Bellabeat Case Study

Scenario

Bellabeat is a high-tech manufacturer of health-focused products for women. These products are Bellabeat app, Leaf, Time, Spring and Bellabeat membership. The company is small but successful and they have the potential to become a larger player in the global smart device market. As the Bellabeat marketing analytics team we have been asked to analyze smart device data to gain insight into how consumers using their smart devices. The insights we discover will then help guide marketing strategy for the company.

Act

Business task

Analyze data from FitBit Fitness Tracker to gain insight and help guide Bellabeat's marketing strategy and unlock growth opportunities for the company.

Key stakeholders

Primary stakeholders: Urška Sršen and Sando Mur, executive team members
Secondary stakeholders: Bellabeat marketing analytics team

Prepare

The data for this analysis will come from Fitbit Fitness Tracker Data stored on Kaggle. This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 12.04.2016-12.05.2016. It contains personal fitness tracker from 30 Fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits. Let's see if this data source follow a ROCCC analysis:

- Reliable: dataset was collected from only 30 Fitbit users
- Original: thirty party data collect via Amazon Mechanical Turk
- Comprehensive: dataset contain multiple and various information for daily activity and its intensity, daily steps and sleep time, calories, heart rate, weight etc., but all these fields are not recorded for all these 30 participants of this survey
- Current: data are not so current because are from 6 years ago and the last years people change their habits and the way they live and exercise
- Cited: Unknown

Limitations of the dataset:

This dataset has low reliability and originality. It would be preferred a larger sample size with more characteristics, such as all fields completed from all these 30 Fitbit users and

information such as women's menstruation.

Process

First of all dailySteps, dailyCalories datasets are checked if they have the same number of observations and the same ids with the dailyActivity dataset. These all three datasets have the same size and identical ids, so it can be concluded that the first two datasets contained in dailyActivity. So, for the data analysis the datasets that used are dailyActivity, hourlySteps and sleepDay. WeightLogInfo has only data from 8 distinct people so for a more accurate analysis it is not used.

The any() and is.na() functions are used to check for NA values. The any() and duplicated() functions are used to check for any duplicated values in our datasets and delete them as shown below.

```
any(duplicated(sleepDay))

#Remove duplicated values from sleepDay
sleepDay <- sleepDay[!duplicated(sleepDay), ]
```

Change the data type of ActivityDate and convert to date format, and add one more column to correspond each day to a weekday. We did the same for the column SleepDay of sleepDay dataset.

```
dailyActivity$ActivityDate <- format(as.Date(dailyActivity$ActivityDate, "%m/%d/%y"), "%d/%m/%y")
dailyActivity$ActivityDate <- as.Date(dailyActivity$ActivityDate, "%d/%m/%y")
dailyActivity$WeekDay <- weekdays(dailyActivity$ActivityDate)
```

The merge() function is used to merge dailyActivity and sleepDay datasets in one by Id and ActivityDate to use it later.
Confirm that TotalTimeInBed is greater or equal to TotalMinutesAsleep for each record.
Limit decimal digits VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryMinutes and TotalDistance.

```
data[, 'VeryActiveDistance'] = round(data[, 'VeryActiveDistance'], 2)
data[, 'ModeratelyActiveDistance'] = round(data[, 'ModeratelyActiveDistance'], 2)
data[, 'LightActiveDistance'] = round(data[, 'LightActiveDistance'], 2)
data$TotalDistance = data$VeryActiveDistance + data$ModeratelyActiveDistance + data$LightActiveDistance
```

Order from Monday to Sunday to plot later.

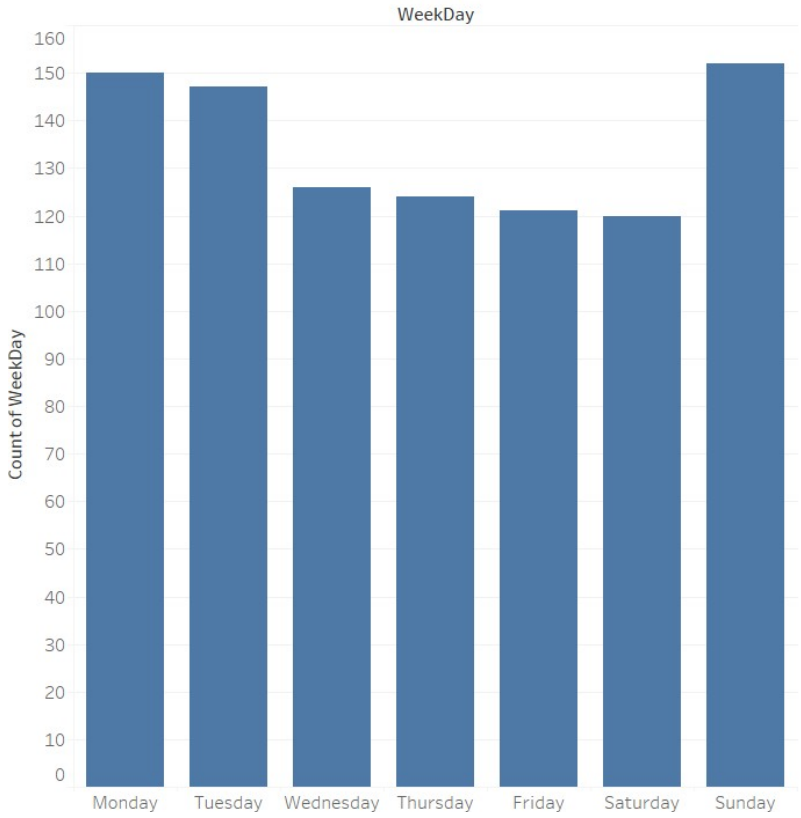
```
weekday_order <- c("Δευτέρα", "Τρίτη", "Τετάρτη", "Πέμπτη", "Παρασκευή", "Σάββατο", "Κυριακή")
data <- data %>% arrange(match(WeekDay, weekday_order))
```

Change the format of Activity hour of hourlySteps dataset and get the hour into a new column.

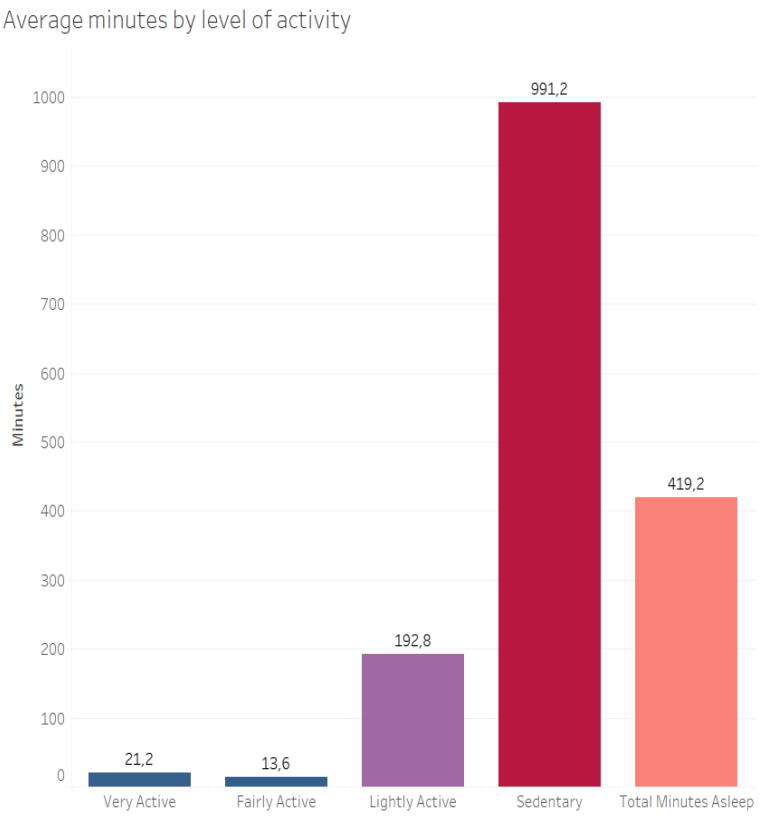
```
#Change the format of ActivityHour of hourlySteps and get the hour into a new column
str(hourlySteps)
hourlySteps$ActivityHour <- parse_date_time(hourlySteps$ActivityHour, "%m/%d/%y %I:%M:%S %p")
hourlySteps$Hour <- format(as.POSIXct(hourlySteps$ActivityHour), format = "%H")
hourlySteps$ActivityHour
write_csv(hourlySteps, "hourlySteps_merged.csv")
```

Analyze

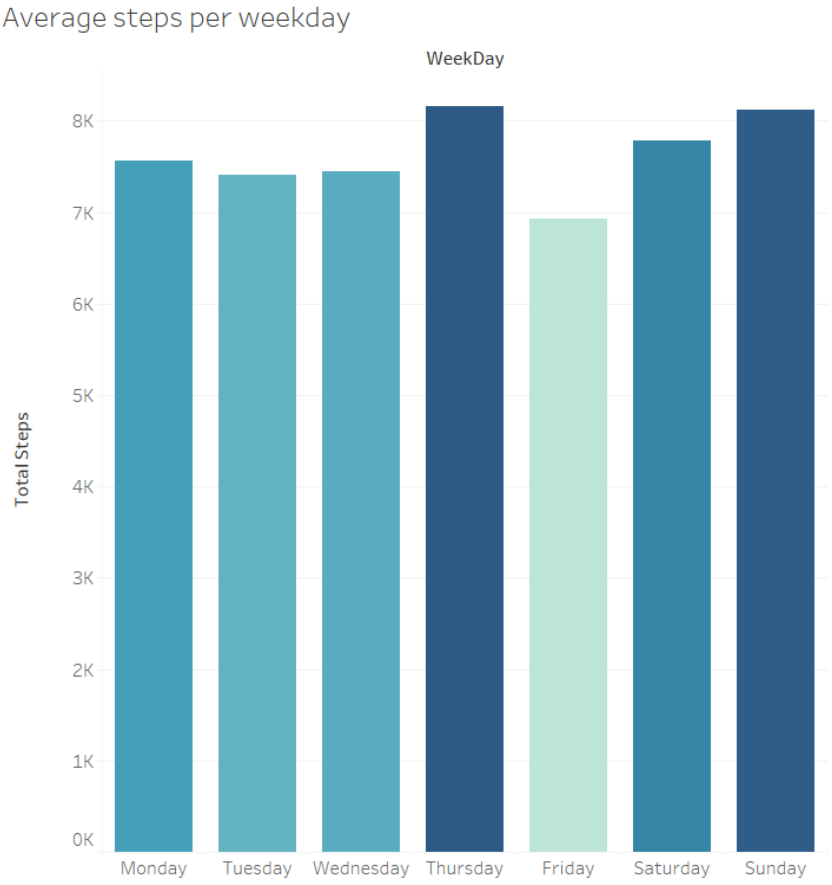
Before we start the analysis phase we need to see the records per weekday. As we can see from the bar graph below the data records are more from Sunday to Tuesday. So, the data are not so comprehensive for an accurate analysis. Therefore, where is necessary, in order to have more accurate results, we don't calculate the sum but the average of the corresponding fields.



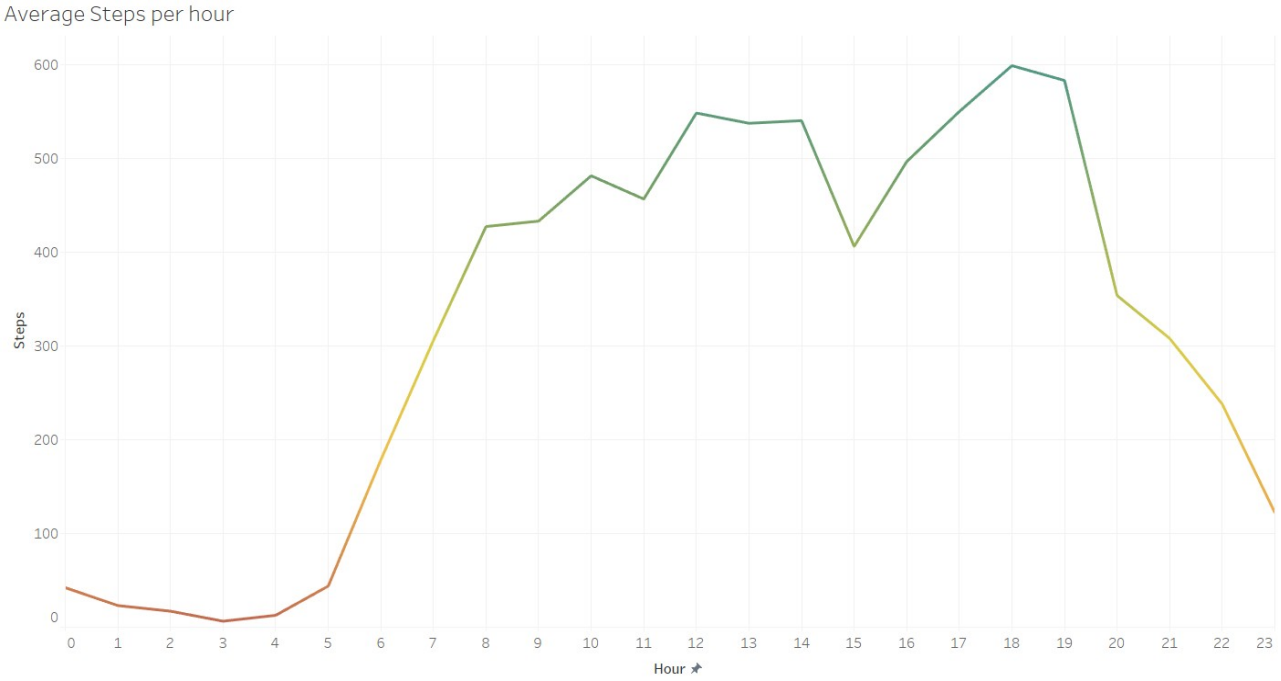
First of all, we discover how the minutes are distributed among the different levels of activity and minutes for sleep. From the bar graph below we can see that Sedentary minutes are the most in a day which makes sense since the app track our activity all the day and we don't exercise all the day. The interesting thing is that the most active minutes are in the level of Lightly Active minutes and the least are in Fairly Active minutes, which are only a little amount of active minutes in general. FitBit users spend about 7 hours sleeping, only about 35 minutes of very and fairly level of activity, 3 hours lightly active and the rest of the day in sedentary minutes.



We now want to calculate which days users do more steps. The day with the most days is Wednesday and follow Sunday and Saturday.



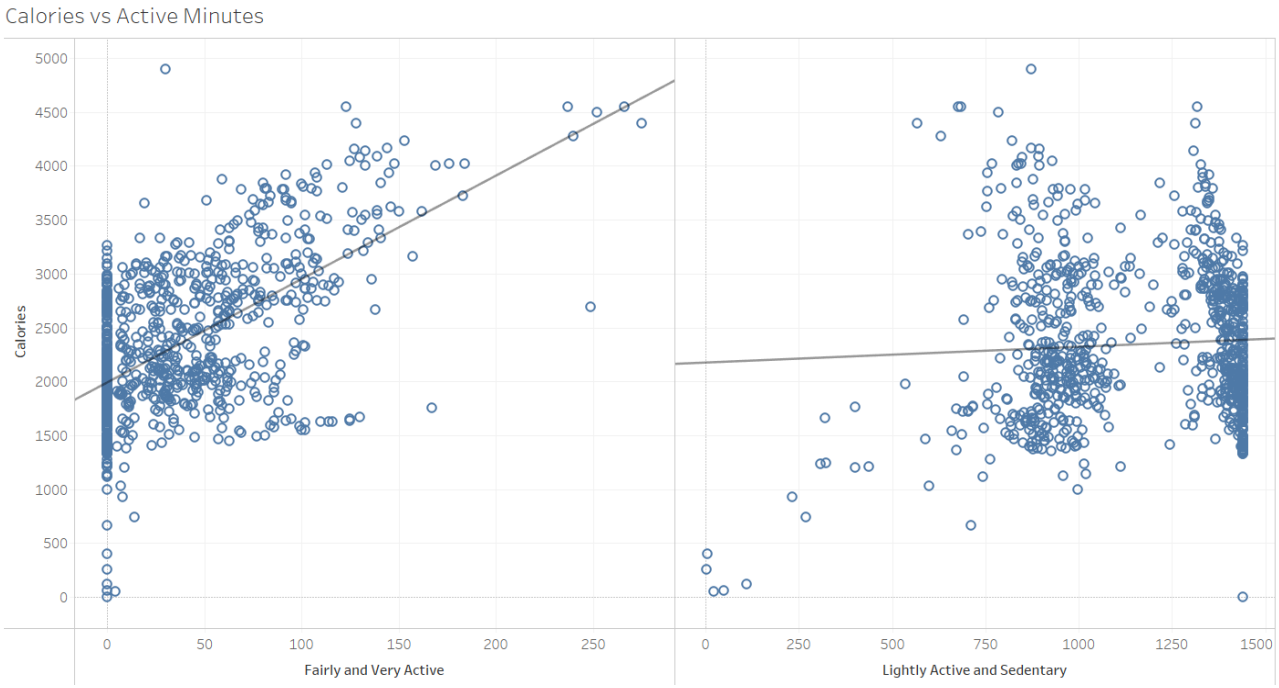
Average steps per hour shows that more active hours are 5-7 PM and follow 12-14 PM.



As we can see from the next figure we conclude that the more steps users take the more calories users burn, something that we expected. The more green circles mean that we have more Very Active minutes. We see that this happens when the number of total steps (and secondarily calories) increase.

Next, correlating calories with very and fairly active minutes comparing to lightly active

and sedentary minutes we see that the first trend line has a positive slope, which means that the more minutes users spend on higher level activity the more calories they burn, while on the other hand the second trend line (and low R-Squared of this) shows that lower levels of activity doesn't play a significant role on how much calories burnt.



Final, we examine the correlation between total steps users take and the different levels of activity. We conclude that the more high intensity of activity the more a positive slope we have. On sedentary minutes we see that the more users spend on this level of activity the less steps they take.

Share

Tableau Presentation

Act

Recommendations for Bellabeat's marketing strategy:

- Record and other useful measurements such women's period, quality of sleep etc. Expand the audience not only to women.
- Encourage users to create a complete profile (age, sex, weight etc) and the goals they want to achieve. Enable alert notifications to encourage users to meet these goals.
- Enable alert notifications to remind users spend more time on high level activity and less on sedentary minutes.
- For Bellabeat membership, users that achieve their goals can rewarded with a discount on Bellabeat's products.