Case Study - Bellabeat

PHASE 1: ASK

Summary of the business task

Key tasks:

1. Identify the business task

- **About the company**: Founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women. The company provides several products, including the Bellabeat app, “Leaf” wellness watch, “Time” wellness watch, “Spring” water bottle, and Bellabeat membership for users.
- **Business objective**: This case study identifies growth opportunities for Bellabeat from available consumer data and addresses 3 questions:
    1. how are people using their smart devices?
    2. how could these trends apply to Bellabeat customers?
    3. how could these trends inform Bellabeat marketing strategy?
- **Limitations of this project**: This is a 1-week long project performed by only 1 analyst. Given the time and resources, the analysis will be most exploratory. It can be revisited in the future for a deeper dive.

2. Consider key stakeholders

- **Executive team**: Urška Sršen (Bellabeat’s co-founder and Chief Creative Officer); Sando Mur (Mathematician and Bellabeat’s co-founder).
- **Bellabeat marketing analytics team**: the group that takes the lead on this analysis.

PHASE 2: PREPARE

Description of all data sources used

Key tasks:

1. Download data and store it appropriately.

- Download *FitBit Fitness Tracker Data* from Kaggle (CC0: Public Domain, dataset made available through Mobius)
- Use Excel and SQL to initially explore data. Identify key data files to work with in **bold**.

File	ID and Date	Key content
dailyActivity_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Steps, Distance, Intensities, Active minutes, Calories
dailyCalories_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Calories (duplicated info as in `dailyActivity_merged.csv`)
dailyIntensities_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensities (duplicated info as in `dailyActivity_merged.csv`)
dailySteps_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Steps (duplicated info as in `dailyActivity_merged.csv`)
heartrate_seconds_merged.csv	7 users over 2 days; 2 users provide limited data	Heart rate (second)
hourlyCalories_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Calories (hourly)

File	ID and Date	Key content
hourlyIntensities_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensities (hourly; total and avg)
hourlySteps_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Steps (hourly; total)
minuteCaloriesNarrow_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Calories (minute)
minuteCaloriesWide_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Calories (minute; wide of `minuteCaloriesNarrow_merged.csv`)
minuteIntensitiesNarrow_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensity (minute)
minuteIntensitiesWide_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensity (minute; wide of `minuteIntensitiesNarrow_merged.csv`)
minuteMETsNarrow_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	METs = metabolic equivalents (minute)
minuteSleep_merged.csv	24 users over 31 days; 11 users provide data less than 25 days	Sleep stage (minute)
minuteStepsNarrow_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensity (minute)
minuteStepsWide_merged.csv	33 users over 31 days; 4 users provide data less than 25 days*	Intensity (minute; wide of `minuteStepsNarrow_merged.csv`)
sleepDay_merged.csv	24 users over 31 days; 14 users provide data less than 25 days	Sleep stage (minute)
weightLogInfo_merged.csv	8 users over 31 days; 7 users provide data less than 25 days	Weight, Body fat, BMI

2. Identify how it's organized.

- 18 CSV files, in both long and wide format.
- Some data files are an aggregate of other files. For example, `dailyActivity_merged.csv` is an aggregate of 3 files: `dailyCalories_merged.csv` , `dailyIntensities_merged.csv` , and `dailySteps_merged.csv` .

3. Sort and filter the data.

- Use PIVOT TABLE and SQL to sort and filter the data.
- Check for a unique identifier in each table – "Id". Notice that 4 users consistently provide data for shorter periods. They can be identified as Id: 2347167796, 3372868164, 4057192912, 8253242879.

4. Determine the credibility of the data.

- **Does this data ROCCC?**
    1. **Reliable**: Yes, the data is reliable. 30 eligible Fitbit users generated the data in a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016.
    2. **Original**: Yes, the original data can be found here (https://zenodo.org/record/53894#.X9oeh3Uzaao).
    3. **Comprehensive**: Yes, the data is comprehensive and relevant to this analysis. It contains personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.
    4. **Current**: No, this is historical data that covers only 1 month in 2016 (6 years ago).
    5. **Cited**: Yes, this data is cited and vetted.
- **Data limitations that could lead to bias:**
    1. **Outdated data.** The current trend of using a wellness device might differ from 2016.
    2. **Small sample size.** For most files, data is available for only 33 participants over 1 month. In some files, information is available for less than 10 people.

3. **No demographic information.** For this analysis, I will assume the participants are a good representative of the targeted customers of Bellabeat.

PHASE 3: PROCESS

Documentation of any cleaning or manipulation of data

Key tasks:

1. Check the data for errors.

- Always to check for data errors — a great thing to always keep in mind.

2. Choose your tools.

- **Excel**: initial exploration, quick check in data content.
- **SQL**: data cleaning, management, and compiling.
- **R**: data visualizations.

3. Transform the data so you can work with it effectively.

- I upload all data files into BigQuery, transform data using SQL, and then move the cleaned data into R for analysis and visualizations.

4. Document the cleaning process.

1. Upload data files under a new project, "Bellabeat", in BigQuery cloud space. The date and time-related variables cannot be auto-detected, so I import them as STRING first.
2. Check for primary and foreign keys to build the relationships among tables. Common identifiers: "Id", and time columns.
3. Standardize time columns using REGEXP.
4. Create a master file at the "day" level by merging `dailyActivity_merged.csv` , `dailyCalories_merged.csv` , `dailyIntensities_merged.csv` , `dailySteps_merged.csv` , and `sleepDay_merged.csv` based on Id and time.
5. Create a master file at the "hour" level by merging `hourlyCalories_merged.csv` , `hourlyIntensities_merged.csv` , `minuteSleep_merged.csv` , and `hourlySteps_merged.csv` based on Id and time.
6. Create "day", "day of week", and "weekend" columns as identifiers for later use.
7. Export the clean data from BigQuery to the local folder for later use. The data files for this project are small enough for download.

PHASE 4: ANALYZE

Summary of the analysis

Key tasks:

1. Aggregate your data so it's useful and accessible.

- Aggregate minute level information to hourly in `minuteSleep_merged.csv` . Roll up sleep minutes, and record sleep stages within an hour.

2. Organize and format your data.

- Organize data in long format initially. Later, when plotting for the heatmap, selected data is changed into a wide format.
- Prepare two master files for analysis: one at the daily level, including steps, distance, intensity, calories, and sleep; the other one at the hourly level, including steps, intensity, calories, and sleep.

3. Perform calculations.

- Upload local cleaned data files into R studio cloud folder.
- Load R packages and cleaned data tables and prepare for analysis.
- Perform merge and data transformation when needed.

```
#install.packages('tidyverse')
library(tidyverse)
library(dplyr)

DayActivity <- read_csv("DayActivity_master.csv")
HourActivity <- read_csv("HourActivity_master.csv")
```

4. Identify trends and relationships.

- Explore the daily data first. Assume information is automatically collected once the participant wears the device, including steps, calories, activity intensity, and sleep. Use of the device is defined as “minutes” tracked during the day.

- ***The first business objective is to find out how people use their smart devices**. Based on the data, I interpret that question as **how long do people wear the device each day** during the testing period.

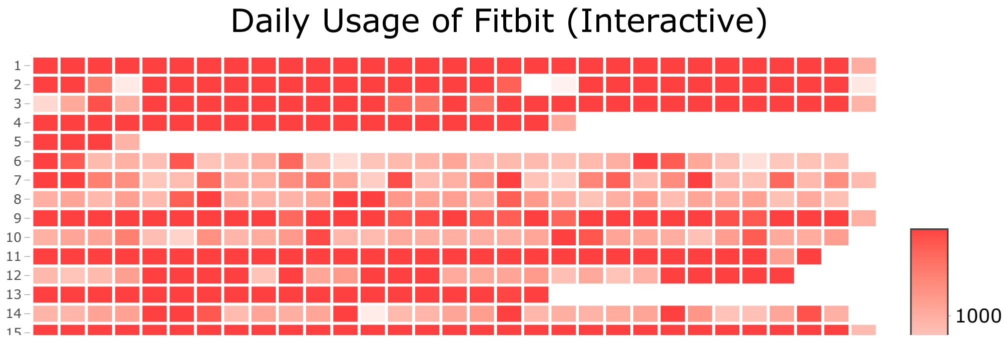
- I find the best way to present that information is to show a plot that visualizes each person’s daily usage over the testing period.

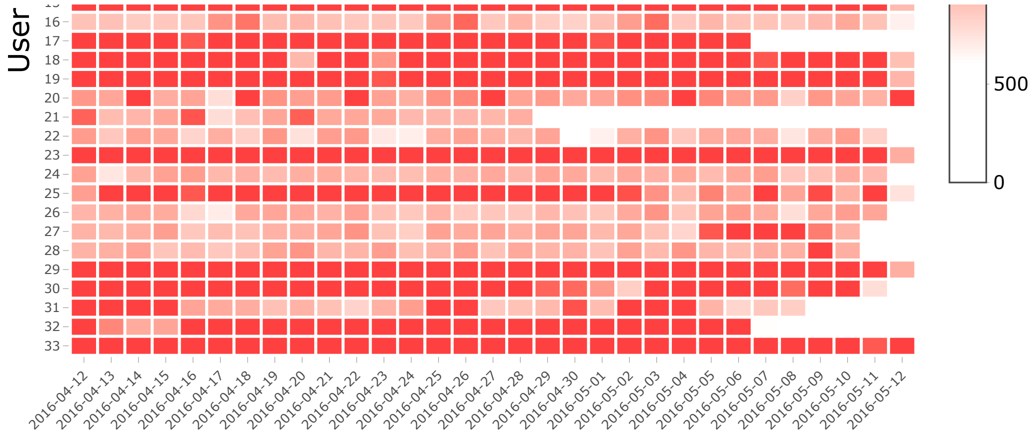
```
# install.packages('heatmaply')
library(heatmaply)

# generate total minutes tracked by fitbit
DayActivity_1 <- DayActivity %>%
  mutate(UseMin = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes) %>%
  distinct()

# change long format to wide, and creat matrix format
mat <- DayActivity_1 %>%
  select(Id, ActivityDate, UseMin) %>%
  arrange(ActivityDate) %>%
  pivot_wider(names_from = ActivityDate, values_from = UseMin) %>%
  select(-"Id") %>%
  as.matrix()

# heatmap
heatmaply(
  mat,
  dendrogram = "none",
  xlab = "Dates", ylab = "User",
  main = "Daily Usage of Fitbit (Interactive)",
  grid_color = "white",
  grid_width = 0.00001,
  titleX = FALSE,
  #hide_colorbar = TRUE,
  label_names = c("User", "Date:", "Tracked Minutes"),
  fontsize_row = 6, fontsize_col = 6,
  labCol = colnames(mat),
  labRow = rownames(mat),
  scale_fill_gradient_fun = ggplot2::scale_fill_gradient2(
    low = "white",
    high = "brown1",
    midpoint = 600,
    limits = c(0, 1440)),
  heatmap_layers = theme(axis.line=element_blank())
)
```





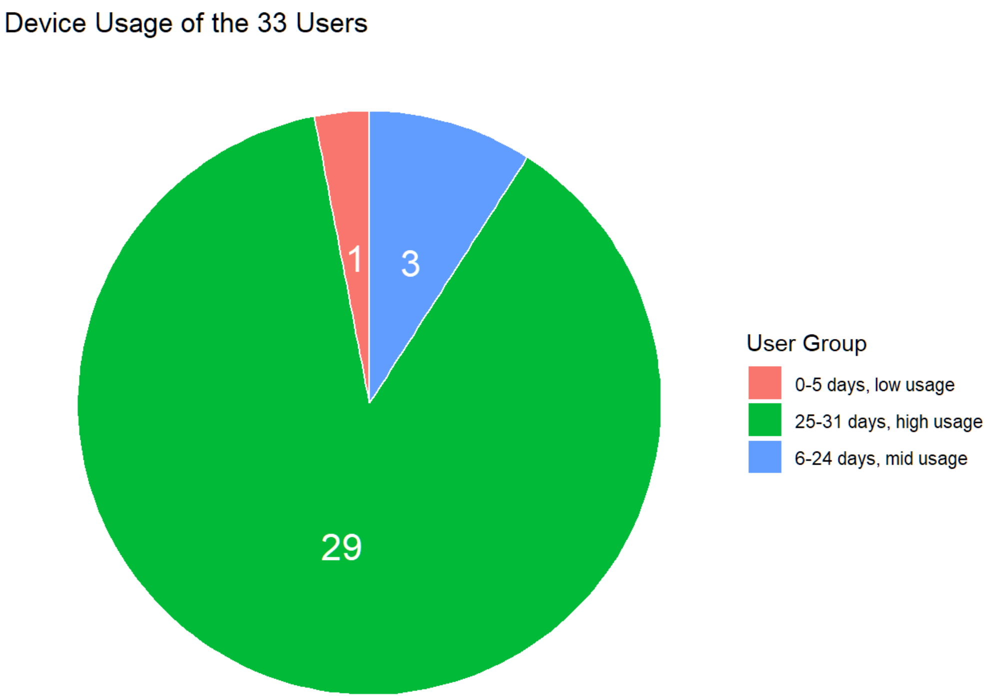
- In this heatmap, each row represents 1 of the 33 participants, and each column represents 1 of the 31 testing days. The color of a cell gets darker when the daily usage is higher, and wise versa. It is an interactive heatmap, so each cell's value shows when the mouse hovers over it. The max value for each cell is 1440 minutes, equivalent to a 24-hour day. The min value for the cell is 0 minutes, and the cell will be a blank in that case.
- Overall, the wearing behavior is consistent over time. That is, when users wear the device "full time", they keep that trend (shown in darker red, such as user no.1), and when they wear it casually, they keep wearing it only half of the time (displayed in a pale coral color, such as user no.6). Notably, there is 1 user stops wearing the device after 4 days of intensive usage (i.e. user no.5). Interviews with this user on the reason of stop-wearing could potentially be beneficial. For this analysis, the focus will be on more general trends.
- There are two primary usage behaviors: (1) continue wearing the device for 31 days and (2) full-time wearing the device once you start wearing it in a day. I prioritize behavior (1) over behavior (2). The rationale is that always returning to the device the next day is more important, even when the user might be wearing it only a couple of hours each day. Coming back to the device implies some parts of the need are satisfied. So it is fair to categorize users into different groups based on the number of days of usage (as opposed to the number of minutes used during a day) and conduct the following analysis on trends.

```
# create group, number of days used, and minutes used per hour
User <- DayActivity_1 %>%
  distinct() %>%
  group_by(Id) %>%
  summarise(num_dayuse = n(), avg_UseMin = round(mean(UseMin)/24)) %>%
  mutate(use_group = case_when(
    between(num_dayuse, 0,5) ~ "0-5 days, low usage",
    between(num_dayuse, 6,24) ~ "6-24 days, mid usage",
    between(num_dayuse, 25,31) ~ "25-31 days, high usage"
  ))

User_bar <- User %>%
  group_by(use_group) %>%
  summarise(num_user = n())

# Compute the position of labels
User_bar <- User_bar %>%
  arrange(desc(use_group)) %>%
  mutate(prop = num_user / sum(User_bar$num_user) *100) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop )

# Basic piechart
ggplot(User_bar, aes(x="", y=prop, fill=use_group)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  labs(fill = "User Group", title = "Device Usage of the 33 Users") +
  geom_text(aes(y = ypos, label = num_user), color = "white", size=6)
```



- As shown in the pie chart, 29 users wear the device for over 25 days in the 31-day testing period and are classified as “high usage” users. 1 user wears only 4 days, classified as “low usage”; the rest 3 users are “mid usage” users.

5. Data integrity checks again.

```
# Compare hourly data to daily data
# install.packages("formattable")
library(formattable)

User_hour <- HourActivity %>%
  distinct() %>%
  group_by(Id) %>%
  summarise(num_hour = n(), num_dayuse_hour = length(unique(ActivityDate))) %>%
  mutate(avg_usemin_hour = (num_hour*60)/(num_dayuse_hour*24)) %>%
  left_join(User, by = "Id") %>%
  mutate(day_diff = num_dayuse - num_dayuse_hour)

User_tbl <- User_hour %>%
  arrange(desc(day_diff)) %>%
  select("Id", "num_dayuse_hour", "num_dayuse", "day_diff") %>%
  filter(day_diff != 0)

colnames(User_tbl) <- c("User ID", "Used Day from Daily Data", "Used Day from Hourly Data", "Difference (Day)")

formattable(User_tbl, list("Difference (Day)" = color_bar("pink")))
```

User ID	Used Day from Daily Data	Used Day from Hourly Data	Difference (Day)
1503960366	30	31	1
3977333714	29	30	1
6290855005	28	29	1
8253242879	18	19	1
8583815059	30	31	1
8792009665	28	29	1

- When moving from daily to hourly data, the first thing to do is to check data integrity by comparing the daily and hourly data. I pick “number of days of usage” as the test because it is critical information that will be used in later analysis and can be derived from both datasets.
- As shown in the table, there is a slight 1-day difference (3%) for 6 users (18%). The discrepancy might come from aggregation processes but can be ignored for now.
- Note: the “Used Day from Daily Data” is calculated based on files `dailyActivity_merged.csv` , `dailyCalories_merged.csv` , `dailyIntensities_merged.csv` , `dailySteps_merged.csv` , and `sleepDay_merged.csv` ; “Used Day from Hourly Data” is calculated based on files `hourlyCalories_merged.csv` , `hourlyIntensities_merged.csv` , `minuteSleep_merged.csv` , and `hourlySteps_merged.csv` .

PHASE 5: SHARE

Supporting visualizations and key findings

Key tasks:

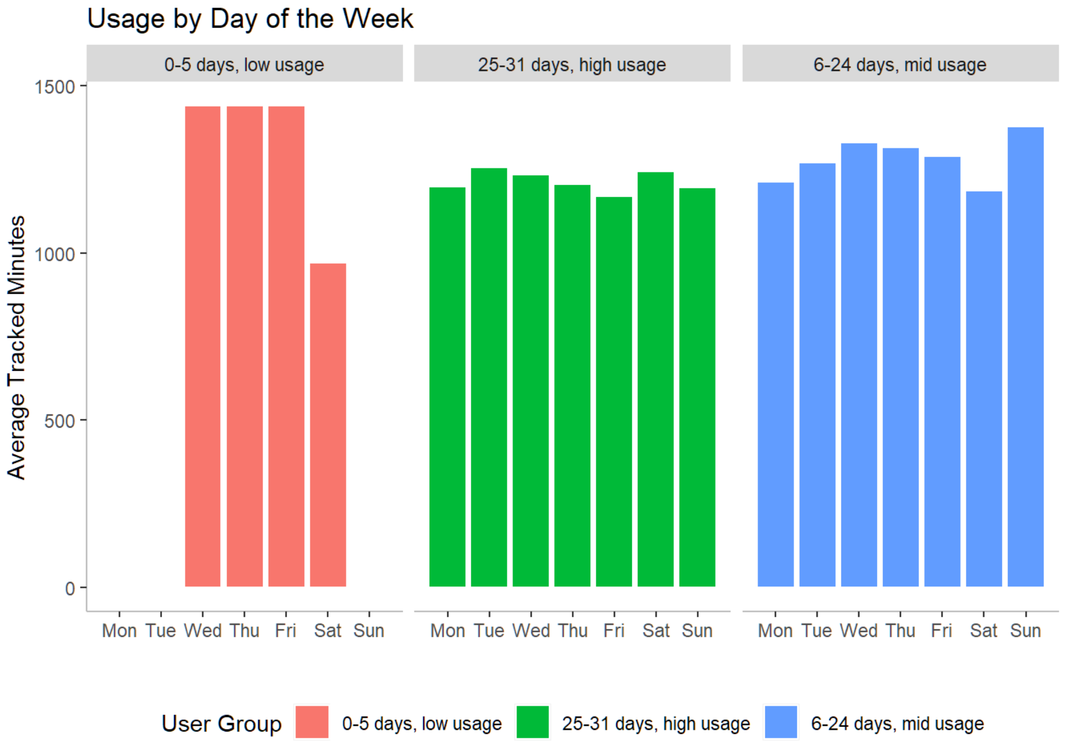
1. Determine the best way to share your findings.
2. Create effective data visualizations.
3. Present your findings.
4. Ensure your work is accessible.

Do people use the device more on certain days of the week?

- I choose bar charts to illustrate the device usage in a day and compare it across different days of the week.

```
# aggregate data by day of the week (Usage)
bar_data <- DayActivity_1 %>%
  left_join(User, by = "Id") %>%
  group_by(use_group, dow, dow_number) %>%
  summarise(avg_UseMin = mean(UseMin)) %>%
  arrange(dow_number)

# stacked bar chart (Usage)
ggplot(bar_data,
       aes(fill=use_group, y=avg_UseMin, x=dow_number)) +
    facet_wrap(~use_group) +
  geom_bar(position=position_dodge(preserve = "single"), stat="identity", color="white") +
  labs(title = "Usage by Day of the Week", fill = "User Group",
       y = "Average Tracked Minutes", x="") +
  scale_x_continuous(breaks=0:6,
                     labels=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun")) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "grey")) +
  theme(legend.position="bottom")
```



- As shown in the chart, device use is slightly higher during the weekend and lower on Monday. Also, people in the “mid usage” group like to “take a break” from the device on Saturdays but then pick it up on Sundays. People in the “high usage” group wear the device consistently throughout the week and might have developed the habit of wearing Fitbit.
- It is also worth noting that the only person in the “low usage” wears the device for almost 24 hours for 3 days in a row, then decreases to half of the time on the 4th day, before dropping off. It would be interesting to find out the reason for quitting.
- Compare the “high usage” and “mid usage” groups, the data also shows that the average daily usage for the “high usage” group is not necessarily higher than the other group. It is surprising because I would expect that behavior (1) and (2) are positively correlated.

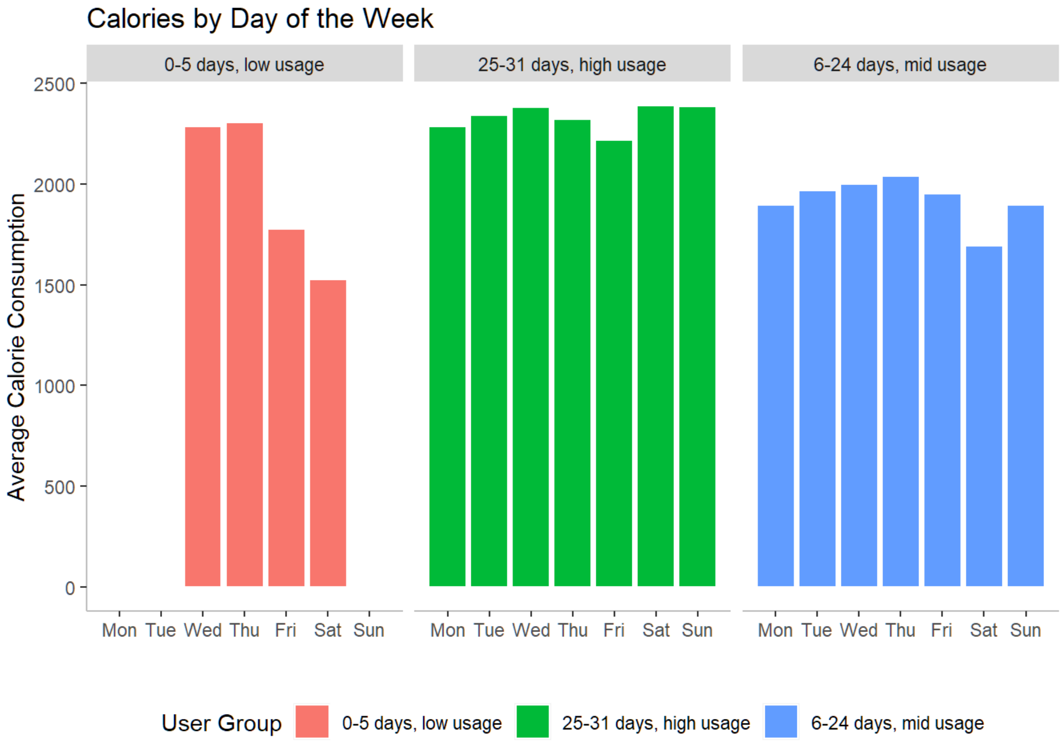
  1. In the 31-day testing period, continue wearing the device.
  2. Within a day, keep wearing the device once you start wearing it.

So if people who wear the device for more days wear it for a shorter period each day, what motivates them to keep wearing the device?

- I decide to plot another bar chart to illustrate the calorie consumption – another way of “using” the device – in a day and compare it across different days of the week.

```
# aggregate data by day of the week (Calories)
bar_data_calorie <- DayActivity_1 %>%
  left_join(User, by = "Id") %>%
  group_by(use_group, dow, dow_number) %>%
  summarise(avg_calorie = mean(Calories)) %>%
  arrange(dow_number)

# stacked bar chart (Calories)
ggplot(bar_data_calorie,
       aes(fill=use_group, y=avg_calorie, x=dow_number)) +
    facet_wrap(~use_group) +
  geom_bar(position=position_dodge(preserve = "single"), stat="identity", color="white") +
  labs(title = "Calories by Day of the Week", fill = "User Group",
       y = "Average Calorie Consumption", x="") +
  scale_x_continuous(breaks=0:6,
                     labels=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun")) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "grey")) +
  theme(legend.position="bottom")
```



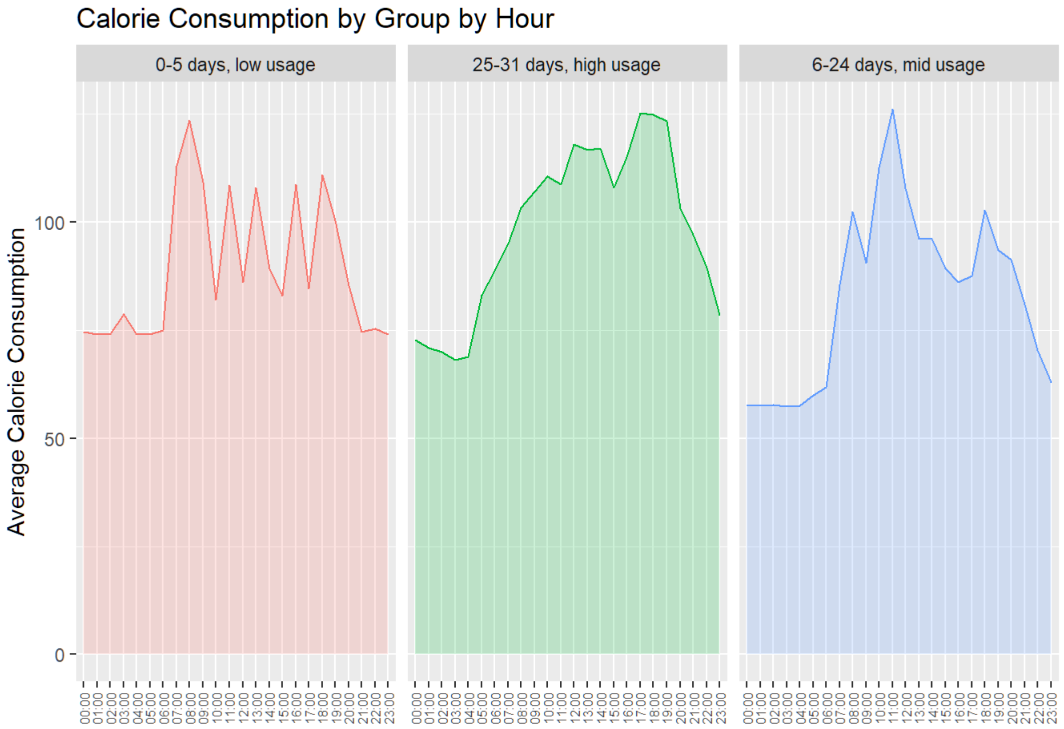
- The calorie consumption for the “high usage” group is noticeably higher than that for the “mid usage” group. By eyeballing, the difference is, on average, 300-400 calories per day. With a lower average minute of wearing, the activity intensity for the “high usage” group has to be higher than that of the “mid usage” group.

So if people who wear the device for more days consume more calories generally, what time are they most active during a day?

- I choose the area chart to show the calories consumed during the day and compare it among different user groups.

```
# Calculate average calories tracked per hour
bar_hour <- HourActivity %>%
  left_join(User, by = "Id") %>%
  mutate(ActHour = format(as.POSIXct(ActHour), format = "%H:%M")) %>%
  group_by(ActHour, use_group) %>%
  summarise(avg_calories = mean(Calories), .groups = "drop") %>%
  arrange(ActHour, use_group)

ggplot(bar_hour,
       aes(x = ActHour, y = avg_calories, group = use_group, color = use_group, fill = use_group)) +
  facet_wrap(~use_group) +
  #geom_line() +
  geom_area(alpha=0.2) +
  labs(title = "Calorie Consumption by Group by Hour", fill = "User Group",
       y = "Average Calorie Consumption", x="") +
  #theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        #panel.background = element_blank(), axis.line = element_line(colour = "grey")),
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=0.5, size=6)) +
  theme(legend.position="none")
```



- According to the data, calorie consumption is more in the “high usage” group than the “mid usage” group throughout the day, implying a greater metabolic rate and potentially healthier lifestyle.
- The only exception is around 11 am lunchtime when the calorie consumption in the “mid usage” group spikes up high. It might be going out to crab food, or a noon workout because the spike is so significant compared to the calorie consumption level at sleep (i.e. 1 am - 6 am).
- There is another calorie consumption peak between 5 and 7 pm for both the “high usage” and “mid usage” groups. This might indicate exercising or intensive activities, and the high-calorie consumption remains for a while before dropping.

PHASE 6: ACT

Top high-level content recommendations based on analysis

Key tasks:

1. Create your portfolio.
2. Add your case study.
3. Practice presenting your case study to a friend or family member.

Key takeaways:

1. How are people using their smart devices?

- “High Usage” group: 29 people (88%) wear the device for 25-31 days
    1. Good compliance in continuing to wear the device.
    2. Consume more calories during the day. This group might consist of male users or users with strong basal metabolism rates.
    3. Are more active during 12 - 2 pm and 5 - 7 pm.
- “Mid Usage” group: 3 people (9%) wear the device for 6-24 days
    1. Wear the device for a longer time during the day.
    2. Consume fewer calories during the day. This group might consist of female users or users with moderate basal metabolism rates.
    3. Are more active during 8 am, 11 am, and 6 pm.
- “Low Usage” group: 1 person (3%) wears the device for 0-5 days
    1. Wear the device almost “full-time” until dropping off.
    2. Good calorie consumption in the first 2 days of wearing. This might be a male user or a user with a strong basal metabolism rate.

3. Is more active during 8 am.

2. How could these trends apply to Bellabeat customers?

- Because Bellabeat focuses on the female market, the “mid usage” group might be the most relevant. The users wear the device throughout the day and especially love to wear the device on Sundays. Bellabeat’s fashionable design might strengthen this usage pattern. Female users are wearing a health device that could also serve as a piece of jewelry during gathering with family and friends on the weekend.
- Also, the “mid usage” group tends to wear the device for longer during the day. If 24-hour wearing is what Bellabeat wants to encourage users to do, then the shape of the device should remove all pointed corners and sharp edges. That way, users won’t get hurt wearing the device to sleep or holding a baby in their arms.

3. How could these trends inform Bellabeat marketing strategy?

- The “mid group” users consume the most calories during lunch break. It could be young moms who have to take of the family after work and have only lunchtime for self-care. Given this lifestyle, Bellabeat might promote noon workout classes online or partner with local health centers and office gyms.

Limitations:

- The data used for this analysis is the tracking record of 33 participants over 31 days in 2016. Conclusions based on a few data points 6 years ago might cause bias.

Next steps:

- Collect more data.
- Collect updated data.
- Survey people on experience using the device, including reason for wearing and not wearing, likes and dislikes about the device, etc.
- Collect demographic data since Bellabeat’s focus is on female users.