ASK

- **Guiding Questions**
    - What is the problem I'm trying to solve?
        - What are some trends in smart device usage?
        - How could these trends apply to our Bellabeat customers?
        - How could these trends help with marketing?
    - How can your insights drive business decisions?
        - Can help choose a marketing technique that will encourage the right type of casual riders to get memberships. Focus on specific area, or people who are frequent riders
- **Key Tasks**
    - Identify the business task
        - Identify usage trends among other, non Bellabeat, smart device users
    - Consider key stakeholders
        - Urska Srsen - cofounder and CCO
        - Sando Mur - cofounder; on executive team (mathematician)
- **Deliverable**
    - Using user data from other smart devices, we will identify trends that will provide us insight on how people use such devices, and how this can be used to market Bellabeat devices

PREPARE

- **Guiding Questions**
    - Data stored
        - In a spreadsheet
    - Format
        - Long format
    - Bias & Credibility
        - The sample size is too small at 30. There may be bias based on location, age, and it's not an accurate representation of the total population of Fitbit users. It is also over 6 years old.
    - Licensing, privacy, security & accessibility
        - Public Domain dataset available on Kaggle
        - Removed all personally identifiable information. Individual users are given unique IDs
    - Integrity
        - Unfortunately, there is no way to verify the integrity as it was obtained by a 3rd party and is 6+ years old.
    - Answer Question

        - This is will help me answer the capstone business task by giving me data to identify trends
    - Any Problems
        - It is over 6+ years old
        - Very small sample size; large margin of error and low confidence level - bias is likely

- **Deliverable**
    - Public Domain on Kaggle - https://www.kaggle.com/datasets/arashnic/fitbit
        - Obtained through 3rd party crowd sourcing- https://zenodo.org/record/53894#.YMoUpnVKiP9
    - Data set obtained through survey via Amazon Mechanical Turk between 3/12/2016 - 5/12/2016 (2 months)
    - Total of **30 participants**
    - Data sets used
        - **dailyActivity_merged: steps, distance, varying active times and distance, calories**
        - **sleepDay_merged: sleep records, minutes asleep, minutes in bed**

PROCESS

- **Guiding Questions**
    - Tools
        - Spreadsheet - convert to uniform formatting, check for duplicates, missing data points
        - SQL - for easier analysis and practice
    - Data Integrity
        - Data integrity is unable to be ensured, but for this example case study is sufficient
- **Cleaning process**
    - Created a copy of each sheet before changing any data
    - Expanded and bolded column headers on all sheets
    - Checked for duplicates - **sleep_perday** had 3 duplicate rows
    - Converted numbers (minutes and distance) to 2 decimal point integers; rounding the data created a small difference between my calculated total and the entered total
    - Converted days to standard yyyy-mm-dd format
    - Used conditional formatting to highlight any missing data - none present
    - Ensured all IDs were the correct (10 digits in length)
    - Calculated the varying intensity distance amounts and compared it to the *TotalDistance* column data in **daily_activity**

ANALYZE

- **Guiding Questions**
  - How should you organize your data?
    - By ID and Day/Month
  - Has the data been properly formatted?
    - Yes. Each data point is the correct data type
  - What surprises did you discover in the data?
  - What trends or relationships did you find?
- **Analysis**

Sleep Data: sleeptime_day

Summary Statistics

```
SELECT
  MIN(TotalMinutesAsleep) AS min_asleep,
  MAX(TotalMinutesAsleep) AS max_asleep,
  MIN(TotalTimeInBed) AS min_inbed,
  MAX(TotalTimeInBed) AS max_inbed,
  AVG(TotalMinutesAsleep) AS avg_asleep,
  AVG(TotalTimeInBed) AS avg_inbed
FROM
  `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`;
```

	Average	**Minimum**	**Maximum**
Asleep	**419.17 min**	**58 min**	**796 min**
In Bed	**458.48 min**	**61 min**	**961 min**

410 total rows of data

Total Users - I wanted to see how many users participated in the sleep portion of the data collection

```
SELECT
COUNT(DISTINCT(Id))
FROM
  `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`
```

24 Total users; 9 less than the 33 total stated participants

Deeper Analysis

I wanted to see how many users got the recommended 8+ hours of sleep per day recorded. Only 19 users met this quota.

```
SELECT
  Id,
  COUNT(TotalMinutesAsleep),
FROM
  `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`
WHERE
  (TotalMinutesAsleep/60) >= 8
GROUP BY
  Id;
```

Sleep Data: sleep_data

My original hypothesis was that people get the most sleep on the weekends. The data generated does not support this argument.

```
SELECT
  day_of_week,
  ROUND(AVG(TotalMinutesAsleep)) AS avg_min_asleep
FROM
  (
  SELECT *,
  CASE
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 1) THEN 'Monday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 2) THEN 'Tuesday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 3) THEN 'Wednesday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 4) THEN 'Thursday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 5) THEN 'Friday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 6) THEN 'Saturday'
  WHEN (EXTRACT(DAYOFWEEK FROM SleepDay)= 7) THEN 'Sunday'
  END AS day_of_week
    FROM `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`
  )
GROUP BY day_of_week
```

Activity Data: daily_activity

Summary Statistics

```
SELECT
  ROUND(AVG(VeryActiveMinutes),2) AS avg_very_active,
  ROUND(AVG(FairlyActiveMinutes),2) AS avg_fairly_active,
  ROUND(AVG(LightlyActiveMinutes),2) AS avg_lightly_active,
  ROUND(AVG(SedentaryMinutes),2) AS avg_sedentary,
  MAX(VeryActiveMinutes) AS max_very_active,
  MAX(FairlyActiveMinutes) AS max_fairly_active,
  MAX(LightlyActiveMinutes) AS max_lightly_active,
  MAX(SedentaryMinutes) AS max_sedentary,
  MIN(VeryActiveMinutes) AS min_very_active,
  MIN(FairlyActiveMinutes) AS min_fairly_active,
  MIN(LightlyActiveMinutes) AS min_lightly_active,
  MIN(SedentaryMinutes) AS min_sedentary,
 FROM
  `bellabeat-case-study-361117.bellabeat_data.daily_activity`;
```

	Average	Maximum	Minimum
Very Active	21.16	210.0	0
Fairly Active	13.56	143.0	0
Lightly Active	192.81	518.0	0
Sedentary	991.21	144.0	0

Total Users - I wanted to see how many users participated in the sleep portion of the data collection

```
SELECT
Count(DISTINCT(Id))
FROM
  `bellabeat-case-study-361117.bellabeat_data.daily_activity`
```

33 Total users; All users participated in gathering daily activity data

Deeper Analysis

The summary stats show that the average user is mostly sedentary throughout the day. This is proven true by the results of the following query.

```
SELECT
  SUM(VeryActiveMinutes) AS very_active,
  SUM(FairlyActiveMinutes) AS fairly_active,
  SUM(LightlyActiveMinutes) AS lightly_active,
  SUM(SedentaryMinutes) AS sedentary
  FROM
```

```
`bellabeat-case-study-361117.bellabeat_data.daily_activity`
```

I also wanted to test the correlation between activity and calories lost…

```
SELECT
  Id,
  SUM(VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes) AS
total_activity,
  SUM(Calories) AS total_calories
  FROM
`bellabeat-case-study-361117.bellabeat_data.daily_activity`
GROUP BY
  Id
ORDER BY
  total_calories DESC,
  total_activity DESC;
```

….As well as steps. While there is a positive correlation between more active minutes spent and calories lost, the correlation for steps and calories is not as strong.

```
SELECT
  Id,
  SUM(TotalSteps) AS sum_steps,
  SUM(Calories) AS sum_calories
FROM
  `bellabeat-case-study-361117.bellabeat_data.daily_activity`
GROUP BY
  Id
ORDER BY
  sum_steps DESC;
```

I also wanted to see if the day of week had any effect on how active a person was. My first hypothesis was that weekends should have the greatest spurt of activity….

```
SELECT
  day_of_week,
  ROUND(AVG(VeryActiveMinutes),2) as avg_veryactive,
  ROUND(AVG(FairlyActiveMinutes),2) as avg_fairlyactive,
  ROUND(AVG(LightlyActiveMinutes),2) as avg_lightlyactive,
  ROUND(AVG(SedentaryMinutes),2) as avg_sedentary,
FROM
  (
  SELECT *,
  CASE
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 1) THEN 'Monday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 2) THEN 'Tuesday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 3) THEN 'Wednesday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 4) THEN 'Thursday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 5) THEN 'Friday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 6) THEN 'Saturday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 7) THEN 'Sunday'
  END AS day_of_week
    FROM `bellabeat-case-study-361117.bellabeat_data.daily_activity`
  )
GROUP BY
  day_of_week;
```

….As well as steps. As for activity I was incorrect, Tuesday had the most activity on average and followed up closely by Saturday. My hypothesis was confirmed as Sunday had the most steps on average, followed closely by Wednesday.

```
SELECT
  day_of_week,
  ROUND(AVG(TotalSteps)) AS avg_steps
FROM
  (
  SELECT *,
  CASE
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 1) THEN 'Monday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 2) THEN 'Tuesday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 3) THEN 'Wednesday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 4) THEN 'Thursday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 5) THEN 'Friday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 6) THEN 'Saturday'
  WHEN (EXTRACT(DAYOFWEEK FROM ActivityDate)= 7) THEN 'Sunday'
  END AS day_of_week
    FROM `bellabeat-case-study-361117.bellabeat_data.daily_activity`
  )
GROUP BY
  day_of_week;
```

I furthered my analysis by trying to find trends by combining information from both datasets.

Are users who get more sleep each night, more active than most? The trend shown on the graph shows that people who get more sleep overall, are more active.

```
       SELECT
```

```
  Id,
  SUM(TotalSteps) AS total_steps,
  SUM(VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes) AS
total_activity,
  SUM(Calories) AS total_calories,
  SUM(TotalMinutesAsleep) AS total_sleep
FROM
`bellabeat-case-study-361117.bellabeat_data.daily_activity`
INNER JOIN `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`
USING (Id)
GROUP BY
  Id
ORDER BY
  total_sleep DESC;
```

Do they take more steps per day? The data proves that this is the case.

```
SELECT
  Id,
  SUM(TotalSteps) AS total_steps,
  SUM(TotalMinutesAsleep) AS total_sleep,
  SUM(TotalTimeInBed) AS time_in_bed
FROM
`bellabeat-case-study-361117.bellabeat_data.daily_activity`
INNER JOIN `bellabeat-case-study-361117.bellabeat_data.sleeptime_day`
USING (Id)
GROUP BY
  Id
ORDER BY
  time_in_bed DESC;
```

SHARE

- **Guiding Questions**
    - Were you able to answer the business questions?
        - The business question focused on finding out how users used their smart devices. The data I obtained shows this
    - What story does your data tell?

  - The story I saw is about how active users tend to use their devices more
- How do your findings relate to your original question?
  - My findings show trends in usage, which is one of the main questions asked in the business task
- Who is your audience?
  - Urska Srsen - cofounder and CCO
  - Sando Mur - cofounder; on executive team (mathematician)
- What is the best way to communicate with them?
  - I visuals mixed with raw data (table) will be the best way in a slideshow presentation
- Can data visualization help you share your findings?
  - A data viz will make it much easier to see any trends and present the data in a clear way.
- Is your presentation accessible to your audience?
  - The font is large enough to be read, clear enough to use text-to-speech, and the colors are color blind accessible