Case Study Documentation

Ann Mariya

2024-02-15

Bellabeat Case Study

1 Introduction

1.1 Scenario I am a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights I discover will then help guide marketing strategy for the company. I will present your analysis to the Bellabeat executive team along with my high-level recommendations for Bellabeat's marketing strategy.

1.2 Bellabeat's Products

1. **Bellabeat app**: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
2. **Leaf**: Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.
3. **Time**: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellnes.
4. **Spring**: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.
5. **Bellabeat membership**: Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

2 Business Task

The objective is to analyze smart device usage data to understand consumer trends and how they can be applied to Bellabeat's marketing strategy for one of their products.

3 Key Stakeholders

- Urska Srsen, Cofounder and Chief Creative Officer
- Sando Mur, Cofounder and key member of executive team
- Executive Team: CMO, Product Manager
- Marketing Analytics Team

4 Data

4.1 Dataset used Fitbit Fitness Tracker Data. Click here to get the dataset.

4.2 Data Source This dataset, sourced from Kaggle, comprises personal fitness tracker data from 33 who use Fitbit. These datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016–05.12.2016. These 33 Fitbit users willingly provided their personal tracker data, including minute-by-minute details on physical activity, heart rate, and sleep monitoring. The dataset encompasses data on daily activity, step counts, and heart rate, offering insights into users' behavioral patterns and habits.

4.3 License [CC0: Public Domain] Reviewing the metadata of our dataset confirms its open-source. The contributor has generously placed the work in the public domain by renouncing all rights worldwide under copyright law, including any associated and neighboring rights, to the fullest extent permissible by law. This grants us the ability to freely copy, modify, distribute, and utilize the dataset for any purpose, including commercial endeavors, without the need for explicit authorization.

4.4 Data Organization The Fitbit dataset contains 18 csv files that are all organized in long format where each row contains a single data point for a particular item. Therefore, each item will have multiple rows of data. For example, each user ID will have multiple rows of data as data is tracked by date and time.

4.5 Limitations of Data

1. **Sample Size**: The dataset encompasses information from only 30 Fitbit users, potentially leading to sampling bias and limiting the generalizability of findings to Bellabeat's broader customer base.
2. **No Demographic Data**: The absence of demographic data further compounds the sampling bias, as insights derived may not accurately reflect the diverse characteristics of Bellabeat's target audience.
3. **Data collection period**: Additionally, the data collection period spans just one month (12/04/2016 – 12/05/2016), potentially restricting the depth of analysis and overlooking seasonal variations or long-term trends in consumer behavior.

5 Processing Data for Analysis

5.1 Tools used:

- Excel
- SQL
- SQL Server Management Studio

5.2 Tables we will need:

- Daily Activity
- Daily Calories
- Daily Intensities
- Daily Steps
- Daily Sleep
- Hourly Calories
- Hourly Intensity
- Hourly Steps
- Weights

We are trying to gather insights that would help market the product 'Time'. Time tracks activity, sleep, and stress. Therefore, we only need the above-mentioned tables for our analysis.

Next, I used SQL & Excel to explore, clean, and merge the data for analysis.

5.3 Data Exploration:

- The Daily Activity Table already contains data regarding calories, intensities, and steps. Hence, we will not use the Calories, Intensities, and Steps tables.
- All tables had 33 unique customer Ids except for Daily Sleep table and Weights table. Daily Sleep table had 24 unique customer Ids while Weights only had 8 unique customer Ids. The sample size is too small for weights and hence we will not be using this table.
- I merged Daily Activity table and Daily Sleep table together.
- I merged Hourly Calories table, Hourly Intensities table, and Hourly Steps table together.
- I split the ActivityHour column into 'Date' and 'ActivityHour'.

5.4 Merged Tables:

1. Daily_Data.csv: 943 rows, 17 columns
2. Hourly_Data.csv: 22099 rows, 7 columns

5.5 Using Excel for further cleaning:

1. Daily_Data.csv:
   - Found and removed 3 duplicates
   - Dropped 'Logged_Distance' column as it only contained 0 and NULL values.
   - Replaced NULL values in columns 'Total_Sleep_Records', 'Total_Minutes_Asleep', and 'Total_Bed_Time' with 0.
2. Hourly_Data.csv:
   - No duplicates
   - Changed 'ActivityHour' format

5.6 Final Tables for Analysis:

1. **Daily_Data.csv**: 940 rows, 16 columns
2. **Hourly_Data.csv**: 22099 rows, 7 columns

6 Data Analysis using R

6.1 Loading necessary libraries & Importing Data: Loading libraries:

```
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   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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(skimr)
library(lubridate)
```

Importing Data:

```
# import daily data
daily_data <- read_csv('Daily_Data.csv')
```

```
## Rows: 940 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
## dbl (15): Id, Total_Steps, Total_Distance, Tracker_Distance, Very_Active_Dis...
##
## 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.
```

```
# import hourly data
hourly_data <- read_csv('Hourly_Data.csv')
```

```
## Rows: 22099 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
## dbl  (5): Id, Calories, Total_Intensity, Average_Intensity, Total_Steps
## time (1): Activity_Hour
##
## 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.
```

Viewing first few rows of dataset

daily_data

```
head(daily_data)
```

```
## # A tibble: 6 x 16
##        Id Date  Total_Steps Total_Distance Tracker_Distance Very_Active_Distance
##     <dbl> <chr>       <dbl>          <dbl>            <dbl>                 <dbl>
## 1  1.50e9 12/0~       13162           8.5              8.5                   1.88
## 2  1.50e9 13/0~       10735           6.97             6.97                  1.57
## 3  1.50e9 14/0~       10460           6.74             6.74                  2.44
## 4  1.50e9 15/0~        9762           6.28             6.28                  2.14
## 5  1.50e9 16/0~       12669           8.16             8.16                  2.71
## 6  1.50e9 17/0~        9705           6.48             6.48                  3.19
## # i 10 more variables: Moderately_Active_Distance <dbl>,
## #   Sedentary_Active_Distance <dbl>, Very_Active_Minutes <dbl>,
## #   Fairly_Active_Minutes <dbl>, Lightly_Active_Minutes <dbl>,
## #   Sedentary_Minutes <dbl>, Calories <dbl>, Total_Sleep_Records <dbl>,
## #   Total_Minutes_Asleep <dbl>, Total_Bed_Time <dbl>
```

hourly_data

```
head(hourly_data)
```

```
## # A tibble: 6 x 7
##           Id Date       Activity_Hour Calories Total_Intensity Average_Intensity
##        <dbl> <chr>      <time>           <dbl>           <dbl>             <dbl>
## 1 1503960366 12/04/2016 00:00               81              20             0.333
## 2 1503960366 12/04/2016 01:00               61               8             0.133
## 3 1503960366 12/04/2016 02:00               59               7             0.117
## 4 1503960366 12/04/2016 03:00               47               0             0
## 5 1503960366 12/04/2016 04:00               48               0             0
## 6 1503960366 12/04/2016 05:00               48               0             0
## # i 1 more variable: Total_Steps <dbl>
```

6.2 Exploring Datasets: Checking the number of unique users in each dataset:

```
n_distinct(daily_data$Id)
```

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

```
n_distinct(hourly_data$Id)
```

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

Checking the number of observations in each dataset:

```
nrow(daily_data)
```

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

```
nrow(hourly_data)
```

```
## [1] 22099
```

Summarizing Data:

Summary of `daily_data`

```
glimpse(daily_data)
```

```
## Rows: 940
## Columns: 16
## $ Id                          <dbl> 1503960366, 1503960366, 1503960366, 1503960~
## $ Date                        <chr> "12/04/2016", "13/04/2016", "14/04/2016", "~
## $ Total_Steps                 <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 130~
## $ Total_Distance              <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9~
## $ Tracker_Distance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9~
## $ Very_Active_Distance        <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3~
## $ Moderately_Active_Distance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1~
## $ Sedentary_Active_Distance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Very_Active_Minutes         <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66,~
## $ Fairly_Active_Minutes       <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, ~
## $ Lightly_Active_Minutes      <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205~
## $ Sedentary_Minutes           <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 8~
## $ Calories                    <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2~
## $ Total_Sleep_Records         <dbl> 1, 2, 0, 1, 2, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1~
## $ Total_Minutes_Asleep        <dbl> 327, 384, 0, 412, 340, 700, 0, 304, 360, 32~
## $ Total_Bed_Time              <dbl> 346, 407, 0, 442, 367, 712, 0, 320, 377, 36~
```

```
colnames(daily_data)
```

```
##  [1] "Id"                          "Date"
##  [3] "Total_Steps"                 "Total_Distance"
##  [5] "Tracker_Distance"            "Very_Active_Distance"
##  [7] "Moderately_Active_Distance" "Sedentary_Active_Distance"
##  [9] "Very_Active_Minutes"         "Fairly_Active_Minutes"
## [11] "Lightly_Active_Minutes"     "Sedentary_Minutes"
## [13] "Calories"                   "Total_Sleep_Records"
## [15] "Total_Minutes_Asleep"       "Total_Bed_Time"
```

Summary of `hourly_data`

```
glimpse(hourly_data)
```

```
## Rows: 22,099
```

```
## Columns: 7
## $ Id                <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503~
## $ Date              <chr> "12/04/2016", "12/04/2016", "12/04/2016", "12/04/201~
## $ Activity_Hour     <time> 00:00:00, 01:00:00, 02:00:00, 03:00:00, 04:00:00, 0~
## $ Calories          <dbl> 81, 61, 59, 47, 48, 48, 48, 47, 68, 141, 99, 76, 73,~
## $ Total_Intensity   <dbl> 20, 8, 7, 0, 0, 0, 0, 0, 13, 30, 29, 12, 11, 6, 36, ~
## $ Average_Intensity <dbl> 0.333333, 0.133333, 0.116667, 0.000000, 0.000000, 0.~
## $ Total_Steps       <dbl> 373, 160, 151, 0, 0, 0, 0, 0, 250, 1864, 676, 360, 2~
```

```
colnames(hourly_data)
```

```
## [1] "Id"                "Date"              "Activity_Hour"
## [4] "Calories"          "Total_Intensity"   "Average_Intensity"
## [7] "Total_Steps"
```

Statistical Summary:

Statistical summary of `daily_data`

```
daily_data %>%
  select(Total_Steps,
         Total_Distance,
         Calories,
         Total_Minutes_Asleep,
         Sedentary_Minutes,
         Very_Active_Minutes,
         Fairly_Active_Minutes,
         Lightly_Active_Minutes) %>%
  summary()
```

```
##   Total_Steps    Total_Distance      Calories    Total_Minutes_Asleep
##  Min.   :    0   Min.   : 0.000   Min.   :   0   Min.   :  0.0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.:1828   1st Qu.:  0.0
##  Median : 7406   Median : 5.245   Median :2134   Median :  0.0
##  Mean   : 7638   Mean   : 5.490   Mean   :2304   Mean   :182.8
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:2793   3rd Qu.:416.2
##  Max.   :36019   Max.   :28.030   Max.   :4900   Max.   :796.0
##  Sedentary_Minutes Very_Active_Minutes Fairly_Active_Minutes
##  Min.   :   0.0    Min.   :  0.00      Min.   :  0.00
##  1st Qu.: 729.8    1st Qu.:  0.00      1st Qu.:  0.00
##  Median :1057.5    Median :  4.00      Median :  6.00
##  Mean   : 991.2    Mean   : 21.16      Mean   : 13.56
##  3rd Qu.:1229.5    3rd Qu.: 32.00      3rd Qu.: 19.00
##  Max.   :1440.0    Max.   :210.00      Max.   :143.00
##  Lightly_Active_Minutes
##  Min.   :  0.0
##  1st Qu.:127.0
##  Median :199.0
##  Mean   :192.8
##  3rd Qu.:264.0
##  Max.   :518.0
```

Statistical summary of `hourly_data`

```
hourly_data %>%
  select(Total_Steps, Calories) %>%
  summary()
```

```
##   Total_Steps         Calories
##  Min.   :    0.0   Min.   : 42.00
##  1st Qu.:    0.0   1st Qu.: 63.00
##  Median :   40.0   Median : 83.00
##  Mean   :  320.2   Mean   : 97.39
##  3rd Qu.:  357.0   3rd Qu.:108.00
##  Max.   :10554.0   Max.   :948.00
```

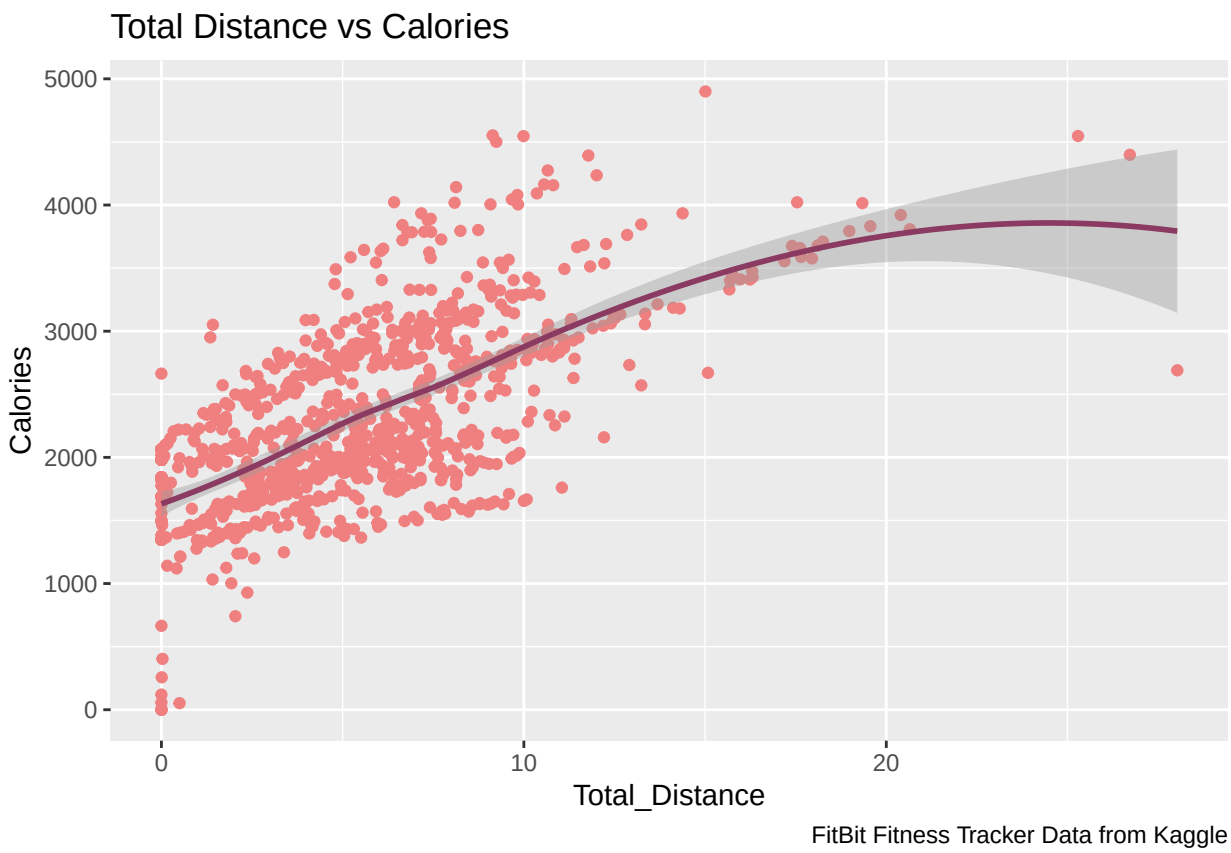
Observations:

- Average sedentary minutes per day: 991 (approximately 17 hours), constituting 71% of the entire day; Physical activity accounts for only 29% of the day or 7 hours.
- Average daily step count: 7638 steps, equivalent to 320 steps per hour.
- Predominance of light activities observed on average, indicating a focus on less strenuous physical exertion.

6.3 Data Visualization: Lets check the relationship between `Total_Distance` and `Calories`

```
ggplot(data = daily_data) +
  geom_point(mapping = aes(x = Total_Distance, y = Calories), color = "lightcoral") +
  geom_smooth(mapping = aes(x = Total_Distance, y = Calories), color = "hotpink4") +
  labs(title = "Total Distance vs Calories",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```

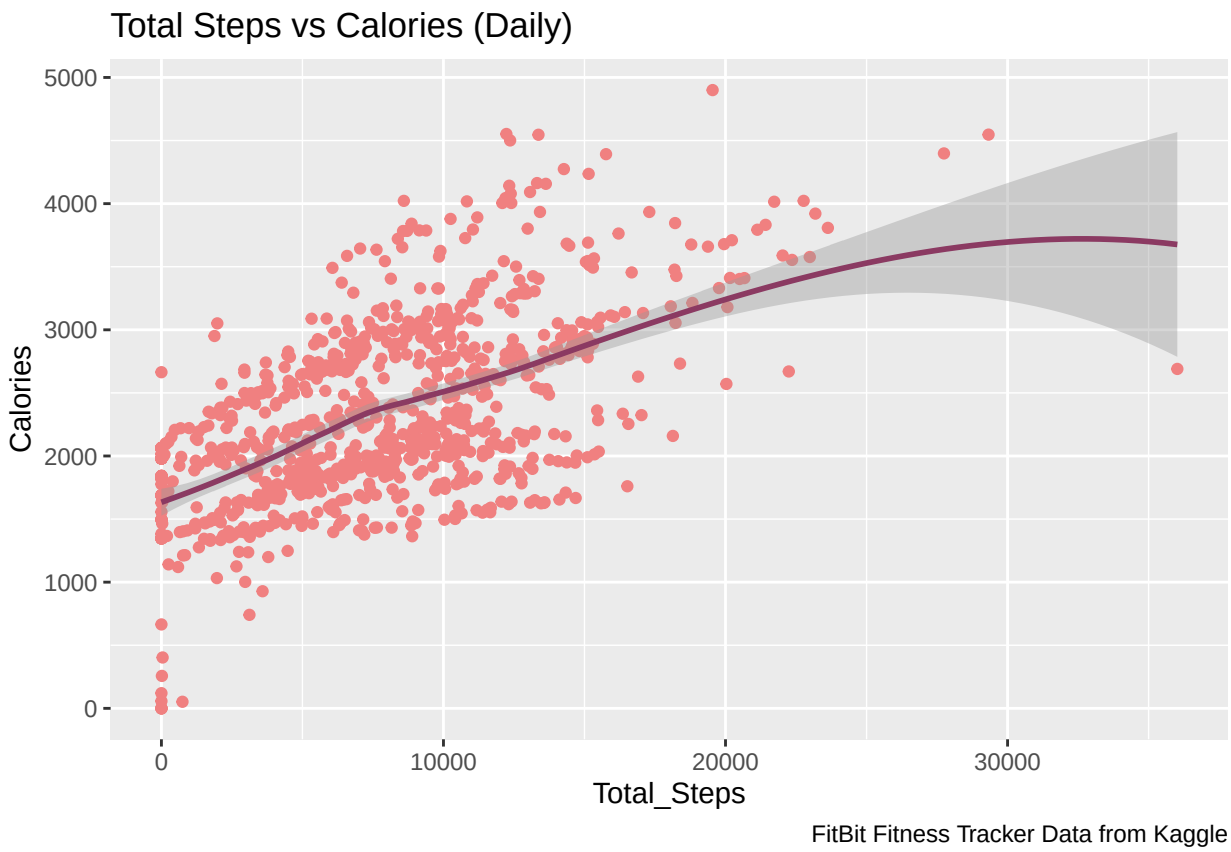
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Lets check the relationship between `Total_Steps` and `Calories` on `daily_data`

```
ggplot(data = daily_data) +
  geom_point(mapping = aes(x = Total_Steps, y = Calories), color = "lightcoral") +
  geom_smooth(mapping = aes(x = Total_Steps, y = Calories), color = "hotpink4") +
  labs(title = "Total Steps vs Calories (Daily)",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```

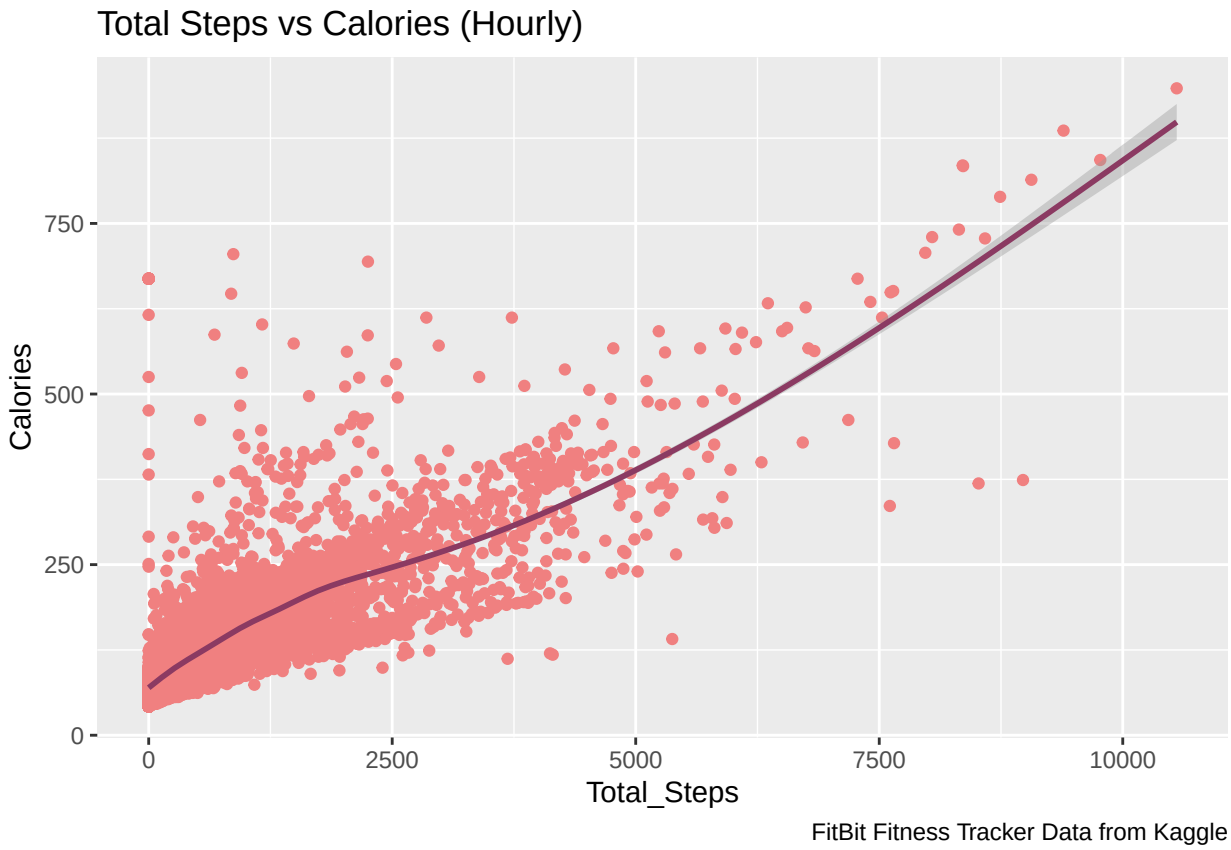
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Lets check the relationship between `Total_Steps` and `Calories` on `hourly_data`

```
ggplot(data = hourly_data) +
  geom_point(mapping = aes(x = Total_Steps, y = Calories), color = "lightcoral") +
  geom_smooth(mapping = aes(x = Total_Steps, y = Calories), color = "hotpink4") +
  labs(title = "Total Steps vs Calories (Hourly)",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Observations:

- Demonstrates a positive correlation, indicating that higher step counts result in increased calorie expenditure.
- Marketing strategies can be tailored towards individuals seeking weight loss, emphasizing how tracking steps with Bellabeat's products aids in achieving calorie burn goals.

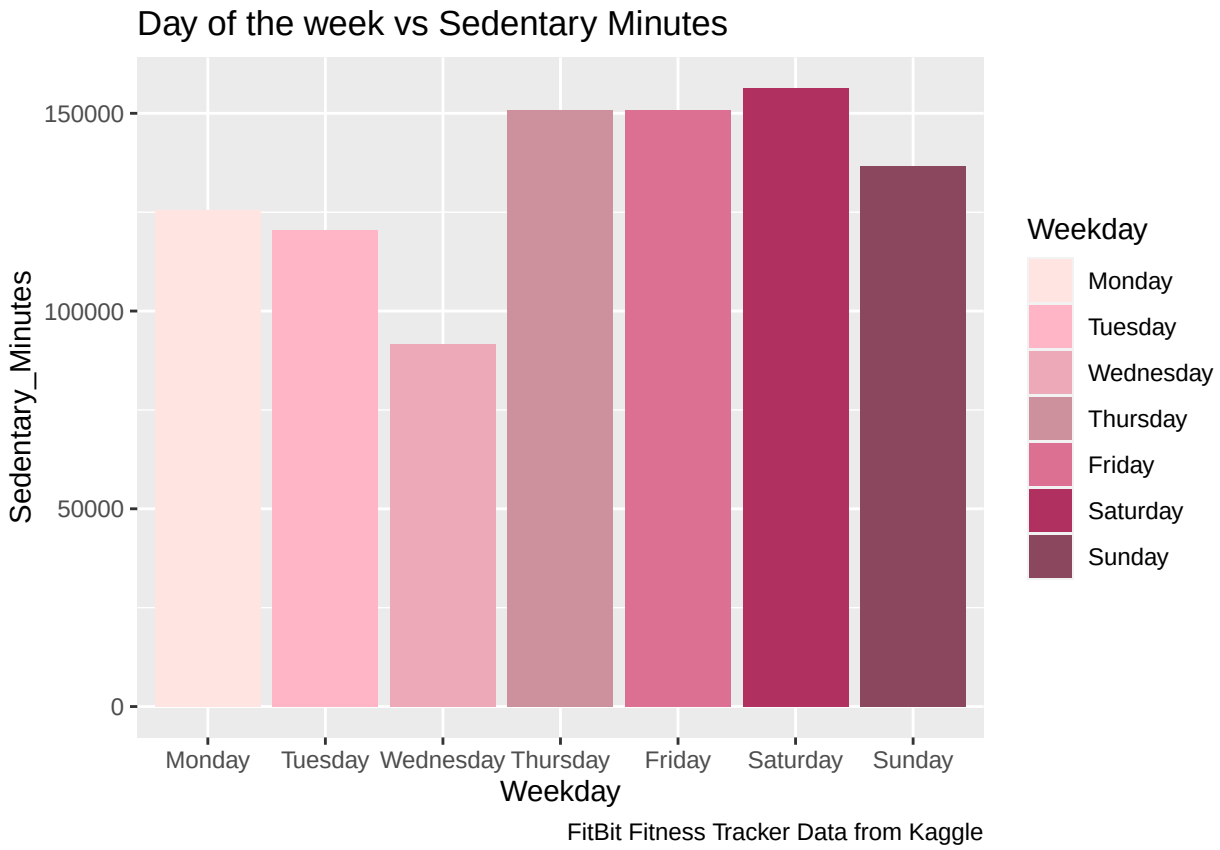
Let us now create a new column `Weekday` in `daily_data` to check trends across different days of the week.

```
daily_data_with_weekday <- mutate(daily_data, Weekday = wday(daily_data$Date, week_start = 1, label=TRU
```

Lets check how `Sedentary_Minutes` varies across different days of the week.

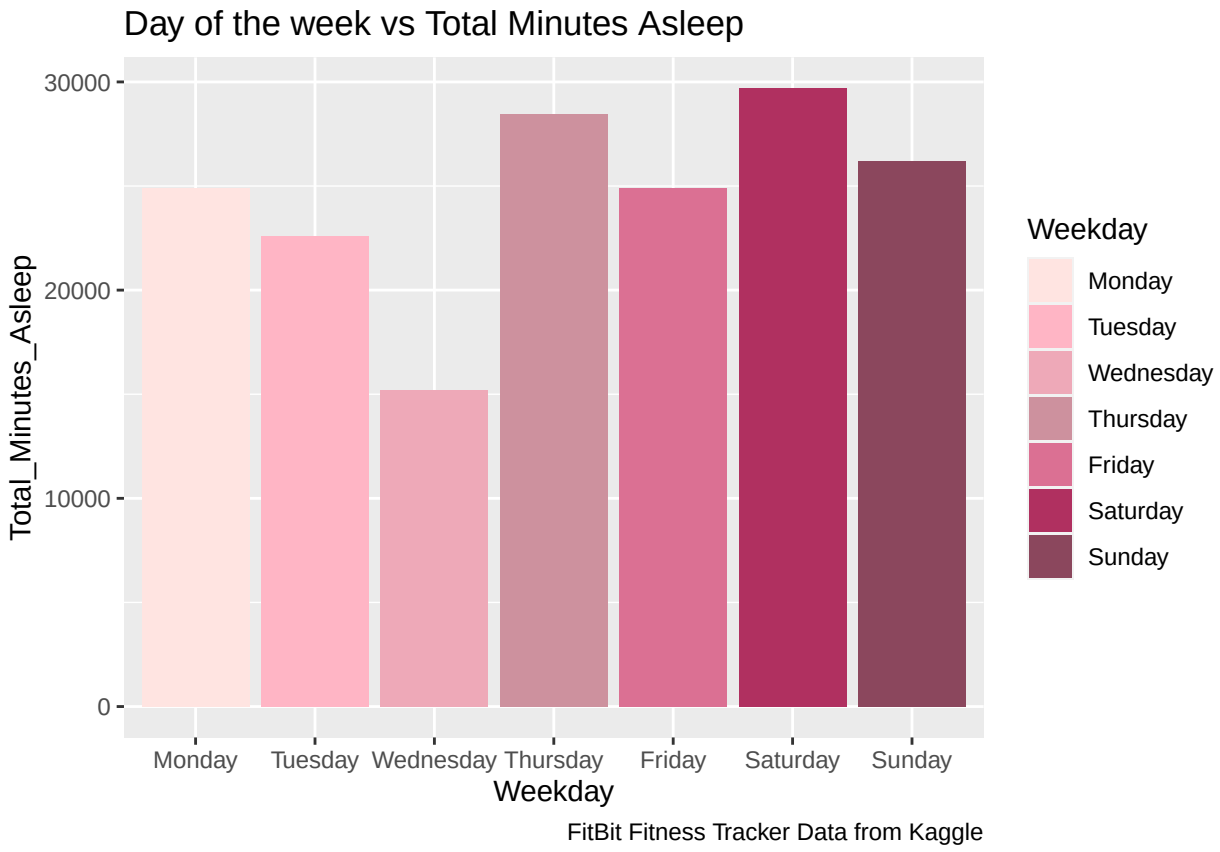
```
# creating a color palette to be used in the chart
mycolors <- c("mistyrose", "pink1", "pink2", "pink3","palevioletred",  "maroon", "palevioletred4")

ggplot(data = daily_data_with_weekday ) +
  geom_col(mapping = aes(x = Weekday, y= Sedentary_Minutes, fill = Weekday)) +
  scale_fill_manual(values=mycolors) +
  labs(title = "Day of the week vs Sedentary Minutes",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```



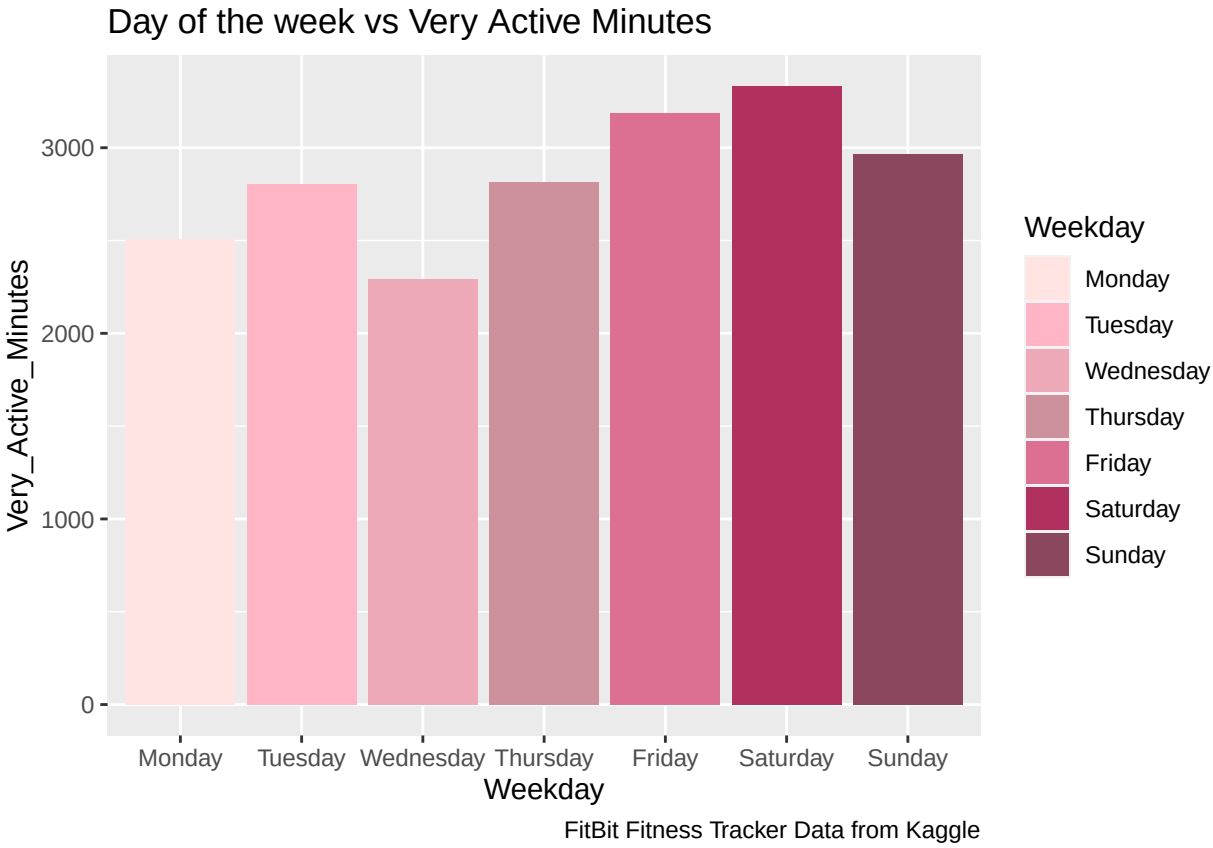
Lets check how `Total_Minutes_Asleep` varies across different days of the week.

```
ggplot(data = daily_data_with_weekday ) +
  geom_col(mapping = aes(x = Weekday, y= Total_Minutes_Asleep, fill = Weekday)) +
  scale_fill_manual(values=mycolors) +
  labs(title = "Day of the week vs Total Minutes Asleep",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```



Lets check how `Very_Active_Minutes` varies across different days of the week.

```
ggplot(data = daily_data_with_weekday ) +
  geom_col(mapping = aes(x = Weekday, y= Very_Active_Minutes, fill = Weekday)) +
  scale_fill_manual(values=mycolors) +
  labs(title = "Day of the week vs Very Active Minutes",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```



Observations:

- Sedentary minutes peak during weekends, likely due to increased rest or sleep.
- Similarly, very active minutes exhibit a weekend trend, suggesting more time for physical activity.
- Friday and Saturday emerge as the most active days of the week.
- Targeted marketing efforts should prioritize weekends to capitalize on increased activity levels.

Lets check the distinct time values in `hourly_data`

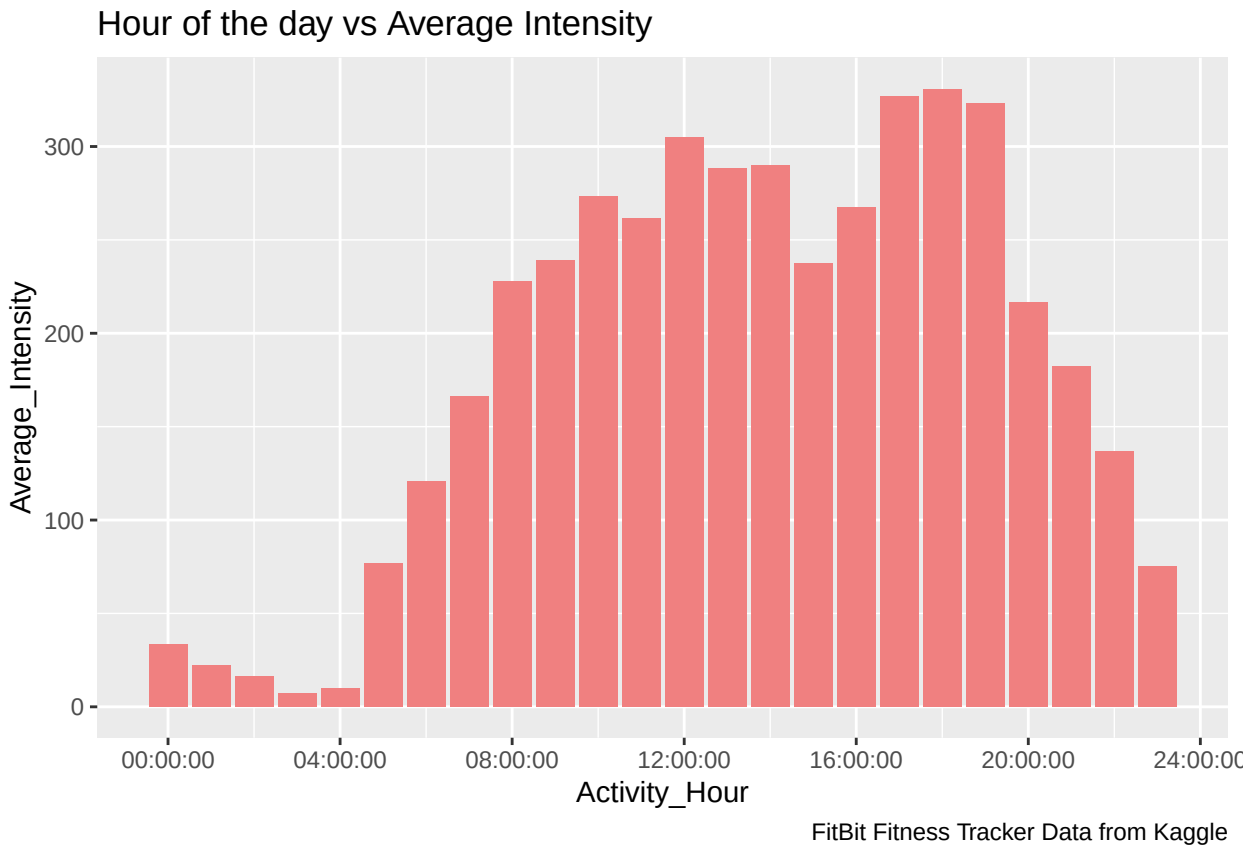
```
unique(hourly_data$Activity_Hour)
```

```
## 00:00:00
## 01:00:00
## 02:00:00
## 03:00:00
## 04:00:00
## 05:00:00
## 06:00:00
## 07:00:00
## 08:00:00
## 09:00:00
## 10:00:00
## 11:00:00
## 12:00:00
## 13:00:00
## 14:00:00
## 15:00:00
## 16:00:00
## 17:00:00
```

```
## 18:00:00
## 19:00:00
## 20:00:00
## 21:00:00
## 22:00:00
## 23:00:00
```

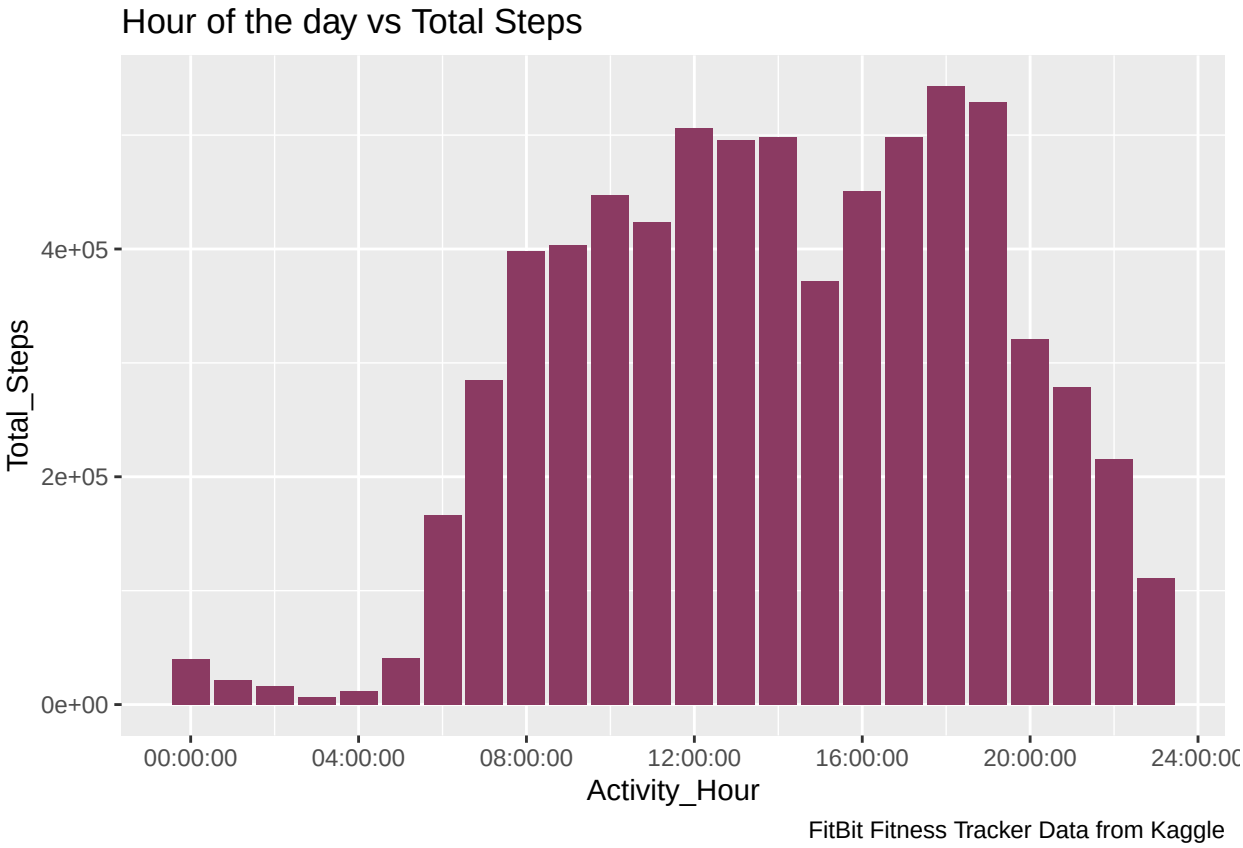
Lets check the `Average_Intensity` throughout the day.

```
ggplot(data = hourly_data) +
  geom_col(mapping = aes(x = Activity_Hour, y= Average_Intensity), fill = "lightcoral") +
  labs(title = "Hour of the day vs Average Intensity",
       caption = "FitBit Fitness Tracker Data from Kaggle")
```



Lets check the `Total_Steps` taken throughout the day.

```
ggplot(data = hourly_data) +
  geom_col(mapping = aes(x = Activity_Hour, y= Total_Steps), fill = "hotpink4") +
  labs(title = "Hour of the day vs Total Steps",
      caption = "FitBit Fitness Tracker Data from Kaggle")
```



Observations:

- Peak intensity occurs between 8:00 AM to 8:00 PM, indicating heightened activity levels during this timeframe.
- Total steps peak between 5:00 PM to 7:00 PM, suggesting increased physical activity in the evening.
- The evening surge in activity may indicate higher energy levels post-work hours or more leisure time for exercise.
- Marketing efforts should target these active hours to engage users during their most active and available times.

7 Recommendations - Marketing Strategies

I have developed marketing strategies focusing on Bellabeat's product **'Time'**.

Promotion of Light Activity Features

- Highlight the benefits of light activities for overall well-being and stress reduction.
- Showcase how Bellabeat Time monitors activities such as light walking, stretching, and gentle yoga sessions, promoting a holistic approach to fitness.
- Emphasize how Bellabeat Time encourages users to stay active throughout the day, even during low-intensity activities, fostering a healthier lifestyle overall.

Targeted Weight Loss Campaigns

- Develop marketing campaigns specifically aimed at individuals with weight loss goals, emphasizing Time's ability to track and monitor progress.

- Showcase success stories and testimonials from users who have achieved their weight loss goals with the help of Bellabeat's Time.
- Highlight features such as step tracking, calorie burn monitoring, and personalized activity recommendations to support users on their weight loss journey.

Weekend Wellness Promotion

- Position Bellabeat Time as an essential tool for weekend wellness, encouraging users to maintain their activity levels even during leisure time.
- Launch weekend-specific challenges or promotions to engage users and motivate them to stay active with Bellabeat Time.
- Highlight how Bellabeat Time helps users strike a balance between relaxation and physical activity, ensuring they make the most of their weekends while prioritizing their health.

Evening Activity Focus

- Tailor marketing efforts to highlight Bellabeat Time's role in supporting evening workouts and relaxation routines.
- Showcase features such as sleep tracking, guided breathing exercises, and relaxation reminders to help users unwind and prepare for a restful night.
- Offer promotions or content focusing on evening activities, such as sunset walks or post-dinner yoga sessions, demonstrating how Bellabeat Time enhances the evening wellness experience.

8 Conclusion

The analysis provided valuable insights into consumer trends using smart device usage data. However, limitations include:

- Sampling bias from data collected from only 33 users.
- Lack of demographic data.
- Data collected over a one-month period in 2016, potentially outdated.

Hence, I propose additional analysis using more relevant data, ideally sourced from Bellabeat's own smart devices. This approach ensures greater accuracy and alignment with our objectives.