Bellbet case of study

Introduccion

This case study is presented as a final assignment to the Google Data Analytics Certificate, where we are going to analyze activity wellness tracking data to find out insights to the company Bellabeat.

The data analysis process consists of six steps:

1. Ask
2. Prepare
3. Process
4. Analyze
5. Share
6. Act

Scenario

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. You have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy

1.Ask

Business task Give insights about how consumers use their smart devices in order to understand trends and behaviours to guide future marketing strategies, like so looking for new growth opportunities for the company.

Key stakeholders

- **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer
- **Sando Mur**: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team
- **Bellabeat** marketing analytics team

2.Prepare

Dataset description The data was generated by respondents to a survey via Amazon Mechanical Turk from March-2016 to May-2016. The datasets are public data from FitBit Fitness Tracker Data and include the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.

ROCCC dataset?

- Reliable: It is not very reliable due to the low numbers of participants (Thirty eligible Fitbit users)
- Original: It was not generated by the company. Is a Third party dataset
- Comprehensive: Is not very comprehensive, because there is no metadata for the datasets
- Current: The source of the data has time on 2016

- Cited: The data is well documented and cited

3.Process

Setting up my environment

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Loading your CSV files

Here we'll create a dataframe named 'daily_activity' and read in one of the CSV files from the dataset.

```
daily_activity <- read.csv("./Dataset/dailyActivity_merged.csv")

sleep_day <- read.csv("./Dataset/sleepDay_merged.csv")

weightLogInfo <- read.csv("./Dataset/weightLogInfo_merged.csv")
```

Exploring a few key tables and column names

Identify all the columns in the daily_activity data.

```
colnames(daily_activity)
```

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

Identify all the columns in the sleep data.

```
colnames(sleep_day)
```

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

Identify all the columns in the weightLogInfo data.

```
colnames(weightLogInfo)
```

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

Looking for NA values for a key datasets

Looking for NA values in the daily_activity data

```
colSums(is.na(daily_activity))
```

```
##                       Id             ActivityDate                TotalSteps
##                        0                        0                         0
##            TotalDistance          TrackerDistance LoggedActivitiesDistance
##                        0                        0                         0
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                        0                        0                         0
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                        0                        0                         0
##     LightlyActiveMinutes         SedentaryMinutes                  Calories
##                        0                        0                         0
```

Looking for NA values in the sleep data

```
colSums(is.na(sleep_day))
```

```
##                 Id           SleepDay  TotalSleepRecords TotalMinutesAsleep
##                  0                  0                  0                  0
##     TotalTimeInBed
##                  0
```

Looking for NA values in the weightLogInfo

```
colSums(is.na(weightLogInfo))
```

```
##             Id           Date       WeightKg   WeightPounds            Fat
##              0              0              0              0             65
##            BMI IsManualReport          LogId
##              0              0              0
```

As it was shown there is no NA values for the datasets and also they are well named (camel case), so we can continue with analyze.

4.Analyze

Understanding some summary statistics

How many unique participants are there in each dataframe?

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

```
n_distinct(weightLogInfo$Id)
```

```
## [1] 8
```

```
nrow(daily_activity)
```

```
## [1] 940
```

```
nrow(sleep_day)
```

```
## [1] 413
```

```
nrow(weightLogInfo)
```

```
## [1] 67
```

What are some quick summary statistics we'd want to know about each data frame?

For the daily activity dataframe:

```
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8
##  Median : 7406   Median : 5.245   Median :1057.5
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5
##  Max.   :36019   Max.   :28.030   Max.   :1440.0
```

For the sleep dataframe:

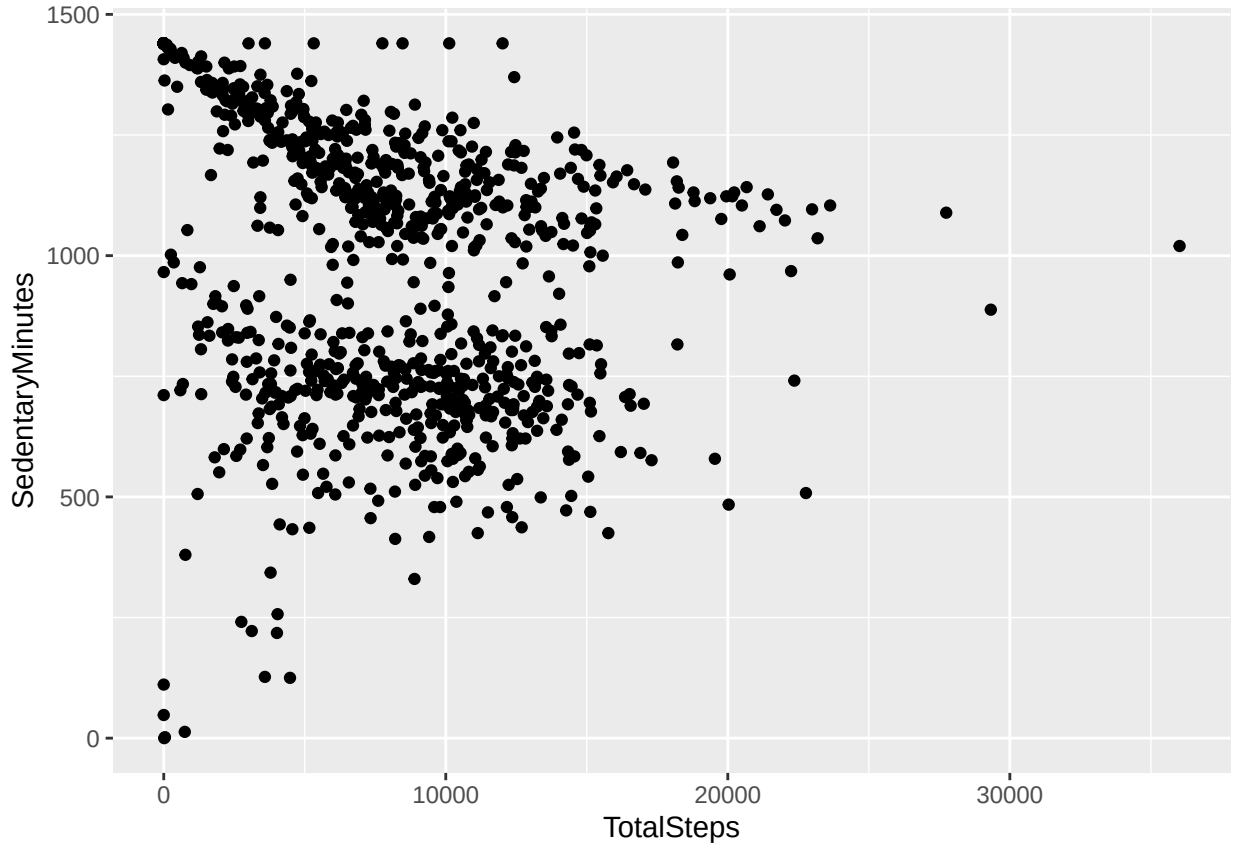
```
sleep_day %>%
  select(TotalSleepRecords,
  TotalMinutesAsleep,
  TotalTimeInBed) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

Plotting a few explorations

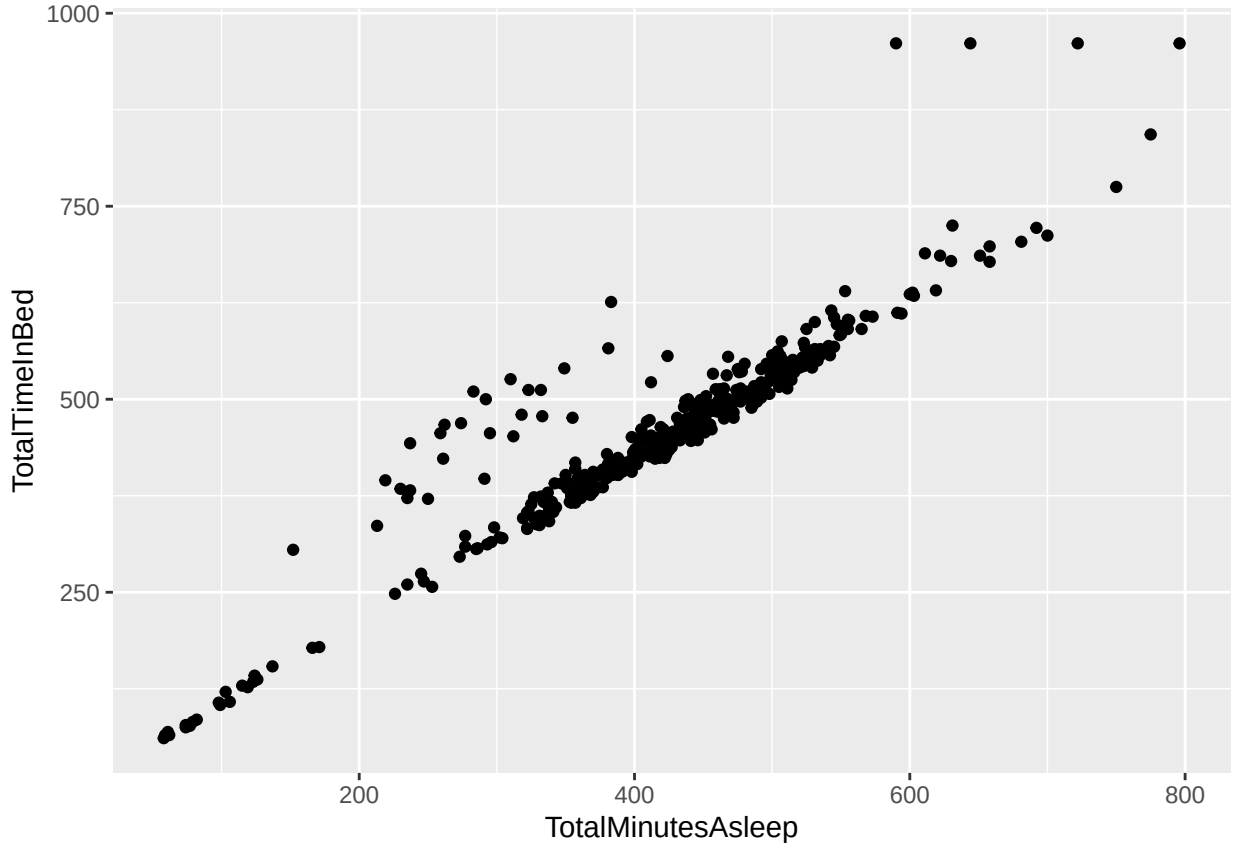
What's the relationship between steps taken in a day and sedentary minutes?

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point()
```



What's the relationship between minutes asleep and time in bed?

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



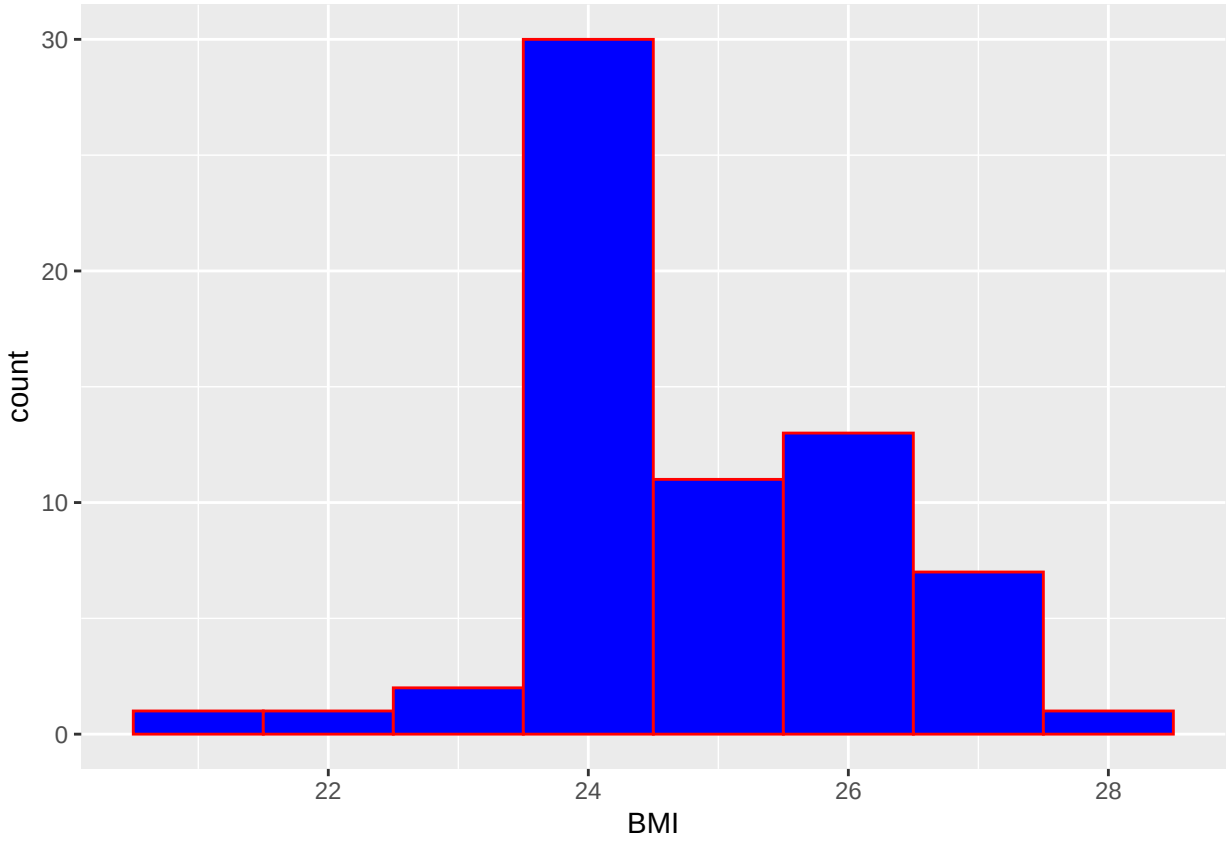

```
weightLogInfo2 <- weightLogInfo[-c(3),]
```

Now we are going to take a look on BMI values. Acording with the National Heart, Lung and Blood Institute, and consider the next **BMI Categories**:

- Underweight = <18.5
- Normal weight = 18.5–24.9
- Overweight = 25–29.9
- Obesity = BMI of 30 or greater

We clearly that the majority of this people are in the "Overweight category".

```
ggplot(data=weightLogInfo2, aes(x = BMI) ) + geom_histogram(binwidth=1, fill = "blue", color = "red")
```



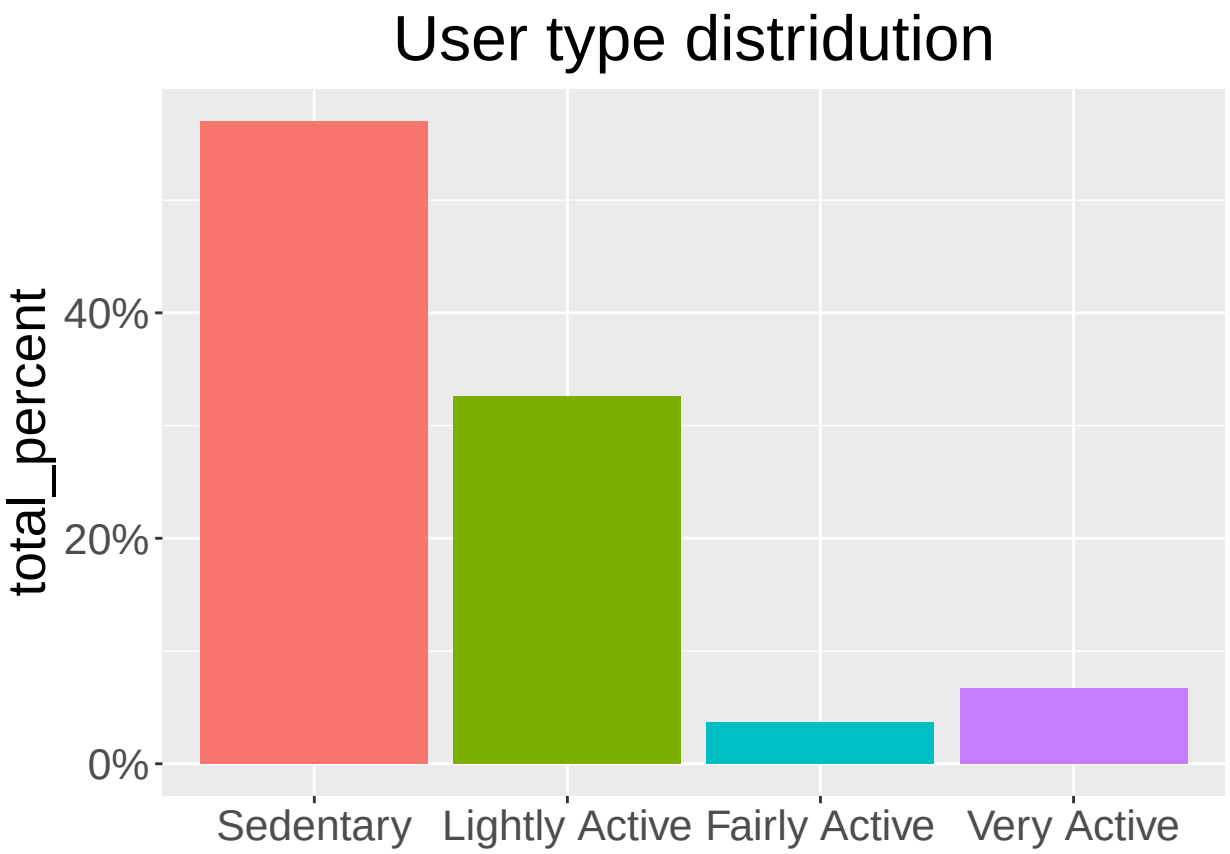
In order to have a better idea about how people behavior we can create categories according with the different activation/sedentary minutes.

```
data_by_usertype <- daily_activity  %>%

  summarise(
    user_type = factor(case_when(

      SedentaryMinutes > mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fa

      SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes > mean(LightlyActiveMinutes) & Fa

      SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fa

      SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fa

  ),
levels=c("Sedentary", "Lightly Active", "Fairly Active", "Very Active")), Calories, .group=Id) %>%
  drop_na()
```

It easy to see that most of the participants who have a FitBit Fitness Tracker are not actually having a fitness life style.

```
data_by_usertype %>%
group_by(user_type) %>%
summarise(total = n()) %>%
mutate(totals = sum(total)) %>%
```

```
group_by(user_type) %>%
summarise(total_percent = total / totals) %>%
ggplot(aes(user_type,y=total_percent, fill=user_type)) +
    geom_col()+
    scale_y_continuous(labels = scales::percent) +
    theme(legend.position="none") +
    labs(title="User type distridution", x=NULL) +
    theme(legend.position="none", text = element_text(size = 20),plot.title = element_text(hjust = 0.5))
```

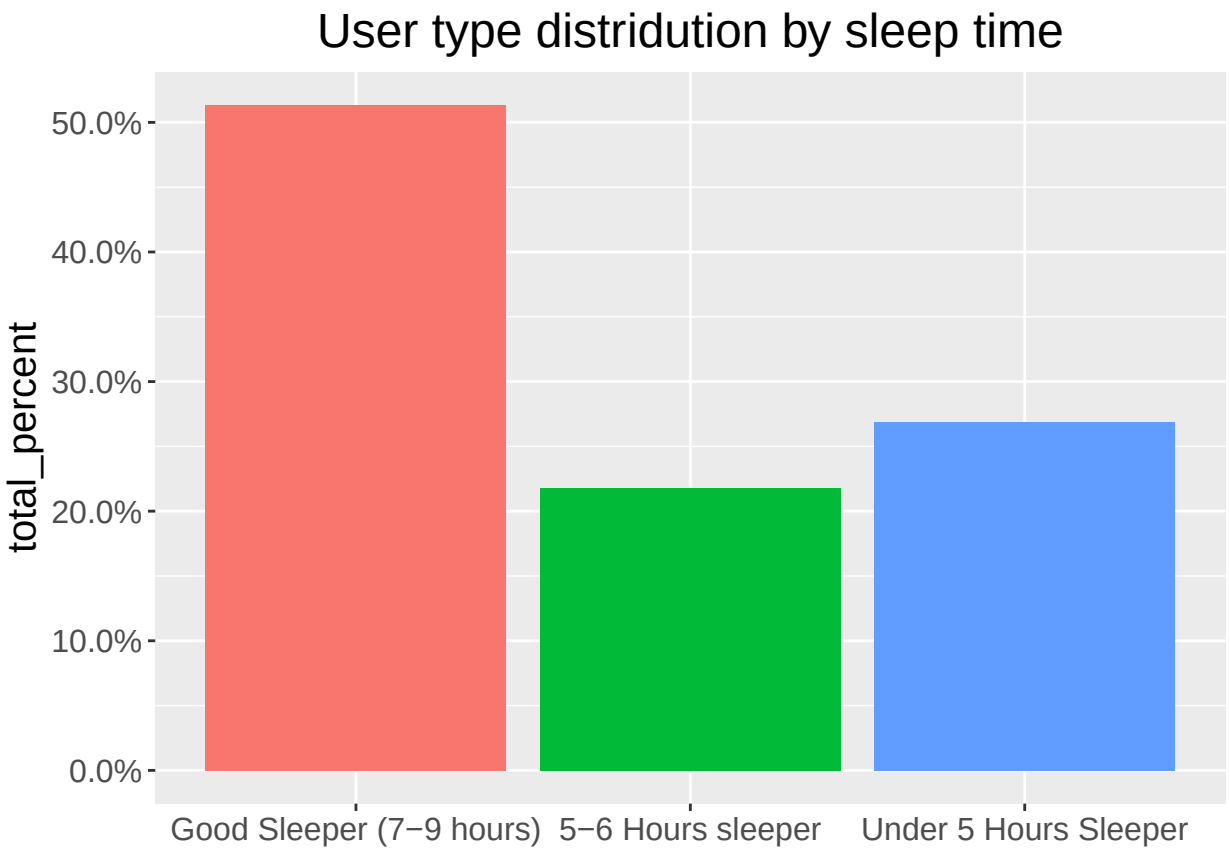


Now we are going to focus on the sleep time, and see how well o how bad are the sleep time for our participants. According with Sleep Foundation the recomended hours of sleep for Adults are between 7-9 hours.

```
data_by_usersleep <- sleep_day  %>%

  summarise(
    user_type = factor(case_when(

      TotalMinutesAsleep > 420 & TotalMinutesAsleep < 540 ~ "Good Sleeper (7-9 hours)",

      TotalMinutesAsleep < 420 & TotalMinutesAsleep > 360 ~ "5-6 Hours sleeper",

      TotalMinutesAsleep < 360 ~ "Under 5 Hours Sleeper",


  ),
levels=c("Good Sleeper (7-9 hours)", "5-6 Hours sleeper", "Under 5 Hours Sleeper")), TotalMinutesAsleep
```

```
  drop_na()
```

```
data_by_usersleep %>%
group_by(user_type) %>%
summarise(total = n()) %>%
mutate(totals = sum(total)) %>%
group_by(user_type) %>%
summarise(total_percent = total / totals) %>%
ggplot(aes(user_type,y=total_percent, fill=user_type)) +
    geom_col()+
    scale_y_continuous(labels = scales::percent) +
    theme(legend.position="none") +
    labels(title="User type distridution by sleep time", x=NULL) +
    theme(legend.position="none", text = element_text(size = 15),plot.title = element_text(hjust = 0.5))
```



It is good to know than the half of our participants sleep the recommended hours, but a huge percent have sleep problems.

5. Share

Link for a presentation

6.Act

Recommendations And Conclusion

The key recommendations that I can make for the stakeholders Urška Sršen, Sando Mur, and the Bellabeat's marketing analytics team are the next ones:

1. Try to create an internal dataset for customers using the Leaf smart device, at this way you can create a comparison between your product and the FitBit Fitness Tracker and the behavior of your users.
2. Try to increse the number of participants in the dataset, in some cases with the FitBit Fitness Tracker Dataset i been working just with 24 people, these can produce some bias on the analisys process.
3. I found some limitations on the data, for this the obtained results must NOT be using for the marketing team or it could be used carefully.

Conclusion Even the limitations of the dataset, for these case of study some interesting finds are found. The most interesting and valuable information is that the most interested people on the FitBit Fitness Tracker are in the "Overweight" category and also are people with a "sedentary" life style, this behavior reflects that most of them had the intention of change their life style but for whatever reason they can't. and at the end the FitBit Fitness Tracker ends like an accessory. May be try to encourage the users to have a more active life using notifications or an app could be a good idea.

Finally in order to answers the business task, the problems with the dataset make difficult try to give an unbiased insight for the marketing team. But the key finding are mention above. Try to change the focus on the Bellabeat Leaf product audience for "Sedentary persons" instead of athletes and encourage them to have a more healthy/active life style.