**Business Tasks**

* Problems
  1. What are some trends in smart device usage?
  2. How could these trends apply to Bellabeat customers?
  3. How could these trends help influence Bellabeat marketing strategy?
* Objectives
  1. Find if the number of usage day has relationship between sleep, heart rate, calories or distance.
* Business Tasks
  1. Use an analysis to reveal more opportunities for growth.
  2. Focus on a Bellabeat product and analyze smart device usage data in order to gain insight into how people are already using their smart devices?
  3. How these trends can inform Bellabeat marketing strategy.
* Stakeholders
  1. Urska Srsen
  2. Sando Mur
  3. Bellabeat marketing analytics team

**Data Collection**

* Data source
  1. [FitBit Fitness Tracker Data](https://www.kaggle.com/arashnic/fitbit) is public data on Kaggle.
  2. The dataset is made available through [Mobius](https://www.kaggle.com/arashnic).
  3. License by CC0: [Public Domain](https://creativecommons.org/publicdomain/zero/1.0/)
* Original data organization
  1. The product is Leaf.
  2. Data is saved in csv file.
  3. Total 18 csv files
  4. The original data folder is named “Fitabase Data 12.04.16-12.05.16”. I renamed the folder to “data”.
  5. All folders have different number of columns but all have Id.
  6. Some data is organized in long format and some is wide format.
* Issues and bias
  1. Only 30 candidates and only 30 days of data so sample is quite small.
  2. Although data is provided by users from their devices, we cannot ensure all data is raw data without any editing.
  3. Cannot be sure It is original.
  4. These datasets include daily activity, steps and heart rate but cannot determine if they are comprehensive or not and cannot determine if important values are missing.
  5. It is not current. The latest data is 2016/05/12.
  6. Currently I do not see any bias (will be discover deeper in data cleaning phases).
  7. Age is probably an important factor of daily activity and use of smart phone, and the datasets seem not provide age information.
* Licensing, privacy, security and accessibility
  1. All licensing information is included in License URL.
  2. Use the data for personal study only.
  3. Mark the source of the material and data.
  4. ID does not contain any personal information and if ID has any other issue about privacy, ID will be blocked if the data needs to be viewed by others
* Data integrity: Data integrity will be considered for all process. Currently only look at original data.

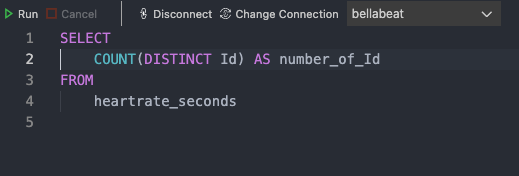
1. Accuracy
   * TO-DO
   * .
2. Completeness
   * TO-DO
   * .
3. Consistency
   * TO-DO
   * .
4. Trustworthiness
   * TO-DO
   * .

* How does the data help to answer the questions?
  1. The datasets include information about daily activity, steps, heart rate and calories.
* Sort and filter
  1. Sorting and filtering will be implemented in data cleaning phase.

**Process**

* Tools

1. R:
   * Do initially data cleaning (bellabeat\_initial\_cleaning.R) in order to import to BigQuery because BigQuery is unable to read datetime format of original datasets.
     + Round all double to second decimal
     + Reformat datetime values.
   * Do secondary data cleaning (bellabeat\_secondary\_cleaning.R)
     + Check all value which should not be negative but is negative
2. MSSQL
   * Check Id numbers



* + - heartrate\_seconds only has 14 Ids
    - minuteSleep only has 24 Ids
    - sleepDay only has 24 Ids
    - weightLogInfo only has 8 Ids
    - Other tables have 33 Ids
* Cleaning data process

1. Step 1:
   * R
     + bellabeat\_initial\_cleaning.R will be included in the file
     + Reformat daytime and day format in order to import to BigQuery and MSSQL.
     + ROUND all floats to the second decimal.
     + Input files are original files and output files end with “\_v2”.
2. Step 2:
   * R
     + bellabeat\_secondary\_cleaning.R will be included in the file.
     + Check all values are not negative which they should not, for example, steps, minutes or intensities.
     + Original I would like to use BigQuery and MSSQL to do this but some files are wide format. There are too many columns. So I wrote a R script to clean all datasests.
     + If there were any negative values, index of row and sub datasets would be shown. The script allows users to choose to remove the rows or change to positive. It depends on users decision.
     + Input files are files ending with “\_v2” and output files end with “\_v3”.
3. Step 3:
   * BigQuery/MSSQL
     + Check if there is any value that does not make sense.
     + There are some values that does not make sense but I did not clean them. I will put them into analysis because these non-sense value might be devices malfunction or users usage habits. It does not mean users physical problems.
4. Step 4:
   * TO-DO
   * .
5. Step 5:
   * TO-DO
   * .
6. Step 6:
   * TO-DO
   * .

* Verify data

1. TO-DO
2. .

* Data transformation
  1. TO-DO
  2. .

**Analyze**

* Tools
  1. TO-DO
  2. .
* Format data
  1. To-DO
  2. .
* Calculations
  1. TO-DO
  2. .
* Trends and data visualization
  1. TO-DO
  2. .

**Share**

* Tools
  1. TO-DO
  2. .
* Findings and stories
  1. Findings:
     + TO-DO
     + .
  2. Stories:
     + TO-DO
     + .
* Conclusion:
  1. TO-DO
  2. .
* Audience:
  1. TO-DO
  2. .