Bellabeat case study

Amit

3/22/2022

## Case Study - How can a Wellness Technology Company play it smart?

### Business Task

Bellabeat, a small company in the smart-device market, has tasked me to analyze smart device usage data in order to gain insights into how consumers use non-Bellabeat smart devices.These insights will help guide the marketing strategy for the company.

### Key stakeholders:

Urška Sršen: Bellabeat’s cofounder and Chief Creative Officer  
Sando Mur: Mathematician and Bellabeat’s co-founder; key member of the Bellabeat executive team

### Packages used for the cleaning, analysis and visualization

* tidyverse
* sqldf
* janitor
* dplyr
* ggplot2
* cowplot

library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

library(skimr)  
library(janitor)

##   
## Attaching package: 'janitor'

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

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
  
library(cowplot)

### Prepare the data

The data used in this analysis is the Fitbit Fitness Tracker Data made available by Mobius stored on Kaggle.

ROCCC analysis -

Reliability : LOW – dataset was collected from 30 individuals whose gender is unknown.  
Originality : LOW – third party data collect using Amazon Mechanical Turk.  
Comprehensive : MEDIUM – dataset contains multiple fields on daily activity intensity, calories used, daily steps taken, daily sleep time and weight record.  
Current : MEDIUM – data is 5 years old but the habit of how people live does not change over a few years.  
Cited : HIGH – data collector and source is well documented.

#### Selecting the data

The following files from the dataset has been selected:  
dailyActivity\_merged.csv  
dailyCalories\_merged.csv  
dailyIntensities\_merged.csv  
dailySteps\_merged.csv  
sleepDay\_merged.csv  
weightLogInfo\_merged.csv

dailyActivity <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\Fitabase Data 4.12.16-5.12.16\\dailyActivity\_merged.csv")  
dailyCalories <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\Fitabase Data 4.12.16-5.12.16\\dailyCalories\_merged.csv")  
dailyIntensities <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\Fitabase Data 4.12.16-5.12.16\\dailyIntensities\_merged.csv")  
dailySteps <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\Fitabase Data 4.12.16-5.12.16\\dailySteps\_merged.csv")  
sleepDay <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\New folder\\sleepDay\_merged.csv")  
weightLogInfo <- read.csv("C:\\Users\\USER\\Desktop\\case study2\\New folder\\weightLogInfo\_merged.csv")

#### Data Review

colnames(dailyActivity)

## [1] "Id" "ActivityDate"   
## [3] "TotalSteps" "TotalDistance"   
## [5] "TrackerDistance" "LoggedActivitiesDistance"  
## [7] "VeryActiveDistance" "ModeratelyActiveDistance"  
## [9] "LightActiveDistance" "SedentaryActiveDistance"   
## [11] "VeryActiveMinutes" "FairlyActiveMinutes"   
## [13] "LightlyActiveMinutes" "SedentaryMinutes"   
## [15] "Calories"

glimpse(dailyActivity)

## Rows: 940  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036~  
## $ ActivityDate <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~  
## $ TotalSteps <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019~  
## $ TotalDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~  
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ VeryActiveMinutes <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~  
## $ FairlyActiveMinutes <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~  
## $ LightlyActiveMinutes <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~  
## $ SedentaryMinutes <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~  
## $ Calories <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~

colnames(dailyCalories)

## [1] "Id" "ActivityDay" "Calories"

glimpse(dailyCalories)

## Rows: 940  
## Columns: 3  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366~  
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/~  
## $ Calories <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775~

colnames(dailyIntensities)

## [1] "Id" "ActivityDay"   
## [3] "SedentaryMinutes" "LightlyActiveMinutes"   
## [5] "FairlyActiveMinutes" "VeryActiveMinutes"   
## [7] "SedentaryActiveDistance" "LightActiveDistance"   
## [9] "ModeratelyActiveDistance" "VeryActiveDistance"

glimpse(dailyIntensities)

## Rows: 940  
## Columns: 10  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036~  
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~  
## $ SedentaryMinutes <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~  
## $ LightlyActiveMinutes <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~  
## $ FairlyActiveMinutes <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~  
## $ VeryActiveMinutes <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~

colnames(dailySteps)

## [1] "Id" "ActivityDay" "StepTotal"

glimpse(dailySteps)

## Rows: 940  
## Columns: 3  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366~  
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/~  
## $ StepTotal <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019, 15506, 1054~

colnames(sleepDay)

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

glimpse(sleepDay)

## Rows: 413  
## Columns: 5  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~  
## $ SleepDay <chr> "4/12/2016", "4/13/2016", "4/15/2016", "4/16/2016",~  
## $ TotalSleepRecords <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~  
## $ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~  
## $ TotalTimeInBed <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~

colnames(weightLogInfo)

## [1] "Id" "Date" "WeightKg" "WeightPounds"   
## [5] "Fat" "BMI" "IsManualReport" "LogId"

glimpse(weightLogInfo)

## Rows: 67  
## Columns: 8  
## $ Id <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212~  
## $ Date <chr> "5/2/2016", "5/3/2016", "4/13/2016", "4/21/2016", "5/12~  
## $ WeightKg <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, ~  
## $ WeightPounds <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6~  
## $ Fat <int> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,~  
## $ BMI <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,~  
## $ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~  
## $ LogId <dbl> 1.46223e+12, 1.46232e+12, 1.46051e+12, 1.46128e+12, 1.4~

### Process the Data

Microsoft Excel and R will be used to process the data as the tool functionality fits the purpose.

#### Data Cleaning:

sleepDay\_merged.csv and weightLogInfo\_merged.csv are loaded into Excel for data cleaning.The following steps has been done:  
- The date column has been select and format to ‘Date’ using spreadsheet function  
- Time in the column has been removed as time is irrelevant in this analysis

Data Integrity:  
The following queries have been run to check the number of unique Id in each dataset

length(unique(dailyActivity$Id))

## [1] 33

length(unique(dailyCalories$Id))

## [1] 33

length(unique(dailyIntensities$Id))

## [1] 33

length(unique(dailySteps$Id))

## [1] 33

length(unique(sleepDay$Id))

## [1] 24

length(unique(weightLogInfo$Id))

## [1] 8

The result shows the dataset is inconsistent as we expect 30 unique Id on all tables.

### Data Analysis

In order to find out the relation, following steps has been taken to aggregate the data for analysis:

#### For finding activity level and calories burnt:

dailyAct1 <- dailyActivity %>%  
 filter(VeryActiveDistance+ModeratelyActiveDistance+LightActiveDistance > 0,VeryActiveMinutes+SedentaryMinutes+LightlyActiveMinutes >0 ) %>%  
 select(Id, ActivityDate,TotalSteps,TotalDistance,TrackerDistance,LoggedActivitiesDistance,VeryActiveDistance,ModeratelyActiveDistance,LightActiveDistance,SedentaryActiveDistance,VeryActiveMinutes,FairlyActiveMinutes,LightlyActiveMinutes,SedentaryMinutes,Calories)

totaldata <- dailyActivity %>%   
 mutate(TotalActiveDistance = VeryActiveDistance+ModeratelyActiveDistance+LightActiveDistance,TotalActiveMinutes = VeryActiveMinutes+FairlyActiveMinutes+LightlyActiveMinutes) %>%  
 select(Id, ActivityDate,Calories, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance,SedentaryActiveDistance,SedentaryMinutes)

TotalActiveDistance is the sum of the VeryActiveDistance, ModeratelyActiveDistance and LightActiveDistance that can be useful to find the relation between calories burnt and activity level.

#### For finding relationship between activity level and sleep time:

sleepQuality <- sqldf("select dailyActivity.Id, ActivityDate,Calories, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance, (VeryActiveDistance + ModeratelyActiveDistance) AS ActiveDistance, (LightActiveDistance+SedentaryActiveDistance) AS non\_ActiveDistance, (VeryActiveMinutes+FairlyActiveMinutes) AS ActiveMinutes, (LightlyActiveMinutes+SedentaryMinutes) AS non\_ActiveMinutes from dailyActivity inner join sleepDay on dailyActivity.Id = sleepDay.Id AND dailyActivity.ActivityDate = SleepDay.SleepDay")

ActiveDistance, non-ActiveDistance, ActiveMinutes and non-ActiveMinutes have been calculated to find out the relationship on sleep quality versus a person’s activity in a day.

#### For finding relationship between activiy and weight/BMI:

weightActivityRel <- sqldf("select dailyActivity.Id, Calories, BMI, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance,SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, (VeryActiveMinutes+FairlyActiveMinutes) AS ActiveMinutes, (LightlyActiveMinutes+SedentaryMinutes) AS non\_ActiveMinutes from dailyActivity inner join weightLogInfo on dailyActivity.Id = weightLogInfo.Id AND dailyActivity.ActivityDate = weightLogInfo.Date")

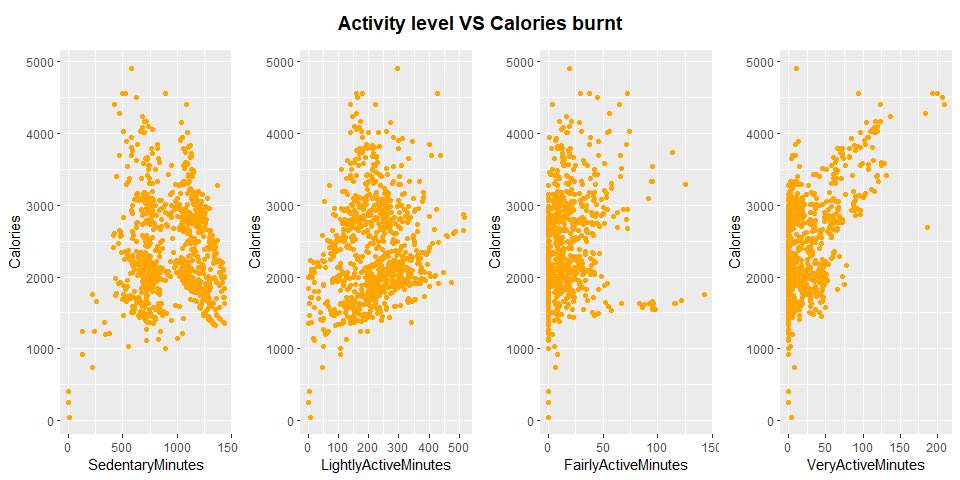
mean\_BMI <- aggregate(BMI ~ Id, data=weightActivityRel, FUN=mean)  
  
mean\_non\_ActMin <- aggregate(non\_ActiveMinutes ~ Id, data=weightActivityRel, FUN=mean)  
  
mean\_ActMin <- aggregate(ActiveMinutes ~ Id, data=weightActivityRel, FUN=mean)  
  
mean\_ini <- sqldf("select mean\_BMI.Id, BMI, non\_ActiveMinutes from mean\_BMI inner join mean\_non\_ActMin on mean\_BMI.ID = mean\_non\_ActMin.ID")  
  
  
mean\_data <- sqldf("select mean\_ini.Id, BMI, non\_ActiveMinutes, ActiveMinutes from mean\_ini inner join mean\_ActMin on mean\_ini.ID = mean\_ActMin.ID")

When comparing the relationship between activity and weight, using BMI is a more consistent metric as weight can also be affected by height. BMI also accounts for height, which is a more universal metric to compare whether a person is underweight or overweight.

### Data visualization

#### 1 - Activity level and calories burnt relation

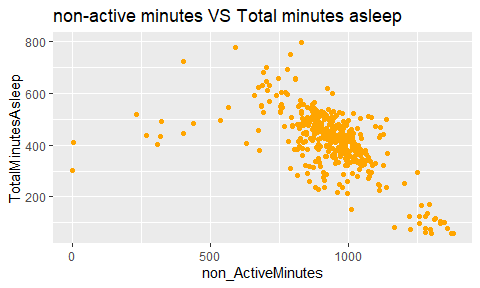
sc <- ggplot(data = dailyAct1) + geom\_point(mapping = aes(x = SedentaryMinutes,y = Calories),color="orange")  
lac <- ggplot(data = dailyAct1) + geom\_point(mapping = aes(x = LightlyActiveMinutes,y = Calories),color="orange")  
fac <- ggplot(data = dailyAct1) + geom\_point(mapping = aes(x = FairlyActiveMinutes,y = Calories),color="orange")  
vac <- ggplot(data = dailyAct1) + geom\_point(mapping = aes(x = VeryActiveMinutes,y = Calories),color="orange")  
  
p <- plot\_grid(sc, lac, fac, vac, ncol = 4, nrow = 1)  
title <- ggdraw() + draw\_label("Activity level VS Calories burnt", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



From the above chart we can see that a person who has higher active minutes tends to burn more calories in a day, the more time they spend inactive, the lower calories they tend to burn in a day.

#### 2 - Activity level and sleep quality relation

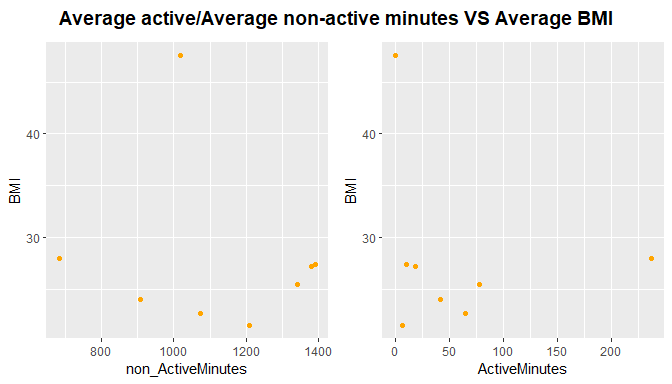
ggplot(data = sleepQuality) +   
 geom\_point(mapping = aes(x = non\_ActiveMinutes,y = TotalMinutesAsleep),color="orange") +  
 labs(title = "non-active minutes VS Total minutes asleep")



This visualisation compares a person’s non-active minutes versus the minutes asleep. We can see that the more time a person spent non-active, the lower the time they are asleep in bed. It is a negative relation which implies non-activity negatively impact sleep quality.

#### 3 - Activity level and BMI relation

anam <- ggplot(data = mean\_data) +   
 geom\_point(mapping = aes(x = non\_ActiveMinutes,y = BMI),color="orange")  
  
aam <- ggplot(data = mean\_data) +   
 geom\_point(mapping = aes(x = ActiveMinutes,y = BMI),color="orange")  
  
p1 <- plot\_grid(anam, aam, ncol = 2, nrow = 1)  
title <- ggdraw() + draw\_label("Average active/Average non-active minutes VS Average BMI", fontface='bold')  
plot\_grid(title, p1, ncol=1, rel\_heights=c(0.1, 1))



The visualisation above compares the average active minutes and average non-active minutes versus the average BMI of the users. It shows a relation that a person who has a higher average non-active minutes tends to have a higher average BMI. The small size sample in this particular comparison hinders the accuracy though.

### Recommendation

#### Conclusion

From the analysis result, it is clear that there is a clear trend in non-active people having a negative lifestyle. The three relations we found during the analysis includes:  
\* Very-active minutes has a positive relation to calories burnt  
\* Active person has a positive relation to sleep quality  
\* Non-active person is more likely to have a high BMI

#### Recommendations to business

We can apply these relations to make data-driven decisions on Bellabeat future products/functionality:  
1. Bellabeat can include function in Bellabeat app to alert user who tends to have a high number to sedentary minutes  
2. Bellabeat can use the relation between high sedentary minutes and BMI to promote an active lifestyle can reduce body fat and create better health with Bellabeat products  
3. Bellabeat can include timely notification in Leaf/Time to motivate user to move around regularly to reduce their sedentary minutes