# Milestone 1 Report

Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

## Project Objective

The goal is to develop a heart disease risk prediction model and interactive dashboard based on stress levels, sleep duration, and gym exercise patterns. By integrating multiple datasets, this model aims to identify key factors contributing to heart disease risk and provide actionable insights.

## Datasets Used

1. Sleep Health and Lifestyle Dataset:

- Source: Kaggle (https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset)
- Attributes: Person ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, Sleep Disorder

2. Heart Disease Dataset

- Source: Kaggle (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)
- Attributes: Age, gender, chest pain type, blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, heart disease diagnosis.

3. Gym Members Exercise Tracking Dataset

- Source: Kaggle (https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset)
- Attributes: Age, Gender, Workout frequency, calories burned, BMI, hydration levels, heart rate, daily steps.

Common Attributes are Age and Gender and Heart Rate

**Tech Stack**

Programming Language

- Python

Libraries & Frameworks

- **Data Manipulation**: Pandas, NumPy
- **Visualization**: Matplotlib, Seaborn

**Data Preprocessing**

Steps taken for data preprocessing include:

- Handling missing values through median imputation.
- Splitting 'Blood Pressure' into 'Systolic_BP' and 'Diastolic_BP'.
- Standardizing column names for consistency.
- Encoding categorical variables such as gender and BMI categories.
- Detecting and handling outliers using the IQR method.
- Merging datasets on age and gender to form a unified dataset.
- Removing duplicate rows to ensure data integrity.

**Exploratory Data Analysis (EDA)**

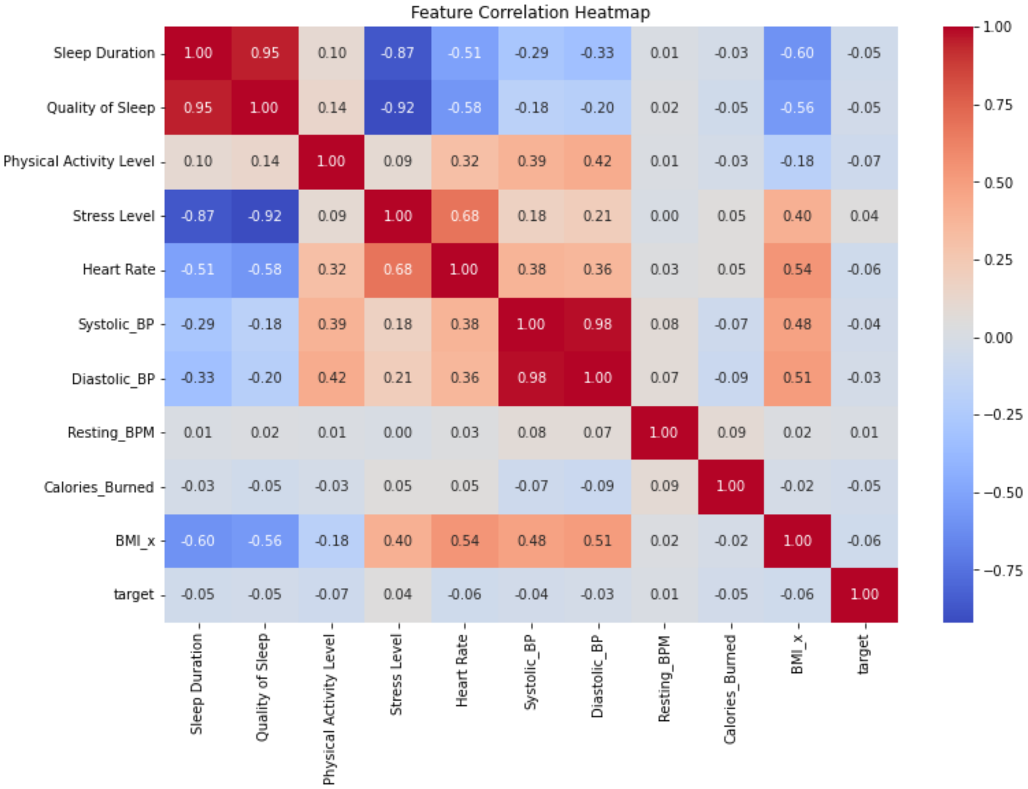
The following analyses were performed:

- Descriptive statistics: Summary statistics computed for all numerical variables.

- shape of balanced dataset is 6386 rows and 39 columns

## Important Visualizations:

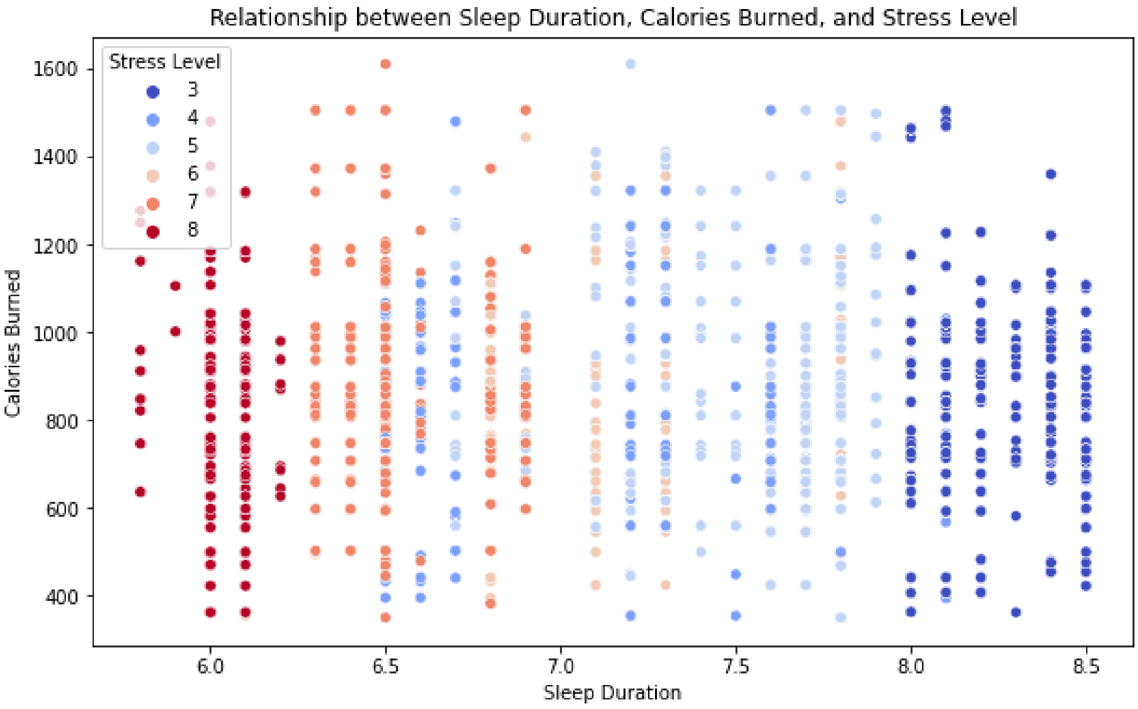
- Correlation Heatmap of merged dataset to identify feature relationships. This heatmap tells that **stress levels, sleep duration, and physical activity are key factors** in determining heart health.



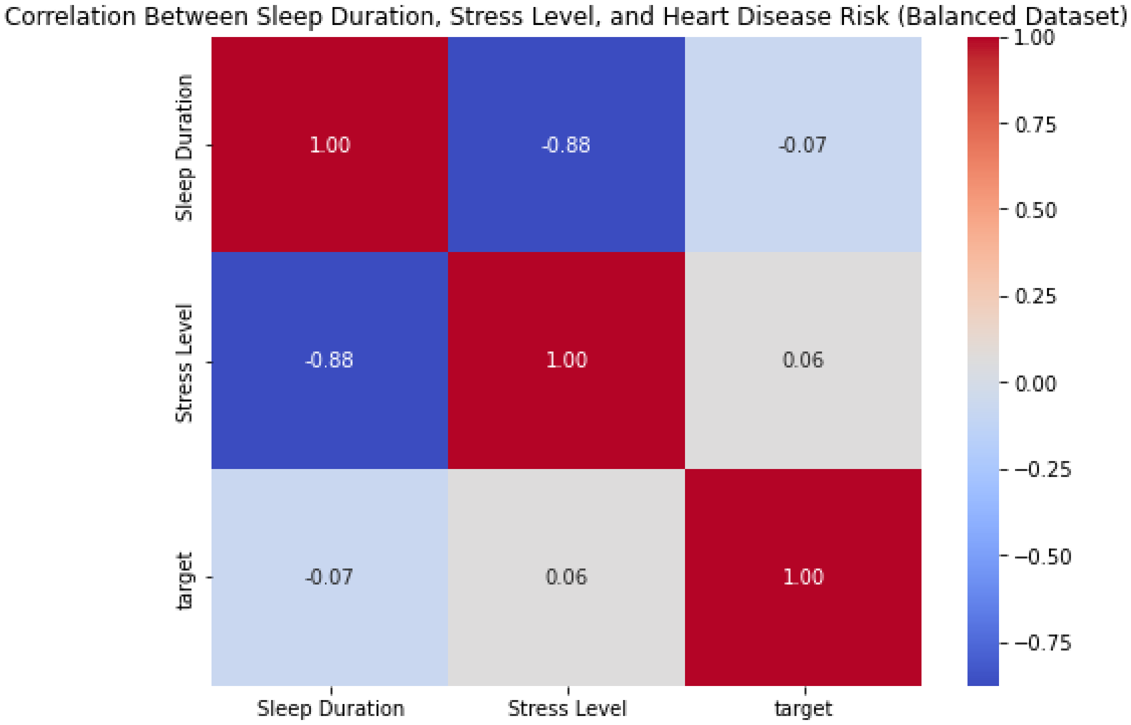
| | Sleep Duration | Quality of Sleep | Physical Activity Level | Stress Level | Heart Rate | Systolic_BP | Diastolic_BP | Resting_BPM | Calories_Burned | BMI_x | target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sleep Duration | 1.00 | 0.95 | 0.10 | -0.87 | -0.51 | -0.29 | -0.33 | 0.01 | -0.03 | -0.60 | -0.05 |
| Quality of Sleep | 0.95 | 1.00 | 0.14 | -0.92 | -0.58 | -0.18 | -0.20 | 0.02 | -0.05 | -0.56 | -0.05 |
| Physical Activity Level | 0.10 | 0.14 | 1.00 | 0.09 | 0.32 | 0.39 | 0.42 | 0.01 | -0.03 | -0.18 | -0.07 |
| Stress Level | -0.87 | -0.92 | 0.09 | 1.00 | 0.68 | 0.18 | 0.21 | 0.00 | 0.05 | 0.40 | 0.04 |
| Heart Rate | -0.51 | -0.58 | 0.32 | 0.68 | 1.00 | 0.38 | 0.36 | 0.03 | 0.05 | 0.54 | -0.06 |
| Systolic_BP | -0.29 | -0.18 | 0.39 | 0.18 | 0.38 | 1.00 | 0.98 | 0.08 | -0.07 | 0.48 | -0.04 |
| Diastolic_BP | -0.33 | -0.20 | 0.42 | 0.21 | 0.36 | 0.98 | 1.00 | 0.07 | -0.09 | 0.51 | -0.03 |
| Resting_BPM | 0.01 | 0.02 | 0.01 | 0.00 | 0.03 | 0.08 | 0.07 | 1.00 | 0.09 | 0.02 | 0.01 |
| Calories_Burned | -0.03 | -0.05 | -0.03 | 0.05 | 0.05 | -0.07 | -0.09 | 0.09 | 1.00 | -0.02 | -0.05 |
| BMI_x | -0.60 | -0.56 | -0.18 | 0.40 | 0.54 | 0.48 | 0.51 | 0.02 | -0.02 | 1.00 | -0.06 |
| target | -0.05 | -0.05 | -0.07 | 0.04 | -0.06 | -0.04 | -0.03 | 0.01 | -0.05 | -0.06 | 1.00 |

- Scatter plot to analyze relationships between sleep duration, calories burned, and stress level.

- This scatter plot visualizes stress levels across different individuals. Higher stress levels (darker red) tend to cluster on the left side, while lower stress levels (blue) are more evenly spread across the distribution.
- The separation between high-stress and low-stress groups suggests a strong impact of lifestyle habits on stress regulation. High-stress individuals may require more intervention strategies to improve heart health.



- Correlation Heatmap between "Sleep Duration, Stress Level, and Heart Disease Risk (Balanced Dataset) interprets strong relationships between stress levels, sleep duration, and heart rate. **High stress correlates with increased heart rate and reduced sleep duration.**



- Class imbalance handling: The dataset was balanced using undersampling to improve model performance.

Balanced Dataset details:
- No missing values after preprocessing.
- Outliers handled appropriately.
- Balanced dataset for improved model performance in the next stages.