Questions about Hourly_merged.csv

1. Dataset Understanding

- What kind of data are we dealing with?
  - Tabular Data, Time Series
  - Measurements of physical activity as well as the physiological responses
  - Different Patients(Id) data grouped by hour, given there Intensity, Steps and burned calories
- What are the key features in the dataset?
  - Calories, Intensity, Steps per hours

2. Summary Statistics

- What insights can you gather from the summary statistics?
  - Number of data points, (mean, std, min, max) of the values for each features
- Are there any features that have extremely high or low values?
  - TotalSteps high value with a max of 10554, high std as well
  - With respect to the mean, we have for Calories and TotalIntensity also a high discrepancy
- How would you interpret the standard deviation for features like Calories and StepTotal?
  - The high std in StepTotal indicates a large variability. This could be interpreted that patients have very different activity levels or that each individual patient have significant variations on activity from hour to hour
  - For the Calories its not that obvious what to interpret from just those numbers. There is a baseline for each patient dependent on certain factors how much calories get burned. This normally increases then while doing an activity like walking.

3. Data Visualization

- What patterns or trends can you observe from the time series plots?
  - Since the plots show data from different patients without considering a specific ordering, (df.head imply that the data is initially ordered patient by patient with orderd time), its hard to make statements about deeper pattern or trends
  - Nevertheless we can see, what we have observed already from the statistics, high variation from hour to hour in the data as well as clusters of activity and rest
  - The Histograms shows us that we have a strong data shift to zero values
    - This could indicate a large number of inactive periods
- Are there any noticeable outliers?
  - There seems an event around datapoint 8000, where the calories, Intensity and StepTotal dropped for a period of time
- How would you interpret the variability in Calories, TotalIntensity, and Step- Total?
  - Different patients, Different Activity(Behaviour) and Conditions, and a daily cycle e.g. a lot of people don't have a activity intensive job, so in there daily work there is no 8 hour moderate activity anymore, just doing sports before or after it for example. Back to the cluster argument

4. Missing Values

- How would you handle missing values if there were any?
  - It depends on how many data points, and how they are ordered through time e.g. is a full day or just an hour missing
  - For the 1hour missing situation, we could use an average of the hour before and after
  - For a day using Imputations e.g. fill the missing day with the average hour mean or median of the surrounding days.
- Why is it important to deal with missing values before proceeding to model training?
  - Many machine learning algorithms, including LSTMs, require a complete dataset as they cannot handle missing values intrinsically. These models need a full sequence of data to learn the temporal dependencies.

5. Data Scaling/Normalization

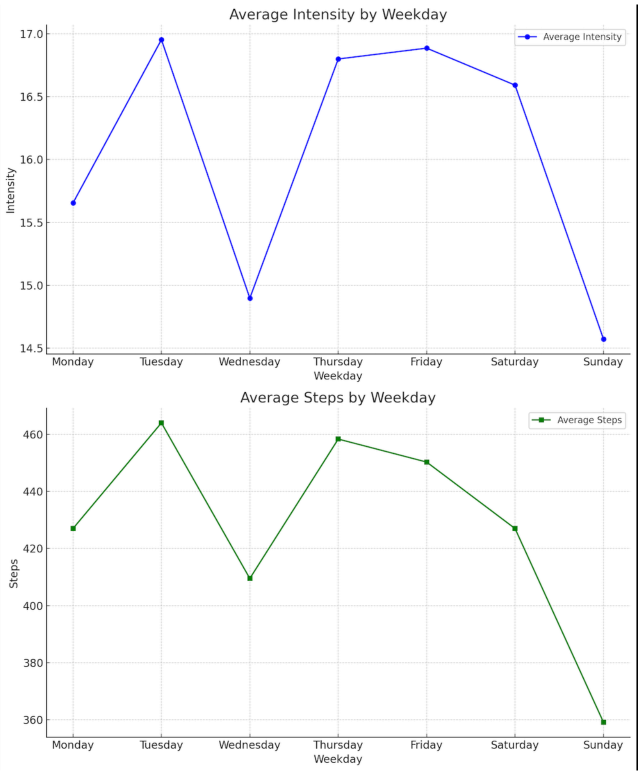
- Why is data scaling important in time-series analysis, especially when using LSTMs?
  - Converge faster when data is scaled
    - Normalisation helps making training more stable
  - LSTMS are sensitive to the scale of input data

- What are some other scaling methods you could use, and how do they differ from Min-Max scaling?
  - Standardisation
    - Scales data based on the standard deviation and mean of the dataset
    - Less sensitive to outliers compared to Min-Max scaling
    - The resulting distribution has a mean of 0 and a standard deviation of 1
    - Gives u less visual insights because we scale natural positive values like Steps and Calories into negative ranges
  - Robust Scaling
    - Use robust statistics that are not affected by outliers
    - Robust scaling centers and scales data using the median and interquartile range, reducing the influence of outliers.
      - whereas Min-Max scaling uses the minimum and maximum values, making it sensitive to outliers

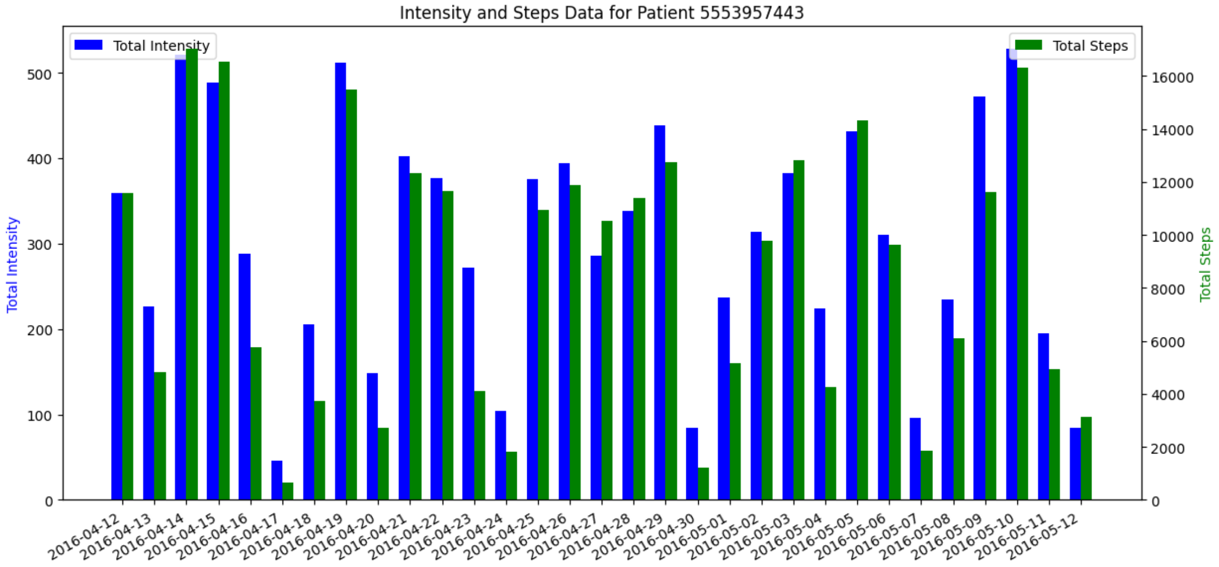
Programming Challenge

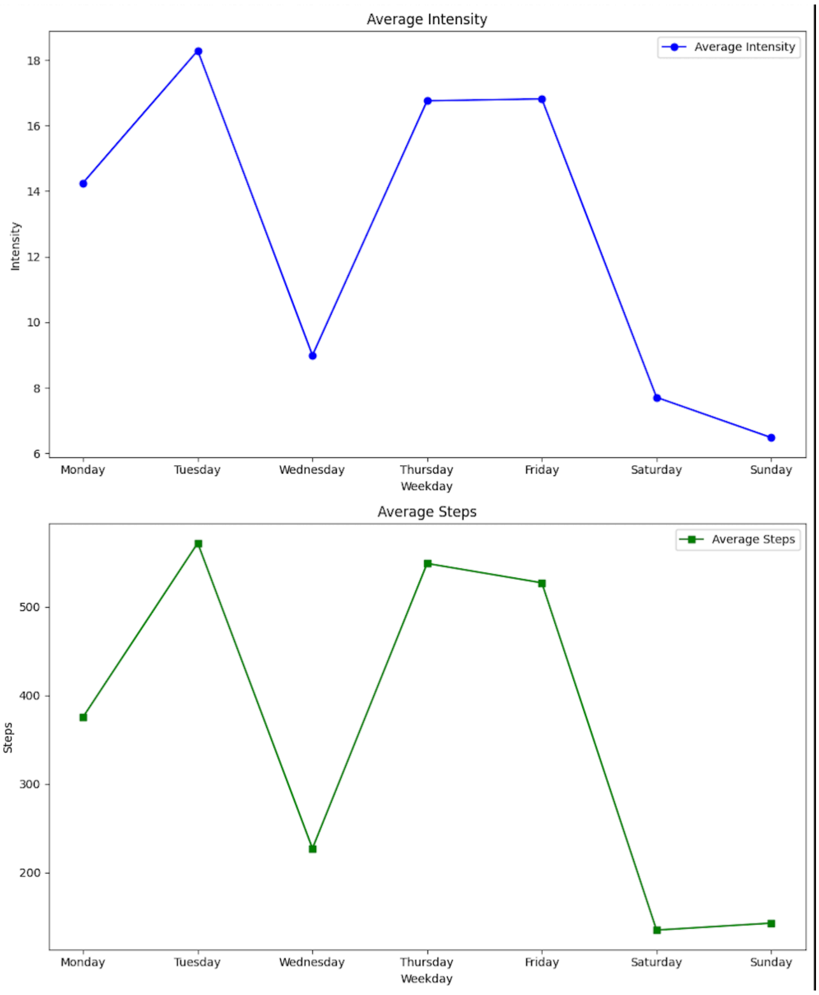
1. Data Exploration

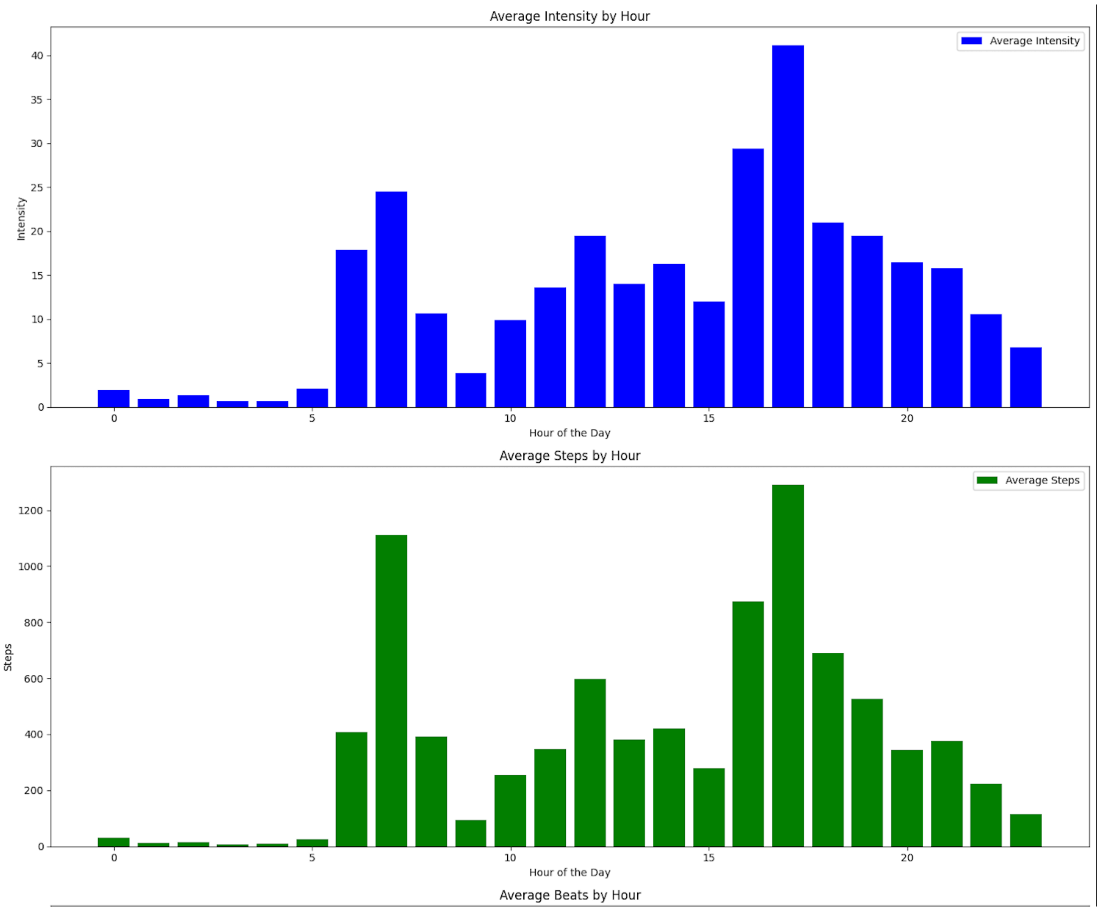
- How do the trends and patterns in the 2022 04 22 hour heartbeat merged.csv dataset compare to those observed in the Hourly merged.csv dataset?
  - They have the exact same max value for TotalSteps.
  - From the statistic side there is a smiliar pattern in variance for specific features like total Steps.
  - The activity variance is here also high, supporting the assumption that people have a higher activity just for specific time frames, (days, hours)
- Based on the time series plots of the two selected features, can we identify any recurring patterns or anomalies?
  - I have chosen Total Intensity and Total Steps.
  - Recurrent pattern, Intensity and Total Steps go down over the night obviously, but also on average on Wednesday and Sunday as u can see in the plots



Example for 1 Patient







2. Data Preprocessing

- After handling missing values, how does the distribution of the dataset change?
  - It depends how we interpret “missing” data, because from a time frame perspective not all patients start and end at the same time. Considering just the missing values within the time series for each patient (there are 6 datapoints with missing values).
    - For 1 patient there are 3 data points with missing values, at the end of the time series. Filled them with zero values to follow the pattern of this day
    - For the rest 3, I applied a forward fill, meaning I took the values from the hour before
    - Because of there are 6 values, the distribution get not really influenced
- Are there significant shifts in mean or variance?
  - The change in mean and variance is not significant.
- How does the scaling of training and testing data separately impact the range and distribution of the scaled values? If needed, compare it with scaled values without splitting the data.
  - They might have there own min and max value for a specific feature, so data gets scaled differently
- How might scaling the entire dataset (training + testing) introduce potential biases?
  - When the model is trained on data scaled with values of the test set, it may perform well on the test data because the scaling parameters were derived using the test set as well. This leads to an optimistic estimate of the model's performance on unseen data, which may not hold in practice.
    - So potentially biased towards the test set