Bellabeat Case Study: Smart Device Usage Insights

Esteban Segura

2025-07-25

1 Business Task

Bellabeat wants to unlock new growth opportunities by leveraging data from smart device usage. As a junior data analyst on the marketing analytics team, I have been tasked with analyzing usage data from non-Bellabeat smart devices to identify trends and user behaviors. The goal is to apply these insights to one Bellabeat product and generate high-level, data-driven recommendations to inform and support Bellabeat's marketing strategy for the Leaf.

2 Executive Summary

This analysis explores Fitbit data from 33 users to uncover behavioral patterns in physical activity, heart rate, and sleep. Using PostgreSQL, R, and Python, the dataset was cleaned, explored, and visualized to identify user habits and business opportunities for Bellabeat's Leaf product. The findings highlight several actionable trends, including low average step counts, suboptimal sleep, and identifiable user tiers. Based on these insights, we recommend targeted strategies to improve user engagement, promote healthy routines, and personalize marketing campaigns.

3 Data Sources

Primary dataset:

- **Name:** FitBit Fitness Tracker Data
- **Source:** Kaggle (Mobius Lab)
- **License:** Public Domain (CC0)
- **Period:** April 12, 2016 – May 12, 2016
- **Format:** CSV files

- **Sample size:** 33 users

Organization:

Data is split across two time windows:

- March 12 – April 11, 2016
- April 12 – May 12, 2016

For this analysis, only the April 12 – May 12, 2016 dataset (18 files) was used, as the March dataset was incomplete.

Files are structured in wide format for most day-level data (e.g., daily activity) and long format for minute/second-level logs (e.g., heart rate, steps, sleep).

Files used in this analysis:

- `dailyActivity_merged_april.csv` : Aggregated user activity per day
- `sleepDay_merged_april.csv` : Minutes asleep vs restless
- `minuteSleep_merged_april.csv` : Minute-level data of user sleep states
- `heartrate_seconds_merged_april.csv` : Heart rate by second

Limitations:

- Only 33 users: Small sample size may limit generalization
- Self-reported data (e.g., weight) may contain inconsistencies
- Some users contributed very little data (as few as 4 days)
- Short data collection window (~1 month used)
- MTurk-based sampling may not represent Bellabeat’s audience
- Only 11 of the 18 expected files were available for the first month (March 12 – April 11, 2016), limiting the ability to analyze certain behaviors across both months.
- While the dataset includes 33 unique users with daily activity data, only 24 users provided sleep data and 14 users provided heart rate data. Therefore, different analyses use different sample sizes depending on data availability.

Integrity and Suitability

- Data was reviewed for completeness, structure, and timestamp consistency
- Data types and formats were validated before loading into the analysis environment (PostgreSQL via pgAdmin4)
- The dataset helps answer the key business questions by offering insight into users’ *daily habits*, such as activity levels, heart rate patterns, sleep duration, and more — all of which are relevant for Bellabeat’s smart wellness products, particularly the Leaf tracker, which monitors activity, sleep, and stress.

---

4 Documentation of Data Cleaning and Transformation (SQL)

For this analysis, I used PostgreSQL as the primary tool, accessed through pgAdmin 4. This environment was selected because:

- Several datasets exceeded the row limit or processing capacity of spreadsheet tools like Excel.
- SQL allows efficient querying, joining, filtering, and transformation of large datasets.
- Using a centralized database provides consistency across all datasets and supports scalable, reproducible analysis.

4.1 **Steps Taken to Process the Data**

4.1.1 File Management and Import Preparation

- All original CSV files were extracted and organized into a clearly named directory: `D:/Data Analysis/.../fitbit_csvs`
- Files were renamed to indicate their date range for traceability (e.g., `dailyActivity_merged_april.csv`).
- Duplicates or inconsistencies across folders were noted — only the April 12 to May 12, 2016 files were used due to completeness.

4.1.2 Uploading Data into PostgreSQ

Python with pandas, sqlalchemy, and psycopg2 were used to:

- Create a `raw_data schema` in the fitbit_data PostgreSQL database.
- Automatically scan, read, and import each CSV file into a corresponding SQL table.
- Clean table names (remove spaces, convert to lowercase) during upload for easier querying.
  - Example: `minuteStepsNarrow_merged.csv` → `raw_data.minutestepsnarrow_merged`

```
import os
import pandas as pd
from sqlalchemy import create_engine, text

# PostgreSQL connection info
username = '...' # Replace with actual username
password = '...'  # Replace with actual PostgreSQL password
host = '...' # Replace with actual host
port = '...' # Replace with actual port
database = 'fitbit_data'
schema = 'raw_data'

# Create SQLAlchemy connection
engine = create_engine(f'postgresql+psycopg2://{username}:{password}@{host}:{port}/{database}')

# Create schema if it doesn’t exist
with engine.connect() as connection:
    connection.execute(text("CREATE SCHEMA IF NOT EXISTS raw_data;"))

# Folder with CSVs
csv_folder = r'...' # Replace with actual location of the folder

# Loop through and upload CSVs
for filename in os.listdir(csv_folder):
    if filename.endswith('.csv'):
        file_path = os.path.join(csv_folder, filename)
        table_name = filename.replace('.csv', '').lower().replace('-', '_').replace(' ', '_')
        print(f'📥 Importing {filename} as table "{schema}.{table_name}"...')
        try:
            df = pd.read_csv(file_path)
            df.to_sql(table_name, con=engine, schema=schema, if_exists='replace', index=False, method='multi')
            print(f'✅ Done: {schema}.{table_name}')
        except Exception as e:
            print(f'❌ Error importing {filename}: {str(e)}')

print("\n✅ All files processed.")
```

4.2 Data Cleaning and Transformation (in SQL)

All cleaning and processing steps were documented through SQL scripts in pgAdmin4 using PostgreSQL. The actions below outline each major step followed by the relevant SQL code.

4.2.1 Standardizing Column Names and Converting Date Columns

To ensure consistency and correct data types, columns were renamed where necessary, and all date columns were converted from text to SQL `DATE` or `TIMESTAMP` types.

```sql
CREATE TABLE clean.daily_activity AS
SELECT
  "Id",
  TO_DATE("ActivityDate", 'MM/DD/YYYY') AS date,
  "TotalSteps", "TotalDistance", "VeryActiveDistance",
  "ModeratelyActiveDistance", "LightActiveDistance",
  "SedentaryActiveDistance", "VeryActiveMinutes",
  "FairlyActiveMinutes", "LightlyActiveMinutes",
  "SedentaryMinutes", "Calories"
FROM raw_data.dailyactivity_merged_april
WHERE "Id" IS NOT NULL;
```

4.2.2 Cleaning Sleep Day Data

Converted `SleepDay` column to date and removed rows with missing IDs.

```sql
CREATE TABLE clean.sleep_day AS
SELECT
  "Id",
  TO_DATE("SleepDay", 'MM/DD/YYYY') AS date,
  "TotalSleepRecords", "TotalMinutesAsleep", "TotalTimeInBed"
FROM raw_data.sleepday_merged_april
WHERE "Id" IS NOT NULL;
```

4.2.3 Extracting Timestamps and Cleaning Heart Rate Data

Separated date and time from full timestamp for minute-level analysis.

```sql
CREATE TABLE clean.hr_sec AS
SELECT
  "Id",
  TO_DATE("time", 'YYYY/MM/DD') AS date,
  TO_TIMESTAMP("time", 'YYYY/MM/DD HH24:MI:SS')::TIME AS time,
  "value"
FROM raw_data.heartrateseconds_merged_april
WHERE "Id" IS NOT NULL;
```

4.2.4 Extracting Time and Date from Minute-Level Sleep Logs

Converted datetime strings to `DATE` and `TIME` for time-based analysis.

```sql
CREATE TABLE clean.sleep_min AS
SELECT
  "Id",
  TO_DATE("date", 'YYYY-MM-DD') AS date,
  TO_TIMESTAMP("date", 'YYYY-MM-DD HH24:MI:SS')::TIME AS time,
  "value"
FROM raw_data.minutesleep_merged_april
WHERE "Id" IS NOT NULL;
```

4.2.5 Handling Missing Values (NULLs)

- Rows with `NULL` IDs were excluded.
- For activity and sleep, `NULL` s were retained when meaningful (e.g., no steps recorded).
- For attributes like weight or height, rows with missing critical fields were excluded.

4.3 Validating Numeric Data

Verified there were no negative values in key numeric columns.

```
SELECT *
FROM raw_data.dailyactivity_merged_april
WHERE "TotalSteps" < 0 OR "Calories" < 0;
-- No invalid records found
```

4.3.1 Removing Duplicates

Removed duplicate rows (exact copies) in logs such as sleep or heart rate.

```
CREATE TABLE clean.sleep_day AS
SELECT DISTINCT *
FROM raw_data.sleepday_merged_april;
```

4.3.2 Trimming Whitespace from Categorical Fields

Trimmed whitespace in non-numeric columns to avoid joining issues.

```
SELECT TRIM("Id") AS trimmed_id
FROM raw_data.sleepday_merged_april;
```

4.3.3 Validating Date Range

Validated that all user records in the dataset fell within the expected date range (April 12 – May 12, 2016).

```
SELECT
    "Id",
    MIN(date) AS earliest_date,
    MAX(date) AS latest_date
FROM
    clean.daily_activity
GROUP BY
    "Id"
ORDER BY
    earliest_date,
    latest_date
-- No invalid records found
```

4.3.4 Data Integrity Checks

- Verified row counts before and after each transformation to ensure complete data loads.
- Spot-checked SQL outputs against original CSV files to validate accuracy.
- Queried unique user IDs and dates to confirm the dataset covered the intended time frame.
- Ensured all timestamps fell within the selected analysis period: April 12 – May 12, 2016.

Key Point

The Fitbit datasets are primarily composed of numeric sensor logs. As a result, no extensive text cleaning or reclassification was required beyond ensuring correct data types, logical ranges, and structural integrity. This made the data suitable for direct analysis after transformation.

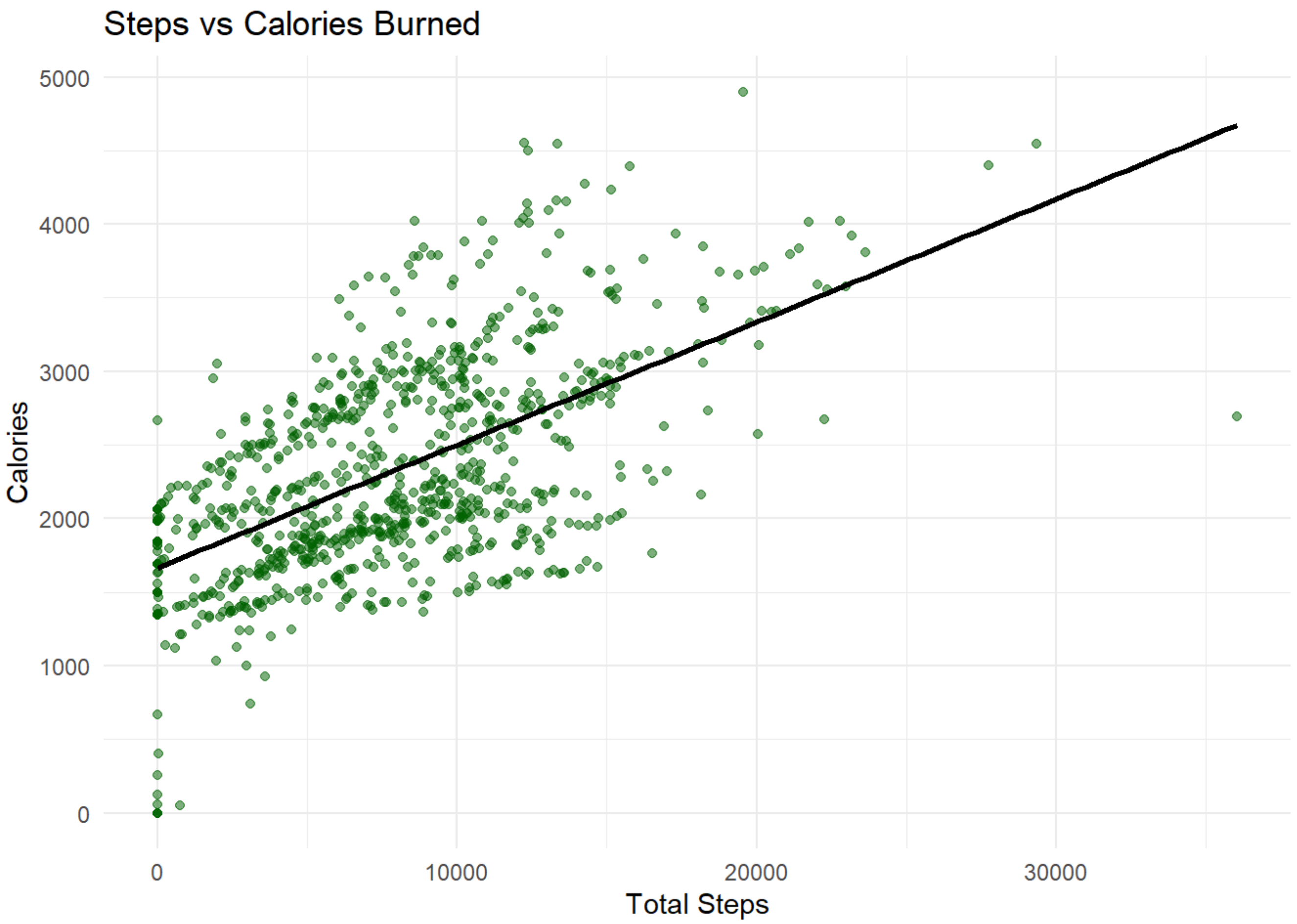
5 Data Exploration and Visualization

The visualizations below are designed to help the Bellabeat team quickly grasp key behavioral patterns among smart device users. Each plot includes a brief summary to guide interpretation and tie the insight back to the business question: How do users interact with wellness trackers, and how can this inform Bellabeat's marketing strategy for one of its products?

5.1 Physical Activity Patterns

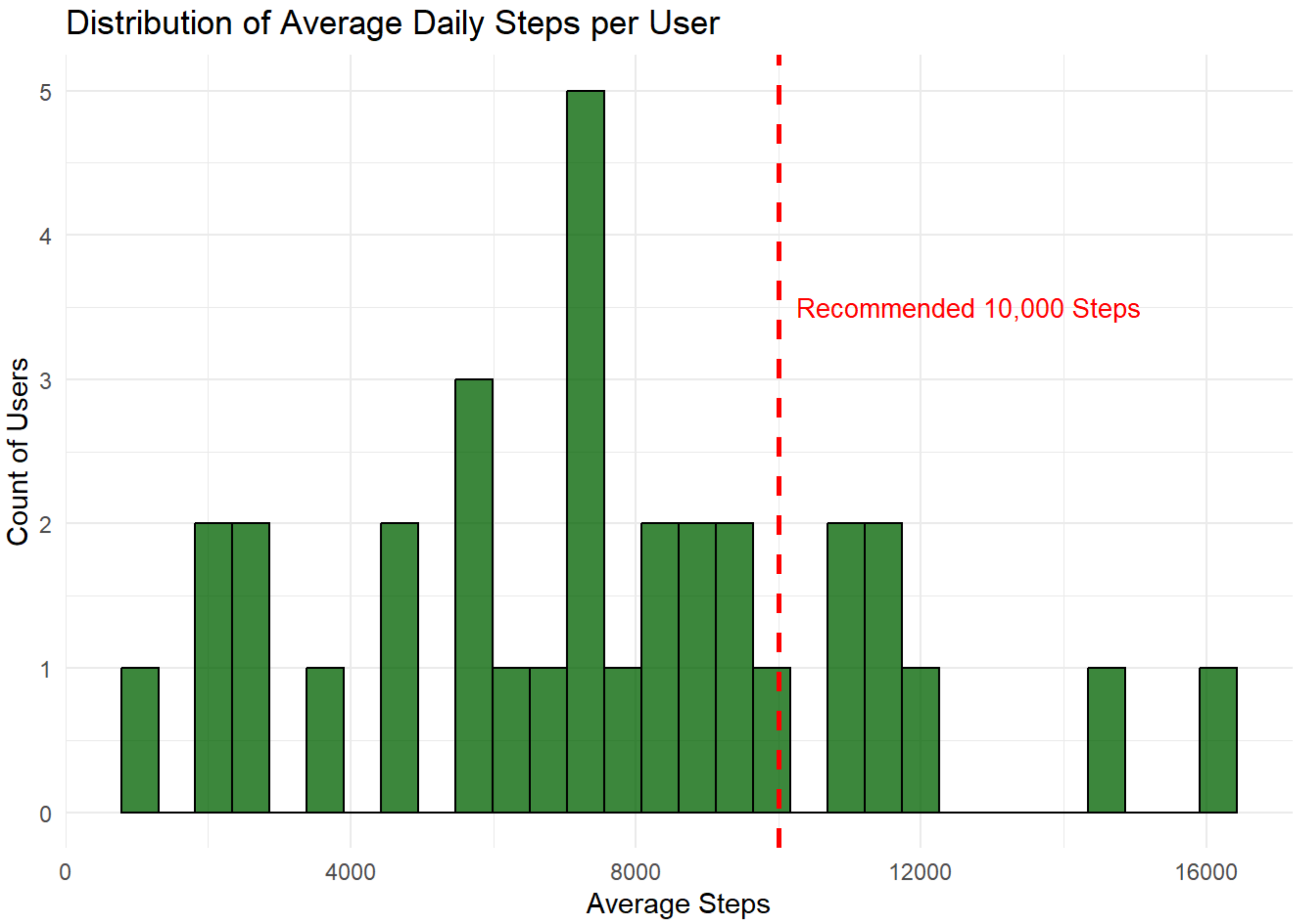
5.1.1 Relationship Between Steps and Calories

Users who take more steps tend to burn more calories, showing a strong positive correlation. This validates the connection between physical activity and calorie expenditure — an insight Bellabeat can emphasize when promoting its activity-tracking features.



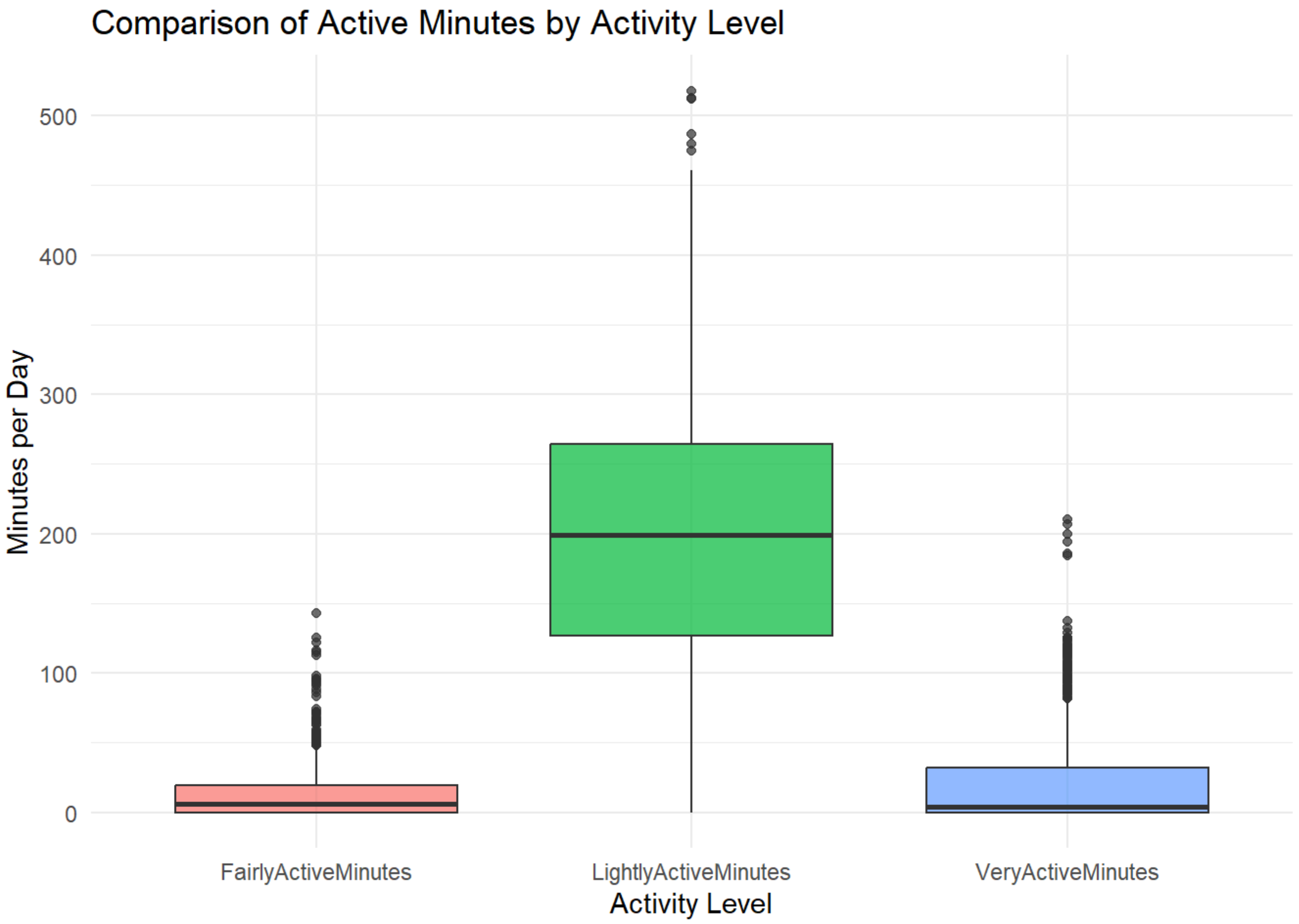
5.1.2 Average Daily Steps per User

Most users average fewer than 10,000 steps per day, indicating a generally moderate to sedentary user base. Bellabeat could focus its messaging on motivating users to reach daily movement goals using the Leaf device's reminders or rewards.



5.1.3 Active Minutes Comparison

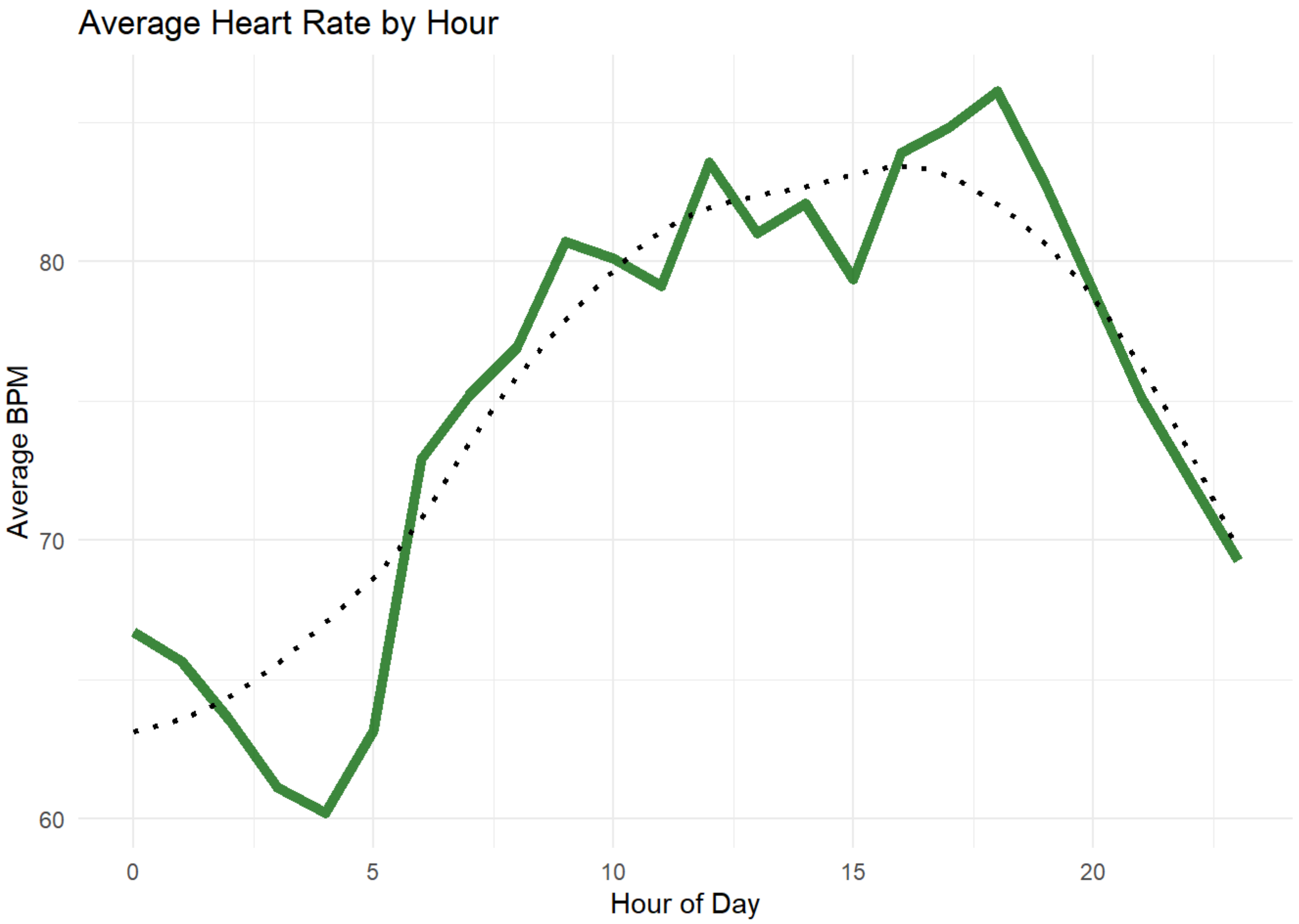
Light activity dominates daily routines, while very active minutes are limited. This suggests an opportunity for Bellabeat to differentiate itself by offering challenges or guided routines to increase vigorous activity.



5.2 Heart Rate Trends

5.2.1 Average Heart Rate by Hour

Heart rate tends to rise in the morning and stabilize in the afternoon/evening. Bellabeat could use this information to schedule stress-management nudges (e.g., breathing reminders) during morning peaks or early work hours.



5.2.2 Heart Rate Range Summary

The average heart rate among users was ~77 BPM, ranging from 36 to 203 BPM. These numbers fall within typical adult ranges, supporting the device's accuracy and potential for wellness monitoring.

Minimum Heart Rate	Maximum Heart Rate	Average Heart Rate	Median Heart Rate
36	203	77.32842	73

5.2.3 User Segmentation or Clustering (Even Simple Tiers)

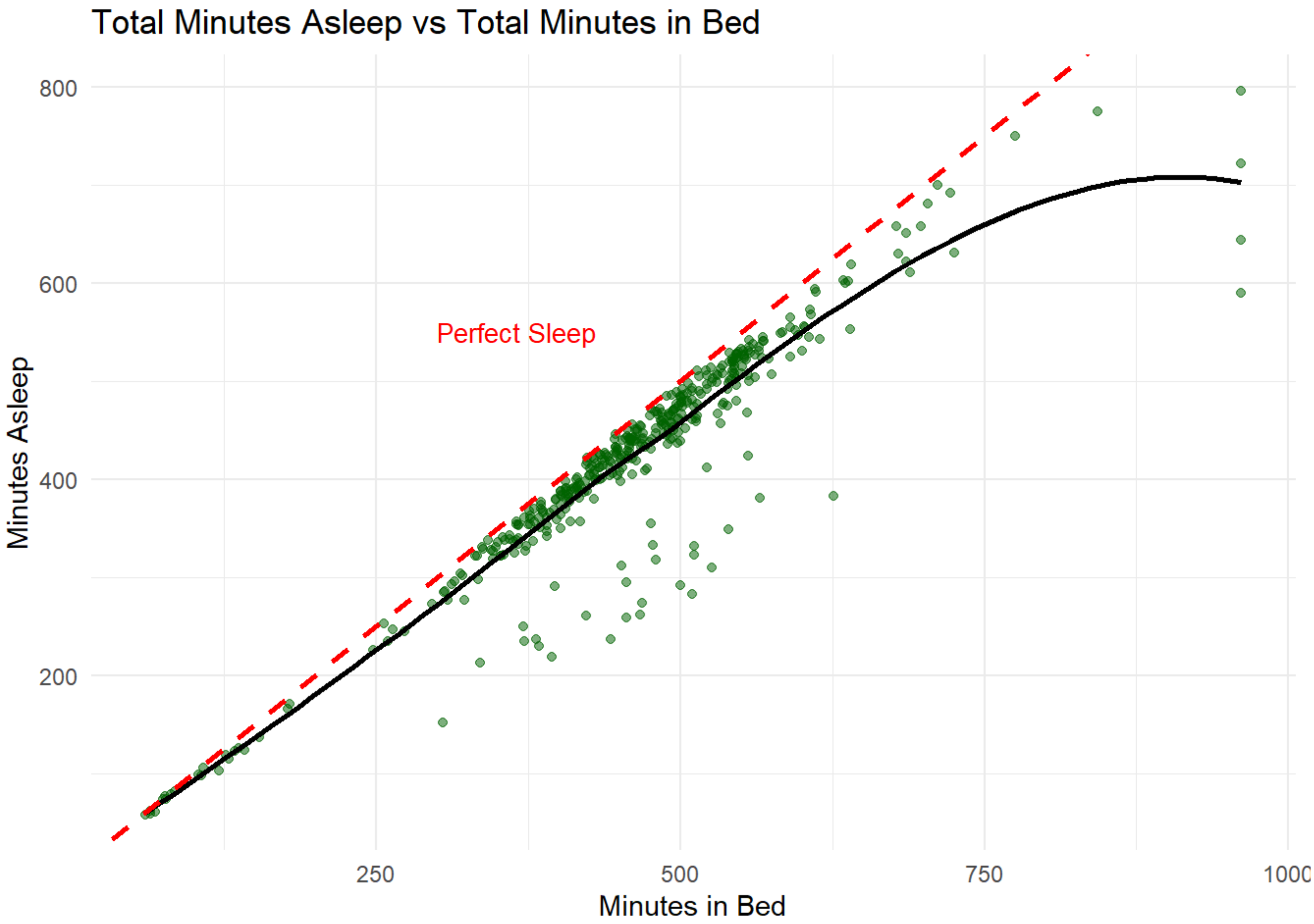
Users can be segmented into three tiers: sedentary (24%), moderately active (55%), and highly active (21%). Bellabeat should tailor marketing strategies and in-app features to appeal to each group — for example, offering beginner-friendly goals to sedentary users and performance features to active users.

Activity Level	Number of Users
Highly Active	7
Moderately Active	18
Sedentary	8

5.3 Sleep Behaviors

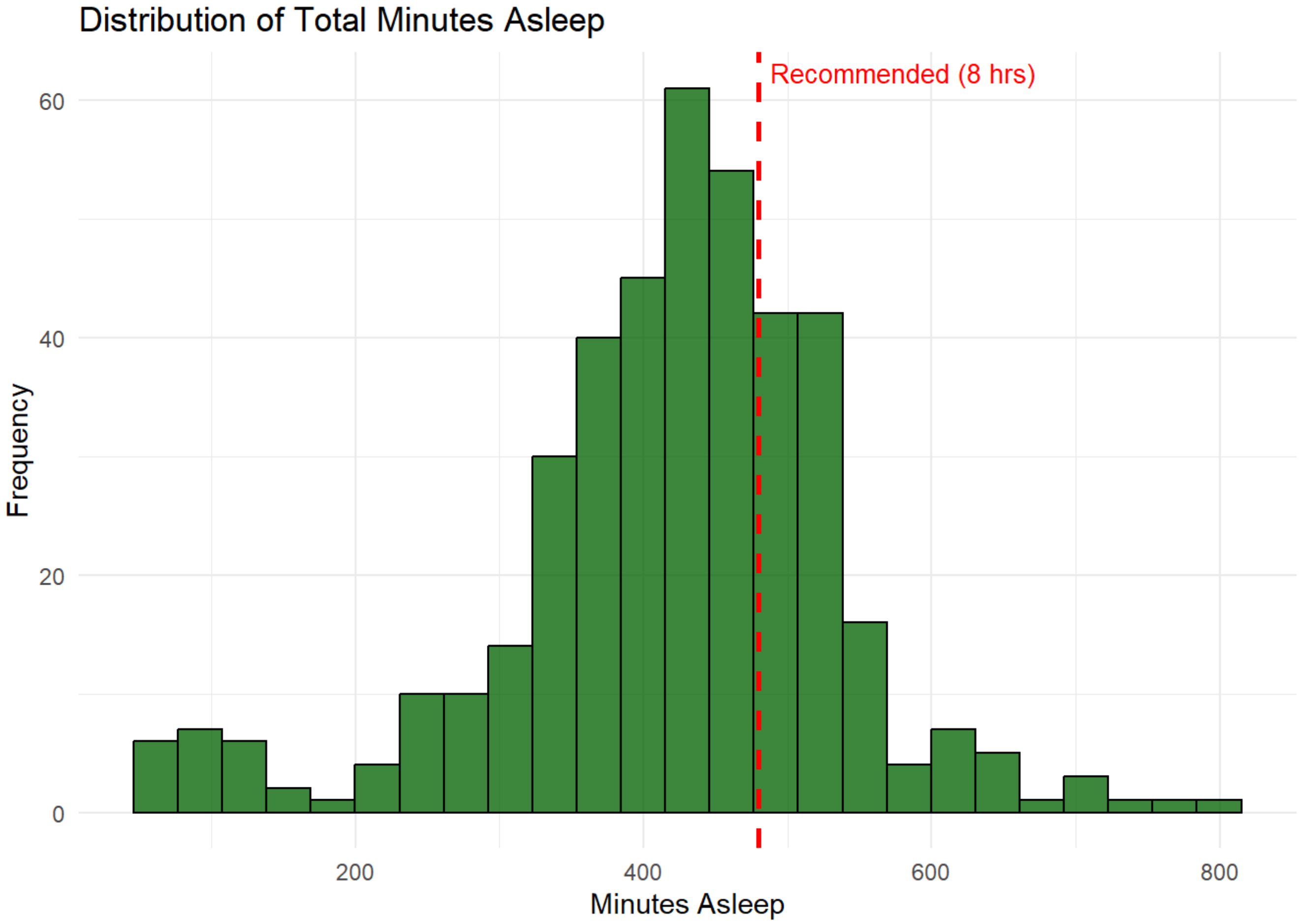
5.3.1 Sleep Duration vs Time in Bed

There is a strong relationship between time in bed and time asleep, but with some inefficiency. Bellabeat could highlight the Leaf's sleep tracking and mindfulness features to help users improve sleep quality, not just duration.



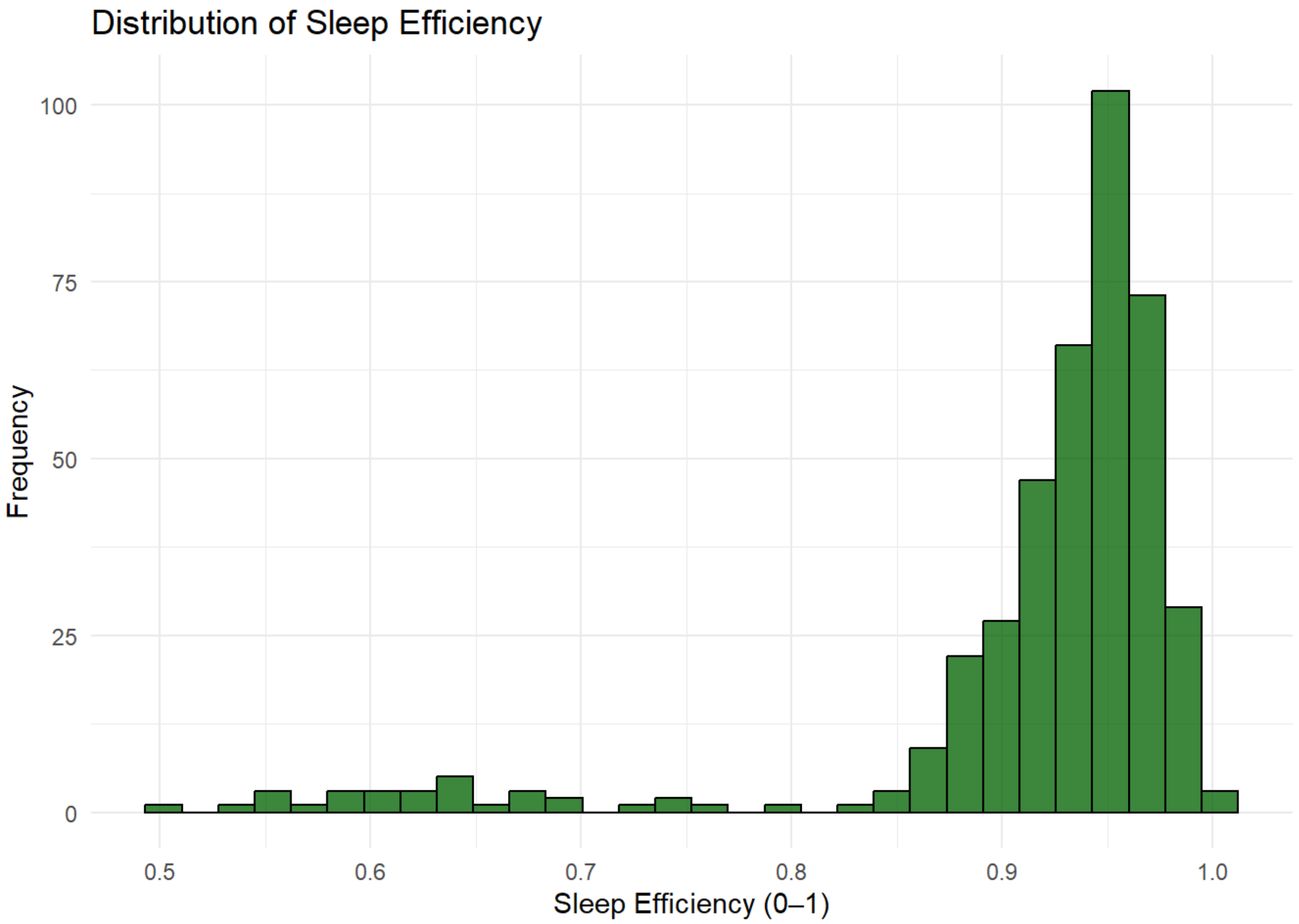
5.3.2 Distribution of Sleep Duration

Most users sleep between 300 and 450 minutes (5–7.5 hours), below the recommended 8 hours. This opens up an opportunity for Bellabeat to promote features that encourage better rest and recovery.



5.3.3 Sleep Efficiency Metric

Sleep efficiency (sleep time ÷ time in bed) varies across users. By focusing on improving this metric, Bellabeat can position itself as not only a tracker of sleep, but as a tool to enhance its effectiveness through actionable feedback.



6 Recommendations

After analyzing smart device usage data across physical activity, heart rate, and sleep behaviors, several behavioral trends have emerged. These patterns present Bellabeat with clear opportunities to enhance user engagement and position its product offerings more effectively in the wellness market.

6.1 Top High-Level Insights

Top High-Level Insights and Strategic Opportunities

Area	Insight	Business Opportunity
Steps	Most users average fewer than 10,000 steps per day	Promote step-tracking features and in-app goal reminders to increase movement
Activity Intensity	Light activity dominates daily routines; vigorous activity is limited	Encourage more vigorous activity through guided workouts or challenge features
Heart Rate	Heart rate tends to peak in the morning and stabilize after noon	Use heart rate trends to schedule stress-reducing nudges like breathing reminders

Area	Insight	Business Opportunity
Sleep Duration	Most users sleep between 5–7.5 hours, below the recommended 8	Emphasize Leaf's ability to support better rest and recovery through mindfulness tools
Sleep Efficiency	Sleep efficiency (asleep ÷ in bed) varies significantly between users	Position Bellabeat as a coach that helps improve sleep quality, not just monitor it
User Segmentation	Users fall into clear tiers: sedentary, moderately active, and highly active	Tailor app messaging and marketing by user activity tier for personalization

6.2 Final Conclusion

Smart device users tend to show moderate activity, short sleep durations, and inefficiencies in sleep quality. These habits highlight a gap between wellness goals and actual behaviors — a space where Bellabeat can deliver value through personalized coaching, reminders, and health insights.

6.3 How Bellabeat Can Apply the Insights

The marketing and product teams can act on these findings by focusing on four strategic areas:

1. Encourage Daily Movement Through Goal-Oriented Nudges
- Most users do not consistently reach the 10,000 steps/day benchmark.
- *Action:* Promote the Leaf's goal-tracking and reminder features (e.g., "You're halfway to 10,000 steps!") in marketing messages.
- Include motivational campaigns or gamified challenges in the app to foster daily habits.

2. Empower Users to Improve Sleep Duration and Quality
- Many users sleep fewer than the recommended 8 hours and show inefficiencies (time in bed > time asleep).
- *Action:* Position Bellabeat's sleep tracking, meditation guidance, and sleep scores as tools to enhance both duration and efficiency.
- Encourage setting bedtime routines via personalized in-app suggestions.

3. Leverage Morning Heart Rate Patterns to Offer Wellness Support
- Heart rate tends to spike in the morning — a common time for stress or high activity.
- *Action:* Use Leaf's physiological tracking to send morning breathing or mindfulness prompts at peak heart rate hours.
- Market Leaf as a proactive stress management assistant, not just a passive tracker.

4. Tailor Messaging by Activity Tier
- Users fall into three tiers (sedentary, moderately active, highly active).
- *Action:* Create segmented campaigns:
- *Sedentary:* Emphasize small, achievable movement wins.
- *Moderate:* Reinforce progress with intermediate-level tips.
- *Highly active:* Promote advanced features like detailed HR analysis or performance monitoring.

6.4 Next Steps for Stakeholders

- Use these insights to **refine campaign messaging**, both in-app and on external platforms.

- Collaborate with product to **introduce customizable goals**, tailored sleep tips, and activity challenges.
- Prioritize **user segments** that show potential for higher engagement (e.g., moderately active users who are closest to becoming “highly active”).
- Consider A/B testing push notifications or reminder campaigns based on user tier.

6.5 Opportunities for Further Analysis

- Explore differences in behavior by **gender, age, or device type** if more granular data becomes available.
- Study **longitudinal changes** (e.g., does user activity improve over time with reminders?).
- Analyze app usage data (if available) to correlate **feature engagement with behavior improvement**.