1. Introduction

This example case study was created for the Google Data Analytics Professional Certificate on Coursera. The purpose of this assignment is to reinforce the data analytics processes, practice the topics that were taught in the course, and add it to a data analytics portfolio. The main tool that will be showcased in this process is R. While Tableau and SQL were also taught in this course, there are other projects in my portfolio that would showcase these skills. Moving forward, this case study will follow the six data analytics processes; ask, prepare, process, analyze, share, and act. This process is the cornerstone of many data analytics tasks and is used in one way or another to guide the discovery process.

The scenario for this case study is to answer data-related business questions from a company called Bellabeat. Bellabeat a real company that is a high-tech manufacturer of health-focused products for women. They have provided open-source data for analysis and the case study revolves around analyzing their smart devices to unlock new growth for the company and guide marketing strategy.

2. Ask

2.1 Case Study Questions

The main concern being asked for this case study is to "analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices." Finally, only one Bellabeat product is necessary for the final presentation. Unlike a real-life business scenario, additional information can't be provided during this stage. However, the "Ask" section of the data analytics process revolves around learning more about the business task, asking measurable questions, and understanding the problem at hand.

Instead, there are key questions that are provided by Google's case study which are the following:

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?

Since these are the tasks that will guide the entire case study, there should be careful consideration and planning for each question. First, finding trends in smart device usage is a data visualization task which can be used to detect trend lines or other correlations. Since the data is not available at this point, the key trends and features are not yet apparent. But, these trends need to stay tuned with the business task. Second, the trends that are found need to also be applied to the customers themselves. Finally, the trends need to be actionable and must be related to influencing Bellabeat's marketing strategy.

2.2 Deliverables

To successfully complete the case study, a list of deliverables need to be finished. The following chart are the deliverables being asked by Google.

Table 2.1

Project 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

2.3 Business Task

The problem that this data analytics process is attempting to solve it the discovery of actionable and meaningful trends that can impact customers and assist the marketing department. During the analytics process, the complete exploration of data is necessary to find trends that can be impactful and useful.

3. Prepare

The next stage of the process is the prepare stage. In this stage, data collection and a high-level understanding of the data is performed here. By the end of this phase, analysts should have the data they need to solve their business problem and understand how that data is organized. Preliminary exploration is done here to determine how credible the data is as well as determine whether the data has all the necessary features to answer the business task.

3.1 Data Source Description

The data that is recommended by Google is hosted on Kaggle. Kaggle is a popular data repository that houses a wide variety of datasets for public use. The specific dataset that will be analyzed is the FitBit Fitness Tracker Data dataset shared by Möbius (2020). According to its Kaggle page, this dataset was collected for a total of two months in 2016 and tracked the activities of thirty consenting Fitbit users. The dataset is also available on the file share and research website, Zenodo, where Robert Furberg and others were responsible for conducting research for this dataset under the Amazon Creative Commons license.

This dataset is a very popular dataset for physical fitness monitoring, and had been widely accepted as open-source dataset for multiple research papers. It is likely that this dataset is credible as it is used in research publications and approved by Amazon. While the dataset has data collected from 2016, there isn't much change in the advancement of human beings that this dataset would be outdated. A special consideration can be made on the technology and accuracy of how the data is collected.

3.2 Dataset Structure

The data is organized into 18 csv files, where some have repeated information. The file names indicate the topic of what is being stored. For example, "daily" contains daily information about each participant. Most are stored in a long format where each participant has their own entry for specific time slots. For example, for each day there are 20 unique entries because there are 20 participants.

Not all files have important or unique information that can be used for analysis. Using Google Sheets, a high-level review of each dataset reveals that not all datasets should be uploaded into R. Firstly, the minute-granularity datasets contain a wide version which splits each row into 60 minute columns. This format is unnecessary as the long format is similar to the other datasets and would facilitate easier comparisons. Finally, the minute granularity has more fine tuned information, but doesn't seem necessary for an analysis on trends. Table 3.1 outlines the files that will be primarily analyzed in R as well as a short description of each.

Table 3.1

Dataset Review

File Name (.csv)	Description
dailyActivity_merged	This file contains daily steps, distances split into into how much exercise was involved, and how long activity is recorded for that day. This file merges information from 3 other files.
hourlyCalories_merged hourlyIntensities_merged hourlySteps_merged	These three files are combinable that have the total amount of calories burned, steps, and average intensity for that hour.

sleepDay_merged	This file contains daily information for the amount of minutes of sleep time a user had for that day
weightLogInfo_merged	This file contains logged information on a participants weight and BMI. Only two participants logged their information more than five times. And only seven logged their information even once.

After reviewing the dataset, only a select number of files would likely be important for the analysis. While the others will be reserved if this subset of the data does not answer the business problem.

3.3 Data Integrity

After exploring the datasets, there is areas where data integrity could be improved. For instance, there are a lot of unrecorded values for each hour. Some datasets don't include all days and some users have more records than others. The data integrity attached to the data itself seems decent as all values of calories burned, time asleep, and other measurable variables are within range of normal behavior. Although there are some instances where zero steps and activity are recorded in a day. So, this will need to be explored further and addressed during the data cleaning step. Overall, the data is usable to explore trends.

4. Process

The “process” stage of data analysis revolves around exploration and data cleaning steps to ensure your dataset is ready for analysis. This step will give a comprehensive overview of all the data and start to give the analyst deeper insights and exploratory ideas for the analysis phase as well. In R, this step will include creating summary statistics, removing outlier, and basic visualizations for the purpose of exploration. At this point, analysts should also understand the steps to should be taken to analyze the data as well.

R is a great tool for this step because it will help with data integration as well as easily create quick charts to find areas where data cleaning is needed. R can also perform those data cleaning tasks if required.

The first step in the data cleaning process is to find summary statistics and look at some sample data to determine the steps required to clean the data. Since there are four total datasets, we can outline the cleaning and data processing steps taken for each one.

4.1 Daily Activity Data

Using R, there are a few oddities with this table as the minimum values for all values in the dataset, including steps taken and distance, are zero. It is unlikely that someone had taken zero steps and had zero active minutes, so upon further investigation there were 77 days where values were not recorded. This could be due to the device being charged or not worn. But, since these values wouldn't be helpful for analysis they were removed.

Secondly, the columns "TotalDistance" and "TrackerDistance" are nearly identical with only 15 rows where they were not. These are also the only rows where the "LoggedActivitiesDistance" column is not zero. However, adding the distances from "VeryActive", "ModeratelyActive", "LightActive", and "SedentaryActive" seems to equal the "TotalDistance" column. So we can remove "TrackerDistance" and "LoggedActivitiesDistance" from the analysis as they could be redundant.

And finally, the "Calories" column doesn't have enough meaning, so it should be changed to "CaloriesBurned". And, the date column is a string value which is harder to interpret for R. So, using the Lubridate package, all tables can be be converted into a similar date and time format for further analysis.

Table 4.1 outlines the cleaning changes done to the Daily Activity data.

Table 4.1

Daily Activity Changes

1. Removed 77 rows where only zero values were recorded
2. Removed “TrackerDistance” and “LoggedActivitiesDistance” columns
3. Renamed “Calories” column to “CaloriesBurned”
4. Converted the “ActivityDate” column into a different date format

4.2 Hourly Activity Data

The hourly activity data is separated into three files, so the first step to processing this is to combine the three files. Then, using R’s skim and head functions, it was clear the the “TotalIntensity” column was the “AverageIntsensityColumn” multiplied by 60. So, since this is an hourly dataset with little need to convert to intensity per minute, the “AverageIntsensityColumn” was removed. Then, this dataset also had a “Calories” column that needed to be renamed to “CaloriesBurned”. Table 4.2 outlines these changes.

Table 4.2

Hourly Activity Changes

1. Combine the three files into one dataset
2. Removed “AverageIntsensityColumn” columns
3. Renamed “Calories” column to “CaloriesBurned”
4. Converted the “ActivityHour” column into a different date time format

4.3 Daily Sleeping Data

When compared to the other daily dataset, the Daily Sleeping data has significantly less rows. There are only 24 unique Id values compared to the 33 unique values in other tables. Upon further investigation, it was found that only 15 of those Ids recorded more than ten days. Combining the two datasets with most users not recording their sleeping times could result in trends that don’t capture the whole picture.

Furthermore, a new column can be created using the “TotalMinutesAsleep” and “TotalTimeInBed”. By finding the difference between these two, a new calculated feature can be created that measures minutes a person is restless in bed. This could lead to new insights not already present in the data.

Table 4.3

Daily Sleeping Changes

1. Remove rows where the user has ten or less records.
2. Create a “TotalTimeInBed” column
3. Converted the “SleepDay” column into a different date time format

4.4 Daily Weight Log Data

Unfortunately, this dataset is even less represented than the daily sleeping data. Only two users have over ten records for recording weight and BMI. Since there were only two participants in this data, it would be challenging to create an unbiased trend using only two samples. As such, this table is no longer being considered for analysis.

5. Analyze

In the analyze phase, visualizations and other data calculations are made in an attempt to dive into the data and find insights that are not readily apparent. Likely, there will be a lot of insights and trends to be found, but analysts should still focus on their business task at hand. In this section, answers to the business problem should be found and discussed while making note of anything that could be related to the task. Unlike the processing phase where general cleaning steps are taken to organize and ensure the data is usable, the analyze phase aggregates and filters data to find trends.

5.1 Missing Rows and Interactivity

Many of the datasets have areas where users interactivity with the features are lacking. There are many days where sleep data is not record and only two users interacted with the weight and BMI tracking features. This may be due to many reasons. One possible reason for this behavior is that users are using night time to charge their devices or don't wish to wear the device at night. For the lack of weight records, many users likely don't have time or wish to input their weight every day. Since these are trends and recommendations are for the marketing department, this analysis need to find a way to solve these issues. One solution is to find the day and time where users are least active to allow for charging. According to the data, Sundays are the days with the least average amount of steps.

On top of this, searching for the least active hours is necessary. By averaging the amount of steps for all the users and grouping them into each hour of the day, the least amount of steps occur in the middle of the night of course. However, that's not very helpful to increase interactivity. So, by examining the change in averages in each hour, there is a large decrease of steps taken at 7 AM, 3 PM, and 8 PM. These are the times where users are still likely awake as well.

5.2 Minutes Restless Feature

In the processing phase, a column that measured the amount of time a user was restless in bed was created by subtracting time in bed and time asleep. Multiple visualizations were used to test its correlation with features such as "TotalSteps", "TotalDistance", "SedentaryMinutes", “CaloriesBurned”, and more. However, none of these other columns seemed to reveal any trends with this calculation. Unfortunately, this wouldn’t turn into a useful analysis.

5.3 Type of Exercise and Calories Burned

The dataset measures activity into one of four categories, sedentary, lightly active, fairly active, and very active. Initial analysis attempted to discover trends with intensity of an exercise and the amount of calories burned. However, no trends were detected. Instead of looking at these four categories separately, combining them into two categories may reveal additional insights. By adding the values of lightly active, fairly active, and very active into a similar category of “active”, they amount of categories is reduced to two.

Finally, trends between calories burned and average active minutes are seen. There is a positive correlation between being active and burning calories as most would expect. As a specific calculation, people burn 1000 more calories in a day if they go from 200 minutes of activity to 400 minutes. The type of activity is much less important, as long as you are not at rest.

6. Share

This share section will be in the form of screenshots to convey what is being shown as well as how the visualizations relate to the business question. Instead of a presentation, as with most share phases, this part of the project will focus on the visualizations that were created and talk through them. Below each visualization will be examples of script a presenter could say when showing a visualization.

6.1 Introduction

To start and introduce the topic, a repeat of the business task is necessary to remind stakeholders what is being expected. The assignment was to find some important trends among users, how they could apply to Bellabeat users, and what trends could be important for the marketing department.

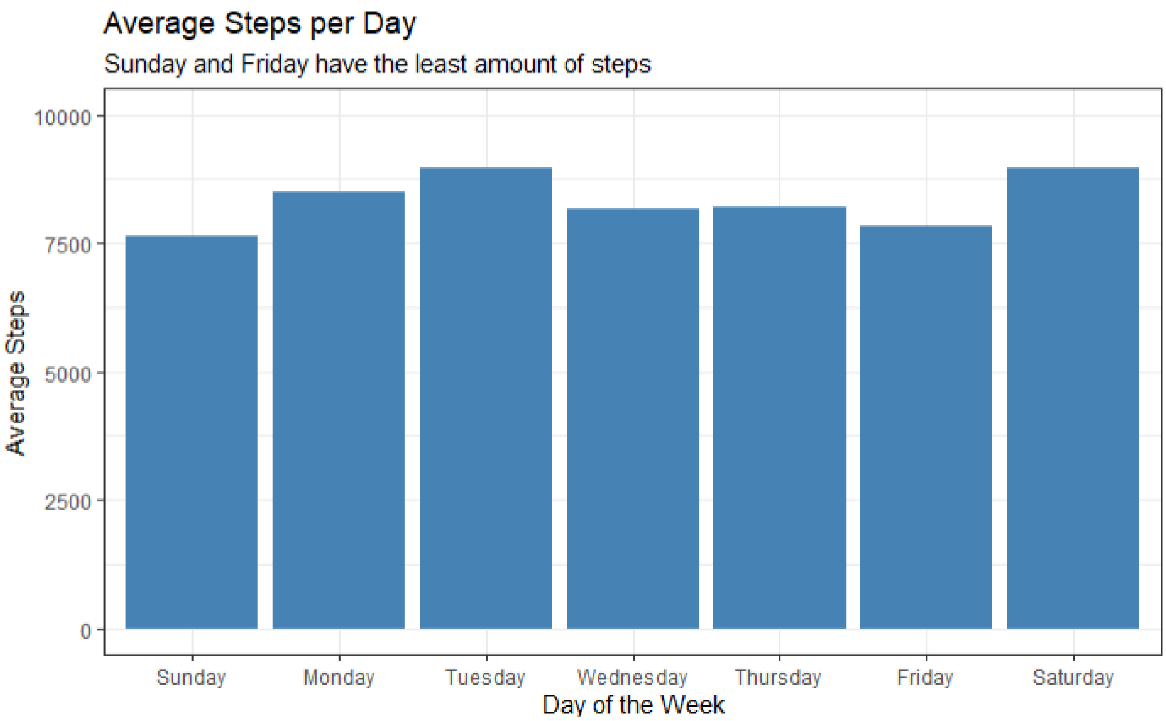
As a focus, a central theme of these trends are towards recommendations for the Bellabeat moving forward. Each of these visualizations have a purpose in telling that story. So, before showing visualizations, a slide could be dedicated to showing those recommendations in a quick summarized manner.

6.2 Daily Usage Trends

The first set of visualizations focuses on the usage patterns of users through days of the week and hour of the day. Figure 6.1 is a bar chart that outlines the average number of steps taken depending on the day of the week.

Figure 6.1

Averages Steps per Day of the Week

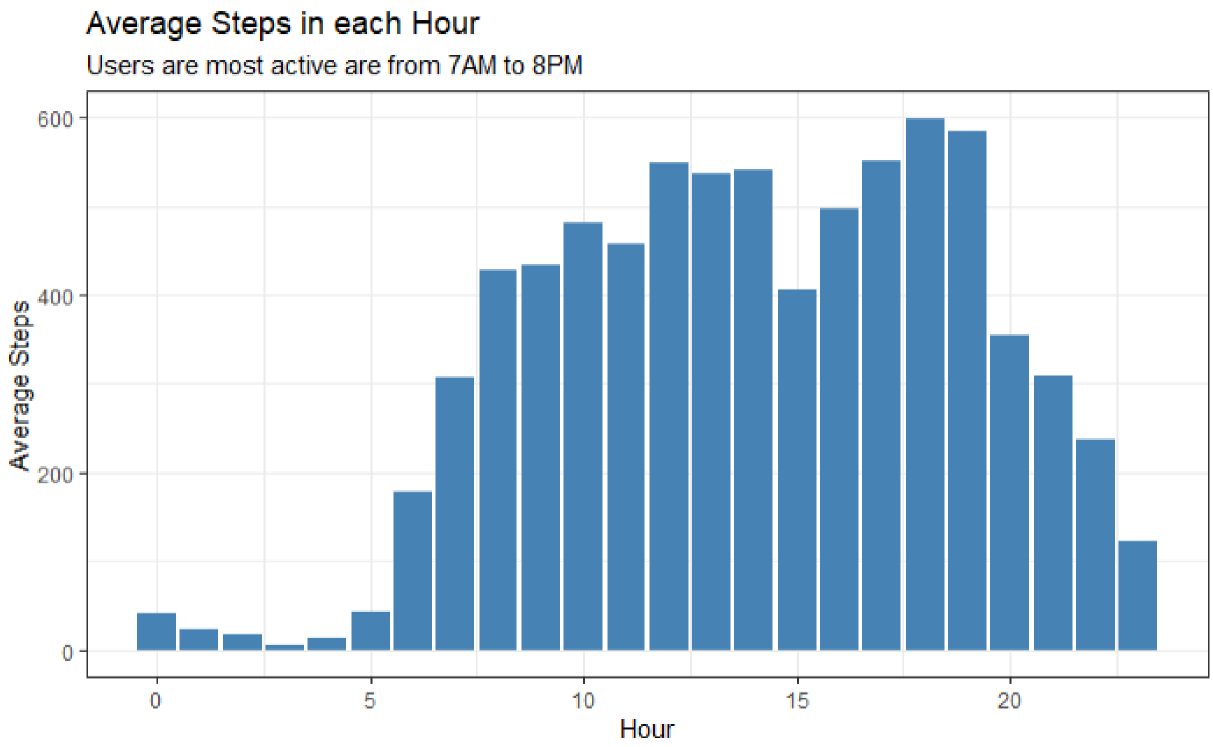


This slide measures physical activity throughout the week. We can see that Sunday and Friday are days which users on average have less steps. Meanwhile, Tuesday and Saturday are the days where users use the Fitbit's step tacking ability the most. A large portion of users have missing days for tracking, sleep logs, and weight logging. So, knowing when users are most and least active is important for implementing new features.

In a continuation to Figure 6.1, Figure 6.2 shows the average number of steps taken per hour of time. These trends are also important to know because it will determine when users are most and least active.

Figure 6.2

Averages Steps per Hour of the Day



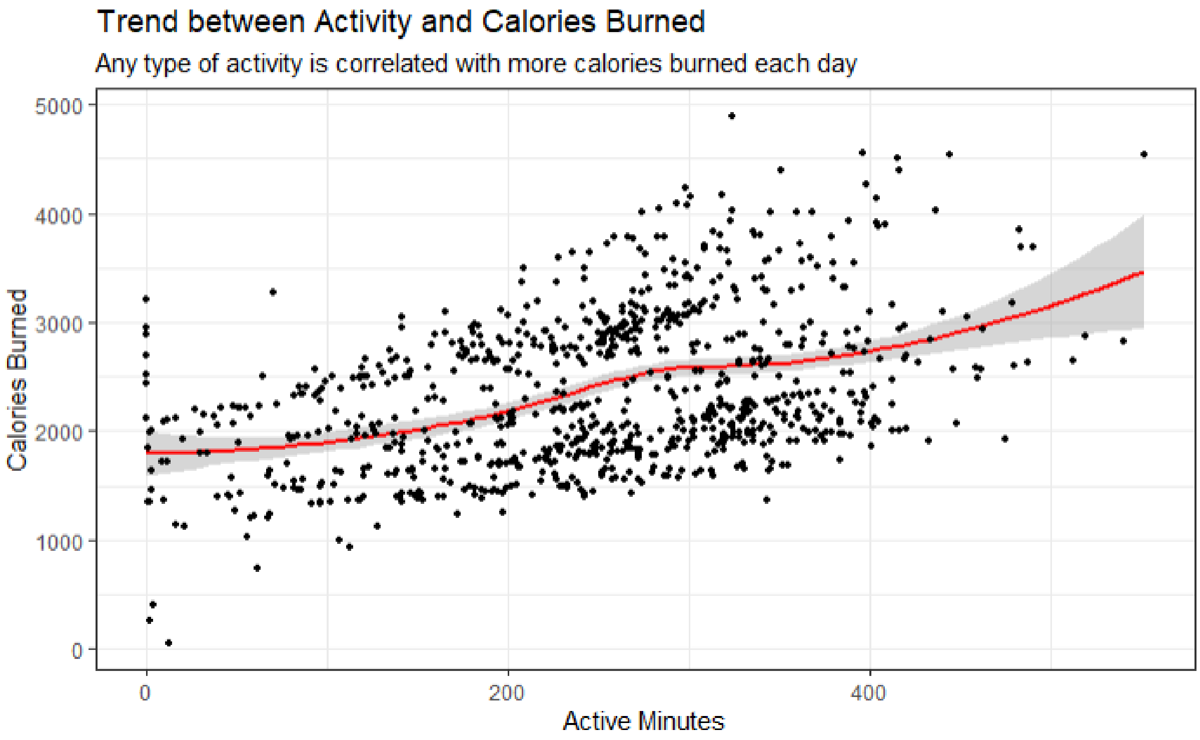
Users are most active from the 7th hour to the 20th hour, or 7 AM and 8 PM. The change of physically active users sharply increases at 6 AM and begins to sharply decrease at 9PM where no activity is present. The purpose of the previous two trends may reveal areas where users can be encouraged to charge their devices to increase the amount of involvement with sleep logging.

6.3 Activity Categories and Calories Burned

Many users decide to buy a fitness monitor to lose weight and burn more calories. By reminding users of their goals, more helpful interaction between the devices and users is created. Figure 6.3 shows a trend between the amount of active minutes in a day and the number of calories burned.

Figure 6.3

Calories Burned vs Activity



Clearly, there is a positive correlation between how active a user is and the amount of calories burned. This is important because the type of activity doesn't matter, so users could be encouraged to maintain any amount of activity to see positive results in much many calories are being burned.

7. Act

In the above pseudo-presentation, a group of visualizations that share a common goal were shared. However, in the act phase of the project, actual recommendations can be created and shared as well. In this phase, analysts recommend next steps for the stakeholders, and they decide whether or not to take that course of action.

For this specific project, it was noticed that there were users who interact with the fitness monitor less than others. An increase in users interacting with a company's systems and offers usually means an increase in profitability as users begin recommending your product or seeking other products that company's products. So, the main goal for the recommendations to the marketing department is to increase the amount of users interacting with all the features of the tracker as well as increase the consistency of that interaction.

There were two areas where users interactions were low; sleep and weight logging. There are multiple benefits to the user for both, but it's likely for multiple reason that these are the areas that are lacking. First, users don't want to waste time using these features. Second, users may spend their nights charging their fitness tracker.

This is where the trends detected earlier are important. Bellabeat and the marketing department could assist in the usage of the automated sleep tracking of their device by recommending times in which the user can charge their device. It would have to be times where usage is generally the lowest. So, Sundays and Fridays at 3 PM or 8 PM could be times where charging notifications could appear more frequently and before it's needed. These are the days and times which usage is generally the lowest.

Finally, to increase weight logging usage, the marketing team at Bellabeat could choose to use the information provided by the final chart. By reminding users that any type of exercise can help burn calories it would also remind them of their overarching goal and purpose of buying a fitness monitor in the first place. Users want to monitor their health, and a great way to do so is by monitoring your BMI. Each notification that recommends increasing activity could also ask or remind the user to interact more with their other features. If possible, a scale that connects to the device could aid the user in quickly logging their progress without the need to do it manually.