Table of Contents

1.0 Summary

In this case study, I am going to analyze data from FitBit users, which is a personal health tracking device. The objective is to gain insights from Fitbit secondary data, to drive business decisions for another health tracker company called BellaBeat - a high-tech manufacturer of health-focused products for women. This data analysis can help guide BellaBeat's marketing strategies, particularly for two of their products Leaf (tracker bracelet) and Time (wellness watch). Their main feature is tracking and measuring user wellness activities by connecting to BellaBeat app. Then the app provides users with insights into their daily wellness, using attributes such as sleep, weight, calories burnt, menstrual cycle, and mindfulness habits. After analyzing the datasets, I will provide recommendations on how to increase BellaBeat's membership program sales. The datasets and some instructions were provided by Google Data Analytics, which is a course on Coursera, developed by Google. The datasets are not perfect and have some limitations. The purpose is not to pass over these limitations, and make it seem like the analysis is perfectly accurate, but rather to face the limitations and discuss some possible ways to avoid them in the future. These limitations commonly occur in many similar datasets in practice. Hence, it may be useful to discuss them thoroughly.

So, the business task is, essentially, to provide BellaBeat with recommendations for their digital marketing strategy. To reach that goal, the main approach used in this case study is to follow these steps: first to develop a scenario, then understand what the stakeholders would be interested in (deliverable), pre-determine some guiding questions, then proceed with a data analysis process to find answers that lead to the desired recommendations for the marketing team. Here the analysis processes involve the following phases: Data Preparation, Cleaning and Processing, Analyzing and Visualizing. This case study is divided into sections and subsections (Table of Content).

Following this summary each section can be viewed as one of the data analyses phases (in the same order), which commonly seen in practice. These phases are sometimes referred to as ask questions, prepare data, process/clean data, share and act. Bellow each figure/table there is a brief summary of insights.

1.1 Questions

1. What are some trends in smart device usage?
2. How could these trends apply to BellaBeat customers?
3. How could these trends help influence BellaBeat marketing strategy?
4. Increase BellaBeat's membership program sales by 10%

1.2 Deliverables

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top high-level content recommendations based on your analysis

2.0 Data Preparation

In this step, we need to prepare data for processing and analyzing. We will need to take a close look at the datasets, summarize them and discover some data quality characteristics. Originally there where 18 available FitBit datasets that were obtained from the **FitBit Fitness Tracker Data via Kaggle but only 6 are used for this analysis.**

2.1 Data Summary

The figure below shows some notable quantitative metadata of the 18 datasets obtained from the **FitBit Fitness Tracker Data.**

Figure 1.0: Dataset Metadata

Datasets	Variables	Num.of.Unique.Ids	Num.of.Variables	Num.of.Rows	Missing.Values
dailyActivity_merged.csv	Id, ActivityDate, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDist	33	15	940	0

	ance, SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories				
heartrate_seconds_merged.csv	Id, Time, Value	14	3	2483658	0
hourlyCalories_merged.csv	Id, ActivityHour, Calories	33	3	22099	0
hourlyIntensities_merged.csv	Id, ActivityHour, TotalIntensity, AverageIntensity	33	4	22099	0
hourlySteps_merged.csv	Id, ActivityHour, StepTotal	33	3	22099	0
minuteCaloriesNarrow_merged.csv	Id, ActivityMinute, Calories	33	3	1325580	0
minuteIntensitiesNarrow_merged.csv	Id, ActivityMinute, Intensity	33	3	1325580	0
minuteMETsNarrow_merged.csv	Id, ActivityMinute, METs	33	3	1325580	0
minuteSleep_merged.csv	Id, date, value, logId	24	4	188521	0
minuteStepsNarrow_merged.csv	Id, ActivityMinute, Steps	33	3	1325580	0
sleepDay_merged.csv	Id, SleepDay, TotalSleepRecords, TotalMinutesAsleep,	24	5	413	0

	TotalTimeInBed				
weightLogInfo_merged.csv	Id, Date, WeightKg, WeightPounds, Fat, BMI, IsManualReport, LogId	8	8	67	65

2.2 Data Limitations

Since this is a case study, and we do not have control over these limitations, we will still proceed to the analysis. But had this been a real-life project that would intend to define BellaBeat's marketing strategy, the following limitations would have been addressed before starting the analysis phrase.

- The datasets are not comprehensive. It only contains a very limited amount of data (33 unique users).
- The datasets are incomplete. Only 8 users logged weight, 12 users logged heart rate, and 24 users logged sleep entries.
- The datasets are not original. The data comes from FitBit users which can be classified as second source and may lead to inaccurate insights since user behavior and the data distribution of FitBit is not the same as that of BellaBeat.
- The datasets are not current. The dates at which the data was collected ranges from 4/12/2016 to 5/12/2016, which was about 6 years before the time of this case study.
- The datasets are not reliable. The dates at which the data was collected ranges for 30 days only. Moreover, we are expecting (30x33) 990 rows of data but got only 940 in the main dataset (Daily Dataset). These suggest that either some users did not enter the information, were not wearing the tracker or the device did not collect the data properly. These and some other complications might result in a **biased** data.

In real-world scenarios, as a data analyst, here are of the questions that we need to keep in mind to ensure that we have the best data for analysis.

- Why some users generated more data rows then others. Is it a device data collection system or did they turn them off?
- Did users voluntarily contributed data or at their convenience or were they told how often and when to use the app?
- Is it possible to know what measures were taken to eliminate the sampling bias?
- Is it possible to obtain the newer version of these datasets?
- Is it possible to obtain similar dataset from BellaBeat for originality?

2.3 Data Privacy

BellaBeat's website provides a thorough data privacy section. Also comparing with its competitors, it seems to be well developed. Just looking at the first glance it is seems that BellaBeat can collect valuable data from its users and develop marketing strategies based on the data users provide, without revealing their identity or selling to another company. It needs further research to determine the validity of these assumptions. I will leave it to the reader and in the References section provide links that might help.

3.0 Data Cleaning and Processing

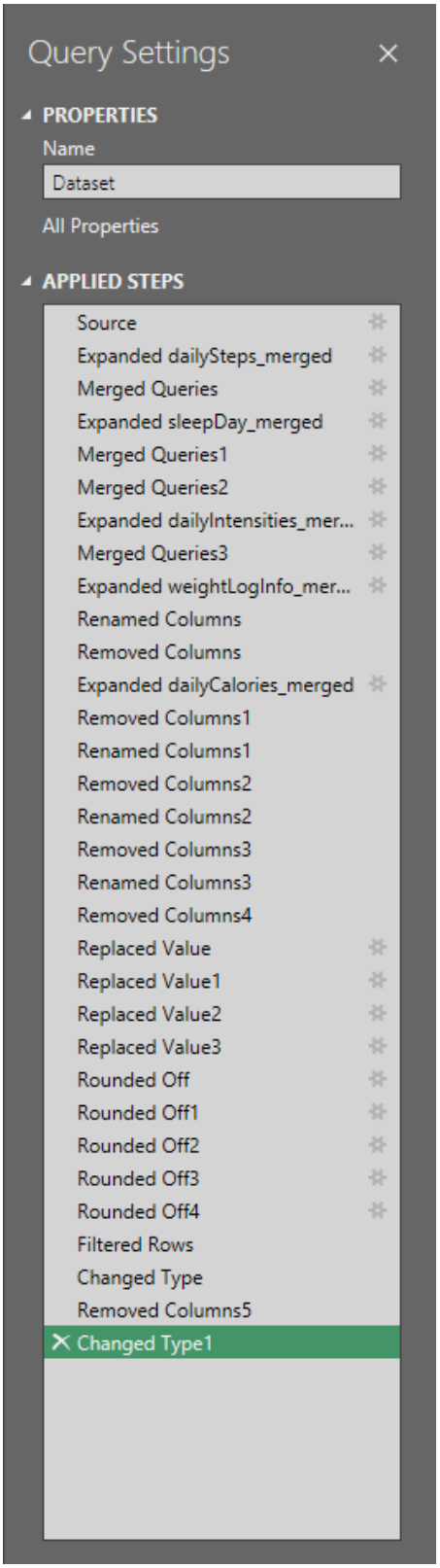
So, we already prepared ground for data processing in the data preparation section. Since we decided that we should proceed with the analysis, despite the limitations, we will start the cleaning process. Our focus here should be on discussing data integrity and some risk that may rise if it's violated. One should note that most likely some integrity issues cannot be resolved in these case study. This is because we won't be able to communicate with the stakeholders and request what we need. For instance, we found that there were only 8 users who had inputs for weight information. This is a huge problem, since it may lead to wrong conclusions. However, we will determine those issues and ignore them for learning purposes. But keep in mind that in a real-life situation, for a strong analysis, a data analyst would have to do what it takes to ensure data integrity, and only then proceed with the analysis phrase.

3.1 Steps taken for data cleaning:

The following steps are done in Big Query in Microsoft Excel as seen in **Figure 1.1.** Moreover, the entire case study from Ask Phase to Analyze Phase was done in Microsoft Excel as I want to test my knowledge with this tool. Also, I truly believe that Microsoft Excel is sufficient for the scope of this case study.

1. Duplicate datasets
2. Load and Transform
3. Clean Data
    - Duplicates
    - Null and Blanks
    - Invalid values
    - Formatting
4. Append and Merge Datasets
5. Remove unnecessary Column

Figure 1.1: Data Cleaning Steps in Excel



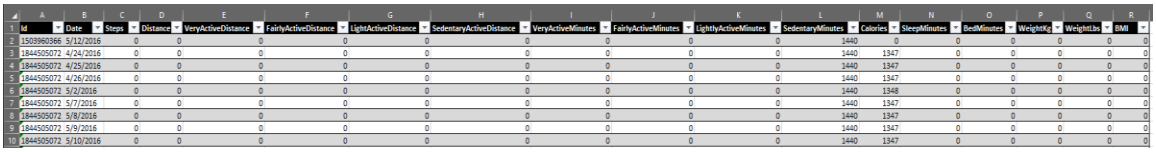
3.2 Columns and Descriptions

After the cleaning the datasets, we came up with the final dataset which is composed of 1048576 rows and 18 columns. The figures below show the column descriptions and first 10 rows of the dataset which we will be using moving forward.

Figure 1.2: Column Descriptions

Column Name	Description
Id	Unique ID per users
Date	Date of Activity
Steps	Total steps per day
Distance	Total distance per day (KM)
VeryActiveDistance	Total Very Active distance per day (KM)
ModeratelyActiveDistance	Total Moderately Active distance per day (KM)
LightActiveDistance	Total Lightly Active distance per day (KM)
SedentaryActiveDistance	Total Sedentary Active distance per day (KM)
VeryActiveMinutes	Total Very Active Minutes per day
FairlyActiveMinutes	Total Fairly Active Minutes per day
LightlyActiveMinutes	Total Lightly Active Minutes per day
SedentaryMinutes	Total Sedentary Active Minutes per day
Calories	Total Calories per day
SleepMinutes	Total Minutes asleep per day
BedMinutes	Total Minutes in bed per day
WeightKg	Total Weight in KG per day
WeightLbs	Total Weight in LBS per day
Fat	Total Fat per day
BMI	BMI per day

Figure 1.3: Cleaned Dataset



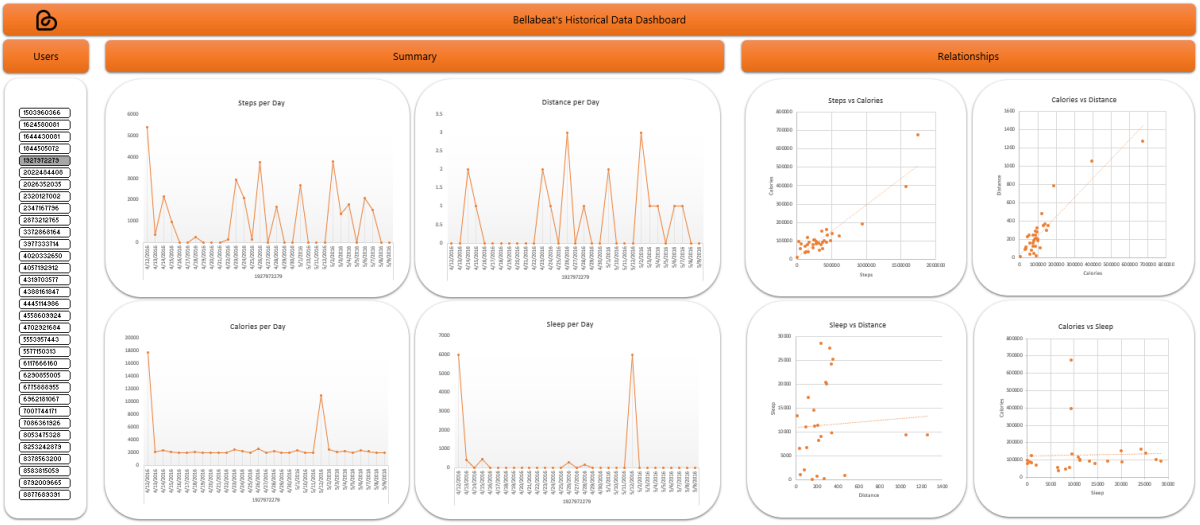
Id	Date	Steps	Distance	VeryActiveDistance	FairlyActiveDistance	LightActiveDistance	SedentaryActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories	SleepMinutes	BedMinutes	WeightKg	WeightLbs	BMI
1503960366	5/12/2016	0	0	0	0	0	0	0	0	0	1440	0	0	0	0	0	0
1844505072	4/24/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	4/25/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	4/26/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	5/2/2016	0	0	0	0	0	0	0	0	0	1440	1348	0	0	0	0	0
1844505072	5/7/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	5/8/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	5/9/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0
1844505072	5/10/2016	0	0	0	0	0	0	0	0	0	1440	1347	0	0	0	0	0

4.0 Analysis and Visualization

We already prepared data for exploration and processed it to make sure that it is clean. Now it is time for heart of the process: the actual analysis. The goal of this procedure is to identify trends and relationships within the data to answer questions that we addressed in the summary section. We will determine some relationships between different variables to see if they can lead to useful recommendations on the marketing strategy of BellaBeat.

The visualization in the form of a dashboard is made using Microsoft Excel as with previous phases. Below each figure we will provide table insights, gained using visualization tools, as well as their interpretations. Then, at the end of this document I give high level recommendation that tend to improve BellaBeat's marketing strategy.

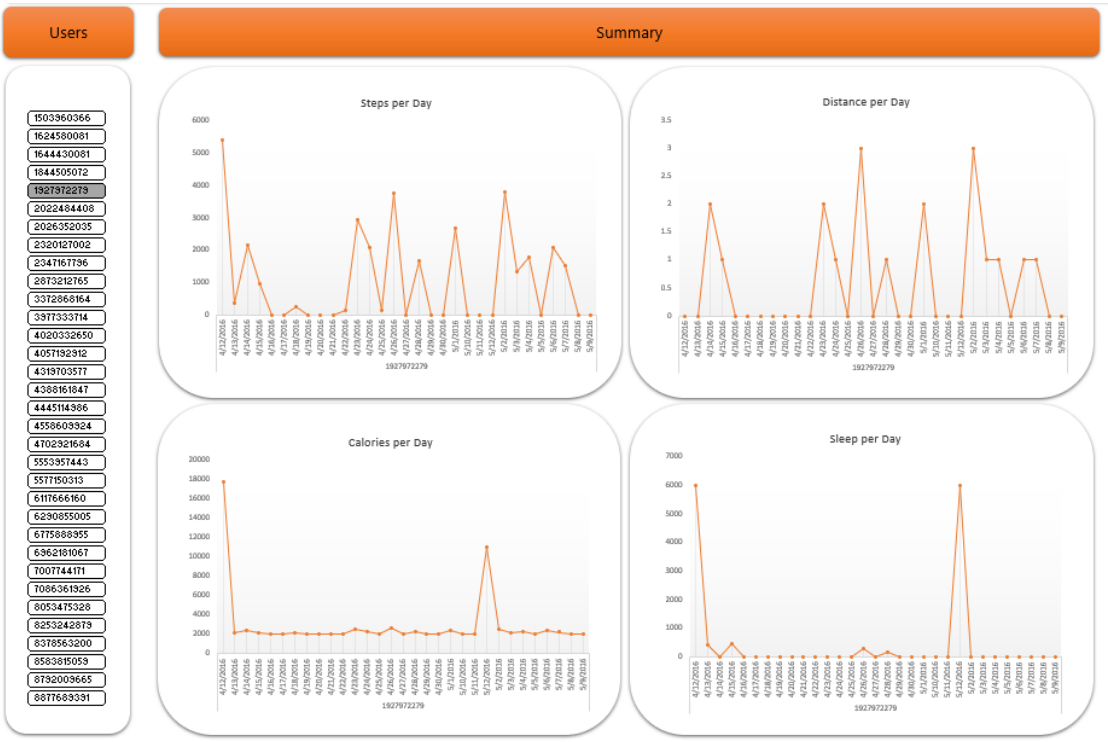
Figure 1.4: Dashboard



4.1 Summary

This section of the dashboard is interactive. We can filter the daily totals of Steps, Distance, Calories, and Sleep per user. In this way, we can have a quick glance at the user data and find patterns and trends easily.

Figure 1.5: Summary



4.2 Correlations

We already found patterns in the dataset that helped us determine some of the daily habits of the users. Now we will analyze the relationship between some of the other variables present in the dataset. Here I will use some tools from traditional methods such as regression analysis.

Figure 1.6: Correlations

Steps vs Calories

- Conclusion: Correlation - 0.91
- Correlation of 0.91 between Total number of steps and Total Calories burned
- This means that there's a strong correlation between these 2 variables.
- The insight makes sense because, in theory, any physical activities require energy and calorie is the unit of energy. Therefore, more movement (steps) means more burned calories.

Figure 1.7 - Steps vs Calories

Distance vs Calories

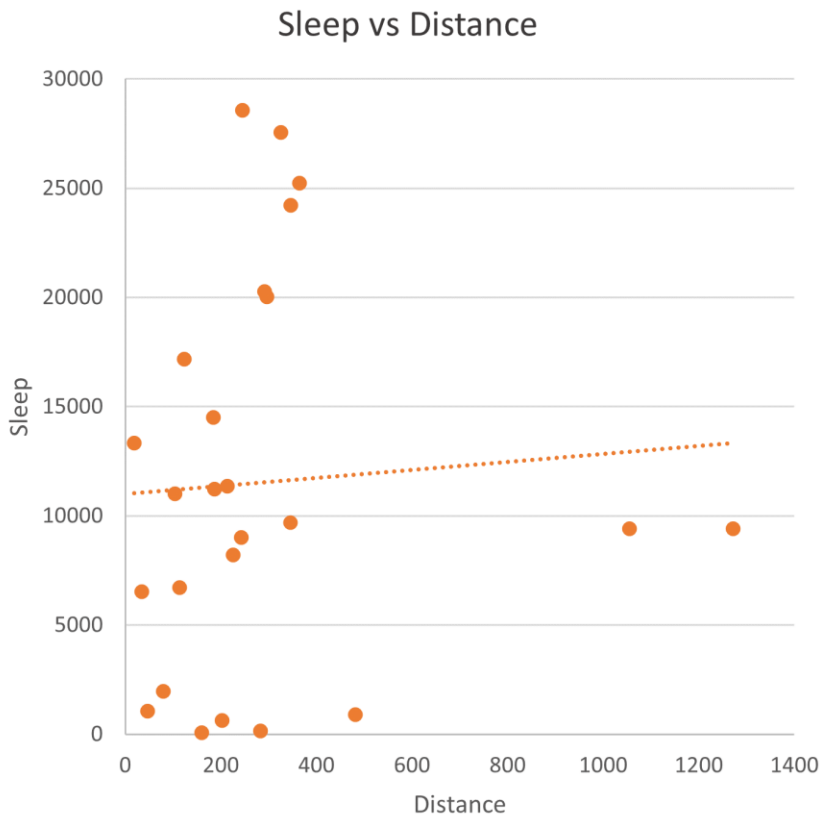
- Conclusion: Correlation - 0.91
- Correlation of 0.91 between Total number of Distance traveled and Total Calories burned
- This means that there's a strong correlation between these 2 variables.
- The insight makes sense because, in theory, any physical activities require energy and calorie is the unit of energy. Therefore, more traveled distance means more burned calories.

Figure 1.8 – Calories vs Distance

Sleep vs Distance

- Note: Due to lack of Sleep data. Following IDs have been removed: 2022484408 2873212765 3372868164 4057192912 6290855005 8253242879 8583815059 8877689391
- Conclusion: Correlation - 0.06
- Correlation of 0.06 between Total number of Distance and Total Sleep Minutes
- This means that there's a weak relationship between the 2 variables
- The insight suggests that:
    - The Total traveled distance is not related to the quality of sleep
    - Participants who sleep more minutes have fewer distance travel/steps

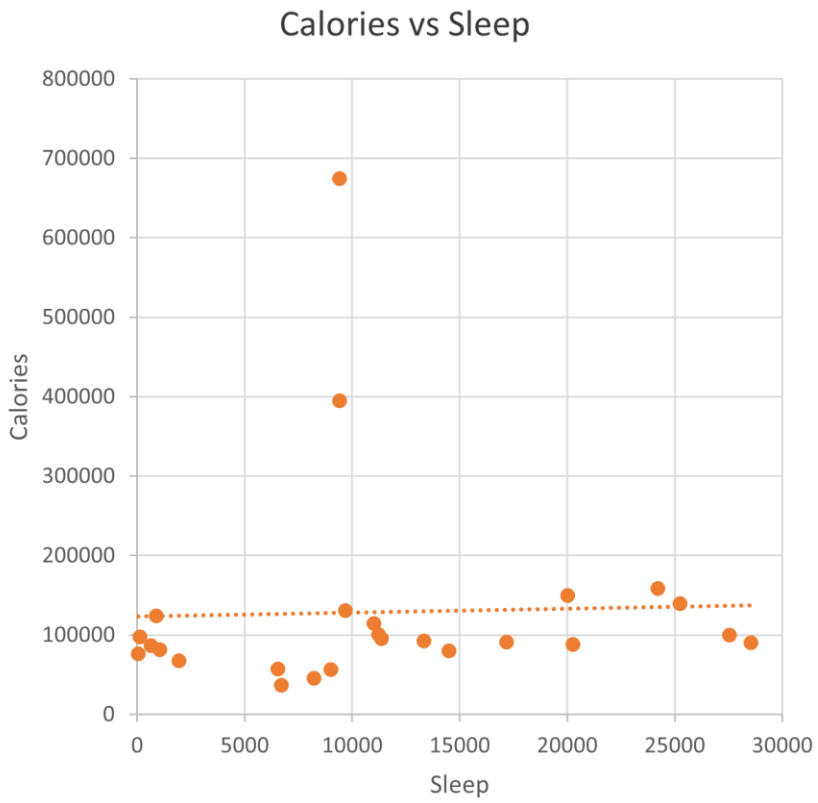
Figure 1.9 - Sleep vs Distance



Calories vs Sleep

- Note: Due to lack of Sleep data. Following IDs have been removed: 2022484408 2873212765 3372868164 4057192912 6290855005 8253242879 8583815059 8877689391
- Conclusion: Correlation - 0.03
- Correlation of 0.03 between Total number of Calories burned and Total Sleep Minutes
- This means that there's a weak relationship between the 2 variables.
- The insight suggests that:
    - Since the 2 variables have weak correlation, High sleep minutes mean low number of burned calories and vice versa.
    - Sleeping doesn't involve a lot of physical movement which supports our insight.

Figure 2.0 - Calories vs Sleep



5.0 Discussion

In this section, we will discuss the key insights from the analysis above.

Steps vs Calories

- The insight makes sense because, in theory, any physical activities require energy and calorie is the unit of energy. Therefore, more movement (steps) means more burned calories.

Sleep vs Distance

- The insight suggests that the Total traveled distance is not related to the quality of sleep of participants who sleep more minutes have fewer distance travel/steps.

Distance vs Calories

- The insight makes sense because, in theory, any physical activities require energy and calorie is the unit of energy. Therefore, more traveled distance means more burned calories.

Calories vs Sleep

- The insight suggests that since the 2 variables have weak correlation, High sleep minutes mean low number of burned calories and vice versa and sleeping doesn't involve a lot of physical movement which supports our insight.

6.0 Recommendations

- Improve application to increase customer sales.
    - Create a weekly summary of user data:
        - Distance traveled
        - Calories burned
        - Steps taken
        - Sleeping hours
    - Add a personal goal section in the application, and with this data, we can create notifications based of their output.
        - Motivate users who fail to commit to their goals
        - Appraise users who achieve their goals
        - Recommend users to avail BellaBeat membership
        -

7.0 References

Case Study Files:

https://drive.google.com/drive/folders/1F0eQFwAnDBqBrWYSKXbOs0OxOPOHz5BA?usp=share_link