Progress Report #2

Group 9

Eric Wu
Dongyeon Kang
Sunwoo Kim
Sifei Zhou

Task and Visualization Tools

Histograms are a visualization tool that we will use for showcasing the distribution of continuous variables, such as "Sleep Duration" and "Heart Rate." They provide a visual representation of the frequency or density of values within these attributes, helping users understand the central tendencies and variations. Bar charts will be employed to compare categorical variables, like 'BMI Categories,' and 'Types of Sleep Disorders.' bar charts are effective in presenting the prevalence of various categories and conditions within the dataset. Box plots are useful for identifying outliers and gaining insights into the overall spread and distribution of data. We will utilize box plots to visualize attributes like "Daily Steps" and "Heart Rate," providing a clear representation of data variability. Our website will feature interactive data exploration tools that allow users to filter and select specific subsets of the dataset based on criteria such as age, gender, or occupation. Users can dynamically explore the relationships between sleep health and lifestyle factors.

Data

The dataset we are using, 'Sleep health and lifestyle,' acquired from Kaggle.com, contains a diverse set of attributes that serve as the foundation for our project's exploration of sleep and lifestyle factors. This dataset includes critical information such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, and the presence of sleep disorders. Notably, our initial data processing revealed that some attributes initially categorized as numeric variables, like 'Quality of Sleep' and 'Stress Level,' are more suitably considered categorical variables due to their discrete nature. Notably, our initial data processing revealed that some attributes initially categorized as numeric variables, like 'Quality of Sleep' and 'Stress Level,' are more suitably considered categorical variables due to their discrete nature. Additionally, we separated the 'Blood Pressure' attribute into systolic and diastolic values to enable more precise analysis and visualization. In our data processing phase, we also took care to identify potential outliers and created effective strategies for dealing with them, ensuring the dataset's reliability. These attributes will be crucial for extracting meaningful insights about the relationships between sleep health and various lifestyle factors, serving as the basis for our website's interactive data exploration features.

Data Processing

Our dataset was originally from Kaggle.com as we mentioned in our first progress report. The dataset was clean, but there is still something we want to process before using it. For our initial data processing, we decided to use R and R Studio. We started off by looking at the types of attributes and whether the data set contains any null values (After importing the CSV file). Gladly, the number of null values contained in the dataset is zero, and we also know how many numerical and categorical attributes we have.

Afterward, we draw some important points regarding the dataset during the process. For example, attributes like “quality of sleep,” and “Stress level” are actually categorical variables despite being numeric input. In addition, the “Blood pressure” attribute is a categorical variable that stores both the systolic and diastolic blood pressure of a record. We decided to separate the “Blood pressure” variable into two numeric values to make it easier for us to visualize the blood pressure.

Group Activity Table

Team Member	Responsibilities
Eric Wu	Data Processing Section
Dongyeon Kang	Task and Visualization Tools section
Sunwoo Kim	Data section
Sifei Zhou	Proofreading the entire report

Group Activity Log

Date	Activity	Attendance
10/10/2023	Discuss the format and details regarding progress report #2.	All team members