**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.

**Scenario**

Working on the marketing analyst team at Bellabeat a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analysing smart device fitness data could help unlock new growth opportunities for the company. **You have been asked to focus on one of Bellabeat’s products and analyse smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company.**

Sršen asks you to **analyse smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants you to select one Bellabeat product to apply these insights to in your presentation.**

These questions will guide your analysis:

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**

With all the information we have the business task is to analyse data from non-Bellabeat smart devices to see how the consumers are using it and with the insights found in that data apply it to a Bellabeat’s product to help the marketing team to find a strategy for the company.

The analysis and recommendations will be presented to the Bellabeat executive team and Marketing team.

**About Dataset**

This Kaggle dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 12.03.2016 - 12.05.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.

**File-Naming Convention**

I used camelcase followed by the dates for the folders containing the data.

Examples:

* fitbitData12\_03\_2016-11\_04\_2016
* dailyActivity2

To streamline the workflow and because there are only 2 different months of data I appended number 1 and 2 at the end of each file name to indicate the month when exporting the files.

1 = 12/03/16 - 11/04/16

2 = 12/04/16 - 12/05/16

**FitBit Fitness Tracker Data (**[**Dataset**](https://www.kaggle.com/datasets/arashnic/fitbit))

License:

([FitBit Fitness Tracker Data](https://creativecommons.org/publicdomain/zero/1.0/), CC0: Public Domain, dataset made available through Mobius)

You can download the data from Kaggle as a “.zip” folder, inside you can find two folders with data from two different months, FROM 12/03/2016 TO 11/04/2016 then FROM 12/04/2016 TO 12/05/16.

Inside each month folder’s there are CSV files with different data measurements like daily Activity, heart rate, sleep monitoring, weight, steps in minutes and hourly, Calories in minutes and hourly, intensities in minutes and hourly.

Some files can be found in long or wide format.

Original Data Dictionary: ([Data dictionary](https://www.fitabase.com/media/1930/fitabasedatadictionary102320.pdf)) (The dictionary was found in a Thread doing a research about the dataset). Find the meaning of any column name that is not clear.

The data is from a public dataset platform, should be considered Second-Party. In this case got provided from Kaggle, a trusted public dataset platform but we can’t say it’s original. Could be a signal to be extra careful with the data. On the other hand Kaggle score for completeness and credibility is 100%.

Not all files within the dataset will be utilized, as some are subsets of others.

Based on the Central Limit Theorem (CLT), a sample size of 30 is considered the minimum. However, since the company's primary customers are women, and this dataset lacks gender information, this sample may not accurately represent the target demographic.

Upon conducting a unique ID count in the dailyActivityMerged file, it reveals a total of 35 unique IDs, whereas the data source indicates there are 30 users.

The data spans from March 2016 to May 2016, meaning user behaviors may have changed since then, so current habits may differ.

**The data will be used for this case study, but conducting complementary analyses with additional and diverse datasets could potentially validate the marketing strategy for the company. This approach could serve as an initial step in the process.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

After a research I found 2 years of raw data. It could be use for additional analysis. Includes more participants and different types of trackers. The data from the second half of 2020 is data during Covid-19 lockdown.

*Data were collected from 113 participants, who shared their physical activity (PA) data using privately owned smart watches and activity trackers from Garmin and Fitbit.*

*This data set consists of two data files: "data.csv" and "data raw.csv":*

* *The first file ("data.csv") contains daily averages for steps, total energy expenditure (TEE), activity energy expenditure (AEE), moderate-to-vigorous physical activity (MVPA), light PA (LPA), moderate PA (MPA), vigorous PA (VPA), and sedentary time, grouped by month. In addition, daily averages for the whole year of 2019 and 2020 are included. Finally, separate variables for the first and second half of March 2020 (pre- and post COVID-19 lockdown in Norway) are included.*
* *The second file ("data raw.csv") contains raw daily values for steps, TEE, AEE, MVPA, LPA, MPA, VPA, sedentary time, and non-wear time.*

*(2021-09-15)*

*Data obtained from:  
Henriksen, André; Johannessen, Erlend; Hartvigsen, Gunnar; Grimsgaard, Sameline; Hopstock, Laila Arnesdatter, 2021, "Replication Data for: Dataset of Consumer-Based Activity Trackers as a Tool for Physical Activity Monitoring in Epidemiological Studies During the COVID-19, Pandemic",*[*https://doi.org/10.18710/TGGCSZ*](https://doi.org/10.18710/TGGCSZ)*, DataverseNO, V3.*

**Process:**

Overview: All the files could be merged to make a big data frame with all the information together but this could be very time consuming due the amount of data in each file. Using some files individually and merging others will give similar insights.

First I used spreadsheets to have a notion of the data in each file. The amount of data in **some** CSV files is too big to work with Spreadsheets. Using R would be ideal tool for the cleaning, transforming and analysis on the larger files.

**Cleaning Data**

First I created a folder called “copy” in the same folder just to make sure I have the raw data of every CSV file in case any issues with the data occurs or for any double-check during the whole analysis process.

**Spreadsheets**

I used spreadsheets to clean and transform 2 CSV files to practice with spreadsheets as it is the most familiar tool. The CSV files are dailyActivity1, dailyActivity2, weightLogInfo1 and weightLogInfo2.

**DailyActivity Data**

**1.** I **imported** the (dailyActivity1 and dailyActivity2) CSV files to Google Sheets, 2 files imported.

**2.** For a quick and better view of the data I use View, Freeze, 1 Row.

**3.** (**Find Blank Cells**) I use conditional formatting to highlight blank cells to find missing information and add it if possible. Non found.

**4.** (**Remove empty Rows**) Create a filter with all the data to remove empty rows, checking on each column for blanks. No blanks found within the data. I removed the extra rows at the end of the spreadsheet and removed the filter applied.

**5.** (**Remove Duplicates**) Data, Data clean-up, Remove duplicates, Data has header row.

No duplicate rows were found.

**6.** (**Check Spelling**) Tools, Spelling, Spell check. Used on the headers. Units added to the headers.

One more tool could be Use Find & Replace to correct misspelling.

**7.** (**Format Consistent**) Select the **Activity Date** column, Format, Number, Date. In this case some dates are not detected as dates and the format option doesn’t work. I go through manually using split text to columns, input the delimiter “/” and then adding a column to concatenate all of them in the correct order and lastly copying and then pasting only values in a new column so the column can be formatted to date correctly.

**8.** Decreased decimal places in all the kilometers columns. Two digits after the decimal (e.g.7.22).

**9.** (**Find & Replace to trim whitespace**) Edit, Find and Replace. In find I typed 2 spaces and replaced with 1 space. Non found.

Another way → Select the data range, click data tab, Data cleanup, Trim Whitespace

**10.** **(Sorted the data)** Data, Sort range, Advanced range sorting options. Sorted by Id and Activity Date.

**11.** **(Merging time)** I merged both sheets as dailyActivityMerged. Created a new sheet and with the QUERY function merged both spreadsheets. Selected the data, added both spreadsheets in the query function, then write the query *“select \* where Col1 is not null”* and lastly chose the number of headers, 0 in this case because I copied and paste the headers line before.

=QUERY({'dailyActivity\_12\_03\_2016-11\_04\_2016'!2:458;'DailyActivity\_12\_04\_2016-12\_05\_2016'!2:941},"select \* where Col1 is not null",0)

Then copied and pasted Values Only in a new sheet, called “dailyActivityMergedValuesOnly”.

To keep this data save after the cleaning and merging, I make a duplicate of the “dailyActivityMergedValuesOnly” to work within. The duplicate is called dailyActivityMVO.

**12. (Data mapping testing phase)**

- Inspect a sample piece of data to confirm that I’s clean and properly formatted.

- Spot checks in like number of nulls.

- Data validation, Conditional formatting, COUNTIF, Sorting, Filtering.

- I Summed the number of rows from both sheets using the COUNT function and compared it to the number of rows on the merged sheet.

- Removed Duplicates. Non Found.

**Analysing Data**

**1.** Sorted the data by Activity Date to check that date entries are in the time frame mentioned.

**2.** On the right side of the data sheet I do a quick Count Unique on the column Id and I found that there’s data from 35 users as mentioned before.

**3.** Added a column to the data with the day of the week for every entry. This way we can make charts to see general trends in a week.

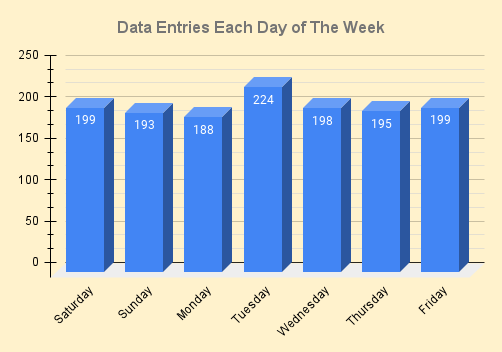
**4.** I created a pivot table with Summary Statistics from Daily Activities data. Contains: Id, Days Activity per Id, Average Steps, Average Distance (Km), Count of Logged Activities, Average DISTANCE of Activity Intensity (3 columns), Average Calories, Max Calories, Average MINUTES of Activity Intensity(4 columns).

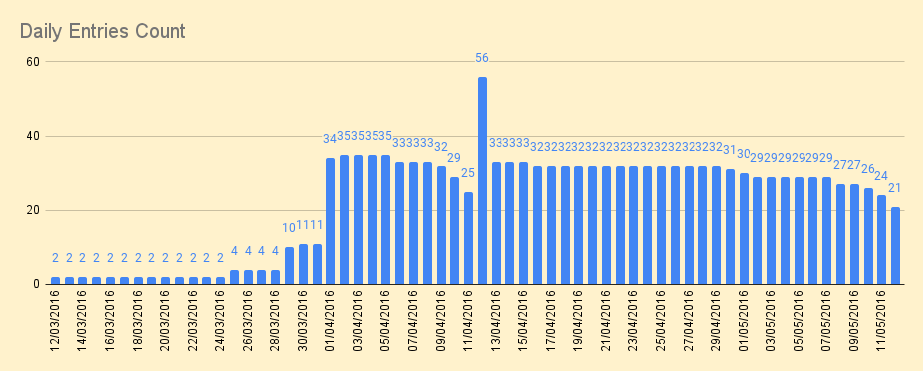
**5.(Exceeded Data Entries)**

I found that one user has 63 entries in 62 days of data.

User 4020332650 on day 12/04/2016 has two different entries.

I created a new sheet called “Data Entries Days”, to check the data entries per day. I counted data entries per day from the main sheet using COUNTIF function. The graph indicates that Tuesday has around 20 more entries than any other day of the week. I made a graph with Daily Entries Count. On the graph I found that data entries on **Tuesday** 12/04/2016 are double the entries from the rest of the days. Makes you think that could be a promotion run because the days before are slowly decreasing, but the days after that **Tuesday** the number of entries are back to similar numbers from days before.





I found that 24 User on Tuesday 12/04/16 have double data entries.

I can’t add the two data entries on each user because the number of minutes summed exceeded the minutes in a day, 1440 minutes.

On a quick view, the first entry in each user are numbers way lower in each column or near zero in general terms. To confirm that theory and delete the first entry in each user, I created a new column called Sum Minutes, adding all the minutes from the locums Very Active minutes, Fairly Active minutes, Light Active minutes and Sedentary minutes.

The number of rows with 1440 min completed is 711 rows out of 1398 total rows.The average of Sum Minutes is 1211 minutes and the median of Sum Minutes is 1440 minutes.

- “If both measures are considerably different, this indicates that the **data are skewed** (i.e. they are far from being normally distributed) and the median generally gives a more appropriate idea of the data distribution.”

- Outliers don’t have such an effect on the median. Therefore, here the median gives a more **realistic picture** of the data.

(There’s 24 users with double entries on 12/04/16.)

The Average Sum Minutes of the first rows out of 24 users is 525 Sum Minutes. The Average Sum Minutes of the second rows out of 24 users is 1206 Sum Minutes.

**Due all the measurements taken I decided to delete the first rows of each doubled data entry on Tuesday 12/04/2016.**

Now the Data Entries Each Day chart it’s steady each day of the week except on Monday that there’s a decrease of 4% below the average.

**6.** On the same sheet I calculated the max entries available and the total data entries. I see that we have only the 63% of the data possible.

Could be because some users didn't have the tracker yet or because they didn't use it those days?

The first 17 out 62 days of data, we have data from 2 - 4 users. On Friday 01/04/16 we can see the number of users increasing to around 33 users and keeping this number of users tracked until almost the end, where we can see a decrease on the last 5 days.

We can confirm on the Daily Entries Count graph that we only have the 63% of the total data entries available because users didn't have the tracker yet. Also could be that a promotion was run on Friday 01/04/16 due the increase of users and continuity on time.

**7.** I created a column with the summed minutes and I found that with the count(CountIf) of total days with 1440 ‘min’ completed divided the total rows. Only 52% of the all data entries completed the 1440 ‘min’ in a day. With an Average 1223 minutes, around 20,5 hours wearing the tracker. This means users wore the tracker for 85% of the day.

In other words, users are tracked most of the day, averaging 20,5 hours, but only 52% of the them completed a full day with the tracker.

This raises questions about why and when they remove the tracker. Is it to charge the tracker?

Further analysis of battery data could lead to product improvements and greater accuracy in data summaries for customers.

**8.** I created a new pivot table with Average Activities Intensities in MINUTES per User; Very Active, Fairly Active, Light Active and Sedentary. I added the units in hours for easier reading.

Created 2 graphs with all the information. One with each user Average Percentage in a stack column chart and the second one a pie chart with the Averaged Percentage Minutes Activity.

The data indicates that users spend an AVERAGE of 81.9% of their time on Sedentary minutes, 15.4% Light Active minutes, 1.6% Very Active minutes and 1.1% Fairly Active minutes.

This translates to an AVERAGE of 16,7 hours on Sedentary activity, approximately 3 hours on Light activity, 20 minutes on Very Active activity and 14 minutes on Fairly Active activity each day.

The users have an AVERAGE of 34 minutes of ‘Very or Fairly Active sport’ per day.

In terms of activity count, users prefer Light Intensity activities 62.7% of the time, followed by Very Active Intensity activities at 21.7%, and Fairly Active Intensity at 15.6%.

Also added charts from the main data sheet. A chart with the Average Steps per Day of the week and a chart with Average Intensity Activity in kilometers per Day of the week.

In the "Average Intensity Activity in Kilometers per Day" graph, there is a noticeable trend. On Monday, Tuesday, Wednesday, and Saturday, the volume of kilometers is higher, with Saturday being the day when people exercise the most. Sunday has the lowest activity, suggesting it is generally a rest day, even though the decrease is not extreme.

The "Steps per Day of the Week" graph confirms the trends observed in the "Average Intensity Activity in Kilometers per Day" graph.

**9.** I created a pivot table to check for Logged Activities percentage. With the total activity count and the logged activity count, I got that from all the activities registered only 2.55% were with a logged activity.

In a second graph comparing the sum of total distance and the sum of tracked distance, I found that there’s only a 35 km difference between them. This means that most of the kilometers tracked were while users were wearing the tracker.

**10.** Added a new pivot table with Average Calories and Average Total Steps per person and made a scatter plot chart.

I added a trend line to see the positive correlation between the variables.

In this same pivot table I calculated some percentages with the average steps.

10,000 is the recommended steps target that was launched by a Japanese Marketing campaign and then proven by science and medicine.([www.nuffieldhealth.com](http://www.nuffieldhealth.com/)).

I used the next guideline on steps and activity levels. ([This initiative is funded by the Queensland Government through Health and Wellbeing Queensland and Preventive Health SA, Government of South Australia.](https://www.10000steps.org.au/articles/healthy-lifestyles/counting-steps/))

* Sedentary is less than 5,000 steps per day
* Low active to Somewhat active is 5,000 to 9,999 steps per day
* Active is more than 10,000 steps per day
* Highly active is more than 12,500

*\*Each Percentage is measured for total user\**

I found that **80% of the users don’t reach the average 10,000 steps**. Around the 50% are between 5,000 – 10,000 steps and 31% of the users are below 5,000 steps.

|  |  |
| --- | --- |
| **Steps Recommended per Day** | 10,000 |
| **Users** | 35 |
| **More 12,500 steps** | 6% |
| **More 10,000 steps** | 20% |
| **Less 10,000 steps** | 80% |
| **Between 5,000 & 10,000 steps** | 49% |
| **Less 5,000 steps** | 31% |

**WeightLogInfo Data**

**1.** I **imported** the (wieghtLogInfo1 and weightLogInfo2) CSV files to Google Sheets, 2 files imported.

**2.** **(Format Consistent)** Split to columns the Date column, and remove the hours and everything that is not useful, with another split to column. After concatenate with the day first and adding a "0" after the forward slash. This way the month gets detected as a month. Copy the column and paste values only to format the column as date.

**3. (Merged)** There’s 34 rows in one sheet and 68 rows in the other sheet. Due the size of the datasets I merged them to do the data cleaning afterward.

Created a new sheet and with the QUERY function merged both spreadsheets. Selected the data, added both spreadsheets in the query function, then write the query *“select \* where Col1 is not null”* and lastly chose the number of headers, 0 in this case because I copied and paste the headers line before.

=QUERY({'weightLogInfo 1 month'!A2:H34;'weightLogInfo 2 month'!A2:H68},"select \* where Col1 is not null",0)

Then copied and pasted Values Only in a new sheet, called "weightLogInfoValuesOnly", to work within.

**4.** (**Find Blank Cells**) Selected all the data, Format tab, Conditional formatting, IS EMPTY.

In the column called Fat we have only 4 entries out of 101. Instead of deleting each row with empty cells, I deleted the Fat column because the information that we can get from there it won’t be relevant.

**5.** (**Remove Duplicates**) Data, Data clean-up, Remove duplicates, Data has header row.

2 duplicate rows found and removed. 98 unique rows remain.

**6.** **(Remove empty Rows)** Create a filter with all the data to remove empty rows, checking on each column for blanks. No blanks found within the data. I removed the extra rows at the end of the spreadsheet and removed the filter applied.

**7.** **(Trim Whitespace)** Select all the data, Data tab, Data clean-up, Trim whitespace.

No selected cells had whitespace trimmed.

**8.** (**Check Spelling**) Tools, Spelling, Spell check. Used on the headers. Units added to the headers.

**9.** Decreased decimal places in both Weight columns and the BMI column. One digit after the decimal (e.g.7.2).

10. **(Data mapping testing phase)**

- Inspected a sample piece of data to confirm that is cleaned and properly formatted.

- Spot checks in like number of nulls.

- Data validation, Conditional formatting, COUNTIF, Sorting, Filtering.

- I Summed the number of rows from both sheets using the COUNT function and compared it to the number of rows on the merged sheet.

- Removed Duplicates. Non Found.

**Analysing Data**

I created a pivot table with the merged data. Users Quantity, Data Entries, Total Days, Average Weight, Average BMI.

I created another pivot table to see the IsManualReport column (Boolean), to see the numbers for each user.

The weightLogInfo data appears to be of poor quality. The analysis of weight information cannot be considered reliable due to insufficient data, which is incomplete and lacks consistency.

Out of 13 users, only 2 have regular habits of entering weight data. There are a total of 98 data entries: 63 entries are manually reported, while 35 are automatic, but 32 of these entries belong to just one user. Only 3 users use automatic reporting, 10 add the data manually, and 22 users have no weight data at all.

Several questions arise from this analysis:

1. Why do we have such a limited amount of weight data? Are users not interested in tracking their weight, or is it difficult for them to measure it easily?
2. Could the manual and automatic reporting options be contributing to the low amount of data?
3. Could introducing a scale with Wi-Fi or Bluetooth connectivity be a viable product for users?
4. What other parameters could a scale measure besides weight?
5. Could a scale be designed differently, perhaps resembling a decorative carpet for placement next to the bed, where users typically stand at least twice a day?

These questions could guide further investigation and potentially lead to product improvements or new design considerations for weight tracking devices.

**R Programming Language**

The datasets minuteSleep1 and minuteSleep2 were cleaned and transformed using R to utilize various tools during this analysis. A PDF containing the complete analysis conducted with R on these two datasets is available.

1. In this part of the analysis with R, the merged “minute Sleep” CSV file has a total of 25 user tracked. The whole case study data has a total of 35 users as found previously in the Excel data activity.

* This means that from the whole case study data 71% of the users tracked their sleep but at the same time we can see that from the 71% tracked users there is an average of 57% of days tracked.

1. We have a range between 11 and 18 users per day tracking their sleep with a mean of 14 user per day.
2. We can see that users sleep an average of 6,5 hours and around 30 minutes in restless state. It seems just below the recommended amount of sleep, 7-9 hours.

[National Heart,Lung and Blood Institute, An official website of the United States government] (<https://www.nhlbi.nih.gov/health/sleep/how-much-sleep>)

* We can see that 3 of the users have between 20%-30% of the sleeping time in restless and awake state. But all the rest are around 90% of the sleeping time in sleeping state.
* We can say that 3 of the user are suffering from some kind sleep disorder.

**I'm unable to conduct a more extensive analysis with additional files, like heart rate, that could be analyzed using R, as the free plan restricts usage to 25 hours per month.**

**ACT**

**Daily Activity Data**

* Users wore the tracker for 85% of the day. Further analysis of battery data could lead to product improvements and greater accuracy in data summaries for customers.
* Develop notifications or activity advice to reduce the percentage of Sedentary minutes per day.
* The percentage of logged activity is very low and knowing the preferred activity by users. Introducing Light Intensity Activity exercise routines on the app may encourage users to engage in more daily activities and diversify their exercise routines.
* There is a noticeable decrease in activity on Sunday as general trend. This raise the question: could a marketing campaign promoting familiar activities increase the average kilometers and overall activity levels on Sundays?
* The analysis shows that 80% of the total users don't reach the average target of 10,000 steps per day. Implementing a point/rewards system could motivate users to achieve this goal.

**Weight Data**

* The analysis of weight information cannot be considered reliable due to insufficient data, which is incomplete and lacks consistency.

Several questions arise from this analysis:

1. Why do we have such a limited amount of weight data? Are users not interested in tracking their weight, or is it difficult for them to measure it easily?
2. Could the manual and automatic reporting options be contributing to the low amount of data?
3. Could introducing a scale with Wi-Fi or Bluetooth connectivity be a viable product for users?
4. What other parameters could a scale measure besides weight?
5. Could a scale be designed differently, perhaps resembling a decorative carpet for placement next to the bed, where users typically stand at least twice a day?

These questions could guide further investigation and potentially lead to product improvements or new design considerations for weight tracking devices.

**Sleep Data**

* Users sleep an average of 6,5 hours and around 30 minutes in restless state. It seems just below the recommended amount of sleep, 7-9 hours.

A marketing campaign highlighting the importance of adequate sleep and its benefits could improve users’ activity performance.

* Three users appear to suffer from some form of sleep disorder. Providing a report of this data to the users could encourage them to visit a doctor, potentially improving their sleep disorders.

**The final and broder question raised by this analysis is what additional metrics could be measured to improve the device?**