Ask

The purpose of this analysis is to gain insights into consumer behavior and usage patterns of non-Bellabeat smart devices. By examining usage data from various smart devices, I aim to identify trends and key factors that influence user engagement and satisfaction. This analysis will provide actionable insights for Bellabeat to better understand the competitive landscape, inform product development strategies, and enhance marketing efforts to effectively position Bellabeat's offerings in the market.

1.2 Questions for analysis

What are some trends in smart device usage?

How could these trends apply to Bellabeat customers?

How could these trends help influence Bellabeat marketing strategy?

1.3 Stakeholders

Urška Sršen: Bellabeat's cofounder and Chief Creative Officer

Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Prepare

2.1 Dataset used

I'll be using the FitBit Fitness Tracker Data provided by Mobius on Kaggle.

2.2 Accessibility & Privacy

The licensing of this dataset confirms that it is open source and that the author has dedicated this dataset to the public domain. You can copy, modify, distribute and perform the work, even for commercial purposes, without asking for permission.

2.3 Information on dataset

This dataset is generated by respondents to a distributed survey via Amazon Mechanical Turk between 3.12.2016 - 5.12.2016. 30 eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. We will be looking at the 4.12.16-5.12.16 timeframe in particular.

2.4. Data Organization

Available are 29 .CSV documents. Each document the data is arranged in long format form with the rows containing one time point per subject. Therefore, one subject can have multiple rows of data, each containing a unique ID to the user, along with different day and time.

2.5. Data Credibility & Integrity

Since there are only 30 eligible Fitbit users included in these datasets, the sample size is very small and no further information is given on each subject. Therefore, there are certain unknown variables in this analysis, such as the age, height, weight, and activity level of each subject. As these will play a crucial role in determining whether the activity level is appropriate and proportional to the size of the individual. As present, sampling bias might be a factor in this case.

Process

3.1 Check the data for errors

After searching through websites and commentary on the Kaggle page for 'FitBit Fitness Tracker Data' I was able to find a PDF of a data dictionary that related to the data sets used.

A first scan of the data sets doesn't reveal any errors like wrong entry. Any faulty entry such as mixing user ids cannot be detected as there is no raw material in hand to compare.

3.2 Cleaning the data

We first install the necessary libraries

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
install.packages("ggpubr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(ggpubr)
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(janitor)
```

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

```
install.packages("ggrepel")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(ggrepel)
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(dplyr)
install.packages("tidyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyr)
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

Importing Datasets

```
daily_activity <- read.csv("dailyActivity_merged.csv")
hourly_calories <- read.csv("hourlyCalories_merged.csv")
hourly_intensity <- read.csv("hourlyIntensities_merged.csv")
hourly_steps <- read.csv("hourlySteps_merged.csv")
weight <- read.csv("weightLogInfo_merged.csv")
```

Previewing data

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    3/25/2016      11004          7.11            7.11
```

```
## 2 1503960366    3/26/2016      17609        11.55          11.55
## 3 1503960366    3/27/2016      12736         8.53          8.53
## 4 1503960366    3/28/2016      13231         8.93          8.93
## 5 1503960366    3/29/2016      12041         7.85          7.85
## 6 1503960366    3/30/2016      10970         7.16          7.16
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              2.57                     0.46
## 2                        0              6.92                     0.73
## 3                        0              4.66                     0.16
## 4                        0              3.19                     0.79
## 5                        0              2.16                     1.09
## 6                        0              2.36                     0.51
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                4.07                       0                33
## 2                3.91                       0                89
## 3                3.71                       0                56
## 4                4.95                       0                39
## 5                4.61                       0                28
## 6                4.29                       0                30
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  12                  205              804     1819
## 2                  17                  274              588     2154
## 3                   5                  268              605     1944
## 4                  20                  224             1080     1932
## 5                  28                  243              763     1886
## 6                  13                  223             1174     1820
```

```
head(hourly_calories)
```

```
##           Id          ActivityHour Calories
## 1 1503960366 3/12/2016 12:00:00 AM       48
## 2 1503960366  3/12/2016 1:00:00 AM       48
## 3 1503960366  3/12/2016 2:00:00 AM       48
## 4 1503960366  3/12/2016 3:00:00 AM       48
## 5 1503960366  3/12/2016 4:00:00 AM       48
## 6 1503960366  3/12/2016 5:00:00 AM       48
```

```
head(hourly_intensity)
```

```
##           Id          ActivityHour TotalIntensity AverageIntensity
## 1 1503960366 3/12/2016 12:00:00 AM              0                0
## 2 1503960366  3/12/2016 1:00:00 AM              0                0
## 3 1503960366  3/12/2016 2:00:00 AM              0                0
## 4 1503960366  3/12/2016 3:00:00 AM              0                0
## 5 1503960366  3/12/2016 4:00:00 AM              0                0
## 6 1503960366  3/12/2016 5:00:00 AM              0                0
```

```
head(hourly_steps)
```

```
##          Id          ActivityHour StepTotal
## 1 1503960366 3/12/2016 12:00:00 AM         0
## 2 1503960366  3/12/2016 1:00:00 AM         0
## 3 1503960366  3/12/2016 2:00:00 AM         0
## 4 1503960366  3/12/2016 3:00:00 AM         0
## 5 1503960366  3/12/2016 4:00:00 AM         0
## 6 1503960366  3/12/2016 5:00:00 AM         0
```

```
head(weight)
```

```
##          Id                Date WeightKg WeightPounds Fat   BMI
## 1 1503960366 4/5/2016 11:59:59 PM     53.3     117.5064  22 22.97
## 2 1927972279 4/10/2016 6:33:26 PM    129.6     285.7191  NA 46.17
## 3 2347167796 4/3/2016 11:59:59 PM     63.4     139.7731  10 24.77
## 4 2873212765 4/6/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 4/7/2016 11:59:59 PM     57.2     126.1044  NA 21.65
## 6 2891001357 4/5/2016 11:59:59 PM     88.4     194.8886  NA 25.03
##   IsManualReport       LogId
## 1           True 1.459901e+12
## 2          False 1.460313e+12
## 3           True 1.459728e+12
## 4           True 1.459987e+12
## 5           True 1.460074e+12
## 6           True 1.459901e+12
```

3.4 Clean the data

Clean the column names so they're all lowercase and unify the date and time column names for when we later format the date.

```
daily_activity <- daily_activity %>%
  rename(date = ActivityDate) %>%
  clean_names()

hourly_calories <- hourly_calories %>%
  rename(date_time = ActivityHour) %>%
  clean_names()

hourly_intensity <- hourly_intensity %>%
  rename(date_time = ActivityHour) %>%
  clean_names()

hourly_steps <- hourly_steps %>%
  rename(date_time = ActivityHour) %>%
  rename(total_steps = StepTotal) %>%
  clean_names()
```

Count the number of participants

Counting the number of IDs will give us the count of the participants for each dataframe.

```
n_distinct(daily_activity$id)
```

```
## [1] 35
```

```
n_distinct(hourly_calories$id)
```

```
## [1] 34
```

```
n_distinct(hourly_intensity$id)
```

```
## [1] 34
```

```
n_distinct(hourly_steps$id)
```

```
## [1] 34
```

```
n_distinct(weight$Id)
```

```
## [1] 11
```

As we see, the number of unique ids doesn't match the number of users surveyed. Two potential reasons: - User switched device - Faulty entry in the data set

We cannot clean this issue as we don't have the original data to compare to and we don't have an information whether some users changes their devices or settings.

Only 11 users reported their weight log. We can already assume this is a weak sample to analyze.

Check for duplicates

```
sum(duplicated(daily_activity))
```

```
## [1] 0
```

```
sum(duplicated(hourly_calories))
```

```
## [1] 0
```

```
sum(duplicated(hourly_intensity))
```

```
## [1] 0
```

```
sum(duplicated(hourly_steps))
```

```
## [1] 0
```

```
sum(duplicated(weight))
```

```
## [1] 0
```

Luckily, we don't have any duplicates.

Adjust format

Adjust the date format so that there are no inconsistencies. Then separate the date and time of the hourly data frames into separate columns so we can use it for visualization.

```
# Format date
daily_activity$date <- as.Date(daily_activity$date, format = "%m/%d/%Y")

# Format date time
hourly_calories$date_time = as.POSIXct(hourly_calories$date_time, format="%m/%d/%Y
%I:%M:%S %p", tz=Sys.timezone())
hourly_calories$date <- format(hourly_calories$date_time, format = "%m/%d/%y")
hourly_calories$time <- format(hourly_calories$date_time, format = "%H:%M:%S")


hourly_intensity$date_time = as.POSIXct(hourly_intensity$date_time, format="%m/%d/%Y
%I:%M:%S %p", tz=Sys.timezone())
hourly_intensity$date <- format(hourly_intensity$date_time, format = "%m/%d/%y")
hourly_intensity$time <- format(hourly_intensity$date_time, format = "%H:%M:%S")

hourly_steps$date_time = as.POSIXct(hourly_steps$date_time, format="%m/%d/%Y
%I:%M:%S %p", tz=Sys.timezone())
hourly_steps$date <- format(hourly_steps$date_time, format = "%m/%d/%y")
hourly_steps$time <- format(hourly_steps$date_time, format = "%H:%M:%S")
```

Analyze

```
head(daily_activity)
```

```
##          id      date total_steps total_distance tracker_distance
## 1 1503960366 2016-03-25       11004          7.11            7.11
## 2 1503960366 2016-03-26       17609         11.55           11.55
## 3 1503960366 2016-03-27       12736          8.53            8.53
## 4 1503960366 2016-03-28       13231          8.93            8.93
## 5 1503960366 2016-03-29       12041          7.85            7.85
## 6 1503960366 2016-03-30       10970          7.16            7.16
##   logged_activities_distance very_active_distance moderately_active_distance
## 1                          0                 2.57                       0.46
## 2                          0                 6.92                       0.73
## 3                          0                 4.66                       0.16
## 4                          0                 3.19                       0.79
## 5                          0                 2.16                       1.09
## 6                          0                 2.36                       0.51
##   light_active_distance sedentary_active_distance very_active_minutes
## 1                  4.07                        0                   33
## 2                  3.91                        0                   89
## 3                  3.71                        0                   56
## 4                  4.95                        0                   39
## 5                  4.61                        0                   28
```

```
## 6              4.29                  0              30
##   fairly_active_minutes lightly_active_minutes sedentary_minutes calories
## 1                    12                   205              804     1819
## 2                    17                   274              588     2154
## 3                     5                   268              605     1944
## 4                    20                   224             1080     1932
## 5                    28                   243              763     1886
## 6                    13                   223             1174     1820
```

```
summary(select(daily_activity, -id, -date))
```

```
##   total_steps    total_distance   tracker_distance logged_activities_distance
##  Min.   :    0   Min.   : 0.000   Min.   : 0.00    Min.   :0.0000
##  1st Qu.: 1988   1st Qu.: 1.410   1st Qu.: 1.28    1st Qu.:0.0000
##  Median : 5986   Median : 4.090   Median : 4.09    Median :0.0000
##  Mean   : 6547   Mean   : 4.664   Mean   : 4.61    Mean   :0.1794
##  3rd Qu.:10198   3rd Qu.: 7.160   3rd Qu.: 7.11    3rd Qu.:0.0000
##  Max.   :28497   Max.   :27.530   Max.   :27.53    Max.   :6.7271
##  very_active_distance moderately_active_distance light_active_distance
##  Min.   : 0.000       Min.   :0.0000             Min.   : 0.00
##  1st Qu.: 0.000       1st Qu.:0.0000             1st Qu.: 0.87
##  Median : 0.000       Median :0.0200             Median : 2.93
##  Mean   : 1.181       Mean   :0.4786             Mean   : 2.89
##  3rd Qu.: 1.310       3rd Qu.:0.6700             3rd Qu.: 4.46
##  Max.   :21.920       Max.   :6.4000             Max.   :12.51
##  sedentary_active_distance very_active_minutes fairly_active_minutes
##  Min.   :0.000000          Min.   :  0.00      Min.   :  0.00
##  1st Qu.:0.000000          1st Qu.:  0.00      1st Qu.:  0.00
##  Median :0.000000          Median :  0.00      Median :  1.00
##  Mean   :0.001904          Mean   : 16.62      Mean   : 13.07
##  3rd Qu.:0.000000          3rd Qu.: 25.00      3rd Qu.: 16.00
##  Max.   :0.100000          Max.   :202.00      Max.   :660.00
##  lightly_active_minutes sedentary_minutes    calories
##  Min.   :  0.0          Min.   :  32.0    Min.   :   0
##  1st Qu.: 64.0          1st Qu.: 728.0    1st Qu.:1776
##  Median :181.0          Median :1057.0    Median :2062
##  Mean   :170.1          Mean   : 995.3    Mean   :2189
##  3rd Qu.:257.0          3rd Qu.:1285.0    3rd Qu.:2667
##  Max.   :720.0          Max.   :1440.0    Max.   :4562
```

CDC recommends 10000 steps per day for adults. Yet, we see that the average daily steps are way below that at 6547 steps. This is an indication that the majority of people surveyed don't live an active lifestyle.

This is also noticeable in the summary of sedentary_minutes where at least 85% of people are resting more than 12 sedentary hours. If we exclude 8 hours of sleep, that means at least 4 hours of sitting or laying down. According to the charity Just Stand, the following thresholds determine a person's risk of developing health problems due to sitting: Low risk: Sitting for less than 4

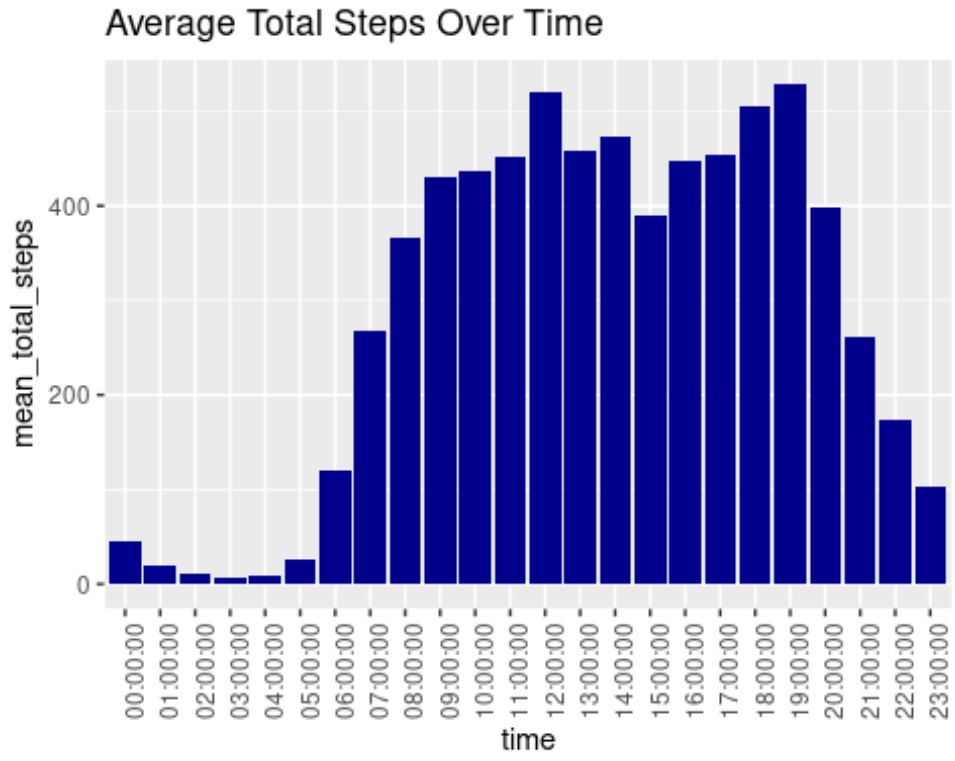
hours per day. Medium risk: Sitting for 4–8 hours per day. High risk: Sitting for 8–11 hours per day.

The majority of our sample is classified in either medium or high risk.

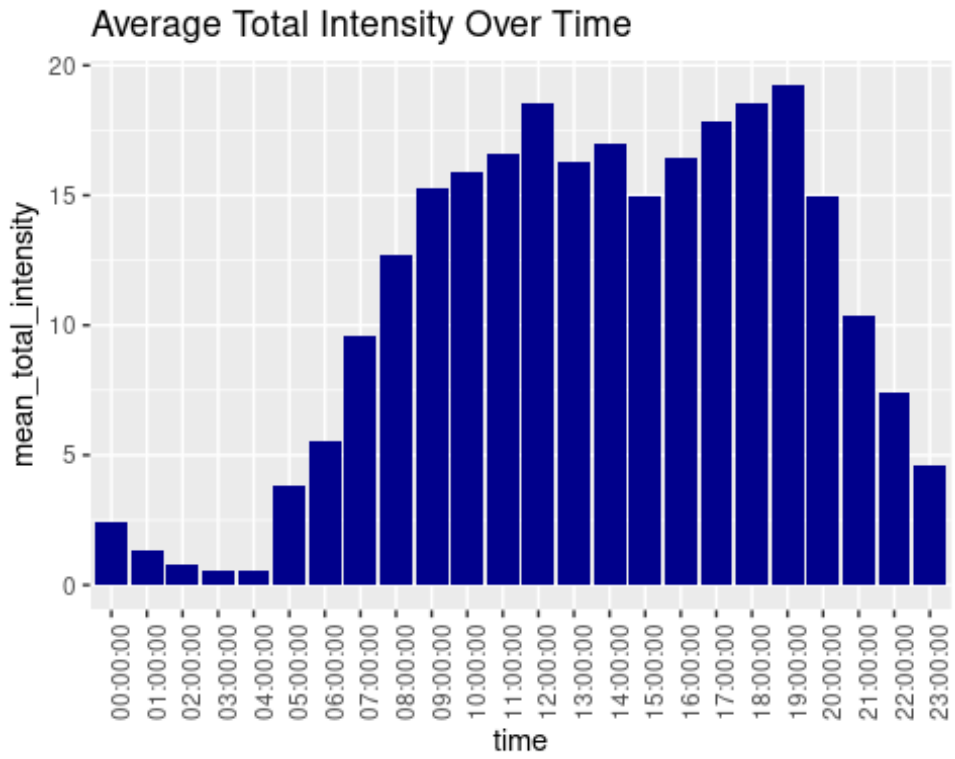
This observation prompts further inquiry into the frequency of individuals' walking habits throughout the day and at which times they exert the most intensity. Understanding these patterns can provide valuable insights into how users engage with physical activity and inform strategies to promote healthier lifestyles.

Hourly Steps and Intensity Over Time

```
steps_over_time <- hourly_steps %>%
  group_by(time) %>%
  drop_na() %>%
  summarize(mean_total_steps = mean(total_steps))

ggplot(data=steps_over_time, aes(x=time, y=mean_total_steps)) + geom_histogram(stat =
"identity", fill='darkblue') +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Average Total Steps Over Time")

## Warning in geom_histogram(stat = "identity", fill = "darkblue"): Ignoring
## unknown parameters: `binwidth`, `bins`, and `pad`
```



```r
intensity_over_time <- hourly_intensity %>%
  group_by(time) %>%
  drop_na() %>%
  summarize(mean_total_intensity = mean(total_intensity))

ggplot(data=intensity_over_time, aes(x=time, y=mean_total_intensity)) + geom_histogram(stat
= "identity", fill='darkblue') +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Average Total Intensity Over Time")

## Warning in geom_histogram(stat = "identity", fill = "darkblue"): Ignoring
## unknown parameters: `binwidth`, `bins`, and `pad`
```



The hourly steps and intensity data exhibit a striking similarity, suggesting a strong correlation between the number of steps taken and the intensity of physical activity. This observation leads to the conclusion that as individuals accumulate a higher number of steps throughout the day, they also tend to engage in activities of higher intensity.

Furthermore, the data reveals a noteworthy trend: the peak in both steps and intensity occurs consistently between the hours of 5-7pm. This insight presents an opportunity for further investigation into the factors driving activity levels during this particular timeframe.

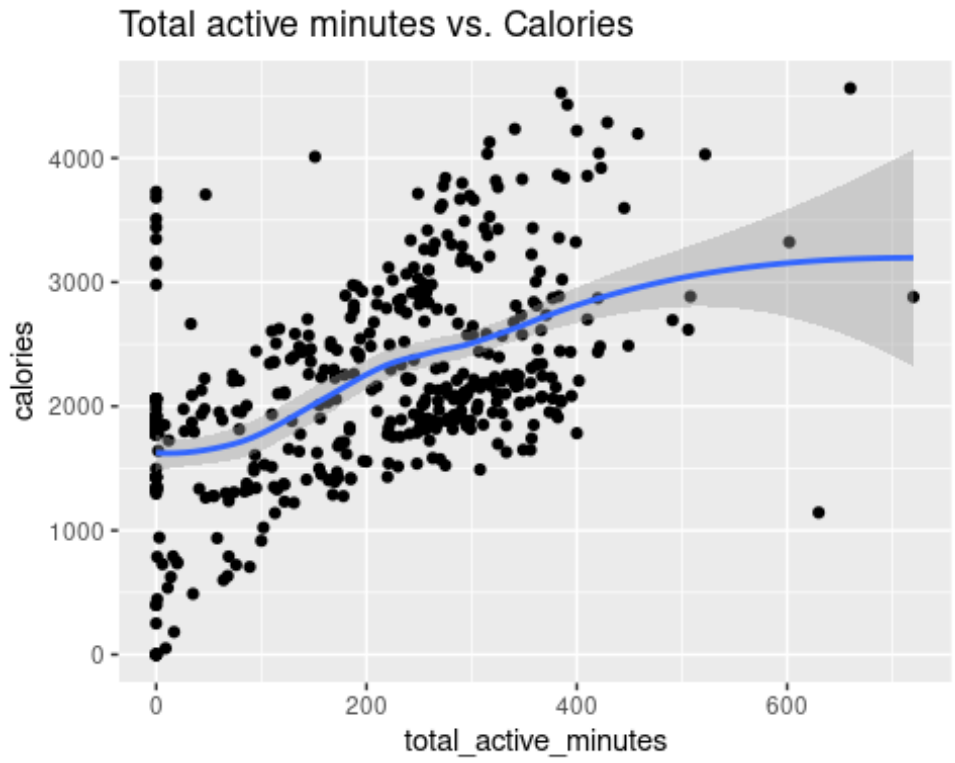
Understanding these dynamics could empower individuals to make informed decisions about their lifestyle habits and support them in achieving optimal health outcomes.

There is also an important note that there are two main peaks in hourly steps which is at 12pm and at 6pm. These two hours seem to be a good time for users to move more or maybe a strong reason to do so: lunch time and shift end. User may need to walk to the restaurant, do some post-lunch jogging to digest and later in the evening, they might tend to have a way to walk more back home ro to the gym as we notice an intensity increase in both hours.

** Comparison of Activity Levels with Calories Burned**

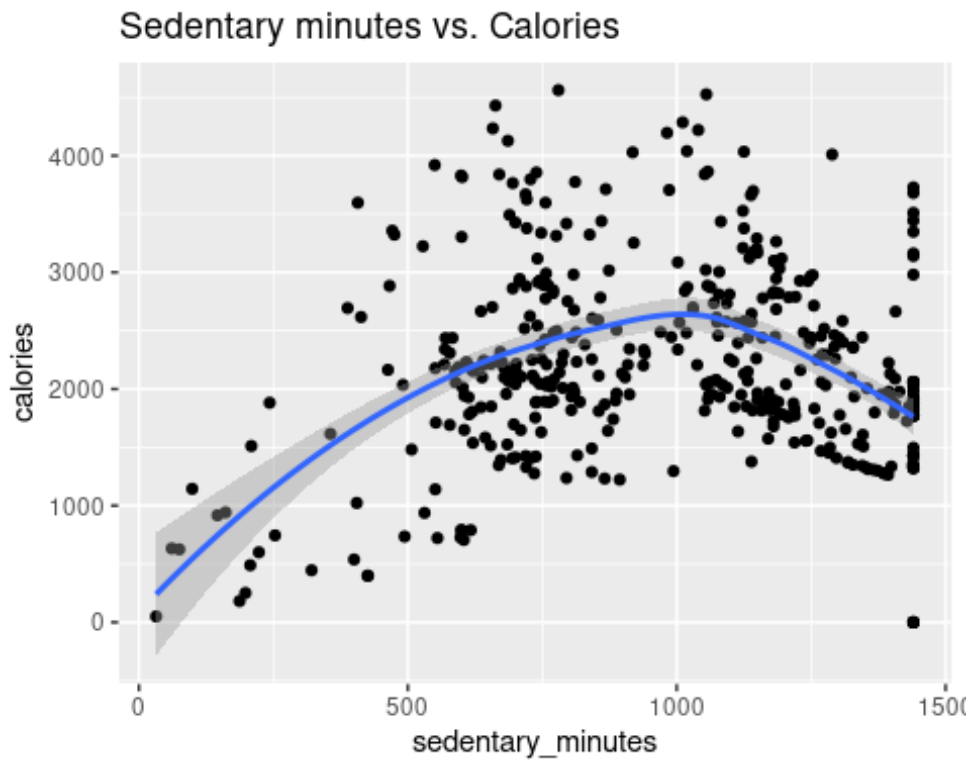
```
#Combine all of the active minutes
daily_activity$total_active_minutes <- daily_activity$very_active_minutes +
daily_activity$fairly_active_minutes + daily_activity$lightly_active_minutes

ggplot(data=daily_activity, aes(x=total_active_minutes, y=calories)) +
  geom_point() + geom_smooth() + labs(title = "Total active minutes vs. Calories")
```

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




```
ggplot(data=daily_activity, aes(x=sedentary_minutes, y=calories)) +
  geom_point() + geom_smooth() + labs(title = "Sedentary minutes vs. Calories")
```

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



As expected, the analysis reveals a positive correlation between active minutes and calories burned. This finding underscores the intuitive connection between physical activity levels and energy expenditure, indicating that individuals who engage in more active minutes are likely to burn a greater number of calories.

Share and Act

Recommendations:

- Guided Evening Workouts: Introduce guided workout sessions or routines tailored to the evening hours, accessible through the Bellabeat app or compatible devices. These workouts could include quick and effective exercises designed to fit into busy evening schedules.
- Motivational Features: Implement features that motivate users to increase their daily activity levels. This could involve setting personalized activity goals based on individual fitness levels and providing rewards or incentives for reaching milestones. Additionally, integrating social sharing capabilities could allow users to compete with friends and share achievements.
- Consider the typical day structure: the app content from workouts to tips shoud consider the daily habit of the majority of adults. We can use the information that users tend to be more active after lunchtime and after workday to provide workout programs fitting their lifestyle. This will easen the use of the app during their daily tasks and they feel it is more accustomed to their needs.

By leveraging data analytics and innovative product features, Bellabeat has the opportunity to position itself as a leader in the global smart device market, offering users personalized solutions for tracking activity, improving sleep quality, and optimizing calorie burning. As the company continues to innovate and expand its product offerings, it has the potential to make a significant impact on the lives of its users, helping them achieve their health and wellness goals and live happier, healthier lives.