You 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 you discover will then help guide marketing strategy for the company

**Background Information:**

As a junior data analyst, you are working on the analyst marketing team at Bellabeat.

**Company Information:**

Bellabeat is a high-tech health company that manufactures health-focused smart products including but not limited to, wellness watches, smart water bottle that tracks daily water intake, and a classic wellness tracker than can be worn as a bracelet, necklace, or a clip.

They pride themselves on being a tech-driven wellness company for women.

Bellabeat is a small successful company, but they have the potential to become a bigger company in the global smart device market.

**Business Task:**

Analyze smart device data to gain insight to how consumers are using their smart devices. These insights will then help guide marketing strategy for the company.

**Data:**

1 month of data (4-12-16 – 5-12-16) was used via [FitBit Fitness Tracker Data](https://www.kaggle.com/datasets/arashnic/fitbit) (CC0: Public Domain, dataset made available through Mobius). This data set from Kaggle 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.

* Data Limitations:
  + The data set is for one month of data. This is not enough to provide a comprehensive analysis and it calls into question the credibility of the provided data.

I used pgAdmin 4 to import the data sets. SQL was then used to analyze the data.

Imported Data Sets:

Daily\_Activity

Sleep\_Log

Weight\_Log

**Data Verification:**

The data sets were first opened in Microsoft Excel to verify the initial count of rows.

I then used the COUNT SQL function in pgAdmin 4 to match the data in Microsoft Excel.

SELECT COUNT(Id) FROM daily\_activity

SELECT COUNT(Id) FROM sleep\_day

SELECT COUNT(Id) FROM weight\_log

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Data Cleaning:**

After analyzing the tables, I noticed there were some users who did not input their data (they had ‘0’ for total steps, total distance, and tracker distance). These users were deleted from the data set as they could skew the results and provide an inaccurate analysis for the client.

A FULL OUTER JOIN was performed in pgAdmin to congregate the data.

SELECT \* FROM daily\_activity

FULL OUTER JOIN sleep\_day

ON daily\_activity.id = sleep\_day.id

FULL OUTER JOIN weight\_log

ON sleep\_day.id = weight\_log.id

Graphical user interface, application, table

Description automatically generated

I filtered out the users who did not have any information for total steps, total distance, and tracker distance. Placing the count at 862.

SELECT \* FROM daily\_Activity

WHERE totaldistance != '0' AND trackerdistance != '0' AND totalsteps != '0'

SELECT COUNT(Id) FROM daily\_Activity

WHERE totaldistance != '0' AND trackerdistance != '0' AND totalsteps != '0'

Graphical user interface, text, application

Description automatically generated

The number of unique ID’s is not consistent across the three data sets which reveals that the data may not be valid for accurate analysis.

SELECT COUNT(DISTINCT(Id)) FROM sleep\_day

SELECT COUNT(DISTINCT(Id)) FROM daily\_activity

SELECT COUNT(DISTINCT(Id)) FROM weight\_log

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

Continuing to clean the data, duplicate values were checked for in each data set. I used a self-join to compare the trackerdistance values as they extend 8 decimal places, making the values less likely to be duplicated. The ID count was 33 which is the same ID count from the original daily activity data set. This confirms there are no duplicate entries in the daily activity dataset.

SELECT DISTINCT(da1.id), da1.activitydate

FROM daily\_activity AS da1

JOIN daily\_activity AS da2

ON da1.id = da2.id

WHERE da1.trackerdistance = da2.trackerdistance

GROUP BY da1.Id, da1.activitydate

ORDER BY da1.activitydate

There were duplicate rows in the sleep day dataset. I used the GROUPBY statement and DISTINCT feature in SQL to remove the duplicates, leaving 410 rows.

SELECT DISTINCT(id), sleepday

FROM sleep\_day

GROUP BY sleepday, id

ORDER BY sleepday

**Analyze:**

I utilized Microsoft excel for data analysis. I checked all data sets for minimum, maximum, average, and medium as well as outliers. The average weight was 158lbs and BMI, 25. The average total number of steps taken was 8,319 and calories burned was 2361. The average user was sedentary for 966 minutes per day which equates to about 16 hours per day. Users were lightly active an average of 210 minutes per day. BellaBeat users slept for an average of 419 minutes per night, averaging 7 hours.

Calulcations in Excel

Weight:

=MIN(weight\_log!F2:F68)

=MAX(weight\_log!F2:F68)

=AVERAGE(weight\_log!F2:F68)

=MEDIAN(weight\_log!F2:F68)

BMI:

=MIN(weight\_log!H2:H68)

=MAX(weight\_log!H2:H68)

=AVERAGE(weight\_log!H2:H68)

=MEDIAN(weight\_log!H2:H68)

Calories Burned:

=MIN(daily\_activity!O2:O864)

=MAX(daily\_activity!O2:O864)

=AVERAGE(daily\_activity!O2:O864)

=MEDIAN(daily\_activity!O2:O864)

Total Steps:

=MIN(daily\_activity!C2:C864)

=MAX(daily\_activity!C2:C864)

=AVERAGE(daily\_activity!C2:C864)

=MEDIAN(daily\_activity!C2:C864)

Sedentary Minutes:

=MIN(daily\_activity!N2:N864)

=MAX(daily\_activity!N2:N864)

=AVERAGE(daily\_activity!N2:N864)

=MEDIAN(daily\_activity!N2:N864)

Lightly Active Minutes:

=MIN(daily\_activity!M2:M864)

=MAX(daily\_activity!M2:M864)

=AVERAGE(daily\_activity!M2:M864)

=MEDIAN(daily\_activity!M2:M864)

Minutes Asleep:

=MIN(sleep\_day!D2:D414)

=MAX(sleep\_day!D2:D414)

=AVERAGE(sleep\_day!D2:D414)

=MEDIAN(sleep\_day!D2:D414)

Active Minutes:

The four active minutes groups, very active, fairly active, lightly active, and sedentary, were taken as a percentage and the results are shown below. Users were sedentary for 79.4% of the day and only lightly active for 17.4% of their total day.

Total Steps:

The chart below shows the total number of steps taken per weekday for the month of data that was provided. BellaBeat users took the most steps on Tuesdays and the least number of steps on Sundays. The data shows a spike in steps on Saturdays.

Act:

Data Analysis Conclusions:

It is recommended that BellaBeat expand globally by following the below recommendations: