Bellabeat Analysis

Alessandro Talia

Contents

Introduction

Bellabeat is a wellness technology company specializing in health-focused smart devices designed for women. Co-founded by Urška Sršen and Sando Mur, Bellabeat leverages beautifully designed technology to empower women with knowledge about their health and habits. Since its founding in 2013, the company has grown rapidly, offering innovative products like the Bellabeat app, Leaf tracker, and Time watch.

In this analysis, the focus is on exploring smart device data to uncover consumer habits and trends. The insights generated will guide marketing strategies for one Bellabeat product, providing actionable recommendations to help the company grow in the competitive global smart device market.

Ask Phase

Business Task The goal is to analyze smart device usage data for one of Bellabeat's products to generate insights that guide the company's marketing strategy. This analysis focuses on the Bellabeat app, which provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. These insights empower users to better understand their routines and make healthier decisions.

Stakeholders Urška Sršen: Cofounder and Chief Creative Officer of Bellabeat. Sando Mur: Cofounder and Mathematician. Bellabeat Marketing Analytics Team: Responsible for data analysis and strategy development. Customers: End users of Bellabeat's products, whose behaviors and needs are central to the analysis.

Prepare Phase

Data Organization

The dataset includes data from smart device users, specifically focusing on:

Daily Activities: Records of overall daily activity levels.

Hourly Steps: Data on steps taken at an hourly level.

Daily Sleep: Information about sleep duration and quality.

This data was collected from a survey distributed via Amazon Mechanical Turk between March 12, 2016, and May 12, 2016. Thirty-three eligible Fitbit users consented to share personal tracker data, including minute-level outputs for physical activity, heart rate, and sleep monitoring.

Data Sources and Characteristics

Source: Fitbit Fitness Tracker Data, contributed by participants through Amazon Mechanical Turk.

Characteristics: The dataset contains detailed minute-level data on various health metrics, making it suitable for understanding trends in user activity and sleep behaviors.

Credibility of Data

Strengths: The dataset is derived from real Fitbit users, providing a reliable foundation for analysis. The granularity of the data allows for detailed insights into user habits.

Limitations: The sample size is relatively small (33 users), which might limit the generalizability of the findings. Additionally, since the data was self-reported via consented tracking, there may be biases or inconsistencies related to user behavior or device accuracy.

Process Phase

In the process phase, R was used to clean, manipulate, and transform the data. The following steps were taken to prepare the datasets for analysis:

Installing and Loading Packages

Essential R packages for data analysis and visualization were installed and loaded, including:

- Tidyverse
- Janitor
- Skimr
- Here

```
library("tidyverse")
library("janitor")
library("here")
library("skimr")
```

Importing Datasets

The following datasets were imported into R:

Daily Activity: Containing daily activity metrics like steps and calories.

Hourly Steps: Recording steps taken each hour.

Sleep Day: Detailing sleep duration and quality.

```
daily_activity <- read_csv(here("data", "dailyActivity_merged.csv"))
hourly_steps <- read_csv(here("data", "hourlySteps_merged.csv"))
sleep_day <- read_csv(here("data", "SleepDay_merged.csv"))
```

```
head(daily_activity)
str(daily_activity)

head(hourly_steps)
str(hourly_steps)

head(sleep_day)
str(sleep_day)
```

Data Cleaning

- **Duplicate Removal**: Identified and removed duplicate rows to ensure data accuracy.
- **Column Name Cleaning**: Renamed and standardized column names for consistency.
- **Date/Time Formatting**: Adjusted the date/time format to ensure compatibility across datasets.

```r
n_unique(daily_activity$Id)
```

```
## [1] 33
```

```r
n_unique(hourly_steps$Id)
```

```
## [1] 33
```

```r
n_unique(sleep_day$Id)
```

```
## [1] 24
```

The sleep day dataset contains 24 participants, while the others contain 33.

Check if there are any null or duplicate values within the datasets:

```r
sum(duplicated(daily_activity))
```

```
## [1] 0
```

```r
sum(duplicated(hourly_steps))
```

```
## [1] 0
```

```r
sum(duplicated(sleep_day))
```

```
## [1] 3
```

Remove the duplicate:

```r
daily_activity <- drop_na(daily_activity)
hourly_steps <- drop_na(hourly_steps)
sleep_day <- sleep_day %>%
  distinct() %>%
  drop_na()

sum(duplicated(sleep_day))
```

```
## [1] 0
```

Clean and rename columns:

```r
daily_activity <- clean_names(daily_activity)
hourly_steps <- clean_names(hourly_steps)
sleep_day <- clean_names(sleep_day)

daily_activity <- rename_with(daily_activity, tolower)
hourly_steps <- rename_with(hourly_steps, tolower)
sleep_day <- rename_with(sleep_day, tolower)
```

Adjusting date-time format:

```r
daily_activity <- daily_activity %>%
  rename(date = activity_date) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))

sleep_day <- sleep_day  %>%
  rename(date = sleep_day) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y %I:%M:%S %p"))

hourly_steps <- hourly_steps %>%
  rename(date_time = activity_hour) %>%
  mutate(date_time = as.POSIXct(date_time, format = "%m/%d/%Y %I:%M:%S %p"))

hourly_steps <- hourly_steps %>%
  separate(date_time, into = c("date", "time"), sep = " ") %>%
  mutate(date = ymd(date))

hourly_steps <- hourly_steps %>%
  mutate(time =
           ifelse(is.na(time), "00:00:00",
                  time))
```

Data Transformations

- **Dataset Merging**: Combined the Daily Activity and Sleep Day datasets into a new dataset called Daily Activity Sleep.
- **Feature Creation**: Calculated daily averages for steps, calories burned, and sleep duration. These metrics were incorporated into a new dataset called User Type, classifying users based on their average daily activity levels.

```r
daily_activity_sleep <- merge(daily_activity, sleep_day, by = c("id", "date"))
glimpse(daily_activity_sleep)

daily_average <- daily_activity_sleep %>%
  group_by(id) %>%
  summarise(mean_daily_steps = mean(total_steps),
            mean_daily_calories = mean(calories),
            mean_daily_sleep = mean(total_minutes_asleep))
head(daily_average)

user_type <- daily_average %>%
  mutate(user_type = case_when(
    mean_daily_steps < 5000 ~ "sedentary",
    mean_daily_steps >= 5000 & mean_daily_steps < 7499 ~ "Low active",
    mean_daily_steps >= 7500 & mean_daily_steps < 9999 ~ "Somewhat active",
    mean_daily_steps >= 10000 ~ "Active"
    ))

user_type <- user_type %>%
  mutate(mean_daily_sleep_hr = mean_daily_sleep /60)
```

Since we don't have any demographic variables from our sample we want to determine the type of users

with the data we have. We can classify the users by activity considering the daily amount of steps. We can categorize users as follows:

- **Sedentary**: Less than 5000 steps a day.
- **Low active**: Between 5000 and 7499 steps a day.
- **Somewhat active**: Between 7500 and 9999 steps a day.
- **Active**: More than 10000 steps a day.

Classification has been made per the following article: Classification

Verification

To validate the cleaning and transformation processes:

Used head ( ) to preview the first rows of the datasets.

Employed str ( ) to examine the structure and ensure the integrity of the data.

```
head(daily_activity)
str(daily_activity)

head(sleep_day)
str(sleep_day)

head(hourly_steps)
str(hourly_steps)

head(user_type)
str(user_type)

head(daily_activity_sleep)
str(daily_activity_sleep)
```

Analyze Phase

```
daily_activity %>%
  select(total_steps,
         total_distance,
         sedentary_minutes,
         calories) %>%
  summary()
```

```
##   total_steps    total_distance   sedentary_minutes    calories
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0    Min.   :   0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8    1st Qu.:1828
##  Median : 7406   Median : 5.245   Median :1057.5    Median :2134
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2    Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5    3rd Qu.:2793
##  Max.   :36019   Max.   :28.030   Max.   :1440.0    Max.   :4900
```

```r
sleep_day %>%
  select(total_minutes_asleep,
         total_time_in_bed) %>%
  summary()
```

```
##  total_minutes_asleep total_time_in_bed
##  Min.   : 58.0        Min.   : 61.0
##  1st Qu.:361.0        1st Qu.:403.8
##  Median :432.5        Median :463.0
##  Mean   :419.2        Mean   :458.5
##  3rd Qu.:490.0        3rd Qu.:526.0
##  Max.   :796.0        Max.   :961.0
```

- Averages indicate that participants take 7,638 steps per day.
- Sedentary time is significant, with an average of 991 minutes daily (16.5 hours).
- Active participants fall primarily into the "Somewhat Active", meaning they take between 7,500 and 10,000 steps daily.
- Participants sleep an average of 7 hours per night.

Correlation Analysis

```r
cor(daily_activity$total_steps, daily_activity$calories)
```

```
## [1] 0.5915681
```

```r
cor(daily_activity_sleep$total_steps, daily_activity_sleep$total_minutes_asleep)
```

```
## [1] -0.1903439
```

```r
cor(daily_activity_sleep$total_minutes_asleep, daily_activity_sleep$total_time_in_bed)
```

```
## [1] 0.9304224
```

```r
cor(daily_activity_sleep$sedentary_minutes, daily_activity_sleep$total_minutes_asleep)
```

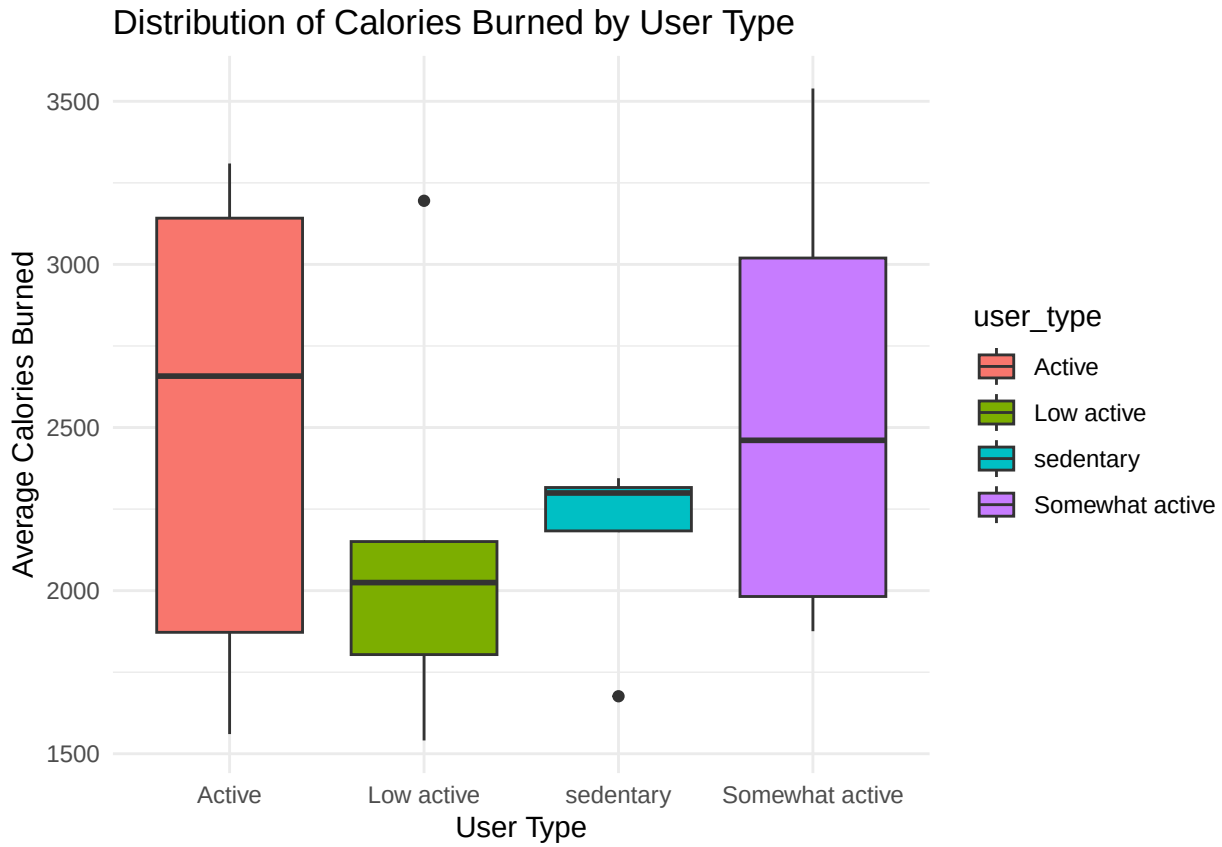
```
## [1] -0.6010731
```

- *Steps vs. Calories*: A moderate positive correlation of 0.59 shows that increased physical activity is strongly associated with higher calorie burn.
- *Sleep vs. Calories*: Participants burn the most calories when sleeping 6–7 hours per night, with a decline after exceeding 7 hours.
- *Sleep vs. Steps*: A weak negative correlation of -0.19 suggests that individuals who walk more tend to sleep slightly less. Minutes in Bed vs. Minutes Asleep: A very strong positive correlation of 0.93 confirms that the majority of time spent in bed is dedicated to sleep.
- *Sedentary Time vs. Sleep Duration*: A moderately strong negative correlation of -0.60 indicates that higher sedentary time during the day corresponds to shorter sleep duration

Share Phase

Distribution of Calories Burned by User Type

```
ggplot(user_type, aes(x = user_type, y = mean_daily_calories, fill = user_type)) +
  geom_boxplot() +
  labs(title = "Distribution of Calories Burned by User Type",
       x = "User Type",
       y = "Average Calories Burned") +
  theme_minimal()
```



More active users tend to burn more calories on average.

"*Active*" and "*Somewhat Active*" groups show a wider distribution, indicating greater variability in calorie expenditure.

Sedentary users have a more consistent calorie burn, with a median around 2100 calories and a narrow distribution.

Low Active users show slightly more variability, with some individuals burning calories at levels closer to Active users.

"*Low Active*" and "*Sedentary*" groups have fewer outliers, indicating a more predictable calorie burn.

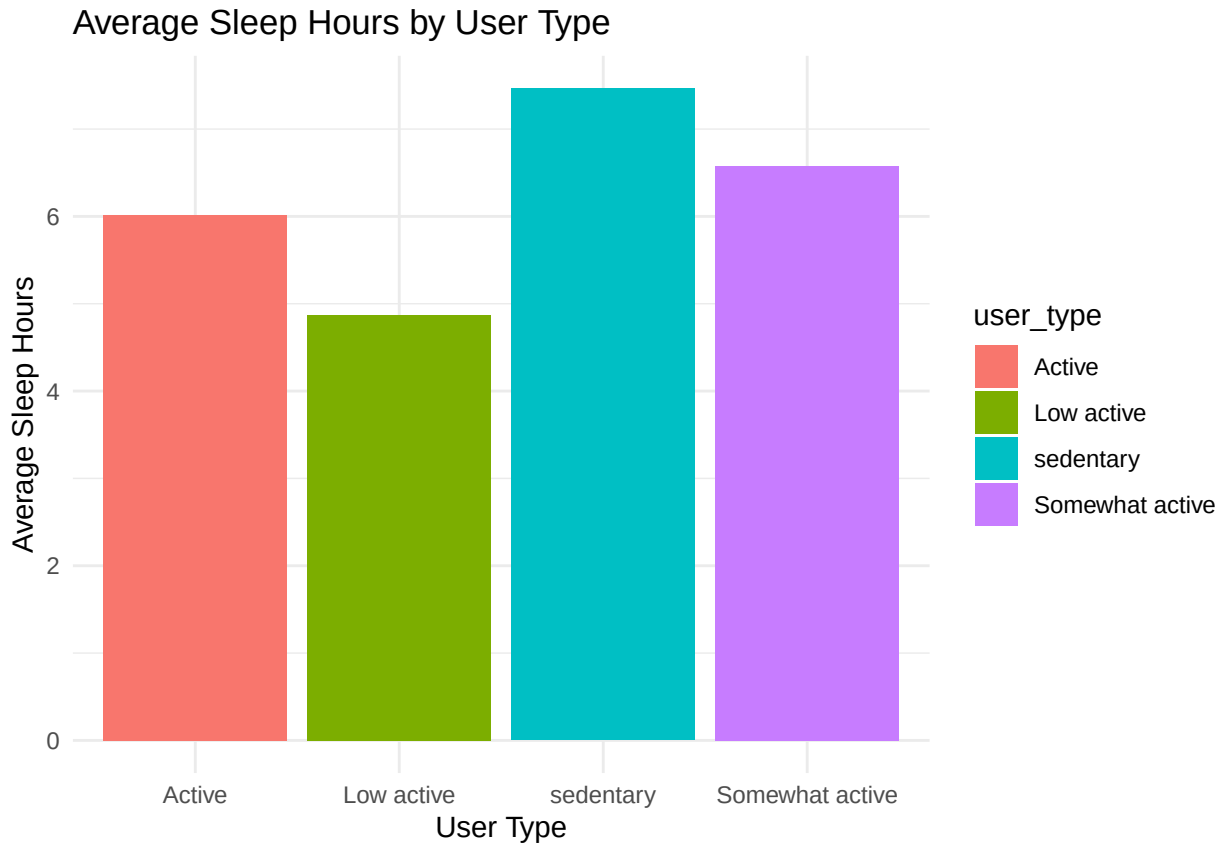
Higher activity levels are correlated with higher and more variable calorie expenditure.

Sedentary users tend to have a more stable and lower calorie burn.

This analysis suggests that activity level significantly influences calorie burn, with more active individuals experiencing greater variability and higher expenditure.

Average Sleep Hours by User Type

```
ggplot(user_type, aes(x = user_type, y = mean_daily_sleep_hr, fill = user_type)) +
  stat_summary(fun = mean, geom = "bar") +
  labs(title = "Average Sleep Hours by User Type",
       x = "User Type",
       y = "Average Sleep Hours") +
  theme_minimal()
```



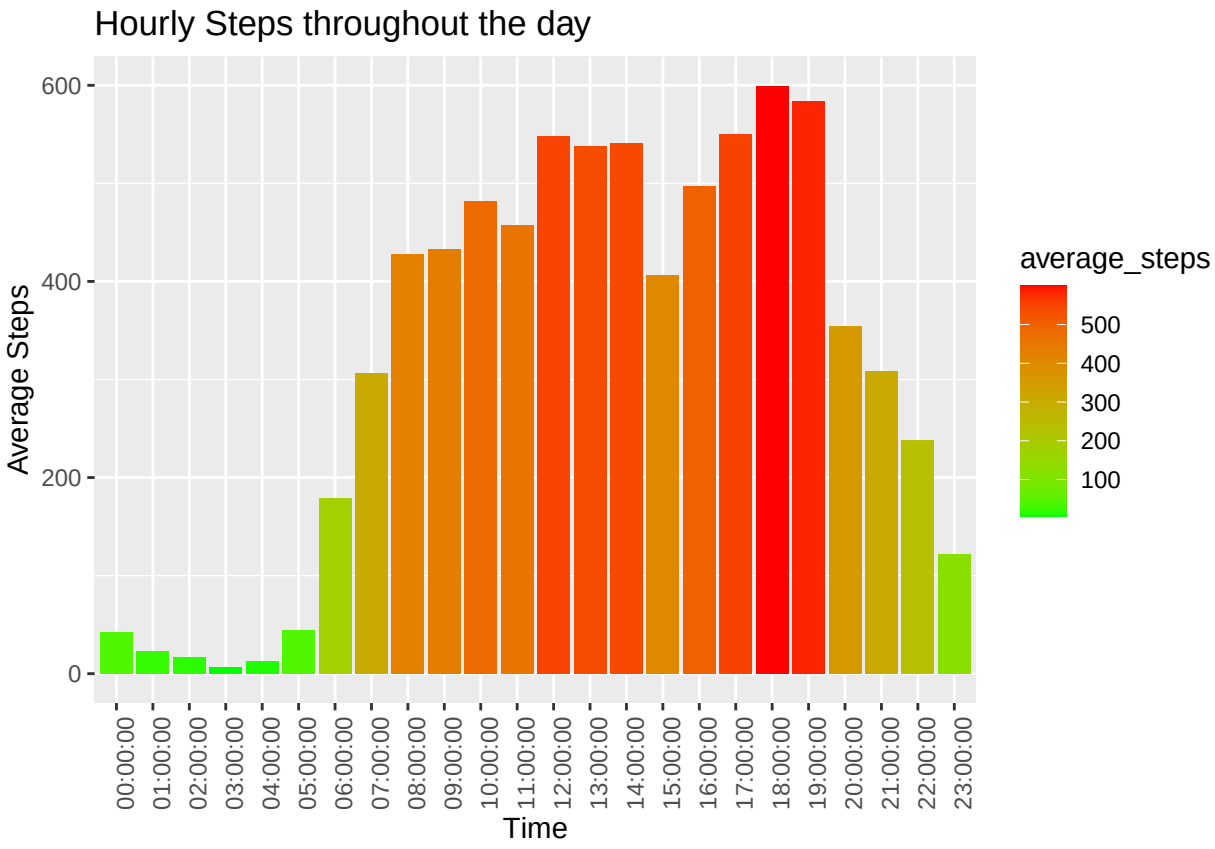
The bar chart illustrates the average sleep hours across different user activity levels.

- **Sedentary users** have the highest average sleep duration, suggesting they might have more time for rest or less physically demanding routines.
- **Somewhat active users** also get a relatively high amount of sleep, but slightly less than sedentary users.
- **Active users** sleep more than low active users but less than sedentary ones, possibly due to their busy schedules or higher energy expenditure.
- **Low active users** sleep the least on average, which may indicate a lifestyle pattern where minimal activity is linked to shorter sleep duration.

This visualization suggests that higher physical activity does not necessarily correlate with more sleep.

Hourly Steps throughout the Day

```
hourly_steps %>%
  group_by(time) %>%
  summarize(average_steps = mean(step_total)) %>%
  ggplot() +
  geom_col(mapping = aes(x = time, y = average_steps, fill = average_steps))+
  labs(title = "Hourly Steps throughout the day",
       x = "Time",
       y = "Average Steps") +
  scale_fill_gradient(low = "green", high = "red") +
  theme(axis.text.x = element_text (angle = 90))
```



The bar chart displays the average number of steps taken per hour, revealing key activity trends throughout the day.

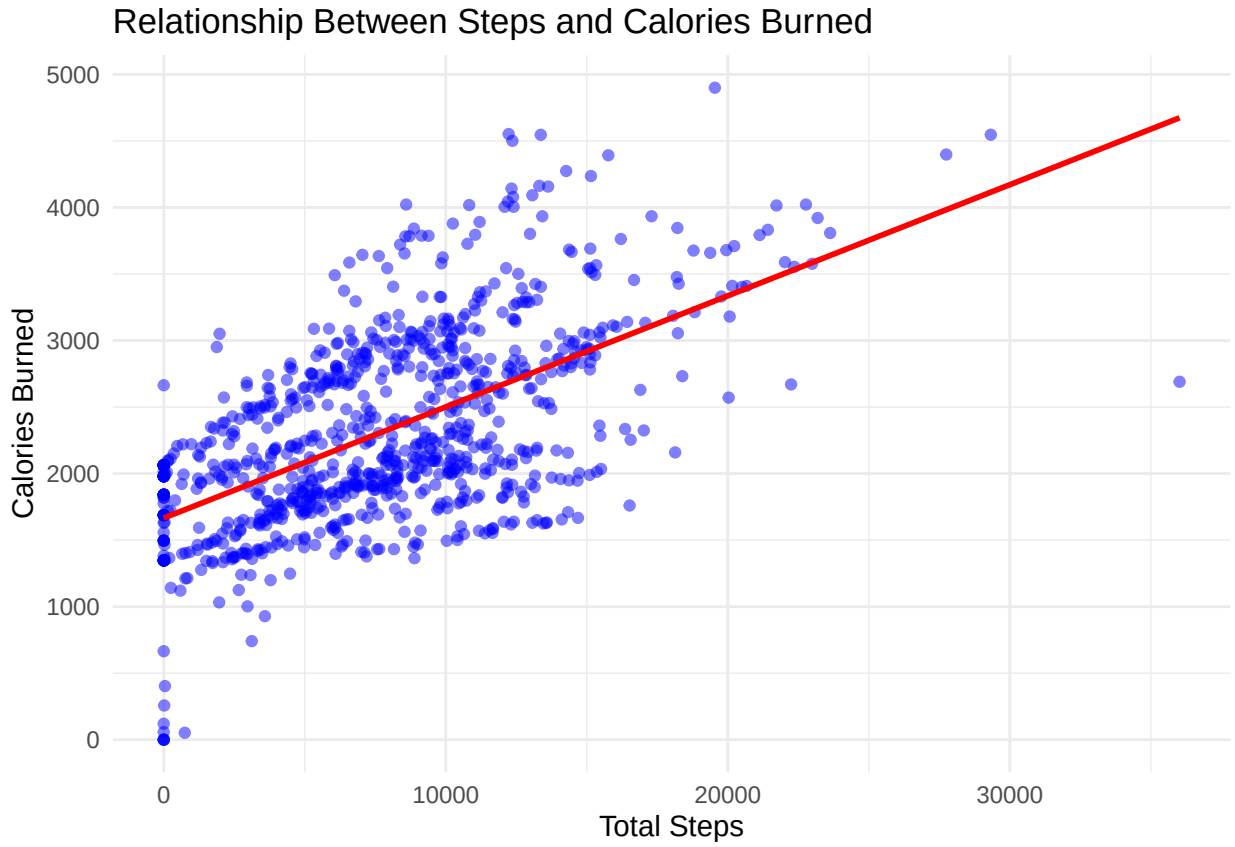
- *Minimal activity during night time* (00:00 - 05:00), indicating expected sleep hours with very few steps recorded.
- *A gradual increase in steps starts around 06:00*, likely as users begin their morning routines.
- *Peak activity occurs between 10:00 and 18:00*, with the highest step count observed around 17:00 - 18:00. This suggests that most users engage in walking or physical activities during midday and early evening.
- *After 19:00, step count declines*, reflecting a wind-down period as users likely transition to less active evening routines.

These insights suggest that users are most active in the late morning and afternoon, with a secondary peak in the early evening.

Steps vs Calories Burned

```
ggplot(daily_activity, aes(x = total_steps, y = calories)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Relationship Between Steps and Calories Burned",
       x = "Total Steps",
       y = "Calories Burned") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatter plot illustrates the correlation between total steps taken and calories burned, providing key observations:

- **Strong Positive Correlation**: The upward trend suggests that as the number of steps increases, the calories burned also increase. This aligns with expectations, as walking or running more results in higher energy expenditure.
- **Variability in Calories Burned**: While the general trend is clear, there is some dispersion in calorie burn for the same step counts. This variation could be due to factors such as walking speed, intensity, individual metabolism, or additional activities influencing calorie expenditure.

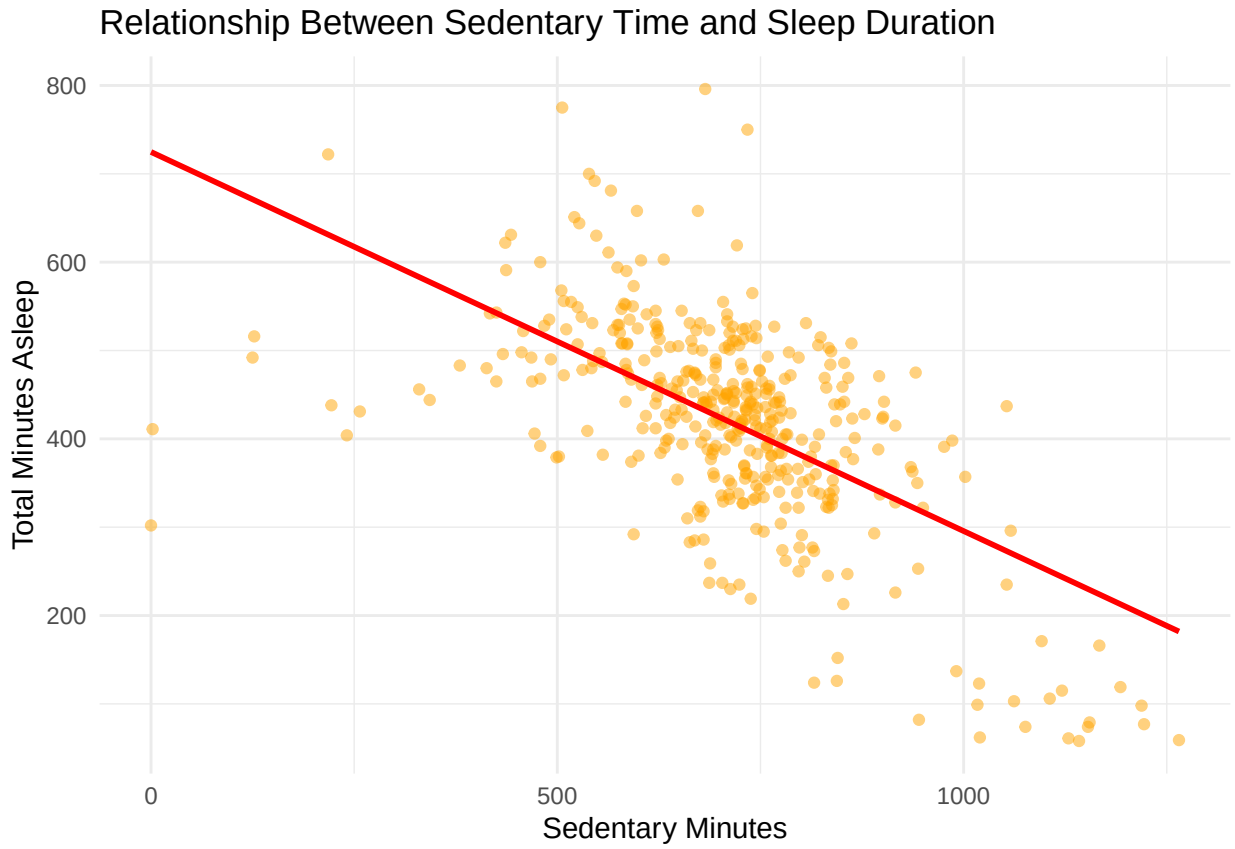
- *Baseline Calories Burned*: Even with very few steps, individuals still burn calories, likely due to resting metabolic rate (RMR), which accounts for energy used at rest.
- *Outliers*: Some points deviate significantly from the trend, possibly representing days with intense workouts or sedentary behavior despite high step counts.

These insights confirm that daily step count is a strong predictor of calorie expenditure, through individual differences play a role.

Sedentary Minutes vs. Total Minutes Asleep

```
ggplot(daily_activity_sleep, aes(x = sedentary_minutes, y = total_minutes_asleep)) +
  geom_point(alpha = 0.5, color = "orange") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Relationship Between Sedentary Time and Sleep Duration",
       x = "Sedentary Minutes",
       y = "Total Minutes Asleep") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatter plot shows how sedentary minutes relate to total minutes asleep, revealing key observations:

- **Negative Correlation**: The downward trend suggests that more sedentary time is associated with less sleep duration. This could indicate that a sedentary lifestyle negatively affects sleep quality or duration.
- **High Variability**: Although the trend is negative, there is considerable dispersion. Some individuals with high sedentary time still get a good amount of sleep, while others with low sedentary time have poor sleep.

Possible Explanations:

- Sedentary behavior may reduce physical tiredness, making it harder to fall asleep.
- High screen time (common in sedentary lifestyles) may disrupt sleep due to blue light exposure.
- Other lifestyle factors, like stress or irregular schedules, could contribute to both high sedentary time and poor sleep.

This data suggests that reducing sedentary behavior could contribute to better sleep patterns.

Act Phase

Based on the analysis of user behavior and the identified trends, several actionable recommendations can be implemented to enhance the Bellabeat app. These suggestions aim to improve user engagement, promote healthier habits, and align the app's features with user needs.

Personalized Notifications Based on User Type

Objective: Motivate users to improve their habits according to their activity level.

Examples of personalized notifications:

Sedentary users: "Have you taken a break today? A 5-minute walk could boost your energy!"

Low active users: "Just 1,000 more steps and you'll reach the 'Somewhat Active' category! You're almost there!"

Somewhat active users: "Great job! A little extra effort and you could enter the 'Active' category."

Active users: "You're a real athlete! Keep up the pace and try to beat your weekly record."

Gamification and Badge System

Objective: Encourage users to stay active by rewarding them with achievements and incentives.

Examples of badges:

"Early Riser": Earned by reaching 2,000 steps before 9:00 AM.

"Evening Mover": Earned by maintaining activity after 7:00 PM.

"Step Master": Earned by surpassing 10,000 steps for seven consecutive days.

"Sleep Hero": Earned by maintaining an average of seven hours of sleep for a week.

"Sedentary Breaker": Earned by reducing sedentary time by at least 30 minutes per day.

Implementation:

Create a progress dashboard displaying unlocked badges and those yet to be achieved.

Integrate weekly challenges, such as "Reach 50,000 steps this week" with virtual rewards.

Improving Sleep Quality: Notifications and Blue Light Filter

Objective: Help users sleep better by reducing sedentary time and blue light exposure.

Solutions:

Pre-sleep notifications: "Get ready for a restful sleep! Avoid screens and try a mindfulness session."

Detection of negative habits: If a user has high sedentary time and poor sleep, the app suggests personalized solutions.

Auto-activating blue light filter: If the user is using their phone before bedtime, Bellabeat can recommend enabling the night mode.

Conclusion: How This Strategy Improves Bellabeat

- Increases user engagement through personalized experiences.
- Encourages healthier behavior with reward mechanisms.
- Enhances sleep quality with intelligent notifications and blue light reduction.

To optimize this strategy, an A/B test can be implemented to determine which features drive the most user engagement and health improvements.