Case Study : How Can a Wellness Technology Company Play It Smart? Documentation

❖ Ask Phase

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

The business task is to analyze smart device usage data in order to

- gain insight into how consumers use non-Bellabeat smart devices.
- selecting one Bellabeat product to apply these insights to in your presentation.

❖ Prepare

dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences. (CC0: Public Domain, dataset made available through **Mobius**)

the data form is long format, as it not refers to a structured representation of data where each row represents a unique observation with variables spread across multiple columns, the data is Reliable, Original, Comprehensive, Current and Cited. The collected data contains values within the time range of April and May 2016.

We checked Data integrity and found it complete, accurate, consistent and safe throughout its entire lifecycle in the following ways. And no problems found in data. Its consist of 18 separate files.

❖ Process

During the process phase, using spreadsheet and SQl language, We ensured Data Integrity and Cleaned / Transformed Data. Handled Missing Values by deleting the null values, Checked/ Converted data types and Standardized data formats to ensure consistency.

Then we install and load the necessary packages required for this process to continue using R language which would be: **Tidyverse, Janitor, Lubridate & Skimr**

```
#Installing the packages
install.packages('tidyverse')
install.packages('janitor')
install.packages('lubridate')
install.packages('skimr')#Loading the packages
library(tidyverse)
library(janitor)
library(lubridate)
library(skimr)
```

After this, we would need to import the datasets into RStudio using **read.csv().** I will also be making slight name changes as well

```
#df_name <- read.csv(dataset_location)
daily_activity <- read.csv("Data/dailyActivity_merged.csv")
daily_sleep <- read.csv("Data/sleepDay_merged.csv")
weight_log <- read.csv("Data/weightLogInfo_merged.csv")
```

Inspecting our data to see if there are any errors with formatting by using **str()**

```
#str(dataframe_name)
str(daily_activity)
str(daily_sleep)
str(weight_log)
```

After a brief view of the output, there are a few issues that we need to address:

- The naming of the column names (camelCase)
- daily_activity$ActivityDate — Is formatted as CHR not as a date format
- daily_sleep$SleepDay — Is formatted as CHR not as a date format
- weight_log$Date — Is formatted as CHR not as a date format
- weight_log$IsManualReport is formated as CHR not logical (for boolean values)

To clean the column names, we would use clean_names()

```
daily_activity <- clean_names(daily_activity)
daily_sleep <- clean_names(daily_sleep)
weight_log <- clean_names(weight_log)
```

then format daily_activity$ActivityDate, daily_sleep$SleepDay, weight_log$Date into the proper date format. using **as.Date() & as.POSIXct()**

For weight_log$date, it's a little tricky because if you look closely, there's the PM indicator at the end. POSIX.ct does not recognize this and will return all values as NA, so we will need to use **parse_date_time** from Lubridate.

And to format weight_log$is_manual_report to a logical format, we will use **as.logical()**

After a quick look at our current data, let's add a day of the week, sedentary hours & total active hours column for further analysis in daily_activity. I will not be adding a month column since the dataset only provides information collected within a month.

Let's also add new columns which convert the current minutes of collection to hours and round it using **round()** in daily_sleep. I will also be adding a column to indicate the time taken to fall asleep in daily_sleep as well.

We will also be removing weight_log$fat, as it has little to no context and would not be helpful during the analysis phase by using **select(-c())**

Lastly, I will also be adding a new column in weight_log called bmi2 which will indicate whether the user is underweight, healthy, or overweight by using a line of code I recently learned about which is **case_when**!

-)) **Analysis & Findings**

let's remove rows in which the total_active_hours & calories burned are 0. The reasoning behind this is that we're using data collected from Fitbits, which are wearables. If they don't wear their smart devices it doesn't collect information, hence we will remove the clutter from the data frame. Users might have also disabled GPS/accelerometer functions that allow for the collection of steps taken.

```
#In laymans term, '!' means is not equals to
daily_activity_cleaned <-
daily_activity[!(daily_activity$calories<=0),]daily_activity_cleaned <-
daily_activity_cleaned[!(daily_activity_cleaned$total_active_hours<=0.00),]
```

If you're using an external visualization tool such as Tableau or PowerBI, we need to export our dataframe using

```
write.csv(daily_activity_cleaned, file ='fitbit_daily_activity.csv')
write.csv(daily_sleep, file = 'fitbit_sleep_log.csv')
write.csv(weight_log, file = 'fitbit_weight_log.csv')
```

I will be using ggplot for this section of the analysis phase. I will also be including another section in which I used Tableau instead.

As per usual, let's revisit our business task to ensure we are not plotting or trying to hypothesize information/relationships which will not help in solving the business task which are:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers
3. How could these trends help influence Bellabeat marketing strategy?

After having a brief view of the current data, I will be plotting a few observations revolving around:

1. The average: Steps taken, sedentary hours, very active minutes & total hours asleep.
2. Which days are users the most active.
3. The relationship between total active hours, total steps taken, and sedentary hours against calories burned.
4. The relationship between weight, total active hours & steps taken
5. The number of overweight users

Let's have a quick look at the average steps taken, sedentary hours, very active minutes & total hours of sleep using **summary()**.

```
> summary(daily_activity_cleaned$total_steps)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0 4920 8053 8319 11100 36019

> summary(daily_activity_cleaned$sedentary_hours)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 12.02 17.00 15.87 19.80 23.98

> summary(daily_activity_cleaned$very_active_minutes)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 0.00 7.00 23.21 36.00 210.00

> summary(daily_sleep$hours_asleep)
Min. 1st Qu. Median Mean 3rd Qu. Max.
    0.970 6.020 7.220 6.992 8.170 13.270
```

- The average number of steps per day were 8319, which is within the 6000–8000 recommended steps per day, however, 25% of people do not hit that recommended quota.
- The average sedentary hours were 15.87 hours, which is absurdly high, shattering the recommended limit of 7–10 hours
- The average very active minutes also falls short of the recommended 30 minutes of vigorous exercise every day. Only 25% of people manage to hit this quota
- The average hours spent asleep (6.9) also barely hits the quota of the recommended sleep time of 7–9 hours

Now let's have a look at which days are users most active:

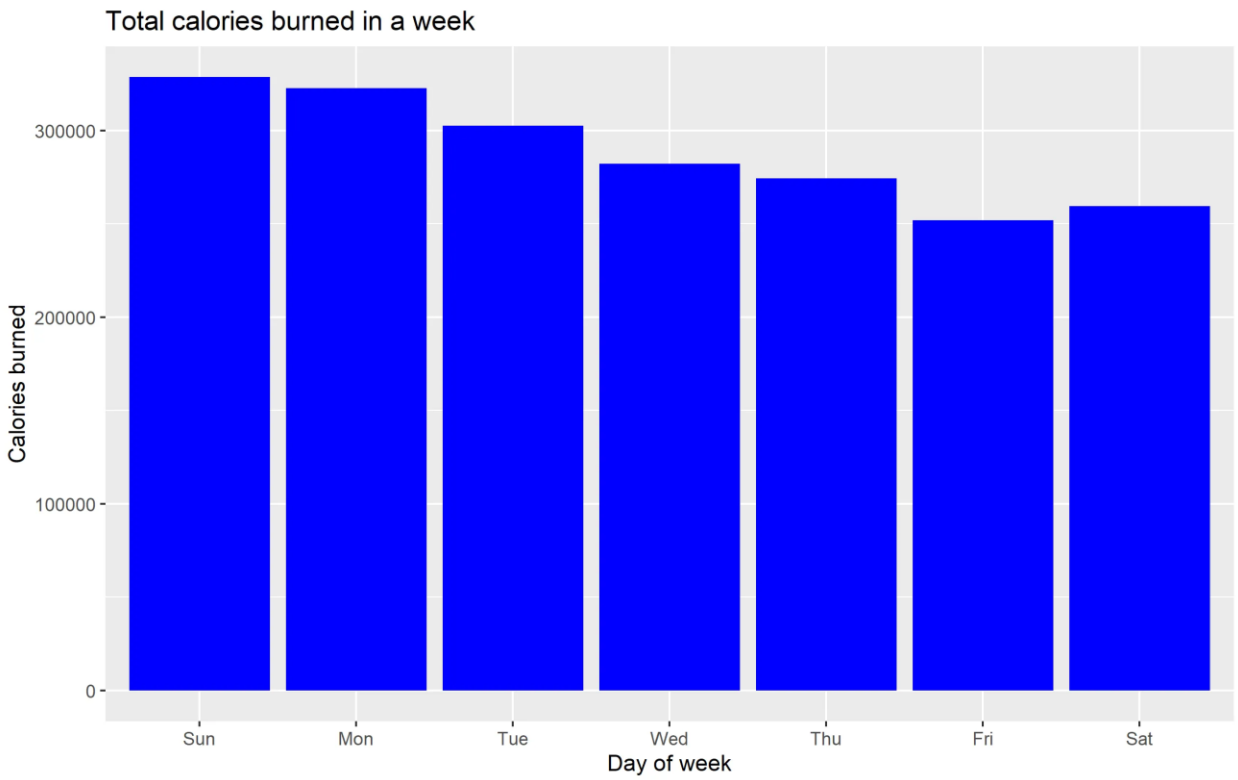
```
#options(scipen=) will remove any scientific notations

options(scipen = 999)
ggplot(data = daily_activity_cleaned) +
aes(x = day_of_week, y = total_steps) +
geom_col(fill = 'blue') +
labs(x = 'Day of week', y = 'Total steps', title = 'Totap steps taken in a week')
ggsave('total_steps.png')

ggplot(data = daily_activity_cleaned) +
aes(x = day_of_week, y = very_active_minutes) +
geom_col(fill = 'red') +
labs(x = 'Day of week', y = 'Total very active minutes', title = 'Total activity in a week')
ggsave('total_activity.png')

ggplot(data = daily_activity_cleaned) +
aes(x = day_of_week, y = calories) +
geom_col(fill = 'brown') +
labs(x = 'Day of week', y = 'Calories burned', title = 'Total calories burned in a week')
ggsave('total_calories.png')
```

the most active days for the Fitbit users were on Sunday, with a slow decline throughout the week. This could be due to motivation levels being fairly high during the end of the week.

Next, let's investigate the relationship between total active hours, total steps taken, and sedentary hours against calories burned by using the following:

the most active days for the Fitbit users were on Sunday, with a slow decline throughout the week. This could be due to motivation levels being fairly high during the end of the week.

Next, let's investigate the relationship between total active hours, total steps taken, and sedentary hours against calories burned by using the following:

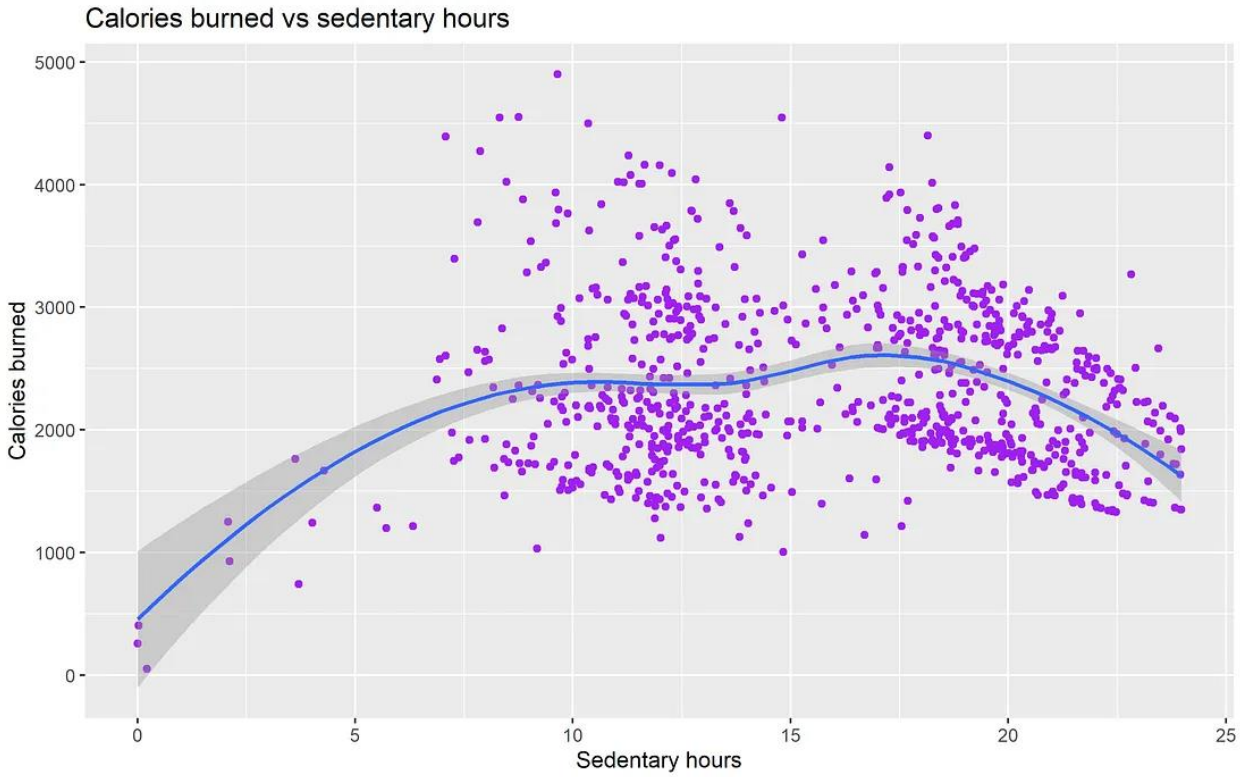


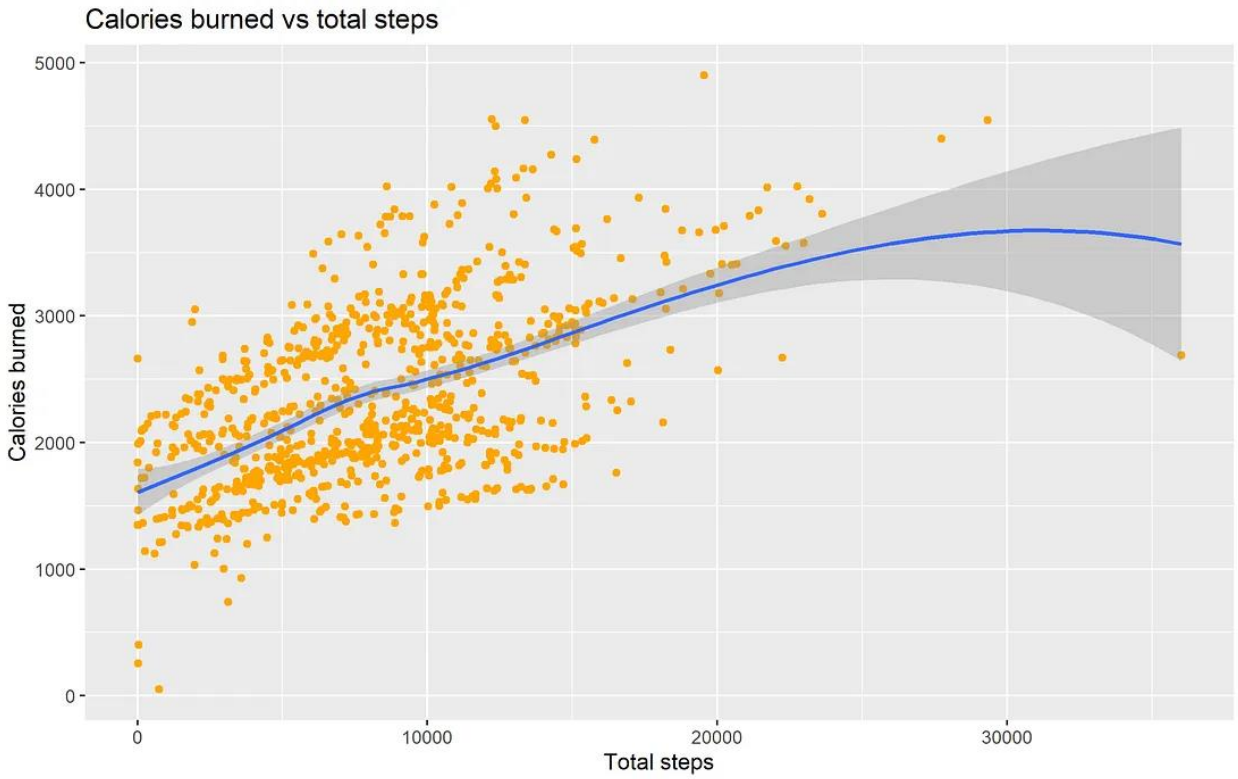
the most active days for the Fitbit users were on Sunday, with a slow decline throughout the week. This could be due to motivation levels being fairly high during the end of the week.

Next, let's investigate the relationship between total active hours, total steps taken, and sedentary hours against calories burned by using the following:

```
ggplot(data = daily_activity_cleaned) +
aes(x= total_active_hours, y = calories) +
geom_point(color = 'red') +
geom_smooth() +
labs(x = 'Total active hours', y = 'Calories burned', title = 'Calories burned vs active hours')
ggsave('calories_burned_vs_active_hours.png')

ggplot(data = daily_activity_cleaned) +
aes(x= total_steps, y = calories) +
geom_point(color = 'orange') +
geom_smooth() +
labs(x = 'Total steps', y = 'Calories burned', title = 'Calories burned vs total steps')
ggsave('calories_burned_vs_total_steps.png')
```

```
ggplot(data = daily_activity_cleaned) +
aes(x= sedentary_hours, y = calories) +
geom_point(color = 'purple') +
geom_smooth() +
labs(x = 'Sedentary hours', y = 'Calories burned', title = 'Calories burned vs sedentary hours')
ggsave('sedentary_hours_vs_calories_burned.png')
```





- Users show a consistent pattern of increased activity during weekdays compared to weekends.
- Steps and activity levels are highest in the morning hours, gradually decreasing throughout the day.
- Heart rate tends to be higher during periods of intense activity, such as in the afternoon and early evening.

- Bellabeat can leverage the observed trends to encourage and promote physical activity during weekdays, especially in the morning hours.
- The insights on heart rate patterns can help Bellabeat develop targeted interventions to manage stress and promote relaxation techniques.
- Understanding the differences in activity levels between weekdays and weekends can guide the development of customized workout plans and fitness challenges for Bellabeat customers.

❖ Act

Based on my findings after my analysis, I would like to share my hypothesis on this matter.

Users spend more time engaged in physical activity specifically on Sundays, which then proceeds to wane throughout the week with a slight peak on Thursdays which then sees a slow climb on Saturdays.

I suspect that: Now to answer the final business task, I would like to share my recommendations based on my findings to help influence Bellabeat's marketing strategy.

1. Bellabeat could host events limited to those that are enrolled in their Bellabeat memberships which would reward users who engage in a healthy lifestyle(IE 8k steps a day, less than 7 hours sedentary etc.) with points. With enough points, users could then use points to purchase products that help supplement a healthy lifestyle.
2. Bellabeat could partner with brands (IE wellness, sports, health) to reward users who consistently engage in a healthy lifestyle with coupons/store discounts.
3. With the 2 previous points combined, Bellabeat could select previously unhealthy individuals (who are now healthy), interview them and publish motivational videos as to how Bellabeat encouraged them to have a change in lifestyle.

Next, I would provide some general recommendations to further improve Bellabeat's products:

1. Bellabeat could implement personalized milestones, to encourage users to slowly engage in a more healthy lifestyle. A simple way of doing this is to create some sort of AI companion on the app/product that would be grumpy/sad if the user does not hit the milestone.
2. Bellabeat could implement a simple reminder to inform users that they've been sedentary for too long by **indefinitely vibrating the device** until the device picks up movement/increase in heart rate, which would indicate that they've engaged in some sort of physical activity.

Additional remarks:

- Bellabeat should require users to input their height and their activity levels so that BMR calculations and a more accurate calculation of TDEE would be possible. This would aid future analysis as well.
- Bellabeat should create devices that would track sleep more sophisticatedly (IE REM sleep tracking, deep sleep tracking) to provide more insights into sleep health, as in the dataset provided, we only had the quantity of sleep, not the quality of sleep.