Data Dictionary

Task: I will analyze the columns in each file individually to determine how they align with the business task and whether they can answer the question at hand.

Primary Datasets for Analysis

dailyActivity_merged.csv: This appears to be a merger of different columns from other tables. This is the main table I will be using for analysis.

- Note: I need to change the file name after uploading it to R
- Columns:
    - Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs. There are 33 Unique Users.
    - ActivityDate: States the date when each data was collected—multiple entries for the same user.
    - TotalSteps: Total number of steps per data entry (daily), not unique ID. The step goal for an average adult is 7,000 to 10,000 steps per day.
    - TotalDistance: Total distance walked daily. No unit (miles or km) mentioned, but I assume it is km. It is the sum of very, moderately, light and sedentary distances. Will need to validate this.
    - TrackerDistance: Same as TotalDistance. Not needed.
    - LoggedActivities: Not much data here. I don't know what it's for. Not needed for analysis. Will remove it.
    - VeryActiveDistance: Distance covered during rigorous exercise. Shows dedicated exercise time. For example, running and high-intensity workouts.
    - ModeratelyActiveDistance: Distance covered during moderate exercise. Shows intentional but not intense activity. For example, light jogging, brisk walking, leisurely cycling, etc.
    - LightActiveDistance: Distance covered during light daily activities. Shows general daily movement patterns. For example, slow walking, household chores, and casual movement.
    - SedentaryActiveDistance: Distance covered during minimal movement. Shows how much users move even during sedentary time, which in this case is almost no movement. It doesn't give much information. Will remove it. For example, moving from desk to printer, short walks, getting coffee, etc.
    - Note: Total distance per user per day will be the sum of these last four columns.
    - SedentaryMinutes: Minutes spent sitting or lying down with minimal movement. High sedentary time = health risks to address. There are 1440 minutes, so the sum of sedentary, lightly, fairly, and very active minutes should not surpass this number. If this value is 1440 minutes (and other values are 0), there is a chance the watch was not worn. I should search up the average time a person spends doing nothing a day.
    - LightlyActiveMinutes: Number of minutes spent doing light activities/exercises.
    - FairlyActiveMinutes: Minutes of moderate exercise. Shows intentional exercise habits.
    - VeryActiveMinutes: Minutes of vigorous exercise. Indicates fitness enthusiasts.
    - Calories: Most likely the amount of calories burnt daily, instead of the amount consumed, as this is mainly a fitness tracker and has no mention of calorie tracking regarding food intake. If the value is 0, I will remove the data because it is likely that the data wasn't

imputed that day. The number of calories you burn in a day varies, but averages around 1,300–2,400 for an inactive person, influenced by factors like age, sex, height, weight, and muscle mass. Anything more than 2400 can mean exercise. Anything less than 1300 can be seen as an error. Or perhaps the column can be interpreted as calories burned during intense exercise, but that is very unlikely.

DailyCalories_merged.csv:
Shows the amount of calories burned daily. I will most likely use this as opposed to the hourlyCalories as it aligns with the daily data in the dailyActivity_merged table.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityDate: States the date when each data was collected—multiple entries for the same user.
- Calories: Most likely the amount of calories burnt daily, instead of the amount consumed, as this is mainly a fitness tracker and has no mention of calorie tracking regarding food intake. If the value is 0, I will remove the data because it is likely that the data wasn't imputed that day. The number of calories you burn in a day varies, but averages around 1,300–2,400 for an inactive person, influenced by factors like age, sex, height, weight, and muscle mass. Anything more than 2400 can mean exercise. Anything less than 1300 can be seen as an error. Or perhaps the column can be interpreted as calories burned during intense exercise, but that is very unlikely.

dailyIntensities_merged.csv: Shows the minutes users spend daily on exercise and the distance they travel during their time moving. I will use this and combine with hourlyIntensities if the number of unique IDs matches, as it includes data on the intensity values.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityDate: States the date when each data was collected—multiple entries for the same user.
- SedentaryMinutes: Minutes spent sitting or lying down with minimal movement. High sedentary time = health risks to address. There are 1440 minutes, so the sum of sedentary, lightly, fairly, and very active minutes should not surpass this number. If this value is 1440 minutes (and other values are 0), there is a chance the watch was not worn. I should search up the average time a person spends doing nothing a day.
- LightlyActiveMinutes: Number of minutes spent doing light activities/exercises.
- FairlyActiveMinutes: Minutes of moderate exercise. Shows intentional exercise habits.
- VeryActiveMinutes: Minutes of vigorous exercise. Indicates fitness enthusiasts.
- SedentaryActiveDistance: Distance covered during minimal movement. Even sedentary time has some movement. Hardly any data, so it's safe to remove this column.
- LightActiveDistance: Distance from light activities. Do the lightlyActiveMinutes and distance correlate? The distance travelled for the same minutes spent may be different, depending on the activity they did. Some may focus on running, but others on weight lifting or jump roping, for example.
- ModeratelyActiveDistance: Distance from moderate exercise. Do the moderatelyActiveDistance and fairlyActiveMinutes correlate?
- VeryActiveDistance: Distance from vigorous exercise. Do the veryActiveMinutes and distance correlate?

- Questions to ask for analysis:
    - What's the average sedentary time? (Bellabeat can target reductions)
    - Do users prefer light, moderate, or intense activity? (Product features)
    - Are distance and time metrics consistent? (Data validation)
    - How do these patterns relate to sleep and calories? (Holistic wellness)

dailySteps_merged.csv: Shows the total number of steps taken on a specific day. Data is included in the dailyActivity merged. I will most likely use this as opposed to the hourlySteps as it aligns with the daily data in the dailyActivity_merged table.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityDay: States the date when each data was collected—multiple entries for the same user.
- StepTotal: Shows the total number of steps taken per day. The average number of steps globally is 5000 daily. Anything less than 1000 should be removed (consider). It can mean that they did not wear the device all day, or they truly stayed in bed all day for whatever reason.

hourlyIntensities_merged.csv: Shows the hourly intensity of users as well as the average intensity, which is the hourly intensity divided by 60 (minutes). Includes values for intensities when the dailyIntensities table does not, so I can use it if all the numbers of unique IDs match.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityHour: States the date AND timestamp for each activity, on an hourly basis. I need to find out how many unique IDs there are.
- TotalIntensity: Sum of intensity values for that hour. Intensity typically represents activity level on a scale of 0-100, where: 0: Completely sedentary/sleeping, 1-39: Light activity (casual walking, household chores), 40-59: Moderate activity (brisk walking, light exercise), 60-100: Vigorous activity (running, intense workouts).
- AverageIntensity: Average intensity level during that hour. I will find peak intensity hours. If both TotalIntensity and AverageIntensity are 0, it's safe to assume that the user is asleep, of course, if the number of 0 entries is consistent.
- Better Than Just Steps Data: Steps alone don't distinguish between casual walking and running. Intensity data tells you the quality of movement. Combined, they give a complete picture. I will see if the number of unique users for both tables is consistent.

sleepDay_merged.csv: Shows the sleep sessions and length for each unique user daily. I will use this combined with the minutesSleep_merged data, as that table shows the quality of sleep (if the number of unique IDs is consistent).

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- SleepDay: States the timestamp for each activity, daily. I need to find out how many unique IDs there are.
- TotalSleepRecords: Number of sleep records/sessions for that day. Typically 1 (one night's sleep), but could be more if naps are recorded.
- TotalMinutesAsleep: Shows the total minutes the user spent asleep daily.

- TotalTimeInBed: Shows how long the user spent in bed daily, inclusive of when they were asleep.

Secondary Datasets for Potential Use

Heartrate_seconds_merged.csv: Shows the heart rate of each unique user after 5 seconds. This is the only table with heart rate data, so I will consider using it.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- Time: States the date and timestamp when each data was collected—multiple entries for the same user. Records value after 5 seconds (so it's very long…). I will try to find out how many unique users are in this table, and if I can find the mean heartbeat for users. Though I'm not quite sure how to find out how to group the heartbeats of the users throughout the day. I will separate date and time.
- Value: I believe it shows the heart rate in beats per minute (BPM) at that specific 5-second interval. Typical range: Resting 60-100 BPM, exercise up to 180+ BPM.
- Key Business Questions for Bellabeat:
    - What are typical resting heart rates? (Indicates cardiovascular health)
    - When do users exercise? (Heart rate spikes indicate workout times)
    - How stressed are users? (Consistently high heart rate might indicate stress)
    - Sleep quality correlation? (Low overnight heart rate = better sleep)
- Important Considerations:
    - Sample Size: Heart rate monitoring might have fewer users than step tracking
    - Data Quality: Some users might not wear devices consistently
    - Privacy: Heart rate is sensitive health data - Bellabeat must handle it carefully
- Value: This heart rate data is extremely valuable for Bellabeat because it provides insights into cardiovascular health, stress levels, exercise intensity, sleep quality, and overall wellness beyond just activity.

minuteMETsNarrow_merged.csv: Contains Metabolic Equivalent of Task (MET) values at a minute-by-minute level. This is actually one of the most scientifically meaningful datasets for understanding activity intensity. I will do more analysis and research on this to see if it can be used.

- MET Definition: MET = Metabolic Equivalent of Task. 1 MET = Energy expended while sitting at rest. Scale: 1.0 (resting) to 10+ (very vigorous activity).
- Value: This is a one-minute-level dataset that might be worth using because METs provide unique scientific insights not available in other datasets. Health guidelines use METs (30 minutes of ≥3 MET activity daily). I can aggregate to hourly/daily for high-level insights. It adds medical credibility to your Bellabeat recommendations.
- Id: Shows the unique user identifier. I need to check how many there are.
- ActivityMinute: Displays the timestamp for each minute.
- METs: Shows the Metabolic Equivalent value for that minute.
- Note: If using this dataset, I will try to identify how it is used. There is a severe limitation to this dataset, as a lot of the MET values are 10 or higher, even during the time that most people are asleep. I will check to see if it is the same user who has the MET value constantly at 10.

minuteSleep_merged.csv: Contains data on the quality of sleep an individual has every minute. I may use it as it contains data on the sleep state of individuals, which is useful for analysis.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- date: States the date AND timestamp for each activity, on a minute basis. I need to find out how many unique IDs there are.
- value: Shows the sleep state at that time. Typical values: 1 = Asleep, 2 = Restless, 3 = Awake. This is the most important column - it shows sleep/wake patterns.
- logId: Shows the unique identifier for each sleep session. It groups all minutes that belong to the same sleep period. It is useful for analyzing complete sleep sessions from start to end.
- Plan: I will start with daily sleep data for high-level trends and use minute data for 1-2 key insights about sleep quality. I will aggregate minute data to the session-level for manageable analysis.

Datasets Not Planned for Use

hourlyCalories_merged.csv: Shows the hourly calories burned per day per unique ID. I may not use, as we have dailyCalories data.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityHour: States the date AND timestamp for each activity, on an hourly basis. I need to find out how many unique IDs there are.
- Calories: Shows the amount of calories burned every hour. I will need to group it. The average calories burned in an hour vary greatly depending on the activity, with a 160-pound person burning roughly 314 calories walking at 3.5 mph and up to 606 calories running at 5 mph. For a 170-pound person, standing burns about 186 calories per hour, while a moderate-paced walk burns around 324 calories. Factors like weight, age, and muscle mass significantly impact the exact number of calories burned, with heavier individuals generally burning more. It will be hard to determine the average calories per individual then because we have no data on the sex, age, muscle mass, etc. I will not use it as there are only 8 unique users. I could find the mean of the calories burned during each hour of the day for all participants to be able to see if there is a trend there. I can check to see if users have a consistent pattern daily. Is there a pattern with the steps data and the calories burned?

hourlySteps_merged.csv: Shows the number of steps a user takes hourly. I may not use as we have dailySteps data.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- ActivityHour: States the date AND timestamp for each activity, on an hourly basis. I need to find out how many unique IDs there are.
- StepTotal: Shows the total number of steps the individual takes during that hour. If steps are 0, the user is either sitting or asleep. I will find the average number of steps a user takes daily. An

average of more than 5000 can be seen as a highly active person. Any hourly step more than 5000 should be treated as exercise. I will consider combining with the hourlyIntensities_merged data, as it better tells the kind and quality of activity the user is doing, and can help produce better insight for Bellabeat. Intensity data tells you what kind of activity users are doing: Low intensity + high steps: Casual walking throughout the day, High intensity + moderate steps: Short, intense workouts, Variable intensity: Mixed activity types. This is crucial for Bellabeat because it helps understand not just how much users move, but how they move - which is essential for designing effective wellness products for women.

minuteCaloriesNarrow_merged.csv: Shows the number of calories burned every minute. I do not think I will be using this, though. In case I am, I will identify the number of unique users first. I will not be using it, as we have both daily and hourly data on calories.

minuteCaloriesWide_merged.csv: I will not be using this dataset either. This is the wide-format version of the minute-level calories data. It doesn't seem that the number of unique users in both datasets is equal, hence, limiting its use. I will not be using it.

minuteIntensitiesNarrow_merged.csv: Shows the intensity of the user every minute. There is no need for this data, as the hourlyIntensity data is enough.

minuteIntensitiesWide_merged.csv: This is the wide-format version of the minute intensity dataset. I will not be using this data.

minuteStepsNarrow_merged.csv: Shows the number of steps a user takes every minute. I may not use it as the hourlySteps_merged dataset is enough.

minuteStepsWide_merged.csv: This is the wide-format version of the minute steps dataset; however, I will not be using it, as the hourlySteps dataset is sufficient.

weightLogInfo_merged.csv: Shows the data regarding the weight of users daily (it is supposed to anyway). The data log for every participant is not consistent, and there isn't enough data to use in the analysis. I will not be using this data as it is not sufficient.

- Id: Unique ID for each user. In long data format. Each row can contain data on the same ID (user). I need to identify the number of unique IDs.
- Date: States the timestamp for each activity, daily. I need to find out how many unique IDs there are.
- WeightKg: Shows the weight in Kg (daily) for each unique user. The data log for every user is not consistent, so I will not be able to do much with the data.
- WeightPounds: Shows the weight in pounds (daily) for each unique user. The data log for every user is not consistent, so I will not be able to do much with the data.
- Fat: I am not sure what this column is for, as there are only 2 values imputed (Update: it means body fat percentage, and it is sparse because most users don't have a smart scale to measure body fat). If I am using this table, I will take out this column.

- BMI: Shows the BMI for each unique user daily. However, as with the rest of the data, the imputation is not consistent per user per day, so not much can be done.
- IsManualReport: Shows whether the data was imputed manually… I think. There are only 2 values: TRUE and FALSE. TRUE = User manually entered the weight, FALSE = Automatically synced from a smart scale.
- LogId: Shows the unique identifier for each weight checking session.
- Note: I will simply note in the report: "Weight data was excluded due to insufficient sample size (8 users) and inconsistent logging patterns."

TLDR; Summary

I AM using these tables:

- dailyActivity_merged.csv
- DailyCalories_merged.csv
- dailyIntensities_merged.csv
- dailySteps_merged.csv
- hourlyIntensities_merged.csv
- sleepDay_merged.csv

I MAY use these tables, depending on whether I can conclude them:

- minuteMETsNarrow_merged.csv
- minuteSleep_merged.csv
- Heartrate_seconds_merged.csv

I will NOT use these tables:

- hourlyCalories_merged.csv
- hourlySteps_merged.csv
- minuteCaloriesNarrow_merged.csv
- minuteCaloriesWide_merged.csv
- minuteIntensitiesNarrow_merged.csv
- minuteIntensitiesWide_merged.csv
- minuteStepsNarrow_merged.csv
- minuteStepsWide_merged.csv
- weightLogInfo_merged.csv