Bellabeat Company

Final Report on Bellabeat Case Study

Presenter: Nguyen Truc Linh

October 6, 2024

Content

1 Problem Statement

The primary goal is to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices by following these questions: 1. What are some trends in smart device usage? 2. How could these trends apply to Bellabeat customers? 3. How could these trends help influence Bellabeat marketing strategy?

2 Data Description

The dataset used for this analysis comes from public sources on Kaggle, including Fitbit fitness tracker data, and is stored locally and in cloud services such as Google BigQuery. The data, collected from 30 Fitbit users, includes minute-level output on activity, heart rate, and sleep monitoring. The attributes in the dataset are as follows:

- **Id**: Unique identifier for each user.
- **SleepDay**: Date of the sleep record.
- **TotalSleepRecords**: Number of sleep sessions recorded in a day.
- **TotalMinutesAsleep**: Total number of minutes the user was asleep.
- **TotalTimeInBed**: Total number of minutes the user spent in bed.
- **ActivityDate**: Date of recorded activity.
- **TotalSteps**: Total number of steps taken by the user.
- **TotalDistance**: Total distance covered by the user.
- **TrackerDistance**: Distance recorded by the fitness tracker.
- **LoggedActivitiesDistance**: Distance covered in logged activities.
- **VeryActiveDistance**: Distance covered during very active periods.
- **ModeratelyActiveDistance**: Distance covered during moderately active periods.
- **LightActiveDistance**: Distance covered during light activity.
- **SedentaryActiveDistance**: Distance covered during sedentary activity.
- **VeryActiveMinutes**: Number of minutes spent in very active physical activity.
- **FairlyActiveMinutes**: Number of minutes spent in fairly active physical activity.
- **LightlyActiveMinutes**: Number of minutes spent in light physical activity.
- **SedentaryMinutes**: Number of minutes spent in sedentary activities.
- **Calories**: Number of calories burned through physical activities.

3 Data Cleaning and Manipulation

Data cleaning and manipulation were executed using Google Sheets and BigQuery with SQL to ensure data integrity and analytical robustness. The process involved:

- **Data Loading and Structure Setup**: Imported the raw data from the Fitbit dataset and constructed tables in Google Sheets, which were subsequently combined and analyzed in BigQuery.

```sql
CREATE TABLE `first-project-435214.Fitbit.JoinedSleepDailyActivity` AS
SELECT
    da.*,
    sd.SleepDay,
    sd.TotalMinutesAsleep,
    sd.TotalSleepRecords,
    sd.TotalTimeInBed
FROM `first-project-435214.Fitbit.sleeyDay` sd
JOIN `first-project-435214.Fitbit.dailyActivity` da
ON sd.Id = da.Id;
```

Combine 2 Dataset

- **Cleaning Process**: Removed null values, corrected erroneous entries, and standardized data formats across all attributes to maintain consistency.

```sql
-- Display rows containing NULL values or negative values
SELECT *
FROM `first-project-435214.Fitbit.JoinedSleepDailyActivity`
WHERE
    -- Check for NULL values in key columns
    TotalSteps IS NULL
    OR TotalDistance IS NULL
    OR TotalMinutesAsleep IS NULL
    OR TotalTimeInBed IS NULL

    -- Check for negative values in key columns
    OR TotalSteps < 0
    OR TotalDistance < 0
    OR SedentaryMinutes < 0
    OR VeryActiveMinutes < 0
    OR FairlyActiveMinutes < 0
    OR LightlyActiveMinutes < 0
    OR TotalMinutesAsleep < 0
    OR TotalTimeInBed < 0;
```

Check for null, negative value

```sql
-- Step 1: Create a cleaned version of the joined dataset
CREATE TABLE `first-project-435214.Fitbit.
    CleanedJoinedSleepDailyActivity` AS
```

```
WITH cleaned_data AS (
    -- Step 1: Remove Null Values
    SELECT *
    FROM `first-project-435214.Fitbit.JoinedSleepDailyActivity`
    WHERE Id IS NOT NULL
      AND TotalSteps IS NOT NULL
      AND TotalDistance IS NOT NULL
      AND TotalMinutesAsleep IS NOT NULL
      AND TotalTimeInBed IS NOT NULL
      AND TrackerDistance IS NOT NULL
      AND LoggedActivitiesDistance IS NOT NULL
      AND VeryActiveDistance IS NOT NULL
      AND ModeratelyActiveDistance IS NOT NULL
      AND LightActiveDistance IS NOT NULL
      AND SedentaryActiveDistance IS NOT NULL
      AND VeryActiveMinutes IS NOT NULL
      AND FairlyActiveMinutes IS NOT NULL
      AND LightlyActiveMinutes IS NOT NULL
      AND SedentaryMinutes IS NOT NULL
      AND Calories IS NOT NULL
)

, corrected_data AS (
    SELECT *,
        -- Step 2: Correct Erroneous Entries by ensuring all values are non-negative
        CASE WHEN TotalSteps < 0 THEN 0 ELSE TotalSteps END AS CorrectedTotalSteps,
        CASE WHEN TotalDistance < 0 THEN 0 ELSE TotalDistance END AS CorrectedTotalDistance,
        CASE WHEN TrackerDistance < 0 THEN 0 ELSE TrackerDistance END AS CorrectedTrackerDistance,
        CASE WHEN LoggedActivitiesDistance < 0 THEN 0 ELSE LoggedActivitiesDistance END AS CorrectedLoggedActivitiesDistance,
        CASE WHEN VeryActiveDistance < 0 THEN 0 ELSE VeryActiveDistance END AS CorrectedVeryActiveDistance,
        CASE WHEN ModeratelyActiveDistance < 0 THEN 0 ELSE ModeratelyActiveDistance END AS CorrectedModeratelyActiveDistance,
        CASE WHEN LightActiveDistance < 0 THEN 0 ELSE LightActiveDistance END AS CorrectedLightActiveDistance,
        CASE WHEN SedentaryActiveDistance < 0 THEN 0 ELSE SedentaryActiveDistance END AS CorrectedSedentaryActiveDistance,
        CASE WHEN VeryActiveMinutes < 0 THEN 0 ELSE VeryActiveMinutes END AS CorrectedVeryActiveMinutes,
        CASE WHEN FairlyActiveMinutes < 0 THEN 0 ELSE FairlyActiveMinutes END AS CorrectedFairlyActiveMinutes,
```

```
            CASE WHEN LightlyActiveMinutes < 0 THEN 0 ELSE
                LightlyActiveMinutes END AS CorrectedLightlyActiveMinutes,
            CASE WHEN SedentaryMinutes < 0 THEN 0 ELSE SedentaryMinutes END
                AS CorrectedSedentaryMinutes,
            CASE WHEN TotalMinutesAsleep < 0 THEN 0 ELSE TotalMinutesAsleep
                END AS CorrectedTotalMinutesAsleep,
            CASE WHEN TotalTimeInBed < 0 THEN 0 ELSE TotalTimeInBed END AS
                CorrectedTotalTimeInBed,
            CASE WHEN Calories < 0 THEN 0 ELSE Calories END AS
                CorrectedCalories
        FROM cleaned_data
    )

    -- Step 3: Standardize Data Formats and Create Final Cleaned Table
    SELECT
        Id,
        -- Standardize SleepDay format using SAFE.PARSE_DATE with
            conditional fallback
        COALESCE(
            SAFE.PARSE_DATE('%m/%d/%Y', CAST(SleepDay AS STRING)),
            SAFE.PARSE_DATE('%Y-%m-%d', CAST(SleepDay AS STRING))
        ) AS SleepDay,
        COALESCE(
            SAFE.PARSE_DATE('%m/%d/%Y', CAST(ActivityDate AS STRING)),
            SAFE.PARSE_DATE('%Y-%m-%d', CAST(ActivityDate AS STRING))
        ) AS ActivityDate,

        TotalSleepRecords,
        CorrectedTotalSteps AS TotalSteps,
        CorrectedTotalDistance AS TotalDistance,
        CorrectedTrackerDistance AS TrackerDistance,
        CorrectedLoggedActivitiesDistance AS LoggedActivitiesDistance,
        CorrectedVeryActiveDistance AS VeryActiveDistance,
        CorrectedModeratelyActiveDistance AS ModeratelyActiveDistance,
        CorrectedLightActiveDistance AS LightActiveDistance,
        CorrectedSedentaryActiveDistance AS SedentaryActiveDistance,
        CorrectedVeryActiveMinutes AS VeryActiveMinutes,
        CorrectedFairlyActiveMinutes AS FairlyActiveMinutes,
        CorrectedLightlyActiveMinutes AS LightlyActiveMinutes,
        CorrectedSedentaryMinutes AS SedentaryMinutes,
        CorrectedTotalMinutesAsleep AS TotalMinutesAsleep,
        CorrectedTotalTimeInBed AS TotalTimeInBed,
        CorrectedCalories AS Calories
    FROM
        corrected_data;
```

Create a clean version

- **Feature Engineering**: Derived new variables to enrich the dataset:

  - **Total_active_minutes**: Total minutes engaged in moderate-to-high physical activity.
  - **sleep_efficiency**: A composite score reflecting sleep quality, calculated by integrating sleep duration and heart rate metrics during sleep.
  - **day_of_week**: Categorization of date into Monday, Tuesday, ... to easily examine.
  - **week_day**: Ordering data into week to discover trends over time.

- **Pivot Tables**: Generated aggregate metrics (e.g., SUM, AVERAGE) for steps and calories burned for each user. Analyzed the distribution of users by active minutes.

Id	SUM of TotalSteps	SUM of TotalActiveMinutes	SUM of Calories
1503960366	375619	8612	56309
1624580081	178061	5207	45984
1644430081	218489	6282	84339
1844505072	79982	3623	48778
1927972279	28400	1261	67357
2022484408	352490	9706	77809
2026352035	172573	7967	47760
2320127002	146223	6266	53449
2347167796	171354	5158	36782
2873212765	234229	10175	59426
3372868164	137233	6823	38662
3977333714	329537	7648	45410
4020332650	70284	2712	73960
4057192912	15352	421	7895
4319703577	225334	7585	63168
4388161847	335232	8459	95910
4445114986	148693	6741	67772
4558609924	238239	9581	63031
4702921684	265734	8328	91932
5553957443	266990	7521	58146
5577150313	249133	7953	100789
6117666160	197308	8175	63312
6290855005	163837	6786	75389
6775888955	65512	1715	55426
6962181067	303639	8901	61443
7007744171	294409	8529	66144
7086361926	290525	6565	79557
8053475328	457662	7617	91320
8253242879	123161	2883	33972
8378563200	270249	6976	106534
8583815059	223154	5275	84693
8792009665	53758	2807	56907
8877689391	497241	9632	106028
Grand Total	**7179636**	**213890**	**2165393**

(a) Sum of features by users

Id	AVERAGE of TotalSteps	AVERAGE of TotalActiveMinutes	AVERAGE of Calories
1503960366	12116.74194	277.8064516	1816.419355
1624580081	5743.903226	167.9677419	1483.354839
1644430081	7282.966667	209.4	2811.3
1844505072	2580.064516	116.8709677	1573.483871
1927972279	916.1290323	40.67741935	2172.806452
2022484408	11370.64516	313.0967742	2509.967742
2026352035	5566.870968	257	1540.645161
2320127002	4716.870968	202.1290323	1724.16129
2347167796	9519.666667	286.5555556	2043.444444
2873212765	7555.774194	328.2258065	1916.967742
3372868164	6861.65	341.15	1933.1
3977333714	10984.56667	254.9333333	1513.666667
4020332650	2267.225806	87.48387097	2385.806452
4057192912	3838	105.25	1973.75
4319703577	7268.83871	244.6774194	2037.677419
4388161847	10813.93548	272.8709677	3093.870968
4445114986	4796.548387	217.4516129	2186.193548
4558609924	7685.129032	309.0645161	2033.258065
4702921684	8572.064516	268.6451613	2965.548387
5553957443	8612.580645	242.6129032	1875.677419
5577150313	8304.433333	265.1	3359.633333
6117666160	7046.714286	291.9642857	2261.142857
6290855005	5649.551724	234	2599.62069
6775888955	2519.692308	65.96153846	2131.769231
6962181067	9794.806452	287.1290323	1982.032258
7007744171	11323.42308	328.0384615	2544
7086361926	9371.774194	211.7741935	2566.354839
8053475328	14763.29032	245.7096774	2945.806452
8253242879	6482.157895	151.7368421	1788
8378563200	8717.709677	225.0322581	3436.580645
8583815059	7198.516129	170.1612903	2732.032258
8792009665	1853.724138	96.79310345	1962.310345
8877689391	16040.03226	310.7096774	3420.258065
Grand Total	**7637.910638**	**227.5425532**	**2303.609574**

(b) Average of features by users

Figure 1

- **BigQuery Analysis**: Executed SQL queries to segment users by activity level, day of week. Aggregated data by week_day, active_minutes, and sleep_quality for deeper insights.

```sql
-- Aggregated query to analyze user activity by week_day, week_number,
    active_minutes, and sleep efficiency
WITH user_activity_summary AS (
    SELECT
        Id,
        DayOfWeek,
        WeekNumber,

        -- Metrics to retain
```

```
            VeryActiveMinutes,
            FairlyActiveMinutes,
            LightlyActiveMinutes,
            TotalActiveMinutes,
            SleepEfficiency,
            TotalMinutesAsleep,
            TotalTimeInBed,
            TotalSteps,
            Calories
        FROM
            `first-project-435214.Fitbit.EnrichedJoinedSleepDailyActivity`
    )

    -- Final aggregation query
    SELECT
        DayOfWeek,
        WeekNumber,

        -- Aggregated metrics by day of the week and week number
        COUNT(DISTINCT Id) AS user_count,
        AVG(VeryActiveMinutes) AS avg_very_active_minutes,
        AVG(FairlyActiveMinutes) AS avg_fairly_active_minutes,
        AVG(LightlyActiveMinutes) AS avg_lightly_active_minutes,
        AVG(TotalActiveMinutes) AS avg_total_active_minutes,
        AVG(SleepEfficiency) AS avg_sleep_efficiency,
        SUM(TotalSteps) AS total_steps_taken,
        SUM(Calories) AS total_calories_burned,
        AVG(TotalMinutesAsleep) AS avg_minutes_asleep,
        AVG(TotalTimeInBed) AS avg_time_in_bed

    FROM
        user_activity_summary

    GROUP BY
        DayOfWeek,
        WeekNumber

    ORDER BY
        WeekNumber,
        DayOfWeek;
```

Aggregated data

Row	DayOfWeek	WeekNumb	user_cour	avg_very_active_minutes	avg_fairly_active_minutes	avg_lightly_active_minute	avg_total_active_minutes	avg_sleep_efficiency	total_steps_taken	total_calories_burne	avg_minutes_asleep	avg_time_in_bed
1	3	15	24	26.002421307506062	8.5375302663438255	171.95399515738507	206.49394673123473	0.916769206515...	3285251	979255	419.4673123486...	458.6392251815...
2	4	15	24	18.56416464891042	12.842615012106537	162.75786924939462	194.16464891041156	0.916769206515...	2777627	925665	419.4673123486...	458.6392251815...
3	5	15	24	27.585956416464867	15.932203389830516	192.42130750605341	235.93946731234857	0.916769206515...	3208886	990490	419.4673123486...	458.6392251815...
4	6	15	24	22.753026634382561	16.644067796610159	228.96125907990327	268.35835351089588	0.916769206515...	3632910	1002364	419.4673123486...	458.6392251815...
5	7	15	24	30.639225181598068	18.162227602905563	205.34382566585953	254.14527845036338	0.916769206515...	3580852	991160	419.4673123486...	458.6392251815...
6	1	16	24	18.084745762711879	13.917675544794188	144.01937046004846	176.02179176755439	0.916769206515...	2204182	885149	419.4673123486...	458.6392251815...
7	2	16	24	30.300242130750615	21.392251815980629	173.5084745762712	225.20096852300262	0.916769206515...	3311593	978840	419.4673123486...	458.6392251815...
8	3	16	24	34.922518159806273	21.084745762711872	198.33656174334135	254.34382566585938	0.916769206515...	3795631	1007122	419.4673123486...	458.6392251815...
9	4	16	24	26.537530266343811	18.997578692493931	205.90314769975788	251.43825665859546	0.916769206515...	3510179	988422	419.4673123486...	458.6392251815...
10	5	16	24	30.254237288135595	17.740920096852292	177.06537530266343	225.06053268765098	0.916769206515...	3537350	1029462	419.4673123486...	458.6392251815...
11	6	16	24	32.278450363196129	17.295399515738517	194.76271186440667	244.33656174334129	0.916769206515...	3328983	987267	419.4673123486...	458.6392251815...
12	7	16	24	16.230024213075062	20.970944309927361	262.09685230024212	299.29782082324448	0.916769206515...	3945753	1003929	419.4673123486...	458.6392251815...
13	1	17	24	20.646489104116217	18.789346246973363	185.14043583535121	224.57627118644075	0.916769206515...	2935567	938267	419.4673123486...	458.6392251815...
14	2	17	24	34.41646489104118	14.648910411622277	199.9733656174335	249.03874092009698	0.916769206515...	3642411	994820	419.4673123486...	458.6392251815...
15	3	17	24	25.968523002421325	19.864406779661028	212.90072639225184	258.73365617433427	0.916769206515...	3738163	983986	419.4673123486...	458.6392251815...
16	4	17	24	19.123486682808711	14.590799031477003	222.70944309927359	256.423728813559	0.916769206515...	3592935	971879	419.4673123486...	458.6392251815...
17	5	17	24	24.334140435835337	13.108958837772386	216.74818401937054	254.19128329297834	0.916769206515...	3489257	978661	419.4673123486...	458.6392251815...

Figure 2: Aggregated data by TotalActiveMinutes, SleepEfficiency, DayOfWeek, and WeekNumber

4 Analysis Summary

Following data cleaning and feature engineering, an extensive statistical analysis was conducted using Python:

- Created descriptive statistical summaries for all users, including key metrics such as mean activity level, and average sleep duration.

Activity Level Summary:

	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	TotalActiveMinutes
count	12441.000000	12441.000000	12441.000000	12441.000000
mean	23.973555	17.352222	199.907081	241.232859
std	34.911291	23.020340	97.241286	107.800645
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	144.000000	193.000000
50%	8.000000	10.000000	200.000000	258.000000
75%	36.000000	24.000000	258.000000	313.000000
max	210.000000	143.000000	518.000000	540.000000

(a) Sum of features by users

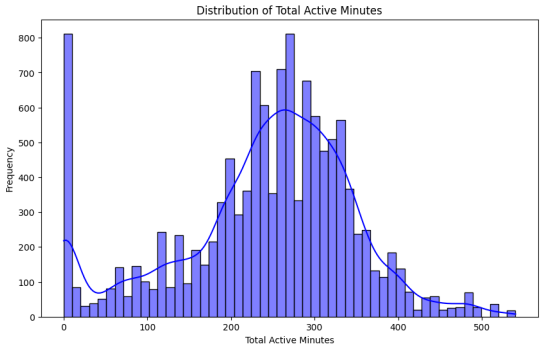
Sleep Metrics Summary:

	TotalMinutesAsleep	TotalTimeInBed	SleepEfficiency
count	12441.000000	12441.000000	12441.000000
mean	419.405996	458.360502	0.917201
std	118.643717	127.506066	0.086926
min	58.000000	61.000000	0.498361
25%	361.000000	402.000000	0.912740
50%	432.000000	463.000000	0.943231
75%	492.000000	526.000000	0.960714
max	796.000000	961.000000	1.000000

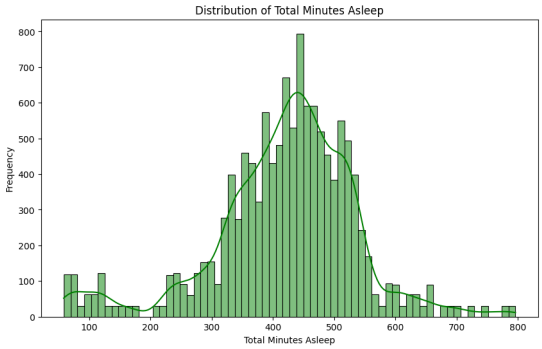
(b) Average of features by users

Figure 3

- Evaluated Distribution of Total Active Minutes and Minutes Asleep



(a) Sum of features by users



(b) Average of features by users

Figure 4

- Track User Activity across day of week

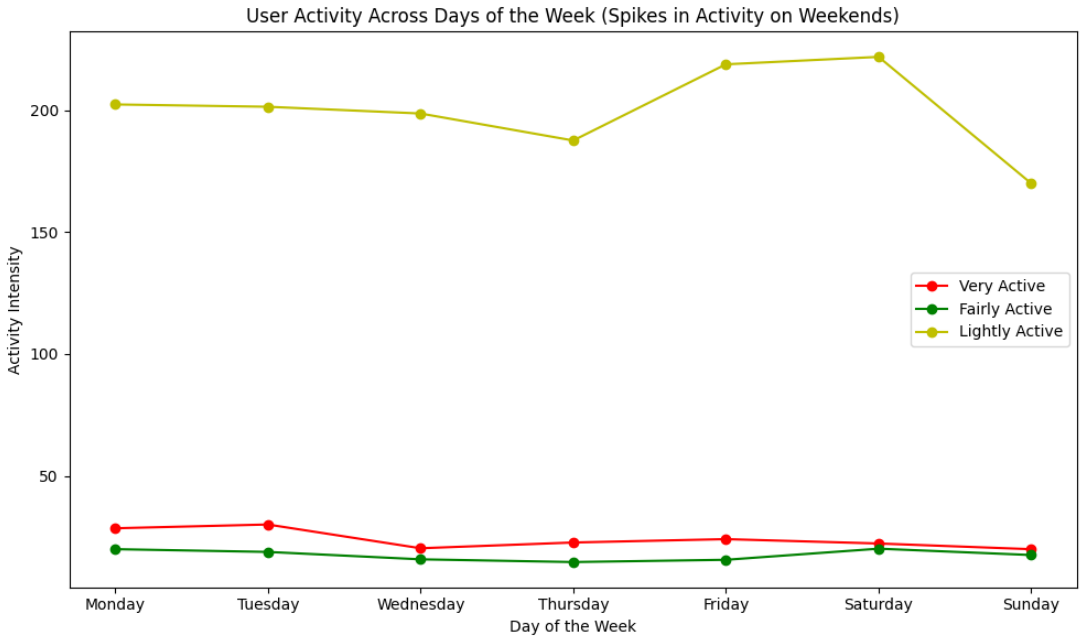


Figure 5: User Activity

Key Analysis Steps:

- Constructed correlation matrices to explore relationships between physical activity, heart rate, and sleep patterns.

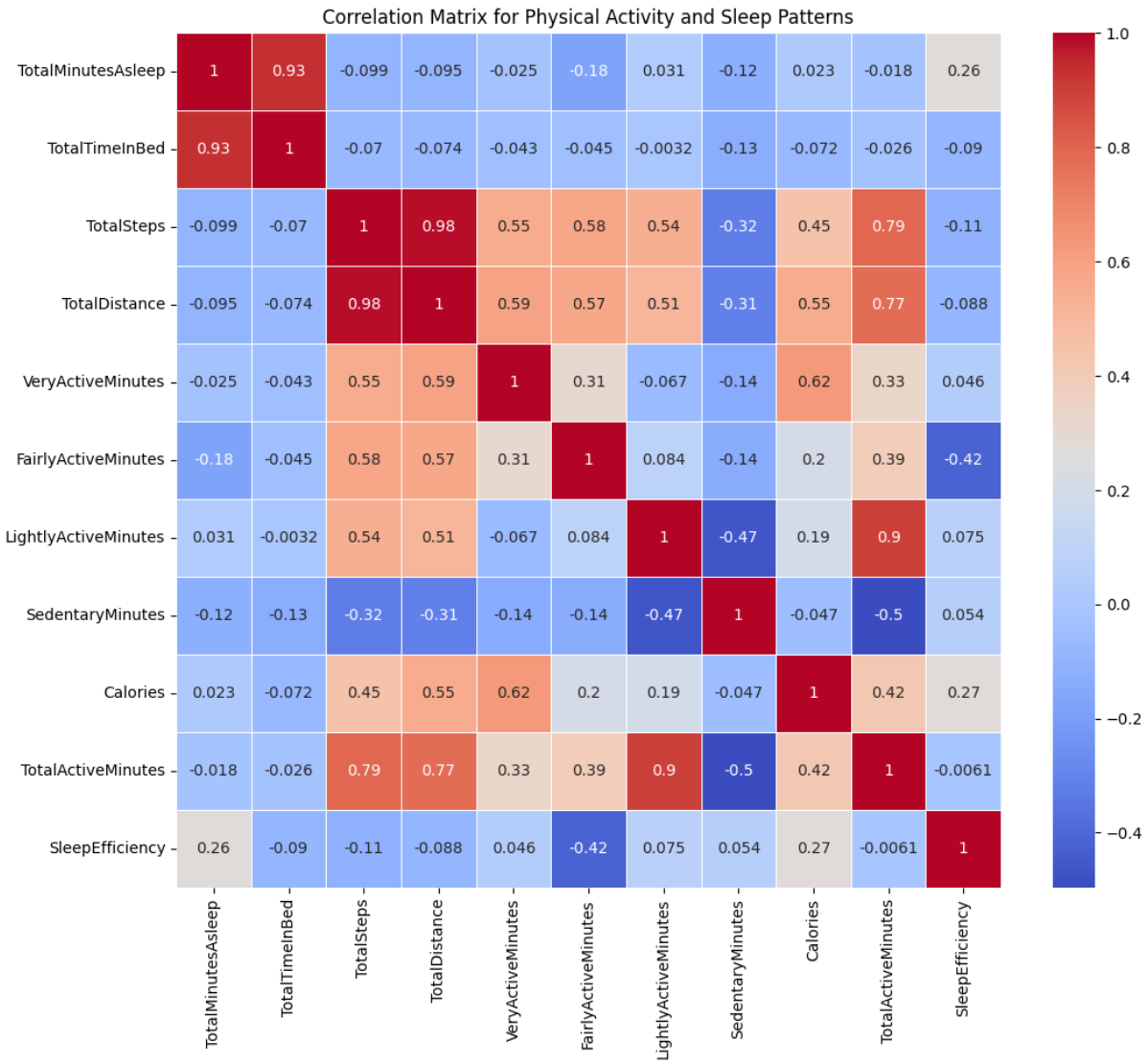


Figure 6: Caption

- Conducted hypothesis testing on whether users with higher sedentary minutes (SedentaryMinutes) have different sleep efficiencies compared to users with lower sedentary minutes? Two-Sample T-Test Results: T-statistic: 5.044741559242541 P-value: 4.605401108094 95e-07 Reject the null hypothesis: There is a significant difference in sleep efficiency between high and low sedentary minute groups.

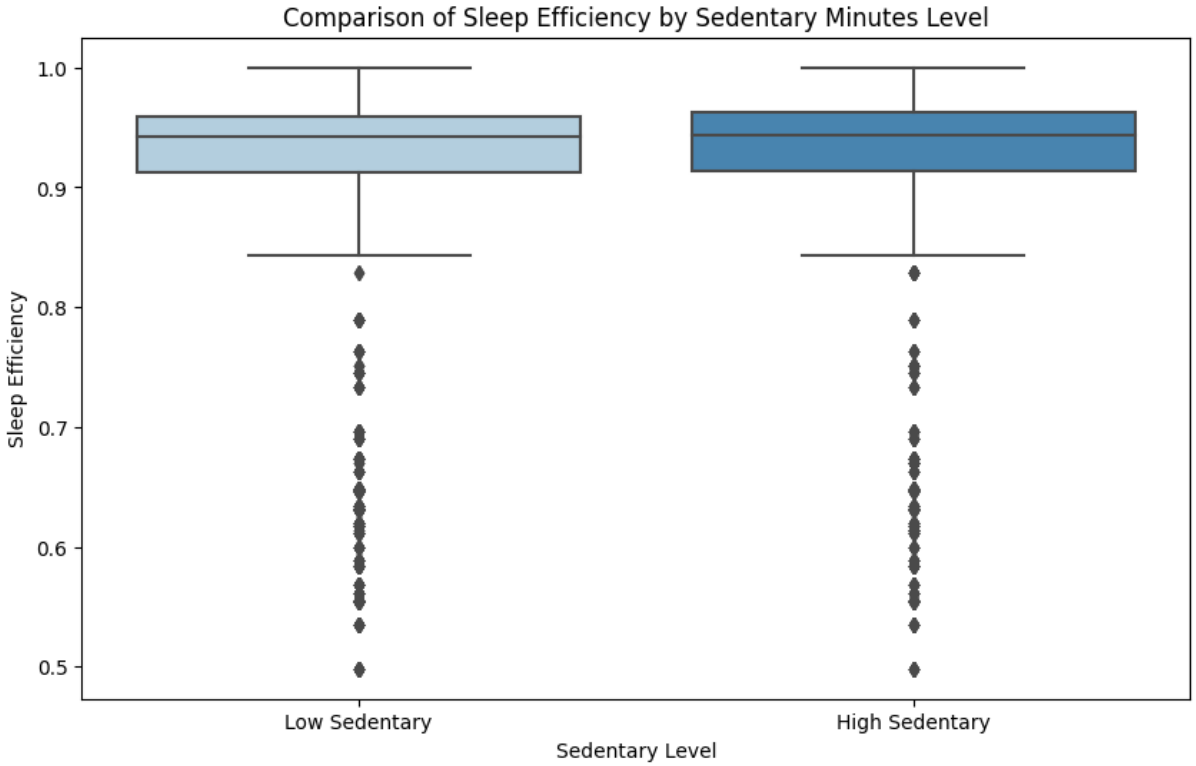


Figure 7: Caption

5 Visualization and Key Findings

The following key findings were identified from the analysis:

- **Device Usage Trends**: Most user are in serentary minutes
- Sleep Patterns: On average, Bellabeat members reported higher sleep quality scores, characterized by extended durations in restorative sleep phases. Sleep tracking was found to be most popular among Bellabeat Time users.
- Activity Level: Users who logged a higher number of daily active minutes were predominantly Bellabeat members, suggesting a strong association between physical activity levels and membership engagement.
- **Device Preferences by Time of Day**: Analysis indicated that Leaf device users are more active during morning hours, while users of the Time and Spring devices exhibit peak activity in the afternoon.

Figure 8: Dashboard

6 Recommendations

From the dataset, only 8 distinct users logged in their weight, 24 distinct users tracked their sleeping pattern and 33 distinct users tracked their activity.

- Only 24 fitbit users tracked their sleep, it is indicative that users may not feel comfortable wearing the smart device to sleep. Having a smart device which is wearable as a necklace will be a plus point. Since bellabeat has a similar product (Leaf), we should focus and market it as an aesthetic and hassle-free smart devices.
- 81notification to remind and encourage user to move around, or recommend types of activity to user.
- 8experiencing low battery level. Having a smart devices which requires short charging time and longer battery life, will encourage user to purchase and wear smart devices more frequently.

Due to the limitations of the data, we would recommend to conduct a study with a larger sample group and a more detailed data collections.