Stakeholders

* **Urška Sršen**: Bellabeat’s cofounder and Chief Creative Officer
* **Sando Mur**: Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team
* **Bellabeat marketing analytics team**: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat’s marketing strategy.

1.3 Business Task

Analyze usage data from a non-Bellabeat smart device for insight into trends we can apply to Bellabeat’s app and customers, with a focus on opportunities for growth and Bellabeat’s marketing strategy.

1.4 Deliverables

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top high-level content recommendations based on your analysis
7. 2 Prepare phase
8. 2.1 Dataset
9. We’ll be using the [FitBit Fitness Tracker Dataset](https://www.kaggle.com/datasets/arashnic/fitbit) made available through [Möbius](https://www.kaggle.com/arashnic) on Kaggle with a [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/) license, allowing us to freely use it.
10. 2.2 Description

In Möbius’s own words: “This dataset was generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.”

.3 Overview

* The data was organized into 18 csv files and each of them organized in long format, meaning that each participant has data in multiple rows based on varying time points.
* Each file represents either one of the eight metrics or one of multiple time-frames for the metrics.
* After examining each in Excel, most datasets include 33 participants (not the 30 stated by Möbius above), though some of the smart device’s features weren’t used by all the volunteers.
* Also, the data begins on 4-12-2016 and lasts for 31 days, rather than beginning on 3-12 as Möbius said.

Here’s a list of the features/metrics, a brief description, and the varying time-frames for each feature/metric:

1. Activity — The most comprehensive dataset, with data on daily total steps, distance, sedentary minutes, and calories, and the active minutes and active distance variables both broken into three levels of intensity (very, moderate, and light).
   * Daily
2. Calories — The number of calories burned in each time-frame by each participant.
   * Daily
   * Hourly
   * Minutes — Wide and narrow
3. Intensity — The total intensity level of activity each participant exhibited within each time-frame.
   * Daily
   * Hourly
   * Minutes — Wide and narrow
4. Steps — The number of steps each participant took within each time-frame.
   * Daily
   * Hourly
   * Minutes — Wide a
5. Heartrate — Each participant’s heartrate at each second of each day.
   * Seconds
6. METs — MET refers to metabolic equivalent, with a MET of 1 when at rest. The number determines the user’s activity levels that are used in the Daily Activity dataset (such as very, moderately, or lightly active minutes/distance).
   * Minutes
7. Sleep — **Only 24 participants used this feature**, which tracks the minutes they slept, minutes spent in bed, and how many sleep sessions they took each day.
   * Daily
   * Minutes
8. Weight log — **Only 8 participants used this feature**, which recorded the user’s BMI and weight in both Kilograms and pounds.
   * Per use of feature

import os

# List the files in the directory to understand the structure

files = os.listdir('fitabeat\_case\_study/data used on fitabeat case study')

print(files)

The directory contains the following files:

* **dailySteps\_Merged.csv**
* **houlyIntensities\_Merged.csv**
* **dailyCalories\_merged.csv**
* **SleepDay\_Merged.csv**
* **DailyActivity\_Merged.csv**
* **DailyIntensities\_Merged.csv**

Let's start by loading and inspecting the **DailyActivity\_Merged.csv** file for analysis.

# Load the necessary libraries

library(dplyr)

library(ggplot2)

# Load the daily activity data

activity\_data <- read.csv('fitabeat case study/data used on fitabeat case study/dailyActivity\_merged.csv')

Next, let's analyze the data to identify trends and insights that can be applied to Bellabeat's app and customers. We'll focus on the following aspects:

1. **Daily Steps and Activity Levels**: Understanding the distribution of daily steps and activity levels.
2. **Calories Burned**: Analyzing the calories burned in relation to activity levels.
3. **Active Minutes**: Examining the distribution of very active, fairly active, and lightly active minutes.

Let's start with the analysis of daily steps and activity levels.

# Convert ActivityDate to Date type

activity\_data$ActivityDate <- as.Date(activity\_data$ActivityDate, format='%m/%d/%Y')

# Extract hour from ActivityDate

activity\_data$Hour <- format(as.POSIXct(activity\_data$ActivityDate, format='%Y-%m-%d %H:%M:%S'), '%H')

# Summarize activity levels by hour

activity\_by\_hour <- activity\_data %>%

group\_by(Hour) %>%

summarise(TotalSteps = sum(TotalSteps),

TotalDistance = sum(TotalDistance),

Calories = sum(Calories))

# Plot the trends in activity levels based on the time of day

p1 <- ggplot(activity\_by\_hour, aes(x = as.numeric(Hour), y = TotalSteps)) +

geom\_line() +

labs(title = 'Total Steps by Hour of the Day', x = 'Hour of the Day', y = 'Total Steps')

p2 <- ggplot(activity\_by\_hour, aes(x = as.numeric(Hour), y = TotalDistance)) +

geom\_line() +

labs(title = 'Total Distance by Hour of the Day', x = 'Hour of the Day', y = 'Total Distance')

p3 <- ggplot(activity\_by\_hour, aes(x = as.numeric(Hour), y = Calories)) +

geom\_line() +

labs(title = 'Calories Burned by Hour of the Day', x = 'Hour of the Day', y = 'Calories Burned')

# Print the plots

print(p1)

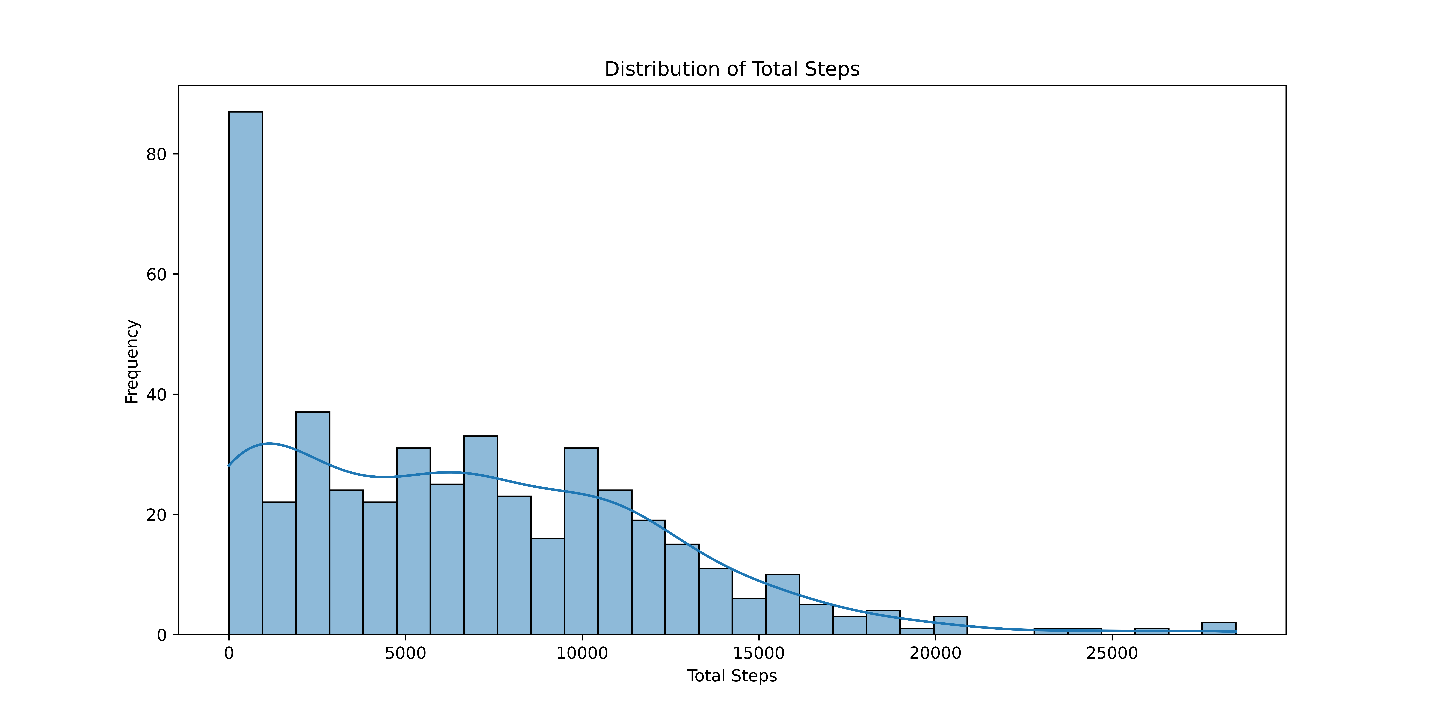
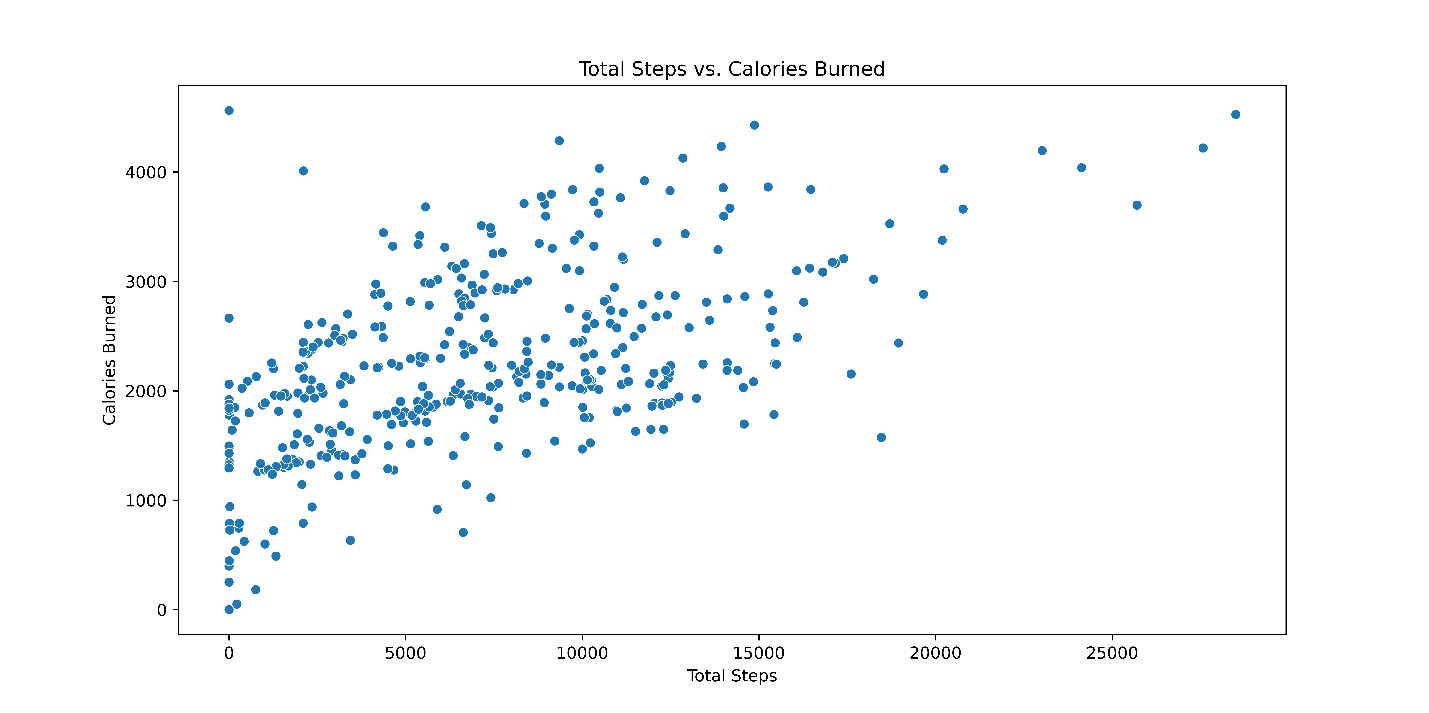
print(p2)

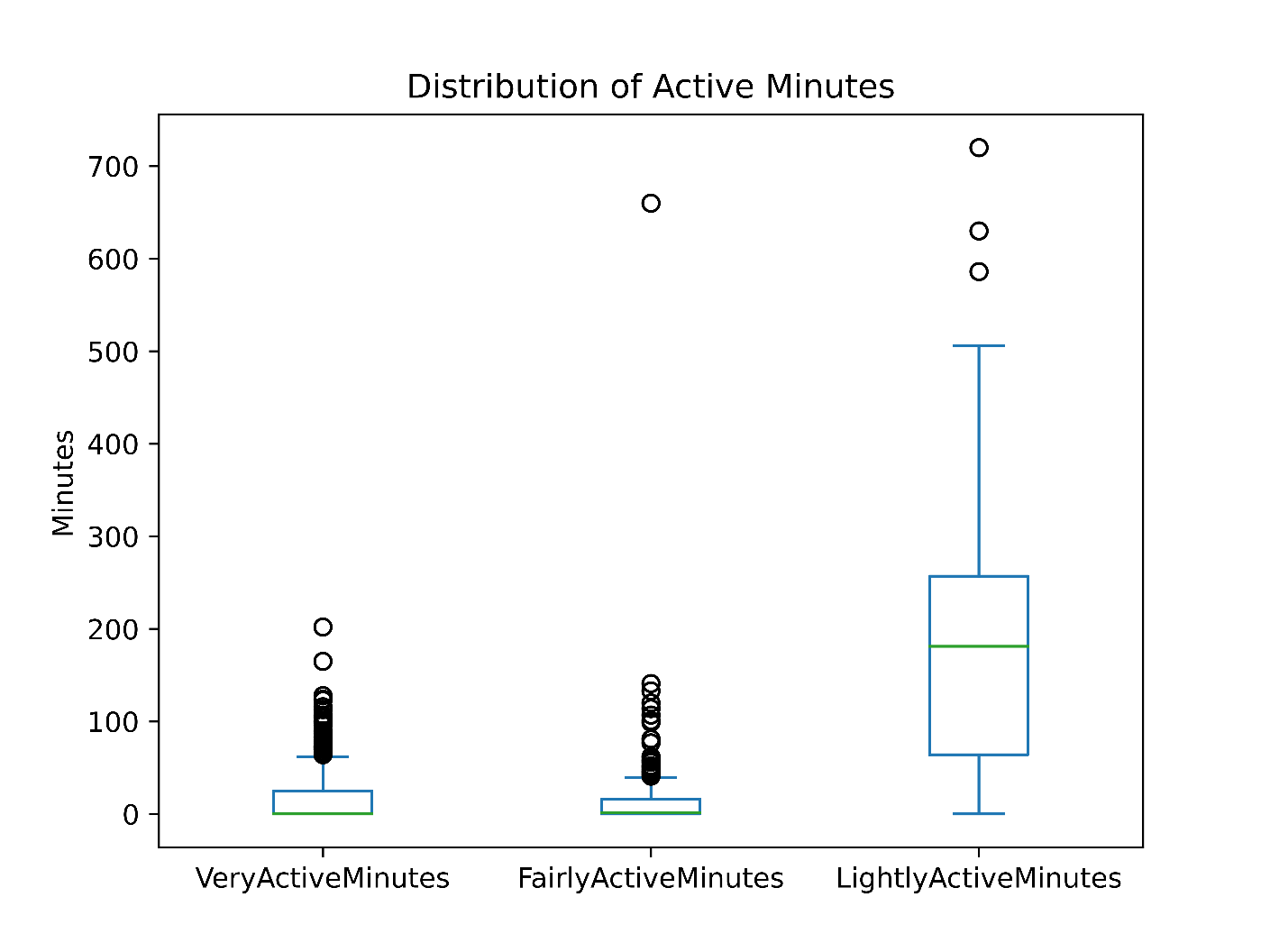
print(p3)

A blue dots on a white background

Description automatically generated

Here are the visualizations for the relationship between total steps, very active minutes, and calories burned:





Correlation between Sleep Duration and Calories Burned:

# Load necessary libraries

library(dplyr)

library(ggplot2)

library(readr)

# Load the data

activity\_data <- read\_csv("DailyActivity\_Merged.csv")

calories\_data <- read\_csv("dailyCalories\_merged.csv")

intensities\_data <- read\_csv("DailyIntensities\_Merged.csv")

sleep\_data <- read\_csv("SleepDay\_Merged.csv")

# Convert date columns to Date type

activity\_data$ActivityDate <- as.Date(activity\_data$ActivityDate, format="%m/%d/%Y")

calories\_data$ActivityDay <- as.Date(calories\_data$ActivityDay, format="%m/%d/%Y")

intensities\_data$ActivityDay <- as.Date(intensities\_data$ActivityDay, format="%m/%d/%Y")

sleep\_data$SleepDay <- as.Date(sleep\_data$SleepDay, format="%m/%d/%Y")

# Check if 'Calories' column exists and its data type

if ('Calories' %in% colnames(merged\_data)) {

# Check the data type and non-numeric values

print(unique(merged\_data$Calories))

# Convert 'Calories' to numeric

merged\_data$Calories <- as.numeric(merged\_data$Calories)

# Handle missing values if any

merged\_data <- merged\_data %>% filter(!is.na(Calories))

} else {

print("The 'Calories' column does not exist in the merged data.")

}

# Verify the conversion

str(merged\_data$Calories)

# Verify column names in each dataframe before merging

colnames(activity\_data)

colnames(calories\_data)

colnames(intensities\_data)

colnames(sleep\_data)

# Verify the column names and structure after merging

colnames(merged\_data)

str(merged\_data)

# Check if any columns contain 'Calories' in their name

grep("Calories", colnames(merged\_data), value = TRUE)

# Assuming the correct column name for Calories is identified, e.g., 'Calories\_Calories'

correct\_calories\_column <- "Calories" # Replace with the correct column name if needed

# Check if the identified column exists

if (correct\_calories\_column %in% colnames(merged\_data)) {

# Ensure it's numeric

merged\_data[[correct\_calories\_column]] <- as.numeric(merged\_data[[correct\_calories\_column]])

# Handle missing values

merged\_data <- merged\_data %>% filter(!is.na(merged\_data[[correct\_calories\_column]]))

# Plotting

ggplot(merged\_data, aes\_string(x="TotalMinutesAsleep", y=correct\_calories\_column)) +

geom\_point(alpha=0.5) +

geom\_smooth(method="lm", col="red") +

labs(title="Sleep Duration vs. Calories Burned", x="Total Minutes Asleep", y="Calories Burned") +

theme\_minimal()

} else {

print(paste("The column", correct\_calories\_column, "does not exist in the merged data."))

}

# Merge the datasets with clear column names

merged\_data <- activity\_data %>%

left\_join(calories\_data %>% rename(Calories\_calories = Calories),

by=c("Id" = "Id", "ActivityDate" = "ActivityDay")) %>%

left\_join(intensities\_data, by=c("Id" = "Id", "ActivityDate" = "ActivityDay")) %>%

left\_join(sleep\_data, by=c("Id" = "Id", "ActivityDate" = "SleepDay"))

# Verify the column names and structure after merging

colnames(merged\_data)

str(merged\_data)

# Ensure 'Calories\_calories' column is numeric and handle missing values

if ('Calories\_calories' %in% colnames(merged\_data)) {

merged\_data$Calories\_calories <- as.numeric(merged\_data$Calories\_calories)

merged\_data <- merged\_data %>% filter(!is.na(Calories\_calories))

} else {

print("The 'Calories\_calories' column does not exist in the merged data.")

}

# Verify the conversion

str(merged\_data$Calories\_calories)

# Scatter plot with regression line for Sleep Duration vs. Calories Burned

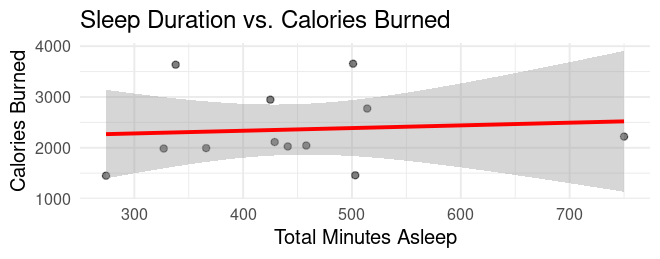
ggplot(merged\_data, aes(x=TotalMinutesAsleep, y=Calories\_calories)) +

geom\_point(alpha=0.5) +

geom\_smooth(method="lm", col="red") +

labs(title="Sleep Duration vs. Calories Burned", x="Total Minutes Asleep", y="Calories Burned") +

theme\_minimal()



# Visualize the relationship between sleep quality and activity levels

ggplot(merged\_data, aes(x=TotalMinutesAsleep, y=TotalSteps)) +

geom\_point() +

geom\_smooth(method="lm", col="blue") +

labs(title="Sleep Duration vs. Total Steps", x="Total Minutes Asleep", y="Total Steps")

A graph with a line and a blue line

Description automatically generated

# Correlation analysis

cor(merged\_data$TotalMinutesAsleep, merged\_data$Calories\_calories, use="complete.obs")

cor(merged\_data$TotalMinutesAsleep, merged\_data$TotalSteps, use="complete.obs")

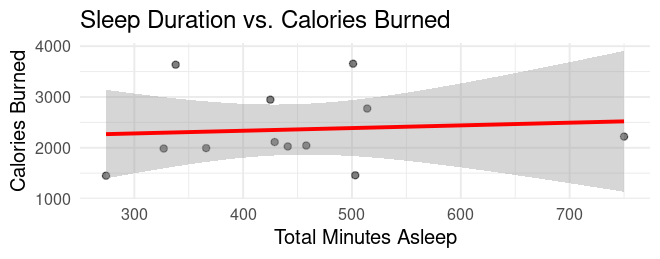
# Linear regression analysis

model1 <- lm(Calories ~ TotalMinutesAsleep, data=merged\_data)

summary(model1)

model2 <- lm(TotalSteps ~ TotalMinutesAsleep, data=merged\_data)

summary(model2)



cor(merged\_data$TotalMinutesAsleep, merged\_data$Calories\_calories, use="complete.obs")

[1] 0.08773128

*  The correlation coefficient is approximately 0.088, indicating a very weak positive correlation between sleep duration and calories burned. This suggests that there's almost no linear relationship between the amount of sleep and the number of calories burned.

 **Correlation between Sleep Duration and Total Steps:**

cor(merged\_data$TotalMinutesAsleep, merged\_data$TotalSteps, use="complete.obs")

[1] -0.3412517

The correlation coefficient is approximately -0.341, indicating a moderate negative correlation between sleep duration and total steps. This suggests that as sleep duration increases, the number of steps taken tends to decrease

**Linear Regression Models**

**Model 1: Predicting Calories Burned from Sleep Duration**

model1 <- lm(Calories\_calories ~ TotalMinutesAsleep, data=merged\_data) summary(model1)

* **Intercept (422.7324)**: This is the expected number of calories burned when the total minutes asleep is zero. However, since this scenario isn't practical, it's more about the starting point of the regression line.
* **Slope (0.5027)**: For every additional minute of sleep, the calories burned are expected to increase by 0.5027 calories. However, this increase is not statistically significant (p-value: 0.701), indicating that the relationship is not reliable.
* **R-squared (0.01533)**: Only about 1.5% of the variability in calories burned can be explained by the sleep duration. This indicates a very poor fit of the model.
* **P-value (0.7015)**: Since the p-value is much greater than 0.05, the model is not statistically significant.

**Model 2: Predicting Total Steps from Sleep Duration**

R

Copy code

model2 <- lm(TotalSteps ~ TotalMinutesAsleep, data=merged\_data)

summary(model2)

**Output:**

plaintext

Copy code

Call:

lm(formula = TotalSteps ~ TotalMinutesAsleep, data = merged\_data)

Residuals:

Min 1Q Median 3Q Max

-1668.4 -1232.0 -724.3 619.6 4205.7

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3611.602 2102.741 1.718 0.117

TotalMinutesAsleep -5.258 4.580 -1.148 0.278

Residual standard error: 1866 on 10 degrees of freedom

(12 observations deleted due to missingness)

Multiple R-squared: 0.1165, Adjusted R-squared: 0.0281

F-statistic: 1.318 on 1 and 10 DF, p-value: 0.2777

* **Intercept (3611.602)**: This is the expected number of total steps when the total minutes asleep is zero. Again, this is more about the starting point of the regression line.
* **Slope (-5.258)**: For every additional minute of sleep, the total steps are expected to decrease by 5.258 steps. However, this decrease is not statistically significant (p-value: 0.278).
* **R-squared (0.1165)**: About 11.65% of the variability in total steps can be explained by sleep duration, which is still quite low.
* **P-value (0.2777)**: Since the p-value is greater than 0.05, the model is not statistically significant.

**Summary**

1. **Correlations**:
   * The relationship between sleep duration and calories burned is very weak and positive.
   * The relationship between sleep duration and total steps is moderately negative.
2. **Linear Models**:
   * Neither model shows a significant relationship between sleep duration and calories burned or total steps.
   * The R-squared values indicate that the models explain very little of the variance in the dependent variables.
   * The p-values indicate that the models are not statistically significant.

**Interpretation**

* **Marketing Strategy**: Since neither sleep duration nor activity levels (calories burned and steps) show significant relationships, the marketing strategy might benefit from focusing on other factors that could influence user engagement and product appeal.
* **Opportunities for Growth**: Bellabeat could explore additional features or metrics that might be more closely related to user behavior and wellness goals. Collecting more comprehensive data and considering other variables such as diet, stress levels, or social engagement could provide better insights.
* **User Insights**: Personalized insights might be more effective than general trends. Providing users with tailored advice based on their unique data could increase engagement and perceived value.