Ehtesham Hanifa

Bellabeat Case Study

Bellabeat is a wellness company founded in 2014. It was founded by CEO Urška Sršen and Sandro Mur and it's main focus is on the health and wellbeing of women. This involves using wearables that collects data on activity, stress, sleep, steps, calories burned, heartrate and many others. This data provides the user with the knowledge of their own heath. Since the start of Bellabeat's journey, the company has grown and has become one of the main tech-driven wellness company for women.

The CEO 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 analyse smart device usage data in order to gain insight into how people are already using their smart devices. Then, using this information, she would like high-level recommendations for how these trends can inform Bellabeat marketing strategy and this tell the marketing team which products to push more.

The business task for this case study is to analyse the data available, see which features the users interact with more to determine trends and this will help the marketing team push the products forward.

Step 1: ASK

The first stage in this case study is to ask the right questions. What are some trends in smart device usage? How could these trends apply to Bellabeat customers? How could these trends help influence Bellabeat marketing strategy?

The stakeholders who I will present the findings to will be the Urška Sršen the CEO, Sando Mur the co founder and the marketing department.

Step 2: PREPARE

The data is from for this project is free data accessible to the public from Kaggle and it is generated by the responders to a survey by Amazon Mechanical Turk. The dataset contains personal fitness data from thirty Fitbit users who have given consent to share their personal fitness tracker data. This includes daily activity, steps, calories, walking distance, heart rate, sleep and many other.

Data is stored in a file specifically made for the case study. It was created in a CVS file and it is wide format. Each ID has data in multiple rows.

There is bias because we have such small dataset from 30 users. If we had data from more users, it would've been less bias because the larger dataset would've been more representative of the general population. Also, the absence of gender and age will have an effect since Bellabeat is a fitness company directed towards women.

The licensing, privacy, security and accessibility came from a public domain. That domain mentions I can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

The data has issues with integrity and bias because we were told via Zenodo.org that 30 users consented to tracking of their data but it shows we have 33 users. Using the unique and count function, I was able to determine this.

Using the ROCCC system to determine the reliability, originality, comprehensiveness, current and citation of the data.

- Reliability – LOW. 30 users have responded. It is a small sample size and doesn't accurately reflect the entire population of the Fitbit users who are women.
- Original – LOW. Amazon Mechanical Turk is a third party data provider so the data being used isn't original.
- Comprehensive – LOW. The dataset doesn't mention anything about gender, age, any underlying health issues.
- Current – LOW. The data is from 2016.
- Cited – LOW. Data obtained from a third party.

Step 3: PROCESS

The datasets used were dailyActivity_merged, sleepDay_merged, average, hourly_activity_cleaned and heartrate_seconds_merged. I used the programming language R on Posit Cloud and also used Google Sheets for data cleaning, manipulation, analysis, and visualization.

I first installed and loaded the packages I will use in R:

```
install.packages("tidyverse")
install.packages("skimr")
install.packages("here")
install.packages("janitor")
install.packages("ggplot2")
install.packages("lubridate")
install.packages("dplyr")
install.packages("sqldf")
install.packages("plotrix")
library(tidyverse)
library(skimr)
library(here)
library(janitor)
library(ggplot2)
library(lubridate)
library(dplyr)
library(sqldf)
library(plotrix)
```

Then, I imported the datasets:

```
> averages <- read.csv("average.csv")
> houractivity <- read.csv("hourly_activity_cleaned.csv")
> sleepday <- read.csv("sleepDay_merged_cleaned.csv")
> heartrate <- read.csv("heartrate_seconds_merged.csv")
> dailyactivity <- read.csv("dailyActivity_merged.csv")
> sleep_day <- read.csv("sleepDay_merged.csv")
> weight_info_log <- read.csv("weightLogInfo_merged.csv")
```

I then counted the number of rows in the dailyactivity, sleep_day and weight_info_log dataset. Daily_activity has 940 rows of data, sleep_day has 413 and weight_info_log has 67.

```
> nrow(daily_activity)
[1] 940
> nrow(sleep_day)
[1] 413
> nrow(weight_info_log)
[1] 67
```

Ehtesham Hanifa

I checked to see if there's any duplicate rows.

```
> nrow(daily_activity[duplicated(daily_activity),])
[1] 0
> nrow(sleep_day[duplicated(sleep_day),])
[1] 3
> nrow(weight_info_log[duplicated(weight_info_log),])
[1] 0
```

Sleep_day contained 3 duplicate rows. These duplicate rows were removed and we are left with 410 rows instead of 413.

```
> sleep_day <- unique(sleep_day)
> nrow(sleep_day)
[1] 410
```

For the sleep_day and weight_info_log, I split the date and time into separate column. This was done with the separate() function.

```
> sleep_day_new <- sleep_day %>%
+ separate(col = 2, into = c("Date", "Time"), sep = " ")
Warning message:
Expected 2 pieces. Additional pieces discarded in 410 rows [1, 2, 3, 4,
5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
> weight_info_new <- weight_info_log %>%
+ separate(col = 2, into = c("Date", "Time"), sep = " ")
Warning message:
Expected 2 pieces. Additional pieces discarded in 67 rows [1, 2, 3, 4, 5,
6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

The data type was changed from char to the date format.

```
> daily_activity$ActivityDate = as.Date(daily_activity$ActivityDate, "%m/%d/%
Y")
> sleep_day_new$Date = as.Date(sleep_day_new$Date, "%m/%d/%Y")
> View(sleep_day_new)
> weight_info_new$Date = as.Date(weight_info_new$Date,"%m/%d/%Y")
> View(weight_info_new)
```

Then using the distinct function, I was able to determine the number of unique users for these datasets. This shows that 33 users were unique in the daily_activity data, 24 for sleep_day and 8 for weight_info.

```
> n_distinct(daily_activity$Id)
[1] 33
> n_distinct(sleep_day_new$Id)
[1] 24
> n_distinct(weight_info_new$Id)
[1] 8
```

Step 4: ANALYSE

It was time to begin analysing the dataset to find trends and the best way to start analysis is to find the mean, medium, max and the quartiles of the dataset and I did that for all three datasets.

```
> daily_activity %>%
+ select(TotalSteps, TotalDistance, VeryActiveMinutes, FairlyActiveMinutes, Lig
htlyActiveMinutes, SedentaryMinutes, Calories) %>%
+ summary()
   TotalSteps    TotalDistance    VeryActiveMinutes FairlyActiveMinutes
 Min.   :    0   Min.   : 0.000   Min.   :  0.00    Min.   :  0.00
 1st Qu.: 3790   1st Qu.: 2.620   1st Qu.:  0.00    1st Qu.:  0.00
 Median : 7406   Median : 5.245   Median :  4.00    Median :  6.00
 Mean   : 7638   Mean   : 5.490   Mean   : 21.16    Mean   : 13.56
 3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.: 32.00    3rd Qu.: 19.00
 Max.   :36019   Max.   :28.030   Max.   :210.00    Max.   :143.00
 LightlyActiveMinutes SedentaryMinutes    Calories
 Min.   :  0.0        Min.   :   0.0   Min.   :   0
 1st Qu.:127.0        1st Qu.: 729.8   1st Qu.:1828
 Median :199.0        Median :1057.5   Median :2134
 Mean   :192.8        Mean   : 991.2   Mean   :2304
 3rd Qu.:264.0        3rd Qu.:1229.5   3rd Qu.:2793
 Max.   :518.0        Max.   :1440.0   Max.   :4900
> sleep_day_new %>%
+ select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%
+ summary()
 TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
 Min.   :1.00      Min.   : 58.0      Min.   : 61.0
 1st Qu.:1.00      1st Qu.:361.0      1st Qu.:403.8
 Median :1.00      Median :432.5      Median :463.0
 Mean   :1.12      Mean   :419.2      Mean   :458.5
 3rd Qu.:1.00      3rd Qu.:490.0      3rd Qu.:526.0
 Max.   :3.00      Max.   :796.0      Max.   :961.0
> weight_info_new %>%
+ select(WeightKg, WeightPounds, BMI) %>%
+ summary()
    WeightKg       WeightPounds        BMI
 Min.   : 52.60   Min.   :116.0   Min.   :21.45
 1st Qu.: 61.40   1st Qu.:135.4   1st Qu.:23.96
 Median : 62.50   Median :137.8   Median :24.39
 Mean   : 72.04   Mean   :158.8   Mean   :25.19
 3rd Qu.: 85.05   3rd Qu.:187.5   3rd Qu.:25.56
 Max.   :133.50   Max.   :294.3   Max.   :47.54
```

Observations for the daily activity dataset:

- The average Fitbit user walked 5.490 km and did on average, 7638 steps per day. The Centers for Disease Control and Prevention (CDC) recommends adults to do at least 10000 steps per day, which is around 8km or 5 miles. This is to improve muscle strength, range on motion, blood flow, breathing, keeping the heart healthy and other benefits.
- The average user spent 991 minutes being inactive within 24 hours. This is over 16.5 hours of being inactive. This can increase the chances of diseases such as high blood

pressure and high cholesterol. This can also increase weight and size due to less calories being burned.

- The average calories burned was 2304 calories. The minimum calories an adult should use per day varies on gender, age, weight, height and activity. We don't have information on gender, age and height so we can't determine if that the users are hitting their calorie goals.
- The average user were highly active for 21.16 minutes. The recommended amount of high active minutes is 30 minutes per day so this is a lot less than the recommended amount.
- The average user was fairly active for 13.56 minutes and the World's Health Organization (WHO) recommends at least 60 minutes a day of moderate exercise.

Observation for the sleep dataset:

- The average user slept for 419 minutes, which is just under 7 hours. The National Institution of Health (NIH) said the daily recommended sleep amount should be at least 7 hours.
- The average user spent 458 minutes in bed, which is around 7 hours and 40 minutes in bed.

Observation for the weight log dataset:

- The average user weighs 72kg.
- The average BMI is 25.19. According to the CDC, the average BMI of an adult should be between 18.5 to 24.9 so the average BMI of the users is slightly higher and is considered overweight.

Conclusion:

- The average user is doing less steps per day than the recommended amount.
- The average user is spending too much time being inactive throughout the day.
- The average user burns 2304 calories per day.
- The average user isn't being active enough.
- The average user is getting the right amount of sleep.
- The average user spends 40 minutes in bed before going to sleep.
- The average weight of users 72kg and the average weight of adults varies depending on height and age.
- the average BMI is slightly higher but the average BMI also varies depending on height, muscle mass and body mass, which is not supplied in the dataset. Also, only 8 users shared their weight data and that is not enough to come to a conclusion.

Step 5: SHARE

The days the users used Fitbit the most:

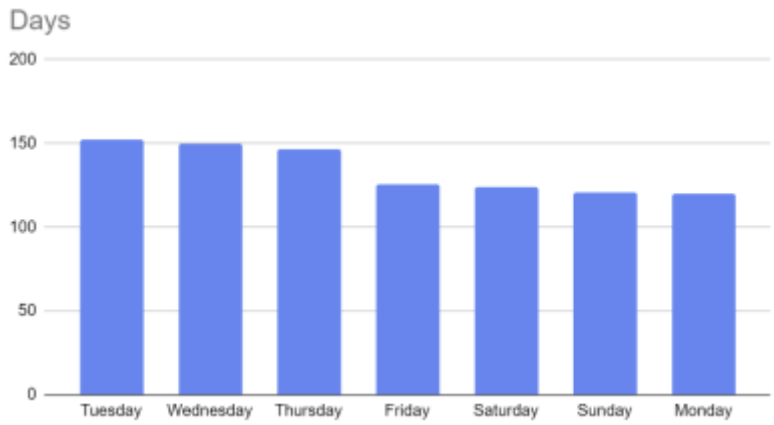
On Google Sheets, I imported the daily activity and worked out which days the users were most active.

I used this formula to get the days from the date and used the graph tool to draw a graph:

Ehtesham Hanifa

=switch(WEEKDAY(B2),1,"Sunday",2,"Monday",3,"Tuesday",4, "Wednesday",5, "Thursday",6,"Friday",7,"Saturday")

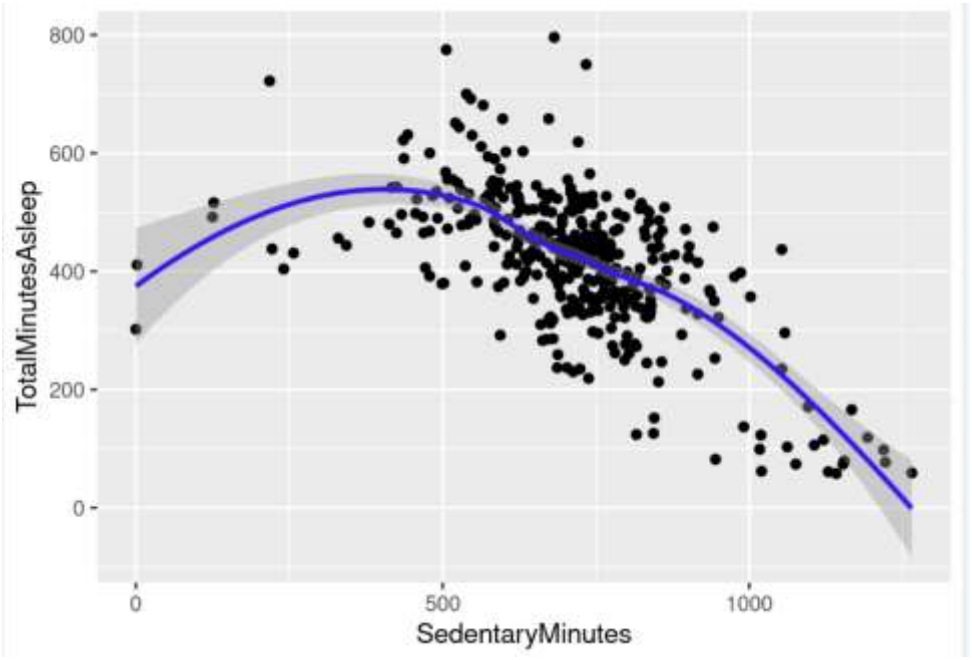
Id	ActivityDate	day
1503960366	4/12/2016	Tuesday
1503960366	4/13/2016	Wednesday
1503960366	4/14/2016	Thursday
1503960366	4/15/2016	Friday
1503960366	4/16/2016	Saturday



The users were most active on Tuesday, Wednesday and Thursday. They were least active on Monday and that could be because it is the first day of the working week and after work, they just want to go home and relax. Friday and Saturday is lower than Tuesday, Wednesday and Thursday because Friday is the last day of the working week and people usually go out after work. Sunday is low because that is the last day before the new working week starts and people usually use Sunday to get ready for work on Monday.

Sleep Vs Sedentary time

```
> ggplot(data = sleep_sedentary_cor) +
+ geom_point(mapping = aes(x = SedentaryMinutes, y = TotalMinutesAsleep)) +
+ geom_smooth(mapping = aes(x = SedentaryMinutes, y = TotalMinutesAsleep), meth
od = 'loess', formula = y ~ x, color = "blue")
```



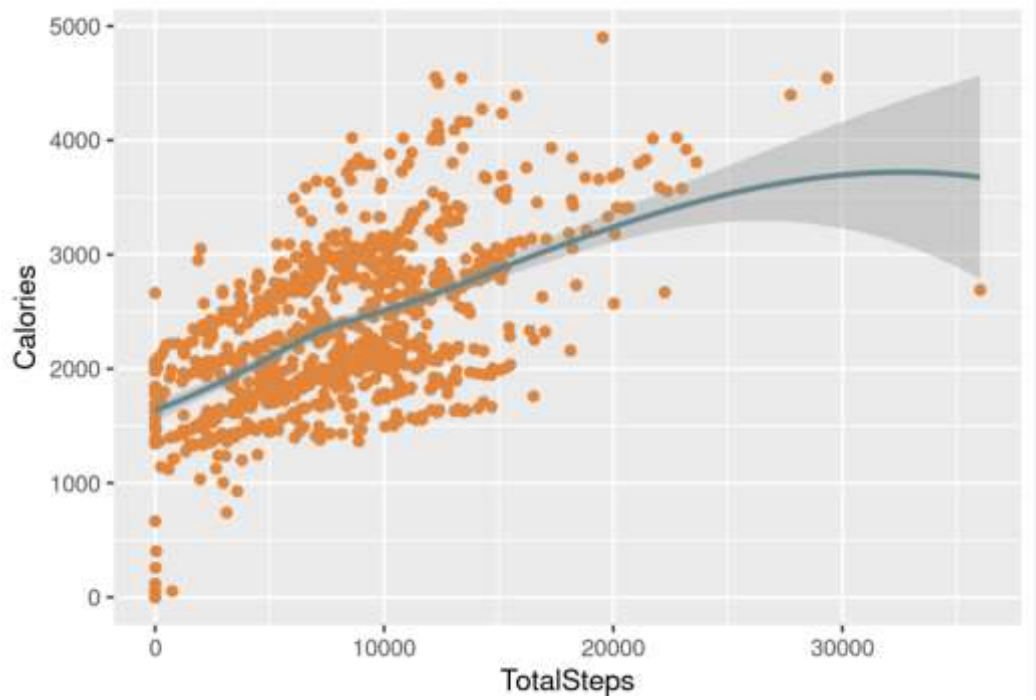
This shows the correlation between total minutes asleep and sedentary minutes. Between 0 and 500 minutes, the total minutes asleep increases but from 500 to 1500 minutes, there is a massive decrease in sleep.

I also calculated the correlation index between total minutes asleep and sedentary minutes and this proves that there is a negative correlation (-0.6) because as the sendetary minutes increase, the total minutes of sleep decreases.

```
> cor(sleep_sedentary_cor$SedentaryMinutes, sleep_sedentary_cor$TotalMinutesAsl
eep)
[1] -0.6010731
```

Total steps Vs Calories burned

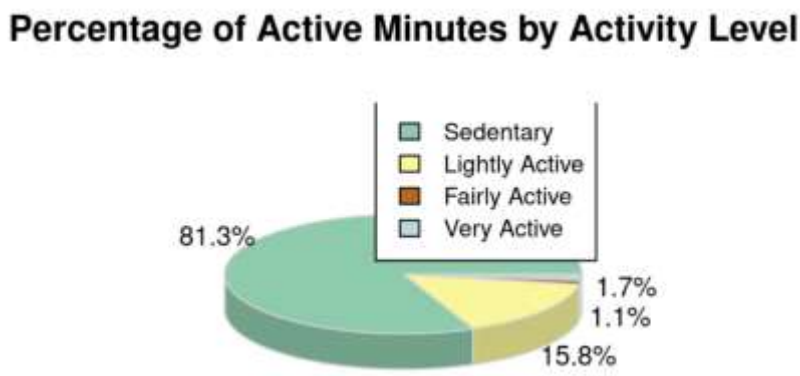
```
> ggplot(data= daily_activity, aes(x=TotalSteps, y=Calories)) +
+ geom_point(color="Chocolate1") +
+ geom_smooth(color = "cadetblue4")
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This graph shows that as the total number of steps increases, so does the calories burned. This makes perfect sense because walking requires calories so the more the user walks, the more calories they burn. This is a positive correlation.

Percentage of activity levels

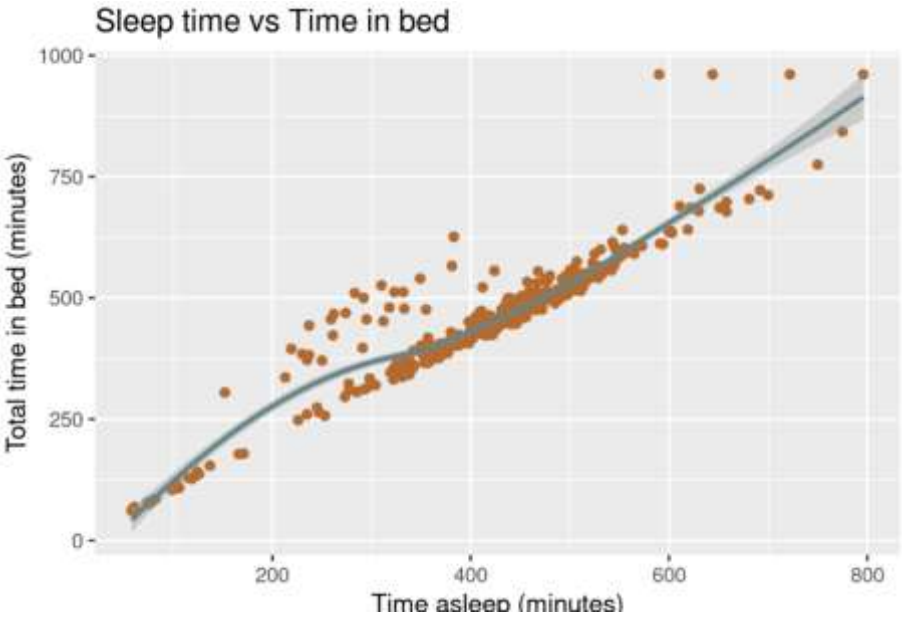
```
> Sedentary <- sum(daily_activity$SedentaryMinutes)
> Lightly <- sum(daily_activity $LightlyActiveMinutes)
> Fairly <- sum(daily_activity $FairlyActiveMinutes)
> Active <- sum(daily_activity $VeryActiveMinutes)
> activity_minutes <- c(Sedentary,Lightly, Fairly, Active)
> activity_percent <- round(activity_minutes/sum(activity_minutes)*100,1)
> legend_labels <- c("Sedentary","Lightly Active", "Fairly Active","Very Activ
e")
> pie3D(activity_percent, labels=paste0(activity_percent,"%"), main="Percentage
of Active Minutes by Activity Level", col=c("aquamarine3","khaki1", "darkorange
3","lightblue"), border="lightgrey", labelcex = 0.9)
> legend("topright", legend_labels, cex=0.8, fill=c("aquamarine3","khaki1", "da
rkorange3","lightblue"))
```



This pie chart shows the percentage of activity levels for the users. 81.3% of the day, the user is in sedentary. This could be because of work where they are sitting at a desk and hardly moving. It could also be lifestyle choices such as sitting down and watching a TV series or eating. 15.8% is lightly active, which could be travelling to and from work. Fairly active (1.1%) and very active (1.7%) accounts for the remaining amount of active minutes.

Sleep Vs Time in bed

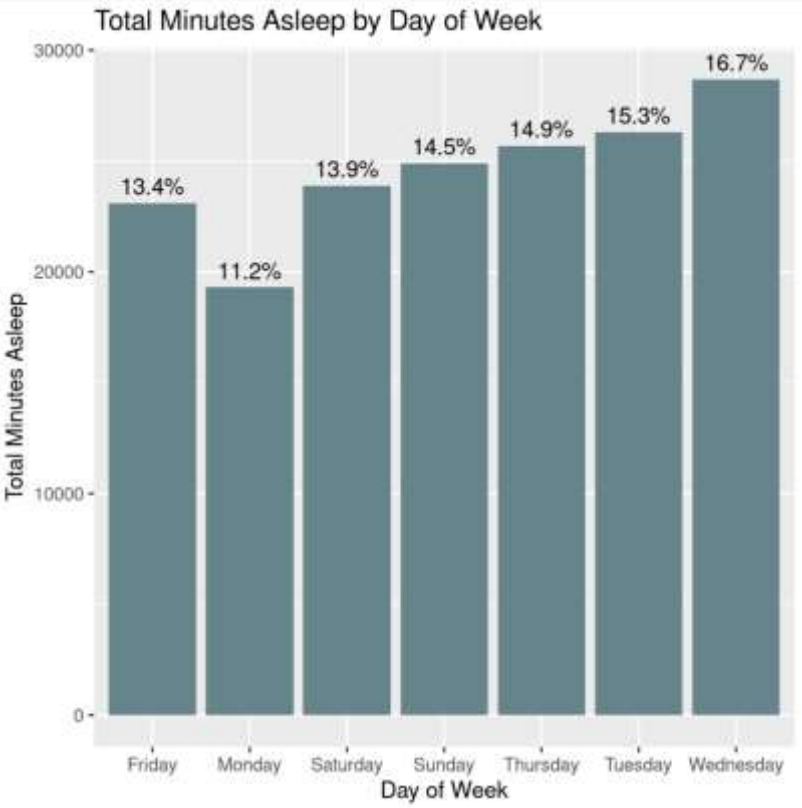
```
> ggplot(data= sleep_day_new, aes(x= TotalMinutesAsleep, y=TotalTimeInBed)) +
+     geom_point(color="Chocolate3") +
+     geom_smooth(color = "cadetblue4")+
+     labs(title="Sleep time vs Time in bed ", x="Time asleep (minutes)", y="To
tal time in bed (minutes)")
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The scatterplot shows the correlation between sleep time and the time spent in bed. It shows a positive correlation because the longer someone is asleep, the more time they spend in bed.

Total minutes asleep by day

```
> sleep_day_new$DayOfWeek <- weekdays(sleep_day_new$Date)
> View(sleep_day_new)
>
> sleep_day_new %>%
+     group_by(DayOfWeek) %>%
+     summarize(total_minutes_asleep = sum(TotalMinutesAsleep)) %>%
+     arrange(desc(total_minutes_asleep)) %>%
+     head(1)
# A tibble: 1 × 2
  DayOfWeek total_minutes_asleep
  <chr>                    <int>
> sleep_day_new %>%
+     group_by(DayOfWeek) %>%
+     summarize(total_minutes_asleep = sum(TotalMinutesAsleep)) %>%
+     mutate(percent_of_total = round(total_minutes_asleep / sum(total_minutes_
asleep) * 100, 1)) %>%
+     ggplot(aes(x = DayOfWeek, y = total_minutes_asleep)) +
+     geom_col(fill = "cadetblue4") +
+     geom_text(aes(label = paste0(percent_of_total, "%")), vjust = -0.5) +
+     labs(x = "Day of Week", y = "Total Minutes Asleep",
+          title = "Total Minutes Asleep by Day of Week")
```



The graph above shows the total amount of sleep per day and it shows that Wednesday is the day where the users sleep the most and Monday has the least amount of sleep per day.

Ehtesham Hanifa

Step 6: ACT

This is the last step in the case study and here, I will provide the observations found with the data and suggestions on how Bellabeat can market their products to help the user achieve their fitness and health goals.

One of the main questions for this case study is “What are some trends in smart device usage?”. Here are the findings:

1. More people used Fitbit to track their daily activities, less people used it to track their sleep and even fewer people used it to track their weight.
2. The average user spent around 16.5 hours inactive a day.
3. The average user did 7638 steps per day instead of the recommended 10,000 steps and walked approximately 5.5km instead of the recommended 8km.
4. The average user burned 2304 calories per day. The recommended amount varies depending on age, gender, height, weight and activity levels.
5. The average user waw active for 21.16 minute per day instead of the recommended 30 minutes.
6. The average user slept on average 7 hours every night, which is within the recommended range.
7. The negative correlation between total sleep time and time in bed may suggest the quality of sleep is low.
8. The average user weighs 72kg and the recommended amount varies depending on height, age, weight, gender and other factors.
9. The average user’s BMI was 25.19, which is considered overweight.
10. Users were most active on Tuesday and least active on Monday.
11. There is a very strong correlation between total steps and calories burned.
12. Users spent 81.3% of the time in sedentary.
13. Users slept the most on Wednesday and the slept the least on Monday.

The other questions for this case study is “How could these trends apply to Bellabeat customers? How could these trends help influence Bellabeat marketing strategy?”. This can be done by focusing on the needs and behavior patterns of the users. Bellabeat can tailor its marketing strategy to focus more on the needs of the users and help them achieve their fitness goals.

Here are suggestions on how Bellabeat can achieve that:

1. Given that the users spend 81.3% in sedentary, Bellabeat should focus on promoting a healthier lifestyle. Bellabeat can send notifications to the user’s Bellabeat device telling them to move around when they are inactive for too long.
2. Encourage the users by setting targets for them. These targets could be step targets, calorie targets and sleep targets.
3. Making these targets into a challenge with other users can also increase the chances of the users hitting those targets and at the end of the week, reward the user with the most steps with badges.
4. For weight control, Bellabeat should focus on features that can help users track their calories such as a food diary. The food diary will let the user know the nutritional information about the food they are eating.

5. A social feature can also be integrated into Bellabeat app where users can share their healthy foods and dishes for others to try.
6. An AI type feature can also be integrated into the app. This feature will ask the user questions and their main goals and the AI feature will come up with meals that will allow the user to reach those calorie targets and weight goals.
7. Since sleep is a major factor in health and wellbeing, Bellabeat can focus on promoting sleep. This can be done by sending notifications to the user telling them it is time to go to sleep to get the recommended amount and this can help promote a regular sleeping pattern.
8. Bellabeat can monitor the users sleep, giving them a detail analysis on their sleep and ways to improve their sleep.