ASK

Business Task:

I want to analyze Bellabeat smart device usage data to gain insight into how consumers use non-Bellabeat smart devices and apply these insights to understand the type of customers that use Bellabeat and to influence the marketing strategy to expand the business.

Stakeholders:

- Urška Sršen: Bellabeat's co-founder and Chief Creative Officer
- Sando Mur: Bellabeat's cofounder; a key member of the Bellabeat executive team
- Bellabeat marketing analytics team

PREPARE

Dataset used:

The dataset source is the Fitbit Fitness Tracker Data on Kaggle by Mobius (https://www.kaggle.com/datasets/arashnic/fitbit). It is an open-source dataset that contains data for 33 fitbit users collected in the space of 1 month; April 12, 2016, to May 12, 2016. It describes data that track fitness data such as steps walked, Intensity, activity levels and time, heart rate, and weight log.

The data is stored in both wide and long formats. The data is partially reliable because it is gotten from the readings of the different Fitbit products but from an external source from the Company. I however have the below reservations about the data

Data Bias:

- Sample Bias – I am unclear if the data is truly representative of the population
- Incomplete dataset – Some of the datasets do not have data on the entire sample example, the Heart rate dataset has data for only 7 users and the Weight log dataset contains data for just 8 out of the 30 users
- The dataset does not contain current data.
- The parameters are not properly described so I have to make assumptions as to what they mean and the units they were calculated in.

Details of the data:

Table Name	Description	Validity
dailyActivity_merged	It is a collection of the daily data of customers including steps, types of activities, calories, Distance etc.	Valid for use
dailyCalories_merged	It contains the daily calories of users	The Same data contained DailyActivity_Merged
dailyIntensities_merged	Contains the daily intensity variations of users	The Same data contained DailyActivity_Merged
dailySteps_merged	It tracks the daily steps	The Same data is contained DailyActivity_Merged
heartrate_seconds_merged	It contains incomplete data sample size	Invalid
hourlyCalories_merged	Tracks calories per hour	Valid
hourlyIntensities_merged	It Tracks intensity variations per hour	Valid
hourlySteps_merged	It counts the steps taken within an hour	Valid
minuteCaloriesNarrow_merged	Contains calories per minute in a narrow format	Valid
minuteCaloriesWide_merged	Contains calories per minute in a wide format	Same data is contained in the Narrow format.
minuteIntensitiesNarrow_merged	Contains intensity variations per minute in a narrow format	Valid
minuteIntensitiesWide_merged	Contains intensity variation per minute in a wide format	Same data is contained in the Narrow format.
minuteMETsNarrow_merged	It contains incomplete data sample size	Invalid
minuteSleep_merged	It contains incomplete data sample size	invalid
minuteStepsNarrow_merged	Contains count of steps per minute in a narrow format	Valid
minuteStepsWide_merged	Contains count of steps per minute in a Wide format	Same data is contained in the Narrow format.
sleepDay_merged	It contains incomplete data sample size sample size	Invalid
weightLogInfo_merged	It contains incomplete data sample size sample size	invalid

PROCESS

I will be using the Big Query SQL to view and clean the data available

To Load the data on Big query, I used the below schema

DailyActivity_Merged


```
[
  {
    "description": "Id",
    "mode": "NULLABLE",
    "name": "Id",
    "type": "INTEGER"
  },
  {
    "description": "ActivityDate",
    "mode": "NULLABLE",
    "name": "ActivityDate",
    "type": "STRING"
  },
  {
    "description": "TotalSteps",
    "mode": "NULLABLE",
    "name": "TotalSteps",
    "type": "FLOAT"
  },
  {
    "description": "TotalDistance",
    "mode": "NULLABLE",
    "name": "TotalDistance",
    "type": "FLOAT"
  },
  {
    "description": "TrackerDistance",
    "mode": "NULLABLE",
    "name": "TrackerDistance",
    "type": "FLOAT"
  },
  {
    "description": "LoggedActivitiesDistance",
    "mode": "NULLABLE",
    "name": "LoggedActivitiesDistance",
    "type": "FLOAT"
  },
  {
    "description": "VeryActiveDistance",
    "mode": "NULLABLE",
    "name": "VeryActiveDistance",
    "type": "FLOAT"
  },
  {
    "description": "ModeratelyActiveDistance",
    "mode": "NULLABLE",
    "name": "ModeratelyActiveDistance",
    "type": "FLOAT"
  },
  {
    "description": "LightActiveDistance",
    "mode": "NULLABLE",
    "name": "LightActiveDistance",
    "type": "FLOAT"
  },
  {
    "description": "SedentaryActiveDistance",
    "mode": "NULLABLE",
    "name": "SedentaryActiveDistance",
    "type": "FLOAT"
  },
  {
    "description": "VeryActiveMinutes",
    "mode": "NULLABLE",
    "name": "VeryActiveMinutes",
    "type": "FLOAT"
  },
  {
```

	"description": "FairlyActiveMinutes",
	"mode": "NULLABLE",
	"name": "FairlyActiveMinutes",
	"type": "FLOAT"
	},
	{
	"description": "LightlyActiveMinutes",
	"mode": "NULLABLE",
	"name": "LightlyActiveMinutes",
	"type": "FLOAT"
	},
	{
	"description": "SedentaryMinutes",
	"mode": "NULLABLE",
	"name": "SedentaryMinutes",
	"type": "FLOAT"
	},
	{
	"description": "Calories",
	"mode": "NULLABLE",
	"name": "Calories",
	"type": "FLOAT"
	}
	]

Minute Tables
 	
minuteCaloriesNarrow_merged
minuteIntensitiesNarrow_merged
minuteStepsNarrow_merged
	[
	{
	"description": "Id",
	"mode": "NULLABLE",
	"name": "Id",
	"type": "INTEGER"
	},
	{
	"description": "ActivityMinute",
	"mode": "NULLABLE",
	"name": "ActivityMinute",
	"type": "STRING"
	},
	{
	"description": "Calories",
	"mode": "NULLABLE",
	"name": "Calories",
	"type": "FLOAT"
	}
	]
	--Repeat schema for each file specifying the header names

Hourly Tables
 	
hourlyIntensities_merged
hourlyCalories_merged
hourlySteps_merged
	[
	{
	"description": "Id",
	"mode": "NULLABLE",
	"name": "Id",
	"type": "INTEGER"
	},
	{
	"description": "ActivityHour",
	"mode": "NULLABLE",
	"name": "ActivityHour",
	"type": "STRING"
	},
	{
	"description": "Calories",
	"mode": "NULLABLE",
	"name": "Calories",
	"type": "FLOAT"
	}
	]
	--Repeat schema for each file specifying the header names

Under Advanced, Header Rows to Skip was 1

- I then checked if the User Ids are the same across the three tables

Minute Tables	SELECT Id FROM `temporal-field-360018.Capstone_Project.Minute_Intensities` EXCEPT DISTINCT

	SELECT Id FROM `temporal-field-360018.Capstone_Project.Minute_Steps` EXCEPT DISTINCT SELECT Id FROM `temporal-field-360018.Capstone_Project.Minutes_Calories`
Hourly Tables	SELECT Id FROM `temporal-field-360018.Capstone_Project.Hourly_Calories` EXCEPT DISTINCT SELECT Id FROM `temporal-field-360018.Capstone_Project.Hourly_Intensities` EXCEPT DISTINCT SELECT Id FROM `temporal-field-360018.Capstone_Project.Hourly_Steps`

- I merged the data in the three tables into one using the below query:

MinuteActivity_Merged	CREATE TABLE Capstone_Project.MinuteActivity_Merged AS SELECT hc.Id, hc.ActivityMinute, hc.Steps, ROUND (ha.Calories, 2) AS Calories, ROUND (hb.Intensity, 2) AS Intensity From `temporal-field-360018.Capstone_Project.Minute_Steps`hc JOIN `temporal-field-360018.Capstone_Project.Minute_Intensities`hb USING (Id, ActivityMinute) JOIN `temporal-field-360018.Capstone_Project.Minute_Calories` ha USING (iD, ActivityMinute)
HourlyActivity_Merged	CREATE TABLE Capstone_Project.HourlyActivity_Merged AS SELECT hc.Id, hc.ActivityHour, hc.StepTotal, ROUND (ha.Calories, 2) AS Calories, ROUND (hb.TotalIntensity, 2) AS TotalIntensity, ROUND (hb.AverageIntensity, 2) AS AverageIntensity From `temporal-field-360018.Capstone_Project.Hourly_Steps`hc JOIN `temporal-field-360018.Capstone_Project.Hourly_Intensities`hb USING (Id, ActivityHour) JOIN `temporal-field-360018.Capstone_Project.Hourly_Calories`ha USING (iD, ActivityHour)

- Convert the Date loaded as “STRING” to “DATETIME” (To ensure that the date format is uniform and readable, “PARSE” the date and time)

Daily Table	CREATE TABLE Capstone_Project.DailyActivity_Parsed AS SELECT DISTINCT Id, PARSE_DATE("%m/%d/%Y", ActivityDate) AS ActivityDate, TotalSteps, ROUND(TotalDistance, 1) AS TotalDistance, ROUND(TrackerDistance, 1) AS TrackerDistance, ROUND(VeryActiveDistance, 1) AS VeryActiveDistance, ROUND(ModeratelyActiveDistance, 1) AS ModeratelyActiveDistance, ROUND(LightActiveDistance, 1) AS LightActiveDistance, ROUND(SedentaryActiveDistance, 1) AS SedentaryActiveDistance, ROUND(VeryActiveMinutes) AS VeryActiveMinutes, ROUND(FairlyActiveMinutes) AS FairlyActiveMinutes, ROUND(LightlyActiveMinutes) AS LightlyActiveMinutes, ROUND(SedentaryMinutes) AS SedentaryMinutes, ROUND(Calories, 2) AS Calories FROM `temporal-field-360018.Capstone_Project.DailyActivity_Merged`
Minute Table	CREATE TABLE Capstone_Project.MinuteActivity_Parsed AS SELECT DISTINCT Id, PARSE_DATETIME("%m/%d/%Y %I:%M:%S %p", ActivityMinute) AS ActivityMinute, ROUND(Calories, 2) AS Calories, Steps, ROUND(Intensity, 2) AS Intensity FROM `temporal-field-360018.Capstone_Project.MinuteActivity_Merged`
Hourly Table	CREATE TABLE Capstone_Project.HourlyActivity_Parsed AS SELECT DISTINCT Id, PARSE_DATETIME("%m/%d/%Y %I:%M:%S %p", ActivityHour) AS ActivityHour, ROUND(Calories, 2) AS Calories,

```
  StepTotal,
  ROUND(TotalIntensity, 2) AS TotalIntensity,
  ROUND(AverageIntensity, 2) AS AverageIntensity
FROM `temporal-field-360018.Capstone_Project.HourlyActivity_Merged`
```

- Remove duplicates By using “DISTINCT” in my queries

To Split the date from the time:

```
SELECT
  DISTINCT Id,
  EXTRACT (DATE FROM ActivityMinute) AS Date,
  EXTRACT(TIME FROM ActivityMinute) AS Time,
  Calories,
  Steps,
  Intensity
FROM
`temporal-field-360018.Capstone_Project.MinuteActivity_Parsed`
ORDER BY Id,
Date,
Time
```

```
CREATE TABLE Capstone_Project.HourlyActivity_Split AS
SELECT
  DISTINCT Id,
  EXTRACT (DATE FROM ActivityHour) AS Date,
  EXTRACT(TIME FROM ActivityHour) AS Time,
  Calories,
  AverageIntensity AS Average_Intensity,
  TotalIntensity AS Total_Intensity,
  StepTotal AS Total_Step
FROM
`temporal-field-360018.Capstone_Project.HourlyActivity_Parsed`
ORDER BY Id,
Date,
Time
```

ANALYZE AND SHARE

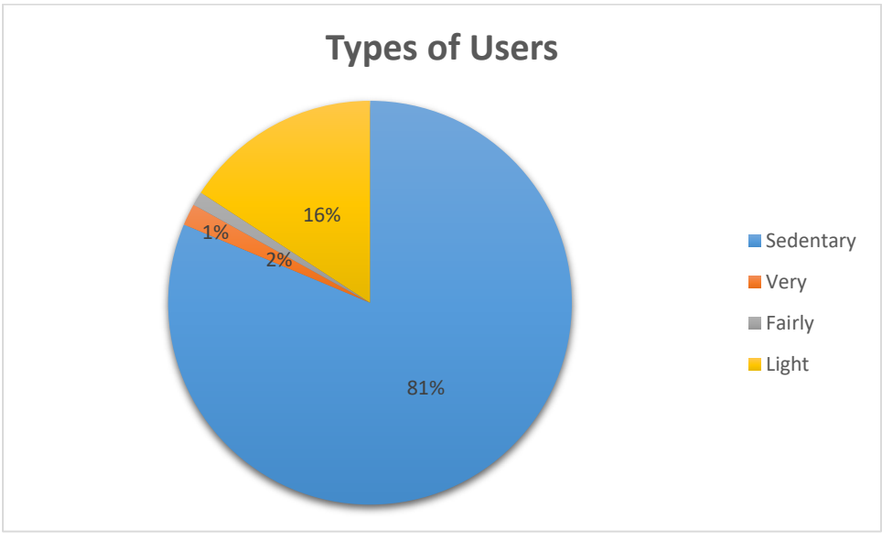
- **THE TYPES OF USERS OF BELLABEAT PRODUCTS**

  I used the below query to get the Total of users spend active from the DailyActivity Table:

```
SELECT
SUM(SedentaryMinutes)  AS SMA,
SUM(VeryActiveMinutes) AS VAM,
SUM(FairlyActiveMinutes) AS FAM,
SUM(LightlyActiveMinutes) AS LAM,
SUM(SedentaryMinutes + VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes) AS TAM
FROM `temporal-field-360018.Capstone_Project.DailyActivity_Parsed`
```

  Result:

Sedentary	Very	Fairly	Light	Total
931,738	19,895	12,751	181,244	1,145,628



Conclusion: *This indicates that 81% of the Users on Bellabeat Products are people who do not do a lot of activities that require them to move around*

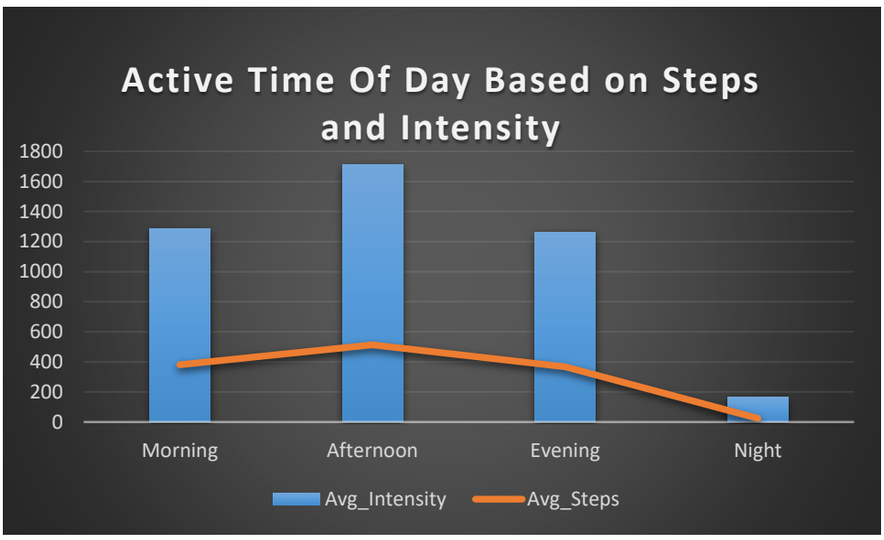
- MOST ACTIVE TIME OF DAY

To get the time of day Users are most active by analyzing the sum of the Average Intensity against the Average number of steps taken at different times of the day. Assuming that Intensity means the amount of energy used.

```
SELECT
  ROUND (SUM(Average_Intensity),2) AS Avg_Intensity,
  ROUND (AVG(Total_Step),2) AS Avg_Steps,
  CASE
    WHEN TIME(Time) BETWEEN '12:00:00' AND '17:00:00' THEN 'Afternoon'
    WHEN TIME(Time) BETWEEN '18:00:00' AND '23:00:00' THEN 'Evening'
    WHEN TIME(Time) BETWEEN '00:00:00' AND '05:00:00' THEN 'Night'
    WHEN TIME(Time) BETWEEN '06:00:00' AND '11:00:00' THEN 'Morning'
    END AS Time_of_Day
FROM
`temporal-field-360018.Capstone_Project.HourlyActivity_Split`
GROUP BY
 CASE
    WHEN TIME(Time) BETWEEN '12:00:00' AND '17:00:00' THEN 'Afternoon'
    WHEN TIME(Time) BETWEEN '18:00:00' AND '23:00:00' THEN 'Evening'
    WHEN TIME(Time) BETWEEN '00:00:00' AND '05:00:00' THEN 'Night'
    WHEN TIME(Time) BETWEEN '06:00:00' AND '11:00:00' THEN 'Morning'
    END
```

Result:

Time_of_Day	Avg_Intensity	Avg_Steps
Morning	1288.58	380.57
Afternoon	1715.28	513.44
Evening	1264.57	367.65
Night	166.19	24.23

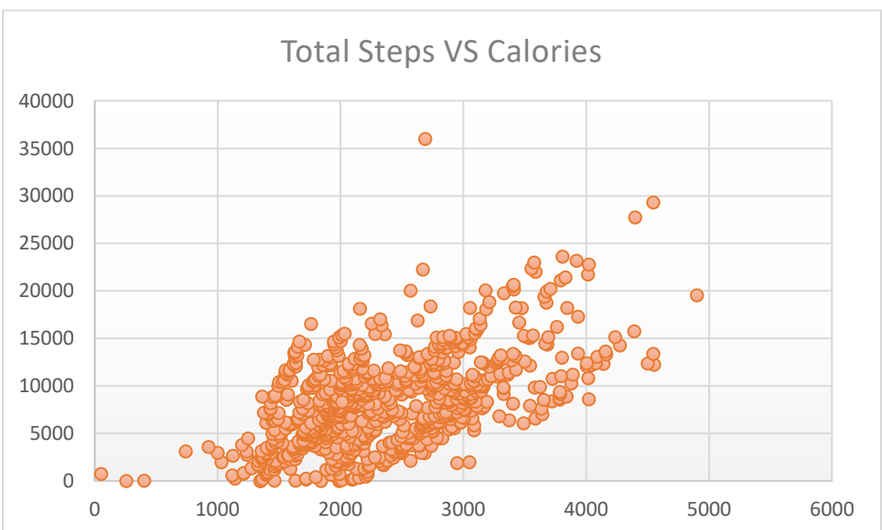


Conclusion: *Although most users are sedentary, they are most active in the afternoon time (between 12 PM and 5 PM). It is safe to assume they are people who work*

-

I queried the relationship between the total steps taken and the calories burned

```
SELECT
  ROUND(Calories,2) AS Calories,
  TotalSteps,
FROM
`temporal-field-360018.Capstone_Project.DailyActivity_Parsed`
WHERE
TotalSteps <> 0 and Calories <> 0
```



Conclusion: *There is a Positive correlation between the steps taken and the calories burned although most users are clustered less than average. I am also able to see the users who are outliers (E.g., Some users took a lot of steps but did not burn as many calories*

ACT

The below insights can be gotten from the analysis above:

- From the analysis above we discovered that most of our users are not active in terms of physical activities.
- Their most active time of the day is in the afternoon
- A relationship exists between the number of steps taken and the number of calories burned

The Marketing Team should take the below actions:

- Educate our users on the importance of physical activity on the body. This can be done through social media and communication platforms and at intervals (e.g. daily in the morning time)

- The watches can be updated to notify users to engage in physical activities when they have been sedentary for a number of minutes or hours

- The users should be able to set target steps or calories for each day and should be rewarded or celebrated when they achieve this consistently

To improve and expand my findings, the data below can help:

- The demography of the users can help understand the users (Race, Sex, etc.)
- Complete statistics of the users such as Weight, heart rate, height, age
- The full disclosure and explanation of the parameters and metrics used