Data Analysis Project – bellabeat

Data Analyst: Juan Camilo Castro

Client/Sponsor: bellabeat

Purpose:

Bellabeat is a small tech company that develop products (apps, tech merch) for women, their principal market is e-commerce and want to increase their sells supplying effectible solutions for their users and bring a high customizable environment, giving to their client products that accompaniment their routines. Ms Srsen would like to know high-level recommendations for how trends can inform Bellabeat marketing strategy, using smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices.

Stake holders:

Primary

- Urska Srsen: Bellabeat's Cofounder and Chief Creative Officer.

Secundary

- Sando Mur: Mathematician and Bellabeat´s cofounder, key member of the Bellabeat executive team.
- Bellabeat's marketing analytics team — a team of data analysts.

Scope / Major Project Activities:

Activity	Description
Verify the dataset that Srsen provided	It's necessary to check the dataset that

(dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016) - https://www.kaggle.com/datasets/arashnic/fitbit?resource=download-directory&select=mturkfitbit_export_4.12.16-5.12.16	Srsen provided, verify if the data could show different trends, if this information are measurable, accurate etc.
Discover additional data	Verify if it's possible to find additional datasets that contributes another trend.
Clean data	It's relevant to clean the data that found, confirm that all data is clear and necessary for the process and analyze.
Process and analyze	Identify patterns or insights that could describe a necessity or behavior from the person who use tech technology in their daily routines.
Graphic	At the least, its important to show this results in presentation that describe the possible recommendation for the marketing department.

This project does not include:

- This project doesn't involve the "devices" that the person uses in their routine days.
- The dataset that MS Srsen provided doesn't have any category related with "genre", with that, its not possible to identify if each report is from a women or men. Consider that the company develop solutions for women exclusively.

Deliverables:

A specific list of things that your project will deliver.

Deliverable	Description/ Details
1. A clear summary of the business task 2. A description of all data sources used	Apart of these deliverables, it's necessary to develop an summary with all SQL's query, and R code that I used for these Capstone.

3. Documentation of any cleaning or manipulation of data 4. A summary of your analysis 5. Supporting visualizations and key findings 6. Your top high-level content recommendations based on your analysis	

PROCESS

- I select SQL tool, because the dataset includes 18 tables from different files, SQL offer a facility to manipulate different types of files and number of rows and columns. Also, I select Tableau public to graph the results and compare the insights.

I explored the resource and the study that Ms. Srsen provide, I Download the CSV files and uploaded them in SQL (Big Query) to take a look how the dataset was created, and understand each field in the study. However, I found the characteristics above;

- The dates in minutes and seconds rates are integer format, I used spreadsheet to separate the date in format MM/DD/YY and time with INT() and MOD(). The rows that don't respect the format was deleted.

Heart rate_seconds : Delete rows 529821

Hourly Calories: Delete Rows 13821

Hourly Intensities: Delete Rows 13821

Hourly steps: Delete Rows 13821

Sleepday: 252

WeightLogInfo: 38

- I omitted the “minutes” tables, it’s a very detailed information and we could consider the rate in hour and daily.

I continued analyze the data from each file in SQL, first of all I checked unique values with the instruction COUNT (DISTINCT Id) for all the tables, where the results are;

Daily activity: Unique values (33)

Daily Calories: Unique values (33)

Daily intensities: Unique values (33)

Sleep day: Unique values (21)

Weight: Unique values (6)

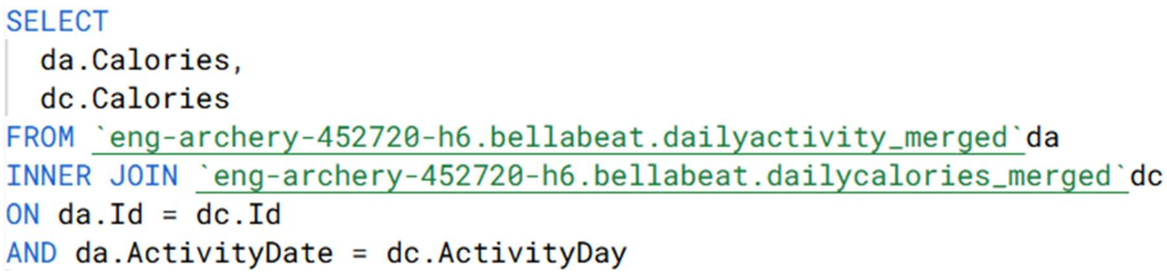
As we obtain results where the Id (users) are lower than 30 (minimum recommend sample size for preserve the confident level), these datasets was ignored by the study (Sleep day, Weight Log Info). With that, I focused on Dailys timeframe, Hourly Calories, Intensities, steps, and heartrate seconds.

ANALYZE

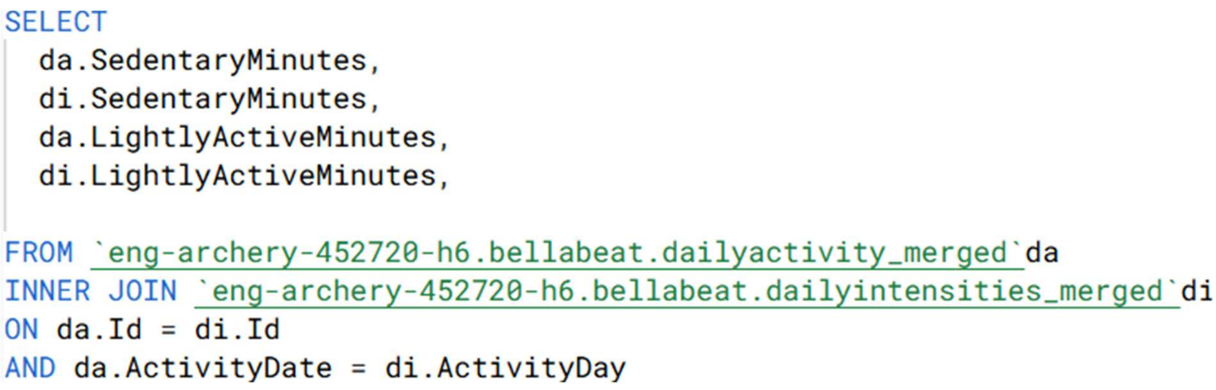
First approach by timeframe;

Daily

I compared the daily activity dataset with the rest daily datasets, looks that this dataset contains all the rest of daily datasets (calories, intensities and steps)

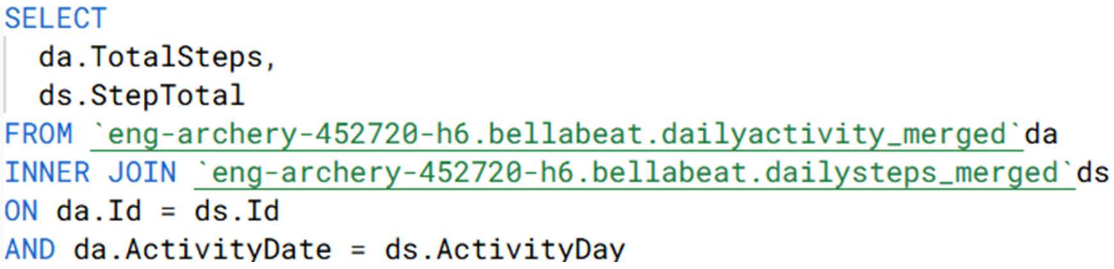


```
SELECT
  da.Calories,
  dc.Calories
FROM `eng-archery-452720-h6.bellabeat.dailyactivity_merged`da
INNER JOIN `eng-archery-452720-h6.bellabeat.dailycalories_merged`dc
ON da.Id = dc.Id
AND da.ActivityDate = dc.ActivityDay
```



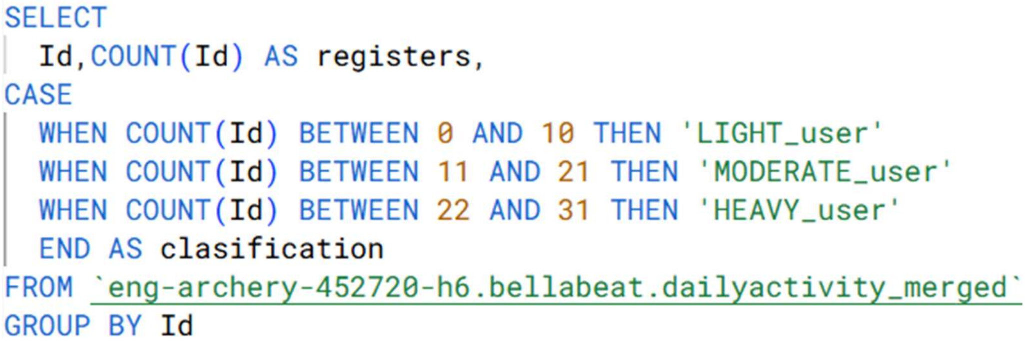
```
SELECT
  da.SedentaryMinutes,
  di.SedentaryMinutes,
  da.LightlyActiveMinutes,
  di.LightlyActiveMinutes,

FROM `eng-archery-452720-h6.bellabeat.dailyactivity_merged`da
INNER JOIN `eng-archery-452720-h6.bellabeat.dailyintensities_merged`di
ON da.Id = di.Id
AND da.ActivityDate = di.ActivityDay
```



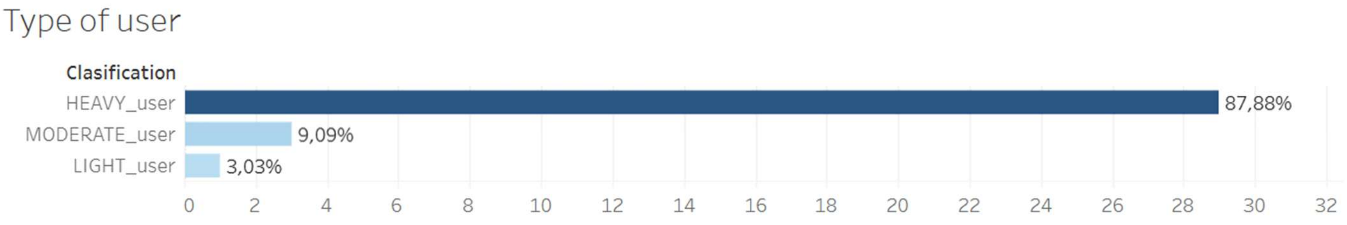
```
SELECT
  da.TotalSteps,
  ds.StepTotal
FROM `eng-archery-452720-h6.bellabeat.dailyactivity_merged`da
INNER JOIN `eng-archery-452720-h6.bellabeat.dailysteps_merged`ds
ON da.Id = ds.Id
AND da.ActivityDate = ds.ActivityDay
```

So, I explored the type of user that each Id are, with the next query;

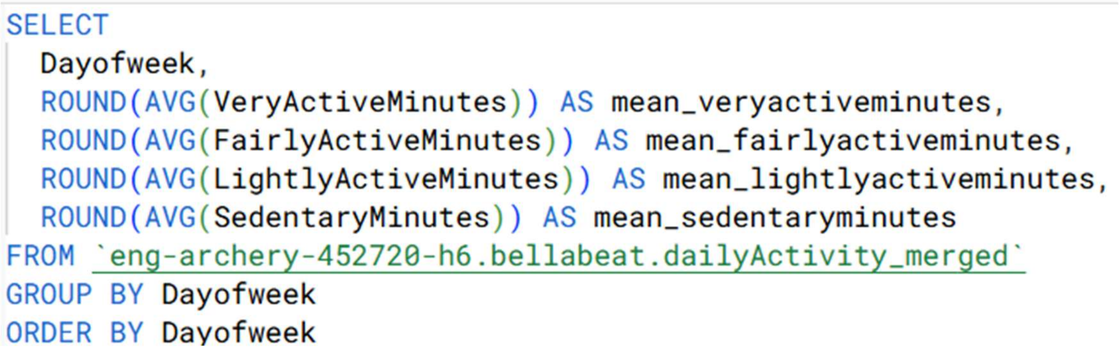


```
SELECT
  Id,COUNT(Id) AS registers,
CASE
  WHEN COUNT(Id) BETWEEN 0 AND 10 THEN 'LIGHT_user'
  WHEN COUNT(Id) BETWEEN 11 AND 21 THEN 'MODERATE_user'
  WHEN COUNT(Id) BETWEEN 22 AND 31 THEN 'HEAVY_user'
  END AS clasification
FROM `eng-archery-452720-h6.bellabeat.dailyactivity_merged`
GROUP BY Id
```

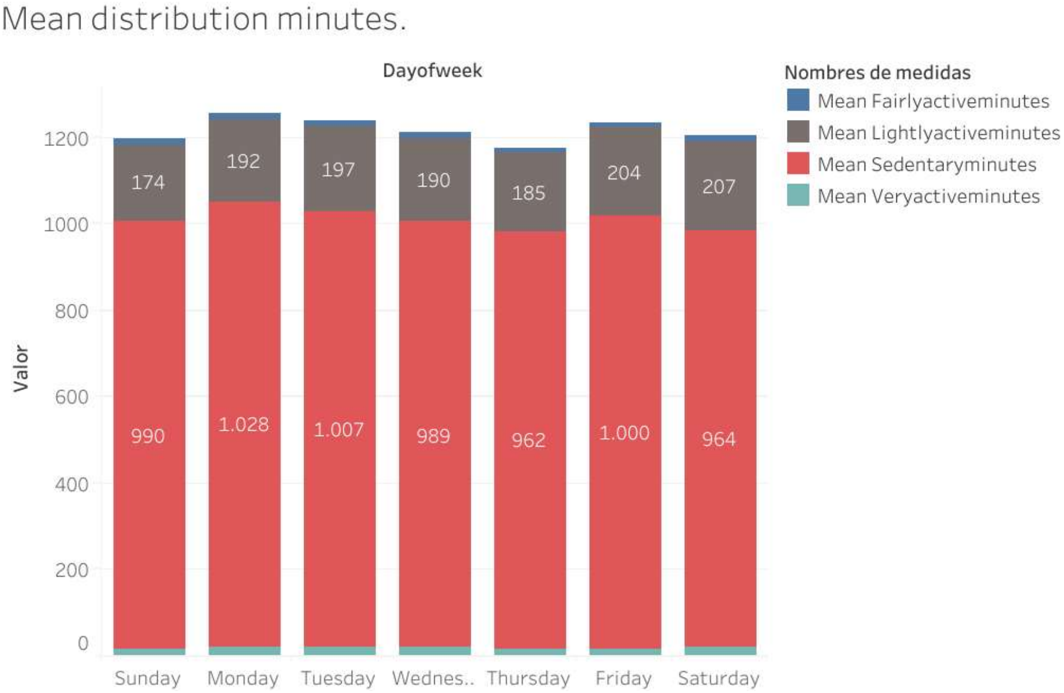
The result was exported via SQL and graph via Tableau Public;



That shows that the users of this study are in big part from heavy and moderate users. On the other hand, I checked the relation of the active minutes in the week;



```
SELECT
  Dayofweek,
  ROUND(AVG(VeryActiveMinutes)) AS mean_veryactiveminutes,
  ROUND(AVG(FairlyActiveMinutes)) AS mean_fairlyactiveminutes,
  ROUND(AVG(LightlyActiveMinutes)) AS mean_lightlyactiveminutes,
  ROUND(AVG(SedentaryMinutes)) AS mean_sedentaryminutes
FROM `eng-archery-452720-h6.bellabeat.dailyActivity_merged`
GROUP BY Dayofweek
ORDER BY Dayofweek
```

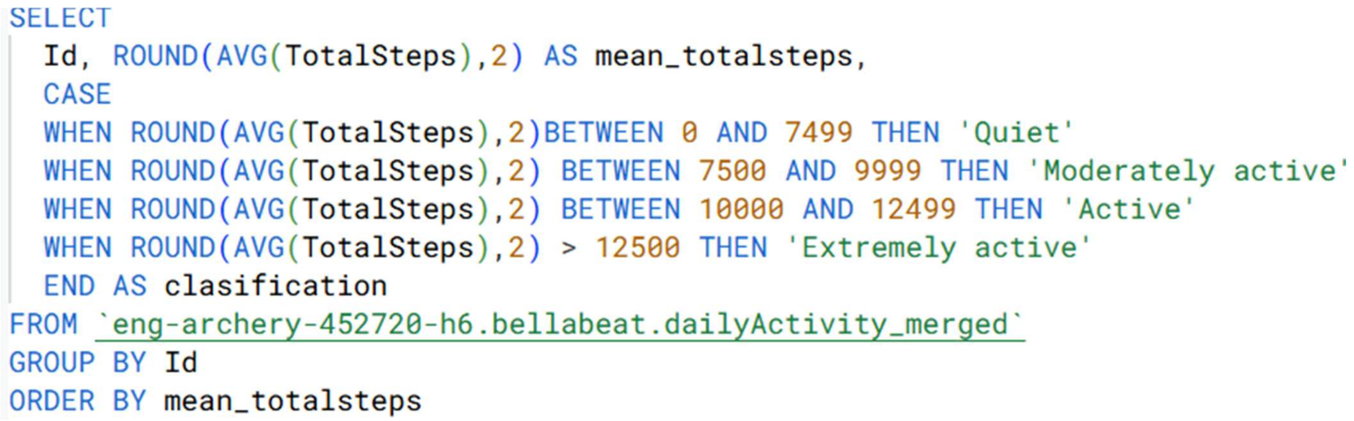


Mean distribution minutes.

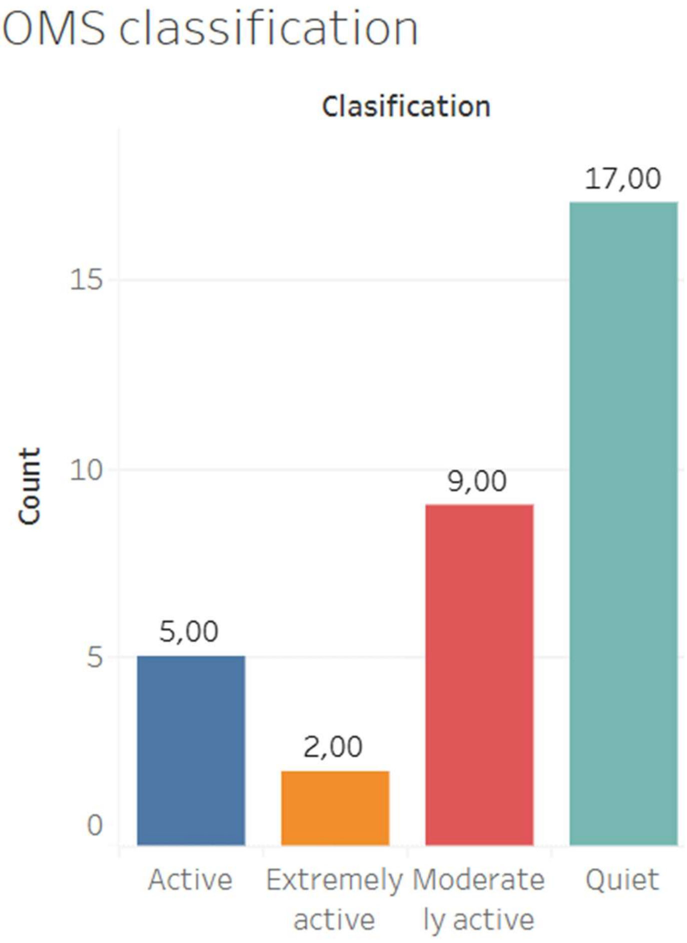
That's show that the person in the study spend more time in "sedentary" activities. It's important to check about the OMS walk recommendation, they stablish the categories above and suggest that a person must to walk 30 minutes as minimum.

- 5000 – 7499 steps per day – Quiet
- 7500 – 9999 steps per day - moderately active

- 10000 – 12499 steps per day – Active
- 12500 and more – extremely active

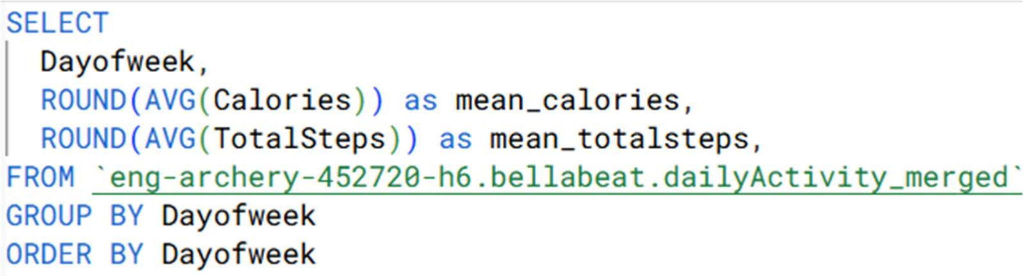


```
SELECT
  Id, ROUND(AVG(TotalSteps),2) AS mean_totalsteps,
  CASE
  WHEN ROUND(AVG(TotalSteps),2)BETWEEN 0 AND 7499 THEN 'Quiet'
  WHEN ROUND(AVG(TotalSteps),2) BETWEEN 7500 AND 9999 THEN 'Moderately active'
  WHEN ROUND(AVG(TotalSteps),2) BETWEEN 10000 AND 12499 THEN 'Active'
  WHEN ROUND(AVG(TotalSteps),2) > 12500 THEN 'Extremely active'
  END AS clasification
FROM `eng-archery-452720-h6.bellabeat.dailyActivity_merged`
GROUP BY Id
ORDER BY mean_totalsteps
```

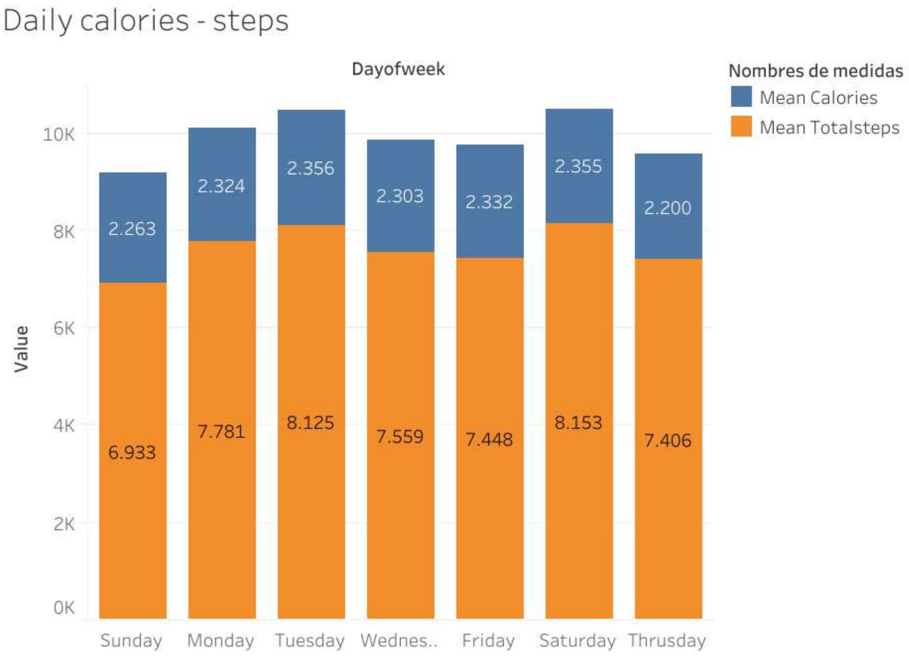


OMS classification

So, a big part of the size is person who don't walk the standard recommendation for the OMS, is a risk healthy. I wanted to know which day per week is the most "active" for the users;



```sql
SELECT
  Dayofweek,
  ROUND(AVG(Calories)) as mean_calories,
  ROUND(AVG(TotalSteps)) as mean_totalsteps,
FROM `eng-archery-452720-h6.bellabeat.dailyActivity_merged`
GROUP BY Dayofweek
ORDER BY Dayofweek
```



That shows that the days when the users are more "active" is in Tuesday and Saturday.

SHARE

Well, it's important to mention that the purpose of these study permit to understand possible necessities or insights that helps a woman routine with bellabeat products. Our woman friends, families, etc. would have a better compression about their performance, healthy and more. This study just create a starting point to understand how the IT, apps, devices could motivate to the woman to think and act in their well-being, unfortunately the data of this study doesn't provide the particular answer that we put as objective, the information it's not discriminate by genre and the insights just correspond in person who use a healthy track device. It's important to continue with the investigation and maybe create a form that just include our objective population and their behavior.

ACT

These are my high-level recommendation for how these trends can inform Bellabeat marketing strategy;

- Most users are quiet and spend a lot of their time in sedentarism activities, it's important to understand and explore how bellabeat's products could motivate their users to be more active and practice healthy habits.
- The user in the study is "heavy" and "moderate", they don't increase their healthy activities and practice good behaviors for herself although use a tool that permit the track of their daily routines. It's important to understand how the tool could be more efficient to motivate the user and use some types of alerts that remember the user to be more active.
- Tuesday and Saturday are the days that the users are more "active", maybe, the marketing team would focus on the rest days of the week, launch campaigns in the others days to increase the use of the IT devices / products.
- As I mentioned, these datasets don't fit about company market. It's necessary to make another study that shows a woman behavior and their daily routines, also, integrate the menstruating and try to understand how this affect to the woman user performance and which bellabeat products could be useful for them.