ASK

Business Task
Analyze smart device usage data to gain insight into how consumers use other smart devices in order to find trends for Bellabeat marketing strategies.

Stakeholders
Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
Sando Mur: Bellabeat's cofounder: key member of the Bellabeat executive team
Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

PREPARE

The data used is located in Kaggle and consists of personal tracker data from 30 FitBit users, generated by respondents to a distributed survey via Amazon Mechanical Turk. The dataset is organized into CSV files according to the type of the data. Also there are two folders (each for one month). Activity, sleep and weight are some of the types of data in the folders. Some limitations are the following:

- Third-party collection: Amazon Mechanical Turk
- Old data: collected 8 years ago
- A small sample: only 30 people.

PROCESS

I chose to examine the activity and sleep CSVs and the steps I took for preprocessing and cleaning the data are the following:

- Duplicate rows Examination: I used the "TEXTJOIN" function (to concatenate all columns) in a new column. Then, I counted the number each cell appeared with the "COUNTIF" function and filtered the column to see if there were values bigger than one. After spotting the duplicates I removed them. I chose to check the duplicates first before deleting them, to evaluate them.
- Wrong Values Examination: I used the "Find and Replace" feature to check for negative values or double spaces.
- Null Values Examination: I used the "COUNTBLANK" function to calculate the number of Null Values.
- Unique IDs Examination: I used the "COUNT" combined with the "UNIQUE" function for the ID column to see how many unique IDs there are.
- Format Dates: I used a combination of formatting, power query and text to columns to properly format the dates.
- Activity and sleep data merge: I merged the two files with power query using right outer join in order to keep all sleep records matching the activity records with the primary key being the columns: id and date. Then I deleted the extra columns of IDs and Dates.

- Extra Columns: 1)Number of sleep days recorded: calculated with the “COUNTIF” function for every ID
  2)Total awake minutes: summing up the very, fairly, light active and the sedentary minutes.
  3)Total awake hours: Minutes converted to hours

File: dailyActivity

- No duplicates found
- No null values found
- No wrong values found
- Found 33 IDs instead of 30 IDs

File: sleepDay

- 3 Duplicates found and removed
- No null-values found
- No wrong values found
- 24 unique IDs found

Problems:

1. There is unequal usage of the trackers, so it is difficult to draw conclusions (for example if someone records only 1 hour from his 5 sedentary hours in a day the results will not correspond to reality).
2. In some cases, the sum of awake minutes equals 24 hours, which is likely incorrect. Additionally, the sedentary minutes seem too high, suggesting possible recording errors.

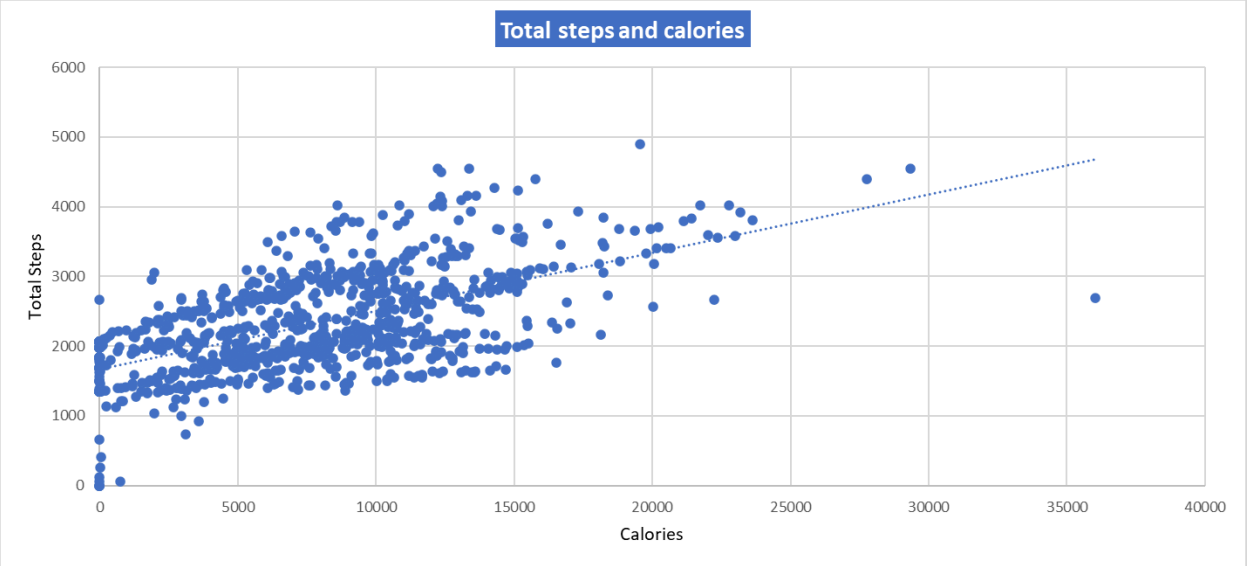
File: Merged Activity and sleep data

- I filtered out the IDs with 5 or fewer sleep days recorded. I added a column where I used the “COUNTIF” function and counted the IDs’ rows filtering out those with numbers 1,2,3,4,5.

Analyze - Share

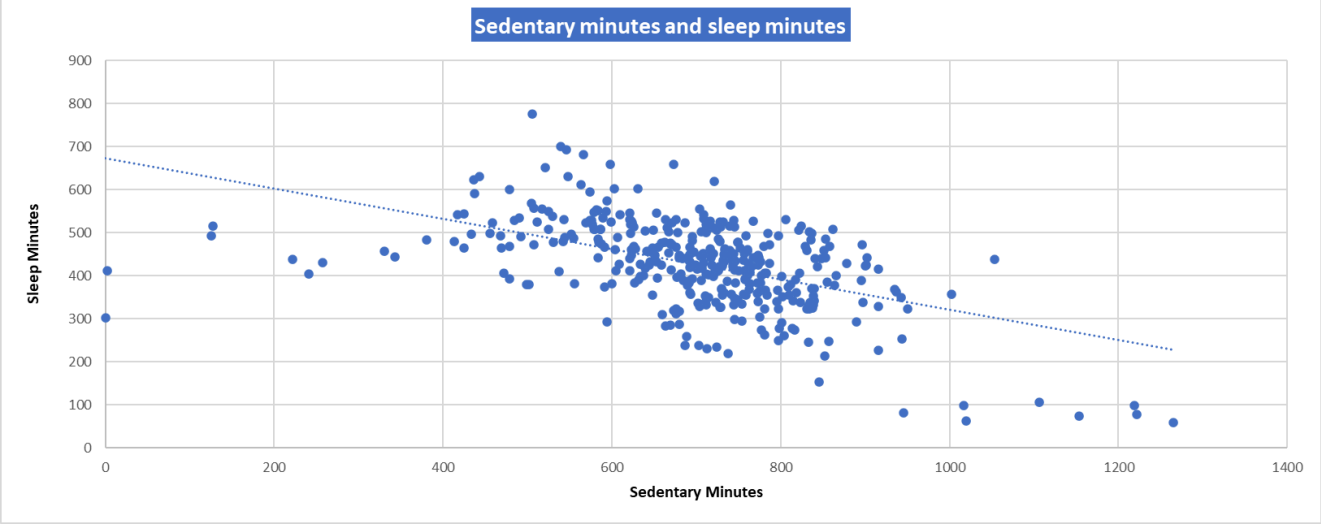
1. CORRELATION

I calculated the correlation with the “CORREL” function between total steps and calories and as expected we have a positive correlation around 0,6. The scatter plot created is the following:



As we can see, the more steps a person takes the more calories they burn.

I also calculated the correlation between sedentary minutes and sleep minutes and I found a negative correlation of around -0.6. So the more sedentary minutes a person spends a day the less sleep they get. The scatter plot created is the following:



2. Descriptive Statistics

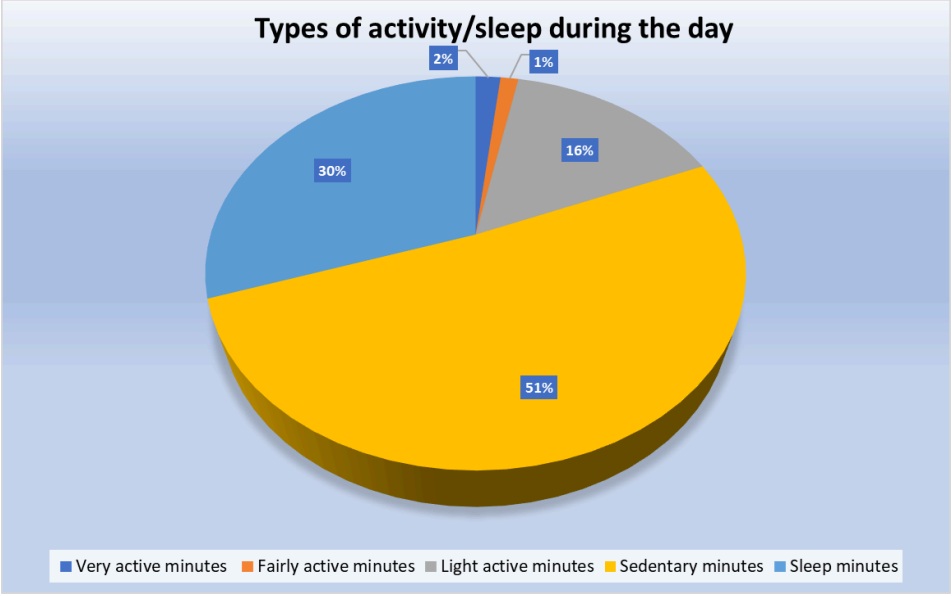
I calculated some descriptive statistics with the data analysis Excel feature and the averages that I found valuable are the following:

Sleep Data	**Activity Data**	
TotalMinutesAsleep	**Total Steps**	**Sedentary Minutes**
419.17	7637.91	991.21

- There is a very high value of the Average Sedentary Minutes. This anomaly likely stems from the absence of sleep records on certain days, which may have caused sleep minutes to be incorrectly categorized as sedentary.
- The average sleep time is a little less than 7 hours, which is a little lower than the recommendation of the CDC (7-8)
- The average daily total steps are also lower than the CDC recommendation (10000).

I also calculated the average of each type of activity and sleep recorded across IDs (in the merged file) and created a pie chart to visualize this distribution.

Very active minutes	**Fairly Active Minutes**	**Light active minutes**	**Sedentary minutes**	**Sleep minutes**
25.05	17.92	216.54	712.10	419.17



- We can see that sedentary activities take up the biggest part of a person's day (around 11-12 hours!)

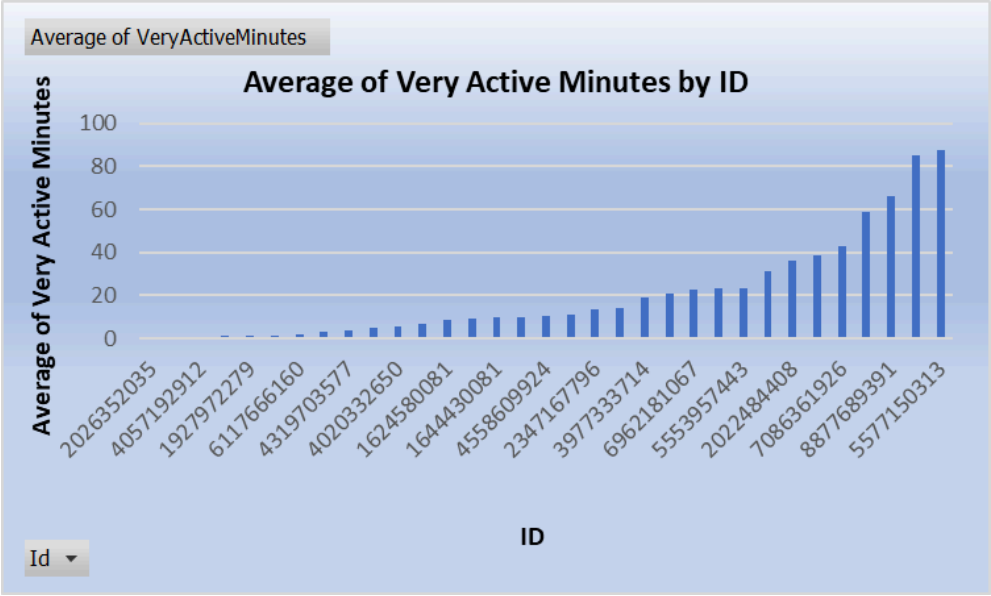
3. **Pivot Table**

I created a pivot table to see how the Average of Very Active Minutes differs across IDs.

Row Labels	Average of VeryActiveMinutes
2026352035	0.096774194
1844505072	0.129032258
4057192912	0.75
8792009665	0.965517241
1927972279	1.322580645
2320127002	1.35483871
6117666160	1.571428571
6290855005	2.75862069
4319703577	3.580645161
4702921684	5.129032258
4020332650	5.193548387
4445114986	6.612903226
1624580081	8.677419355
3372868164	9.15
1644430081	9.566666667
8583815059	9.677419355
4558609924	10.38709677
6775888955	11
2347167796	13.5
2873212765	14.09677419
3977333714	18.9
8253242879	20.52631579

6962181067	**22.80645161**
4388161847	**23.16129032**
5553957443	**23.41935484**
7007744171	**31.03846154**
2022484408	**36.29032258**
1503960366	**38.70967742**
7086361926	**42.58064516**
8378563200	**58.67741935**
8877689391	**66.06451613**
8053475328	**85.16129032**
5577150313	**87.33333333**
Grand Total	**21.16489362**

The bar chart created from the pivot table is the following:



We can see that the very active minutes differ a lot from person to person. Some people have almost zero very active minutes and some others are very active for more than an hour daily.

ACT

- Bellabeat could encourage its users to reach their daily step goal by giving them recommendations and by allowing them to create reminders.
- Bellabeat could recommend to its users to engage in more physical activity if they want to improve their sleep (and if too little sleep is detected), by giving them recommendations according to their physical condition and reminders if the sedentary minutes of the day pass a specific threshold.
- Bellabeat could create different activity/exercise plans according to the physical condition of its users and help them reach their goals with reminders and personalized recommendations. Customer segmentation is a good practice to divide customers into similar groups based on their physical condition.

All of these features will enhance customer experience and satisfaction, and if promoted in marketing campaigns, they will attract more customers.