**Marketing Project Capstone for Bellabeat Company**

8/8/2023

Amir Fikri

The goal of the capstone project is to use a data set made up of information from 33 Fitbit users to perform the data analysis steps of Ask, Prepare, Process, Analyse, Share, and Act. The information includes specifics about tracking physical activity, heart rate, and sleep. The objective is to offer the marketing team strategic recommendations based on the learnings from the analysis.

# **About Company**

The wellness technology industry has witnessed significant growth in recent years, driven by an increasing focus on personal health and well-being. As a prominent player in this market, Bellabeat is a high-tech manufacturer of health-focused products for women. The company aims to leverage smart device fitness data to unlock new growth opportunities and guide their marketing strategy effectively. In this case study, we analyze user engagement and behavior using data to gain insights that can inform Bellabeat's marketing approach.

# **Ask**

The first phase of the data analysis process is asking the right questions. We should clearly understand why we are doing this analysis and what kind of problem you are solving.

* What is the problem you are trying to solve?
* How can your insights drive business decision?
* What is the overall business task?

# **Prepare**

The second phase is to prepare the data. Preparing data means collecting or using data that relevant to the problem we are trying to solve.

The dataset, obtained from Kaggle.com, comprises data from 30 FitBit users who willingly shared their personal tracker information. This includes detailed minute-level data on physical activity, heart rate, and sleep monitoring. The data was collected through a survey distributed via Amazon Mechanical Turk during the period from 03/12/2016 to 05/12/2016.

## **Data Limitation**

* Data comprehensiveness:
  + The data includes only 30 Fitbit users and covers a 2-month period. Comprehensiveness would be greatly enhanced by a larger sample size and a longer user time frame.
* Data credibility:
  + Because the data is regarded as third-party information, we must analyse it cautiously.
* Data update:
  + The data is currently five years old. Since then, variables like user types or technological advancements may have altered. A more recent date for the data would have been preferable.
* Data bias:
  + To determine whether there is bias in the sample data, we are unable to access crucial details like location, gender, or age.

## **Tables Used**

* dailyActivity\_merged.csv
* Mets\_merged.csv
* sleepDay\_merged.csv
* heartrate\_merged.csv
* weightloginfo\_merged.csv

# **Process**

The third phase is to process the data. Data processing is to find various inaccuracies, errors, inconsistencies in the data and get rid of them so that our primary business problem is not affected.

## **Load & Preview Data**

Firstly, I will check the data in Microsoft Excel/Google Sheets to verify the tables are related so that I will be able to understand the data context. Since excel/sheets are not able to load >1M rows, then, I import those tables into SQL Big query to do data cleaning and data transformation. This dataset is prepared for the data visualization to do analysis and find the pattern.

## **Clean Data**

### **Standardize Format**

During importing the data into SQL Big query, I found there is errors on the date column as it said error “Failed to parse input string "4/25/2016 9:37:30 AM". Then I added the schema manually and fill the date column as string type rather than date type.

As importing all the tables successful, then I created a new table with new name. I spotted some problems with the timestamp data. So before analysis, I need to convert it to date time format and split to date and time. Then I split the date and time column to extract the hours from the time column.

* Example for METs table:

CREATE OR REPLACE TABLE Fitbit2.METs2 AS (

  SELECT

    Id,

    DATE(PARSE\_TIMESTAMP('%m/%d/%Y %I:%M:%S %p', Date)) AS Date,

    TIME(PARSE\_TIMESTAMP('%m/%d/%Y %I:%M:%S %p', Date)) AS Time,

    EXTRACT(HOUR FROM PARSE\_TIMESTAMP('%m/%d/%Y %I:%M:%S %p', Date)) AS Hour,

METs

  FROM `new-390608.Fitbit2.METs`

)

The exact same code structure was repeated for other 4 tables.

Changes Include:

* activity\_data - “ActivityDate” chr –> date
* sleep\_data - “SleepDay” chr –> date and remove time
* weight\_data - “Date” chr –> date , “Weightpounds” –> NULL because we will use “WeightKg”, “Fat” –> NULL because it has 65 missing cells.

### **Identify Missing Value**

Then I identify if there is missing value in the table by writing this code. And this code is written for all tables and there is no missing value in METs2 table. If there is missing value in the table, we will either remove the column (if the data is not representative) or develops reasonable guesses for missing data (if the data is representative).

* Example for METS table:

SELECT \*

FROM Fitbit.METs2

WHERE Id IS NULL OR Date IS NULL OR Time IS NULL OR Hour IS NULL OR METs IS NULL;

The exact same code was repeated for other 4 tables.

### **Check Duplicate Value**

Then I find for any duplicate data that might have occurred in the data.

SELECT Id, Date, Time, Hour, METs, COUNT(\*) AS Count

FROM Fitbit2.METs2

GROUP BY Id, Date, Time, Hour, METs

HAVING COUNT(\*) > 1;

The exact same code was repeated for other 4 tables.

### **Consistent Labelling**

Then before I start exporting the data for analysis, I make sure the labelling for each table is consistent. For instance, all tables must be in small letters and put “\_” symbol to fill space.

# **Analyze**

In this phase, we will think analytically about the data, sort or format the data to understand it deeply, try to make sense of the data and we will try to find out what the data is telling us.

Before the data were uploaded into visualization tools, I am analysing the statistical information of the data using Big query.

## **Activity Analysis**

The sample population is divided into 4 activity level categories:

Inactive - average steps less than 5,000

Lightly Active - average steps less than 7500

Active - average steps less than 10,000

Very active - average steps more than 10,000

Then I create a new table and count how many ID were fall down into each of the category

SELECT

  StepCategory,

  COUNT(\*) AS IdCount

FROM (

  SELECT

    Id,

    AVG(TotalSteps) AS avg\_steps,

    CASE

      WHEN AVG(TotalSteps) < 5000 THEN 'Inactive'

      WHEN AVG(TotalSteps) BETWEEN 5001 AND 7500 THEN 'Lightly Active'

      WHEN AVG(TotalSteps) BETWEEN 7501 AND 10000 THEN 'Active'

      ELSE 'Very Active'

    END AS StepCategory

  FROM

    `case-study-fitness-tracker.Fitbit.ActivityDaily\_NT`

  GROUP BY

    Id

) AS subquery

GROUP BY

  StepCategory;

|  |  |  |
| --- | --- | --- |
| No | Category | No. of users |
| 1 | VeryActive | 7 |
| 2 | Active | 9 |
| 3 | LightActive | 9 |
| 4 | Inactive | 8 |

Approximately 21% of BellaBeat's users are categorized as "very active," indicating a high level of physical activity. Around 27% fall into the "light active" or "active" category, while 24% are considered "inactive." Overall, the analysis suggests that BellaBeat has a diverse user base in terms of activity levels.

## **METs Analysis**

The sample population is divided into 2 METs Categories:

Inadequate physical activity - Mean METs less than 7.5

Adequate physical activity - Mean METs more than 7.5

  SELECT

    Id,

    AVG(METs) AS avg\_METs,

    CASE

      WHEN AVG(METs) < 7.5 THEN 'Inadequate physical activity'

      ELSE 'Adequate physical activity'

    END AS METsCategory

  FROM

    `case-study-fitness-tracker.Fitbit.METs2`

  GROUP BY

    Id

|  |  |  |
| --- | --- | --- |
| Row | METsCategory | IdCount |
| 1 | Adequate physical activity | 33 |

Even though there is variety in category of users’ total steps, we found that all the users are actively engage with their physical daily activity. Means that for the inactive category, they might have lowest steps due to their working environment might be in office or something else. But when they had chances to exercise, they will do it. Take note that this is just an assumption, and we will need a more comprehensive data find the support evidence of our statement.

## **Sleep Analysis**

The sample population is divided into 4 sleep categories:

Bad - Mean percent asleep less than 70%

OK - Mean percent asleep between 70% and 90%

Good - Mean percent asleep between 90% and 95%

Great - Mean percent asleep greater than 95%

WITH SleepPercentages AS (

  SELECT

    Id,

    AVG(TotalMinutesAsleep / TotalTimeInBed) \* 100 AS PercentAsleep

  FROM

    `case-study-fitness-tracker.Fitbit.Sleep\_NT`

  GROUP BY

    Id

)

SELECT

  SleepQuality,

  COUNT(\*) AS Count

From (

  SELECT

  Id,

  CASE

    WHEN (SELECT PercentAsleep FROM SleepPercentages WHERE Id = s.Id) < 70 THEN 'Bad'

    WHEN (SELECT PercentAsleep FROM SleepPercentages WHERE Id = s.Id) BETWEEN 70 AND 90 THEN 'OK'

    WHEN (SELECT PercentAsleep FROM SleepPercentages WHERE Id = s.Id) BETWEEN 90 AND 95 THEN 'Good'

    ELSE 'Great'

  END AS SleepQuality

FROM

  SleepPercentages s

) t

GROUP BY SleepQuality;

| Row | SleepQuality | Count |
| --- | --- | --- |
| 1 | OK | 2 |  |
| 2 | Great | 6 |  |
| 3 | Good | 14 |  |
| 4 | Bad | 2 |

The results demonstrate that sleep habits respectively include 2 Bad, 2 OK, 14 Good, and 6 Great. Over 90% of the users’ time in bed needs to be asleep to be considered Good, so 20 of the 22 users are sleeping considerably well.

## **Weight Analysis**

The sample population is divided into 4 BMI categories:

Underweight - BMI less than 18.5

Healthy - BMI between 18.5 and 24.9

Overweight - BMI between 25.0 and 29.9

Obese - BMI more than 30.0

SELECT

  BMICategory,

  COUNT(\*) AS IdCount

FROM (

  SELECT

    Id,

    AVG(BMI) AS avg\_BMI,

    CASE

      WHEN AVG(BMI) < 18.4 THEN 'Underweight'

      WHEN AVG(BMI) BETWEEN 18.5 AND 24.9 THEN 'Healthy'

      WHEN AVG(BMI) BETWEEN 25.0 AND 29.9 THEN 'Overweight'

      ELSE 'Obese'

    END AS BMICategory

  FROM

    `case-study-fitness-tracker.Fitbit.Weight\_NT`

  GROUP BY

    Id

) AS subquery

GROUP BY

  BMICategory;

According to the 8 recorded users, 4 of them are in the “overweight” and 3 of them are “healthy”. The data demonstrates that BellaBeat devices users dominantly consist of overweight and healthy individuals. Again, this data would have been more accurate with a larger population and longer time frame.

# **Share**

## **Visualization**

The fifth phase is to share our data findings. We can do this with the help of visualization because putting information in the image can help people understand the analysis easily.

We are using power bi tools to do data visualization and get insight of the data overview. For Bi language, we are using DAX query to manipulate and transform data for our overview analysis and insight.

Kindly refer link attached for the data visualization:

## **Findings**

The analysis of user behaviour reveals valuable insights as following which can guide Bellabeat's marketing strategy effectively:

* Target Audience Identification: The findings suggest that most users exhibit higher activity levels on weekdays and prioritize rest and relaxation on weekends, particularly on Sundays. This indicates that Bellabeat's target audience may primarily consist of working adults who have active lifestyles during the week and seek relaxation and recovery during their days off.
* Peak Activity Hours: The peak activity hours shows that most users tend to do activities at 5-8 p.m. (after office hours) during weekdays and at 11-2 p.m. during weekend. This can help Bellabeat optimize the timing of marketing campaigns and push notifications. By sending relevant messages during users' most active periods, the company can increase the likelihood of engagement and response.
* Activity Preferences: Analysing users' activity types and preferences can guide Bellabeat in updating available features or developing new features that align with what users enjoy most. For instance, 100% users have the data for Daily and METs but only 73.72%, 57.58% and 24% users have data for sleep, heart rate and weight respectively. This means that daily and Mets is possibly the default features while the others are not. The company need to consider for further study why are the other features are not fully utilized. Possibly the method to key in data is time consuming or there is complicated process to use the features which unable to solve user’s issue.

# **Act**

The sixth phase of the data is to act. In this phase, we will use everything we have learned from our analysis and act upon it. We need to provide recommendations to the stakeholder on how to solve the business problem and help them make a good decision.

## **Recommendation**

* Weekday Fitness Solutions: With the understanding that users are more active during weekdays, Bellabeat can tailor marketing campaigns to promote features that support and enhance fitness and wellness during busy workdays. For example, they can emphasize the benefits of tracking steps, METs value, and heart rate during office hours or lunch breaks, offering solutions to stay active amidst a busy work schedule.
* Personalized Recommendations: Leveraging the data on users' unique activity patterns, Bellabeat can develop personalized recommendations based on individual needs and preferences. For example, they can send tailored notifications or emails suggesting activities or workouts that align with users' activity levels and goals for each specific day of the week.
* Promote Work-Life Balance: Based on the observation that users have varying activity levels on weekdays and weekends, Bellabeat can emphasize the importance of work-life balance in their marketing messaging. They can position their products as tools to help users maintain a healthy balance between their professional commitments and personal well-being.
* Social Engagement: Since Sundays record the highest sleep duration and lower activity levels, Bellabeat can encourage users to engage with social features within their app or community. This can include sharing sleep achievements, setting weekend relaxation challenges, or connecting with others for mutual support and motivation.