

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?
 4. People wearing these smart devices should wear them all the time so that the smart devices can record all data. But WHY ARE THERE LIGHT USERS?
- Objectives
 1. Find if the number of usage day has relationship between sleep, heart rate, calories or distance.
 2. Frequency index is calculated below.
- Business Tasks
 3. Use an analysis to reveal more opportunities for growth.
 4. 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?
 5. 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](#) is public data on Kaggle (**cited**).
 2. The dataset is made available through [Mobius](#). (**cited**)
 3. License by CC0: [Public Domain](#)
- 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". All other files are named as the topic of each data sets.
 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 33 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 (**not current**).
 6. Currently I do not see any bias (will be discover deeper in data cleaning phases).
 7. Age and occupation are probably important factors of daily activity and use of smart devices. But all datasets do not provide these two information (**not comprehensive**).
- 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 or transformed into another forms 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
 - Data samples are not enough (**known from data analysis phases**).
 - Data is not evenly distributed (**known from data analysis phases**).
 - Samples should be able to represent the whole population of U.S but some average figure of samples are well below U.S average.
 2. Completeness
 - Data samples are not enough.
 - There are 33 Ids but some datasets contains only few Ids.
 3. Consistency
 4. Trustworthiness

- Original data is cleaned and merged by creators. It is not raw data. So trustworthiness might not be high.
- It is not raw data so it might contains some bias.
- How does the data help to answer the questions?
 1. The datasets include information about daily activity, steps, heart rate and calories.
 2. The datasets include day of records and hours per day. This can be used to calculate frequencies.
 3. Compare frequencies with daily activities, steps, heart rate and calories to figure out if there is any trends.
- Sort and filter
 1. Sorting and filtering will be implemented in data cleaning phase.

Process

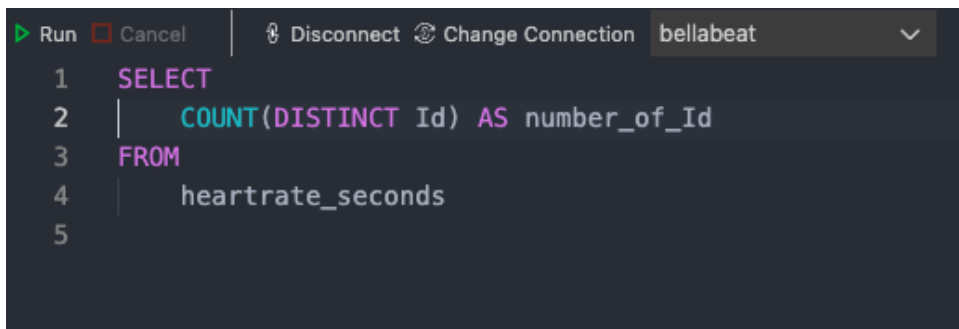
- Tools

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



```
Run Cancel | Disconnect Change Connection bellabeat
1 SELECT
2   COUNT(DISTINCT Id) AS number_of_Id
3 FROM
4   heartrate_seconds
5
```

- ◆ 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 datasets.
 - ◆ 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:

- Import data into BigQuery and MSSQL.
- The reason I use R to do data cleaning first is because BigQuery and MSSQL cannot read formats of date and datetime. Even if successfully import into BigQuery and MSSQL, the format is incorrect. Also, one R script can automatically transform all date and datetime format for all datasets.
- Jump to data verification

● Verify data

1. BigQuery/MSSQL (after successfully import)

- Use LENGTH() to check the length of value for date and datetime.
- Use GROUP BY and COUNT() to check all Id and number of Id for each datasets.
- Recheck all values that should not be below 0.
- 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.

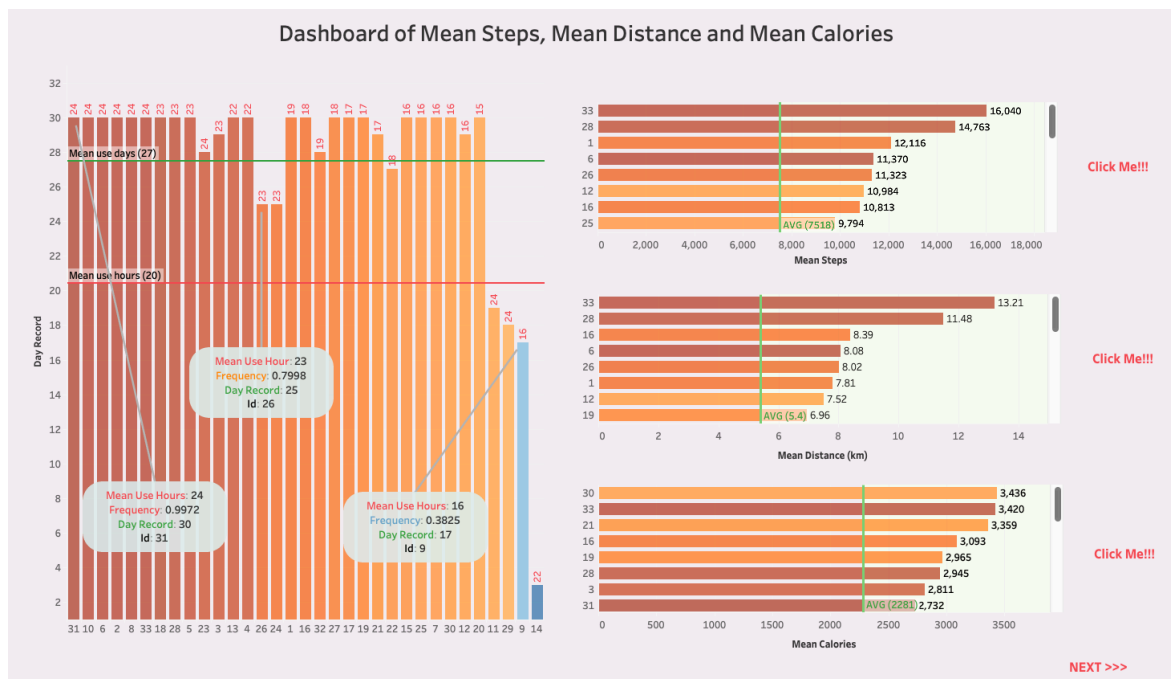
2. All datasets were cleaned and merged initially by the creator. Therefore, there are not too many problems.

● Data transformation

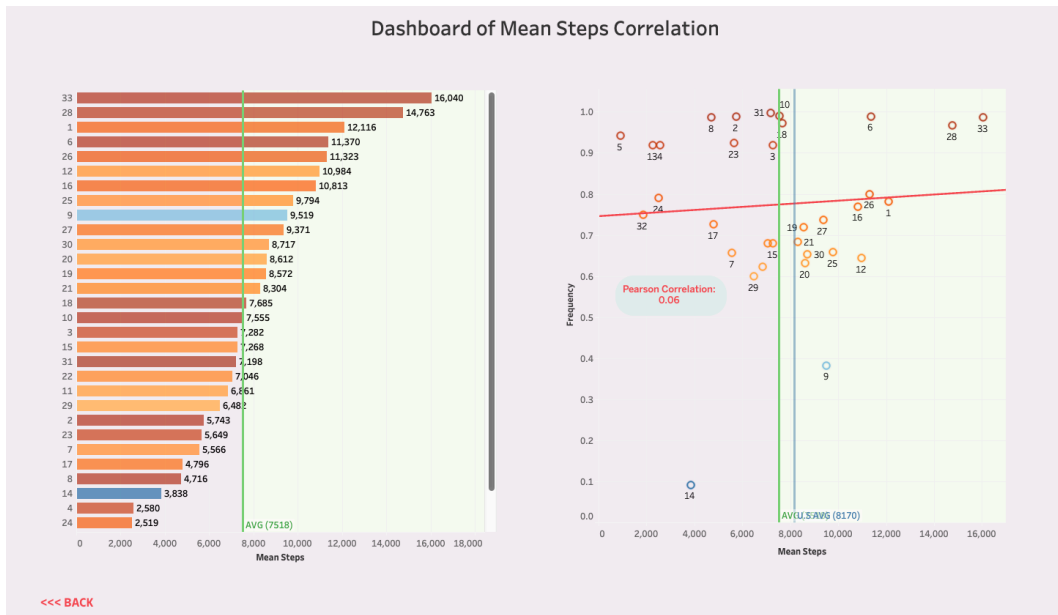
1. Data transformation (date, datetime and other floats) was done in data cleaning process using R.
2. Data transformation was also done after importing into MSSQL and BigQuery (data migration) to make sure data can be compare and visualize.
3. Data transformation is conducted when importing into Tableau for future data analysis process (data migration).
 - ◆ Change Id from int type to string type.
 - ◆ Change long Id into short Id, starting from 1 and end at 33.

Analyze

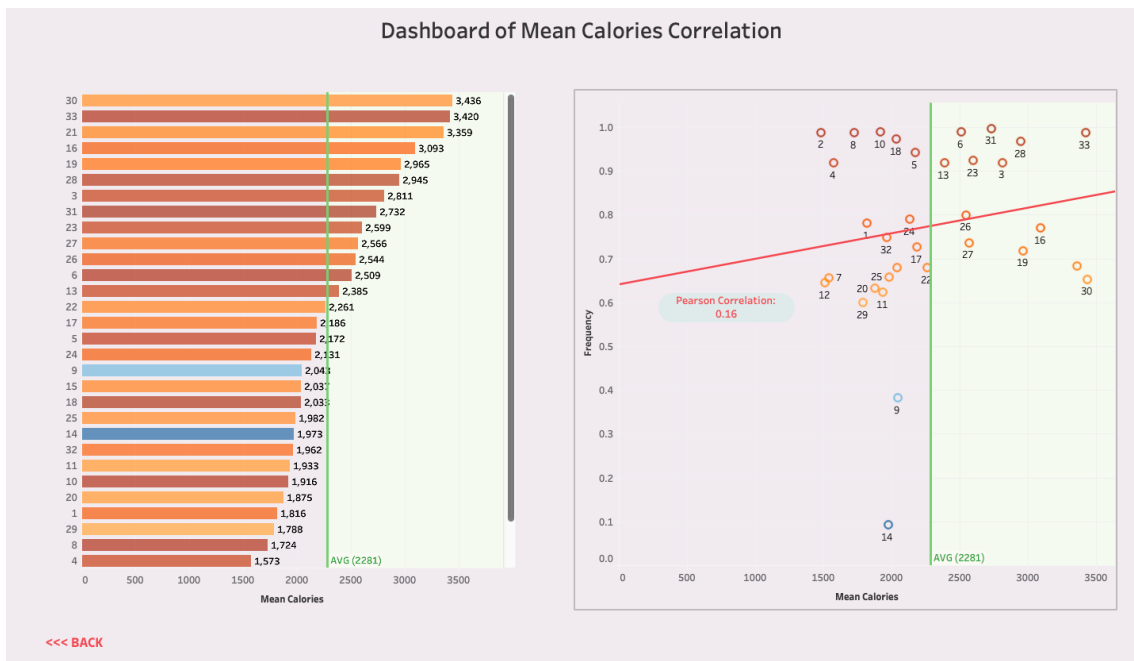
- Tools
 1. MSSQL to calculate average, sum or count.
 2. MSSQL exports new csv file and import into Tableau.
 3. Data visualizes by using Tableau to figure out if there are some trends.
- Format data
 1. All data is properly formatted by using MSSQL for further calculation.
 2. All data types are properly defined by using Tableau.
- Calculations
 1. New csv files are exported from MSSQL. Most of them contain AVG, COUNT or SUM.
 2. Frequency index is calculated in Tableau.
 - ◆ $\text{Frequency index} = (\text{days of use}/30) * (\text{hours of use}/24)$
 - ◆ I break all users into very high frequent users (frequency index > 0.8), frequent users (0.6 ~ 0.8), mid users (0.4 ~ 0.6) and light users (0 ~ 0.4).
- Trends and data visualization
 1. Tableau
 - ◆ Use data visualization to visualize calories, steps and sleeps for each Id with frequency index.
 - ◆ From the chart figure out trends and calculate Pearson correlation.
 2. General information: Frequency vs calories, steps and distance



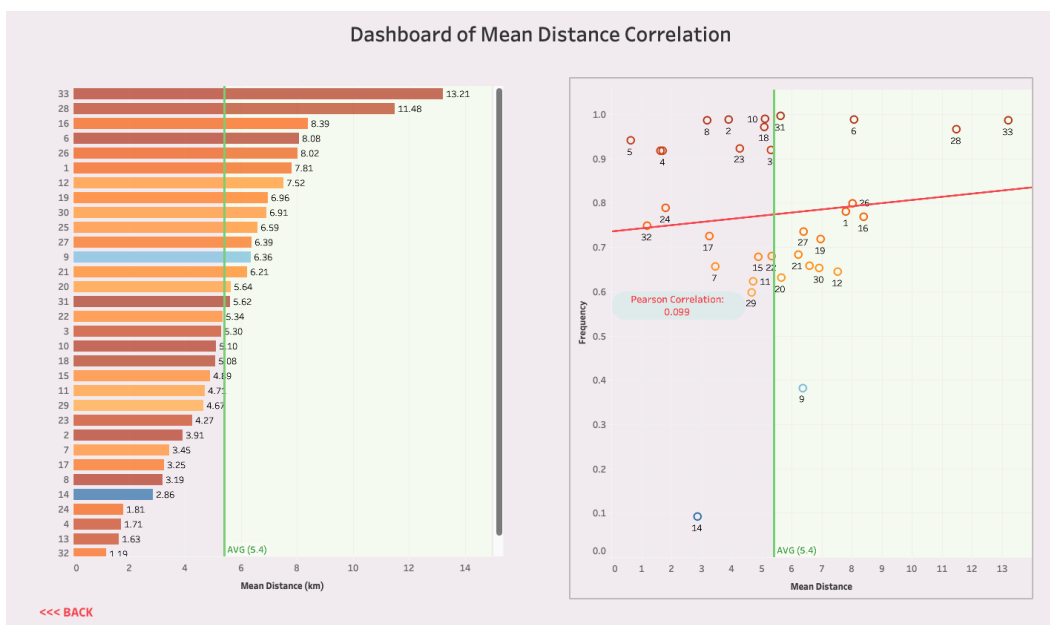
3. Mean steps correlation with frequency



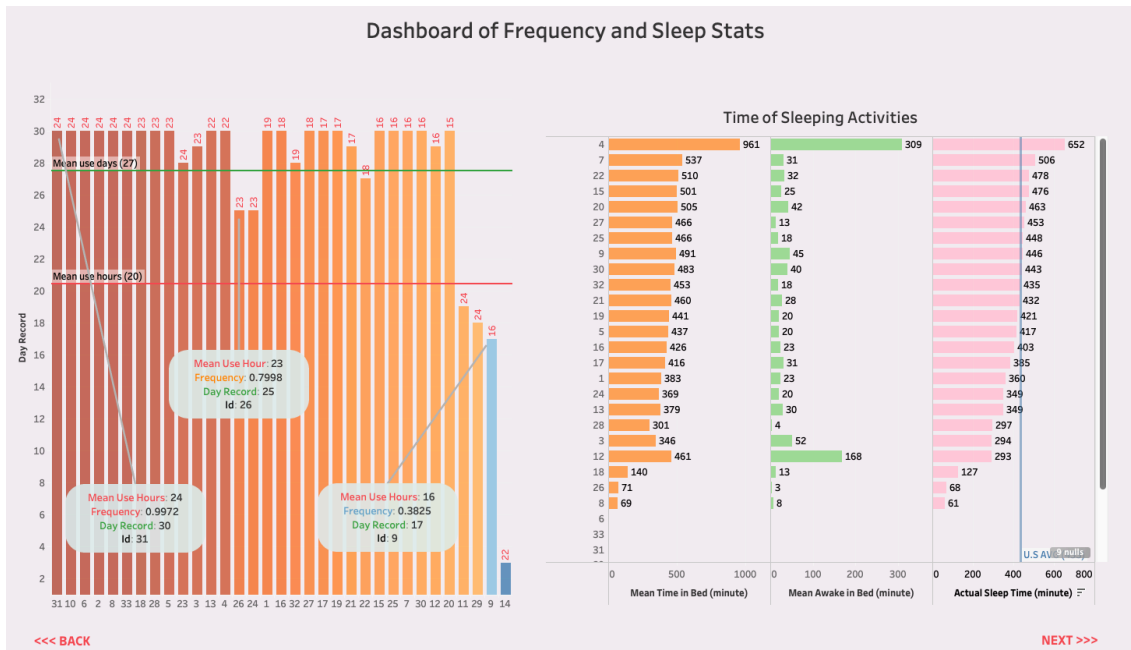
4. Mean calories correlation with frequency



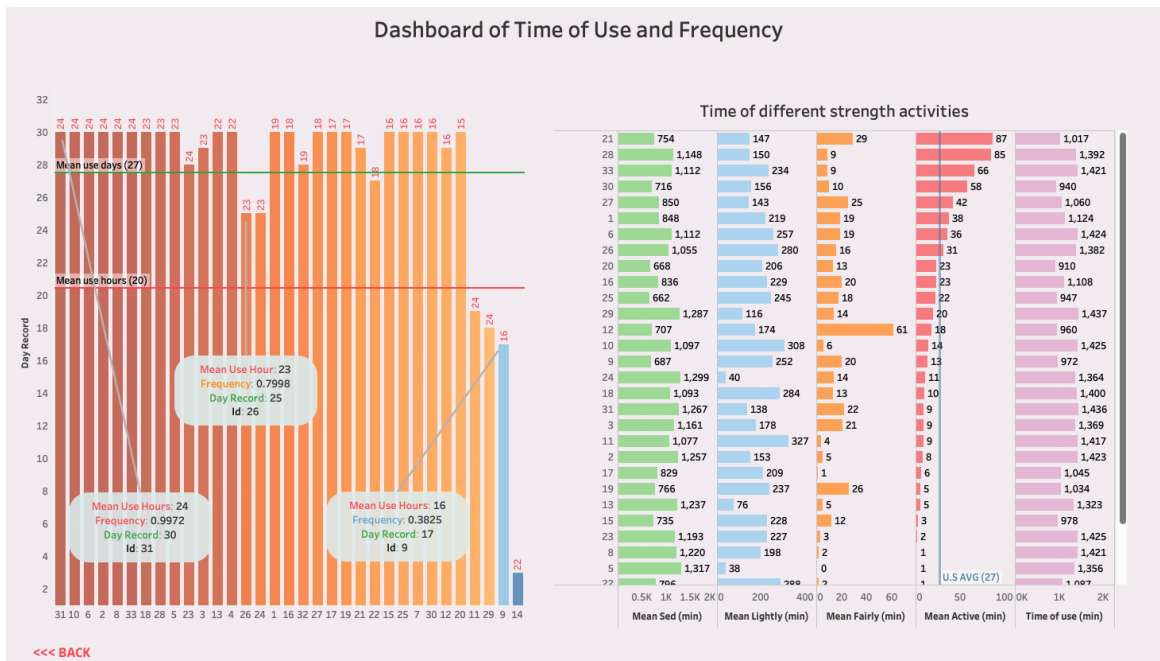
5. Mean distance correlation with frequency



6. Sleeping mode for each Id



7. Time of use for each Id



● Analysis

1. Time of use and frequency

- All users' active time higher than U.S average are very frequent users or frequent users. However, those who has very low active time are also very frequent users or frequent users. Only few mid users and one low users.
- Possible reasons:
 - Samples are not evenly distributed (90% are frequent and very frequent users) (**not comprehensive, bias**).

2. Steps and correlation

- Trend line shows that mean steps slightly increases as frequency increases but we are not sure they have positive correlation because samples are not enough, average is well below U.S average (**not comprehensive**) and samples cannot represent the whole population.

- U.S average figures are from fitbit website (<https://www.fitbit.com/global/us/activity-index>).
- 3. Calories and correlation (**not comprehensive**)
 - Trend line shows that mean steps slightly increases as frequency increases but we are not sure they have positive correlation because samples are not enough.
- 4. Distance and correlation (**not comprehensive**)
 - Trend line shows that mean steps slightly increases as frequency increases but we are not sure they have positive correlation because samples are not enough.
- 5. Frequency and sleeping hours (**inaccurate**)
 - Number 4 users seems to have sleeping problems. But is this her purpose of using smart devices?
 - In some cases, sleeping hours are not accurate because some people do not wear smart devices when they sleep. These cases include null value users or very low sleeping hours users (they might fall asleep for a while and take them off then go to bed).

Share

- Tools
 1. Powerpoint for start of the report including tasks, objects and so on.
 2. Data visualization using Tableau and shared by using URL.
- Findings and stories
 1. Findings:
 - My hypothesis is that people who uses smart devices less frequent use for specific purpose. However, samples are not enough so that we are not able to conclude and prove this hypothesis. More research and data are needed.
 2. Stories:
 - Stories are shared by using Tableau.
- Conclusion:
 1. Trends of frequency index, calories, distance, steps and sleeping hours are not obvious because of data itself.
 2. Data are not completed and comprehensive. This leads data inaccurate. More researches and data are needed.