Ask:

What is the problem you are trying to solve and how can these insights drive business decisions?

Our primary objective is to address the issue of understanding the usage patterns of non-Bellabeat smart devices and leveraging that insight to inform Bellabeat's marketing strategy for their own products. The insights derived from the analysis may reveal trends and patterns in consumer smart device usage data. This information can guide our business decisions by allowing us to align with consumer preferences or identify features offered by competitors that appeal to consumers.

Prepare:

Where is your data stored? How is the data organized? Are there issues with bias or credibility in this data? Does your data ROCCC? How are you addressing Licensing, privacy, security and accessibility?

The data originally sourced from Kaggle was downloaded and saved locally on the computer. Subsequently, the data was imported into BigQuery SQL Server and Tableau for further analysis and visualization. All the data being used is structured in a long format, with each file containing specific details such as calories burned, time spent sleeping, and weight progression, which are organized separately. ROCCC is an acronym that represents the following characteristics: Reliable, Original, Comprehensive, Current, and Cited. The reliability of the data is uncertain as its source from a third party is unknown. It is comprehensive and well-organized. However, the data is not current, spanning from 3/12/16 to 5/12/16. The only citation available is from a survey conducted via Amazon Mechanical Turk. The data originates from a publicly accessible dataset, ensuring that there are no licensing concerns as anyone can download it. As the data does not contain any personal information or names, there are no privacy concerns associated with it.

Deliverable:

What data sources are used?

The data for this analysis will be sourced from the fitabase dataset available on Amazon Mechanical Turk. Specifically, we will be utilizing the following files: dailyActivity\_Merged, heartrate\_seconds\_merged, SleepDay\_merged, and weightLogInfo\_merged.

Process:

What tools are you choosing and why? Have you ensured your data's integrity? What steps have you taken to ensure that your data is clean? How can you verify that your data is clean and ready to analyze?

For this analysis, I will utilize SQL and Tableau as the primary tools. The data cleaning process will primarily take place within SQL, where we will address any inconsistencies or discrepancies in the data. The consistency of the data remains uniform across the entire dataset. Initially, the focus was on verifying the correctness of the column types, which proved to be accurate. Next, the analysis involved identifying any duplicate values, which were indeed present. Subsequently, an examination for null values was conducted, and it was discovered that a column containing such values was irrelevant for the intended analysis. To ensure data cleanliness, one approach involved taking small samples. Additionally, SQL functions were executed, and the schema in BigQuery was inspected to confirm the correctness of the data types.