Selected the topic: Life style factors &their impact on grades/gpa

Why select this topic? Gpt add the details.

1st got the data from Kaggle.

It was clean.

We uncleaned it. Added some outliers. Removed some values. Total 2010+ rows with 8 columns

Started preprocessing. Loaded the data in to R. command used for navigating to the directory and then reading the file.

1. Get current working directory:

getwd()

2. Set a new working directory:

setwd("C:/path/to/your/folder") # Windows

3. List files in the current directory:

list.files()

data <- read.csv("lifestyle\_gpa.csv") # Load the dataset

View(data) # View it in a spreadsheet format

str(data) # Understand data types

summary(data) # View basic stats

colSums(is.na(data)) # Checked missing values per column

got issues with the data read. Two of the columns were read in character but they were in numerical values. So fixed them.

# Convert numeric columns that are mistakenly read as characters

data$Study\_Hours\_Per\_Day <- as.numeric(data$Study\_Hours\_Per\_Day)

data$Grades <- as.numeric(data$Grades)

here we also got a warning saying that missing will be replaced by NA. so we will be replacing them with mean value as we move forward with the analysis.

then converted categorical data to factors

# Convert categorical columns to factors

data$Stress\_Level <- as.factor(data$Stress\_Level)

data$Gender <- as.factor(data$Gender)

then we moved to the next step

Step 2: Handle Missing Valuesv

Replaced all the NA values

# Replace NA in numeric columns with column mean

data$Extracurricular\_Hours\_Per\_Day[is.na(data$Extracurricular\_Hours\_Per\_Day)] <- mean(data$Extracurricular\_Hours\_Per\_Day, na.rm = TRUE)

data$Sleep\_Hours\_Per\_Day[is.na(data$Sleep\_Hours\_Per\_Day)] <- mean(data$Sleep\_Hours\_Per\_Day, na.rm = TRUE)

data$Social\_Hours\_Per\_Day[is.na(data$Social\_Hours\_Per\_Day)] <- mean(data$Social\_Hours\_Per\_Day, na.rm = TRUE)

data$Physical\_Activity\_Hours\_Per\_Day[is.na(data$Physical\_Activity\_Hours\_Per\_Day)] <- mean(data$Physical\_Activity\_Hours\_Per\_Day, na.rm = TRUE)

now the next step was to remove the duplicates.

Step 3: Remove Duplicate Records

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

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

then I used the summary(data) command to get info about the current situation of the data set, and got amazing things. The gender and the stress level where there as categorical data, some additional values were also present as mentioned in snapshots.

I used the commands

**# Convert to lowercase for consistency**

**data$Gender <- tolower(data$Gender)**

**# Fix common typos and unknowns**

**data$Gender[data$Gender %in% c("femal", "female", "femaale", "f")] <- "female"**

**data$Gender[data$Gender %in% c("male", "m")] <- "male"**

**data$Gender[!data$Gender %in% c("male", "female")] <- NA # Mark others as missing**

**# Optional: Fill NA with most frequent value**

**most\_common\_gender <- names(sort(table(data$Gender), decreasing = TRUE))[1]**

**data$Gender[is.na(data$Gender)] <- most\_common\_gender**

# Convert to lowercase

**data$Stress\_Level <- tolower(data$Stress\_Level)**

**# Replace typos and spaces**

**data$Stress\_Level <- trimws(data$Stress\_Level) # remove extra spaces**

**data$Stress\_Level[data$Stress\_Level %in% c("high", "hig", "h")] <- "high"**

**data$Stress\_Level[data$Stress\_Level %in% c("moderate", "mod", "moder")] <- "moderate"**

**data$Stress\_Level[data$Stress\_Level %in% c("low", "l")] <- "low"**

**data$Stress\_Level[!data$Stress\_Level %in% c("low", "moderate", "high")] <- NA**

**# Optional: Fill NA with most frequent level**

**most\_common\_stress <- names(sort(table(data$Stress\_Level), decreasing = TRUE))[1]**

**data$Stress\_Level[is.na(data$Stress\_Level)] <- most\_common\_stress**

now since the data is now cleaned and we the rows are reduced do 1912, after the removel of duplicates, outliers, and other stuff. We saved the file so that we can have a copy of both the clean and unclean data sets. Below was the command used.

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

We used summary tags to get insights about the descriptive details of the data set.

**Summary(data)**

**# Save summary output to a text file [just for data backup]**

**sink("summary\_output.txt")**

**summary(data)**

**sink()**

On suggestion of chatgpt we used the other command that actually comes with another package.

**install.packages("psych") # if not already installed**

**library(psych)**

**Describe(data)** the out put of both are mentioned in snapshots.

**# Save describe output to a text file**

sink("describe\_output.txt")

print(describe(data))

sink()

instead of using normal histrograms and charts I decided to use some advance libraries for better and interactive graphs. For tha tpuropse I used these libraries

**install.packages("ggplot2")**

**install.packages("plotly")**

**library(ggplot2)**

**library(plotly)**

here are the relationships:

1. Interactive grade distribution x and y

**p1 <- ggplot(data, aes(x = Grades)) +**

**geom\_histogram(binwidth = 1, fill = 'skyblue', color = 'black') +**

**labs(title = "Distribution of Grades", x = "Grades", y = "Count")**

**ggplotly(p1)**

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

1. Interactive Boxplot: Grades by Gender

**p2 <- ggplot(data, aes(x = Gender, y = Grades, fill = Gender)) +**

**geom\_boxplot() +**

**labs(title = "Grades by Gender", x = "Gender", y = "Grades")**

**ggplotly(p2)**

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

1. Interactive Boxplot: Grades by Stress Level

**p3 <- ggplot(data, aes(x = Stress\_Level, y = Grades, fill = Stress\_Level)) +**

**geom\_boxplot() +**

**labs(title = "Grades by Stress Level", x = "Stress Level", y = "Grades")**

**ggplotly(p3)**

**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.**

1. **Study Hours vs Grades:**

**p4 <- ggplot(data, aes(x = Study\_Hours\_Per\_Day, y = Grades)) +**

**geom\_point(color = "steelblue") +**

**geom\_smooth(method = "lm", se = FALSE, color = "red") +**

**labs(title = "Study Hours vs Grades", x = "Study Hours", y = "Grades")**

**ggplotly(p4)**

**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 study hours alone do not determine grades — efficiency, comprehension, and individual learning styles also play important roles.**

1. **Sleep Hours vs Grades**

**p5 <- ggplot(data, aes(x = Sleep\_Hours\_Per\_Day, y = Grades)) +**

**geom\_point(color = "purple") +**

**geom\_smooth(method = "lm", se = FALSE, color = "red") +**

**labs(title = "Sleep Hours vs Grades", x = "Sleep Hours", y = "Grades")**

**ggplotly(p5)**

**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.**

1. **Social Hours vs Grades**

**p6 <- ggplot(data, aes(x = Social\_Hours\_Per\_Day, y = Grades)) +**

**geom\_point(color = "orange") +**

**geom\_smooth(method = "lm", se = FALSE, color = "red") +**

**labs(title = "Social Hours vs Grades", x = "Social Hours", y = "Grades")**

**ggplotly(p6)**

**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 on 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.**

1. **Physical Activity vs Grades:**

**p7 <- ggplot(data, aes(x = Physical\_Activity\_Hours\_Per\_Day, y = Grades)) +**

**geom\_point(color = "green") +**

**geom\_smooth(method = "lm", se = FALSE, color = "red") +**

**labs(title = "Physical Activity vs Grades", x = "Physical Activity Hours", y = "Grades")**

**ggplotly(p7)**

**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.**

1. **Extracurricular Hours vs Grades**

**p8 <- ggplot(data, aes(x = Extracurricular\_Hours\_Per\_Day, y = Grades)) +**

**geom\_point(color = "coral") +**

**geom\_smooth(method = "lm", se = FALSE, color = "red") +**

**labs(title = "Extracurricular Hours vs Grades", x = "Extracurricular Hours", y = "Grades")**

**ggplotly(p8)**

**From the scatter plot titled "Extracurricular Hours vs Grades", we can draw the following inferences:**

1. **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.**
2. **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.**
3. **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.**
4. **Conclusion: Participating in extracurricular activities, based on this data, does not have a significant impact (positive or negative) on academic grades.**
5. **Stress Level by Gender:**

**P9 <- ggplot(data, aes(x = Stress\_Level, fill = Gender)) +**

**geom\_bar(position = "dodge") +**

**labs(title = "Stress Level Count by Gender", x = "Stress Level", y = "Count")**

**ggplotly(p9)**

**Inference: Male experience noticeably more stress—this might affect performance or well-being.**

1. **Study Hours by Stress Level:**

**p10 <- ggplot(data, aes(x = Stress\_Level, y = Study\_Hours\_Per\_Day, fill = Stress\_Level)) +**

**geom\_boxplot() +**

**labs(title = "Study Hours by Stress Level", x = "Stress Level", y = "Study Hours")**

**ggplotly(p10)**

**Inference: Students with high stress tend to study the most, with a relatively widespread.**

**Inference: Moderately stressed students study less than highly stressed ones but more than low-stress students. Spread is narrower.**

**Inference: Students with low stress study the least, with the tightest distribution and lowest hours overall.**

1. **Sleep Hours by Stress Level**

**p11 <- ggplot(data, aes(x = Stress\_Level, y = Sleep\_Hours\_Per\_Day, fill = Stress\_Level)) +**

**geom\_boxplot() +**

**labs(title = "Sleep Hours by Stress Level", x = "Stress Level", y = "Sleep Hours")**

**ggplotly(p11)**

**more sleep hours when stress level is low, and less sleep hours when stress is really high.**

1. **Physical Activity by Gender**

**p12 <- ggplot(data, aes(x = Gender, y = Physical\_Activity\_Hours\_Per\_Day, fill = Gender)) +**

**geom\_boxplot() +**

**labs(title = "Physical Activity by Gender", x = "Gender", y = "Physical Activity")**

**ggplotly(p12)**

**Inference: Males tend to be more physically active than the other gender. As majority of them fall above the median while the females majority falls below the median.**

1. **Social Hours by Gender**:

p13 <- ggplot(data, aes(x = Gender, y = Social\_Hours\_Per\_Day, fill = Gender)) +

geom\_boxplot() +

labs(title = "Social Hours by Gender", x = "Gender", y = "Social Hours")

ggplotly(p13)

**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.

1. Combined Activity Time vs Grades

data$Total\_Activity <- data$Study\_Hours\_Per\_Day + data$Extracurricular\_Hours\_Per\_Day + data$Physical\_Activity\_Hours\_Per\_Day

p15 <- ggplot(data, aes(x = Total\_Activity, y = Grades)) +

geom\_point(color = "brown") +

geom\_smooth(method = "lm", se = FALSE, color = "blue") +

labs(title = "Total Productive Activity vs Grades", x = "Total Activity Hours", y = "Grades")

ggplotly(p15)

**Key Observations:**

1. **Positive Trend**:
   * The **blue regression line** has a **slight positive slope**, indicating that **more productive activity hours are generally associated with slightly higher grades**.
   * However, the trend is very **weak**, suggesting **total activity hours have only a modest impact** on grades.
2. **Spread of Data**:
   * There's **significant scatter** in the data, showing that grades vary widely at nearly every level of total activity hours.
   * This implies that **other factors** likely have stronger effects on grades than productive time alone.
3. **Range**:
   * Total productive hours mostly fall between **8 to 18 hours per day**.
   * Grades cluster around **7.5 to 8.5**, with few extremes.
4. **Outliers**:
   * A few students with low activity hours still achieve high grades, and vice versa—this supports the idea that **efficiency, not just quantity**, matters.

**Conclusion:**

While there's a slight positive correlation between productive activity hours and grades, it's not strong. Productive time helps, but it's likely **not the sole predictor** of academic success.

**Heat Map:**

**Step 1: Correlation Analysis**

We’ll compute the **correlation matrix** for numeric columns (e.g., Grades, Study\_Hours\_Per\_Day, etc.), and visualize it using a **heatmap**.

# Load necessary libraries

library(dplyr)

library(plotly)

# Select numeric columns (excluding Student\_ID)

numeric\_data <- data[, sapply(data, is.numeric)]

numeric\_data <- subset(numeric\_data, select = -Student\_ID)

# Compute correlation matrix

cor\_matrix <- cor(numeric\_data, use = "complete.obs")

# Create interactive correlation heatmap

heatmap\_plot <- plot\_ly(

x = colnames(cor\_matrix),

y = rownames(cor\_matrix),

z = cor\_matrix,

type = "heatmap",

colorscale = "Viridis"

)

# Display the heatmap

heatmap\_plot

see heat-map.png in the correlation directory.

**Step 7: Regression Analysis**

We’ll begin with a **multiple linear regression** to understand how various lifestyle factors affect grades.

# Linear regression: Grades predicted by lifestyle factors

model <- lm(Grades ~ Study\_Hours\_Per\_Day + Sleep\_Hours\_Per\_Day +

Social\_Hours\_Per\_Day + Extracurricular\_Hours\_Per\_Day +

Physical\_Activity\_Hours\_Per\_Day, data = data)

# Summary of the model

summary(model)

this model explains about **51.5%** of the variance in grades, meaning **48.5% is unexplained**, possibly due to:

**Factors Not Included:**

* **Mental health** (beyond stress level)
* **Motivation, learning environment, socioeconomic status**
* **Teaching quality, course difficulty**
* **Family background or support**
* **Diet, screen time, etc.**

These are **non-numeric** or **not captured** in your dataset, so the model can't account for them.

**What This Means Practically:**

* **Study time** remains the strongest and most consistent predictor of Grades.
* **More sleep**, **extracurriculars**, and **physical activity** tend to slightly reduce grades, but not drastically.

**Interpretation of Key Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Estimate | Impact on Grades | Significance |
| Study\_Hours\_Per\_Day | +0.358 | Higher study time = ↑ grades | \*\*\* Highly significant |
| Sleep\_Hours\_Per\_Day | -0.038 | Weak negative link | . Marginally significant |
| Extracurricular\_Hours\_Per\_Day | -0.047 | Negative impact | \* Significant |
| Others | Minor negative or no impact | Not significant |  |

Residual vs fitted:

# Residuals vs Fitted

plot(model$fitted.values, model$residuals,

xlab = "Fitted Values", ylab = "Residuals",

main = "Residuals vs Fitted")

abline(h = 0, col = "red")

**What your plot shows**:

* **Fairly good randomness**: No strong curve or pattern — suggests linearity is *mostly* satisfied.
* **Some spread changes**: Slightly more vertical spread around higher fitted values (~8.5), hinting at **mild heteroscedasticity**.
* No strong outliers or funnel shapes — which is good.

**# Q-Q Plot**

**qqnorm(model$residuals)**

**qqline(model$residuals, col = "blue")**

**SIMPLE LINEAR REGRESSION:**

1. **Grades ~ Study\_Hours\_Per\_Day**

**Call:**

**lm(formula = Grades ~ Study\_Hours\_Per\_Day, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-1.8981 -0.3456 -0.0060 0.3445 2.2145**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.926589 0.065184 75.58 <2e-16 \*\*\***

**Study\_Hours\_Per\_Day 0.383775 0.008565 44.81 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.5295 on 1910 degrees of freedom**

**Multiple R-squared: 0.5125, Adjusted R-squared: 0.5122**

**F-statistic: 2008 on 1 and 1910 DF, p-value: < 2.2e-16**

1. **SLR of Grades ~ Sleep\_Hours\_Per\_Day**

**Call:**

**lm(formula = Grades ~ Sleep\_Hours\_Per\_Day, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.1762 -0.5529 -0.0027 0.5224 2.2066**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 7.82018 0.09135 85.606 <2e-16 \*\*\***

**Sleep\_Hours\_Per\_Day -0.00315 0.01196 -0.263 0.792**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7584 on 1910 degrees of freedom**

**Multiple R-squared: 3.634e-05, Adjusted R-squared: -0.0004872**

**F-statistic: 0.06941 on 1 and 1910 DF, p-value: 0.7922**

1. **Grades ~ Social\_Hours\_Per\_Day**

**Call:**

**lm(formula = Grades ~ Social\_Hours\_Per\_Day, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.24789 -0.56373 0.00139 0.52558 2.27735**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 7.89806 0.03254 242.717 < 2e-16 \*\*\***

**Social\_Hours\_Per\_Day -0.03771 0.01024 -3.682 0.000238 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7557 on 1910 degrees of freedom**

**Multiple R-squared: 0.007046, Adjusted R-squared: 0.006526**

**F-statistic: 13.55 on 1 and 1910 DF, p-value: 0.0002383**

1. **Grades ~ Physical\_Activity\_Hours\_Per\_Day**

**Call:**

**lm(formula = Grades ~ Physical\_Activity\_Hours\_Per\_Day, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.18691 -0.52711 0.00385 0.50344 2.14173**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 8.225886 0.032766 251.05 <2e-16 \*\*\***

**Physical\_Activity\_Hours\_Per\_Day -0.099756 0.006592 -15.13 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7167 on 1910 degrees of freedom**

**Multiple R-squared: 0.1071, Adjusted R-squared: 0.1066**

**F-statistic: 229 on 1 and 1910 DF, p-value: < 2.2e-16**

1. **Grades ~ Extracurricular\_Hours\_Per\_Day**

**Call:**

**lm(formula = Grades ~ Extracurricular\_Hours\_Per\_Day, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.19826 -0.55320 0.00226 0.51926 2.21678**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 7.84017 0.03451 227.162 <2e-16 \*\*\***

**Extracurricular\_Hours\_Per\_Day -0.02190 0.01499 -1.461 0.144**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.758 on 1910 degrees of freedom**

**Multiple R-squared: 0.001117, Adjusted R-squared: 0.0005939**

**F-statistic: 2.136 on 1 and 1910 DF, p-value: 0.1441**

1. **Grades ~ Total\_Activity**

**lm(formula = Grades ~ Total\_Activity, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-2.19478 -0.55774 -0.00297 0.52063 2.24339**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 7.395971 0.117854 62.756 < 2e-16 \*\*\***

**Total\_Activity 0.029084 0.008464 3.436 0.000603 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.7561 on 1910 degrees of freedom**

**Multiple R-squared: 0.006144, Adjusted R-squared: 0.005624**

**F-statistic: 11.81 on 1 and 1910 DF, p-value: 0.0006026**