bellabeat

We are going to follow the Google data analysis methodology in order to reply the different stakeholders needs:

1) Ask:

In order to help Bellabeat influence digital strategies, we will analyze smart device usage data for non-Bellabeat customers to look for trends that could be applied to our company customers.

2) Prepare:

Our data is stored in **Kaagle** which is a website for data scientist studies and researches. It's about **18 csv files** that contains personal fitness tracker from thirty 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. We don't have enough information about the data bias, it's only a **30-person records** and we don't know if they are randomly selected or not, they are all sports professionals or not, male, female, we don't have enough information to verify if our data is biased and credible.

The data is in public domain at Kaagle so that anyone can use it for learning or other purposes. It's about daily, activity, calories, intensities, and steps, sleep also per hour, minutes and heart rate by seconds.
We can look for the daily/sleep activities of the persons in order to answer the questions. There are many problems in the data there are missing values and duplicates for different persons. For simplicity reasons, the following CSV files were focused on:

- dailyActivity_merged.csv
- sleepDay_merged.csv

3) Process:

For the analysis we used Big Query SQL and Google sheets.

First, the CSV files have been uploaded to Google sheets, I had an overview of the data, by looking on columns stats and checking IDS integrity (string of 10).

Next I uploaded the files to Big Query. I have created dataset digital-maker-329900.Bellabeat and I used the following queries to verify the total number of users per table:

```
SELECT DISTINCT Id FROM digital-maker-329900.Bellabeat.DailyActivities
```

```
SELECT DISTINCT Id FROM digital-maker-329900.Bellabeat.DSleep
```

Results:

- dailyActivity_merged.csv = 33 unique user IDs
- sleepDay_merged.csv = 24 unique user IDs

We can see that 24 user IDs are consistent across both tables due to some feature not being used.

4) Analyze:

I tried to look for this users habits, so i used the following queries to get the average of the different features from both tables grouped by ID :

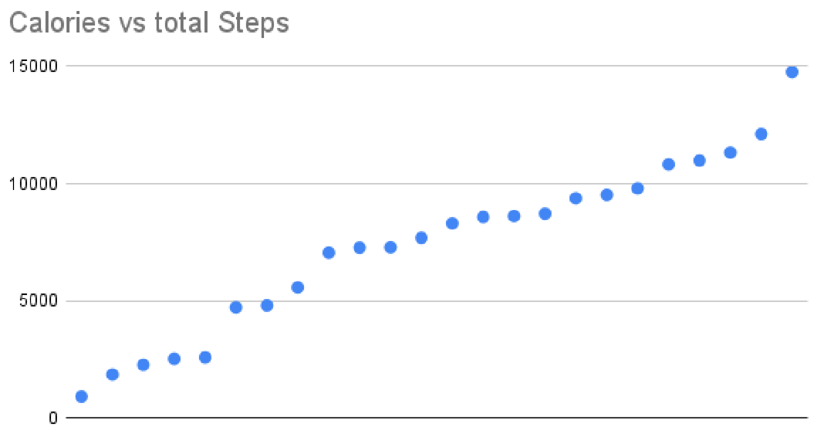
```
-- We will first select the average of steps and distance per person

            SELECT avg(TotalSteps) AS Tsteps,
            avg(TotalDistance) AS Tdistance,

-- Then we will select the average of active/moderately and lite distances per person

            avg(VeryActiveDistance) AS VeryActdis,
            avg(ModeratelyActiveDistance) AS MordActdis,
            avg(LightActiveDistance) AS LightActdis,

-- Now we will select the average of its minute per person

            avg(VeryActiveMinutes) As VeryActmin,
            avg(FairlyActiveMinutes) AS MordActmin,
            avg(LightlyActiveMinutes) AS LightActmin,

-- Here we will add the average of TotalMinutesAsleep and TotalTimeInBed in hours and the
minutes in bed with no sleep per person

            avg(dslep.TotalMinutesAsleep/60) AS TminSleep,
            avg(dslep.TotalTimeInBed/60) AS TminBed,
            avg(dslep.TotalTimeInBed -dslep.TotalMinutesAsleep) AS minNoSleep,

-- We will finally select the average sedentary hours and calories per person

            avg(SedentaryMinutes /60) AS SedentaryHours,
            avg(Calories)  AS AvgCalories
            FROM `digital-maker-329900.Bellabeat.DailyActivities` AS dact
            INNER JOIN  `digital-maker-329900.Bellabeat.DSleep` as dslep
            ON dact.Id = dslep.Id
            GROUP BY ((dslep.Id))
```

A simpler and more efficient way of analyzing this data is to save the results to a Google Sheets document.

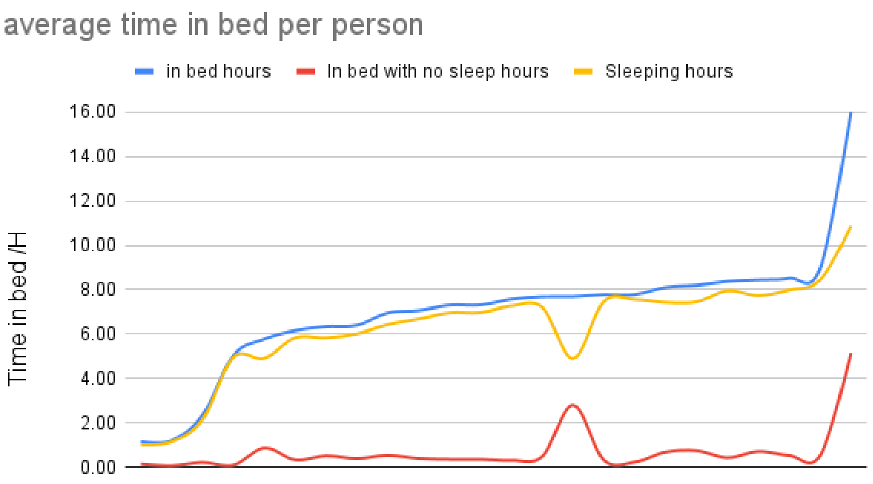
The spreadsheet consists of **24 rows** and **14 columns** showing averages for all **24 users** in the categories of the average of steps and distance, active/moderately and lite distances and its minute, TotalMinutesAsleep and TotalTimeInBed in hours, the minutes in bed with no sleep, the average sedentary hours and calories.

5) Share:

- Tsteps vs. Calories: More distance equaling more calories is not as accurate as one may think. Users inputting weight, and logging activities will provide more accurate results.



- Time in Bed while no sleeping: by subtracting the time in bed and the time of sleeping we found that users spend an average of 42 minutes in bed which can be important to deliver more features for them.



6) Act:

- Simplify the process where the user is required to enter information (activities, weight, etc.)
- Show the benefit of entering information will provide more accurate results for the user.
- Market the product on the key features and benefits that consumers are using. Steps, total distance, calories, and sleep

Limitations:

- Having a larger sample size could provide better insights
- Having demographics like age and gender/sex can provide more detail of the data
- The participants volunteered and were not randomly selected
- All variables were self-reported/automatically collected through their trackers which could have led to over reporting or underreporting their experiences