Bellabeat_Project

Ebenezer Nii Okai Mensah

2023-07-15

About the Company

Bellabeat ia a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market.

Analysis Questions

What are some trends in smart device usage? How could these trends apply to Bellabeat customers? How could these trends help influence Bellabeat's marketing strategy?

Data Source:

Data: Kaggle-FitBit Fitness Tracker Data: link here Author: Mobius

ASK PHASE

Business Plan

Is to analyze smart device usage and identify trends and insight to be applied to Bellabeat marketing strategies to unlock new growth opportunities.

Stakeholders

Urška Sršen - Bellabeat cofounder and Chief Creative Officer Sando Mur - Mathematician and Bellabeat's cofounder

PREPARE PHASE

Data and Organization

Data contains 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. There are 18 csv files in the folder. The data has been organized in a long format.

Credibility and Integrity

Using the ROCCC, Reliable,Original,Comprehensive, Current and Cited - The data won't pass the mark of reliable because it is limited to only 30 fitbit users and this data was originally not collected by the bellabeat company. This data is not current, it was updated 3 years ago. There is a sampling bias based on demographic selection. In terms of gender selection we don’t know the percentage of each gender in the sampling selection. There is limitation of time because the data being used was collected within 1 month (4-12-2016 - 5-12-2016). The date format is not consistent throughout the files, daily activity has only date and sleep and heartrate has date with time format. The datatype to all the dates in all 3 files are character formats, this needs to be changed to date format.

Licensing, privacy and accessibility

The data is public data and the owner has voluntarily waived all rights worldwide under copyright law. This makes this dataset free to use, modify and distribute without any prior permission.

Installing and loading common packages and libraries

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.1

## ── 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_conflict
s() ──
## ✕ 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
```

Loading the source for the files

```
setwd("C:/Users/USER/Desktop/Capstone Project/Fitbit_tracker_data/Fitabase Da
ta")
```

Loading your CSV files

I will be working with this files - dailyActivity_merged.csv - sleepDay_merged.csv - weightLogInfo_merged.csv

Dataframe for the daily activity data.

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

Dataframe for the sleep data.

```
sleep_day <- read.csv("sleepDay_merged.csv")
```

Dataframe for the weight data.

```
weight <- read.csv("weightLogInfo_merged.csv")
```

Sorting and Filtering datasets for NA

Data in all 3 files are in ascending order and I also checked for N/A in the id to remove any observation without Id's.

```
daily_activity <- filter(daily_activity,Id != " ")
sleep_day <- filter(sleep_day,Id != " ")
weight <- filter(weight,Id != " ")
```

PROCESS PHASE

Tool being used

I am using Rstudio for this analysis because it has the tools for processing, analysis, visualization an documentation.

Exploring the 3 files

Data and column names in daily_activity data.

```
head(daily_activity)

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"
```

Data and column names in sleep_day data.

```
head(sleep_day)

colnames(sleep_day)

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

Data and column names in weight data

```
head(weight)

colnames(weight)

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

Installing and loading common packages and libraries

```
library(skimr)
library(janitor)

library(here)
```

Check the data for errors

Cleaning

Time to cleanup! First we will check for the number of participants in each dataset if it truly matches the 30 fitbit members mentioned earlier.

```
n_distinct(daily_activity$Id)

## [1] 33

n_distinct(sleep_day$Id)

## [1] 24

n_distinct(weight)

## [1] 67
```

Observations for each dataframe

```
nrow(daily_activity)

## [1] 940

nrow(sleep_day)

## [1] 413

nrow(weight)

## [1] 67
```

Looking for duplicates

```
sum(duplicated(daily_activity))

## [1] 0

sum(duplicated(sleep_day))

## [1] 3

sum(duplicated(weight))

## [1] 0
```

Remove duplicates

You realize that there are duplicates in the sleep_day data frame. It has 413 observations.

```
sleep_day <- sleep_day %>%
  distinct()%>%
  drop_na()
```

Checking whether duplicate is gone

```
sum(duplicated(sleep_day))

## [1] 0
```

Renaming Columns names

I will rename all column names to a specific format thus in lowercase, which will help with merging of data frames in the coming steps of the analysis.

```
clean_names(daily_activity)

daily_activity <- rename_with(daily_activity,tolower)

clean_names(sleep_day)

sleep_day <- rename_with(sleep_day,tolower)

clean_names(weight)

weight <- rename_with(weight,tolower)
```

Cleaning Date and Time

Looking at the data we have in daily activity, sleep and weight, you will realize that, the formats are in characters instead of date. The column ActivityDate in dailyactivity has only date and the column Sleepday and date in sleepday and weight has date and time. For consistency in working with all files we need to format these columns into one format %M/%d/%y. We will ignore the time in the sleeepday and date column, we only require the date from both datasets.

Installing and loading common packages and libraries

```
library(lubridate)
```

changing column names to one consistent name

```
daily_activity <- daily_activity %>%
  rename(date = activitydate)
sleep_day <- sleep_day %>%
  rename(date = sleepday)
weight <- weight %>%
  rename(date = date)
```

Changing character format to date format for Consistency

```
daily_activity <- daily_activity %>%
  mutate(date = as.Date(date, format = "%m/%d/%Y"))
head(daily_activity)
```

```
##           id       date totalsteps totaldistance trackerdistance
## 1 1503960366 2016-04-12      13162          8.50            8.50
## 2 1503960366 2016-04-13      10735          6.97            6.97
## 3 1503960366 2016-04-14      10460          6.74            6.74
## 4 1503960366 2016-04-15       9762          6.28            6.28
## 5 1503960366 2016-04-16      12669          8.16            8.16
## 6 1503960366 2016-04-17       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
```

```
sleep_day <- sleep_day %>%
  mutate(date = as.Date(date, format = "%m/%d/%Y"))
head(sleep_day)
```

```
##           id       date totalsleeprecords totalminutesasleep totaltimeinbe
d
## 1 1503960366 2016-04-12                 1                327            34
6
## 2 1503960366 2016-04-13                 2                384            40
7
## 3 1503960366 2016-04-15                 1                412            44
2
## 4 1503960366 2016-04-16                 2                340            36
7
## 5 1503960366 2016-04-17                 1                700            71
2
## 6 1503960366 2016-04-19                 1                304            32
0
```

```
weight <- weight %>%
  mutate(date = as.Date(date, format = "%m/%d/%Y"))
head(weight)
```

```
##           id       date weightkg weightpounds fat   bmi ismanualreport
## 1 1503960366 2016-05-02     52.6     115.9631  22 22.65           True
## 2 1503960366 2016-05-03     52.6     115.9631  NA 22.65           True
## 3 1927972279 2016-04-13    133.5     294.3171  NA 47.54          False
## 4 2873212765 2016-04-21     56.7     125.0021  NA 21.45           True
## 5 2873212765 2016-05-12     57.3     126.3249  NA 21.69           True
## 6 4319703577 2016-04-17     72.4     159.6147  25 27.45           True
##          logid
## 1 1.462234e+12
## 2 1.462320e+12
## 3 1.460510e+12
## 4 1.461283e+12
## 5 1.463098e+12
## 6 1.460938e+12
```

Verification of Clean data

Data has been cleaned properly.

Merge two datasets first, then 1 after

```
dasw_combined <- merge(daily_activity,sleep_day, by=c("id","date"))
glimpse(dasw_combined)
```

```
## Rows: 410
## Columns: 18
## $ id                       <dbl> 1503960366, 1503960366, 1503960366, 15039
6036…
## $ date                     <date> 2016-04-12, 2016-04-13, 2016-04-15, 2016
-04-…
## $ totalsteps               <int> 13162, 10735, 9762, 12669, 9705, 15506, 1
0544…
## $ totaldistance            <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68,
6.3…
## $ trackerdistance          <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68,
6.3…
## $ loggedactivitiesdistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, …
## $ veryactivedistance       <dbl> 1.88, 1.57, 2.14, 2.71, 3.19, 3.53, 1.96,
1.3…
## $ moderatelyactivedistance <dbl> 0.55, 0.69, 1.26, 0.41, 0.78, 1.32, 0.48,
0.3…
## $ lightactivedistance      <dbl> 6.06, 4.71, 2.83, 5.04, 2.51, 5.03, 4.24,
4.6…
## $ sedentaryactivedistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, …
## $ veryactiveminutes        <int> 25, 21, 29, 36, 38, 50, 28, 19, 41, 39, 7
3, 3…
```

```
## $ fairlyactiveminutes      <int> 13, 19, 34, 10, 20, 31, 12, 8, 21, 5, 14,
23,…
## $ lightlyactiveminutes     <int> 328, 217, 209, 221, 164, 264, 205, 211, 2
62, …
## $ sedentaryminutes         <int> 728, 776, 726, 773, 539, 775, 818, 838, 7
32, …
## $ calories                 <int> 1985, 1797, 1745, 1863, 1728, 2035, 1786,
177…
## $ totalsleeprecords        <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, …
## $ totalminutesasleep       <int> 327, 384, 412, 340, 700, 304, 360, 325, 3
61, …
## $ totaltimeinbed           <int> 346, 407, 442, 367, 712, 320, 377, 364, 3
84, …
```

```
dasw_combined <- merge(dasw_combined,weight, by=c("id","date"))
glimpse(dasw_combined)
```

```
## Rows: 35
## Columns: 24
## $ id                       <dbl> 1503960366, 1503960366, 1927972279, 45586
0992…
## $ date                     <date> 2016-05-02, 2016-05-03, 2016-04-13, 2016
-05-…
## $ totalsteps               <int> 14727, 15103, 356, 3428, 12231, 10199, 56
52, …
## $ totaldistance            <dbl> 9.71, 9.66, 0.25, 2.27, 9.14, 6.74, 3.74,
1.0…
## $ trackerdistance          <dbl> 9.71, 9.66, 0.25, 2.27, 9.14, 6.74, 3.74,
1.0…
## $ loggedactivitiesdistance <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0
.000…
## $ veryactivedistance       <dbl> 3.21, 3.73, 0.00, 0.00, 5.98, 3.40, 0.57,
0.0…
## $ moderatelyactivedistance <dbl> 0.57, 1.05, 0.00, 0.00, 0.83, 0.83, 1.21,
0.0…
## $ lightactivedistance      <dbl> 5.92, 4.88, 0.25, 2.27, 2.32, 2.51, 1.96,
1.0…
## $ sedentaryactivedistance  <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.0…
## $ veryactiveminutes        <int> 41, 50, 0, 0, 200, 50, 8, 0, 0, 50, 5, 13
, 35…
## $ fairlyactiveminutes      <int> 15, 24, 0, 0, 37, 14, 24, 0, 0, 3, 13, 42
, 41…
## $ lightlyactiveminutes     <int> 277, 254, 32, 190, 159, 189, 142, 86, 217
, 28…
## $ sedentaryminutes         <int> 798, 816, 986, 1121, 525, 796, 548, 862,
837,…
## $ calories                 <int> 2004, 1990, 2151, 1692, 4552, 1994, 1718,
146…
```

```
## $ totalsleeprecords        <int> 1, 1, 1, 1, 1, 1, 3, 2, 1, 1, 1, 1, 1, 1,
1, …
## $ totalminutesasleep       <int> 277, 273, 398, 115, 549, 366, 630, 508, 3
70, …
## $ totaltimeinbed           <int> 309, 296, 422, 129, 583, 387, 679, 535, 3
86, …
## $ weightkg                 <dbl> 52.6, 52.6, 133.5, 69.9, 90.7, 62.5, 62.1
, 61…
## $ weightpounds             <dbl> 115.9631, 115.9631, 294.3171, 154.1031, 1
99.9…
## $ fat                      <int> 22, NA, NA, NA, NA, NA, NA, NA, NA, NA, N
A, N…
## $ bmi                      <dbl> 22.65, 22.65, 47.54, 27.32, 28.00, 24.39,
24.…
## $ ismanualreport           <chr> "True", "True", "False", "True", "False",
"Tr…
## $ logid                    <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12,
1.4…
```

View combined data

```
head(dasw_combined)
```

ANALYZE PHASE

It's now time to identify trends and relationship in our data. The cleaned and prepared data will now be used to bring out new insights that will help improve Bellabeat marketing strategy.

Count for the total number of people in the combined dataset

Before I start the analyses I will want to check if the fitbit members are equal to the 30 we were told about earlier.

```
n_distinct(dasw_combined)

## [1] 35
```

After checking, I have 35 unique members that I will be running my analyses on, which does not go with our 30 fitbit members.

Summary on activity

```
dasw_combined %>%
  select(totalsteps,totaldistance,sedentaryminutes,calories) %>%
  summary()

##    totalsteps    totaldistance    sedentaryminutes    calories
##  Min.   :  356   Min.   : 0.250   Min.   : 127.0   Min.   : 928
##  1st Qu.: 5780   1st Qu.: 3.825   1st Qu.: 635.5   1st Qu.:1852
##  Median :10524   Median : 6.960   Median : 689.0   Median :2039
##  Mean   : 9687   Mean   : 6.523   Mean   : 688.5   Mean   :2052
##  3rd Qu.:12484   3rd Qu.: 8.730   3rd Qu.: 736.0   3rd Qu.:2168
##  Max.   :20031   Max.   :13.240   Max.   :1121.0   Max.   :4552
```

Summary on sleeep

```
dasw_combined %>%
  select(totalsleeprecords,totalminutesasleep,totaltimeinbed,weightpounds) %>
%
  summary()
```

```
##  totalsleeprecords totalminutesasleep totaltimeinbed   weightpounds
##  Min.   :1.000     Min.   :115.0      Min.   :129.0   Min.   :116.0
##  1st Qu.:1.000     1st Qu.:399.0      1st Qu.:420.0   1st Qu.:134.9
##  Median :1.000     Median :442.0      Median :455.0   Median :135.6
##  Mean   :1.086     Mean   :430.3      Mean   :449.8   Mean   :141.5
##  3rd Qu.:1.000     3rd Qu.:472.5      3rd Qu.:494.0   3rd Qu.:136.5
##  Max.   :3.000     Max.   :630.0      Max.   :679.0   Max.   :294.3
```

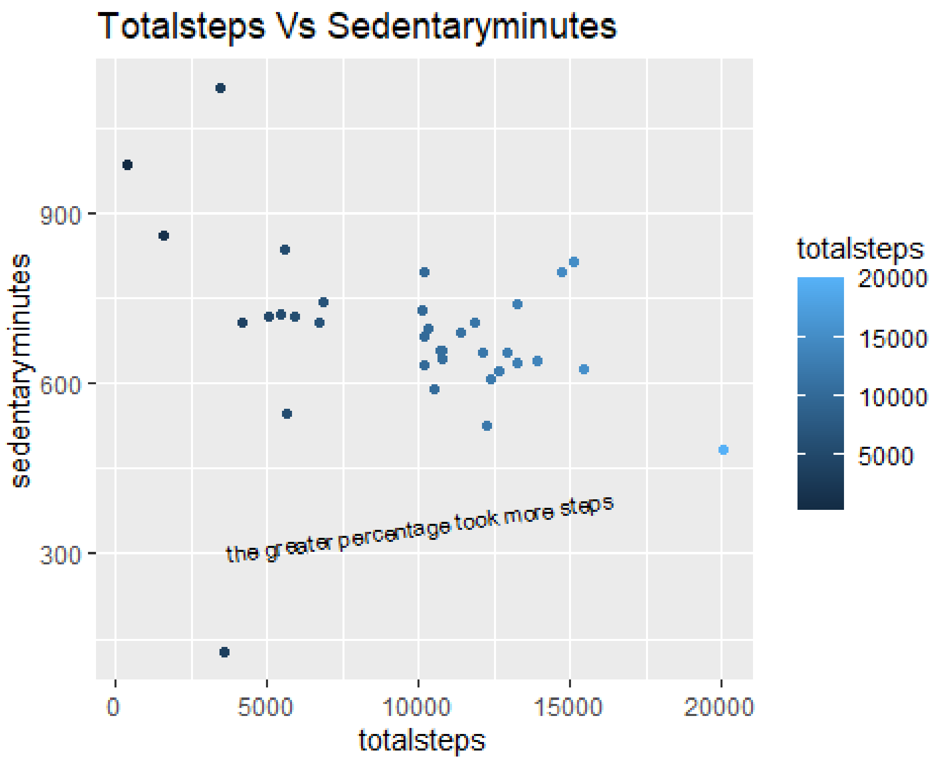
After running the summary, we have been provided with the minimum, maximum and average results of selected columns(totalsteps, totaldistance, sedentaryminutes, calories,totalsleeprecords, totalminutesasleep, totaltimeinbed and weightpounds. We will later plot this down and check if some of them correspond or correlate to each other.

SHARE

Now we will use visualization to draw out trends in our analyses and later share our finding with our stakeholders. I will be using gglot in R for my visualization.

Relatiobnship b/n Totalsteps and Sedentaryminutes

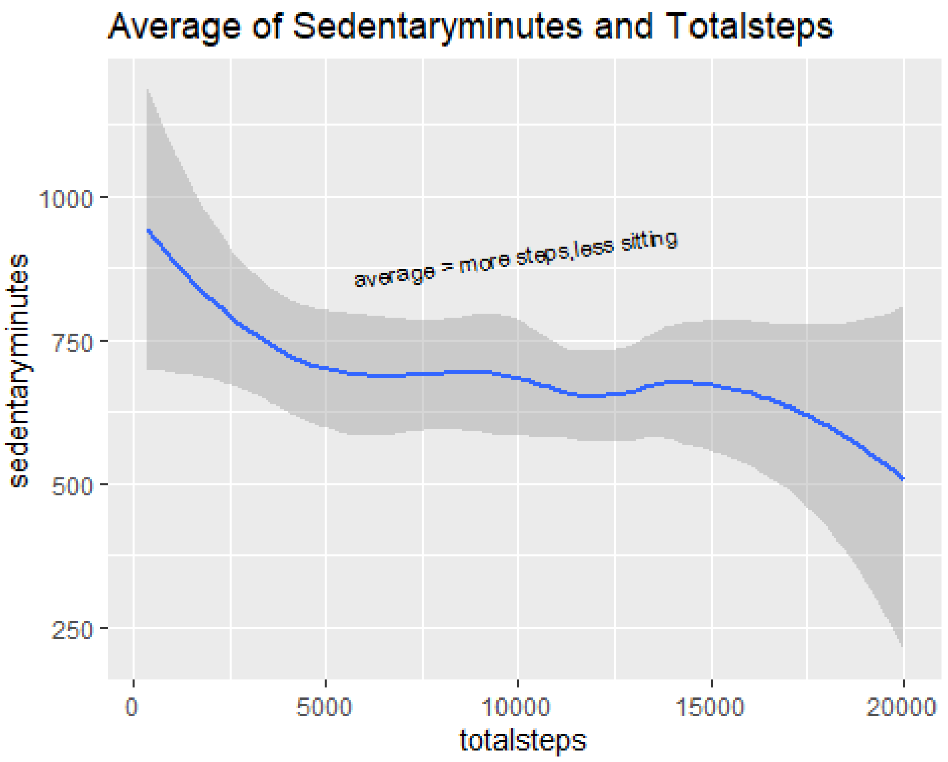
```
ggplot(data = dasw_combined)+
  geom_point(mapping = aes(x=totalsteps, y=sedentaryminutes,color=totalsteps)
)+labs(title = "Totalsteps Vs Sedentaryminutes")+annotate("text",x=10000,y=35
0,label="the greater percentage took more steps",size=3,angle=8)
```



Based on the scatter plot, the total steps is more than the sedentary minutes. You also realize that the higher the total steps the lower the sedentary minutes and vice-versa, if the sedentary minutes is higher the total steps is less. The higher percentage fall in the middle of the plot where sedentary minutes and total steps are balanced. More steps were taken compared to sedentary minutes. Lets get the average of both sides to prove our point.

Average of Sedentary and totalsteps

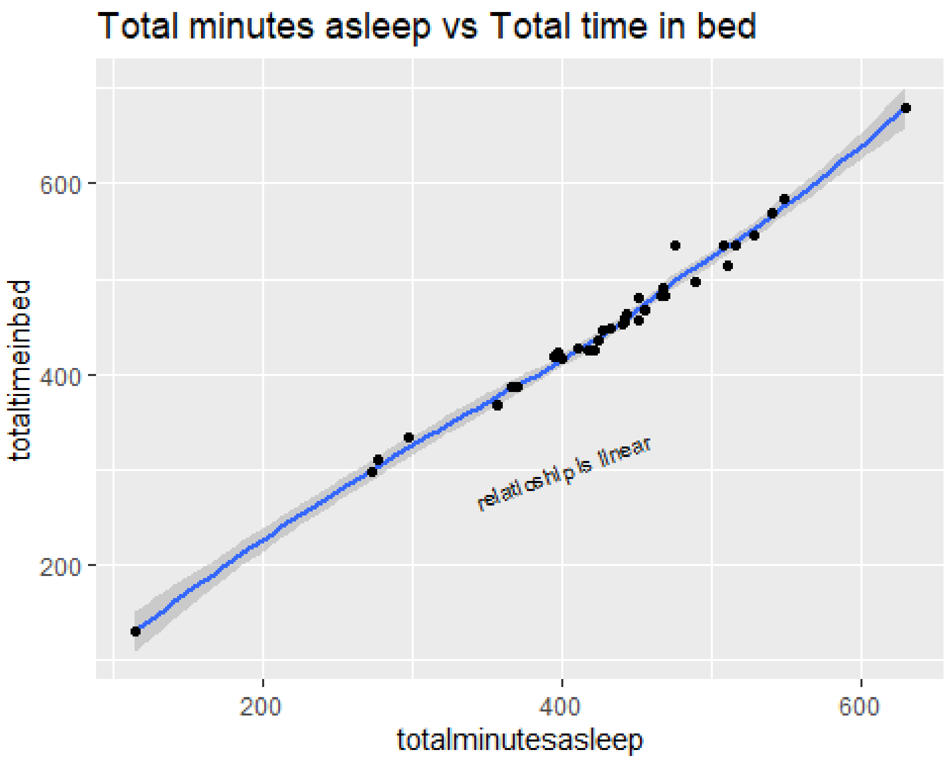
```
ggplot(data = dasw_combined)+
  geom_smooth(mapping=aes(x=totalsteps,y=sedentaryminutes))+labs(title = "Ave
rage of Sedentaryminutes and Totalsteps")+annotate("text",x=10000,y=900,label
="average = more steps,less sitting ",size=3,angle=8,)
```



Based on our average, we can now inform our customer segments on behaviour and physiological ways.

Relatiobnship b/n Totalminutesasleep vs Totaltimeinbed

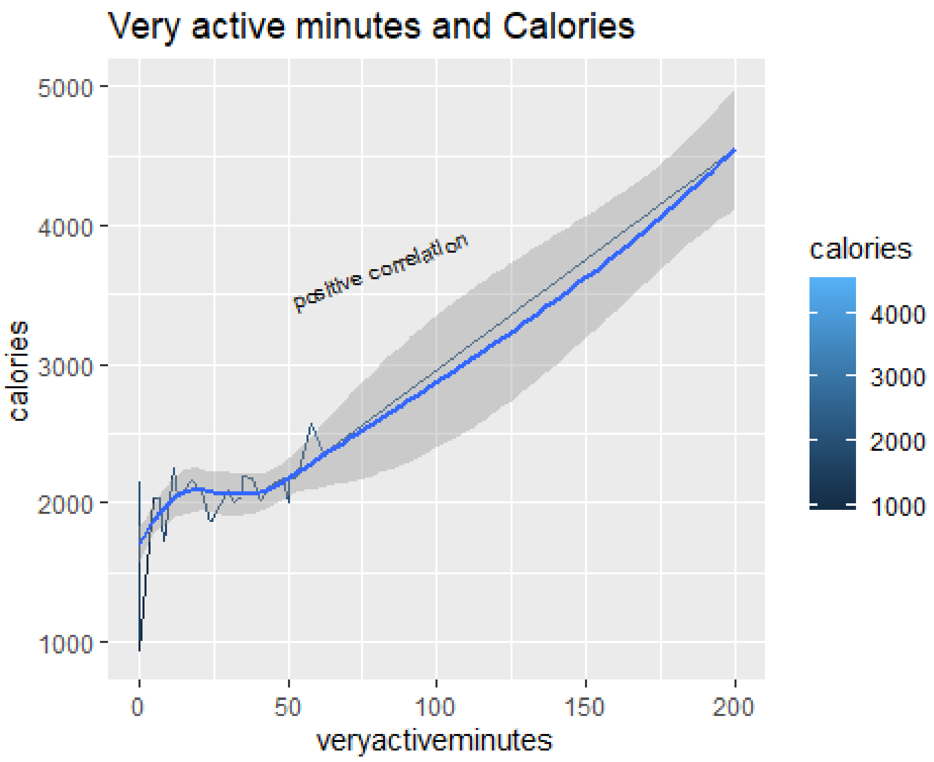
```
ggplot(data = dasw_combined)+
  geom_smooth(mapping = aes(x=totalminutesasleep, y=totaltimeinbed))+ geom_po
int(mapping = aes(x=totalminutesasleep, y=totaltimeinbed))+ labs(title = "Tot
al minutes asleep vs Total time in bed")+annotate("text",x=400,y=300,label="r
elatioship is linear",size=3,angle=20,)
```



The relationship between total minutes asleep and total time in bed is linear. This shows that people sleep at the right time and do not participate in any other activities.

Relatiobnship b/n Veryactiveminutes and Calories

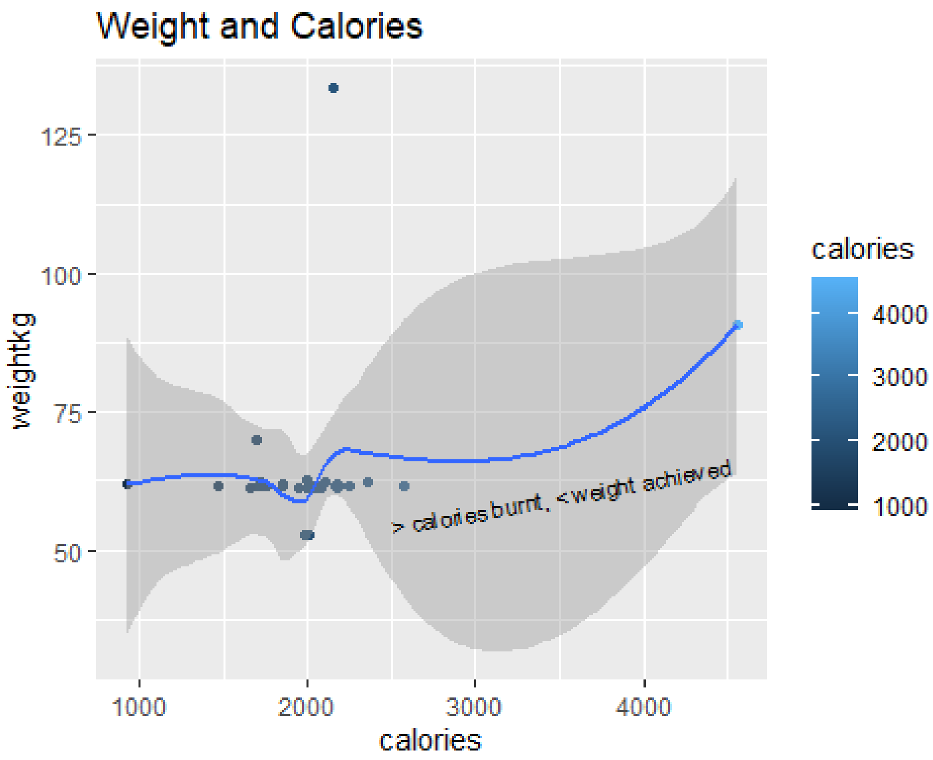
```
ggplot(data = dasw_combined)+
  geom_line(mapping=aes(x=veryactiveminutes,y=calories,color=calories))+geom_
smooth(mapping=aes(x=veryactiveminutes,y=calories))+annotate("text",x=80,y=37
00,label="positive correlation",size=3,angle=20)+labs(title="Very active minu
tes and Calories")
```



Based on the correlation between veryactiveminutes and calories, we can conclude that the more active a person is, the higher chance of burning calories. This insight will be crossed check with calories and weight to confirm if active minutes affecting calories then it should also affect the the weight of a person.

Relatiobnship b/n Weight and Calories

```
ggplot(data = dasw_combined)+
  geom_point(mapping=aes(x=calories,y=weightkg,color=calories))+geom_smooth(mapping=aes(x=calories,y=weightkg))+labs(title="Weight and Calories")+annotate("text",x=3500,y=60,label="> calories burnt, < weight achieved",size=3,angle=10)
```



This plot explains the theory behind more calories being burnt, weight loss being achieved. We noticed from the earlier plot that, active minutes leads to burning of calories and burning of calories also leads to loss of weight. This is an insight that will benefit the business plan of the Bellabeat Corporation.

ACT

Observation, insights and trends have been drawn out of our analyze phase with calculations in summary(looking at our maximum, minimum and average) of a few selected columns in the dataset. Graphs were also created in the share phase to throw more light on relationships between data.

Some few recommendations have been made below to answer our business task: Analyze smart devices and identify trends and insights to be applied to Bellabeat marketting strategies to unlock new growth opporunities.

Recommendations:

- Daily, weekly and monthly update on steps taken to keep encouraging those who hit their daily mark steps and set new goals for them.
- Identify and remind those within the sedentary portion on the benefits of taking walks and recommend ways on taking daily steps.e.g. taking dogs for walks, parking further away and using the stairs more.

- Reward systems can also be created to boost to reward those who continuously hit their daily steps mark.

- Bellabeat can use the analyses under sleep to improve peoples sleep time by sending notifications when time for sleep is up and set off alarms to wake them up after a specified sleep time.

- Being active leads to burning calories and Bellabeat can use this information to set up programs for its members such jogging, community athletics programs, sports and etc, and after update its participants on how many calories they have burnt in the few hours spent.

- Lastly based on insight and trend confirmed on the fact that active minutes affects burning of calories then to loss of weight. Bellabeat can use this information to inform its members on their loss of weight and gain of weight periodically and this can help members to check themselves properly.

END OF ANALYZING BELLABEAT MARKETING ANALYSIS CASE STUDY