Data description

1 Data general information

The dataset used is available on the Kaggle platform, a site internationally recognized for offering open datasets intended for research, predictive model development, and academic studies. This dataset is titled "FitBit Fitness Tracker Data" and is published by the user Möbius.

The dataset contains information collected through Fitbit devices and records physical activity data from 30 individuals. Participants monitor their performance in terms of daily steps, calories burned, minutes spent at different activity intensity levels (light, moderate, vigorous), hours of sleep, and periods of inactivity. Each user is represented by a unique identifier, thereby protecting their personal identity and ensuring anonymity within the sample. The data consists of daily records gathered between April and May 2016.

The author, Möbius, is an active contributor to Kaggle who shares various datasets primarily related to data science applications in health and personal wellness. Although no specific information is provided regarding their real identity or institutional affiliation, their consistent participation and the positive reception of their work within the Kaggle community add credibility to their contributions.

Regarding usage rights, the dataset is released under the Creative Commons CC0 (Public Domain Dedication) license. This license means that anyone is free to use, modify, distribute, and share the dataset without the need for attribution or prior permission from the original author. This policy promotes broad access for educational, research, commercial, or personal purposes.

The table below evaluates the "FitBit Fitness Tracker Data" dataset based on the ROCCC criteria.

ROCCC	Comments
Reliable	The data was collected through a distributed survey and not directly from Fitbit, which may affect its reliability.
Original	It is not an original dataset provided by Fitbit; it was generated by users based on their own device data.
Comprehensive	The sample includes only 30 individuals, limiting the representativeness and breadth of the information.
Current	The data was collected in 2016 and may not reflect current behaviors or trends.
Cited	No additional verifiable sources or formal data collection methodology are detailed, aside from mentioning that the data comes from a survey.

The table analysis shows that the "FitBit Fitness Tracker Data" does not fully meet the ROCCC criteria. However, this dataset will be used as it is the recommended one for the project and provides a useful foundation for analysis, despite its limitations.

2 Data

This dataset is divided into two sets, each containing data for one month. Each set includes the following CSV files:

Table Name	Description	In two tables?
dailyActivity_merged	Traking daily, Steps, Distance, Intensities, Calories	Yes
dailyCalories_merged	dailyCalories	No
dailyIntensities_merged	Daily Calories dividing in 4 categories	No
dailySteps_merged	Daily Steps	No
heartrate_seconds_merged	Heartrate	Yes
hourlyCalories_merged	Hourly Calories	Yes
hourlyIntensities_merged	Total and average intensity	Yes
hourlySteps_merged	Hourly Steps	Yes
minuteCaloriesNarrow_merged	Calories burned every minute	Yes
minuteCaloriesWide_merged	Calories burned every minute	No
minuteIntensitiesNarrow_merged	Intensity counted by minute	Yes
minuteIntensitiesWide_merged	Intensity counted by minute	No

Table Name	Description	In two tables?
minuteMETsNarrow_merged	Ratio of the energy you are using in a physical activity compared to the energy you would use at rest	Yes
minuteSleep_merged	Log Sleep by Minute	Yes
minuteStepsNarrow_merged	Steps tracked every minute	Yes
minuteStepsWide_merged	Steps tracked every minute	No
sleepDay_merged	Daily sleep	No
weightLogInfo_merged	Weight track	Yes

Certain tables are not present in the March–April dataset. First, the daily data is missing, but we can calculate it using the minute-level data. Second, the "Wide" data is missing, so I will use only the "Narrow" data instead.

To verify that the daily data and hourly data match, pivot tables in Excel were used. The dates were set as rows, and calories and steps were set as values. From this, the average percentage differences were calculated. The summary table is shown below.

	Calories	Steps
Max % Difference	4,719%	9,303%
Min % Difference	0,003%	0,013%
Avg % Difference	0,708%	1,629%
% Difference of the Total	0,613%	1,463%

As seen in the table, the differences are not significant enough to conduct a separate analysis for each data set. For this reason, only the data from the hourly tables will be used, as they are present in both files.

In the case of the intensity table, it represents different data in the daily data table and the hourly data table, which is why separate analyses will be conducted. However, the daily table only exists for the May and April files, so only this data will be analyzed.

It was also checked whether the Narrow and Wide formats contain the same data. An attempt was made to apply the same analysis, but in this case, the percentage differences turned out to be significantly higher. In other contexts, it would be advisable to consult the data source to understand the reason behind these discrepancies. However, for the sake of practicality and alignment with the project's goals, only the Narrow format data will be used, as it is available for both months.

Additionally, the daily activity table summarizes the information from the daily calories, daily intensity, and daily steps tables, so it was decided to work exclusively with the daily activity table.

On the other hand, I believe that minute-level data does not provide particularly relevant information for a marketing campaign, except in certain specific cases. For instance, intensity data offers different insights compared to the hourly data. Additionally, the METs and Sleep tables are only available in minute-level format. Finally, the CSV files to be used are presented below.

Table Name	Observation
dailyActivity_merged	
hourlyCalories_merged	
hourlyIntensities_merged	
hourlySteps_merged	
minuteIntensitiesNarrow_merged	
minuteMETsNarrow_merged	
minuteSleep_merged	
sleepDay_merged	Only with April-May data
weightLogInfo_merged	

There were issues when importing the time-based tables into BigQuery because the data was in a 12-hour format, which BigQuery did not recognize. To resolve this, the format was converted to 24-hour time in Excel. This was done using functions such as splitting columns, CONCATENATE, and TEXT.

An attempt was made to find additional information from other datasets, particularly regarding water consumption and menstrual cycle tracking, which are not covered in this dataset. However, no free and sufficiently reliable datasets were found that collect this type of data from smart devices

3 Conclusions

- The dataset has some level of recognition on the Kaggle platform and by the user who published it.
- The dataset does not fully meet the ROCCC criteria, due to limitations such as the small sample size of only 30 users and the data being quite outdated.
- Data from both the March-April and April-May files will be used.
- Only the data considered useful for identifying trends relevant to the company's marketing strategy will be utilized.
- The dataset does not cover certain features of some products, such as menstrual cycle tracking and hydration.