# BELLABEAT CASE STUDY

## INTRODUCTION

Bellabeat is a successful company founded in 2013, focused on producing high-tech health gadgets basically for women, to allow them monitor and collect data on their daily activity, sleep pattern, stress, and also help them better understand their reproductive health so as to cultivate a healthier lifestyle.

The cofounder and Chief Creative Officer of Bellabeat , Urška Sršen, believes that analyzing smart device fitness trackers could help unlock new growth opportunities for the budding company which has the potential of becoming a larger player in the global smart device market.

The company's major products are,

- ***The Bellabeat app*** which connects to the other line of wellness products to provide users with health data related to their activity, sleep pattern, menstrual cycle, and other things related to women.
- ***The Leaf*** is a classic wellness tracker which can be worn as a necklace, a bracelet or a clip. It connects to ***the Bellabeat app*** to track activity, sleep, and stress.
- **T*he Time***, which is a wellness watch with smart technology that also tracks the user's daily wellness, working just as ***the Leaf*** only with the extra-telling time.
- **T*he Spring***, a water bottle that tracks daily water intake using smart technology too, ensuring that the user is appropriately hydrated throughout the day. ***The Spring*** connects to ***the Bellabeat app*** to track hydration levels.

As a junior data analyst, in the marketing analyst team at Bellabeat, I have been asked to develop a marketing strategy using one of the company's products by analysing smart device data to gain insight on how consumers are using their smart devices.

In order to achieve this goal, we will use a case study provided, following standard steps of the data analysis process which are, ***ask, prepare, process, analyze, share,*** and ***act.***

## 1. ASK PHASE:

- **Business task-**To analyze smart device usage data for insight into how other consumers use non-Bellabeat smart devices. For the purpose of this presentation, we will focus primarily on the "Bellabeat's Leaf" which tracks activity, sleep and stress in women.

Using the following questions as a guide;

a) What are some trends in smart device usage?
b) How could these trends apply to Bellabeat customers?
c) How could these trends help influence Bellabeat marketing strategy?

- **Key stakeholders-**
  a) **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer.
  b) **Sando Mur**: Bellabeat's cofounder, mathematician and a key member of the Bellabeat executive team.
  c) **Bellabeat Marketing analytics team**: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

## 2. PREPARE PHASE:

**Data location:** The dataset used for this case study, FitBit Fitness Tracker Data(CC0: Public Domain), is located in Kaggle made available through Möbius , a data scientist, in Melbourne, Victoria, Australia. Since this dataset is open sourced, we can use it without seeking permission from its owner.

There are 18 .csv files in the Fitabase data, compiled between 12th April 2016 and 12th May 2016. The files are in the Long format because of values that are repeated in the first column and the many rows that contain personal fitness tracker data for only thirty (30) Fitbit users.

Thirty eligible Fitbit users consented to the submission of personal tracker data, which include minute-level output for daily physical activity, heart rate, weight information and sleep monitoring that can be used to explore users' habits.

To make sure our dataset is credible and without bias, we use the **ROCCC** (reliable, original, comprehensive, current, and well cited) approach.

- **Reliability**: Due to the relatively small sample size, the dataset may not be as reliable. It was noticed that some of the files contained information for either less or more than the 30 (unique IDs) respondents expected which is the minimum sample size according to Central Limit Theorem.
- **Originality**: The information is sourced from a third party survey distributed by Amazon Mechanical Turk. So it is considered not fully original.
- **Comprehensive:** Apart from the age, gender and location details, most of the data type corresponds to those collected by the Bellabeat's products.
- **Current:** Though the dataset was last updated three (3) years ago and sourced in 2016, it is still relevant for this case study (May not be recommended in real time for certain business analysis).
- **Cited:** FitBit Fitness Tracker Data (kaggle.com) MÖBIUS on Kaggle. Generated by respondents to a distributed survey via Amazon Mechanical Turk.

**3. PROCESS PHASE:** In this case study, “R” is used as the main tool for data processing to prepare, clean, analyze, and visualize making it possible to explore the entire dataset in one environment. To ensure the integrity of the data, getting it ready for the next phase (analyze), the following steps were taken.

**#Installation and loading of packages:** Before installing and loading the requir ed packages, the work environment was set;

```
# set as working directory
setwd("/cloud/project/Fitabase Data 4.12.16-5.12.16")
list.files()

#Install packages
install.packages("tidyverse")
install.packages("here")
install.packages("readr")
install.packages("dplyr")
install.packages("skimr")
install.packages("janitor")
install.packages("ggplot2")
install.packages("lubridate")
install.packages("ggforce")
install.packages("ggpubr")

# Load packages
library(tidyverse)
library(here)
library(readr)
library(dplyr)
library(skimr)
library(janitor)
library(ggplot2)
library(lubridate)
library(ggforce)
library(data.table)
```

**#Imported some files from the folder after unzipping them and storing them in a folder on our desktop.**
**Unzip the files.**

```
dailyActivity<-read.csv("dailyActivity_merged.csv")
sleepDay<-read.csv("sleepDay_merged.csv")
Mets<-read.csv("minuteMETsNarrow_merged.csv")
dailySteps<-read.csv("dailySteps_merged.csv")
```

**#Previewed the files to see what we would be working on.**

```
head(dailyActivity)
```

```
str(dailyActivity)

head(dailySteps)
str(dailySteps)

head(Mets)
str(Mets)

head(sleepDay)
str(sleepDay)
```

**#Cleaning and formatting the data: checking for the unique Id and the number of unique Id in each file.**

```
n_distinct(dailyActivity$Id)
n_distinct(sleepDay$Id)
n_distinct(Mets$Id)
n_distinct(dailySteps$Id)

dailyActivity<-dailyActivity%>%
  distinct()%>%
  drop_na()

sleepDay<-sleepDay%>%
  distinct()%>%
  drop_na()

Mets<-Mets%>%
  distinct()%>%
  drop_na()

dailySteps<-dailySteps%>%
  distinct()%>%
  drop_na()
```

**#Check for duplicates or n/a have been removed to make sure the data has int egrity.**

```
Sum(duplicated(dailyActivity))
sum(duplicated(sleepDay))
sum(duplicated(Mets))
sum(duplicated(dailySteps))
```

**#Organize the data frames, by changing the date field in the data frames ,converting dates to days, and making copies of some of the data frames so we would not loose important data and renaming some data frames .**

```
library(data.table)
sleepDay2<- copy(sleepDay)
sleepDay$SleepDay1<- as.Date(sleepDay$SleepDay,format = "%m-%d-%y %H:%M:%S")

head(sleepDay$sleepDay1)
sleepDay$SleepDay1<- mdy_hms(sleepDay$SleepDay)
dailyActivity$ActivityDate1<- mdy(dailyActivity$ActivityDate)
View(dailyActivity)

Mets$metsDate<-mdy(Mets$ActivityMinute)
View(Mets)  #Converting Activitydate column from date to weekdays Monday-Saturday from the dailyActivity_merged.csv

dailyActivity<- dailyActivity %>%
  mutate(weekday1 = weekdays(as.Date(ActivityDate, "%m/%d/%Y")))

dailySteps<- dailySteps %>%
  mutate(weekday1 = weekdays(as.Date(ActivityDay, "%m/%d/%Y")))

MetsCopy<-copy(Mets)
Mets2<-Mets%>%
View(Mets2)

Mets<- Mets %>%
  mutate(weekday1 = weekdays(as.Date(ActivityMinute, "%m/%d/%Y")))

sleepDay<- sleepDay%>%
  mutate(weekday1=weekdays(as.Date(SleepDay, "%m/%d/%Y")
Mets$ActivityMinuteDate2<- as.Date(Mets$ActivityMinute, '%m/%d/%Y')
```

**#Check and clean the relevant data in the Heartrate file.**

```
heartRate<- read.csv("heartrate_seconds_merged.csv")
str(heartRate)
head(heartRate)
View(heartRate)
n_distinct(heartRate$Id)
heartRate<-heartRate%>%
  distinct()%>%
  drop_na()
 sum(duplicated(heartRate))
```

**#Rename columns with date variables to Date.**

```
 colnames(dailyActivity)[colnames(dailyActivity) == 'ActivityDate1'] <- 'date'
```

```
colnames(Mets2)[colnames(Mets2) == 'ActivityMinuteDate2'] <- 'date'
colnames(sleepDay)[colnames(sleepDay) == 'sleepDay1'] <- 'date'
```

**4. ANALYZE PHASE:** In this phase, we took the following steps to see what we could deduce from the given data, finding relationships between the relevant variable . We could not use the weight and heart rate data though quite important, because th e data sample size was too small to draw meaningful conclusions.

**#Look for patterns and trends, merge three of the data sets we cleaned to de duce relationships between the variables, excluding the ones that were duplica tes of what we already had ("The Calories file").**

```
dailyActivityMetsSleep<- merge(dailyActivityMets,sleepDay, by=c("Id", "date"))
glimpse(dailyActivityMetsSleep)
```

**#create a new column for weekdays in "dailyActivityMets"**

```
dailyActivityMets<- dailyActivityMets%>%
  mutate(weekday1=weekdays(as.Date(date, "%m/%d/%Y")))
```

**# Calculate the daily average mean**

```
dailyAverage <-dailyActivityMetsSleep%>%
  group_by(Id)%>%
  summarise (mean_dailySteps=mean(TotalSteps),mean_dailyCalories=mean(Calor
ies),mean_dailyMets=mean(Avg_METs),mean_dailyMinutesAsleep=mean(TotalMinu
tesAsleep,mean_dailyMinutesInBed=mean(TotalMinutesInBed)))

head(dailyAverage)
```

**# Categorize the respondents according to the average daily steps taken.**

```
userType <- dailyAverage %>%
  mutate(userType = case_when(
    mean_dailySteps < 5000 ~ "sedentary",
    mean_dailySteps >= 5000 & mean_dailySteps < 7499 ~ "lightly active",
    mean_dailySteps >= 7500 & mean_dailySteps < 9999 ~ "fairly active",
    mean_dailySteps >= 10000 ~ "very active"))

head(userType)
```

**#Calculate % of each user type**

```
userTypePercent <- userType %>%
  group_by(userType) %>%
  summarise(total = n()) %>%
  mutate(totals = sum(total)) %>%
  group_by(userType) %>%
  summarise(totalPercent = total / totals) %>%
  mutate(labels = scales::percent(totalPercent))
```

```
 userTypePercent$userType <- factor(userTypePercent$userType , levels = c("very active", "fairly active", "lightly active", "sedentary"))

  head(userTypePercent)
```

**# Deduce the trends in dataset through visuals**

```
 userTypePercent %>%
   ggplot(aes(x="",y=totalPercent, fill=userType)) +
   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")) +
   scale_fill_manual(values = c("green","green 4", "olive Drab", "aquamarine")) +
   geom_text(aes(label = labels),
             position = position_stack(vjust = 0.5))+
   labs(title="User type distribution")
```

The Pie chart shows that the majority of users are “lightly active,” while the other thre e categories each make up around 21% of the total.

**# Data frame with sleep, METs, steps and weekday**

```
 weekdayActivityMetsSleep <- dailyActivityMetsSleep %>%
   mutate(weekday1 = weekdays(date))
   weekdayActivityMetsSleep$weekday <-ordered(dailyActivityMetsSleep$weekday1, levels=c("Monday", "Tuesday", "Wednesday", "Thursday”, "Friday", "Saturday", "Sun day"))

head(weekdayActivityMetsSleep)

dailyAverageCopy<-data.frame(dailyAverage)
 str(dailyAverageCopy)
 dailyAverage2 <- dailyAverageCopy%>%
   group_by(Id) %>%
   summarize(dailySleep = mean(mean_dailyMinutesAsleep), dailymets=mean(mean _dailyMets,daily_steps=mean(mean_dailySteps),daily_calories=mean(mean_dailyC alories)))
 glimpse(dailyAverage2)
```

**# Create more visuals to see the types of correlations within the various Daily Averages.**

Average daily steps vs Average daily calories.

- ggplot(data=dailyAverage)+
  geom_point(mapping=aes(x=mean_dailySteps,y=mean_dailyCalories),color="red")+
  geom_smooth(mapping=aes(x=mean_dailySteps,y=mean_dailyCalories))+
  labs(title="Average daily Steps vs Average daily Calories" )

**#Rename the Avg_METs column to mets1**

```
colnames(weekdayActivityMetsSleep)[colnames(weekdayActivityMetsSleep) == "Avg_METs"] <- "mets1"
```

**Calories and METs**

- ggplot(data=weekdayActivityMetsSleep)+
  geom_point(mapping=aes(x=Calories, y=mets1),color="purple")+
  labs(title="Calories and METs")

**Average daily sleep time vs Average daily METs.**

- ggplot(data=dailyAverage)+
  geom_point(mapping=aes(x=mean_dailyMinutesAsleep,y=mean_dailyMets, color="red"))+
  geom_smooth(mapping=aes(x=mean_dailyMinutesAsleep,y=mean_dailyMets))+
  labs(title="Average daily sleeptime vs Average daily METs")

**Average daily steps vs Average daily METs**

- ggplot(data=dailyAverage)+
  geom_point(mapping=aes(x=mean_dailySteps,y=mean_dailyMets))+
  geom_smooth(mapping=aes(x=mean_dailySteps,y=mean_dailyMets), color="green")+
  labs(Title="Average daily Steps vs Average daily METs")

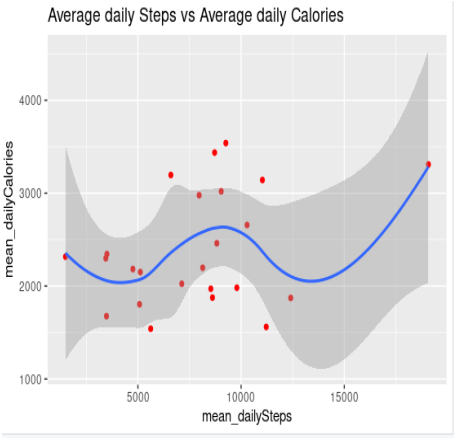
**Daily steps according to week day.**

- ggplot(data=weekdayMetsSteps)+
  geom_bar(mapping = aes(x=dailySteps, fill=weekday ))+
  labs(title="Daily Steps", subtitle = ("According to weekday"))

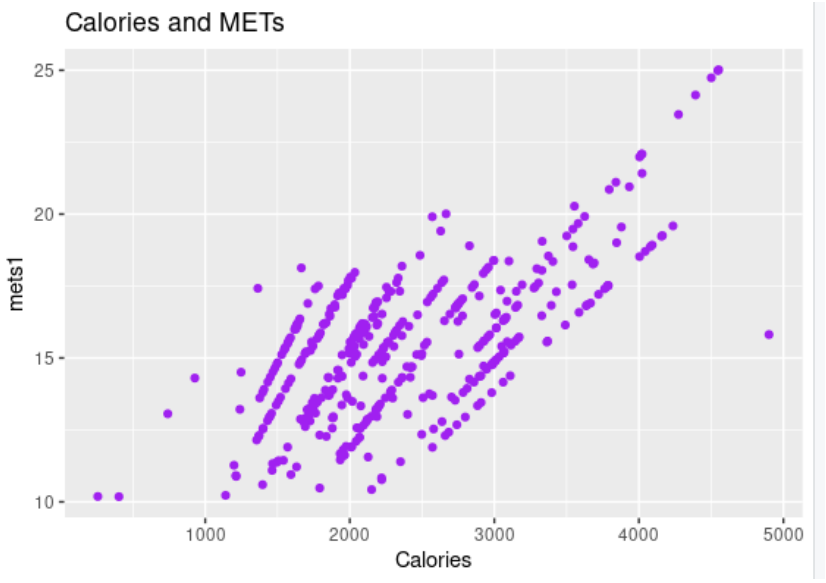
**Daily METs according to week day.**

- ggplot(data=weekdayMetsSteps)+
  geom_bar(mapping = aes(x=dailymets, fill=weekday ))+
  labs(title="Daily METs", subtitle = ("According to weekday")

### 5. **SHARE PHASE:**

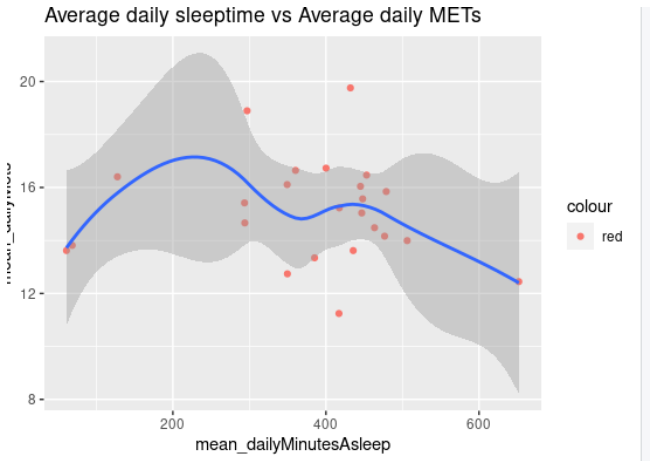



Based on the visual, the correlation between the two variables, **average daily steps** and **average daily calories**, appears to be positive but nonlinear. The nonlinear outcome may likely have been influenced by other factors like feeding, health status or other activity.
The number of mean daily steps increases as the mean daily burnt calories also increase. The blue line indicates a trend where there is an initial decrease in calories as steps increase up to a certain point, after which the calories begin to increase with an increasing number of steps.

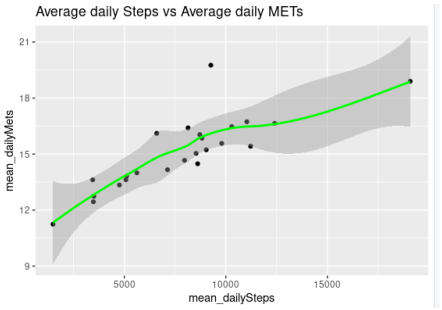



From this scatter plot, we see a positive relationship between the **calories** burned and the **METs** of the users because the number of calories burned increases as the METs also increase. Note that this positive correlation does not necessarily imply causation.

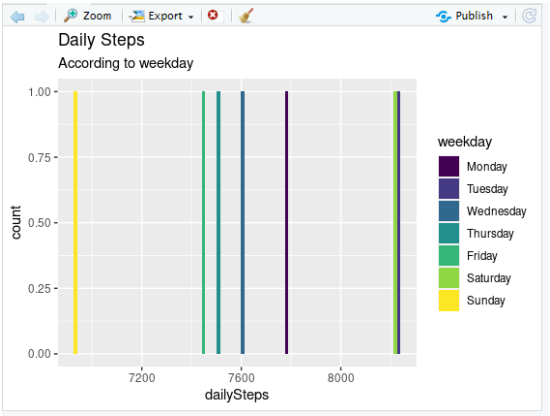
METs stands for the metabolic equivalent of task. One MET is the amount of energy used while sitting quietly. Physical activities may be rated using METs to indicate their intensity. For example, reading may use about 1.3 METs while running may use 8-9 METs. It's advised that healthy adults (between the ages of 18 and 65) should expend between 450 to 750 METs min per week.

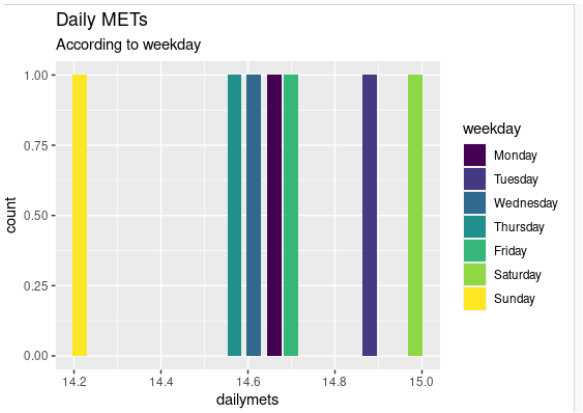



In the above plot we see an inverse relationship though not a very strong one, between average daily sleep time and average daily METs, as sleep increases the METs decrease and vice versa. This trend is as expected because more sleep generally implies fewer calories burnt.




From the above plot we notice a positive correlation between the two variables, average daily steps and average daily METs. The mean daily steps increase as the mean daily METs also increase. We see an almost linear and positive trend meaning that the steps are effective when measured in METs (which is the basic unit for calories burnt).




From the two bar graphs we noticed that Sunday (yellow) had the lowest daily steps and METs while Tuesday and Saturdays show the highest steps and METs. The trend is similar to a setting where most users have a relaxed day on Sundays and more active on Saturdays.

## 6. ASK PHASE:

**Recommendations:**

Based on the focus of the Bellabeat company to improve the health of women, providing users with the relevant data to both understand and improve their health will go a long way in realising this goal.

1. The Bellabeat app should incorporate user settings to include personal goals based on individual age, weight, gender and general health status since all these factors affect the actual calorie burn requirements of an individual.

2. From the average daily steps and average METs analysis, Bellabeat app can also inspire users by presenting a real time awareness in graph form or trend so that progress can be visualised for calories burnt and actual METs.

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?

Share Once you have completed your analysis, create your data visualizations. The visualizations should clearly communicate your high-level insights and recommendations. Use the following Case Study Roadmap as a guide: Case Study Roadmap - Share Guiding questions

- Were you able to answer the business questions?
- What story does your data tell?
- How do your findings relate to your original question?
- Who is your audience? What is the best way to communicate with them?
- Can data visualization help you share your findings?
- Is your presentation accessible to your audience? Key tasks

1. Determine the best way to share your findings.
2. Create effective data visualizations.
3. Present your findings.
4. Ensure your work is accessible. Deliverable Supporting visualizations and key findings

Follow these steps:
1. Take out a piece of paper and a pen and sketch some ideas for how you will visualize the data.
2. Once you choose a visual form, open your tool of choice to create your visualization. Use a presentation software, such as PowerPoint or Google Slides; your spreadsheet program; Tableau; or R.
3. Create your data visualization, remembering that contrast should be used to draw your audience’s attention to the most important insights. Use artistic principles including size, color, and shape.
4. Ensure clear meaning through the proper use of common elements, such as headlines, subtitles, and labels.
5. Refine your data visualization by applying deep attention to detail

We did not have age, gender and other data sets that could have brought better insight.

/cloud/project/Fitabase Data 4.12.16-5.12.16/case_study2_Bellabeat.html (posit.cloud)

Focusing on 10,000 steps a day could be a misstep | Diet and Nutrition | Heart | Prevention | UT Southwestern Medical Center (utswmed.org) UTSouthwestern article

Metabolic Equivalent of Task (MET) values are a way to estimate how many calories are burned during a specific physical activity, according to the American Council on Exercise. This table MET Values for 800+ Activities - Golf - ProCon.org