BELLABEAT

FITBIT DATA ANALYSIS REPORT

A CASE STUDY

Prepared by:

Edun Joshua, Junior Data Analyst.

Bellabeat Marketing Analyst Team

HOW CAN A WELLNESS TECHNOLOGY COMPANY PLAY IT SMART?

About the company

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Sršen 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 analyze 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.

A. BUSINESS TASK:

1.To find trends in the fitness data of BellaBeat Users.
2.To examine how these trends apply to BellaBeat customers.
3.To provide recommendations based on trends discovered to help guide BellaBeat marketing strategy.

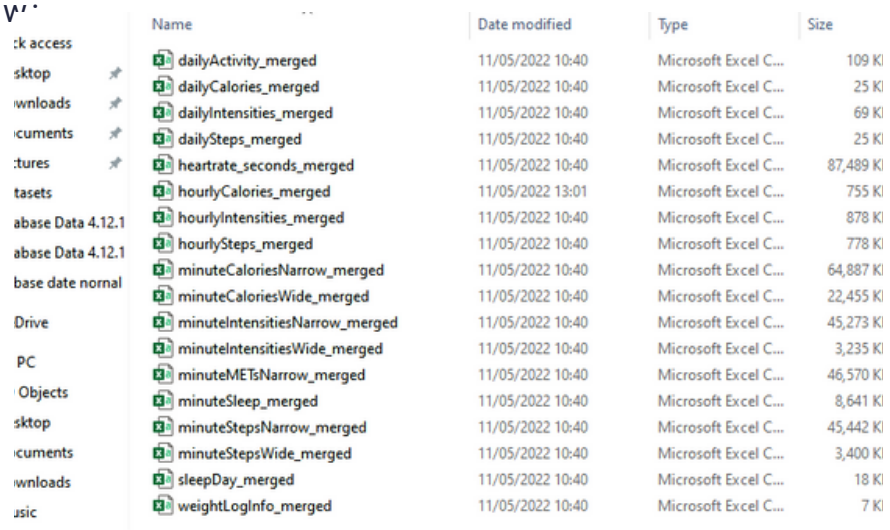
B.PREPARE

#Where is your data stored?

The data resides in the Public Domain (Kaggle), made available through Mobius.
The data set contains personal fitness tracker from thirty 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.

#How is the data organized? Is it in long or wide format?

The data was organized in the long format. It consists of 18 tables as shown below:



#Are there issues with bias or credibility in this data?

The data was collected ethically as all thirty Fitbit users consented to the submission of their data.

#How are you addressing licensing, privacy, security, and accessibility?

All datasets and data analysis tools were password protected and only members of the marketing analyst team were able to access the data.

C.PROCESS

1. Check the data for errors.

On preliminary exploration on Excel, I discovered that date values on some of the tables were inconsistently formatted. Therefore all date fields have to be homogenized going forward

2. Choose your tools.

I employed **R**'s versatile date formatting function to set all date fields to the correct format. I paste the script below.

```
# cleaning daily_calories.csv
write.csv(calories, "C:/Users/Public/Datasets/Fitbase date
nornal/daily_calories_norm.csv")
calories <- read.csv("C:/Users/Public/Datasets/Fitabase Data 4.12.16-
5.12.16/dailyCalories_merged.csv")
head(calories)
calories$ActivityDay=as.POSIXct(calories$ActivityDay,
format="%m/%d/%Y")
head(calories)

#cleaning daily_intensities.csv
intensities <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/dailyIntensities_merged.csv")
intensities$ActivityDay=as.POSIXct(intensities$ActivityDay,
format="%m/%d/%Y")
write.csv(intensities, "C:/Users/Public/Datasets/Fitbase date
nornal/daily_intensities_norm.csv")

#cleaning daily_steps
daily_steps <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/dailySteps_merged.csv")
daily_steps$ActivityDay=as.POSIXct(daily_steps$ActivityDay,
format="%m/%d/%Y")
write.csv(daily_steps, "C:/Users/Public/Datasets/Fitbase date
nornal/daily_steps.csv")

#cleaning heartrate_seconds.csv
heartrate_seconds <- read.csv("C:/Users/Public/Datasets/Fitabase
Data 4.12.16-5.12.16/heartrate_seconds_merged.csv")
heartrate_seconds$Time=as.POSIXct(heartrate_seconds$Time,
format="%m/%d/%Y %I:%M:%S %p")
write.csv(heartrate_seconds, "C:/Users/Public/Datasets/Fitbase date
nornal/heartrate_seconds.csv")
```

```
#cleaning hours_calories.csv
hour_calories <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/hourlyCalories_merged.csv")
hour_calories$ActivityHour=as.POSIXct(hour_calories$ActivityHour,
format="%m/%d/%Y %I:%M:%S %p")
head(hour_calories)
edit(hour_calories)
write.csv(hour_calories, "C:/Users/Public/Datasets/Fitbase date
nornal/hourlyCalories_merged.csv")

#cleaning hour_intensities.csv
hour_intensities <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/hourlyIntensities_merged.csv")
hour_intensities$ActivityHour=as.POSIXct(hour_intensities$ActivityHo
ur, format="%m/%d/%Y %I:%M:%S %p")
write.csv(hour_intensities, "C:/Users/Public/Datasets/Fitbase date
nornal/hour_intensities.csv")

#cleaning hourly_steps.csv
hourly_steps <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/hourlySteps_merged.csv")
hourly_steps$ActivityHour=as.POSIXct(hourly_steps$ActivityHour,
format="%m/%d/%Y %I:%M:%S %p")
write.csv(hourly_steps, "C:/Users/Public/Datasets/Fitbase date
nornal/hourly_steps.csv")

#cleaning minutes_calories.csv
minutes_calories <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/minuteCaloriesNarrow_merged.csv")
minutes_calories$ActivityMinute=as.POSIXct(minutes_calories$Activit
yMinute, format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_calories, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_calories.csv")


#cleaning minutes_intensities.csv
minutes_intensities <- read.csv("C:/Users/Public/Datasets/Fitabase
Data 4.12.16-5.12.16/minuteIntensitiesNarrow_merged.csv")
minutes_intensities$ActivityMinute=as.POSIXct(minutes_intensities$A
ctivityMinute, format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_intensities, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_intensities.csv")
```

```
#cleaning minutes_intensities.csv
minutes_intensities <- read.csv("C:/Users/Public/Datasets/Fitabase
Data 4.12.16-5.12.16/minuteIntensitiesNarrow_merged.csv")
minutes_intensities$ActivityMinute=as.POSIXct(minutes_intensities$A
ctivityMinute, format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_intensities, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_intensities.csv")

#cleaning minutes_METs.csv
minutes_METs <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/minuteMETsNarrow_merged.csv")
minutes_METs$ActivityMinutes=as.POSIXct(minutes_METs$ActivityMin
ute, format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_METs, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_METs.csv")

#cleaning minutes_sleep.csv
minutes_sleep <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/minuteSleep_merged.csv")
minutes_sleep$date=as.POSIXct(minutes_sleep$date,
format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_sleep, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_sleep.csv")

#cleaning minutes_steps.csv
minutes_steps <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/minuteStepsNarrow_merged.csv")
head(minutes_steps)
minutes_steps$ActivityMinute=as.POSIXct(minutes_steps$ActivityMin
ute, format="%m/%d/%Y %I:%M:%S %p")
write.csv(minutes_steps, "C:/Users/Public/Datasets/Fitbase date
nornal/minutes_steps.csv")

#cleaning sleep_day.csv
sleep_day <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/sleepDay_merged.csv")
sleep_day$SleepDay=as.POSIXct(sleep_day$SleepDay,
format="%m/%d/%Y %I:%M:%S %p")
write.csv(sleep_day, "C:/Users/Public/Datasets/Fitbase date
nornal/sleep_day.csv")

#cleaning weight_log_info.csv
weight_log_info <- read.csv("C:/Users/Public/Datasets/Fitabase Data
4.12.16-5.12.16/weightLogInfo_merged.csv")
weight_log_info$Date=as.POSIXct(weight_log_info$Date,
format="%m/%d/%Y %I:%M:%S %p")
write.csv(weight_log_info, "C:/Users/Public/Datasets/Fitbase date
nornal/weight_log_info.csv")
```

D. ANALYZE

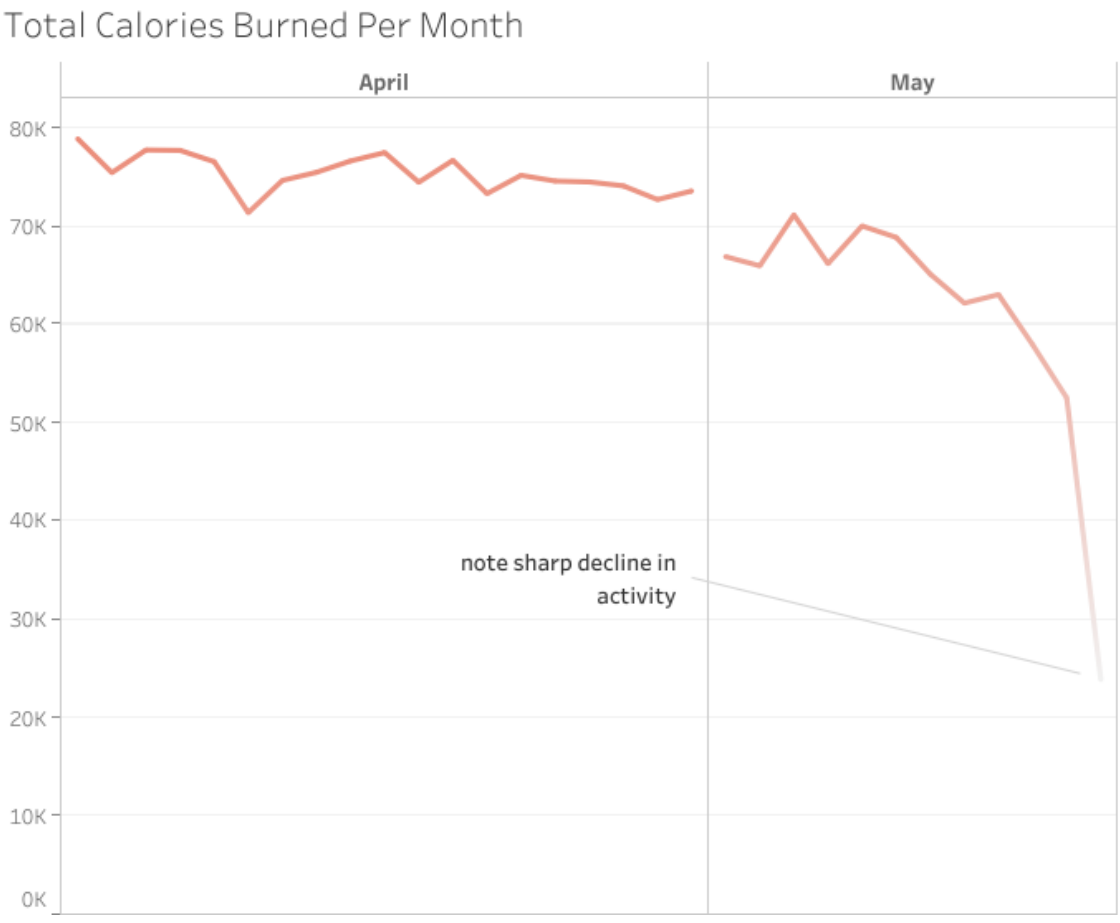
I aggregated the data using SQL (PostGres) according to the following script:

```
--1.0 exploring the tables

--Check for total number of rows
SELECT COUNT (id)
FROM calories_hourly
--22099

--Checking for incorrect ids
SELECT length(CAST(id as TEXT)) AS length_of_char,
COUNT(*)
FROM calories_hourly
GROUP BY 1
--no incorrect ids

--Check for nulls
SELECT COUNT (id)
FROM calories_hourly
WHERE id IS NULL
-- no nulls

--How many customers are in the calories_hourly table?
SELECT COUNT (DISTINCT id) AS no_of_customers
FROM calories_hourly;
--33 customers

--What is the period over which this data was collected
SELECT MAX(activity_hour), MIN(activity_hour)
FROM calories_hourly

SELECT MAX(activity_date), MIN(activity_date)
FROM daily_calories
--a period of one month, from 12th May, 2016 to 12th April, 2016

--What is the average, max & min calories burned per hour
SELECT id as customers,
round(AVG(calories), 2) as average_calories_burned,
MAX(calories),
MIN(calories)
FROM calories_hourly
GROUP BY 1
ORDER BY 2 DESC
```

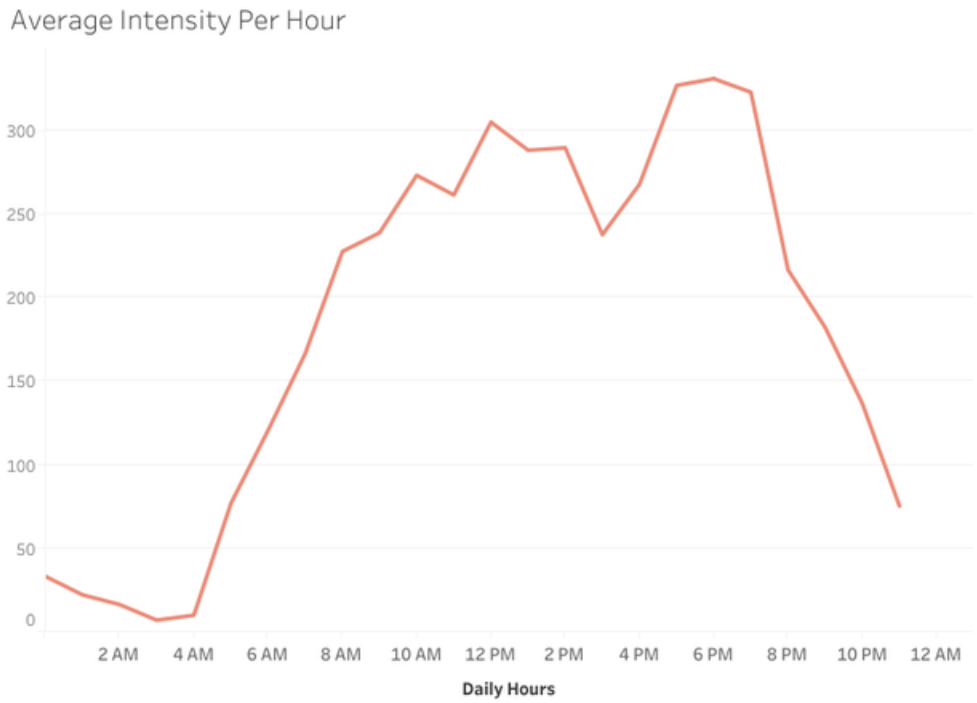
```
--What is the average, max & min calories burned per day
SELECT id as customers,
round(AVG(calories), 2) as average_calories_burned,
MAX(calories),
MIN(calories)
FROM daily_calories
GROUP BY 1
ORDER BY 2 DESC

--further investigate the above two by viz
--do viz showing percentage of customers below WHO standard

--2.0
--Check for total number of rows in daily_activity
SELECT COUNT (id)
FROM daily_activity
--940

--Checking for incorrect ids
SELECT length(CAST(id as TEXT)) AS length_of_char,
COUNT(*)
FROM  daily_activity
GROUP BY 1
--no incorrect ids

--Check for nulls
SELECT COUNT (id)
FROM daily_activity
WHERE id IS NULL
-- no nulls

--How many customers are in the calories_hourly table?
SELECT COUNT (DISTINCT id) AS no_of_customers
FROM daily_activity;
--33 customers

--Join daily activity and weight log
SELECT d.id, d.activity_date, d.total_steps, d.total_distance,
d.very_active_distance, d.very_active_minutes, d.calories,
w.weight_kg, w.bmi
FROM daily_activity d INNER JOIN weight_log w
ON d.id = w.id

--Join heart rate and weight log
SELECT h.id, h.time, h.value, w.weight_kg, w.bmi
FROM heart_rate_seconds h INNER JOIN weight_log w
ON h.id = w.id
```

```
--Summary of weight_log
SELECT round(MAX(weight_kg), 1) AS MAX_WEIGHT,
round(AVG(weight_kg), 1) AS AVG_WEIGHT,
round(MIN(weight_kg), 1) AS MIN_WEIGHT,
MAX(bmi) AS MAX_BMI,
round(AVG(bmi), 1) AS AVG_BMI,
MIN(bmi) AS MIN_BMI
FROM weight_log
```

VISUALIZATION:



Using Tableau, I discovered and visualized the following trends:
1. A sharp decline in activity in May. Historical data would be needed to determine if this is an annual trend or an anomaly.



2. BellaFit users were mostly active from 5pm - 7pm daily. I recommend BellaBeat provide notifications at these times to remind users to stay hydrated.

3.A positive correlation exists between total minutes asleep and actitivy. I infer that the more sedentary users are not sleeping enough

ACT

1.Notification reminders between 5pm - 7pm would be beneficial in reminding Bellabeat users to stay hydrated. Bellabeat's product, The Spring: which is a water bottle that tracks daily water intake using smart technology can be leveraged to ensure users are appropriately hydrated.

2.Less active users are not getting enough sleep. The Bellabeat app can be used to remind users to get their good night rest.

3.Encourage users to spend more time engaging with the app