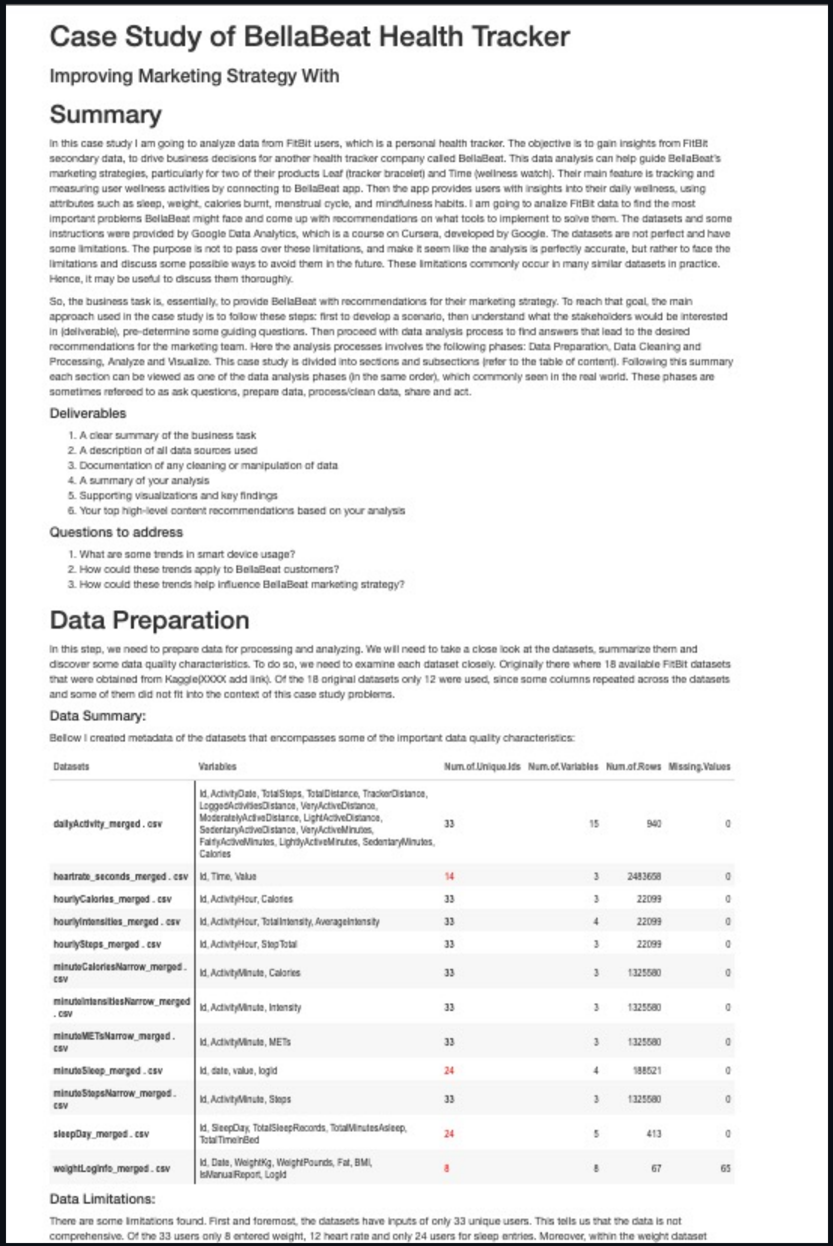


Case Study of BellaBeat Health Tracker

Improving Marketing Strategy With

Summary

In this case study I am going to analyze data from FitBit users, which is a personal health tracker. The objective is to gain insights from FitBit secondary data, to drive business decisions for another health tracker company called BellaBeat. 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. I am going to analize FitBit data to find the most important problems BellaBeat might face and come up with recommendations on what tools to implement to solve them. The datasets and some instructions were provided by Google Data Analytics, which is a course on Cursera, 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 marketing strategy. To reach that goal, the main approach used in the 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 data analysis process to find answers that lead to the desired recommendations for the marketing team. Here the analysis processes involves the following phases: Data Preparation, Data Cleaning and Processing, Analyze and Visualize. This case study is divided into sections and subsections (refer to the table of content). Following this summary each section can be viewed as one of the data analysis phases (in the same order), which commonly seen in the real world. These phases are sometimes refereed to as ask questions, prepare data, process/clean data, share and act.

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

Questions to address

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?

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. To do so, we need to examine each dataset closely. Originally there where 18 available FitBit datasets that were obtained from Kaggle(XXXX add link). Of the 18 original datasets only 12 were used, since some columns repeated across the datasets and some of them did not fit into the context of this case study problems.

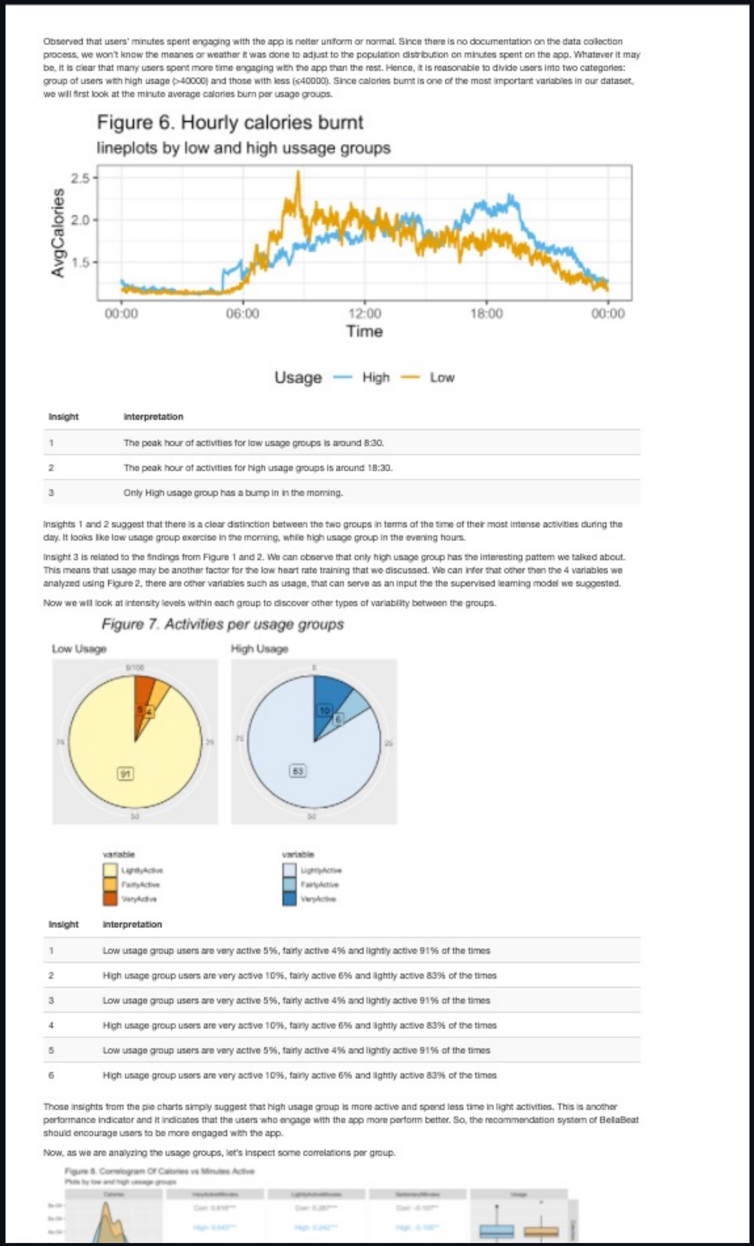
Data Summary:

Bellow I created metadata of the datasets that encompasses some of the important data quality characteristics:

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, LightActiveDistance, SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories	33	15	940	0
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, TotalTimeInBed	24	5	413	0
weightLogInfo_merged . csv	Id, Date, WeightKg, WeightPounds, Fat, BMI, IsManualReport, LogId	8	8	67	65

Data Limitations:

There are some limitations found. First and foremost, the datasets have inputs of only 33 unique users. This tells us that the data is not comprehensive. Of the 33 users only 8 entered weight, 12 heart rate and only 24 users for sleep entries. Moreover, within the weight dataset



Observed that users' minutes spent engaging with the app is neiter uniform or normal. Since there is no documentation on the data collection process, we won't know the meanes or weather it was done to adjust to the population distribution on minutes spent on the app. Whatever it may be, it is clear that many users spent more time engaging with the app than the rest. Hence, it is reasonable to divide users into two categories: group of users with high usage (>40000) and those with less (≤40000). Since calories burnt is one of the most important variables in our dataset, we will first look at the minute average calories burn per usage groups.

Figure 6. Hourly calories burnt

lineplots by low and high ussage groups

Insight	Interpretation
1	The peak hour of activities for low usage groups is around 8:30.
2	The peak hour of activities for high usage groups is around 18:30.
3	Only High usage group has a bump in in the morning.

Insights 1 and 2 suggest that there is a clear distinction between the two groups in terms of the time of their most intense activities during the day. It looks like low usage group exercise in the morning, while high usage group in the evening hours.

Insight 3 is related to the findings from Figure 1 and 2. We can observe that only high usage group has the interesting pattern we talked about. This means that usage may be another factor for the low heart rate training that we discussed. We can infer that other then the 4 variables we analyzed using Figure 2, there are other variables such as usage, that can serve as an input the the supervised learning model we suggested.

Now we will look at intensity levels within each group to discover other types of variability between the groups.

Figure 7. Activities per usage groups

Insight	Interpretation
1	Low usage group users are very active 5%, fairly active 4% and lightly active 91% of the times
2	High usage group users are very active 10%, fairly active 6% and lightly active 83% of the times
3	Low usage group users are very active 5%, fairly active 4% and lightly active 91% of the times
4	High usage group users are very active 10%, fairly active 6% and lightly active 83% of the times
5	Low usage group users are very active 5%, fairly active 4% and lightly active 91% of the times
6	High usage group users are very active 10%, fairly active 6% and lightly active 83% of the times

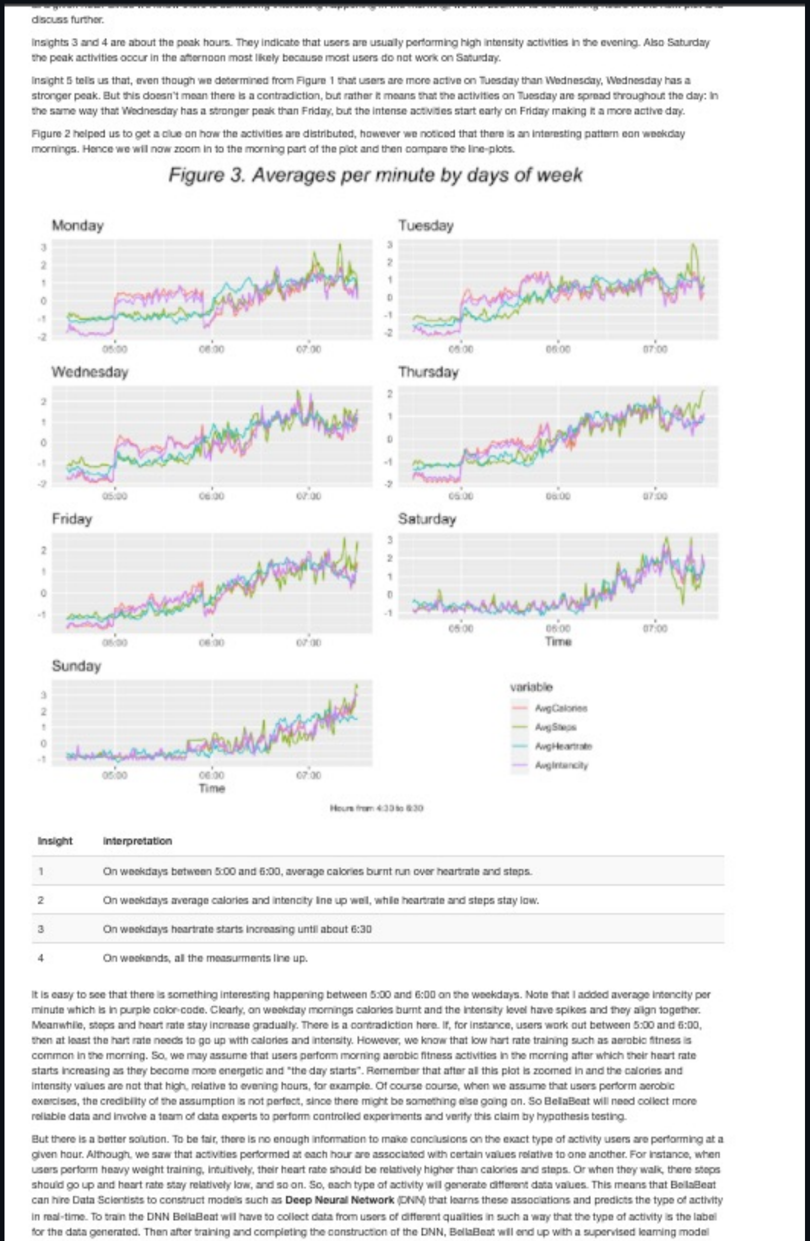
Those insights from the pie charts simply suggest that high usage group is more active and spend less time in light activities. This is another performance indicator and it indicates that the users who engage with the app more perform better. So, the recommendation system of BellaBeat should encourage users to be more engaged with the app.

Now, as we are analyzing the usage groups, let's inspect some correlations per group.

Figure 8. Correlogram Of Calories vs Minutes Active

Plots by low and high ussage groups

Calories
Usage



discuss further.

Insights 3 and 4 are about the peak hours. They indicate that users are usually performing high intensity activities in the evening. Also Saturday the peak activities occur in the afternoon most likely because most users do not work on Saturday.

Insight 5 tells us that, even though we determined from Figure 1 that users are more active on Tuesday than Wednesday, Wednesday has a stronger peak. But this doesn't mean there is a contradiction, but rather it means that the activities on Tuesday are spread throughout the day: In the same way that Wednesday has a stronger peak than Friday, but the intense activities start early on Friday making it a more active day.

Figure 2 helped us to get a clue on how the activities are distributed, however we noticed that there is an interesting pattern eon weekday mornings. Hence we will now zoom in to the morning part of the plot and then compare the line-plots.

Figure 3. Averages per minute by days of week

Hours from 4:30 to 8:30

Insight	Interpretation
1	On weekdays between 5:00 and 6:00, average calories burnt run over heartrate and steps.
2	On weekdays average calories and intencity line up well, while heartrate and steps stay low.
3	On weekdays heartrate starts increasing until about 6:30
4	On weekends, all the measurments line up.

It is easy to see that there is something interesting happening between 5:00 and 6:00 on the weekdays. Note that I added average intencity per minute which is in purple color-code. Clearly, on weekday mornings calories burnt and the intensity level have spikes and they align together. Meanwhile, steps and heart rate stay increase gradually. There is a contradiction here. If, for instance, users work out between 5:00 and 6:00, then at least the hart rate needs to go up with calories and intensity. However, we know that low hart rate training such as aerobic fitness is common in the morning. So, we may assume that users perform morning aerobic fitness activities in the morning after which their heart rate starts increasing as they become more energetic and "the day starts". Remember that after all this plot is zoomed in and the calories and intensity values are not that high, relative to evening hours, for example. Of course course, when we assume that users perform aerobic exercises, the credibility of the assumption is not perfect, since there might be something else going on. So BellaBeat will need collect more reliable data and involve a team of data experts to perform controlled experiments and verify this claim by hypothesis testing.

But there is a better solution. To be fair, there is no enough information to make conclusions on the exact type of activity users are performing at a given hour. Although, we saw that activities performed at each hour are associated with certain values relative to one another. For instance, when users perform heavy weight training, intuitively, their heart rate should be relatively higher than calories and steps. Or when they walk, there steps should go up and heart rate stay relatively low, and so on. So, each type of activity will generate different data values. This means that BellaBeat can hire Data Scientists to construct models such as **Deep Neural Network** (DNN) that learns these associations and predicts the type of activity in real-time. To train the DNN BellaBeat will have to collect data from users of different qualities in such a way that the type of activity is the label for the data generated. Then after training and completing the construction of the DNN, BellaBeat will end up with a supervised learning model