Bellatbeat_Case_Study_md

Leon

2025-06-25

1.Backgorund

Bellabeat is a cutting-edge company specializing in smart products designed to support women's health and wellness. Drawing on her artistic background, Sršen created elegantly crafted technology that educates and motivates women globally. By gathering data on activity, sleep, stress, and reproductive health, Bellabeat empowers women to better understand their bodies and lifestyles. Since its launch in 2013, the company has experienced rapid growth, establishing itself as a leading tech-focused wellness brand for women.

Sršen believes that analyzing Bellabeat's existing consumer data could uncover valuable opportunities for growth. She has tasked the marketing analytics team with examining usage data from one of Bellabeat's smart products to better understand how users are currently engaging with their devices. Based on these insights, she is seeking strategic, high-level recommendations on how emerging usage trends can shape and enhance Bellabeat's marketing approach.

2.Ask Phase

Business task: Identify potential opportunities for growth and recommendations for the Bellabeat marketing strategy improvement based on trends in smart device usage.

3.Prepare Phase

Dataset used:

The data source used for our case study is FitBit Fitness Tracker Data. This dataset is stored in Kaggle and was made available through Mobius

Table Name	Type	Description
dailyActivity_merged	Microsoft Excel CSV	Daily Activity over 31 days of 33 users. Tracking daily: Steps, Distance, Intensities, Calories
dailyCalories_merged	Microsoft Excel CSV	Daily Calories over 31 days of 33 users

Table Name	Type	Description
dailyIntensities_merged	Microsoft Excel CSV	Daily Intensity over 31 days of 33 users. Measured in Minutes and Distance, dividing groups in 4 categories: Sedentary, Lightly Active, Fairly Active, Very Active
dailySteps_merged	Microsoft Excel CSV	Daily Steps over 31 days of 33 users
heartrate_seconds_merged	Microsoft Excel CSV	Exact day and time heartrate logs for just 7 users
hourlyCalories_merged	Microsoft Excel CSV	Hourly Calories burned over 31 days of 33 users
hourlyIntensities_merged	Microsoft Excel CSV	Hourly total and average intensity over 31 days of 33 users
hourlySteps_merged	Microsoft Excel CSV	Hourly Steps over 31 days of 33 users
minuteCaloriesNarrow_merged	Microsoft Excel CSV	Calories burned every minute over 31 days of 33 users (Every minute in single row)
minuteCaloriesWide_merged	Microsoft Excel CSV	Calories burned every minute over 31 days of 33 users (Every minute in single column)
minuteIntensitiesNarrow_merged	Microsoft Excel CSV	Intensity counted by minute over 31 days of 33 users (Every minute in single row)
minuteIntensitiesWide_merged	Microsoft Excel CSV	Intensity counted by minute over 31 days of 33 users (Every minute in single column)
minuteMETsNarrow_merged	Microsoft Excel CSV	Ratio of the energy you are using in a physical activity compared to the energy you would use at rest. Counted in minutes

Table Name	Type	Description
minuteSleep_merged	Microsoft Excel CSV	Log Sleep by Minute for 24 users over 31 days. Value column not specified
minuteStepsNarrow_merged	Microsoft Excel CSV	Steps tracked every minute over 31 days of 33 users (Every minute in single row)
minuteStepsWide_merged	Microsoft Excel CSV	Steps tracked every minute over 31 days of 33 users (Every minute in single column)
sleepDay_merged	Microsoft Excel CSV	Daily sleep logs, tracked by: Total count of sleeps a day, Total minutes, Total Time in Bed
weightLogInfo_merged	Microsoft Excel CSV	Weight track by day in Kg and Pounds over 30 days. Calculation of BMI. 5 users report weight manually, 3 users not. In total there are 8 users

Data Credibility and Integrity:

Given the small sample size (30 users) and lack of demographic data, the dataset may suffer from sampling bias, making it unclear whether the sample accurately represents the broader population. Additionally, the data is not recent, and the observation period is limited to just two months. Due to these constraints, our analysis will take an operational approach rather than attempting broad generalizations.

4.Process Phase

Loading packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
```

Importing datasets

Given the available datasets, we will upload those most relevant to addressing our business objectives. Our analysis will concentrate on the following datasets: * dailyActivity_merged * daily_intensities * daily_step * daily_sleep * hourlyIntensities_merged * hourlySteps_merged

```
daily_activity <- read.csv("data/dailyActivity_merged.csv")
daily_intensities <- read.csv("data/daily_intensities.csv")
daily_step <- read.csv("data/daily_step.csv")
daily_sleep <- read.csv("data/daily_sleep.csv")
hour_intensities <- read.csv("data/hourlyIntensities_merged.csv")
hour_steps <- read.csv("data/hourlySteps_merged.csv")
```

Standardize column names, Remove duplicates and NA

```
daily_activity <- clean_names(daily_activity) %>%
  distinct() %>%
  drop_na()

daily_intensities <- clean_names(daily_intensities) %>%
  distinct() %>%
  drop_na()

daily_step <- clean_names(daily_step) %>%
  distinct() %>%
  drop_na()

daily_sleep <- clean_names(daily_sleep) %>%
  distinct() %>%
  drop_na()

hour_intensities <- clean_names(hour_intensities) %>%
  distinct() %>%
  drop_na()
```

```
hour_steps <- clean_names(hour_steps) %>%
  distinct() %>%
  drop_na()
```

Verify any duplicate remained

```
sum(duplicated(daily_activity))
```

```
## [1] 0
```

```
sum(duplicated(daily_intensities))
```

```
## [1] 0
```

```
sum(duplicated(daily_step))
```

```
## [1] 0
```

```
sum(duplicated(daily_sleep))
```

```
## [1] 0
```

```
sum(duplicated(hour_intensities))
```

```
## [1] 0
```

```
sum(duplicated(hour_steps))
```

```
## [1] 0
```

Convert it to date time format and split to date and time.

```
hour_intensities <- hour_intensities %>%
  mutate(activity_hour = mdy_hms(activity_hour),
         time = format(activity_hour, "%H:%M:%S"),
         date = as_date(activity_hour))
```

```
hour_steps <- hour_steps %>%
  mutate(activity_hour = mdy_hms(activity_hour),
         time = format(activity_hour, "%H:%M:%S"),
         date = format(activity_hour, "%m/%d/%y"))
```

```
daily_sleep <- daily_sleep %>%
  mutate(sleep_day = mdy_hms(sleep_day),
         time = format(sleep_day, "%H:%M:%S"),
         date = format(sleep_day, "%m/%d/%y"))
```

```
daily_step <- daily_step %>%
  rename(date = activity_day) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))
```

```
daily_activity <- daily_activity %>%
  rename(date = activity_date) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))
```

```
daily_intensities <- daily_intensities %>%
  rename(date = activity_day) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))
```

5.Analyze Phase and Share Phase

We will analyze activity patterns of the users of FitBit and determine if that can help us on getting some insights or opportunity of BellaBeat's marketing strategy.

5.1 Sleep Pattern Analysis

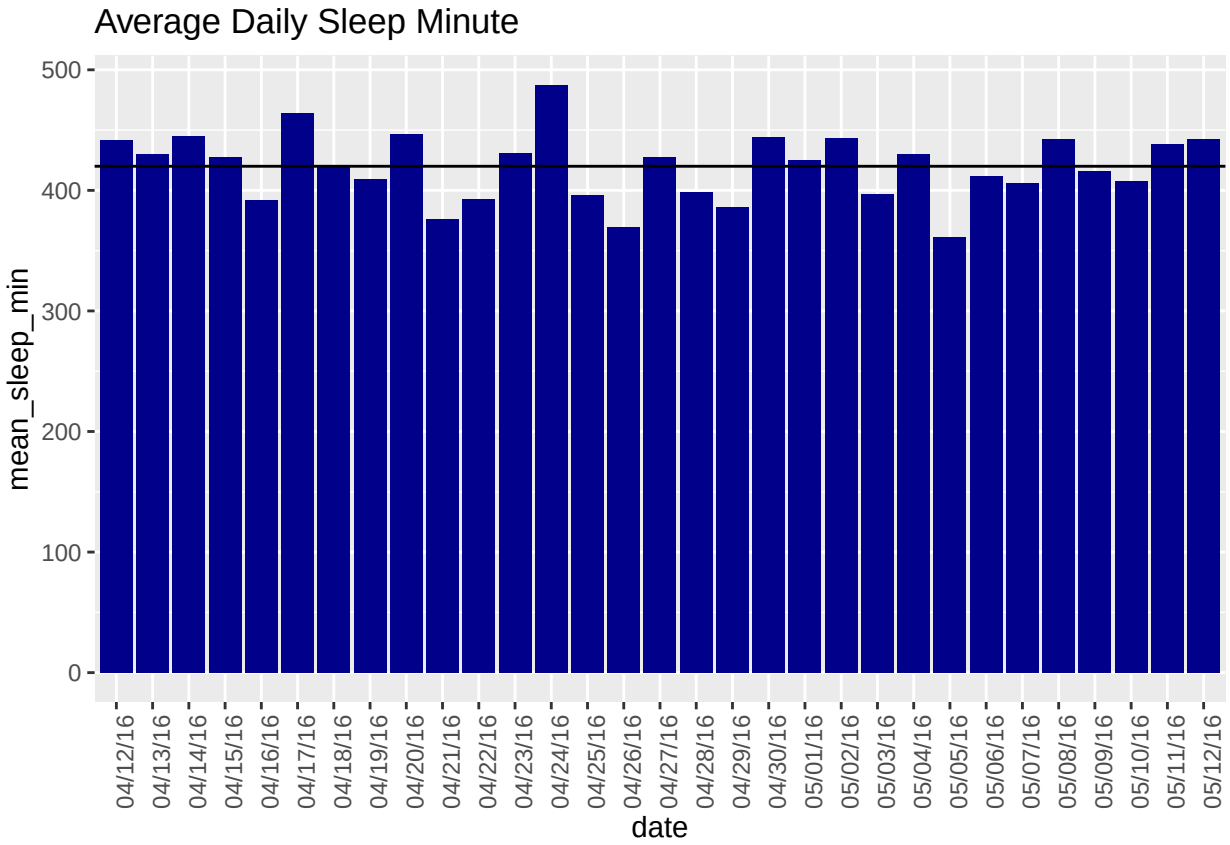
Based on the website of National Library of Medicine, joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society suggest: Adults should sleep 7 or more hours per night on a regular basis to promote optimal health. https://pmc.ncbi.nlm.nih.gov/articles/PMC4434546/

```
daily_sleep_copy <- daily_sleep %>%
     group_by(date) %>%
     summarise(mean_sleep_min=mean(total_minutes_asleep))
daily_sleep_copy
```

```
## # A tibble: 31 x 2
##    date     mean_sleep_min
##    <chr>             <dbl>
##  1 04/12/16           442.
##  2 04/13/16           430.
##  3 04/14/16           445.
##  4 04/15/16           427.
##  5 04/16/16           392.
##  6 04/17/16           464.
##  7 04/18/16           420.
##  8 04/19/16           409.
##  9 04/20/16           446.
## 10 04/21/16           376
## # i 21 more rows
```

```
ggplot(data=daily_sleep_copy,aes(x=date,y=mean_sleep_min))+
  geom_col(stat="identity", fill="darkblue")+
  geom_hline(yintercept = 420)+
  labs(title = "Average Daily Sleep Minute")+
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning in geom_col(stat = "identity", fill = "darkblue"): Ignoring unknown
## parameters: `stat`
```



From the graph, about half of the dates, the users average daily sleep hour is below recommended hour (7 hours per day)

5.2 Average Daily Steps

According to the PubMed Central (PMC) article, pedometer-determined phsical activity in healthy adults is classified as: * Sedentary: Less than 5000 steps a day. * low active: 5000-7499 steps/day * somewhat active: 7500-9999 steps/day * active: >=10000 steps/day * highly active: >12500 steps/day https://pubmed.ncbi.nlm.nih.gov/14715035/

First, we calculate the average daily steps for each user

```
daily_step_copy <- daily_step %>%
     group_by(id) %>%
     summarise(mean_daily_steps=mean(step_total))
daily_step_copy
```

```
## # A tibble: 33 x 2
##            id mean_daily_steps
##         <dbl>            <dbl>
##  1 1503960366           12117.
##  2 1624580081            5744.
##  3 1644430081            7283.
##  4 1844505072            2580.
##  5 1927972279             916.
##  6 2022484408           11371.
```

```
##  7 2026352035            5567.
##  8 2320127002            4717.
##  9 2347167796            9520.
## 10 2873212765            7556.
## # i 23 more rows
```

Then, users are categoried into different active level types

```
daily_step_copy <- daily_step_copy %>%
          mutate(user_type = case_when(
                mean_daily_steps < 5000 ~ "sedentary",
                mean_daily_steps >= 5000 & mean_daily_steps<7500 ~ "low active",
                mean_daily_steps >= 7500 & mean_daily_steps<10000 ~ "somewhat active",
                mean_daily_steps >= 10000 & mean_daily_steps<=12500 ~"active",
                mean_daily_steps > 12500 ~"highly active"
             ))
daily_step_copy
```

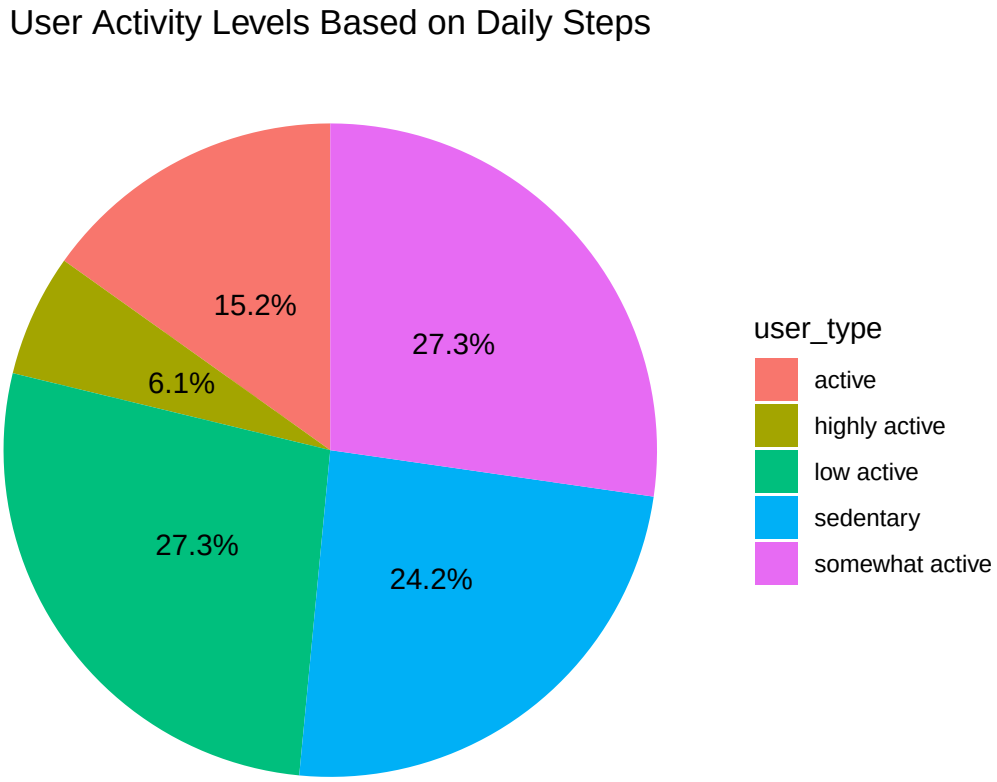
```
## # A tibble: 33 x 3
##            id mean_daily_steps user_type
##         <dbl>            <dbl> <chr>
##  1 1503960366           12117. active
##  2 1624580081            5744. low active
##  3 1644430081            7283. low active
##  4 1844505072            2580. sedentary
##  5 1927972279             916. sedentary
##  6 2022484408           11371. active
##  7 2026352035            5567. low active
##  8 2320127002            4717. sedentary
##  9 2347167796            9520. somewhat active
## 10 2873212765            7556. somewhat active
## # i 23 more rows
```

Compute the percentage of each type

```
step_summary <- daily_step_copy %>%
     group_by(user_type) %>%
     summarise(sub_total=n()) %>%
     mutate(total = sum(sub_total)) %>%
     group_by(user_type) %>%
     summarise(total_percent = sub_total / total) %>%
     mutate(percentage = scales::percent(total_percent))
step_summary
```

```
## # A tibble: 5 x 3
##   user_type       total_percent percentage
##   <chr>                   <dbl> <chr>
## 1 active                 0.152  15.2%
## 2 highly active          0.0606 6.1%
## 3 low active             0.273  27.3%
## 4 sedentary              0.242  24.2%
## 5 somewhat active        0.273  27.3%
```

```
ggplot(step_summary, aes(x = "", y = total_percent, fill = user_type)) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar("y", start = 0) +
    geom_text(aes(label = percentage),
              position = position_stack(vjust = 0.5)) +
    labs(title = "User Activity Levels Based on Daily Steps") +
    theme_void() +
    theme(plot.title = element_text(hjust = 0.5))
```



we can see the distribution of the different active levels. More than half of the users are below 7500 daily steps(low active and sedentary) which is a unhealthy life style needed to be improved We further check during the data period if the average daily steps computed from all users meet the recommended steps (7500)

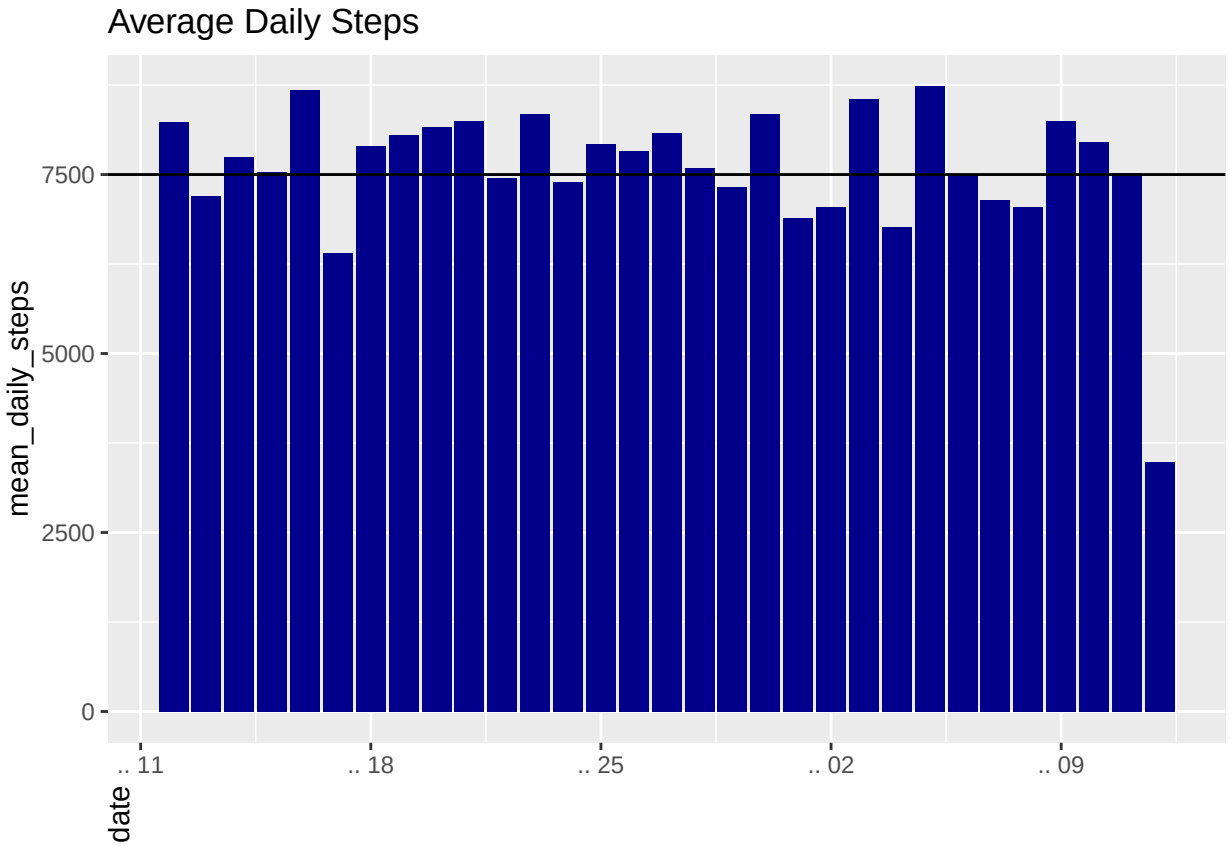
```
average_daily_step <- daily_step %>%
    group_by(date) %>%
    summarise(mean_daily_steps=mean(step_total))
average_daily_step
```

```
## # A tibble: 31 x 2
##    date       mean_daily_steps
##    <date>                <dbl>
##  1 2016-04-12            8237.
##  2 2016-04-13            7199.
##  3 2016-04-14            7744.
##  4 2016-04-15            7534.
```

```
##  5 2016-04-16           8679.
##  6 2016-04-17           6409.
##  7 2016-04-18           7897.
##  8 2016-04-19           8049.
##  9 2016-04-20           8163.
## 10 2016-04-21           8244.
## # i 21 more rows
```

```
ggplot(data = average_daily_step,aes(x=date,y=mean_daily_steps))+
  geom_col(stat="identity",fill="darkblue")+geom_hline(yintercept=7500)+
  labs(title="Average Daily Steps")+
  theme(axis.title.x = element_text(angle=90))
```

```
## Warning in geom_col(stat = "identity", fill = "darkblue"): Ignoring unknown
## parameters: `stat`
```



From the graph, about 1/3 of the dates, users did not meet the recommended daily steps

5.3 User Intensities over Weekdays

Change the system's locale settings, so that the outputs weekday names will be English instead of Chinese

```
Sys.setlocale("LC_TIME", "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8"
```

```
hour_intensities_copy <- hour_intensities %>%
  group_by(id, date) %>%
  drop_na() %>%
  summarise(daily_intensities = sum(total_intensity), .groups = "drop") %>%
  mutate(weekday = weekdays(date)) %>%
  group_by(weekday) %>%
  summarise(mean_intensities = mean(daily_intensities))

hour_intensities_copy
```

```
## # A tibble: 7 x 2
##   weekday   mean_intensities
##   <chr>                <dbl>
## 1 Friday                257.
## 2 Monday                249.
## 3 Saturday              264.
## 4 Sunday                251.
## 5 Thursday              269.
## 6 Tuesday               232.
## 7 Wednesday             270.
```

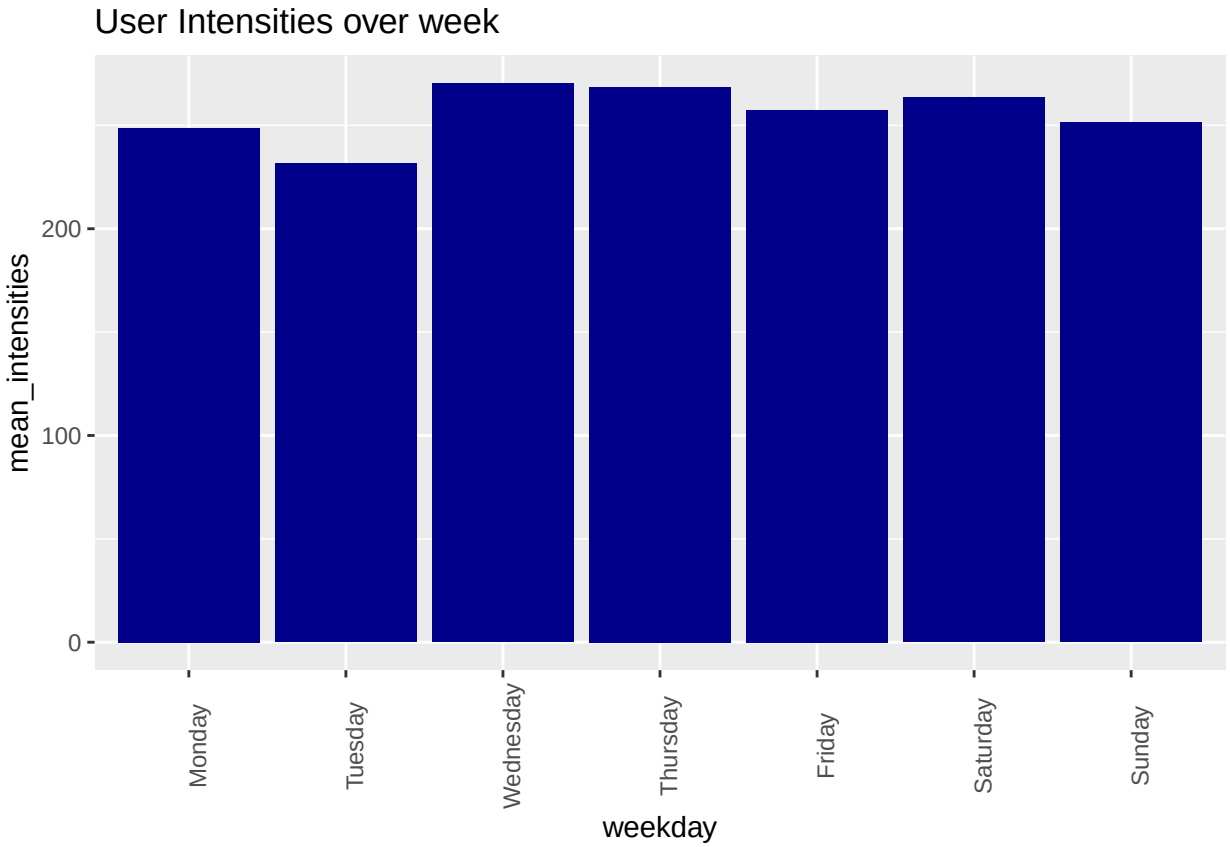
Reorder the Weekdays as normal order

```
hour_intensities_copy$weekday<-ordered(hour_intensities_copy$weekday, levels=c("Monday", "Tuesday", "Wed
```

Plot the graph to visualize the average user intensities over differen weekdays

```
ggplot(data=hour_intensities_copy, aes(x=weekday,y=mean_intensities))+
  geom_histogram(stat="identity", fill="darkblue")+
  theme(axis.text.x=element_text(angle=90))+
  labs(title="User Intensities over week")
```

```
## Warning in geom_histogram(stat = "identity", fill = "darkblue"): Ignoring
## unknown parameters: `binwidth`, `bins`, and `pad`
```



From the graph we can see on Saturaday and Sunday, users have the highest intensities and lowest intensities respectively.

5.4Frequency of usage

We further investigate how frequent the users wear the Bellabeat gadgets. We will determine the number of users who use their smart devices daily, categorizing our sample into three groups based on a 31-day observation period. * Frequent Users: Active on 21 to 31 days * Regular Users: Active on 10 to 20 days * Occasional Users: Active on 1 to 10 days

```
daily_usage <- daily_activity %>%
  group_by(id) %>%
  summarise(days = sum(n())) %>%
  mutate(usage = case_when(
    days <=10 ~"occasional user",
    days >10 & days <=20 ~"regular user",
    days >20 ~"frequent user"
  ))
daily_usage
```

```
## # A tibble: 35 x 3
##            id  days usage
##         <dbl> <int> <chr>
##  1 1503960366    19 regular user
##  2 1624580081    19 regular user
##  3 1644430081    10 occasional user
```

```
##  4 1844505072    12 regular user
##  5 1927972279    12 regular user
##  6 2022484408    12 regular user
##  7 2026352035    12 regular user
##  8 2320127002    12 regular user
##  9 2347167796    15 regular user
## 10 2873212765    12 regular user
## # i 25 more rows
```

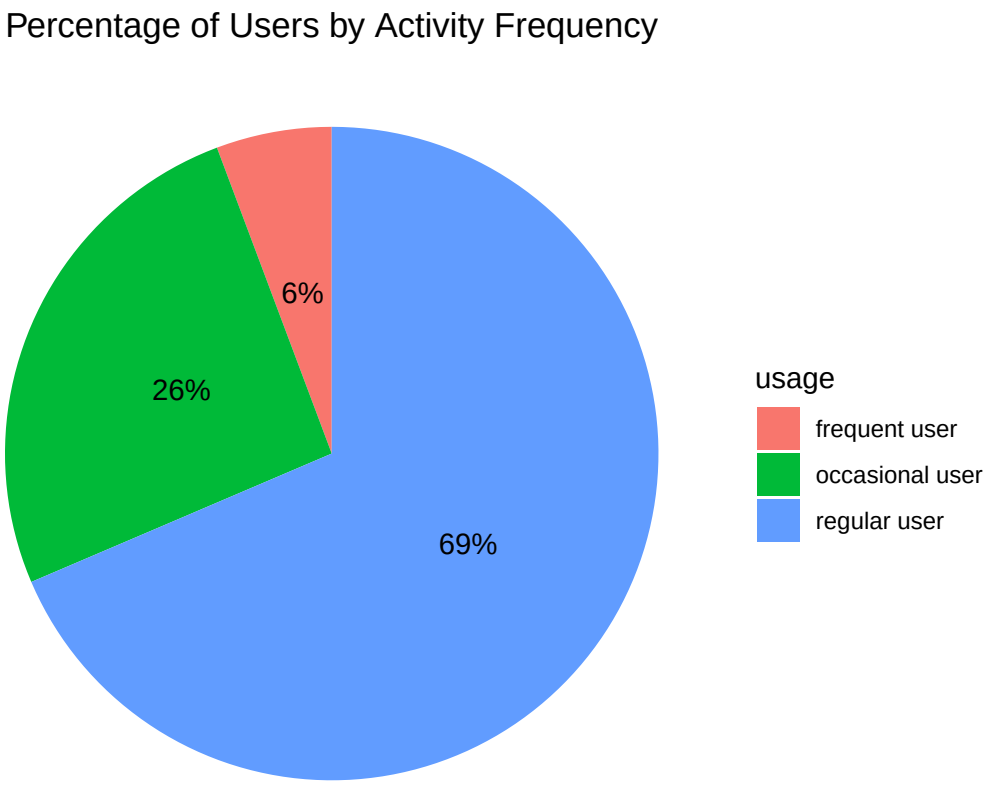
Compute their percentage

```
daily_usage_percentage <- daily_usage %>%
  group_by(usage) %>%
  summarise(sub_total=n()) %>%
  mutate(total = sum(sub_total)) %>%
  group_by(usage) %>%
  summarise(total_percent=sub_total/total) %>%
  mutate(percentage=scales::percent(total_percent))
daily_usage_percentage
```

```
## # A tibble: 3 x 3
##   usage           total_percent percentage
##   <chr>                   <dbl> <chr>
## 1 frequent user          0.0571 6%
## 2 occasional user        0.257  26%
## 3 regular user           0.686  69%
```

Then plot the pie chart to visualize the distribution of user types

```
ggplot(daily_usage_percentage, aes(x = "", y = total_percent, fill = usage)) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar("y", start = 0) +
    geom_text(aes(label = percentage),
              position = position_stack(vjust = 0.5)) +
    labs(title = "Percentage of Users by Activity Frequency") +
    theme_void() +
    theme(plot.title = element_text(hjust = 0.5))
```



We can see that * 6% users frequently wear and use the Bellatbeat devices(21-31days) * 69% users regularly use the devices(10 to 20days) * 26% users only use the devices less than 10 days

6. Act Phase

Summary of Key Insights

Sleep Patterns

- On approximately half of the recorded days, users averaged less than the recommended 7 hours of sleep.

Physical Activity

- On about one-third of the days, users failed to meet the recommended daily step count.
- Over 50% of users consistently logged fewer than 7,500 steps per day, placing them in the low-active or sedentary category—an indicator of unhealthy lifestyle habits.

Activity Intensity by Day

- Users showed the highest activity intensities on Saturdays, while Sundays had the lowest.

Device Engagement

- 88% of users are highly engaged, using their Bellabeat devices for 21–31 days.
- 9% are moderately engaged (10–20 days).
- 3% show low engagement, using the device fewer than 10 days.

Recommendations for Bellabeat

Promote Sleep Health

- Introduce personalized sleep coaching features or reminders based on sleep tracking data.
- Offer educational content on the importance of sleep and how to improve sleep hygiene.

Encourage Physical Activity

- Implement motivational nudges or gamified challenges to help users reach at least 7,500 steps daily.
- Provide tailored activity goals based on user history and gradually increase targets.

Weekend Optimization

- Leverage high Saturday engagement by launching weekend wellness campaigns or guided workouts.
- Address low Sunday activity with gentle reminders or relaxing movement suggestions (e.g., yoga, walking).

Boost Engagement for Low-Use Users

- Send personalized re-engagement messages or offer incentives for consistent use.
- Simplify on-boarding and highlight key benefits to encourage regular usage.

Segmented Marketing Strategy

- Use engagement tiers (frequent, regular, occasional) to tailor messaging and product recommendations.
- For highly engaged users, promote premium features or community challenges.
- For less engaged users, focus on ease-of-use and habit formation.