Bellabeat Case Study

Business Questions:

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. Do competing smart devices fill a niche not covered by Bellabeat?

Deliverables:

- A clear summary of the business task
- A description of all data sources used
- Documentation of any cleaning or manipulation of data
- A summary of your analysis
- Supporting visualizations and key findings
- Your top high-level content recommendations based on your analysis

ASK:

Exploring Bellabeat smart products versus similar products in the market to evaluate gaps in marketing.

Looking for uses or features that appear to be directly correlated to sales, positively or negatively.

Study is being presented primarily to primary stakeholders and marketing teams for use in marketing direction.

Marketing centered around female fitness and a female first approach to health. Focus is put on female specific needs when it comes to health and fitness.

PREPARE:

Given FitBit Fitness Tracker dataset on Kaggle.
https://www.kaggle.com/arashnic/fitbit/metadata

Data stored in wide and narrow formats in CSV files.

Data comes from verified Kaggle dataset as a secondary source, primary source comes from Zenodo as a part of a larger dataset which was collected by RTI International with a creative commons attribution 4.0 International Public License.

Furberg, R., Brinton, J., Keating, M., & Ortiz, A. (2016).
Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016 [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.53894

There is additional data to analyze if necessary:
https://zenodo.org/record/53894#.Yg6GOejMJD9

Data appears to be sorted by three categories of related data: measures by the minute, measures by the hour and daily measures. To identify trends I am going to ignore the data gathered by the minute as it is summarized in the other data sets.

PROCESS:
Initial thoughts to guide cleaning:
Looking for trends in highly active times or high calorie burn to see if there are any meaningful insights to when the sample population is exercising.

Look for daily or weekly trends, and if people tend to have consistent routines or not.

Look at highly active minutes as percentage of overall activity versus total calories and total steps.

Look at weight loss trends of logged and explore why only 8 out of 30 logged weight.

Explore which features appear to be used the most.

Look for average calorie burn versus time using smart devices.

Initial questions about missing data:
Daily activity has 33 unique ids
Weight loss info has 8 unique ids
Sleep day has 24 unique ids

Cleaning Process:
First establishing which data sets I am going to clean. Based on the structure of the data I am presented with I am going to focus on six of the eighteen tables. I am omitting twelve of the data tables because they are summarized in one or more of the six tables that I will be looking at.

Therefore the tables that I will be cleaning for my analysis includes:Hourly_Intensities, Hourly_Steps, Hourly_Calories, Sleep_Day, Weight_Log_Info, Daily_Activity

To start cleaning I am first verifying important pieces of the data against what is known of the data set and some of the other tables of which I am omitting.

First notice is that the metadata claims records of 30 individuals however the Ids record 33 unique individuals. This is something to consider during analysis.

Next I trimmed any whitespace and removed any duplicates from all six tables. Then I combined all of the hourly table measures into one table and renamed all of my tables to accurately represent the data contained within them.

Then I used column stats in sheets to check for empty values and take an initial look for any outliers in my data set. I found no empty values, but I found some values in the sleep data that will need to be examined if I need to use them in my analysis. There was also an outlier in my weight log data although upon further examination it appears to be accurate data that should be included in analysis.

Data Cleaning Change Log:

- Created a copy of all 6 sheets to combine into a single workbook.
- Removed 3 duplicate rows from the Daily_Sleep sheet.
- Combine HourIntensity, HourStep, and HourCalorie into one table.

ANALYZE:

Initial hypothesis about trends:

Customers tend to be more interested in smart products to view calories burned and workout activity than to see weight loss or sleep activity.

Customers tend to not wear smart devices all day long and instead for specific activities.

Customers who use smart devices more often burn more calories on average.

Analysis Process:

My first thought for analysis was to see on a daily basis what time people were working out. Upon creating a pivot table to compare average steps, average intensity (derived from heart rate) and average calories burned versus time sorted by hour I initially saw no interesting results. All three graphs told relatively the same story which was that activity steadily climbed from 4 AM to about 8 AM where it then stayed within a small range until about 7 PM when activity would drop off for the night. Not satisfied with that conclusion I filtered my data to show average intensity that was greater than the average daily intensity which was 12 and I noticed something strange. The purpose of the filter was to get rid of inactive data points in order to look for trends for when people actually were active. So when filtered there was a major spike in activity at 5 AM which then dropped off at 6 AM which was surprising, it seemed that people were exercising at 5 AM. I looked further into this data by filtering it by only the 5 AM data points to find that the spike at 5 AM actually came from a single Id (person) and more than that, the calories and intensity values were the exact same high value (669 and 165 respectively) across 16 days with the rest of the days being close to 0 for both. Yet more interesting than that was the fact that such high calorie burn was achieved without recording any steps and so I filtered the data by this single id to investigate the surrounding data points only to find minimal steps on either of the high calorie burn which led me to think I must just exclude this id from my analysis. To be cautious of any other strange data I decided to examine other extreme values. So I filtered my same pivot table by max values instead of averages to find some extremely high but not impossible values. To further investigate these max values I moved from my hourly activity breakdown to my daily activity breakdown and created a new pivot table consisting of max daily steps, max tracker distance, max daily calories burned and max very active time. These results left me extremely shocked because the entire data set did not seem to make any logical sense. One such discrepancy comes from the id with the highest total step count, 36019, and distance, 28.03, burning a maximum of 2690 calories in a day versus the id with the maximum burned calories, 4900, only having a step count of 19542 and distance of 15.01. There are 16 of the 33 ids that have a higher max calories burned than the id with the highest activity metrics with one shocking user claiming 3101 calories burned from 5.35 distance and 8360 steps. Concerned with the validity of the data or my cleaning process I cross referenced my confusing entries with my original copy as well as the original data source to find that I made no mistakes during my cleaning. From here my next thought was to compare my work with my peers working on the same case study to which I could not find anyone who found the same discrepancies, but instead used the data with these weird points included. So I reached out to a mentor to run my questions on the data set and he suggested analyzing the correlations between my variables. Upon doing that I found very poor correlation between any metric you could choose and calories burned, the highest correlation being a r coefficient of 0.645 between tracker distance and calories burned. Doing further research on the way that calories are burned, I learned that there are many other factors that need to be considered like weight, sex, and muscle mass to name a few. While still unsure if these other factors could have an impact as great as the discrepancies found in the data I would find it to be unhelpful to analyze the data for any further trends without verifying the accuracy of the data which can not be done with the data set provided.

SHARE:

Interactive dashboard in tableau to allow for easy viewing of data concerns.

ACT:

My final conclusion for the stakeholders and marketing team would be to evaluate what measurements Bellabeat smart devices offer and how they can more accurately be used to assess health and calories burned when compared to competing brands which appear to have inaccurate readings. I think it would be a good idea to recreate a more controlled case study around the provided metrics to determine the validity of the provided data set. The insights I gained from this analysis are that current fitness metrics measured by fitbit devices poorly correlate together giving a confusing story for a user looking to burn more calories. My suggestions for the stakeholders would be to focus on branding towards accurate readings and more specialized metrics to the individual. And one final observation was the lack of data available for sleep tracking and weight tracking which would lead me to hypothesize that either users found it too difficult to use and therefore did not or that these features are not as important to users which could be explored by surveying users on feature preferences. One potential future solution to capitalize on this market would be to offer pairing products like a weight scale and a sleeping pillow which would sync up to gather these metrics automatically, eliminating difficulty of use from the equation.

FURTHER ANALYSIS:

Upon looking further into the data and really exploring it through the use of Tableau I was able to make a few more observations. The biggest observation that I found was that when comparing the scatterplots for total steps vs calories and total active minutes vs calories, when I would select the most correlated points on the active minutes graph, the correlation on the steps vs calories graph would go up significantly. My thought behind this is that when the fitbit tracker is most accurately tracking the activity of someone then it would register as anything other than sedentary activity. This is important because I hypothesize that the reason for the poor tracking data might be due in part to bad readings although for the data that seems to be more accurate I think some of the poor quality might come from how users would generally wear their smart device. As a personal observation on my own smart device I get a worse pickup on my tracked activity at the gym if my watch is not fastened tightly which I suspect makes it harder to get readings like steps and heart rate.

Analyzing another measure of calories burned per minute vs the total active minute (not sedentary) you can clearly see absolutely no correlation (r squared value of .148) which is concerning when active minutes generally should increase the calories you burn in a minute.