**THE BELLABEAT CASE STUDY REPORT**

The case study consists of 6 phases: Ask, Prepare, Process, Analyse, Share and Act.

## **Ask**

### **1.Introduction**

Bellabeat is a high-tech manufacturer of health-focused products for women. They focus mainly on developing wearables and accompanying products that monitor biometric and lifestyle data to help women better understand how their bodies work, and as a result, make healthier lifestyle choices.

**Products**-

**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.

**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.

**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 wellness.

**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.

### **2.Business task**

### Identify trends in how consumers use non-Bellabeat smart devices to apply insights into Bellabeat’s marketing strategy.

**3.Key Stakeholders**

* **Urška Sršen —** Bellabeat’s cofounder and Chief Creative Officer.
* **Sando Mur —** 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.

**Prepare**

The data source used for our case study is FitBit Fitness Tracker Data. This dataset is stored in Kaggle and was made available through Mobius.

**Accessibility and privacy of data:**

Verifying the metadata of our dataset we can confirm it is open-source. The owner has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighbouring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

**Credibility of the data:**

The data is public data from FitBit Fitness Tracker Data stored and widely available in Kaggle and is in long format. It's a dataset from 30 fitbit users that includes minute-level output for physical activity, heart rate, and sleep monitoring. It's a good database segmented in several tables with different aspects of the data of the device with lots of details about the user behaviour. I have no doubt of its source integrity.

**Information about the dataset:**

These datasets were generated by customers to a distributed survey via Amazon Mechanical Turk between 12th April-12th May 2016. There are a total of 18 CSV documents made available for the analysis. Each document represents different quantitative data tracked by Fitbit.

**Does the data ROCCC?**

* Reliable - Med - It has a large amount of data, but could induce a sampling bias as there are only 30 users, we are not sure if the sample is representative of the population as a whole.
* Original - High - It is a first-party data.
* Comprehensive - High - It matches most of Bellabeat's product parameters.
* Current - Low - 12th April-12th May 2016, no real time data and consists of only 2 months of data.
* Cited - High- It is a first-party data.

That is why we will give our case study an operational approach.

# **Process**

I used RStudio to conduct my analysis due to the efficiency and accessibility of the program, its ability to handle huge amount of data and being able to create the visualisations.

### **1.Installing packages and library**

install.packages(“tidyverse”)

install.packages(“dplyr”)

install.packages(“ggplot2”)

install.packages(“lubridate”)

install.packages(“janitor”)

install.packages(“ggpubr”)

install.packages(“skimr”)

install.packages(“here”)

install.packages(“ggrepel”)

install.packages(“readr”)

install.packages(“scales”)

library(tidyverse)

library(dplyr)

library(ggplot2)

library(lubridate)

library(janitor)

library(ggpubr)

library(skimr)

library(here)

library(ggrepel)

library(readr)

library(scales)

I chose these packages to help me with my analysis

### **2.Importing and preparing the data**

daily\_activity <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/dailyActivity\_merged.csv")

daily\_steps <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/dailySteps\_merged.csv")

daily\_sleep <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/sleepDay\_merged.csv")

hourly\_steps <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/hourlySteps\_merged.csv")

hourly\_calories <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/hourlyCalories\_merged.csv")

hourly\_intensities <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/hourlyIntensities\_merged.csv")

weight <- read.csv("C:/Users/Computer/Desktop/12Apr-12May/weightLogInfo\_merged.csv")

### **3.Data cleaning and formatting**

**Check the number of participants for each data set**

I will make sure to check the number of unique users per data frame using n\_unique() before coming on with the cleaning process.

n\_unique(daily\_activity$Id)  
n\_unique(daily\_sleep$Id)  
n\_unique(daily\_steps$Id)  
n\_unique(hourly\_calories$Id)  
n\_unique(hourly\_intensities$Id)  
n\_unique(hourly\_steps$Id)  
n\_unique(weight$Id)

All datasets have 33 participants each, except for the daily\_sleep and weight datasets, which have 24 and 8 participants respectively. I will exclude the weight dataset because 8 participants are too small a sample size to draw meaningful conclusions and make recommendations.

I used **rm(weight)** to remove the weight table.

## **Check for duplicates and Missing Values**

I will check for any duplicates in the data

sum(duplicated(daily\_activity))  
sum(duplicated(daily\_sleep))  
sum(duplicated(daily\_steps))  
sum(duplicated(hourly\_calories))  
sum(duplicated(hourly\_intensities))  
sum(duplicated(hourly\_steps))

I found **3** duplicates in daily\_sleep table.

I used distinct () to remove the said duplicates.

daily\_sleep <- daily\_sleep %>% distinct()

then, I proceeded to remove missing values from all the tables using drop\_na() to ensure

accuracy.

daily\_activity<-daily\_activity %>% drop\_na()

daily\_sleep<-daily\_sleep %>% drop\_na()

daily\_steps<-daily\_steps %>% drop\_na()

hourly\_calories<-hourly\_calories %>% drop\_na()

hourly\_intensities<-hourly\_intensities %>% drop\_na()

hourly\_steps<-hourly\_steps %>% drop\_na()

-I cleaned and formatted the column names of every table to make sure of its accuracy and consistency.

clean\_names(daily\_activity)  
daily\_activity <- rename\_with(daily\_activity, tolower)  
clean\_names(daily\_sleep)  
daily\_sleep <- rename\_with(daily\_sleep, tolower)  
clean\_names(daily\_steps)  
daily\_steps <- rename\_with(daily\_steps, tolower)  
clean\_names(hourly\_calories)  
hourly\_calories <- rename\_with(hourly\_calories, tolower)  
clean\_names(hourly\_intensities)  
hourly\_intensities <- rename\_with(hourly\_intensities, tolower)  
clean\_names(hourly\_steps)  
hourly\_steps <- rename\_with(hourly\_steps, tolower)

## 

## **Make date and time columns consistent**

daily\_activity <- daily\_activity %>%

+ rename(date = activitydate) %>%

+ mutate(date = mdy(date))

daily\_sleep <- daily\_sleep %>%

rename(date = sleepday) %>%

mutate(date = mdy\_hms(date, tz = Sys.timezone()))

hourly\_calories <- hourly\_calories %>%

rename(date\_time = activityhour) %>%

mutate(date\_time = as.POSIXct(date\_time, format = "%m/%d/%Y %I:%M:%S %p", tz = Sys.timezone()))

hourly\_intensities <- hourly\_intensities %>%

rename(date\_time = activityhour) %>%

mutate(date\_time = as.POSIXct(date\_time, format = "%m/%d/%Y %I:%M:%S %p", tz = Sys.timezone()))

hourly\_steps <- hourly\_steps %>%

rename(date\_time = activityhour) %>%

mutate(date\_time = as.POSIXct(date\_time, format = "%m/%d/%Y %I:%M:%S %p", tz = Sys.timezone()))

# **Merging data sets**

# I will merge the daily\_activity and daily\_sleep data sets together.

daily\_activity\_sleep <- merge(daily\_activity, daily\_sleep, by = c("id", "date"))

# **Analyze and Share (Visualization)**

I will analyze trends of the users of FitBit and determine if that can help me on BellaBeat’s marketing strategy.

**1.Summarize and explore each data set**

daily\_activity %>%

select(totalsteps,

totaldistance,

sedentaryminutes, calories) %>%

summary()

daily\_activity\_sleep %>%

select(totalminutesasleep,) %>% summary()

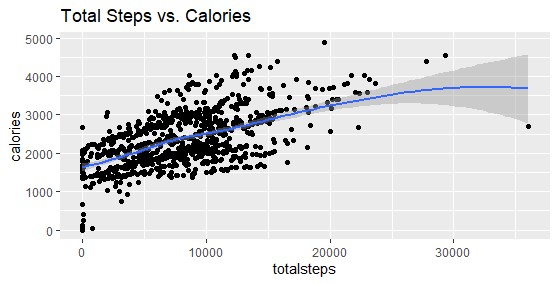
* Average total steps are 7638 in a day. The daily recommended number of steps to be taken per day is 7500.
* Sedentary minutes on an average is 991(~17 hours). This needs to be reduced.
* Majority of the participants are light users.
* Participants sleep for an average of 419 minutes (~7 hours).
* A total of 97 calories is burned per hour on an average.

**2. Steps taken and Calories burned**

I made a visualization to check if there’s a correlation between steps taken and the amount of calories burned.

ggplot(data = daily\_activity, aes(x = totalsteps, y = calories)) +

geom\_point() + geom\_smooth() + labs(title ="Total Steps vs. Calories")



I see a positive correlation between total steps taken and the amount of calories burned.​

**3. Steps per weekday**

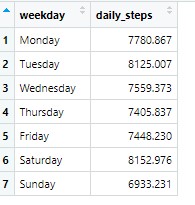
To know what day of the week users are more active.

weekday\_steps <- daily\_activity %>%

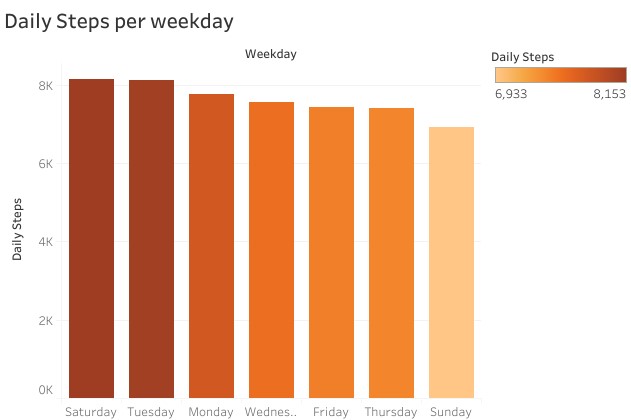
mutate(weekday = weekdays(date))

weekday\_steps$weekday <-ordered(weekday\_steps$weekday, levels=c("Monday", "Tuesday", "Wednesday", "Thursday","Friday", "Saturday", "Sunday"))

weekday\_steps <-weekday\_steps %>%  
 group\_by(weekday) %>%summarize (daily\_steps = mean(totalsteps))



Now that we have gotten the steps taken for each day of the week. I will make a visualization to better understand the data using Tableau.

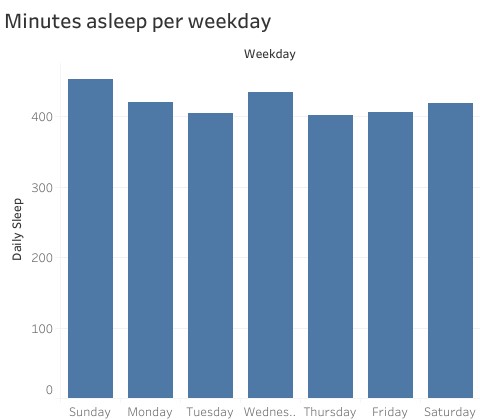


- Users take recommended number of 7500 steps a day excepts for Sundays.​

**4. Sleep per weekday**

weekday\_sleep <- daily\_sleep %>%  
 mutate(weekday = weekdays(date))

weekday\_sleep$weekday <-ordered(weekday\_sleep$weekday, levels=c("Monday", "Tuesday", "Wednesday", "Thursday","Friday", "Saturday", "Sunday"))  
  
weekday\_sleep <-weekday\_sleep %>% group\_by(weekday) %>%  
 summarize (daily\_sleep = mean(totalminutesasleep))



-In the graphs above we can deduce that users don’t take the recommended amount of sleep of 8 hours​.

**5.Hourly intensities throughout the day**

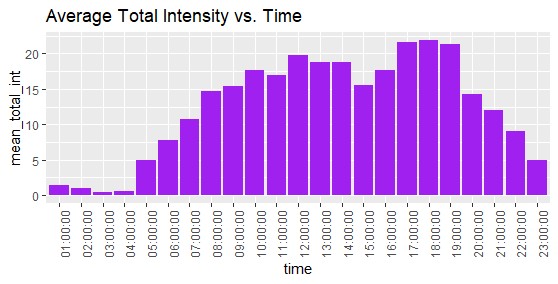
**​-**Split the datetime column into date and time columns

hourly\_intensities <- hourly\_intensities %>% separate(date\_time, into = c("date", "time"), sep= " ")

head(hourly\_intensities)

hourly\_intensities <- hourly\_intensities %>%

group\_by(time) %>%drop\_na() %>% summarise(mean\_total\_int = mean(totalintensity))

-I will make a visualization of the hourly intensities data.

ggplot(data = hourly\_intensities, aes(x = time, y = mean\_total\_int)) + geom\_histogram(stat = "identity", fill='purple') +

theme(axis.text.x = element\_text(angle = 90)) + labs(title="Average Total Intensity vs. Time")

* I found out that people are more active between 5am and 11pm.
* Most of the activity happens between 5 pm and 7 pm — I suppose that when people are done with work for the day, they go to the gym or take a walk. We can use this period of time to remind and motivate users to go for a run or walk using the Bellabeat app.

# **6 Hourly steps throughout the day**

-Separating the datetime column into date and time column.

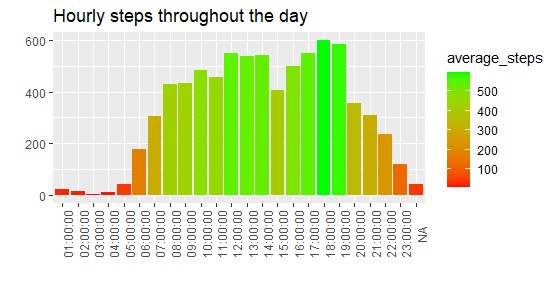
hourly\_steps <- hourly\_steps %>%

separate(date\_time, into = c("date", "time"), sep= " ") %>%mutate(date = ymd(date))

head(hourly\_steps)

-I will make a visualization of the hourly steps throughout the day.

hourly\_steps %>%  
 group\_by(time) %>%  
 summarize(average\_steps = mean(steptotal)) %>%  
 ggplot() +  
 geom\_col(mapping = aes(x=time, y = average\_steps, fill = average\_steps)) +   
 labs(title = "Hourly steps throughout the day", x="", y="") +   
 scale\_fill\_gradient(low = "red", high = "green")+  
 theme(axis.text.x = element\_text(angle = 90))



* Users are more active from 8am to 7pm and they walk more steps from 12pm to 2pm and 5pm to 7pm.
* I assume that most users are working class women and the time more steps are recorded suggests that they have their lunch break (12pm-2pm) and close for the day (5pm-7pm) during those periods.

**7. Type of user based on the number of days smart device was used**

Calculate the number of users that use their smart device on a daily basis, classifying our sample into three categories knowing that the duration of the survey is 31 days:

* High user — users who use their device for 21–31 days
* Moderate user — users who use their device for 10–20 days
* Low user — users who use their device for 1–10 days

Next, I will create a new data frame grouping by id, calculate the number of days smart device was used and create a new column with the classification explained above.

daily\_use <- daily\_activity\_sleep %>%

group\_by(id) %>%

summarize(days\_used=sum(n())) %>%

mutate(user\_type= case\_when(

days\_used >= 1 & days\_used <= 10 ~ "low user",

days\_used >= 11 & days\_used <= 20 ~ "moderate user",

days\_used >= 21 & days\_used <= 31 ~ "high user",

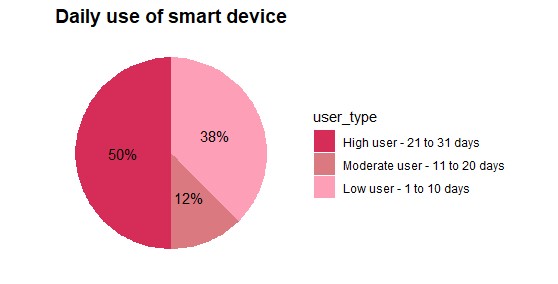
))

head(daily\_use)

-Create a percentage data frame to better visualize the results in the graph

daily\_use\_percent <- daily\_use %>%  
 group\_by(user\_type) %>%  
 summarise(total = n()) %>%  
 mutate(totals = sum(total)) %>%  
 group\_by(user\_type) %>%  
 summarise(total\_percent = total / totals) %>%  
 mutate(labels = scales::percent(total\_percent))  
  
daily\_use\_percent$user\_type <- factor(daily\_use\_percent$user\_type, levels = c("high user", "moderate user", "low user"))  
  
head(daily\_use\_percent)

**Make a visualization of the smart device usage per user**

daily\_use\_percent %>%  
 ggplot(aes(x = "",y = total\_percent, fill = user\_type)) +  
 geom\_bar(stat = "identity", width = 1)+  
 coord\_polar("y", start=0)+  
 theme\_minimal()+  
 theme(axis.title.x= element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),   
 panel.grid = element\_blank(),   
 axis.ticks = element\_blank(),  
 axis.text.x = element\_blank(),  
 plot.title = element\_text(hjust = 0.5, size=14, face = "bold")) +  
 geom\_text(aes(label = labels),  
 position = position\_stack(vjust = 0.5))+  
 scale\_fill\_manual(values = c( "#d62d58","#db7980","#fc9fb7"),  
 labels = c("High user - 21 to 31 days",  
 "Moderate user - 11 to 20 days",  
 "Low user - 1 to 10 days"))+  
 labs(title="Daily use of smart device")

* 50% of the users of our sample use their device frequently — between 21 to 31 days.
* 12% are moderate users (they use their device for 11 to 20 days).
* 38% of our sample rarely used their device.

**Act**

**Conclusion and Recommendations**

After analysing the FitBit Fitness data, I will respond to the business task of helping Bellabeat on its marketing strategy based on my results;

I would advise that for further analysis, we use tracking data from Bellabeat’s device.

The datasets used have a small sample and can be biased since we didn’t have any demographic details of users.

* **Daily notification on steps taken**

Bellabeat can encourage people to take at least 8, 000 explaining the benefits for their health. Sending notifications daily at different times will make users conscious of the number of steps achieved so far and encourage them to meet the set target of 8000 steps according to CDC. The app can also educate users on the health benefits of walking the daily recommended number of steps.

* **Daily notification on sleep**

From the results of my analysis, I can see that users have less than the recommended amount of sleep in a day. We can enable a feature on the app that allows user set up a desired time to go to sleep and receive a notification minutes before to prepare to sleep or set up an alarm to sleep.

* **Notification based on user needs**

If users want to lose weight, controlling daily calorie consumption is a good idea. The Bellabeat can suggest some ideas for low-calorie food and tips to lose weight to such users.

* **Menstrual Cycle tracker**

Since, it’s mainly targeted to women, it’ll be great to add a feature which tracks menstrual cycle monthly, and provide educational tips and knowledge about the PMS, Menstrual cycle and even Pregnancy planning, empowering them to make important health choices.

* **Reward system**

We can develop a reward system based on users' activity levels on the app. The app could feature different stages that users progress through, determined by their daily step count. Users would need to sustain their activity level for a specified duration, such as a month, to advance to the next stage. Upon reaching each stage, users would earn a certain number of stars, which could be redeemed for discounts on other Bellabeat products.

* **Recommendation for the online campaign**

Make sure the online campaign portrays the Bellabeat app more than just a fitness activity app. It should be seen as a guide that empowers women to strike a balance in their personal and professional life and their health habits by educating and motivating them through daily app recommendations.