Project Report:
Analysis, Observations & Recommendations

Data Analyst: William Rodrigues

Capstone Case Study
Google Data Analytics
January 14, 2022

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

A. Purpose and Scope

Bellabeat is a successful high-tech manufacturer of health-focused products for women. Bellabeat is looking to unlock new growth opportunities for the company to become a larger player in the global smart device market. Bellabeat believes that this goal can be achieved by boosting their market share in the smart watch segment of the wearables category.

The purpose of the project is to analyze smart device data to gain insight into how consumers are using their smart devices. These insights will then help guide marketing strategy for Bellabeat's smart wellness watch, Bellabeat Time, which is a hybrid watch featuring an analog watch face with smart technology built in to track user activity, sleep, and stress.

B. Background

Bellabeat makes wearables and associated applications that are geared towards women to help them better understand how their bodies work and make healthier choices by monitoring their biometric and lifestyle data. Bellabeat products are available through the company's own website and other online. The Bellabeat Time watch is an elegant timepiece with a traditional look that syncs with the Bellabeat app on a mobile device to provide details into daily activities without a touchscreen or other addons on the device itself.

In the next nine years, the global market for smart wearables is predicted to grow three times when compared to the current market. The demand for smart watches is a growing trend within this category. In 2020, despite the pandemic, the estimated increase in demand of fitness trackers surged by 20%, with the smartwatch segment accounting for little less than 50% of the revenue share. In the US, it is reported that 57% of females use wearables in their normal daily activities. This indicates that the steady market demand and opportunity for high penetration rates worldwide could be key in propelling sales of the Bellabeat Time.

C. Results and Conclusions

The outcomes from this data analysis study applicable to Bellabeat's customers are the risk of declining user participation over time, the benefits for sleep quality from limiting the non-active time and the opportunity to ensure a more regular level of activity with a tracker to aid with weight loss.

The recommendations, based on these insights, for the marketing strategy for the Time wellness watch include improving sustained use of the device, boosting technological reliability, and promoting personalized guidance.

D. Organization of this Report

Section II of this report provides an overview of Bellabeat's facts and circumstances, including an industry and market analysis and the basis on which this project is founded. Section III details the data checking and cleaning process. Section IV presents data analysis and results gathered from the data set. Section V outlines the findings and conclusions of the data analysis and discusses the strategy for the Marketing Team. Appendix A contains the scripts used during data processing and analysis. Appendix B is the Data Log which details the deletions and outliers in the original data and column metadata in the cleaned data sets. Appendix C includes the Team's presentation to the stakeholders of Bellabeat, an integral part of this project.

II. PROJECT BASIS

A. Business Overview

Bellabeat was founded in 2013 by Urška Sršen, the Chief Creative Officer and Sandro Mur, a key member of the Bellabeat executive team. The company is headquartered in San Francisco with offices in Zagreb, Hong Kong and London. Bellabeat is a data-driven wellness company with a collection of products and services focused on women's health. Bellabeat has raised close to $19M over several rounds of funding and is one of the fastest-growing wellness subscription services for women centered on some of the most fashionably designed tech-powered wellness products. The company recently turned profitable and has had over 100 percent growth from 2018 to 2020.

Bellabeat's product range includes the following products and accessories:

- Bellabeat app: The app connects users to their line of smart wellness products and provides them with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits.
- Leaf: This wellness tracker can be worn as a bracelet, necklace, or clip and tracks activity, sleep, and stress.
- Time: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress.
- Spring: This is a water bottle that tracks the user's daily water intake using smart technology to ensure appropriate hydration levels.
- Bellabeat membership: This is a subscription-based membership program for users which gives 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

Bellabeat sells their products through a diverse mix of distribution channels, both direct and indirect. The company market their products online on their own website as well as through other online retailers. Advertising is carried out primarily through digital platforms like Google and active accounts on Facebook, Instagram and Twitter. The Leaf pendant, one of Bellabeat's most successful products, has sold more than 2.2 million units and is billed as the best overall smart jewelry product and the best smart necklace in Business Insider. The Leaf trackers are so popular that they have spawned online marketplaces for handmade bracelets. Business Insider also expressed a favorable first impression of the Time smart watch. Bellabeat also has the fastest-growing wellness subscription app with around 10 million basic users and 2.5 million paying members as of February 2021.

B. Industry Analysis

The global fitness tracker market size was valued at $36 billion in 2020 and is expected to expand at a annual growth rate of 15% from 2021 to 2028. This growth is fueled by rising popularity and acceptance of fitness trackers to track daily personal information such as physical activities, fitness, sleep and physiological data. The technology to monitor this information wirelessly on a connected device such as a phone or computer at any time has improved substantially making it easier for consumers to integrate these trackers into their daily life. Advancements in sensor and battery/charging technology will likely further augment the growth of this market segment. The ability of these devices to monitor multiple health indicators in real-time without the risk of low battery life will lead to increased adoption in the healthcare industry. Threats to market expansion prospects are from factors such as data privacy concerns and high initial cost.

In terms of distribution channels, the online sales channel has been dominant for fitness trackers in the global market. The convenience and instant service offered by e-commerce platforms and the steady growth in penetration of the internet and smart phones, means that online sales will continue to be the preferred channel of distribution in the market over the coming years.

In February 2021, Bellabeat announced the launch of their new corporate wellness program. The stated goal of this program is to help businesses improve employee engagement and productivity. In 2020, the global corporate wellness market, which was valued at $53 billion in size, was expected to expand at an annual growth rate of 7% over the next six years. Awareness of the benefits of employee health and wellbeing and the influences on corporate productivity, operational costs and revenue is boosting the market demand for corporate health and wellness programs.

C. Company Mission and Strategy

The company's mission is to "empower women to reconnect with themselves, unleash their inner strengths and be what they were meant to be". Bellabeat's aim is to provide women with the tools to live in harmony with themselves by focusing on the following aspects:

- Smart Insights: Use data to help women understand and transform themselves
- Women-centric: Use feminine design and technology adjusted to female bodies
- Holistic Approach: Help women achieve their personal goals by syncing body, mind and cycle
- Body Positivity: Help women discover their own definition of beauty and health.

D. Stakeholders

The main stakeholders for this project are:

- Urška Sršen: Bellabeat's Chief Creative Officer
- Sando Mur: Key member of the Bellabeat executive team
- Bellabeat Marketing Analytics Team: Team of data analysts responsible for collecting, analyzing, and reporting data that helps guide the marketing strategy.

E. Business Task

The primary business task for this project is to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. This insight will then guide the formulation of the marketing strategy for the Time wellness watch.

The Time tracker watch is a stylish timepiece that not only has a familiar watch face but also built-in technology with highly sensitive movement sensors to track activity, sleep, reproductive health and stress resistance. The data from the sensors is wirelessly synced and monitored via the Bellabeat app, which also includes inactivity alerts and silent alarms, and is available on Google Play and Apple App Store. The watch is water resistant and has a 14-day memory that can store the data on the tracker itself. The watch and the sensor circuits are powered by an internal power source, eliminating the need for charging the watch. The battery life is six months and can be easily replaced with a tool, included in the package. Time is available in two finishes including gold and rose stainless steel with a wide range of interchangeable strap accessories to fit wearers' personal style. Besides basic features like time telling, activity tracking and sleep tracking, advanced features include tracking meditation and menstrual cycle tracking, and stress resistance prediction.

To carry out the business task described above effectively, the analysis will address the following topics:

- Trends in smart device usage
- Application of these trends to Bellabeat customers
- Influence of these trends on the Bellabeat marketing strategy for the Time wellness watch

F. Project Data Sources and Characteristics

The dataset for this project is made available by Möbius on Kaggle. This Fitbit Fitness Tracker Dataset was generated by respondents to a distributed survey via Amazon Mechanical Turk over 30 days in 2016.

Personal tracker data, including minute-level output for physical activity, heart rate, and sleep, of 30 eligible Fitbit users was collected during this period.

The data is licensed under CC0: Public Domain which means that the data can be copied, modified, distributed, and used to perform the work, all without asking permission of the author. The dataset is organized in long format in a total of 18 files (.csv format). The data will be stored locally on a password protected computer.

The focus of this project is on the usage and the individual tracking behaviors/preferences of the users which can provide a picture of user adoption and benefits, if any, realized over the short period of monitoring. Therefore, only the following files from the dataset has been selected for analysis:

- "dailyActivity_merged.csv"
- "sleepDay_merged.csv"
- "weightLogInfo_merged.csv"

The following is the ROCCC analysis for this dataset:

- Reliability: This is low since the dataset was collected from 30 users whose gender is not documented. This is especially important since the business task is to identify trends in usage for a product geared towards women and it cannot be determined if the dataset is biased. The data on sleep and weight is incomplete resulting from the use of different types of Fitbit trackers. Though the data files and how they are organized are self-explanatory, there is no metadata provided for the dataset.
- Originality: This is also low as this is third party data collected on Amazon Mechanical Turk and cannot be validated with the original source.
- Comprehensive: This is medium because though the dataset contains multiple fields on daily activity, calories consumed, total steps, sleep durations and weight records, crucial demographic data on participants such as age, gender, fitness and lifestyle, is omitted. Moreover, the sample size is small and may not be large enough to represent the target population.
- Current: This dimension is at a medium level even though data is almost 6 years old. Human behaviors and habits tend to remain uniform and not vary much over a limited period. This premise may have changed over the last couple of years during the pandemic but will average out over time. The duration of the data collection is too small to account for seasonal changes in behavioral patterns over the year.
- Cited: This is high as the data collector and source is well documented and cited.

As the Chief Creative Officer, Urška Sršen, noted this dataset has limitations. An alternative data source is the replication data from a multi-modal data collection study conducted at University of Texas at Austin to measure health, behavior and living environment of study participants. While the datasets are more recent, from November 2018, have good diversity of participants with age and gender included, and are distributed under the Creative Commons Attribution-CC0 License, the duration of the Fitbit data assimilation was only 2 weeks and only around a quarter of the 31 participants submitted data for more than 12 days. Since the comprehensiveness of this data set is not much better than the original Fitbit Fitness Tracker Dataset, the data sets from this source were not used in this project for Bellabeat.

A literature survey yielded a study in the Journal of Medical Internet Research which analyzed data from 26,935 Withings-connected device users (wearable activity trackers and digital scales) to assess the initial physical activity characteristics and their 6-month changes. The data sets used in this study are not available for replication, but an important observation provided in the study is that both a greater physical activity level and higher regularity level are significantly associated with weight loss. The limitations of the study included a study sample primarily composed of men and lack of information on height of users which prevented the calculation of BMI.

III. DATA PROCESSING

The data processing for the data sets was carried out in Microsoft SSMS. The SQL script use to check, clean and process the data is included in the Appendix. Here are some the issues encountered when processing the data:

- Duplicate Data: The SleepRecord table had 3 rows in total that showed up more than once. The duplicate rows were dropped to avoid inaccurate counts.
- Outdated Data: All data was within the date range specified and was consistent across the three data sets.
- Incomplete/Missing Data: The DailyActivity table is missing some records where the total daily steps are reported as zero. It is assumed that the users forgot to wear their Fitbits on those days and hence, these null records have been deleted to prevent inaccurate insights. This table provides a breakdown of the records deleted.
- Incorrect/inaccurate Data: Since there is no metadata available, the units for the distances in the DailyActivity table are unclear. According to the Fitabase Data Dictionary, the units for the distances tracked is kilometers. The conversion of steps to distance is dependent on the stride length and pace of the user and hence it is difficult to substantiate the veracity of this data. Though this data is complete, it should be used with caution and as a comparison between users or over time, rather than an absolute number.
- Inconsistent Data: The date data uses different formats across the tables. To rectify this issue, the date data in the SleepRecord and WeightLog tables were split into separate date and time columns to ensure consistency across the tables. Also, the character lengths of the user IDs were verified and there were no disparities across the tables.
- Outliers: The data sets were reviewed individually for outliers by checking the count, minimum and maximum values for principal fields in each set. The results for these queries can be found in the appendix. In brief, the DailyActivity table had one user with only three records, and six users with days where there were less than 100 steps. In the same table, one user had a low estimated calorie expenditure with low activity minutes indicating that they may have not used the Fitbit all day or had battery issues with an incomplete record for the day. The SleepRecord table has nine users with ten or less records and seven with sleep records less than 90 minutes. The WeightLog had five users out of eight with two or less weight records. Only two users had more than twenty records over the recording period. The outliers in the datasets have not been scrubbed, but necessary care will be taken when using this data in the analysis.

- Structural Errors and Contaminated Data: None of the inaccuracies caused by these flaws were found in the data sets. The data has been logged correctly during measurement, and the data transfer and download were completed fully.

The table below contains some metadata about the data sets and actions taken when cleaning the data.

Table	Rows	Unique IDs	Character Length of IDs	Start Date	End Date	Duplicate Rows		Zero Records	
						No.	Action	No.	Action
DailyActivity	940	33	10	4/12/2016	5/12/2016	0	None	77	Drop
SleepRecord	413	24	10	4/12/2016	5/12/2016	3	Drop	0	None
WeightLog	67	8	10	4/12/2016	5/12/2016	0	None	0	None

The cleaned data sets were saved in a separate folder and the metadata on the fields has been included in the Appendix.

IV. DATA ANALYSIS AND RESULTS

A. Data Summary

The clean data files created in SQL were imported into RStudio and new data frames were created. The RStudio script for the entire data analysis process can be found in the Appendix.

The first step was to generate descriptive statistics for the data frames.

Descriptive Statistics

activity_df
N: 863

	TotalSteps	Sedentary Minutes	Lightly Active Minutes	Fairly Active Minutes	Very Active Minutes
Mean	8319.39	955.75	210.02	14.78	23.02
Min	4	0	0	0	0
Q1	4920	721	146	0	0
Median	8053	1021	208	8	7
Q3	11100	1189	272	21	35
Max	36019	1440	518	143	210
Std.Dev	4744.97	280.29	96.78	20.43	33.65
CV	0.57	0.29	0.46	1.38	1.46

sleep_df
N: 410

	Total Minutes Asleep	Total Time In Bed
Mean	419.17	458.48
Min	58	61
Q1	361	403
Median	432.5	463
Q3	490	526
Max	796	961
Std.Dev	118.64	127.46
CV	0.28	0.28

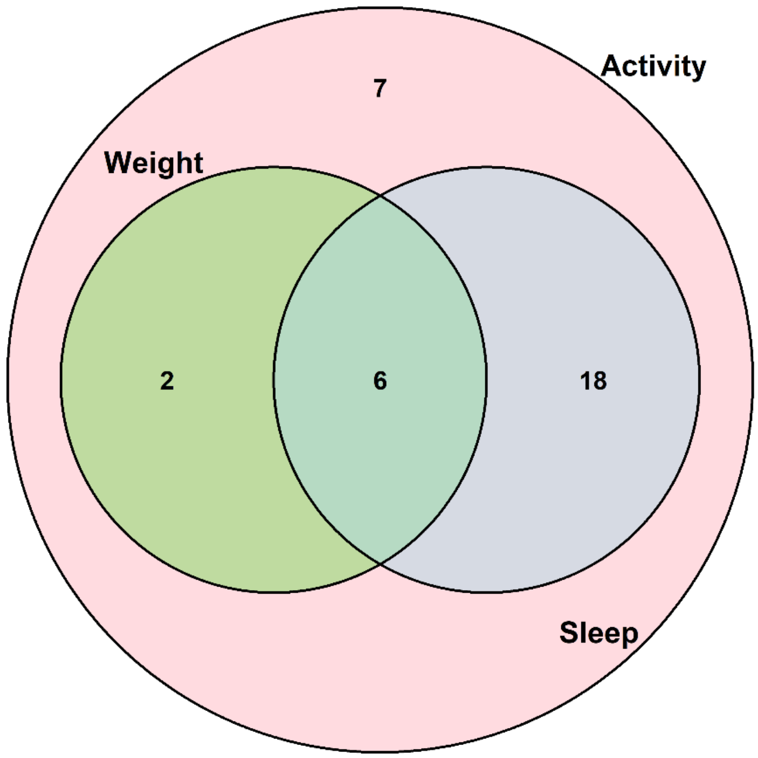
weight_df
N: 67

	Weight Kg	BMI
Mean	72.04	25.19
Min	52.6	21.45
Q1	61.4	23.96
Median	62.5	24.39
Q3	85.1	25.56
Max	133.5	47.54
Std.Dev	13.92	3.07
CV	0.19	0.12

The summary shows the average user in these datasets takes 8319 steps a day, which is lower than the generally recommended 10,000 steps. On average, users have 37.8 minutes a day of fairly or very active minutes. Participants are averaging 458 minutes in bed, during 39 minutes of which they are awake. Average BMI is 25.2, which is considered overweight. However, the credibility of this statistic is low considering that only a small number of participants logged their weight during the study.

B. Smart Device Usage

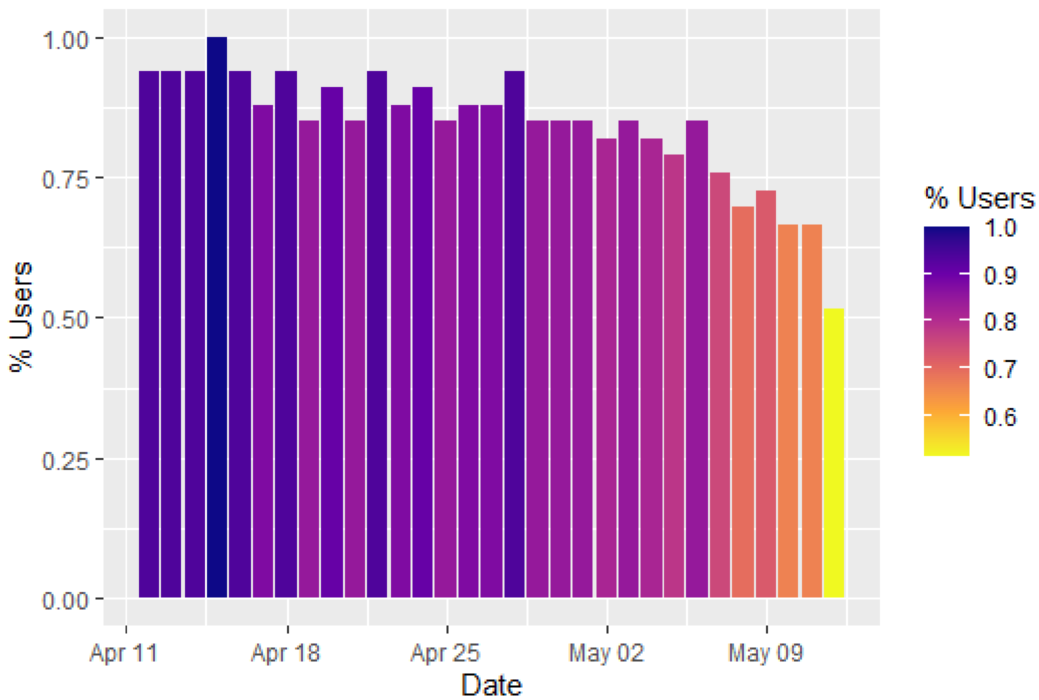
Due to the differences in the features of the Fitbits used by the study participants and their preferences, there is a variation in the data logged by the users. This can be seen in the following Venn diagram that maps the intersection of tracking data reported for the users.



Only 6 participants out of the total of 33 recorded IDs reported all three data types. Around 73% registered both sleep and activity data, and merely 24% charted their weight along with their activity. This implies that the sample size of the data is comparatively small to draw meaningful conclusions, especially when analyzing the sleep and weight data sets.

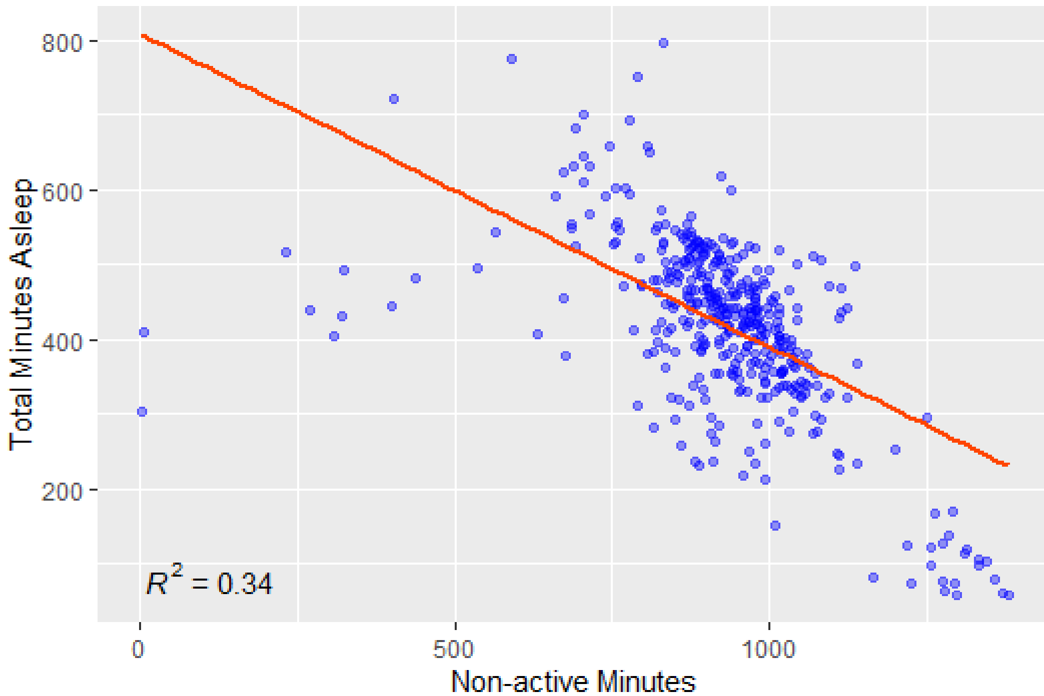
C. Trends in Daily Usage

Most long-term studies have reported a decline in usage of activity trackers over time. Given that the duration of the study was limited to only a month of data, it was expected that there may not be a noticeable trend in daily usage. However, the following graph showing the distribution of participant's daily tracker use seems to show a steady decline in usage over the period of the study and a significant drop towards the end of the study.



As seen in the graph, the usage drops to an average of around 68% in the last week of the data logging period, which is significant even though the duration is not that long. A reason for this decline could be that participants knew that the study was ending and assumed that they have provided enough data already. Even though this decline is exceptionally drastic, it generally agrees with long term study numbers that half of the users stop using their trackers after six months.

Majority of the sleep studies have found little or no relationship between active minutes and sleep duration daily. Then again, some studies have noted a negative association between daily sedentary minutes and sleep duration. This following visualization compares the participants' non-active minutes and the total minutes asleep. Non-active minutes was calculated as the sum of sedentary minutes and lightly active minutes. From this plot, it is evident that the more time that a user was non-active, the lower the total time spent asleep. The coefficient of determination (R^2) is 0.34, indicating a minor correlation between non-activity and sleep, and the relationship is negative, implying that sleep quality is negatively impacted by non-activity.



The relationship between greater mean physical activity and weight loss is well documented. A further relationship that is also observed is that greater weight loss is to a large extent associated with lower variability in physical activity levels. The following table enumerates the data available for Daily Active Minutes (Very Active Minutes + Fairly Active Minutes) and Total Weight Loss for participants with a total reporting period greater than or equal to 15 days.

ID	Total Period in Days	Weight Loss	% Weight Loss	Mean Active Minutes	Standard Deviation Active Minutes
2873212765	21	-0.60	-1.06%	6.00	8.49
4319703577	17	0.10	0.14%	10.50	14.85
4558609924	21	0.60	0.86%	18.60	17.26
6962181067	30	0.60	0.96%	41.27	29.54
8877689391	30	1.80	2.10%	77.88	36.88

While the sample size is particularly small to gain any insights, it can be seen that the user (ID: 8877689391) with the greatest mean active minutes lost the most weight. Another insight is that User ID 4558609924 had almost the same % weight loss as User ID 6962181067 though their Mean Active Minutes are lower than that of the other, but with greater regularity (low standard deviation).

V. CONCLUSIONS AND RECOMMENDATIONS

A. Application of these Trends to Bellabeat customers

The primary trends from this data analysis that are applicable to Bellabeat customers are:

- Declining Participation: There is a steady decline in user participation over time, the main causes of which, according to the study in the Journal of Medical Internet Research, are motivation, lost device, and technological reliability. Reasons for technical issues were mostly due to battery problems and the user experience/ease of use.
- Sleep Quality Benefits: Users can benefit from limiting the time that they are non-active during the day and have enhanced quality of sleep with longer sleep durations. Better sleep leads to users being more energetic the following day resulting in reduced non-active time, which further improves sleep time.
- Weight Loss Aid: Greater physical activity with consistent regularity leads to higher weight loss. Customers have an opportunity to monitor and ensure a more regular level of activity with the assistance of a tracker.

B. Influence of these trends on the Marketing Strategy for the Time Wellness Watch

Based on the trends detailed above, the marketing strategy for the Time wellness watch can take into consideration the following recommendations:

- Improve Sustained Use: Maintaining user motivation is a key strategy to increase the chance of sustained tracker use. Progress visualization provides users with feedback on their activity and ability to observe their improvement, encouraging them to keep moving towards their long-term goals. Social interaction by means of competition and experience sharing is known to improve user interest for the product. Wellness challenges and rewards, like badges and incentives, create recognition for users, motivating them further.
- Boost technological reliability: The Bellabeat app is a crucial component for the success of the Time wellness watch. Ease of use and simplicity are essential features for the app so that the time and effort required of the user to interact with the app is kept to a minimum. One of the ways to achieve this is to limit manual entry of data. This also means the easiness to change devices and monitor data on multiple platforms. The timing and frequency of reminders to enter manual data or notifications to encourage activities need to be easily modified, otherwise users will ignore them. Application of wireless battery charging technology in the Time wellness tracker could be considered, since changing batteries is a known barrier to product reliance.

- Promote personalized guidance: Providing customers with personalized information based on their data is essential for participants to find information that is relevant to them. This may also include basic information on guidance of using associated apps and the right approach to get the best value. The app may offer personalized plans tailored to the users' needs and data and assist them with realistic goal setting. These plans could not only have targets for activities or weight loss, but also guidance for food choices and a positive lifestyle, including sleep health and mental wellness.

VI. REFERENCES

Bellabeat. (2018, December 6). *Bellabeat Introduces Time*. PRN Newswire. Retrieved January 10, 2022, from https://www.prnewswire.com/news-releases/bellabeat-introduces-time-300761038.html

Bellabeat. (2021, September 22). *Bellabeat Time - Wellness Smartwatch*. Retrieved January 10, 2022, from https://bellabeat.com/time/

Bellabeat. (2021b, October 7). *About Us – Bellabeat*. Retrieved January 10, 2022, from https://bellabeat.com/about/

Bellabeat - Crunchbase Company Profile & Funding. (n.d.). Crunchbase. Retrieved January 10, 2022, from https://www.crunchbase.com/organization/bellabeat

el Fatouhi, D., Delrieu, L., Goetzinger, C., Malisoux, L., Affret, A., Campo, D., & Fagherazzi, G. (2021). Associations of Physical Activity Level and Variability With 6-Month Weight Change Among 26,935 Users of Connected Devices: Observational Real-Life Study. *JMIR MHealth and UHealth*, *9*(4). https://doi.org/10.2196/25385

Fitabase. (2018, May 2). *Fitabase Data Dictionary*. https://www.Fitabase.Com/Media/1930/Fitabasedatadictionary102320.Pdf Retrieved January 17, 2022, from https://www.fitabase.com/resources/knowledge-base/exporting-data/data-dictionaries/

Fortune Business Insights. (2021, May). *Fitness Tracker Market Size, Share & COVID-19 Impact Analysis, By Device Type (Smart Watches, Fitness Bands, Smart Glasses, Smart Clothing, and Others), By Application (Heart Rate Tracking, Sleep Measurement, Glucose Measurement, Sports, Running, Cycling Tracking), By Distribution Channel (Online, Retail, and Others) and Regional Forecast, 2021–2028*. Retrieved January 20, 2022, from https://www.fortunebusinessinsights.com/fitness-tracker-market-103358

Gokey, M., & Lord, S. (2021, March 31). *The best smart jewelry in 2021, from necklaces to rings and watches*. Business Insider. Retrieved January 20, 2022, from https://www.businessinsider.com/guides/tech/best-smart-jewelry?international=true&r=US&IR=T#the-best-overall-1

Grand View Research. (2021, March). *Corporate Wellness Market Size & Share Report, 2021–2028*. Retrieved January 20, 2022, from https://www.grandviewresearch.com/industry-analysis/corporate-wellness-market

Hermsen, S., Moons, J., Kerkhof, P., Wiekens, C., & de Groot, M. (2017). Determinants for Sustained Use of an Activity Tracker: Observational Study. *JMIR MHealth and UHealth*, *5*(10), e164. https://doi.org/10.2196/mhealth.7311

Koksal, I. (2020, December 19). *Bellabeat's Ivy Analyzes Wellness Data By Wearables*. Forbes. https://www.forbes.com/sites/ilkerkoksal/2020/12/18/bellabeats-ivy-analyzes-wellness-data-by-wearables/?sh=23160540358b

Liao, Y., Robertson, M. C., Winne, A., Wu, I. H. C., Le, T. A., Balachandran, D. D., & Basen-Engquist, K. M. (2020). Investigating the within-person relationships between activity levels and sleep duration using Fitbit data. *Translational Behavioral Medicine*, *11*(2), 619–624. https://doi.org/10.1093/tbm/ibaa071

Meoli, C. (2021, May 5). *Why Investors Should Start to Care about Femtech and Women's Health*. Stableton | Marketplace for Alternative Investments. Retrieved January 20, 2022, from https://stableton.widelab.co/insights/why-investors-should-start-to-care-about-femtech-and-womens-health/

SoftwareFindr. (2021, June 8). *Wearable Statistics And Facts For 2021*. SaaS Scout (Formerly SoftwareFindr). Retrieved January 10, 2022, from https://saasscout.com/statistics/wearable-statistics/#Global_wearables_statistics_in_2020

Wu, C., Fritz, H., Bastami, S., Maestre, J. P., Thomaz, E., Julien, C., Castelli, D. M., de Barbaro, K., Bearman, S. K., Harari, G. M., Cameron Craddock, R., Kinney, K. A., Gosling, S. D., Schnyer, D. M., & Nagy, Z. (2021). Multi-modal data collection for measuring health, behavior, and living environment of large-scale participant cohorts. *GigaScience*, *10*(6), giab044. https://doi.org/10.1093/gigascience/giab044

APPENDIX A: DATA PROCESSING AND ANALYSIS SCRIPTS

SQL Script

```
#### BELLABEAT STUDY - CHECK, CLEAN AND PROCESS DATA IN MICROSOFT SSMS

## Check DailyActivity Table
# Number of rows in table
SELECT
  COUNT(*) AS row_count
FROM
  DailyActivity
;

# Get Number of unique user IDs and check that all IDs have the same number of characters
SELECT
  DISTINCT Id,
  LEN(Id) AS id_len
FROM
  DailyActivity
;

# Start and End Date of data
SELECT
  MIN(ActivityDate) AS start_date,
  MAX(ActivityDate) AS end_date
FROM
  DailyActivity
;

# Find duplicate rows in table
SELECT
  Id,
  ActivityDate
FROM
  DailyActivity
GROUP BY
  Id,
  ActivityDate
HAVING
  COUNT(Id) > 1
;

# Check for records with 0 in TotalSteps
SELECT
  SUM(num_days_no_steps) AS total_days_no_steps
FROM (
  SELECT
            COUNT(*) AS num_days_no_steps
  FROM
            DailyActivity
  WHERE
            TotalSteps = 0
  ) AS nd
;

# List participants and number of days with 0 in TotalSteps
SELECT
  Id, COUNT(*) AS num_days_no_steps
FROM
  DailyActivity
```

```
WHERE
  TotalSteps = 0
GROUP BY
  Id
ORDER BY
  num_days_no_steps
DESC
;

# Delete rows where TotalSteps = 0; (Users most likely forgot to use Fitbit on those days)
DELETE FROM
  DailyActivity
WHERE
  TotalSteps = 0
;

# Check for outliers in TotalSteps and Calories;
SELECT
  Id,
  COUNT(TotalSteps) AS rec_count,
  MIN(TotalSteps) AS min_steps,
  MAX(TotalSteps) AS max_steps,
  MIN(Calories) AS min_calories,
  MAX(Calories) AS max_calories
FROM DailyActivity
GROUP BY Id
;

# Number of columns in the table
SELECT
  COUNT(*) AS col_count
FROM
  INFORMATION_SCHEMA.COLUMNS
WHERE
  TABLE_NAME ='DailyActivity'
;

# List names of columns and data type in the table
SELECT
  COLUMN_NAME,
  DATA_TYPE
FROM
  INFORMATION_SCHEMA.COLUMNS
WHERE
  TABLE_NAME = 'DailyActivity'
;


## Check SleepRecord Table
# Number of rows in table
SELECT
  COUNT(*) AS row_count
FROM
  SleepRecord
;

# Get Number of unique user IDs and check that all IDs have the same number of characters
SELECT
  DISTINCT Id,
  LEN(Id) AS Id_len
FROM
  SleepRecord
;

```

```
# Start and End Date of data
SELECT
          MIN(SleepDay) AS start_date,
          MAX(SleepDay) AS end_date
FROM
          SleepRecord
;

# Find duplicate rows in table
SELECT
          Id,
          SleepDay
FROM
          SleepRecord
GROUP BY
          Id,
          SleepDay
HAVING
          COUNT(Id) > 1
;

# Create new SleepRecord table with duplicate rows deleted
SELECT
          DISTINCT *
INTO
          dbo.SleepRecordV2
FROM
          SleepRecord
SP_RENAME
          'SleepRecord', 'SleepRecordV1'
DROP TABLE IF EXISTS
          SleepRecordV1
SP_RENAME
          'SleepRecordV2', 'SleepRecord'
;

# Check for records with 0 in TotalSleepRecords
SELECT
          SUM(num_days_no_sleep) AS total_days_no_sleep
FROM
          (
                    SELECT
                              COUNT(*) AS num_days_no_sleep
                    FROM
                              SleepRecord
                    WHERE
                              TotalSleepRecords = 0
          ) AS nd
;

# Split SleepDay column into Date and Time
SELECT
          *,
          CAST(SleepDay AS date) AS SleepDate,
          CAST(SleepDay AS time) AS SleepTime
INTO SleepRecordV2
FROM SleepRecord
ALTER TABLE SleepRecordV2
DROP COLUMN Date
SP_RENAME
          'SleepRecord', 'SleepRecordV1'
DROP TABLE IF EXISTS
          SleepRecordV1
SP_RENAME
```

```
			'SleepRecordV2', 'SleepRecord'
	;

	# Check for outliers in TotalMinutesAsleep and TotalTimeInBed;
	SELECT
			Id,
			COUNT(TotalMinutesAsleep) AS rec_count,
			MIN(TotalMinutesAsleep) AS min_sleep,
			MAX(TotalMinutesAsleep) AS max_sleep,
			MIN(TotalTimeInBed) AS min_bedtime,
			MAX(TotalTimeInBed) AS max_bedtime
	FROM SleepRecord
	GROUP BY Id
	;

	# Number of columns in the table
	SELECT
			COUNT(*) AS col_count
	FROM
			INFORMATION_SCHEMA.COLUMNS
	WHERE
			TABLE_NAME ='SleepRecord'
	;

	# List names of columns and data type in the table
	SELECT
			COLUMN_NAME, DATA_TYPE
	FROM
			INFORMATION_SCHEMA.COLUMNS
	WHERE
			TABLE_NAME = 'SleepRecord'
	;

	## Check WeightLog Table
	# Number of rows in table
	SELECT
			COUNT(*) AS row_count
	FROM
			WeightLog
	;

	# Get Number of unique user IDs and check that all IDs have the same number of characters
	SELECT
			DISTINCT Id,
			LEN(Id) AS Id_len
	FROM
			WeightLog
	;

	# Start and End Date of data
	SELECT
			MIN(Date) AS start_date,
			MAX(Date) AS end_date
	FROM WeightLog
	;

	# Find duplicate rows in table
	SELECT
			Id,
			Date
	FROM
			WeightLog
	GROUP BY
			Id,
```

```
            Date
HAVING
            COUNT(Id) > 1
;

# Check for records with 0 in WeightKg
SELECT
            SUM(num_days_no_weight) AS total_days_no_weight
FROM
            (
                        SELECT
                                    COUNT(*) AS num_days_no_weight
                        FROM
                                    WeightLog
                        WHERE
                                    WeightKg = 0
            ) AS nd
;

# Split Date column into Date and Time
SELECT
            *,
            CAST(Date AS date) AS WeightDay,
            CAST(Date AS time) AS WeightTime
INTO WeightLogV2
FROM WeightLog
ALTER TABLE WeightLogV2
DROP COLUMN Date
SP_RENAME
            'WeightLog', 'WeightLogV1'
DROP TABLE IF EXISTS
            WeightLogV1
SP_RENAME
            'WeightLogV2', 'WeightLog'
;

# Check for outliers in WeightKg and BMI;
SELECT
            Id,
            COUNT(WeightKg) AS rec_count,
            MIN(WeightKg) AS min_weight,
            MAX(WeightKg) AS max_weight,
            MIN(BMI) AS min_BMI,
            MAX(BMI) AS max_BMI
FROM WeightLog
GROUP BY Id

# Number of columns in the table
SELECT
            COUNT(*) AS col_count
FROM
            INFORMATION_SCHEMA.COLUMNS
WHERE
            TABLE_NAME ='WeightLog'
;

# List names of columns and data type in the table
SELECT
            COLUMN_NAME, DATA_TYPE
FROM
            INFORMATION_SCHEMA.COLUMNS
WHERE
            TABLE_NAME = 'WeightLogInfo'
;
```

```

# Inner join DailyActivity and SleepRecord
SELECT da.*, sr.TotalMinutesAsleep, sr.TotalTimeInBed
INTO Acivity_Sleep
FROM DailyActivity AS da
JOIN SleepRecord AS sr
ON da.ActivityDate = sr.SleepDay AND da.Id = sr.Id
ORDER BY da.Id, da.ActivityDate
;

# Inner join DailyActivity and WeightLog
SELECT da.*, wl.WeightKg, wl.BMI, wl.Fat
INTO Acivity_Weight
FROM DailyActivity AS da
JOIN WeightLog AS wl
ON da.ActivityDate = wl.Date AND da.Id = wl.Id
ORDER BY da.Id, da.ActivityDate
;
```

R Script

```
 ## Set up environment
library(tidyverse)
library(dplyr)
library(summarytools)
library(VennDiagram)
library(ggplot2)
library(ggpmisc)

## Import CSV and view data frame

activity_df <-read_csv("DailyActivity.csv")
View(activity_df)
sleep_df <-read_csv("SleepRecord.csv")
View(sleep_df)
weight_df <-read_csv("WeightLog.csv")
View(weight_df)
act_slp_df <-read_csv("Activity_Sleep.csv")
View(act_slp_df)
act_wgt_df <-read_csv("Activity_Weight.csv")
View(act_wgt_df)

## Get key statistics about data frames
act_stats <- activity_df %>%
  select(TotalSteps, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryActiveMinutes) %>%
  descr(., stats = c("mean", "min", "Q1", "med", "Q3", "max", "sd", "CV"))
view(act_stats, file = "~/act_stats.html")

slp_stats <- sleep_df %>%
  select(TotalMinutesAsleep, TotalTimeInBed) %>%
  descr(., stats = c("mean", "min", "Q1", "med", "Q3", "max", "sd", "CV"))
view(slp_stats, file = "~/slp_stats.html")

wgt_stats <- weight_df %>%
  select(WeightKg, BMI) %>%
  descr(., stats = c("mean", "min", "Q1", "med", "Q3", "max", "sd", "CV"))
```

```
view(wgt_stats, file = "~/wgt_stats.html")


## Plot Venn diagram of users
act_ids <- unique(activity_df$Id)
slp_ids <- unique(sleep_df$Id)
wgt_ids <- unique(weight_df$Id)

venn.diagram(
  x = list(act_ids, slp_ids, wgt_ids),
  category.names = c("Activity", "Sleep","Weight"),
  filename = "Usage_venn.png",
  output=TRUE,
  fill = c("light pink", "light blue", "green"),
  cex = 1.2,
  fontfamily = "sans",
  fontface = "bold",
  cat.cex = 1.4,
  cat.fontfamily = "sans",
  cat.fontface = "bold",
  cat.default.pos = "outer",
  cat.pos = c(-30, 40, 155))

## Get number of users that used their devices each day and
## Plot Bar Chart of activity tracker use
num_users <- activity_df %>%
  group_by(ActivityDate) %>%
  summarise(users_perday = sum(n())) %>%
  mutate(percent_users = users_perday/max(users_perday))

ggplot(data = num_users) +
  geom_col(mapping = aes(x = ActivityDate, y = percent_users, fill = percent_users )) +
  scale_fill_viridis_c(option = "plasma",direction = -1) +
  labs(x = "Date", y = "% Users",fill = "% Users")

## Get active and non-active minutes for users and
## Plot relationship between activity and sleep
activity_sleep <- act_slp_df %>%
  mutate(non_active_mins = (LightlyActiveMinutes+SedentaryMinutes) )

ggplot(data = activity_sleep, mapping = aes(x = non_active_mins, y = TotalMinutesAsleep)) +
  geom_point(color = 'blue', alpha = 0.4) +
  geom_smooth(method = 'lm', se=FALSE, color = 'orangered', formula = y ~ x) +
  stat_poly_eq(formula = y ~ x, aes(label = paste(..rr.label..)), label.y = "bottom", label.x = "left") +
  labs(x = "Non-active Minutes", y = "Total Minutes Asleep")

## Get active and non-active minutes for users and
## Plot relationship between activity and weight
weight_loss_df <- act_wgt_df %>%
  group_by(Id) %>%
  mutate (active_mins = (VeryActiveMinutes+FairlyActiveMinutes)) %>%
  summarize(period = (last(ActivityDate) - first(ActivityDate)),
            weight_loss = (first(WeightKg) - last(WeightKg)),
            percent_wl = (weight_loss/first(WeightKg))*100,
            mean_act = mean(active_mins),
            sd_act = sd(active_mins)) %>%
  filter(period >= 15)
View(weight_loss_df)
```

APPENDIX B: DATA LOG

Deletions

This table provides a breakdown of the records deleted from the DailyActivity table.

ID	No of days with zero steps
1927972279	14
4020332650	14
1844505072	10
8792009665	10
6775888955	9
6117666160	5
6290855005	5
7007744171	2
5577150313	2
7086361926	1
8253242879	1
8583815059	1
1503960366	1
4057192912	1
4702921684	1

Outliers

For DailyActivity table:

ID	Count	Minimum Steps	Maximum Steps	Minimum Calories	Maximum Calories
1644430081	30	1223	18213	1276	3846
7086361926	30	31	14560	1199	2997
4702921684	30	1664	15126	1240	3691
3977333714	30	746	16520	52	1760
6290855005	24	4562	9837	2175	3327
3372868164	20	3077	9715	1237	2124
8877689391	31	4790	29326	1849	4547
2026352035	31	254	12357	1141	1926
5577150313	28	3421	15764	1665	4552
4319703577	31	17	13658	257	2530
4020332650	17	16	11728	1120	3879
4057192912	3	3984	5974	1527	2306
8053475328	31	1170	22988	1505	3589
4445114986	31	768	9105	1212	2499
6962181067	31	1551	20031	928	2571
8792009665	19	144	8360	1720	3101
2022484408	31	3292	18387	1848	3158
8378563200	31	2132	16208	1976	4236
1927972279	17	149	3790	2093	2638
2320127002	31	772	10725	1125	2124
5553957443	31	655	17022	741	2335
7007744171	24	3761	20067	2051	3180
1503960366	30	9705	18134	1728	2159
8583815059	30	3008	15168	2439	3513
4558609924	31	3428	13743	1452	2666
2873212765	31	2524	9685	1431	2241
4388161847	31	3369	22770	1623	4022
6775888955	17	9	10771	1032	3727
1624580081	31	1510	36019	1002	2690
1844505072	21	4	8054	1348	2130
6117666160	23	2997	19542	1248	4900
2347167796	18	42	22244	403	2670
8253242879	18	2672	11268	1580	2218

For SleepRecord table:

ID	Count	Minimum Sleep	Maximum Sleep	Minimum Bedtime	Maximum Bedtime
1503960366	25	245	700	264	712
1644430081	4	119	796	127	961
1844505072	3	590	722	961	961
1927972279	5	166	750	178	775
2026352035	28	357	573	380	607
2320127002	1	61	61	69	69
2347167796	15	374	556	386	602
3977333714	28	152	424	305	626
4020332650	8	77	501	77	541
4319703577	26	59	692	65	722
4388161847	23	62	619	65	641
4445114986	28	98	502	107	542
4558609924	5	103	171	121	179
4702921684	27	253	591	257	612
5553957443	31	322	775	353	843
5577150313	26	74	603	78	634
6117666160	18	336	658	350	698
6775888955	3	235	423	260	441
6962181067	31	298	630	334	679
7007744171	2	58	79	61	82
7086361926	24	322	681	333	704
8053475328	3	74	486	75	493
8378563200	31	323	611	355	689
8792009665	15	339	531	360	552

For WeightLog table:

ID	Count	Minimum Weight	Maximum Weight	Minimum BMI	Maximum BMI
1503960366	2	52.6	52.6	22.65	22.65
1927972279	1	133.5	133.5	47.54	47.54
2873212765	2	56.7	57.3	21.45	21.69
4319703577	2	72.3	72.4	27.38	27.45
4558609924	5	69.1	70.3	27	27.46
5577150313	1	90.7	90.7	28	28
6962181067	30	61	62.5	23.82	24.39
8877689391	24	84	85.8	25.14	25.68

<u>Columns</u>

For cleaned DailyActivity table:

COLUMN_NAME	DATA_TYPE
Id	bigint
ActivityDate	date
TotalSteps	int
TotalDistance	float
TrackerDistance	float
LoggedActivitiesDistance	tinyint
VeryActiveDistance	float
ModeratelyActiveDistance	float
LightActiveDistance	float
SedentaryActiveDistance	float
VeryActiveMinutes	tinyint
FairlyActiveMinutes	tinyint
LightlyActiveMinutes	smallint
SedentaryMinutes	smallint
Calories	smallint

For cleaned SleepRecord table:

COLUMN_NAME	DATA_TYPE
Id	bigint
SleepDate	date
SleepTime	time
TotalSleepRecords	tinyint
TotalMinutesAsleep	smallint
TotalTimeInBed	smallint

For cleaned WeightLog table:

COLUMN_NAME	DATA_TYPE
Id	bigint
WeightDate	date
WeightTime	time
WeightKg	float
WeightPounds	float
Fat	tinyint
BMI	float
IsManualReport	bit
LogId	bigint

APPENDIX C: POWERPOINT PRESENTATION

Data Analysis

Smart Device Usage

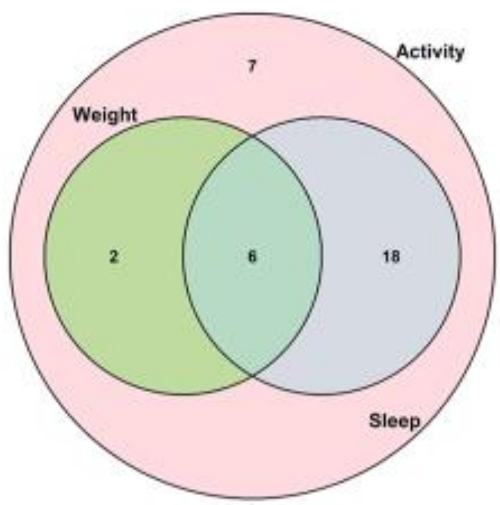
Trends in Daily Usage

Smart Device Usage

18% reported all three data types

73% registered both sleep and activity

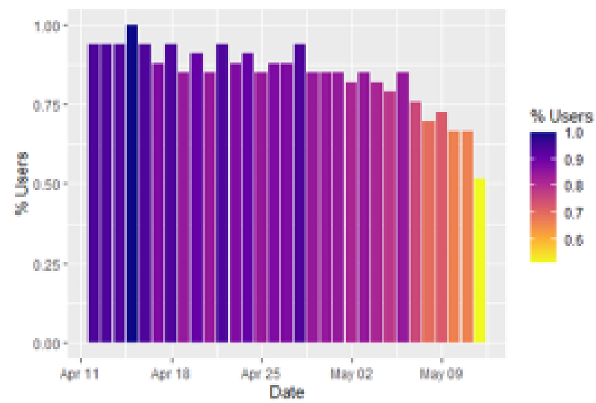
24% charted their weight along with activity



Decline in usage over time

Usage drops to an average of around 68% in last week of reporting period

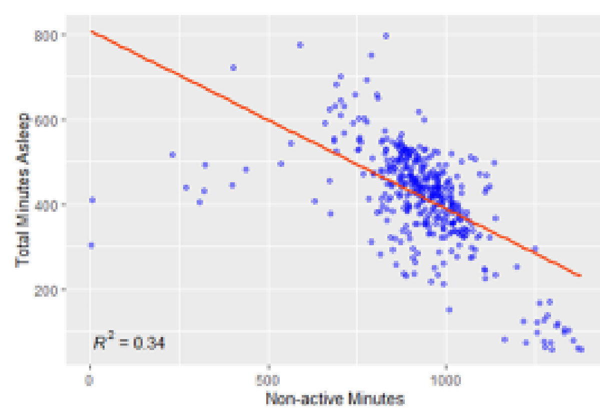
Studies note that half of the users stop using their trackers after 6 months



Sedentary minutes and sleep

Minor correlation between non-activity and sleep

Negative relationship implying that sleep quality is negatively impacted by non-activity



Active minutes and weight loss

Sample size is too small to gain any insight

Greater physical activity with consistent regularity leads to higher weight loss

ID	Total Period in Days	Weight Loss	% Weight Loss	Mean Active Minutes	Standard Deviation Active Minutes
2873212765	21	-0.60	-1.06%	6.00	8.49
4319703577	17	0.10	0.14%	10.50	14.85
4558609924	21	0.60	0.86%	18.60	17.26
6962181067	30	0.60	0.96%	41.27	29.54
8877689391	30	1.80	2.10%	77.88	36.88

Recommendations

Improve sustained use by maintaining user motivation, providing social interaction and setting up wellness challenges, rewards, and recognition for users.

Promote technological reliability by ensuring ease of use and simplicity in the app, allowing users to modify timing and frequency of reminders and considering wireless battery charging technology.

Promote personalized guidance by providing customers with personalized information based on their data, offer personalized plans tailored to the users' needs and assist them with realistic goal setting.