Cleaning or manipulation of data

1 Data cleaning process

First, I verified that the data in each file fell within the date range specified in the documentation. To do this, I sorted the rows both in ascending and descending order by the date column, and then performed a visual check. An example of the query used for this review is as follows:

```
SELECT *
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_041116`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_051216`
)
ORDER BY
  ActivityDay ASC
```

The maximum and minimum values were also identified, and then it was checked whether these values were realistic. Below is an example of the code used for this review:

```
SELECT
  MAX(Calories) AS max_calories,
  MIN(Calories) AS min_calories
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_041116`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_051216`
)
```

In addition, the presence of null values was checked, and below is an example of the code used for this verification:

```
SELECT *
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_041116`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyCalories_051216`
)
WHERE
  Calories IS NULL
```

It was verified that all IDs were valid by checking that they contained only numeric values and by counting the number of characters. The query used was as follows:

```
SELECT DISTINCT
  LENGTH(CAST(Id AS string)) AS id_lenght
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_051216`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_041116`
)
```

In some cases, the number of records with extremely low or high values was counted. The query used was as follows:

```
SELECT
  TotalSteps,
  COUNT(TotalSteps) AS records
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_051216`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_041116`
)
GROUP BY
  TotalSteps
ORDER BY
  TotalSteps ASC
```

It was also verified that the sum of the different distance fields in the daily activity table matched the total distance. Since exact matches were not found, the percentage difference was used to assess consistency. The code used is shown below:

```
SELECT
  TotalDistance,
  TrackerDistance,
  LoggedActivitiesDistance + VeryActiveDistance +
  ModeratelyActiveDistance + LightActiveDistance +
  SedentaryActiveDistance AS calculateDistance,
  100*ABS(TotalDistance - (
    LoggedActivitiesDistance + VeryActiveDistance +
    ModeratelyActiveDistance + LightActiveDistance +
    SedentaryActiveDistance
  )) / TotalDistance AS PorcentualDifference
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_051216`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_041116`
)
WHERE
  TotalDistance != 0 AND
  100*ABS(TotalDistance - (
    LoggedActivitiesDistance + VeryActiveDistance +
    ModeratelyActiveDistance + LightActiveDistance +
    SedentaryActiveDistance
  )) / TotalDistance > 30
```

2 Data cleaning results

In the daily activity table, several data-cleaning actions were performed to ensure consistency and analytical relevance. A total of 138 records with 0 steps were removed. Although a value of 0 is not inherently implausible—such as in cases of complete rest—the number of such records was excessive and could interfere with trend analysis.

Similarly, 141 records were removed where the total distance traveled was 0, as they provide limited insight and may suggest tracking issues. In addition, 64 records were discarded due to a percentage discrepancy greater than 30% between the total distance and the sum of the individual distance components.

Finally, 9 entries were excluded where no calories were recorded as burned throughout the day. The code used for these filtering operations is shown below:

```
SELECT *
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_051216`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.dailyActivity_041116`
)
WHERE
  (TotalSteps != 0 OR
  TotalDistance != 0 OR
  Calories != 0) AND
  (TotalDistance != 0 AND
  100*ABS(TotalDistance - (
    LoggedActivitiesDistance + VeryActiveDistance +
    ModeratelyActiveDistance + LightActiveDistance +
    SedentaryActiveDistance
  )) / TotalDistance < 30)
```

Finally, the daily activity table contained 1,192 records after the data cleaning process.

In the hourly intensity table, 20,143 records were found with an intensity value of 0. This is a realistic value, as it is possible that no movement was recorded during certain hours. However, the number of records with a value of 0 represents a significant portion—nearly half of the total. In this case, it was decided to keep these data, as identifying periods of low activity will be important for detecting patterns and trends in physical behavior.

Records with METs values equal to 0 were removed, resulting in the exclusion of 13 entries. Consequently, the METs table now contains a total of 2,770,607 records.

Several issues were found in the weight table related to missing decimal points in the values for weight in kilograms, weight in pounds, and BMI. The first step in correcting these problems was to adjust the WeightPounds column since all the values appeared to lack a decimal point. This was done using the CONCAT and SUBSTR functions to insert the decimal point after the third digit from the left

Once the weight in pounds was corrected, it was used to recalculate the weight in kilograms using the standard conversion factor. These newly calculated values were then compared with the original WeightKg values to verify their consistency

For the BMI column, a similar approach was applied. However, a CASE statement was added to ensure that only values with obvious errors—specifically those greater than 100—were adjusted by inserting the decimal point

Lastly, the Fat column was not included in the analysis because it contained only four non-null entries

The SQL code used for all corrections and validations is presented below

```
SELECT
  Id,
  Date,
  ROUND(CAST(
    CONCAT(
      SUBSTR(CAST(WeightPounds AS STRING), 1, 3),
      '.',
      SUBSTR(CAST(WeightPounds AS STRING), 4)
    ) AS FLOAT64
  )*0.45359237, 2) AS WeightKg,
  CAST(
    CONCAT(
      SUBSTR(CAST(WeightPounds AS STRING), 1, 3),
      '.',
      SUBSTR(CAST(WeightPounds AS STRING), 4)
    ) AS FLOAT64
  ) AS WeightPounds,
  CASE
    WHEN BMI > 100 THEN CAST(
      CONCAT(
        SUBSTR(CAST(BMI AS STRING), 1, 2),
        '.',
        SUBSTR(CAST(BMI AS STRING), 3)
      ) AS FLOAT64
    )
    ELSE BMI
  END AS BMI,
  IsManualReport,
  LogId
FROM (
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.weightLogInfo_041116`
  UNION ALL
  SELECT * FROM `bellabeat-
456023.fitbit_fitness_tracker.weightLogInfo_051216`)
ORDER BY Id, Date
```