Bellabeat Case Study

Precious Molokwu

2023-05-04

Bellabeat Case Study

Introduction

Background Information

Bellabeat is a high-tech company that manufactures health-focused smart products for women that incorporate beautifully designed technology. While experiencing notable growth, Bellabeat is always looking for ways in which to improve their share of the global smart device market. As such, this case study uses publicly available and user-provided data to paint a better picture of the customer segment and share relevant insights which can help guide marketing strategy for the company.

Problem Statement

Identifying trends in smart device usage and how these trends can apply to Bellabeat customers and subsequent marketing opportunities.

Data Used

Eligible Fitbit users willingly submitted their personal tracker data and the FitBit Fitness Tracker Data is a publicly available dataset found on Kaggle and made available through Mobius.

This data was collected through an Amazon survey and includes minute and hour-level output for physical activity, heart rate, and sleep monitoring.

More detailed information on the dataset can be found here[1].

Data Documentation

I cleaned several datasets and selected four to further examine for this case study. I removed duplicates, checked for spelling errors, missing values and formatting issues. I also combined datasets where viable, omitted columns that would skew outcomes and renamed columns for congruity between datasets. An in-depth look at my data cleaning process can be found in the changelog.

Analysis

Setting up my environment. Installing and loading the "here", "tidyverse", "lubridate", "skimr", "janitor", and "knitr" packages.

```
install.packages("here")
library(here)
install.packages("tidyverse")
library(tidyverse)
library(lubridate)
install.packages("skimr")
```

```
library(skimr)
install.packages("janitor")
library(janitor)
install.packages("knitr")
library(knitr)
```

Importing cleaned .csv files (and re-labeling them for simplicity).

```
activity_daily <- read_csv("dailyActivity_merged1.csv")
sleep_daily <- read_csv("sleepDay_merged1.csv")
weight_daily <- read_csv("weightLogInfo_merged1.csv")
steps_intensities_hourly <-read_csv("hourlyStepsAndIntensities_merged1.csv")
intensities_daily <- read_csv("dailyIntensities_merged1.csv")
```

Taking a closer look at the data using the colnames(), head(), str(), and skim_without_charts() functions.

```
colnames(activity_daily)
```

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

```
head(steps_intensities_hourly)
```

```
## # A tibble: 6 x 5
##           Id ActivityHour   TotalIntensity AverageIntensity StepTotal
##        <dbl> <chr>                   <dbl>            <dbl>     <dbl>
## 1 1503960366 4/12/2016 0:00             20            0.333       373
## 2 1503960366 4/12/2016 1:00              8            0.133       160
## 3 1503960366 4/12/2016 2:00              7            0.117       151
## 4 1503960366 4/12/2016 3:00              0            0             0
## 5 1503960366 4/12/2016 4:00              0            0             0
## 6 1503960366 4/12/2016 5:00              0            0             0
```

```
str(sleep_daily)
```

```
## spc_tbl_ [410 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Date              : chr [1:410] "4/12/2016" "4/13/2016" "4/15/2016" "4/16/2016" ...
##  $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Renaming column "SedentaryActiveDistance" to "SedentaryDistance" in activity_daily data frame.

```
activity_daily <- activity_daily %>%
  rename(SedentaryDistance = SedentaryActiveDistance)
```

Using summary statistics to see the number of participants per data frame.

We see that activity_daily and steps_intensities_hourly data frames both have 33 participants, the sleep_daily data frame has 24 participants and the weight_daily data frame has 8 participants.

```
activity_participants <- activity_daily %>%
  summarize(n_distinct(activity_daily$Id),
            n_distinct(sleep_daily$Id),
            n_distinct(weight_daily$Id),
            n_distinct(steps_intensities_hourly$Id))

as_tibble(activity_participants)
```

```
## # A tibble: 1 x 4
##   `n_distinct(activity_daily$Id)` n_distinct(sleep_dail~1 n_distinct(weight_da~2
##                             <int>                   <int>                  <int>
## 1                              33                      24                      8
## # i abbreviated names: 1: `n_distinct(sleep_daily$Id)`,
## #   2: `n_distinct(weight_daily$Id)`
## # i 1 more variable: `n_distinct(steps_intensities_hourly$Id)` <int>
```

We can also get a more detailed look at daily activity and sleeping habits with summary statistics. We see that most people walk around 7,500 steps a day and are sedentary for roughly 17-18 hours a day (including sleep). We also see that the average person sleeps once a day and does so for around 7 hours each sleep cycle.

```
activity_daily %>%
  select(TotalSteps,
         TotalTrackerDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalTrackerDistance 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
```

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

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.00      Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.00      1st Qu.:361.0      1st Qu.:403.8
##  Median :1.00      Median :432.5      Median :463.0
##  Mean   :1.12      Mean   :419.2      Mean   :458.5
##  3rd Qu.:1.00      3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.00      Max.   :796.0      Max.   :961.0
```

Combining data to get a clearer picture of how different variables interact with each other. Creating a new data frame ("activity_daily_with_sleep") and dropping all N/A values in order to only see daily entries that had matching records for total steps, total minutes slept, and all other factors of interest.

```
activity_daily_with_sleep <- merge(sleep_daily, activity_daily, by=c("Id","Date"), all =
TRUE) %>%
  drop_na()
```

Using n_distinct() to get a look at unique participant count (24 participants).

```
activity_daily_with_sleep %>%
  summarize(n_distinct(activity_daily_with_sleep$Id))
```

```
##   n_distinct(activity_daily_with_sleep$Id)
## 1                                       24
```

Creating a new column (ActiveIntensity) representing VeryActiveMinutes count divided by 60 to convert it to hours.

```
intensities_daily$ActiveIntensity <- (intensities_daily$VeryActiveMinutes)/60
```

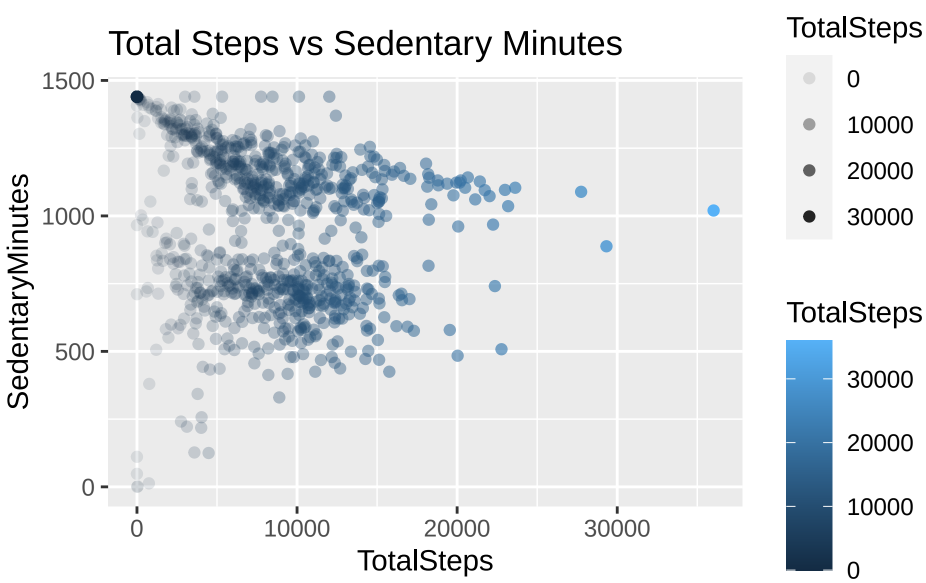
Creating a merged data frame that reflects the relationship between weight and intensity while dropping all N/A values and creating a new column titled "Active Intensity".

```
weight_and_intensities <- merge(weight_daily, intensities_daily, by="Id", all=TRUE) %>%
  drop_na()
```

Visualizations

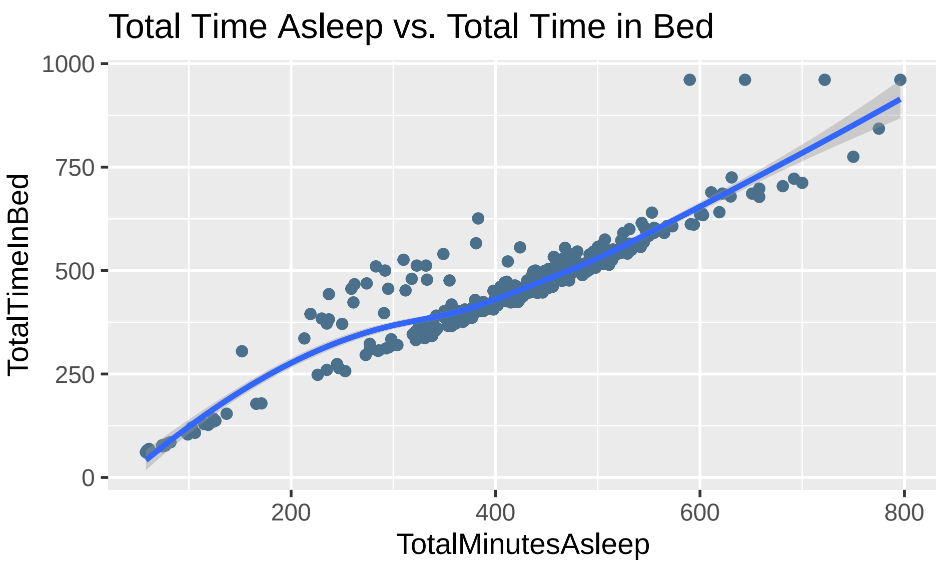
We can use visualizations to help us get a better picture of the data and tell a clearer story.

```
ggplot(data = activity_daily) + geom_point(mapping=aes(x=TotalSteps, y=SedentaryMinutes,
alpha=TotalSteps, color=TotalSteps)) + labs(title="Total Steps vs Sedentary Minutes")
```



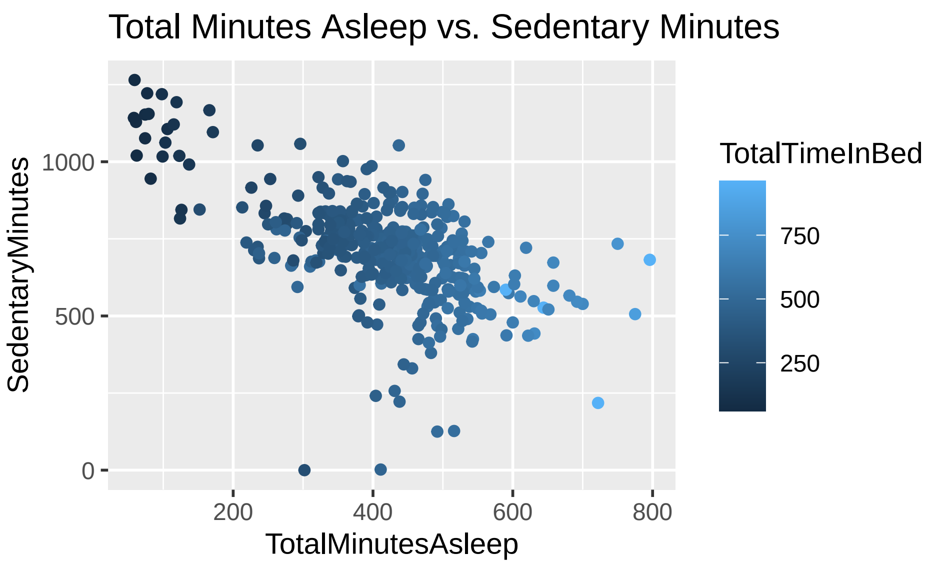
This scatterplot shows the relationship between total steps taken in a day (TotalSteps) and sedentary minutes in a day (SedentaryMinutes) while increasing the opacity of point based on total steps (the darker blue the point, the more steps recorded).

```
ggplot(data = sleep_daily) + geom_point(mapping=aes(x=TotalMinutesAsleep,
y=TotalTimeInBed), color="skyblue4") + geom_smooth(mapping=aes(x=TotalMinutesAsleep,
y=TotalTimeInBed)) + labs(title="Total Time Asleep vs. Total Time in Bed")
```



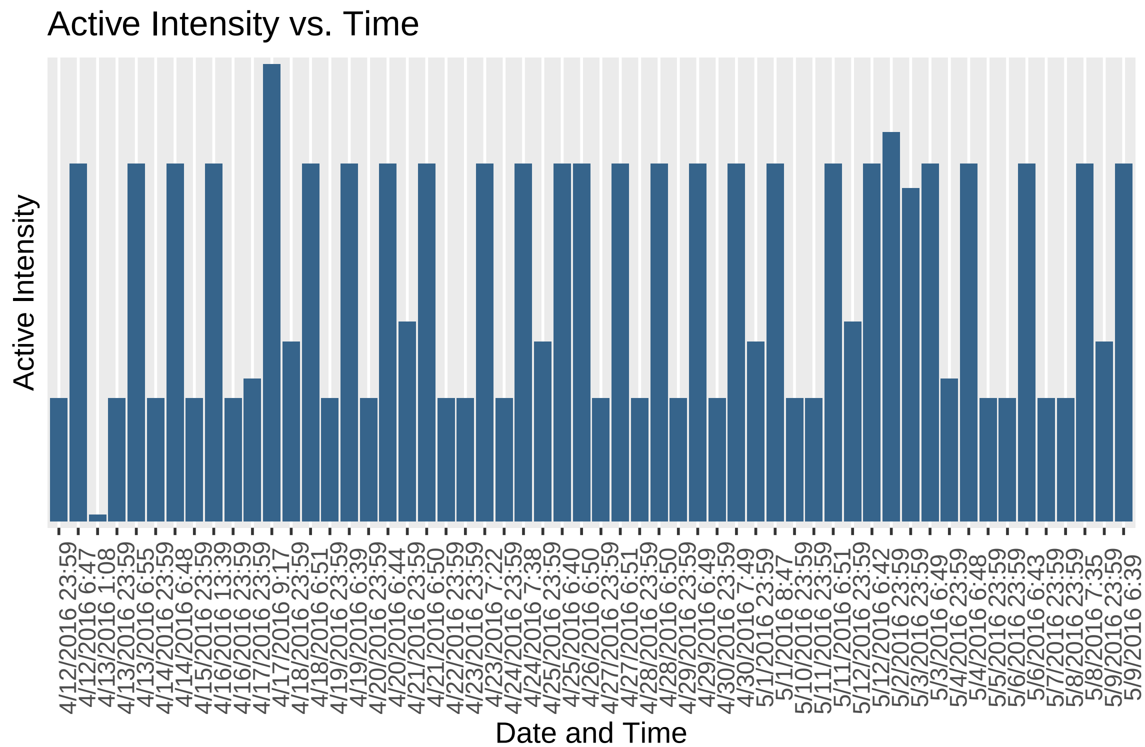
This scatterplot shows the relationship between time spent asleep and time spent in bed with a trend line.

```
ggplot(data = activity_daily_with_sleep) + geom_point(mapping=aes(x=TotalMinutesAsleep,
y=SedentaryMinutes, color=TotalTimeInBed)) + labs(title="Total Minutes Asleep vs.
Sedentary Minutes")
```



This scatterplot shows the relationship between time spent asleep and sedentary time, highlighting which areas have the most and least total time spent in bed.

```
ggplot(data=weight_and_intensities, aes(x=Date.x, y=ActiveIntensity)) +
geom_histogram(stat = "identity", fill='steelblue4') +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Active Intensity vs. Time ") +
  scale_x_discrete(name="Date and Time") +
  scale_y_discrete(name="Active Intensity")
```



This histogram shows the relationship between active intensity and time.

Key Findings

Examining the relationship between total steps taken in a day and associated sedentary time, we find that there are two main groups of people who roughly take the same amount of steps but have largely different amounts of sedentary time in a day. This may indicate that one group sleeps less or that the other group may have obligations (work, school, commuting, etc.) that may keep them sedentary for some part of the day and they may make up their steps through exercise or other means. Another explanation for sedentary behavior could be related to the increase of time spent sitting around watching television or consuming other forms of media[2]. Research also shows that many affluent people today have become less active compared to their predecessors.

Tracking "Active Intensity" over time, there seems to be a pattern of high active intensity during earlier periods of the day and lower active intensity during later periods in the day. In terms of marketing, advertisements are more likely to catch the eye during low active intensity periods.

Looking at the provided sleep data, we can see that while there generally is a positive relationship between time spent asleep and time spent in bed, there is a customer segment that spends more time in bed, but not asleep, than others. Findings show that the most sedentary time is not spent while in bed, indicating that most inactive time occurs during the day.

Conclusion

Recommendations

Investigation using summary statistics shows that most people take around 7,500 steps a day (75% of the CDC-recommended 10,000 steps a day[3]), have one daily sleep cycle and sleep roughly 7 hours during it. With around 1,000 sedentary minutes (times of inactivity) in the average day, and roughly 10 hours of those spent awake once accounting for sleep, there seems to be ample time to reach the recommended 10,000 steps a day.

The Bellabeat team may want to look into implementing reminder notifications that are triggered once a certain threshold of sedentary non-sleeping time has been reached.

With an increase in sedentary media consumption (especially among the affluent), Bellabeat should also target high-income young adults, and young adults in general, due to their high lifetime value as consumers. Areas to target for advertising may include Social Media, Gaming, Video and Audio as younger generations are more engaged in those channels than their older counterparts[4].

While the general expectation of there being a completely linear relationship with total time asleep and total time in bed holds up, there are some outliers. Interestingly, there is a cluster of customers who spend more time awake in bed relative to how long everyone else is asleep in bed. This customer segment could potentially be focused on with in-app notifications during the times they are in bed but not asleep, especially during the mornings, encouraging them to use this free time exercise.

It seems that most sedentary time occurs outside of bed and often later in the day, probably sitting down consuming media, commuting, or stationary while at work or in school. There are potential marketing opportunities which will allow Bellabeat to reach a larger target audience by advertising during the times that most people are awake but least active, such as the evenings. Bellabeat should also focus their advertising efforts in areas such as social media, gaming, and video. These are areas where they can consistently reach a very valuable customer segment.

Sources

[1]FitBit Fitness Tracker Data (Kaggle Dataset)

[2]Mayo Clinic Study on Sedentary Behavior

[3]CDC Lifestyle Coach Facilitation Guide

[4]"Media Habits Are Changing Rapidly For Young Adults, Making Ad Targeting More Challenging"