Bellabit\_Report

2023-01-19

# Case Brief

## Bellabit- case Study

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products.

### Ask

Sršen asks to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants you to select one Bellabeat product to apply these insights to in your presentation.

Case Study Roadmap -

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

### Prepare

Importing 2 three datasets:

1. dailyActivity\_merged
2. sleepDay\_merged

### Process

Install packages and Load Libraries

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(janitor)

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

library(dplyr)  
library(ggplot2)  
library(readr)  
library(forcats)  
library(lubridate)

## Loading required package: timechange  
##   
## Attaching package: 'lubridate'  
##   
## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(tinytex)

Import datasets

daily\_activity <- read\_csv("C:/Users/tamon/Desktop/Google Analytics Course/Case Study/Dataset/dailyActivity\_merged.csv")

## Rows: 940 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDate  
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...  
##   
## 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.

sleep\_day <- read\_csv("C:/Users/tamon/Desktop/Google Analytics Course/Case Study/Dataset/sleepDay\_merged.csv")

## Rows: 413 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): SleepDay  
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed  
##   
## 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.

## Cleaning & Formatting  
daily\_activity %>%  
 duplicated() %>%  
 sum()

## [1] 0

sleep\_day %>%  
 duplicated() %>%  
 sum()

## [1] 3

## Remove duplicate rows  
sleep\_day <- sleep\_day %>%  
 distinct() %>%  
 drop\_na()  
  
sleep\_day %>%  
 duplicated() %>%  
 sum()

## [1] 0

## Checking distinct IDs  
n\_distinct(daily\_activity$Id)

## [1] 33

n\_distinct(sleep\_day$Id)

## [1] 24

## Make all columns lower case  
daily\_activity <- rename\_with(daily\_activity, tolower)  
sleep\_day <- rename\_with(sleep\_day,tolower)  
  
## make all columns as date for ease of merge  
daily\_activity <- daily\_activity %>%  
 rename(date = activitydate)  
head(daily\_activity)

## # A tibble: 6 × 15  
## id date total…¹ total…² track…³ logge…⁴ verya…⁵ moder…⁶ light…⁷ seden…⁸  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12… 13162 8.5 8.5 0 1.88 0.550 6.06 0  
## 2 1.50e9 4/13… 10735 6.97 6.97 0 1.57 0.690 4.71 0  
## 3 1.50e9 4/14… 10460 6.74 6.74 0 2.44 0.400 3.91 0  
## 4 1.50e9 4/15… 9762 6.28 6.28 0 2.14 1.26 2.83 0  
## 5 1.50e9 4/16… 12669 8.16 8.16 0 2.71 0.410 5.04 0  
## 6 1.50e9 4/17… 9705 6.48 6.48 0 3.19 0.780 2.51 0  
## # … with 5 more variables: veryactiveminutes <dbl>, fairlyactiveminutes <dbl>,  
## # lightlyactiveminutes <dbl>, sedentaryminutes <dbl>, calories <dbl>, and  
## # abbreviated variable names ¹​totalsteps, ²​totaldistance, ³​trackerdistance,  
## # ⁴​loggedactivitiesdistance, ⁵​veryactivedistance, ⁶​moderatelyactivedistance,  
## # ⁷​lightactivedistance, ⁸​sedentaryactivedistance

sleep\_day <- sleep\_day %>%  
 rename(date = sleepday)  
head(sleep\_day)

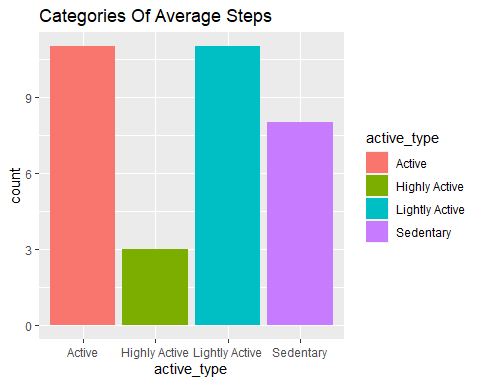
## # A tibble: 6 × 5  
## id date totalsleeprecords totalminutesasleep totalt…¹  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 12:00:00 AM 1 327 346  
## 2 1503960366 4/13/2016 12:00:00 AM 2 384 407  
## 3 1503960366 4/15/2016 12:00:00 AM 1 412 442  
## 4 1503960366 4/16/2016 12:00:00 AM 2 340 367  
## 5 1503960366 4/17/2016 12:00:00 AM 1 700 712  
## 6 1503960366 4/19/2016 12:00:00 AM 1 304 320  
## # … with abbreviated variable name ¹​totaltimeinbed

daily\_activity <- daily\_activity %>%  
 mutate(date = as.Date(date, format = "%m/%d/%Y"))  
  
sleep\_day <- sleep\_day %>%  
 mutate(date = as.Date(date,format ="%m/%d/%Y %I:%M:%S %p" , tz=Sys.timezone()))

### Analyze/Share

1. The below bar chart shows there is not a strong majority of light or sedentary users as per the step activity level.

stepcategories <- daily\_activity %>%   
 group\_by(id) %>%   
 summarise(avg\_step= mean(totalsteps)) %>%   
 mutate (active\_type=case\_when (  
 avg\_step <5000 ~ "Sedentary",  
 avg\_step >=5000 & avg\_step< 8000 ~"Lightly Active",  
 avg\_step>=8000 & avg\_step <12000~"Active",  
 avg\_step >=12000 ~ 'Highly Active'))  
  
ggplot(data = stepcategories)+  
 geom\_bar(mapping = aes(x=active\_type,fill=active\_type))+  
 labs(title = "Categories Of Average Steps")

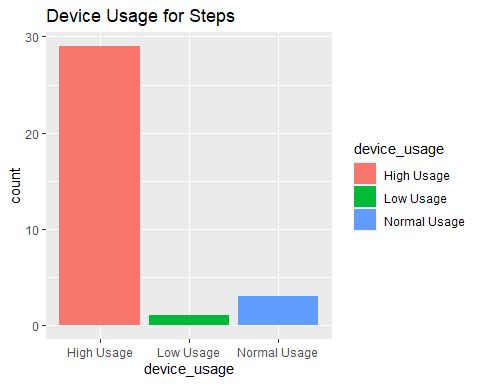


1. The visuals show that more users are using their devices to track steps than sleep. More than 25 users have high usage of the device for tracking steps while only 9 of users have high usage of the device for tracking sleep.

##step\_daily\_active  
step\_daily\_activity <- daily\_activity %>%  
 group\_by(id) %>%  
 summarize(days\_step = n\_distinct(date))  
head(step\_daily\_activity)

## # A tibble: 6 × 2  
## id days\_step  
## <dbl> <int>  
## 1 1503960366 31  
## 2 1624580081 31  
## 3 1644430081 30  
## 4 1844505072 31  
## 5 1927972279 31  
## 6 2022484408 31

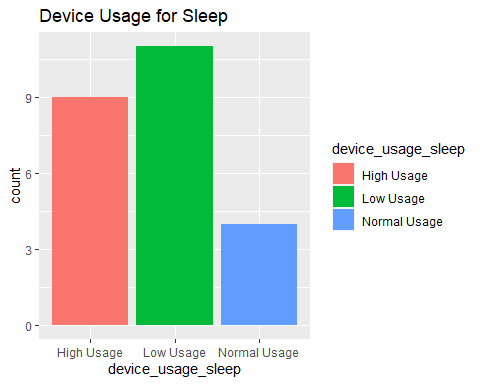
daily\_step\_by\_category <- step\_daily\_activity %>%  
 mutate(device\_usage = case\_when(  
 days\_step <=15 ~ 'Low Usage',  
 days\_step >15 & days\_step <= 25 ~ 'Normal Usage',  
 days\_step >25 ~ 'High Usage',  
 TRUE~'Otherwise'))  
  
p <- ggplot(data = daily\_step\_by\_category)+  
 geom\_bar(mapping = aes(x = device\_usage, fill=device\_usage))+  
 labs(title = "Device Usage for Steps")  
p



##Usage\_sleep\_day  
usage\_sleep\_day <- sleep\_day %>%  
 group\_by(id) %>%  
 summarize(days\_sleep = n\_distinct(date))  
head(usage\_sleep\_day)

## # A tibble: 6 × 2  
## id days\_sleep  
## <dbl> <int>  
## 1 1503960366 25  
## 2 1644430081 4  
## 3 1844505072 3  
## 4 1927972279 5  
## 5 2026352035 28  
## 6 2320127002 1

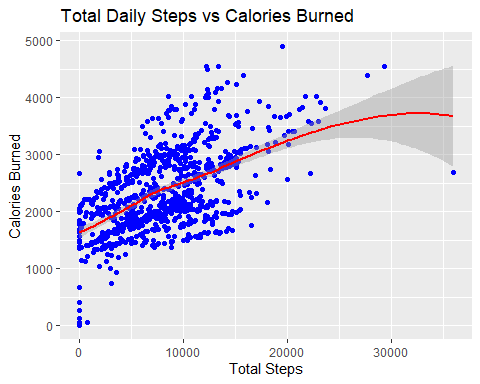
daily\_sleep\_by\_category <- usage\_sleep\_day %>%  
 mutate(device\_usage\_sleep = case\_when(  
 days\_sleep <=15 ~ 'Low Usage',  
 days\_sleep >15 & days\_sleep <= 25 ~ 'Normal Usage',  
 days\_sleep >25 ~ 'High Usage',  
 TRUE~'Otherwise'))  
  
q <- ggplot(data = daily\_sleep\_by\_category)+  
 geom\_bar(mapping = aes(x = device\_usage\_sleep, fill=device\_usage\_sleep))+  
 labs(title = "Device Usage for Sleep")  
q



1. There is a positive correlation between steps and calories burned, but the correlation is weaker when calories burned is correlated to total active minutes in a day.

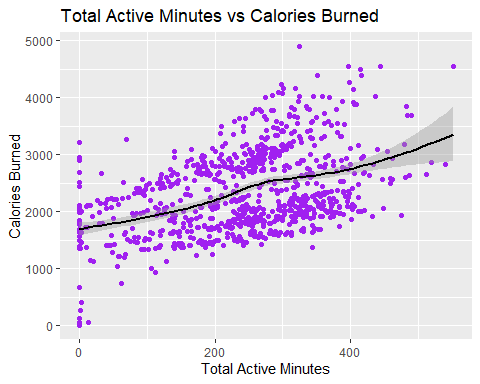
##Total Steps and total calories burned  
  
a<-ggplot(data = daily\_activity)+  
 geom\_point(mapping = aes(x = totalsteps, y = calories),color="blue")+  
 geom\_smooth(mapping = aes(x = totalsteps, y = calories),color="red")+  
 labs(title = "Total Daily Steps vs Calories Burned",x="Total Steps", y="Calories Burned")  
a

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



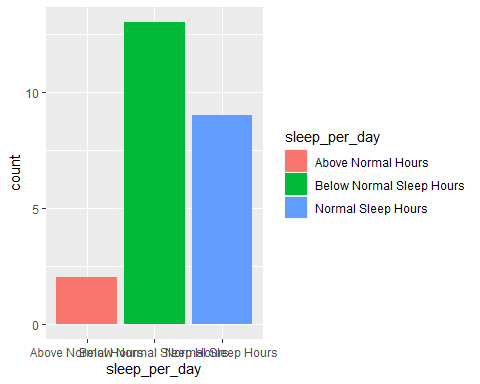
##Total active minutes and total calories burned  
  
daily\_activity\_new <-daily\_activity %>%   
 mutate(total\_active\_minutes= veryactiveminutes+fairlyactiveminutes+lightlyactiveminutes)  
  
b<-ggplot(data = daily\_activity\_new)+  
 geom\_point(mapping = aes(x = total\_active\_minutes, y = calories),color="purple")+  
 geom\_smooth(mapping = aes(x = total\_active\_minutes, y = calories),color="black")+  
 labs(title = "Total Active Minutes vs Calories Burned",x="Total Active Minutes", y="Calories Burned")  
  
b

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



1. There are significant number of users who are getting below normal sleep due to stress/other issues

## Below Normal Sleep Hours  
  
avg\_minutes\_asleep <- sleep\_day %>%  
 group\_by(id) %>%  
 summarize(avg\_asleep = mean(totalminutesasleep))  
  
  
avg\_minutes\_asleep1 <- avg\_minutes\_asleep %>%  
 mutate(sleep\_per\_day = case\_when(  
 avg\_asleep <=420 ~ 'Below Normal Sleep Hours',  
 avg\_asleep >421 & avg\_asleep <= 480 ~ 'Normal Sleep Hours',  
 avg\_asleep >480 ~ 'Above Normal Hours',  
 TRUE~'Otherwise'))  
  
ggplot(data = avg\_minutes\_asleep1)+  
 geom\_bar(mapping = aes(x = sleep\_per\_day, fill=sleep\_per\_day))



### Act

As per the results achieved, it is recommended to promote Bellabeat app with leaf…