Capstone case study 2

Emmanuel Nwosu

2024-07-30

### Broad Objective

This case study is designed to use fitness data to inform the stakeholders at Bellabeat, a high-tech company that manufactures health-focused smart products, especially for women. The Analysis of the fitness data would help the company gain insights into the usage of their products and the outcomes. This analysis follows the six steps of the data analysis process: ask, prepare, process, analyze, share, and act.

####  Step One: ASK

In this phase, the objective is to analyze smart device usage data to gain insight into how consumers use non-Bellabeat smart devices. The guiding questions are as follows:

##### a) Guiding Questions:

• What is the problem you are trying to solve? • How can your insights drive business decisions?

##### b) Key Tasks:

• Identification of the business task: To use the information from the data to help the company target the fitness market properly.

##### • The stakeholders identified:

The main stakeholders of interest in this analysis are: Urška Sršen: Bellabeat’s co-founder and Chief Creative Officer; and Sando Mur: Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team. Others are the Bellabeat marketing analytics team, which is a team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat’s marketing strategy.

##### c) The business task:

The business task is to understand the usage pattern of these smart devices namely the Bellabeat app, Leaf, Time and Spring and use the information to better inform the stakeholders on making better marketing decisions.

#### Step Two: Prepare

##### Deliverable: A description of all data sources used

In this step, the data used for the analysis is a public dataset called FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius). It is stored as a Kaggle data set containing personal fitness trackers from thirty Fitbit users. The data is stored in different files detailing daily activities, hourly activities, and so on, which are stored as CSV files to make data merging and cleaning easier since the data contains primary and secondary keys.

##### Sort and Filtering of Data:

The focus in this analysis is the daily data instead of the hourly data because once the hourly or minute data are aggregated into daily data, one would get average information about user behaviour. I did some merging of data where necessary, although some of them were not utilized such as merging of hourly data.

### Setting up My Environment

NB: The packages that are needed for data cleaning and analysis are installed and loaded. These include tidyverse, ggplot2, etc as shown in the code chunk

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)  
library(dplyr)  
library(skimr)  
library(janitor)

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

library(here)

## here() starts at C:/Users/USER/Documents/RStudio Exercises and Data

### Data Processing: The datasets are imported and cleaned for Analysis

NB: They are imported as .csv files from the unzipped folder

hourlyCalories\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/hourlyCalories\_merged.csv")  
View(hourlyCalories\_merged)  
head(hourlyCalories\_merged)

## Id ActivityHour Calories  
## 1 1503960366 4/12/2016 12:00:00 AM 81  
## 2 1503960366 4/12/2016 1:00:00 AM 61  
## 3 1503960366 4/12/2016 2:00:00 AM 59  
## 4 1503960366 4/12/2016 3:00:00 AM 47  
## 5 1503960366 4/12/2016 4:00:00 AM 48  
## 6 1503960366 4/12/2016 5:00:00 AM 48

hourlyIntensities\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/hourlyIntensities\_merged.csv")  
#Correct merging requires using both Id and ActivityHour  
new\_data\_merges<- merge(hourlyCalories\_merged, hourlyIntensities\_merged, by=c("Id", "ActivityHour"))  
#the other data is imported from .csv  
hourlySteps\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/hourlySteps\_merged.csv")  
new\_data\_merges2<- merge(new\_data\_merges, hourlySteps\_merged, by=c("Id", "ActivityHour"))  
head(new\_data\_merges2)

## 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

#Next we import and merge daily activity data  
dailyActivity\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv")  
dailyCalories\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/dailyCalories\_merged.csv")  
dailyIntensities\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/dailyIntensities\_merged.csv")  
dailySteps\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/dailySteps\_merged.csv")  
#Next the minute by minute data are imported for merging  
sleepDay\_merged <- read.csv("~/RStudio Exercises and Data/archive (2)/mturkfitbit\_export\_4.12.16-5.12.16/Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv")  
sleep\_dailyactivity\_data\_merged <- merge(dailyActivity\_merged, sleepDay\_merged, by="Id")

### Data Analysis: Data Cleaning and Descriptive Statistics

filtered\_daily\_activity\_m <- dailyActivity\_merged %>%  
 group\_by(Id) %>%  
 summarise(mean\_totsteps1 = mean(TotalSteps, na.rm = T))  
View(filtered\_daily\_activity\_m)  
group\_mean <- sleepDay\_merged %>%  
 group\_by(Id) %>%  
 summarise\_at(vars(TotalMinutesAsleep),  
 list(Mean\_TotalMinutesAsleep = mean))  
View(group\_mean)  
View(group\_mean)  
# Summarizing by ID   
new\_data\_merges2 %>%   
 group\_by(Id) %>%   
 drop\_na() %>%   
 summarize( max\_calor = max(Calories ), mean\_calor = mean(Calories), max\_totalintensity= max(TotalIntensity), mean\_totalintensity=mean(TotalIntensity), cor(Calories,TotalIntensity))

## # A tibble: 33 × 6  
## Id max\_calor mean\_calor max\_totalintensity mean\_totalintensity  
## <dbl> <int> <dbl> <int> <dbl>  
## 1 1503960366 318 78.5 159 16.2   
## 2 1624580081 374 62.5 169 8.04  
## 3 1644430081 456 119. 130 10.5   
## 4 1844505072 184 66.6 67 5.02  
## 5 1927972279 265 91.5 63 1.86  
## 6 2022484408 486 105. 157 17.0   
## 7 2026352035 171 64.9 56 10.8   
## 8 2320127002 191 72.6 67 8.74  
## 9 2347167796 334 88.7 132 14.5   
## 10 2873212765 410 80.2 148 15.1   
## # ℹ 23 more rows  
## # ℹ 1 more variable: `cor(Calories, TotalIntensity)` <dbl>

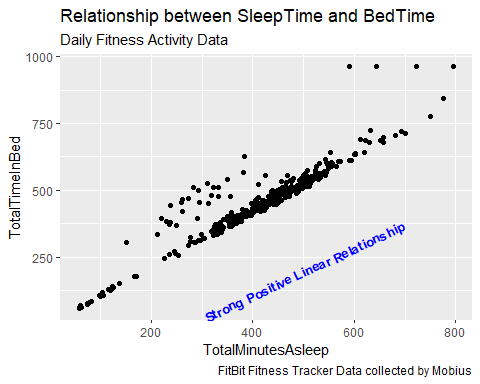
# Summarizing merged\_daily\_sleep\_data\_merged by ID  
sleep\_dailyactivity\_data\_merged %>%   
 group\_by(Id) %>%   
 drop\_na() %>%   
 summarize( mean\_calor = mean(Calories), mean\_TotalSteps=mean(TotalSteps), mean\_TotalDistance=mean(TotalDistance), mean\_SedentaryMinutes=mean(SedentaryMinutes), mean\_TotalMinutesAsleep= mean(TotalMinutesAsleep),  
 mean\_TotalTimeInBed=mean(TotalTimeInBed))

## # A tibble: 24 × 7  
## Id mean\_calor mean\_TotalSteps mean\_TotalDistance mean\_SedentaryMinutes  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1.50e9 1816. 12117. 7.81 848.  
## 2 1.64e9 2811. 7283. 5.30 1162.  
## 3 1.84e9 1573. 2580. 1.71 1207.  
## 4 1.93e9 2173. 916. 0.635 1317.  
## 5 2.03e9 1541. 5567. 3.45 689.  
## 6 2.32e9 1724. 4717. 3.19 1220.  
## 7 2.35e9 2043. 9520. 6.36 687.  
## 8 3.98e9 1514. 10985. 7.52 708.  
## 9 4.02e9 2386. 2267. 1.63 1237.  
## 10 4.32e9 2038. 7269. 4.89 736.  
## # ℹ 14 more rows  
## # ℹ 2 more variables: mean\_TotalMinutesAsleep <dbl>, mean\_TotalTimeInBed <dbl>

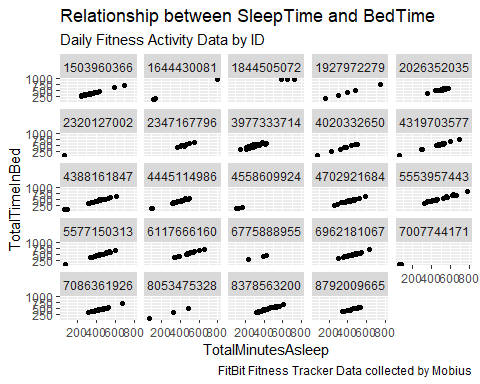
## The Data Visualizations

NB: This shows pictorially how some variables are related

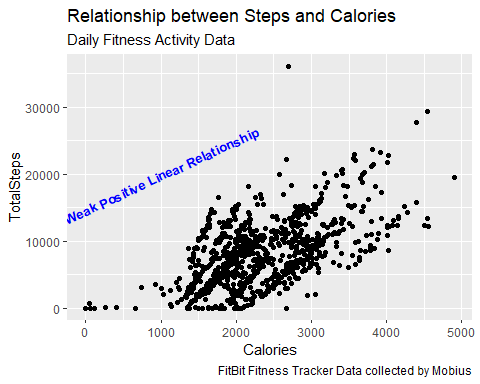
#There appears to be almost a perfect linear relationship between TotalMinutesAsleep and TotalTimeInBed by respondents ID  
ggplot(data = sleepDay\_merged, mapping = aes(x=TotalMinutesAsleep, y =TotalTimeInBed))+ geom\_point() + labs(title = "Relationship between SleepTime and BedTime", subtitle="Daily Fitness Activity Data", caption = "FitBit Fitness Tracker Data collected by Mobius") + annotate("text", x=500, y=200, label="Strong Positive Linear Relationship", color="blue", fontface="bold", size=3.5, angle=25)



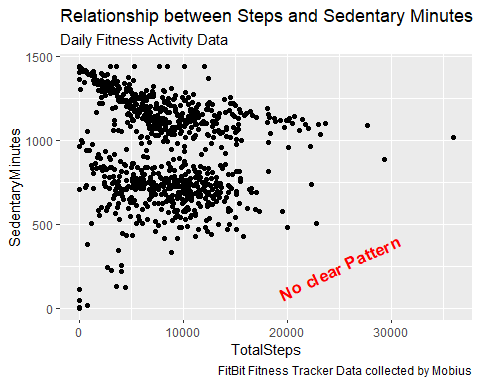
ggplot(data = sleepDay\_merged, mapping = aes(x=TotalMinutesAsleep, y =TotalTimeInBed))+ geom\_point() + facet\_wrap(~Id) + labs(title = "Relationship between SleepTime and BedTime", subtitle="Daily Fitness Activity Data by ID", caption = "FitBit Fitness Tracker Data collected by Mobius")



ggplot(dailyActivity\_merged, mapping = aes(x = Calories, y = TotalSteps)) +geom\_point() + labs(title = "Relationship between Steps and Calories", subtitle="Daily Fitness Activity Data", caption = "FitBit Fitness Tracker Data collected by Mobius") + annotate("text", x=1000, y=20000, label="Weak Positive Linear Relationship", color="blue", fontface="bold", size=3.5, angle=25)



ggplot(data=dailyActivity\_merged, aes(x=TotalSteps, y=SedentaryMinutes)) + geom\_point() + labs(title = "Relationship between Steps and Sedentary Minutes", subtitle="Daily Fitness Activity Data", caption = "FitBit Fitness Tracker Data collected by Mobius") + annotate("text", x=25000, y=250, label="No clear Pattern", color="red", fontface="bold", size=4.5, angle=25)



#### Share and Act: Insights

The findings show that total steps are achieved per day is highly correlated with the calories burned either by looking at the individual data. But aggregate data seems to suggest a positive relationship but weak correlation.

There is no clear pattern between total steps and sedentary minutes.

Total sleep time and time in bed are highly correlated. This means that individuals on average, go to bed only when they are about to sleep. They also get sufficient sleep that commensurates with the bedtime.