BELLA BEAT CASE STUDY

My first case study

(Due to case study roadmap)

MY PROBLEM SOLVING STRATEGY

This is my first data analytics case study and obviously I find it hard to do it on my own based on the meager wisdom that I've obtained recently. So mentioning that the first step of learning is imitation; I started off my case study by looking at many other case studies on kaggle or personal blogs in order to find a meaningful pattern to employ on my own study. ((a special shout out to Gary Yiu and his personal blog which guided me a lot on this case))

And then after finishing my case study once and gaining the primary insights, I will entirely focus on the google case study roadmap to improve my method and gain new insights on both the dataset and the accuracy and validity of the path I've taken.

SUMMARY

Bella Beat is a high-tech company which provides health-focused apps and smart devices (Smart watches and etc.) for women. Bella beats smart devices gather data on how its consumers use them and tries to help women make better use of what Bella Beat provides in order the help them live a better and healthier life.

In this case study we are about to identify trends in how random consumers use smart devices, so that we can gain and apply insights into Bella Beat's marketing strategy. The main focus of this case is Bella Beat app which provides user with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits.

THE ASK PHASE

1. What is the problem you are trying to solve?

The company thinks they can have more users after knowing how people use smart devices in their lives. So we should try to debug the present problems of the app and the smart devices and make better use of them so that it'll become what people would actually need.

2. The three questions Bella Beat wants to be answered are;

A. What are the trends in smart device usage?

B. How could these trends apply to Bella Beat?

C. How could these trends help Bella Beat marketing strategy?

3. Identify the business task

Analyzing usage of non-Bella Beat smart device data in order to find trends and gain insight on how to produce, improve or promote Bella Beat smart devices.

4. Consider key stakeholders

- **Urška Sršen:** Bella beats cofounder and Chief Creative Officer
- **Sando Mur:** Mathematician and Bella beats cofounder; key member of the Bella beat executive team
- **Bella beat marketing analytics team**

THE PREPARE PHASE

1. Credibility of data

Our data source is Fit Bit fitness tracker data. This data is stored on Kaggle and is made available through Mobius. It's an open-source dataset from thirty Fit Bit users that includes minute-level output for physical activity, heart rate and sleep monitoring. This dataset is under CC0: Public Domain license meaning the creator has waive his right to the work under the copyright law.

The dataset has in total 18 files in .csv format organized in long format.

The data has been stored as long format because each subject has data in multiple rows.

The data we're using has two critical flaws; first one is being outdated and the second one is using an extremely small population, so we are not sure if the sample is representative of the population as a whole.

The data is mainly about ((TIME)) and ((ACTION)) like daily or hourly activity so we can take it as a sign of integrity of the data.

In conclusion drilling down to explore user behaviors to identify trends would be a good starting point for our analysis.

2. ROCC analysis

Reliability: LOW – dataset was collected from 30 individuals whose gender is unknown.

Originality: LOW – third party data collect using Amazon Mechanical Turk.

Comprehensive: MEDIUM – dataset contains multiple fields such as daily activity, calories used, daily steps and etc.

Current: MEDIUM – data is 6years old and it's not new enough but we can say that people's habits won't normally change much within 6years.

Cited: HIGH – data is well documented

3. Data selection

I will focus on daily usage of smart devices as it provides a better insight on the usage pattern of the devices, Thus the following files from the dataset will be used;

- dailyActivity_merged.csv
- dailyCalories_merged.csv
- dailyIntensities_merged.csv
- dailySteps_merged.csv
- sleepDay_merged.csv
- weightLogInfo_merged.csv

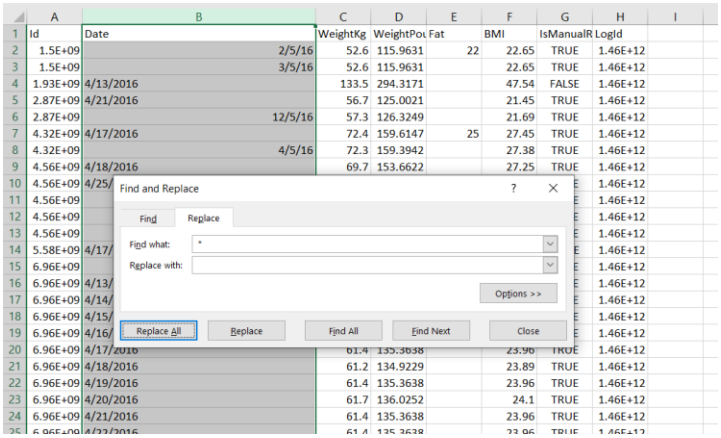
THE PROCESS PHASE

Microsoft Excel and MySQL will be used to process the data as they fit our purpose, and finally we will use PowerBI to visualize what we have found.

DATA CLEANING ((Excel))

After loading sleepDay_merged.csv and weightLogInfo_merged.csv into Excel for data cleaning, we can see that the "SleepDay" and "Date" columns are not correctly formatted, So;

After selecting the mentioned column we use ctrl-F to replace the data using a space and a *:



DATA CLEANING ((MySQL))

"Importing dataset"

After importing all the csv formatted data into MySQL, I double checked the imported data in order to find out if any of them is missing. The final number of imported data were:

Dailyactivity_merged: 940rows

Dailycalories_merged: 940rows

Dailyintensities_merged: 940rows

Dailysteps_merged: 940rows

Sleepday_merged: 413rows

Weightloginfo_merged: 67rows ((NO DATA WAS MISSING))

“Number of users”

THEN by using “ SELECT DISTINCT id “ query we can find out how many unique IDs we have in each table. Result:

Dailyactivity_merged: 33

Dailycalories_merged: 33

Dailyintensities_merged: 33

Dailysteps_merged: 33

Sleepday_merged: 24

Weightloginfo_merged: 8

These numbers show us that there are some missing data in the dataset according to the missing IDs that we faced while analyzing the last two tables ((Sleepday_merged & Weightloginfo_merged)) with 6 and 22 missing IDs.

“Duplicates & N/A”

I looked for duplicates in each table in order to find out if each user has successfully done their task of importing the data each day. Results:

```
1   SELECT id, COUNT(*)
2   FROM dailyactivity_merged
3   GROUP BY id
4   HAVING COUNT(*)>1
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	COUNT(*)
1844505072	31
1927972279	31
2022484408	31
2026352035	31
2320127002	31
2347167796	18
2873212765	31
3372868164	20
3977333714	30
4020332650	31
4057192912	4
4319703577	31
4388161847	31
4445114986	31
4558609924	31
4702921684	31
5553957443	31
5577150313	30
6117666160	28
6290855005	29
6775888955	26
6962181067	31
7007744171	26
7086361926	31
8053475328	31
8253242879	19
8378563200	31
8583815059	31

Results show that many of users haven't done the task of tracking and importing their activity info every day, meaning that the data is incomplete and it would be hard to analyze the data.

The result is the same with other tables either.

I should also mention that the way that data has been formatted makes it hard for data analyzer to find the duplicated data and omit them

Then I checked for all the Null IDs and I found none:

```
1 •  SELECT *
2    FROM bellabeat.`weightloginfo_merged 2`
3    WHERE id is null
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Id	Date	WeightKg	WeightPounds	Fat	BMI	IsManualReport	LogId

weightloginfo_merged 2 6

Output

Action Output

#	Time	Action	Message
1	09:48:19	SELECT * FROM bellabeat.dailyactivity_merged LIMIT 0, ...	940 row(s) returned
2	09:53:33	SELECT * FROM bellabeat.dailycalories_merged LIMIT 0, ...	940 row(s) returned
3	09:54:00	SELECT * FROM bellabeat.dailyintensities_merged LIMIT ...	940 row(s) returned
4	10:05:34	SELECT * FROM dailyactivity_merged WHERE id is null L...	0 row(s) returned
5	10:05:44	SELECT * FROM dailycalories_merged WHERE id is null ...	0 row(s) returned
6	10:05:53	SELECT * FROM dailyintensities_merged WHERE id is nu...	0 row(s) returned
7	10:06:00	SELECT * FROM dailysteps_merged WHERE id is null LI...	0 row(s) returned
8	10:06:25	SELECT * FROM bellabeat.`sleepday_merged 2` WHERE...	0 row(s) returned
9	10:06:33	SELECT * FROM bellabeat.`weightloginfo_merged 2` WH...	0 row(s) returned

THE ANALYZE PHASE

For this stage my goal is to write down as many hypothesis as I can only by looking at the data I have, then I'll be analyzing them and as we go forward I may omit the hypothesis that wouldn't make sense anymore.

- There is a direct relationship between the daily activity and calories burnt. (approved)
- More daily activity needs more sleep, also more sleep can lead to more daily activity. (approved)
- There is a direct relationship between daily steps and daily activity. (approved)
- There is a reverse relationship between daily activity and weight kg. (unable to find out due to shortage of data)
- Heavier people need more sleep. (not approved)
- Total sleep time has a reverse relationship with the sedentary minutes.(approved)

1##

In order to find the relationship between activity level and calories:

```
4 •  SELECT id, ActivityDate, TotalSteps, Calories, VeryActiveMinutes,LightlyActiveMinutes, FairlyActiveMinutes, SedentaryMinutes
5    FROM dailyactivity_merged
6    WHERE VeryActiveMinutes+FairlyActiveMinutes+LightlyActiveMinutes <> 0
7    ORDER BY TotalSteps desc
```

we can see that total steps doesn't have much of a meaningful relationship with calories burnt, but the sum of active minutes plus total steps would definitely lead to higher calorie burn, a fun fact was that lightly active minutes actually has a huge impact on the amount of daily burnt calories which I didn't expect.

And to have it all together we can:

```
SELECT id, ActivityDate, TotalSteps, Calories, (VeryActiveMinutes+ LightlyActiveMinutes+ FairlyActiveMinutes) AS total_active_minutes, SedentaryMinutes
FROM dailyactivity_merged
WHERE VeryActiveMinutes+FairlyActiveMinutes+LightlyActiveMinutes <> 0
ORDER BY calories desc
```

So we can obviously say that total activity level and calories burnt have a direct relationship. And also that that lightly active minutes actually has a huge impact on the amount of daily burnt calories.

2##

NOW it’s time to find the relationship between the activity level and sleep time

I should mention that in these kinds of analyses it'd better to analyze person by person, so that we can see people’s life style and then find the pattern in it.

```
SELECT activity.id, ActivityDate, TotalSteps, Calories, VeryActiveMinutes,LightlyActiveMinutes, totalsleeprecords, TotalMinutesAsleep
FROM dailyactivity_merged AS activity
INNER JOIN `sleepday_merged 2` AS sleep
ON activity.id = sleep.id AND activity.activitydate = sleep.sleepday
```

What this query shows us is that there isn't a meaningful relationship between total sleep and calories burnt by looking at the highest to lowest amount of calorie burnt or sleep time in general.

But if we look closer to each person’s data, we can see that according to the regular sleep minutes of each person, there's an increase in calorie burn volume in the day after the night that he/she sleeps more than usual. or at least we can say that if the person has have a high volume of calorie burn in two to three days, He/ She would spend more time in bed than usual in one of the mentioned days night in order to get refreshed.

```
SELECT activity.id, ActivityDate, TotalSteps, Calories, VeryActiveMinutes,LightlyActiveMinutes, totalsleeprecords, TotalMinutesAsleep
FROM dailyactivity_merged AS activity
INNER JOIN `sleepday_merged 2` AS sleep
ON activity.id = sleep.id AND activity.activitydate = sleep.sleepday
```

3##

For the relationship between activity level and weight using BMI (A weight measuring metric that includes height too)

```
SELECT activity.id, ActivityDate, TotalSteps, Calories, VeryActiveMinutes,LightlyActiveMinutes, BMI
FROM dailyactivity_merged AS activity
INNER JOIN `weightloginfo_merged 2` AS weight
ON activity.id = weight.id AND activity.activitydate = weight.date
```

Due to shortage of data of data it's hard to draw a conclusion from this table, also weight is an element that should be tracked in long run and no interesting result can be extracted in a short period of time.

4##

AND for the relationship between sleep time and weight

```
SELECT sleep.id,  totalsleeprecords, TotalMinutesAsleep, BMI
FROM `sleepday_merged 2` AS sleep
INNER JOIN `weightloginfo_merged 2` AS weight
ON sleep.id = weight.id AND sleep.sleepday = weight.date
```

We couldn’t find any relationship between these two tables due to shortage of data and inconsistency.

5##

The relationship between sedentary minutes and sleep:

```
SELECT activity.id, ActivityDate, TotalSteps, Calories, VeryActiveMinutes,LightlyActiveMinutes, totalsleeprecords, TotalMinutesAsleep, SedentaryMinutes
FROM dailyactivity_merged AS activity
INNER JOIN `sleepday_merged 2` AS sleep
ON activity.id = sleep.id AND activity.activitydate = sleep.sleepday
ORDER BY TotalMinutesAsleep desc
```

As we can see there's a reverse relationship between these two items. So as the results show us; more activity during the day would lead to more quality sleep during bedtime.

THE VISUALIZATION PHASE

At the beginning I imported the three mentioned data sets into Power BI;

- Daily activity
- Sleep day
- Weight log

At this point we should first find out what we would like to see in the Visualizations, which tables? Which relationships? And which numbers?

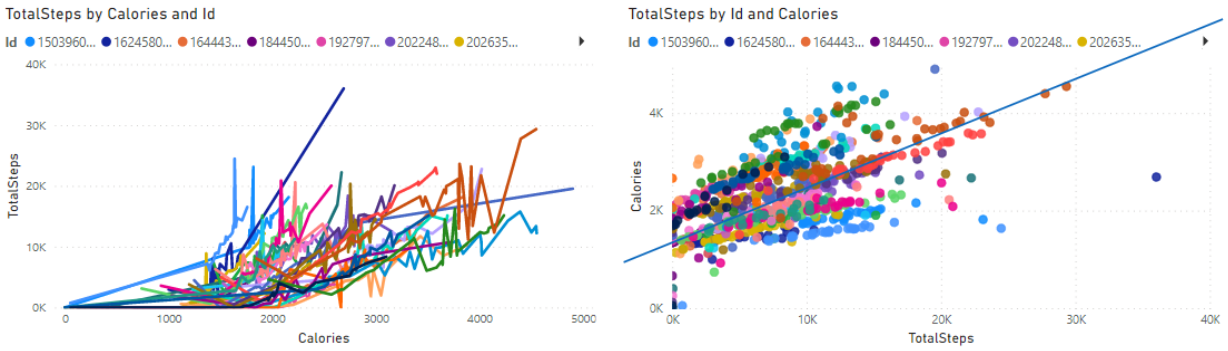
We will visualize the data according to the hypothesis we've had at the beginning. We may also add up some new visualizations in order to find new relationships.

First we prepare the data in the table and model section of PowerBI, Then we can make links between the tables, for instance we linked all the IDs in each table to the other to make a many-to-many relationship.

We will now take each step according to our hypothesis:

- There is a direct relationship between the daily activity and calories burnt. (approved)

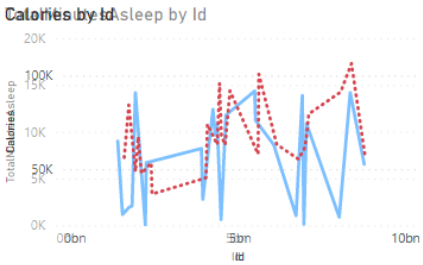
The relationship between activity and calories:



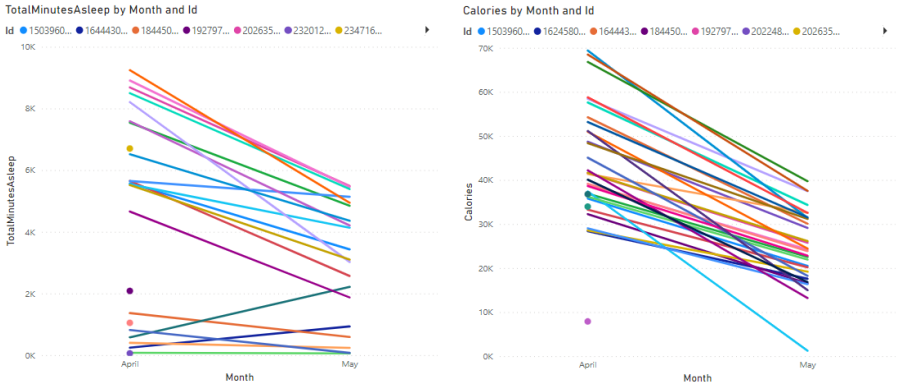
- More daily activity needs more sleep, also more sleep can lead to more daily activity. (approved)

The relationship between activity and sleep-time:

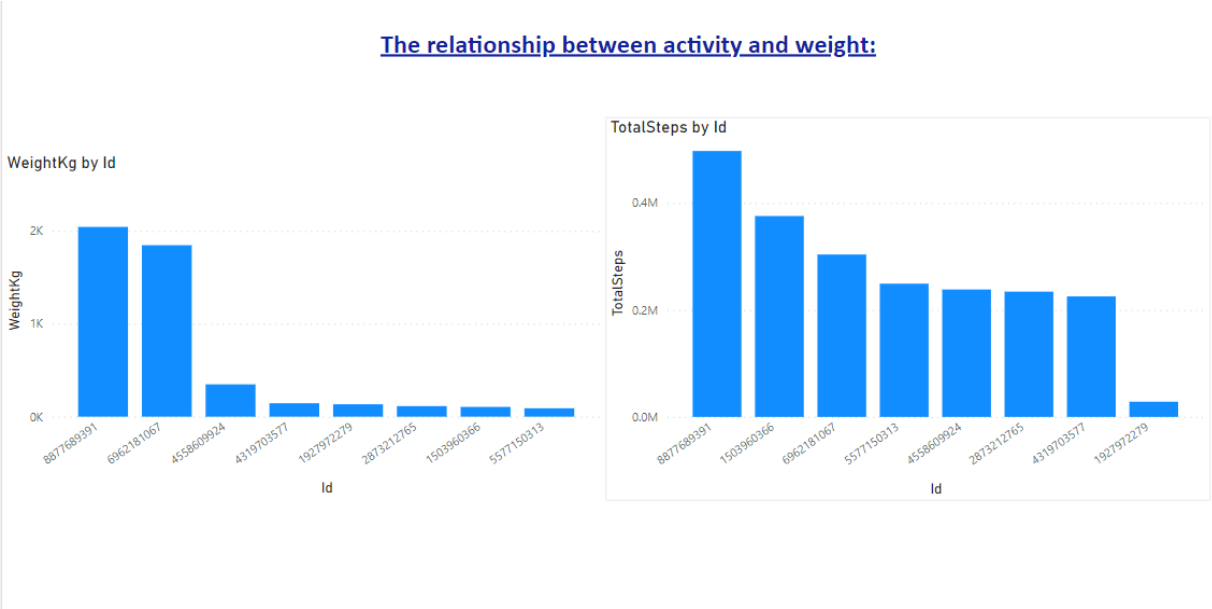
And by putting them together we can see that:



The relationship between sleep and calories:



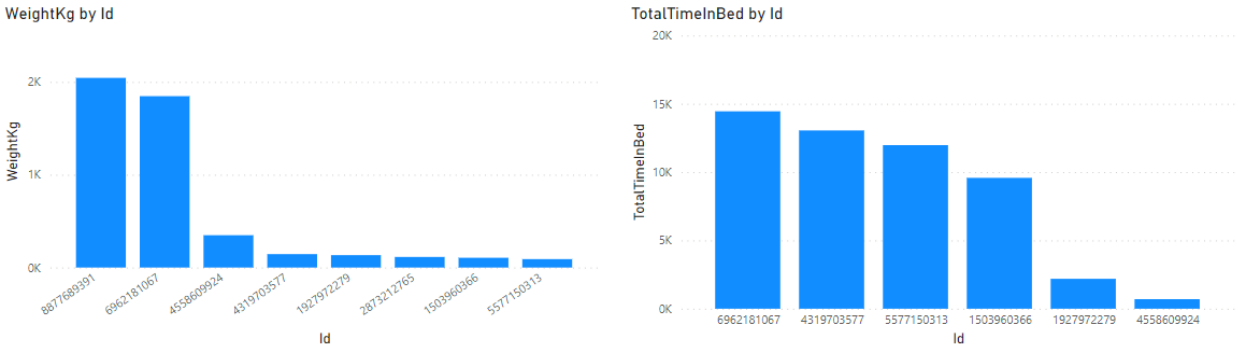
- There is a reverse relationship between daily activity and weight kg. (unable to find out due to shortage of data)



We can say that according to the data we have there could be a relationship, but due to shortage of data it's not so vivid and trust worthy to rely on.

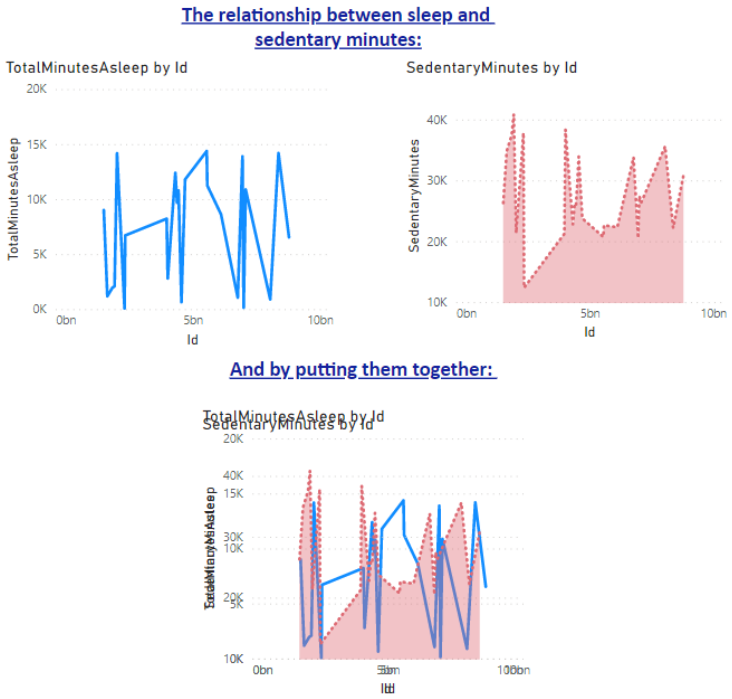
- Heavier people need more sleep. (not approved)

The relationship between sleep and weight:



No relationship has been witnessed because there is no correlation in the order of the IDs from the highest to lowest in the two visualized tables.

- Total sleep time has a reverse relationship with the sedentary minutes.(approved)



Here we can obviously witness the reverse relationship among these two items.

CONCLUSION

The results show us that there's a clear trend in non-active people having a negative lifestyle,

The analysis on my hypothesis has led us to draw a few major conclusions:

- More activity will surely lead to more calories burnt.

- More activity would lead to better health and lifestyle, is also increases the quality of sleep. (according to my last hypothesis)

- Non-active people would probably have a higher BMI

RECOMMENDATION

According to the pattern and trends that we noticed studying smart devices, there are a few recommendations that would help BellaBeat improve its smart devices and app:

- 1) BellaBeat can track peoples sleep habits and hint them on how much they would need to sleep in order to modify their lifestyle and health.

- 2) BellaBeat can include functions that track sedentary minutes, so that it can alert the consumer on how to reform their sedentary time and advise them on what they can do during this time.(like taking a walk with friends or cycling to work)

- 3) BellaBeat can track consumers BMI and recommend them some daily activities or healthy food so that they'd be able to keep their body healthy and shaped.

- 4) BellaBeat can aware people on how big time sedentary minutes would decrease the efficiency of your lifestyle and advise consumers on how to modify it by meditating and yoga or maybe even dancing

We can apply these recommendations to make data-driven decisions on Bellabeats future products/functionality.