How Can Bellabeat Improve Their Marketing for Their Fitness Products?

Submitted by: Sebastián Capeáns
Course: Google Data Analytics Certificate
Date: May 20, 2025

Summary: I assumed the role of a junior analyst in the marketing 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.

I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights discovered will then help guide marketing strategy for the company. I will present my analysis to the Bellabeat executive team along with my high-level recommendations for Bellabeat's marketing strategy.

Tools used: Excel, SQL, Tableau

ASK PHASE

Business Task

The goal is to identify trends and patterns in the usage of non-Bellabeat health smart devices in order to use these findings to create data-driven recommendations for Bellabeat's Leaf product marketing strategy.

Stakeholders
○ Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
○ Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

PREPARE PHASE

The dataset being utilized is a Kaggle dataset titled "FitBit Fitness Tracker Data." This dataset is in the public domain, and it contains personal fitness tracker from thirty FitBit users. Thirty eligible FitBit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.
This dataset was made available by Kaggle user Möbius, a Generative AI & LLM-empowered Systems Tech Lead at Healthcare. It contains 11 CSV files stored in a folder for the date range 3.12.16-4.11.16 and 18 CSV files for the date range 4.12.16-5.12.16. This means we have two months of data recordings.

This dataset is original and comprehensive, and it appears to be reliable and cited, since these datasets are said to be generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. However, it has its limitations. Firstly, it only contains data on 30 users, which is a very small sample size. We also don't know how the sample was selected, and if there was any bias. In addition, the data is not current, since it's from 2016.

With these limitations in mind, the dataset can help us get a general picture on how people use their FitBit devices, and their recorded activity, heart rate, weight and sleep levels. The dataset is in long format, so it would be wise to aggregate and average recordings for individual users into a single row.

PROCESS PHASE

I started by importing the datasets into Excel, since most of them are relatively small, to get a feel for the data I'll be working with. I will be using 7 of the CSV files for my analysis, the Daily Activity, Heartbeat Seconds, and Weight Log Info files from the first period (3.12.16-4.11.16), and the Daily Activity, Sleep Day, Heartbeat Seconds and Weight Log Info files for the second period (4.12.16-5.12.16). I didn't include the Sleep Day data for the first period because it wasn't included in the files. I chose those because the Daily Activity, Sleep Day and Weight Log Info files are basically an aggregation of all the other files into condensed tables, except for the heartbeat data, which I will be using separately. With these tables I can do a pretty thorough analysis.

All of the datasets, except for the Heartbeat Seconds, are relatively small, so I can start by cleaning them in Excel. I added all datasets, except for the Heartbeat Seconds, as tables in individual sheets in the same workbook.

I started with some basic cleaning, including:

- Setting all dates to the appropriate format (MM-DD-YY). I also removed the time info on the Sleep Day table, since I found it redundant, having all logs starting at 12:00 AM.
- Removing duplicates (3 were found in the Sleep Day table)
- Checking to see if there were values in the ID columns with different character lengths, using the LEN() function.

Then I did some basic functions to find some general insights about the data:

- =MIN(B:B) to confirm the START DATE of the study was correct
- =MAX(B:B) to confirm the END DATE of the study was correct
- =COUNT(UNIQUE(A:A)) to find the number of participants for each dataset, by counting how many times unique ids appear on the table

I found that there were no errors in the dates, since the start and end dates were as established. I also found that, for month 1, 35 users submitted information on their daily activity and 11 users submitted information on their weight logs. For month 2, 31 participants submitted information on their daily activity, 24 submitted information on their sleep days, and 8 submitted weight log information. This contradicts what was established from the data, where it said that the study used 30 subjects.

I finished by using conditional formatting on the tables to find any blank cells, which I did in the Weight Log datasets. There were a lot of blanks in the column "Fat." This may indicate that most users didn't have a scale that showed fat mass as a separate value.

I exported my cleaned spreadsheets as CSVs, which I later uploaded into BigQuery, since I will be analyzing the data using SQL and Tableau.

ANALYZE PHASE

I started the analysis by checking the schemas and previews of the created tables in BigQuery, to see if there were any problems with data types or formatting. After confirming that everything was correct, I proceeded with my analysis. The tables created in BigQuery are the following:

- daily_activity_month_1
- daily_activity_month_2
- heartrate_month_1
- heartrate_month_2
- sleep_day_month_2
- weight_log_month_1
- weight_log_month_2

I wanted to have a general view of the data to figure out the possible next steps in my analysis. I aggregated the data in my tables, grouping by ID, and selecting the main columns to have a general idea of the data trends.

Daily activity for the whole period, grouped by subject id.

```
SELECT Id, COUNT(ActivityDate) AS activity_logs, AVG(TotalSteps) AS avg_steps,
AVG(TotalDistance) AS avg_distance, AVG(Calories) AS avg_calories
FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
GROUP BY Id
```

Id	activity_logs	avg_steps	avg_distance	avg_calories
1624580081	50	5167.2	3.471000002	1433.78
1644430081	40	7780.925	5.658750016	2837.575
2022484408	43	11595.09302	8.276744172	2500.302326
2347167796	33	9647.121212	6.426363642	2033.393939
3977333714	42	10321.52381	7.028333279	1480.642857
4319703577	43	7422.813953	4.993720922	2025.55814
4388161847	39	8595.692308	6.67153853	2829.538462
4702921684	46	8367.065217	6.789130421	2918.565217
5577150313	41	8385.926829	6.276341433	3343.707317

6775888955	35	3301.228571	2.371999974	2284.257143
6962181067	45	10679.88889	7.227333355	2015.377778
7007744171	38	11619.28947	8.281052621	2570.236842
7086361926	43	8459.813953	5.747906959	2457.697674
8253242879	31	4898.064516	3.510645178	1662.193548
8583815059	39	6346.615385	4.951282015	2662.128205
8792009665	41	2217	1.419024387	1994.902439
1844505072	43	2876.023256	1.901627916	1585.325581
1927972279	43	1269.069767	0.8790697649	2195.465116
2026352035	43	4960.139535	3.077674417	1488.976744
2320127002	43	4276.372093	2.890465086	1670.55814
2873212765	43	7299.255814	4.925116265	1855.232558
3372868164	30	6616.933333	4.543333356	1908.833333
4020332650	63	4049.761905	2.904920635	2736.063492
4057192912	36	2103.972222	1.552777781	1911.333333
4445114986	46	4632.369565	3.135217388	2160.630435
4558609924	43	7154.930233	4.729767446	1976.581395
5553957443	43	8540.627907	5.589069798	1855.255814
6117666160	38	7363	5.575000035	2218.552632
6290855005	39	4615.846154	3.49051283	2488.333333
8053475328	42	14784.52381	11.50166659	2932.02381
8378563200	43	8555.162791	6.784651207	3414.139535
8877689391	43	16424.32558	13.45790688	3428.883721
1503960366	50	11935.78	7.73279999	1808.74
2891001357	8	773.625	0.6037500054	2273.375
6391747486	9	1336.888889	1.074444466	1763.111111

Average heartrate and number of logs per subject

```
SELECT
  Id, AVG(value) AS avg_heartrate, COUNT(*) AS num_of_logs
FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_2`
)
GROUP BY Id
ORDER BY avg_heartrate DESC
```

Id	avg_heartrate	num_of_logs
6775888955	94.98125856	67871
7007744171	90.73395739	198378
2026352035	89.58006145	2929
8877689391	84.69282638	313956
6391747486	84.11609287	3747
6117666160	83.68530567	212565
4020332650	82.10177161	569255
4558609924	81.11138899	261507
2022484408	80.62621624	210587
6962181067	78.67732107	392201
2347167796	76.43388534	273487
8792009665	73.88197151	192928
5553957443	68.85078661	352971
5577150313	68.37521899	336209
4388161847	66.13299806	249748

Number of sleep logs, average minutes slept, and average time in bed for all subjects

```
SELECT
  Id, COUNT(TotalSleepRecords) AS num_of_records, ROUND(AVG(TotalMinutesAsleep),1) AS
avg_minutes_slept, ROUND(AVG(TotalTimeInBed),1) AS avg_time_in_bed
FROM
  `bellabeat-case-study-457914.FitBit_Dataset.sleep_month_2`
WHERE Id IS NOT NULL
GROUP BY
Id
ORDER BY
avg_minutes_slept DESC
```

Id	num_of_records	avg_minutes_slept	avg_time_in_bed
1844505072	3	652	961
2026352035	28	506.2	537.6
6117666160	18	478.8	510.2
4319703577	26	476.7	502
5553957443	31	463.5	505.9
7086361926	24	453.1	466.4
6962181067	31	448	466.1
2347167796	15	446.8	491.3
8378563200	31	445.1	485.9
8792009665	15	435.7	453.8
5577150313	26	432	460.6
4702921684	27	417.5	438.2
1927972279	5	417	437.8
4388161847	23	400.2	423.2
4445114986	28	385.2	416.8
1503960366	25	360.3	383.2
6775888955	3	349.7	369
4020332650	8	349.4	379.8
8053475328	3	297	301.7
1644430081	4	294	346

3977333714	28	293.6	461.1
4558609924	5	127.6	140
7007744171	2	68.5	71.5
2320127002	1	61	69

Number of weight logs, average weights, average fat and BMI for all subjects

```
SELECT
  Id, COUNT(DISTINCT LogID) AS num_of_logs, ROUND(AVG(WeightKg),1) AS avg_weight_kg,
AVG(Fat) AS avg_fat, ROUND(AVG(BMI),1) AS avg_bmi
FROM  (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_2`
)
GROUP BY Id
ORDER BY num_of_logs DESC
```

Id	num_of_logs	avg_weight_kg	avg_fat	avg_bmi
6962181067	43	61.7		24.1
8877689391	32	85.3		25.5
4558609924	6	69.6		27.2
2873212765	4	57		21.6
1503960366	3	52.8	22	22.8
4319703577	2	72.4	25	27.4
1927972279	2	131.6		46.9
2347167796	1	63.4	10	24.8
8253242879	1	75.6		29.5
2891001357	1	88.4		25
4445114986	1	92.4		35
4702921684	1	99.7		26.1
5577150313	1	90.7		28

Usage of devices analysis

After having a feel for the data and average values, I can start doing some analysis. First, I want to explore how consistent are users when it comes to using their smart devices. I will be exploring how often, in the 60-day period, did users wear their devices. Ideally, users should wear their devices every day, to really understand their health metrics. I will divide users into four categories:

- Minimally active users: 0-15 logs in the 2-month period
- Occasionally active users: 16-30 logs in the 2-month period
- Fairly active users: 31-45 logs in the 2-month period
- Highly active users: 46-60 (or more) logs in the 2-month period

```sql
SELECT
  Id,
  COUNT(Id) AS num_of_logs,
  CASE
    WHEN COUNT(Id) BETWEEN 0 AND 15 THEN 'Minimally active'
    WHEN COUNT(Id) BETWEEN 16 AND 30 THEN 'Occasionally active'
    WHEN COUNT(Id) BETWEEN 31 AND 45 THEN 'Fairly active'
    ELSE 'Highly active'
  END AS activity_level
FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
GROUP BY Id
ORDER BY num_of_logs DESC;
```

Id	num_of_logs	activity_level
4020332650	63	Highly active
1624580081	50	Highly active
1503960366	50	Highly active
4445114986	46	Highly active
4702921684	46	Highly active
6962181067	45	Fairly active
1927972279	43	Fairly active
2320127002	43	Fairly active
8378563200	43	Fairly active

5553957443	43	Fairly active
2026352035	43	Fairly active
7086361926	43	Fairly active
4558609924	43	Fairly active
4319703577	43	Fairly active
8877689391	43	Fairly active
1844505072	43	Fairly active
2873212765	43	Fairly active
2022484408	43	Fairly active
3977333714	42	Fairly active
8053475328	42	Fairly active
5577150313	41	Fairly active
8792009665	41	Fairly active
1644430081	40	Fairly active
8583815059	39	Fairly active
4388161847	39	Fairly active
6290855005	39	Fairly active
7007744171	38	Fairly active
6117666160	38	Fairly active
4057192912	36	Fairly active
6775888955	35	Fairly active
2347167796	33	Fairly active
8253242879	31	Fairly active
3372868164	30	Occasionally active
6391747486	9	Minimally active
2891001357	8	Minimally active

From the results above, we see that there are:

- 5 highly active users
- 27 fairly active users
- 1 occasionally active user

- 2 minimally active users

Most of the users fall within the “Fairly Active” category, meaning that users wear their devices most days. However, it would be better if more users were “Highly Active,” since only 14.29% were so.

Physical activity data analysis

I then wanted to know how physically active users were, on average. I wrote the following query:

```
SELECT
  ROUND(AVG(TotalSteps),1) AS avg_steps,
  ROUND(AVG(TotalDistance),2) AS avg_distance,
  ROUND(AVG(SedentaryMinutes),1) AS avg_sedentary,
  ROUND(AVG(LightlyActiveMinutes),1) AS avg_lightly_active,
  ROUND(AVG(FairlyActiveMinutes),1) AS avg_fairly_active,
  ROUND(AVG(VeryActiveMinutes),1) AS avg_very_active,
  ROUND(AVG(Calories),1) AS avg_calories,

FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
```

avg_steps	avg_distance	avg_sedentary	avg_lightly_active	avg_fairly_active	avg_very_active	avg_calories
7280.9	5.22	992.5	185.4	13.4	19.7	2266.3

Most of the time for logged activities is sedentary, on average, which makes sense, since most of the day for most people is spent sleeping and working (mostly sedentary jobs). However, I wanted to know what proportion of subjects followed the CDC guidelines weekly physical activity.

Based on data retrieved from the CDC, https://www.cdc.gov/physical-activity-basics/guidelines/adults.html, "*…adults need 150 minutes of moderate-intensity physical activity a week. This can also be 75 minutes of vigorous-intensity or an equivalent combination of moderate- and vigorous-intensity physical activity. In addition, adults need at least 2 days of muscle-strengthening activity each week.*"

I don't have access to muscle-strengthening activity data from my source, but I could see if users reached the recommended amount of weekly physical activity. 150 minutes weekly of moderate-intensity physical activity is roughly equivalent to 1,200 minutes of the same activity for the 2-month (8 weeks) period. This can also be 75 minutes of vigorous-intensity physical activity, or 600 minutes for the whole 8 weeks period. In addition, the CDC stated that it can be an equivalent combination of both. Since the CDC recommendation for vigorous-intensity

physical activity is half of the recommendation for moderate-intensity physical activity, I'll use the following formula:

$0.5 * (moderate\ activity\ minutes) + vigorous\ activity\ minutes > 600$ minutes

At least 1,200 minutes of moderate-intensity physical activity
At least 600 minutes of vigorous-intensity physical activity
At least 600 minutes of a combination of moderate-intensity physical activity AND vigorous-intensity physical activity

In this case, I'll assign the moderate activity minutes to the column "Fairly Active Minutes" in our dataset, and the vigorous activity minutes to the column "Very Active Minutes." This results could be a bit misleading, since I believe the CDC is considering slight rest breaks during their recommended minutes of activity, which I believe the FitBit in our dataset isn't, it's just totaling the total minutes where the heart rate is elevated over a certain threshold or were the walking or running speed is over a certain threshold. However, this can give us an idea to how close the subjects were to meeting the CDC guidelines.

```
SELECT
  Id,
  SUM(VeryActiveMinutes) AS VigorousActivityMinutes,
  SUM(FairlyActiveMinutes) AS ModerateActivityMinutes,
  CASE WHEN SUM(VeryActiveMinutes) > 600 THEN 'Meets CDC guidelines'
  WHEN SUM(FairlyActiveMinutes) > 1200 THEN 'Meets CDC guidelines'
  WHEN SUM(VeryActiveMinutes) + ((0.5) * SUM(FairlyActiveMinutes)) > 600 THEN 'Meets
CDC guidelines'
  ELSE 'Doesnt Meet CDC guidelines' END AS cdc
FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
GROUP BY Id
ORDER BY cdc
```

Id	VigorousActivityMinutes	ModerateActivityMinutes	cdc
6290855005	158	217	Doesnt Meet CDC guidelines
1624580081	283	191	Doesnt Meet CDC guidelines
1927972279	41	44	Doesnt Meet CDC guidelines
4020332650	291	411	Doesnt Meet CDC guidelines
4445114986	281	68	Doesnt Meet CDC guidelines
6391747486	46	6	Doesnt Meet CDC guidelines

2026352035	3	8	Doesnt Meet CDC guidelines
2320127002	53	93	Doesnt Meet CDC guidelines
6117666160	49	98	Doesnt Meet CDC guidelines
8792009665	46	182	Doesnt Meet CDC guidelines
4319703577	190	601	Doesnt Meet CDC guidelines
2891001357	0	660	Doesnt Meet CDC guidelines
3372868164	301	117	Doesnt Meet CDC guidelines
4057192912	43	175	Doesnt Meet CDC guidelines
1844505072	13	49	Doesnt Meet CDC guidelines
5577150313	3520	1205	Meets CDC guidelines
3977333714	713	2213	Meets CDC guidelines
6775888955	465	828	Meets CDC guidelines
8378563200	2484	442	Meets CDC guidelines
8053475328	3573	422	Meets CDC guidelines
6962181067	1202	980	Meets CDC guidelines
2873212765	498	262	Meets CDC guidelines
5553957443	1009	603	Meets CDC guidelines
4558609924	374	458	Meets CDC guidelines
2022484408	1606	870	Meets CDC guidelines
7007744171	1343	618	Meets CDC guidelines
1503960366	1881	895	Meets CDC guidelines
7086361926	1635	998	Meets CDC guidelines
2347167796	420	717	Meets CDC guidelines
1644430081	435	1076	Meets CDC guidelines
4702921684	203	1066	Meets CDC guidelines
8877689391	2850	489	Meets CDC guidelines
4388161847	718	631	Meets CDC guidelines
8253242879	460	328	Meets CDC guidelines
8583815059	305	703	Meets CDC guidelines

From the results we observe that 20 of the subjects met CDC guidelines for weekly activity, while 15 didn't. This is alarming, since only slightly more than half of all subjects met physical activity requirements.

Heart rate data analysis

Now I'm going to take a look at the heartrate data. There's not many variables in this dataset, but I want to know if the subjects' resting heart rates fall within the healthy range, which is from 60 to 100 beats per minute, according to Harvard Medical School, .

The main problem is that in the heartrate datasets there are millions of rows, since the FitBit devices seem to take heartrate information every 5 seconds, in many cases. I will create an approximate resting heartrate by finding the average minimum heartrate for every day they wore their devices. I wrote the following query:

```
WITH daily_min_hr AS (
  SELECT
    Id,
    DATE(Time) AS day,
    MIN(Value) AS min_hr
  FROM (
    SELECT *
    FROM
    `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_1`
    UNION ALL
    SELECT *
    FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_2`

  )
  GROUP BY
    Id, day
),
avg_resting_hr AS (
  SELECT
    Id,
    ROUND(AVG(min_hr), 1) AS approx_resting_hr
  FROM
    daily_min_hr
  GROUP BY
    Id
)
SELECT
  Id,
  approx_resting_hr,
  CASE
    WHEN approx_resting_hr BETWEEN 60 AND 100 THEN 'Healthy Resting HR'
    ELSE 'Outside Healthy Range'
  END AS status
FROM
  avg_resting_hr
```

```
ORDER BY
  approx_resting_hr;
```

Id	approx_resting_hr	status
5577150313	42.1	Outside Healthy Range
4388161847	44.4	Outside Healthy Range
8792009665	49.4	Outside Healthy Range
8877689391	50.5	Outside Healthy Range
5553957443	50.6	Outside Healthy Range
2022484408	51.8	Outside Healthy Range
2347167796	53.8	Outside Healthy Range
6962181067	54.3	Outside Healthy Range
6117666160	54.9	Outside Healthy Range
4558609924	55.1	Outside Healthy Range
4020332650	55.4	Outside Healthy Range
6391747486	60.7	Healthy Resting HR
7007744171	62.8	Healthy Resting HR
6775888955	65.5	Healthy Resting HR
2026352035	66	Healthy Resting HR

According to the query results, 4 of the subjects had healthy approximate resting heart rates, while 11 of the subjects had unhealthy approximate resting heart rates. This result, however, should be taken with a grain of sand, because those that fall below the “healthy range,” have values ranging from 40 to 60, which is a normal value during sleep and for athletic and physically active individuals. The main purpose of this query was to find any minimum values that were **over** the “healthy range,” which is a more significant result.

Sleep data analysis

Next, I'll be checking out the sleep dataset. Here there is only data for 1 month, but it should give me meaningful insights. I wanted to find out if subjects were getting enough sleep, on average. According to Harvard Medical School, https://www.health.harvard.edu/blog/how-much-sleep-do-you-actually-need-202310302986, most healthy adults should aim for at least seven hours of sleep every night, even though some individuals might require less or more. I wrote the following query:

```
SELECT
  Id,
  ROUND(AVG(TotalMinutesAsleep),1) AS avg_minutes_asleep,
  CASE WHEN ROUND(AVG(TotalMinutesAsleep),1) > 420 THEN 'Enough sleep'
  WHEN ROUND(AVG(TotalMinutesAsleep),1) < 420 THEN 'Not enough sleep'
  END AS outcome

FROM
  `bellabeat-case-study-457914.FitBit_Dataset.sleep_month_2`
WHERE TotalMinutesAsleep IS NOT NULL
GROUP BY
Id
ORDER BY avg_minutes_asleep DESC
```

Id	avg_minutes_asleep	outcome
1844505072	652	Enough sleep
2026352035	506.2	Enough sleep
6117666160	478.8	Enough sleep
4319703577	476.7	Enough sleep
5553957443	463.5	Enough sleep
7086361926	453.1	Enough sleep
6962181067	448	Enough sleep
2347167796	446.8	Enough sleep
8378563200	445.1	Enough sleep
8792009665	435.7	Enough sleep
5577150313	432	Enough sleep
4702921684	417.5	Not enough sleep
1927972279	417	Not enough sleep
4388161847	400.2	Not enough sleep

4445114986	385.2	Not enough sleep
1503960366	360.3	Not enough sleep
6775888955	349.7	Not enough sleep
4020332650	349.4	Not enough sleep
8053475328	297	Not enough sleep
1644430081	294	Not enough sleep
3977333714	293.6	Not enough sleep
4558609924	127.6	Not enough sleep
7007744171	68.5	Not enough sleep
2320127002	61	Not enough sleep

According to my results, 11 of the 24 subjects were getting enough sleep, while 13 were not. This is also concerning, since more than half of the subjects were not sleeping enough. Additionally, there are 3 really big outliers. The 3 users with the least average sleep had 127.6, 68.5 and 61 minutes of sleep on average, which is between 1 and 2 hours approximately. This seems really extreme, so I went to look into it. I wrote the following query:

```
SELECT
  Id,
  SleepDay,
  TotalMinutesAsleep
FROM
  `bellabeat-case-study-457914.FitBit_Dataset.sleep_month_2`
WHERE Id = 7007744171 OR Id = 2320127002 OR Id = 4558609924
```

Id	SleepDay	TotalMinutesAsleep
2320127002	4/23/2016	61
4558609924	4/21/2016	126
4558609924	4/26/2016	103
4558609924	4/29/2016	171
4558609924	5/1/2016	115
4558609924	5/8/2016	123
7007744171	4/16/2016	79
7007744171	5/1/2016	58

It it shown that one of the least sleepers had only one sleep day logged, another had only two, and the other had five. It could be likely that, at least for the two with the least logs, they happened to log their sleeps on days when they got significantly below average amount of sleep, and therefore our results. However, for the third user, it is extremely unlikely that they happened to log their sleep on their days with significantly below average amount of sleep five times in a row, so I would say that this person most likely averages a similar amount of sleep to what is logged, which is very concerning, since they average only a little over 2 hours a day.

I then wanted to see how the amount of time in bed related to the amount of time asleep. I wrote the following query:

```
SELECT
  Id,
  ROUND(AVG(TotalMinutesAsleep),1) AS avg_minutes_asleep,
  ROUND(AVG(TotalTimeInBed),1) AS avg_minutes_in_bed,
  ROUND(AVG(TotalTimeInBed) - AVG(TotalMinutesAsleep), 1) AS in_bed_not_sleeping,
  CASE WHEN ROUND(AVG(TotalMinutesAsleep),1) > 420 THEN 'Enough sleep'
  WHEN ROUND(AVG(TotalMinutesAsleep),1) < 420 THEN 'Not enough sleep'
  END AS outcome

FROM
  `bellabeat-case-study-457914.FitBit_Dataset.sleep_month_2`
WHERE TotalMinutesAsleep IS NOT NULL
GROUP BY
Id
ORDER BY in_bed_not_sleeping DESC
```

Id	avg_minutes_asleep	avg_minutes_in_bed	in_bed_not_sleeping	outcome
1844505072	652	961	309	Enough sleep
3977333714	293.6	461.1	167.5	Not enough sleep
1644430081	294	346	52	Not enough sleep
2347167796	446.8	491.3	44.5	Enough sleep
5553957443	463.5	505.9	42.4	Enough sleep
8378563200	445.1	485.9	40.8	Enough sleep
4445114986	385.2	416.8	31.6	Not enough sleep
2026352035	506.2	537.6	31.5	Enough sleep
6117666160	478.8	510.2	31.4	Enough sleep
4020332650	349.4	379.8	30.4	Not enough sleep
5577150313	432	460.6	28.6	Enough sleep
4319703577	476.7	502	25.3	Enough sleep

4388161847	400.2	423.2	23	Not enough sleep
1503960366	360.3	383.2	22.9	Not enough sleep
1927972279	417	437.8	20.8	Not enough sleep
4702921684	417.5	438.2	20.7	Not enough sleep
6775888955	349.7	369	19.3	Not enough sleep
8792009665	435.7	453.8	18.1	Enough sleep
6962181067	448	466.1	18.1	Enough sleep
7086361926	453.1	466.4	13.3	Enough sleep
4558609924	127.6	140	12.4	Not enough sleep
2320127002	61	69	8	Not enough sleep
8053475328	297	301.7	4.7	Not enough sleep
7007744171	68.5	71.5	3	Not enough sleep

Most of the subjects had less than an hour, on average, of time in bed while not sleeping. This is relatively normal, considering the amount of time individuals take to fall asleep, and the amount of time indivudals take to get out of bed after waking up. However, here there are two outliers. There is a subject that had 309 minutes, on average, of time in bed not sleeping; this is over 5 hours! This also happens to be the subject with the most sleep, on average. This user averaged over 16 hours in bed daily — an unusually high figure that may indicate a data anomaly or unique circumstance, so I wanted to check out their Id on the Daily Activty, to find out if it could be possible that they were injured during this period and had to rest, or if this was an error.

I wrote the following query:

```
SELECT
  Id,
  ActivityDate,
  TotalSteps,
  VeryActiveMinutes,
  FairlyActiveMinutes,
  LightlyActiveMinutes,
  Calories

FROM (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
```

```
WHERE Id = 1844505072
```

Id	ActivityDate	TotalSteps	VeryActive Minutes	FairlyActive Minutes	LightlyActive Minutes	Calories
1844505072	4/1/2016	6847	9	9	251	1969
1844505072	4/2/2016	5367	0	0	263	1889
1844505072	4/3/2016	2841	0	0	136	1636
1844505072	4/4/2016	0	0	0	0	1347
1844505072	4/5/2016	0	0	0	0	1347
1844505072	4/6/2016	7	0	0	1	1349
1844505072	4/7/2016	6344	0	0	290	1965
1844505072	4/8/2016	5316	0	0	221	1824
1844505072	4/9/2016	4979	0	0	184	1807
1844505072	4/10/2016	6556	0	0	305	1971
1844505072	4/11/2016	5430	0	0	251	1888
1844505072	4/12/2016	0	0	0	0	399
1844505072	4/12/2016	6697	0	0	339	2030
1844505072	4/13/2016	4929	0	0	248	1860
1844505072	4/14/2016	7937	0	0	373	2130
1844505072	4/15/2016	3844	0	0	176	1725
1844505072	4/16/2016	3414	0	0	147	1657
1844505072	4/17/2016	4525	2	8	199	1793
1844505072	4/18/2016	4597	0	12	217	1814
1844505072	4/19/2016	197	0	0	10	1366
1844505072	4/20/2016	8	0	0	1	1349
1844505072	4/21/2016	8054	2	13	308	2062
1844505072	4/22/2016	5372	0	0	220	1827
1844505072	4/23/2016	3570	0	0	139	1645
1844505072	4/24/2016	0	0	0	0	1347
1844505072	4/25/2016	0	0	0	0	1347

1844505072	4/26/2016	0	0	0	0	1347
1844505072	4/27/2016	4	0	0	1	1348
1844505072	4/28/2016	6907	0	0	302	1992
1844505072	4/29/2016	4920	0	0	247	1856
1844505072	4/30/2016	4014	0	0	184	1763
1844505072	5/1/2016	2573	0	7	75	1541
1844505072	5/2/2016	0	0	0	0	1348
1844505072	5/3/2016	4059	0	0	184	1742
1844505072	5/4/2016	2080	0	0	87	1549
1844505072	5/5/2016	2237	0	0	120	1589
1844505072	5/6/2016	44	0	0	2	1351
1844505072	5/7/2016	0	0	0	0	1347
1844505072	5/8/2016	0	0	0	0	1347
1844505072	5/9/2016	0	0	0	0	1347
1844505072	5/10/2016	0	0	0	0	1347
1844505072	5/11/2016	0	0	0	0	1347
1844505072	5/12/2016	0	0	0	0	665

Here I can see that there isn't a big consecutive period of days of no activity, but there are some chunks of time were the user probably didn't put on their device, or for very little time, so the probability of the huge amount of time in bed previously observed being derived from an injury is pretty low. The user was probably on vacation during the second month of the study, and also used their device less during this time.

Weight data analysis

Next, I wanted to take a look at the weight log datasets. I wanted to see how often users were logging their weight, what whas their starting and end weights, and how was their avergae BMI. According to the National Heart, Lung and Blood Institute, https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm, the BMI categories are the following:

Underweight = <18.5
Normal weight = 18.5–24.9
Overweight = 25–29.9
Obesity = BMI of 30 or greater

I wrote the following query:

```
WITH main AS (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_2`
),
first_last_weights AS (
  SELECT
    Id,
    FIRST_VALUE(WeightPounds) OVER (PARTITION BY Id ORDER BY Date ASC) AS
starting_weight,
    FIRST_VALUE(WeightPounds) OVER (PARTITION BY Id ORDER BY Date DESC) AS end_weight
  FROM main
)
SELECT
  m.Id,
  COUNT(DISTINCT m.Date) AS num_of_logs,
  ROUND(flw.starting_weight,1) AS start_weight,
  ROUND(flw.end_weight,1) AS end_weight,
  ROUND(AVG(m.BMI),1) AS avg_bmi,
  CASE
    WHEN AVG(m.BMI) < 18.5 THEN 'Underweight'
    WHEN AVG(m.BMI) >= 18.5 AND AVG(m.BMI) < 25 THEN 'Normal weight'
    WHEN AVG(m.BMI) >= 25 AND AVG(m.BMI) < 30 THEN 'Overweight'
    ELSE 'Obese'
  END AS bmi_category
FROM main AS m
INNER JOIN first_last_weights AS flw
  ON m.Id = flw.Id
GROUP BY m.Id, flw.starting_weight, flw.end_weight
ORDER BY avg_bmi;
```

Id	num_of_logs	start_weight	end_weight	avg_bmi	bmi_category
2873212765	4	125	126.3	21.6	Normal weight
1503960366	3	117.5	116	22.8	Normal weight
6962181067	43	135.6	136.5	24.1	Normal weight
2347167796	1	139.8	139.8	24.8	Normal weight
2891001357	1	194.9	194.9	25	Overweight
8877689391	32	188.5	185.2	25.5	Overweight
4702921684	1	219.8	219.8	26.1	Overweight
4558609924	6	153	152.3	27.2	Overweight
4319703577	2	159.6	159.4	27.4	Overweight
5577150313	1	200	200	28	Overweight
8253242879	1	166.7	166.7	29.5	Overweight
4445114986	1	203.7	203.7	35	Obese
1927972279	2	285.7	294.3	46.9	Obese

We can see that very few users logged their weight, and most of those that logged their weight, did it only a few times. Only 13 users logged their weight during this period, 6 of those logging it only once, 2 logged twice, 1 logged three times, 1 logged four times, 1 logged six times, 1 logged thirty-two times and the last one logged forty-three times.

Out of the 13 subjects, 4 were, on average, normal weight, 7 were, on average, overweight, and 2 were, on average, obese. This means that most of the subjects were over the designated healthy weight. However, it needs to be stated that some individuals may have a BMI indicating that they are overweight while having a realtively low body fat. This can happen if the individual has a significant amount of muscle mass, causing them to be very heavy for their height. Since only two of the measurements in this dataset had body fat information, I didn't include it in my query, and we can't be sure what the actual body composition of these individuals looks like.

Comparing and relating datasets

Having had a look at all the datasets, I want to make some comparisons between them, to see if there exists any correlation between different factors. I want to start by seeing if there is any correlation between BMI and resting heart rate.

```
WITH main AS (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_2`
),
daily_min_hr AS (
  SELECT
    Id,
    DATE(Time) AS day,
    MIN(Value) AS min_hr
  FROM (
    SELECT *
    FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_1`
    UNION ALL
    SELECT *
    FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_2`
  )
  GROUP BY Id, day
),
avg_resting_hr AS (
  SELECT
    Id,
    ROUND(AVG(min_hr), 1) AS approx_resting_hr
  FROM daily_min_hr
  GROUP BY Id
)

SELECT
  m.Id,
  COUNT(DISTINCT m.Date) AS num_of_logs,
  ROUND(AVG(m.BMI), 1) AS avg_bmi,
  CASE
    WHEN AVG(m.BMI) < 18.5 THEN 'Underweight'
    WHEN AVG(m.BMI) >= 18.5 AND AVG(m.BMI) < 25 THEN 'Normal weight'
    WHEN AVG(m.BMI) >= 25 AND AVG(m.BMI) < 30 THEN 'Overweight'
    ELSE 'Obese'
  END AS bmi_category,
  hr.approx_resting_hr,
  CASE
```

```
    WHEN hr.approx_resting_hr BETWEEN 60 AND 100 THEN 'Healthy Resting HR'
    ELSE 'Outside Healthy Range'
  END AS status
FROM main AS m
INNER JOIN avg_resting_hr AS hr
  ON m.Id = hr.Id
GROUP BY m.Id, hr.approx_resting_hr
ORDER BY avg_bmi;
```

Id	num_of_logs	avg_bmi	bmi_category	approx_resting_hr	status
6962181067	43	24.1	Normal weight	54.3	Outside Healthy Range
2347167796	1	24.8	Normal weight	53.8	Outside Healthy Range
8877689391	32	25.5	Overweight	50.5	Outside Healthy Range
4558609924	6	27.2	Overweight	55.1	Outside Healthy Range
5577150313	1	28	Overweight	42.1	Outside Healthy Range

We can obverse that there is very little overlap between the users that logged their weight and the users that logged their heartrate, there are only five subjects in this results table. This won't lead to any significant conclusions. As we can see, all five of the subejcts were outside the healthy heartrate range, having three of them being overweight and two of them being normal weight.

Next, I'll see if there is any correlation between BMI and sleep.

```
SELECT
  m.Id,
  ROUND(AVG(m.BMI), 1) AS avg_bmi,
  CASE
    WHEN AVG(m.BMI) < 18.5 THEN 'Underweight'
    WHEN AVG(m.BMI) >= 18.5 AND AVG(m.BMI) < 25 THEN 'Normal weight'
    WHEN AVG(m.BMI) >= 25 AND AVG(m.BMI) < 30 THEN 'Overweight'
    ELSE 'Obese'
  END AS bmi_category,
  ROUND(AVG(s.TotalMinutesAsleep),1) AS avg_minutes_asleep,
  CASE WHEN ROUND(AVG(s.TotalMinutesAsleep),1) > 420 THEN 'Enough sleep'
  WHEN ROUND(AVG(s.TotalMinutesAsleep),1) < 420 THEN 'Not enough sleep'
  END AS sleep_outcome

FROM
  `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_2` AS m
INNER JOIN
  `bellabeat-case-study-457914.FitBit_Dataset.sleep_month_2` AS s
ON m.Id = s.Id
```

```
GROUP BY
  m.Id
ORDER BY
  avg_bmi
```

Id	avg_bmi	bmi_category	avg_minutes_asleep	sleep_outcome
1503960366	22.6	Normal weight	360.3	Not enough sleep
6962181067	24	Normal weight	448	Enough sleep
4558609924	27.2	Overweight	127.6	Not enough sleep
4319703577	27.4	Overweight	476.7	Enough sleep
5577150313	28	Overweight	432	Enough sleep
1927972279	47.5	Obese	417	Not enough sleep

In this case, like the last, we also only have a few subjects that had both sleep logs and weight logs. Here we have 6 subjects, which is not enough to make meaninfgul conclusions. In addition, with the little data we have, there doesn't seem to be a correlation between sleep and BMI.

Next, I wanted to see if it was true, for this sample, that more active people tend to have lower resting heart rates.

```
WITH daily_min_hr AS (
  SELECT
    Id,
    DATE(Time) AS day,
    MIN(Value) AS min_hr
  FROM (
    SELECT *
    FROM
    `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_1`
    UNION ALL
    SELECT *
    FROM `bellabeat-case-study-457914.FitBit_Dataset.heartrate_month_2`

  )
  GROUP BY
    Id, day
),
avg_resting_hr AS (
  SELECT
    Id,
    ROUND(AVG(min_hr), 1) AS approx_resting_hr
  FROM
```

```
    daily_min_hr
  GROUP BY
    Id
),
activity_combined AS (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
)
SELECT
  act.Id,
  SUM(act.VeryActiveMinutes) AS VigorousActivityMinutes,
  SUM(act.FairlyActiveMinutes) AS ModerateActivityMinutes,
  SUM(act.VeryActiveMinutes) + ((0.5) * SUM(act.FairlyActiveMinutes)) AS
overall_level,
  hr.approx_resting_hr
FROM activity_combined AS act
INNER JOIN
avg_resting_hr AS hr
ON act.Id = hr.Id
GROUP BY Id, hr.approx_resting_hr
ORDER BY overall_level
```

Id	VigorousActivityMinutes	ModerateActivityMinutes	overall_level	approx_resting_hr
2026352035	3	8	7	66
6391747486	46	6	49	60.7
6117666160	49	98	98	54.9
8792009665	46	182	137	49.4
4020332650	291	411	496.5	55.4
4558609924	374	458	603	55.1
2347167796	420	717	778.5	53.8
6775888955	465	828	879	65.5
4388161847	718	631	1033.5	44.4
5553957443	1009	603	1310.5	50.6
7007744171	1343	618	1652	62.8
6962181067	1202	980	1692	54.3
2022484408	1606	870	2041	51.8

8877689391	2850	489	3094.5	50.5
5577150313	3520	1205	4122.5	42.1

On a simple view of the table, there appears to be a slight correlation between overall activity level and approximate resting heart rate. We would see this more clearly when we create some visualizations.

Finally, I'm going to compare activity levels to BMI. It makes sense that more active subjects would tend to have a healthier BMI. Let's see if that's true for our sample.

```
WITH activity_combined AS (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.daily_activity_month_2`
),
weight_combined AS (
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_1`
  UNION ALL
  SELECT *
  FROM `bellabeat-case-study-457914.FitBit_Dataset.weight_log_month_2`
)
SELECT
  a.Id,
  SUM(a.VeryActiveMinutes) + ((0.5) * SUM(a.FairlyActiveMinutes)) AS overall_level,
  ROUND(AVG(w.BMI), 1) AS avg_bmi,
  CASE
    WHEN AVG(w.BMI) < 18.5 THEN 'Underweight'
    WHEN AVG(w.BMI) >= 18.5 AND AVG(w.BMI) < 25 THEN 'Normal weight'
    WHEN AVG(w.BMI) >= 25 AND AVG(w.BMI) < 30 THEN 'Overweight'
    ELSE 'Obese'
  END AS bmi_category,


FROM activity_combined AS a
INNER JOIN
weight_combined AS w
ON a.Id = w.Id
GROUP BY Id
ORDER BY overall_level
```

Id	overall_level	avg_bmi	bmi_category
1927972279	126	46.9	Obese

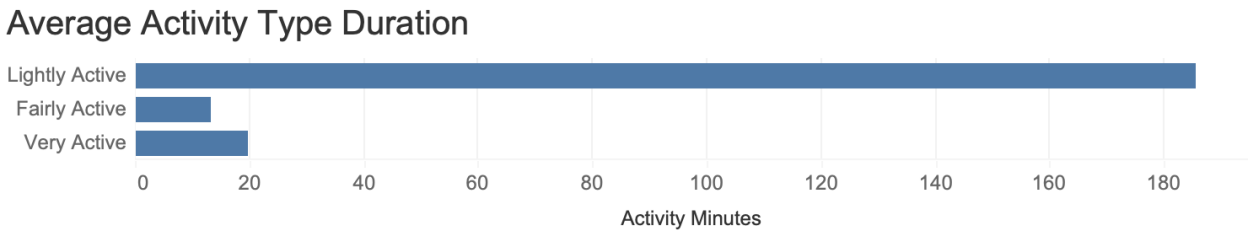
4445114986	315	35	Obese
2891001357	330	25	Overweight
8253242879	624	29.5	Overweight
4702921684	736	26.1	Overweight
2347167796	778.5	24.8	Normal weight
4319703577	981	27.4	Overweight
2873212765	2516	21.6	Normal weight
4558609924	3618	27.2	Overweight
5577150313	4122.5	28	Overweight
1503960366	6985.5	22.8	Normal weight
6962181067	74448	24.1	Normal weight
8877689391	102118.5	25.5	Overweight

In these results we have 13 subjects. There appears to be a slight correlation between overall activity level and BMI, indicating that more active individuals tend to have a healthier BMI. We would see this relationship with more detail in the visualizations.

Analysis Conclusions:

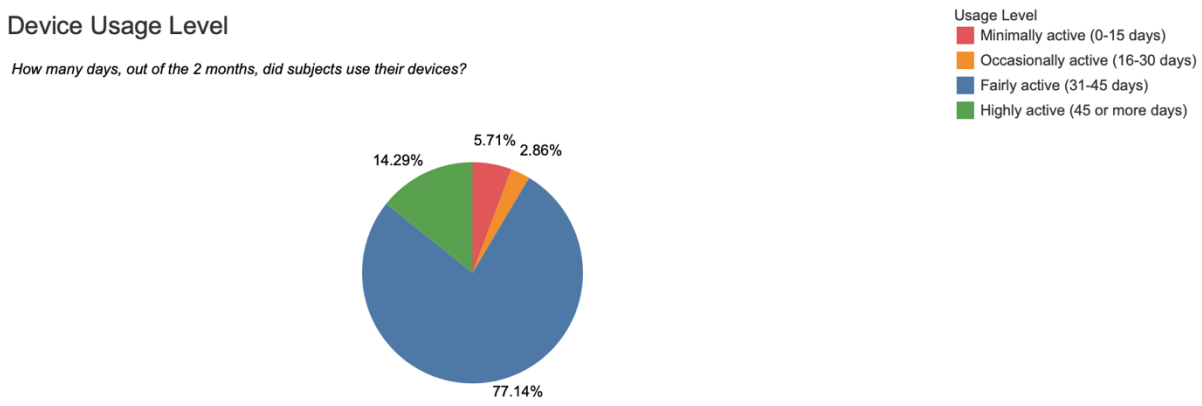
- Most users wore their devices most of the days of the study, but only a small group wore them nearly every day, as intended.
- Only 20 out of 35 users met the CDC guidelines for recommended weekly activity.
- Not many users logged their heart rates or weight info.
- Only 11 of 24 subjects were getting enough sleep.
- Out of the 13 subjects that logged their weight, 4 were, on average, normal weight, 7 were, on average, overweight, and 2 were, on average, obese.
- There appears to be a slight correlation between overall activity level and approximate resting heart rate. More active users tend to have a lower resting heart rate.
- More active individuals tended to have a healthier BMI.

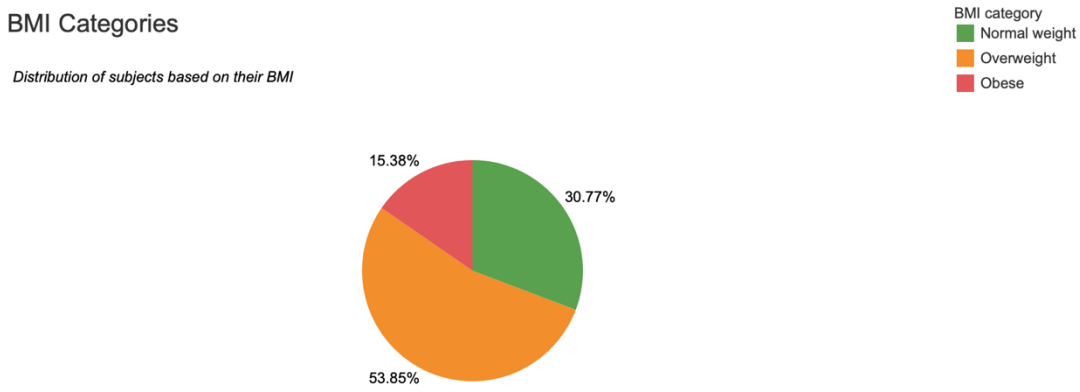
SHARE PHASE



From the above visualization, we can see that:

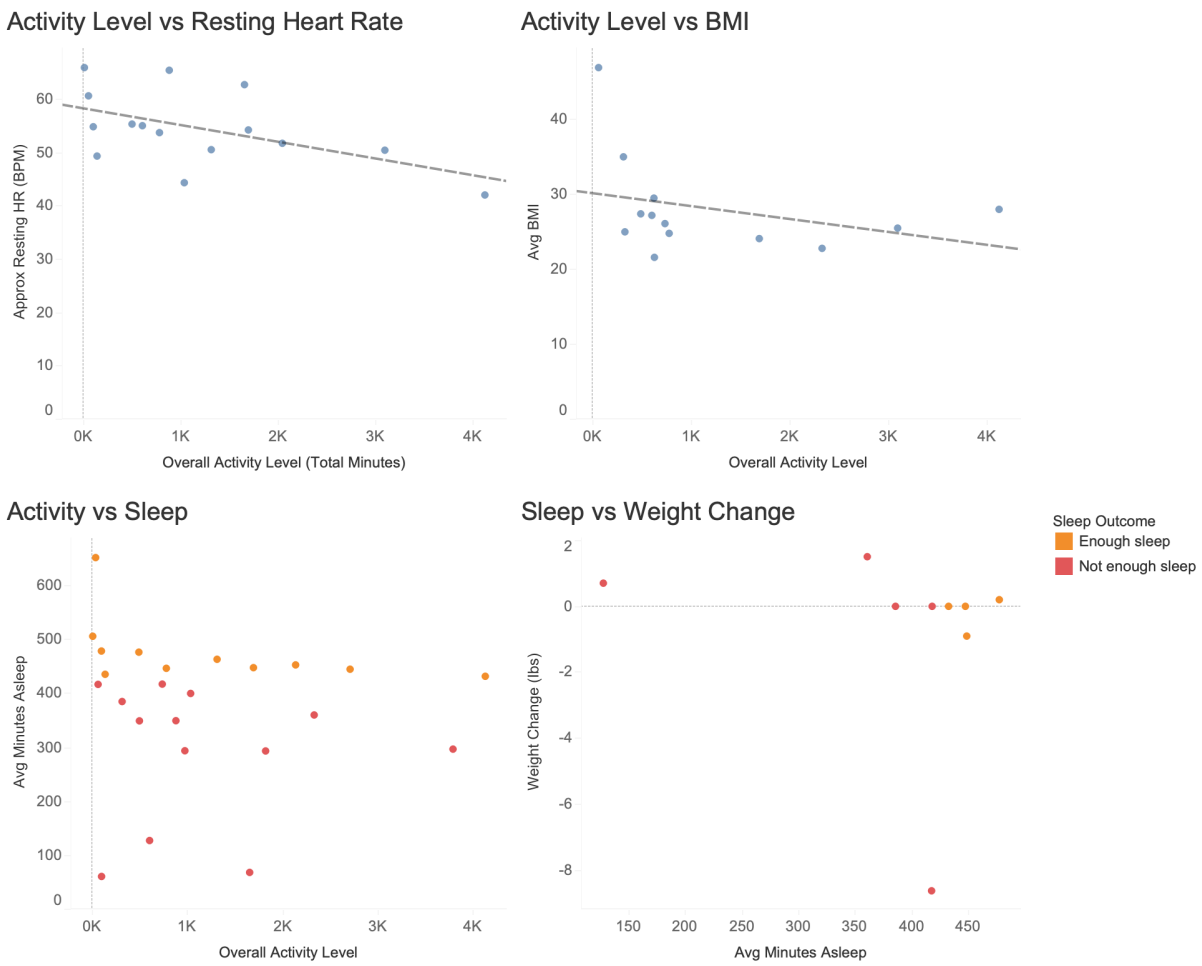
- Users tend to do very little moderate and vigorous intensity activity, in comparison to light activity. Users, on average, did less than 20 minutes of both moderate and vigorous intensity activity for the whole 2 months.





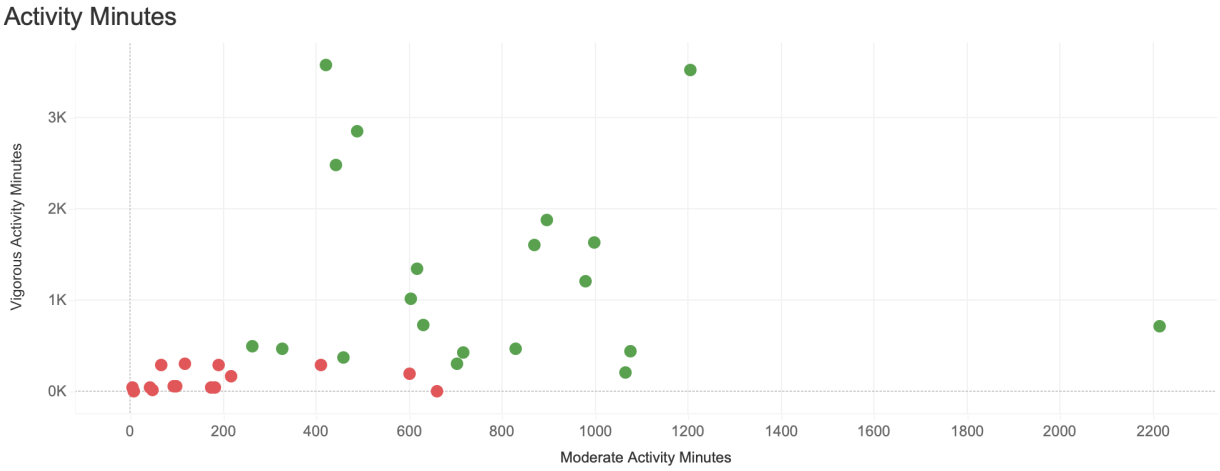
From the above visualizations, we can see that:

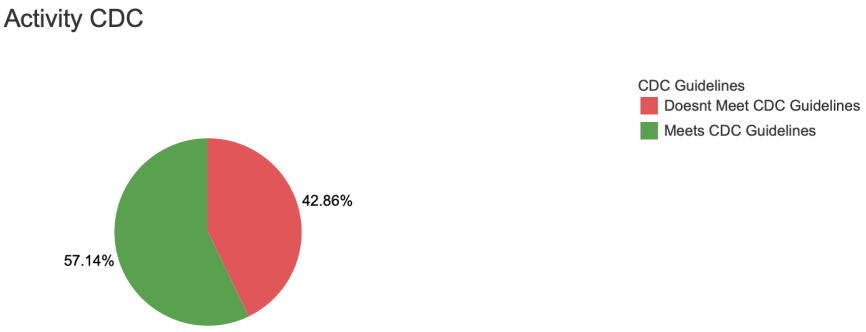
- Most users wore their devices for more than half of all days, but very little of them wore them for 45 days or more.
- Most of the users are above a healthy weight; only 30% had, on average, a normal weight for their height.



From the above visualizations, we can see that:

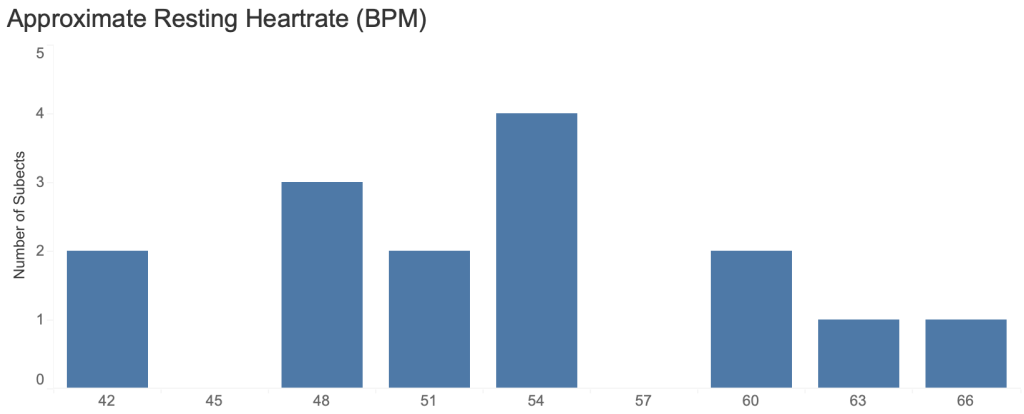
- There is a slight negative correlation between overall activity level and resting heart rate, with an R-Squared of 0.28 and p-value of 0.04, making it statistically significant.
- There appears to be close to no correlation between activity level and BMI, but the data is too small to tell.
- There appears to be no correlation between activity level and sleep, but, again, the data is too small to tell.
- There are only a couple subjects that had sleep and weight data, not enough to study a possible correlation.
- Very few people logged their sleep and weight.

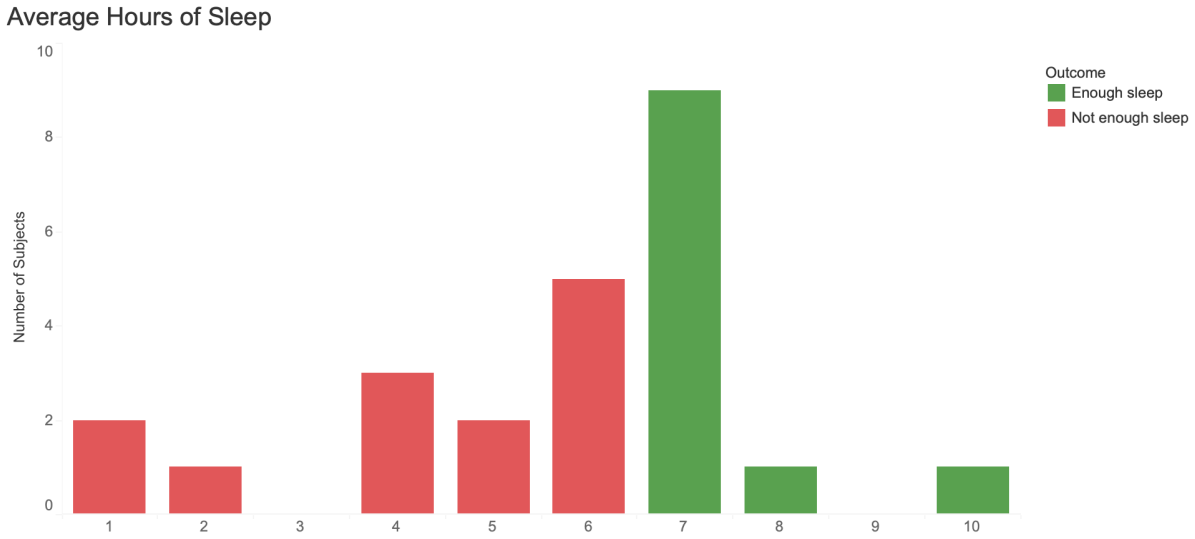




From the above visualizations, we can see that:

- Users that had more moderate activity minutes, tended to also have more vigorous activity minutes.
- Only a little more than half of all subjects met the CDC requierement for minimum activity.





From the above visualizations, we can see that:

- Users had low resting heart rates, compared to the general health range, which is 60-100 BPM. However, more active individuals tend to have lower resting heart rates, so this is not negative.
- Most individuals were getting less than enough sleep.

ACT PHASE

There are a lot of key insights that we can extract from our findings, in order to lead Bellabeat's marketing decisions, based on this 2-month study on FitBit data.

Firstly, when it comes to the subjects wearing their devices, I noticed that nearly no individuals wore their device everyday, or very close to everyday. For smart health devices to be used to their full potential, they need to be worn close to everyday. Bellabeat should implement a marketing strategy that emphasizes the importance of daily device usage. A helpful strategy might be to accentuate the sleek nature of Bellabeat's devices, which can be worn seemlessly everyday, even for more formal situations. In addition, it will be great for there to be reminders on user's phones, as notifications, to wear their devices.

In regards to activity, users should be incentivized to do physical activity more often, specially more intense types of physical activity, as it helps with heart health and stamina. Only slighly more than half of subjects met the CDC guidelines for weekly activity. This is supported by the visualization that shows that more active individuals had a lower average resting heart rate. We could implement a sort of game, with points earned on how high heart rate elevates during activity, and that to be compared to other users' data. In addition, Bellabeat should emphasize the importance of physical activity, while making the point that these smart devices help to stay in check with the requiered physical activity.

In addition, more than half of subjects were overweight, which is concerning. The previous recommendation on activity could be useful here. However, I noticed that very little users logged their weight. There should be an incentive on weight logging, since that is the most efficient way of tracking weight loss or gain, and help users stay on track and monitor their progress, or lack thereof. Bellabeat devices should push users to log their weight daily, if possible.

In regards to sleep, the results were very concerning, since most users were having less than enough sleep. It will be smart to implement some sort of point or reward system, similar to what I suggested for activity, to incentivize users to improve their sleep.

Overall, the data was very limited in most of the datasets, especially the weight log dataset. I also had only one month worth of data for the sleep dataset. If I had more sleep data, I could have found more meaningful conclusions in relation to sleep's effect on other metrics, and vice versa.

If we could push users to log more of their data, relating to weight and sleep, and wear their devices more often, we could have better data for a later analysis.