This document includes all the steps taken into the ETL process in BigQuery.

! First of all we are going to check the credability of the data, using the ROCCC acronym.! :

R – Realible – The data is uploaded by a Kaggle’s user. Thus, we should be careful in considering it appropriate sources. However, it have a license. Although, the license and that the data includes 18 different .csv files, the sample size is small and the duration of the study is less than one quarter of a years' worth of data.

O – Original – As mentioned above, the data is not uploaded by the institution, which stands behind it. However, it have a license, which approves its originality.

C – Comprehensive - Bellabeat is marketed to women; the data included in this study does not indicate the sex, age, geographical information of the participants. This could be considered as a comprehensiveness.

C – Current – Unfortunately the data is not a current one. The data is from 2016, which is now six years old-may not be relevant to the fitness device users of today (2021)

C – Cited - The data is cited, and collected by a third party.

Also, please note that we are going to validate the correctness of the data in R Studio. This could be found in the R Notebook – Cleaning. There I will load the data, check it for duplicates and so on.

With the help of R, I also found the correlation between the each daily activity from the dailyActivity\_merged.csv. The results are very important and show things that could be used in very optimistic way into our tracker. :

- The people who are walking most and record the longest distance are not the ones that burn most calories.

- The people that burn most calories are the ones that are doing extremely intensive exercise.

Outcome : We could focus on the active, intensive exercising and even promote it.

The graph could be found in the image “corr.png”

1.Steps

- Loading the Steps data into BigQuery.

- Counted the total days in which the trackers were counting steps – 940

SELECT COUNT(StepTotal) FROM `ninth-matter-317418.12345678.steps`

- Counted the total days in which the trackers monitored more than 10k steps per day – 303

SELECT COUNT(StepTotal) FROM `ninth-matter-317418.12345678.steps`

WHERE StepTotal > 10000;

Outcome : only 1/3 of the days the users are reaching the minimum steps, which the doctors are advising for normal way of life and good health condition.

! This could be used in a commercial, where the data shows that more than half of the people are inactive. Also, it is good to be noted that this conclusion is drawn from people that already have taken the first step in their healthy life (buying a tracking gadget). This means that the real proportion of people, who not only do not look after their health, but also do not monitor their steps is significantly bigger. !

https://www.healthline.com/health/how-many-steps-a-day

2. The mood of the distances :

It is very important to know how the users are taking their steps. For instance, if a user is walking slowly, running, or jogging. The result of this separation will be beneficial to the content of our ads. If we notice more runners within this survey, it could means that the majority of people using those trackers prefer to run. In that case we can emphasize it in our commercial.

! Please note that the resulting tables will be saved so later they can be used in Tableau, where the appropriate visualizations will be developed. !

- First lets produce the table :

SELECT Distinct(id),

SUM(VeryActiveDistance) as very\_active,

SUM(ModeratelyActiveDistance) as moderately\_active,

SUM(LightActiveDistance) as Light\_active

FROM `ninth-matter-317418.12345678.activity`

GROUP BY id

- Next we are going to use Google Sheets, to write an statement in ending column, where based on the peace with which a user is taking his/her steps, we will conclude if he/she is taking it more or less intensively.

The Google Sheet formula is : =IF(AND(B2>C2, B2>D2), 1, 0), and respectively similar for the other two columns. The added columns are named : “very”, “moderately”, and “light”. Each of them represent the peace of an user. If he/she falls in any of those category the number 1 is present, otherwise 0. This techniques is called “dummy variables”.

! I will add fourth column for Quality Assurance. This column will calculate the previous 3. If a number higher than 1 is present in any of the cells it will meant somewhere we have an error. !

- Everything seems alright. We can save the data frame and use it later in our Tableau dashboard.

3.Activity

Here we are going to find how much of the users are actually doing high intensity training.

This is very important to understand the behaviour of the customers, as a lot of people are buying similar products hopping that they will start to look after their health more often. However, the majority of them are still procrastinating even after buying a tracker. The outcome from the proving of this hypothesis (whether or not the people are actively exercising) is important for future life of our company’s product, as well. This is because, if we prove the hypothesis that the people are not using them for training purposes, than for ordinary monitoring we can focus to purpose in our own product. Otherwise, if we reject the hypothesis we can focus on developing more features directed to the active sports and claim them in our advertising.

- First we are going to see how many people are intensively exercising more than 60 minutes, in a day. Here is very important to count the distinctive users, as in the table we have many records from a sing user :

SELECT COUNT(Distinct(Id)) FROM `ninth-matter-317418.12345678.activity`

WHERE VeryActiveMinutes > 60

Result : 18

- This is not that bad, bearing in mind the overall number of the users in this survey is – 33

SELECT COUNT(Distinct(Id)) FROM `ninth-matter-317418.12345678.activity`

Thus, we do not have 80 % dominance to either side. This mean we cannot take a decision in approving or disapproving our hypothesis. We are going to use both aim/goals in our advertising . Both for people who are actively doing exercises and those that are just trying to keep appropriate health.

4. Sleep .

This is very important as our tracker will also monitor the time during the user is asleep.

This time the cleaning and analysis is happening directly into Tableau.

- First we are going to see whether or not the majority of the users are not taking off their tracker while sleeping. This was done not only with loading the appropriate data into Tableau, but also with doing some calculation within the software. For instance, I had to transform the minutes into hours, as they were present only in minutes in the data frame. :

COUNT([minuteSleep\_merged.csv]) / 60

- Next, I have done another report, with the average of the sleeping hours per day.

COUNT([minuteSleep\_merged.csv]) / 60 / 30

The difference in the code is not that big, but the difference in the graphs is significant. For instance, the first one shows that the majority of the people actually, never take off their trackers even during a sleep. This is good, because now we know they appreciate this functionality. The second graph, however, shows that only two users are having more or equal to 8 hours of sleep which is the recommendation from the doctors. This mean the majority of people have to improve their sleeping routines. The first step toward this goal is to use a tracker. This worth to be emphasized in the commercial for our own product, as well.

5.Weight

We are going to add the weight data to see if any of the users succeeded to drop a few kilos.

This will be done in google spreadsheets.

- To find whenever someone succeeded of not, I used a formula in a new column. The formula compares their input in the beginning and from the last record.

=IF(C2<C3, TRUE, FALSE)

- After that a column for the TRUE cases (when someone succeeded to lower their weight), with the total of these occurrences : 25

- After that a column for the FALSE cases (when someone did not succeed to lower their weight), with the total of these occurrences : 42

6.Process :

Before continuing to tableau and the design of the dashboard, we are going to transform the data that we will work with a little bit more :

- The steps data set – I want to know the average steps every user is doing :

SELECT DISTINCT(Id) as user,

AVG(StepTotal) as avg\_steps

FROM `ninth-matter-317418.12345678.steps`

GROUP BY Id

- I also want to see the overall average step taken per day :

SELECT AVG(StepTotal) as avg\_steps

FROM `ninth-matter-317418.12345678.steps`

The result is 7637, which is under the threshold for a normal healthy life

- I will also use the heart rate data. Before that, we have to change the date format in excel, because BigQuery do not allow to upload the data set in its raw format.

This is the biggest data in FitBit dataset. This is because the records are for every single minute for each person in the duration of days. To take the exact number we need to know we should use the following querry :   
  
SELECT Id,

count(distinct(Time)) as days ,

count(Value) as heart\_beats,

count(Value) / 1440 as beats\_per\_minute FROM `ninth-matter-317418.12345678.heart\_rate`

WHERE Time = "2016-04-22"

group by ID

Unfortunately, the result is not satisfying. I was trying to find how many beats per minute in a day which almost all the users wore their devices, was present. Although the complex querry, the result is bad. Around 4/5 beats per minute, which is not logical nor normal. The only logical explanation is that we already have a flaw in the FitBit trackers. They do not record accurate the heart beating of a person. This is something that we can use. First of all to be sure that a correct monitoring is present in our devices. Second of all to show a new functionality that seems not present in our exisiting competitors.