Bellabeat Project

Ayodele

2025-09-04

1. Business Task (Ask) Bellabeat, a high-tech wellness technology company, aims to expand its presence in the global smart device market. The goal of this analysis is to identify trends in smart device usage (based on Fitbit data) and translate them into actionable insights that can inform Bellabeat's marketing strategy.

   Key Questions:

   a. What are the trends in smart device usage?
   b. How could these trends apply to Bellabeat customers?
   c. How could these insights influence Bellabeat's marketing strategy?

2. Data Sources (Prepare) Fitbit Fitness Tracker Dataset (CC0: Public Domain, Kaggle) Includes daily activity, sleep, steps, weight, and hourly records. Covers 30 users over ~30 days. Limitations: Small sample size (not fully representative). No demographic details (gender, age, location). Short timeframe (~1 month).

3. Data Cleaning Process (Process) Imported CSVs into R. Standardised date formats (mdy, mdy_hms). Cleaned column names (janitor::clean_names). Removed duplicates, handled missing values. Joined datasets (daily activity + sleep).

```
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────────────
tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ──────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(lubridate)
library(janitor)

##
## Attaching package: 'janitor'
```

```
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

daily_activity <- read_csv("dailyActivity_merged.csv") %>% clean_names()

## Rows: 940 Columns: 15
## ── Column specification ──────────────────────────────────────────
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance,
LoggedActivitiesDi...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.

sleep_day <- read_csv("sleepDay_merged.csv") %>% clean_names()

## Rows: 413 Columns: 5
## ── Column specification ──────────────────────────────────────────
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

Convert dates

```
daily_activity$activity_date <- mdy(daily_activity$activity_date)
sleep_day$sleep_day <- mdy_hms(sleep_day$sleep_day)
```

Merge datasets

```
activity_sleep <- merge(daily_activity, sleep_day,
                        by.x = "activity_date",
                        by.y = "sleep_day")
```

Analysis & Visuals (Analyse + Share) A. Daily Activity Levels

```
ggplot(daily_activity, aes(x = total_steps)) +
  geom_histogram(binwidth = 1000, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Daily Steps", x = "Steps", y = "Count")
```

Distribution of Daily Steps
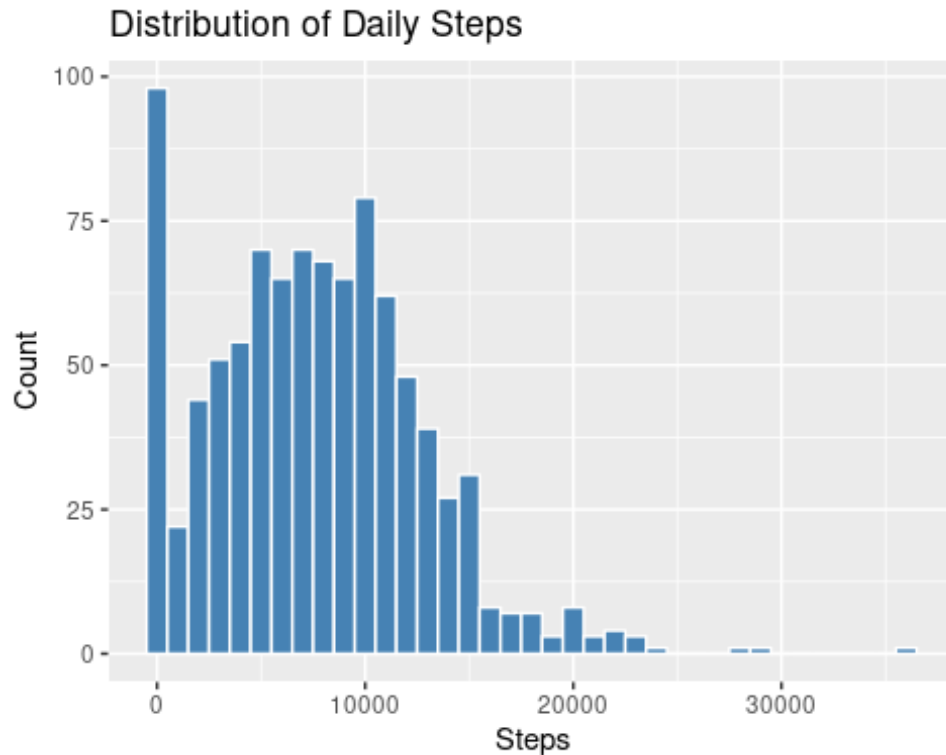
Insight: Average steps ~7,500/day, below the recommended 10,000. Sedentary time dominates daily routines.

B. Weekday vs Weekend Activity

```r
daily_activity %>%
  mutate(day_of_week = weekdays(activity_date)) %>%
  group_by(day_of_week) %>%
  summarise(avg_steps = mean(total_steps, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(day_of_week, avg_steps), y = avg_steps, fill =
day_of_week)) +
  geom_col() +
  labs(title = "Average Steps by Day of Week", x = "Day", y = "Average
Steps") +
  theme(legend.position = "none")
```
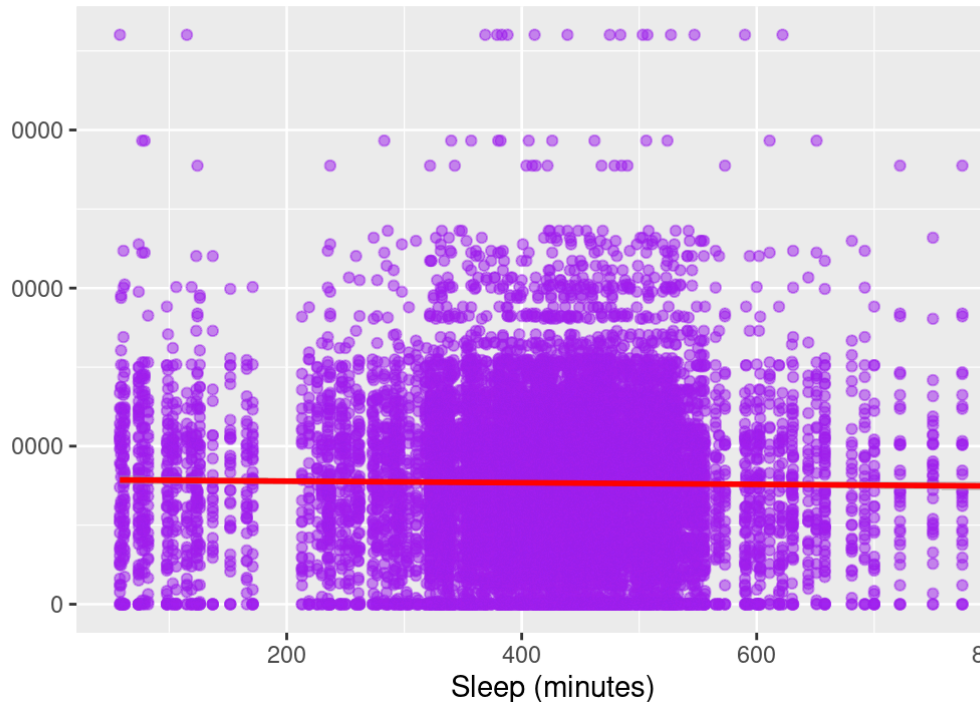
Average Steps by Day of Week
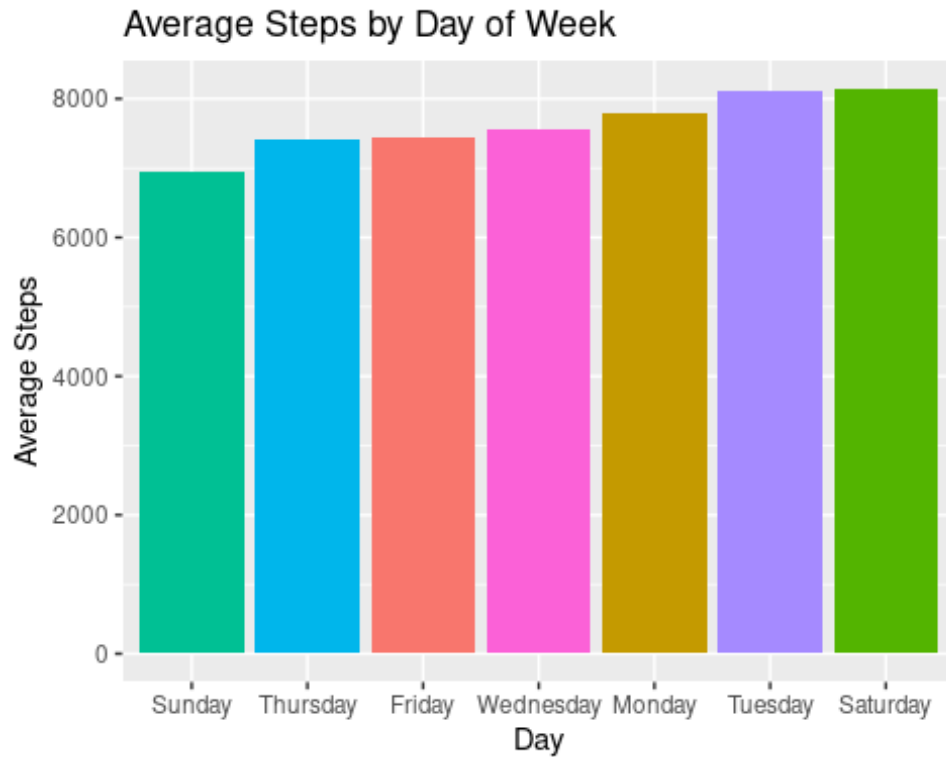
Insight: Users are more active on weekdays than at weekends.

C. Sleep Patterns

```r
ggplot(activity_sleep, aes(x = total_minutes_asleep, y = total_steps)) +
  geom_point(alpha = 0.5, color = "purple") +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Relationship Between Sleep & Activity",
       x = "Sleep (minutes)", y = "Steps")
```

Relationship Between Sleep & Activity

Insight: Weak correlation between sleep and activity. Average sleep is ~7 hours, slightly below recommended.

D. Hourly Activity Trends

```
hourly_steps <- read_csv("hourlySteps_merged.csv") %>% clean_names()

## Rows: 22099 Columns: 3
## ── Column specification ─────────────────────────────────────────────
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, StepTotal
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

hourly_steps$activity_hour <- mdy_hms(hourly_steps$activity_hour)

hourly_steps %>%
  mutate(hour = hour(activity_hour)) %>%
  group_by(hour) %>%
  summarise(avg_steps = mean(step_total, na.rm = TRUE)) %>%
  ggplot(aes(x = hour, y = avg_steps)) +
  geom_line(color = "darkgreen", size = 1.2) +
  labs(title = "Average Steps by Hour of Day", x = "Hour", y = "Steps")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

Average Steps by Hour of Day
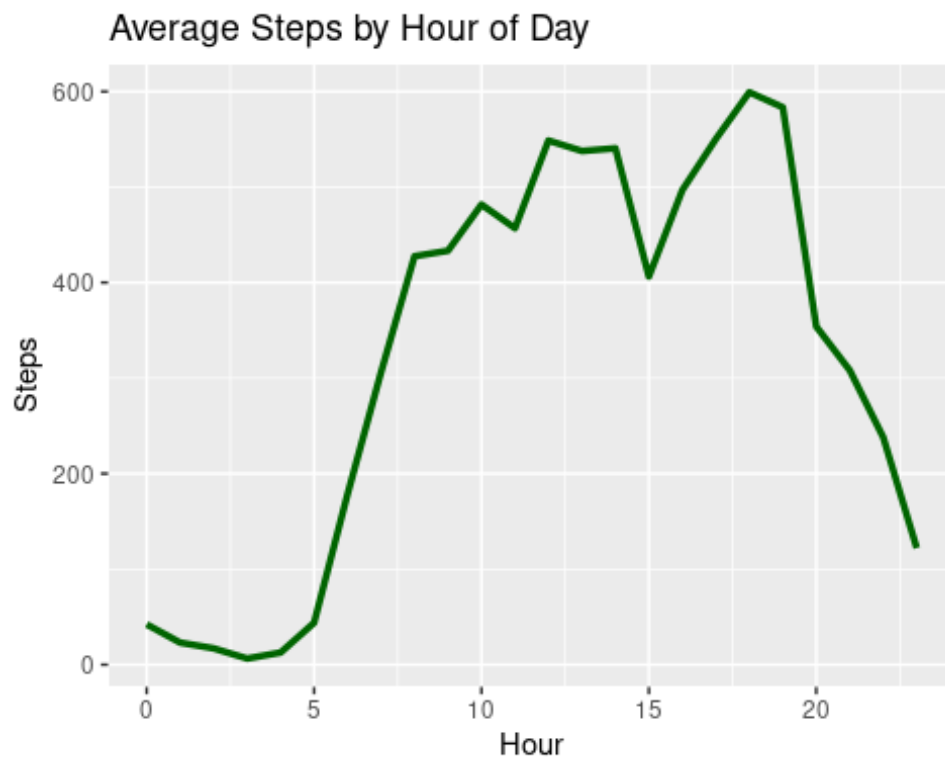
Insight: Peak activity in mornings (7–9 AM) and evenings (5–7 PM).

E.Sedentary vs Active Time Sedentary minutes (~960/day, or 16 hours) dominated users' days. Very active minutes averaged less than an hour per day. This imbalance shows the opportunity for smart devices to encourage short, regular activity breaks.

```
daily_activity %>%
  summarise(
    avg_sedentary = mean(sedentary_minutes, na.rm = TRUE) / 60,
    avg_active = mean(very_active_minutes, na.rm = TRUE) / 60
  )

## # A tibble: 1 × 2
##   avg_sedentary avg_active
##           <dbl>      <dbl>
## 1          16.5      0.353

daily_activity %>%
  summarise(
    sedentary = mean(sedentary_minutes, na.rm = TRUE),
    active = mean(very_active_minutes, na.rm = TRUE)
```
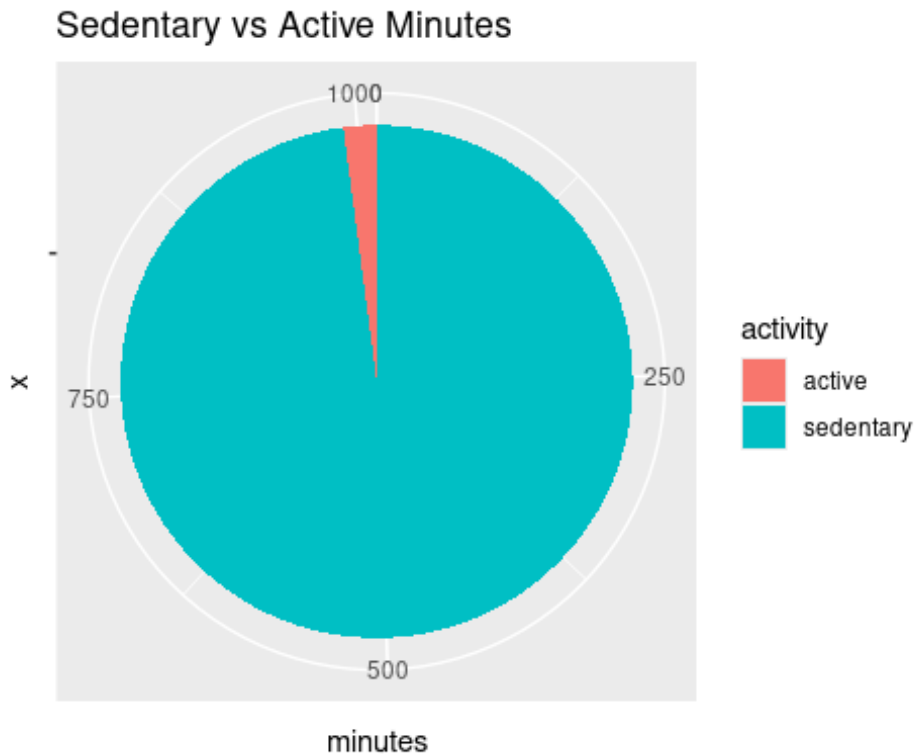
```
  ) %>%
  pivot_longer(cols = everything(), names_to = "activity", values_to =
"minutes") %>%
  ggplot(aes(x = "", y = minutes, fill = activity)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Sedentary vs Active Minutes")
```

Sedentary vs Active Minutes

5. Key Insights (Summary) Users average ~7,500 steps/day → below 10,000 target. Sedentary time dominates (~16 hrs/day). Weekday activity > weekend activity. Sleep ~7 hrs/night → slightly lower than recommended. Weak correlation between sleep & steps. Peak activity mornings/evenings; midday sedentary.

6. Recommendations (Act) Promote holistic wellness → market Bellabeat as more than a step tracker (hydration, sleep, stress). Weekend wellness campaigns → encourage activity during low-activity days. Push midday activity reminders → notifications during sedentary hours. Highlight unique features → hydration (Spring), sleep insights (Leaf/Time), mindfulness (app). Personalised marketing: tailored goals, reminders, and feedback for women's lifestyle needs.