Bellabeat Data Analytics Case Study

By Saw Jing Xien on 6-Feb-2023

1.0 Introduction

Bellabeat is a smart tech company that manufactures healthcare smart products including the Bellabeat app, Leaf, Time and Spring for women. Collecting data on activity, sleep, stress and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. The co-founder Sršen knows that an analysis of Bellabeat's available consumer data would reveal more opportunities for growth.

2.0 Summary of Business Task (Ask Phase)

The invention of smart devices has changed people's lifestyle tremendously and as a smart tech company, Bellabeat can benefit from analysing the trend of non-Bellabeat smart devices usage data so that our key stakeholders Urška Sršen and Sando Mur can make data-driven business strategies to promote opportunities for growth.

Below are some good points to get started:-

1. What are the non-Bellabeat products with similar functionalities?
2. What are some trends in smart devices usage?
3. How could these trends apply to Bellabeat customers?
4. How could these trends help influence Bellabeat marketing strategy?

3.0 Prepare the Data (Prepare Phase)

FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius on Kaggle) contains 33 consented eligible users' personal fitness tracker data including up to minute level output of heartrate, calories, intensities and METs (shorts for metabolic equivalent of task which measures activity intensities) which helps to reveal users' habits. It is believed that it's been mistyped at data source description that there are only 30 eligible users and when checked using BigQuery there are 33 distinct Id which represents each user.

This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Variation between output represents use of different types of Fitbit trackers and individual tracking behaviours / preferences. The dataset has 18 .csv files where 15 are in long format and 3 in wide format. Since the wide format data is repetitive to the long format, we will only be using the 15 long format data in the analysis.

When doing ROCCC analysis, some findings are observed as below:-

1. Reliability (R) – No information of gender, location, and age to assess bias and target markets.
2. Originality (O) - Third party data collected via Amazon Mechanical Turk, not open-source data from Fitbit. Fitbit has no open-source fitness data available.
3. Comprehensive (C) - Fitness data fields comprising steps, distance, intensity, calories, heartrate, METs, sleep, weight with multi-level of time output up to minutes. However there is lack of definitions of the data fields provided in the metadata. The sleep data are

recorded in terms of date without time and no differentiation of sleep sessions available. The date range of the dataset is from 13/4/2016 – 12/5/2016, which is a month of data.

4. Current (C) - Data is almost 7 years old but human habits and behaviour do not change that much over some years.
5. Cited (C) – Data is well documented and cited on Kaggle.

Beyond the limitations highlighted above, the data integrity has been checked and corrected where required. It is recognized that more data source should be identified for a more comprehensive study, however this case study will be based on only the limited FitBit data available.

This report will present data cleaning, data wrangling and data visualization using Google Cloud BigQuery, Google Sheet and Tableau.

4.0 Data Cleaning (Process Phase)

Environment setup

The BigQuery environment is setup by defining *"rare-scout-365317"* as the default project ID and *"fitabase_data"* as the dataset name. All 15 tables in long format are uploaded into this dataset.

Cleaning date-time format

It is noticed that the date time format in the original .csv files is not supported on BigQuery as BigQuery does not recognize the AM and PM format. Hence the files are uploaded and formatted to 24 hours format in Google Sheet before uploading to BigQuery. "daySleep" table has time info but all it contained is 00:00:00 hence it is converted to date instead of date-time.

Data integrity check for distinct Id count

The distinct Id which represents individual users is checked in BigQuery as follow:-

```
SELECT "daily_activity" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.daily_activity UNION ALL
SELECT "daily_calories" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.daily_calories UNION ALL
SELECT "daily_intensities" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare
    -scout-365317.fitabase_data.daily_intensities UNION ALL
SELECT "daily_steps" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.daily_steps UNION ALL
SELECT "heartrate" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-scout-
    365317.fitabase_data.heartrate UNION ALL
SELECT "hourlyCalories" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.hourlyCalories UNION ALL
SELECT "hourlyIntensities" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare
    -scout-365317.fitabase_data.hourlyIntensities UNION ALL
SELECT "hourlySteps" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.hourlySteps UNION ALL
SELECT "minuteCalories" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.minuteCalories UNION ALL
SELECT "minuteIntensities" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare
    -scout-365317.fitabase_data.minuteIntensities UNION ALL
```

```
SELECT "minuteMETs" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-scout-
    365317.fitabase_data.minuteMETs UNION ALL
SELECT "minuteSleep" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.minuteSleep UNION ALL
SELECT "minuteSteps" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.minuteSteps UNION ALL
SELECT "sleepDay" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-scout-
    365317.fitabase_data.sleepDay UNION ALL
SELECT "weightLogInfo" AS table_name, COUNT(DISTINCT Id) AS Id_count FROM rare-
    scout-365317.fitabase_data.weightLogInfo;
```

Output:

table_name	Id_count
minuteCalories	33
hourlySteps	33
daily_activity	33
minuteSteps	33
minuteSleep	24
weightLogInfo	8
sleepDay	24
minuteMETs	33
minuteIntensities	33
hourlyCalories	33
daily_intensities	33
hourlyIntensities	33
daily_steps	33
heartrate	14
daily_calories	33

All tables have 33 distinct users except “minuteSleep”, “sleepDay”, “weightLogInfo” and “heartrate”. This shows that the dataset is inconsistent with the data source description which states that 30 user data are available. This impacts the analysis result as there is incomplete data. However, they could be used for individual analysis considering them as smaller group of sample size.

Data integrity check for unique column names, data types and frequency appeared in the dataset

```
SELECT column_name, data_type, COUNT(table_name) AS count
FROM `rare-scout-365317.fitabase_data.INFORMATION_SCHEMA.COLUMNS`
GROUP BY column_name, data_type
ORDER BY count DESC, column_name;
```

Output:

column_name	data_type	count
Id	INT64	15
ActivityMinute	TIMESTAMP	4
ActivityDay	DATE	3
ActivityHour	TIMESTAMP	3
Calories	INT64	3
FairlyActiveMinutes	INT64	2
LightActiveDistance	FLOAT64	2
LightlyActiveMinutes	INT64	2
ModeratelyActiveDistance	FLOAT64	2
SedentaryActiveDistance	FLOAT64	2
SedentaryMinutes	INT64	2
StepTotal	INT64	2
VeryActiveDistance	FLOAT64	2
VeryActiveMinutes	INT64	2
ActivityDate	DATE	1
AverageIntensity	FLOAT64	1
BMI	FLOAT64	1
Calories	FLOAT64	1
Date	TIMESTAMP	1
Fat	INT64	1
Intensity	INT64	1
IsManualReport	BOOL	1
LogId	INT64	1
LoggedActivitiesDistance	FLOAT64	1
METs	INT64	1
SleepDay	DATE	1
Steps	INT64	1
Time	TIMESTAMP	1
TotalDistance	FLOAT64	1
TotalIntensity	INT64	1
TotalMinutesAsleep	INT64	1
TotalSleepRecords	INT64	1
TotalSteps	INT64	1
TotalTimeInBed	INT64	1
TrackerDistance	FLOAT64	1
Value	INT64	1
WeightKg	FLOAT64	1
WeightPounds	FLOAT64	1
date	TIMESTAMP	1
logId	INT64	1
value	INT64	1

There are 40 distinct column names across all tables and all tables contain “Id” field.

5.0 Data Analysis (Analysis Phase)

We first dive into data on a daily level to study the behavior of users during active sessions.

a) User behavior on logged activity function

The "LoggedActivitiesDistance" column represents manually logged activity on the smart device such as a running session. This data will give us insight on how often users use this function which requires manual user input.

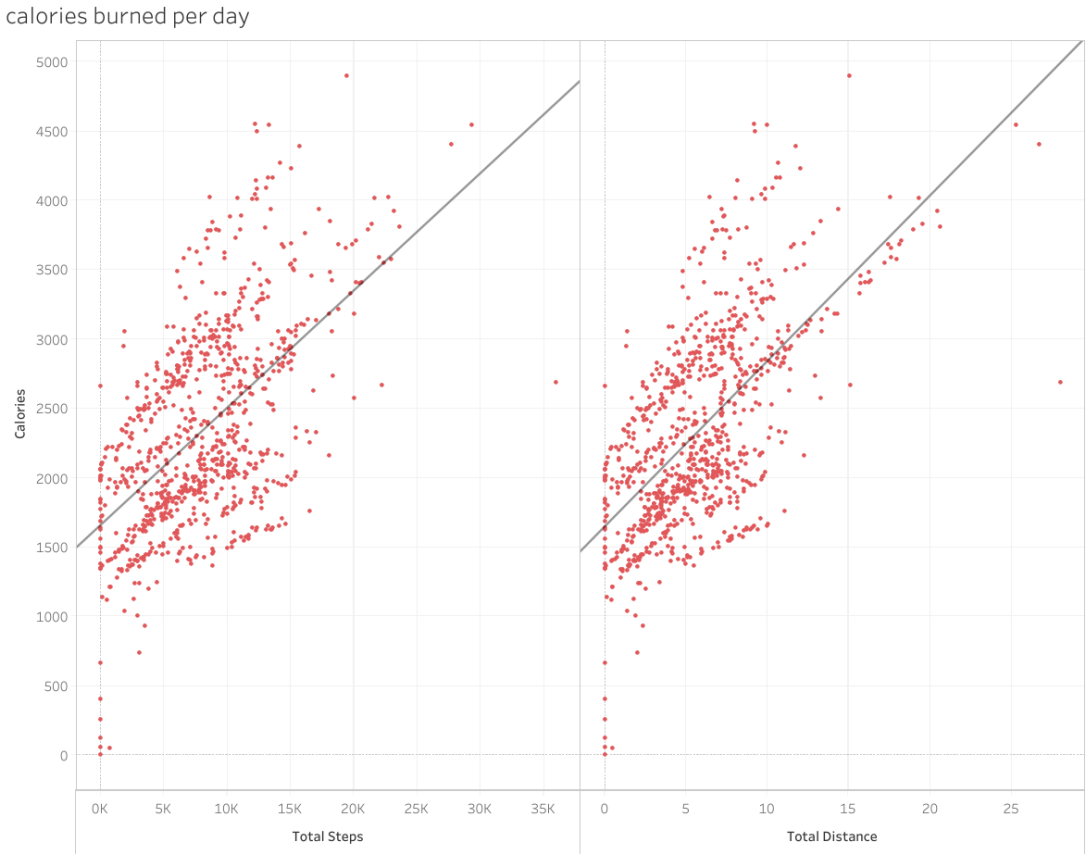
```
SELECT COUNT(DISTINCT Id) as logged_activity_no, total activity
FROM `rare-scout-365317.fitabase_data.daily_activity`
WHERE LoggedActivitiesDistance <> 0
```

Row	logged_activity_no
1	4

It turns out that out of 33 users only 4 users had logged at least one activity between 13/4/2016 – 12/5/2016. This function seems to be underwhelmed by existing Fitbit users in our dataset. The hypothesis is that this function is not intuitive enough in the app interface, as when a user forgets to log an activity before a workout, he/she could do so inside the app but it may be too cumbersome to do so.

It is suggested that a smart alert to automatically detect an increase of intensity and remind users to view backtracked data so they can log an activity post-workout and this process should as seamless and intuitive as possible to encourage users to log their activities.

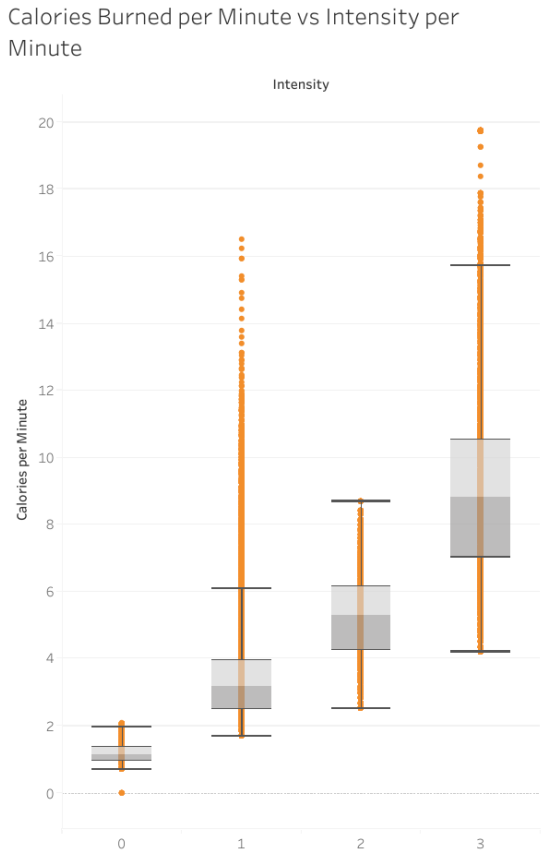
b) Total calories burned per day

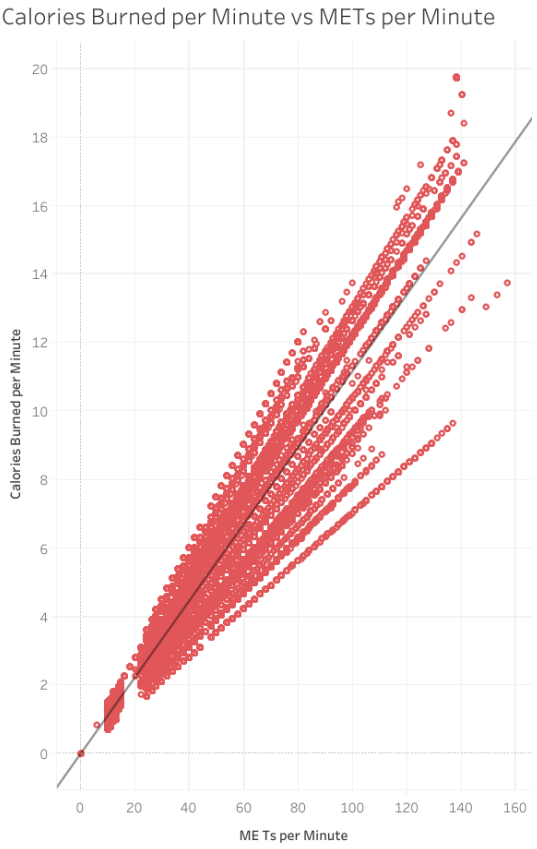


Source: *https://public.tableau.com/views/caloriesburnedperday/Dashboard2?:language=en-US&:display_count=n&:origin=viz_share_link*

The viz above shows total calories burned per day increase as total steps/total distance taken per day increase. The R^2 is at 0.33 and 0.4 for total steps and total distance respectively. There is positive correlation of total calories burned and total steps/distance taken per day. A more active walker burns more calories per day.

c) Calories burned vs Intensity / METs



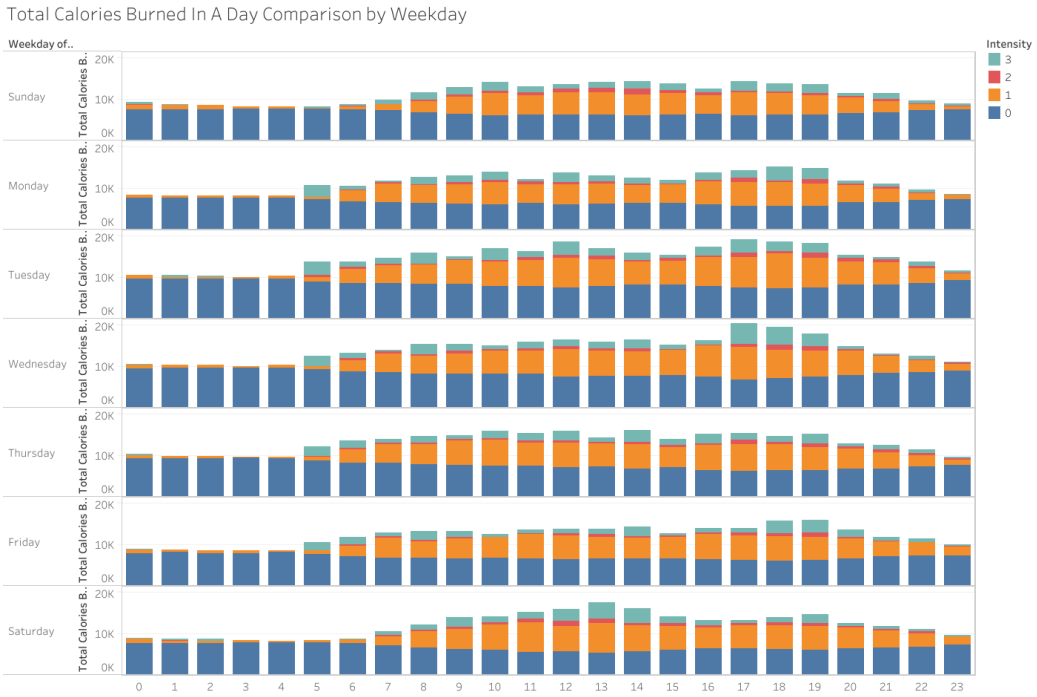


Source:
https://public.tableau.com/views/Calories_16780261225300/Calories?:language=en-US&:display_count=n&:origin=viz_share_link

The viz above was created by joining “minuteCalories”, “minuteIntensity” and “minuteMETs” tables in Tableau.

Activity intensity is measured from a scale of 0 to 3, where 0 is sedentary, 1 is light activity, 2 is moderately active and 3 is very active. On average the calories expenditure gets higher as the user gets into more intensive activities.

It is interesting to note that the outliers in Intensity = 1 (light activity) may give us some insights on how user calories burned during light activity. Above 9 calories burned per minute gives us the same calories burned on average while being very active.



Source:
https://public.tableau.com/views/TotalCaloriesBurnedInADayComparisonbyWeekday/TotalCaloriesBurnedInADayComparisonbyWeekday?:language=en-US&:display_count=n&:origin=viz_share_link

A closer look at the individual data can reveal user's behavior across the week. This information can be used to provide push alerts to users on custom fitness suggestions at the appropriate time of day. This function provides possibilities for more user interactions with personal touch which adds value to the users.

d) Sleep pattern

In table "minuteSleep", each session of sleep is logged by a logId so this can be used to identify separate sleep sessions.

```
SELECT
    Id,
    logId,
    MIN(date) AS sleep_start,
    MAX(date) AS sleep_end,
    TIME(TIMESTAMP_DIFF(MAX(date),MIN(date),HOUR),
    MOD(TIMESTAMP_DIFF(MAX(date),MIN(date),MINUTE),60),

    MOD(MOD(TIMESTAMP_DIFF(MAX(date),MIN(date),SECOND),3600),60) ) AS time_slee
ping,
    CAST(TIMESTAMP_DIFF(MAX(date),MIN(date),HOUR)+MOD(TIMESTAMP_DIFF(MAX(dat
      e),MIN(date),MINUTE),60)/60 AS FLOAT64) AS time_diff
 FROM `rare-scout-365317.fitabase_data.minuteSleep`
 WHERE value=1
 GROUP BY Id, logId
```

Output (first 10 rows):

Id	logId	sleep_start	sleep_end	time_sleeping	time_diff
1503960366	11380564589	2016-04-12 02:49:30	2016-04-12 08:32:30	5:43:00	5.71666667
1503960366	11388770715	2016-04-13 03:08:30	2016-04-13 08:20:30	5:12:00	5.2
1503960366	11388770716	2016-04-13 20:10:00	2016-04-13 21:43:00	1:33:00	1.55
1503960366	11402722600	2016-04-15 03:03:00	2016-04-15 10:20:00	7:17:00	7.28333333
1503960366	11421831252	2016-04-16 02:14:00	2016-04-16 06:57:00	4:43:00	4.71666667
1503960366	11421831253	2016-04-16 07:02:00	2016-04-16 08:19:00	1:17:00	1.28333333
1503960366	11421831254	2016-04-16 23:27:00	2016-04-17 11:18:00	11:51:00	11.85
1503960366	11439580762	2016-04-19 02:06:30	2016-04-19 07:23:30	5:17:00	5.28333333
1503960366	11447640793	2016-04-20 02:01:00	2016-04-20 08:17:00	6:16:00	6.26666667
1503960366	11455720858	2016-04-21 02:38:30	2016-04-21 08:34:30	5:56:00	5.93333333

Sleep pattern differs by individual and lifestyle choice.

6.0 Recommendations & Conclusion (Act Phase)

1. In-app functionality could be improved to be more intuitive and seamless based on user's historical data as the current trend is that user lacks motivation to log activities on app.
2. The lack of BMI and weight data suggests that most users do not have this habit of recording these data in app. A weekly or monthly reminder could be sent to the users so that more insights can be brought in-app to the more health-conscious users.
3. Light activities that are calories burning above 9 calories per minute could be suggested in-app to users who are keen on exploring new methods to burn calories.