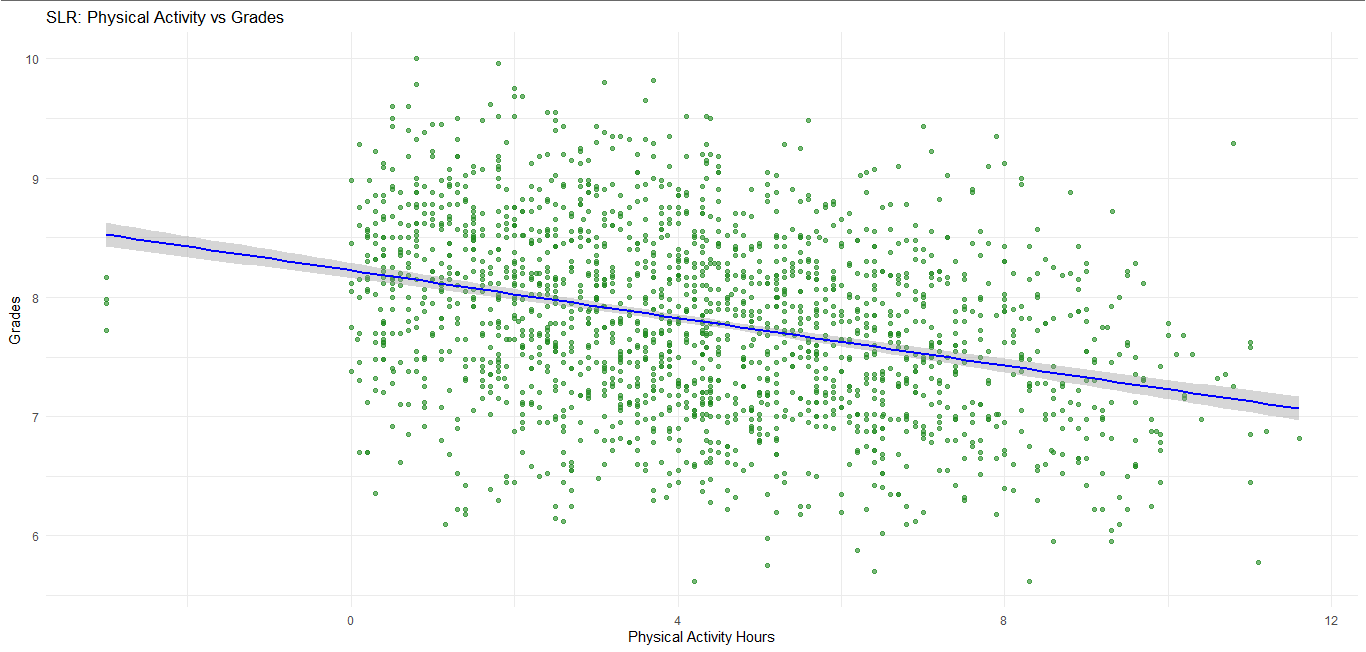


Figure (Plot SLR MODEL 4 GRADES vs PA / HRS)

### 5. Grades – Extracurricular Hours Per Day:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **Slope**: -0.0219 → Slight negative relationship.
* **R²**: 0.0011 → Nearly no explanatory power.
* **Not significant**.

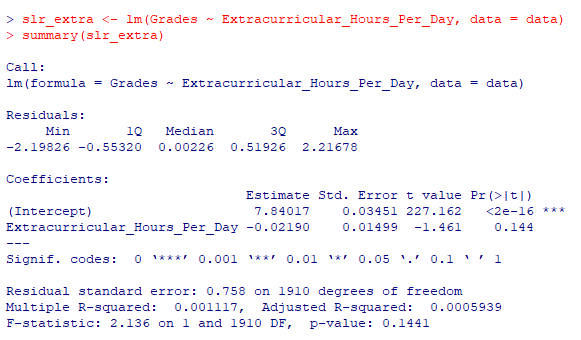


Figure (Summary SLR MODEL 5 GRADES vs EXTRAS)

#### Plot:

A screen shot of a graph

AI-generated content may be incorrect.

Figure (Plot SLR MODEL 5 GRADES vs EXTRAS)

### 6. Grades - Total Activity:

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **Slope**: +0.0291 → Slight positive effect.
* **R²**: 0.0061 → Minimal variance explained.
* **Statistically significant**, but weak.

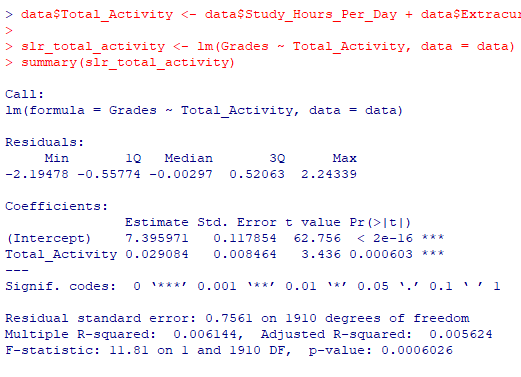


Figure (Summary SLR MODEL 6 GRADES vs ALL ACTIVITIES)

#### Plot:

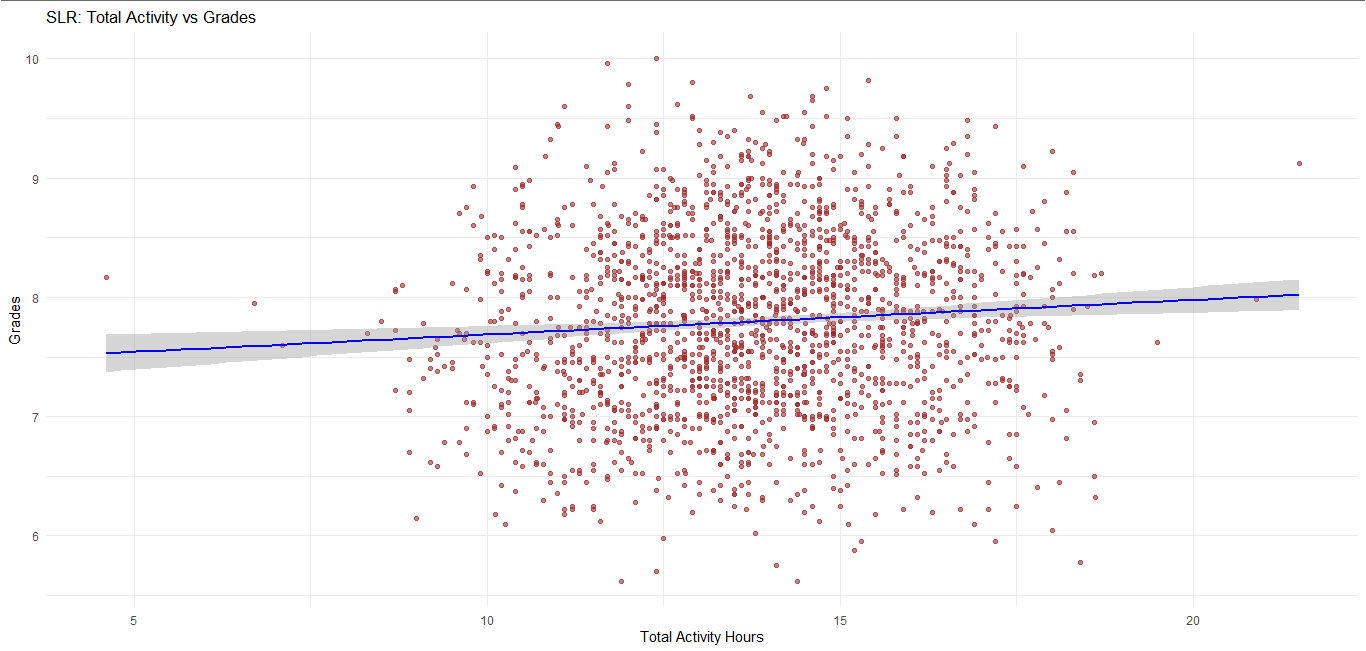


Figure (Plot SLR MODEL 6 GRADES vs ALL ACTIVITIES)

### Comparison Table: SLR Models Predicting Grades

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Predictor Variable | Estimate (Slope) | Std. Error | p-value | R² | Adj. R² | Residual SE | Significance |
| Study Hours Per Day | **+0.384** | 0.00857 | **<**  **2e-16** | **0.5125** | **0.5122** | 0.5295 | \*\*\* Strongest predictor |
| Sleep Hours Per Day | -0.003 | 0.01196 | 0.792 | 0.00004 | -0.0005 | 0.7584 | Not significant |
| Social Hours Per Day | -0.0377 | 0.01024 | **0.00024** | 0.0070 | 0.0065 | 0.7557 | \*\* Weak negative |
| Physical Activity Hours | -0.0998 | 0.00659 | **< 2e-16** | 0.1071 | 0.1066 | 0.7167 | \*\* Strong negative |
| Extracurricular Hours | -0.0219 | 0.01499 | 0.144 | 0.0011 | 0.0006 | 0.7580 | Not significant |
| Total Activity | +0.0291 | 0.00846 | **0.00060** | 0.0061 | 0.0056 | 0.7561 | \*\* Weak positive |

#### Conclusion:

* **Best predictor**: Study Hours Per Day (high effect size, R², and significance).
* **Second strongest**: Physical Activity Hours Per Day, but in a **negative** direction.
* **Negligible or no impact**: Sleep Hours Per Day, Extracurricular Hours, Total Activity (despite significance, effect is minimal).
* **Social time** slightly **hurts** grades, but very weakly.

## MLR (Multi Linear Regression):

First, we will be doing individual model interpretation. And then will do a collective as well. To get the overall summary of relation of grades with all the activities. There are 3 models with 2, 3 and 4 variables. And the final complete 4th model has all the 5 variables. That are not categorical. For incorporating the non-categorical variables, we have made the 5th model which is faceted into parts based on male and female (genders) for better understanding.

**Note: All these models will be related to grades. Thanks.**

### Model 1: Study Hours + Sleep Hours

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **R² = 0.5130** (51.3% variance explained)
* **Study Hours** are highly significant (**p < 2e-16**), positive impact.
* **Sleep Hours** is **not significant** (p = 0.148), suggesting it adds little after controlling for study time.

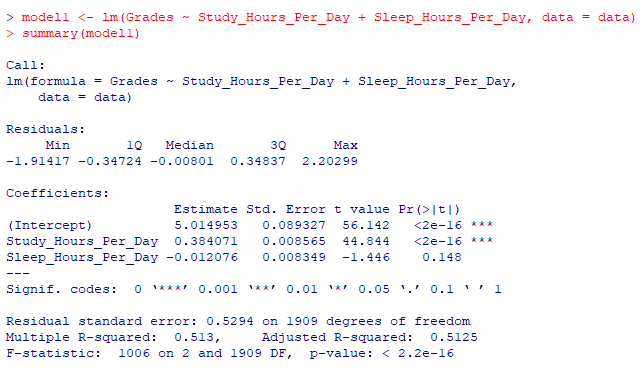


Figure (Summary MLR MODEL 1 GRADES + Study Hours + Sleep Hours)

#### Plot:

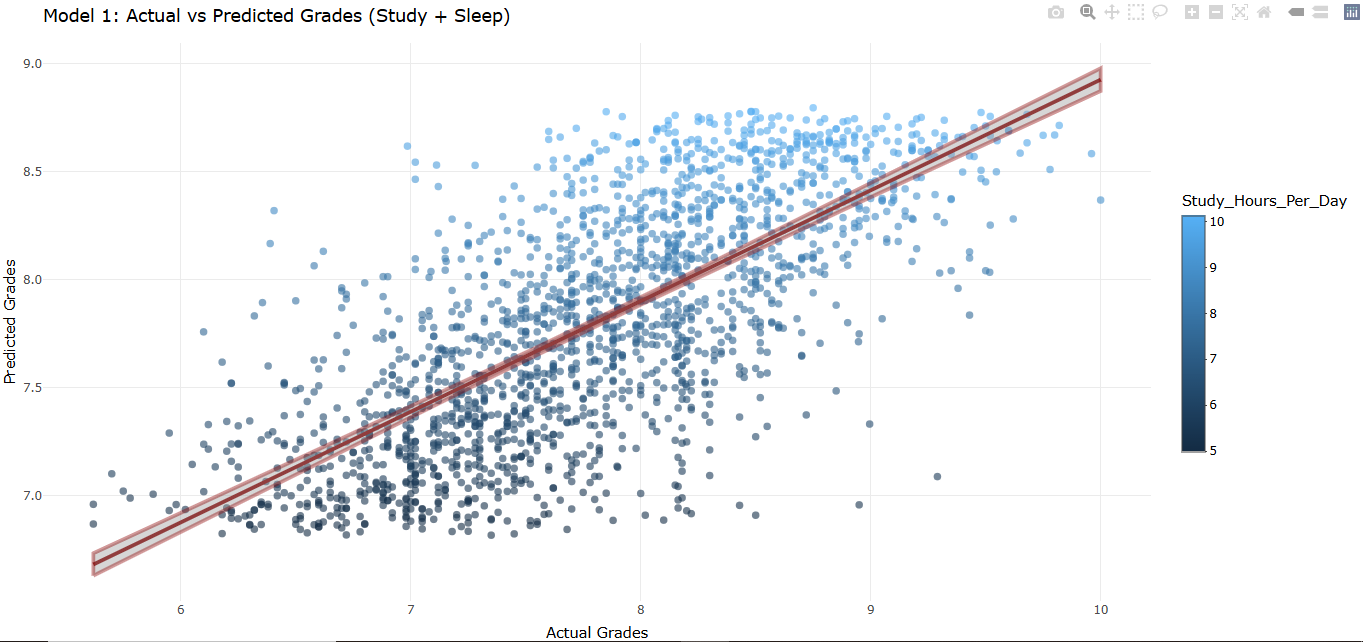


Figure (Plot MLR MODEL 1 GRADES + Study Hours + Sleep Hours)

### Model 2: Study Hours + Sleep Hours + Social Hours

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **R² barely increases** to 0.5133.
* **Social Hours** are **not significant** (p = 0.285).
* No practical improvement from MLR 1.

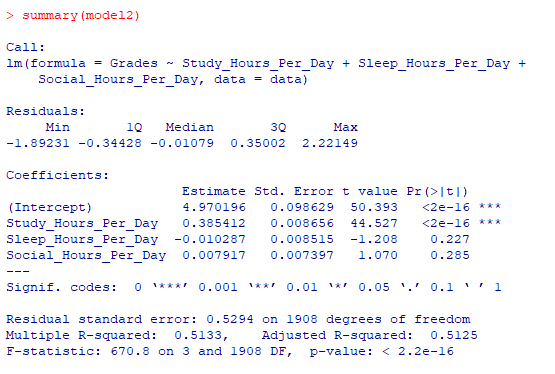


Figure (Summary MLR MODEL 2 GARDES + Study Hours + Sleep Hours + Social Hours)

#### Plot:

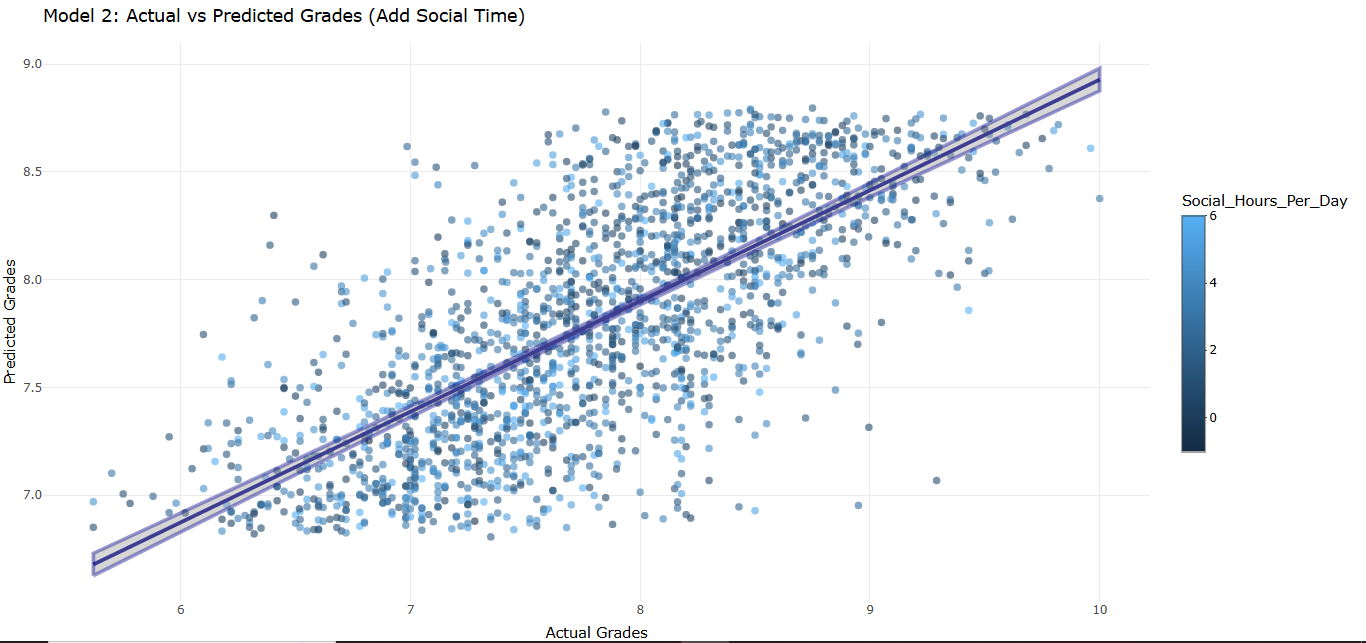


Figure (Plot MLR MODEL 2 GARDES + Study Hours + Sleep Hours + Social Hours)

### Model 3: Study Hours + Sleep Hours + Social Hours + Extracurricular Time

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **R² = 0.5142**, a slight bump.
* **Extracurricular** becomes **marginally significant** (p = 0.0596).
* Still, **Study Hours** remain the dominant factor.

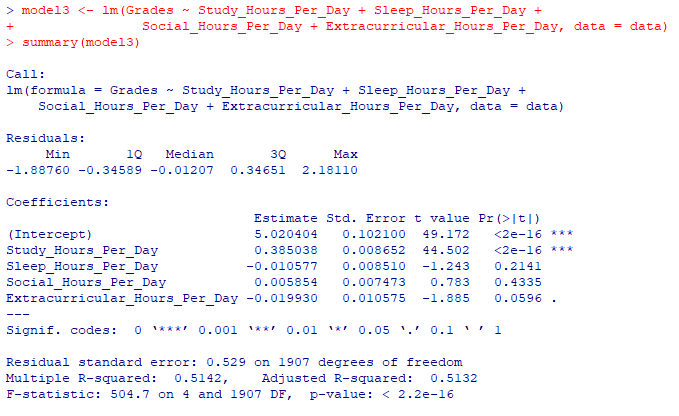


Figure (Summary MLR MODEL 3 GRADES + Study Hours + Sleep Hours + Social Hours + Extracurricular Time)

#### Plot:

A line of blue dots

AI-generated content may be incorrect.

Figure (Plot MLR MODEL 3 GRADES + Study Hours + Sleep Hours + Social Hours + Extracurricular Time)

### Model 4: Full Model (Add Physical Activity)

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

* **R² = 0.5148**, **Adj. R² = 0.5136** – again, minimal improvement.
* **Study Hours** are still **dominant and significant**.
* **Extracurricular Hours** becomes **statistically significant (p = 0.020)** but has a small negative impact.
* Other predictors: either **not significant** or very weak.

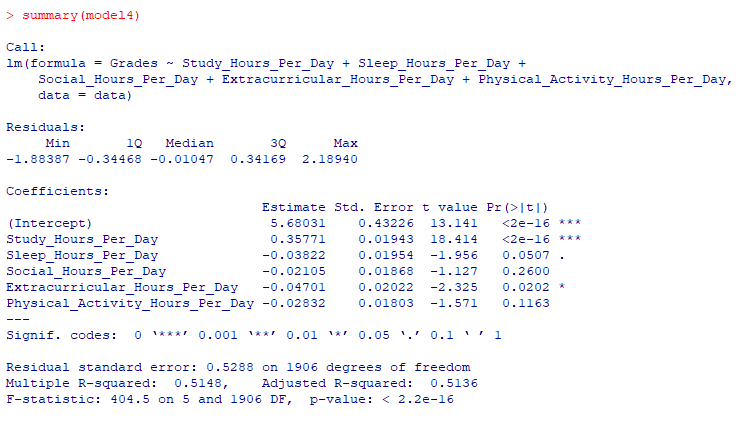


Figure (Summary MLR MODEL 4 - Full Model (Add Physical Activity))

#### Plot:

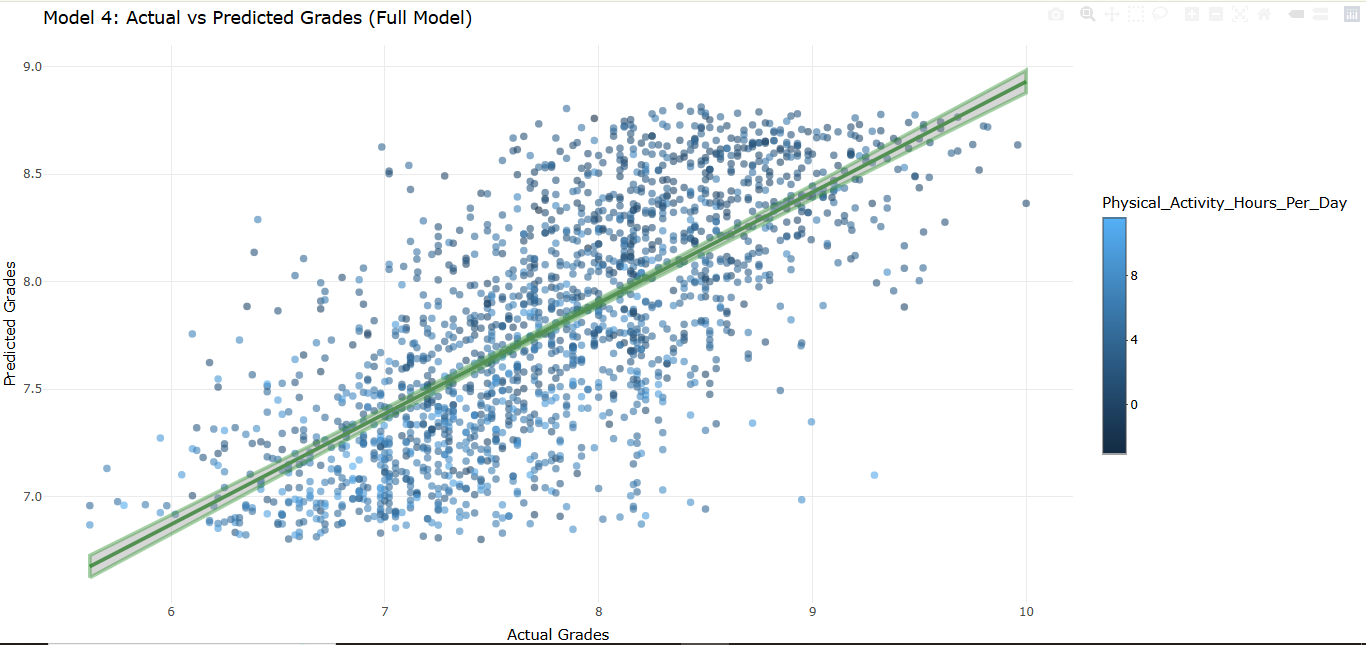


Figure (Plot MLR MODEL 4 - Full Model (Add Physical Activity))

### Model 5: Full Model + Gender + Stress Level

First, we will see the model summary, and then we will plot it to analyze it visually.

#### Summary:

**R² = 0.5161**, **Adjusted R² = 0.5141**: Slight improvement from Model 4.

**Significant Predictors:**

* Study Hours Per Day (**p < 2e-16**): Still the strongest, most consistent predictor.
* Extracurricular Hours Per Day (*p = 0.0179*): Negative impact, now statistically significant.

**Marginal/Not Significant Predictors:**

* Sleep Hours Per Day: Weak evidence (p = 0.0776).
* Gender, Stress Level, Social Hours Per Day, Physical Activity Hours Per Day: **Not significant**.

**Model Complexity vs. Gain**: Despite adding 3 new predictors, the **improvement in explanatory power is minimal (ΔAdj.R² ≈ 0.0005)**.

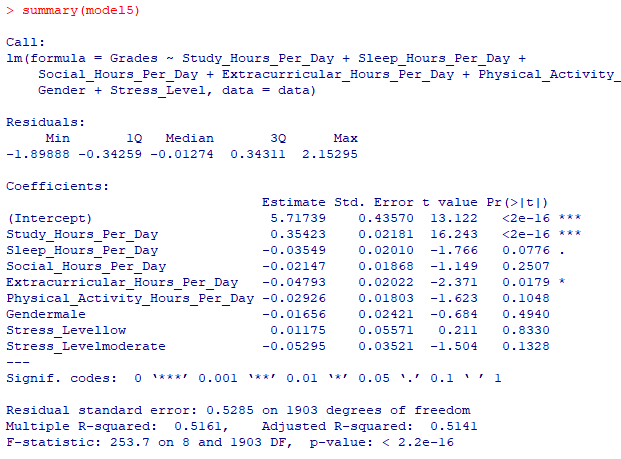


Figure (Summary MLR MODEL 5 - Full Model + Gender + Stress Level)

#### Plot:

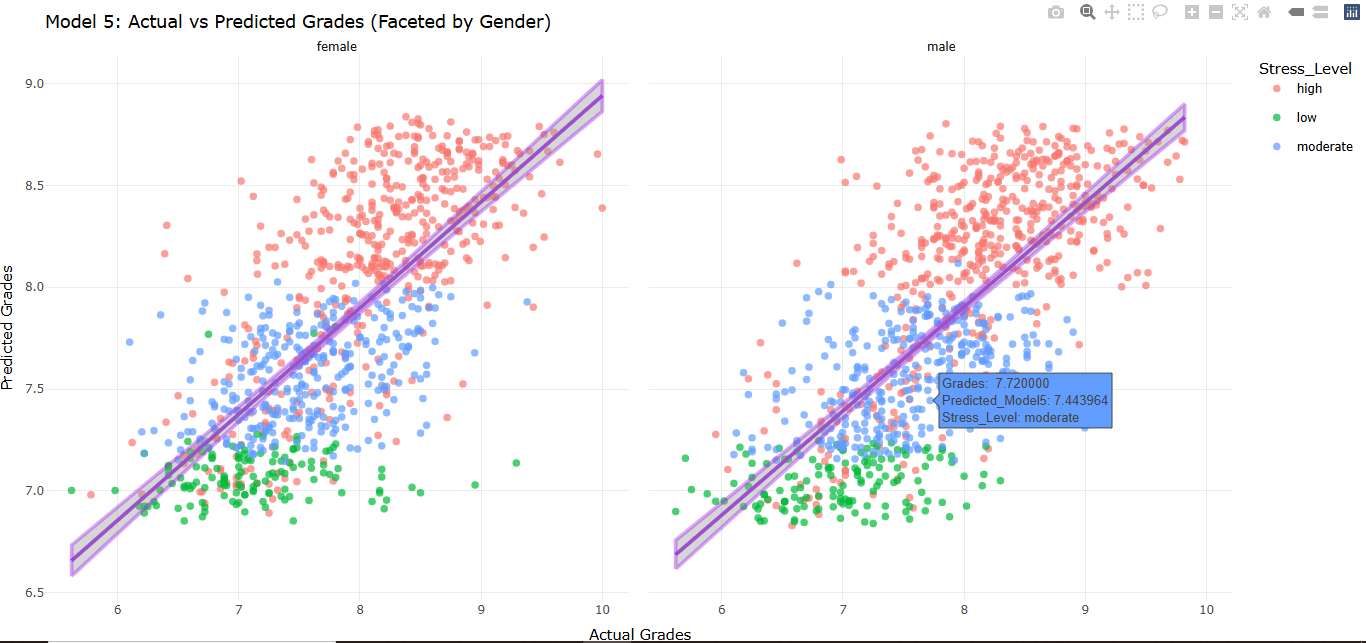


Figure (Plot MLR MODEL 5 - Full Model + Gender + Stress Level)

### Comparison Table: MLR Models 1–5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Predictors | R² | Adj. R² | Residual SE | Significant Predictors | Notes |
| MLR 1 | Study + Sleep | 0.5130 | 0.5125 | 0.5294 | Study (\*\*\*) | Baseline model. Sleep not significant. |
| MLR 2 | + Social Hours | 0.5133 | 0.5125 | 0.5294 | Study (\*\*\*) | Social Hours not significant. |
| MLR 3 | + Extracurricular | 0.5142 | 0.5132 | 0.5290 | Study (\*\*\*), Extracurricular (.) | Slight improvement. Extracurricular marginal (p = 0.0596). |
| MLR 4 | + Physical Activity | 0.5148 | 0.5136 | 0.5288 | Study (\*\**), Extracurricular (*) | Physical activity is not significant. Minor R² gain. |
| MLR 5 | + Gender + Stress Level (as factors) | 0.5161 | 0.5141 | 0.5285 | Study (\*\**), Extracurricular (*); others not significant | Gender and stress levels not significant; minimal R² gain overall. |

### Key Takeaways:

Here are some key takeaways from these models (MLR Model 1-5):

* **Study Hours remain the only robust predictor** throughout all models.
* **Extracurricular hours consistently show a slight negative effect**, becoming significant in later models.
* **Stress Level and Gender do not meaningfully contribute** after accounting for the other variables.
* **Model 5 is the most comprehensive** but offers **diminishing returns** in terms of model performance.

# Conclusion:

We have divided the conclusion into 3 main parts. All of them are mentioned below.

## Summary of Our Study:

Our main goal was to explore **how lifestyle factors affect students’ grades**. We worked with a dataset from Kaggle, which had info on things like how much students study, sleep, socialize, play sports, etc. We cleaned the data, added outliers for testing, and then cleaned it again to see the full impact.

After preprocessing, we had a nice clean dataset of **1912 rows** and **8 variables**. We used **descriptive statistics**, **exploration data analysis (EDA)**, **Simple Linear Regression (SLR)**, and **Multiple Linear Regression (MLR)** to understand the relationships.

## Key Findings:

* **Study hours had a positive impact on grades,** the more a student studied, the better the result.
* **Sleep hours** and **physical activity** also had a slight positive connection with performance.
* **Stress levels** seemed to play a big role. Students with **high stress** usually had **lower grades**.
* **Extracurricular and social hours** didn’t show a very strong correlation, but extremes (too low or too high) affected performance slightly.
* **Gender** didn't have a very strong effect overall — the average grades were similar after cleaning and scaling.

From both SLR and MLR, we saw that **study hours** were the most significant predictor of grades. It consistently showed a strong and positive relationship.

## Limitations of Our Study:

* **Sample Size**: Even though we had 2000+ rows, it’s still not enough to represent *all* students across Pakistan or globally.
* **Fake or Modified Data**: Since we added outliers and missing values ourselves for testing, it's not 100% real-world data anymore.
* **Subjectivity in Stress Levels**: Stress level is self-reported and can vary from person to person. One student’s “moderate” stress could be another’s “high”.
* **No Course Difficulty Factor**: We didn’t know what type of courses students were taking, some could be harder than others, which can affect grades regardless of lifestyle.
* **No Time Dimension**: Our data was static. We didn’t track how a student's performance changed over time — so no trend analysis.

## Final Thoughts:

Even with the limitations, this study gave us a solid idea of how student habits and lifestyle impact their academic performance. It also showed us how important **data cleaning** and **preprocessing** are because messy data can lead to completely wrong insights.

This project also helped us get hands-on experience with R, visualizations, and building basic regression models. And honestly, it was quite fun too now we understand how real data analysis works, instead of just reading about it in books!