Health Tracker Analytics

Kevin Ajith

2023-07-29

# Summary

Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company.

# Ask

### Business Task

Sršen asks us to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants us to select one Bellabeat product to apply these insights to in our presentation. We have to find the following:

1. Understanding how customers use smart devices
2. Inform Bellabeat’s marketing strategy for growth
3. Get insights on key features to include in design of Bellabeat app or product to help with growth

### Stakeholder

1. Urška Sršen: Bellabeat’s Co-founder and Chief Creative Officer
2. Sando Mur: Mathematician and Bellabeat’s Co-founder; key member of the Bellabeat Executive team
3. Bellabeat marketing analytics team

# Prepare

### Overview of the Data

* Data is open-sourced
* Data is taken from [Zendo](https://zenodo.org/record/53894#.YMoUpnVKiP9)
* Data was uploaded by MORIUS
* Data was downloaded from [Kaggle](https://www.kaggle.com/datasets/arashnic/fitbit)
* There are 18 CSV files containing data
* The data has 33 unique Ids.
* Data contains fields like Intensities, Calories, Steps Taken, Distance Traveled and Sleep.
* Data has been measured daily, hourly as well as in minutes.

### Quality of Data

* The data is inconsistent since the data was last updated in 2016.
* The validation of data becomes difficult because of the above reason.
* The data did not have duplicate or null values.
* The data does not have identifiers for users as to protect the user information.
* The User Ids are inconsistent in some CSV files since it is much less than 30 user Ids.

# Process

### Filtering CSV

There is a lot of data and according to our requirement we need to have 30 ids for analysis. So we can remove data which do not meet the criteria. We can also remove data which are redundant or not required for analysis.

The Table for the CSV are as follows:

| Table Name | Number of Unique Ids | Is Removed | Reason for Removal |
| --- | --- | --- | --- |
| dailyActivity\_merged | 33 | F |  |
| dailyCalories\_merged | 33 | T | Redundant Data |
| dailyIntensities\_merged | 33 | T | Redundant Data |
| dailySteps\_merged | 33 | T | Redundant Data |
| heartrate\_seconds\_merged | 7 | T | Insufficient data |
| hourlyCalories\_merged | 33 | F |  |
| hourlyIntensities\_merged | 33 | F |  |
| hourlySteps\_merged | 33 | F |  |
| minuteCaloriesNarrow\_merged | 27 | T | Not Required |
| minuteCaloriesWide\_merged | 33 | T | Not Required |
| minuteIntensitiesNarrow\_merged | 27 | T | Not Required |
| minuteIntensitiesWide\_merged | 33 | T | Not Required |
| minutesMETsNarrow\_merged | 27 | T | Not Required |
| minuteSleep\_merged | 24 | T | Insufficient data |
| minuteStepsNarrow\_merged | 27 | T | Not Required |
| minuteStepsWide\_merged | 33 | T | Not Required |
| sleepDay\_merged | 28 | F |  |
| weightLogInfo | 8 | T | Insufficient data |

After excluding the datasets which are not required, we are left with the following:

1. dailyActivity\_merged
2. sleepDay\_merged
3. hourlyCalories\_merged
4. hourlyIntensities\_merged
5. hourlySteps\_merged

### Load The Libraries

The Libraries required for the project are:

1. tidyverse
2. dplyr

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

### Loading Dataset

dailyActivity <- read.csv("C:/Users/kevin/Desktop/DA/Case Study DA/2/data/Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv")  
sleepDay <- read.csv("C:/Users/kevin/Desktop/DA/Case Study DA/2/data/Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv")  
hourlyCalories <- read.csv("C:/Users/kevin/Desktop/DA/Case Study DA/2/data/Fitabase Data 4.12.16-5.12.16/hourlyCalories\_merged.csv")  
hourlyIntensities <- read.csv("C:/Users/kevin/Desktop/DA/Case Study DA/2/data/Fitabase Data 4.12.16-5.12.16/hourlyIntensities\_merged.csv")  
hourlySteps <- read.csv("C:/Users/kevin/Desktop/DA/Case Study DA/2/data/Fitabase Data 4.12.16-5.12.16/hourlySteps\_merged.csv")

# Analyse

In the Analyse Stage, We will be organize and format the data as well as analyse it.

### Merge Hourly Activity Into hourlyActivity

The hourly activity is required because having different hourly data is redundant and becomes difficult to manage when it comes to analysis and it becomes easy to compare different fields in hourly activity.

hourlyActivity <- merge(hourlyCalories, hourlyIntensities, by = c("Id","ActivityHour"))  
hourlyActivity <- merge(hourlyActivity, hourlySteps, by = c("Id","ActivityHour"))  
rm(hourlySteps,hourlyCalories,hourlyIntensities)

### Displaying Data

Before Analysis, Lets display the data in each of the data frame.

#Head  
head(dailyActivity)

## Id ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 1503960366 4/12/2016 13162 8.50 8.50  
## 2 1503960366 4/13/2016 10735 6.97 6.97  
## 3 1503960366 4/14/2016 10460 6.74 6.74  
## 4 1503960366 4/15/2016 9762 6.28 6.28  
## 5 1503960366 4/16/2016 12669 8.16 8.16  
## 6 1503960366 4/17/2016 9705 6.48 6.48  
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance  
## 1 0 1.88 0.55  
## 2 0 1.57 0.69  
## 3 0 2.44 0.40  
## 4 0 2.14 1.26  
## 5 0 2.71 0.41  
## 6 0 3.19 0.78  
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes  
## 1 6.06 0 25  
## 2 4.71 0 21  
## 3 3.91 0 30  
## 4 2.83 0 29  
## 5 5.04 0 36  
## 6 2.51 0 38  
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories  
## 1 13 328 728 1985  
## 2 19 217 776 1797  
## 3 11 181 1218 1776  
## 4 34 209 726 1745  
## 5 10 221 773 1863  
## 6 20 164 539 1728

head(hourlyActivity)

## Id ActivityHour Calories TotalIntensity AverageIntensity  
## 1 1503960366 4/12/2016 1:00:00 AM 61 8 0.133333  
## 2 1503960366 4/12/2016 1:00:00 PM 66 6 0.100000  
## 3 1503960366 4/12/2016 10:00:00 AM 99 29 0.483333  
## 4 1503960366 4/12/2016 10:00:00 PM 65 9 0.150000  
## 5 1503960366 4/12/2016 11:00:00 AM 76 12 0.200000  
## 6 1503960366 4/12/2016 11:00:00 PM 81 21 0.350000  
## StepTotal  
## 1 160  
## 2 221  
## 3 676  
## 4 89  
## 5 360  
## 6 338

head(sleepDay)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep  
## 1 1503960366 4/12/2016 12:00:00 AM 1 327  
## 2 1503960366 4/13/2016 12:00:00 AM 2 384  
## 3 1503960366 4/15/2016 12:00:00 AM 1 412  
## 4 1503960366 4/16/2016 12:00:00 AM 2 340  
## 5 1503960366 4/17/2016 12:00:00 AM 1 700  
## 6 1503960366 4/19/2016 12:00:00 AM 1 304  
## TotalTimeInBed  
## 1 346  
## 2 407  
## 3 442  
## 4 367  
## 5 712  
## 6 320

The data is loaded, the hourly activities are compiled into one and the datasets are displayed for the same.

### Data Transformation

The Id and date types are inconsistent since Id data type is number but we require it to be categorical data and date type is also inconsistent. Hence, we will streamline the data and then perform other analysis.

#Change ID to category type  
dailyActivity$Id <- as.character(dailyActivity$Id)  
hourlyActivity$Id <- as.character(hourlyActivity$Id)  
sleepDay$Id <- as.character(sleepDay$Id)  
  
#Date and Time Transformation  
dailyActivity$ActivityDate <- as.POSIXct(dailyActivity$ActivityDate, format = "%m/%d/%Y")  
  
hourlyActivity$ActivityHour <- as.POSIXct(hourlyActivity$ActivityHour, format = "%m/%d/%Y %I:%M:%S %p", tz = Sys.timezone())  
hourlyActivity$date <- format(hourlyActivity$ActivityHour, format = "%m/%d/%y")  
hourlyActivity$time <- format(hourlyActivity$ActivityHour, format = "%H:%M:%S")  
  
sleepDay$SleepDay <- as.POSIXct(sleepDay$SleepDay, format = "%m/%d/%Y %I:%M:%S %p", tz = Sys.timezone())

### Summaries

After the data transformation, we found that there are no duplicates as well as missing data in the chosen dataset. The summairs for their filds are as follows:

#Summary  
summary(dailyActivity)

## Id ActivityDate TotalSteps   
## Length:940 Min. :2016-04-12 00:00:00.00 Min. : 0   
## Class :character 1st Qu.:2016-04-19 00:00:00.00 1st Qu.: 3790   
## Mode :character Median :2016-04-26 00:00:00.00 Median : 7406   
## Mean :2016-04-26 06:53:37.01 Mean : 7638   
## 3rd Qu.:2016-05-04 00:00:00.00 3rd Qu.:10727   
## Max. :2016-05-12 00:00:00.00 Max. :36019   
## TotalDistance TrackerDistance LoggedActivitiesDistance VeryActiveDistance  
## Min. : 0.000 Min. : 0.000 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 2.620 1st Qu.: 2.620 1st Qu.:0.0000 1st Qu.: 0.000   
## Median : 5.245 Median : 5.245 Median :0.0000 Median : 0.210   
## Mean : 5.490 Mean : 5.475 Mean :0.1082 Mean : 1.503   
## 3rd Qu.: 7.713 3rd Qu.: 7.710 3rd Qu.:0.0000 3rd Qu.: 2.053   
## Max. :28.030 Max. :28.030 Max. :4.9421 Max. :21.920   
## ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance  
## Min. :0.0000 Min. : 0.000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.: 1.945 1st Qu.:0.000000   
## Median :0.2400 Median : 3.365 Median :0.000000   
## Mean :0.5675 Mean : 3.341 Mean :0.001606   
## 3rd Qu.:0.8000 3rd Qu.: 4.782 3rd Qu.:0.000000   
## Max. :6.4800 Max. :10.710 Max. :0.110000   
## VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes  
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:127.0 1st Qu.: 729.8   
## Median : 4.00 Median : 6.00 Median :199.0 Median :1057.5   
## Mean : 21.16 Mean : 13.56 Mean :192.8 Mean : 991.2   
## 3rd Qu.: 32.00 3rd Qu.: 19.00 3rd Qu.:264.0 3rd Qu.:1229.5   
## Max. :210.00 Max. :143.00 Max. :518.0 Max. :1440.0   
## Calories   
## Min. : 0   
## 1st Qu.:1828   
## Median :2134   
## Mean :2304   
## 3rd Qu.:2793   
## Max. :4900

summary(hourlyActivity)

## Id ActivityHour Calories   
## Length:22099 Min. :2016-04-12 00:00:00.00 Min. : 42.00   
## Class :character 1st Qu.:2016-04-19 01:00:00.00 1st Qu.: 63.00   
## Mode :character Median :2016-04-26 06:00:00.00 Median : 83.00   
## Mean :2016-04-26 11:46:42.58 Mean : 97.39   
## 3rd Qu.:2016-05-03 19:00:00.00 3rd Qu.:108.00   
## Max. :2016-05-12 15:00:00.00 Max. :948.00   
## TotalIntensity AverageIntensity StepTotal date   
## Min. : 0.00 Min. :0.0000 Min. : 0.0 Length:22099   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.: 0.0 Class :character   
## Median : 3.00 Median :0.0500 Median : 40.0 Mode :character   
## Mean : 12.04 Mean :0.2006 Mean : 320.2   
## 3rd Qu.: 16.00 3rd Qu.:0.2667 3rd Qu.: 357.0   
## Max. :180.00 Max. :3.0000 Max. :10554.0   
## time   
## Length:22099   
## Class :character   
## Mode :character   
##   
##   
##

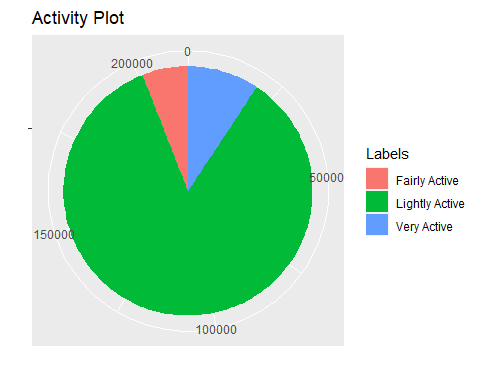
summary(sleepDay)

## Id SleepDay TotalSleepRecords  
## Length:413 Min. :2016-04-12 00:00:00.00 Min. :1.000   
## Class :character 1st Qu.:2016-04-19 00:00:00.00 1st Qu.:1.000   
## Mode :character Median :2016-04-27 00:00:00.00 Median :1.000   
## Mean :2016-04-26 12:40:05.80 Mean :1.119   
## 3rd Qu.:2016-05-04 00:00:00.00 3rd Qu.:1.000   
## Max. :2016-05-12 00:00:00.00 Max. :3.000   
## TotalMinutesAsleep TotalTimeInBed   
## Min. : 58.0 Min. : 61.0   
## 1st Qu.:361.0 1st Qu.:403.0   
## Median :433.0 Median :463.0   
## Mean :419.5 Mean :458.6   
## 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :796.0 Max. :961.0

The data was found to be clean as well as consistent data. The data type was converted to the required type for analysis.

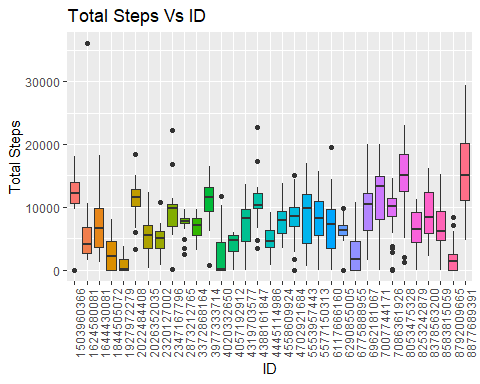
### Data Visualization

VeryActive <- sum(dailyActivity$VeryActiveMinutes)  
FairlyActive <- sum(dailyActivity$FairlyActiveMinutes)  
LightlyActive <- sum(dailyActivity$LightlyActiveMinutes)  
  
  
Activity\_df = data.frame(  
 "Activity\_SUM" = c(VeryActive,FairlyActive,LightlyActive),  
 "Labels" = c("Very Active", "Fairly Active", "Lightly Active")  
 )  
  
ggplot(Activity\_df, aes(x = "", y = Activity\_SUM, fill = Labels)) +   
 geom\_bar(stat="identity", width=1) +   
 coord\_polar(theta = "y") +  
 labs(x = "", y = "", title = "Activity Plot")



From the chart, we could infer that the users mostly are lightly activity.

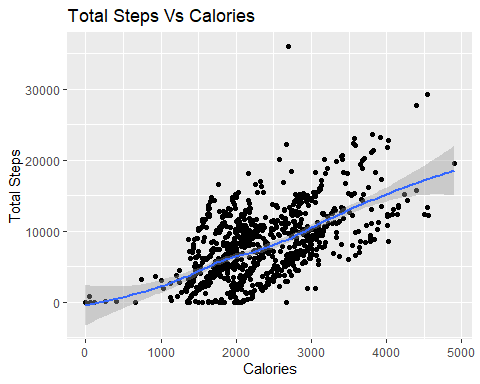
ggplot(dailyActivity) +   
 geom\_boxplot(aes(x = Id, y = TotalSteps, fill = Id)) +   
 labs( x = "ID", y = "Total Steps", title = "Total Steps Vs ID") +  
 theme(axis.text.x = element\_text(angle = 90), legend.position = "none")



From the above plot we can infer that, most of the users takes less than 10,000 steps on an average daily.

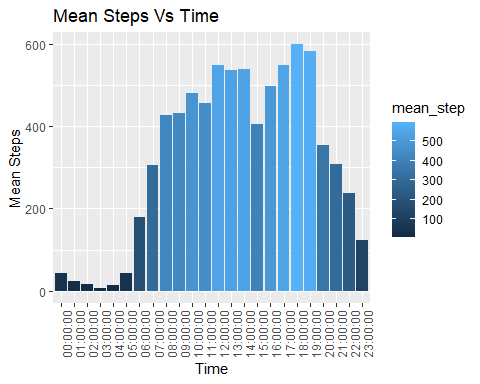
ggplot(dailyActivity, aes(x = Calories, y = TotalSteps)) +   
 geom\_point() +   
 labs( x = "Calories", y = "Total Steps", title = "Total Steps Vs Calories") +  
 geom\_smooth()

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



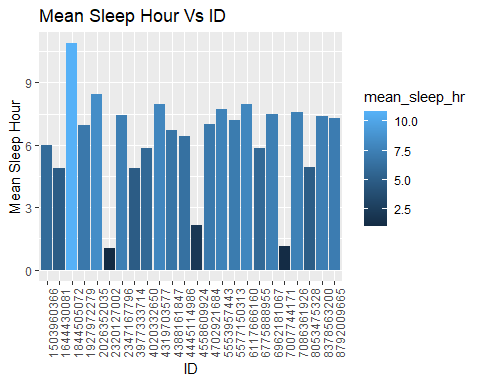
The above plot tells us that the TotalSteps has a positive correlation with Calories, i.e, The more steps you take more calories your burn, even though it may vary from user to user because of their metabolism.

ggplot(hourlyActivity %>% group\_by(time) %>% summarise(mean\_step=mean(StepTotal))) +   
 geom\_bar(aes(x = time, y = mean\_step, fill = mean\_step),stat = "identity") +   
 labs( x = "Time", y = "Mean Steps", title = "Mean Steps Vs Time") +  
 theme(axis.text.x = element\_text(angle = 90))



The average step count highest between 4 PM to 7 PM.

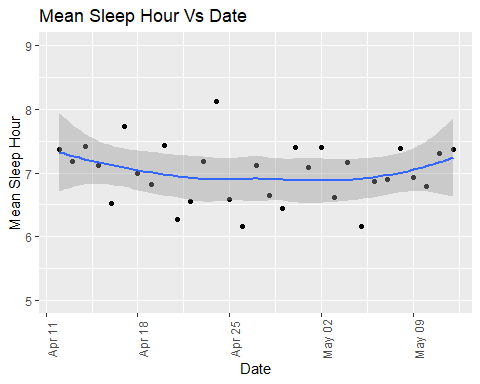
ggplot(sleepDay %>% group\_by(Id) %>% summarise(mean\_sleep\_hr = mean(TotalMinutesAsleep)/60)) +   
 geom\_bar(aes(x = Id, y = mean\_sleep\_hr, fill = mean\_sleep\_hr),stat = "identity") +   
 labs( x = "ID", y = "Mean Sleep Hour", title = "Mean Sleep Hour Vs ID") +  
 theme(axis.text.x = element\_text(angle = 90))



According to the National Sleep Foundation, on an average an adult human requires sleep of 7 to 8 hours. Though the sample for sleep is available for 24 users only, It can be said that on average they sleep for 6.98 hours, from which we can say that there are users who do not meet the required standard when it comes to sleep.

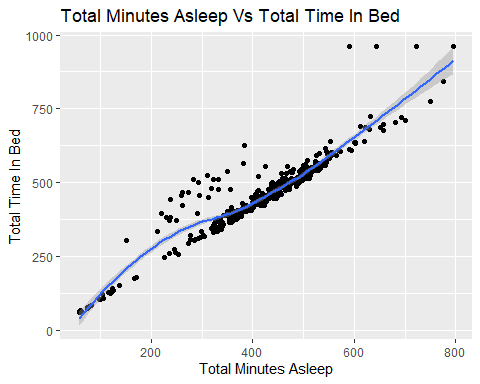
ggplot(sleepDay %>% group\_by(SleepDay) %>% summarise(mean\_sleep\_hr = mean(TotalMinutesAsleep)/60),  
 aes(x = SleepDay, y = mean\_sleep\_hr)) +   
 geom\_point() +   
 geom\_smooth() +  
 labs( x = "Date", y = "Mean Sleep Hour", title = "Mean Sleep Hour Vs Date") +  
 theme(axis.text.x = element\_text(angle = 90)) +  
 ylim(5,9)

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



ggplot(sleepDay, aes(x = TotalMinutesAsleep, y = TotalTimeInBed)) +   
 labs( x = "Total Minutes Asleep", y = "Total Time In Bed",   
 title = "Total Minutes Asleep Vs Total Time In Bed") +  
 geom\_point() + geom\_smooth()

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



The data specifies that there is a correlation between sleep duration as well as time in bed, but in some cases people spend a lot time in bed before sleeping. Though it can be various number of reasons, we can assume that some of them may find it difficult to sleep.