Exploratory Data Analysis (EDA) Report

Machine Learning Assignment

Datasets: Sleep Health and Lifestyle Dataset & Student Performance Factors Dataset
Source: Kaggle (provided for the assignment)

1. Introduction

This report presents the overall findings from an exploratory data analysis (EDA) aimed at learning and applying various techniques in data exploration.

Two datasets were analyzed: the Sleep Health and Lifestyle Dataset and the Student Performance Factors Dataset. The purpose of this analysis is to uncover insights and test a series of hypotheses regarding sleep, lifestyle, and academic performance. By performing detailed EDA, we intend to better understand the underlying patterns, distributions, and relationships among features, which will guide future data-driven decisions and modeling.

2. Data Overview and Preprocessing

2.1. Sleep Health and Lifestyle Dataset

Features & Descriptions:

- **Person ID:** Unique identifier for each individual.
- **Gender:** Male or Female.
- **Age:** Age of the person in years.
- **Occupation:** Profession of the individual.
- **Sleep Duration (hours):** Number of hours slept per day.
- **Quality of Sleep (scale: 1–10):** Subjective sleep quality rating.
- **Physical Activity Level (minutes/day):** Daily minutes of physical activity.
- **Stress Level (scale: 1–10):** Subjective stress rating.
- **BMI Category:** Categorization based on BMI (e.g., Underweight, Normal, Overweight).
- **Blood Pressure (systolic/diastolic):** Recorded as two separate values.
- **Heart Rate (bpm):** Resting heart rate.

- **Daily Steps:** Count of daily steps.
- **Sleep Disorder:** Presence of sleep disorder (None, Insomnia, Sleep Apnea).

  Note: Missing values in Sleep Disorder were replaced with "No Disorder" based on the understanding that they indicate the absence of a sleep disorder.

Data Issues & Cleaning:

- **Missing Values:** Handled by replacing or deleting depending on the case; for Sleep Disorder, missing values were converted to "No Disorder."
- **Outliers:** Notably, heart rate outliers were detected via box plots.
- **Categorical Variables:** Encoding was deemed unnecessary for the EDA stage.

2.2. Student Performance Factors Dataset

Features & Descriptions:

1. **Hours_Studied:** Weekly hours spent studying.
2. **Attendance:** Percentage of classes attended.
3. **Parental_Involvement:** Level (Low, Medium, High).
4. **Access_to_Resources:** Availability of educational resources (Low, Medium, High).
5. **Extracurricular_Activities:** Participation (Yes, No).
6. **Sleep_Hours:** Average nightly sleep.
7. **Previous_Scores:** Scores from past exams.
8. **Motivation_Level:** Student's motivation (Low, Medium, High).
9. **Internet_Access:** Availability (Yes, No).
10. **Tutoring_Sessions:** Number of sessions per month.
11. **Family_Income:** Income level (Low, Medium, High).
12. **Teacher_Quality:** Quality of teaching (Low, Medium, High).
13. **School_Type:** (Public, Private).
14. **Peer_Influence:** (Positive, Neutral, Negative).
15. **Physical_Activity:** Weekly hours of physical activity.
16. **Learning_Disabilities:** (Yes, No).
17. **Parental_Education_Level:** Highest education level of parents (High School, College, Postgraduate).
18. **Distance_from_Home:** (Near, Moderate, Far).
19. **Gender:** Male or Female.
20. **Exam_Score:** Final exam score.

Data Issues & Cleaning:

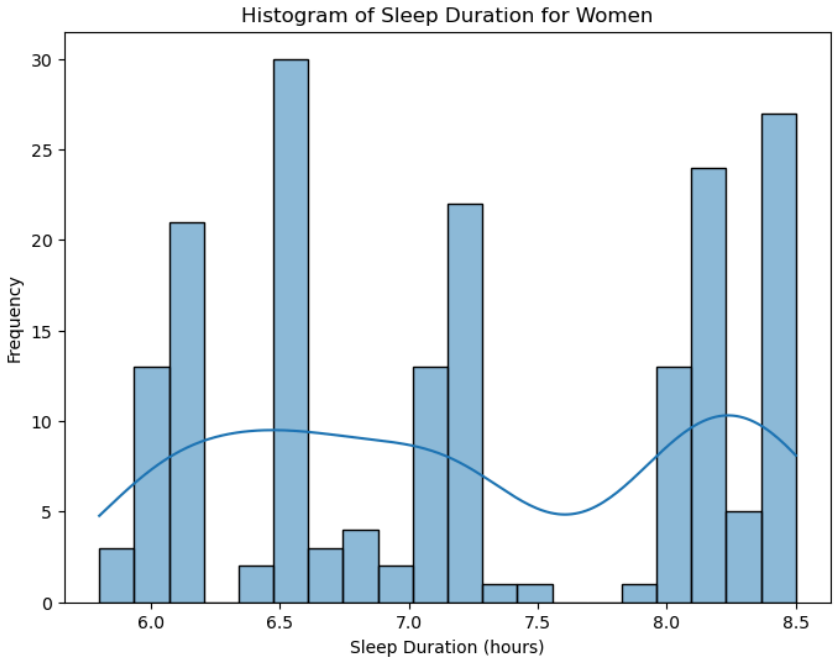
- **Missing Values:** Some missing values were observed in *Teacher_Quality*, *Parental_Education_Level*, and *Distance_from_Home*. Given the low number of missing entries relative to the dataset size, rows with null values were deleted.
- **Outliers:** Outliers were detected in *Hours_Studied* and *Tutoring_Sessions* via visual methods (box plots).
- **Preprocessing:** Encoding and normalization were not applied, as the focus is solely on EDA.

3. Exploratory Data Analysis & Hypotheses Testing

3.1. Sleep Health and Lifestyle Dataset

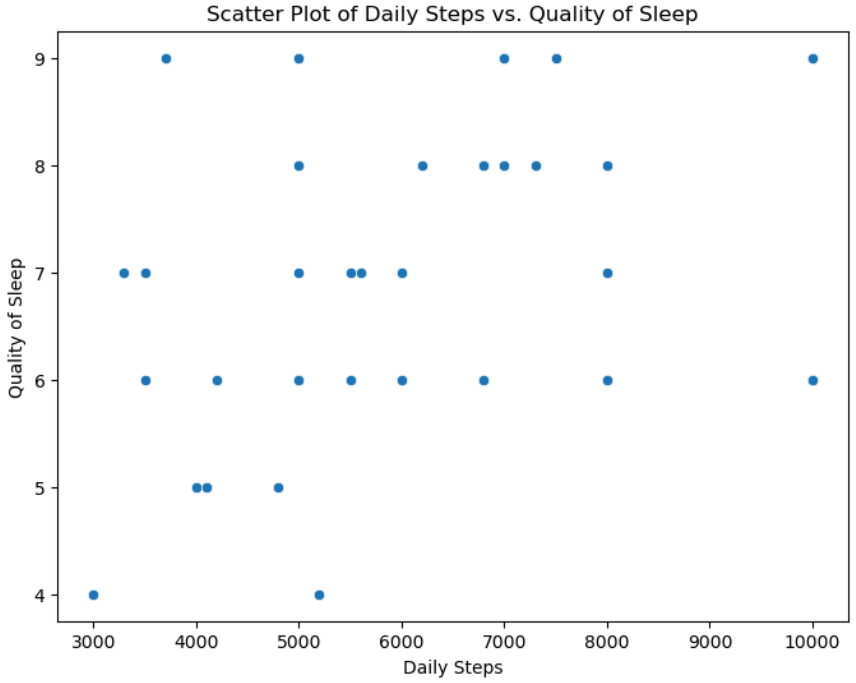
Hypothesis 1: Women's Sleep Duration Distribution

- **Question:** Does women's sleep duration follow a normal distribution?
- **Finding:** Analysis of the distribution (using histograms and normality tests) indicates that the sleep duration of women is not normally distributed.



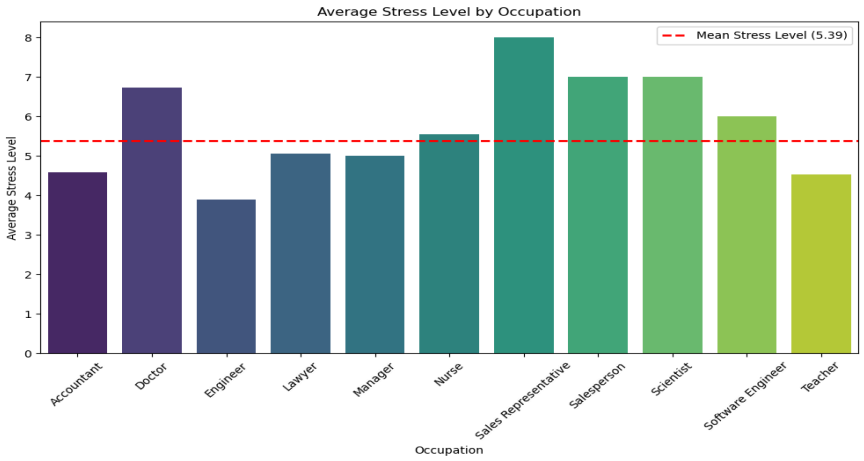
Hypothesis 2: Daily Steps and Quality of Sleep

- **Question:** Is having higher daily steps a contributing factor to better sleep quality?
- **Finding:** A correlation analysis between Daily Steps and Quality of Sleep reveals that an increase in daily steps does not correspond to better sleep quality.



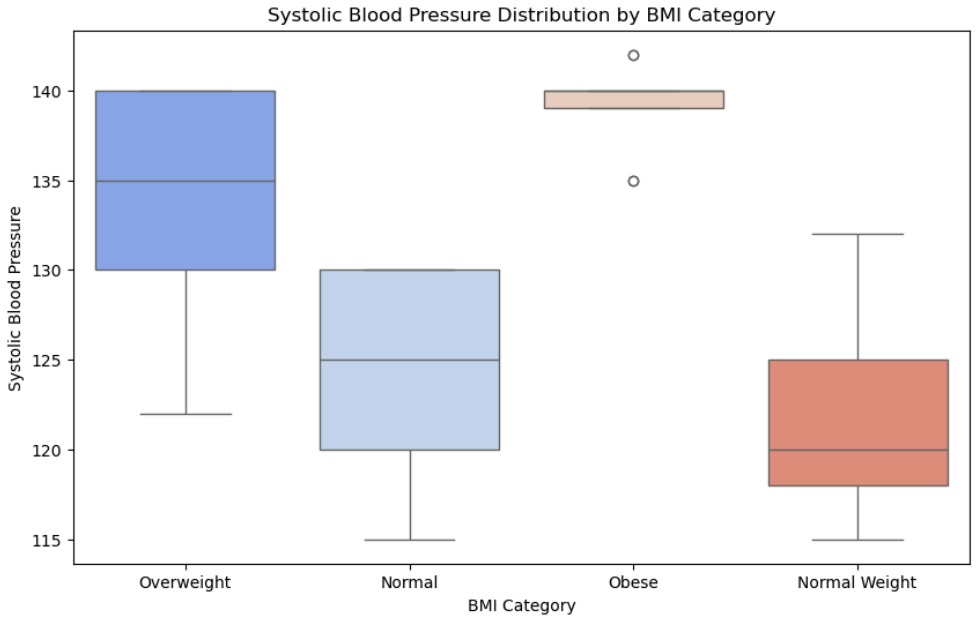
Hypothesis 3: Stress Level Across Occupations

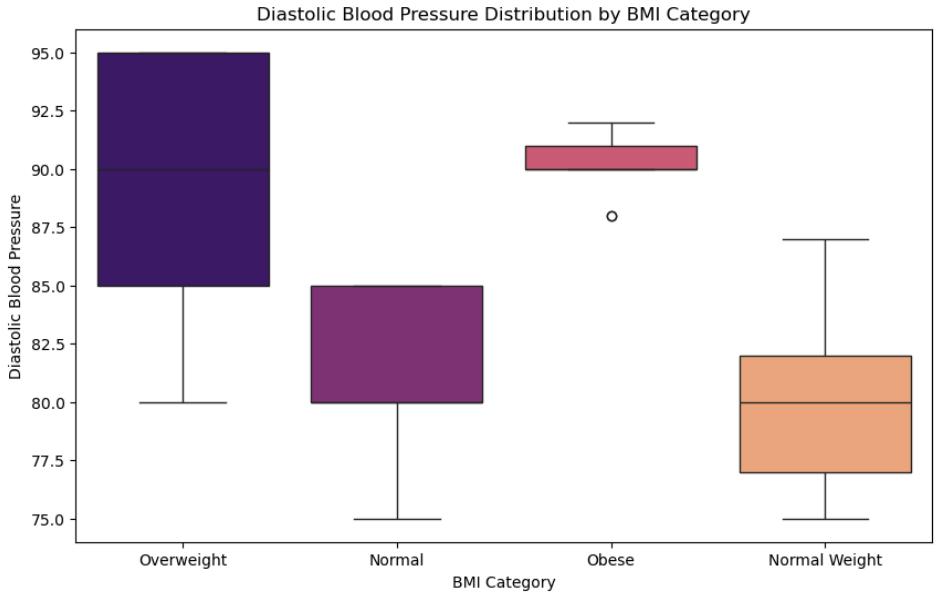
- **Question:** Is the stress level significantly different among various occupations?
- **Approach:**
    - Perform a statistical test (e.g., ANOVA) to compare stress levels across different occupations.
    - Compute and visualize the average stress level using bar charts.
- **Finding:** At least one occupation exhibits a significantly different mean stress level compared to others.



Hypothesis 4: BMI Categories and Blood Pressure

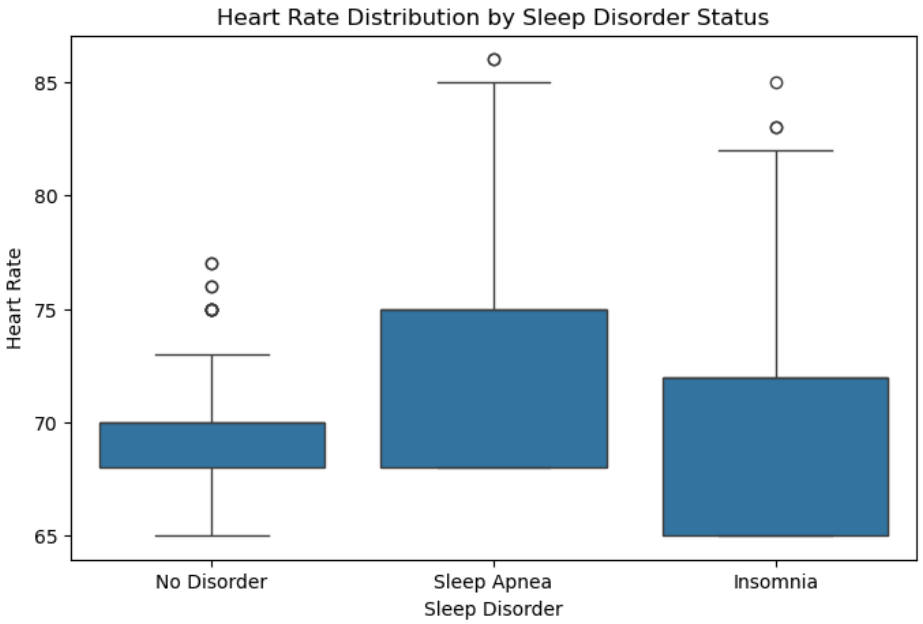
- **Question:** Are there significant differences in blood pressure across different BMI categories?
- **Approach:**
    - Split the Blood Pressure feature into systolic and diastolic columns.
    - Apply statistical tests (e.g., t-tests or ANOVA) for each BMI category.
- **Finding:** There is a statistically significant difference in blood pressure for at least one BMI category.





Hypothesis 5: Sleep Disorders and Heart Rate

- **Question:** Do individuals with sleep disorders have higher heart rates compared to those without?
- **Finding:** Statistical comparisons reveal a significant difference in heart rates between individuals with and without sleep disorders.



Hypothesis 6 (Optional): Age and Sleep Duration Relationship

- **Question:** Is there a significant relationship between Age and Sleep Duration?
- **Finding:** Both linear and monotonic relationship tests indicate that there is no significant relationship between age and sleep duration.

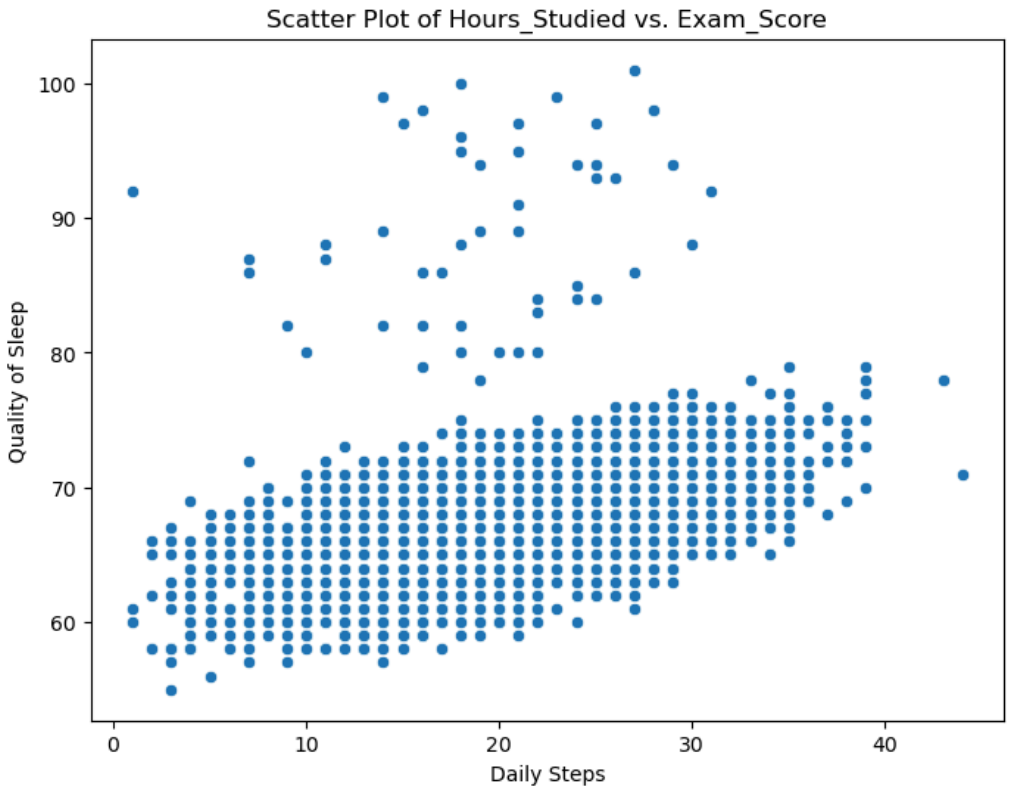
Hypothesis 7 (Optional): Physical Activity and Quality of Sleep

- **Question:** Does higher Physical Activity Level correlate with better Quality of Sleep?
- **Finding:** Analysis shows no linear or monotonic relationship between physical activity levels and quality of sleep.

3.2. Student Performance Factors Dataset

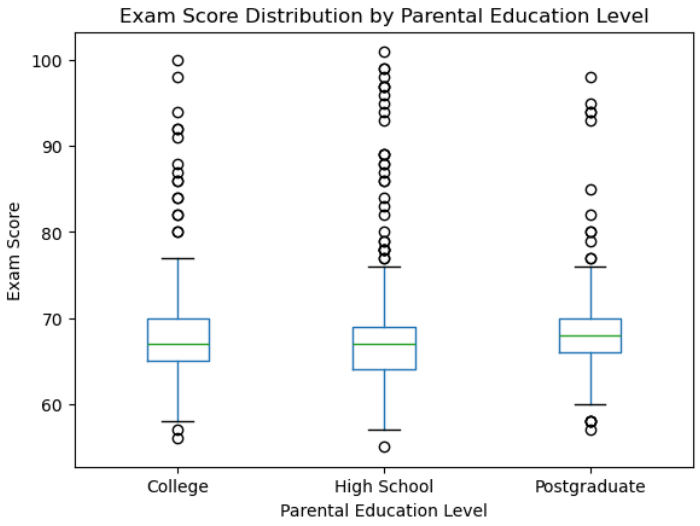
Hypothesis 1: Relationship Between Hours Studied and Exam Score

- **Question:** Is there any linear relationship between Hours_Studied and Exam_Score?
- **Finding:** Correlation analysis shows that there is no linear relationship between hours studied and the exam score.



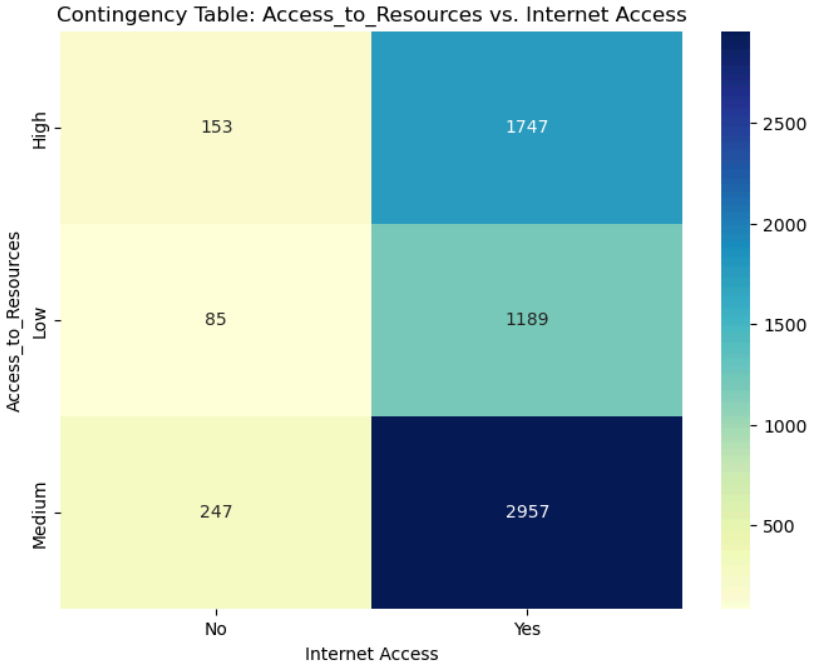
Hypothesis 2: Exam Scores Across Parental Education Levels

- **Question:** How do mean Exam_Scores differ across groups of Parental Education Level?
- **Approach:**
    - Compute mean Exam_Scores for each parental education category.
    - Visualize differences using bar charts.
- **Finding:** There is a clear difference in mean Exam_Scores among different parental education levels.



Hypothesis 3: Association Between Access to Resources and Internet Access

- **Question:** Is there an association between Access_to_Resources and Internet_Access?
- **Finding:** The analysis indicates no statistically significant association between the two variables.

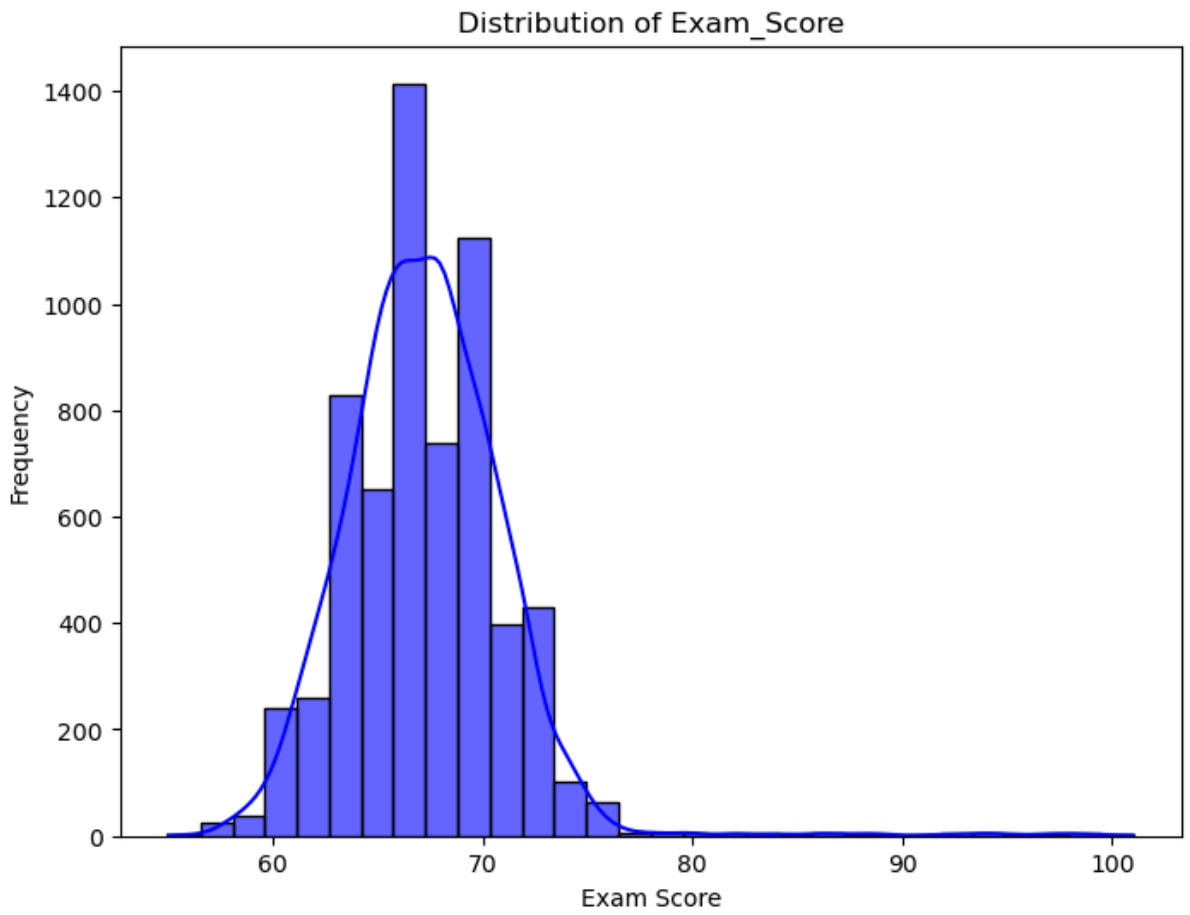


Hypothesis 4: Monotonic Relationship Between Motivation and Exam Score

- **Question:** Is there a monotonic relationship between Motivation_Level and Exam_Score?
- **Finding:** The tests for monotonic trends suggest that there is no significant monotonic relationship.

Hypothesis 5: Normality of Exam Scores

- **Question:** Do Exam_Scores follow a normal distribution?
- **Finding:** Normality tests confirm that the Exam_Score distribution is not normally distributed.



4. Conclusion and Discussion

The EDA investigation provided valuable insights into both datasets. Key conclusions include:

- **Sleep Health and Lifestyle Dataset:**
    - Women's sleep duration does not follow a normal distribution.
    - Daily steps do not positively impact sleep quality, and occupation appears to influence stress levels significantly.
    - Significant differences were also noted across BMI categories with respect to blood pressure, while age and physical activity showed no meaningful relationship with sleep metrics.
    - Individuals with sleep disorders have a significantly higher heart rate.
- **Student Performance Factors Dataset:**
    - No linear relationship was found between the number of hours studied and exam performance.
    - Parental education levels affect exam scores, and despite thorough testing, no associations were observed between access to resources and internet access or between motivation levels and exam scores.
    - The exam score distribution deviates from normality, suggesting potential implications for subsequent predictive modeling.

These findings underscore the importance of tailored data exploration techniques for different types of data. The analysis not only validated several hypotheses but also highlighted areas for further investigation and potential model improvement.

This report demonstrates the process of EDA, reinforcing the learning objectives by systematically investigating, visualizing, and testing the underlying assumptions of each dataset. The outcomes inform not only the current understanding of the data but also guide further analyses in subsequent machine learning tasks.