BellaBeat Capstone Project

Ayush Ram

2025-11-17

Introduction

This R Markdown document presents the Bellabeat data analysis case study through the Google Data Analysis Course. In this document, we will explore the business task that is proposed by Bellabeat, a high-tech manufacturer of health products for women. Bellabeats aims to increase their global presence in the smart device market by understanding user behaviour that can guide product strategies and marketing decisions.

Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that by analysing existing competitor data (Fitbit) and comparing it to Bellabeat's Time Product (a wellness watch that tracks users' activity, sleep, and stress), we can apply existing trends to the device, benefiting both Bellabeat and helping support their customers better.

Business Task

Sršen has requested an analysis of smart device usage to gain insight into how customers use non-Bellabeat smart devices.

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat's marketing strategy?

Source

The public dataset was provided by Sršen, FitBit Fitness Tracker Data. This dataset is the raw, unedited Kaggle dataset containing usage data from thirty Fitbit users.

This document will focus mainly on:

- `dailyActivity_merged`
- `sleepDay_merged`
- `dailySteps_merged`
- `weightLogInfo_merged`

Data Limitations

The data itself is presented with a few limitations that are noteworthy. There are only 30 Fitbit users, which is a relatively small sample. The data itself does not represent Bellabeats users; this data is only an implication of what can be applied to them. Some data is self-reported, such as Weight, which can lead to inconsistencies or incorrect inputs. Missing some potential information, such as gender and age, which could have provided more insight. No context of what people do in bed before they are asleep.

Cleaning

For the Cleaning of the Data, Google Sheets was used before exporting the data into R Studio, as the data was not too large and the process would be quick and fast, leading to a faster response time to the Stakeholders' Business Task.

The following is a breakdown of key cleaning that was conducted on Google Sheets.

- Standardised date format from `MM/DD/YYYY` to `YYYY/MM/DD`, allowing data to be consistent and universally readable.
- Applied filter to `sleepDay_merged` Column D, Sort A-Z, applied syntax in column G `=IF(COUNTIFS(D:D, D2, E:E, E2) > 1, "Duplicate", "Unique")`. Showing duplicates, Applied Data cleanup to remove the duplicates. 3 data entries removed from `sleepDay_merged`. Other data entries returned results of no duplicates.
- Applied conditional formatting to see if any empty cells, removed Column `Fat` from `weightLogInfo_merged`, as there were only 2 completed entries.
- Removed non-representative days in `dailySteps_merged` where `StepTotal` < 50 steps, as these days were days of inactivity of the Fitbit by users.
- Applied `=IF(SUM(C2:M2)=0, "Non-Wear", "Valid")` to `DailyActivity_merged` to determine exactly which days had no activity on the Fitbit, applied a filter to show `Non-Wear` days, and removed all those days.
- Verified column names and formats to ensure a seamless import into

  R.
- Clean Data was then uploaded to R Studio with the same titles, and the inclusion of "Cleaned" was added to each file.

Download Packages

install.packages("readr")

install.packages("dplyr")

install.packages("ggplot2")

install.packages("tidyr")

Loading Packages

Loading the necessary packages

```r
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.5.2
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.5.2
```

Loading Datasets

Renaming clean data to friendly, readable names.

```r
Activity <- read.csv("Fitabase_Data_cleaned/dailyActivity_merged_cleaned.csv")
Sleep <- read.csv("Fitabase_Data_cleaned/sleepDay_merged_cleaned.csv")
Steps <- read.csv("Fitabase_Data_cleaned/dailySteps_merged_cleaned.csv")
Weight <- read.csv("Fitabase_Data_cleaned/weightLogInfo_merged_cleaned.csv")
```

CSV files can be found in attachment files. All data can be found in "Fitabase_Data_cleaned" in folder.

Review data with head() Showing the first few rows of the datasets.

```r
head(Activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   2016-04-12      13162          8.50            8.50
## 2 1503960366   2016-04-13      10735          6.97            6.97
## 3 1503960366   2016-04-14      10460          6.74            6.74
## 4 1503960366   2016-04-15       9762          6.28            6.28
```

```
## 5 1503960366   2016-04-16      12669          8.16            8.16
## 6 1503960366   2016-04-17       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

```
head(Sleep)
```

```
##           Id   SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 2016-04-12                 1                327            346
## 2 1503960366 2016-04-13                 2                384            407
## 3 1503960366 2016-04-15                 1                412            442
## 4 1503960366 2016-04-16                 2                340            367
## 5 1503960366 2016-04-17                 1                700            712
## 6 1503960366 2016-04-19                 1                304            320
```

```
head(Steps)
```

```
##           Id ActivityDay StepTotal
## 1 4020332650  2016-04-18        62
## 2 4020332650  2016-04-14       108
## 3 8792009665  2016-04-21       144
## 4 1927972279  2016-04-22       149
## 5 1927972279  2016-04-25       152
## 6 1844505072  2016-04-19       197
```

```
head(Weight)
```

```
##           Id       Date WeightKg WeightPounds   BMI IsManualReport        LogId
## 1 1503960366 2016-05-02     52.6     115.9631 22.65           TRUE 1.462234e+12
## 2 1503960366 2016-05-03     52.6     115.9631 22.65           TRUE 1.462320e+12
## 3 1927972279 2016-04-13    133.5     294.3171 47.54          FALSE 1.460510e+12
## 4 2873212765 2016-04-21     56.7     125.0021 21.45           TRUE 1.461283e+12
## 5 2873212765 2016-05-12     57.3     126.3249 21.69           TRUE 1.463098e+12
## 6 4319703577 2016-04-17     72.4     159.6147 27.45           TRUE 1.460938e+12
```

```
sum(duplicated(Sleep))
```

Checking for any duplicates in datasets

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

```
sum(duplicated(Activity))
```

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

```
sum(duplicated(Steps))
```

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

```
sum(duplicated(Weight))
```

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

no duplicates found, data was cleaned prior.

```
colSums(is.na(Activity))
```

Check for missing values in each dataset

```
##                       Id             ActivityDate                TotalSteps
##                        0                        0                         0
##            TotalDistance          TrackerDistance LoggedActivitiesDistance
##                        0                        0                         0
##       VeryActiveDistance ModeratelyActiveDistance      LightActiveDistance
##                        0                        0                         0
##  SedentaryActiveDistance        VeryActiveMinutes      FairlyActiveMinutes
##                        0                        0                         0
##     LightlyActiveMinutes         SedentaryMinutes                 Calories
##                        0                        0                         0
```

```
colSums(is.na(Sleep))
```

```
##                 Id           SleepDay  TotalSleepRecords TotalMinutesAsleep
##                  0                  0                  0                  0
##     TotalTimeInBed
##                  0
```

```
colSums(is.na(Steps))
```

```
##          Id ActivityDay   StepTotal
##           0           0           0
```

```
colSums(is.na(Weight))
```

```
##             Id           Date       WeightKg   WeightPounds            BMI
##              0              0              0              0              0
## IsManualReport          LogId
##              0              0
```

Activity <- Activity %>% select(-X.1)

removed extra column "X.1 which was an empty placeholder from CSV export.

```
n_distinct(Activity$Id)
```

Inspecting how many unique ID are in the dataset

```
## [1] 33
```

```
n_distinct(Steps$Id)
```

```
## [1] 33
```

```
n_distinct(Sleep$Id)
```

```
## [1] 24
```

```
n_distinct(Weight$Id)
```

```
## [1] 8
```

As shown, both Steps and Activity have the same number of distinct ID, due to them using the same data pool. Verify this, as both have ID 1624580081 Total steps: 36019 on 2016-05-01. Weight dataset is smaller sets as fewer people recorded this data themselves.

Analysis of Data

Summary of all data to note down key information

```
Activity %>%
  select(TotalSteps, TotalDistance, ModeratelyActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, L
  summary()
```

Activity

```
##    TotalSteps    TotalDistance    ModeratelyActiveDistance VeryActiveMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000           Min.   :  0.00
##  1st Qu.: 4910   1st Qu.: 3.367   1st Qu.:0.0000           1st Qu.:  0.00
##  Median : 8027   Median : 5.585   Median :0.3050           Median :  7.00
##  Mean   : 8310   Mean   : 5.973   Mean   :0.6175           Mean   : 23.03
##  3rd Qu.:11089   3rd Qu.: 7.895   3rd Qu.:0.8625           3rd Qu.: 35.00
##  Max.   :36019   Max.   :28.030   Max.   :6.4800           Max.   :210.00
##  FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes    Calories
##  Min.   :  0.00      Min.   :  0.0        Min.   :   0.0   Min.   :  52
##  1st Qu.:  0.00      1st Qu.:146.0        1st Qu.: 721.8   1st Qu.:1856
##  Median :  8.00      Median :208.0        Median :1021.0   Median :2220
##  Mean   : 14.76      Mean   :209.8        Mean   : 956.3   Mean   :2362
##  3rd Qu.: 21.00      3rd Qu.:272.0        3rd Qu.:1189.2   3rd Qu.:2832
##  Max.   :143.00      Max.   :518.0        Max.   :1440.0   Max.   :4900
```

This dataset will focus on the activity levels of the Fitbit users. The data shows that users spend significantly more time in a sedentary state in comparison to the other activity levels. SedentaryMinutes has the largest Mean and Max values, indicating that most users are in states of inactivity for large periods. There is a large variance between steps, indicating that there is a lot of diversity in people's exercise habits.

```
Sleep %>%
  select(TotalMinutesAsleep, TotalTimeInBed) %>%
  summary()
```

Sleep

```
##  TotalMinutesAsleep TotalTimeInBed
##  Min.   : 58.0      Min.   : 61.0
##  1st Qu.:361.0      1st Qu.:403.8
##  Median :432.5      Median :463.0
##  Mean   :419.2      Mean   :458.5
##  3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :796.0      Max.   :961.0
```

This data shows that most people do spend time in bed before they go to sleep, the Mean value difference being 39.3 minutes. Meaning people are awake in bed for a period of time before falling asleep. There is not enough information provided to know what they are doing during this time.

```
Weight %>%
  select(WeightKg, BMI) %>%
  summary()
```

Weight

```
##     WeightKg           BMI
##  Min.   : 52.60   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:23.96
```

```
##  Median : 62.50   Median :24.39
##  Mean   : 72.04   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:25.56
##  Max.   :133.50   Max.   :47.54
```

Information in the dataset is fairly self-explanatory. This shows the Weight (kg) and BMI of users who logged this information. The "Fat" column was removed, as it had too many missing values to be meaningful.

Dataset Steps was not included, as it was already in Activity Summary.

Visualisation of Activity, Sleep and Weight

```
ggplot(data= Activity) +
  geom_point(mapping = aes(x= TotalSteps, y = Calories), alpha= 0.6)+
  geom_smooth(aes(x = TotalSteps, y = Calories))+
  labs(
    title = "Total Steps Vs Calories Burned",
    subtitle = "With Trend Line",
    x = "Total Steps",
    y = "Calories Burned"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

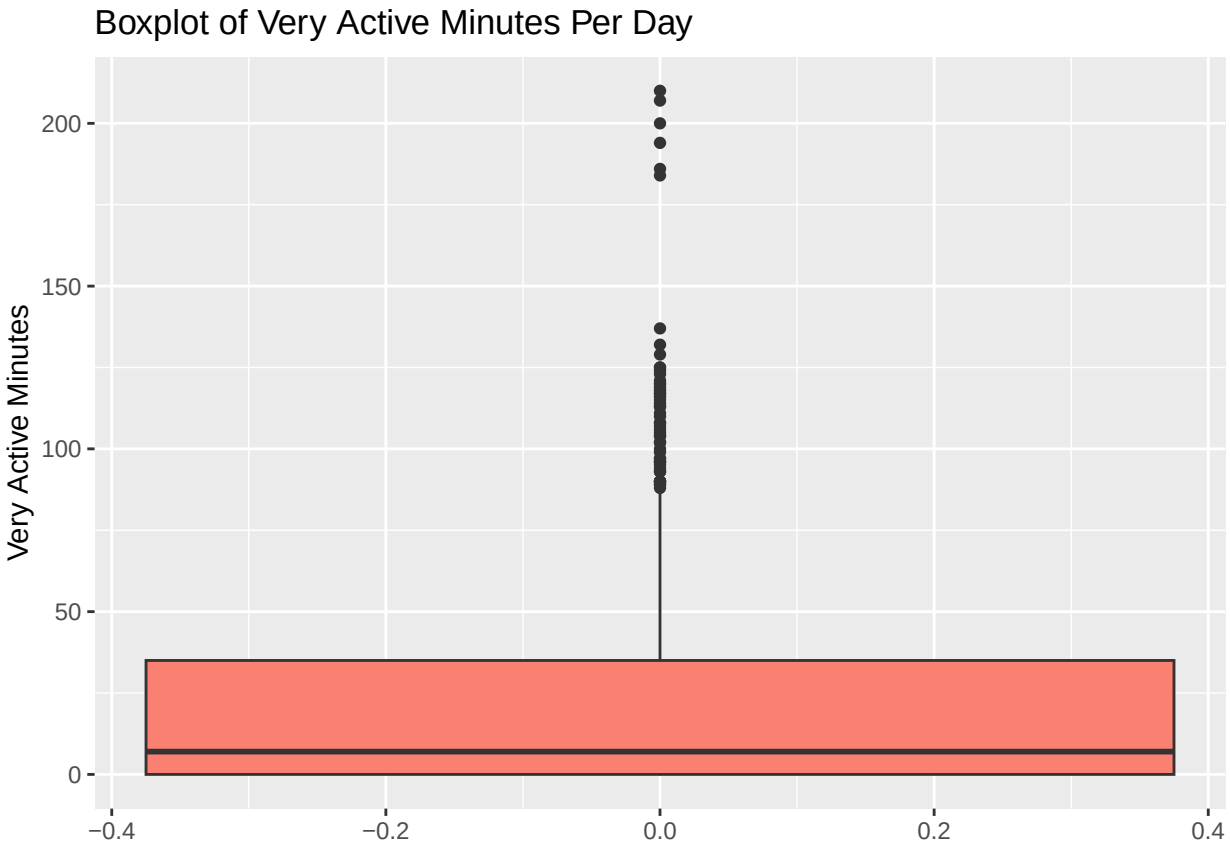
This visualisation shows the relation between the TotalSteps and Calories burned in the dataset activity. There is a clear positive correlation, so it can be observed, those who have a higher total step count will see their expenditure of calories increase.

```
cor(Activity$TotalSteps, Activity$Calories, use = "complete.obs")
```

```
## [1] 0.5600494
```

The positive correlation of r= 0.560 confirms this.

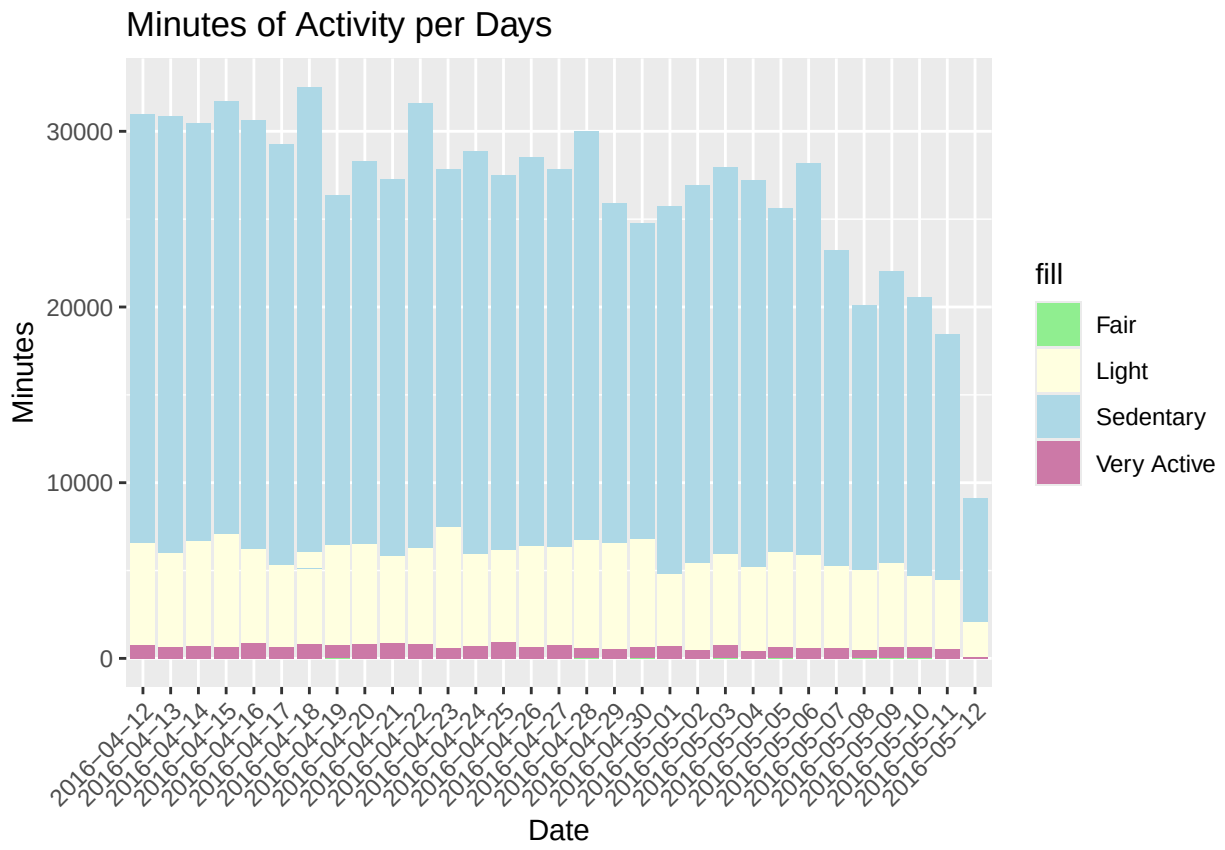
```
ggplot(Activity, aes(y = VeryActiveMinutes)) +
  geom_boxplot(fill= "salmon") +
  labs(title = "Boxplot of Very Active Minutes Per Day",
       y= "Very Active Minutes")
```



This visualisation shows that most users spend relatively few minutes per day in highly active minutes; only a few users are highly active past 100 minutes, and activity past ~180 minutes is not common.

```
ggplot(Activity, aes (x = ActivityDate)) +
  geom_bar(aes(y= SedentaryMinutes, fill = "Sedentary"), stat = "identity") +
  geom_bar(aes(y= LightlyActiveMinutes, fill = "Light"), stat = "identity") +
  geom_bar(aes(y= FairlyActiveMinutes, fill = "Fair"), stat = "identity") +
  geom_bar(aes(y= VeryActiveMinutes, fill = "Very Active"), stat = "identity")+
  labs(
    title= "Minutes of Activity per Days",
```

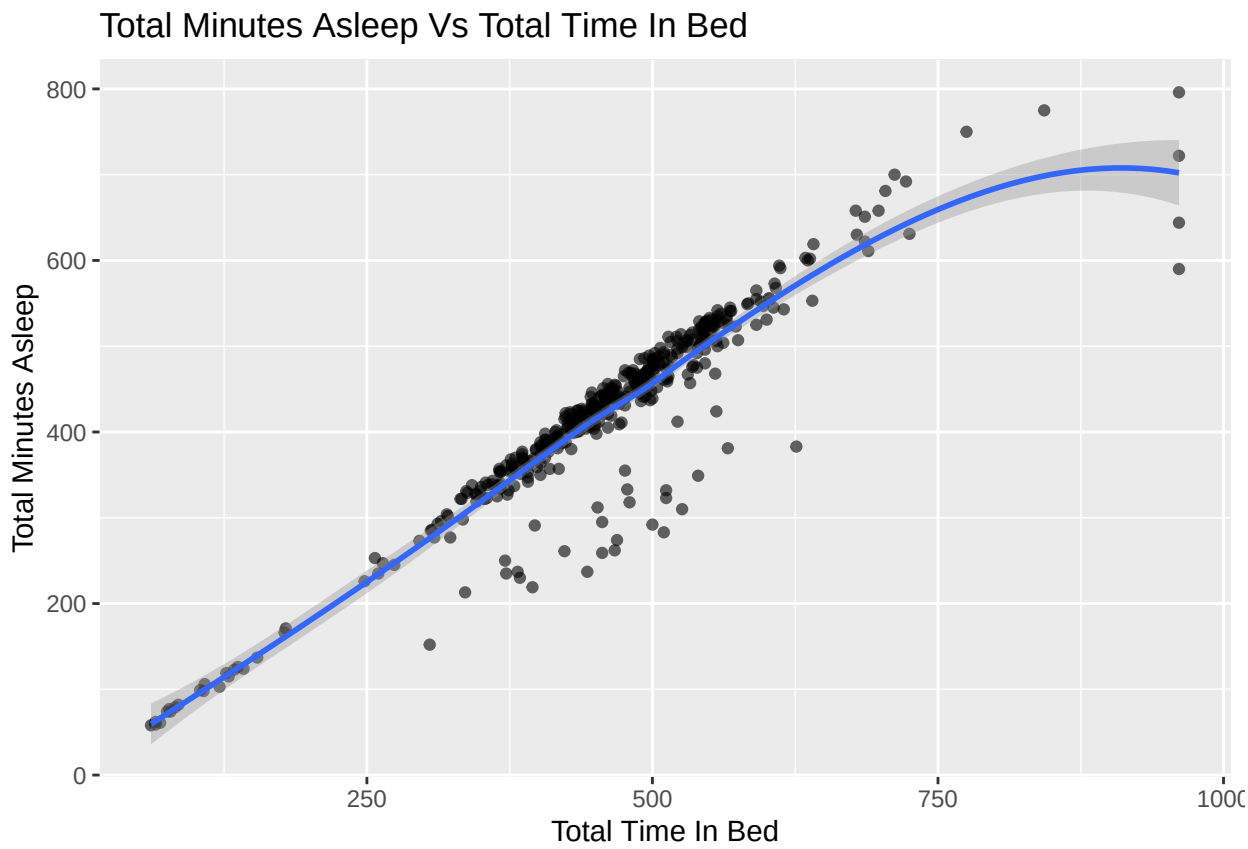
```
    x= "Date",
    y= "Minutes"
  ) +
  scale_fill_manual(values = c("Sedentary" = "lightblue", "Light" = "lightyellow", "Fair" = "lightgreen
  theme(axis.text.x= element_text(angle = 45, hjust= 1))
```



This shows the Minutes of Activity per day for the users; from this information, we can see that most of the users spend their time in a sedentary state, and while active, most users are in a light state of activity, followed by very active and finally fair activity. The reason that "Fair" is almost non-existent in the graph may be due to the way the types of data are considered. If we were to apply this to our Bellabeats product, we could note that most users like to be in a light state of activity, so reminders to simply get up and go for a walk instead of being in a sedentary state, or even a message that would show up when you are in a light state of activity or higher, stating that you are doing better than "X" amount of other users as motivation.

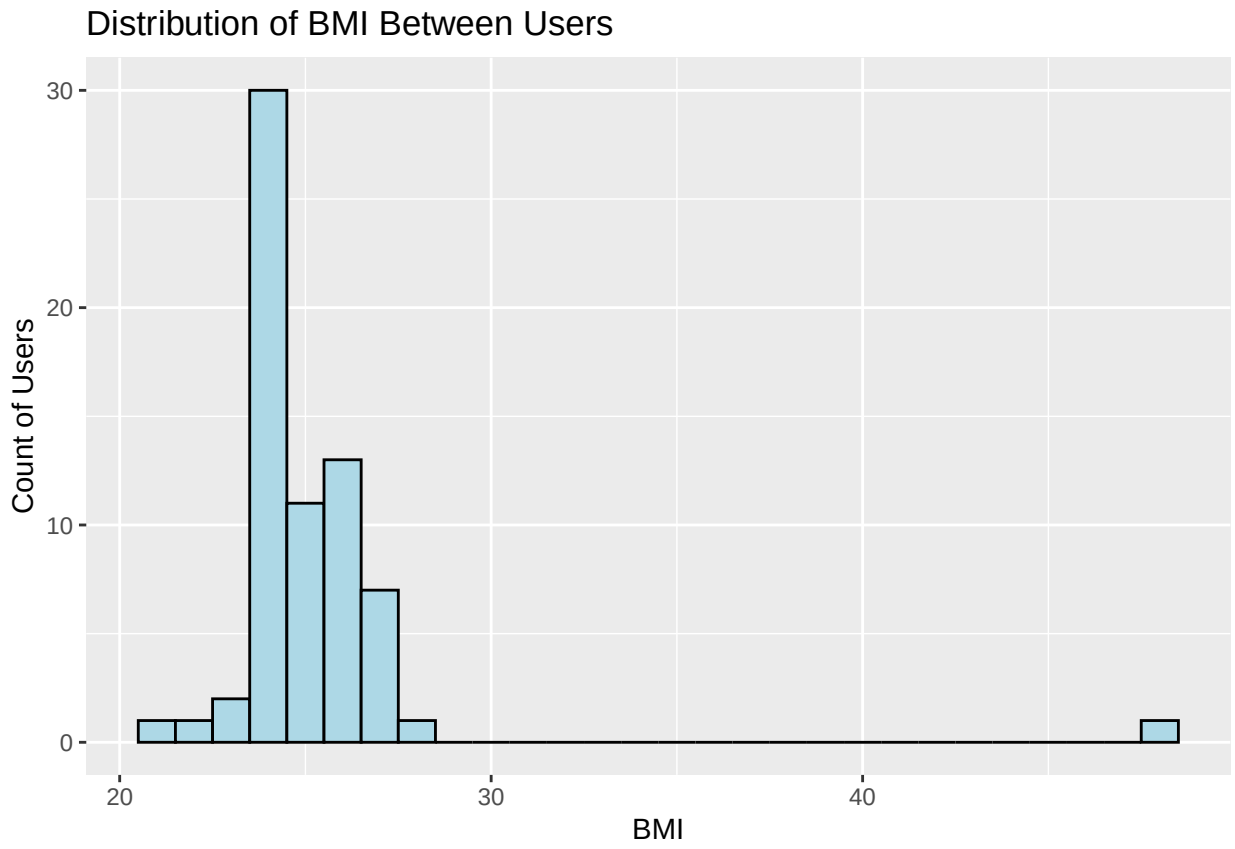
```
ggplot(data= Sleep) +
  geom_point(mapping = aes(x= TotalTimeInBed, y= TotalMinutesAsleep), alpha= 0.6)+
  geom_smooth(aes(x= TotalTimeInBed, y= TotalMinutesAsleep))+
  labs(
    title=  "Total Minutes Asleep Vs Total Time In Bed",
    x= "Total Time In Bed",
    y= "Total Minutes Asleep"
  )
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



There is a positive correlation here between time spent in bed and time spent asleep. As expected, Those who spend more time in bed tend to sleep for longer. However, Most users don't fall asleep immediately. Spending only roughly 80%-90% of the time asleep. Suggesting there is a gap between going to bed and actually being asleep. This can be due to winding down or having difficulty falling asleep.

```
ggplot(Weight, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "lightblue", colour = "Black") +
  labs(
    title= "Distribution of BMI Between Users",
    x= "BMI",
    y= "Count of Users"
  )
```




```
summary(Weight$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.45   23.96   24.39   25.19   25.56   47.54
```

This graph shows the distribution of the BMI among the users. The spread here indicates that most users fall between 20 to 30, BMI verified by the summary below the table, which shows a Mean of 25.19 BMI. With anything in between 18.5 to 24.9 BMI being considered normal weight, whereas 25.0-29.9 BMI being overweight. There is a small population that falls in the overweight division, with the Max being 47.54 BMI.

Comparing Datasets

```
Sleep_Weight <- merge(Sleep, Weight, by.x = c("Id", "SleepDay"), by.y = c("Id", "Date"))

cor(Sleep_Weight$TotalMinutesAsleep, Sleep_Weight$WeightKg, use = "complete.obs")
```

Test correlation of Sleep Vs Weight

```
## [1] 0.03348514
```

With a correlation of 0.033, there is little to no relationship between these two variables. There will be no further investigation of this.

```r
Steps_Sleep <- merge(Steps, Sleep, by.x =c("Id", "ActivityDay"), by.y = c("Id", "SleepDay"))

cor(Steps_Sleep$StepTotal, Steps_Sleep$TotalMinutesAsleep, use = "complete.obs")
```

Test correlation of Steps Vs Sleep

```
## [1] -0.197921
```

Cross-dataset relationship shows weak to no correlation, suggesting that user activity, sleep, and body weight are likely independent in the dataset.

Recommendations

From the information we can infer from this document, there are a lot of trends in smart device usage for BellaBeats to stand out from Fitbit and the rest of the competition.

- Encourage movement with motivating goals

Users who are moving more tend to burn more calories. BellaBeats can use this information and create in-app step goals, badges and rewards for activity across multiple days, such as an activity streak and milestones. Marketing campaigns can focus on the more movement, the more calories burned.

- Support low-activity users with a gentle reminder

Many users stay in lower activity ranges (sedentary/low activity). On average, ordinary people spend 66.5% (956.30 minutes) a day in a sedentary state, followed by 14.4% (208 minutes) in a fairly light activity state. leaving only 19.1% (275 minutes) for fair activity or higher.

Bellabeats product, the Time could introduce gentle reminders for less active users to encourage movement, and then give those who are doing better than others positive reinforcement. A visual could show the users' activity as a percentage compared to the population to drive competition and growth.

For example, a message could appear on the app after users pass 4 hours and 30 minutes, "Fun fact, most people are only active for about 4 hours and 30 minutes, keep pushing to beat the rest!"

- Improve sleep awareness and the collection of that data

Users often spend 250-600 minutes in bed, but some do not spend that time asleep; some users are only asleep for a total time between 200- 400 minutes. So for those users, we could allow them to turn on a gentle notification system when users are detected as being in bed and still awake. Also, due to this dataset having 8 distinct users, whereas the other datasets has 24-33 users, there is less user information provided, we could offer in-app benefits for wearing the BellaBeats Time to sleep to gather more data about this, and be ahead of the competition in user understanding.

For example, users who wear the Time in bed receive in-app currency for an avatar that they can show off to their friends and family, providing a low-cost initiative for users.

- Provide personalised data based on BMI

Users BMi values are very diverse. most users fall in a BMI range of 20 to 30 With a single outlier of a BMI of 47.54, Rather than exclude a user we can offer personalised recommendations, make custom recommended steps, activity plans. etc

Because the reality is most people will have large differences in their bodies and fitness levels, and we should be inclusive of a sensitive topic and not try to exclude people as outliers. This can help position BellaBeats as a personalised platform that shows genuine care for its users' differences.

- Avoid connections of meaningless datasets

Marketing should treat sleep, weight and activity as different metrics, as the correlation between sleep and weight is low (r=0.033) and steps vs sleep is extremely low (r=-0.198). With these findings the company can focus records into more actionable insights

Summary

As shown in the analysis, there are some clear trends in smart device usage. Users are more active when they move around more throughout the day. Most people remain in a sedentary state for long periods, and most people stay in bed longer than they are actually sleeping. That, along with BMI and weight, varies and needs personalisation for users.

Together, there are many opportunities for BellaBeats to stand out from the rest. To look at this data and go for a personalised, gentle, one-on-one approach and focus on deep and meaningful motivation can help users feel valued and more engaged with the BellaBeat product.