Bellabeat Analysis - Google Capstone Project

Yuri Faria Frota

31/10/2021

# Bellabeat

Bellabeat is a a high-tech manufacturer of health-focused products for women, founded by Urška Sršen and Sando Mur. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women. Sršen knows that an analysis of Bellabeat’s available consumer data would reveal more opportunities for growth. She has asked the marketing analytics team to focus on a Bellabeat product and analyze smart device usage data in order to gain insight into how people are already using their smart devices. Using this information, she would like high-level recommendations for how these trends can inform Bellabeat marketing strategy.

Link to Webpage for more information <https://bellabeat.com/>

## PHASE 1 – ASK

By the data collected from smart devices (as calories burn, time of activities, daily data about sleep quality, among others), we gonna be able to get some insights about how to improve Bellabeat devices, and better meet the customers demands. Desta forma, os clientes Bellabeat terão mais conhecimento sobre seus hábitos, seu corpo, e sobre como promover uma vida saudável, proporcionando melhor qualidade de vida. Doing that, Bellabeat customers will have more knowledge about their habits, your body, and about how to promote a healthy life, providing better life quality. The analysis from this dataset can help bellabeat to stand out, being able to offer demands never offered before.

The main objetives of this analysis, are to offer: \* Bring new insights from other smart devices data analysis \* Pattern and tendencies coming from data. \* Proposals for improvements in Bellabeat devices. \* Possibilidades de novas demandas, a partir da análise dos dados \* Possibilities of new demands, arrising through data analysis \* Valuable informations for Bellabeat marketing, and strategy department.

We gonna use **Leaf**, from the website catalog, as a reference for this study.

## PHASE 2 – PREPARE

This phase consists on the beginning of the analysis process. We gonna check which kind of data is available, your formats, credibility, and also check with kind of information we could get from it. For knowing about the quality of the data set used, we gonna utilize ROCCC methodology, on which focus on five basic criteria to ensure quality. For this, the data set must have:

#### Reliability:

The small sample set (around 30 users), and some statistical variables missing, restrict the possibility for further analysis.

#### Originality:

Not from original source, it is a third party data, collected from Amazon Mechanical Turk.

#### Comprehensiveness:

The data doesn’t give sex, age, and other specific conditions about the users, on which could result in more accurated results. Do not consider those elements can lead to biased results.

#### Current:

The data was collected in 2016. Even though it’s an outdated dataset, people habits doesn’t tend to change quickly in few years.

#### Cited:

Data collected from Amazon Mechanical Murk, and well documented.

Through the observations made, we can consider a lack of data quantity and quality, making more accurate analysis difficult.

Saying that, we gonna get some insights from this analysis, and turn out our findings into a starting point for next analysis.

## Setting up the environment.

Installing and loading the packages we gonna use.

library(readr)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v dplyr 1.0.7  
## v tibble 3.1.5 v stringr 1.4.0  
## v tidyr 1.1.4 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(sqldf)

## Carregando pacotes exigidos: gsubfn

## Carregando pacotes exigidos: proto

## Carregando pacotes exigidos: RSQLite

library(ggplot2)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Importing data

sleepDay\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/sleepDay\_merged.csv")  
weightLogInfo\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/weightLogInfo\_merged.csv")  
minuteSleep\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteSleep\_merged.csv")  
minuteStepsNarrow\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteStepsNarrow\_merged.csv")  
minuteIntensitiesWide\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteIntensitiesWide\_merged.csv")  
minuteMETsNarrow\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteMETsNarrow\_merged.csv")  
minuteCaloriesWide\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteCaloriesWide\_merged.csv")  
minuteIntensitiesNarrow\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteIntensitiesNarrow\_merged.csv")  
hourlyCalories\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/hourlyCalories\_merged.csv")  
hourlyIntensities\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/hourlyIntensities\_merged.csv")  
hourlySteps\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/hourlySteps\_merged.csv")  
minuteCaloriesNarrow\_merged <- read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/minuteCaloriesNarrow\_merged.csv")  
dailyactivity\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/dailyActivity\_merged.csv")  
dailyCalories\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/dailyCalories\_merged.csv")  
dailyIntensities\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/dailyIntensities\_merged.csv")  
dailySteps\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/dailySteps\_merged.csv")  
heartrate\_seconds\_merged <-read\_csv("C:/Users/Yuri/Documents/capstone project/fitbit\_fitnesstrack\_data/heartrate\_seconds\_merged.csv")

## Checking data

We gonna use the dataframe ‘dailyactivity\_merged’ to show a major inspection of the general data.

colnames(dailyactivity\_merged)

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

head(dailyactivity\_merged)

## # A tibble: 6 x 15  
## Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/2016 13162 8.5 8.5 0  
## 2 1.50e9 4/13/2016 10735 6.97 6.97 0  
## 3 1.50e9 4/14/2016 10460 6.74 6.74 0  
## 4 1.50e9 4/15/2016 9762 6.28 6.28 0  
## 5 1.50e9 4/16/2016 12669 8.16 8.16 0  
## 6 1.50e9 4/17/2016 9705 6.48 6.48 0  
## # ... with 9 more variables: VeryActiveDistance <dbl>,  
## # ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,  
## # SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,  
## # FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,  
## # SedentaryMinutes <dbl>, Calories <dbl>

nrow(dailyactivity\_merged)

## [1] 940

glimpse(dailyactivity\_merged)

## Rows: 940  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036~  
## $ ActivityDate <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~  
## $ TotalSteps <dbl> 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 <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~  
## $ FairlyActiveMinutes <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~  
## $ LightlyActiveMinutes <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~  
## $ SedentaryMinutes <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~  
## $ Calories <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~

**Using same functions for all the sets below**

*(sleepDay\_merged)*

*(dailyIntensities\_merged)*

*(dailyCalories\_merged)*

*(weightLogInfo\_merged)*

*(dailySteps\_merged)*

## Verifying data integrity

**Let’s verify how many diferente people are in each data set.** (Separating the distinct Id, and counting the resulting lines).

teste1 <- sqldf ('SELECT DISTINCT Id FROM "dailyCalories\_merged"')  
nrow(teste1)

## [1] 33

teste2 <- sqldf ('SELECT DISTINCT Id FROM "dailyactivity\_merged"')  
nrow(teste2)

## [1] 33

teste3 <- sqldf ('SELECT DISTINCT Id FROM "dailyIntensities\_merged"')  
nrow(teste3)

## [1] 33

teste4 <- sqldf ('SELECT DISTINCT Id FROM "dailySteps\_merged"')  
nrow(teste4)

## [1] 33

teste5 <- sqldf ('SELECT DISTINCT Id FROM "sleepDay\_merged"')  
nrow(teste5)

## [1] 24

teste6 <- sqldf ('SELECT DISTINCT Id FROM "weightLogInfo\_merged"')  
nrow(teste6)

## [1] 8

**Verifying how many days of data collected**

n\_distinct(dailyactivity\_merged$ActivityDate)

## [1] 31

**During this check, we could observe that both tables below had fewer registers than the other data frames.**

*(weightLogInfo\_merged)*

nrow(weightLogInfo\_merged)

## [1] 67

n\_distinct(weightLogInfo\_merged$Id)

## [1] 8

*(dailySteps\_merged)*

nrow(sleepDay\_merged)

## [1] 413

n\_distinct(sleepDay\_merged$Id)

## [1] 24

**Both sets above are incomplete, as they don’t include all participants, and also have less registers. So, we won’t include those data frames into this analysis.** **We gonna proceed utilizing the daily tables, and verify if the columns with same name, has equal data between them.** **For this, we gonna use the following instructions:**

**‘dailyActivity\_merged’ ∩ ‘dailyCalories\_merged’**

daily\_activity1 <- dailyactivity\_merged %>% select(Id, ActivityDate, Calories)  
check1 <- sqldf ('SELECT \* FROM daily\_activity1 INTERSECT SELECT \* FROM dailyCalories\_merged')  
nrow(check1)

## [1] 940

n\_distinct(check1$Id)

## [1] 33

**‘dailyActivity\_merged’ ∩ ‘dailyIntensities\_merged’**

daily\_activity2 <- dailyactivity\_merged %>% select(Id, ActivityDate, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryActiveMinutes, SedentaryActiveDistance, LightActiveDistance, ModeratelyActiveDistance, VeryActiveDistance)  
check2 <- sqldf('SELECT \* FROM daily\_activity2 INTERSECT SELECT \* FROM dailyIntensities\_merged')  
nrow(check2)

## [1] 940

n\_distinct(check2$Id)

## [1] 33

**‘dailyActivity\_merged’ ∩ ‘dailySteps\_merged’**

daily\_activity3 <- dailyactivity\_merged %>% select(Id, ActivityDate, TotalSteps)  
check3 <- sqldf('SELECT \* FROM daily\_activity3 INTERSECT SELECT \* FROM dailySteps\_merged')  
nrow(check3)

## [1] 940

n\_distinct(check3$Id)

## [1] 33

\*\* The 940 rows between each of the intersections (check1, check2, e check3) confirm the hipothesis that all columns with same name between tables have the same data. Therefore we can use only the ‘dailyActivity\_merged’ as a summary from all the data from other ‘daily’ tables. \*\*

## PHASE 3 AND 4 – PROCESS AND ANALYZE

**Let’s now create a summary of some general characteristics from daily habits to start our analysis**

dailyactivity\_merged %>% summary()

## Id ActivityDate TotalSteps TotalDistance   
## Min. :1.504e+09 Length:940 Min. : 0 Min. : 0.000   
## 1st Qu.:2.320e+09 Class :character 1st Qu.: 3790 1st Qu.: 2.620   
## Median :4.445e+09 Mode :character Median : 7406 Median : 5.245   
## Mean :4.855e+09 Mean : 7638 Mean : 5.490   
## 3rd Qu.:6.962e+09 3rd Qu.:10727 3rd Qu.: 7.713   
## Max. :8.878e+09 Max. :36019 Max. :28.030   
## TrackerDistance LoggedActivitiesDistance VeryActiveDistance  
## Min. : 0.000 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 2.620 1st Qu.:0.0000 1st Qu.: 0.000   
## Median : 5.245 Median :0.0000 Median : 0.210   
## Mean : 5.475 Mean :0.1082 Mean : 1.503   
## 3rd Qu.: 7.710 3rd Qu.:0.0000 3rd Qu.: 2.053   
## 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

sleepDay\_merged %>% summary ()

## Id SleepDay TotalSleepRecords TotalMinutesAsleep  
## Min. :1.504e+09 Length:413 Min. :1.000 Min. : 58.0   
## 1st Qu.:3.977e+09 Class :character 1st Qu.:1.000 1st Qu.:361.0   
## Median :4.703e+09 Mode :character Median :1.000 Median :433.0   
## Mean :5.001e+09 Mean :1.119 Mean :419.5   
## 3rd Qu.:6.962e+09 3rd Qu.:1.000 3rd Qu.:490.0   
## Max. :8.792e+09 Max. :3.000 Max. :796.0   
## TotalTimeInBed   
## Min. : 61.0   
## 1st Qu.:403.0   
## Median :463.0   
## Mean :458.6   
## 3rd Qu.:526.0   
## Max. :961.0

**By the summany, we could reach some mean values:** Através do resumo, podemos obter a média diária de alguns dados:\*\*

* Daily Steps : 7638
* Distance: 5.490m / 5.49 km
* Calories :2304 kcal
* Sedentary minutes : 991.2 / 16.52 hours
* Active minutes : 227.52 / 3.8 hours
* Sleep minutes : 419.5 / 7 hours

**With this quick summary, we are able to predict some users general behaviors and characteristics:**

**In general, they are sedentary (over 16 hours/day doing ‘sedentary activities’)** \*\*They sleep little (Under 8 hours of sleep recommendation)

**Let’s see some plots, to reach some valuable insights:**

daily\_activity4 <- sqldf ('SELECT Id, ActivityDate,Calories, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance, (VeryActiveDistance+ModeratelyActiveDistance+LightActiveDistance) AS ActiveDistanceTotal,SedentaryActiveDistance, (VeryActiveMinutes+FairlyActiveMinutes+LightlyActiveMinutes) AS ActiveMinutesTotal, SedentaryMinutes FROM "dailyactivity\_merged"')  
  
colnames(daily\_activity4)

## [1] "Id" "ActivityDate"   
## [3] "Calories" "TotalSteps"   
## [5] "TotalDistance" "TrackerDistance"   
## [7] "LoggedActivitiesDistance" "ActiveDistanceTotal"   
## [9] "SedentaryActiveDistance" "ActiveMinutesTotal"   
## [11] "SedentaryMinutes"

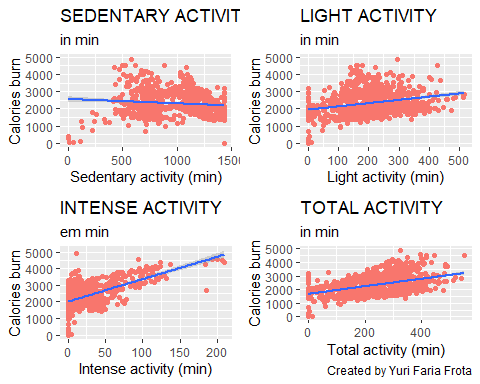
nrow(daily\_activity4)

## [1] 940

distanciamaximaplot<- ggplot(data=dailyactivity\_merged) +  
 geom\_point(mapping=aes(x=TotalDistance, y= Calories, color = 'Orange'))+  
 labs (title = "TOTAL DISTANCE", subtitle = "in km ",   
 x = "Total distance (km)", y = "Calories burn")+   
 theme(legend.position = "none") +  
 geom\_smooth(method=lm,mapping = aes (x=TotalDistance, y= Calories ))  
  
  
tempoleveplot <- ggplot(data=dailyactivity\_merged) +  
 geom\_point(mapping=aes(x=LightlyActiveMinutes, y= Calories, color = 'Orange'))+  
 labs (title = "LIGHT ACTIVITY", subtitle = "in min ", x = "Light activity (min)", y = "Calories burn") +  
 theme(legend.position = "none") +  
 geom\_smooth(method=lm, mapping = aes (x=LightlyActiveMinutes, y= Calories ))  
  
  
tempointensoplot <-ggplot(data=dailyactivity\_merged) +  
 geom\_point(mapping=aes(x=VeryActiveMinutes, y= Calories, color = 'Orange'))+  
 labs (title = "INTENSE ACTIVITY", subtitle = "em min ", x = "Intense activity (min)", y = "Calories burn") +   
 theme(legend.position = "none") +  
 geom\_smooth(method=lm,mapping = aes (x=VeryActiveMinutes, y= Calories ))  
  
  
tempoesedentarioplot<-ggplot(data=dailyactivity\_merged) +  
 geom\_point(mapping=aes(x=SedentaryMinutes, y= Calories, color = 'Orange'))+  
 labs (title = "SEDENTARY ACTIVITY", subtitle = "in min ", x = "Sedentary activity (min)", y = "Calories burn") +  
 theme(legend.position = "none") +  
 geom\_smooth(method=lm, mapping = aes (x=SedentaryMinutes, y= Calories ))  
  
tempototalplot<-ggplot(data=daily\_activity4) +  
 geom\_point(mapping=aes(x=ActiveMinutesTotal, y= Calories, color = 'Orange'))+  
 labs (title = "TOTAL ACTIVITY", subtitle = "in min ",   
 caption= "Created by Yuri Faria Frota", x = "Total activity (min)", y = "Calories burn") +  
 theme(legend.position = "none") +  
 geom\_smooth(method=lm, mapping = aes (x=ActiveMinutesTotal, y= Calories ))

grid.arrange(tempoesedentarioplot, tempoleveplot, tempointensoplot, tempototalplot, nrow = 2)

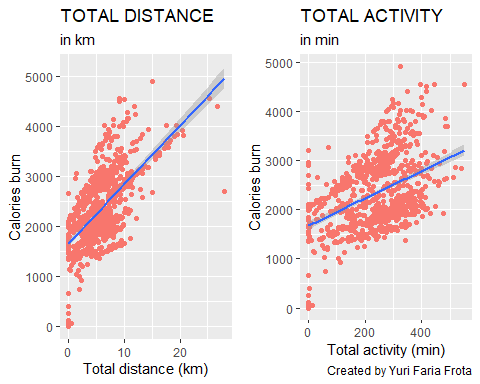
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



**Note the difference in calories burn among different types of activity. It’s clear the importance of keep doing activities to reach an efficient calorie burn rate.**

grid.arrange(distanciamaximaplot, tempototalplot, nrow = 1)

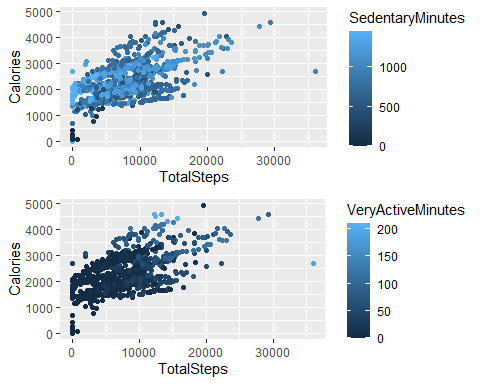
## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'



**Long walks, under light or fair activity, are also efficient to calories burn.**

sedentario <- ggplot(data=dailyactivity\_merged) + geom\_point(mapping = aes (x=TotalSteps, y=Calories, color = SedentaryMinutes))  
  
leve <- ggplot(data=dailyactivity\_merged) + geom\_point(mapping = aes (x=TotalSteps, y=Calories, color = LightlyActiveMinutes))  
  
moderado <- ggplot(data=dailyactivity\_merged) + geom\_point(mapping = aes (x=TotalSteps, y=Calories, color = FairlyActiveMinutes))  
  
intenso <- ggplot(data=dailyactivity\_merged) + geom\_point(mapping = aes (x=TotalSteps, y=Calories, color = VeryActiveMinutes))

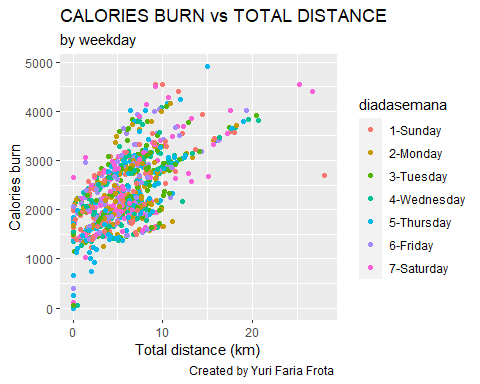
grid.arrange(sedentario, intenso, nrow = 2)



**The users longest time in activity is sedentary. Note that more activity leads to greater caloric consumption**

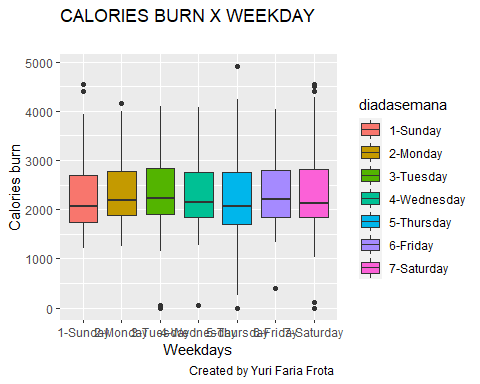
semanasono <- sqldf ('SELECT Id, SleepDay , TotalMinutesAsleep, CASE WHEN   
 SleepDay = "4/18/2016 12:00:00 AM" OR SleepDay = "4/25/2016 12:00:00 AM" OR SleepDay = "5/2/2016 12:00:00 AM" OR SleepDay = "5/9/2016 12:00:00 AM" THEN "2-Monday"  
 WHEN SleepDay = "4/12/2016 12:00:00 AM" OR SleepDay = "4/19/2016 12:00:00 AM" OR SleepDay = "4/26/2016 12:00:00 AM" OR SleepDay = "5/3/2016 12:00:00 AM" OR SleepDay = "5/10/2016 12:00:00 AM" THEN "3-Tuesday"   
 WHEN SleepDay = "4/13/2016 12:00:00 AM" OR SleepDay = "4/20/2016 12:00:00 AM" OR SleepDay = "4/27/2016 12:00:00 AM" OR SleepDay = "5/4/2016 12:00:00 AM" OR SleepDay = "5/11/2016 12:00:00 AM" THEN "4-Wednesday"   
 WHEN SleepDay = "4/14/2016 12:00:00 AM" OR SleepDay = "4/21/2016 12:00:00 AM" OR SleepDay = "4/28/2016 12:00:00 AM" OR SleepDay = "5/5/2016 12:00:00 AM" OR SleepDay = "5/12/2016 12:00:00 AM" THEN "5-Thursday"   
 WHEN SleepDay = "4/15/2016 12:00:00 AM" OR SleepDay = "4/22/2016 12:00:00 AM" OR SleepDay = "4/29/2016 12:00:00 AM" OR SleepDay = "5/6/2016 12:00:00 AM" THEN "6-Friday"   
 WHEN SleepDay = "4/16/2016 12:00:00 AM" OR SleepDay = "4/23/2016 12:00:00 AM" OR SleepDay = "4/30/2016 12:00:00 AM" OR SleepDay = "5/7/2016 12:00:00 AM" THEN "7-sabado"   
 WHEN SleepDay = "4/17/2016 12:00:00 AM" OR SleepDay = "4/24/2016 12:00:00 AM" OR SleepDay = "5/1/2016 12:00:00 AM" OR SleepDay = "5/8/2016 12:00:00 AM" THEN "1-Sunday"   
 ELSE 0 end as diadasemana FROM "sleepDay\_merged" ')

ggplot(data=semana1) +  
 geom\_point(mapping=aes(x=TotalDistance, y= Calories, color = diadasemana ))+  
 labs (title = "CALORIES BURN vs TOTAL DISTANCE ", subtitle = "by weekday",   
 caption= " Created by Yuri Faria Frota ", x = "Total distance (km)", y = "Calories burn")

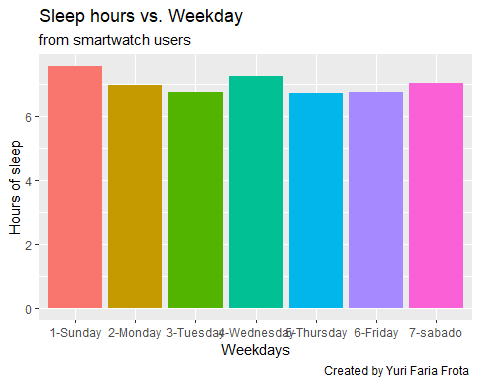


\*\* On a weekday distribution, we can observe that there is no noticed discrepancies. In other words, the users doesn’t have different activity habits during weekdays or weekends.\*\*

ggplot(data=semana1) +  
 geom\_boxplot(mapping=aes(x=diadasemana, y= Calories, fill = diadasemana ))+  
 labs (title = "CALORIES BURN X WEEKDAY ", subtitle = " ",   
 caption= " Created by Yuri Faria Frota ", x = "Weekdays", y = "Calories burn")

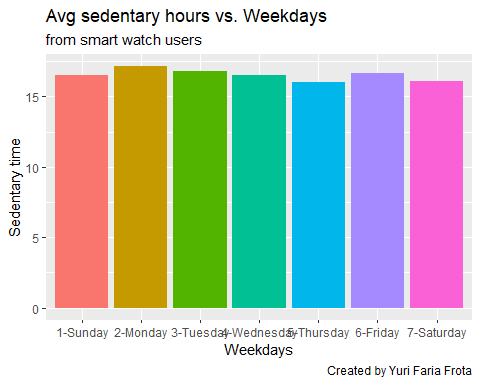


semanasono %>%   
 group\_by(diadasemana) %>%   
 summarise(totalsono = mean(TotalMinutesAsleep)) %>%   
 ggplot(aes(x = diadasemana, y = totalsono/60, fill = diadasemana)) +  
 geom\_col() +  
 labs (title = "Sleep hours vs. Weekday", subtitle = "from smartwatch users ",   
 caption= " Created by Yuri Faria Frota ", x = " Weekdays ", y = "Hours of sleep") + theme(legend.position = "none")



**We can see a slight increase on sleep hours, on Sundays and Wednesdays.**

semana1 %>%   
 group\_by(diadasemana) %>%   
 summarise( meansedentaryminutes= mean(SedentaryMinutes) ) %>%   
 ggplot(aes(x = diadasemana, y = (meansedentaryminutes/60), fill = diadasemana)) +  
 geom\_col() +  
 labs (title = "Avg sedentary hours vs. Weekdays", subtitle = "from smart watch users",   
 caption= "Created by Yuri Faria Frota", x = "Weekdays", y = "Sedentary time") + theme(legend.position = "none")



**The plot above represents the mean sedentary hours, by weekday. Note that there is a slight difference, showing that in general, people keep on sedentary activities during all days, including weekends.**

## PHASE 5 - SHARE

The analysis give us a few conclusions: On average, the users of this smart device are sedentary, and sleep little, with poor activities during all week. This helps us to find out one major characteristic from the users: **They are people who need to improve their life quality**

Through the plots we could see the importance to keep in activity, and also with high activity intensities, generating significant increase in calories consumption. We also observer that the simple fact of walk (light activity) is enough to keep a good rate of activity, resulting in significant impacts on calories consumption. Keep in activity is the major thing!

\*\*Suggestions: \*

\*We want to improve life quality from our users> Emphasize the most important characteristics, and show them the evolution of their daily habits, through data feedback, could make them more engaged with their life style habits, and consequently, with bellabeat services and devices.

\*Creating a well structures database coming from bellabeat devices, to direct the insights into a more efficient way.

\*The calories consumption does not vary accordingly to weekdays. We could encourage them to have more activities or weekends, through “deliverables”, or a “gamification” system.

\*A more completed data about weight and sleep characteristics could bring us important information. It is very important to create more efficient ways to obtain these data, and do a research about the real motive that make this data hard to obtain.