Deliverable 1: Project Initiation, Domain Understanding, Data Exploration

Students Health🩺 & Academic📖 Performance🎗️

Group 14

Sowmya Kanjula
Shubh Almal
Venkata Mahalakshmi Vishnumolakala
Mounika Muddy Paka

Project Introduction

The dataset for this project is a Student Health and Academic Performance Dataset which was obtained from Kaggle.The data consists of various attributes related to customer demographics, service usage patterns, and account information. Key features include customer tenure, service usage metrics, and billing details, among others. The primary objective of this project is to predict customer churn, i.e., identifying customers who are likely to discontinue their service in the near future.

We aim to make this project predictive in nature, as it aims to forecast future churn events based on historical data. For this project, we will use supervised learning techniques and we are planning to train multiple Machine Learning models and compare which model is best for the database.

Methodology

Our project is predictive in nature, aiming to forecast future outcomes based on historical data. We will use supervised learning techniques since our goal involves predicting specific target variables (e.g., academic performance and health outcomes).

Techniques and Models

- **Classification**: To categorize students based on their risk of poor academic performance or adverse health outcomes.
- **Regression**: Specifically, logistic regression, to predict binary outcomes such as whether a student is likely to perform poorly or experience health issues.

- **Model Comparison**: We will train and compare multiple machine learning models, such as logistic regression, decision trees, random forests, and support vector machines, to identify the most effective model for our predictions.

Research Question

Idea - Impact of Mobile Phone Usage on Student Performance and Well-being

The dataset appears to capture various aspects of mobile phone usage among students, including their demographic details (age, gender), type and frequency of mobile phone usage, its impact on their education, and health-related symptoms. The data seems to be collected to understand how mobile phone usage affects students' academic performance and well-being.

So keeping the above aspects in mind, we aim to -

1. Diagnose how different patterns of mobile phone usage relate to students' academic performance and health.
2. Predict potential risk factors for poor academic performance and adverse health outcomes based on mobile phone usage.

Relevant Domain Information

Information about numerous student health and academic performance variables may be found in the "Students Health and Academic Performance" Kaggle dataset. Examining the connections between health-related variables and academic outcomes may be helpful with this dataset.

You might look at studies on how academic performance is affected by health for more pertinent content. These two relevant articles are as follows:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3447545/

https://www.sciencedirect.com/science/article/pii/S0277953613000244

These articles explore the various ways in which students' academic performance might be impacted by health difficulties, including mental health, diet, and physical exercise.

Data Source and Description

Data Source:
You can find the dataset called "Students Health and Academic Performance" on

Kaggle at this link:Students Health and Academic Performance Dataset. Innocent Mfaume gathered information to study how students' health affects their school performance.

Description:

This dataset includes different details about students' health and how well they do in school. The features include many different things about people's age, lifestyle, and education. Here is a simple explanation of the features included in the dataset:

Gender: The student's identification as male or female.
Age: The student's age.
Height: The student's height in centimeters.
Weight: The student's weight is shown in kilograms.
BMI: Body Mass Index, is a number calculated using a person's height and weight.
Smoking: Tells if the student smokes (Yes or No).
Alcohol: Is the student a consumer of alcohol?(eg, Yes or No).
Physical Activity: How often do you participate in physical exercise: is it daily, weekly, or monthly?
Sleep Duration: The usual number of hours you sleep each night.
Stress Level: People describe their stress as Low, Medium, or High.
Health Status: How you feel about your health (Bad, Fair, Good, Excellent)
StudyTime: Hours spent studying each week
School Absences: The total number of days missed from school.
Grades: Grades (A, B, C, D, and F) are used to measure how well someone is doing in school.

Understanding Data and EDA

1. Overview of the Data: The dataset contains a number of characteristics pertaining to students' academic achievement and health, including:
Gender (Male, Female)

Age (years)

Height (cm)

Weight (kg)

BMI (calculated)

Smoking (Yes, No)

Alcohol (Yes, No)

Physical Activity (Daily, Weekly, Monthly)

Sleep Duration (hours per night)

Stress Level (Low, Medium, High)

Health Status (Poor, Fair, Good, Excellent)

Study Time (hours per week)

School Absences (number)

Grades (A, B, C, D, F)

2. Perspectives on EDA:

Age Group Distribution:

Teens make up the majority of students.

BMI Outliers:

Extreme BMI readings are present in certain students.

Distribution of Gender:

men and females distributed almost equally.

Study time and grades:

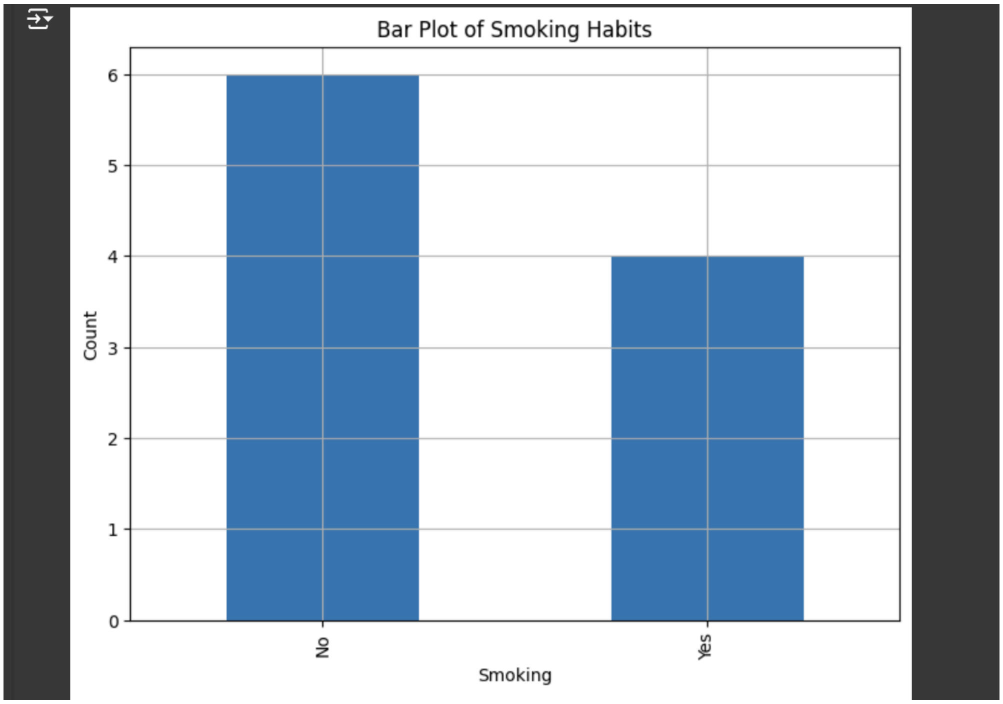
A higher grade is positively correlated with greater study time.

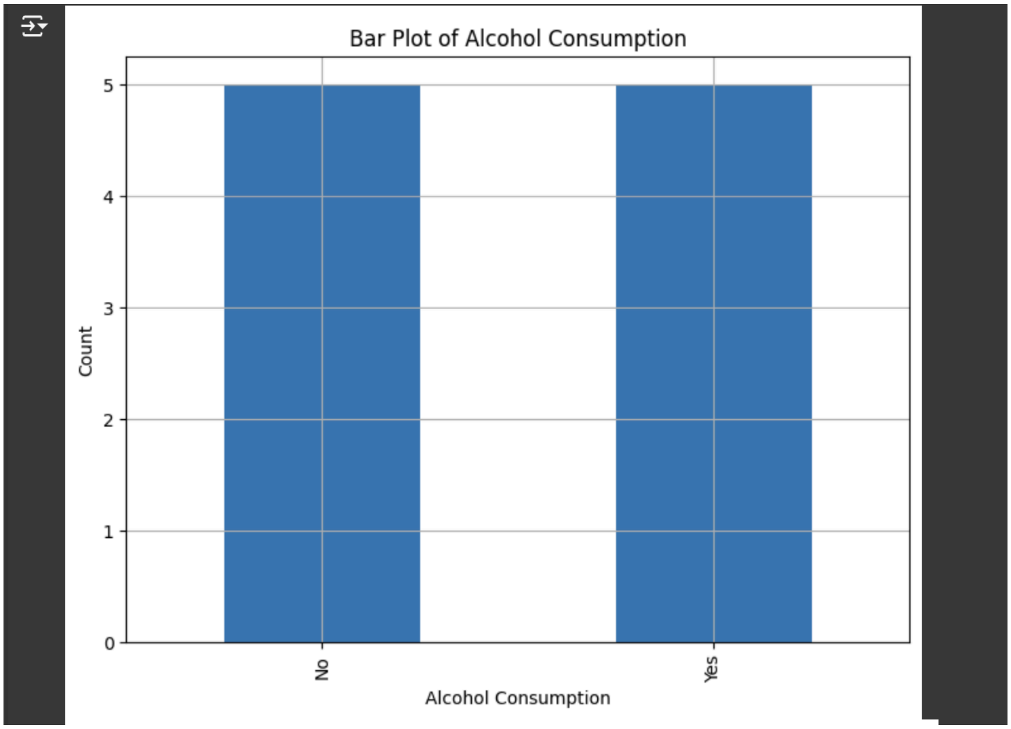
Grade Assignment:

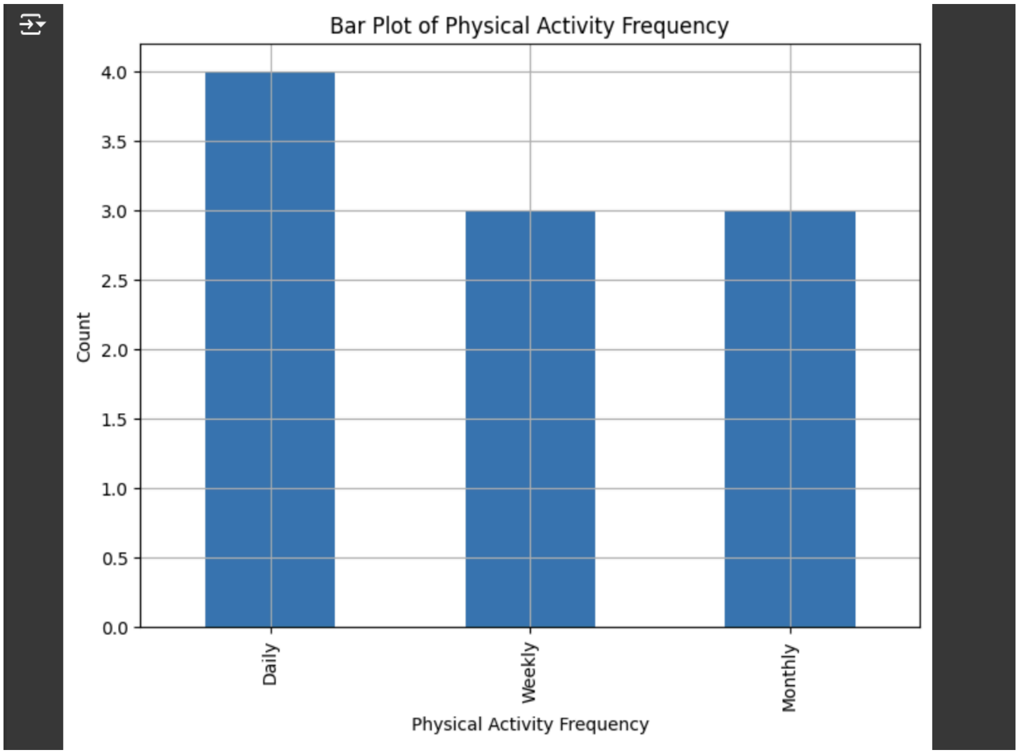
Grades in the B and C level are typical of pupils.

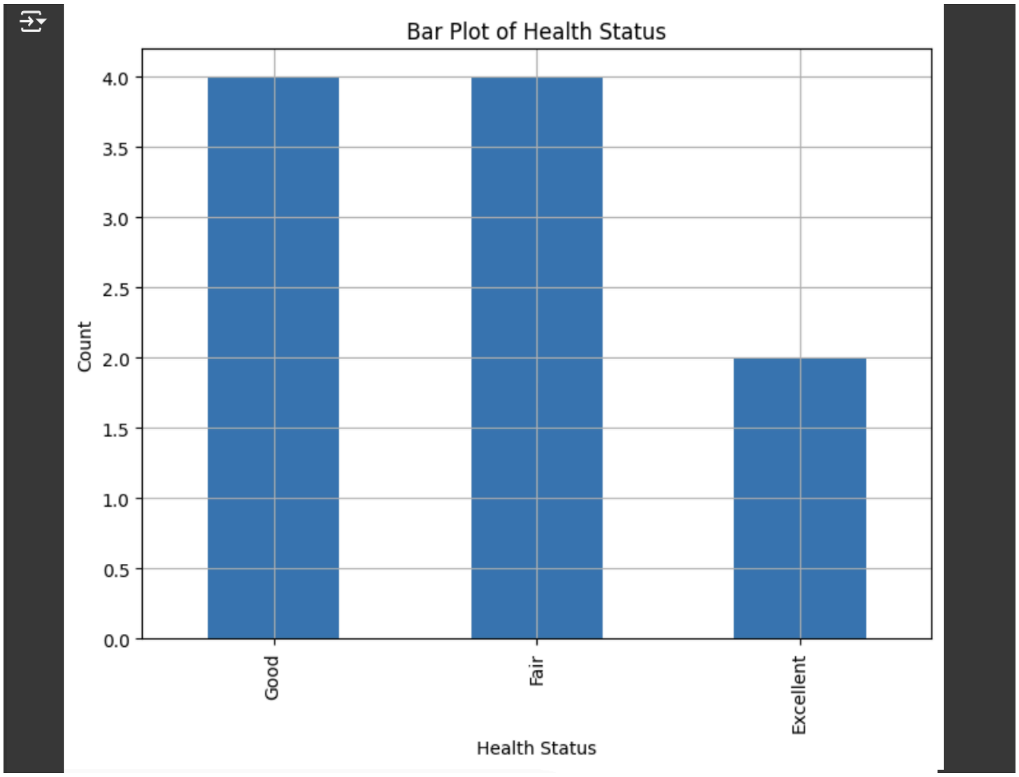
Conclusion:

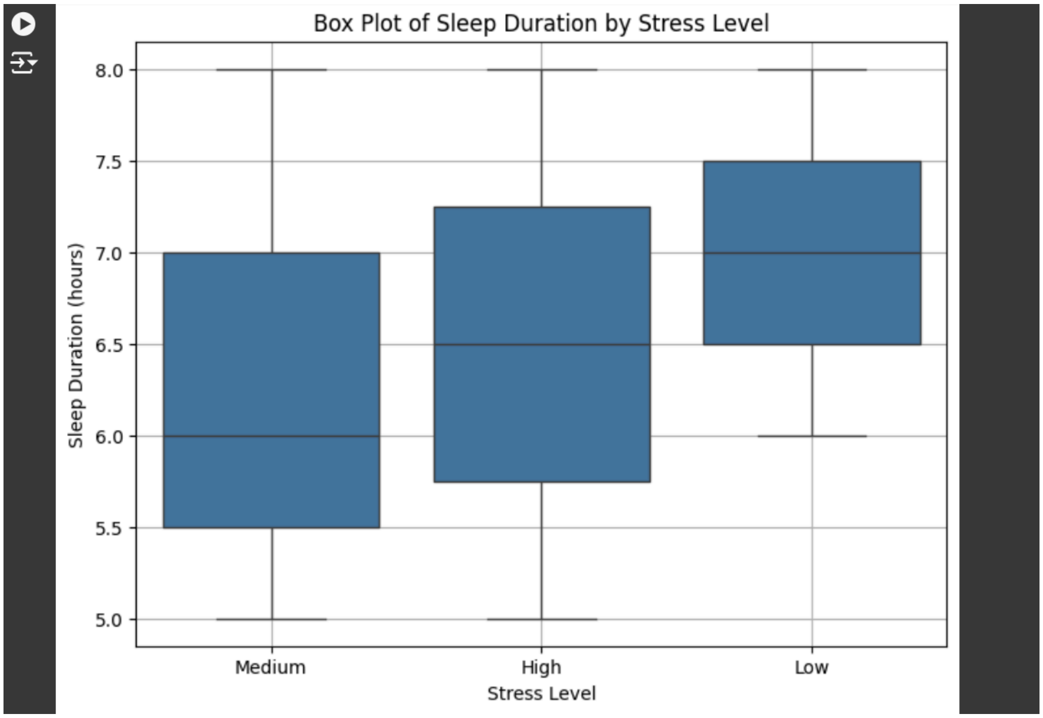
Finally, the EDA provided a strong basis for additional research and the creation of a model that will predict academic success based on lifestyle and health-related variables. It also identified important patterns and linkages among the variables.

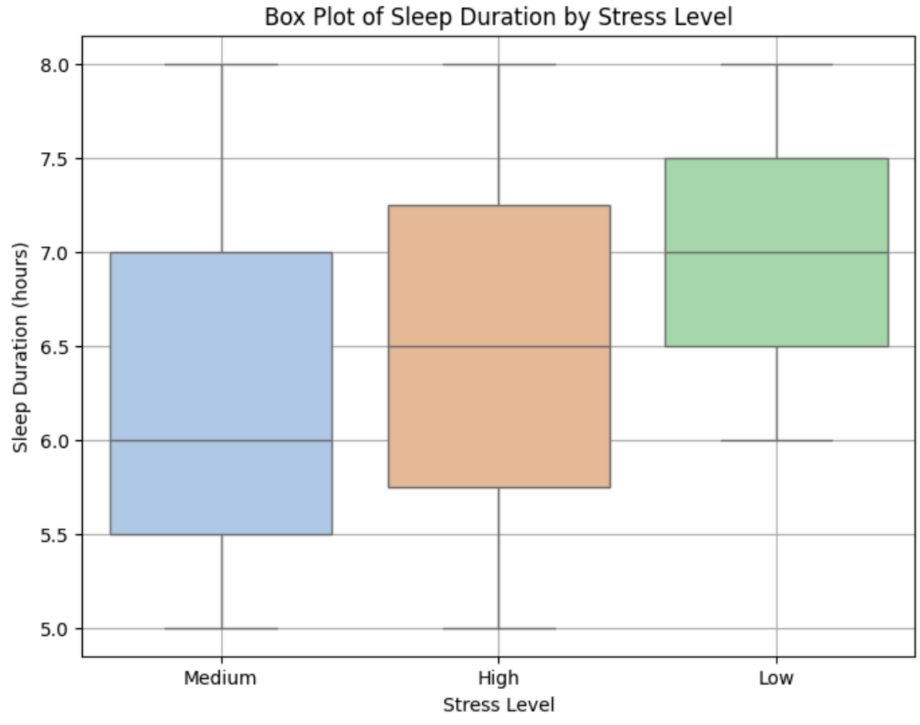

Bar Plot of Smoking Habits
Count
0
1
2
3
4
5
6
No
Yes
Smoking

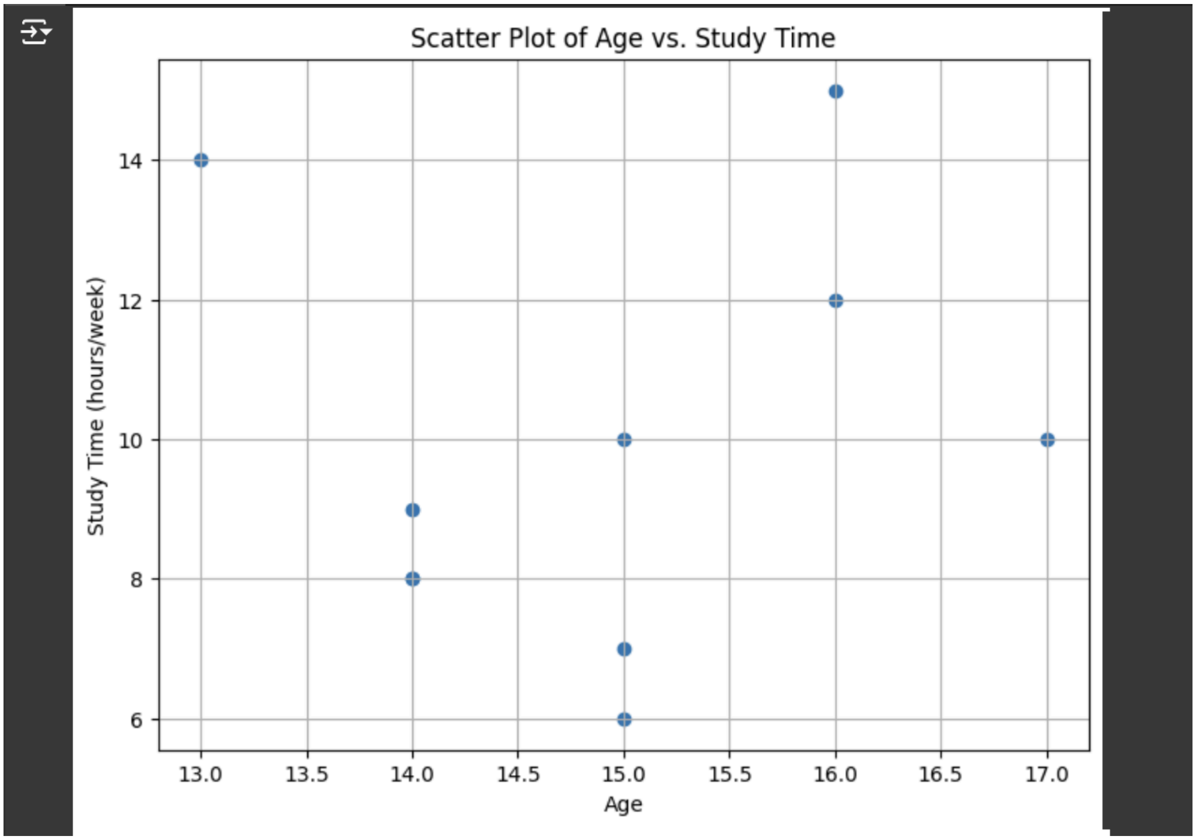

Bar Plot of Alcohol Consumption
Count
0
1
2
3
4
5
No
Yes
Alcohol Consumption

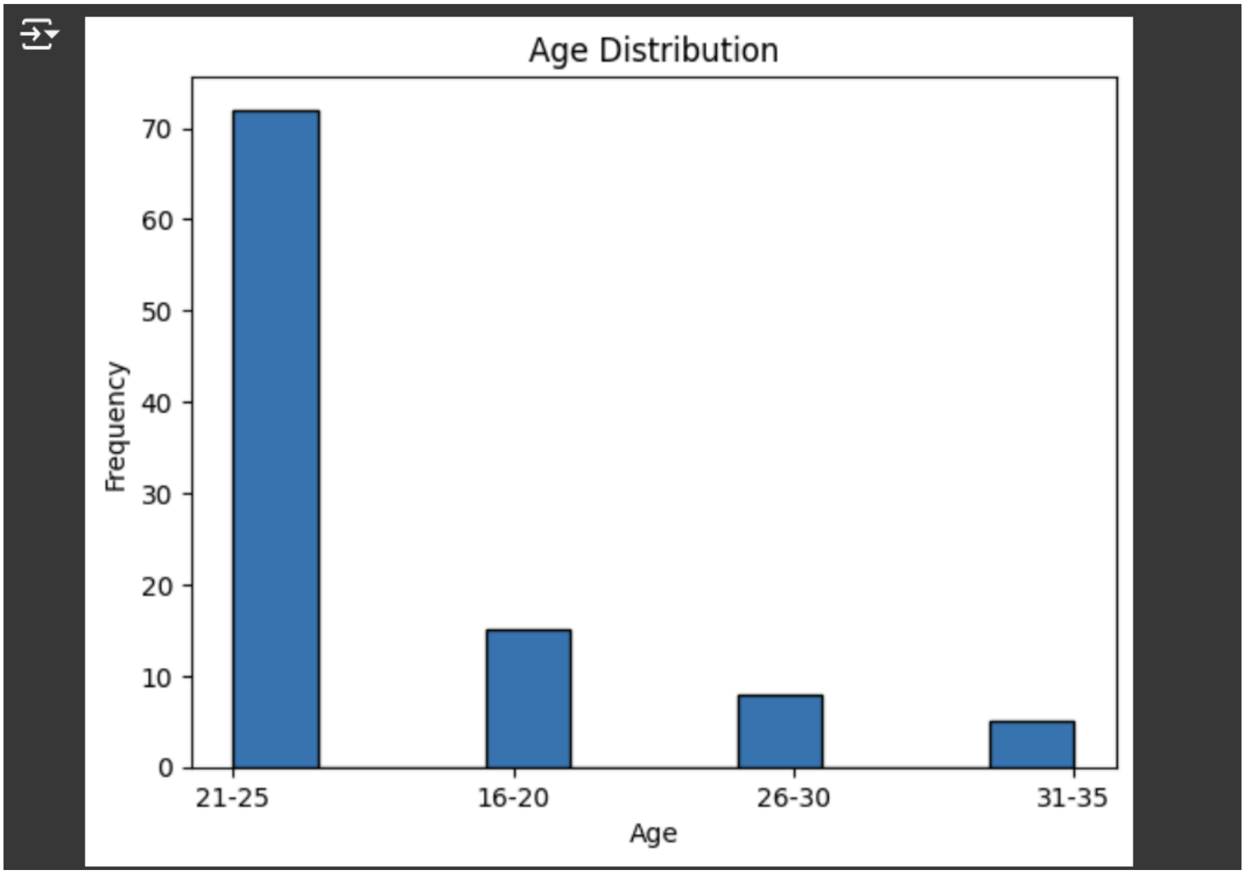

Bar Plot of Physical Activity Frequency
Count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Daily
Weekly
Monthly
Physical Activity Frequency


Bar Plot of Health Status
Count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Good
Fair
Excellent
Health Status


Box Plot of Sleep Duration by Stress Level
Sleep Duration (hours)
8.0
7.5
7.0
6.5
6.0
5.5
5.0
Medium
High
Low
Stress Level


Box Plot of Sleep Duration by Stress Level
Sleep Duration (hours)
8.0
7.5
7.0
6.5
6.0
5.5
5.0
Medium
High
Low
Stress Level


Scatter Plot of Age vs. Study Time
Study Time (hours/week)
14
12
10
8
6
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17.0
Age



<u>Data Preparation - continued in Deliverable 2</u>

Data Preparation

Data Cleaning:

- Dealt with missing data by filling it in or taking it out.
- Removed duplicate entries from the records.
- Modified the data types for consistency.

Data Transformation:

- Modify the figures to ensure they are uniform in scale.
- Changed categorical variables into numbers by using one-hot encoding.
- Made new tools from old ones to get more information. .

Data Integration:

- Merging information: Integrating details from various sources while maintaining the same identifying characteristics, if feasible

Data Splitting:

- Divide the data into two parts: one for training (80% or 70%) and one for testing (20% or 30%).

Feature Selection:

- Selected key attributes through correlation analysis combined with our understanding of the topic.