Bellabeat Case Study

Adam LaFave

2022-03-18

Introduction

This is a case study for Bellabeat, a high-tech manufacturer of fashionable health-focused products for women. Bellabeat currently has an app that works with wearable wellness devices for women that tracks activity, sleep, and stress. Bellabeat's stakeholders have asked for an analysis of currently available smart device data to gain insight into how consumers are using smart devices. The 6 steps of the data analysis process: **ask, prepare, process, analyze, share, and act**, were used. The stakeholder's are seeking high-level recommendations for how these trends can inform the Bellabeat marketing strategy team and stakeholders in positioning a new product.

Step 1: Ask

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

Deliverables

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. High-level content recommendations based on the analysis

Key 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

Step 2: Prepare

Data Source The dataset used is an opensource public dataset generated from thirty eligible Fitbit users who consented to the submission of their personal tracker data from 03-12-2016 to 05-12-2016. The data

includes minute-level output for physical activity, heart rate, and sleep monitoring.

Click Here to View the Data Source

Limitations with the data Bias and credibility - Does the data ROCCC? (Low to High)

- **R**eliable: **Low** - small population sample
- **O**riginal: **Low** - third party source
- **C**omprehensive: **Medium** - matches two of the three categories Bellabeat's trackers use
- **C**urrent: **Low** - Outdated data from 2016
- **C**ited: **Low** - Source of data cited: Amazon Mechanical Turk

Overall, the data source would receive a low score for credibility and it's not recommended to be used for making business suggestions. The data has a small population sample of only thirty Fitbit users which does not represent the overall population. Although cited correctly, it's not directly from the original data source, but from Amazon Mechanical Turk. The data is not current being from 2016.

How can the data help you answer your business task question?

The dataset does contain daily activity and sleep patterns which Bellabeat's technology tracks as well.

Step 3: Process

##Preparing the environment

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.1.3
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

##Importing the datasets and assigning them to a new dataframes

```
daily_activity <- read.csv("dailyActivity_merged.csv")
```

```
sleep <- read.csv("sleepDay_merged.csv")
```

Begin exploring the data ##Previewing the daily_activity data

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      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
```

##Identify all the columns in the daily_activity data

```
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"
```

##Previewing the sleep data

```
head(sleep)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
```

```
## 1 1503960366 4/12/2016 12:00:00 AM                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
```

##Identifying the column names in sleep data

```
colnames(sleep)
```

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

##Identifying how many unique participants there are in each dataframe. Shows us there are 33 more observations in the daily_activity dataframe

```
n_distinct(daily_activity$Id)
```

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

```
n_distinct(sleep$ID)
```

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

A few observations about the data

- There are more entries in the daily_activity data frame than there are in the sleep dataframe
- There were no null or missing values shown in the summary output

Data Wrangling ##Transforming SedentaryMinutes to Sedentary_Hours from daily_Activity for plotting later

```
SedentaryHours <- daily_activity$SedentaryMinutes/60
```

##Transforming TotalMinutesAsleep and TotalTimeInBed to Hours for easier visualization

```
Total_Hours_Asleep <- sleep$TotalMinutesAsleep/60
Total_Hours_In_Bed <- sleep$TotalTimeInBed/60
```

##Merging the two datasets, removing any null values, and removing one of the dates columns which would be duplicated because it has a different column name

```
combined_data <- merge(sleep, daily_activity, by="Id") %>%
   select(-SleepDay) %>%
  drop_na()
```

##Taking a look at how many participants are in the combined dataset

```
n_distinct(combined_data$Id)
```

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

- We could use outer_join to keep the participants that don't have sleep data but do have daily_activity data in the combined dataset, but will leave it out for plotting purposes

#Adding an Hours_In_Bed column to the combined_data set I created to better visualize hours spent in bed instead of minutes spent in bed

```
combined_hourly <- mutate(combined_data, Hours_In_Bed= TotalTimeInBed/60 )
```

Step 4: Analyze

Finding some summary statistics about each dataframe ##Summarizing the TotalSteps, TotalDistance, and SedentaryMinutes columns from the daily_activity data frame

```
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
```

##Summarizing the TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed columns from the sleep dataframe

```
sleep %>%
  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
```

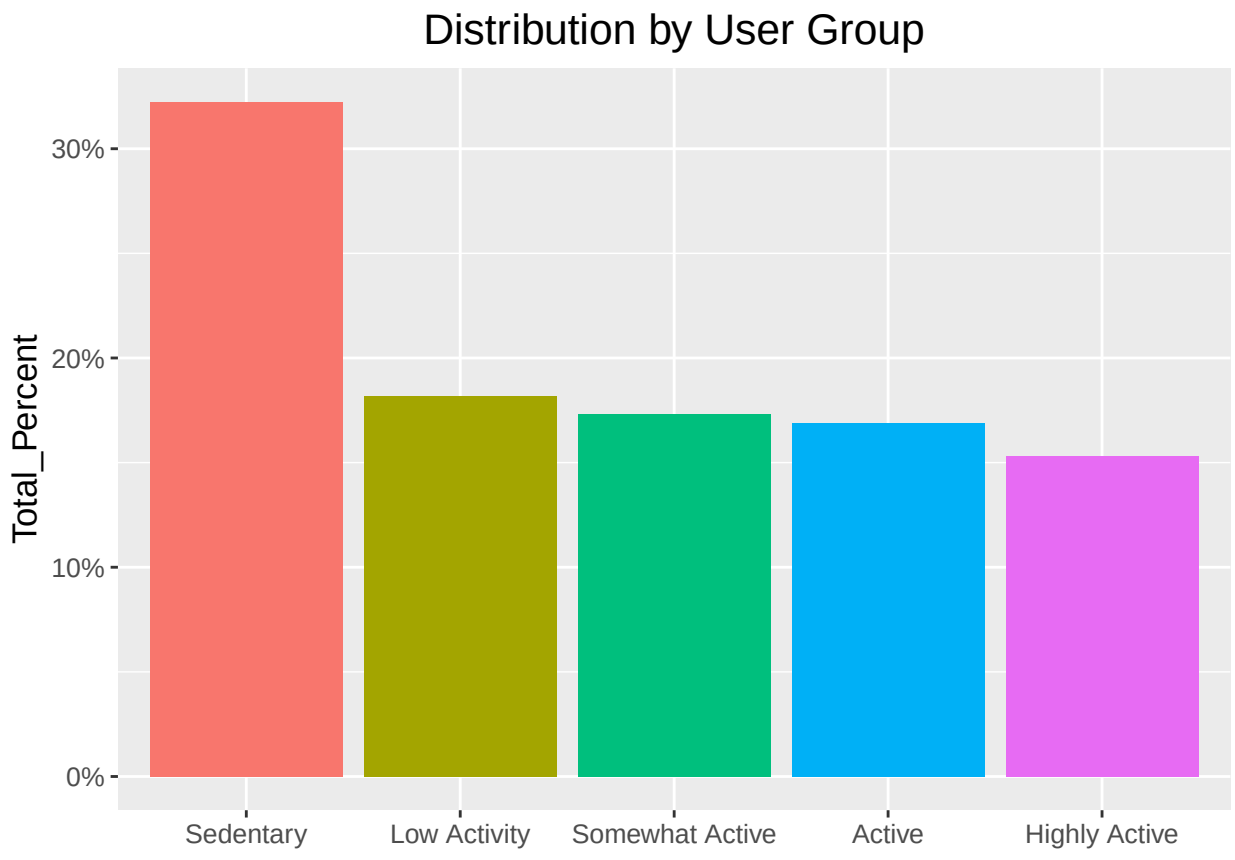
##According to a published article on nih.gov, individuals who have <5,000 steps/day can be classified as having a 'sedentary lifestyle', 5,000-7,499 steps/day as having a 'low activity lifestyle', 7,500-9,999 steps/day as having a 'somewhat active lifestyle', 10,000-12,499 steps/day as having an 'active lifestyle', and >=12,500 steps/day are likely to be classified as 'highly active lifestyle.' We group all users into these 5 categories for a better visualization. Source:nih.gov

```
usergroup_df <- daily_activity %>%
  summarise(
    user_group = factor(case_when(
      TotalSteps < 5000 ~
        "Sedentary",
      TotalSteps >= 5000 & TotalSteps < 7500 ~
        "Low Activity",
      TotalSteps >= 7500 & TotalSteps < 10000 ~
```

```
      "Somewhat Active",
    TotalSteps >= 10000 & TotalSteps < 12500 ~
      "Active",
    TotalSteps >= 12500 ~
      "Highly Active"
  ), levels=c("Sedentary", "Low Activity", "Somewhat Active", "Active","Highly Active")), .group=Id) %
  drop_na()
```

##Now we can show the group distribution to see a percentage of how many users are in each group

```
usergroup_df %>%
group_by(user_group) %>%
summarise(total = n()) %>%
mutate(totals = sum(total)) %>%
group_by(user_group) %>%
summarise(Total_Percent = total/totals) %>%
ggplot(aes(user_group, y = Total_Percent, fill=user_group)) +
  geom_col()+
scale_y_continuous(labels = scales::percent) +
theme(legend.position="none") +
labs(title = "Distribution by User Group", x=NULL) +
theme(legend.position="none", text=element_text(size=13),plot.title=element_text(hjust = 0.5))
```



Key TakeAways & Possible Recommendations

- Evidence suggests that even though most people are wearing active tracking devices they are still living a sedentary lifestyle. This could could be used as a pain point to get users to interact more with the

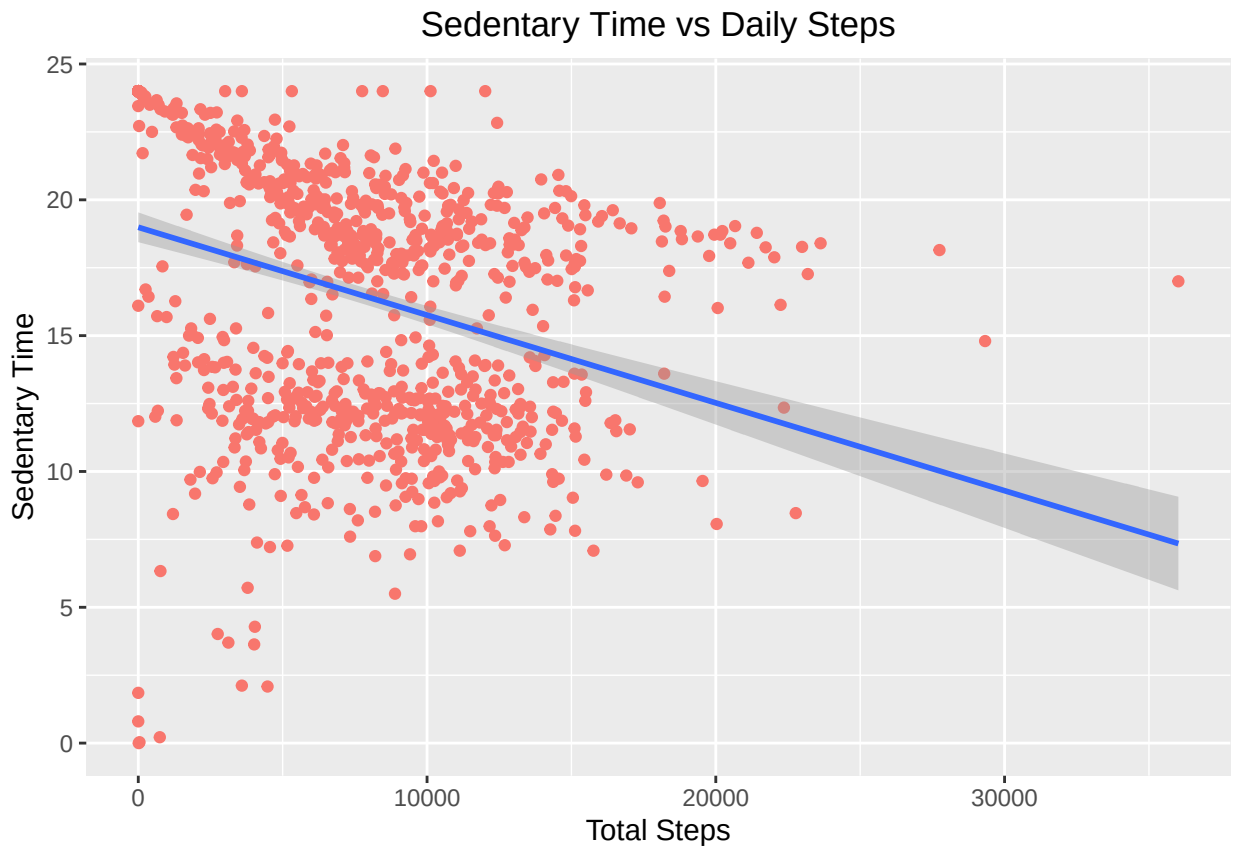
app. This could also be used as a motivator to get users to sign up for the subscription membership where they can get 24/7 access to fully personalized guidance on nutrition, activity, sleep, health, beauty, and mindfulness based on their lifestyle and goals.

Step 5: Share

##Plotting daily sedentary time vs steps walked ##Adding geom_smooth to show the regression line

```
ggplot(data=daily_activity)+
  geom_point(mapping=aes(x=TotalSteps,y=SedentaryHours, color=""))+
  labs(title="Sedentary Time vs Daily Steps", x="Total Steps", y="Sedentary Time") +
  theme(legend.position="none", plot.title=element_text(hjust = 0.5))+
  geom_smooth(method="lm",mapping=aes(x=TotalSteps,y=SedentaryHours))
```

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



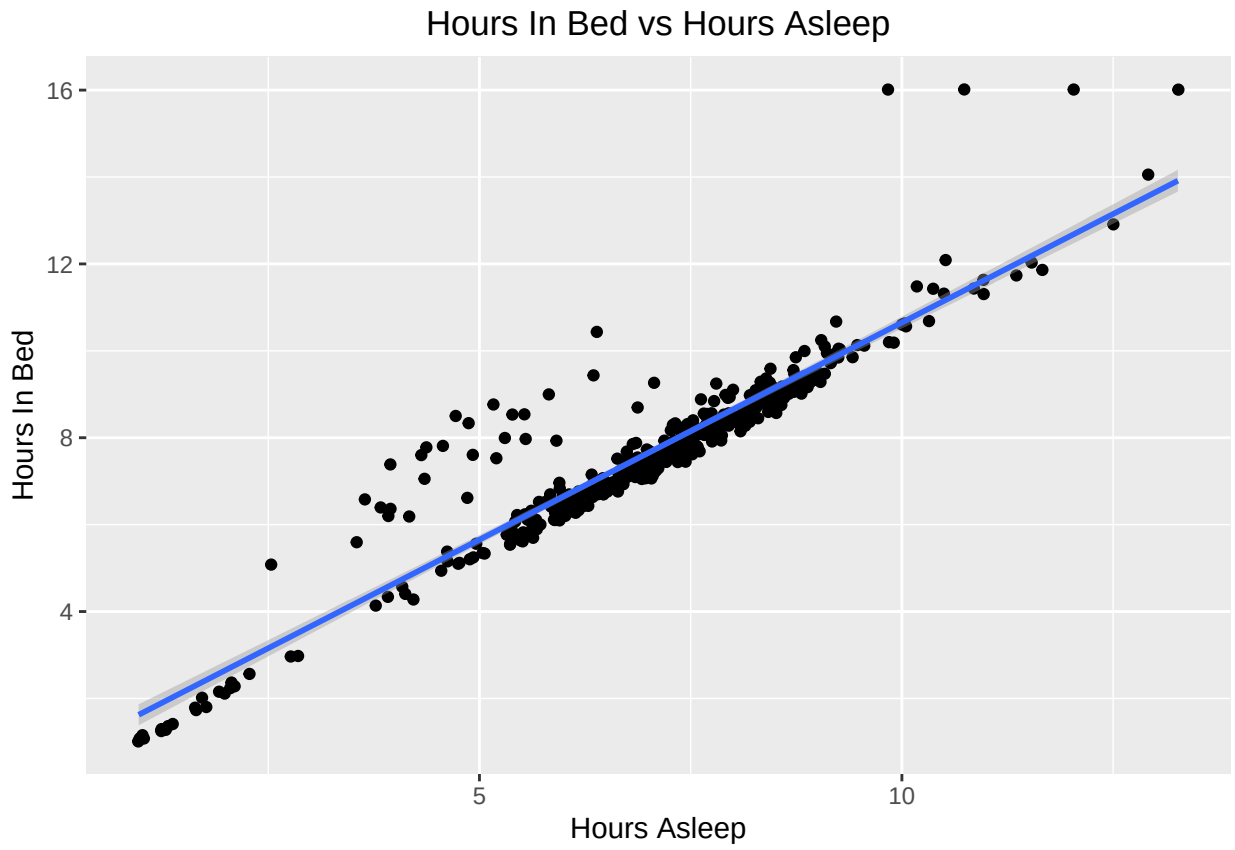
Key TakeAways & Possible Recommendations

- We discover that the he lower the number of steps walked the higher the number of sedentary minutes throughout the day.
- Recommendation: I would recommend customer target marketing for those searching for walking and health benefits and/or emailing current Bellabeat customers using this chart to get them to use the app more often by checking their daily walking habits.

```
ggplot(data=sleep)+
  geom_jitter(mapping=aes(x=Total_Hours_Asleep, y=Total_Hours_In_Bed))+
  labs(title="Hours In Bed vs Hours Asleep", x="Hours Asleep", y="Hours In Bed") +
```

```
  geom_smooth(method="lm",mapping=aes(x=Total_Hours_Asleep,y=Total_Hours_In_Bed)) +
  theme(plot.title=element_text(hjust = 0.5))
```

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

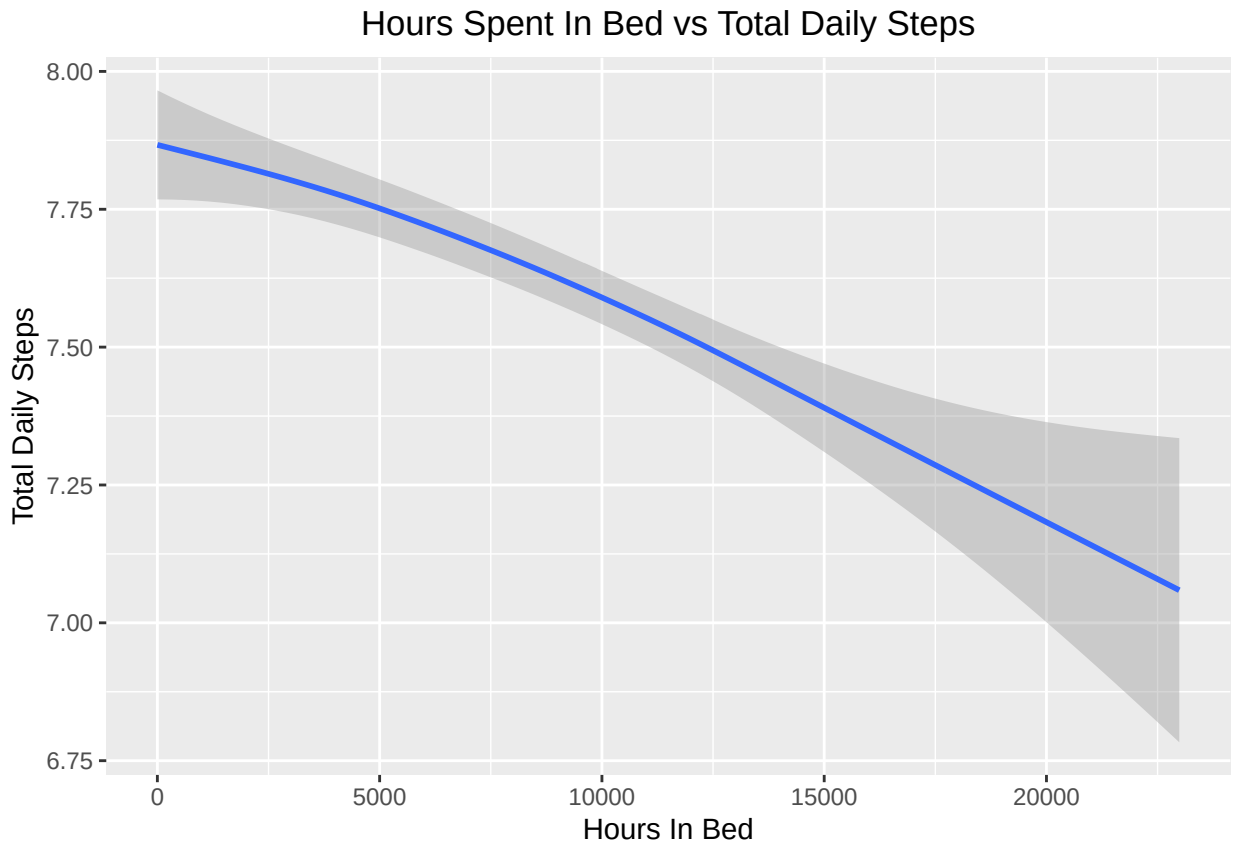


Key TakeAways & Possible Recommendations

- We discovered that the total hours in bed and totals hours of sleep have more than just a few outliers and is not perfectly linear. Noting this, a possible market segment to focus on would be those struggling to get a good night's rest and how Bellabeat's technology can help measure this.

```
ggplot(data=combined_hourly)+
  geom_smooth(mapping=aes(x=TotalSteps,y=Hours_In_Bed,)) +
  labs(title="Hours Spent In Bed vs Total Daily Steps", x="Hours In Bed", y="Total Daily Steps")+
  theme(plot.title=element_text(hjust = 0.5))
```

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



* We discover here that the hours spent in bed clearly decreases the more steps a user walks per day. Again, another target specific marketing strategy for the team.

Step 6: Act

- As mentioned under the distribution table, a major take away is that evidence shows that even though most people are wearing active tracking devices they are still living a sedentary lifestyle and could potentially be a big opportunity for the marketing department to target this message to get users to sign up for the subscription membership where they can get 24/7 access to fully personalized guidance. All other findings in this analysis strongly indicate opportunity to show customers the benefits of being more active and how measuring your activity with Bellabeat can help guide you to a better lifestyle.
- I would also strongly recommend the stakeholders use the current data that they may have from their own devices and try to get a more current analysis with a larger dataset that will represent the greater population. With Bellabeat being a women focused business, the genders of these participants is unknown and the results may change. Either way having a more current, original, and reliable dataset will allow the stakeholders and marketing team to get a much better look at where potential opportunities may lie.