Introduction

Bellabeat is a high-tech company that manufactures health-focused smart products for women. Bellabeat collects data on activity, sleep, stress, and reproductive health to help women monitor their lifestyle and make healthier choices.

Bellabeat has invested in traditional advertising but focuses most of their budget on digital marketing extensively. These includes investing year-round in Google SEO and being active on various social media platforms.

Bellabeat wants to analyze smart device usage data in order to gain insight into how people are already using smart devices to help guide the marketing strategy for the company.

In order to come up with a suitable marketing strategy, we need to know:

- What are some trends in smart device usage?
- How can we apply these trends to Bellabeat users to help influence Bellabeat's marketing strategy?

Preparing the data

Dataset information

- Dataset: Fitbit Fitness Tracker Data from Kaggle
- Dataset contains personal fitness tracker from 30 Fitbit users
- The dataset was generated by respondents to a distributed survey between 03/12/2016 to 05/12/2016
- Dataset contains 18 files (.csv format) with information about activity levels, calories, heart rate, MET and sleep

Sorting and filtering the data

- I will only be focusing on the daily timeframe of the data gathered
- Files being used in the analysis include:
    - dailyCalories_merged
    - dailyActivity_merged
    - dailyIntensities_merged
    - sleepDay_merged
    - weightLogInfo_merged

Processing the data

The tools used to process the data includes Microsoft Excel and Google BigQuery.

Data cleaning

- sleepDay_merged, “sleepday” and weightLogInfo_merged, “date” column was not formatted correctly, some of the data was in string format while some was in date format
- Corrected the format to date in excel
- Removed time from date using the left function in excel to extract the date only as time is not needed in the analysis

Data integrity

All relevant data is uploaded to BigQuery for analysis.

- Double checking if the daily calories recorded in “dailyCalories_merged” matches the calories recorded in “dailyActivities_merged”
  SQL query syntax:

```
SELECT daily_activity_merged.Id, daily_activity_merged.calories, dailycalories_merged.calories,
FROM
my-dataproject-1.fitbit_publicdata.daily_activity_merged
INNER JOIN
my-dataproject-1.fitbit_publicdata.dailycalories_merged
on
daily_activity_merged.id = dailycalories_merged.id and
daily_activity_merged.activitydate = dailycalories_merged.activityday
WHERE
(
CASE
WHEN daily_activity_merged.calories=dailycalories_merged.calories
THEN "true"
ELSE
"false"
END
) = "false"
```

Query returned with no results, which proves that the calories in “dailyCalories_merged” matches the calories shown in “dailyActivity_merged”.

- A quick glance on “dailyIntensities_merged” and “dailyActivity_merged” shows that they contain some similar information. A quick SQL query shows 940 observations for each. In order to confirm both datasets contain similar information, sedentary minutes and very active distance will be used in each dataset.
  SQL query syntax:

```
SELECT
  *
FROM
  my-dataproject-1.fitbit_publicdata.daily_activity_merged AS activity
INNER JOIN
  my-dataproject-1.fitbit_publicdata.dailyintensities_merged AS intensity
ON
  activity.id = intensity.id
  AND activity.activitydate = intensity.activityday
WHERE
  activity.sedentaryminutes = intensity.sedentaryminutes
  AND activity.veryactivedistance = intensity.veryactivedistance
```

Results return 940 observations, which proves that both datasets contain similar information. Thus, “dailyIntensities_merged” will be excluded as it contains lesser information.

- A quick glance on the “weightLogInfo_merged” file shows very few observations as well. Using SQL to find the number of unique IDs in the weight log info file:

```
SELECT DISTINCT
ID
FROM
my-dataproject-1.fitbit_publicdata.weightloginfo_merged
```

Results only show 8 unique IDs. Running the same query for the rest of the dataset to determine the number of participants in each dataset:

- dailyActivity_merged: 33
- dailyCalories_merged: 33
- dailyIntensities_merged: 33
- sleepDay_merged: 24
- weightLogInfo_merged: 8

According to the dataset information, there should be 30 unique IDs across all tables. However, there are 33 unique IDs spread amongst the bulk of the files. It is safe to assume there are 33 users’ data in the “dailyActivity_merged” dataset we are going to use. However, the dataset “weightLogInfo_merged” and “sleepDay_merged” are inconsistent, containing only 8 and 24 users’ data respectively. This is important to note as it will affect the analysis.

Could it be that most people using Fitbit are mostly interested in tracking their calories burnt and activity levels?

Analyzing and visualizing the data

The tools used to analyze and visualize the data are Google BigQuery and Tableau respectively.
I am interested to find out the trends and relationship between:

- Activity level and calories
- Activity level and BMI
- Activity level and sleep time

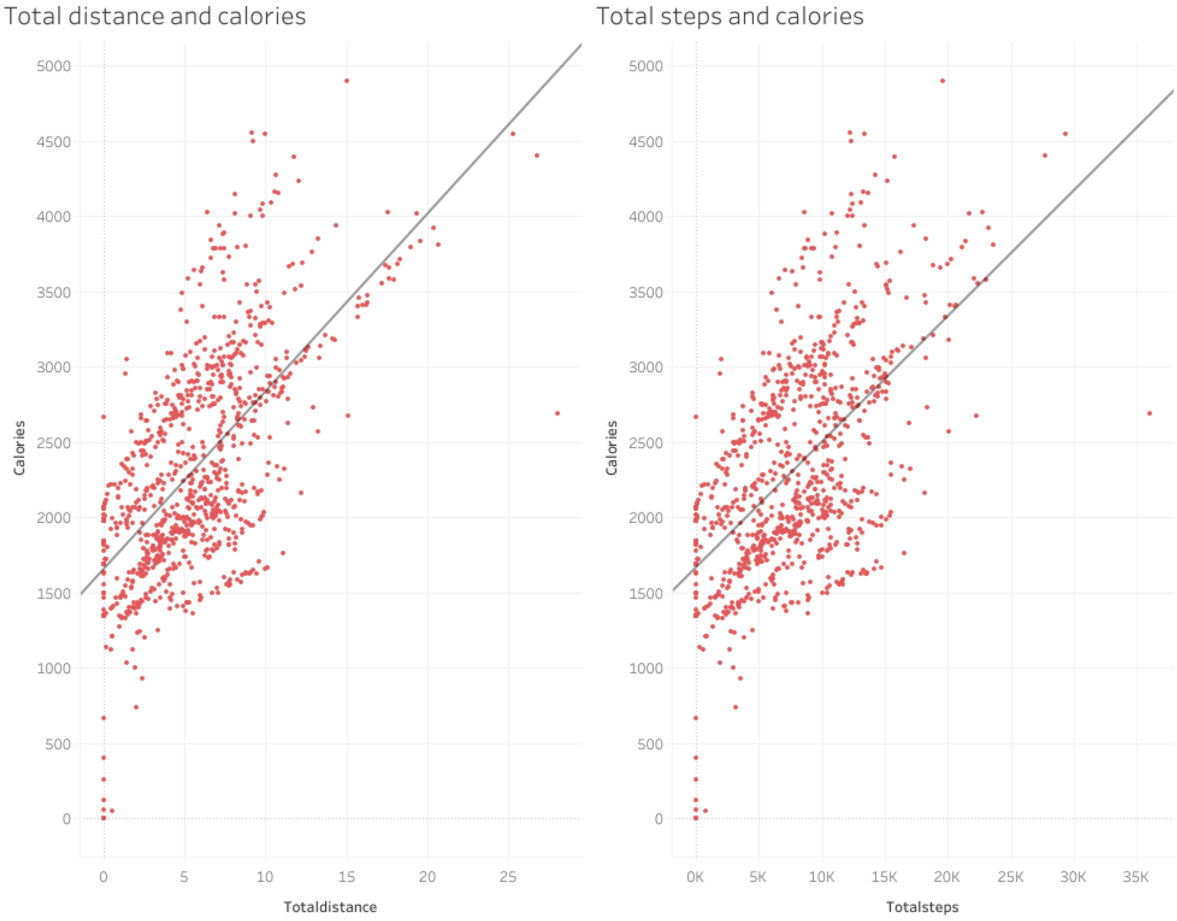
Activity level and calories

- In order to find out the relationship between activity level and calories, I will compare:
  - Total steps taken and calories burnt
  - Total distance travelled and calories burnt
  - Minutes spent on various activity levels and calories burnt

SQL query syntax:

```
SELECT
  ID,
  veryactiveminutes,
  fairlyactiveminutes,
  lightlyactiveminutes,
  sedentaryminutes,
  totalsteps,
  totaldistance,
  calories
FROM
  my-dataproject-1.fitbit_publicdata.daily_activity_merged
```

Exporting the table to Tableau:

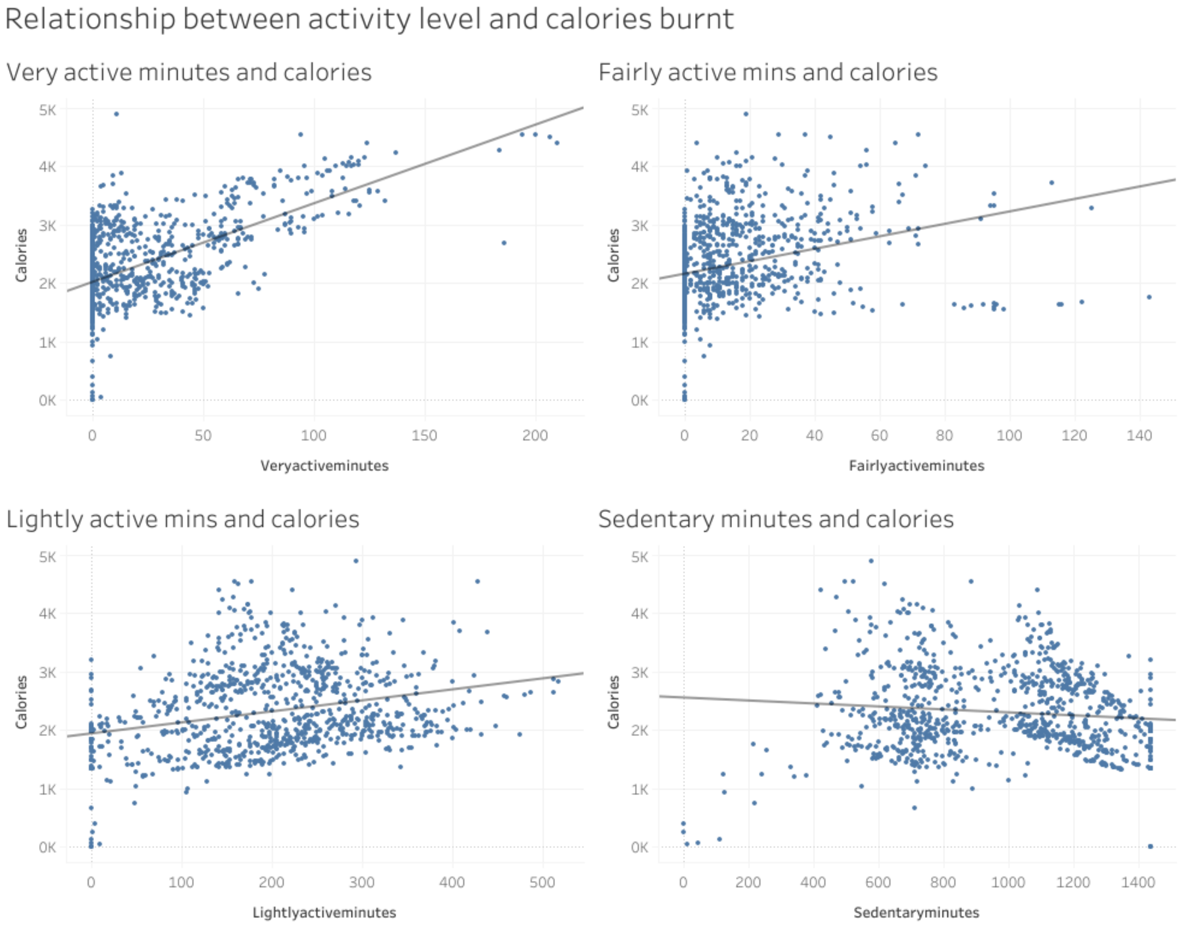


Total distance and calories

Total steps and calories

Analysis

As seen from the dashboard above, it is obvious that the more steps you take and the further the distance you travel daily, the more calories you burn.



Analysis

Unsurprisingly, the higher the intensity of active minutes, the more the calories burnt. Whereas, the more time a person spends sedentary, the lesser the calories burnt.

Activity level and BMI

BMI is used in place of weight as it is a more accurate measure. BMI takes into account height and weight and is a reasonable indicator of body fat. I am curious to know what the average BMI for Fitbit users is.
Using SQL in BigQuery:

```
SELECT
AVG(bmi) AS avg_bmi,
FROM
  my-dataproject-1.fitbit_publicdata.weightloginfo_merged
```

Analysis

The results show the average BMI to be 25.185. According to CDC, a person is considered overweight if their BMI is between 25.0 – 29.9.
Could it be that most of the people tracking their BMI are overweight because they want to lose weight? We would need a larger sample size to be more certain. (Recall the “weightLogInfo_merged” file only has 8 unique IDs)

Next, I am also curious to know how various activity levels will affect an individual’s BMI. For this, I will categorize:
Active minutes = Very active + fairly active + lightly active minutes
Non-active minutes = Sedentary minutes

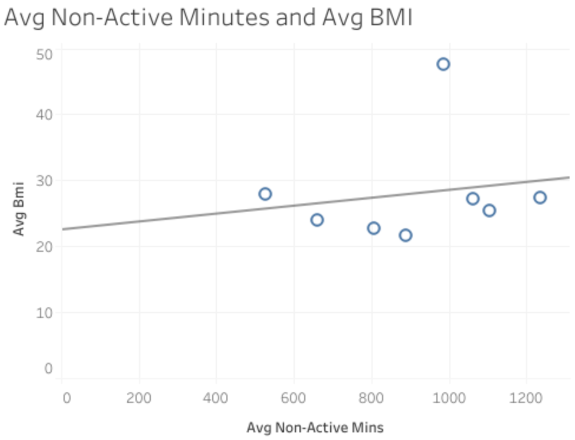
In order to identify the relationship between BMI and activity level, I will aggregate the data so that it shows the average BMI and average activity level per individual.

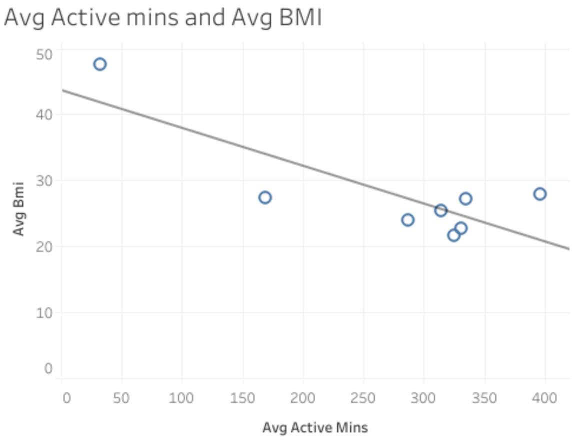
SQL query syntax:

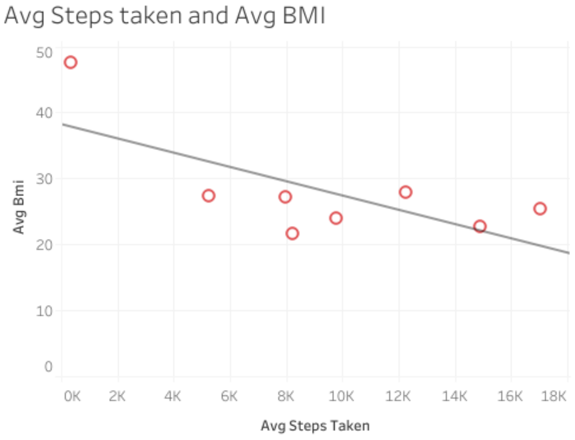
```
SELECT
  dailyactivity.ID,
  AVG(veryactiveminutes+fairlyactiveminutes+lightlyactiveminutes) AS avg_active_mins,
  AVG(sedentaryminutes) AS avg_non_active_mins,
  AVG(totalsteps) AS avg_steps_taken,
  AVG(totaldistance) AS avgdistance,
  AVG(weight.BMI) AS avg_bmi,
FROM
  my-dataproject-1.fitbit_publicdata.daily_activity_merged AS dailyactivity
INNER JOIN
  my-dataproject-1.fitbit_publicdata.weightloginfo_merged AS weight
ON
  dailyactivity.id = weight.id
  AND dailyactivity.activitydate = weight.dateonly
GROUP BY
  dailyactivity.id
```

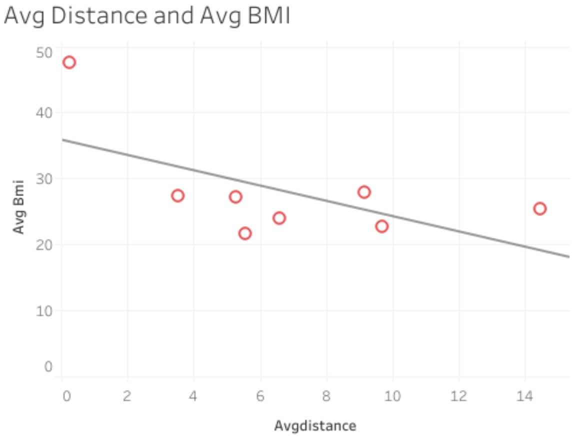
The table show the average active, non-active, steps, distance, and BMI for the 8 individuals.

Exporting the table to Tableau:









Analysis

Unsurprisingly, there is a positive correlation between non-active minutes and BMI. Which means the more time a person spends sedentary on average, the higher their average BMI is likely to be.

The more time a person spends active, the lower their BMI. Similarly, on average, the more steps and distance the individual travels, the lower their average BMI. However, the results of this analysis might not be accurate due to the small sample size (n=8).

Activity level and sleep time

In order to find out the relationship between activity level and sleep time, I will aggregate the data so that it shows the average sleep time and activity level per individual.

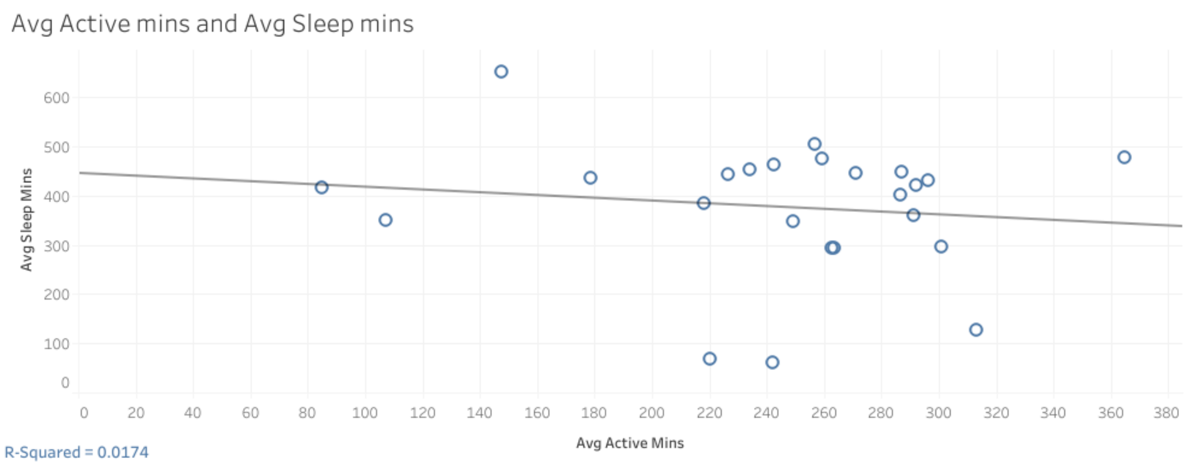
Again, categorizing the activity levels as:
Active minutes = Very active + fairly active + lightly active minutes
Non-active minutes = Sedentary minutes

SQL query syntax:

```
SELECT
  activity.id AS activity_ID,
  AVG(active_mins) AS avg_active_mins,
  AVG(sedentaryminutes) AS avg_nonactive_mins,
  AVG(sleep.totalminutesasleep) AS avg_sleep_mins
```

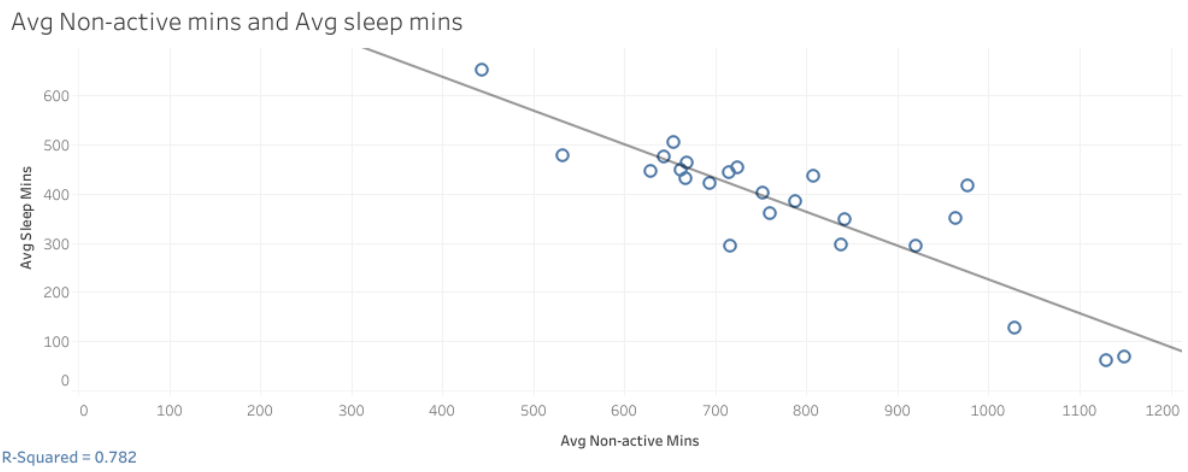
```
## making a table of of active and non active minutes to select from. Yes, the subquery is
not needed but I already had this query open in another tab and decided to copy and paste
from it instead.
FROM (
  SELECT
    id,
    (veryactiveminutes + fairlyactiveminutes + lightlyactiveminutes) AS active_mins,
    sedentaryminutes,
    activitydate,
  FROM
    my-dataproject-1.fitbit_publicdata.daily_activity_merged ) AS activity
INNER JOIN
  my-dataproject-1.fitbit_publicdata.sleep_day_merged_dateonly AS sleep
ON
  sleep.ID = activity.ID
  AND sleep.sleepday = activity.activitydate
GROUP BY
  activity.id
```

Exporting the table to Tableau:



Avg Active mins and Avg Sleep mins

R-Squared = 0.0174



Avg Non-active mins and Avg sleep mins

R-Squared = 0.782

Analysis

From the dashboard, it appears that both active and non-active minutes are negatively related to time asleep. However, slope of the regression line is way steeper in the graph which compares the average non-active minutes and average sleep minutes, and the coefficient of determination indicates a stronger association between the 2 variables that graph as well. This could imply that the more time a person spends non-active on average, the lesser time the person spends asleep. Maybe a more active person would tend to be more exhausted at the end of the day, thus they would have better quality sleep at night?

Conclusion

Trends in smart device usage:

- People mainly utilize smart devices to track their calories burnt and activity level throughout the day, while fewer people use it to track their sleep and only a **selected few** are logging their weight
- The **selected few** consists of overweight individuals. It could be that overweight people are more likely to use smart devices to track their BMI
- There is a positive relationship between time spent active, total steps taken, total distance traveled, and calories burnt
- A non-active person is more likely to have a higher BMI
- A non-active person is likely to spend lesser time asleep

Applying these trends to marketing strategies:

- Bellabeat can encourage more users to log their height and weight in order to encourage more people to take control of their health
- From the data, it is obvious that being non-active results in adverse health effects (Higher BMI, lower sleep quality). Bellabeat can send notifications to users who have a high amount of non-active minutes to encourage users to be more active
- Bellabeat can emphasize on the positive relationship between activity level and calories burnt, encouraging users to purchase their products to track their activity levels