Google Data Analytics Capstone Project

Case Study 2: How Can a Wellness Technology Company Play It Smart?

SCENARIO

You are a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analysing smart device fitness data could help unlock new growth opportunities for the company. You have been asked to focus on one of Bellabeat's products and analyse smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy.

Full background information and guidance for the project can be found here.

BUSINESS TASK

Analyse data from the users of smart fitness tracking devices to identify trends that can be used to improve the utility of Bellabeat products, the company's marketing strategy, and the satisfaction of customers.

DATA OVERVIEW

Data provided for the project included publicly available crowd-sourced Fitbit datasets from Kaggle (https://www.kaggle.com/datasets/arashnic/fitbit).

Available metadata:

- The data was originally sourced from Zenodo [1] and was collected between 2016-03-12 and 2016-05-12 via Amazon Mechanical Turk platform [1, 2];
- The data from 30 participants (each assigned a unique participant ID) were collected but the criteria for the participant selection were not provided;
- All types of Fitbit devices were accepted [2];
- The following variables were collected:
  - Daily: steps, distance, burned calories, sleep time, active versus sedentary time, body weight, body fat, and body mass index (BMI);

- Hourly: steps, burned calories, active versus sedentary time, heart rate and sleep;
- Minutely: steps, burned calories, physical activity intensities, sleep, and metabolic equivalents of task (METs).

Possible biases:

- The data might not be representative of the general public, as only people who had Fitbit devices, or had the financial means to purchase one were able to participate in the study. However, because Bellabeat's target population are women using or interested in using personal fitness tracking devices, this limitation might not be an issue.
- The data was collected in 2016 and thus might not accurately represent current trends. However, due to their biological nature, the relationships between metrics recorded in the provided datasets are unlikely to change with time.
- The gender of the participants was not specified in the provided datasets. As Bellabeat specialises in wellness tracking devices for women, trends observed in this data might not be fully representative of the target population.

DATA ANALYSIS TOOLS

The data were cleaned and analysed using Python programming language on a Jupyter notebook platform. Comprehensive notes documenting the details of the analysis performed for this project can be found here. Graphics displayed in this report were generated using Tableau Public.

DATA CLEANING AND INTEGRITY ASSESSMENT

The following steps were taken to improve the integrity and reliability of the data:

- No missing values were detected apart from the information on body fat which was not used for the analysis.
- Duplicated values were detected, investigated, and removed from 'sleepDay_merged' dataset.
- For most measured variables, data from 33 unique IDs were detected. This is inconsistent with the number of participants indicated in the metadata (30 individuals). Possibly, some people were using several personal fitness tracking devices with individual IDs or more people participated in the study than specified. However, this cannot be verified without additional information about how data was collected.

- Data from less than 30 IDs were present for variables on sleep and body composition. This likely reflects the differences in device capabilities and individual usage preferences.
- The data on body composition included information from 8 participants with only two of them contributing more than 5 entries. Data with a such small sample size is likely unreliable and thus was not used for further analysis.
- A likely mistake was detected in the 'minuteMETsNarrow_merged' dataset. MET is used to describe the intensity of physical activity compared to the resting state [3]. Thus, during sleep, MET is around 1. However, in this dataset it was found to be around 10 during hours people are usually asleep (12 am to 8 am). This suggests that MET values were accidently multiplied by 10 during handling. To correct this for further analysis, MET values were divided by 10.
- To check the integrity of data between datasets, daily steps and burned calories were compared to daily steps and burned calories calculated from values of steps and burned calories measure hourly or minutely. 934 out of 940 entries in 'dailyActivity_merged' dataset contained all of this information. 802 observations were found to have a discrepancy in burned calories data, and 159 in total steps data. These differences might be a result of rounding error or mistakes during data entry and handling. To increase the reliability of the data, entries with more than 5% differences between daily values and daily values calculated from hourly and minutely values for steps and burned calories were removed. This eliminated 38 observations leaving 896 entries in the 'dailyActivity_merged' dataset to be used for further analysis.

Overall, multiple discrepancies were detected in the data suggesting that it can only be used for preliminary analysis and any findings should be verified when data of higher quality is available.

DATA ANALYSIS

Organising the data

From the metrics available in the provided datasets, sleep duration had the highest practical importance for the overall health and wellbeing. Thus, possible correlations between sleep duration and other variables were the main focus of the analysis.

METs were selected as the most objective metric to evaluate the intensity of physical activity. As mentioned earlier, MET is a standard unit defined as a rate of energy expended during an activity compared to the rate of energy expended

during rest [3]. To transform the minutely MET data to a more convenient format, minutely MET values were used to calculate average daily METs per minute for each day for each participant.

To construct the final dataset used for the analysis, data from the 'sleepDay_merged' dataset was merged to the entries from the 'dailyActivity_merged' dataset (cleaned and filtered as described above), and then to the calculated average daily METs per minute. Only observations containing the information on all the metrics of interest were retain. The resulting merged dataset contained 393 entries from 24 participants. 15, 12, and 4 participants provided data for more than 14, 21, and 28 days respectively.

Exploring the Correlations and Trends

A moderate strength (R-squared = 0.39) negative correlation was detected between the total daily sedentary minutes and the total daily minutes asleep (Figure 1). There are two main things to consider regarding this observation:

- Longer time spent sitting during a day might negatively affect the duration of sleep, but it is just as likely that fewer minutes asleep might lead to more time being spent sitting due to lower energy levels.
- Time spent sitting might be associated with sleep duration directly, or simply be a proxy for the overall physical activity and daily energy expenditure.

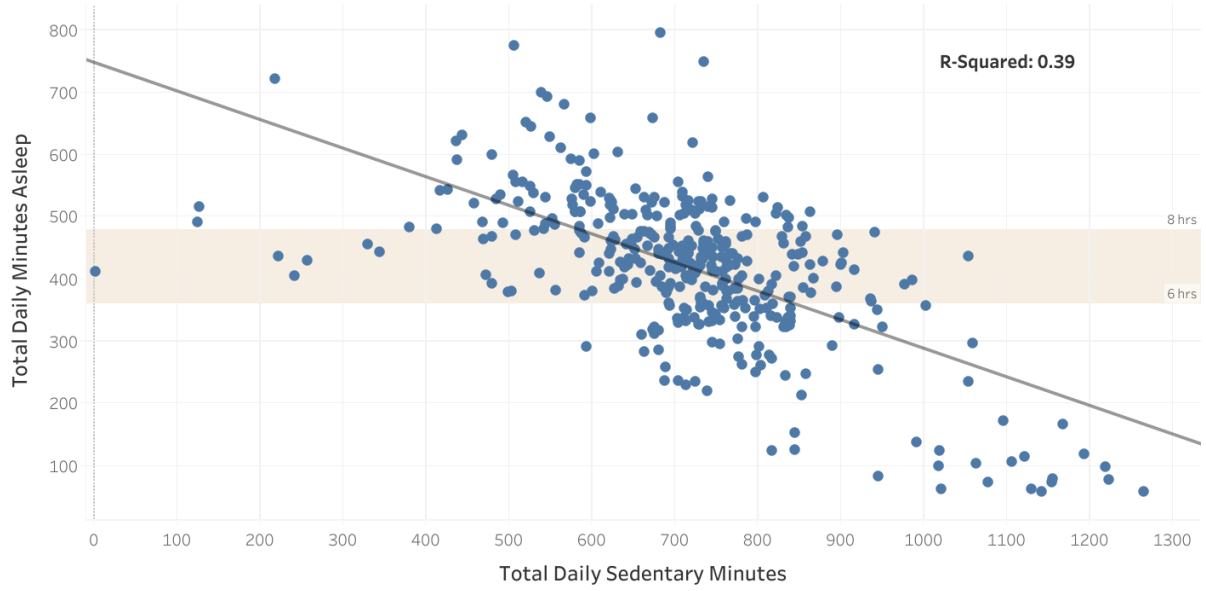

Figure 1. The relationship between sedentary time and sleep duration. The highlighted area of the graph indicates the interval of normal sleep duration (6 – 8 hours).

These ambiguities can be addressed by examining the correlation between daily sedentary minutes and physical activity. Indeed, time spent sitting does not seem to show any strong correlations with total daily steps, burned calories, or average daily METs per minute (Figure 2), indicating that sedentary time is not a great predictor of person's overall physical activity. This suggests that negative effects of long sitting time on sleep duration might be independent of one's physical activity and energy expenditure. For example, longer sedentary time might possibly be associated with longer work hours and thus higher stress levels, or improper posture while sitting might lead to musculoskeletal problems that impact the duration of sleep.

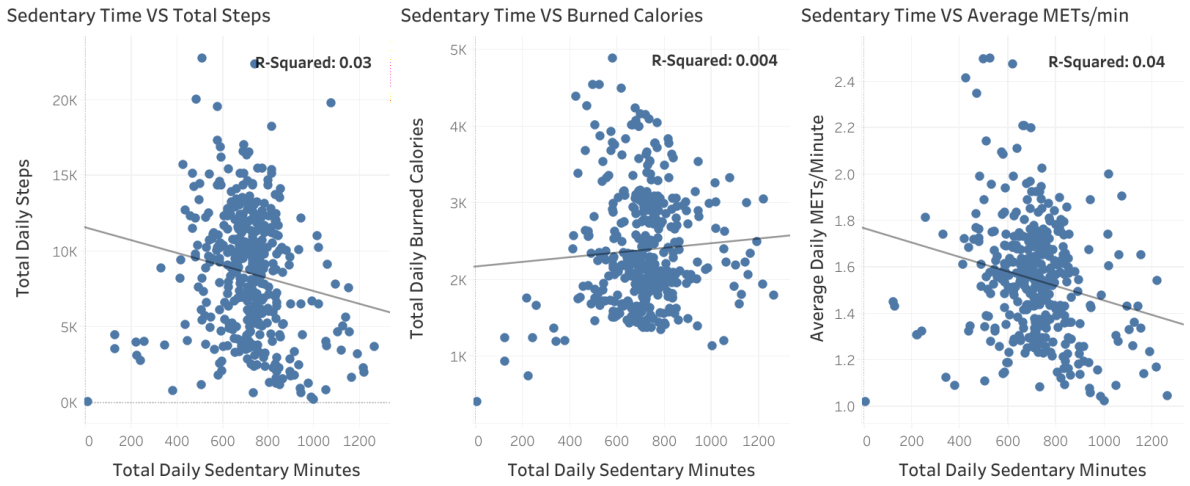

Figure 2. The relationship between sedentary time and physical activity or energy expenditure.

In contrast to sedentary time, physical activity (measured in average daily METs per minute) was not found to be correlated with sleep duration in the linear regression analysis (Figure 3). However, an observation was made, that the spread of data seemed to decrease as physical activity increased. To test this observation, the values of average daily METs per minute were separated into quartiles designated the 'Lowest Activity' (1.02 – 1.36 average daily METs per minute), 'Light Activity' (1.36 – 1.55 average daily METs per minute), 'Moderate Activity' (1.55 – 1.69 average daily METs per minute), and 'Highest Activity' (1.69 – 2.50). The difference in the variance of data between each quartile was then investigated using the Levene's test (Table 1). Indeed, a statistically significant difference in the variability of data was detected between the 'Lowest Activity' and 'Moderate Activity' quartiles ($p=0.05$). The p-value describing the difference in the spread of data between the 'Lowest Activity' and 'Highest Activity' quartiles ($p=0.06$) was just above the threshold of statistical significance strongly suggesting that the observed difference is also real.

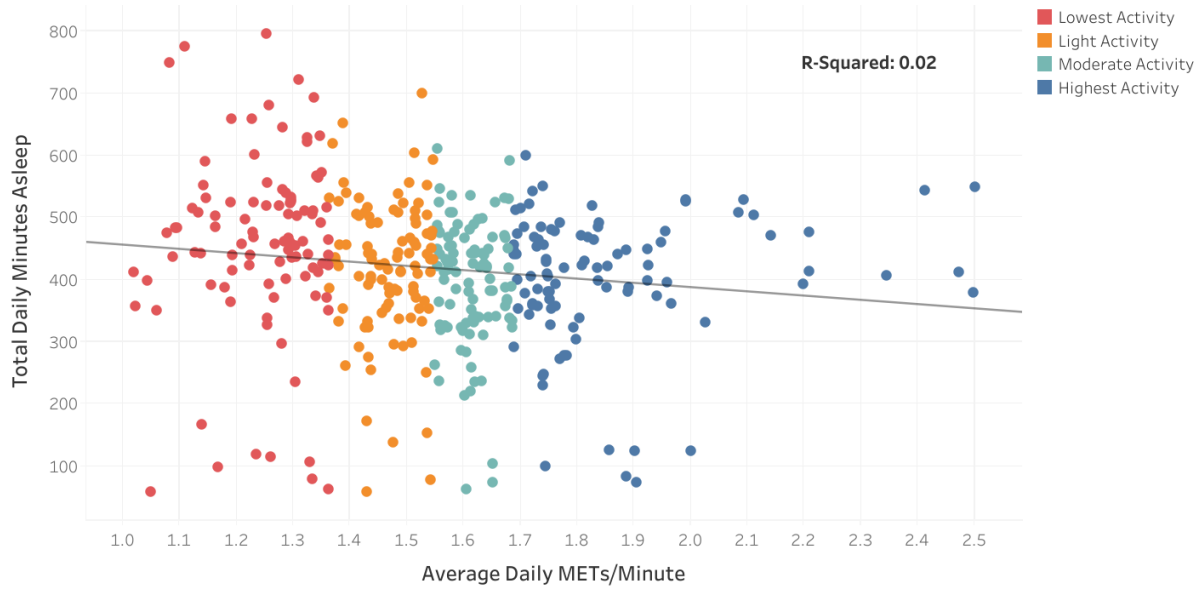

Figure 3. The relationship between average daily METs per minute and sleep duration. MET values are separated into quartiles from the lowest to the highest activity.

Table 1. P-values of the Levene's test comparing the variability of data between different physical activity quartiles.

Quartiles	Lowest Activity	Light Activity	Moderate Activity	Highest Activity
Lowest Activity	1.00	0.14	0.05	0.06
Light Activity	0.14	1.00	0.60	0.63
Moderate Activity	0.05	0.60	1.00	0.99
Highest Activity	0.06	0.63	0.99	1.00

As both undersleeping and oversleeping are associated with negative health outcomes [4], lower variation in the total daily minutes asleep might indicate an increase in sleep events with a normal duration (data points clustering within 6 – 8 hours range). To investigate this further, all sleep duration entries were assigned to 'undersleeping' (<6 hours), 'normal' (6 – 8 hours), or 'oversleeping' (>8 hours) categories [4]. Categories were then expressed as percentages of the total events for each physical activity quartile. A notably higher occurrence of oversleeping events was observed in the 'Lowest Activity' quartile compared to the other quartiles (Figure 4). At first glance it would seem that people with the lowest physical activity were undersleeping less than the other groups, but this is because of their disproportionally higher tendency to oversleep. Even light physical activity (1.36 – 1.55 daily average METs per minute) appeared to positively affect the duration of sleep decreasing the frequency of oversleeping events. Overall, there was an upward trend in the percentage of sleep events with normal duration as physical activity levels increased. This is supported by a statistically significant difference in the frequency of sleep events with normal duration observed between the 'Lowest Activity' and the 'Highest Activity' quartiles (chi-square test; p=0.05).

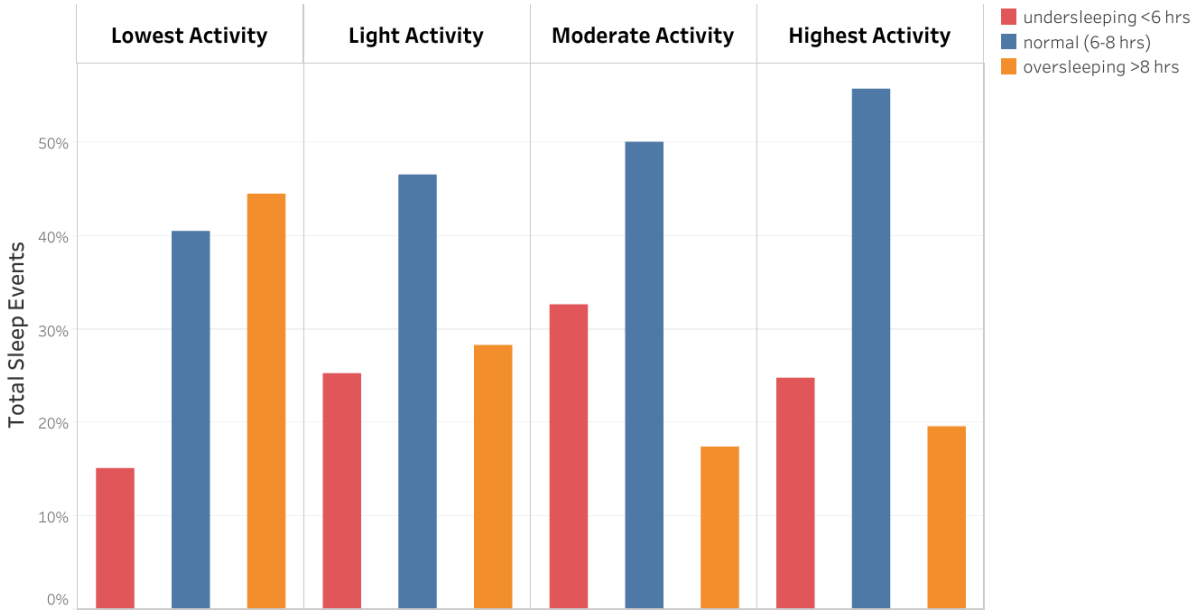

Figure 4. The frequency of undersleeping, oversleeping and normal duration sleep events within different physical activity quartiles.

To gain insight into people's physical activity throughout a day, average hourly intensities from the 'hourlyIntensities_merged' dataset were calculated using all entries from all participants. Two distinct periods of daily physical activity can be identified in the data: from 5 am to 4 pm and from 4 pm to 8 pm (Figure 5). The first period likely reflects work related tasks, and the second, personal errands and purposeful exercise.

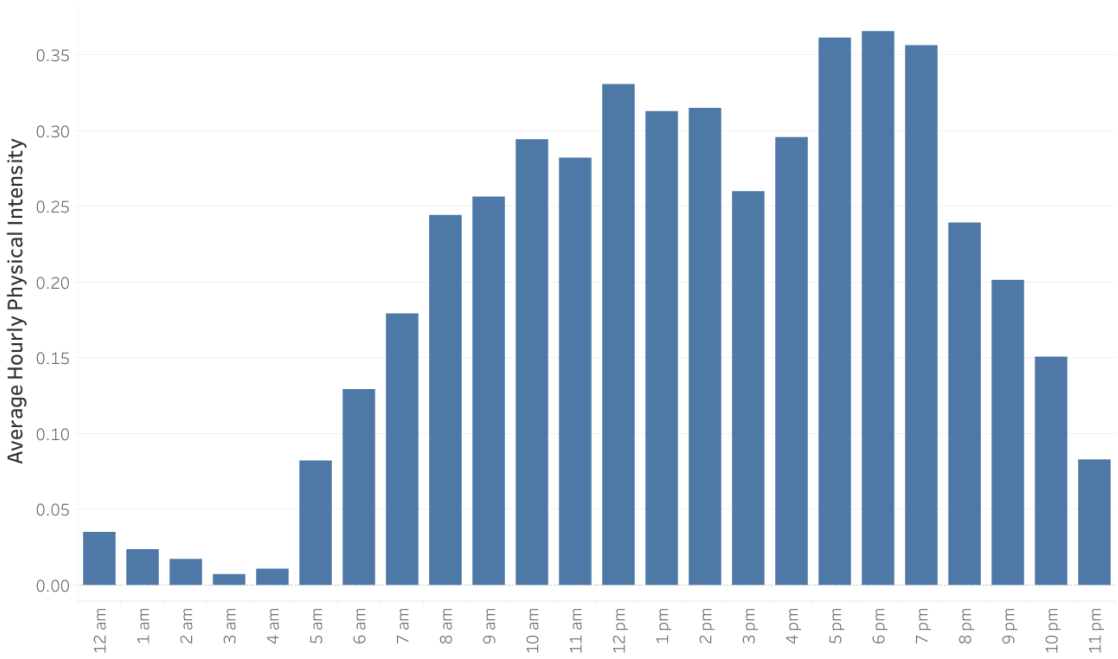

Figure 5. Average hourly physical intensity throughout a day.

KEY TAKEAWAYS

- The quality of data provided is suboptimal, and the results of this analysis should be treated as preliminary.
- Longer time spent sitting was found to be correlated with a shorter sleeping time and this association seems to be independent of the overall physical activity. The linear regression analysis suggests that sedentary time should be limited to 850 minutes. As this includes sleeping time, an average person should not be sitting for more than 430 minutes (~7 hours) a day.
- Physical activity was associated with a higher frequency of sleep events with normal duration (6 – 8 hours), and a disproportionately high tendency to oversleep was observed in the 'Lowest Activity' quartile (1.02 – 1.36 average daily METs per minute).

RECOMMENDATIONS

- A personalised sleep optimisation functionality might be introduced into the Bellabeat app. For users sleeping less than 6 hours or more than 8 hours, the app could remind to take regular breaks from sitting. These reminders might be primarily targeted for the time window between 5 am to 4 pm, when people are at work. From 4 pm to 8 pm the app could encourage people to partake in more intense exercise with the goal of reaching at least 1.36 – 1.55 average daily METs per minute (the range of 'Light Activity' quartile). To attain 1.5 daily average METs per minute, people need to accumulate at least 6 hours of light physical activity, such as slow walking, 2.4 hours of moderate physical activity, such as brisk walking, or 1.5 hours of intense physical activity, such as running [3]. This seems to be quite feasible as an average person in this study was found to attain 1.47 average daily METs per minute. As sleep is a complex physiological process, many factors can affect its duration [4]. Bellabeat app could track the changes in the total sedentary time, average daily METs per minute, and daily sleep duration of users over time, to determine if the changes suggested by the app have any positive effects. Depending on the outcome of such analysis, the app could inform the users about the likelihood of their issues regarding sleep duration being associated with daily sedentary time and physical activity.
- Bellabeat could enhance their marketing strategy by emphasizing the significance of sleep duration and its correlation with daily activity, while highlighting how personalised Bellabeat services can assist women in improving their sleeping experiences.

REFERENCES

1. 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
2. Brinton JE, Keating MD, Ortiz AM, Evenson KR, Furberg RD. Establishing Linkages Between Distributed Survey Responses and Consumer Wearable Device Datasets: A Pilot Protocol. JMIR Res Protoc. 2017 Apr 27;6(4):e66. doi: 10.2196/resprot.6513. PMID: 28450274; PMCID: PMC5427248.
3. Mendes MA, da Silva I, Ramires V, Reichert F, Martins R, Ferreira R, Tomasi E. Metabolic equivalent of task (METs) thresholds as an indicator of physical activity intensity. PLoS One. 2018 Jul 19;13(7):e0200701. doi: 10.1371/journal.pone.0200701. PMID: 30024953; PMCID: PMC6053180.
4. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. Eur Heart J. 2011 Jun;32(12):1484-92. doi: 10.1093/eurheartj/ehr007. Epub 2011 Feb 7. PMID: 21300732.