Bellabeat: Analyzing Fitness Smart Device User Data

Johnathan Ceja

Introduction:

Bellabeat is a high-tech manufacturer of health-focused smart products for women founded in 2013 by Urška Sršen and Sando Mur. The company has high potential for growth and can become a larger player in the global market for smart devices. Bellabeat could use an analysis of competitor smart device data to fuel where to place its own emphasis on future marketing and products,

Business Task:

The business task is to draw insights from non Bellabeat smart device users to guide Bellabeat's marketing strategy toward targeting key smart device user trends and potentially stimulate growth in the company.

Description of data:

The data source originated from a Kaggle dataset uploaded by a user named Mobius. It is an open dataset available for public use. The data was collected from a survey distributed by Amazon Mechanical Turk involving 30 eligible Fitbit users consenting to the submission of their personal tracker data from 03/12/2016 to 05/12/2016. The data was structured into multiple csv files. The csv files analyzed were named dailyActivities_merged.csv, dailyCalories_merged.csv, dailyIntensities_merged.csv, dailySteps_merged.csv, sleepDay_merged.csv, and weightLogInfo_merged.csv.

The 30 eligible Fitbit users fulfills the minimum sample size of 30 participants to achieve reasonable statistical power. However, the credibility of the data is low when testing the reliability, originality, comprehensiveness, currentness, and if it is well-cited. The data lacks reliability because it was collected from 30 individuals with unknown gender and limited information about the individuals. It is not stated what made these individuals eligible for the study. The data is not original either as it is collected from Amazon Mechanical Turk, a third party source and was collected seven years ago in 2016 so it is not very up to date. However, seven years is not a long time in this context because society has not changed the way overall fitness is achieved since then. The data is filled with many different fields and categories to analyze so it can be said to be fairly comprehensive. The best quality in this data is that the data collector and sources are cited well.

The data's integrity was tested by assessing that it was complete, accurate, consistent, and trustworthy. After running a few SQL queries which will be later shown in the report, we found that there were 33 total participants. Furthermore, inspection of the data on Google Sheets revealed that the sleep log data only contained data for 24 out of 33 participants and had gaps in dates where data was logged. The weight log data had even fewer participants with only 8 out of 33 participants logging their weight and wild data inconsistencies were found throughout multiple fields. Overall the data was incomplete and inconsistent. The data source could be said to be trustworthy because Amazon Mechanical Turk is a reputable source of data collection. The data is also accurate because the data itself was collected using personal trackers on Fitbit devices. When assessing the integrity and credibility of the data, we can see that the results of the analysis will be affected by the quality of the data.

Cleaning and Processing Data:

The first step taken was to import the data into a program. For the purposes of this analysis, the data was first imported into Google Sheets. Then,the format of the data was examined and found to be stored in the long format which is best for running SQL queries and transformations. Then, the data was checked for any null or blank values using conditional formatting and filters. None were found. Next, the remove duplicates tool was used to make sure the data was free of any duplicate entries. Three duplicates were found and cleaned on the sleepDay_merged sheet.

The remove whitespaces tool was used on the data but was not needed as the data contained no extra whitespaces. Finally, data types were made consistent across the columns and sheets. The columns named “SleepDay” and “Date” in the sleepDay_merged and weightLogInfo_merged sheets respectively were in the date time format which was unnecessary given the context of the data. These columns’ data were changed to the date type to keep typing consistent across every sheet and allow joins in SQL. It was also noted that the dailyActivity_merged sheet was a compilation of data from the dailyCalories_merged, dailyIntensities_merged, and dailySteps_merged sheets so that sheet alone could be analyzed . After being cleaned, the sleepDay_merged and weightLogInfo_merged sheets were exported into csv files imported into BigQuery along with every other original csv file. They were placed into a BigQuery dataset and given table names that simplify their file names. The schemas of each individual table were examined to get familiar with how they were imported in the context of SQL. A data integrity test was performed to check the number of participants across the different sheets. The SQL query counted the number of different Ids in each table. The following were the queries.

```
-   SELECT COUNT(DISTINCT Id)
    FROM `bellabeat_casestudy.dailyactivity`
-   SELECT COUNT(DISTINCT Id)
    FROM `bellabeat_casestudy.dailycalories
-   SELECT COUNT(DISTINCT Id)
    FROM `bellabeat_casestudy.dailyintensities
-   SELECT COUNT(DISTINCT Id)
    FROM `bellabeat_casestudy.dailysteps
-   SELECT COUNT(DISTINCT Id)
    FROM `bellabeat_casestudy.sleepday
-   SELECT COUNT(DISTINCT Id)
```

```
        FROM `bellabeat_casestudy.weightloginfo
```

The first four queries returned a count of 33 distinct Ids which was different from the data's description of 30 participants. That was noted. The fifth query returned a result of 24 distinct Ids and the sixth query returned a result of 8. The sleepday table only had 24/33 participants included while the weightloginfo table only had 8/33 participants included. Because of this, the weightloginfo table was excluded from the rest of this study as the data itself was very incomplete. That concludes the cleaning and manipulation process for this dataset.

Analyzing Data:

The business task of the analysis is finding out non Bellabeat user trends that can help guide Bellabeat's marketing strategy. So, the data given was inspected for any interesting trends in user behavior that Bellabeat could incorporate in future products and marketing. After examining the fields, an analysis on multiple metrics was run. The metrics were total steps, total distance, distances at the sedentary, lightly active, fairly active, and very active intensities, minutes at the sedentary, lightly active, fairly active, and very active intensities, total calories, total minutes asleep, and total minutes in bed.

Multiple hypotheses were created to test during this analysis. The first hypothesis was that the average total steps had a positive effect on the average total calories burned throughout the time that data was collected. Since averages of fields were likely to be referred to very often during the analysis, a temporary table was created with the following query.

```
WITH
  avg_totals
  AS
  (SELECT Id,AVG(TotalSteps) AS avg_total_steps,
AVG(TotalDistance) AS avg_total_dist,

AVG(TrackerDistance) AS avg_tracker_dist,

AVG(LoggedActivitiesDistance) AS avg_logged_activity_dist,

AVG(VeryActiveDistance) AS avg_very_active_dist,

AVG(ModeratelyActiveDistance) AS avg_moderate_active_dist,

AVG(LightActiveDistance) AS avg_light_active_dist,

AVG(SedentaryActiveDistance) AS avg_sedentary_active_dist,

AVG(VeryActiveMinutes) AS avg_very_active_mins,

AVG(FairlyActiveMinutes) AS avg_fairly_active_mins,

AVG(LightlyActiveMinutes) AS avg_light_active_mins,

AVG(SedentaryMinutes) AS avg_sedentary_minutes,

AVG(Calories) AS avg_calories

  FROM `stunning-yeti-371722.bellabeat_casestudy.dailyactivity`

  GROUP BY Id)

SELECT *

FROM avg_totals
```

This query populated a temporary table with average values from each activity metric to be studied for each participant. From now on, it will be assumed that each query mentioned will contain the WITH statement above it. The next query displayed data for the average steps and calories for each participant in starting with those with the most average steps to the least.

```
SELECT Id,avg_total_steps,avg_calories

FROM avg_totals

ORDER BY avg_total_steps DESC
```

It was noted that those with more steps seemed to have a greater amount of calories burned. Now, total average distance and average calories were compared with the next query.

```
SELECT Id,avg_total_dist,avg_calories
FROM avg_totals
ORDER BY avg_total_dist DESC
```

A similar pattern was found. It seemed that the more average distance that a participant had, the more average calories they burned over the course that data was collected. This would be a positive correlation between the mentioned fields. Now, average time under different exercise intensities was compared to average calories burned for each participant using the following queries.

```
SELECT Id,avg_sedentary_minutes,avg_calories
FROM avg_totals
ORDER BY avg_sedentary_minutes DESC
```

```
SELECT Id,avg_light_active_minutes,avg_calories
FROM avg_totals
ORDER BY avg_light_active_minutes DESC
```

```
SELECT Id,avg_fairly_active_minutes,avg_calories
FROM avg_totals
ORDER BY avg_fairly_active_minutes DESC
```

```
SELECT Id,avg_very_active_minutes,avg_calories
```

```
FROM avg_totals
ORDER BY avg_very_active_minutes DESC
```

Here we see that average sedentary minutes and average light minutes don't seem to have a drastic variance in average calories burned when comparing those who logged the most minutes to the least minutes in the respective fields. It seems as though there is little correlation between these fields. For those who logged more average fairly active minutes however, it seems there may be more average calories burned throughout the time data was collected when compared to those who logged less. There may be a positive correlation here. It also seems as though average very active minutes had the strongest correlation with calories burned than the others as the participants with higher minutes logged seemed to burn lots more calories than those who had the less minutes logged. Later, visualizations on Tableau will be created to clearly demonstrate the correlations and their strength should they exist.

From this point, the average_totals table is no longer required so the WITH statement is no longer added above each query. A hypothesis about average sleep time affecting average calories burned was tested by running the following query.

```
SELECT activity.Id, ,AVG(sleepday.TotalMinutesAsleep) AS average_sleep_mins
AVG(activity.Calories) AS average_calories
FROM `bellabeat_casestudy.dailyactivity` activity
LEFT JOIN `bellabeat_casestudy.sleepday` sleepday
ON activity.Id = sleepday.Id AND activity.ActivityDate = sleepday.SleepDay
GROUP BY Id
HAVING average_sleep_mins IS NOT NULL
ORDER BY average_sleep_mins DESC
```

There didn't seem to be a correlation here when examining the participants' data. Next, a hypothesis on whether overall rest time affected sleep was tested. Data comparing average sedentary time and average total minutes asleep was called with the following query.

```
SELECT activity.Id,AVG(sleepday.TotalMinutesAsleep) AS avg_sleep_mins,
AVG(activity.SedentaryMinutes) AS avg_sedentary_mins
FROM `bellabeat_casestudy.dailyactivity` activity
LEFT JOIN `bellabeat_casestudy.sleepday` sleepday
ON activity.Id = sleepday.Id AND activity.ActivityDate = sleepday.SleepDay
GROUP BY Id
HAVING avg_sleep_mins IS NOT NULL
ORDER BY avg_sleep_mins DESC
```

Examining the data shows that there may be a strong negative correlation with more sleep time resulting in less sedentary minutes. Since there was a correlation, an idea to check if sedentary minutes and time in bed has any type of correlation arose. The next query pulls data pertaining to that.

```
SELECT activity.Id,AVG(sleepday.TotalTimeInBed) AS avg_bedtime,
AVG(activity.SedentaryMinutes) AS avg_sedentary_mins
FROM `bellabeat_casestudy.dailyactivity` activity
LEFT JOIN `bellabeat_casestudy.sleepday` sleepday
ON activity.Id = sleepday.Id AND activity.ActivityDate = sleepday.SleepDay
GROUP BY Id
HAVING avg_bedtime IS NOT NULL
ORDER BY avg_bedtime DESC
```

Once again, there seems to be a negative correlation between these two fields. So it seemed the more average time in bed a participant spent, the less average sedentary time they logged. Now that concludes the SQL portion of the analysis as there are now some key insights that can possibly be used for further visualization through Tableau and recommendations.

Visualizing Data:

Visualizations for the data were produced using Tableau. Only two csv files were uploaded to BigQuery as that is all that would be necessary to produce the visualizations. The files uploaded were dailyActivity_merged.csv and sleepDay_merged_cleaned.csv, The two files were linked by their date fields and Id fields. The first step taken here was to produce a visualization of total steps and total distance vs calories. Because steps and distance are two similar metrics, they were chosen to be presented on the same worksheet. Given the number of points, they were also visualized as a scatterplot. Now, a trendline was added to the scatterplots and it was now clear that there was a positive correlation between both total distance and steps when compared to calories burned. The coefficients of determination were 0.33, and 0.4 respectively indicating that there was at least some reliability to the strength of the relationships. The p values were less than 0.05 as well which shows statistical significance. Next, four scatterplot was created that visualized time spent under each different exercise intensity vs calories burned. Trend lines were also added to these visualizations. The coefficients of determination were 0.001, 0.05, 0.07, and 0.36 for sedentary, lightly active, fairly active, and very active minutes vs calories respectively. With the coefficients of determination being very far away from 1, it can’t be said that any of the relationships are very strong at all besides the very active minutes having some signs of a positive correlation. The p value for the sedentary minutes vs calories was also at 0.29 which is a lot higher than the 0.05 or less that is necessary for statistical significance.The other intensities all did carry a p value less than 0.05 however. There was some clear positive correlation coming from very active minutes and calories burned according to the trend line. The next visualizations created were comparing total time in bed and total minutes asleep vs sedentary minutes. For these two separate worksheets were created with scatter plots and trend lines for each comparison. The trend lines clearly show negative correlations with coefficients of determination of 0.38 and 0.36 for total time in bed vs sedentary minutes and total minutes asleep vs sedentary minutes respectively. The respective p values showed statistical significance being under 0.05. The two worksheets were taken and fit onto a dashboard to show how resting was analyzed clearly as one graphic. The key findings were compiled into a slideshow presentation with recommendations.

Recommendations:

After conducting technical analysis on the dataset, a few insights have been drawn. The first is that total steps and distance have a positive effect on calories burned. The recommendation for Bellabeat is to equip smart devices with smart reminders catered to users that are not logging a high count of total steps and total distance and encouraging them to go on a walk or run. This would help users to burn more calories and get better overall desired fitness results from their Bellabeat smart devices. Smart reminders can also be created for users who are logging less time performing high intensity exercises. Challenges can be part of the smart device reminders that can encourage users with guided high intensity exercise. High Intensity Interval Training (HIIT) can be promoted by these challenges and the exercises can be taught by Bellabeat apps and devices themselves. Users should see an increase in calories burned since very active minutes positively affects calories burned as shown by the analysis. Smart reminders on Bellabeat devices can also be issued to target users that have poor sleeping and bed rest habits. Those who spend less time sleeping or in bed resting should receive reminders telling them to get some proper sleep or rest. The data analysis indicated that proper sleep and rest results in less sedentary time and potentially higher energy levels throughout the day. Users would benefit from adding this feature Lastly, health information based on the importance of rest, sleeping, intense exercise, and moving should be added to Bellabeat's apps and devices so that users are informed on their importance for overall health and fitness goals. It would be very convenient and beneficial to have information natively available on Bellabeat apps and could serve users' fitness journeys. The marketing team can get started by promoting "smart reminders", "smart challenges", and "health information" as selling points once they are implemented into Bellabeat smart devices. This concludes the analysis on non Bellabeat smart device data and there is a high hope that these recommendations will bring about future growth and prosperity for Bellabeat.