The goal of this stage was to extract meaningful features from a dataset collected via a step tracker. These features will serve as inputs for building a machine learning model in later stages of the project.

**Dataset Overview**

The dataset contains daily logs of physical and lifestyle metrics including:

* steps: Number of steps taken
* sleep\_hours: Total hours of sleep
* water\_intake\_liters: Liters of water consumed
* mood: Self-reported emotional state (e.g., *happy*, *sad*, etc.)

**Steps Completed in Feature Extraction**

**1. Data Cleaning**

* **Stripped whitespaces** from column names and **converted them to lowercase** to ensure consistent referencing.

python

CopyEdit

df.columns = df.columns.str.strip().str.lower()

**2. Mood Mapping**

* Mapped the categorical mood column to numerical values to enable analysis and modeling:

| **Mood** | **Numeric** |
| --- | --- |
| Stressed | 1 |
| Sad | 2 |
| Neutral | 3 |
| Happy | 4 |
| Excited | 5 |

* python
* CopyEdit
* mood\_mapping = {
* "stressed": 1,
* "sad": 2,
* "neutral": 3,
* "happy": 4,
* "excited": 5
* }
* df["mood\_numeric"] = df["mood"].map(mood\_mapping)

**3. Feature Engineering**

* **Hydration per 1000 steps**: A new feature was created to evaluate hydration efficiency in relation to physical activity:

hydration\_per\_1000\_steps=water\_intake\_liters(steps1000)\text{hydration\\_per\\_1000\\_steps} = \frac{\text{water\\_intake\\_liters}}{\left(\frac{\text{steps}}{1000}\right)}hydration\_per\_1000\_steps=(1000steps​)water\_intake\_liters​

To avoid division by zero, we replaced any steps = 0 with 1 during computation.

python

CopyEdit

df["hydration\_per\_1000\_steps"] = df["water\_intake\_liters"] / (df["steps"] / 1000).replace(0, 1)

**4. Data Visualization**

* Created a **boxplot** showing the distribution of hydration\_per\_1000\_steps across different moods.

<img src="hydration\_vs\_mood.png" alt="Hydration vs Mood Boxplot" width="500"/>

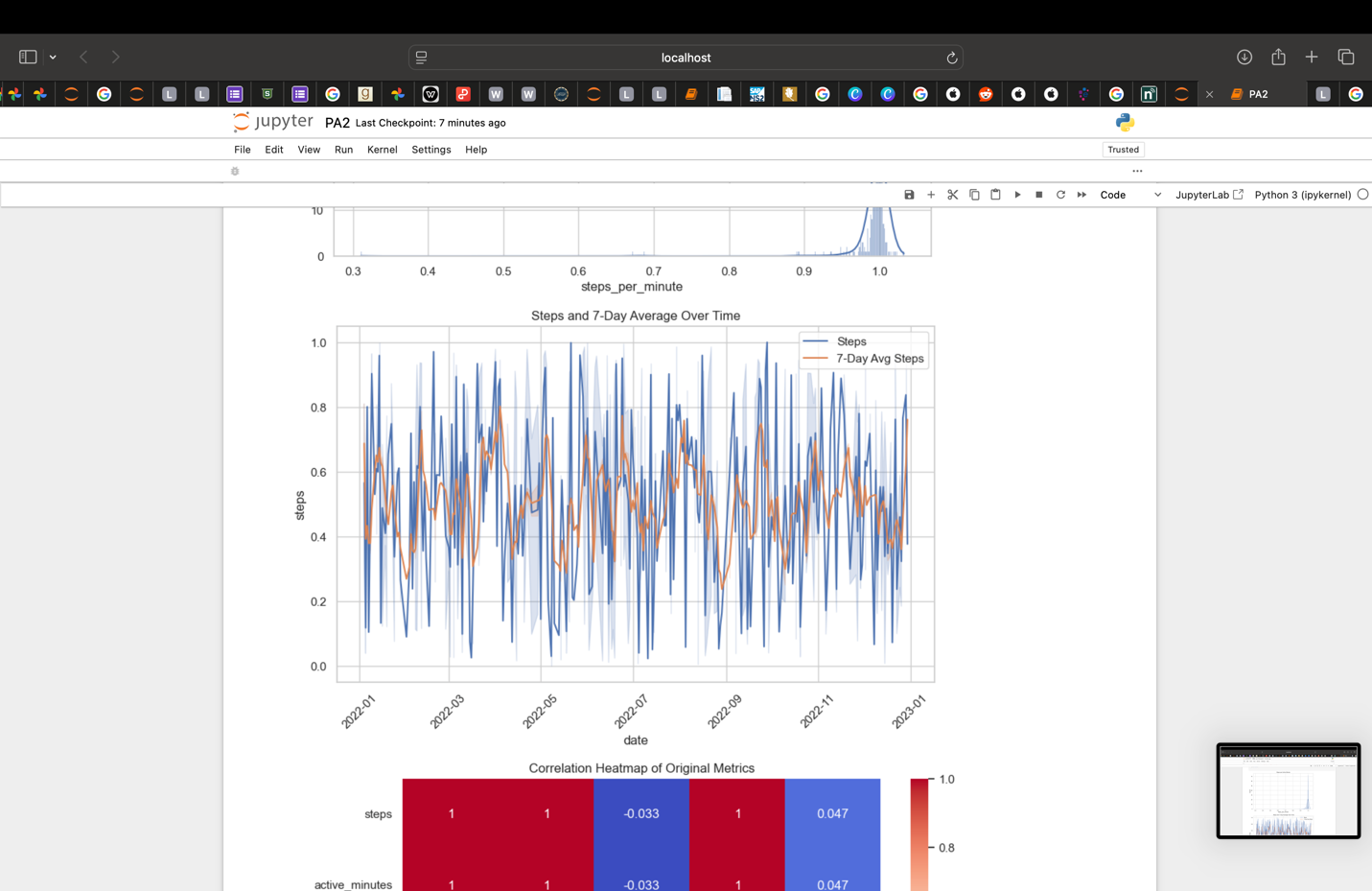
This visualization highlights how hydration patterns vary with mood, helping to identify relationships that may be relevant for classification tasks.

**Output Files**

* **extracted\_features.xlsx**: Contains the original data along with the new features (mood\_numeric, hydration\_per\_1000\_steps).
* **hydration\_vs\_mood.png**: Visualization of hydration patterns across moods.
* **Python Code**: Script used for extraction, following the project’s template structure.

**Conclusion**

We successfully extracted and visualized features that can be used in the next steps of model training and evaluation. These features are expected to help in predicting user mood or lifestyle behavior based on tracked activity data.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.