Bellabeat Case Study

18/04/2024

1. Ask

Business Task:

Bellabeat, a company that manufactures health-focused smart devices, wants to understand how consumers use their fitness trackers. The goal is to analyze data from similar devices and provide insights that will help Bellabeat shape an effective marketing strategy to grow its user base and increase product engagement.

2. Prepare

Data Source:

- Dataset: FitBit Fitness Tracker Data from Kaggle
- Link: https://www.kaggle.com/datasets/arashnic/fitbit
- Time Period: March 12 to May 12, 2016

Credibility Check (ROCCC):

- Reliable: From consistent sources (Fitbit API).
- Original: User-generated data with consent.
- Comprehensive: Covers various health metrics.
- Current: Slightly outdated (2016), but patterns are still relevant.
- Cited: Provided through Kaggle.

Privacy and Ethics:

- Data is anonymized and shared with consent.

3. Process

Tools Used:

- Google Sheets: For quick exploratory data review and cleaning
- R and RStudio: For in-depth analysis and visualizations
- R Packages: tidyverse, ggplot2, dplyr.

Data Cleaning Steps and Preparation:

1. Google Sheets: Reviewed raw CSVs for structure, checked for nulls and duplicates.

   Sample formula used:

   =COUNTBLANK(A2:Z1000) ->To identify missing values =UNIQUE(A2:A1000) ->To identify duplicates

2. RStudio: Imported and cleaned the data programmatically.

Load Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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 error
```

```
library(ggplot2)
library(here)
```

```
## here() starts at /cloud/project
```

Load CSV Files

```
daily_activity <- read.csv(here("case_study_1_fitbit", "daily_activity_merged.csv"))
sleep_day <- read.csv(here("case_study_1_fitbit", "sleepDay_merged.csv"))
```

View First Few Rows

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   04-12-2016      13162          8.50            8.50
## 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
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

```
head(sleep_day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366      04-12-2016 00:00                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

Column Names

```
colnames(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"
```

```
colnames(sleep_day)
```

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

Both datasets have a common key: **"Id"**, which can be used to join them.

Number of Unique Participants and Observations

```
n_distinct(daily_activity$Id)  # 33 participants
```

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

```
n_distinct(sleep_day$Id)       # 24 participants
```

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

```
nrow(daily_activity)  # 940 observations(Records)
```

```
## [1] 940
```

```
nrow(sleep_day)       # 413 observations(Records)
```

```
## [1] 413
```

```
#This shows that fewer users are consistently using the sleep tracker compared to activity tracker. We
```

Summary Statistics

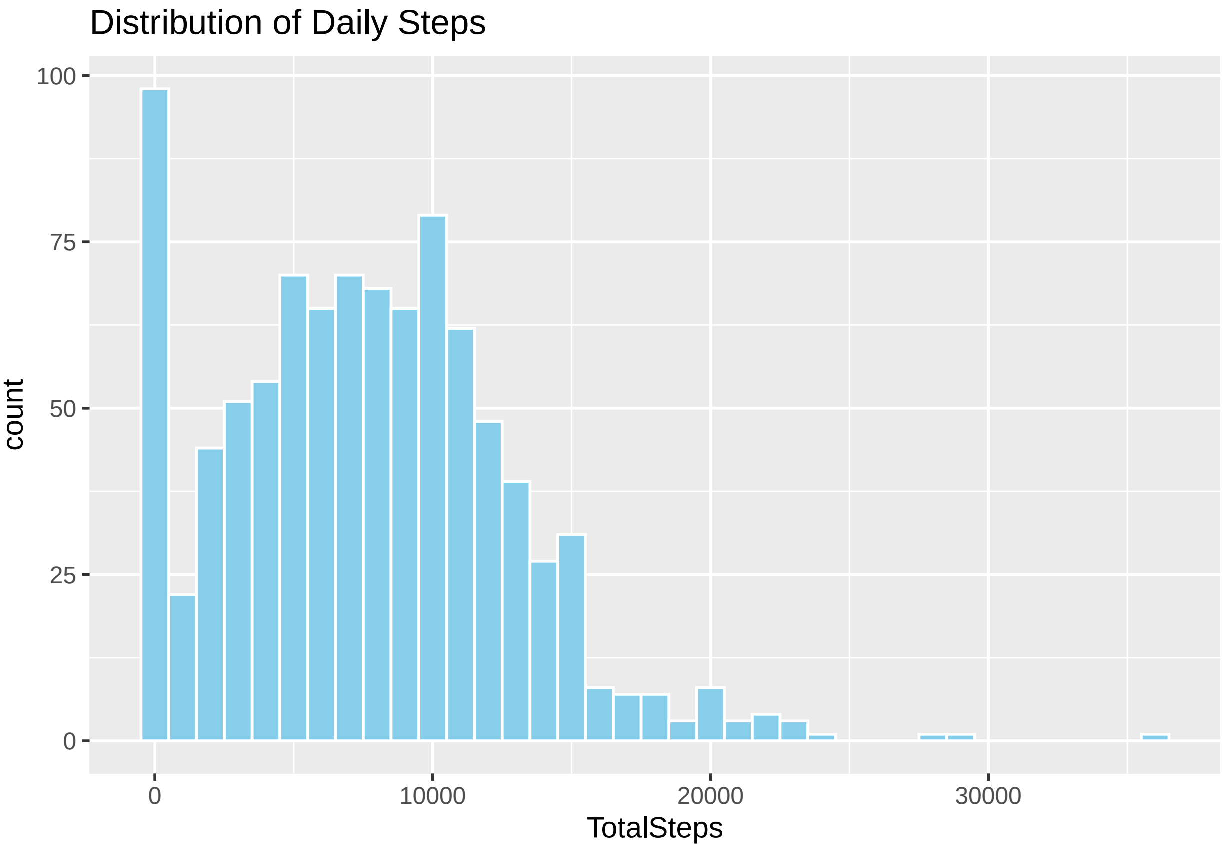
Daily Activity Summary

```
daily_activity %>%
  select(TotalSteps, TotalDistance, SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8
##  Median : 7406   Median : 5.245   Median :1057.5
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5
##  Max.   :36019   Max.   :28.030   Max.   :1440.0
```

##“Average daily steps: ~7,500”

```
ggplot(daily_activity, aes(x = TotalSteps)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "white") +
  labs(title = "Distribution of Daily Steps")
```



Sleep Summary

```
sleep_day %>%
  select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%
  summary()
```

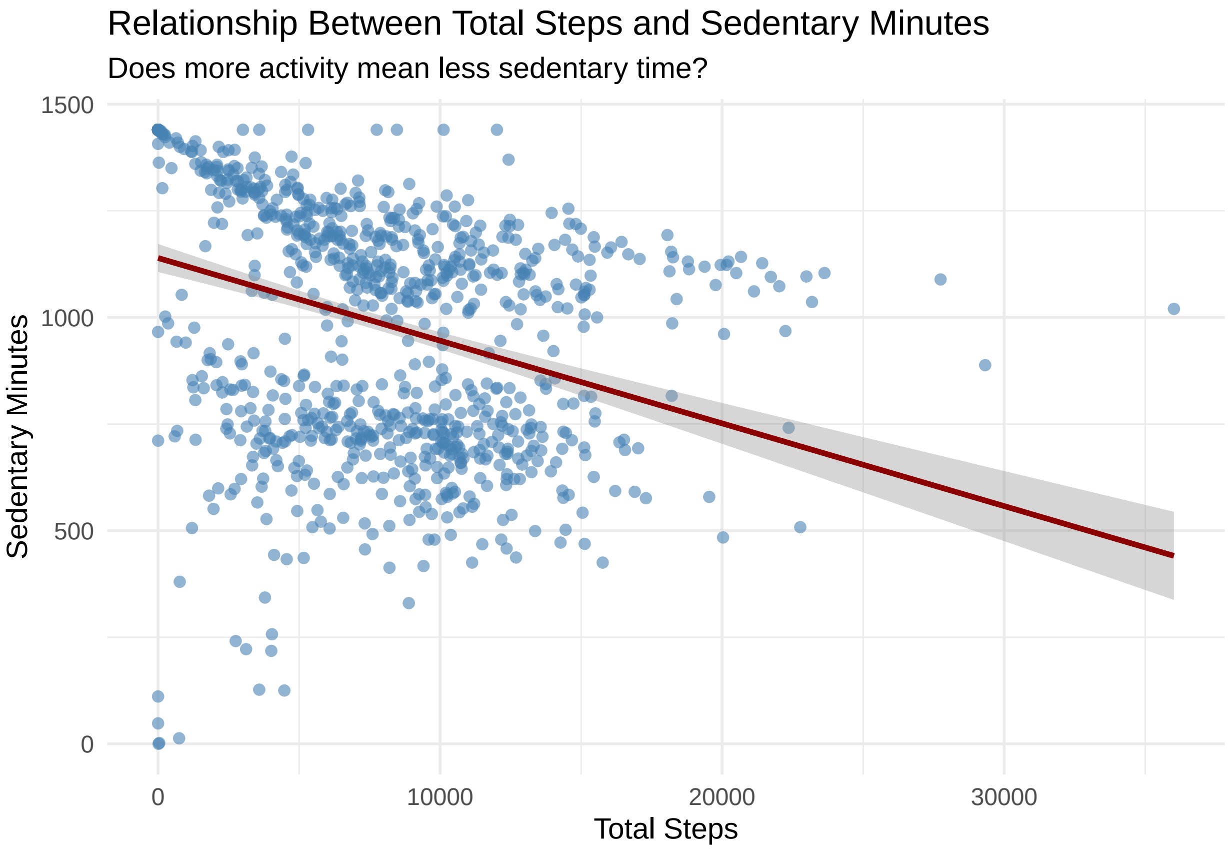
```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0 
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0 
##  Median :1.000     Median :433.0      Median :463.0 
##  Mean   :1.119     Mean   :419.5      Mean   :458.6 
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0 
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

Visualization: Relationships

I. Steps vs. Sedentary Minutes

```
ggplot(data = daily_activity, aes(x = TotalSteps, y = SedentaryMinutes)) +
  geom_point(color = "steelblue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "darkred", se = TRUE) +
  labs(
    title = "Relationship Between Total Steps and Sedentary Minutes",
    subtitle = "Does more activity mean less sedentary time?",
    x = "Total Steps",
    y = "Sedentary Minutes"
  ) +
  theme_minimal()
```

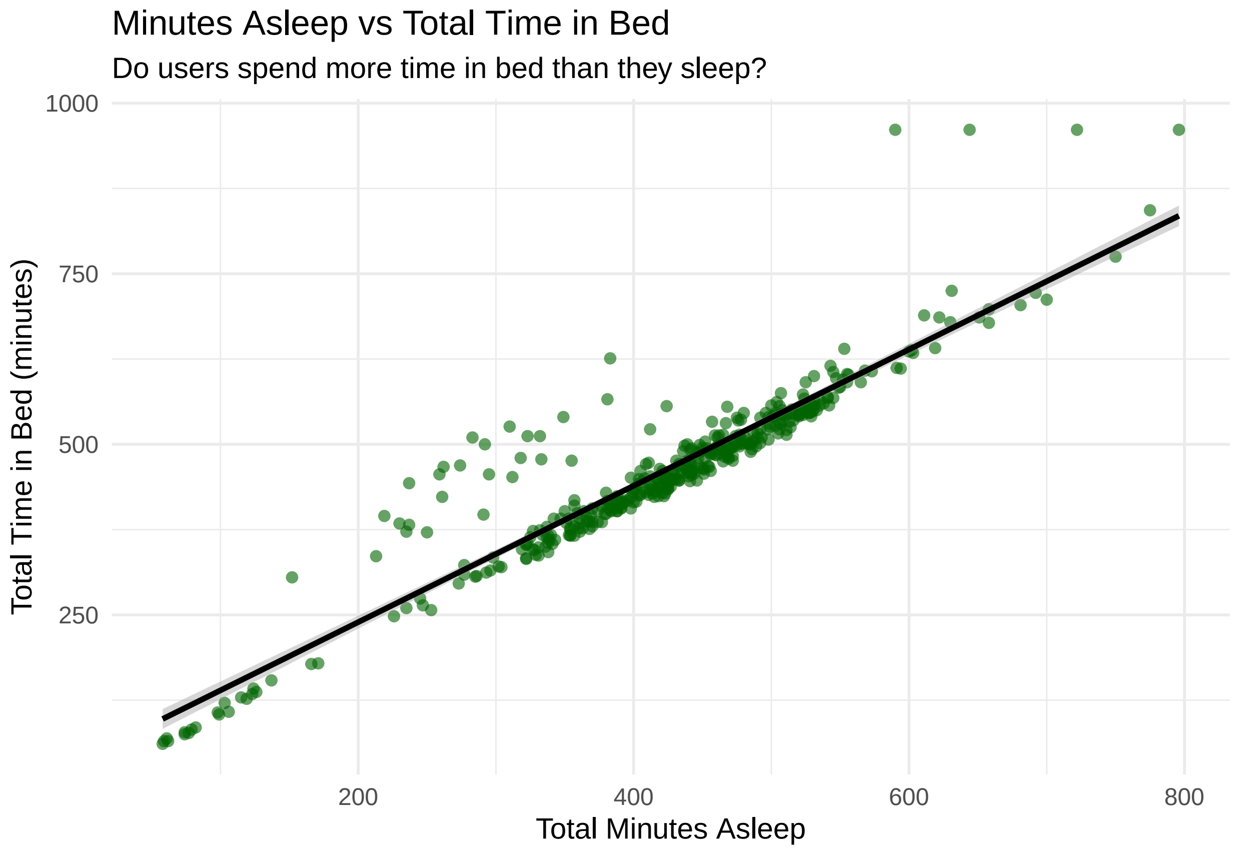
```
## `geom_smooth()` using formula = 'y ~ x'
```



II. Minutes Asleep vs. Time in Bed

```
ggplot(data = sleep_day, aes(x = TotalMinutesAsleep, y = TotalTimeInBed)) +
  geom_point(color = "darkgreen", alpha = 0.6) +
  geom_smooth(method = "lm", color = "black", se = TRUE) +
  labs(
    title = "Minutes Asleep vs Total Time in Bed",
    subtitle = "Do users spend more time in bed than they sleep?",
    x = "Total Minutes Asleep",
    y = "Total Time in Bed (minutes)"
  ) +
  theme_minimal()
```

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



Merge Datasets

```
combined_data <- merge(sleep_day, daily_activity, by = "Id")
write.csv(combined_data, "combined_fitbit_data.csv", row.names = FALSE)

combined_data1 <- full_join(sleep_day, daily_activity, by = "Id")
```

```
## Warning in full_join(sleep_day, daily_activity, by = "Id"): Detected an unexpected many-to-many rela
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
write.csv(combined_data1, "combined_fitbit_data1.csv", row.names = FALSE)

n_distinct(combined_data$Id)    # 24
```

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

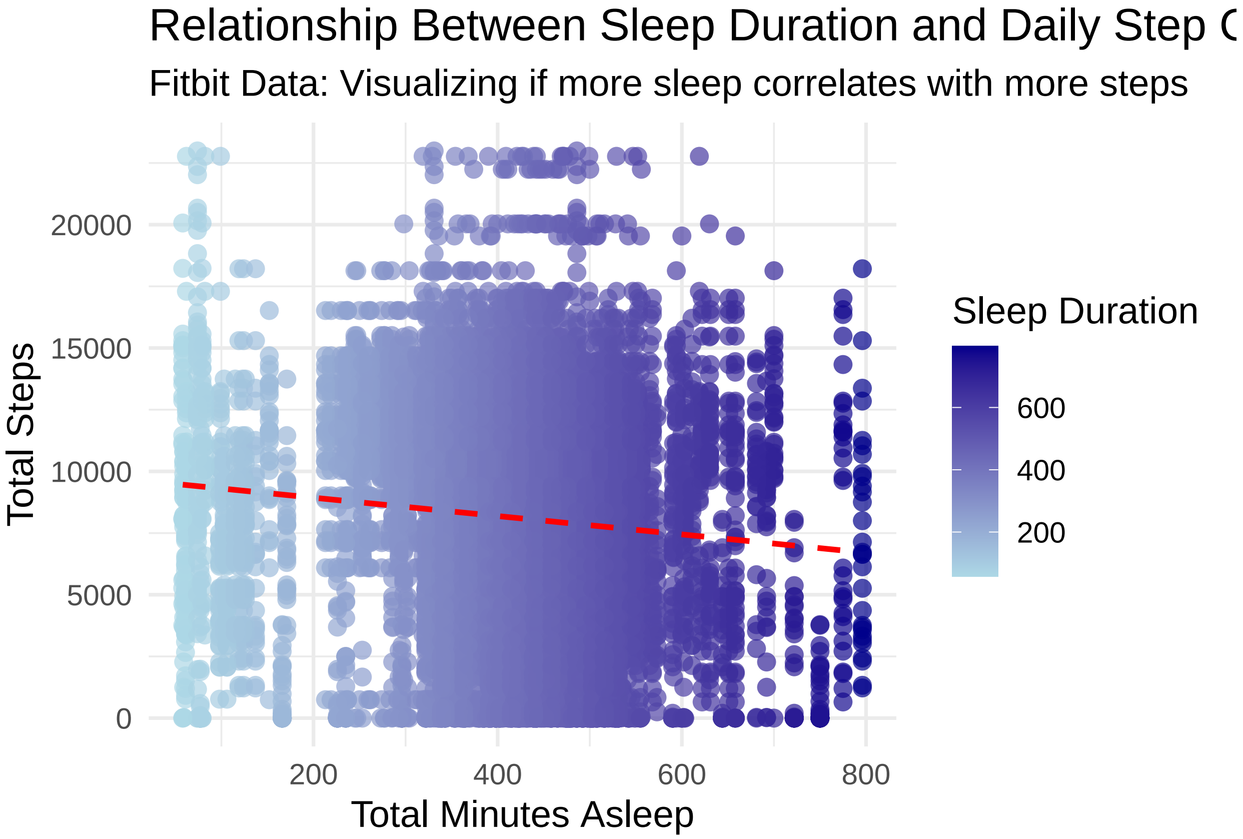
```
n_distinct(combined_data1$Id)   # 33
```

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

III. Sleep vs. Steps with Gradient Coloring

```
ggplot(data = combined_data, aes(x = TotalMinutesAsleep, y = TotalSteps)) +
  geom_point(
    aes(color = TotalMinutesAsleep),
    size = 3,
    alpha = 0.7,
    shape = 16
  ) +
  scale_color_gradient(low = "lightblue", high = "darkblue") +
  geom_smooth(method = "lm", color = "red", se = FALSE, linetype = "dashed") +
  labs(
    title = "Relationship Between Sleep Duration and Daily Step Count",
    subtitle = "Fitbit Data: Visualizing if more sleep correlates with more steps",
    x = "Total Minutes Asleep",
    y = "Total Steps",
    color = "Sleep Duration"
  ) +
  theme_minimal(base_size = 14)
```

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



Additional Visualizations

IV. Calories Burned vs Total Steps

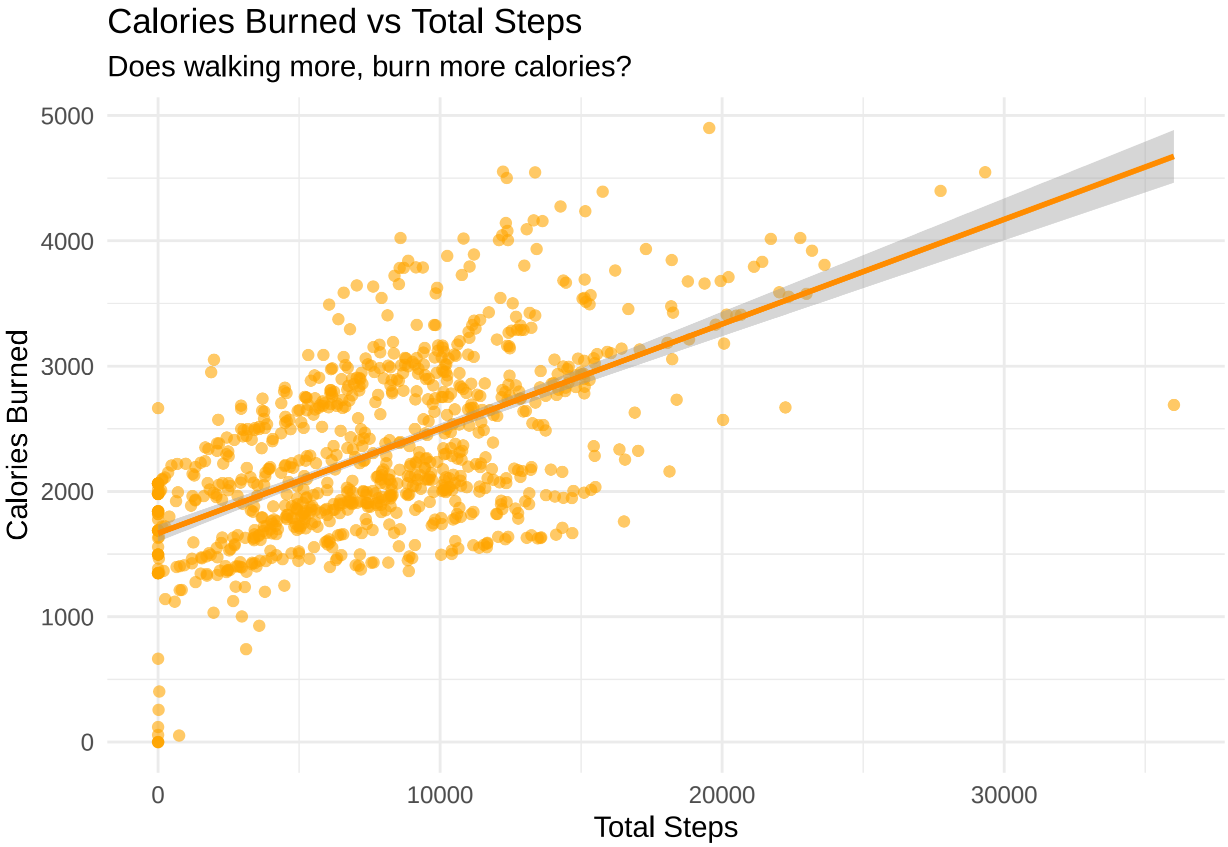
```
# correlation b/w total steps and calories burned

cor(daily_activity$TotalSteps, daily_activity$Calories, use = "complete.obs")
```

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

```
# If this gives a correlation coefficient (r) around 0.5 or higher, it indicates a positive relationshi
```

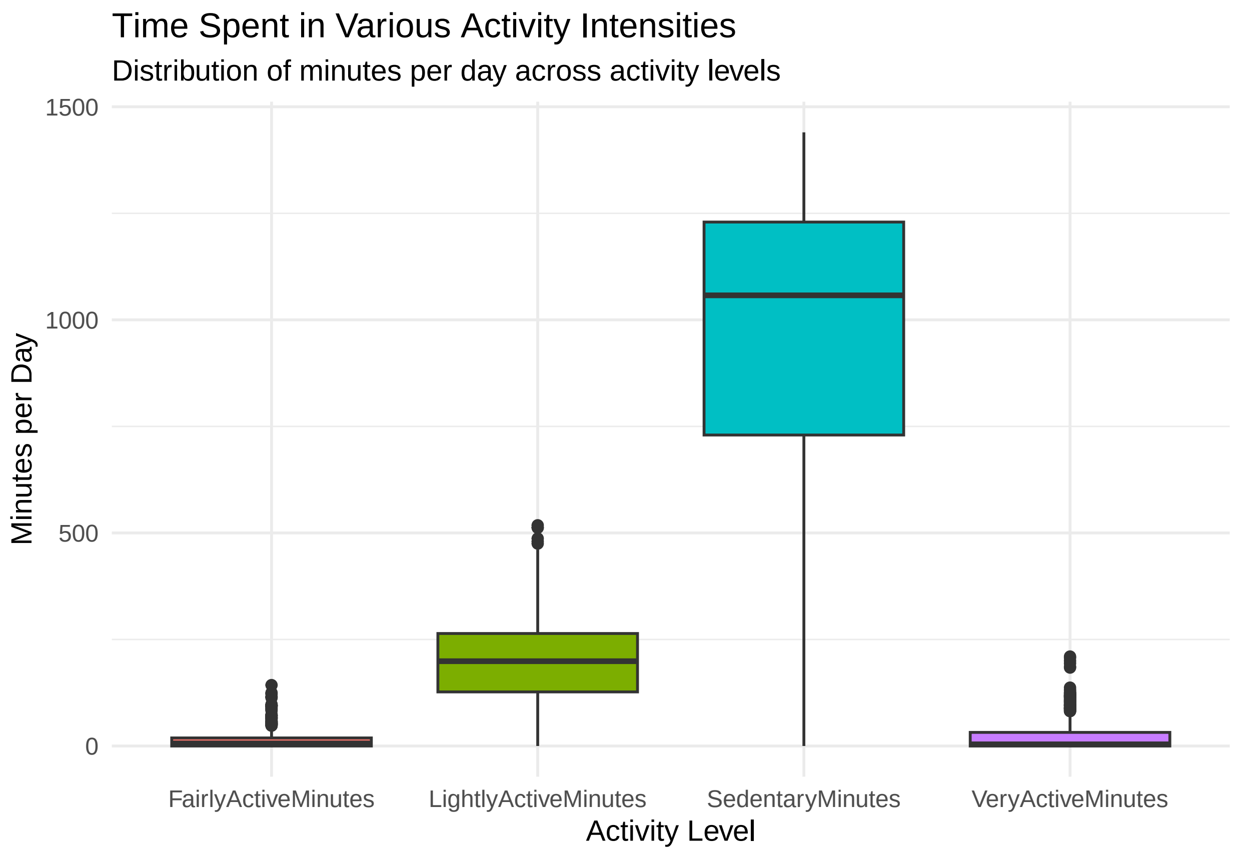
```
ggplot(data = daily_activity, aes(x = TotalSteps, y = Calories)) +
  geom_point(color = "orange", alpha = 0.6) +
  geom_smooth(method = "lm", color = "darkorange", se = TRUE) +
  labs(
    title = "Calories Burned vs Total Steps",
    subtitle = "Does walking more, burn more calories?",
    x = "Total Steps",
    y = "Calories Burned"
  ) +
  theme_minimal()
```

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



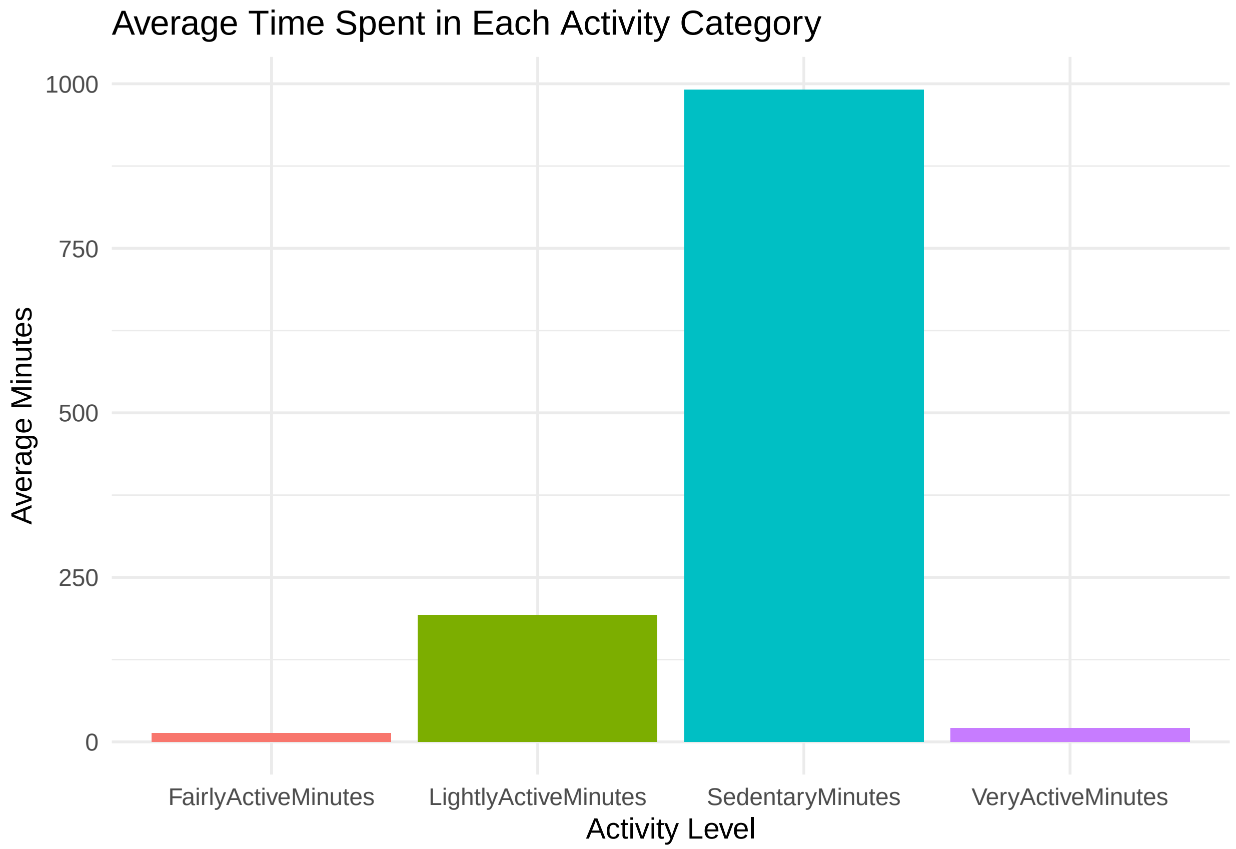
V. Time Spent in Various Activity Intensities

```
intensity_data <- daily_activity %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes) %>%
  pivot_longer(cols = everything(), names_to = "ActivityLevel", values_to = "Minutes")

ggplot(intensity_data, aes(x = ActivityLevel, y = Minutes, fill = ActivityLevel)) +
  geom_boxplot() +
  labs(
    title = "Time Spent in Various Activity Intensities",
    subtitle = "Distribution of minutes per day across activity levels",
    x = "Activity Level",
    y = "Minutes per Day"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



VI. Daily Activity Breakdown - Bar Plot

```
daily_activity_long <- daily_activity %>%
  select(Id, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes) %>%
  pivot_longer(cols = -Id, names_to = "Activity", values_to = "Minutes")

ggplot(daily_activity_long, aes(x = Activity, y = Minutes, fill = Activity)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(
    title = "Average Time Spent in Each Activity Category",
    y = "Average Minutes",
    x = "Activity Level"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Analyze

Findings:

- Average daily steps: ~7,500
- Positive linear correlation between total steps and calories burned
- Sleep tracking is underutilized compared to activity tracking

5. Share

Insights Shared:

- Users walking more than 10,000 steps/day burn significantly more calories.
- Most users are more active on weekdays than weekends.
- Sleep tracking is inconsistent across users.

6. Act

Strategic Recommendations for Bellabeat:

- Encourage consistent step goals, Many users are sedentary; marketing should promote step challenges.
- Promote daily step goals and rewards for 10,000+ steps.
- Highlight calorie-tracking features and how they tie to fitness goals.
- Add push notifications or reminders to wear the device consistently.
- Educate users on the health benefits of regular sleep tracking.