# How Can a Wellness Technology Company Play It Smart?

Gregory Charles

February 24th, 2025

## Introduction

This case study analyzes smart device usage data from publicly available Fitbit datasets to provide insights for Bellabeat, a wellness technology company focused on empowering women through beautifully designed health-focused products. The goal is to understand consumer trends related to activity, sleep, and calorie expenditure, and apply them to Bellabeat's marketing strategy, focusing on the Leaf wellness tracker. This report will cover the data analysis process from data preparation to actionable recommendations.

## 1. Ask

**Business Task:** Analyze smart device usage data to understand consumer trends related to activity, sleep, and calorie expenditure, and provide data-backed marketing recommendations for Bellabeat's Leaf product, specifically focusing on how these trends can inform product positioning and feature promotion.

**Key Stakeholders:** Urška Sršen (Co-founder & CCO, who initiated this analysis), Sando Mur (Co-founder & Mathematician, providing expertise in data interpretation), Bellabeat marketing analytics team (responsible for implementing the recommendations).

**Guiding Questions:**

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

## 2. Prepare

**Data Sources:**

- *Fitbit Fitness Tracker Data (Kaggle):* This dataset contains minute-level data on activity, heart rate, and sleep from 30 Fitbit users. It includes daily activity summaries, calorie burn, intensity levels, steps, heart rate (seconds), hourly data (calories, intensity, steps), minute-level data (calories, intensity, METs, steps), sleep records, and weight logs.

**Data Organization:** The data is provided in multiple CSV files, requiring merging based on user ID and date/time. The format is mostly long, with some wide format files for minute-level data.

**Data Credibility:** The data is from a small sample (30 users) and may not be representative of the broader population. Self-reported data (weight logs) can be subject to bias. While the dataset offers granular data, its limited size and potential biases need to be considered. We will primarily focus on daily summaries to mitigate some of the noise in minute-level data.

**Data Limitations:** The dataset is from Fitbit users, not Bellabeat users. We assume similar usage patterns but this is a potential limitation. The small sample size also limits generalizability.

**Data Integrity:** The data was downloaded from a reputable source (Kaggle) and checked for completeness. Basic data checks (e.g., missing values, impossible values) will be performed during processing.

## 3. Process

### 3.1 Install packages

```
# install.packages("readr")
# install.packages("tidyverse")
# install.packages("lubridate")
# install.packages("ggplot2")

library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x 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(ggplot2)
library(dplyr)
```

### 3.2 Load the data

```
# Define data directory (adjust as needed)
data_dir <- "dataset/" # Replace with the actual path to your data files

daily_activity <- read_csv(paste0(data_dir, "dailyActivity_merged.csv"))
```

```
## 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.
```

```
daily_calories <- read_csv(paste0(data_dir, "dailyCalories_merged.csv"))
```

```
## Rows: 940 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (2): Id, Calories
##
## 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.
```

```r
daily_intensities <- read_csv(paste0(data_dir, "dailyIntensities_merged.csv"))
```

```
## Rows: 940 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (9): Id, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, Ve...
##
## 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.
```

```r
daily_steps <- read_csv(paste0(data_dir, "dailySteps_merged.csv"))
```

```
## Rows: 940 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## 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.
```

```r
heartrate_seconds <- read_csv(paste0(data_dir, "heartrate_seconds_merged.csv"))
```

```
## Rows: 2483658 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Time
## dbl (2): Id, Value
##
## 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.
```

```r
hourly_calories <- read_csv(paste0(data_dir, "hourlyCalories_merged.csv"))
```

```
## Rows: 22099 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, Calories
##
## 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.
```

```r
hourly_intensities <- read_csv(paste0(data_dir, "hourlyIntensities_merged.csv"))
```

```
## Rows: 22099 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (3): Id, TotalIntensity, AverageIntensity
##
## 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.
```

```r
hourly_steps <- read_csv(paste0(data_dir, "hourlySteps_merged.csv"))
```

```
## 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.
```

```
minute_calories_narrow <- read_csv(paste0(data_dir, "minuteCaloriesNarrow_merged.csv"))
```

```
## Rows: 1325580 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityMinute
## dbl (2): Id, Calories
##
## 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.
```

```
minute_intensities_narrow <- read_csv(paste0(data_dir, "minuteIntensitiesNarrow_merged.csv"))
```

```
## Rows: 1325580 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityMinute
## dbl (2): Id, Intensity
##
## 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.
```

```
minute_mets_narrow <- read_csv(paste0(data_dir, "minuteMETsNarrow_merged.csv"))
```

```
## Rows: 1325580 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityMinute
## dbl (2): Id, METs
##
## 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.
```

```
minute_sleep <- read_csv(paste0(data_dir, "minuteSleep_merged.csv"))
```

```
## Rows: 188521 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): date
## dbl (3): Id, value, logId
##
## 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.
```

```
minute_steps_narrow <- read_csv(paste0(data_dir, "minuteStepsNarrow_merged.csv"))
```

```
## Rows: 1325580 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (1): ActivityMinute
## dbl (2): Id, Steps
##
## 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(paste0(data_dir, "sleepDay_merged.csv"))
```

```
## 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.
```

```
weight_log <- read_csv(paste0(data_dir, "weightLogInfo_merged.csv"))
```

```
## Rows: 67 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
##
## 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.
```

### 3.3 Check Data

```
# Check the structure of each dataframe
check_data <- function(df, name) {
  print(paste("Checking", name, ":"))
  print(glimpse(df))
  print(summary(df))
  print(paste("Missing values:", sum(is.na(df))))
  cat("\n")
}

# Apply the function to the dataframes you want to check
check_data(daily_activity, "daily_activity")
```

```
## [1] "Checking daily_activity :"
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ TotalSteps               <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

```
## $ VeryActiveMinutes        <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~
## $ FairlyActiveMinutes      <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~
## $ LightlyActiveMinutes     <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~
## $ SedentaryMinutes         <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~
## $ Calories                 <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~
## # A tibble: 940 x 15
##            Id ActivityDate TotalSteps TotalDistance TrackerDistance
##         <dbl> <chr>             <dbl>         <dbl>           <dbl>
##  1 1503960366 4/12/2016         13162          8.5             8.5
##  2 1503960366 4/13/2016         10735          6.97            6.97
##  3 1503960366 4/14/2016         10460          6.74            6.74
##  4 1503960366 4/15/2016          9762          6.28            6.28
##  5 1503960366 4/16/2016         12669          8.16            8.16
##  6 1503960366 4/17/2016          9705          6.48            6.48
##  7 1503960366 4/18/2016         13019          8.59            8.59
##  8 1503960366 4/19/2016         15506          9.88            9.88
##  9 1503960366 4/20/2016         10544          6.68            6.68
## 10 1503960366 4/21/2016          9819          6.34            6.34
## # i 930 more rows
## # i 10 more variables: LoggedActivitiesDistance <dbl>,
## #   VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
##        Id              ActivityDate         TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:940         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Mode  :character   Median : 7406   Median : 5.245
##  Mean   :4.855e+09                      Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09                      3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09                      Max.   :36019   Max.   :28.030
##  TrackerDistance  LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.000   Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 2.620   1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 5.245   Median :0.0000           Median : 0.210
##  Mean   : 5.475   Mean   :0.1082           Mean   : 1.503
##  3rd Qu.: 7.710   3rd Qu.:0.0000           3rd Qu.: 2.053
##  Max.   :28.030   Max.   :4.9421           Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5675           Mean   : 3.341      Mean   :0.001606
##  3rd Qu.:0.8000           3rd Qu.: 4.782      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0        1st Qu.: 729.8
##  Median :  4.00    Median :  6.00      Median :199.0        Median :1057.5
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8        Mean   : 991.2
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0        3rd Qu.:1229.5
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
##     Calories
##  Min.   :   0
```

```
##  1st Qu.:1828
##  Median :2134
##  Mean   :2304
##  3rd Qu.:2793
##  Max.   :4900
## [1] "Missing values: 0"
```

```
check_data(daily_calories, "daily_calories")
```

```
## [1] "Checking daily_calories :"
## Rows: 940
## Columns: 3
## $ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366~
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/~
## $ Calories    <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775~
## # A tibble: 940 x 3
##            Id ActivityDay Calories
##         <dbl> <chr>          <dbl>
##  1 1503960366 4/12/2016       1985
##  2 1503960366 4/13/2016       1797
##  3 1503960366 4/14/2016       1776
##  4 1503960366 4/15/2016       1745
##  5 1503960366 4/16/2016       1863
##  6 1503960366 4/17/2016       1728
##  7 1503960366 4/18/2016       1921
##  8 1503960366 4/19/2016       2035
##  9 1503960366 4/20/2016       1786
## 10 1503960366 4/21/2016       1775
## # i 930 more rows
##        Id            ActivityDay           Calories
##  Min.   :1.504e+09   Length:940         Min.   :   0
##  1st Qu.:2.320e+09   Class :character   1st Qu.:1828
##  Median :4.445e+09   Mode  :character   Median :2134
##  Mean   :4.855e+09                      Mean   :2304
##  3rd Qu.:6.962e+09                      3rd Qu.:2793
##  Max.   :8.878e+09                      Max.   :4900
## [1] "Missing values: 0"
```

```
check_data(daily_intensities, "daily_intensities")
```

```
## [1] "Checking daily_intensities :"
## Rows: 940
## Columns: 10
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDay              <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ SedentaryMinutes         <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~
## $ LightlyActiveMinutes     <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~
## $ FairlyActiveMinutes      <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~
## $ VeryActiveMinutes        <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~
## # A tibble: 940 x 10
##         Id ActivityDay SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes
```

```
##      <dbl> <chr>                  <dbl>                <dbl>                <dbl>
##  1  1.50e9 4/12/2016                728                  328                   13
##  2  1.50e9 4/13/2016                776                  217                   19
##  3  1.50e9 4/14/2016               1218                  181                   11
##  4  1.50e9 4/15/2016                726                  209                   34
##  5  1.50e9 4/16/2016                773                  221                   10
##  6  1.50e9 4/17/2016                539                  164                   20
##  7  1.50e9 4/18/2016               1149                  233                   16
##  8  1.50e9 4/19/2016                775                  264                   31
##  9  1.50e9 4/20/2016                818                  205                   12
## 10  1.50e9 4/21/2016                838                  211                    8
## # i 930 more rows
## # i 5 more variables: VeryActiveMinutes <dbl>, SedentaryActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, moderatelyActiveDistance <dbl>,
## #   VeryActiveDistance <dbl>
##        Id            ActivityDay        SedentaryMinutes LightlyActiveMinutes
##  Min.   :1.504e+09   Length:940         Min.   :   0.0   Min.   :  0.0
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 729.8   1st Qu.:127.0
##  Median :4.445e+09   Mode  :character   Median :1057.5   Median :199.0
##  Mean   :4.855e+09                      Mean   : 991.2   Mean   :192.8
##  3rd Qu.:6.962e+09                      3rd Qu.:1229.5   3rd Qu.:264.0
##  Max.   :8.878e+09                      Max.   :1440.0   Max.   :518.0
##  FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
##  Min.   :  0.00      Min.   :  0.00    Min.   :0.000000
##  1st Qu.:  0.00      1st Qu.:  0.00    1st Qu.:0.000000
##  Median :  6.00      Median :  4.00    Median :0.000000
##  Mean   : 13.56      Mean   : 21.16    Mean   :0.001606
##  3rd Qu.: 19.00      3rd Qu.: 32.00    3rd Qu.:0.000000
##  Max.   :143.00      Max.   :210.00    Max.   :0.110000
##  LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
##  Min.   : 0.000      Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 1.945      1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 3.365      Median :0.2400           Median : 0.210
##  Mean   : 3.341      Mean   :0.5675           Mean   : 1.503
##  3rd Qu.: 4.782      3rd Qu.:0.8000           3rd Qu.: 2.053
##  Max.   :10.710      Max.   :6.4800           Max.   :21.920
## [1] "Missing values: 0"
```

```
check_data(daily_steps, "daily_steps")
```

```
## [1] "Checking daily_steps :"
## Rows: 940
## Columns: 3
## $ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366~
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/~
## $ StepTotal   <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019, 15506, 1054~
## # A tibble: 940 x 3
##           Id ActivityDay StepTotal
##        <dbl> <chr>           <dbl>
##  1 1503960366 4/12/2016       13162
##  2 1503960366 4/13/2016       10735
##  3 1503960366 4/14/2016       10460
##  4 1503960366 4/15/2016        9762
##  5 1503960366 4/16/2016       12669
##  6 1503960366 4/17/2016        9705
```

```
##  7 1503960366 4/18/2016       13019
##  8 1503960366 4/19/2016       15506
##  9 1503960366 4/20/2016       10544
## 10 1503960366 4/21/2016        9819
## # i 930 more rows
##        Id            ActivityDay          StepTotal
##  Min.   :1.504e+09   Length:940         Min.   :    0
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790
##  Median :4.445e+09   Mode  :character   Median : 7406
##  Mean   :4.855e+09                      Mean   : 7638
##  3rd Qu.:6.962e+09                      3rd Qu.:10727
##  Max.   :8.878e+09                      Max.   :36019
## [1] "Missing values: 0"
```

```
check_data(sleep_day, "sleep_day")
```

```
## [1] "Checking sleep_day :"
## Rows: 413
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~
## $ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "~
## $ TotalSleepRecords  <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~
## $ TotalTimeInBed     <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~
## # A tibble: 413 x 5
##            Id SleepDay       TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##         <dbl> <chr>                      <dbl>              <dbl>          <dbl>
##  1 1503960366 4/12/2016 12:~                1                327            346
##  2 1503960366 4/13/2016 12:~                2                384            407
##  3 1503960366 4/15/2016 12:~                1                412            442
##  4 1503960366 4/16/2016 12:~                2                340            367
##  5 1503960366 4/17/2016 12:~                1                700            712
##  6 1503960366 4/19/2016 12:~                1                304            320
##  7 1503960366 4/20/2016 12:~                1                360            377
##  8 1503960366 4/21/2016 12:~                1                325            364
##  9 1503960366 4/23/2016 12:~                1                361            384
## 10 1503960366 4/24/2016 12:~                1                430            449
## # i 403 more rows
##        Id              SleepDay         TotalSleepRecords TotalMinutesAsleep
##  Min.   :1.504e+09   Length:413         Min.   :1.000     Min.   : 58.0
##  1st Qu.:3.977e+09   Class :character   1st Qu.:1.000     1st Qu.:361.0
##  Median :4.703e+09   Mode  :character   Median :1.000     Median :433.0
##  Mean   :5.001e+09                      Mean   :1.119     Mean   :419.5
##  3rd Qu.:6.962e+09                      3rd Qu.:1.000     3rd Qu.:490.0
##  Max.   :8.792e+09                      Max.   :3.000     Max.   :796.0
##  TotalTimeInBed
##  Min.   : 61.0
##  1st Qu.:403.0
##  Median :463.0
##  Mean   :458.6
##  3rd Qu.:526.0
##  Max.   :961.0
## [1] "Missing values: 0"
```

```
check_data(weight_log, "weight_log")
```

```
## [1] "Checking weight_log :"
## Rows: 67
## Columns: 8
## $ Id             <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212~
## $ Date           <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2~
## $ WeightKg       <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, ~
## $ WeightPounds   <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6~
## $ Fat            <dbl> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ BMI            <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,~
## $ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
## $ LogId          <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,~
## # A tibble: 67 x 8
##            Id Date      WeightKg WeightPounds   Fat   BMI IsManualReport   LogId
##         <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl> <lgl>            <dbl>
##  1 1503960366 5/2/2016~     52.6         116.    22  22.6 TRUE           1.46e12
##  2 1503960366 5/3/2016~     52.6         116.    NA  22.6 TRUE           1.46e12
##  3 1927972279 4/13/201~    134.          294.    NA  47.5 FALSE          1.46e12
##  4 2873212765 4/21/201~     56.7         125.    NA  21.5 TRUE           1.46e12
##  5 2873212765 5/12/201~     57.3         126.    NA  21.7 TRUE           1.46e12
##  6 4319703577 4/17/201~     72.4         160.    25  27.5 TRUE           1.46e12
##  7 4319703577 5/4/2016~     72.3         159.    NA  27.4 TRUE           1.46e12
##  8 4558609924 4/18/201~     69.7         154.    NA  27.2 TRUE           1.46e12
##  9 4558609924 4/25/201~     70.3         155.    NA  27.5 TRUE           1.46e12
## 10 4558609924 5/1/2016~     69.9         154.    NA  27.3 TRUE           1.46e12
## # i 57 more rows
##        Id                Date              WeightKg       WeightPounds
##  Min.   :1.504e+09   Length:67          Min.   : 52.60   Min.   :116.0
##  1st Qu.:6.962e+09   Class :character   1st Qu.: 61.40   1st Qu.:135.4
##  Median :6.962e+09   Mode  :character   Median : 62.50   Median :137.8
##  Mean   :7.009e+09                      Mean   : 72.04   Mean   :158.8
##  3rd Qu.:8.878e+09                      3rd Qu.: 85.05   3rd Qu.:187.5
##  Max.   :8.878e+09                      Max.   :133.50   Max.   :294.3
##
##       Fat              BMI        IsManualReport      LogId
##  Min.   :22.00   Min.   :21.45   Mode :logical   Min.   :1.460e+12
##  1st Qu.:22.75   1st Qu.:23.96   FALSE:26        1st Qu.:1.461e+12
##  Median :23.50   Median :24.39   TRUE :41        Median :1.462e+12
##  Mean   :23.50   Mean   :25.19                   Mean   :1.462e+12
##  3rd Qu.:24.25   3rd Qu.:25.56                   3rd Qu.:1.462e+12
##  Max.   :25.00   Max.   :47.54                   Max.   :1.463e+12
##  NA's   :65
## [1] "Missing values: 65"
```

```
check_data(heartrate_seconds, "heartrate_seconds")
```

```
## [1] "Checking heartrate_seconds :"
## Rows: 2,483,658
## Columns: 3
## $ Id    <dbl> 2022484408, 2022484408, 2022484408, 2022484408, 2022484408, 2022~
## $ Time  <chr> "4/12/2016 7:21:00 AM", "4/12/2016 7:21:05 AM", "4/12/2016 7:21:~
## $ Value <dbl> 97, 102, 105, 103, 101, 95, 91, 93, 94, 93, 92, 89, 83, 61, 60, ~
## # A tibble: 2,483,658 x 3
```

```
##            Id Time                  Value
##         <dbl> <chr>                 <dbl>
##  1 2022484408 4/12/2016 7:21:00 AM    97
##  2 2022484408 4/12/2016 7:21:05 AM   102
##  3 2022484408 4/12/2016 7:21:10 AM   105
##  4 2022484408 4/12/2016 7:21:20 AM   103
##  5 2022484408 4/12/2016 7:21:25 AM   101
##  6 2022484408 4/12/2016 7:22:05 AM    95
##  7 2022484408 4/12/2016 7:22:10 AM    91
##  8 2022484408 4/12/2016 7:22:15 AM    93
##  9 2022484408 4/12/2016 7:22:20 AM    94
## 10 2022484408 4/12/2016 7:22:25 AM    93
## # i 2,483,648 more rows
##        Id                Time              Value
##  Min.   :2.022e+09   Length:2483658     Min.   : 36.00
##  1st Qu.:4.388e+09   Class :character   1st Qu.: 63.00
##  Median :5.554e+09   Mode  :character   Median : 73.00
##  Mean   :5.514e+09                      Mean   : 77.33
##  3rd Qu.:6.962e+09                      3rd Qu.: 88.00
##  Max.   :8.878e+09                      Max.   :203.00
## [1] "Missing values: 0"
```

```
check_data(hourly_calories, "hourly_calories")
```

```
## [1] "Checking hourly_calories :"
## Rows: 22,099
## Columns: 3
## $ Id           <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityHour <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/12/20~
## $ Calories     <dbl> 81, 61, 59, 47, 48, 48, 48, 47, 68, 141, 99, 76, 73, 66, ~
## # A tibble: 22,099 x 3
##            Id ActivityHour          Calories
##         <dbl> <chr>                    <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM       81
##  2 1503960366 4/12/2016 1:00:00 AM        61
##  3 1503960366 4/12/2016 2:00:00 AM        59
##  4 1503960366 4/12/2016 3:00:00 AM        47
##  5 1503960366 4/12/2016 4:00:00 AM        48
##  6 1503960366 4/12/2016 5:00:00 AM        48
##  7 1503960366 4/12/2016 6:00:00 AM        48
##  8 1503960366 4/12/2016 7:00:00 AM        47
##  9 1503960366 4/12/2016 8:00:00 AM        68
## 10 1503960366 4/12/2016 9:00:00 AM       141
## # i 22,089 more rows
##        Id            ActivityHour          Calories
##  Min.   :1.504e+09   Length:22099       Min.   : 42.00
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 63.00
##  Median :4.445e+09   Mode  :character   Median : 83.00
##  Mean   :4.848e+09                      Mean   : 97.39
##  3rd Qu.:6.962e+09                      3rd Qu.:108.00
##  Max.   :8.878e+09                      Max.   :948.00
## [1] "Missing values: 0"
```

```r
check_data(hourly_intensities, "hourly_intensities")
```

```
## [1] "Checking hourly_intensities :"
## Rows: 22,099
## Columns: 4
## $ Id               <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 15039~
## $ ActivityHour     <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/1~
## $ TotalIntensity   <dbl> 20, 8, 7, 0, 0, 0, 0, 0, 13, 30, 29, 12, 11, 6, 36, 5~
## $ AverageIntensity <dbl> 0.333333, 0.133333, 0.116667, 0.000000, 0.000000, 0.0~
## # A tibble: 22,099 x 4
##            Id ActivityHour          TotalIntensity AverageIntensity
##         <dbl> <chr>                          <dbl>            <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM             20            0.333
##  2 1503960366 4/12/2016 1:00:00 AM               8            0.133
##  3 1503960366 4/12/2016 2:00:00 AM               7            0.117
##  4 1503960366 4/12/2016 3:00:00 AM               0            0
##  5 1503960366 4/12/2016 4:00:00 AM               0            0
##  6 1503960366 4/12/2016 5:00:00 AM               0            0
##  7 1503960366 4/12/2016 6:00:00 AM               0            0
##  8 1503960366 4/12/2016 7:00:00 AM               0            0
##  9 1503960366 4/12/2016 8:00:00 AM              13            0.217
## 10 1503960366 4/12/2016 9:00:00 AM              30            0.5
## # i 22,089 more rows
##        Id            ActivityHour       TotalIntensity   AverageIntensity
##  Min.   :1.504e+09   Length:22099       Min.   :  0.00   Min.   :0.0000
##  1st Qu.:2.320e+09   Class :character   1st Qu.:  0.00   1st Qu.:0.0000
##  Median :4.445e+09   Mode  :character   Median :  3.00   Median :0.0500
##  Mean   :4.848e+09                      Mean   : 12.04   Mean   :0.2006
##  3rd Qu.:6.962e+09                      3rd Qu.: 16.00   3rd Qu.:0.2667
##  Max.   :8.878e+09                      Max.   :180.00   Max.   :3.0000
## [1] "Missing values: 0"
```

```r
check_data(hourly_steps, "hourly_steps")
```

```
## [1] "Checking hourly_steps :"
## Rows: 22,099
## Columns: 3
## $ Id           <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityHour <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/12/20~
## $ StepTotal    <dbl> 373, 160, 151, 0, 0, 0, 0, 0, 250, 1864, 676, 360, 253, 2~
## # A tibble: 22,099 x 3
##            Id ActivityHour          StepTotal
##         <dbl> <chr>                     <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM       373
##  2 1503960366 4/12/2016 1:00:00 AM        160
##  3 1503960366 4/12/2016 2:00:00 AM        151
##  4 1503960366 4/12/2016 3:00:00 AM          0
##  5 1503960366 4/12/2016 4:00:00 AM          0
##  6 1503960366 4/12/2016 5:00:00 AM          0
##  7 1503960366 4/12/2016 6:00:00 AM          0
##  8 1503960366 4/12/2016 7:00:00 AM          0
##  9 1503960366 4/12/2016 8:00:00 AM        250
## 10 1503960366 4/12/2016 9:00:00 AM       1864
## # i 22,089 more rows
```

```
##        Id              ActivityHour         StepTotal
##  Min.   :1.504e+09   Length:22099       Min.   :    0.0
##  1st Qu.:2.320e+09   Class :character   1st Qu.:    0.0
##  Median :4.445e+09   Mode  :character   Median :   40.0
##  Mean   :4.848e+09                      Mean   :  320.2
##  3rd Qu.:6.962e+09                      3rd Qu.:  357.0
##  Max.   :8.878e+09                      Max.   :10554.0
## [1] "Missing values: 0"
```

```
check_data(minute_calories_narrow, "minute_calories_narrow")
```

```
## [1] "Checking minute_calories_narrow :"
## Rows: 1,325,580
## Columns: 3
## $ Id             <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960~
## $ ActivityMinute <chr> "4/12/2016 12:00:00 AM", "4/12/2016 12:01:00 AM", "4/12~
## $ Calories       <dbl> 0.7865, 0.7865, 0.7865, 0.7865, 0.7865, 0.9438, 0.9438,~
## # A tibble: 1,325,580 x 3
##            Id ActivityMinute        Calories
##         <dbl> <chr>                    <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM    0.786
##  2 1503960366 4/12/2016 12:01:00 AM    0.786
##  3 1503960366 4/12/2016 12:02:00 AM    0.786
##  4 1503960366 4/12/2016 12:03:00 AM    0.786
##  5 1503960366 4/12/2016 12:04:00 AM    0.786
##  6 1503960366 4/12/2016 12:05:00 AM    0.944
##  7 1503960366 4/12/2016 12:06:00 AM    0.944
##  8 1503960366 4/12/2016 12:07:00 AM    0.944
##  9 1503960366 4/12/2016 12:08:00 AM    0.944
## 10 1503960366 4/12/2016 12:09:00 AM    0.944
## # i 1,325,570 more rows
##        Id              ActivityMinute        Calories
##  Min.   :1.504e+09   Length:1325580     Min.   : 0.0000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 0.9357
##  Median :4.445e+09   Mode  :character   Median : 1.2176
##  Mean   :4.848e+09                      Mean   : 1.6231
##  3rd Qu.:6.962e+09                      3rd Qu.: 1.4327
##  Max.   :8.878e+09                      Max.   :19.7499
## [1] "Missing values: 0"
```

```
check_data(minute_intensities_narrow, "minute_intensities_narrow")
```

```
## [1] "Checking minute_intensities_narrow :"
## Rows: 1,325,580
## Columns: 3
## $ Id             <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960~
## $ ActivityMinute <chr> "4/12/2016 12:00:00 AM", "4/12/2016 12:01:00 AM", "4/12~
## $ Intensity      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## # A tibble: 1,325,580 x 3
##            Id ActivityMinute        Intensity
##         <dbl> <chr>                    <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM         0
##  2 1503960366 4/12/2016 12:01:00 AM         0
##  3 1503960366 4/12/2016 12:02:00 AM         0
##  4 1503960366 4/12/2016 12:03:00 AM         0
```

```
##  5 1503960366 4/12/2016 12:04:00 AM         0
##  6 1503960366 4/12/2016 12:05:00 AM         0
##  7 1503960366 4/12/2016 12:06:00 AM         0
##  8 1503960366 4/12/2016 12:07:00 AM         0
##  9 1503960366 4/12/2016 12:08:00 AM         0
## 10 1503960366 4/12/2016 12:09:00 AM         0
## # i 1,325,570 more rows
##        Id            ActivityMinute       Intensity
##  Min.   :1.504e+09   Length:1325580     Min.   :0.0000
##  1st Qu.:2.320e+09   Class :character   1st Qu.:0.0000
##  Median :4.445e+09   Mode  :character   Median :0.0000
##  Mean   :4.848e+09                      Mean   :0.2006
##  3rd Qu.:6.962e+09                      3rd Qu.:0.0000
##  Max.   :8.878e+09                      Max.   :3.0000
## [1] "Missing values: 0"
```

```
check_data(minute_mets_narrow, "minute_mets_narrow")
```

```
## [1] "Checking minute_mets_narrow :"
## Rows: 1,325,580
## Columns: 3
## $ Id             <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960~
## $ ActivityMinute <chr> "4/12/2016 12:00:00 AM", "4/12/2016 12:01:00 AM", "4/12~
## $ METs           <dbl> 10, 10, 10, 10, 10, 12, 12, 12, 12, 12, 12, 12, 10, 10,~
## # A tibble: 1,325,580 x 3
##            Id ActivityMinute         METs
##         <dbl> <chr>                <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM    10
##  2 1503960366 4/12/2016 12:01:00 AM    10
##  3 1503960366 4/12/2016 12:02:00 AM    10
##  4 1503960366 4/12/2016 12:03:00 AM    10
##  5 1503960366 4/12/2016 12:04:00 AM    10
##  6 1503960366 4/12/2016 12:05:00 AM    12
##  7 1503960366 4/12/2016 12:06:00 AM    12
##  8 1503960366 4/12/2016 12:07:00 AM    12
##  9 1503960366 4/12/2016 12:08:00 AM    12
## 10 1503960366 4/12/2016 12:09:00 AM    12
## # i 1,325,570 more rows
##        Id            ActivityMinute          METs
##  Min.   :1.504e+09   Length:1325580     Min.   :  0.00
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 10.00
##  Median :4.445e+09   Mode  :character   Median : 10.00
##  Mean   :4.848e+09                      Mean   : 14.69
##  3rd Qu.:6.962e+09                      3rd Qu.: 11.00
##  Max.   :8.878e+09                      Max.   :157.00
## [1] "Missing values: 0"
```

```
check_data(minute_sleep, "minute_sleep")
```

```
## [1] "Checking minute_sleep :"
## Rows: 188,521
## Columns: 4
## $ Id    <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366, 1503~
## $ date  <chr> "4/12/2016 2:47:30 AM", "4/12/2016 2:48:30 AM", "4/12/2016 2:49:~
## $ value <dbl> 3, 2, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 2, 1, 1, 1, 1, 1, 1~
```

```
## $ logId <dbl> 11380564589, 11380564589, 11380564589, 11380564589, 11380564589,~
## # A tibble: 188,521 x 4
##            Id date                 value       logId
##         <dbl> <chr>                <dbl>       <dbl>
##  1 1503960366 4/12/2016 2:47:30 AM     3 11380564589
##  2 1503960366 4/12/2016 2:48:30 AM     2 11380564589
##  3 1503960366 4/12/2016 2:49:30 AM     1 11380564589
##  4 1503960366 4/12/2016 2:50:30 AM     1 11380564589
##  5 1503960366 4/12/2016 2:51:30 AM     1 11380564589
##  6 1503960366 4/12/2016 2:52:30 AM     1 11380564589
##  7 1503960366 4/12/2016 2:53:30 AM     1 11380564589
##  8 1503960366 4/12/2016 2:54:30 AM     2 11380564589
##  9 1503960366 4/12/2016 2:55:30 AM     2 11380564589
## 10 1503960366 4/12/2016 2:56:30 AM     2 11380564589
## # i 188,511 more rows
##        Id                date              value           logId
##  Min.   :1.504e+09   Length:188521      Min.   :1.000   Min.   :1.137e+10
##  1st Qu.:3.977e+09   Class :character   1st Qu.:1.000   1st Qu.:1.144e+10
##  Median :4.703e+09   Mode  :character   Median :1.000   Median :1.150e+10
##  Mean   :4.997e+09                      Mean   :1.096   Mean   :1.150e+10
##  3rd Qu.:6.962e+09                      3rd Qu.:1.000   3rd Qu.:1.155e+10
##  Max.   :8.792e+09                      Max.   :3.000   Max.   :1.162e+10
## [1] "Missing values: 0"
```

```
check_data(minute_steps_narrow, "minute_steps_narrow")
```

```
## [1] "Checking minute_steps_narrow :"
## Rows: 1,325,580
## Columns: 3
## $ Id             <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960~
## $ ActivityMinute <chr> "4/12/2016 12:00:00 AM", "4/12/2016 12:01:00 AM", "4/12~
## $ Steps          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## # A tibble: 1,325,580 x 3
##            Id ActivityMinute        Steps
##         <dbl> <chr>                 <dbl>
##  1 1503960366 4/12/2016 12:00:00 AM     0
##  2 1503960366 4/12/2016 12:01:00 AM     0
##  3 1503960366 4/12/2016 12:02:00 AM     0
##  4 1503960366 4/12/2016 12:03:00 AM     0
##  5 1503960366 4/12/2016 12:04:00 AM     0
##  6 1503960366 4/12/2016 12:05:00 AM     0
##  7 1503960366 4/12/2016 12:06:00 AM     0
##  8 1503960366 4/12/2016 12:07:00 AM     0
##  9 1503960366 4/12/2016 12:08:00 AM     0
## 10 1503960366 4/12/2016 12:09:00 AM     0
## # i 1,325,570 more rows
##        Id            ActivityMinute         Steps
##  Min.   :1.504e+09   Length:1325580     Min.   :  0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.:  0.000
##  Median :4.445e+09   Mode  :character   Median :  0.000
##  Mean   :4.848e+09                      Mean   :  5.336
##  3rd Qu.:6.962e+09                      3rd Qu.:  0.000
##  Max.   :8.878e+09                      Max.   :220.000
## [1] "Missing values: 0"
```

### 3.4 Data Cleaning & Transformation

**3.4.1 Convert date columns to appropriate format** Notes: Just to be sure all dates are the same format

```
# Convert date/time columns to appropriate formats
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format = "%m/%d/%Y")
daily_calories$ActivityDay <- as.Date(daily_calories$ActivityDay, format = "%m/%d/%Y")
daily_intensities$ActivityDay <- as.Date(daily_intensities$ActivityDay, format = "%m/%d/%Y")
daily_steps$ActivityDay <- as.Date(daily_steps$ActivityDay, format = "%m/%d/%Y")
sleep_day$SleepDay <- as.Date(sleep_day$SleepDay, format = "%m/%d/%Y")
weight_log$Date <- as.Date(weight_log$Date, format = "%m/%d/%Y")

heartrate_seconds$Time <- mdy_hms(heartrate_seconds$Time)
hourly_calories$ActivityHour <- mdy_hms(hourly_calories$ActivityHour)
hourly_intensities$ActivityHour <- mdy_hms(hourly_intensities$ActivityHour)
hourly_steps$ActivityHour <- mdy_hms(hourly_steps$ActivityHour)
minute_calories_narrow$ActivityMinute <- mdy_hms(minute_calories_narrow$ActivityMinute)
minute_intensities_narrow$ActivityMinute <- mdy_hms(minute_intensities_narrow$ActivityMinute)
minute_sleep$date <- mdy_hms(minute_sleep$date)
minute_steps_narrow$ActivityMinute <- mdy_hms(minute_steps_narrow$ActivityMinute)
```

**3.4.2 Merge daily data**

```
# Load the data (as you've already done)

# 1. Inspect the column names of EACH dataframe *before* any joins
print(names(daily_activity))
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

```
print(names(daily_calories))
```

```
## [1] "Id"          "ActivityDay" "Calories"
```

```
print(names(daily_intensities))
```

```
##  [1] "Id"                       "ActivityDay"
##  [3] "SedentaryMinutes"         "LightlyActiveMinutes"
##  [5] "FairlyActiveMinutes"      "VeryActiveMinutes"
##  [7] "SedentaryActiveDistance"  "LightActiveDistance"
##  [9] "ModeratelyActiveDistance" "VeryActiveDistance"
```

```
print(names(daily_steps))
```

```
## [1] "Id"          "ActivityDay" "StepTotal"
```

```
print(names(sleep_day))
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

```r
print(names(weight_log))
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

```r
print(names(heartrate_seconds))
```

```
## [1] "Id"    "Time"  "Value"
```

```r
print(names(hourly_calories))
```

```
## [1] "Id"           "ActivityHour" "Calories"
```

```r
print(names(hourly_intensities))
```

```
## [1] "Id"               "ActivityHour"     "TotalIntensity"   "AverageIntensity"
```

```r
print(names(hourly_steps))
```

```
## [1] "Id"           "ActivityHour" "StepTotal"
```

```r
print(names(minute_calories_narrow))
```

```
## [1] "Id"             "ActivityMinute" "Calories"
```

```r
print(names(minute_intensities_narrow))
```

```
## [1] "Id"             "ActivityMinute" "Intensity"
```

```r
print(names(minute_sleep))
```

```
## [1] "Id"    "date"  "value" "logId"
```

```r
print(names(minute_steps_narrow))
```

```
## [1] "Id"             "ActivityMinute" "Steps"
```

```r
# 1. Rename columns in `daily_intensities` to avoid conflicts
daily_intensities <- daily_intensities %>%
  rename(
    VeryActiveDistance_intensities = VeryActiveDistance,
    ModeratelyActiveDistance_intensities = ModeratelyActiveDistance,
    LightActiveDistance_intensities = LightActiveDistance,
    SedentaryActiveDistance_intensities = SedentaryActiveDistance,
    VeryActiveMinutes_intensities = VeryActiveMinutes,
    FairlyActiveMinutes_intensities = FairlyActiveMinutes,
    LightlyActiveMinutes_intensities = LightlyActiveMinutes,
    SedentaryMinutes_intensities = SedentaryMinutes
  )

# 2. Select the *exact* columns you need *before* merging. This is much more efficient.
daily_data <- daily_activity %>%
  left_join(daily_calories, by = c("Id", "ActivityDate" = "ActivityDay")) %>%
  left_join(daily_intensities, by = c("Id", "ActivityDate" = "ActivityDay")) %>%
  left_join(daily_steps, by = c("Id", "ActivityDate" = "ActivityDay")) %>%
  select(-Calories.y) %>%
  rename(Calories = Calories.x) %>%
  filter(!is.na(TotalSteps))

merged_data <- sleep_day %>%
```

```
    rename(SleepDate = SleepDay) %>%
    left_join(daily_data, by = c("Id", "SleepDate" = "ActivityDate"))

print(names(merged_data))
```

```
##  [1] "Id"
##  [2] "SleepDate"
##  [3] "TotalSleepRecords"
##  [4] "TotalMinutesAsleep"
##  [5] "TotalTimeInBed"
##  [6] "TotalSteps"
##  [7] "TotalDistance"
##  [8] "TrackerDistance"
##  [9] "LoggedActivitiesDistance"
## [10] "VeryActiveDistance"
## [11] "ModeratelyActiveDistance"
## [12] "LightActiveDistance"
## [13] "SedentaryActiveDistance"
## [14] "VeryActiveMinutes"
## [15] "FairlyActiveMinutes"
## [16] "LightlyActiveMinutes"
## [17] "SedentaryMinutes"
## [18] "Calories"
## [19] "SedentaryMinutes_intensities"
## [20] "LightlyActiveMinutes_intensities"
## [21] "FairlyActiveMinutes_intensities"
## [22] "VeryActiveMinutes_intensities"
## [23] "SedentaryActiveDistance_intensities"
## [24] "LightActiveDistance_intensities"
## [25] "ModeratelyActiveDistance_intensities"
## [26] "VeryActiveDistance_intensities"
## [27] "StepTotal"
```

```
# 1. Check if "ActivityDate" is present.  If not, look for similar names.
if ("ActivityDate" %in% names(merged_data)) {
    # ActivityDate is there, proceed with debugging.
    print("ActivityDate found. Proceeding with debugging.")

    # ... (Your existing debugging code using ActivityDate) ...

} else if ("ActivityDay" %in% names(merged_data)) {  # Example: Check for ActivityDay
    # ActivityDate is missing, but ActivityDay is present. Rename it.
    print("ActivityDate NOT found. But ActivityDay found. Renaming...")
    merged_data <- merged_data %>%
        rename(ActivityDate = ActivityDay)  # Rename ActivityDay to ActivityDate

    # ... (Your debugging code, now using the renamed ActivityDate) ...

} else if ("SleepDate" %in% names(merged_data)){
    print("ActivityDate and ActivityDay NOT found. But SleepDate found. Renaming...")
    merged_data <- merged_data %>%
      rename(ActivityDate = SleepDate)

    # ... (Your debugging code, now using the renamed ActivityDate) ...
```

```
} else {
    # Neither ActivityDate nor ActivityDay is found. Serious problem!
    print("ERROR: Neither ActivityDate nor ActivityDay found in merged_data!")
    # Stop execution or take other corrective action.
    stop("Critical error: Missing date column.")  # Stop execution
}
```

```
## [1] "ActivityDate and ActivityDay NOT found. But SleepDate found. Renaming..."
```

```
# Now calculate Weekday (AFTER ensuring ActivityDate exists)
merged_data$Weekday <- wday(merged_data$ActivityDate, label = TRUE)
```

```
# Calculate overall averages (as before)
overall_averages <- merged_data %>%
  summarize(
    overall_avg_steps = mean(TotalSteps, na.rm = TRUE),
    overall_avg_calories = mean(Calories, na.rm = TRUE),
    overall_avg_sedentary = mean(SedentaryMinutes, na.rm = TRUE),
    overall_avg_active = mean(VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes, na.rm = TR
    overall_avg_sleep = mean(TotalMinutesAsleep, na.rm = TRUE)
  )

print(overall_averages)
```

```
## # A tibble: 1 x 5
##   overall_avg_steps overall_avg_calories overall_avg_sedentary
##               <dbl>                <dbl>                 <dbl>
## 1             8541.                2398.                  712.
## # i 2 more variables: overall_avg_active <dbl>, overall_avg_sleep <dbl>
```

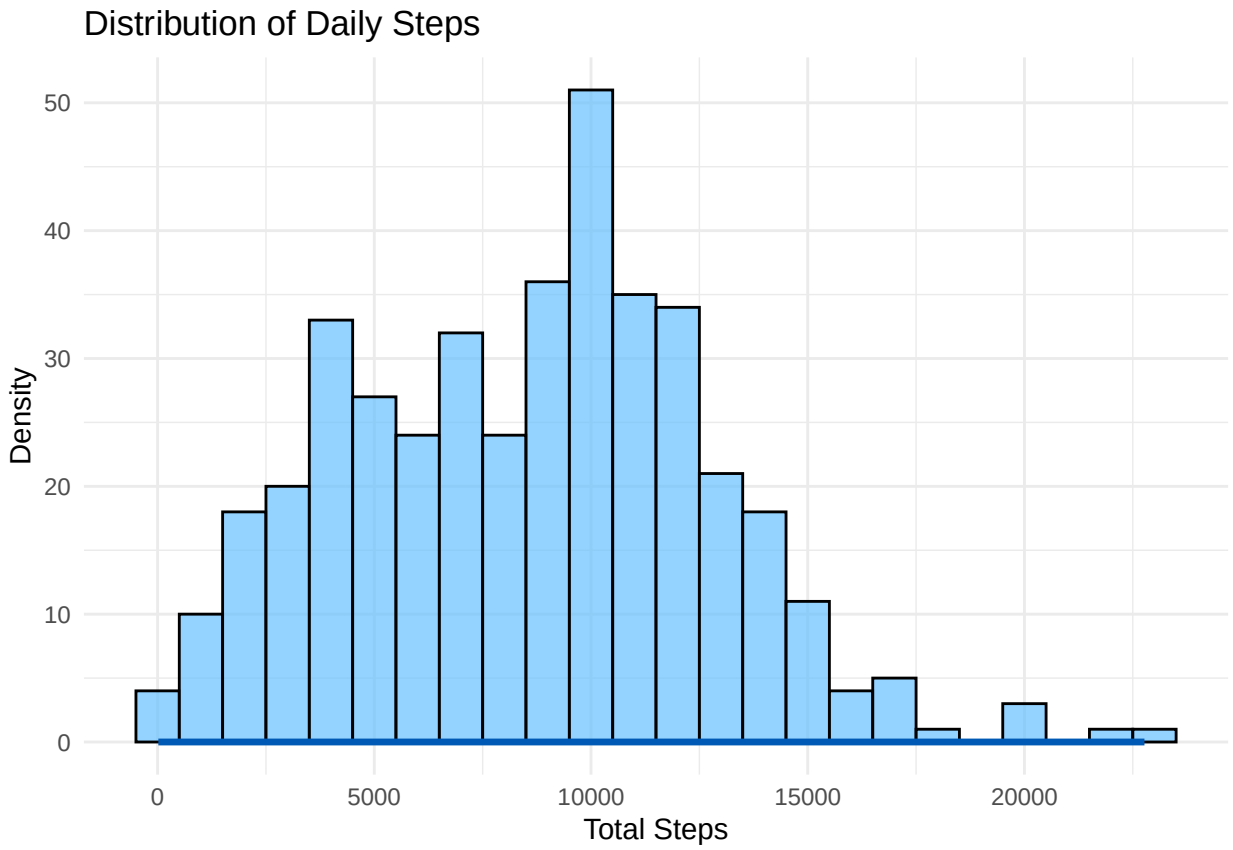
## 4. Analyze

### 4.1 Distribution of Daily Steps

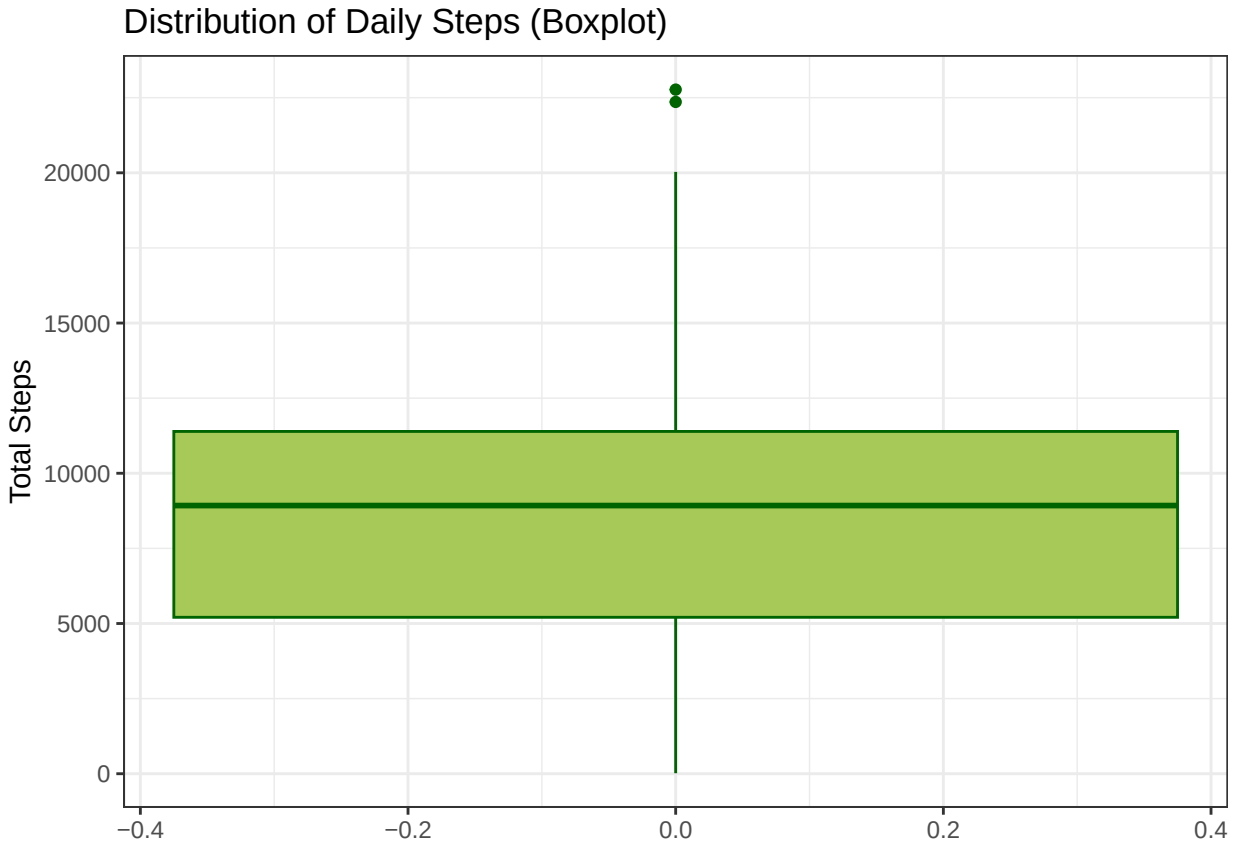
This report analyzes the distribution of daily step counts to understand user activity levels and inform Bellabeat's product and marketing strategies.

**Data and Methods:** We examined the distribution of total daily steps using a histogram and a boxplot. The histogram visualizes the frequency of different step counts, while the boxplot provides a summary of key statistics, including the median, quartiles, and outliers.

```
# 1. Distribution of Daily Steps (Histogram & Boxplot)
ggplot(merged_data, aes(x = TotalSteps)) +
  geom_histogram(binwidth = 1000, fill = "#66c0ff", color = "black", alpha = 0.7) +
  geom_density(aes(y = after_stat(density)), color = "#005ab5", linewidth = 1.2) + # Corrected!
  labs(title = "Distribution of Daily Steps", x = "Total Steps", y = "Density") +
  theme_minimal()
```




```r
ggplot(merged_data, aes(y = TotalSteps)) +
  geom_boxplot(fill = "#a7c957", color = "darkgreen") +
  labs(title = "Distribution of Daily Steps (Boxplot)", y = "Total Steps") +
  theme_bw()
```




**Key Findings** Our analysis reveals the following about daily step counts:

- **Varying Step Counts:** Daily step counts vary widely across users in the dataset. The range spans from a few hundred steps to over 20,000 steps, highlighting diverse activity levels.
- **Central Tendency:** While there is variation, a significant portion of the data clusters around the 10,000-12,000 step range. This likely represents a common activity level amongst the users in this data set.
- **Potential for Improvement:** The data shows that a substantial number of users take fewer than 10,000 steps per day, which is often considered a benchmark for healthy activity. This suggests a significant opportunity to encourage users to increase their step counts.
- **Outliers:** The boxplot identifies several outliers, representing unusually high or low step counts. These outliers may indicate highly active individuals, measurement errors, or unique circumstances on those particular days. Further investigation into these outliers could provide valuable insights.

**Implications for Bellabeat** The observed distribution of daily steps has several important implications for Bellabeat:

- **Targeted Interventions:** The data highlights the need for personalized interventions. Users with lower step counts could benefit from tailored encouragement and strategies for increasing their activity. The Leaf could be used to set personalized step goals and provide reminders or motivation.
- **Motivational Features:** Bellabeat should focus on developing features that motivate users to be more active. Gamification, social challenges, and progress tracking can be effective tools.
- **Marketing Strategies:** Marketing campaigns can emphasize the health benefits of regular physical activity and promote the Leaf as a tool for achieving activity goals. Messaging should be tailored to different user segments, acknowledging the variety in activity levels.

- **Product Development:** Bellabeat could explore integrating the Leaf with other health and fitness apps to provide a more holistic view of user activity. Features that track different types of activity beyond steps, such as cycling, swimming, or strength training, could also be valuable.

### 4.2 Steps vs. Calories Burned

This report examines the relationship between daily step count and calories burned to understand how these metrics correlate and inform Bellabeat's strategies.

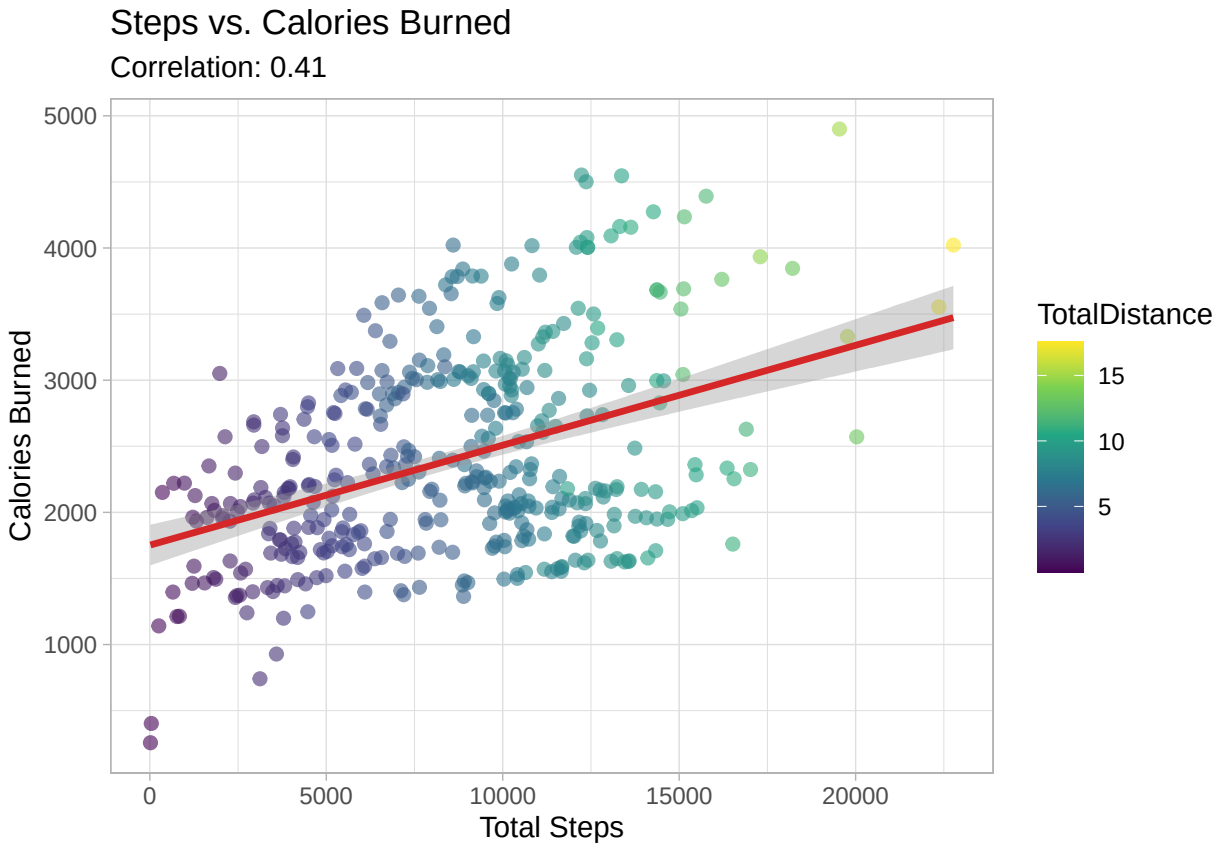
#### Data and Methods:

We analyzed data on total daily steps, calories burned, and total distance. A scatter plot was used to visualize the relationship between steps and calories burned, with total distance represented by color intensity. A trend line was added to indicate any potential correlation.

```
# 2. Steps vs. Calories Burned
merged_data_for_correlation <- merged_data %>%
  filter(!is.na(TotalSteps), !is.na(Calories), is.finite(TotalSteps), is.finite(Calories))

ggplot(merged_data, aes(x = TotalSteps, y = Calories)) +  # Plot on the original data
  geom_point(alpha = 0.6, size = 2, aes(color = TotalDistance)) +
  geom_smooth(method = "lm", color = "#d62728", linewidth = 1.2, data = merged_data_for_correlation) + 
  scale_color_viridis_c(option = "D") +
  labs(title = "Steps vs. Calories Burned", x = "Total Steps", y = "Calories Burned",
       subtitle = paste("Correlation:", round(cor(merged_data_for_correlation$TotalSteps, merged_data_f
  theme_light()
```

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



**Key Findings:** Our analysis reveals the following regarding the relationship between steps and calories burned:

- **Positive Correlation:** The scatter plot and trend line demonstrate a positive correlation between total steps and calories burned. As step count increases, calorie burn also tends to increase. This is an expected and logical relationship.
- **Strength of Correlation:** The subtitle of the plot indicates the correlation coefficient, which quantifies the strength of the linear relationship. A value closer to 1 indicates a stronger positive correlation. The observed correlation suggests a reasonably strong relationship, though not perfect.
- **Influence of Distance:** The color gradient representing total distance reinforces the relationship. As both steps and distance increase, calorie burn also tends to increase. This suggests that distance walked is a contributing factor to calorie expenditure.
- **Variability:** While a positive trend is clear, there's still variability in calorie burn at any given step count. This indicates that other factors (e.g., intensity of activity, individual metabolism, terrain) also influence calorie expenditure.

**Implications for Bellabeat:** The positive correlation between steps and calories burned has several implications for Bellabeat:

- **Reinforcing Step Goals:** The data supports the promotion of increasing step counts as a way to increase calorie expenditure. The Leaf can be positioned as a tool to help users track both steps and estimated calorie burn.
- **Holistic Activity Tracking:** The inclusion of total distance emphasizes the importance of considering various aspects of activity. Bellabeat could further explore incorporating more detailed activity tracking (e.g., different activity types, intensity levels) to provide a more comprehensive picture of calorie expenditure.
- **Personalized Insights:** While the general trend is clear, individual variations exist. The Leaf could provide personalized insights and recommendations based on individual step, distance, and calorie burn patterns.
- **Motivational Strategies:** Marketing campaigns can highlight the link between steps, distance, and calorie burn to motivate users to be more active. The Leaf can be positioned as a tool for achieving both step and calorie goals.
- **Further Research:** Further research could explore the influence of other factors (e.g., age, weight, fitness level, type of activity) on the relationship between steps and calories burned to provide more accurate and personalized calorie estimations.

### 4.3 Analysis of sleep Time Distribution

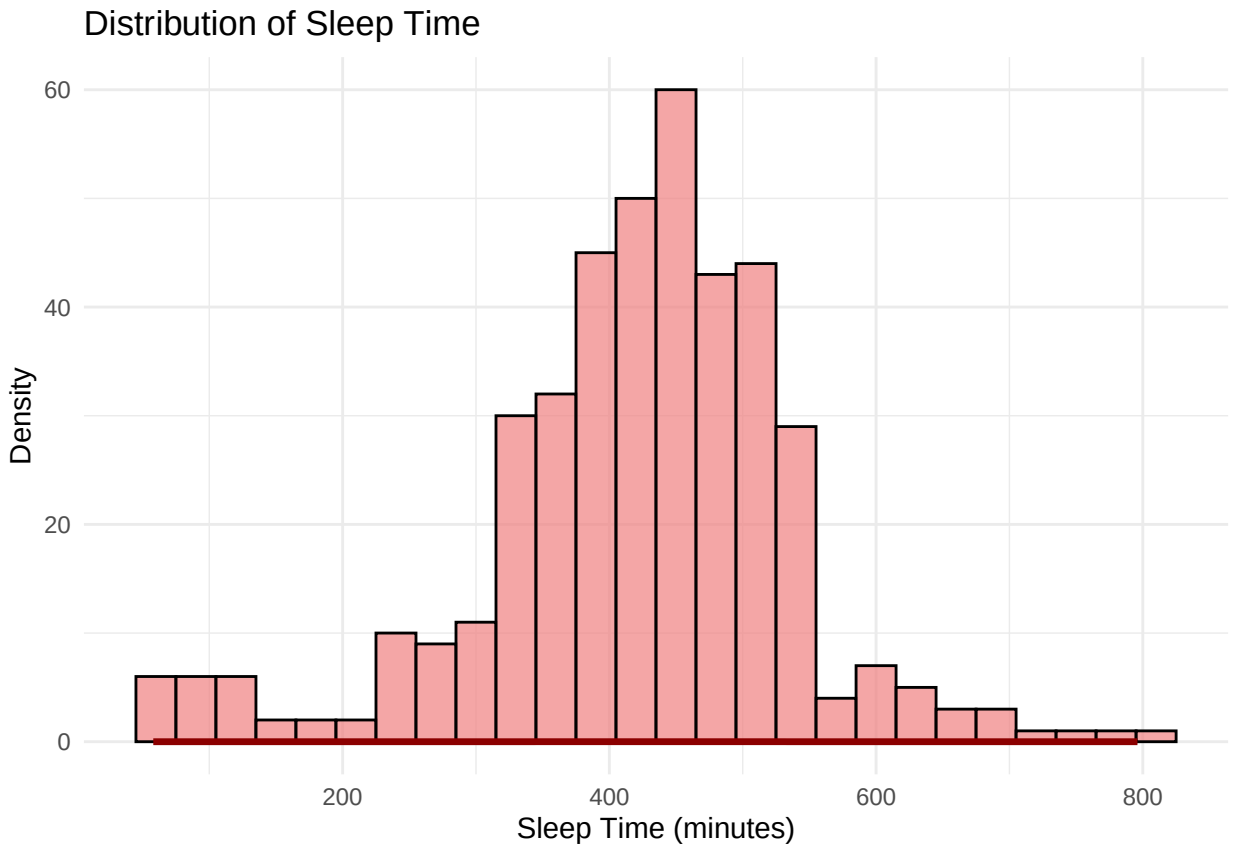
This report presents an analysis of sleep time data collected from fitness trackers. Our goal was to understand sleep patterns and provide insights relevant to Bellabeat's Leaf product.

**Data and Methods:** We analyzed data on total minutes asleep. We used a histogram and a boxplot to visualize the distribution of sleep times. The histogram shows the frequency of different sleep durations, while the boxplot summarizes key statistics like the median, interquartile range (IQR), and outliers.

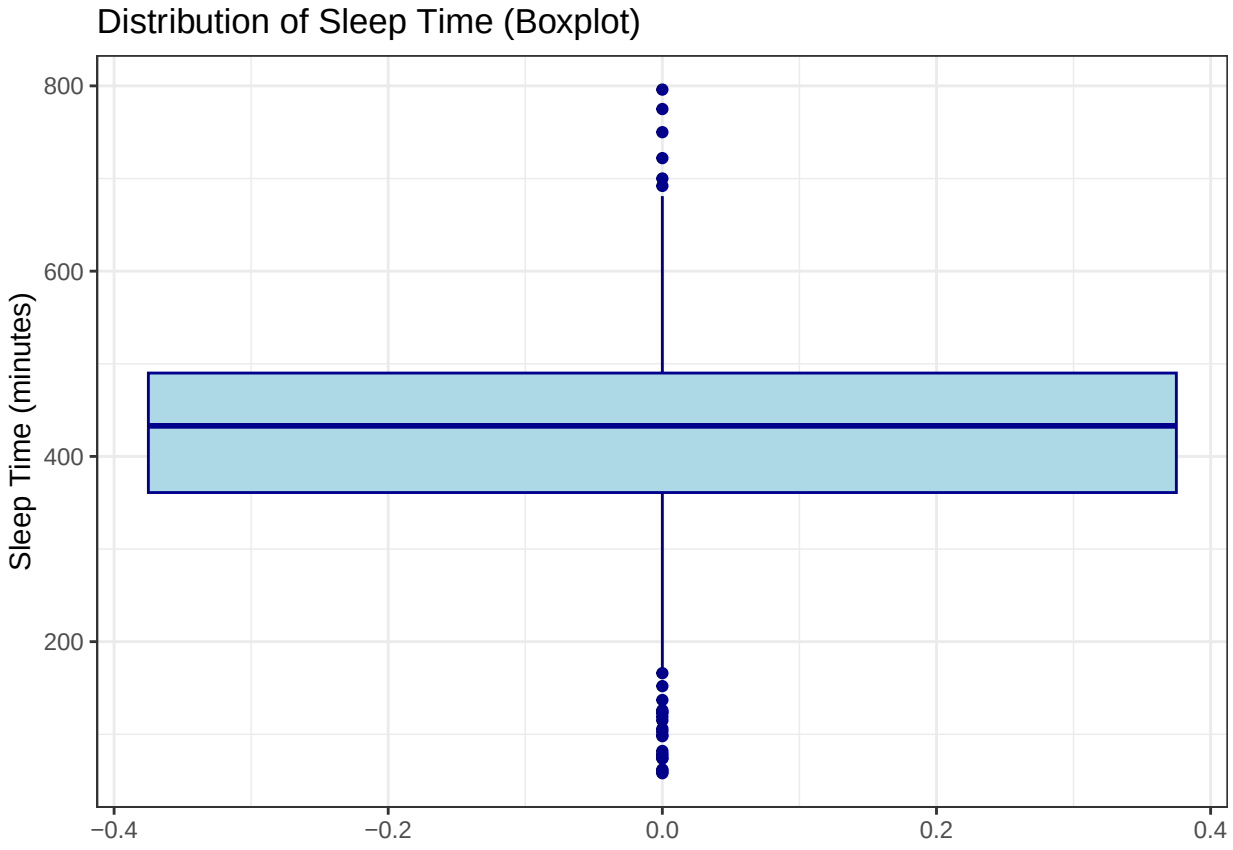
```
# 3. Distribution of Sleep Time
merged_data_sleep_clean <- merged_data %>%
  filter(!is.na(TotalMinutesAsleep), is.finite(TotalMinutesAsleep))

ggplot(merged_data_sleep_clean, aes(x = TotalMinutesAsleep)) +
  geom_histogram(binwidth = 30, fill = "#f08080", color = "black", alpha = 0.7) +
  geom_density(aes(y = after_stat(density)), color = "#8b0000", linewidth = 1.2) + # Corrected!
```

```
  labs(title = "Distribution of Sleep Time", x = "Sleep Time (minutes)", y = "Density") +
  theme_minimal()
```




```
ggplot(merged_data_sleep_clean, aes(y = TotalMinutesAsleep)) + # Use the filtered data
  geom_boxplot(fill = "#add8e6", color = "darkblue") +
  labs(title = "Distribution of Sleep Time (Boxplot)", y = "Sleep Time (minutes)") +
  theme_bw()
```




**Key Findings:** The distribution of sleep time reveals the following:

- **Range and Central Tendency:** Sleep times in the dataset range from a minimum of 58 minutes to a maximum of 796 minutes (approximately 13 hours). However, the majority of users sleep between 300 and 600 minutes (5 to 10 hours), as indicated by the interquartile range (IQR) in the boxplot.
- **Typical Sleep Duration:** While there is variation, a significant portion of the data clusters around the 400-500 minute range (6.6 to 8.3 hours), which likely represents the typical sleep duration for many users in this dataset.
- **Variability:** There's significant variability in sleep duration, as shown by the spread of the histogram and the IQR in the boxplot. Some users sleep considerably less than the average, while others sleep considerably more.
- **Outliers:** The boxplot identifies some outliers, representing unusually short or long sleep durations. These could be due to various factors, including measurement errors, individual sleep patterns, or specific circumstances on those days. Further investigation might be needed to understand the reasons behind these outliers.

**Implications for Bellabeat:** The observed distribution of sleep times has several implications for Bellabeat:

- **Target Audience:** The data reinforces the importance of targeting a broad range of sleep habits. While a substantial portion of users get a "typical" amount of sleep, there are significant segments who sleep less or more.
- **Product Features:** Bellabeat should emphasize the Leaf's ability to track and analyze sleep across this full spectrum. Features that provide personalized insights and recommendations for improving sleep quality, regardless of duration, are crucial.

- **Marketing Strategies:** Marketing campaigns should acknowledge the variability in sleep patterns and promote the Leaf as a tool for understanding and optimizing individual sleep. Messaging could focus on the importance of consistent sleep tracking and personalized feedback. Highlighting the Leaf's ability to identify trends and potential sleep issues, regardless of how much someone sleeps, will resonate with a wider audience.

### 4.4 Average Daily Activity Levels

This report analyzes average daily activity levels, categorized by intensity, to understand user activity patterns and inform Bellabeat's marketing strategies.

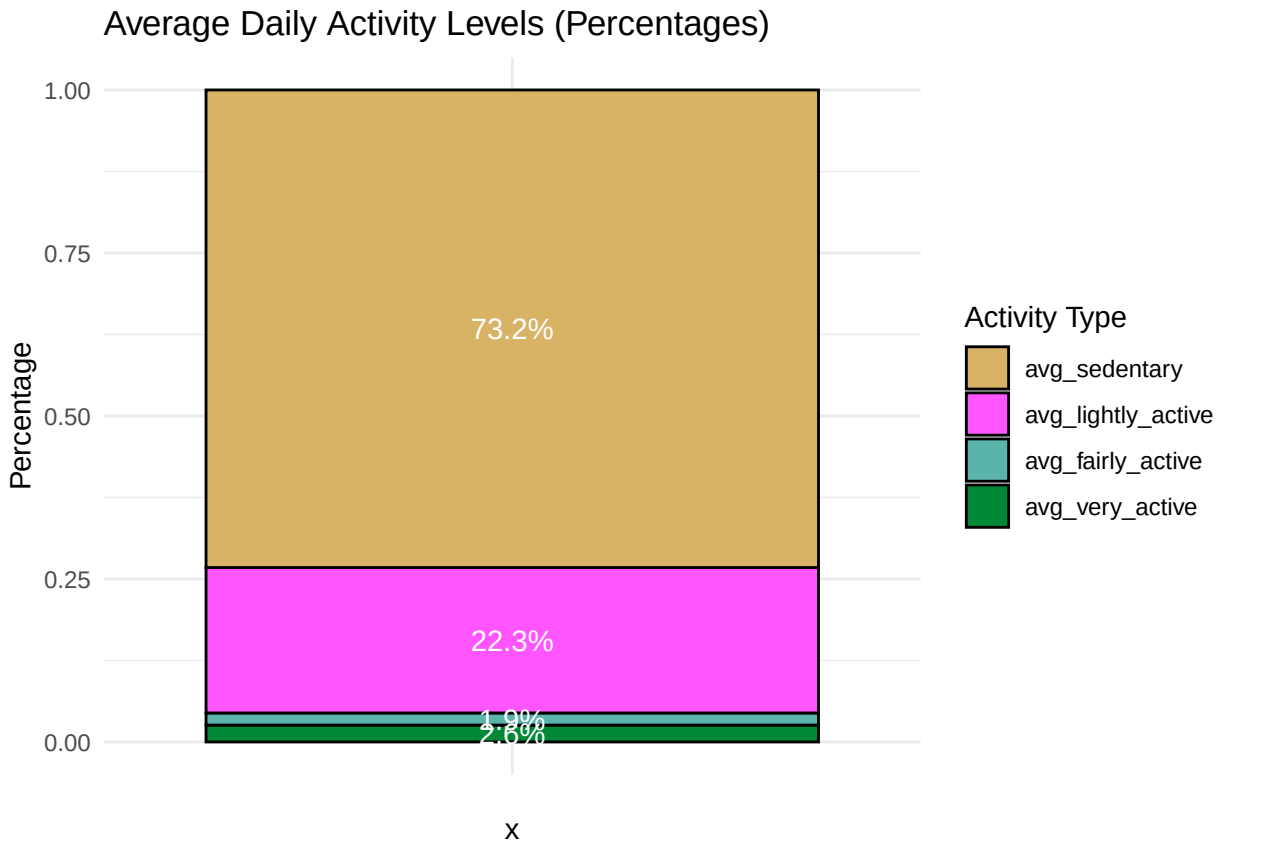
**Data and Methods:**

We calculated the average minutes spent in four activity categories: very active, fairly active, lightly active, and sedentary. The data was visualized using a stacked bar chart, where each segment of the bar represents the proportion of time spent in a particular activity level. The percentages are displayed directly on the chart for easy interpretation.

```
# 4. Average Daily Activity Levels (Stacked Bar Chart)
activity_levels <- merged_data %>%
  summarize(
    avg_very_active = mean(VeryActiveMinutes, na.rm = TRUE),
    avg_fairly_active = mean(FairlyActiveMinutes, na.rm = TRUE),
    avg_lightly_active = mean(LightlyActiveMinutes, na.rm = TRUE),
    avg_sedentary = mean(SedentaryMinutes, na.rm = TRUE)
  )

activity_levels_long <- activity_levels %>%
  pivot_longer(cols = everything(), names_to = "Activity Type", values_to = "Minutes") %>%
  mutate(`Activity Type` = factor(`Activity Type`, levels = c("avg_sedentary", "avg_lightly_active", "av

ggplot(activity_levels_long, aes(x = "", y = Minutes, fill = `Activity Type`)) +
  geom_col(color = "black", position = "fill") +
  geom_text(aes(label = paste0(round(Minutes / sum(Minutes) * 100, 1), "%")),
            position = position_fill(vjust = 0.5), color = "white") +
  labs(title = "Average Daily Activity Levels (Percentages)", y = "Percentage", fill = "Activity Type")
  scale_fill_manual(values = c("#d8b365", "#f5f", "#5ab4ac", "#008837")) +
  theme_minimal()
```




```r
activity_levels <- merged_data %>%
  summarize(
    avg_very_active = mean(VeryActiveMinutes, na.rm = TRUE),
    avg_fairly_active = mean(FairlyActiveMinutes, na.rm = TRUE),
    avg_lightly_active = mean(LightlyActiveMinutes, na.rm = TRUE),
    avg_sedentary = mean(SedentaryMinutes, na.rm = TRUE)
  )
```

**Key Findings:** Our analysis of average daily activity levels reveals the following:

- **Dominance of Sedentary Time:** The majority of the average user's day is spent in sedentary activities (73.2%). This highlights the prevalence of sedentary behavior in the dataset.
- **Lightly Active Minutes:** Lightly active minutes constitute the second-largest portion of the day (22.3%). This suggests that users engage in some level of light activity, but not as much as sedentary behavior.
- **Moderate and Very Active Minutes:** Fairly active and very active minutes make up relatively small portions of the day. Users spend 1.9% in fairly active pursuits and 2.6% in very active pursuits.

**Implications for Bellabeat:** These findings have important implications for Bellabeat's product development and marketing:

- **Targeting Sedentary Users:** Given the high proportion of sedentary time, there's a significant opportunity to target users who are looking to increase their activity levels. The Leaf could be positioned as a tool to help break up sedentary time and encourage more movement.
- **Promoting Light Activity:** While users engage in some light activity, there's potential to encourage more of it. Marketing could focus on the benefits of even small increases in light activity for overall

health and well-being.

- **Motivating Higher Intensity Activity:** Increasing moderate and very active minutes should be a key focus. The Leaf could be promoted as a way to track and achieve fitness goals, with features that motivate users to increase the intensity and duration of their workouts.
- **Feature Development:** Bellabeat could consider developing features that specifically address sedentary behavior, such as reminders to move, personalized activity goals, or integration with other apps that promote active lifestyles. Gamification and social features could also be explored to encourage more active minutes.

### 4.5 Sleep vs Activity

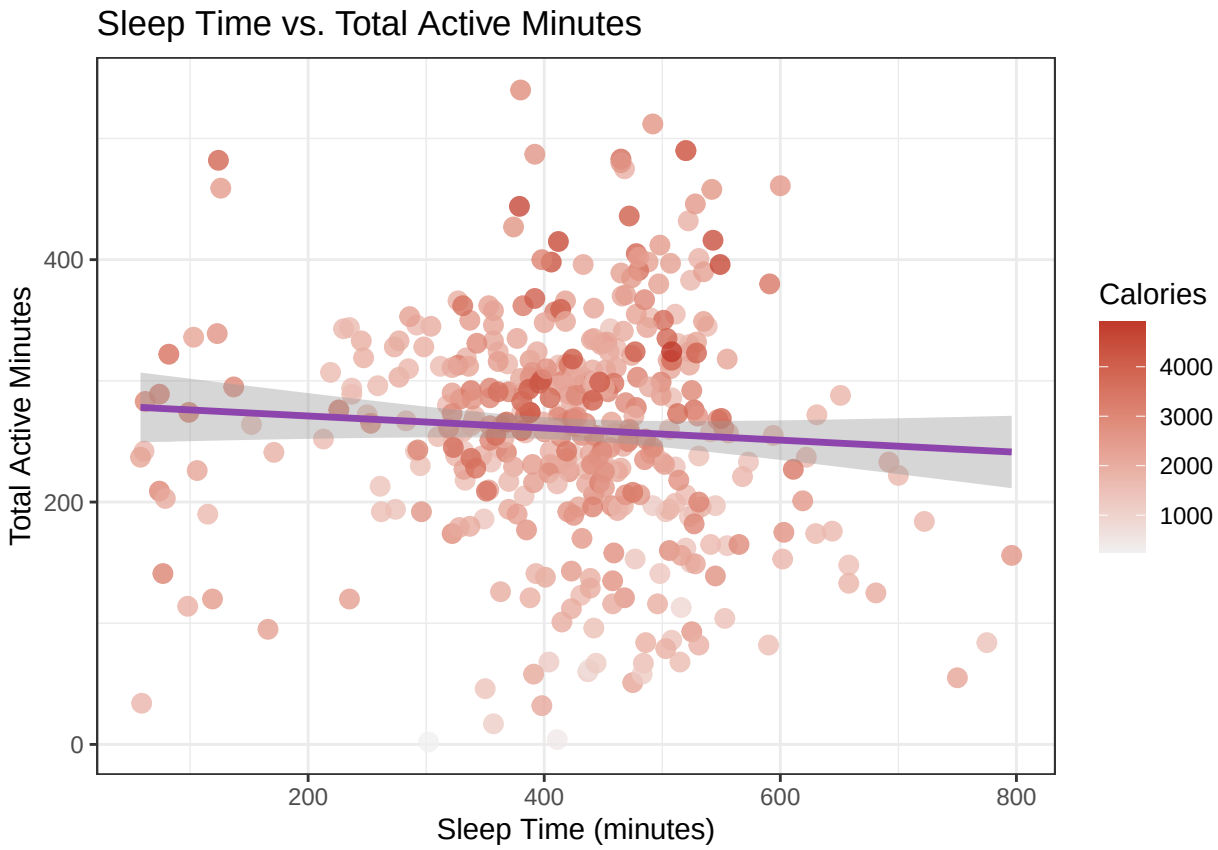
This report investigates the relationship between sleep duration and total daily active minutes, aiming to understand how these factors relate and inform Bellabeat's strategies.

**Data and Methods:** We analyzed data on total minutes asleep and total daily active minutes, calculated by summing very active, fairly active, and lightly active minutes. We used a scatter plot to visualize this relationship, with calorie burn represented by color intensity. A trend line was added to indicate any potential correlation.

```
# 5. Sleep vs. Total Activity
merged_data_sleep_total_activity <- merged_data %>%
  filter(!is.na(TotalMinutesAsleep), !is.na(VeryActiveMinutes), !is.na(FairlyActiveMinutes), !is.na(Ligh
  mutate(TotalActiveMinutes = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes) %>% # Crea
  select(TotalMinutesAsleep, TotalActiveMinutes, Calories) # Select needed columns


ggplot(merged_data_sleep_total_activity, aes(x = TotalMinutesAsleep, y = TotalActiveMinutes, color = Cal
  geom_point(size = 3, alpha = 0.8) +
  geom_smooth(method = "lm", color = "#8e44ad", linewidth = 1.2) +
  scale_color_gradient(low = "#f0f0f0", high = "#c0392b") +
  labs(title = "Sleep Time vs. Total Active Minutes", x = "Sleep Time (minutes)", y = "Total Active Minu
  theme_bw()
```

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




**Key Findings:** Our analysis reveals the following about the relationship between sleep and very active minutes:

- **Weak or No Clear Correlation:** The scatter plot and trend line suggest a weak or non-existent linear relationship between sleep time and total active minutes. The data points are scattered widely, and the trend line is nearly flat, indicating that longer sleep duration does not necessarily translate to more (or less) total activity.
- **Variability in Activity:** There's considerable variability in total active minutes across all sleep durations. Some individuals are highly active regardless of how much they sleep, while others are less active even with longer sleep times.
- **Potential Influence of Other Factors:** Calorie burn, represented by the color gradient, might be related to both sleep and activity, but the relationship is complex. It's likely that other unmeasured factors (e.g., diet, stress, daily routines, job type) contribute significantly to both sleep and activity levels.

**Implications for Bellabeat:** The lack of a strong direct correlation between sleep and total activity has several implications for Bellabeat:

- **Holistic Wellness Approach:** While a simple linear relationship isn't apparent, both sleep and activity are vital for overall wellness. Bellabeat should continue to emphasize the importance of tracking both and promote the Leaf as a comprehensive wellness tool.
- **Personalized Recommendations:** Given the high individual variability, personalized insights are key. The Leaf should be able to identify individual patterns and provide tailored recommendations for optimizing both sleep and activity, rather than relying on general trends.
- **Focus on Individual Patterns:** Marketing should avoid implying a direct causal link between sleep

and activity. Instead, it should focus on how the Leaf can help users understand their own unique sleep and activity patterns, empowering them to make informed choices.

- **Further Research and Integration:** Bellabeat could benefit from further research to explore the influence of other factors on sleep and activity. Integrating data from other apps or sources (e.g., nutrition tracking, stress management apps) could provide a more holistic view of user behavior and enable more effective interventions.

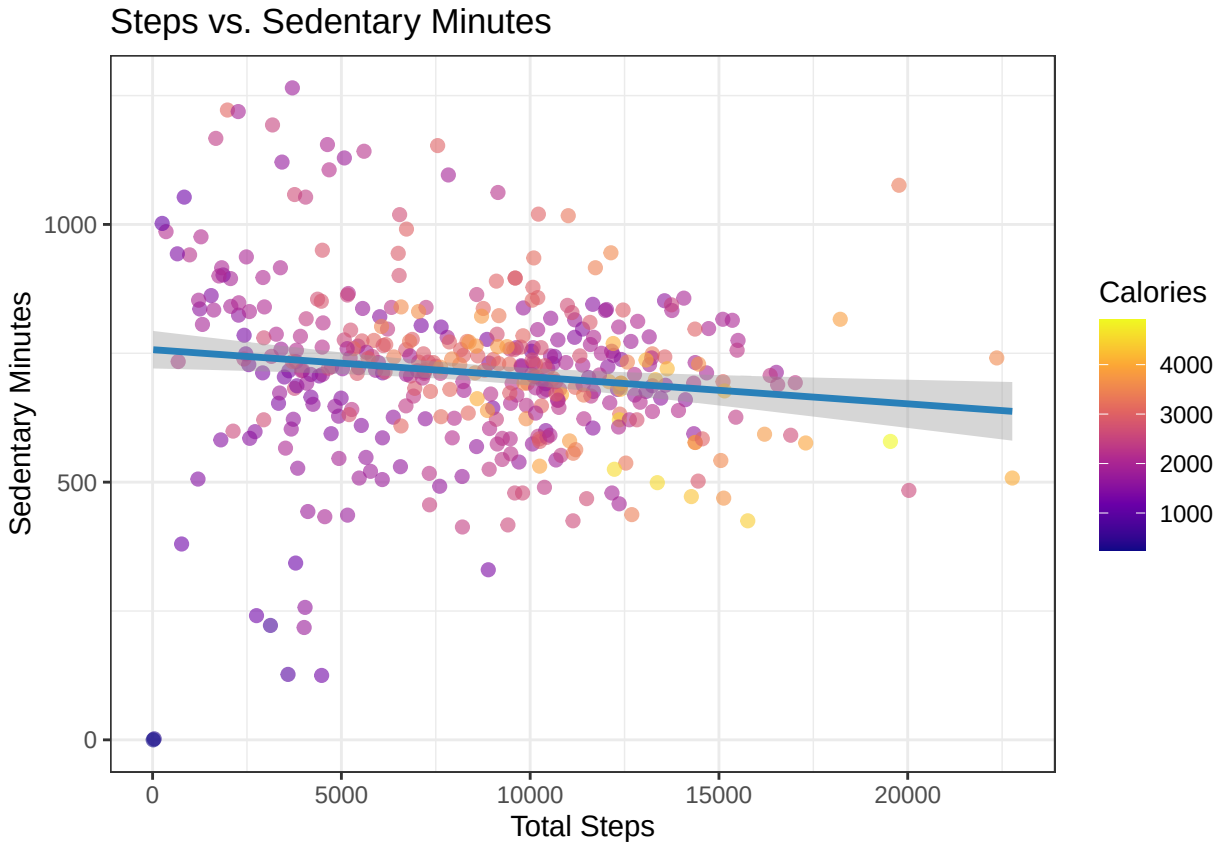
### 4.6 Steps vs. Sedentary Minutes

This report examines the relationship between daily step count and sedentary minutes to understand how these two key activity metrics interact and inform Bellabeat's strategies.

**Data and Methods:** We analyzed data on total daily steps and sedentary minutes. A scatter plot was used to visualize the relationship between these two variables, with calorie burn represented by color intensity. A trend line was added to highlight any potential correlation.

```
# 6. Steps vs. Sedentary Minutes (Scatter Plot)
ggplot(merged_data, aes(x = TotalSteps, y = SedentaryMinutes, color = Calories)) +
  geom_point(alpha = 0.6, size = 2) +
  geom_smooth(method = "lm", color = "#2980b9", linewidth = 1.2) +
  scale_color_viridis_c(option = "C") +
  labs(title = "Steps vs. Sedentary Minutes", x = "Total Steps", y = "Sedentary Minutes", color = "Calo
  theme_bw()
```

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



**Key Findings:** Our analysis reveals the following about the relationship between steps and sedentary minutes:

- **Moderate Negative Correlation:** The scatter plot and trend line suggest a moderate negative correlation between total steps and sedentary minutes. This indicates that, generally, as step count increases, sedentary minutes tend to decrease. However, the relationship isn't extremely strong, meaning there's still a fair amount of variation.
- **Variability in Sedentary Time:** Even at similar step counts, there's a considerable range of sedentary minutes. Some individuals with relatively high step counts still have a significant amount of sedentary time, while others with lower step counts might have less sedentary time.
- **Potential Influence of Other Factors:** Calorie burn, represented by the color gradient, appears related to both steps and sedentary minutes. Individuals with higher step counts and lower sedentary minutes tend to burn more calories. However, as with previous analyses, other factors likely influence this relationship and warrant further investigation.

**Implications for Bellabeat:** The observed negative correlation between steps and sedentary minutes has several implications for Bellabeat:

- **Reinforcing Activity Goals:** The data supports the idea of promoting increased step counts as a way to reduce sedentary time. The Leaf can be positioned as a tool to help users achieve this goal.
- **Personalized Interventions:** Given the variability in sedentary time, personalized interventions are crucial. The Leaf could provide personalized insights and recommendations based on individual step and sedentary patterns. For example, it could suggest strategies for breaking up long periods of sitting, even for users who are already relatively active.
- **Targeted Messaging:** Marketing campaigns can emphasize the health benefits of reducing sedentary behavior in addition to increasing steps. The Leaf could be promoted as a tool for achieving a balanced activity profile, not just hitting a certain step count.
- **Feature Development:** Bellabeat could consider developing features specifically designed to address sedentary behavior, such as alerts or reminders to move after periods of inactivity, or gamified challenges to reduce sitting time. Integration with other health and wellness apps could also be beneficial.

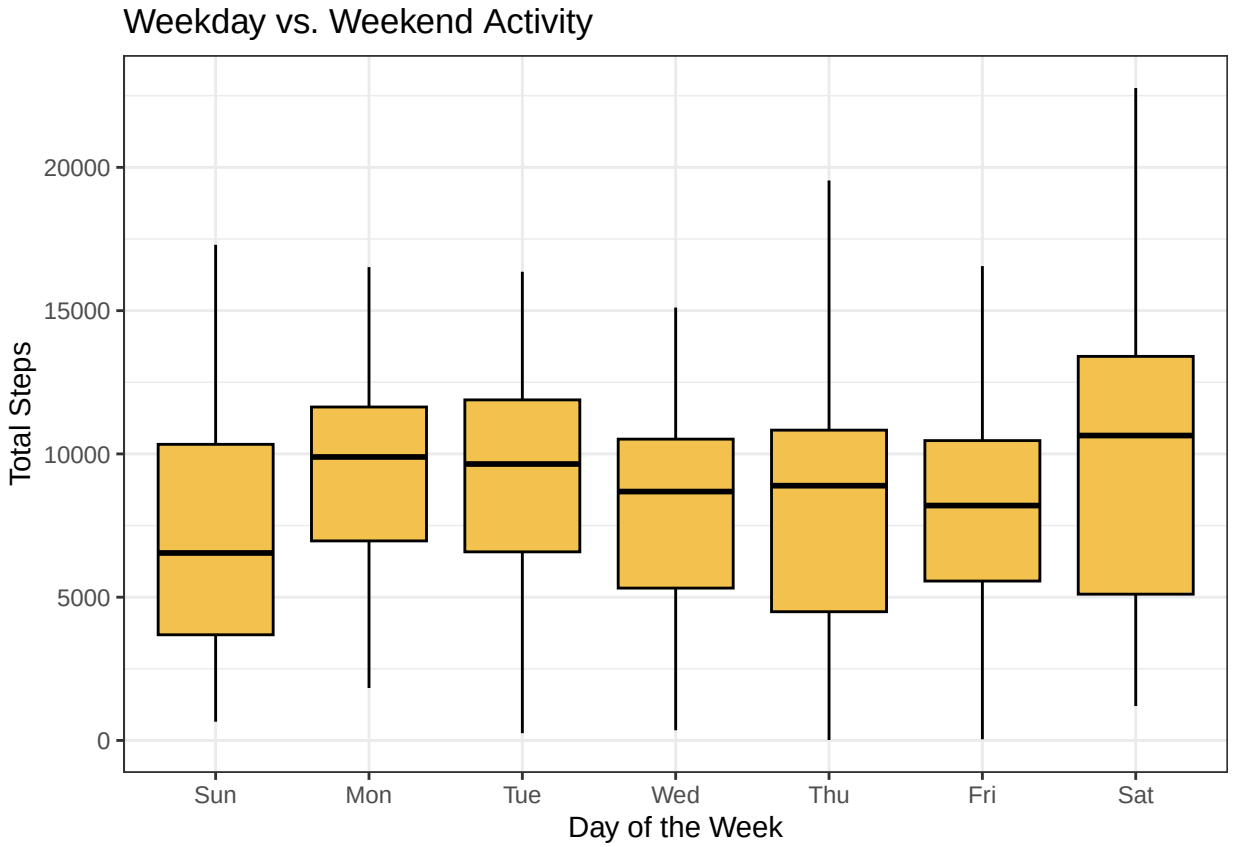
### 4.7 Weekday vs. Weekend Activity

This report analyzes the difference in daily step counts between weekdays and weekends to understand user activity patterns and inform Bellabeat's engagement strategies.

**Data and Methods:** We added a "Weekday" column to the dataset, identifying each day as either a weekday (Monday-Friday) or a weekend (Saturday-Sunday). We used a boxplot to visualize the distribution of total steps for each day of the week.

```
# 7. Weekday vs. Weekend Activity (Boxplot)
merged_data$Weekday <- wday(merged_data$ActivityDate, label = TRUE)  # Add weekday column

ggplot(merged_data, aes(x = Weekday, y = TotalSteps)) +
  geom_boxplot(fill = "#f2c14e", color = "black") +
  labs(title = "Weekday vs. Weekend Activity", x = "Day of the Week", y = "Total Steps") +
  theme_bw()
```




**Key Findings:** Our analysis of weekday vs. weekend activity reveals the following:

- **Potential Differences in Activity:** The boxplot visually compares the distribution of step counts across the days of the week. While a direct comparison is difficult without statistical tests, the boxplot suggests potential differences in activity levels between weekdays and weekends.
- **Variability in Daily Steps:** The boxplot shows the range and distribution of step counts for each day. The spread of the data (represented by the interquartile range or IQR) suggests how consistent or variable activity is on each day of the week.
- **Further Analysis Needed:** The boxplot provides a visual comparison, but further statistical analysis (e.g., t-tests or ANOVA) would be needed to determine if the observed differences between weekday and weekend activity are statistically significant.

**Implications for Bellabeat:** The observed differences (or lack thereof) in weekday vs. weekend activity have several implications for Bellabeat:

- **Targeted Engagement:** If significant differences exist, Bellabeat can tailor engagement strategies to specific days of the week. For example, if weekend activity is lower, targeted promotions or challenges could encourage more weekend movement.
- **Understanding User Behavior:** Understanding how activity patterns change throughout the week provides valuable insights into user behavior. This information can be used to refine product features and marketing messages.
- **Personalized Recommendations:** The Leaf could provide personalized recommendations based on individual weekday/weekend activity patterns. For instance, if a user is consistently less active on weekends, the app could suggest weekend-specific activities.

- **Feature Development:** Bellabeat could consider developing features that cater to different activity patterns. For example, weekday-focused features might include workplace wellness challenges, while weekend-focused features could suggest outdoor activities.
- **Further Research:** As mentioned, additional statistical testing is needed to confirm any observed differences. Further research could also explore why these differences exist (e.g., work schedules, social activities, weather).

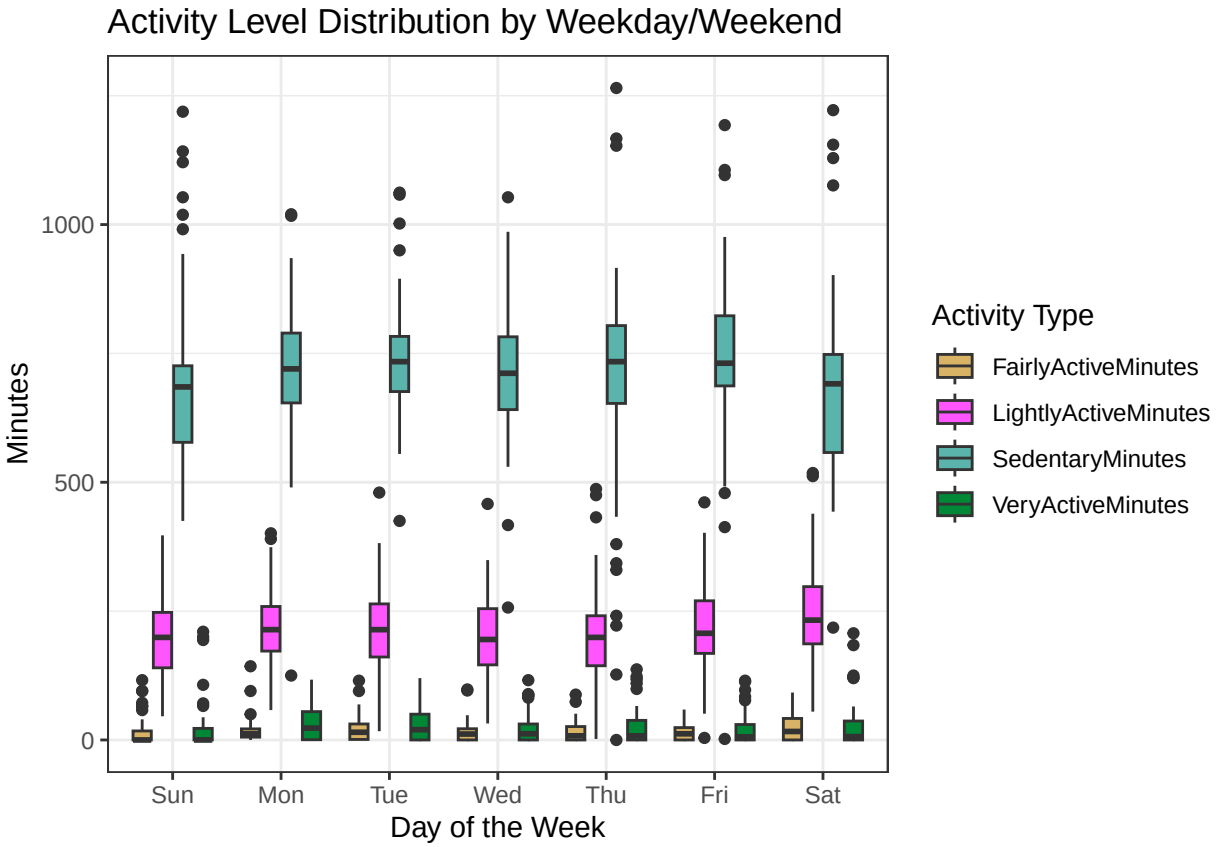
### 4.8 Analysis of Activity Level Distribution by Day of the Week

This report analyzes the distribution of different activity levels (very active, fairly active, lightly active, and sedentary) across weekdays and weekends to understand user activity patterns and inform Bellabeat's engagement strategies.

**Data and Methods:** We examined the distribution of minutes spent in each activity level for each day of the week. We used a boxplot to visualize these distributions, allowing for comparison of activity levels across different days. A summary table of median, mean, and quartile values was also generated to provide further insight into the data.

```
# 8. Activity Level Distribution by Weekday/Weekend (Boxplot)
# Reshape the data for plotting
activity_by_day <- merged_data %>%
    select(Weekday, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes) %>%
    pivot_longer(cols = c(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes
                 names_to = "ActivityType", values_to = "Minutes")

# Create the boxplot
ggplot(activity_by_day, aes(x = Weekday, y = Minutes, fill = ActivityType)) +
    geom_boxplot() +
    labs(title = "Activity Level Distribution by Weekday/Weekend",
         x = "Day of the Week", y = "Minutes", fill = "Activity Type") +
    scale_fill_manual(values = c("#d8b365", "#f5f", "#5ab4ac", "#008837")) + # Use same colors as befor
    theme_bw()
```




**Key Findings:** Our analysis of activity level distribution reveals the following:

- **Variation in Activity Levels:** The boxplot and summary table clearly show that the distribution of minutes spent in different activity levels varies across the days of the week. This highlights the importance of considering daily patterns when analyzing user activity.
- **Sedentary Time:** The data indicates a significant amount of sedentary time across all days of the week. Understanding the daily fluctuations in sedentary behavior is crucial for developing targeted interventions.
- **Active Minutes:** The distributions of very active, fairly active, and lightly active minutes also show variation. Further investigation is needed to explore the reasons behind these differences and how they relate to user behavior.
- **Weekend vs. Weekday Trends:** While visual inspection suggests potential differences between weekday and weekend activity, further statistical analysis is needed to confirm these trends. The summary table provides a good overview of the central tendency and spread of the data for each activity type on each day, enabling more detailed comparisons.

**Implications for Bellabeat:** The observed variations in activity levels have several important implications for Bellabeat:

- **Targeted Interventions:** Understanding daily activity patterns allows for the development of targeted interventions. For example, if sedentary time is higher on certain days, the Leaf could provide prompts or suggestions for breaking up periods of inactivity.
- **Personalized Recommendations:** The Leaf could offer personalized recommendations based on individual activity patterns. If a user's activity levels are consistently lower on specific days, the app could suggest tailored activities or goals.

- **Feature Development:** Bellabeat could consider developing features that cater to different daily activity patterns. For instance, the app could offer weekday-specific challenges or weekend-focused activity suggestions.
- **Marketing Strategies:** Marketing campaigns can be tailored to specific days of the week, promoting activities or features that align with user behavior on those days.
- **Further Research:** Further research is needed to understand the underlying reasons for the observed variations in activity levels. Exploring factors such as work schedules, social activities, and sleep patterns could provide valuable insights. Statistical tests should be used to confirm any visually observed trends.

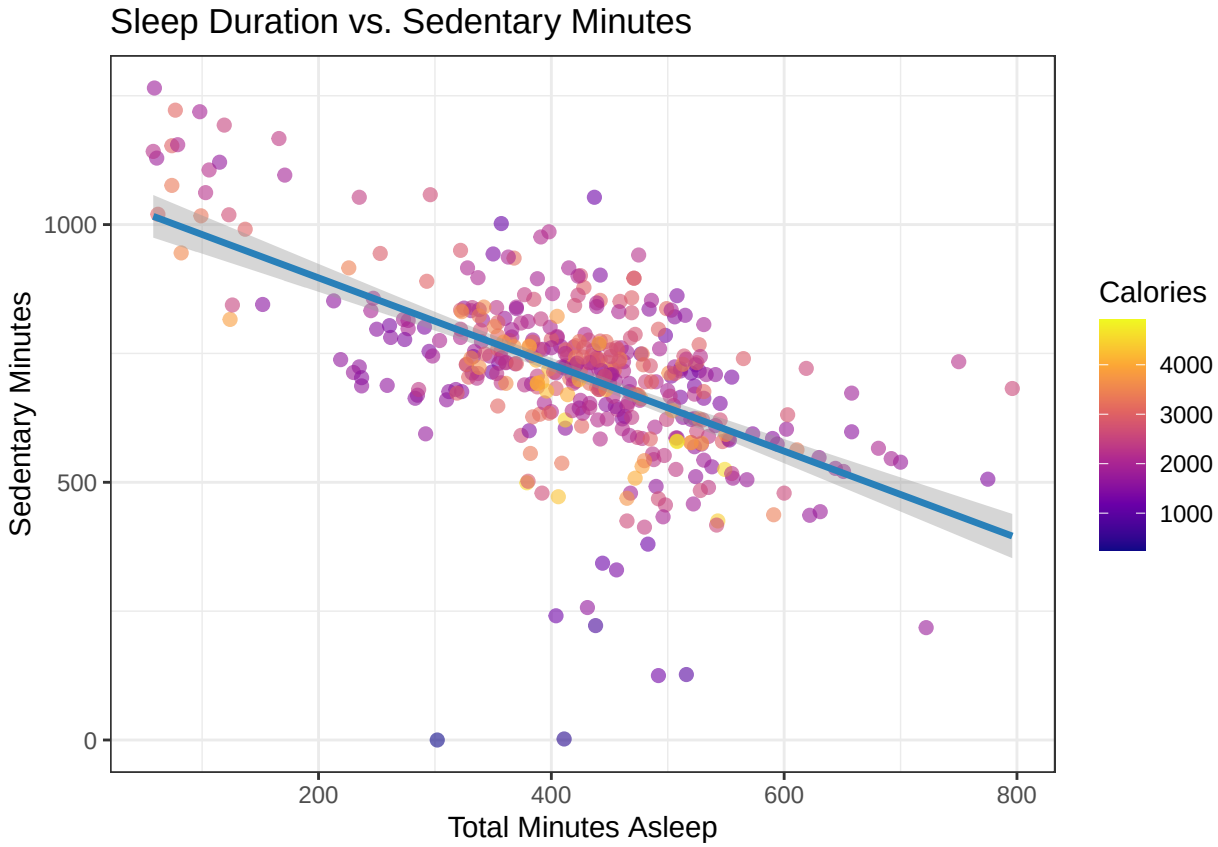
### 4.9 Analysis of Sleep Duration vs. Sedentary Minutes

This report analyzes the relationship between sleep duration and sedentary minutes to understand how sleep patterns may influence sedentary behavior and inform Bellabeat's wellness strategies.

**Data and Methods:** We examined the relationship between total minutes asleep and sedentary minutes, using calorie burn as a visual indicator. A scatter plot with a trend line was used to visualize the correlation between these two variables.

```
# 9. Sleep Duration vs. Sedentary Minutes (Scatter Plot)
ggplot(merged_data, aes(x = TotalMinutesAsleep, y = SedentaryMinutes, color = Calories)) +
    geom_point(alpha = 0.6, size = 2) +
    geom_smooth(method = "lm", color = "#2980b9", linewidth = 1.2) +
    scale_color_viridis_c(option = "C") +
    labs(title = "Sleep Duration vs. Sedentary Minutes",
         x = "Total Minutes Asleep", y = "Sedentary Minutes", color = "Calories") +
    theme_bw()
```

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




**Key Findings:** Our analysis reveals the following regarding the relationship between sleep duration and sedentary minutes:

- **Potential Negative Correlation:** The scatter plot and trend line suggest a potential negative correlation between sleep duration and sedentary minutes. This indicates that, generally, individuals who sleep longer tend to have fewer sedentary minutes during the day. However, the relationship isn't extremely strong, suggesting other factors are at play.
- **Variability in Sedentary Time:** Even with similar sleep durations, there's a considerable range of sedentary minutes. Some individuals with longer sleep times still have a high amount of sedentary time, while others with shorter sleep times might have less sedentary time.
- **Influence of Calories Burned:** Calorie burn, represented by the color gradient, appears related to both sleep duration and sedentary minutes. Individuals who sleep longer and have fewer sedentary minutes tend to burn more calories. However, as with previous analyses, other factors likely influence this relationship and warrant further investigation.

**Implications for Bellabeat:** The potential negative correlation between sleep duration and sedentary minutes has several implications for Bellabeat:

- **Promoting Healthy Sleep:** The data supports the idea of promoting healthy sleep habits as a way to potentially reduce sedentary behavior. The Leaf can be positioned as a tool to help users track and improve their sleep quality.
- **Integrated Wellness Approach:** Bellabeat should emphasize an integrated approach to wellness, highlighting the interconnectedness of sleep, activity, and overall health. The Leaf can be used to track multiple metrics and provide holistic insights.
- **Personalized Insights:** Given the variability in sedentary time, personalized interventions are crucial.

The Leaf could provide personalized insights and recommendations based on individual sleep and sedentary patterns. For example, it could suggest strategies for breaking up long periods of sitting, even for users who are already getting adequate sleep.

- **Further Research:** Further research is needed to establish a stronger causal link between sleep duration and sedentary behavior. Investigating other factors that may influence this relationship, such as stress levels, work schedules, and physical activity levels, would be valuable.

### 4.10 Analysis of Daily Steps vs. Sleep Duration

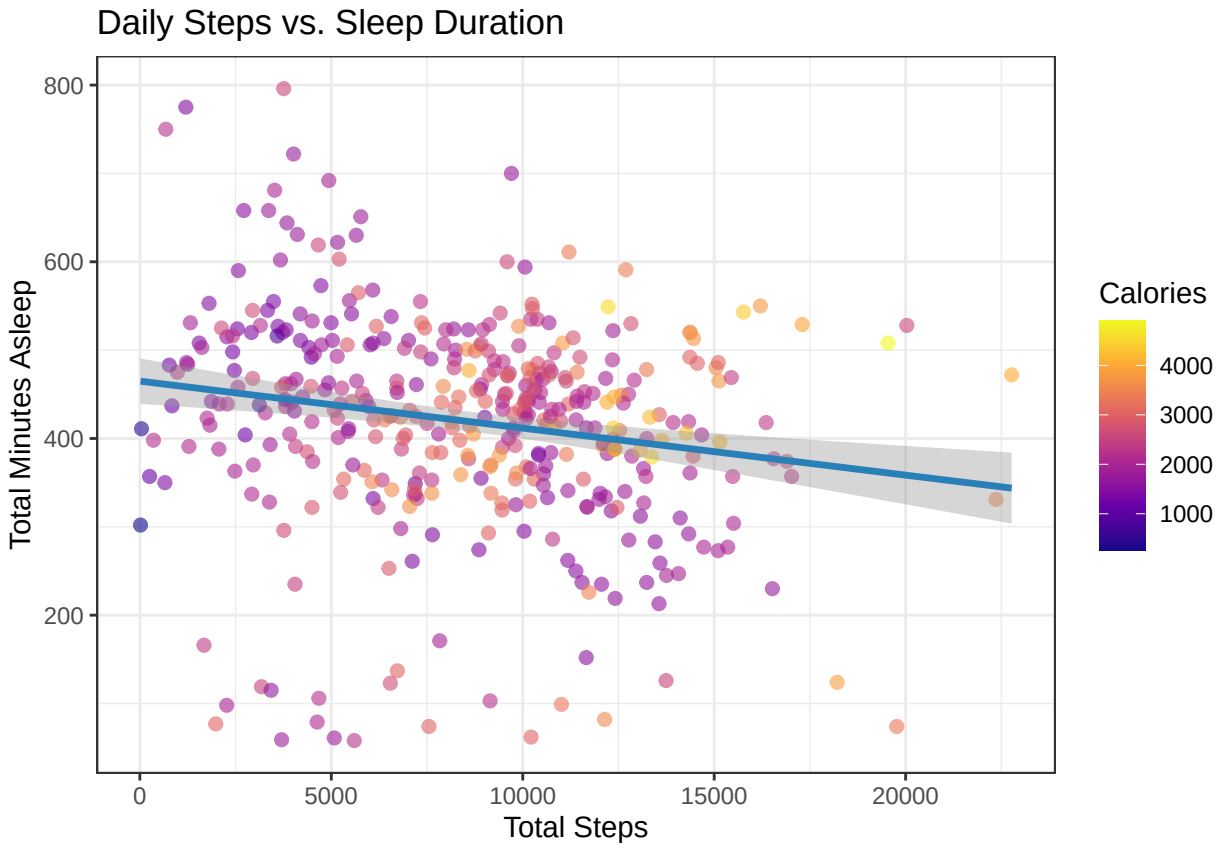
This report examines the relationship between daily step count and sleep duration to understand how physical activity may influence sleep patterns and inform Bellabeat's wellness recommendations.

**Data and Methods:** We analyzed the correlation between total daily steps and total minutes asleep, using calorie burn as a visual indicator. A scatter plot with a trend line was used to visualize the relationship between these two variables.

```
# 10. Daily Steps vs. Sleep Duration (Scatter Plot)

ggplot(merged_data, aes(x = TotalSteps, y = TotalMinutesAsleep, color = Calories)) +
    geom_point(alpha = 0.6, size = 2) +
    geom_smooth(method = "lm", color = "#2980b9", linewidth = 1.2) +
    scale_color_viridis_c(option = "C") +
    labs(title = "Daily Steps vs. Sleep Duration",
         x = "Total Steps", y = "Total Minutes Asleep", color = "Calories") +
    theme_bw()
```

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



**Key Findings:** Our analysis reveals the following regarding the relationship between daily steps and sleep duration:

- **Weak or No Clear Correlation:** The scatter plot and trend line suggest a weak or possibly non-existent correlation between daily step count and sleep duration. While there might be a very slight positive trend, the relationship is not strong. This implies that simply increasing daily steps does not guarantee an increase in sleep duration.
- **Variability in Sleep Duration:** The data shows a wide range of sleep durations for any given step count. This indicates that other factors, beyond just daily steps, significantly influence sleep patterns.
- **Influence of Calories Burned:** Calorie burn, as visualized by the color gradient, doesn't show a clear pattern related to the relationship between steps and sleep. This suggests that while calorie burn is related to both steps and potentially sleep (as discussed in other reports), it doesn't explain the variability in sleep duration at different step counts.

**Implications for Bellabeat:** The weak correlation between daily steps and sleep duration has several implications for Bellabeat:

- **Holistic Wellness Recommendations:** Bellabeat should avoid framing increased step counts as a direct solution for improving sleep duration. Instead, a more holistic approach to wellness is necessary, considering other factors that influence sleep.
- **Targeted Sleep Improvement Strategies:** Bellabeat should focus on providing targeted sleep improvement strategies that go beyond just increasing physical activity. This might include recommendations for sleep hygiene, stress management, and relaxation techniques.
- **Further Research:** Further research is crucial to understand the complex interplay between physical activity, sleep, and other lifestyle factors. Investigating the impact of different types of physical activity, the timing of exercise, and individual sleep needs would be valuable.
- **Personalized Insights:** The Leaf could provide more personalized insights by considering multiple data points, including sleep tracking, activity levels, and other relevant metrics (e.g., heart rate, stress). This would allow for more tailored recommendations for improving both activity and sleep.
- **Managing User Expectations:** It's important to manage user expectations about the impact of steps on sleep. While physical activity is generally beneficial for overall health, it's not a guaranteed solution for sleep problems. Bellabeat should communicate this clearly to users.

## 5. Share (Summary of Key Findings)

This section summarizes the key findings from our analysis of Bellabeat fitness tracker data, focusing on user activity and sleep patterns.

### 5.1 Distribution of Daily Steps:

- Daily step counts vary widely, from a few hundred to over 20,000, indicating diverse activity levels. While a significant portion of users cluster around 10,000-12,000 steps, a substantial number fall below the recommended 10,000-step benchmark, highlighting an opportunity to encourage increased activity. Outliers representing unusually high or low step counts warrant further investigation.

### 5.2 Steps vs. Calories Burned:

- A strong positive correlation exists between daily step count and calories burned, with increased steps and distance traveled leading to higher calorie expenditure. However, variability in calorie burn at any given step count suggests the influence of other factors like activity intensity and metabolism.

**5.3 Sleep Time Distribution:**

- Sleep durations vary considerably, ranging from under an hour to over 13 hours. While a substantial portion of users sleep between 5 and 10 hours, significant segments sleep less or more, emphasizing the diversity in sleep patterns.

**5.4 Average Daily Activity Levels:**

- Sedentary time dominates the average user's day (73.2%), followed by lightly active minutes (22.3%). Moderate and very active minutes constitute only small portions of daily activity, highlighting the prevalence of sedentary behavior.

**5.5 Sleep vs. Activity:**

- No strong linear relationship exists between sleep duration and total daily active minutes, suggesting a complex interplay of factors influencing both sleep and activity levels.

**5.6 Steps vs. Sedentary Minutes:**

- A moderate negative correlation exists between daily step count and sedentary minutes. Increased step counts tend to be associated with decreased sedentary time, but considerable variability exists, indicating that even active individuals may have significant amounts of sedentary time.

**5.7 Weekday vs. Weekend Activity:**

- Visual inspection suggests potential differences in activity levels between weekdays and weekends. Further statistical analysis is required to confirm these trends.

**5.8 Activity Level Distribution by Day of the Week:**

- The distribution of different activity levels varies across days of the week, highlighting the importance of considering daily patterns when analyzing user activity.

**5.9 Sleep Duration vs. Sedentary Minutes:**

- A potential negative correlation exists between sleep duration and sedentary minutes, suggesting that longer sleep might be associated with less sedentary behavior, though the relationship is not very strong.

**5.10 Daily Steps vs. Sleep Duration:**

- A weak or possibly non-existent correlation was found between daily step count and sleep duration, indicating that increasing steps does not necessarily lead to increased sleep duration.

**5.11 Answering the Guiding Questions**

**1. What are some trends in smart device usage?**

- **Diverse Activity Levels with a Sedentary Lean:** Daily step counts exhibit a wide range, from a few hundred to over 20,000, demonstrating diverse activity levels among users. While a significant portion clusters around 10,000-12,000 steps, a substantial number fall below the recommended 10,000-step benchmark. Furthermore, sedentary time dominates the average user's day, accounting for 73.2% of daily activity, followed by lightly active minutes at 22.3%. This highlights the prevalence of sedentary behavior and the need for interventions to encourage increased activity.
- **Positive Correlation between Steps and Calorie Burn:** A strong positive correlation exists between daily step count and calories burned. As step count and distance traveled increase, calorie expenditure also tends to rise. However, it's important to note that calorie burn varies even at similar

step counts, suggesting that other factors, such as activity intensity and individual metabolism, also play a role.

- **Varied Sleep Patterns:** Sleep durations vary considerably, ranging from less than an hour to over 13 hours. While a substantial portion of users sleep between 5 and 10 hours, significant segments sleep less or more, emphasizing the diversity in sleep patterns and the importance of personalized sleep tracking and recommendations.
- **Complex Relationship between Sleep and Activity:** No strong linear relationship exists between sleep duration and total daily active minutes. This suggests that sleep and activity levels are influenced by a complex interplay of factors and that simply sleeping longer does not necessarily translate to increased activity.
- **Inverse Relationship between Steps and Sedentary Time:** A moderate negative correlation exists between daily step count and sedentary minutes. As step count increases, sedentary time tends to decrease. However, considerable variability exists, indicating that even active individuals may still have significant amounts of sedentary time.
- **Potential for Weekday/Weekend Activity Differences:** Visual inspection suggests potential differences in activity levels between weekdays and weekends. Further statistical analysis is needed to confirm the statistical significance of these variations. This could indicate different activity patterns based on work/school schedules versus leisure time.
- **Daily Variations in Activity Level Distribution:** The distribution of different activity levels (very active, fairly active, lightly active, and sedentary) varies across days of the week. This underscores the importance of considering daily patterns when analyzing user activity and highlights the need for personalized activity recommendations that take these daily variations into account.
- **Weak Relationship between Sleep Duration and Sedentary Time:** While a potential negative correlation exists between sleep duration and sedentary minutes, suggesting that longer sleep might be associated with less sedentary behavior, the relationship is not very strong. This implies that other factors are more influential in determining sedentary behavior.
- **Limited Connection between Steps and Sleep Duration:** A weak or possibly non-existent correlation was found between daily step count and sleep duration. This indicates that increasing steps does not necessarily lead to increased sleep duration, highlighting the importance of addressing sleep improvement strategies beyond just physical activity.

**2. How could these trends apply to Bellabeat customers?** These trends are likely relevant to Bellabeat users, who share similar interests in health and wellness. The prevalence of sedentary behavior, the importance of sleep tracking, and the need for personalized insights are all applicable to Bellabeat's target audience. Specifically:

- The high prevalence of sedentary behavior suggests a market for products and services that encourage movement.
- The correlation between steps and calorie burn can be leveraged to motivate users to be more active.
- The diverse sleep patterns reinforce the need for personalized sleep tracking and recommendations.
- The complex relationship between sleep and activity highlights the importance of a holistic approach to wellness.
- The potential for weekday/weekend differences in activity levels can inform targeted marketing campaigns.

**3. How could these trends help influence Bellabeat marketing strategy?**

- **Product Positioning:** Position the Leaf to address sedentary behavior, provide personalized sleep insights, and offer holistic wellness solutions. Emphasize the Leaf's ability to track various activity types and intensity levels, going beyond just steps.

- **Feature Promotion:** Highlight features related to step tracking, calorie estimation, sleep monitoring, personalized recommendations, and sedentary behavior alerts. Showcase the Leaf's ability to integrate with other health and fitness apps.
- **Targeted Messaging:** Emphasize the link between activity and calorie burn, the importance of reducing sedentary time, and the benefits of personalized wellness plans. Tailor messaging to different user segments based on their activity levels, sleep patterns, and weekday/weekend activity variations.
- **Content Creation:** Create content (blog posts, articles, social media posts) about the health risks of sedentary behavior, the benefits of regular physical activity, the importance of sleep hygiene, and how the Leaf can help users achieve their wellness goals.
- **Partnerships:** Partner with fitness influencers or other wellness companies to promote the Leaf and encourage healthy habits.
- **Gamification and Challenges:** Design gamified challenges and social features within the Bellabeat app to motivate users to be more active and reduce their sedentary time. Consider weekday/weekend specific challenges.
- **Personalized Recommendations:** Leverage the insights from the data to provide personalized recommendations for activity, sleep, and overall wellness. This could include suggesting specific types of activities, setting personalized step goals, and offering tailored sleep improvement tips.

## 6. Act (Recommendations)

Based on the key findings summarized above, we recommend the following actions for Bellabeat:

**I. Activity Enhancement:**

- **Targeted Interventions:** Develop personalized interventions for users with low step counts, leveraging the Leaf to set personalized step goals, provide reminders, and offer tailored encouragement. This directly addresses the finding that a substantial number of users fall below the recommended step count.
- **Holistic Activity Tracking:** Enhance activity tracking beyond steps and distance to include different activity types and intensity levels, providing a more comprehensive view of calorie expenditure. This addresses the finding that distance contributes to calorie burn and the observed variability in calorie burn at similar step counts.
- **Sedentary Behavior Reduction:** Develop features specifically designed to reduce sedentary behavior, such as alerts or reminders to move, and personalized strategies for breaking up long periods of sitting. This directly targets the high prevalence of sedentary behavior and the negative correlation between steps and sedentary minutes.

**II. Sleep Optimization:**

- **Personalized Sleep Insights:** Provide personalized insights and recommendations for improving sleep quality, regardless of total sleep duration, recognizing the variability in sleep patterns among users.
- **Holistic Sleep Strategies:** Offer sleep improvement strategies that go beyond just increasing physical activity, including recommendations for sleep hygiene, stress management, and relaxation techniques. This acknowledges the finding that there's no direct correlation between steps and sleep duration.

**III. User Engagement & Personalization:**

- **Weekday/Weekend Targeting:** If further analysis confirms statistically significant differences in weekday/weekend activity, tailor engagement strategies to specific days of the week, offering targeted promotions or challenges.

- **Personalized Recommendations:** Leverage the Leaf to provide personalized recommendations based on individual activity and sleep patterns, acknowledging the complex interplay of factors influencing these behaviors.

**IV. Product Development & Research:**

- **Further Research:** Invest in further research to explore the influence of other factors (e.g., age, weight, fitness level, stress, diet) on activity and sleep patterns. This will enhance the accuracy of personalized insights and recommendations.
- **Integration with Other Apps:** Explore integrating the Leaf with other health and fitness apps to provide a more holistic view of user behavior and enable more effective interventions.

## 7. Next Steps

**I. Activity Enhancement:**

**Targeted Interventions:**

- **Task 1:** Conduct user research (surveys, focus groups) to understand user motivations and barriers to increasing activity levels. *Timeline:* 4 weeks. *Responsibility:* User Research Team. *Resources:* Budget for survey platform and participant incentives.
- **Task 2:** Design and prototype personalized goal-setting and reminder features for the Leaf app. *Timeline:* 6 weeks. *Responsibility:* Product Design Team. *Resources:* Design software, prototyping tools.
- **Task 3:** Develop and implement A/B testing framework for evaluating different intervention strategies. *Timeline:* 4 weeks. *Responsibility:* Engineering Team. *Resources:* Development tools, testing platform.

**Motivational Features:**

- **Task 1:** Brainstorm and prioritize new motivational feature ideas (gamification, social challenges, personalized progress tracking) based on user research and competitor analysis. *Timeline:* 2 weeks. *Responsibility:* Product Management Team & Marketing Team. *Resources:* Market research reports, brainstorming tools. *Metrics:* Number of new feature ideas generated, prioritized feature list.
- **Task 2:** Design, prototype, and user test the top 2-3 feature ideas. *Timeline:* 8 weeks. *Responsibility:* Product Design Team & UX Team. *Resources:* Design software, prototyping tools, user testing platform. *Metrics:* User feedback on new features, usability testing results.
- **Task 3:** Develop and integrate the selected features into the Leaf app. *Timeline:* 12 weeks. *Responsibility:* Engineering Team. *Resources:* Development tools, testing platform. *Metrics:* User adoption rates of new features, user engagement metrics.

**Holistic Activity Tracking:**

- **Task 1:** Research and evaluate potential integrations with other fitness apps and wearables. *Timeline:* 4 weeks. *Responsibility:* Business Development Team & Engineering Team. *Resources:* Market research reports, API documentation. *Metrics:* List of potential integration partners, feasibility assessment for each integration.
- **Task 2:** Develop and implement integration with selected partners. *Timeline:* 8 weeks. *Responsibility:* Engineering Team. *Resources:* Development tools, API integration tools. *Metrics:* Number of successful integrations, user usage of integrated features.
- **Task 3:** Explore the feasibility of adding new activity tracking capabilities (e.g., cycling, swimming, strength training) to the Leaf. *Timeline:* 6 weeks. *Responsibility:* Product Development Team & Engineering Team. *Resources:* Market research, sensor technology research. *Metrics:* Feasibility report, cost estimates for new features.

**Sedentary Behavior Reduction:**

- **Task 1:** Design and prototype alerts, reminders, and gamified challenges to encourage movement and break up sedentary periods. *Timeline:* 6 weeks. *Responsibility:* Product Design Team & UX Team. *Resources:* Design software, prototyping tools. *Metrics:* User feedback on new features, usability testing results.
- **Task 2:** Develop and integrate the selected features into the Leaf app. *Timeline:* 8 weeks. *Responsibility:* Engineering Team. *Resources:* Development tools, testing platform. *Metrics:* User adoption rates of new features, changes in user sedentary behavior.

**II. Sleep Optimization:**

**Personalized Sleep Insights:**

- **Task 1:** Develop algorithms to analyze sleep data and generate personalized insights on sleep quality, including factors like sleep efficiency, sleep stages, and sleep disturbances. *Timeline:* 12 weeks. *Responsibility:* Data Science Team. *Resources:* Data analysis tools, machine learning libraries. *Metrics:* Accuracy of sleep analysis algorithms, user feedback on personalized insights.
- **Task 2:** Design and develop a user interface to present personalized sleep insights in a clear and understandable way. *Timeline:* 6 weeks. *Responsibility:* Product Design Team & UX Team. *Resources:* Design software, prototyping tools. *Metrics:* User feedback on the presentation of sleep insights.

**Holistic Sleep Strategies:**

- **Task 1:** Partner with sleep experts or develop educational content on sleep hygiene, stress management, and relaxation techniques. *Timeline:* 4 weeks. *Responsibility:* Content Marketing Team & Partnerships Team. *Resources:* Sleep research, content creation tools. *Metrics:* Quality and relevance of educational content, partnerships established.
- **Task 2:** Integrate sleep improvement recommendations into the Leaf app, based on individual sleep patterns and user preferences. *Timeline:* 8 weeks. *Responsibility:* Product Management Team & Engineering Team. *Resources:* Content management system, development tools. *Metrics:* User engagement with sleep improvement recommendations, changes in user sleep patterns.

**III. User Engagement & Personalization:**

**Weekday/Weekend Targeting:**

- **Task 1:** Conduct statistical analysis (t-tests, ANOVA) to confirm the statistical significance of observed differences in weekday/weekend activity. *Timeline:* 2 weeks. *Responsibility:* Data Science Team. *Resources:* Data analysis tools. *Metrics:* Statistical significance of differences in activity levels.
- **Task 2:** Develop targeted promotions, challenges, and notifications for weekdays and weekends, based on user activity patterns. *Timeline:* 6 weeks. *Responsibility:* Marketing Team & Product Management Team. *Resources:* Marketing automation tools, content creation tools. *Metrics:* Click-through rates on targeted promotions, user participation in challenges.

**Personalized Recommendations:**

- **Task 1:** Develop machine learning models to analyze user data (activity, sleep, heart rate, etc.) and generate personalized recommendations for activity and sleep improvement. *Timeline:* 16 weeks. *Responsibility:* Data Science Team. *Resources:* Machine learning platform, data analysis tools. *Metrics:* Accuracy of personalized recommendations, user engagement with recommendations.
- **Task 2:** Integrate personalized recommendations into the Leaf app. *Timeline:* 8 weeks. *Responsibility:* Engineering Team. *Resources:* Development tools, API integration tools. *Metrics:* User adoption rates of personalized recommendations, changes in user behavior.

**IV. Product Development & Research:**

**Further Research:**

- **Task 1:** Design and conduct research studies to investigate the influence of other factors (age, weight, fitness level, stress, diet) on activity and sleep patterns. *Timeline:* Ongoing. *Responsibility:* Research Team & Data Science Team. *Resources:* Research grants, data collection tools. *Metrics:* Research publications, insights gained on influencing factors.

**Outlier Analysis:**

- **Task 1:** Develop methods for identifying and analyzing outliers in user data.
  *Timeline:* 4 weeks. *Responsibility:* Data Science Team. *Resources:* Data analysis tools, outlier detection algorithms. *Metrics:* Number of outliers identified, characteristics of outliers.
- **Task 2:** Investigate the reasons for outliers (measurement errors, unique user segments) and develop strategies for addressing them. *Timeline:* 6 weeks. *Responsibility:* Data Science Team & Product Management Team. *Resources:* User research tools, data visualization tools. *Metrics:* Insights gained from outlier analysis, recommendations for product development.

**Integration with Other Apps: (Already covered above in I. Holistic Activity Tracking)**