Bellabeat Case Study

Peter Masternak

17/01/2022

## s

# **ASK**

### INTRO

You are working with 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, co-founder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. You 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 you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat’s marketing strategy.

### BUSINESS TASK

**Find out how customers are using their fitness smart devices by identifying trends and providing the executive team with high-level recommendations.**

## ——————————————————————————–

# **PREPARE**

### ABOUT THE DATA

The dataset was obtained from Kaggle which includes open source Fitbit fitness tracker data available to use for analysis. This dataset was generated by respondents to a distributed survey via Amazon Mechanical Turk between 03-12-2016 and 05-12-2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.

This dataset is organized into 18 CSV files. Most of the data are duplicates and the most relative data can be found in the daily activity and sleep data. We will be focusing on the daily time frame since we are looking for high-level trends. This data can help us uncover key trends and gain insight on how our users like to use their smart fitness devices.

Data License Agreement and datasets used can be found [here](https://www.kaggle.com/arashnic/fitbit/metadata).

Metadata file containing info about the data types and data description can be found [here](https://www.fitabase.com/media/1930/fitabasedatadictionary102320.pdf).

### DATA LIMITATIONS

* Data-privacy issues prohibit us from having access to data from customers who did not consent to the submission of personal tracker data. This limits the total available data for analysis and the data collected is considered a small sample of the overall smart fitness device users. The small sample size of participants may not reveal the big picture.
* It’s unclear for how long each participant had their devices in use. Charging the device would result in unrecorded data; This may have taken place during low activity hours or during sleep, but this is not certain.
* Gender is unknown in the survey. Bellabeat focuses exclusively on women’s health and this data came from Fitbit which has both men and women users. Gender information here is important to provide additional insights during the analysis and recommendations.
* Age and fitness level are unknown in the survey. This limits us to understand what level of fitness each participant is in and set a benchmark according to age.
* There is no data to cover an entire year which would’ve given a better understanding of how individuals use their fitness devices throughout each season of the year.
* Data was collected in 2016 which is not current.

## ——————————————————————————–

# **PROCESS**

### TOOLS USED

We will be using **RStudio** to prepare, clean, analyze and visualize the data.

### SETTING UP THE ENVIRONMENT

Installing and loading packages that we will be using:

install.packages("tidyverse", lib = "/cloud/project")  
library(tidyverse)  
install.packages("skimr", lib = "/cloud/project")  
library(skimr)  
install.packages("janitor", lib = "/cloud/project")  
library(janitor)  
install.packages("lubridate", lib = "/cloud/project")  
library(lubridate)  
install.packages("ggplot2", lib = "/cloud/project")  
library(ggplot2)

Import CSV files:

Activity <- read.csv("dailyActivity\_merged.csv")  
Sleep <- read.csv("sleepDay\_merged.csv")  
Weight <- read.csv("weightLogInfo\_merged.csv")  
Intensities <- read.csv("hourlyIntensities\_merged.csv")

Check column headers and data structure:

# Activity  
str(Activity)

## 'data.frame': 940 obs. of 15 variables:  
## $ Id : num 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDate : chr "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...  
## $ TotalSteps : int 13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...  
## $ TotalDistance : num 8.5 6.97 6.74 6.28 8.16 ...  
## $ TrackerDistance : num 8.5 6.97 6.74 6.28 8.16 ...  
## $ LoggedActivitiesDistance: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ VeryActiveDistance : num 1.88 1.57 2.44 2.14 2.71 ...  
## $ ModeratelyActiveDistance: num 0.55 0.69 0.4 1.26 0.41 ...  
## $ LightActiveDistance : num 6.06 4.71 3.91 2.83 5.04 ...  
## $ SedentaryActiveDistance : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ VeryActiveMinutes : int 25 21 30 29 36 38 42 50 28 19 ...  
## $ FairlyActiveMinutes : int 13 19 11 34 10 20 16 31 12 8 ...  
## $ LightlyActiveMinutes : int 328 217 181 209 221 164 233 264 205 211 ...  
## $ SedentaryMinutes : int 728 776 1218 726 773 539 1149 775 818 838 ...  
## $ Calories : int 1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...

# Sleep  
str(Sleep)

## 'data.frame': 413 obs. of 5 variables:  
## $ Id : num 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ SleepDay : chr "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM" "4/16/2016 12:00:00 AM" ...  
## $ TotalSleepRecords : int 1 2 1 2 1 1 1 1 1 1 ...  
## $ TotalMinutesAsleep: int 327 384 412 340 700 304 360 325 361 430 ...  
## $ TotalTimeInBed : int 346 407 442 367 712 320 377 364 384 449 ...

# Weight  
str(Weight)

## 'data.frame': 67 obs. of 8 variables:  
## $ Id : num 1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...  
## $ Date : chr "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM" "4/21/2016 11:59:59 PM" ...  
## $ WeightKg : num 52.6 52.6 133.5 56.7 57.3 ...  
## $ WeightPounds : num 116 116 294 125 126 ...  
## $ Fat : int 22 NA NA NA NA 25 NA NA NA NA ...  
## $ BMI : num 22.6 22.6 47.5 21.5 21.7 ...  
## $ IsManualReport: chr "True" "True" "False" "True" ...  
## $ LogId : num 1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...

# Intensities  
str(Intensities)

## 'data.frame': 22099 obs. of 4 variables:  
## $ Id : num 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityHour : chr "4/12/2016 12:00:00 AM" "4/12/2016 1:00:00 AM" "4/12/2016 2:00:00 AM" "4/12/2016 3:00:00 AM" ...  
## $ TotalIntensity : int 20 8 7 0 0 0 0 0 13 30 ...  
## $ AverageIntensity: num 0.333 0.133 0.117 0 0 ...

### CLEANING THE DATA

#### Rename Columns

Renaming column headers for consistent and easier readability:

# Activity  
Activity <- Activity %>%   
 rename(  
 ID = Id,  
 Date = ActivityDate,  
 Total\_Steps = TotalSteps,  
 Total\_Distance\_km = TotalDistance,  
 Tracker\_Distance\_km = TrackerDistance,  
 Logged\_Activites\_Distance\_km = LoggedActivitiesDistance,  
 Very\_Active\_Distance\_km = VeryActiveDistance,  
 Moderate\_Active\_Distance\_km = ModeratelyActiveDistance,  
 Light\_Active\_Distance\_km = LightActiveDistance,  
 Sedentary\_Active\_Distance\_km = SedentaryActiveDistance,  
 Very\_Active\_min = VeryActiveMinutes,  
 Fairly\_Active\_min = FairlyActiveMinutes,  
 Lightly\_Active\_min = LightlyActiveMinutes,  
 Sedentary\_min = SedentaryMinutes  
 )  
str(Activity)  
# Sleep  
Sleep <- Sleep %>%   
 rename(  
 ID = Id,  
 Date = SleepDay,  
 Total\_Sleep\_Records = TotalSleepRecords,  
 Total\_Min\_Asleep = TotalMinutesAsleep,  
 Total\_Min\_In\_Bed = TotalTimeInBed  
 )  
str(Sleep)  
# Weight  
Weight <- Weight %>%   
 rename(  
 ID = Id,  
 Weight\_kg = WeightKg,  
 Weight\_lbs = WeightPounds,  
 Is\_Manual\_Report = IsManualReport,  
 Log\_ID = LogId  
 )  
str(Weight)  
# Intensities  
Intensities <- Intensities %>%   
 rename(  
 ID = Id,  
 Hour = ActivityHour,  
 Total\_Intensity = TotalIntensity,  
 Avg\_Intensity = AverageIntensity  
 )  
str(Intensities)

#### Data Type

All Date columns don’t have correct data type — Converting to date time format:

# Activity  
Activity$Date=as.POSIXct(Activity$Date, format="%m/%d/%Y", tz=Sys.timezone())  
str(Activity)  
  
# Sleep  
Sleep$Date=as.POSIXct(Sleep$Date, format="%m/%d/%Y", tz=Sys.timezone())  
str(Sleep)  
  
# Weight  
Weight$Date=as.POSIXct(Weight$Date, format="%m/%d/%Y", tz=Sys.timezone())  
str(Weight)

For *Intensities*, we need to take steps to get the *average total intensity* for each hour of the day, and isolate the hours from the date

# Separating Date\_Hour columns into Date, Hour, and AM/PM columns  
Intensities <-separate(Intensities, Hour, into=c('Date', 'Hour', 'AM/PM'), sep=' ')   
  
## Unite Hour and AM/PM columns together to Hour column  
Intensities$Hour <- paste(Intensities$Hour, Intensities$`AM/PM`)   
  
## Drop Date and AM/PM columns  
Intensities <- Intensities[-c(2,4)]   
  
## Create hourly average for total intensities  
Intensities\_Hour\_Avg <- Intensities %>%   
 group\_by(Hour) %>%   
 summarize(Total\_Intensity = mean(Total\_Intensity))   
  
## Changing data type to date time for Hour column (also changes time to 24hr format)  
Intensities\_Hour\_Avg$Hour=as.POSIXct(Intensities\_Hour\_Avg$Hour, format="%I:%M:%S %p", tz=Sys.timezone())   
  
## Separate date and hour  
Intensities\_Hour\_Avg <-separate(Intensities\_Hour\_Avg, Hour, into=c('Date', 'Hour'), sep=' ')   
  
## Drop Date column  
Intensities\_Hour\_Avg <- Intensities\_Hour\_Avg[-c(1)]   
  
## Sorting hours in order  
Intensities\_Hour\_Avg <- Intensities\_Hour\_Avg %>%   
 arrange(Hour)

## Check  
str(Intensities\_Hour\_Avg)

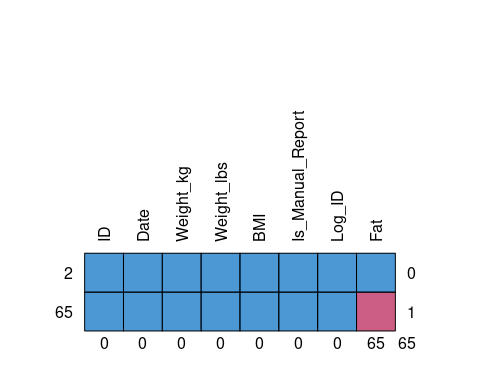
## tibble [24 × 2] (S3: tbl\_df/tbl/data.frame)  
## $ Hour : chr [1:24] "00:00:00" "01:00:00" "02:00:00" "03:00:00" ...  
## $ Total\_Intensity: num [1:24] 2.13 1.419 1.044 0.444 0.633 ...

#### NULL and NA Values

Check data frames for NULL and NA values

# Activity  
sum(is.null(Activity)) # 0 NULL  
sum(is.na(Activity)) # 0 NA  
  
# Sleep  
sum(is.null(Sleep)) # 0 NULL  
sum(is.na(Sleep)) # 0 NA  
  
# Weight  
sum(is.null(Weight)) # 0 NULL  
sum(is.na(Weight)) # 65 NA  
  
# Intensities  
sum(is.null(Intensities)) # 0 NULL  
sum(is.na(Intensities)) # 0 NA

# Install and load "mice" package to help identify which columns the NA values are in  
install.packages("mice")  
library(mice)  
  
# Create viz to show where the missing data is  
md.pattern(Weight, plot = TRUE, rotate.names = TRUE)



## ID Date Weight\_kg Weight\_lbs BMI Is\_Manual\_Report Log\_ID Fat   
## 2 1 1 1 1 1 1 1 1 0  
## 65 1 1 1 1 1 1 1 0 1  
## 0 0 0 0 0 0 0 65 65

All 65 NA values are in the *Fat* column (65 out of 67 data points in the column). Since all of the NA values are in the *Fat* column, and make up 65 out of the 67 data points in that column, a proper analysis cannot be done with this column:

# Create new data frame without the "Fat" column  
Weight <- data.frame(Weight %>% select(-Fat))  
str(Weight)

#### Unique IDs

Identifying number of unique values in ID column in each data frame to confirm number of participants:

# Activity  
length(unique(Activity$ID))# 33 IDs  
# Sleep  
length(unique(Sleep$ID)) # 24 IDs  
# Weight  
length(unique(Weight$ID)) # 8 IDs  
# Intensities  
length(unique(Intensities$ID)) # 33 IDs

A total of 33 unique IDs were identified; Activity: 33, Sleep: 24, Weight: 8, Intensities: 33; There are 3 more unique IDs than stated on the data description; Sleep makes up 73% (24/33) of the overall unique IDs and may still be relevant for useful analysis; Weight makes up 24% (8/33) of the overall unique IDs and is insignificant for useful analysis

#### Duplicates

# Activity  
sum(duplicated(Activity)) # 0 duplicates  
# Sleep  
sum(duplicated(Sleep)) # 3 duplicates  
# Intensities  
sum(duplicated(Intensities)) # 12, 351 duplicates

It makes sense for so many duplicates for Intensities since it’s organized by hours and total intensity can be the same during hours of sleep which can produce duplicates:

## The Sleep column should not have duplicates since ID x Date should be unique; Removing duplicates  
Sleep <- distinct(Sleep)  
  
# Check  
str(Sleep)

#### Merging Data Frames

We will merge data frames for easier analysis:

## Joining the Activity and Sleep data frames with the ID and Date columns for analysis  
Daily\_Data\_Merged <- merge(Activity, Sleep, by=c('ID', 'Date'))

#### Adding Column

Adding Total Activity Minutes column in Daily\_Data\_Merged data frame for analysis:

# Adding up Very\_Active\_min, Fairly\_Active\_min, and Lightly\_Active\_min into a new column  
Daily\_Data\_Merged$Total\_Activity\_Min <- Daily\_Data\_Merged$Very\_Active\_min +  
 Daily\_Data\_Merged$Fairly\_Active\_min +  
 Daily\_Data\_Merged$Lightly\_Active\_min  
  
# Check  
str(Daily\_Data\_Merged)

#### Summary

Looking at summery stats on data:

summary(Daily\_Data\_Merged)

## ID Date Total\_Steps   
## Min. :1.504e+09 Min. :2016-04-12 00:00:00 Min. : 17   
## 1st Qu.:3.977e+09 1st Qu.:2016-04-19 00:00:00 1st Qu.: 5189   
## Median :4.703e+09 Median :2016-04-27 00:00:00 Median : 8913   
## Mean :4.995e+09 Mean :2016-04-26 11:38:55 Mean : 8515   
## 3rd Qu.:6.962e+09 3rd Qu.:2016-05-04 00:00:00 3rd Qu.:11370   
## Max. :8.792e+09 Max. :2016-05-12 00:00:00 Max. :22770   
## Total\_Distance\_km Tracker\_Distance\_km Logged\_Activites\_Distance\_km  
## Min. : 0.010 Min. : 0.010 Min. :0.0000   
## 1st Qu.: 3.592 1st Qu.: 3.592 1st Qu.:0.0000   
## Median : 6.270 Median : 6.270 Median :0.0000   
## Mean : 6.012 Mean : 6.007 Mean :0.1089   
## 3rd Qu.: 8.005 3rd Qu.: 7.950 3rd Qu.:0.0000   
## Max. :17.540 Max. :17.540 Max. :4.0817   
## Very\_Active\_Distance\_km Moderate\_Active\_Distance\_km Light\_Active\_Distance\_km  
## Min. : 0.000 Min. :0.0000 Min. :0.010   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:2.540   
## Median : 0.570 Median :0.4200 Median :3.665   
## Mean : 1.446 Mean :0.7439 Mean :3.791   
## 3rd Qu.: 2.360 3rd Qu.:1.0375 3rd Qu.:4.918   
## Max. :12.540 Max. :6.4800 Max. :9.480   
## Sedentary\_Active\_Distance\_km Very\_Active\_min Fairly\_Active\_min  
## Min. :0.0000000 Min. : 0.00 Min. : 0.00   
## 1st Qu.:0.0000000 1st Qu.: 0.00 1st Qu.: 0.00   
## Median :0.0000000 Median : 9.00 Median : 11.00   
## Mean :0.0009268 Mean : 25.05 Mean : 17.92   
## 3rd Qu.:0.0000000 3rd Qu.: 38.00 3rd Qu.: 26.75   
## Max. :0.1100000 Max. :210.00 Max. :143.00   
## Lightly\_Active\_min Sedentary\_min Calories Total\_Sleep\_Records  
## Min. : 2.0 Min. : 0.0 Min. : 257 Min. :1.00   
## 1st Qu.:158.0 1st Qu.: 631.2 1st Qu.:1841 1st Qu.:1.00   
## Median :208.0 Median : 717.0 Median :2207 Median :1.00   
## Mean :216.5 Mean : 712.1 Mean :2389 Mean :1.12   
## 3rd Qu.:263.0 3rd Qu.: 782.8 3rd Qu.:2920 3rd Qu.:1.00   
## Max. :518.0 Max. :1265.0 Max. :4900 Max. :3.00   
## Total\_Min\_Asleep Total\_Min\_In\_Bed Total\_Activity\_Min  
## Min. : 58.0 Min. : 61.0 Min. : 2.0   
## 1st Qu.:361.0 1st Qu.:403.8 1st Qu.:206.5   
## Median :432.5 Median :463.0 Median :263.5   
## Mean :419.2 Mean :458.5 Mean :259.5   
## 3rd Qu.:490.0 3rd Qu.:526.0 3rd Qu.:315.5   
## Max. :796.0 Max. :961.0 Max. :540.0

Data description says the data set represents data between March 12, 206 to May 12, 2016, however, it only has data from April 12, 206 to May 12, 2016, which is one month less (or 50% of the data) than stated.

### PROBLEMS WITH THE DATA

* Data description says there are 30 participants but the datasets have 33 unique IDs.
* Data description says the dataset represents data between 03/12/206 to 05/12/2016, however, it only has data from 04/12/2016 to 05/12/2016, which is one month less (or 50% of the data) than stated.
* Sleep Value is not clear. We have 1, 2, and 3 for values.
* Units of measurement for distance are unclear if km or miles was used.
* Date time columns do not have correct data type.
* Weight dataset has 65 NA values which were all in the “Fat” column (65 out of 67 of the data points were NA).

### SOLUTION TO THE PROBLEMS

* Used all 33 user IDs as this gave us 3 extra unique IDs to work with.
* We will use the one month of data for analysis, but this is not an adequate amount of data and should be considered.
* Since there is no way of knowing for sure what the Sleep Values are we won’t be using it for the analysis.
* Unit of measurement for distance was determined by using an online calculator to convert steps to miles and km to determine the best possible match with the data. Variables such as gender and height are factors to determine accurate conversion. For this case we used an average person’s height to get a ballpark. Based on this metric we’ve identified that km was used in this data.
* Converted all date time columns to appropriate date time format.
* Since all of the NA values are in the “Fat” column, and make up 65 out of the 67 data points in that column, a proper analysis cannot be done. Knowing this a new data frame was created without the “Fat” column called “Weight\_Clean”.

## ——————————————————————————–

# **ANALYZE**

Looking at the summary of our data in bits that make sense for quick analysis

### STATS

## Looking at minutes of all activity levels and sleep  
Daily\_Data\_Merged %>%   
 select(Very\_Active\_min,  
 Fairly\_Active\_min,  
 Lightly\_Active\_min,  
 Sedentary\_min,  
 Total\_Min\_Asleep,  
 Total\_Min\_In\_Bed) %>%   
 summary()

## Very\_Active\_min Fairly\_Active\_min Lightly\_Active\_min Sedentary\_min   
## Min. : 0.00 Min. : 0.00 Min. : 2.0 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:158.0 1st Qu.: 631.2   
## Median : 9.00 Median : 11.00 Median :208.0 Median : 717.0   
## Mean : 25.05 Mean : 17.92 Mean :216.5 Mean : 712.1   
## 3rd Qu.: 38.00 3rd Qu.: 26.75 3rd Qu.:263.0 3rd Qu.: 782.8   
## Max. :210.00 Max. :143.00 Max. :518.0 Max. :1265.0   
## Total\_Min\_Asleep Total\_Min\_In\_Bed  
## Min. : 58.0 Min. : 61.0   
## 1st Qu.:361.0 1st Qu.:403.8   
## Median :432.5 Median :463.0   
## Mean :419.2 Mean :458.5   
## 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :796.0 Max. :961.0

* A minimum sedentary time of zero minutes may suggest one or more users used the device only for active activities at least once
* Looking at the maximum difference in time between being asleep and being in bed is a 2.75 hour difference, which can suggest poor sleep practices or illness
* The minimum amount of less than an hour may suggest that a user could have charged the device during sleep but took a nap during the day
* As the level of activity increases the average minutes decreases, which makes sense since it’s not expected for someone to be at a Very Active state for majority of the day, however, we do see slight more minute in Very Active compared to Fairly Active with an average difference of 7 minutes.
* On average, people were in bed awake for 39 minutes
* The combination of average minutes between sedentary (712 min, or 11.8 hour) and time in bed (458 minutes, or 7.6 hours) make up 19.5 hours, which may suggest users have desk jobs

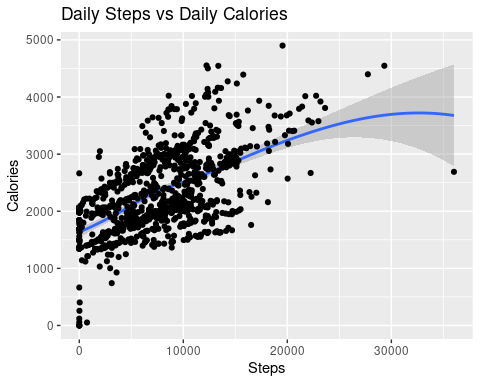
## Looking at distance between all levels of activity and calories  
Daily\_Data\_Merged %>%   
 select(Total\_Steps,  
 Total\_Distance\_km,  
 Very\_Active\_Distance\_km,  
 Moderate\_Active\_Distance\_km,  
 Light\_Active\_Distance\_km,  
 Sedentary\_Active\_Distance\_km,  
 Calories) %>%   
 summary()

## Total\_Steps Total\_Distance\_km Very\_Active\_Distance\_km  
## Min. : 17 Min. : 0.010 Min. : 0.000   
## 1st Qu.: 5189 1st Qu.: 3.592 1st Qu.: 0.000   
## Median : 8913 Median : 6.270 Median : 0.570   
## Mean : 8515 Mean : 6.012 Mean : 1.446   
## 3rd Qu.:11370 3rd Qu.: 8.005 3rd Qu.: 2.360   
## Max. :22770 Max. :17.540 Max. :12.540   
## Moderate\_Active\_Distance\_km Light\_Active\_Distance\_km  
## Min. :0.0000 Min. :0.010   
## 1st Qu.:0.0000 1st Qu.:2.540   
## Median :0.4200 Median :3.665   
## Mean :0.7439 Mean :3.791   
## 3rd Qu.:1.0375 3rd Qu.:4.918   
## Max. :6.4800 Max. :9.480   
## Sedentary\_Active\_Distance\_km Calories   
## Min. :0.0000000 Min. : 257   
## 1st Qu.:0.0000000 1st Qu.:1841   
## Median :0.0000000 Median :2207   
## Mean :0.0009268 Mean :2389   
## 3rd Qu.:0.0000000 3rd Qu.:2920   
## Max. :0.1100000 Max. :4900

* Makes sense for super low (practically zero) for sedentary distance since it’s stationary
* Very active distance maximum of 12.5km may suggest at least one user has relatively good fitness levels, however, this may be a one off since the average is much lower at 1.45km
* Most of the distance is made via lightly active activities at 3.8km, which can be credited to every day walking
* Decent average amount of calories burnt in a day could suggest that fitness is primarily not distance related, but a stationary workout instead

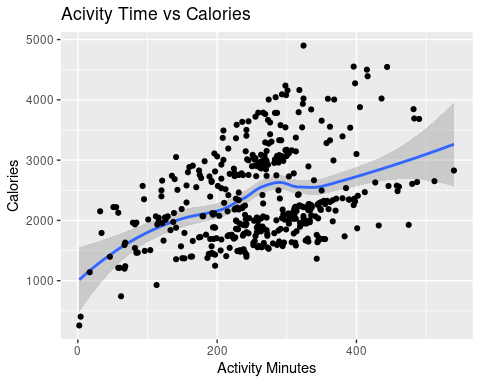
### VISUALIZATIONS

## Steps vs Calories  
ggplot(data = Activity) +   
 geom\_smooth(mapping = aes(x=Total\_Steps, y=Calories)) +  
 geom\_point(mapping = aes(x=Total\_Steps, y=Calories)) +  
 labs(title="Daily Steps vs Daily Calories", x= "Steps", y= "Calories")



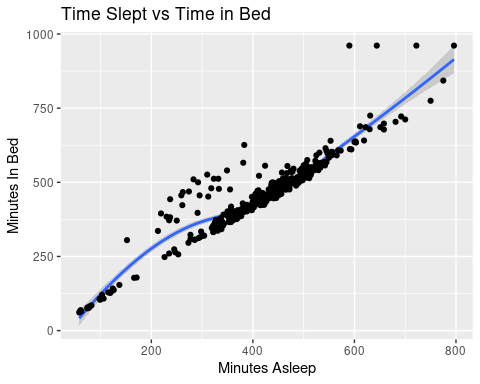
* There is a positive correlation between, which makes sense when you walk you burn calories.
* There are records of 0 steps with calories as high as almost 3,000, which can suggest that some users only used the device for stationary workouts

## Total Activity Time vs. Calories  
ggplot(data = Daily\_Data\_Merged) +   
 geom\_smooth(mapping = aes(x = Total\_Activity\_Min, y = Calories)) +  
 geom\_point(mapping = aes(x = Total\_Activity\_Min, y = Calories)) +  
 labs(title="Acivity Time vs Calories", x= "Activity Minutes", y= "Calories")



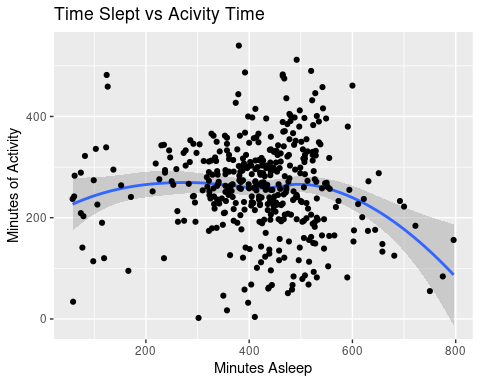
* There is a positive correlation, which makes sense because you burn more calories the longer an activity is

## Sleep vs Time In Bed  
ggplot(data=Daily\_Data\_Merged) +   
 geom\_smooth(mapping=aes(x=Total\_Min\_Asleep, y=Total\_Min\_In\_Bed)) +  
 geom\_point(mapping=aes(x=Total\_Min\_Asleep, y=Total\_Min\_In\_Bed)) +  
 labs(title="Time Slept vs Time in Bed", x= "Minutes Asleep", y= "Minutes In Bed")



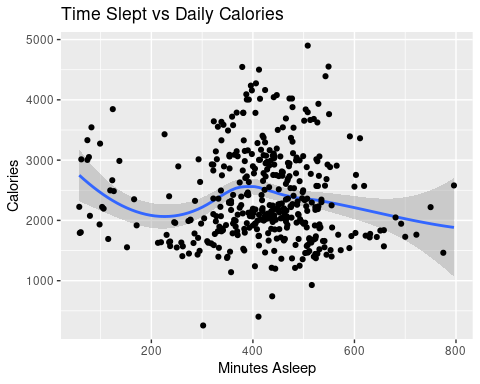
* There is positive correlation which makes sense since you will spend more time in bed awake than asleep

## Sleep vs. Total Activity Time  
ggplot(data=Daily\_Data\_Merged) +   
 geom\_smooth(mapping=aes(x=Total\_Min\_Asleep, y=Total\_Activity\_Min)) +  
 geom\_point(mapping=aes(x=Total\_Min\_Asleep, y=Total\_Activity\_Min)) +  
 labs(title="Time Slept vs Acivity Time", x= "Minutes Asleep", y= "Minutes of Activity")



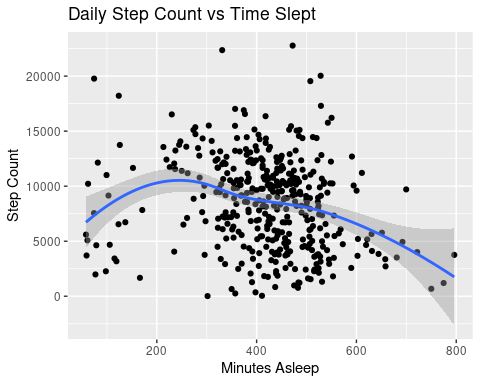
* There is no clear correlation here, however, it does look like those who do sleep more than 600 minutes (10 hrs) have lower activity time, which may be due to the less time in the day for activities
* It does look like when people get between 400 (6.7 hrs) and 600 (10 hrs) of sleep have higher activity minutes

## Sleep vs. Calories  
ggplot(data=Daily\_Data\_Merged) +   
 geom\_smooth(mapping=aes(x=Total\_Min\_Asleep, y=Calories)) +  
 geom\_point(mapping=aes(x=Total\_Min\_Asleep, y=Calories)) +  
 labs(title="Time Slept vs Daily Calories", x= "Minutes Asleep", y= "Calories")



* There’s no correlation here, but there is something interesting here
* There are records of sleep under 100 minutes that have over 2000 calories burned, which can suggest that some users only used the device for stationary workouts and don’t use it for sleep

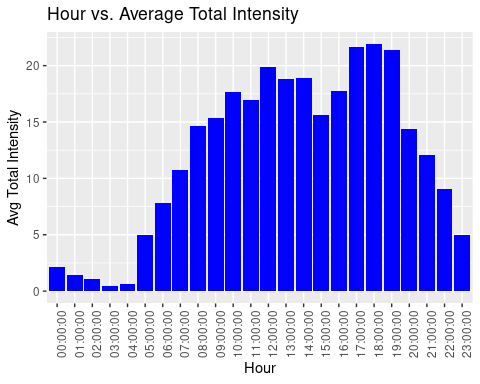
## Sleep vs. Steps  
ggplot(data=Daily\_Data\_Merged) +   
 geom\_jitter(mapping=aes(x=Total\_Min\_Asleep, y=Total\_Steps)) +  
 geom\_smooth(mapping = aes(x = Total\_Min\_Asleep, y = Total\_Steps)) +  
 labs(title="Daily Step Count vs Time Slept", x= "Minutes Asleep", y= "Step Count")



* There are records of sleep over 600 minutes(10hrs) with decreasing step count, which can make sense because there’s less time in the day on the feet

## Hour vs Avg Total Intensity  
ggplot(data=Intensities\_Hour\_Avg, aes(x=Hour, y=Total\_Intensity)) +   
 geom\_col(stat = "identity", fill='blue1') +  
 theme(axis.text.x = element\_text(angle = 90)) +  
 labs(title="Hour vs. Average Total Intensity", x="Hour", y="Avg Total Intensity")

## Warning: Ignoring unknown parameters: stat



* We can see that the highest average total intensity occurs between 5pm and 7pm, and the lowest occurs between 11pm and 5am, which makes sense in terms of when most people sleep and work

## ——————————————————————————–

# **SHARE**

### FINDINGS

* The more steps taken means more calories burned.
* The longer the activity time means more calories burned.
* Since there are multiple records of people burning calories with zero steps and zero minutes of sleep, it can suggest that some people may only use their fitness devices only to track stationary workouts. This can mean that some people don’t wear the watch throughout the day and/or during sleep.
* Those who get between 400 minutes (6.7 hrs) and 600 minutes (10 hrs) of sleep have higher activity minutes throughout the day overall, but between those same times we also have lower activity minutes overall. The lower activity minutes can be due to “rest days” or “busy schedules”.
* We found that those who sleep over 600 minutes (10 hrs) typically have the least amount of steps, which is under 5000. This can be because there’s less time in the day on their feet.
* We can see that the highest average total intensity occurs between 5pm and 7pm, and the lowest occurs between 11pm and 5am, which makes sense in terms of when most people sleep and work

### RECOMMENDATIONS

We want to know how customers are using their fitness smart devices by identifying trends and proving high-level recommendations. Based on our analysis and findings, here are our top recommendations:

#### 1) Evolving/Involving the Community:

Bellabeat is a community of women set on living a healthy lifestyle. We can engage with the community to offer a sense of belonging and motivation.We can evolve the community by involving everyone. Creating scheduled workout times throughout the day where users can workout at the same time as other users and at the end of the scheduled workout (after an hour for instance) we can share to everyone who participated the overall community stats during that workout session. This can make users feel accountable to their workouts and want to contribute to their community where their participation motivates others while also getting motivation from the community. It’s not a surprise to see that the most activity time occurs between 5pm and 7pm, which is normally after work hours. We can take advantage of this by inviting users to participate during this time on the device.

#### 2) Sending Sleep Reminders:

Sleep is important and having the right amount is critical. We can have users input their sleep schedule and let them know if it’s too little or too much sleep based on their input. We can remind users in steps to get them settled down for bed; For example, one hour before scheduled bedtime we can remind users to put down electronic devices, and another reminder 15 minutes before bedtime.

#### 3) Comparing Data:

We’re our own best competition. Giving weekly recaps and comparing it to how users performed the week prior across all metrics can show progress and promote consistency. Additionally, we can have a sense of community for women using Bellabeat devices and provide users regular updates (either daily, weekly, monthly, or yearly) where it shows how they did compare to the overall community during the selected duration.

#### Things to Consider:

There were some limitations discovered along the way with the dataset that must be considered when analyzing this data. Please look at the Data Limitation section for more information.

# THANK YOU