FitPulse – Milestone 1

Data Collection & Preprocessing

Presented by **Divya Sharma**

Milestone Objective

Collect and preprocess comprehensive smartwatch data to enable actionable health insights and support data-driven decision making.

This milestone focuses on creating a **clean and structured dataset** from raw wearable data, including heart rate, sleep, steps, calories, and activity type.It establishes the foundation for building **real-time interactive dashboards**, performing **pattern analysis**, and training **predictive machine learning models** to provide personalized health recommendations.

Dataset Overview

Heart Rate

Beats per minute (bpm) tracked continuously

Sleep Duration

Hours of sleep recorded nightly

Step Count

Daily activity and movement tracking

Calories Burned

Energy expenditure monitoring

Activity Type

Classification of physical activities

Timestamp

5-minute interval recordings

Our dataset captures comprehensive health metrics every 5 minutes over a 7-day period, providing high-resolution insights into daily patterns and physiological responses.

Preprocessing Pipeline

Data Import & Validate

Load smartwatch data files (CSV/JSON) and verify that all required columns are present, data types are correct, and values fall within expected ranges.

Timestamp Normalization

Convert all timestamps into UTC format to ensure consistency across different time zones and devices, enabling accurate temporal analysis.

Missing Values Handling

Address gaps in the dataset using forward/backward fill, linear interpolation, mean/median imputation, or deletion of invalid entries to maintain data integrity.

Detect Outliers

Detect unusual or extreme values caused by sensor errors or anomalies, and treat them by replacing with averages or removing unrealistic entries to prevent skewed analysis.

Resample Data

Aggregate high-resolution data into uniform time intervals (minute, hour, or day) for smoother analysis. Use appropriate aggregation methods such as mean for heart rate or sum for steps.

Quality Assessment

Perform a final review of the dataset, checking for duplicates, missing values, and consistency. Ensure the dataset is clean, complete, and ready for advanced analytics and machine learning.

Step 1: Import & Validation

Data Loading Process

- Import raw CSV files using Pandas library
- Standardize column names for consistency
- Remove duplicate timestamp entries
- Verify data types and record counts

Initial Assessment: Dataset contains ~2,016 records per metric (7 days × 288 5-minute intervals per day), providing comprehensive coverage for analysis.

Steps 2–4: Normalize, Impute & Detect

1

Timestamp Normalization

Convert all timestamp entries to standardized datetime format, ensuring consistent timezone handling and enabling time-based operations.

2

Missing Value Handling

Apply mean imputation for heart rate and calories, median for step count, and mode for categorical activity types to preserve data distribution.

3

Outlier Detection

Identify anomalous readings using Z-score (±3σ) and Interquartile Range (IQR) methods to flag physiologically implausible values for review.

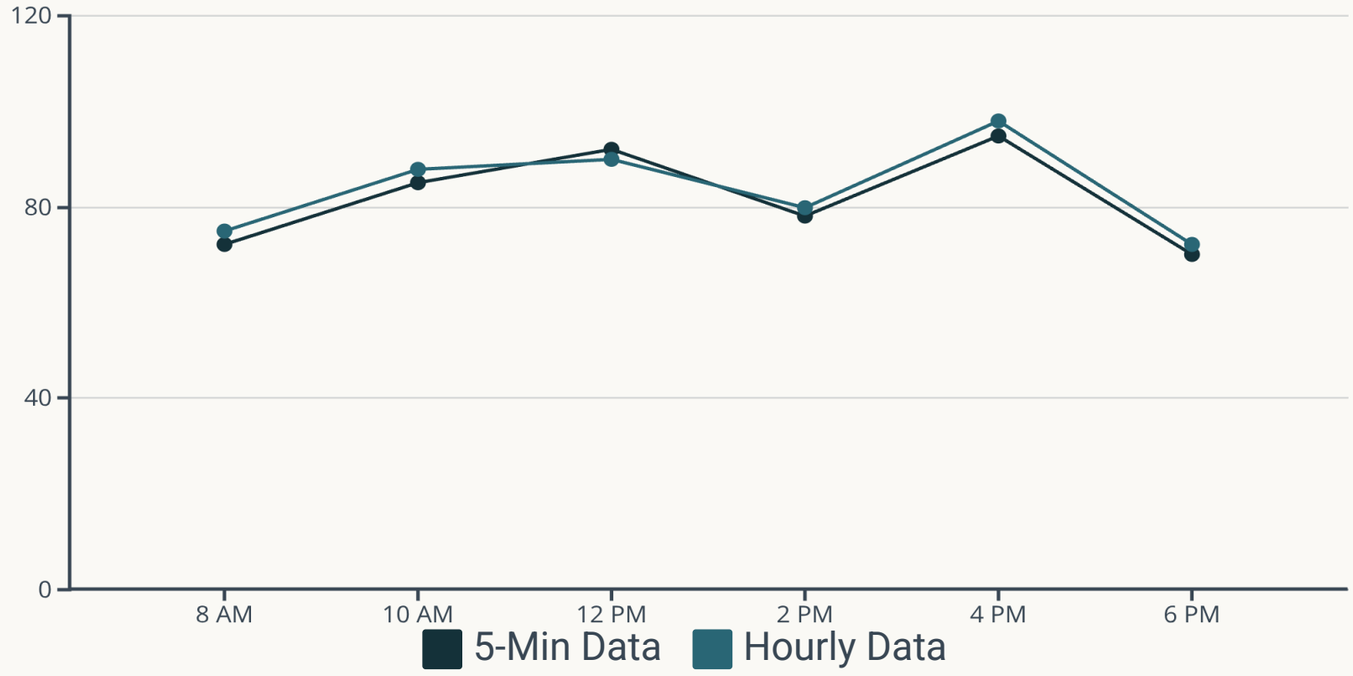
Step 5: Resampling Strategy

5-Minute to Hourly Aggregation

To reduce noise and improve analysis efficiency, we aggregate high-frequency 5-minute data into hourly intervals:

- **Numeric metrics:** Calculate mean (heart rate, calories, steps)
- **Categorical data:** Use mode (activity type)
- **Timestamp indexing:** Set as primary index for time-series operations

This approach preserves trends while reducing dataset size from 2,016 to 168 records per week.



Step 6: Data Quality Assessment

0

Missing Values

Complete dataset with no gaps after imputation

0

Duplicate Timestamps

All temporal duplicates successfully removed

~600

Repeated Values

Expected duplicates during rest/inactivity periods

100%

Data Integrity

Clean, validated, and analysis-ready

Final validation confirms the dataset meets quality standards for downstream analytics. Repeated values in activity data represent legitimate physiological states during sleep and sedentary periods, not data quality issues.

Expected Output Results of Milestone 1

After completing all steps in Module 1 — data ingestion, timestamp normalization, missing value handling, duplicate removal, outlier detection, and resampling — the final output will be:

Cleaned Dataset

- No missing values (handled or imputed)
- No duplicates
- Outliers removed or corrected

Standardized Timestamp

- All timestamps converted to UTC
- Consistent time format everywhere

Uniformly Resampled Data

- Heart rate, steps, sleep, etc. aligned into fixed time intervals (1-minute, hourly, or daily)

Noise-Reduced and Reliable Data

- Motion artifacts and random noise minimized
- Values fall in logical, realistic ranges

Ready-for-Modeling Dataset

- Structured, clean, consistent data
- Directly usable for feature extraction and anomaly detection in Module 2 & 3

Next Steps: From Data to Insights

1

Feature Extraction & Modeling Preparation

Generate time-series features, identify trends, cluster patterns, and prepare data for predictive models

2

Anomaly Detection

Detect unusual events in heart rate, sleep, or activity using statistical and model-based methods

3

Dashboard & Insights

Visualize health metrics in interactive dashboards and provide actionable daily/weekly insights

With a clean dataset, FitPulse is ready to extract insights, detect anomalies, and build dashboards for personalized health recommendations.

Thank You!