Module 3: R Practice

Name: Aishwarya Doni

College of Professional Studies, Northeastern University

ALY6010: Probability Theory and Introductory Statistics

Professor: Kenneth Parker

November 17, 2024

Introduction

Dataset Information

The chosen dataset for this analysis is the **Student Lifestyle Dataset**, containing 2000 entries and 8 columns. This dataset provides information on various aspects of students' academic performance, daily habits, and personal characteristics. Below is a summary of the dataset's key features:

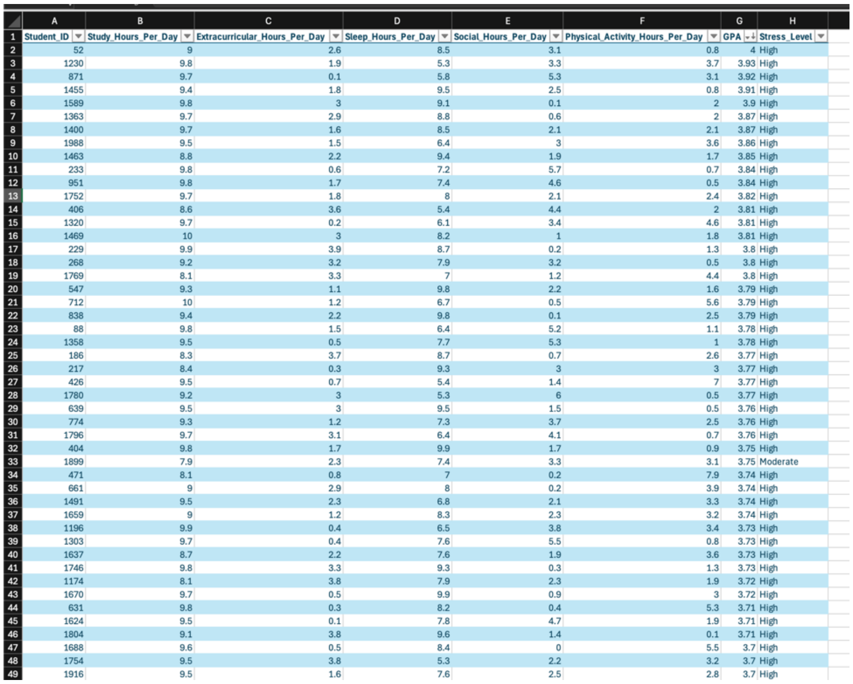
1. **Structure**
    - **Rows** (Observations): 62 students.
    - **Columns** (Variables): 12 attributes capturing demographic, academic, and lifestyle data.

2. **Variables with their description**
    - **Student_ID**: Unique identifier for each student.
    - **Age**: Age of the student (integer).
    - **Gender**: Gender of the student (categorical: Male/Female).
    - **Country**: Country of residence (categorical).
    - **Major**: Academic major (categorical).
    - **GPA**: Grade Point Average (numeric, scale: 0.0 to 4.0).
    - **Study_Hours_Per_Week**: Weekly hours spent studying (numeric).
    - **Sleep_Hours_Per_Day**: Daily hours of sleep (numeric).
    - **Physical_Activity_Hours_Per_Week**: Weekly hours spent on physical activities (numeric).
    - **Part_Time_Job**: Whether the student has a part-time job (categorical: Yes/No).
    - **Social_Activity_Hours_Per_Week**: Weekly hours spent on social activities (numeric).
    - **Device_Usage_Hours_Per_Day**: Daily hours spent using electronic devices (numeric).

3. **Dataset Features**
    - **Diverse Variables:**
        1. Mix of numerical (e.g., GPA, sleep hours) and categorical data (e.g., gender, country).
        2. Covers academic, health, and social dimensions of student life.

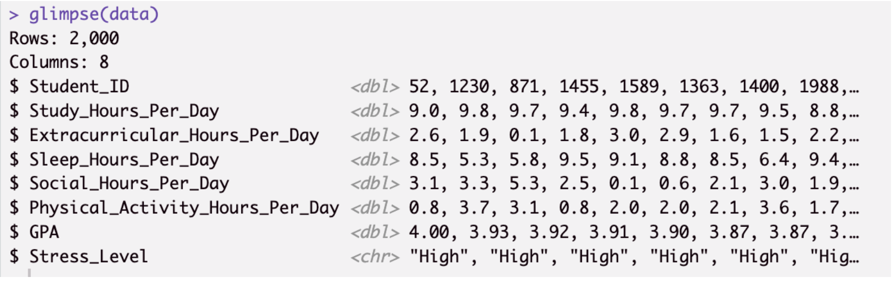
    - **Focus Areas:**
        1. Academic performance (GPA, study hours).
        2. Health-related metrics (sleep hours, physical activity).
        3. Lifestyle and social habits (device usage, social activities).

4. Here’s the glimpse of dataset from excel sheet (which has been imported to R Studio)



	A	B	C	D	E	F	G	H
1	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
2	52	9	2.6	8.5	3.1	0.8	4	High
3	1230	9.8	1.9	5.3	3.3	3.7	3.93	High
4	871	9.7	0.1	5.8	5.3	3.1	3.92	High
5	1455	9.4	1.8	9.5	2.5	0.8	3.91	High
6	1589	9.8	3	9.1	0.1	2	3.9	High
7	1363	9.7	2.9	8.8	0.6	2	3.87	High
8	1400	9.7	1.6	8.5	2.1	2.1	3.87	High
9	1988	9.5	1.5	6.4	3	3.6	3.86	High
10	1463	8.8	2.2	9.4	1.9	1.7	3.85	High
11	233	9.8	0.6	7.2	5.7	0.7	3.84	High
12	951	9.8	1.7	7.4	4.6	0.5	3.84	High
13	1752	9.7	1.8	8	2.1	2.4	3.82	High
14	406	8.6	3.6	5.4	4.4	2	3.81	High
15	1320	9.7	0.2	6.1	3.4	4.6	3.81	High
16	1469	10	3	8.2	1	1.8	3.81	High
17	229	9.9	3.9	8.7	0.2	1.3	3.8	High
18	268	9.2	3.2	7.9	3.2	0.5	3.8	High
19	1769	8.1	3.3	7	1.2	4.4	3.8	High
20	547	9.3	1.1	9.8	2.2	1.6	3.79	High
21	712	10	1.2	6.7	0.5	5.6	3.79	High
22	838	9.4	2.2	9.8	0.1	2.5	3.79	High
23	88	9.8	1.5	6.4	5.2	1.1	3.78	High
24	1358	9.5	0.5	7.7	5.3	1	3.78	High
25	186	8.3	3.7	8.7	0.7	2.6	3.77	High
26	217	8.4	0.3	9.3	3	3	3.77	High
27	426	9.5	0.7	5.4	1.4	7	3.77	High
28	1780	9.2	3	5.3	6	0.5	3.77	High
29	639	9.5	3	9.5	1.5	0.5	3.76	High
30	774	9.3	1.2	7.3	3.7	2.5	3.76	High
31	1796	9.7	3.1	6.4	4.1	0.7	3.76	High
32	404	9.8	1.7	9.9	1.7	0.9	3.75	High
33	1899	7.9	2.3	7.4	3.3	3.1	3.75	Moderate
34	471	8.1	0.8	7	0.2	7.9	3.74	High
35	661	9	2.9	8	0.2	3.9	3.74	High
36	1491	9.5	2.3	6.8	2.1	3.3	3.74	High
37	1659	9	1.2	8.3	2.3	3.2	3.74	High
38	1196	9.9	0.4	6.5	3.8	3.4	3.73	High
39	1303	9.7	0.4	7.6	5.5	0.8	3.73	High
40	1637	8.7	2.2	7.6	1.9	3.6	3.73	High
41	1746	9.8	3.3	9.3	0.3	1.3	3.73	High
42	1174	8.1	3.8	7.9	2.3	1.9	3.72	High
43	1670	9.7	0.5	9.9	0.9	3	3.72	High
44	631	9.8	0.3	8.2	0.4	5.3	3.71	High
45	1624	9.5	0.1	7.8	4.7	1.9	3.71	High
46	1804	9.1	3.8	9.6	1.4	0.1	3.71	High
47	1688	9.6	0.5	8.4	0	5.5	3.7	High
48	1754	9.5	3.8	5.3	2.2	3.2	3.7	High
49	1916	9.5	1.6	7.6	2.5	2.8	3.7	High

5. Here’s the glimpse of dataset after importing in R Studio (using glimpse() function)



```
> glimpse(data)
Rows: 2,000
Columns: 8
$ Student_ID                      <dbl> 52, 1230, 871, 1455, 1589, 1363, 1400, 1988,…
$ Study_Hours_Per_Day             <dbl> 9.0, 9.8, 9.7, 9.4, 9.8, 9.7, 9.7, 9.5, 8.8,…
$ Extracurricular_Hours_Per_Day   <dbl> 2.6, 1.9, 0.1, 1.8, 3.0, 2.9, 1.6, 1.5, 2.2,…
$ Sleep_Hours_Per_Day             <dbl> 8.5, 5.3, 5.8, 9.5, 9.1, 8.8, 8.5, 6.4, 9.4,…
$ Social_Hours_Per_Day            <dbl> 3.1, 3.3, 5.3, 2.5, 0.1, 0.6, 2.1, 3.0, 1.9,…
$ Physical_Activity_Hours_Per_Day <dbl> 0.8, 3.7, 3.1, 0.8, 2.0, 2.0, 2.1, 3.6, 1.7,…
$ GPA                             <dbl> 4.00, 3.93, 3.92, 3.91, 3.90, 3.87, 3.87, 3.…
$ Stress_Level                    <chr> "High", "High", "High", "High", "High", "Hig…
```

Data Cleaning

- Dataset is loaded and its structure is examined to understand the data types, layout, and the initial data entries. This preliminary inspection, using functions like head() and str(), gives us insight into the dataset’s organization and helps plan further analysis steps.
- In the data cleaning phase, any missing values are addressed by checking for NA values and, if necessary, removing rows with missing data using na.omit(). Data cleaning is crucial because it ensures the accuracy, consistency, and reliability of your dataset, which is the foundation for any valid analysis or insights.

Objective

The student lifestyle dataset is intended to explore statistical relationships between variables related to students' daily activities, academic performance, and stress levels. Based on the assignment instructions, the dataset can be used to perform the following statistical analyses:

1. **One-Sample t-Tests for Mean**

   The goal is to test whether the average GPA of students significantly differs from a hypothesized value of 3.0.

   Hypotheses:

   - **Null Hypothesis (H0H0)**: The population mean GPA is 3.0.
     This assumes no difference between the sample mean and the hypothesized mean.

   - **Alternative Hypothesis (HaHa)**: The population mean GPA is not equal to 3.0.

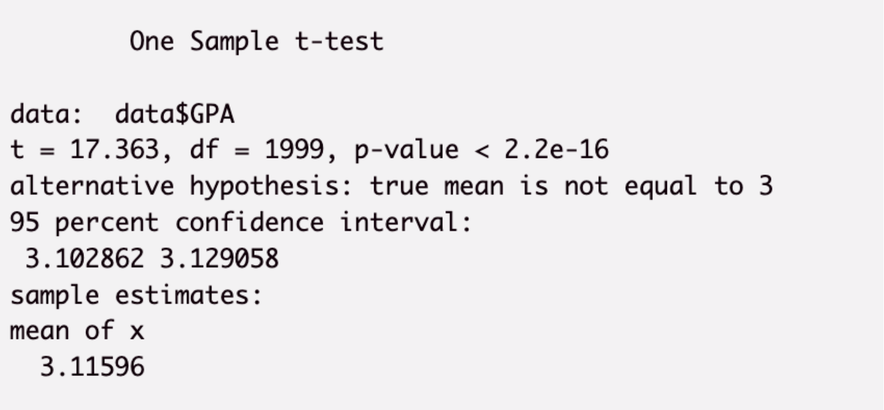
   Output:

   The t-test produces a p-value, a measure of the probability of obtaining results as extreme as the observed, assuming H0H0 is true.

   Decision Rule:

   - If p-value < 0.05, reject H0H0; concluding that the mean GPA significantly differs from 3.0.
   - If p-value ≥ 0.05, fail to reject H0H0; concluding that there is no significant difference.

   R Output



```
        One Sample t-test

data:  data$GPA
t = 17.363, df = 1999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.102862 3.129058
sample estimates:
mean of x
  3.11596
```

Results:

1. **t-Value:**
    - The computed *t*-value is 17.363.
    - This measures how many standard errors the sample mean is away from the hypothesized mean.
2. **Degrees of Freedom (df):**
    - The degrees of freedom for this test are 1999, which corresponds to the sample size minus one.
3. **p-Value:**
    - The p-value is reported as < 2.2e-16 (essentially zero).
    - This is much smaller than the significance level.
4. **Confidence Interval:**
    - The 95% confidence interval for the true mean GPA is [3.102862, 3.129058].
    - This interval does not include the hypothesized mean value of 3.0, further supporting that the mean GPA is different from 3.0.
5. **Sample Mean:**
    - The mean GPA calculated from the data is 3.11596, which is slightly higher than the hypothesized value.

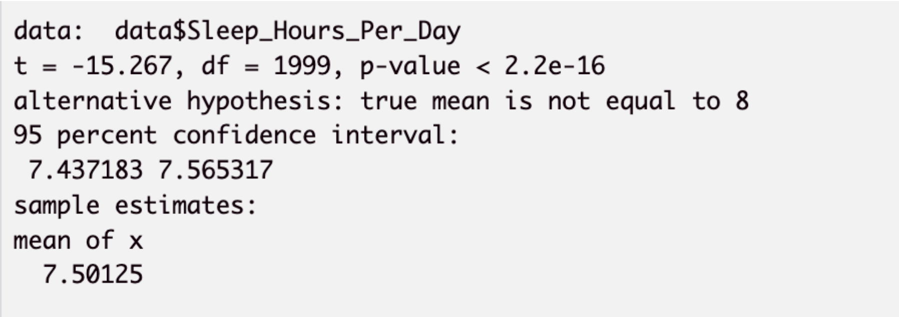
Conclusion

- Since the p-value is less than αα, we reject the null hypothesis (H0H0).
- This means there is strong evidence to conclude that the mean GPA of students significantly differs from 3.0.
- The actual mean GPA is approximately 3.12.

2. **One-Sample t-Test for Sleep Hours**
   The aim of this test is to determine if the mean number of sleep hours per day for students significantly differs from the benchmark value of 8 hours.

   Details:

   - **Null Hypothesis (H0H0):** The mean sleep hours per day is 8 hours.
   - **Alternative Hypothesis (HaHa):** The mean sleep hours per day is not 8 hours.
   - **Significance Level:** 0.05.



```
data:  data$Sleep_Hours_Per_Day
t = -15.267, df = 1999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.437183 7.565317
sample estimates:
mean of x
  7.50125
```

Results

1. **t-Value**:
   - The computed *t*-value is **-15.267**.
   - This indicates the sample mean is far below the hypothesized value in terms of standard errors.
2. **Degrees of Freedom (df)**:
   - The degrees of freedom for this test are **1999**, corresponding to the sample size minus one.
3. **p-Value**:
   - The p-value is reported as **< 2.2e-16** (essentially zero).
   - This is much smaller than the significance level (α=0.05α=0.05).
4. **Confidence Interval**:
   - The 95% confidence interval for the true mean sleep hours is **[7.437183, 7.565317]**.
   - This interval does not include the hypothesized value of 8, further confirming a significant difference.
5. **Sample Mean**:
   - The mean sleep hours calculated from the data is **7.50125**, which is slightly less than the hypothesized value of 8 hours.

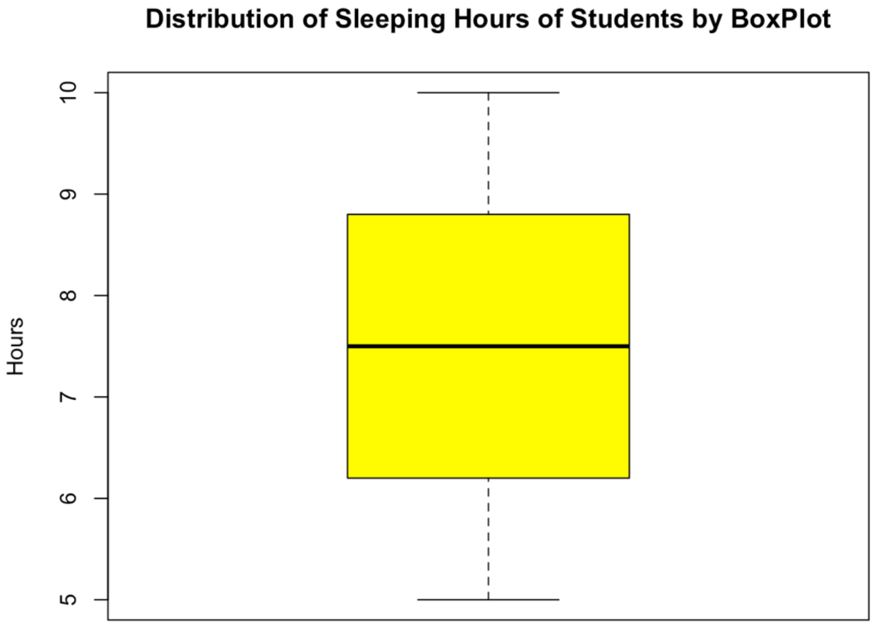
Conclusion

- Since the p-value is less than αα, we **reject the null hypothesis (H0H0)**.
- This means there is strong evidence to conclude that the mean number of sleep hours per day significantly differs from **8 hours**.
- The actual mean sleep hours are approximately **7.50 hours**, which is less than the benchmark.

Visualizations

1. Boxplot

a. Distribution of Sleeping Hours of Students by Boxplot

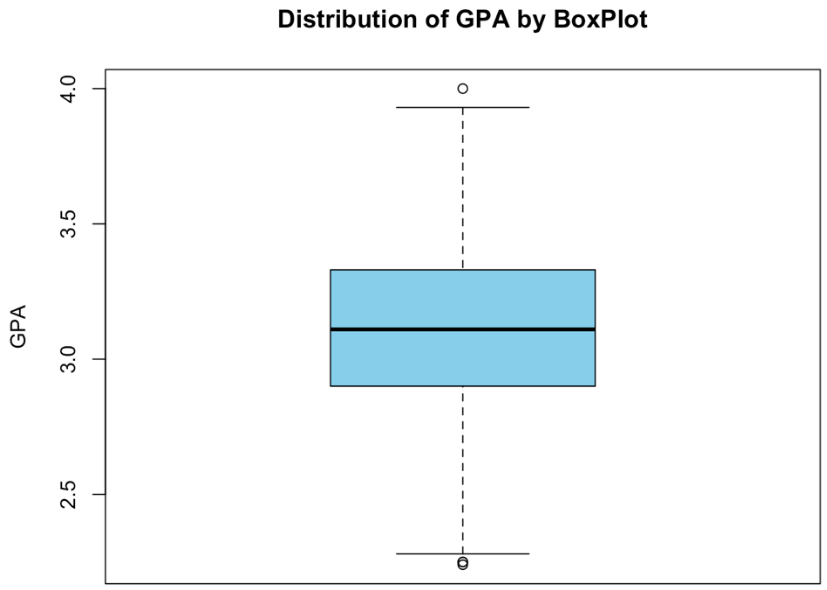


The box plot reveals a distribution of sleeping hours among students with a median of approximately 7.5 hours. The data is moderately spread out, with the middle 50% of students sleeping between 6 and 9 hours. The absence of outliers suggests a relatively consistent sleep pattern among the students.

Additional Insights:

- The IQR of 4 hours indicates a moderate amount of variability in sleeping hours among these students.
- **Central Tendency:** The median being 7.5 hours suggests that a typical student in this group sleeps around 7.5 hours per night.

b. Distribution of GPA by Boxplot



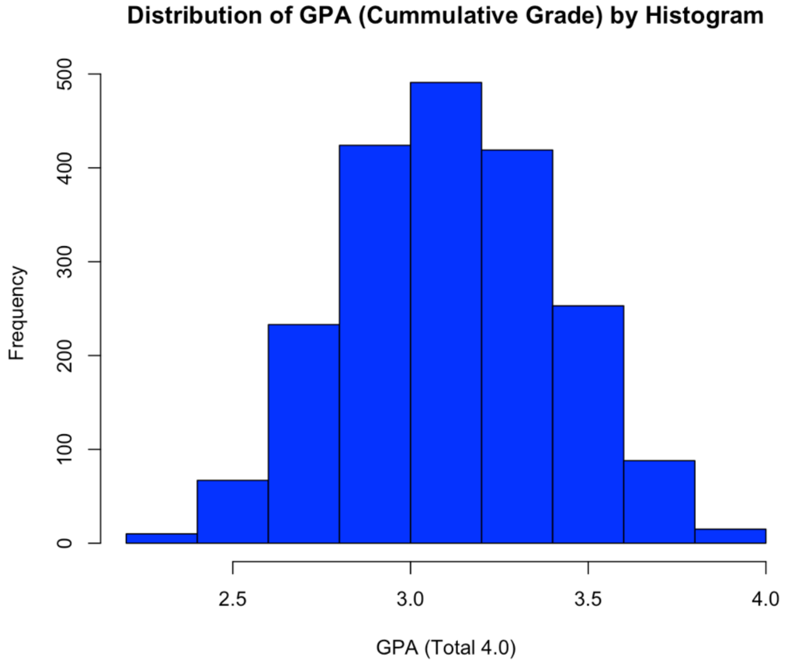
The box plot reveals a distribution of GPAs with a median around 3.1. The data is relatively compact, with the middle 50% of individuals having GPAs between 3.0 and 3.4. Two outliers suggest the presence of individuals with significantly higher or lower GPAs compared to the majority.

Additional Insights:

- **Variability:** The IQR of 0.4 suggests a relatively small amount of variability in GPAs among these individuals.
- **Central Tendency:** The median being 3.1 suggests that a typical individual in this group has a GPA around 3.1.
- **Outliers:** The presence of outliers indicates that there are a few individuals with significantly higher or lower GPAs compared to the majority.

2. Histogram

a. Distribution of GPA (Cumulative Grade) by Histogram



The histogram displays the distribution of GPAs (Cumulative Grade) among a group of individuals. Here's a detailed summary of the graph:

Distribution Shape:

- The histogram shows a roughly bell-shaped distribution, which is characteristic of a normal distribution. This indicates that most of the GPAs are clustered around the central value, with fewer individuals having significantly higher or lower GPAs.

Central Tendency:

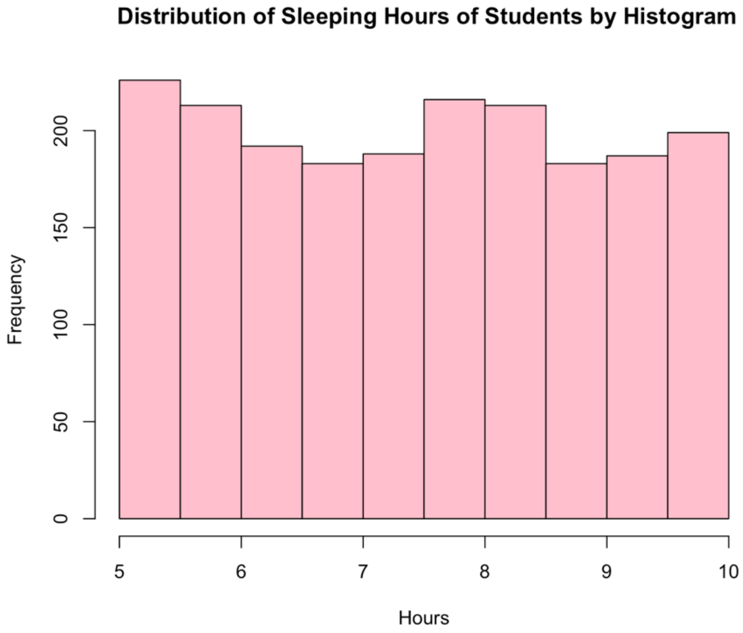
- The peak of the distribution appears to be around 3.0, suggesting that most individuals have GPAs close to 3.0. This value represents the mode, or the most frequent GPA in the dataset.

Spread:

- The bars extend from approximately 2.0 to 4.0, indicating that the GPAs in this dataset range from around 2.0 to 4.0.
- The width of the bars suggests that the GPAs are evenly spread across this range.

The histogram suggests that the GPAs in this dataset are normally distributed, with a central tendency around 3.0. The distribution is relatively symmetric, with a similar number of individuals on either side of the peak. This pattern indicates that most individuals in this group have GPAs within a typical range, with fewer individuals having significantly higher or lower GPA

b. Distribution of Sleeping Hours of Students by Histogram



The histogram displays the distribution of sleeping hours among students. The data appears to be roughly evenly distributed across the range of 5 to 10 hours, with no significant peaks or valleys. This suggests that there is no clear preference for a specific number of sleeping hours among these students.

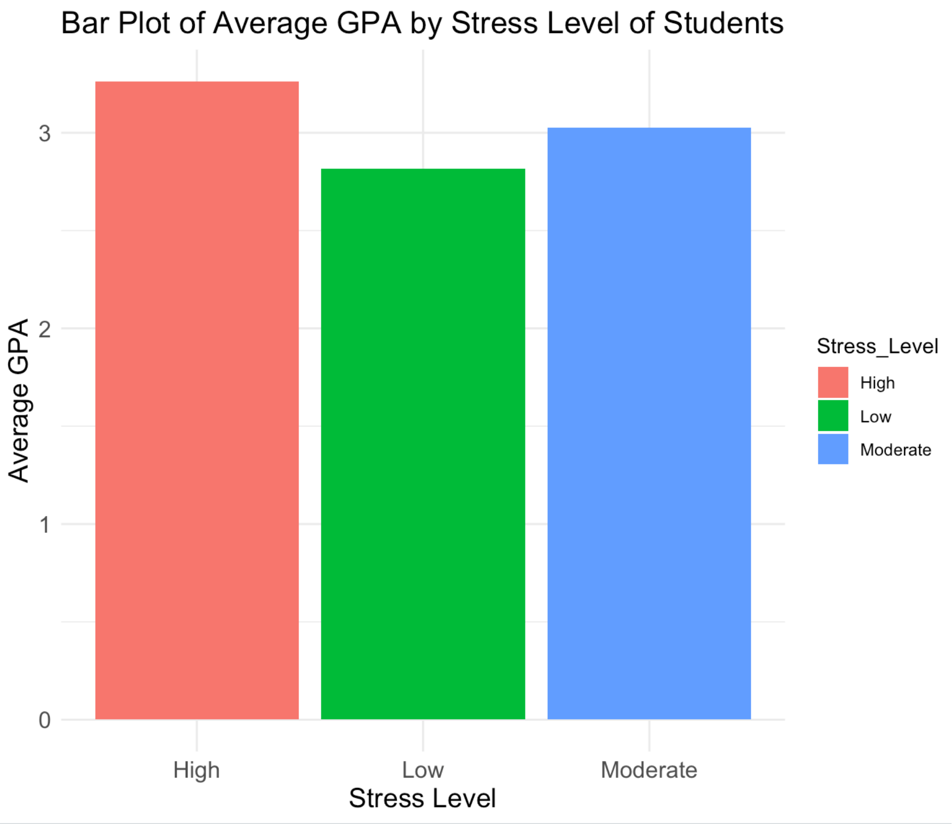
Here are some key observations:

- Even Distribution: The bars have similar heights, indicating that a similar number of students fall within each sleep duration range.
- No Outliers: There are not very tall or very short bars, suggesting that there are no extreme outliers in the data.
- Range: The data spans from around 5 hours to 10 hours, suggesting a wide range of sleeping habits among the students.

Overall, the histogram suggests that the students in this sample have a diverse range of sleep patterns, with no trend or preference for a specific sleep duration.

3. Barplot

Bar Plot of Average GPA by Stress Level of Students



1. **Stress Level and GPA:** The plot reveals a clear relationship between stress level and average GPA. Students with low stress levels tend to have the lowest average GPA, followed by students with moderate stress levels. Students with high stress levels have the highest average GPA.
2. **GPA Differences:** The average GPA for students with high stress levels is significantly higher than that of students with low and moderate stress levels. The difference in average GPA between students with low and moderate stress levels is less pronounced.

The bar plot suggests that higher levels of stress are positively correlated with academic performance, as measured by GPA. Students with higher stress levels tend to achieve higher GPAs. This finding highlights the importance of managing stress for academic success.
There can be multiple other factors involved for high stress apart from academics like not spending enough time in co-curricular activities, etc.
Therefore, learning has to be made easy for students so that they get good grades with less stress.

Conclusion:

The analysis of the Student Lifestyle dataset revealed significant insights into the relationship between sleep, GPA, and stress. Through statistical tests and visualizations, it was found that:

- **Hypothesis Testing:**
  - The average GPA was significantly different from 3.0, indicating a higher average GPA.
  - The average sleep hours were significantly lower than 8 hours, suggesting insufficient sleep.
- **Visualizations:**
  - Box plots showed a moderate spread in sleep hours and GPA.
  - Histograms indicated a normal distribution for GPA and an even distribution for sleep hours.
  - A bar plot highlighted a positive correlation between stress and GPA.

These findings emphasize the importance of adequate sleep and effective stress management for optimal academic performance. Further research can delve into specific stress factors and long-term impacts to develop targeted interventions and improve student well-being.

References:

1. Bluman, A. G. (2017). **Elementary statistics: A step by step approach** (10th ed.). McGraw-Hill Education.
2. Triola, M. F. (2018). **Essentials of statistics** (6th ed.). Pearson.
3. https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset.