Bellabeat Data Cleaning & ETL Report

Prepared by: Daniel Darzi

Date: 2026-01-12
Google Data Analytics Capstone — Bellabeat Case Study

This document outlines the full data cleaning and transformation process applied to the Fitbit datasets used for the Bellabeat case study. The goal was to ensure that all files were accurate, consistent, and ready for reliable analysis. A structured ETL workflow was applied across all datasets, with detailed validation steps and helper columns added for transparency and auditability.

1. Initial Data Examination

The raw data was reviewed to identify structural issues, inconsistencies, and potential errors. Key checks included:

- Removing or flagging non-wear days
- Checking for impossible or illogical values
- Reviewing inconsistent activity records
- Verifying completeness across all fields
- Ensuring logical relationships between steps, distance, and activity minutes
- Adding helper columns to support validation and reproducibility

2. Data Cleaning & Preparation (Transform Stage)

A systematic ETL process was applied to ensure the dataset was clean, consistent, and analysis-ready.

A. Structural Validation

- Confirmed all expected columns were present in each file
- Ensured column names were consistent across datasets

- Validated data types (dates, integers, decimals, text)
- Checked for duplicate rows and removed them where necessary

B. Completeness Checks

- Identified missing values in key fields (steps, distance, calories, activity minutes)
- Flagged rows with incomplete or zeroed-out data
- Counted non-wear days using the rule:
  TotalSteps = 0 AND SedentaryMinutes = 1440

C. Logical Consistency Checks

- **Steps vs Distance:** flagged cases where distance > 0 but steps = 0
- **Activity Minutes vs Activity Distance:** checked for VeryActiveMinutes > 0 but VeryActiveDistance = 0
- **Daily Minutes Validation:** ensured all activity minutes summed to 1440 minutes per day
- **Calories vs Activity:** flagged unusually low calories on days with high activity

D. Outlier Detection

- Reviewed extremely high values in steps, distance, or calories
- Checked for negative or impossible values
- Flagged consecutive non-wear days that may indicate device abandonment

E. Data Quality Actions

- Flagged non-wear days for exclusion from activity-based analysis
- Retained minor inconsistencies for transparency but excluded them from summary statistics
- Added helper columns (e.g., **OriginalOrder = ROW()**) to preserve raw ordering
- Standardized date formats and ensured consistent user IDs

This structured approach ensured the final dataset was clean, reliable, and suitable for meaningful analysis.

3. Handling Minor Inconsistencies

A small number of rows showed inconsistencies between activity minutes and distance. These were retained because:

- They likely represent manual activity logs or Fitbit's internal categorization quirks
- They do not materially affect the analysis
- Transparency is preferred over unnecessary deletion

Some days also showed non-zero **TotalDistance** but zero activity-specific distances. This occurs when Fitbit records movement but does not classify it into intensity categories.
These rows were retained as valid.

4. Detailed Validation Steps

1. Verified Dataset Structure

- Confirmed all expected columns were present (steps, distances, activity minutes, calories, dates, IDs)
- Checked for missing values, nulls, and blank rows — none found
- Ensured each row represented one user-day (no duplicates after validation)

2. Validated Date Fields

- ActivityDate was originally stored as text
- Converted text dates (e.g., "3/25/2016") into real Excel dates using a custom parsing formula
- Verified conversion using **ISNUMBER()** to ensure Excel recognized the values as dates

3. Validated Time-Based Columns

- Confirmed that the sum of:
    - SedentaryMinutes
    - LightlyActiveMinutes

    - FairlyActiveMinutes
    - VeryActiveMinutes
      never exceeded **1440 minutes**
- No rows violated the 24-hour rule

4. Validated Distance Columns

- Confirmed **TotalDistance ≥ sum of activity-specific distances**
- Verified consistency between:
    - VeryActiveMinutes ↔ VeryActiveDistance
    - ModeratelyActiveMinutes ↔ ModeratelyActiveDistance
    - LightlyActiveMinutes ↔ LightActiveDistance
- No negative or impossible values were found

5. Logical Consistency Checks

- Ensured steps > 0 on days with non-zero active minutes
- Ensured calories > 0 for all rows
- Confirmed no negative values in any numeric column
- Verified that user IDs were valid and consistent

6. Data Issue Resolution

- No rows required deletion
- No outliers required removal (Fitbit data naturally varies)
- The primary correction was converting dates from text to real Excel date format

7. Helper Columns Added

- **TotalMinutesCheck** = sum of all activity minutes (used for 24-hour validation)
- All rows passed validation

✓ Cleaning Phase Completed

The dataset is now:

- Structurally clean

- Logically consistent
- Free of impossible values
- Ready for feature engineering and analysis

5. Timestamp Cleaning Across All Files

Many Fitbit timestamps were stored as text and required conversion.

General Timestamp Cleaning

- Identified timestamps stored as text
- Used **Text to Columns** to split and convert date/time components
- Verified conversion using **ISNUMBER()**
- Applied custom datetime formats (e.g., yyyy-mm-dd hh:mm:ss AM/PM)
- Combined date + time using:
  =J2 + K2
- Adjusted cell formatting to ensure proper display

hourlyIntensities_merged

- Converted ActivityHour to real Excel datetime
- Verified TotalIntensity and AverageIntensity were numeric
- No further cleaning required

minuteCaloriesNarrow_merged

- Converted ActivityMinute to real Excel datetime
- Verified minute-level timestamps were accurate
- Numeric columns required no transformation

minuteMETsNarrow_merged

- Converted ActivityMinute to real Excel datetime
- Verified METs values (0–159) were within Fitbit's normal range
- No transformation required

minuteSleep_merged

- Converted sleep timestamps to real Excel datetime
- Identified sleep stage codes:
    - 1 = Asleep
    - 2 = Awake
    - 3 = Restless
- Confirmed only valid codes were present
- Optional sleep stage label created using:
  =IFS(C2=1,"Asleep", C2=2,"Awake", C2=3,"Restless", TRUE,"Unknown")
- Considered feature engineering, not required for cleaning

minuteStepsNarrow_merged

- Converted ActivityMinute to real Excel datetime
- Verified steps column represented steps taken in that minute
- No cleaning required

weightLogInfo_merged

- Converted Date column to real Excel date
- Interpreted **IsManualReport**:
    - TRUE = user manually entered weight
    - FALSE = recorded automatically by Fitbit scale
- Added optional readable label:
  =IF(D2,"Manual","Scale")
- No further cleaning required