Contents

1. Purpose **1**

2. Load Cleaned Dataset **2**

3. Descriptive Overview **2**

4. Daily Activity Trends **3**

5. Weekday vs Weekend Activity **4**

6. Relationship Between Sleep Duration and Activity **4**

7. Relationship Between Sleep Efficiency and Activity **6**

8. Comparing Sleep Duration vs Sleep Efficiency **7**

9. Correlation Matrix **7**

10. Training Intensity & Sleep **8**
10.1 Create an Intensity Score first 8
10.2. Correlating Training Intensity & Sleep Duration 8
10.3. Correlating Training Intensity & Sleep Efficiency 9

11. Activity Type Analysis (Very / Fairly / Lightly Active) **10**
11.1. Mean Time in Each Intensity 10
11.2. Reshaping the data 11
11.3. Correlating Activity Type - Sleep Duration 11
11.4. Correlating Activity Type - Sleep Efficiency 12
11.5. Correlation Summary 13

12. Statistical Significance — Interpreting p-values **15**

13. Key Findings **15**

14. Analyst's Reflection **17**

---

title: "Bellabeat Case Study — Analyze Phase" author: "Francesco De Felice" date: "2025-11-15" output: pdf_document: toc: true toc_depth: '2' latex_engine: xelatex html_document: toc: true toc_depth: 2 theme: flatly —

1. Purpose

This phase explores **behavioral patterns** in activity and sleep using the cleaned dataset. The goal is to derive insights that can guide **Bellabeat App feature development, UX design, and marketing positioning.**

We'll focus on:

- Daily and weekly activity trends
- Sleep behavior and consistency
- Relationships between activity and sleep quality

---

2. Load Cleaned Dataset

```
clean_data <- read_csv("/cloud/project/Bellabeat Case Study/Data/Raw/clean_data.csv")
glimpse(clean_data)
```

```
## Rows: 410
## Columns: 21
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ Date                     <date> 2016-04-12, 2016-04-13, 2016-04-15, 2016-04-~
## $ TotalSteps               <dbl> 13162, 10735, 9762, 12669, 9705, 15506, 10544~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.14, 2.71, 3.19, 3.53, 1.96, 1.3~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 1.26, 0.41, 0.78, 1.32, 0.48, 0.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 2.83, 5.04, 2.51, 5.03, 4.24, 4.6~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveMinutes        <dbl> 25, 21, 29, 36, 38, 50, 28, 19, 41, 39, 73, 3~
## $ FairlyActiveMinutes      <dbl> 13, 19, 34, 10, 20, 31, 12, 8, 21, 5, 14, 23,~
## $ LightlyActiveMinutes     <dbl> 328, 217, 209, 221, 164, 264, 205, 211, 262, ~
## $ SedentaryMinutes         <dbl> 728, 776, 726, 773, 539, 775, 818, 838, 732, ~
## $ Calories                 <dbl> 1985, 1797, 1745, 1863, 1728, 2035, 1786, 177~
## $ TotalSleepRecords        <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep       <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, ~
## $ TotalTimeInBed           <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, ~
## $ TotalActiveMinutes       <dbl> 366, 257, 272, 267, 222, 345, 245, 238, 324, ~
## $ SleepEfficiency          <dbl> 0.9450867, 0.9434889, 0.9321267, 0.9264305, 0~
## $ Weekday                  <chr> "Tuesday", "Wednesday", "Friday", "Saturday",~
```

---

3. Descriptive Overview

```
summary(select(clean_data, TotalSteps, Calories, TotalActiveMinutes, TotalMinutesAsleep, SleepEfficiency
```

```
##    TotalSteps       Calories    TotalActiveMinutes TotalMinutesAsleep
##  Min.   :   17   Min.   : 257   Min.   :  2.0      Min.   : 58.0
##  1st Qu.: 5189   1st Qu.:1841   1st Qu.:206.5      1st Qu.:361.0
##  Median : 8913   Median :2207   Median :263.5      Median :432.5
##  Mean   : 8515   Mean   :2389   Mean   :259.5      Mean   :419.2
##  3rd Qu.:11370   3rd Qu.:2920   3rd Qu.:315.5      3rd Qu.:490.0
##  Max.   :22770   Max.   :4900   Max.   :540.0      Max.   :796.0
##  SleepEfficiency
##  Min.   :0.4984
##  1st Qu.:0.9118
##  Median :0.9426
##  Mean   :0.9165
##  3rd Qu.:0.9606
##  Max.   :1.0000
```

```
n_users <- n_distinct(clean_data$Id)
date_range <- range(clean_data$Date)
cat("Users:", n_users, "\n")
```

```
## Users: 24
```

```
cat("Date range:", date_range, "\n")
```

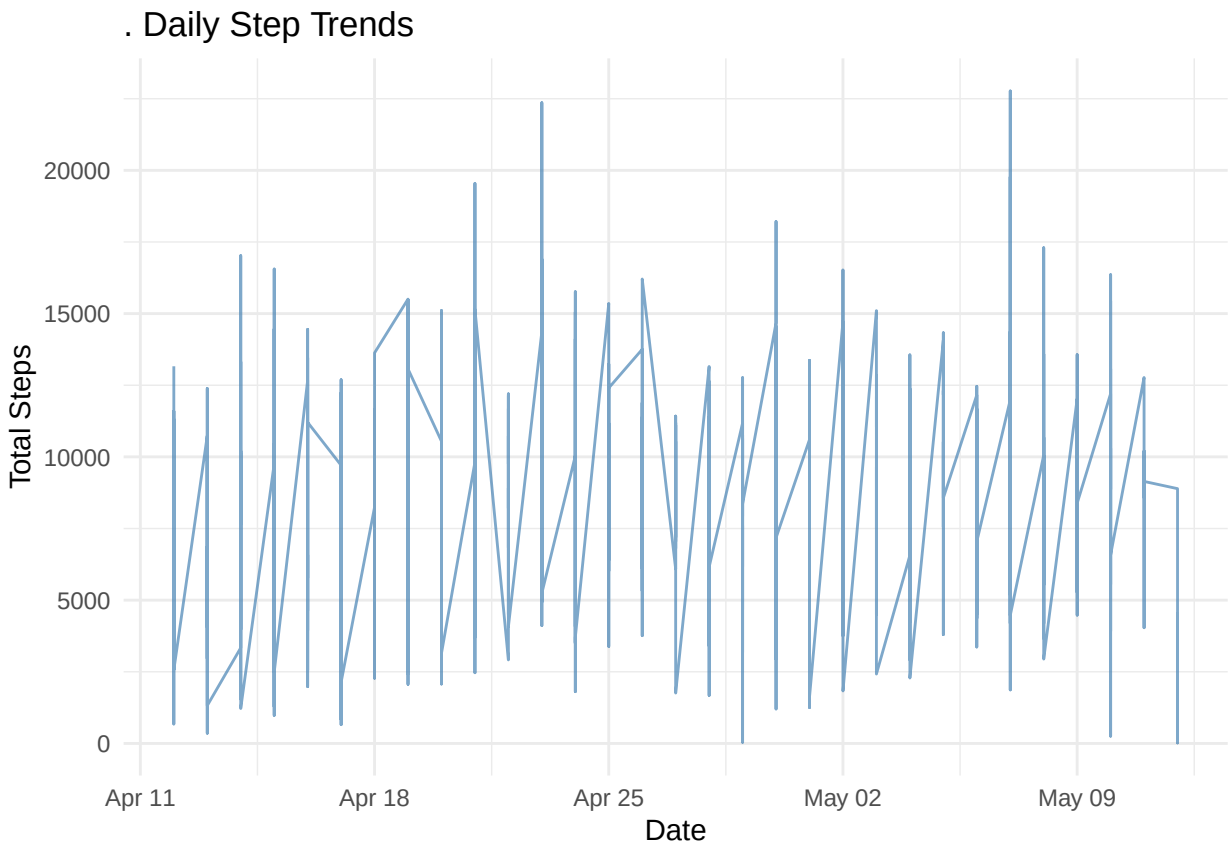
```
## Date range: 16903 16933
```

Dataset includes 24 *users covering the period* 2016-04-12 *to* 2016-05-12.

---

4. Daily Activity Trends

```
ggplot(clean_data, aes(x = Date, y = TotalSteps)) +
  geom_line(color = "steelblue", alpha = 0.7) +
  labs(
    title = "  Daily Step Trends",
    x = "Date",
    y = "Total Steps"
  ) +
  theme_minimal()
```



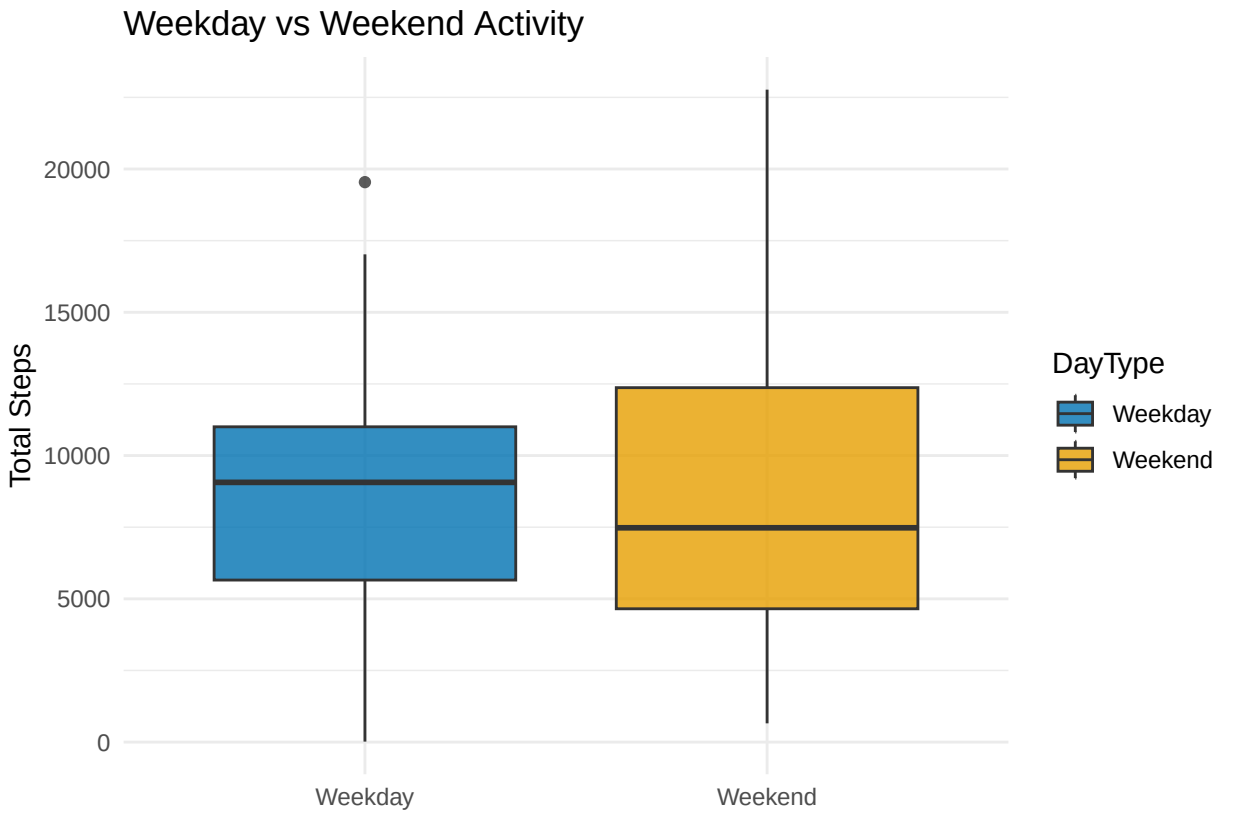
Interpretation: Daily step trends show engagement patterns — dips usually occur on weekends.

---

5. Weekday vs Weekend Activity

```
clean_data <- clean_data %>%
  mutate(
    DayType = ifelse(Weekday %in% c("Saturday", "Sunday"), "Weekend", "Weekday")
  )

ggplot(clean_data, aes(x = DayType, y = TotalSteps, fill = DayType)) +
  geom_boxplot(alpha = 0.8) +
  labs(
    title = "Weekday vs Weekend Activity",
    x = "",
    y = "Total Steps"
  ) +
  scale_fill_manual(values = c("#0072B2", "#E69F00")) +
  theme_minimal()
```



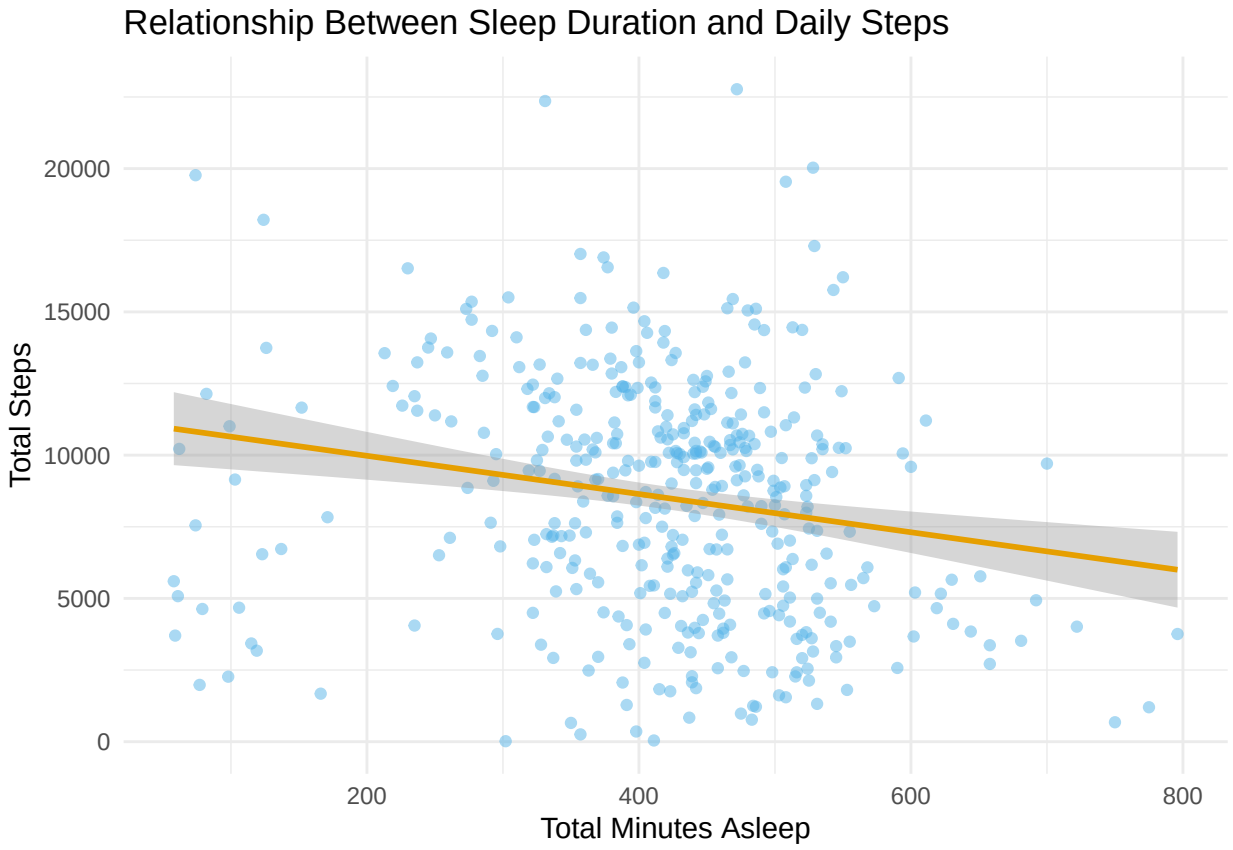
Interpretation: Users tend to walk less on weekends — confirming Hypothesis 1. Bellabeat can promote **mindful weekend challenges** or light-activity prompts.

---

6. Relationship Between Sleep Duration and Activity

```
ggplot(clean_data, aes(x = TotalMinutesAsleep, y = TotalSteps)) +
  geom_point(alpha = 0.5, color = "#56B4E9") +
  geom_smooth(method = "lm", se = TRUE, color = "#E69F00") +
```

```
  labs(
    title = "Relationship Between Sleep Duration and Daily Steps",
    x = "Total Minutes Asleep",
    y = "Total Steps"
  ) +
  theme_minimal()
```




```
# Correlation test
sleep_duration_cor <- cor.test(clean_data$TotalMinutesAsleep, clean_data$TotalSteps)
sleep_duration_cor
```

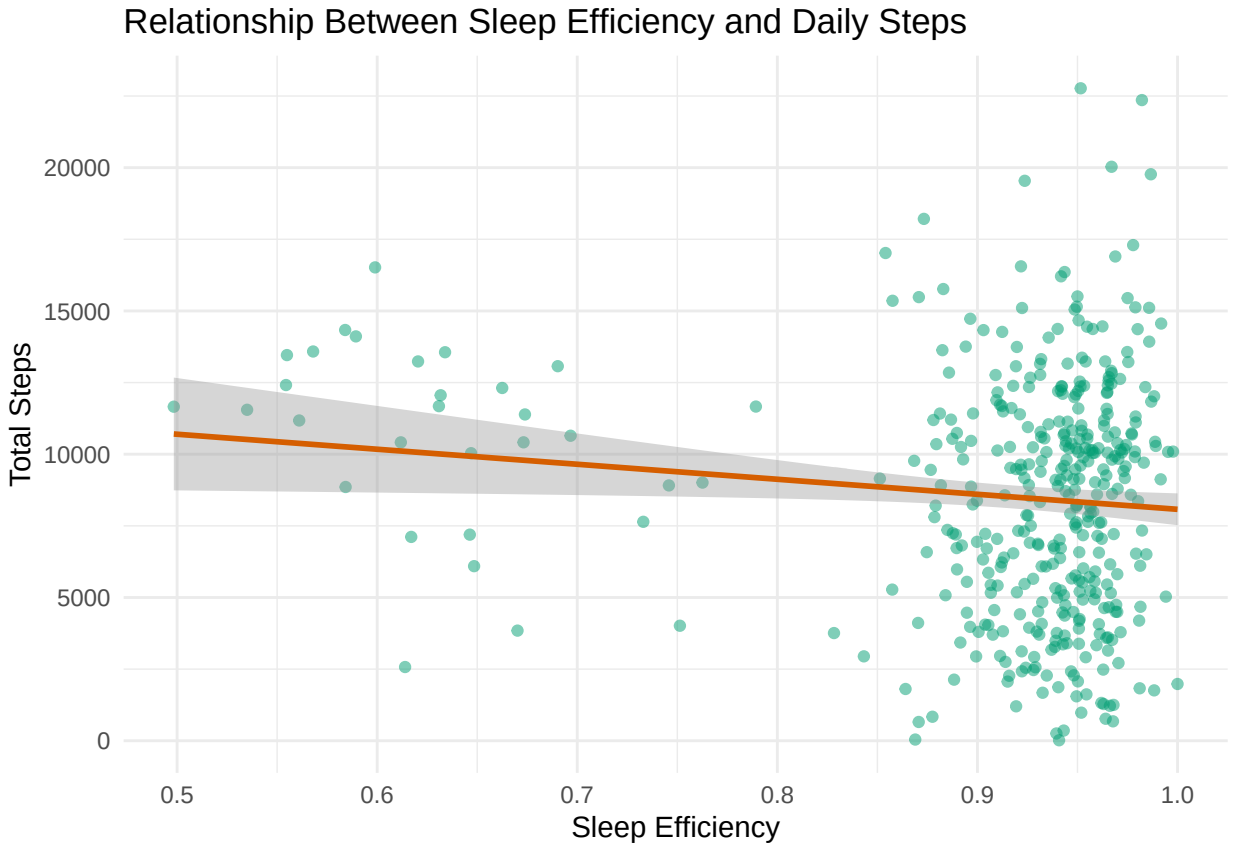
```
##
##  Pearson's product-moment correlation
##
## data:  clean_data$TotalMinutesAsleep and clean_data$TotalSteps
## t = -3.9164, df = 408, p-value = 0.0001054
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.28199288 -0.09525253
## sample estimates:
##        cor
## -0.1903439
```

Interpretation: If correlation is negative and significant (`p < 0.05`), longer sleep correlates with fewer steps — possibly a sign of recovery days. If positive, more rested users tend to move more — a wellness feedback loop.

---

7. Relationship Between Sleep Efficiency and Activity

```
ggplot(clean_data, aes(x = SleepEfficiency, y = TotalSteps)) +
  geom_point(alpha = 0.5, color = "#009E73") +
  geom_smooth(method = "lm", se = TRUE, color = "#D55E00") +
  labs(
    title = "Relationship Between Sleep Efficiency and Daily Steps",
    x = "Sleep Efficiency",
    y = "Total Steps"
  ) +
  theme_minimal()
```




```
# Correlation test
sleep_eff_cor <- cor.test(clean_data$SleepEfficiency, clean_data$TotalSteps)
sleep_eff_cor
```

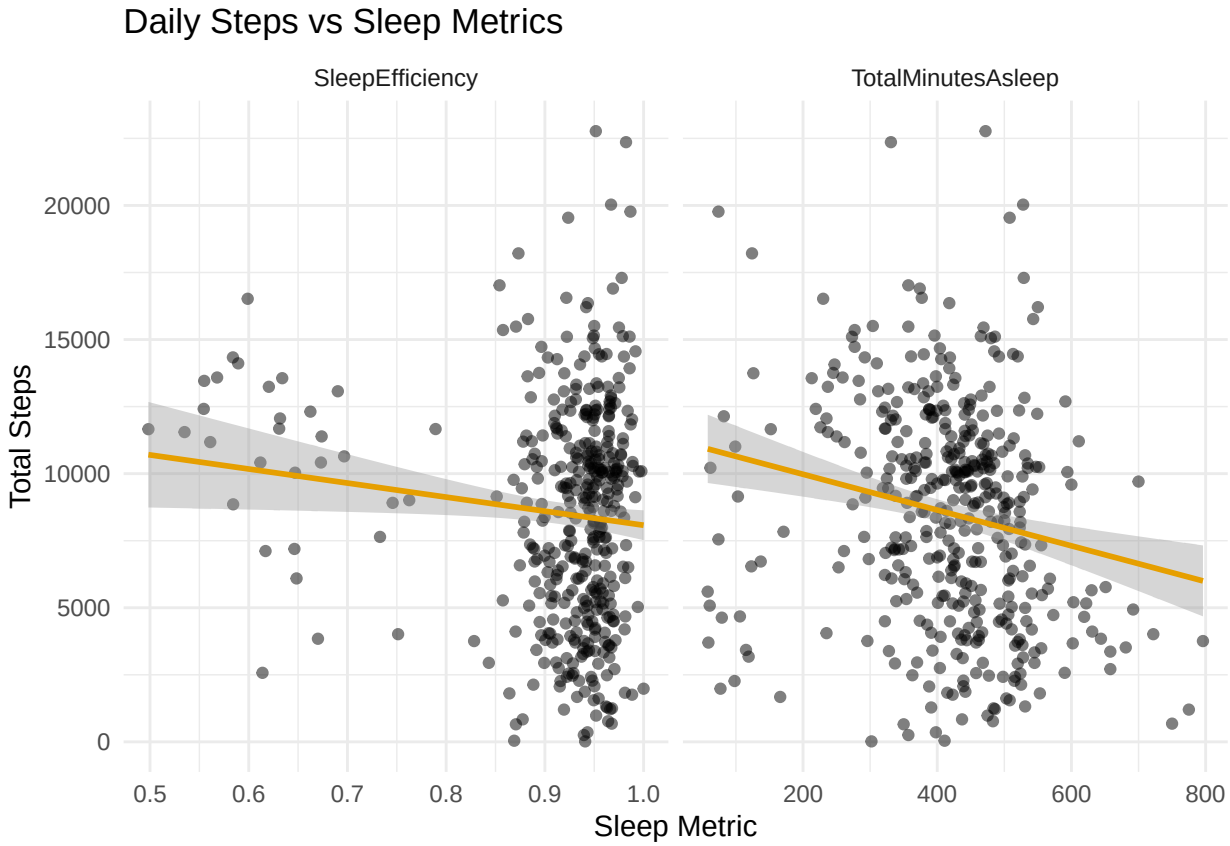
```
##
##  Pearson's product-moment correlation
##
## data:  clean_data$SleepEfficiency and clean_data$TotalSteps
## t = -2.236, df = 408, p-value = 0.02589
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.20469171 -0.01332017
## sample estimates:
##        cor
## -0.1100255
```

Interpretation: An inverse relationship (negative correlation) means high activity may reduce sleep efficiency — reinforcing Bellabeat's opportunity to emphasize **recovery and rest balance**.

---

8. Comparing Sleep Duration vs Sleep Efficiency

```
clean_data %>%
  select(TotalSteps, TotalMinutesAsleep, SleepEfficiency) %>%
  pivot_longer(cols = c(TotalMinutesAsleep, SleepEfficiency),
               names_to = "SleepMetric", values_to = "Value") %>%
  ggplot(aes(x = Value, y = TotalSteps)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "#E69F00") +
  facet_wrap(~ SleepMetric, scales = "free_x") +
  theme_minimal() +
  labs(title = "Daily Steps vs Sleep Metrics", x = "Sleep Metric", y = "Total Steps")
```



Interpretation: Faceted charts highlight how both **sleep duration** and **sleep efficiency** relate differently to activity. You might see a mild inverse correlation for both, supporting the recovery hypothesis.

---

9. Correlation Matrix

```r
clean_data %>%
  select(TotalSteps, Calories, TotalActiveMinutes, TotalMinutesAsleep, SleepEfficiency) %>%
  cor(use = "complete.obs") %>%
  round(2)
```

```
##                     TotalSteps Calories TotalActiveMinutes TotalMinutesAsleep
## TotalSteps                1.00     0.41               0.74              -0.19
## Calories                  0.41     1.00               0.39              -0.03
## TotalActiveMinutes        0.74     0.39               1.00              -0.07
## TotalMinutesAsleep       -0.19    -0.03              -0.07               1.00
## SleepEfficiency          -0.11     0.29               0.04               0.26
##                     SleepEfficiency
## TotalSteps                    -0.11
## Calories                       0.29
## TotalActiveMinutes             0.04
## TotalMinutesAsleep             0.26
## SleepEfficiency                1.00
```

Interpretation:

- Steps & Calories → strong positive correlation
- Sleep metrics → weak or negative correlation with activity
- Suggests Bellabeat can integrate *"Rested Readiness"* feedback loops in the app.

---

10. Training Intensity & Sleep

10.1 Create an Intensity Score first

```r
clean_data <- clean_data %>%
mutate(
IntensityScore = (VeryActiveMinutes * 3) +
(FairlyActiveMinutes * 2) +
(LightlyActiveMinutes * 1)
)

summary(clean_data$IntensityScore)
```

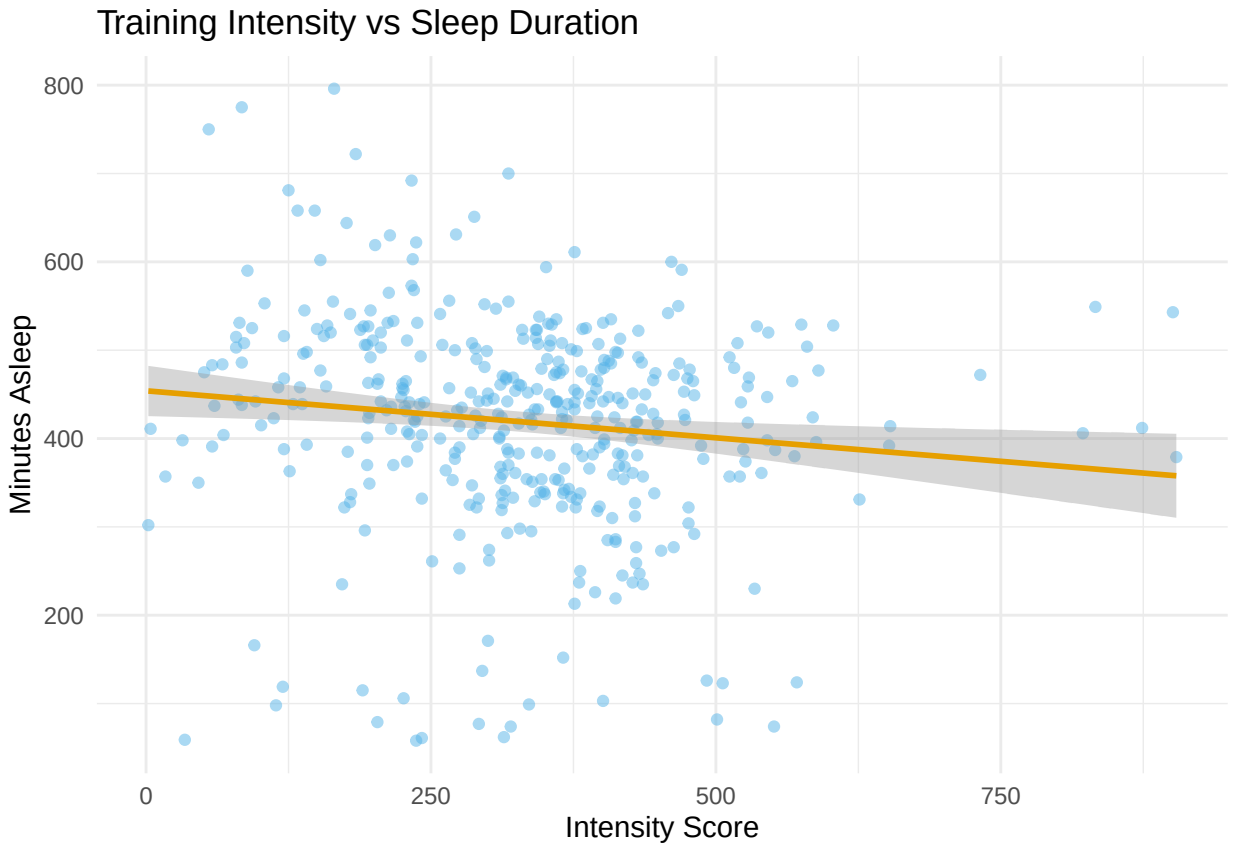
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0   229.0   334.5   327.5   411.8   904.0
```

Idea: This weighted IntensityScore reflects how hard a day's training was. Very active minutes count more than light activity.

10.2. Correlating Training Intensity & Sleep Duration

```r
ggplot(clean_data, aes(x = IntensityScore, y = TotalMinutesAsleep)) +
geom_point(alpha = 0.5, color = "#56B4E9") +
geom_smooth(method = "lm", color = "#E69F00", se = TRUE) +
labs(
title = "Training Intensity vs Sleep Duration",
x = "Intensity Score",
```

```
y = "Minutes Asleep"
) +
theme_minimal()
```



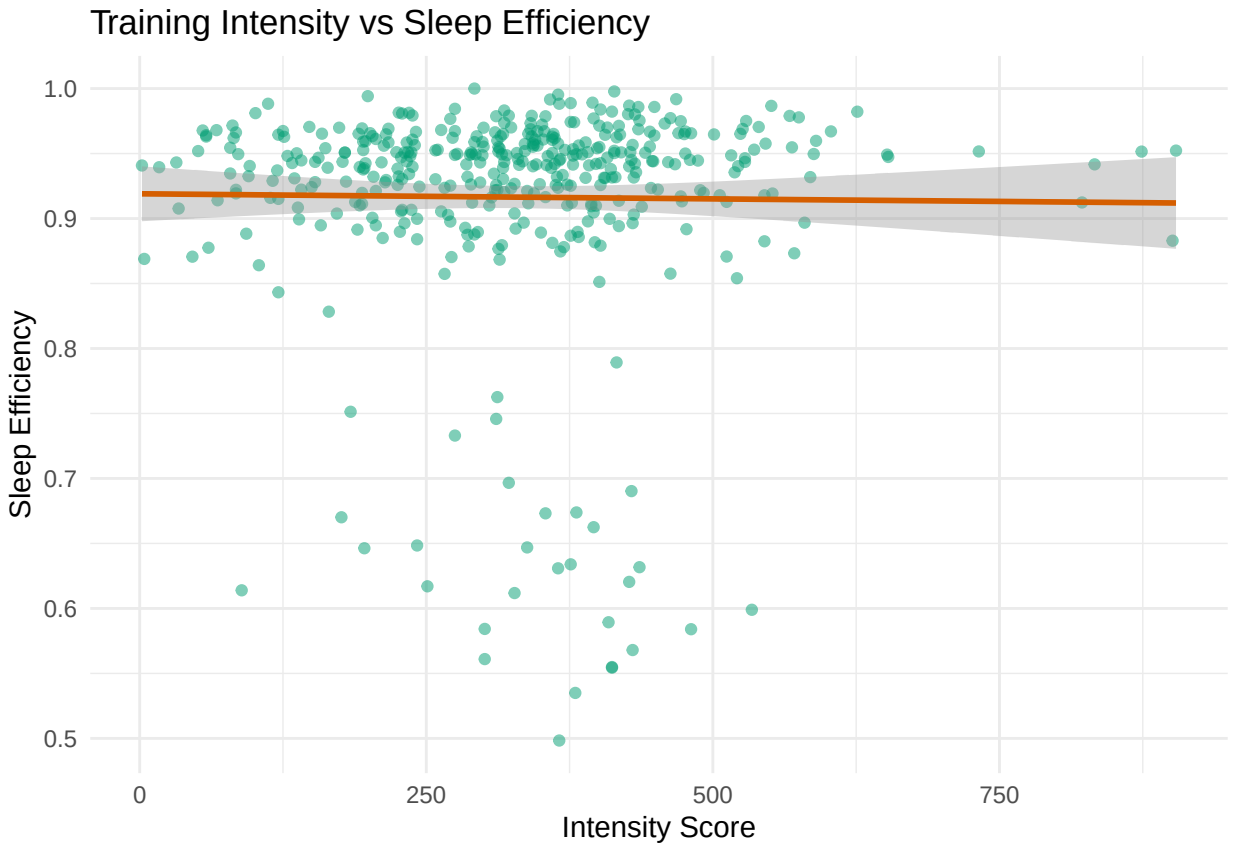

```
cor.test(clean_data$IntensityScore, clean_data$TotalMinutesAsleep)
```

```
##
##  Pearson's product-moment correlation
##
## data:  clean_data$IntensityScore and clean_data$TotalMinutesAsleep
## t = -2.6132, df = 408, p-value = 0.009303
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.22238543 -0.03184987
## sample estimates:
##        cor
## -0.1283014
```

10.3. Correlating Training Intensity & Sleep Efficiency

```
ggplot(clean_data, aes(x = IntensityScore, y = SleepEfficiency)) +
geom_point(alpha = 0.5, color = "#009E73") +
geom_smooth(method = "lm", color = "#D55E00", se = TRUE) +
labs(
title = "Training Intensity vs Sleep Efficiency",
```

```
x = "Intensity Score",
y = "Sleep Efficiency"
) +
theme_minimal()
```




```
cor.test(clean_data$IntensityScore, clean_data$SleepEfficiency)
```

```
##
##  Pearson's product-moment correlation
##
## data:  clean_data$IntensityScore and clean_data$SleepEfficiency
## t = -0.25954, df = 408, p-value = 0.7953
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.10955919  0.08410374
## sample estimates:
##         cor
## -0.01284821
```

---

11. Activity Type Analysis (Very / Fairly / Lightly Active)

11.1. Mean Time in Each Intensity

```
clean_data %>%
summarise(
```

```
MeanVeryActive = mean(VeryActiveMinutes, na.rm = TRUE),
MeanFairlyActive = mean(FairlyActiveMinutes, na.rm = TRUE),
MeanLightlyActive = mean(LightlyActiveMinutes, na.rm = TRUE)
)
```

```
## # A tibble: 1 x 3
##   MeanVeryActive MeanFairlyActive MeanLightlyActive
##            <dbl>            <dbl>             <dbl>
## 1           25.0             17.9              217.
```

11.2. Reshaping the data

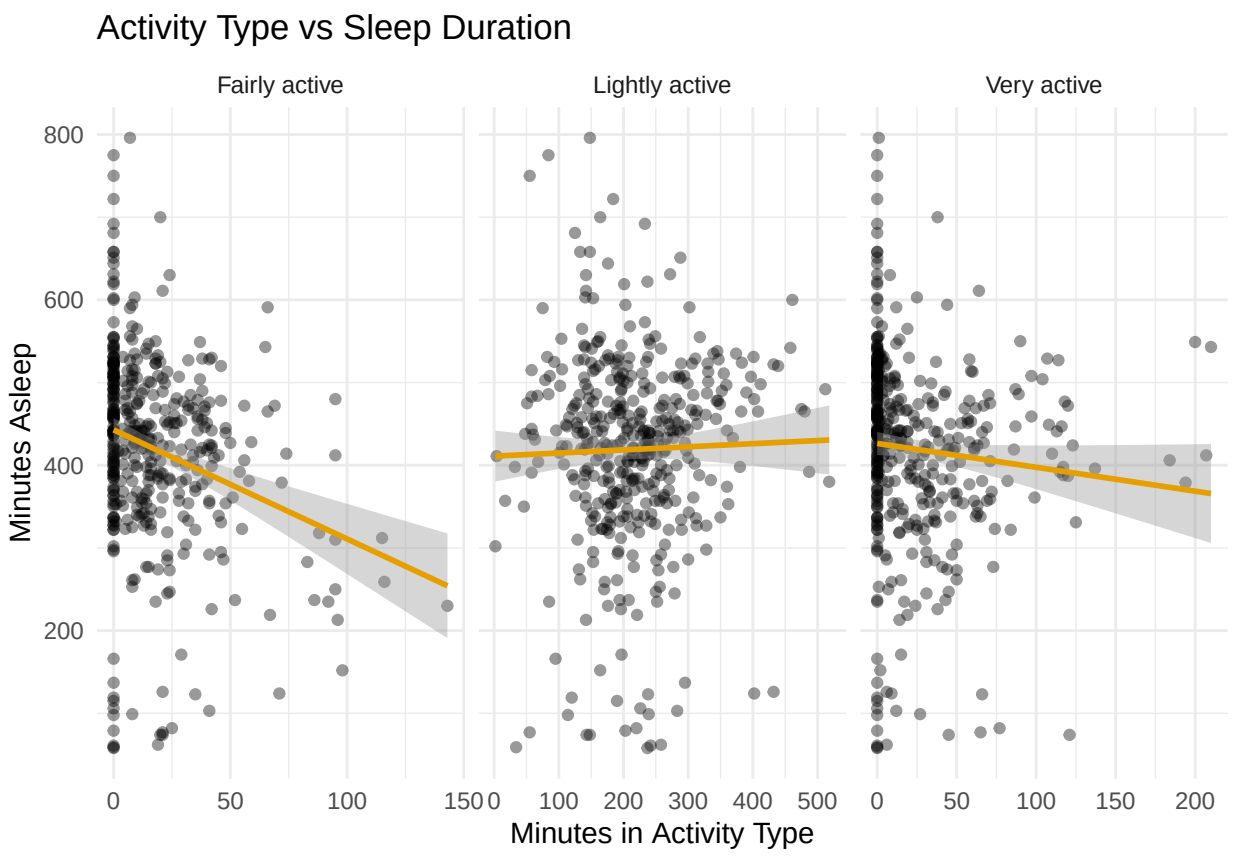
```
activity_sleep_long <- clean_data %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes,
         TotalMinutesAsleep, SleepEfficiency) %>%
  pivot_longer(
    cols = c(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes),
    names_to = "ActivityType",
    values_to = "MinutesActive"
  ) %>%
  mutate(
    ActivityType = recode(
      ActivityType,
      VeryActiveMinutes   = "Very active",
      FairlyActiveMinutes = "Fairly active",
      LightlyActiveMinutes= "Lightly active"
    )
  )

glimpse(activity_sleep_long)
```

```
## Rows: 1,230
## Columns: 4
## $ TotalMinutesAsleep <dbl> 327, 327, 327, 384, 384, 384, 412, 412, 412, 340, 3~
## $ SleepEfficiency    <dbl> 0.9450867, 0.9450867, 0.9450867, 0.9434889, 0.94348~
## $ ActivityType       <chr> "Very active", "Fairly active", "Lightly active", "~
## $ MinutesActive      <dbl> 25, 13, 328, 21, 19, 217, 29, 34, 209, 36, 10, 221,~
```

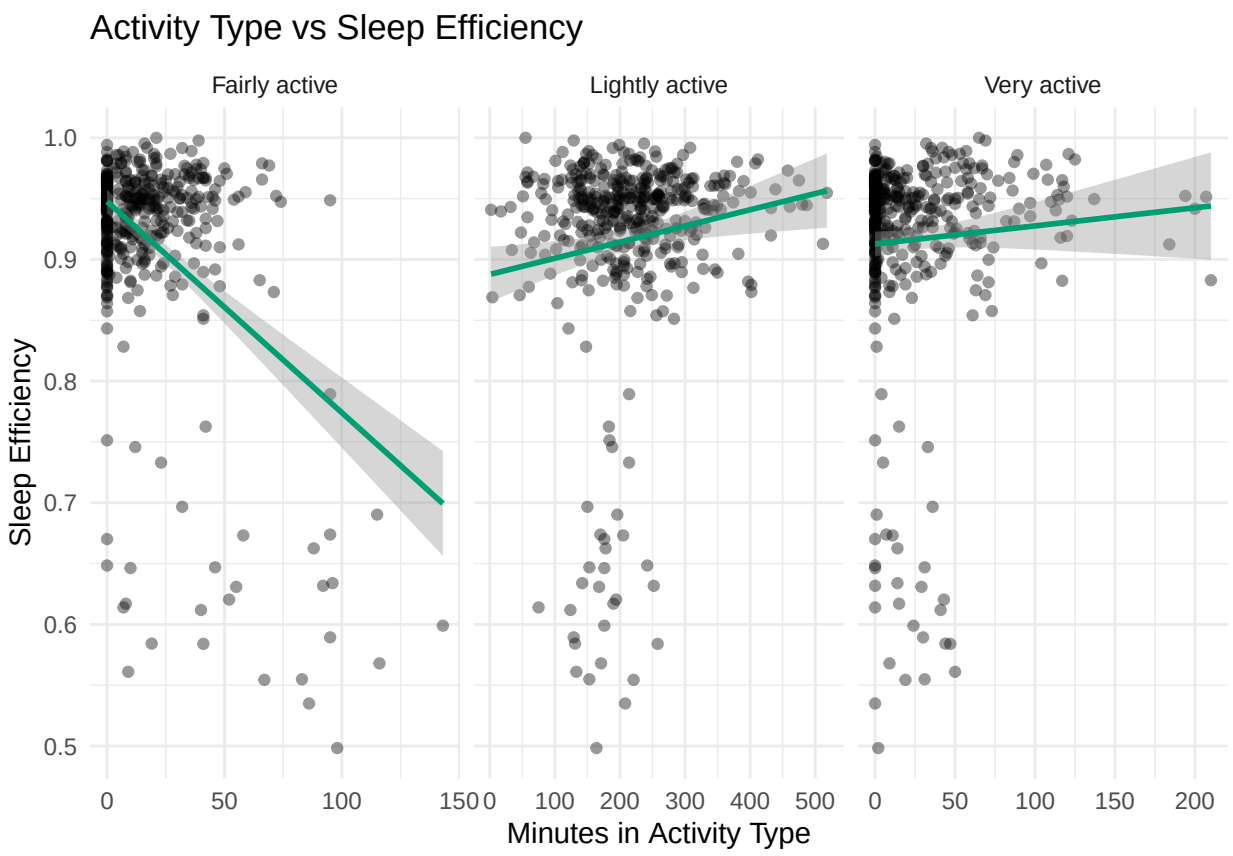
11.3. Correlating Activity Type - Sleep Duration

```
ggplot(activity_sleep_long, aes(x = MinutesActive, y = TotalMinutesAsleep)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = TRUE, color = "#E69F00") +
  facet_wrap(~ ActivityType, scales = "free_x") +
  theme_minimal() +
  labs(
    title = "Activity Type vs Sleep Duration",
    x = "Minutes in Activity Type",
    y = "Minutes Asleep"
  )
```



11.4. Correlating Activity Type - Sleep Efficiency

```
ggplot(activity_sleep_long, aes(x = MinutesActive, y = SleepEfficiency)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = TRUE, color = "#009E73") +
  facet_wrap(~ ActivityType, scales = "free_x") +
  theme_minimal() +
  labs(
    title = "Activity Type vs Sleep Efficiency",
    x = "Minutes in Activity Type",
    y = "Sleep Efficiency"
  )
```



11.5. Correlation Summary

```
list(
VeryActive_Duration = cor.test(clean_data$VeryActiveMinutes, clean_data$TotalMinutesAsleep),
FairlyActive_Duration = cor.test(clean_data$FairlyActiveMinutes, clean_data$TotalMinutesAsleep),
LightlyActive_Duration = cor.test(clean_data$LightlyActiveMinutes, clean_data$TotalMinutesAsleep),

VeryActive_Eff = cor.test(clean_data$VeryActiveMinutes, clean_data$SleepEfficiency),
FairlyActive_Eff = cor.test(clean_data$FairlyActiveMinutes, clean_data$SleepEfficiency),
LightlyActive_Eff = cor.test(clean_data$LightlyActiveMinutes, clean_data$SleepEfficiency)
)
```

```
## $VeryActive_Duration
##
##  Pearson's product-moment correlation
##
## data:  clean_data$VeryActiveMinutes and clean_data$TotalMinutesAsleep
## t = -1.787, df = 408, p-value = 0.07468
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.183408523  0.008795793
## sample estimates:
##         cor
## -0.08812658
##
```

```
##
## $FairlyActive_Duration
##
##  Pearson's product-moment correlation
##
## data:  clean_data$FairlyActiveMinutes and clean_data$TotalMinutesAsleep
## t = -5.1977, df = 408, p-value = 3.195e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3379000 -0.1561288
## sample estimates:
##        cor
## -0.2492079
##
##
## $LightlyActive_Duration
##
##  Pearson's product-moment correlation
##
## data:  clean_data$LightlyActiveMinutes and clean_data$TotalMinutesAsleep
## t = 0.55737, df = 408, p-value = 0.5776
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06944947  0.12409914
## sample estimates:
##        cor
## 0.02758336
##
##
## $VeryActive_Eff
##
##  Pearson's product-moment correlation
##
## data:  clean_data$VeryActiveMinutes and clean_data$SleepEfficiency
## t = 1.2484, df = 408, p-value = 0.2126
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03537135  0.15759307
## sample estimates:
##        cor
## 0.06168727
##
##
## $FairlyActive_Eff
##
##  Pearson's product-moment correlation
##
## data:  clean_data$FairlyActiveMinutes and clean_data$SleepEfficiency
## t = -10.061, df = 408, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5202276 -0.3647450
## sample estimates:
##        cor
```

```
## -0.4458431
##
##
## $LightlyActive_Eff
##
##  Pearson's product-moment correlation
##
## data:  clean_data$LightlyActiveMinutes and clean_data$SleepEfficiency
## t = 2.6908, df = 408, p-value = 0.00742
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0356583 0.2260066
## sample estimates:
##       cor
## 0.1320496
```

12. Statistical Significance — Interpreting p-values

When using `cor.test()`:

Term	Meaning
cor	Correlation coefficient (-1 to +1) — strength & direction
p-value	Probability the result is random
$p < 0.05$	Statistically significant
$p < 0.01$	Highly significant
$p > 0.05$	Not significant (relationship may be due to chance)

---

13. Key Findings

1. **Activity Behavior**

Users show clear weekday–weekend differences:

Weekend activity is consistently lower, with step counts dropping visibly on Saturdays and Sundays.

Daily activity varies widely across individuals, with some days reaching >20,000 steps and others nearly sedentary.

Steps and intensity are strongly linked:

Higher TotalActiveMinutes is strongly correlated with higher steps (+0.74) and calories (+0.41).

2. **Sleep Behavior**

Users sleep an average of 419 minutes (~7 hours) with high efficiency (~91%).

Sleep duration ranges widely (from roughly 1 hour to 13 hours), suggesting variability in habits or data logging.

Sleep efficiency tends to be high overall, but a few low-efficiency outliers indicate inconsistent nights.

3. **Activity - Sleep Relationships**

Sleep Duration vs Total Steps Correlation = -0.19 ($p < 0.001$)

→ Statistically significant negative relationship.

→ Users who sleep longer tend to walk less the next day.

→ Likely reflects recovery days, illness, or low-activity rest days.

Sleep Efficiency vs Total Steps Correlation = -0.11 (p = 0.026)

→ Slight but significant inverse relationship.

→ High-efficiency sleep may follow lower-activity days → users rest more efficiently when less active.

4. **Training Intensity & Sleep**

Intensity Score → Weighted score = (very active × 3) + (fairly active × 2) + (lightly active × 1)

Intensity vs Sleep Duration: Correlation = -0.13 (p = 0.009)

→ Higher-intensity days precede shorter sleep.

Intensity vs Sleep Efficiency: Correlation = -0.01 (p = 0.79)

→ No meaningful link between intensity and efficiency.

→ Users may compensate (e.g., longer time in bed) to maintain efficiency.

5. **Activity Type Breakdown**

Correlations from the summary table:

Sleep Duration

Fairly active minutes show the strongest negative correlation (-0.25, $p < 0.000001$).

→ Even moderate-intensity movement is associated with reduced sleep duration.

Very active: slight, nonsignificant negative (-0.09, $p = 0.07$)

Lightly active: no correlation ($+0.03$, ns)

Sleep Efficiency

Fairly active minutes show a strong negative correlation (-0.45, $p < 0.00000000000000022$).

→ The most influential activity type affecting sleep efficiency.

Very active: slight positive nonsignificant ($+0.06$)

Lightly active: small positive ($+0.13$, $p = 0.007$)

Takeaway:

→ Moderate-intensity activity (fairly active minutes) appears to be the strongest driver of reduced sleep duration and efficiency.

→ Light activity may even support better sleep efficiency.

6. **Overall Behavioral Insights**

Users seem to follow a structured weekday routine with predictable sleep and activity.

Weekends emerge as opportunities for re-engagement, lighter activities, and recovery prompts.

Sleep patterns show that recovery-related behaviors can be anticipated from activity patterns.

Light activity plays a supportive role in sleep, while moderate activity has the largest negative effect.

---

14. Analyst's Reflection

The dataset reveals a nuanced but consistent behavioural pattern: users with higher daily activity, particularly moderate-intensity activity, tend to show shorter sleep duration and reduced sleep efficiency.

This suggests that Bellabeat users may benefit from **recovery-focused insights**, especially on days when their activity patterns indicate high strain.

The weekend drop in activity also stands out as a predictable behavioral rhythm, presenting opportunities for Bellabeat to encourage lighter, restorative activity or mindful habits during these low-engagement periods.

While the correlations are significant in several cases, their magnitudes are generally small, reflecting the complexity of human sleep. Therefore, Bellabeat's strategy should focus on guiding trends rather than strict causal predictions.

Ultimately, the analysis provides Bellabeat with a foundation for:

- Smart notifications that adapt to user fatigue signals
- Balanced habit coaching that integrates activity and sleep insights
- User segmentation based on observed activity-sleep patterns

This strengthens Bellabeat's vision of empowering users with preventative, behaviour-aware wellness guidance.