Case Study: Bellabeat

Artyom Pak

7/4/2021

Introduction

Bella beat - a high-tech company that provides health-focused smart products. The products are designed to collect data on activity, sleep, stress, and reproductive health. Bellabeat empowers women with knowledge about their own health and habits.

Products:

App - provides users with health data related to their activity, so they better understand current habits and make healthy decisions.
Leaf - a tracker can be worn as a bracelet, necklace, or clip. Connects to the app to track activity (sleep and stress)
Time - a smartwatch to track user activity. The product tracks your daily wellness.
Spring - a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The product connects to the app.
Membership - a subscription program. Provides 24/7 fully personalized guidance on nutrition, activity, sleep, health, beauty based on lifestyle and goals.

Stage 1: Ask questions and determine business tasks

The stakeholders:

- Urška Sršen (Bellabeat's cofounder and Chief Creative Officer)
- Sando Mur (Mathematician and Bellabeat's cofounder)

The business tasks:

- Analyze smart device usage data to find patterns and gain insight
- Discover more opportunities for growth

Stage 2: Prepare the data

Dataset: click here
License: CC0: Public Domain
Storage: Oracle Cloud Service (OCID: `ocid1.user.oc1..aaaaaaaa6tr4lsnis2npabube3v7dt76wvk7nkook3b3w4xsggvq2fs`

Description: This data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore habits and patterns.

Stage 3: Process the data

The observation of the data sets will be done using **Oracle RDMBS**, and **Google Spreadsheets**. The cleaning and visualization processes will be conducted via **R programming language** (`tidyverse` and `scales` packages), since it provides more flexibility when it comes to working with big volumes of data.

```
activity <- read_csv('~/Desktop/cs/dailyActivity_merged.csv')
sleep <- read_csv('~/Desktop/cs/sleepDay_merged.csv')
intensities <- read_csv('~/Desktop/cs/hourlyIntensities_merged.csv')
```

3.1 Importing datasets:

```
# Overall activity information
activity <- rename(activity, date=ActivityDate)
activity$date <- as.Date(activity$date, format="%m/%d/%Y")

# Information about sleep
sleep <- rename(sleep, date=SleepDay)
sleep$date = as.POSIXct(sleep$date, format="%m/%d/%Y %I:%M:%S")
sleep$date <- as.Date(sleep$date, format="%m/%d/%Y")

# Intensities. Extracting timestamp
intensities$ActivityHour = as.POSIXct(intensities$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.ti
intensities$time <- format(intensities$ActivityHour, format = "%H:%M:%S")
```

3.2 Formating incorrect dates (changing chr to date):

```
# Finding duplicates
any(duplicated(activity))
any(duplicated(intensities))
any(duplicated(sleep))

# Finding empty cells
any(is.na(activity))
any(is.na(intensities))
any(is.na(sleep))

# Counting rows
```

```
n_distinct(activity$Id)
n_distinct(sleep$Id)
n_distinct(intensities$Id)
```

3.3 General cleaning: ! *Sleep data set contains null values. The problem will be handled during the analysis*

```
# Merging data sets for complex analysis
merged <- merge(sleep, activity, by=c('Id', 'date'))
```

3.4 Merging data sets for complex analysis:

3.5 Summary

- Data type errors have been fixed
- Duplicated rows have not been found
- Empty values will be handled during the analysis
- Data sets are organized and prepared to analyze

Stage 4: Analyze and visualization

The aggregated data below (4.1 - 4.2) shows that there are records with total steps and calories equal to 0. Highly likely the devices were not being used properly or there was a technical issue.

```
activity %>%
  select(TotalSteps, VeryActiveMinutes, SedentaryMinutes, Calories) %>%
  summary()
```

4.1 Activity:

```
##    TotalSteps    VeryActiveMinutes SedentaryMinutes    Calories
##  Min.   :    0   Min.   :  0.00    Min.   :   0.0   Min.   :   0
##  1st Qu.: 3790   1st Qu.:  0.00    1st Qu.: 729.8   1st Qu.:1828
##  Median : 7406   Median :  4.00    Median :1057.5   Median :2134
##  Mean   : 7638   Mean   : 21.16    Mean   : 991.2   Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 32.00    3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :36019   Max.   :210.00    Max.   :1440.0   Max.   :4900
```

```
activity %>%
  select(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance) %>%
  summary()
```

4.2 Distance patterns:

```
##  VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
##  Min.   : 0.000     Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 0.000     1st Qu.:0.0000           1st Qu.: 1.945
##  Median : 0.210     Median :0.2400           Median : 3.365
##  Mean   : 1.503     Mean   :0.5675           Mean   : 3.341
##  3rd Qu.: 2.053     3rd Qu.:0.8000           3rd Qu.: 4.782
##  Max.   :21.920     Max.   :6.4800           Max.   :10.710
```

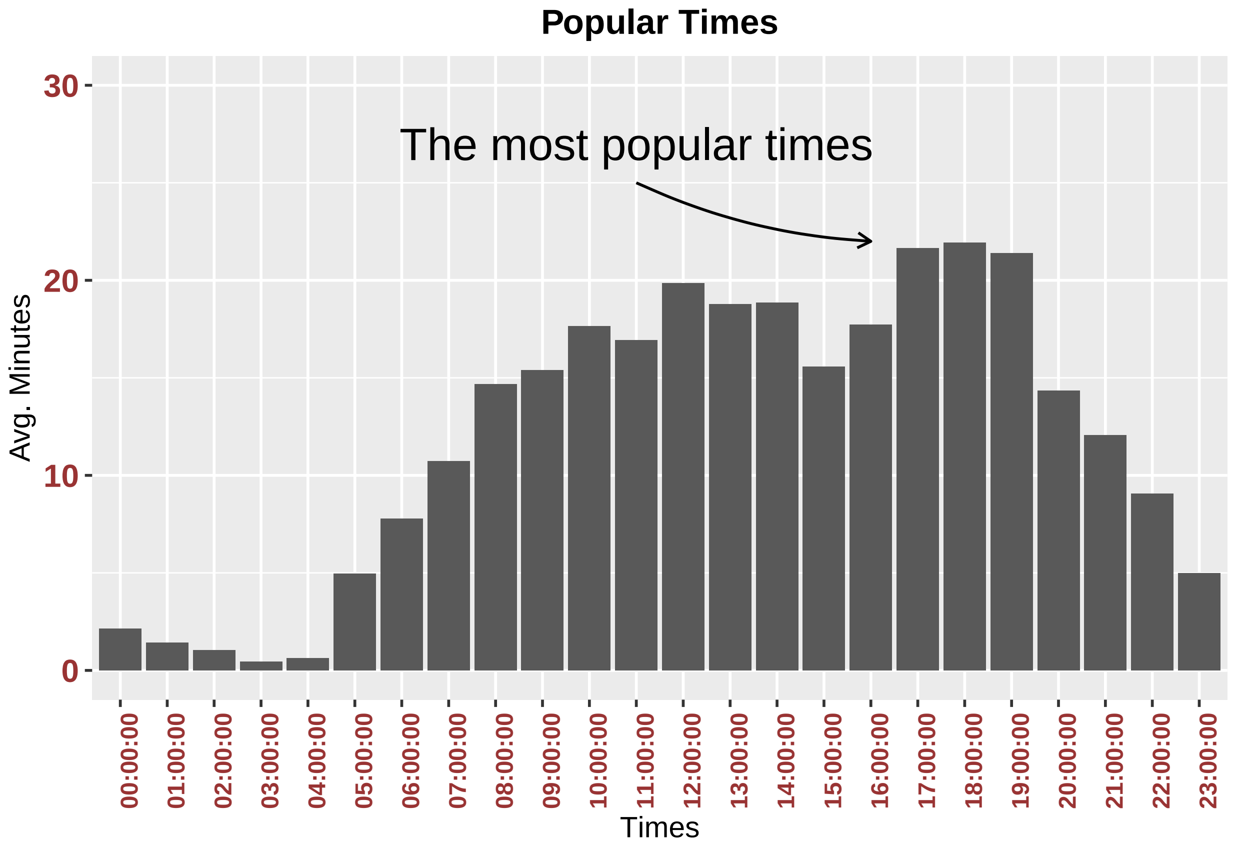
```
sleep %>%
  select(TotalMinutesAsleep, TotalTimeInBed) %>%
  drop_na() %>%
  summary()
```

4.3 Sleeping:

```
##  TotalMinutesAsleep TotalTimeInBed
##  Min.   : 58.0      Min.   : 61.0
##  1st Qu.:361.0      1st Qu.:403.0
##  Median :433.0      Median :463.0
##  Mean   :419.5      Mean   :458.6
##  3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :796.0      Max.   :961.0
```

4.5 Popular times: The following visualization shows that the most popular time period for physical activity is between 5 PM and 8 PM. Activities start to grow as early as 5 AM and slow down by 11 PM.

```
intensities %>%
  # Aggregating the average intensity
  group_by(time) %>%
  summarise(avg_intensities = mean(TotalIntensity)) %>%
  ggplot(aes(x = time, y = avg_intensities)) +
  geom_bar(stat = 'identity') +
  scale_y_continuous(limits = c(0, 30)) +
  # Emphasizing the peak activity
  annotate(geom = 'curve',
           x = '11:00:00', xend = '16:00:00',
           y = 25, yend = 22,
           curvature= 0.1,
           arrow = arrow(length = unit(2, 'mm'))) +
  annotate(geom = 'text',
           x = '11:00:00', y = 27,
           label = 'The most popular times',
           size = 6) +
  # Customizing non-data elements
  theme(axis.text.x = element_text(face = 'bold', color = '#993333', angle = 90, size = 9),
        axis.text.y = element_text(face = 'bold', color = '#993333', size = 12)) +
  labs(title = 'Popular Times',
       x = 'Times',
       y = 'Avg. Minutes') +
  theme(plot.title = element_text(face = 'bold', hjust = 0.5))
```

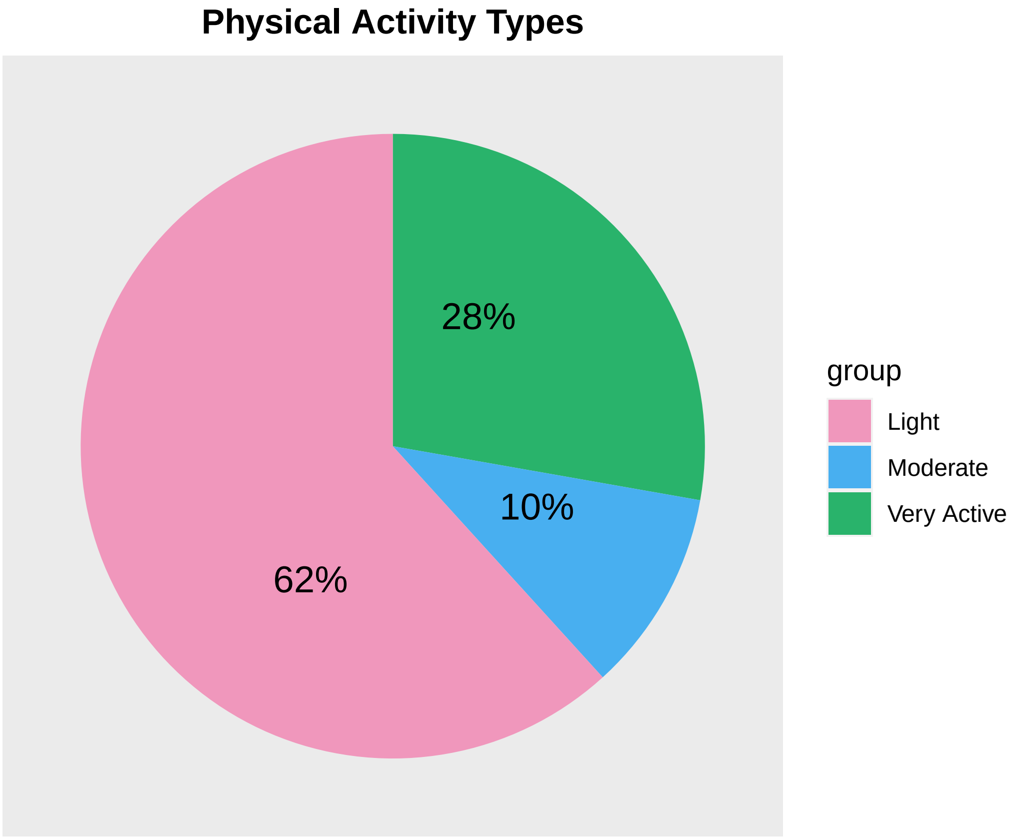


4.6 Types of activity The participants showed that light activities make up **62%** of the entire time spent working out. Although **sedentary activities** were excluded from the sample since they are not physical activities per se, the effect of **sedentary time** will be explored in 4.7.

```
# Calculating the total
distance1 <- activity %>%
  summarise(VeryActive = sum(VeryActiveDistance),
            ModerateActive = sum(ModeratelyActiveDistance),
            LightActive = sum(LightActiveDistance),
            Overall = VeryActive + ModerateActive + LightActive)

# Calculating the percentage
distance2 <- data.frame(group = c('Very Active', 'Moderate', 'Light'),
                        values = c(distance1$VeryActive / distance1$Overall * 100,
                                    distance1$ModerateActive / distance1$Overall * 100,
                                    distance1$LightActive / distance1$Overall * 100))

# Pie chart
distance2 %>%
  ggplot(aes(x = '', y = values, fill = group)) +
  geom_bar(width = 1, stat = 'identity') +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = c("#f097bc", "#48aff0", "#29b36b")) +
  # Clean non-data objects
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid = element_blank(),
        axis.ticks = element_blank(),
        axis.text = element_blank()) +
  # Label percentage
  geom_text(aes(y = values/3 + c(0, cumsum(values)[-length(values)]),
                label = percent(values/100)), size = 5) +  # percent() - library(scale)
  labs(title = 'Physical Activity Types') +
  theme(plot.title = element_text(face = 'bold', hjust = 0.5))
```

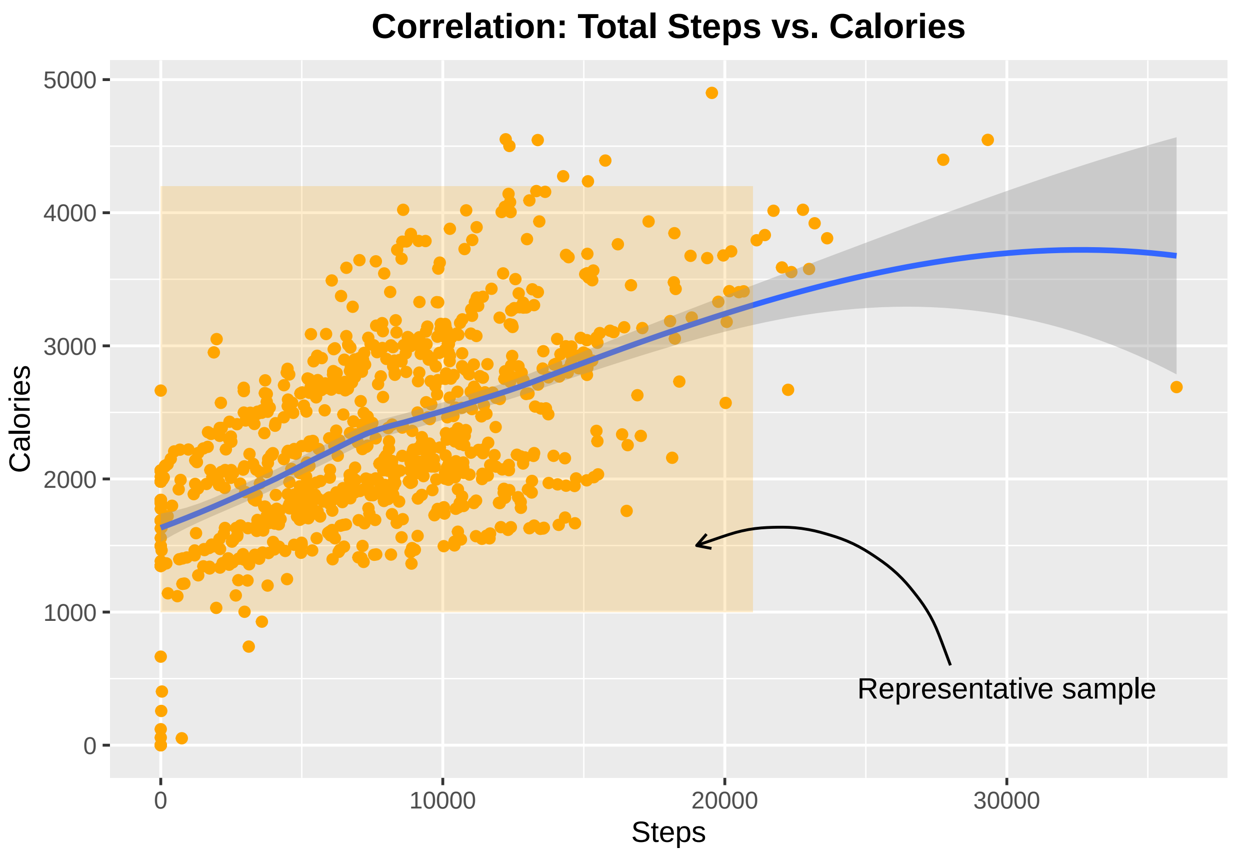


4.7 Correlations The function **cor.test** shows if there is a relationship between variables. Having tested the relationship between steps per day and calories, we can see the index **0.5915**. It means there is a moderate and positive relationship between these two variables.

```r
# Correlation test: Steps vs. Calories
cor.test(activity$TotalSteps, activity$Calories)
```

```
##
##  Pearson's product-moment correlation
##
## data:  activity$TotalSteps and activity$Calories
## t = 22.472, df = 938, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5483688 0.6316184
## sample estimates:
##       cor
## 0.5915681
```

```r
# Visualization
activity %>%
  ggplot(aes(x = TotalSteps, y = Calories)) +
  geom_jitter(color='orange') +
  geom_smooth() +
  # Customization
  annotate(geom = 'rect',
           xmin = 0, xmax = 21000,
           ymin = 1000, ymax = 4200,
           alpha = 0.2,
           fill = 'orange') +
  annotate(geom = 'curve',
           x = 28000, y = 600,
           xend = 19000, yend = 1500,
           curveture = 0.1,
           arrow = arrow(length = unit(2, 'mm'))) +
  annotate(geom = 'text',
           x = 30000, y = 430,
           label = 'Representative sample') +
  labs(title = 'Correlation: Total Steps vs. Calories',
       x = 'Steps',
       y = 'Calories') +
  theme(plot.title = element_text(face = 'bold', hjust = 0.5))
```

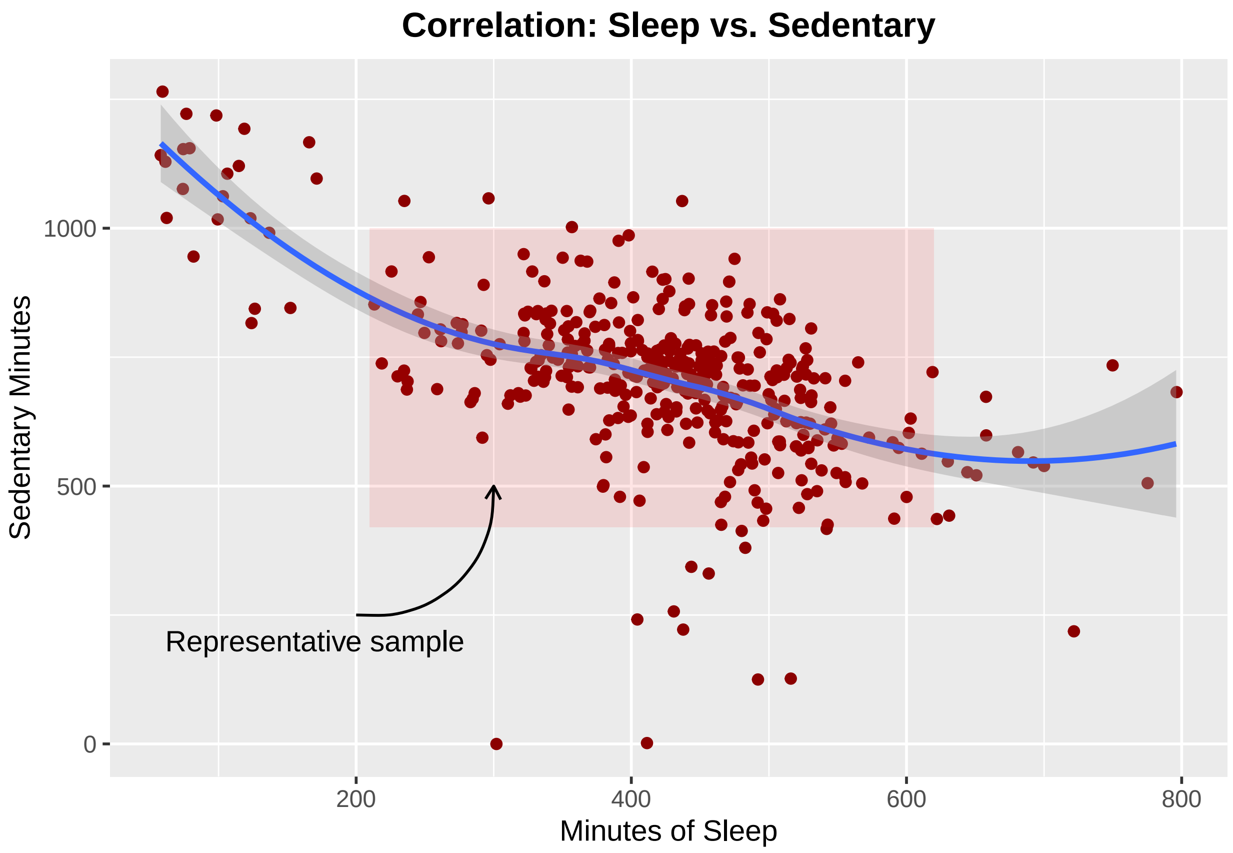


Having applied the same correlation function we can see the negative relationship between sedentary activity and time of sleep. Considering that the correlation is not linear (index = **-0.5993**), there is no distinctive impact of sedentary time on quality of sleep.

```
# Correlation test: Sleep vs. Sedentary
cor.test(merged$TotalMinutesAsleep, merged$SedentaryMinutes)
```

```
##
##  Pearson's product-moment correlation
##
## data:  merged$TotalMinutesAsleep and merged$SedentaryMinutes
## t = -15.181, df = 411, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6578402 -0.5337719
## sample estimates:
##       cor
## -0.599394
```

```
# Visualization
merged %>%
  ggplot(aes(x = TotalMinutesAsleep, y = SedentaryMinutes)) +
  geom_jitter(color = 'darkred') +
  geom_smooth() +
  # Customization
  annotate(geom = 'rect',
           xmin = 210, xmax = 620,
           ymin = 420, ymax = 1000,
           alpha = 0.1,
           fill = 'red') +
  annotate(geom = 'curve',
           x = 200, xend = 300,
           y = 250, yend = 500,
           curveture = 0.3,
           arrow = arrow(length = unit(2, 'mm'))) +
  annotate(geom = 'text',
           x = 170, y = 200,
           label = 'Representative sample') +
  labs(title = 'Correlation: Sleep vs. Sedentary',
       x = 'Minutes of Sleep',
       y = 'Sedentary Minutes') +
  theme(plot.title = element_text(face = 'bold', hjust = 0.5))
```



4.8 Summary:

- Average distance covered per day - 5.49 miles
- Average calories burned per day - 2304 kcal
- Average **sedentary time - 991.2 minutes**

- Participants exercise mostly through **light activity**

- The most popular time for physical activity is **between 5 PM and 8 PM** (After work time).

- Average hours of sleep **- 419.5 minutes**

- Moderate uphill relationship between **total steps** and **calories**
- Moderate downhill relationship between **time of sleep** and **sedentary time**

Recommendations

- As we see in paragraph 4.1, participants spent on average 991.2 minutes in sedentary state which is over 16 hours. Considering this number includes 7 hours of sleep (paragraph 4.3), the overall hours spent in sedentary state is 9 (37.5% of a day). According to the CDC, the harmful consequences of not getting enough physical activity are heart disease, diabetes, cancer, and obesity (link).
  It may be a good idea for Bellabeat to encourage people by implementing daily walking/working out challenges based on user expectations. Also this feature can turn the part of the "Light activity" segment, which is 62% (paragraph 4.6), into "Moderate activity".

- During the analysis we discovered the fact that the average time of sleep was nearly 7 hours (paragraph 4.3). According to the CDC guidance, 7 hours is the lowest threshold needed to contribute to your health (link).
  Bellabeat could track sleep patterns of users and give notifications to go to sleep based on the information collected. Logging the information could be also useful so that users are aware about their quality of sleep.

- Consistency is the key. Lack of motivation can slow down or hold back progress.
  The Bellabeat application can keep people motivated providing weekly reports so that users can see changes and show off their achievements.