**Bellabeat Case Study Report**

Bellabeat Leaf Data Analytics

Prepared by: Somayah Alhabib

Course: Google Data Analytics Capstone

Date: Auguste 2025

**Executive summary**

The analysis explored four key datasets (daily activity, sleep, heart rate, and weight logs). After cleaning and aggregating, daily activity and sleep were merged for deeper insights. The visualizations show:

* Sleep distribution: Most users sleep between 400–500 minutes per night.
* Sleep vs. steps: No strong relationship between total sleep and steps taken per day.
* Sedentary minutes vs. sleep: Higher sedentary hours often coincide with slightly less sleep.

Weight and heart rate logs were too sparse for full integration, but they highlight underused features and opportunities for Bellabeat.

These insights suggest targeted product recommendations, such as encouraging bedtime routines for short sleepers and promoting activity challenges for users with high sedentary time.

**Ask step:**

**Business Task:**

Analyze smart device usage data from non-Bellabeat products to uncover wellness engagement trends and generate actionable insights that inform the marketing strategy for the Bellabeat Leaf.

**Key Stakeholders:**

- Urška Sršen: Bellabeat’s cofounder and Chief Creative Officer

- Sando Mur: Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team

**Business Objective:**

1. What are some trends in smart device usage?

2. How could these trends apply to Bellabeat customers?

3. How could these trends help in influence Bellabeat marketing strategy?

The provided Fitbit dataset folder contains multiple CSV files. For this case study, I selected four datasets that provide the most relevant and actionable insights for Bellabeat Leaf device:

* **dailyActivity\_merged.csv**
* **minuteSleep\_merged.csv**
* **heartrate\_seconds\_merged.csv**
* **weightLogInfo\_merged.csv**

Other files (e.g., hourly steps, minute-level activity logs) were excluded because they contained redundant or overly granular information that was not essential for answering the business question.

**Where is the data stored?**

* The data is provided in a public Fitbit dataset folder and stored locally as .csv files.

**How is the data organized? (long vs. wide format)**

* dailyActivity\_merged.csv – wide format (daily summaries).
* minuteSleep\_merged.csv – initially long format (minute-by-minute), aggregated to daily summaries.
* heartrate\_seconds\_merged.csv – long format (second-level), aggregated to minute-level.
* weightLogInfo\_merged.csv – wide format (daily weight and BMI entries).

**Are there issues with bias or credibility? (ROCCC test)**

* Reliable: Collected from Fitbit devices, but dependent on user compliance.
* Original: Data is device-generated, but sample size is small (30 users).
* Comprehensive: Covers key wellness measures (activity, sleep, heart rate, weight) but lacks demographic details.
* Current: Data is outdated (2016) and may not fully reflect current trends.
* Cited: Dataset is publicly available but not officially released by Fitbit.

**How are licensing, privacy, security, and accessibility addressed?**

* Data is anonymized using IDs, ensuring user privacy.
* Publicly available for educational purposes; no licensing conflicts.
* Stored securely in local project files.

**How was data integrity verified?**

* Checked and removed duplicates.
* Standardized date-time formats.
* Filtered missing values.
* Added logical validation checks (e.g., flagged cases where calories burned with zero steps, verified daily totals ≤ 1,440 minutes).

**How does this data help answer the business question?**

* Provides a holistic view of user activity, sleep, heart health, and weight trends—all directly relevant to Bellabeat’s focus on women’s wellness and lifestyle insights.

**Are there any problems with the data?**

* Limited participant pool (not representative of all populations).
* Missing demographic variables (age, gender, etc.).
* Some unrealistic values (corrected during cleaning).

**Process Step:**

**Guiding Questions**

* **What tools are you choosing and why?**
  + I used **Excel** for datasets where spreadsheet-based cleaning and validation was sufficient (daily activity and weight logs).
  + I used **R** for larger, more complex datasets (sleep and heart rate), as it provides efficient data manipulation, aggregation, and transformation functions.
  + This combination allowed me to leverage Excel’s ease of use for structured data while using R’s power for processing high-volume, time-series data.
* **Have you ensured your data’s integrity?**
  + Yes. All datasets were checked for duplicates, blanks, incorrect formats, and logical consistency.
  + Validation checks were added (e.g., calories vs. steps, daily minute totals ≤ 1,440).
  + Aggregations (minute-to-day, second-to-minute) were tested against raw data to confirm accuracy.
* **What steps have you taken to ensure that your data is clean?**
  + Removed duplicate and blank entries.
  + Standardized date and time formats.
  + Renamed columns for clarity and consistency.
  + Dropped redundant or incomplete variables.
  + Aggregated raw datasets into more usable daily/summary formats.
* **How can you verify that your data is clean and ready to analyze?**
  + Conducted logical checks (e.g., flagged unrealistic values, verified daily totals).
  + Cross-checked before-and-after values during transformations.
  + Ensured datasets had consistent user IDs and date ranges, making them ready for merging.
* **Have you documented your cleaning process?**
  + Yes. Each dataset includes a record of the cleaning steps performed, along with tool selection. This ensures transparency and reproducibility.

**Deliverable: Cleaning & Transformation Summary**

* **dailyActivity\_merged.csv** (*Cleaned in Excel*)
  + Provides daily summaries of steps, calories, and activity minutes.
  + Duplicates removed, blanks filtered, and columns renamed for clarity.
  + Logical validation performed:
    - “Check” column flagged impossible cases (e.g., calories burned with zero steps).
    - “Check minute totals vs day limits” column ensured daily totals ≤ 1,440 minutes.
* **minuteSleep\_merged.csv** (*Cleaned in R*)
  + Originally minute-level data, aggregated into daily summaries for usability.
  + Numeric codes are mapped to descriptive stages (Light, Deep, REM).
  + Duplicates removed and missing values handled.
  + Unrealistic totals flagged (e.g., sleep > 24 hours/day).
  + Result: compact dataset ready for merging with other health data.
  + The dataset was initially too large to be handled directly in **R**. To address this, the dataset was first reduced in **Excel** to make it manageable, and then imported into **R** for further cleaning,
* **heartrate\_seconds\_merged.csv** (*Cleaned in R*)
  + Second-level heart rate readings are converted into standardized date-time format.
  + Duplicates removed and completeness verified.
  + Aggregated into minute-level readings to reduce noise and improve usability.
  + Both raw and summarized datasets are retained for flexible analysis.
* **weightLogInfo\_merged.csv** (*Cleaned in Excel*)
  + Daily weight and BMI records.
  + Standardized date formats and removed duplicates.
  + Redundant/incomplete variables (e.g., Weight Pounds, Fat) dropped.
  + Retained key fields: Id, Date, Weight Kg, BMI, Is Manual Report, LogId.
  + Numeric values formatted consistently.
  + few users recorded it

**Analysis Step:**

**How should you organize your data to perform analysis on it?**

* The four cleaned datasets were aggregated and aligned on User ID and Date to enable comparisons across activity, sleep, heart rate, and weight.
* Datasets were reformatted into consistent daily summaries so they could be merged when needed.

**Has your data been properly formatted?**

* Yes. Dates were standardized, numeric fields were reformatted consistently, and all categorical variables (e.g., sleep stages) were labeled clearly.
* Outliers and unrealistic values were flagged during cleaning to ensure accuracy.

**What surprises did you discover in the data?**

* Many users logged very little sleep data, showing gaps in usage or device tracking limitations.
* Some users recorded high calorie burns despite low activity, suggesting logging errors or anomalies in device readings.
* Only a small portion of users logged weight data, indicating that weight-tracking features may be underutilized.

**What trends or relationships did you find in the data?**

* Steps and Calories: Strong positive correlation—users who walked more consistently burned more calories.
* Sleep and Activity: Some users with high activity showed better sleep duration, but the relationship was inconsistent. “Before merging, we reformatted the date columns in both datasets to ensure they shared the same format. Without this step, the merge would have failed because the dates were stored as character strings.”
* Heart Rate and Activity: Periods of high activity were reflected in increased average heart rates, confirming device accuracy.
* Weight and Activity: Users who logged weight tended to also have higher engagement with activity tracking, suggesting more health-conscious behavior.

**How will these insights help answer your business questions?**

Insights show how users engage with daily activity, sleep, and health tracking, which informs Bellabeat on which features (activity tracking, sleep insights, mindfulness, weight monitoring) resonate most with users.

* Identifying gaps (e.g., underuse of weight logs, inconsistent sleep data) helps Bellabeat understand where user experience improvements may be needed.
* The correlations between activity, calories, and sleep support targeted wellness recommendations that align with Bellabeat’s focus on promoting healthy habits

**Share Step:**

**Who was studied**

* Participants and record counts for sleep & activity (and note weight is sparse).

**What was seen**

* Sleep distribution (how many minutes people actually sleep).
* Relationship: Sleep vs Steps and Sedentary vs Sleep.
* Weekday patterns (avg sleep and steps by weekday).

**Why it matters**

* Short, concrete takeaways for Bellabeat marketing/product.

**What is recommended**

* A few targeted, data-driven suggestions (e.g., bedtime nudges for short sleepers; step streak challenges for low-activity days).

**A graph of a sleep

AI-generated content may be incorrect.**

* Shows how much sleep users actually get, most clustering around 400–500 minutes.

**A graph showing a number of steps

AI-generated content may be incorrect.**

* Checks whether more sleep correlates with more daily activity. It looks scattered, meaning no strong relationship.

**A graph showing red dots

AI-generated content may be incorrect.**

* Suggests that higher sedentary minutes often come with slightly less sleep.

**Act Step:**

**Final Conclusion**

The analysis showed that most participants sleep fewer than the recommended 7–8 hours, and there is a weak or inconsistent relationship between sleep duration and physical activity. Sedentary time remains high, even among participants with moderate step counts. Weight and heart rate data were sparse, suggesting limited user engagement with logging features.

**Deliverable**

* Many users are not consistently achieving healthy sleep durations.
* Higher daily steps do not necessarily translate to better sleep quality or duration.
* Prolonged sedentary periods are common, highlighting opportunities for movement reminders.
* Weight and heart rate data were logged by very few users, suggesting that these features may need better integration or nudging to drive adoption.

**Next Steps for Bellabeat**

* **Product feature ideas**: Introduce bedtime reminders, sleep streak rewards, and gentle nudges for sedentary breaks.
* **Marketing**: Position Bellabeat as not only an activity tracker but also a sleep wellness tool.
* **Future data collection**: Encourage users to record weight and heart rate more consistently to strengthen holistic health insights.