**\*\*\*This an introduction case study of mine for my certificate. I believe I could improve on my analysis, which would lead to better sharing and acting, however it is a good start to show my potential and thought process given a raw data set on Kaggle.\*\*\***

Part 1: Ask

**Smart devices have been an upcoming, popular product for all of society to enjoy. One main purpose of those devices is tracking health qualities. Bellabeat is focused on tracking specific qualities of a person’s health, specifically women. The main topic of exploration in this case study is how Bellabeat can market better products to women in society based on recent smart device usage and trends. This analysis will provide the marketing team and co-founders a better understanding of what Bellabeat customers may invest in, and how they can market products in the future for better ROI.**

Part 2: Prepare

**I used the following dataset provided to prepare for analysis:**

[**FitBit Fitness Tracker Data (kaggle.com)**](https://www.kaggle.com/datasets/arashnic/fitbit/data)

**This data consisted of 30 people who consented to release their information from the months of march through may, with that being said, there were two folders dividing the early half and the later half. I will only be focusing on the second half, as it is more relevant to the time frame we are living in.**

**Out of this data set, I am using 10 of the 18 sheets:**

* **dailyActivity\_merged**
* **dailyCalories\_merged**
* **dailyIntensities\_merged**
* **dailySteps\_merged**
* **hourlyCalories\_merged**
* **hourlyIntensities\_merged**
* **hourlySteps\_merged**
* **minuteSleep\_merged**
* **sleepDay\_merged**
* **weightLogInfo\_merged**

**I chose these data sets, as I believe they cover the basic usage of fitness devices used by society. Some of the wide and narrow sheets seemed confusing to me and unnecessary for analysis.**

**Overviewing each dataset, each set seems to be in long format, as in there are many rows and little columns. Additionally, viewing the background of the collection of this data and the column names, bias and credibility weakness is being shown for this analysis. This is due to the unknown knowledge of gender in this data set. The company focuses on women products, and we don’t even know if this dataset is only women, or represents both genders. Additionally, while it may seem appreciative to have thirty people provide consent to release personal information, I deem it to be too small of a sample size to represent society as a whole. A larger sample set would have sufficed to represent most of society. With this knowledge in mind, we prepared our data.**

Part 3: Process

**I did all of my processing in excel.**

**This included by starting with creating separate sheets in each excel file, and named it the filename+cleaned. This is where I did all of my data cleaning to ensure I had clean data for analyzing. My cleaning process was conducted in the following way:**

1. **Sorting the data by ID number from shortest to longest**
2. **Checking for duplicates (As you can see from screenshots, I had two files that had duplicates)**
3. **Verifying formats of cells and adjusting date columns that had date and time into two different columns.**

****

****

**I then moved to SQL for analysis.**

Part 4: Analyze

**I started by adding all clean tables to BigQuery for analysis.**

****

**I loaded them under a dataset called DailyActivityData and distinguished each CSV file with a proper name. From there, I realized daily\_activity may be a combination of each daily activity from users. To do this, I used a JOIN clause to determine if all data was the same in each table.**

****

**As you can see in the bottom right corner, it pulls up the same amount of rows and same data in each sheet. I followed that same pattern for the other two sheets and ended up getting the same results, which means I can focus on daily\_activity table only.**

**From there, I pursued to see the amount of users was equal across each table within that dataset. This allows me to further confirm that the data I will be further analyzing is accurate among common users.**

****

**Using that same style query throughout the rest of the tables, I was able to disregard the last three tables of my dataset as the user amount was different from the others. From there we had 4 tables left:**

* **Daily\_activity**
* **Hourly\_calories**
* **Hourly\_steps**
* **Hourly\_intensities**

**What can we analyze with these tables to solve our problem?**

**I started out with visualizing how much sedentary time was gathered from each person within the daily\_activity dataset. As you can see in the chart below, many participants were shown to have many minutes in sedentary.**

****

**This is a pretty basic observation but, it is important to take notice of the high amounts of sit down time. I then decided to compare it with how many “fairly-active” minutes each person had total as well.**

****

**Each bar has the total amount of minutes a specific person was sitting down within that logged day, along with the active minutes. As you can observe, the average person in this set has more sit down time than active time. This is a major takeaway from the fitness device recordings.**

Part 5: Share

**Upon viewing the charts and data, it shows that many customers of the fitbit used a lot of sitting time with the device. For number wise, you have almost all of the users above at least 10,000 total minutes of rest time with most users under 1,000 minutes of being fairly active.**

Part 6: Act

**One marketing strategy may be to provide reminders to stand up, or plan workouts for the day. This may be done by sending notifications to users to become active, stand up, etc. when the condition of a user is sitting for too long in one spot. Another strategy would be a notification to plan a workout for the day, that way the fairly active minutes would increase significantly among users.**