# ‘Bellabeat’ Product Analysis: How Can A Wellness Company Play It Smart?

For this case study, I take on the role of a junior data analyst working in the marketing analyst team at *Bellabeat*, a high-tech manufacturer of health-focused products for women. Its founder believes that analyzing user data could help unlock new growth opportunities for the company. Therefore, my team is directed to gain insight into how users are utilizing our products and services in order to present high-level recommendations for Bellabeat’s marketing strategy. I am specifically assigned to focus on our best-selling product, the Leaf smart device.

To accomplish my task, I will follow the steps of the data analysis process: **Ask**, **Prepare**, **Process**, **Analyze**, **Share**, and **Act**.

**Tools**: Spreadsheet (Microsoft Excel), R Programming (RStudio), Tableau

**Visualization**: Go to my [**Tableau profile**](https://public.tableau.com/app/profile/jhermienpaul)**.**

**Documentation**: Visit my[**GitHub profile**](https://github.com/jhermienpaul)**.**

## Step 1: Ask

**Context**

In 2016, Bellabeat launched the following products and services that empower women with knowledge about their health:

1. **Leaf** – a smart tracker worn as a bracelet/necklace which monitors users’ activity and sleep

2. **Time** – a smart watch that tracks users’ pulse rate

3. **Spring** – a smart water bottle that traces users' daily water intake

4. **Bellabeat app** – an application connected to Bellabeat smart devices which provides users with their health data in real-time

5. **Bellabeat membership** – a subscription-based membership program that gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals

**Task**

I will analyze how users are utilizing the Leaf smart device. This will involve finding relevant patterns and trends in whatever information is found in the dataset.

Afterward, I will propose recommendations for Bellabeat’s marketing strategy.

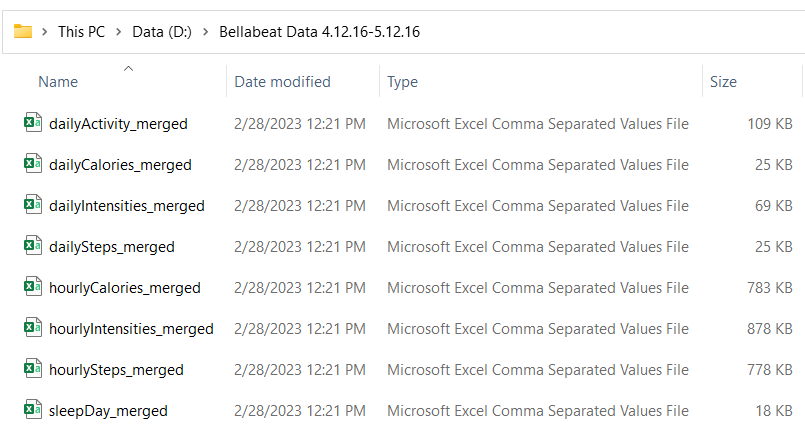
## Step 2: Prepare

**Access the data**

My team was provided with Bellabeat’s product user data from April 12 to May 12 of 2016. Access the dataset through this [link](https://www.kaggle.com/arashnic/fitbit).

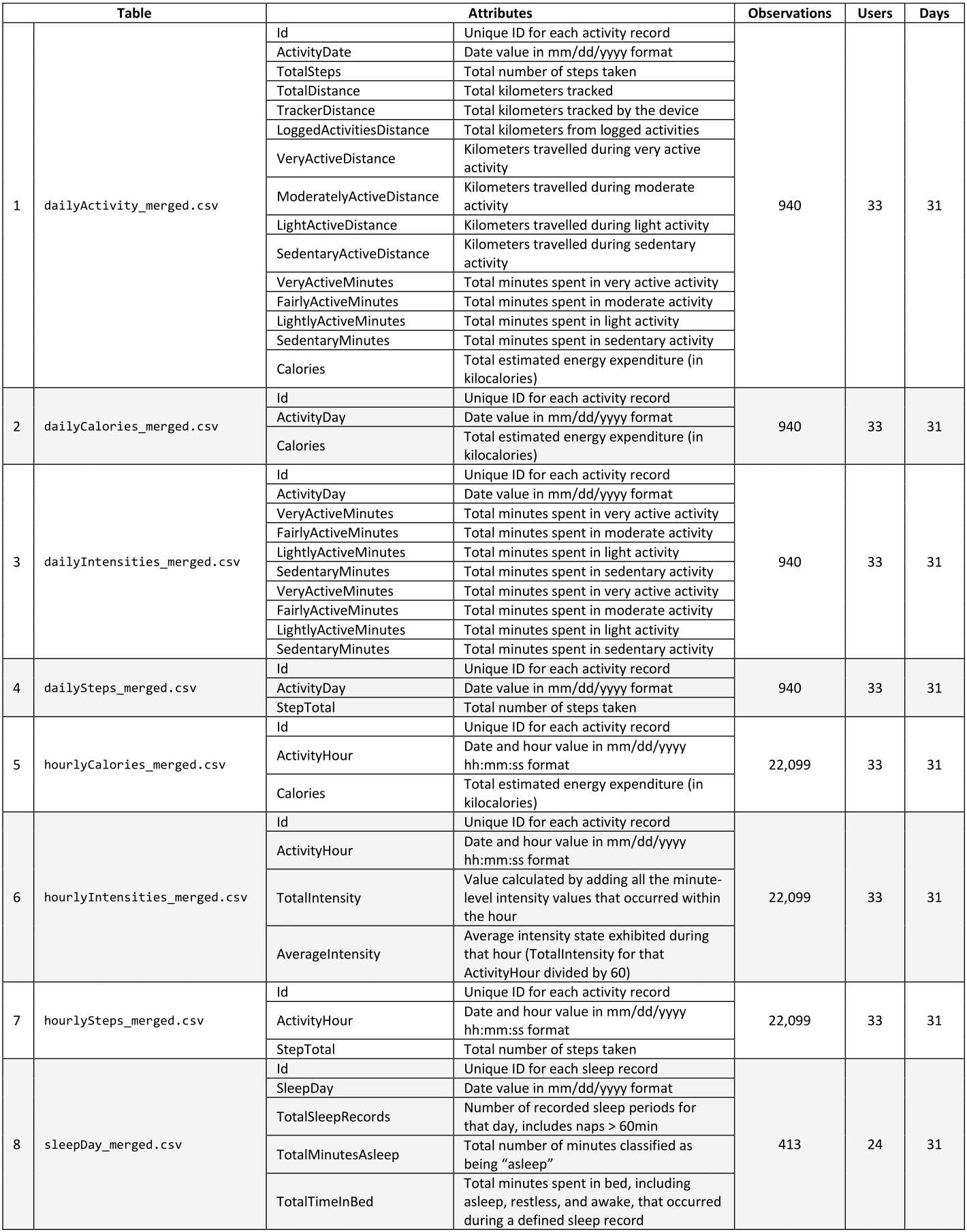
**Collect the data**

The dataset consists of many CSV files with users’ health data in various timeframes. I only downloaded 8 files pertaining to activity and sleep data in daily and hourly intervals to concentrate on larger trends and patterns in Leaf product usage.



**Examine the data**

I inspected the files in Microsoft Excel and found that each file holds the following information:



The description for each attribute is based on this [metadata](https://www.fitabase.com/media/1930/fitabasedatadictionary102320.pdf).

The number of users and days were determined by counting the number of unique values in the **Id** and **ActivityDate/ActivityDay** columns, respectively.

Graphical user interface, text, application

Description automatically generated

For files with **ActivityHour** column (DATE-TIME format), I temporarily inserted a column named **Date** and used the INT function to extract only the DATE portion. Then, I counted the number of unique values in the **Date** column.

Graphical user interface, text, application

Description automatically generated

My manager confirmed only 33 users consented to share their personal data for data analysis purposes.

It appears that **dailyCalories\_merged.csv, dailyIntensities\_merged.csv,** and **dailySteps\_merged.csv** are mere subsets of **dailyActivity\_merged.csv,** so I decided to delete them to avoid redundancy.

**Limitations of the data**

1. **Outdated data**: The data is from 2016, which is seven years ago. The usage patterns, user behaviors, and market trends might have significantly changed since then, and the findings may not be relevant to the current market and user base.

2. **Narrow timeframe**: The data only covers a period of 31 days, which is a short period to make any significant conclusions about the usage of the product.

3. **Small sample size**: The dataset only includes data from 33 users at most. This could limit the representativeness of the data, and the findings may not generalize to the wider population of Leaf users.

4. **No demographic info**: The data lacks any demographic information about the users (such as age, race, income, etc.) which can provide additional insights into how different user groups are utilizing the product.

## Step 3: Process

I used R for data cleaning and wrangling. I started by installing and loading the required packages in R Studio.

# install and load packages

install.packages(c("tidyverse", "lubridate", "dplyr", "tidyr", "janitor","ggplot2", "waffle", "corrplot"))

library(tidyverse) # for data manipulation and visualization

library(lubridate) # for working with dates and times

library(dplyr) # for data manipulation

library(tidyr) # for data tidying and reshaping

library(janitor) # for data cleaning and formatting

library(ggplot2) # for data visualization

library(waffle) # for creating waffle charts

library(corrplot) # for creating correlation plots

Then, I imported the files as data frames.

# import files

daily\_activity <- read\_csv("D:/Bellabeat Data 4.12.16-5.12.16/dailyActivity\_merged.csv")

daily\_sleep <- read\_csv("D:/Bellabeat Data 4.12.16-5.12.16/sleepDay\_merged.csv")

hourly\_calories <- read\_csv("D:/Bellabeat Data 4.12.16-5.12.16/hourlyCalories\_merged.csv")

hourly\_intensities <- read\_csv("D:/Bellabeat Data 4.12.16-5.12.16/hourlyIntensities\_merged.csv")

hourly\_steps <- read\_csv("D:/Bellabeat Data 4.12.16-5.12.16/hourlySteps\_merged.csv")

**Clean the data**

1. I renamed the headers in each data frame using the **clean\_names()** function to ensure they are syntactically valid (i.e., start with a letter/underscore and contain only letters/numbers/underscores).

# rename headers

daily\_activity <- clean\_names(daily\_activity)

daily\_sleep <- clean\_names(daily\_sleep)

hourly\_calories <- clean\_names(hourly\_calories)

hourly\_intensities <- clean\_names(hourly\_intensities)

hourly\_steps <- clean\_names(hourly\_steps)

Look at the new headers:

# show headers

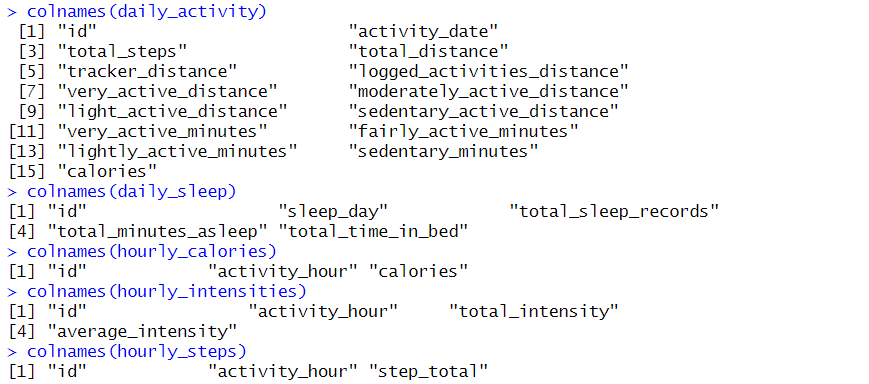
colnames(daily\_activity)

colnames(daily\_sleep)

colnames(hourly\_calories)

colnames(hourly\_intensities)

colnames(hourly\_steps)



2. I checked the data type and format of headers in each data frame using the **glimpse()** function to ensure they are consistent with the metadata.

# show data type and format

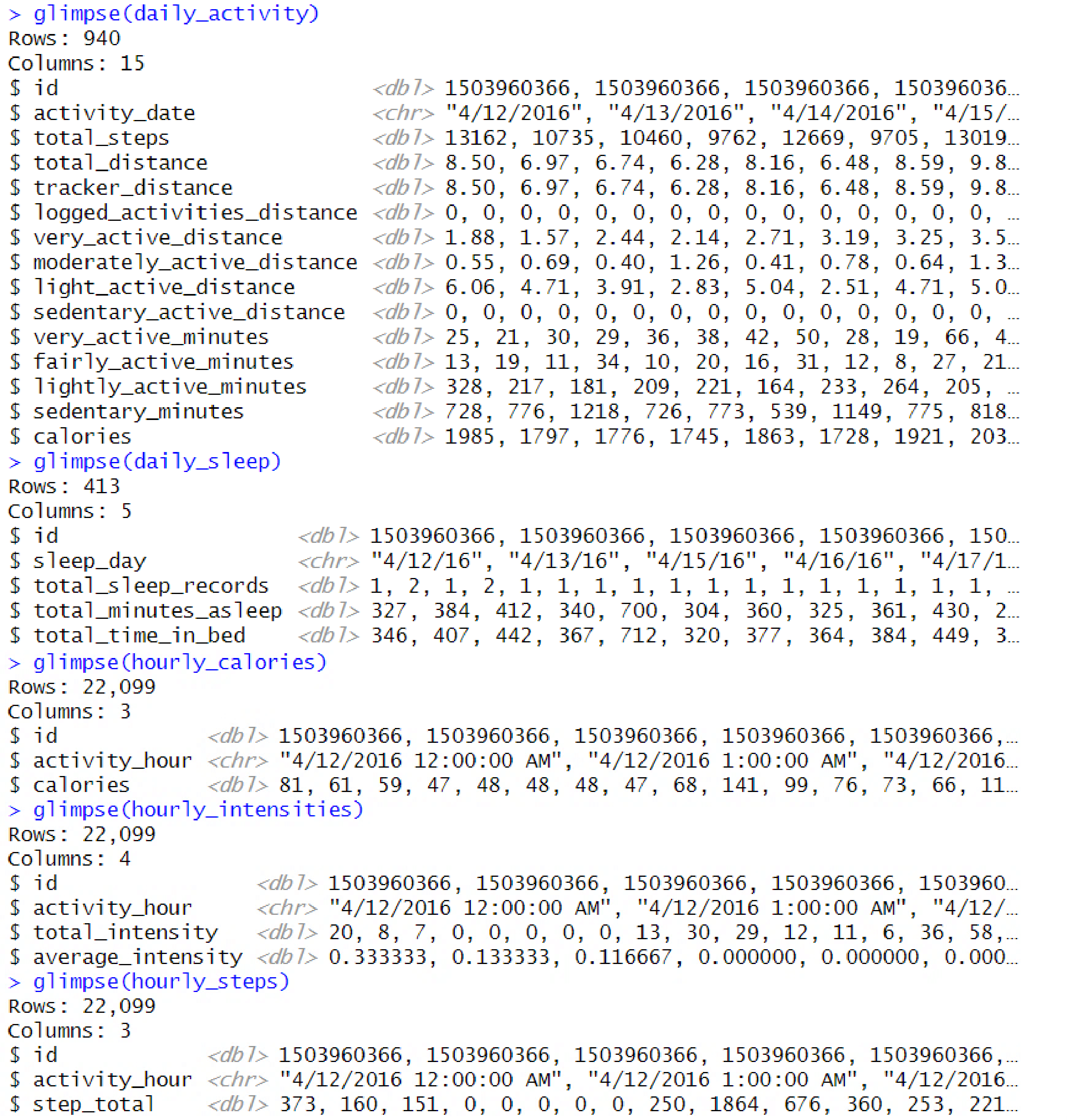
glimpse(daily\_activity)

glimpse(daily\_sleep)

glimpse(hourly\_calories)

glimpse(hourly\_intensities)

glimpse(hourly\_steps)



It appears that **id** is read as FLOAT type, and **activity\_date, sleep\_day,** and **activity\_hour** are read as STRING type. Therefore, I converted these columns into their appropriate formats using the **mutate()** function.

# typecast columns: daily\_activity

daily\_activity <- daily\_activity %>%

mutate(id = as.character(id)) %>%

mutate(activity\_date = mdy(activity\_date))

# typecast columns: daily\_sleep

daily\_sleep <- daily\_sleep %>%

mutate(id = as.character(id)) %>%

mutate(sleep\_day = mdy(sleep\_day))

# typecast columns: hourly\_calories

hourly\_calories <- hourly\_calories %>%

mutate(id = as.character(id)) %>%

mutate(activity\_hour = mdy\_hms(activity\_hour))

# typecast columns: hourly\_intensities

hourly\_intensities <- hourly\_intensities %>%

mutate(id = as.character(id)) %>%

mutate(activity\_hour = mdy\_hms(activity\_hour))

# typecast columns: hourly\_steps

hourly\_steps <- hourly\_steps %>%

mutate(id = as.character(id)) %>%

mutate(activity\_hour = mdy\_hms(activity\_hour))

3. For uniformity, I renamed the headers in DATE format as **date** and those in DATE-TIME format as **date\_time** using the **rename()** function.

# rename headers: daily\_activity

daily\_activity <- daily\_activity %>%

rename(date = activity\_date)

# rename headers: daily\_sleep

daily\_sleep <- daily\_sleep %>%

rename(date = sleep\_day)

# rename headers: hourly\_calories

hourly\_calories <- hourly\_calories %>%

rename(date\_time = activity\_hour)

# rename headers: hourly\_intensities

hourly\_intensities <- hourly\_intensities %>%

rename(date\_time = activity\_hour)

# rename headers: hourly\_steps

hourly\_steps <- hourly\_steps %>%

rename(date\_time = activity\_hour)

4. I checked for observations with missing or null values using the **!complete.cases()** function.

# count incomplete observations

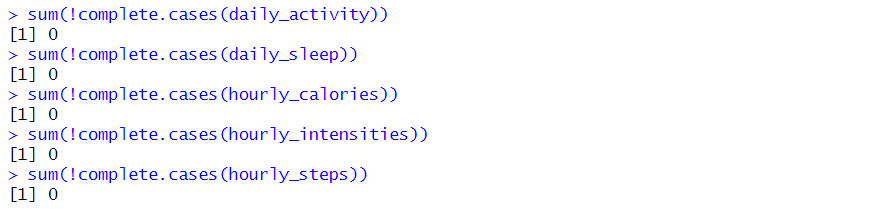
sum(!complete.cases(daily\_activity))

sum(!complete.cases(daily\_sleep))

sum(!complete.cases(hourly\_calories))

sum(!complete.cases(hourly\_intensities))

sum(!complete.cases(hourly\_steps))

****

No incomplete observations were found in all data frames.

5. I checked for duplicated observations using the **duplicated()** function.

# count duplicated observations

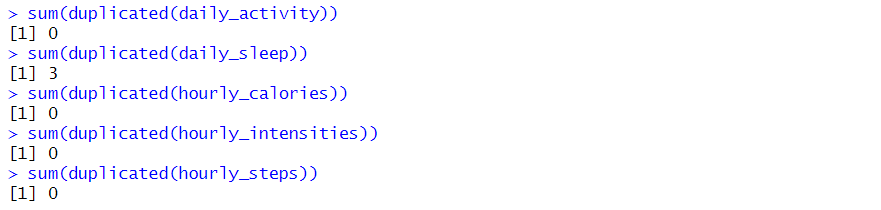
sum(duplicated(daily\_activity))

sum(duplicated(daily\_sleep))

sum(duplicated(hourly\_calories))

sum(duplicated(hourly\_intensities))

sum(duplicated(hourly\_steps))

****

I found three duplicated observations in **daily\_sleep** and removed them using the **distinct()** function.

# remove duplicated observations

daily\_sleep <- daily\_sleep %>%

distinct()

6. If the tracker is utilized, it should record at least one step in a day. Hence, I checked for observations with 0 steps in **daily\_activity.**

# count invalid observations

invalid\_steps <- daily\_activity %>%

filter(total\_steps==0)

nrow(invalid\_steps)

****

I found 77 observations with invalid steps value and removed them using the **filter()** function.

# remove invalid observations

daily\_activity <- daily\_activity %>%

filter(total\_steps!=0)

7. If the tracker works properly, the total minutes spent in four intensity levels of daily activity should not exceed 1440 minutes (24 hours). Thus, I checked for observations with total activity time greater than 1440 minutes in **daily\_activity** by temporarily adding the values from all intensity levels**.**

# create a column for total activity time

daily\_activity\_time <- daily\_activity %>%

mutate(total\_minutes = very\_active\_minutes +

fairly\_active\_minutes +

lightly\_active\_minutes +

sedentary\_minutes)

# count invalid observations

invalid\_time <- daily\_activity\_time %>%

filter(total\_minutes > 1440)

nrow(invalid\_time)

****

No observations with invalid time value were found.

**Transform the data**

1. I decided to combine the cleaned data frames with the same timeframe using the **inner\_join()** function.

I first merged **daily\_activity** and **daily\_sleep** into a new data frame named **daily\_activity\_sleep**. Then, I also merged **hourly\_calories, hourly\_intensities,** and **hourly\_steps** into a new data frame named **hourly\_activity.** This would return only the observations that have matching values.

# merge daily data frames

daily\_activity\_sleep <- inner\_join(daily\_activity, daily\_sleep, by=c("id", "date"))

# merge hourly data frames

hourly\_activity <- inner\_join(hourly\_calories, hourly\_intensities, by = c("id", "date\_time")) %>%

inner\_join(hourly\_steps, by = c("id", "date\_time"))

Let’s check the number of rows and users (unique id) in each data frame.

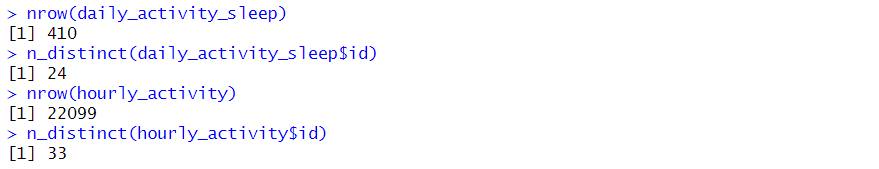
# count rows and users

nrow(daily\_activity\_sleep)

n\_distinct(daily\_activity\_sleep$id)

nrow(hourly\_activity)

n\_distinct(hourly\_activity$id)

****

With the sample size of **daily\_activity\_sleep** significantly reduced to match **daily\_sleep**, I would prefer using **daily\_activity** for finding activity trends due to its broad coverage. However, **daily\_activity\_sleep** would still be valuable for finding sleep trends and correlating activity and sleep variables. Meanwhile, since **hourly\_activity** retained the sample size of the three hourly data frames, I chose to alternatively use the former.

In short, I would use three data frames for my analysis: **daily\_activity, daily\_activity\_sleep,** and **hourly\_activity.**

2. I added some columns that would be helpful in analyzing the data frames using the **mutate()** function.

I first inserted a column for the day of the week (Monday to Sunday) named **day** in all data frames. To do that, I first split the **date\_time** column in **hourly\_activity** and cast the divided columns in their appropriate data types.

# separate date and time in hourly\_activity

hourly\_activity <- hourly\_activity %>%

separate(col = date\_time, into = c("date", "time"), sep = " ", remove = FALSE) %>%

mutate(date = ymd(date), time = hms(time))

# add a column for day

daily\_activity <- daily\_activity %>%

mutate(day = weekdays(date))

daily\_activity\_sleep <- daily\_activity\_sleep %>%

mutate(day = weekdays(date))

hourly\_activity <- hourly\_activity %>%

mutate(day = weekdays(date))

Then, I inserted a column for time spent in non-sedentary activity named **total\_active\_minutes** in **daily\_activity** and **daily\_activity\_sleep** computed as the sum of **very\_active\_minutes, fairly\_active\_minutes,** and **lightly\_active\_minutes.**

# add a column for non-sedentary activity

daily\_activity <- daily\_activity %>%

mutate(total\_active\_minutes = very\_active\_minutes +

fairly\_active\_minutes +

lightly\_active\_minutes)

daily\_activity\_sleep <- daily\_activity\_sleep %>%

mutate(total\_active\_minutes = very\_active\_minutes +

fairly\_active\_minutes +

lightly\_active\_minutes)

Lastly, I inserted a column for awake time in bed named **total\_minutes\_awake** in **daily\_activity\_sleep** computed as **total\_time\_in\_bed** minus **total\_minutes\_asleep**.

# add a column for awake time in bed

daily\_activity\_sleep <- daily\_activity\_sleep %>%

mutate(total\_minutes\_awake = total\_time\_in\_bed - total\_minutes\_asleep)

3. In contrast, I removed a few columns that would unlikely help in my analysis. Such columns are related to distance, which I believe would provide minimal value in understanding product usage and even user health profile.

# remove columns

daily\_activity <- daily\_activity [, -c(4:10)]

daily\_activity\_sleep <- daily\_activity\_sleep [, -c(4:10)]

4. I sorted the columns in each data frame in a custom order.

# sort columns

daily\_activity <- daily\_activity [, c(1, 2, 9, 8, 3, 10, 4:7)]

daily\_activity\_sleep <- daily\_activity\_sleep [, c(1, 2, 12, 8, 3, 13, 4:7, 11, 10, 14, 9)]

hourly\_activity <- hourly\_activity [, c(1:3, 9, 4:8)]

5. I sorted the rows in each data frame first by **id** and then by **date/date\_time.**

# sort rows

daily\_activity <- daily\_activity %>%

arrange(id, date)

daily\_activity\_sleep <- daily\_activity\_sleep %>%

arrange(id, date)

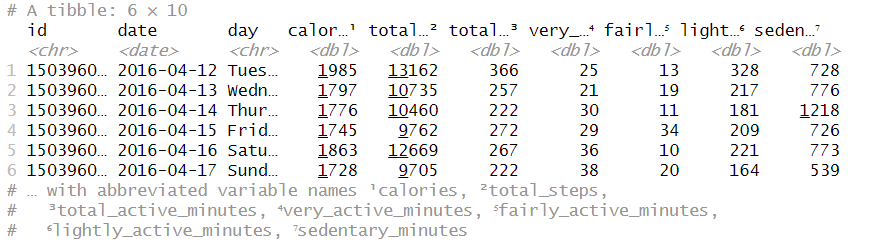
hourly\_activity <- hourly\_activity %>%

arrange(id, date\_time)

Here’s a preview of the **daily\_activity**:

# preview daily\_activity

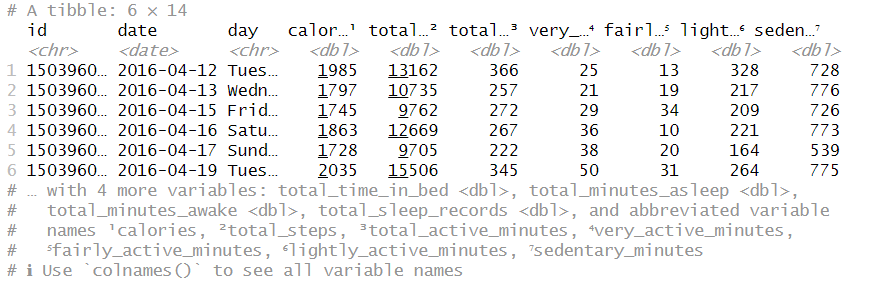
head(daily\_activity)

****

Here’s a preview of the **daily\_activity\_sleep**:

# preview daily\_activity\_sleep

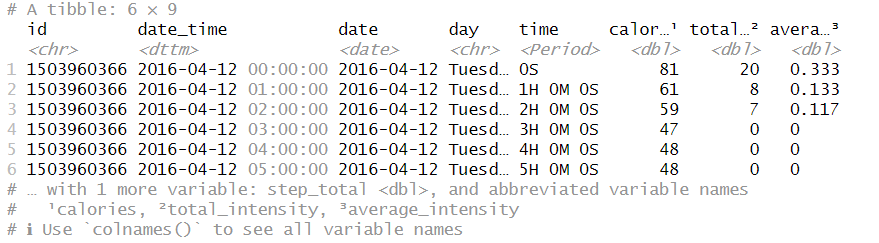
head(daily\_activity\_sleep)

****

Here’s a preview of the **hourly\_activity**:

# preview hourly\_activity

head(hourly\_activity)

****

## Step 4: Analyze and Share

**PRODUCT USAGE ANALYSIS**

Based on the available information in the data frames, I opted to explore the following trends and patterns in Leaf product usage:

* Usage rate of device features
* Usage rate by day in a week
* Usage rate by hour in a day
* Usage rate by hour in a week

**Usage rate of device features**

I created a horizontal bar chart to identify which Leaf smart device feature (Activity or Sleep Tracker) is used more often.

# create a data frame

usage\_by\_features <- data.frame(device\_feature = c("Activity Tracker", "Sleep Tracker"), no\_of\_users = c(n\_distinct(daily\_activity$id), n\_distinct(daily\_activity\_sleep$id)))

# create a horizontal bar chart

ggplot(usage\_by\_features, aes(x = no\_of\_users, y = device\_feature, fill = device\_feature)) +

geom\_bar(stat = "identity") +

geom\_text(aes(label = no\_of\_users), position = position\_stack(vjust = 0.5), color = "white", size = 10) +

labs(x = "Number of Users", y = "Device Feature") +

scale\_fill\_manual(values = c("#946597", "#9A7BB3")) +

scale\_y\_discrete(labels = c("Activity\nTracker", "Sleep\nTracker")) +

theme(panel.background = element\_blank(),

axis.title.x = element\_text(size = 25, color = "#6B2D7C", face = "bold"),

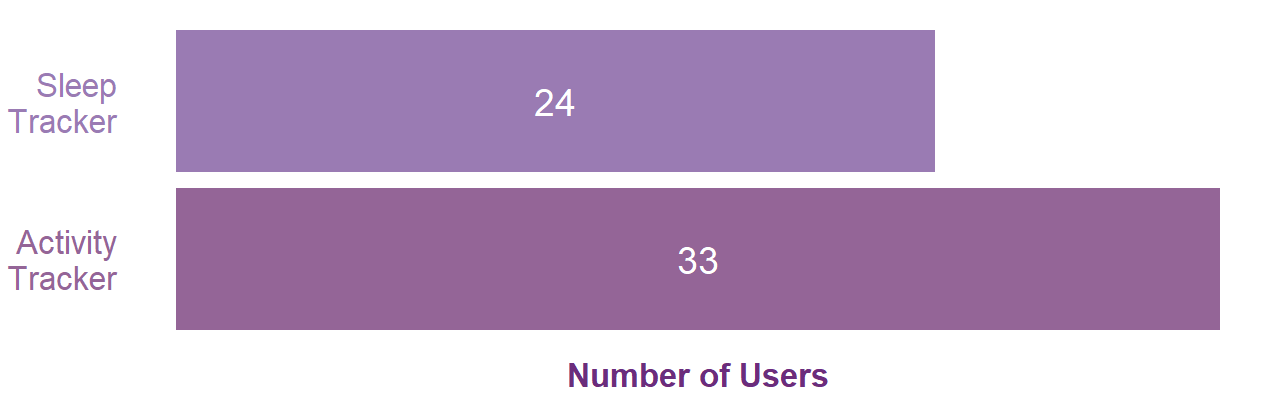
axis.title.y = element\_blank(),

axis.text.x = element\_blank(),

axis.text.y = element\_text(size = 25, color = c("#946597", "#9A7BB3")),

axis.ticks = element\_blank(),

legend.position = "none")



Based on this chart, the Leaf smart device's Activity Tracker feature (monitors calories, steps, and activity levels or intensities) appears to be more often used than its Sleep Tracker feature (monitors asleep and awake time in bed).

This finding suggests that users might be more interested or motivated in tracking their activity levels and monitoring their progress toward fitness or weight loss goals than tracking their sleep patterns.

Users might also perceive the Sleep Tracker feature as less useful or relevant to their overall health goals. They might believe that their sleep quality is not as critical to their health as their physical activity and therefore do not feel the need to track it as often.

Finally, the Leaf device's Activity Tracker feature might also be more user-friendly or convenient to use. Users might find it easier to remember to activate or find the activity data more accessible and useful.

**User type distribution**

I made a waffle chart to know how various Leaf users are distributed based on their usage level (high, mid, or low usage).

# create a data frame

user\_type <- daily\_activity %>%

count(id, name = "days\_used") %>%

mutate(usage\_level = cut(days\_used, c(0, 10, 20, 31), labels = c("Low Usage", "Mid Usage", "High Usage"))) %>%

count(usage\_level, name = "no\_of\_users") %>%

pivot\_wider(names\_from = usage\_level, values\_from = no\_of\_users) %>%

select("High Usage", "Mid Usage", "Low Usage")

# create a waffle chart

waffle(user\_type, row = 3, size = 1, colors = c("#946597", "#9A7BB3", "#E3CBDA"), legend = "left") +

labs(caption = "1 Box = 1 User") +

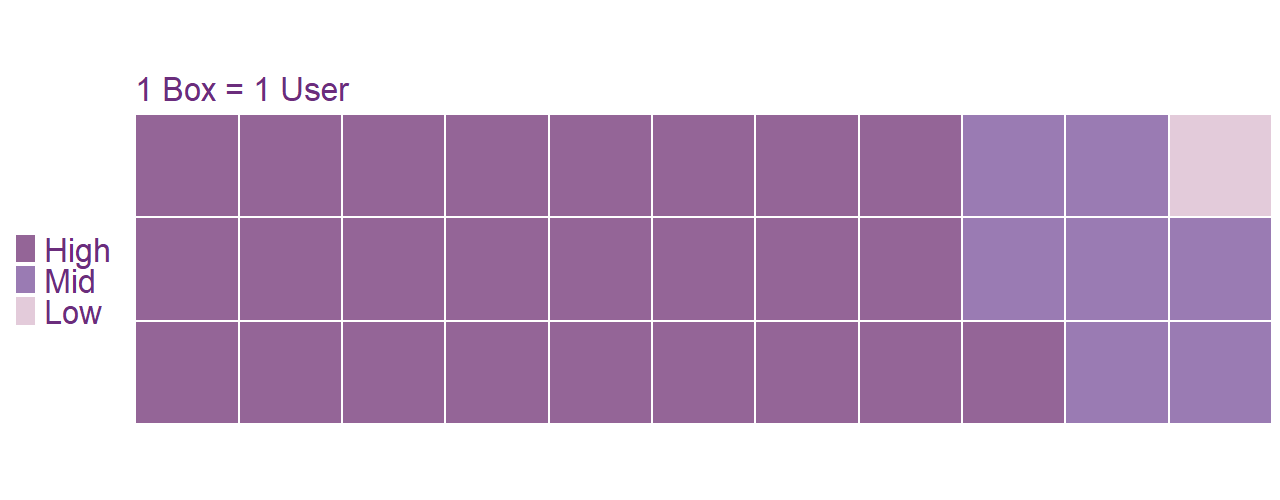
guides(fill = guide\_legend(title = "Usage Level", title.position = "top", title.hjust = 0.5, label.position = "bottom")) +

theme(plot.caption = element\_text(size = 15, color = "#6B2D7C"),

legend.position = "top",

legend.title = element\_text(size = 25, color = "#6B2D7C", face = "bold"),

legend.text = element\_text(size = 15, color = "#6B2D7C"))



Based on this diagram, the majority of Leaf smart device users (75%) demonstrated high-level device usage (21-31 days of usage per month), while 21% and 3% of the users showed mid-level (11-20 days) and low-level (0-10 days) device usage, respectively.

This finding indicates that users might be highly satisfied with the Leaf device's features and performance, leading them to use it regularly. It could be that the device is effectively fulfilling their health tracking and monitoring needs, motivating them to use it daily.

Moreover, users who use the Leaf device more frequently could have more specific health or fitness goals they want to achieve, and the device helps them monitor their progress toward those goals.

Lastly, the users who use the device more often could also have had a better onboarding experience or training when they first received the device. Perhaps they received more personalized guidance or instruction, leading them to use the device more frequently.

**Usage rate by day of the week**

I created a vertical bar chart to determine which day of the week is Leaf device usage highest and lowest.

# create a data frame

usage\_by\_day <- daily\_activity %>%

group\_by(date, day) %>%

summarise(no\_of\_users = n\_distinct(id)) %>%

group\_by(day) %>%

summarise(avg\_no\_of\_users = round(mean(no\_of\_users), 1))

# create a vertical bar chart

ggplot(usage\_by\_day, aes(x = day, y = avg\_no\_of\_users)) +

geom\_bar(aes(fill = day %in% c("Friday")), stat = "identity", width = 0.8) +

geom\_text(aes(label = avg\_no\_of\_users), position = position\_stack(vjust = 0.5), color = "white", size = 10, angle = 90, hjust = 0.5) +

labs(x = "Day", y = "Average Number of Users") +

scale\_fill\_manual(values = c("#9A7BB3", "#946597"), guide = FALSE) +

scale\_x\_discrete(limits = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")) +

theme(panel.background = element\_blank(),

panel.grid.major.x = element\_blank(),

panel.grid.major.y = element\_line(color = "grey"),

panel.grid.minor.x = element\_blank(),

panel.grid.minor.y = element\_line(color = "grey"),

axis.title.x = element\_blank(),

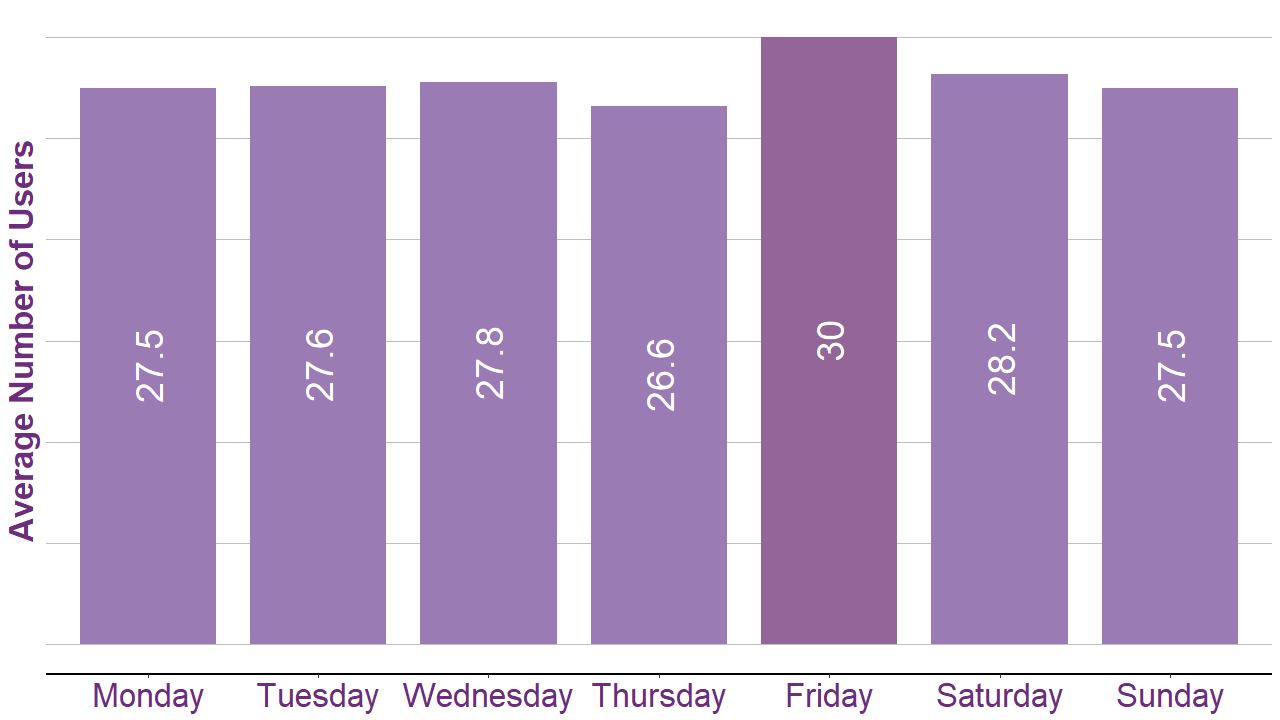
axis.title.y = element\_text(size = 25, color = "#6B2D7C", face = "bold"),

axis.text.x = element\_text(size = 25, color = "#6B2D7C"),

axis.text.y = element\_blank(),

axis.ticks.y = element\_blank(),

axis.line.x = element\_line(size = 1))



Based on this graph, the Leaf device appears to be consistently used each day of the week. However, there is slight a variation in usage pattern as the device is most frequently used on Fridays and least used on Thursdays.

Such a pattern signals that users could be more motivated to use the device on Fridays, knowing they have the weekend ahead and might want to ensure they are meeting their fitness goals or tracking their health metrics effectively. On the other hand, on Thursdays, users might be feeling more tired or less motivated to use the device since they have already been tracking their health for most of the week.

It is also possible that the pattern of device usage could be influenced by users' work or social schedules. For example, users might be more busy or active on Fridays, leading them to use the device more often, while they might have more commitments on Thursdays, leading to less device usage.

Lastly, some users might have formed habits around device usage on specific days of the week. For instance, some users might make it a habit to use the device more frequently on Fridays, as a way to ensure they meet their weekly fitness goals.

**Usage rate by hour in a day**

I made a line chart to track when Leaf device usage is highest and lowest throughout the day.

# create a data frame

usage\_by\_hour <- hourly\_activity %>%

group\_by(date\_time, hour = hour(date\_time)) %>%

summarise(no\_of\_users = n\_distinct(id)) %>%

group\_by(hour) %>%

summarise(avg\_no\_of\_users = round(mean(no\_of\_users), 1))

# create a vertical bar chart

ggplot(usage\_by\_hour, aes(x = hour, y = avg\_no\_of\_users)) +

labs(x = "Hour", y = "Average Number of Users") +

geom\_line(color="#9A7BB3", size=5) +

geom\_point(color="#946597", size=10) +

scale\_x\_continuous(breaks = seq(0, 23), labels = paste0(seq(0, 23))) +

ylim(28, 32) +

theme(panel.background = element\_blank(),

panel.grid.major.x = element\_line(color = "grey"),

panel.grid.major.y = element\_line(color = "grey"),

panel.grid.minor.x = element\_blank(),

panel.grid.minor.y = element\_line(color = "grey"),

axis.title.x = element\_text(size = 25, color = "#6B2D7C", face = "bold"),

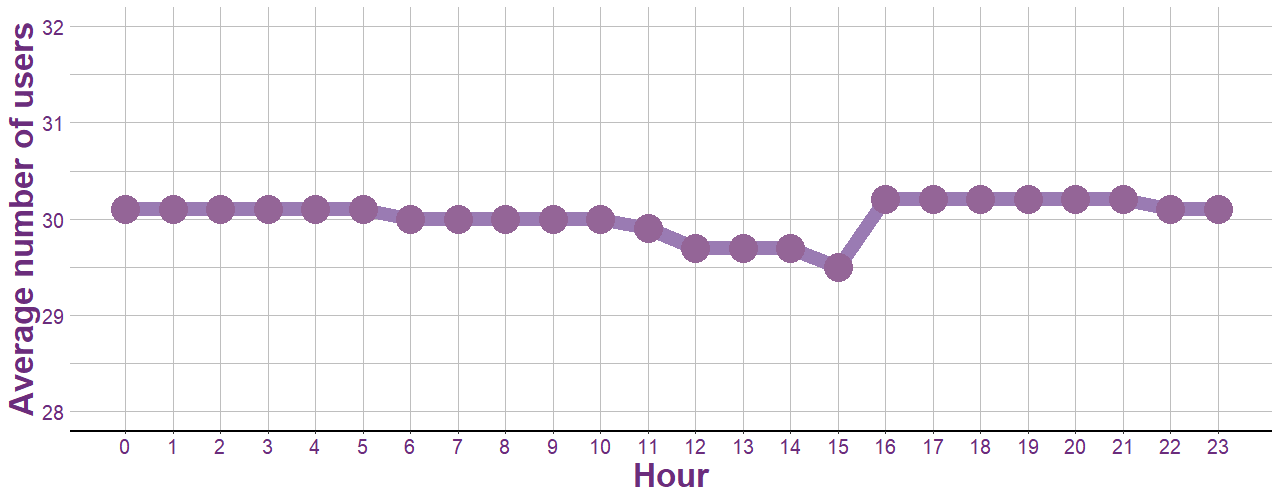
axis.title.y = element\_text(size = 25, color = "#6B2D7C", face = "bold"),

axis.text.x = element\_text(size = 15, color = "#6B2D7C"),

axis.text.y = element\_text(size = 15, color = "#6B2D7C"),

axis.ticks.y = element\_blank(),

axis.line.x = element\_line(size = 1))



Based on this illustration, the Leaf device is found to be consistently utilized throughout the day, with a gradual decline in usage from midnight to 3 PM followed by a spike in usage after 3 PM.

This trend indicates that users might be less likely to use the device during regular working hours and more likely to use the device after work or during leisure or free time.

Users might also be more inclined to engage in certain types of activities during certain times of the day. For instance, some users might prefer to engage in outdoor activities such as jogging or cycling in the late afternoon or early evening when the weather is cooler.

Certainly, Bellabeat's reminder notifications could also be influencing the pattern of device usage. Users might be receiving more reminder notifications during the spike in usage after 3 PM, leading to more device usage.

**USER WELLNESS PROFILE**

Although my primary task is to only analyze how users are utilizing the Leaf device, I believe understanding their health data can offer additional insights to inform the company’s marketing strategy. Therefore, I decided to also probe the following trends and patterns in user wellness status:

* Average calories burned by hour
* Average steps taken by hour
* Average active time by day
* Average time in bed by day

**Average calories burned by hour**

I created an area chart to identify what time of the day Leaf users burn the most and least calories.

# create a data frame

calories\_by\_hour <- hourly\_activity %>%

group\_by(date\_time, hour = hour(date\_time)) %>%

summarise(calories\_burned = mean(calories)) %>%

group\_by(hour) %>%

summarise(avg\_calories\_burned = mean(calories\_burned))

# compute max and min values for annotation

max\_row <- which.max(calories\_by\_hour$avg\_calories\_burned)

min\_row <- which.min(calories\_by\_hour$avg\_calories\_burned)

# create a line chart

ggplot(calories\_by\_hour, aes(x = hour, y = avg\_calories\_burned)) +

geom\_area(fill = "#FCC0C9", alpha = 0.7, color = "#C6687B", size = 5) +

labs(x = "Hour", y = "Average Calories Burned") +

scale\_x\_continuous(breaks = seq(0, 23), labels = paste0(seq(0, 23))) +

ylim(0,170) +

theme(panel.background = element\_blank(),

panel.grid.major = element\_line(color = "grey"),

panel.grid.minor.x = element\_blank(),

panel.grid.minor.y = element\_line(color = "grey"),

axis.title = element\_text(size = 25, color = "#94384A", face = "bold"),

axis.text = element\_text(size = 15, color = "#94384A"),

axis.ticks.y = element\_blank(),

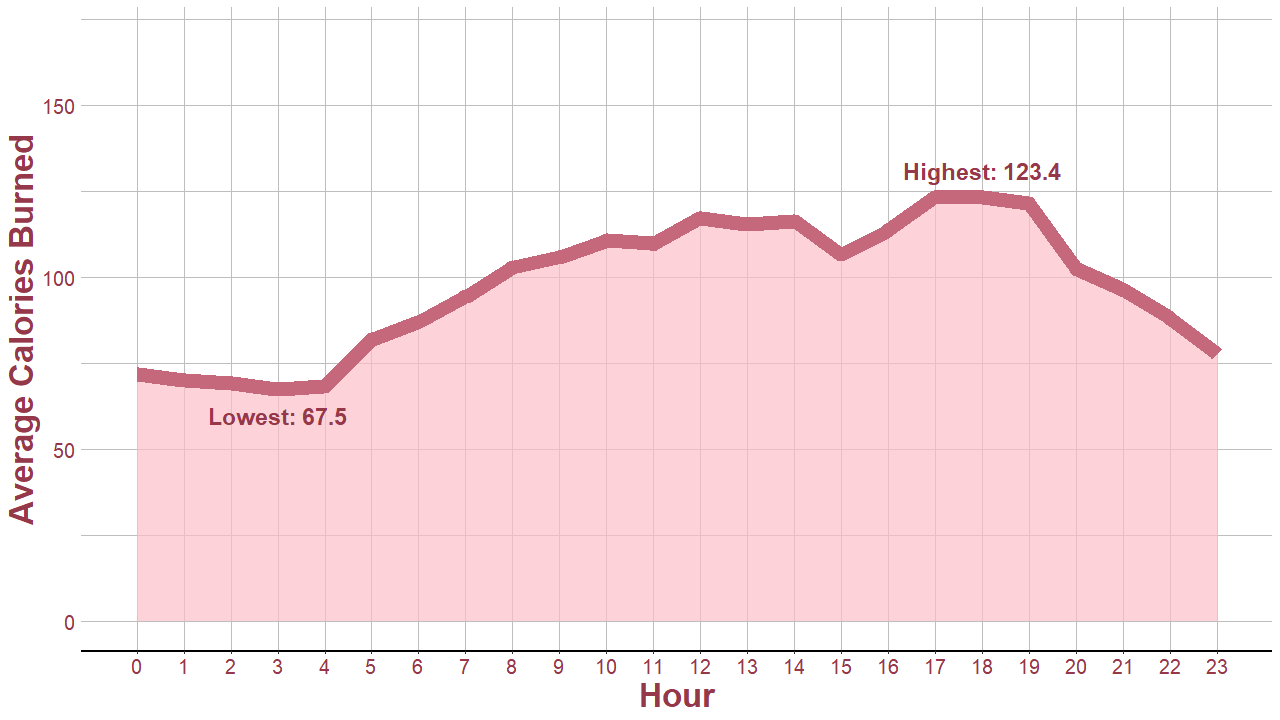
axis.line.x = element\_line(size = 1),

legend.position = "none") +

# add annotation

annotate("text", label = paste0("Highest: ", round(calories\_by\_hour$avg\_calories\_burned[max\_row], 1)), x = calories\_by\_hour$hour[max\_row], y = calories\_by\_hour$avg\_calories\_burned[max\_row], vjust = -1, color = "#94384A", size = 6, fontface = "bold") +

annotate("text", label = paste0("Lowest: ", round(calories\_by\_hour$avg\_calories\_burned[min\_row], 1)), x = calories\_by\_hour$hour[min\_row], y = calories\_by\_hour$avg\_calories\_burned[min\_row]-5, vjust = 1, color = "#94384A", size = 6, fontface = "bold")



Based on this visualization, Leaf users appear to burn fewer calories in the early morning hours and burn more calories in the early evening. This could be due to many factors, such as work schedules, leisure activities, or mealtimes.

Evidently, there is also a gradual increase in calorie expenditure throughout the day. This could be because users become more active as the day progresses brought by the accumulation of energy and motivation.

**Average steps taken by hour**

I made a heat map to see what time of the day Leaf users take the most and least number of steps.

# create a data frame

steps\_by\_hour <- hourly\_activity %>%

group\_by(hour = hour(date\_time), day = weekdays(date)) %>%

reframe(avg\_no\_of\_steps = mean(step\_total))

# create heat map

ggplot(steps\_by\_hour, aes(x = hour, y = day, fill = avg\_no\_of\_steps)) +

geom\_tile() +

labs(x = "Hour", y = "Day", fill = "Average Steps Taken") +

scale\_x\_continuous(breaks = seq(0, 23), labels = paste0(seq(0, 23))) +

scale\_y\_discrete(limits = c("Sunday", "Saturday", "Friday", "Thursday", "Wednesday", "Tuesday", "Monday")) +

scale\_fill\_gradient(low = "#FCC0C9", high = "#94384A", limits = c(2, 780)) +

guides(fill = guide\_colorbar(barwidth = 20, title.position = "top", title.hjust = 0.5,label.position = "bottom")) +

theme(panel.background = element\_blank(), panel.grid = element\_blank(),

axis.title.x = element\_text(size = 25, color = "#94384A", face = "bold"),

axis.title.y = element\_blank(),

axis.text.x = element\_text(size = 15, color = "#94384A"),

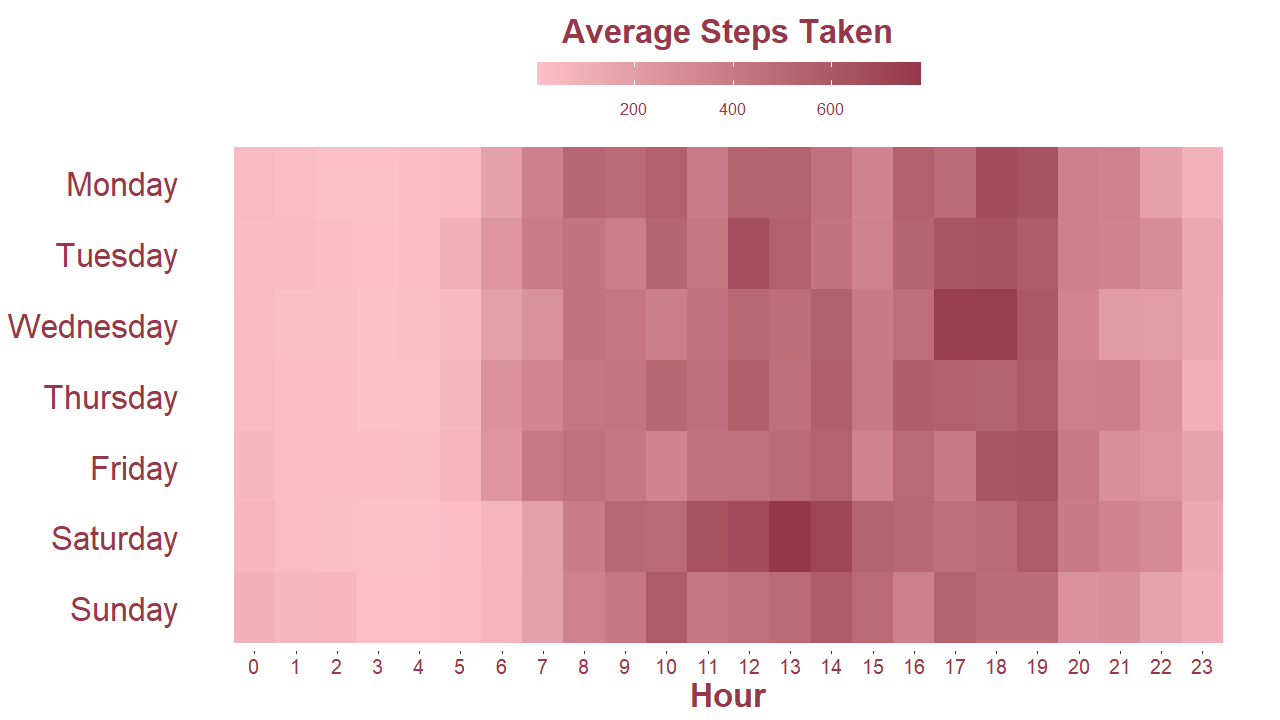
axis.text.y = element\_text(size = 25, color = "#94384A"),

axis.ticks.y = element\_blank(),

legend.title = element\_text(size = 25, color = "#94384A", face = "bold"),

legend.text = element\_text(size = 12, color = "#94384A"),

legend.position = "top")



Based on this heat map, it is apparent that Leaf users take the most steps between 8 AM and 7 PM and the least steps between 12 AM to 4 AM, proving they are more active during the daytime hours and less active during the nighttime hours. This is a common pattern among people who follow a typical workday schedule.

Furthermore, there is a significant increase in step count between 5 to 6 AM on weekdays compared to weekends, suggesting that users are more likely to engage in physical activities like commuting to work or preparing for the day on weekdays than on weekends when they have more leisure time.

**Average active time by day**

I created a vertical stacked bar chart to identify which day of the week Leaf users are most and least active (classified into very, moderately, and lightly active).

# create a data frame

active\_time\_by\_day <- daily\_activity %>%

group\_by(date, day) %>%

summarise(across(c(very\_active\_minutes, fairly\_active\_minutes, lightly\_active\_minutes), mean)) %>%

group\_by(day) %>%

summarise(across(c(very\_active\_minutes, fairly\_active\_minutes, lightly\_active\_minutes), ~ round(mean(.), 1), .names = "avg\_{.col}")) %>%

pivot\_longer(cols = starts\_with("avg\_"), names\_to = "activity\_intensity", values\_to = "avg\_active\_minutes")

# compute overall average for annotation

overall\_avg\_active <- mean(daily\_activity$total\_active\_minutes)

# create a vertical stacked bar chart

ggplot(active\_time\_by\_day, aes(x = day, y = avg\_active\_minutes, fill = activity\_intensity)) +

geom\_bar(stat = "identity", width = 0.8) +

labs(x = "Day", y = "Average Active Time (mins)", fill = "Activity Intensity") +

scale\_x\_discrete(limits = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")) +

scale\_fill\_manual(values = c("#EA7F8E", "#FCC0C9", "#B24459"), labels = c("Fairly Active", "Lightly Active", "Very Active")) +

guides(fill = guide\_legend(title.position = "top", title.hjust = 0.5, label.position = "bottom")) +

theme(panel.background = element\_blank(),

panel.grid.major.x=element\_blank(),

panel.grid.major.y=element\_line(color="grey"),

panel.grid.minor.x=element\_blank(),

panel.grid.minor.y=element\_line(color="grey"),

axis.title.x = element\_blank(),

axis.title.y = element\_text(size = 25, color = "#94384A", face = "bold"),

axis.text.x = element\_text(size = 25, color = "#94384A"),

axis.text = element\_text(size = 15, color = "#94384A"),

axis.ticks.y = element\_blank(),

axis.line.x = element\_line(size = 1),

legend.position = "top",

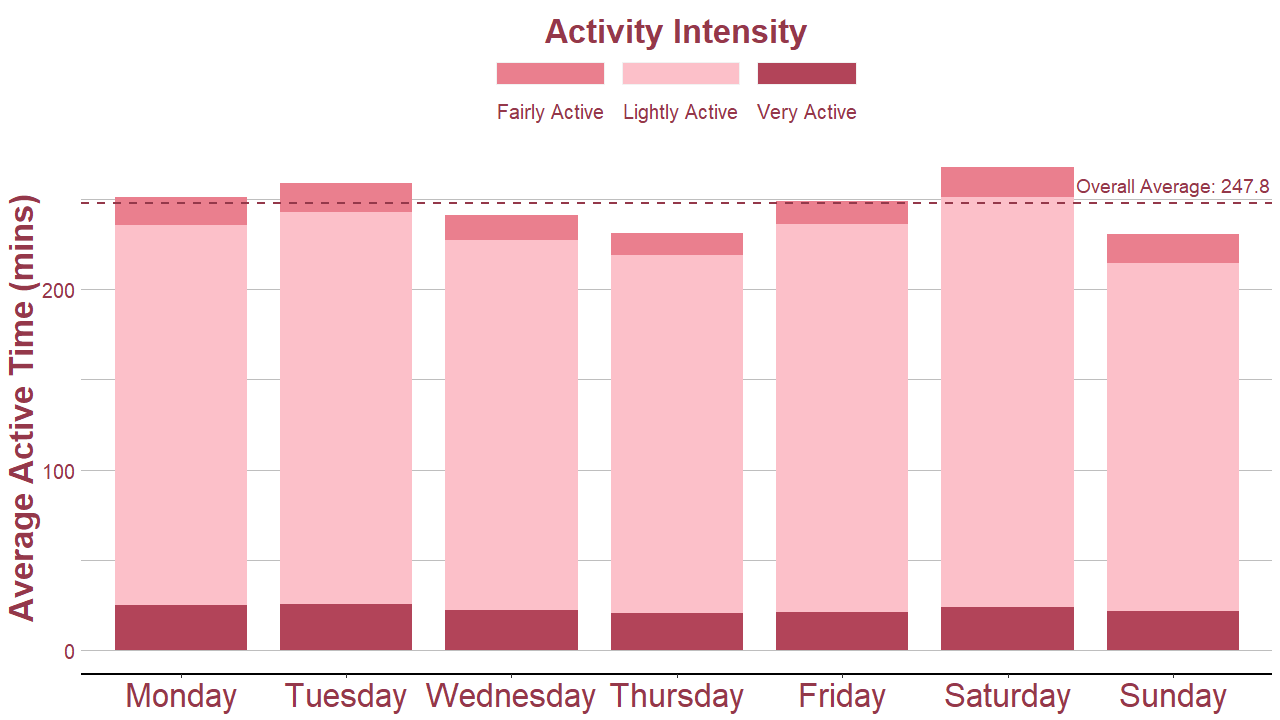
legend.title = element\_text(size = 25, color = "#94384A", face = "bold"),

legend.text = element\_text(size = 15, color = "#94384A")) +

# add annotation

annotate("text", label = paste0("Overall Average: ", round(overall\_avg\_active, 1)), x = 7, y = overall\_avg\_active + 10, color = "#94384A", size = 5) +

geom\_hline(yintercept = overall\_avg\_active, color = "#94384A", linetype = "dashed", size = 1)



Based on this chart, Leaf users tend to have a consistent pattern of activity throughout the week, with higher activity levels on Mondays, Tuesdays, and Saturdays, and lower activity levels on Wednesdays, Thursdays, and Sundays signifying that users might have a specific routine or schedule during these days that influence their activity levels.

Additionally, users also tend to mostly engage in light-intensity activities every day. This suggests that the Leaf device is more suitable for tracking low-intensity activities such as walking or light exercises rather than intense workouts.

**Average time in bed by day**

I created another vertical stacked bar chart to determine what day of the week Leaf users spend the most and least time in bed (including both awake and asleep time).

# create a data frame

bed\_time\_by\_day <- daily\_activity\_sleep %>%

group\_by(date, day) %>%

summarise(across(c(total\_minutes\_asleep, total\_minutes\_awake), mean)) %>%

group\_by(day) %>%

summarise(across(c(total\_minutes\_asleep, total\_minutes\_awake), ~ round(mean(.), 1), .names = "avg\_{.col}")) %>%

pivot\_longer(cols = starts\_with("avg\_"), names\_to = "asleep\_awake", values\_to = "avg\_bedtime\_minutes")

# compute overall average for annotation

overall\_avg\_bedtime <- mean(daily\_activity\_sleep$total\_time\_in\_bed)

# create a vertical stacked bar chart

ggplot(bed\_time\_by\_day, aes(x = day, y = avg\_bedtime\_minutes, fill = asleep\_awake)) +

geom\_bar(stat = "identity", width = 0.8) +

labs(x = "Day", y = "Average Time in Bed (mins)", fill = "Asleep or Awake?") +

scale\_x\_discrete(limits = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")) +

scale\_fill\_manual(values = c("#FCC0C9", "#B24459"), labels = c(" Time Asleep ", " Time Awake ")) +

guides(fill = guide\_legend(title.position = "top", title.hjust = 0.5, label.position = "bottom")) +

theme(panel.background = element\_blank(),

panel.grid.major.x=element\_blank(),

panel.grid.major.y=element\_line(color="grey"),

panel.grid.minor.x=element\_blank(),

panel.grid.minor.y=element\_line(color="grey"),

axis.title.x = element\_blank(),

axis.title.y = element\_text(size = 25, color = "#94384A", face = "bold"),

axis.text.x = element\_text(size = 25, color = "#94384A"),

axis.text = element\_text(size = 15, color = "#94384A"),

axis.ticks.y = element\_blank(),

axis.line.x = element\_line(size = 1),

legend.position = "top",

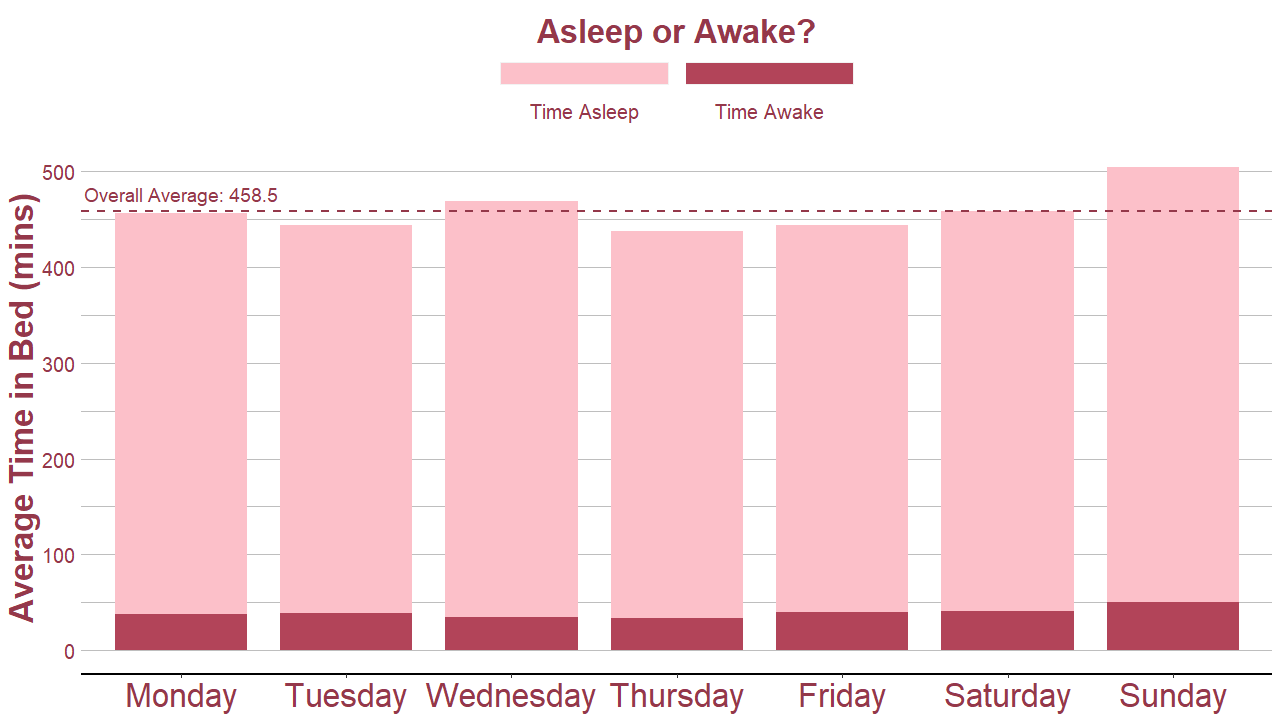
legend.title = element\_text(size = 25, color = "#94384A", face = "bold"),

legend.text = element\_text(size = 15, color = "#94384A")) +

# add annotation

annotate("text", label = paste0("Overall Average: ", round(overall\_avg\_bedtime, 1)), x = 1, y = overall\_avg\_bedtime + 18, color = "#94384A", size = 5) +

geom\_hline(yintercept = overall\_avg\_bedtime, color = "#94384A", linetype = "dashed", size = 1)



Based on this graph, Leaf users' sleep patterns vary throughout the week. They tend to have longer bedtime and sleep durations on Wednesdays and Sundays, indicating that these are the days when they have fewer commitments or responsibilities, allowing them to get more rest.

Conversely, they tend to have shorter bedtime and sleep durations on Tuesdays, Thursdays, and Fridays, suggesting that users may have busier schedules or may be prioritizing other activities over sleep during these days.

**FINDING CORRELATIONS**

Additionally, I decided to further investigate the relationships between various health indicators in the dataset to more or less verify Leaf’s accuracy in recording user health data.

To begin with, I performed a Pearson correlation test for the numeric variables in **daily\_activity\_sleep** and **hourly\_activity** to identify the variables that exhibited a strong relationship. Then, I transformed the test results into a correlation plot for practical purposes.

Here’s the Pearson correlation test for **daily\_activity\_sleep**:

# calculate correlation coefficients

corr\_daily\_activity\_sleep <- daily\_activity\_sleep[, 4:13] %>%

cor(method = "pearson")

# create a correlation plot

corrplot(corr\_daily\_activity\_sleep, type = "upper", order = "hclust",

col = colorRampPalette(c("#b2182b", "#ef8a62", "#fddbc7", "#d1e5f0", "#67a9cf", "#2166ac"), space = "rgb")(100),

tl.col = "black",

tl.cex = 0.8,

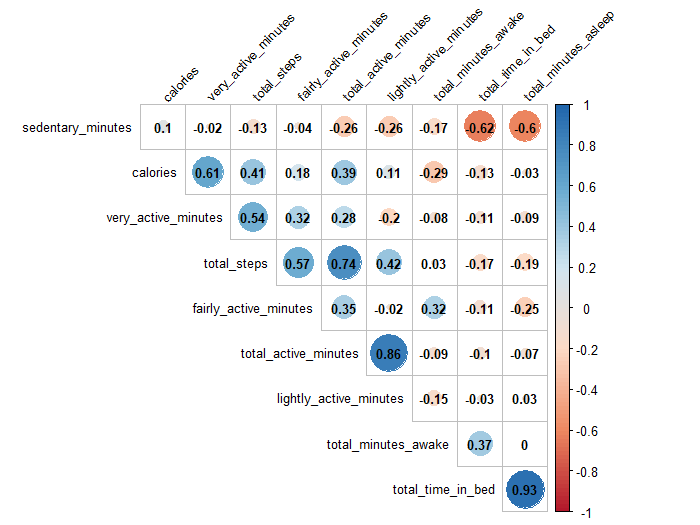
tl.srt = 45,

addCoef.col = "black",

number.cex = 0.8,

diag = FALSE,

method = "circle")



And here’s the Pearson correlation test for **hourly\_activity:**

# calculate correlation coefficients

corr\_hourly\_activity <- hourly\_activity[, 6:9] %>%

cor(method = "pearson")

# create a correlation plot

corrplot(corr\_hourly\_activity, type = "upper", order = "hclust",

col = colorRampPalette(c("#b2182b", "#ef8a62", "#fddbc7", "#d1e5f0", "#67a9cf", "#2166ac"), space = "rgb")(100),

tl.col = "black",

tl.cex = 0.8,

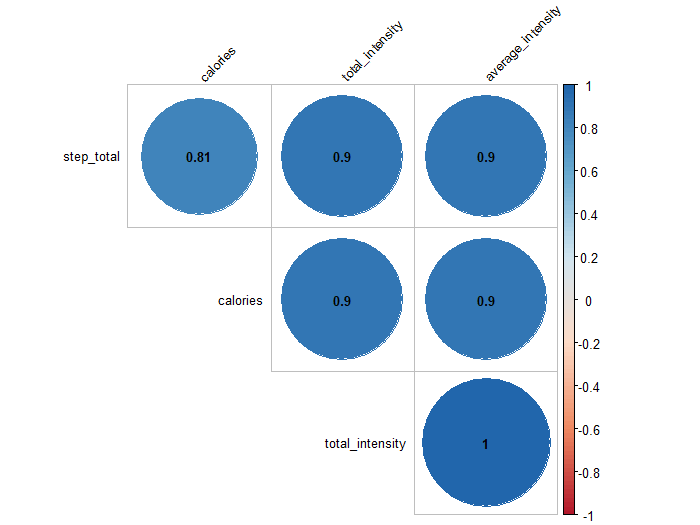
tl.srt = 45,

addCoef.col = "black",

number.cex = 0.8,

diag = FALSE,

method = "circle")



The size and color of the circles reflect the strength and direction of the relationship between the health indicators. A large dark blue circle indicates a strong positive correlation (as one variable increases, the other variable tends to increase as well). Conversely, a large red circle indicates a strong negative correlation (as one variable increases, the other variable tends to decrease).

Based on the test results, I decided to map out the relationship between the following pairs of variables due to their relatively strong correlation:

* Daily sedentary time and sleep time
* Daily active time and steps
* Hourly steps and calories
* Hourly intensity level and calories

**Daily sedentary time and sleep time**

The chart below illustrates the moderate negative correlation (-0.6) between daily sedentary time and sleep time, indicating that as sedentary time increases, sleep time tends to decrease.

# compute correlation coefficient

cor\_coef\_1 <- cor(daily\_activity\_sleep$sedentary\_minutes, daily\_activity\_sleep$total\_minutes\_asleep)

# create a scatter plot

ggplot(data = daily\_activity\_sleep, aes(x = sedentary\_minutes, y = total\_minutes\_asleep, color = total\_minutes\_asleep)) +

geom\_point(size = 6) +

geom\_smooth(color = "#A5420B", size = 2) +

labs(x = "Daily Sedentary Time (mins)", y = "Daily Sleep Time (mins)") +

scale\_color\_gradient(low = "#962212", high = "#F09083") +

theme(panel.background = element\_blank(),

panel.grid.major = element\_line(color = "#D3D3D3"),

panel.grid.minor = element\_line(color="#D3D3D3"),

axis.title = element\_text(size = 25, color = "#962212", face = "bold"),

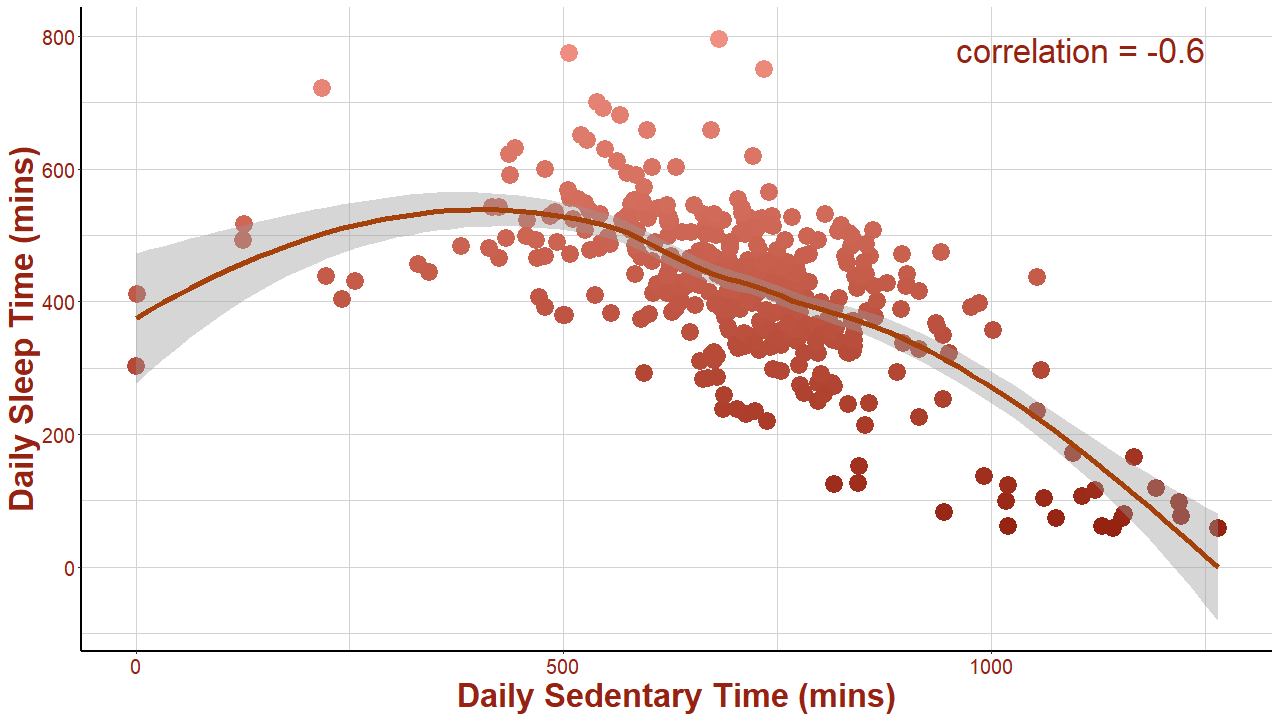
axis.text = element\_text(size = 15, color = "#962212"),

axis.line = element\_line(size = 1),

legend.position = "none") +

# add annotation

annotate(geom = "text", label = paste0("correlation = ", round(cor\_coef\_1, 2)), x = 1250, y = 800, hjust = 1, vjust = 1, color = "#962212", size = 9)



To determine if this correlation is significant, I performed a t-test with a significance level of 0.05 using the **cor.test()** function. Let the null (H0) and alternative (Ha) hypotheses be the following:

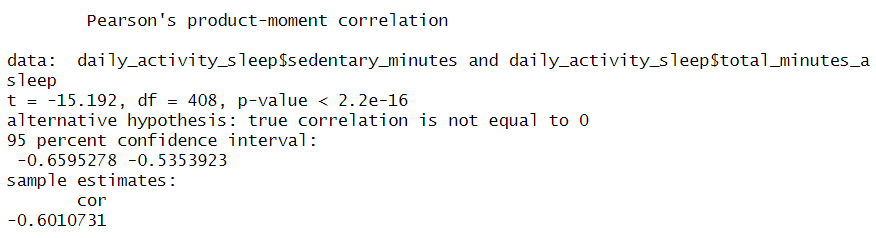
**H0** – The population correlation coefficient is zero (no correlation).

**Ha** – The population correlation coefficient is not zero (there is a correlation).

# perform t-test

cor.test(daily\_activity\_sleep$sedentary\_minutes, daily\_activity\_sleep$total\_minutes\_asleep, significance.level = 0.05)

The t-test would calculate the t-statistic and its corresponding p-value, which is important to ascertain if the correlation is significant. This is the result of the t-test:



Since the p-value (< 2.2e-16) is less than the significance level, we can reject the null hypothesis and conclude that daily sedentary time and sleep time are significantly correlated.

This negative correlation is consistent with studies stating that a sedentary lifestyle is known to have detrimental effects on sleep quality and duration. Hence, it can be inferred that Leaf's tracking features for sedentary time and sleep time are precise and offer meaningful information for users who wish to enhance their health.

**Daily active time and steps**

The chart below shows a strong positive correlation (0.74) between daily active time and steps, hinting that as active time increases, the number of steps taken also tends to increase.

# compute correlation coefficient

cor\_coef\_2 <- cor(daily\_activity\_sleep$total\_active\_minutes, daily\_activity\_sleep$total\_steps)

# create a scatter plot

ggplot(data = daily\_activity\_sleep, aes(x = total\_active\_minutes, y = total\_steps, color = total\_steps)) +

geom\_point(size = 6) +

geom\_smooth(color = "#A5420B", size = 2) +

labs(x = "Daily Active Time (mins)", y = "Daily Steps Taken") +

scale\_color\_gradient(low = "#962212", high = "#F09083") +

theme(panel.background = element\_blank(),

panel.grid.major = element\_line(color = "#D3D3D3"),

panel.grid.minor = element\_line(color="#D3D3D3"),

axis.title = element\_text(size = 25, color = "#962212", face = "bold"),

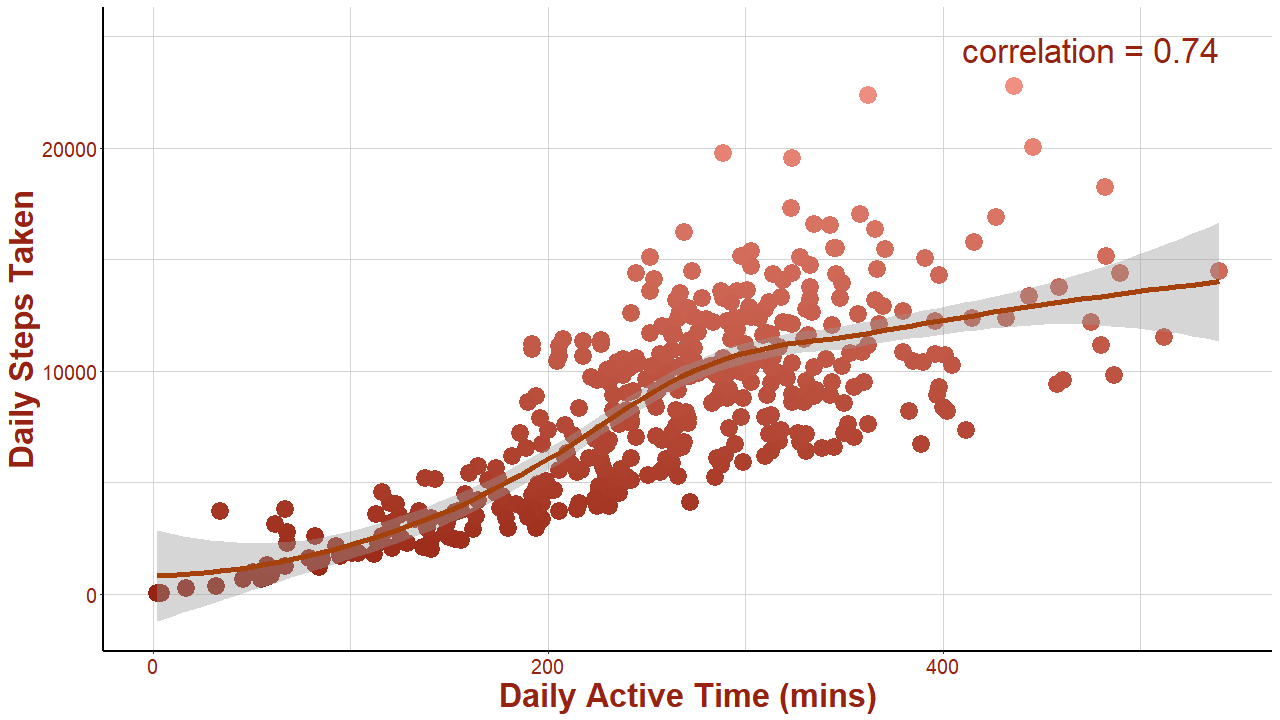
axis.text = element\_text(size = 15, color = "#962212"),

axis.line = element\_line(size = 1),

legend.position = "none") +

# add annotation

annotate(geom = "text", label = paste0("correlation = ", round(cor\_coef\_2, 2)), x = 540, y = 25000, hjust = 1, vjust = 1, color = "#962212", size = 9)

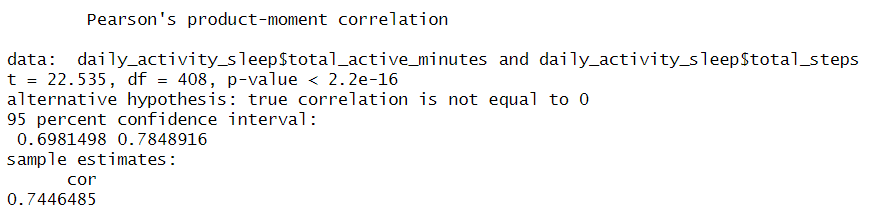


To determine if this correlation is significant, I performed the same procedure above.

# perform t-test

cor.test(daily\_activity\_sleep$total\_active\_minutes, daily\_activity\_sleep$total\_steps, significance.level = 0.05)

This is the result of the t-test:



Since the p-value (< 2.2e-16) is less than the significance level, we can reject the null hypothesis and conclude that daily active time and steps are significantly associated.

This positive association is not surprising as activities such as walking, jogging, or running, which are typical forms of physical activity, require more steps to be taken. This suggests that Leaf's active time and steps tracking features are accurate and can provide valuable insights for users looking to improve their health.

**Hourly steps and calories**

The chart below displays the strong positive correlation (0.81) between hourly steps and calories, implying that as the number of steps taken increases, the number of calories burned also tends to increase.

# compute correlation coefficient

cor\_coef\_3 <- cor(hourly\_activity$step\_total, hourly\_activity$calories)

# create a scatter plot

ggplot(data = hourly\_activity, aes(x = step\_total, y = calories, color = step\_total)) +

geom\_point(size = 6) +

geom\_smooth(color = "#A5420B", size = 2) +

labs(x = "Hourly Steps Taken", y = "Calories Burned (kcal)") +

scale\_color\_gradient(low = "#962212", high = "#F09083") +

theme(panel.background = element\_blank(),

panel.grid.major = element\_line(color = "#D3D3D3"),

panel.grid.minor = element\_line(color="#D3D3D3"),

axis.title = element\_text(size = 25, color = "#962212", face = "bold"),

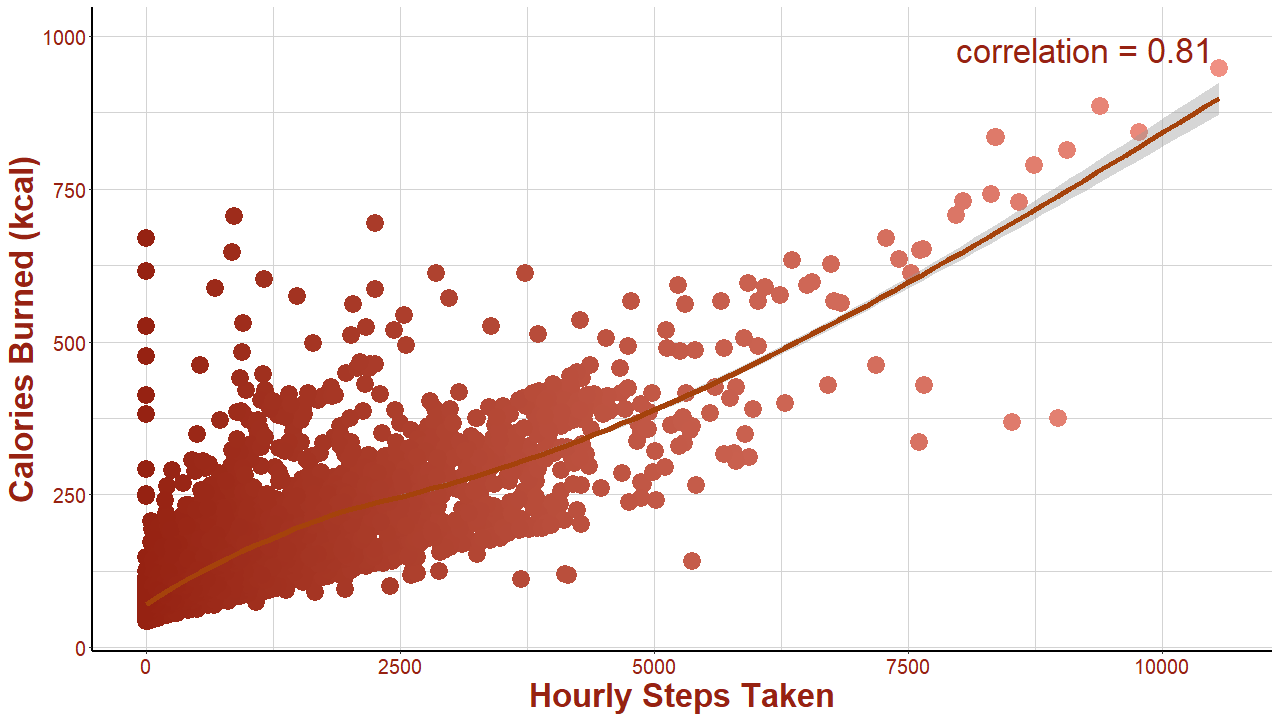
axis.text = element\_text(size = 15, color = "#962212"),

axis.line = element\_line(size = 1),

legend.position = "none") +

# add annotation

annotate(geom = "text", label = paste0("correlation = ", round(cor\_coef\_3, 2)), x = 10500, y = 1000, hjust = 1, vjust = 1, color = "#962212", size = 9)

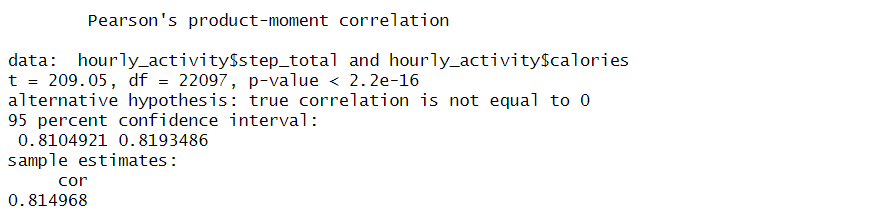


To determine if this correlation is significant, I performed the same procedure above.

# perform t-test

cor.test(hourly\_activity$step\_total, hourly\_activity$calories, significance.level = 0.05)

This is the result of the t-test:



Since the p-value (< 2.2e-16) is less than the significance level, we can reject the null hypothesis and conclude that hourly steps and calories are significantly related.

This positive relationship is expected given that the number of steps taken is a key indicator of physical activity, and calorie expenditure is directly related to the amount of physical activity. This supports the precision of the device's measurement of hourly steps and calories.

**Hourly intensity level and calories**

The chart below portrays the strong positive correlation (0.9) between hourly average intensity level and calories, indicating that as the average intensity level of physical activity increases, the number of calories burned also tends to increase.

# compute correlation coefficient

cor\_coef\_4 <- cor(hourly\_activity$average\_intensity, hourly\_activity$calories)

# create a scatter plot

ggplot(data = hourly\_activity, aes(x = average\_intensity, y = calories, color = average\_intensity)) +

geom\_point(size = 6) +

geom\_smooth(color = "#A5420B", size = 2) +

labs(x = "Average Intensity Level", y = "Calories Burned (kcal)") +

scale\_color\_gradient(low = "#962212", high = "#F09083") +

theme(panel.background = element\_blank(),

panel.grid.major = element\_line(color = "#D3D3D3"),

panel.grid.minor = element\_line(color="#D3D3D3"),

axis.title = element\_text(size = 25, color = "#962212", face = "bold"),

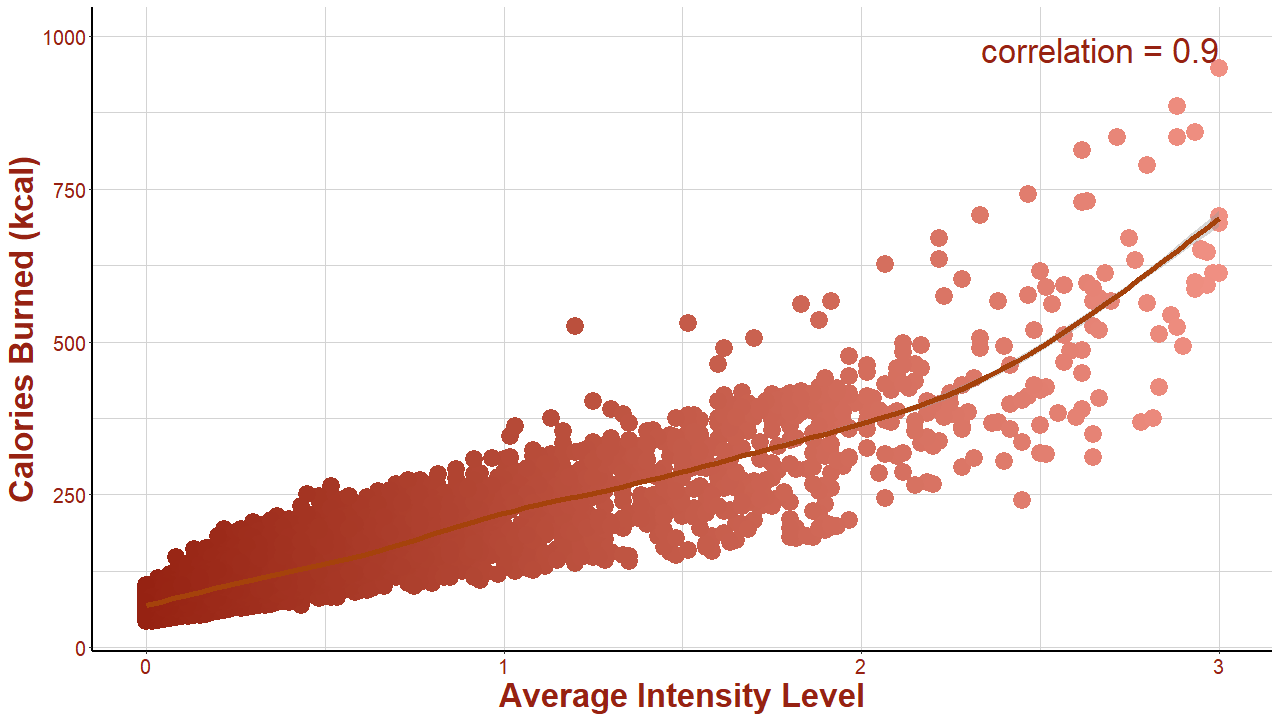
axis.text = element\_text(size = 15, color = "#962212"),

axis.line = element\_line(size = 1),

legend.position = "none") +

# add annotation

annotate(geom = "text", label = paste0("correlation = ", round(cor\_coef\_4, 2)), x = 3, y = 1000, hjust = 1, vjust = 1, color = "#962212", size = 9)

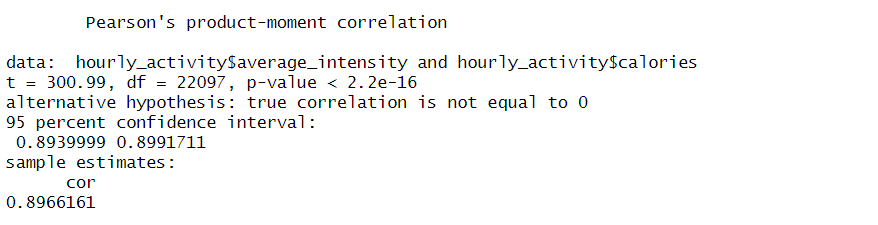


To determine if this correlation is significant, I performed the same procedure above.

# perform t-test

cor.test(hourly\_activity$average\_intensity, hourly\_activity$calories, significance.level = 0.05)

This is the result of the t-test:



Since the p-value (< 2.2e-16) is less than the significance level, we can reject the null hypothesis and conclude that hourly intensity level and calories are significantly connected.

This positive connection is not surprising because the body demands more energy to carry out activities at a higher intensity level. This can mean that the Leaf device is correctly estimating the intensity of physical activity and the number of calories burned.

**Note:** It is important to understand that correlation does not necessarily mean causation, and further analysis is needed to determine the exact nature of the relationship between these variables.

## Step 5: Act

Based on these insights, I would suggest the following approaches to help Bellabeat improve the Leaf smart device and expand its market reach:

**Focus on the Activity Tracker feature**

Given that users seem to use the Activity Tracker feature more frequently, Bellabeat could focus more on marketing this feature and highlighting its benefits to attract more customers. They could also improve the user interface and make it more user-friendly and interactive.

**Enhance onboarding experience**

Since users who use the device more frequently could have had a better onboarding experience, Bellabeat could invest in creating a personalized onboarding experience for new users. This could include offering online tutorials, personalized coaching, and instructional videos to help users get the most out of their devices.

**Target Fridays**

Bellabeat could create promotions or discounts to incentivize users to use the device more frequently on Fridays, which seems to be the day with the highest device usage. This could encourage users to stay motivated and keep up with their fitness goals throughout the weekend.

**Optimize reminder notifications**

To ensure that users are reminded to use the device more often during the day, Bellabeat could optimize the reminder notifications to coincide with the times when users are most active. This could lead to increased device usage and more engagement with the app.

**Highlight calorie tracking**

Given that Leaf users tend to burn more calories in the early evening, Bellabeat could highlight this feature and create marketing campaigns that focus on tracking and monitoring calorie intake and expenditure during different times of the day.

**Promote weekday activities**

Since Leaf users tend to be more active on weekdays, Bellabeat could create campaigns that encourage users to engage in more physical activities during the weekdays. This could include creating challenges, incentives, or workout routines that users could do during their work breaks or after work.

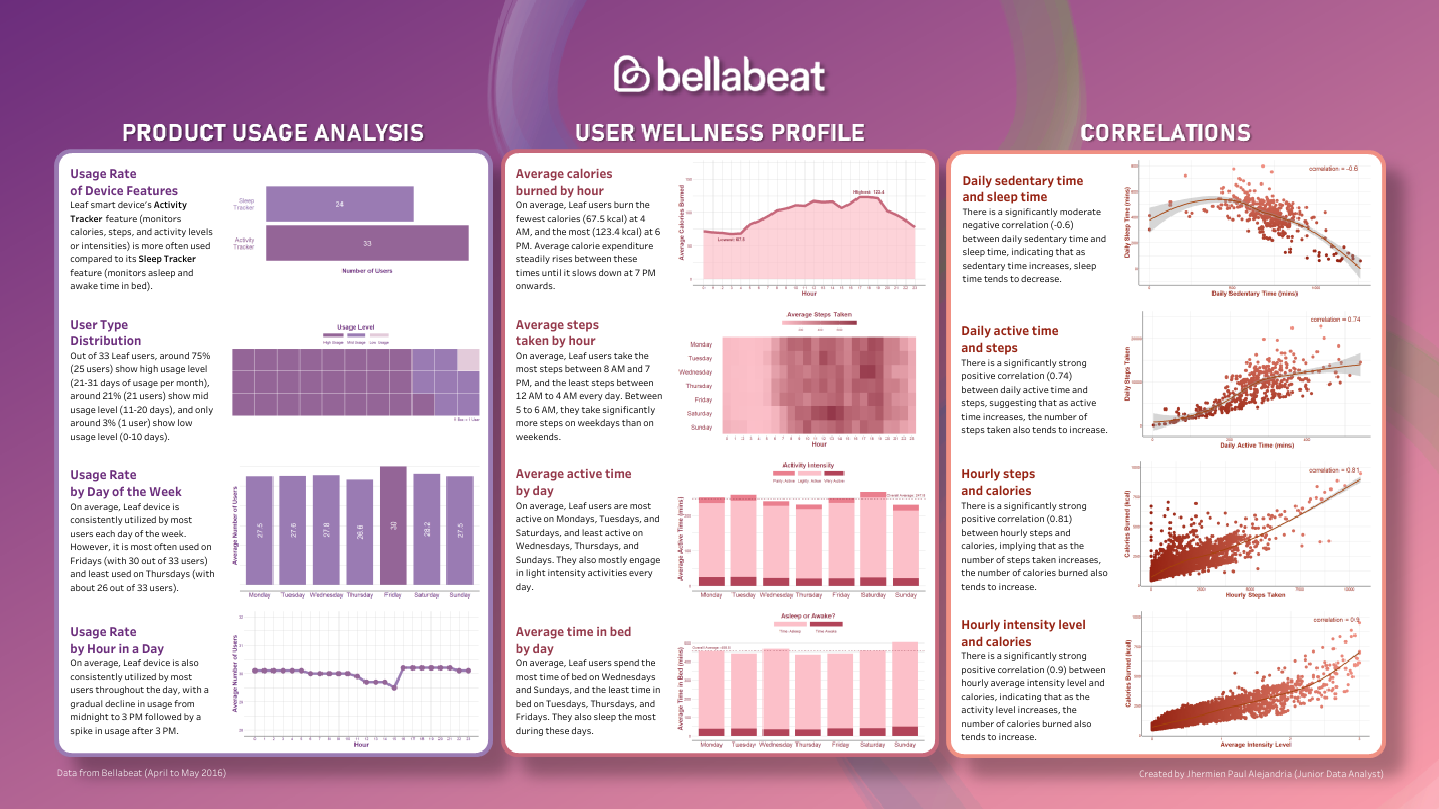
**Emphasize low-intensity activities**

Since Leaf users tend to engage in light-intensity activities, Bellabeat could emphasize this feature and market the device as a low-impact fitness tracker suitable for people who prefer less strenuous activities like walking or yoga.

**Focus on sleep tracking**

Given that users tend to have longer bedtime and sleep durations on Wednesdays and Sundays, Bellabeat could highlight the Sleep Tracker feature and promote the benefits of tracking and monitoring sleep patterns to improve overall health and wellness. They could also improve the user interface of this feature to make it more engaging and informative.

**Note**: My findings and recommendations are based on Bellabeat’s available data, which is subject to several limitations I indicated earlier. The stakeholders are advised to interpret these with caution and consider the limitations when making business decisions.



*This project was completed as part of the requirements for the* [***Google Data Analytics Professional Certificate***](https://www.coursera.org/professional-certificates/google-data-analytics)*program.*