Bellabeat Case Study: FitBit Usage Trend Analysis

Masaki Nakanishi

2022-09-01

Introduction

This project aims to investigate how people use existing smart devices and to give a high-level recommendation for how the findings can inform Bellabeat marketing strategy.

Bellabeat is a high-tech company that manufactures health-focused smart devices including bracelet, necklace, clip, watch, and water bottle. The smart bracelet, necklace, clip, and watch can track users' activity, sleep and stress. The smart water bottle can track users' daily water intake. All the smart devices can be connected to Bellabeat app so that users can review their comprehensive health-related habits and information.

In this project, we, Bellabeat marketing analytics team, analyzed a publicly-available dataset to gain an insight into how people who alreay own smart devices use them, and revealed potential marketing opportunities for Bellabeat's growth.

Questions

To identify some trends in smart device usage, we will aim to answer the following questions:

- Do people use their smart devices for tracking specific information (e.g., activity, sleep, and stress)? For example, do they use their devices for tracking activity rather than stress and sleep, or vice versa?
- Assuming that most people use their devices for tracking their activity, during what kinds of activities (e.g., light, moderate, and intensive activities) do they use them the most?

To address the questions, we will conduct a series of studies.

Study 1

Hypothesis

- People use their smart devices for tracking activity or stress rather than sleep.

  Since sleep quality can be tracked by different devices such as a smartphone, people who bought a wearable smart device might be interested in the information that can only be tracked by wearable devices.

Materials and Methods

We used FitBit Fitness Tracker Data (https://www.kaggle.com/datasets/arashnic/fitbit) to test the hypothesis. FitBit is also a company that manufactures smart devices. Their smart devices can also track users' activity and sleep. The FitBit Fitness Tracker Data contains personal tracker data including physical activity, heart rate, and sleep monitoring obtained from ## participants recorded between 04-12-2016 - 05-12-2016.

Using the data, we compared the participants' usage of the FitBit for monitoring their physical activity and sleep. The metrics we used were: * The number of days each participant used to track his/her physical activity * The number of days each participant used to track his/her sleep

We submitted the two metrics into a paired t-test to test a null-hypothesis: there is no difference in the mean numbers of days used to track phsyical activity and sleep.

Analysis

Data cleaning in SQL We first cleaned the data using SQL. We removed invalided observations where all the columns are 0 (indicating he/she did not use his/her device on that day) and computed the metrics (the numbers of days each participant used to track his/her physical activity and sleep) using the raw daily usage information in Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv and Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv using the following queries.

SELECT ID, COUNT(ActivityDate) AS NumOfActiveDays FROM `bellabeat-case-study-361204.FitBit_Fitness_Trac`
WHERE TotalDistance != 0 GROUP BY ID

SELECT ID, COUNT(SleepDay) AS NumOfSleepDays FROM `bellabeat-case-study-361204.FitBit_Fitness_Tracker_`
WHERE TotalSleepRecords != 0 GROUP BY ID

We then saved the tables as numOfDaysForActivityTrack.csv and numOfDaysForSleepTrack.csv.

Preparation A package that was necessary for the following analysis was installed as follows:

```
install.packages("tidyverse", repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/vx/d33k5vr53c5dl561vq7c67100000gn/T//Rtmpi2lnHS/downloaded_packages
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Two data tables were imported and merged together. In addtion, the data cells that contained NA were replaced by 0, and a new column that contained the participant ID in a character type was added.

```
activity_data = read_csv("Fitabase_Data_4_12_16-5_12_16/numOfDaysForActivityTrack.csv")
```

```
## Rows: 33 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): ID, NumOfActiveDays
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sleep_data = read_csv("Fitabase_Data_4_12_16-5_12_16/numOfDaysForSleepTrack.csv")
```

```
## Rows: 24 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): ID, NumOfSleepDays
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_merged = merge(x = activity_data, y = sleep_data, by = "ID", all = TRUE)
data_merged = mutate(data_merged, char_ID = as.character(data_merged$ID))
data_merged[is.na(data_merged)] <- 0
head(data_merged)
```

```
##           ID NumOfActiveDays NumOfSleepDays    char_ID
## 1 1503960366              30             25 1503960366
## 2 1624580081              31              0 1624580081
## 3 1644430081              30              4 1644430081
## 4 1844505072              20              3 1844505072
## 5 1927972279              17              5 1927972279
## 6 2022484408              31              0 2022484408
```

For the following visualization, the merged data were reformated to a long table with three columns including participant ID, purpose of using the device (i.e., Activity or Sleep), and the number of days they used the device for a purpose.

```
long_data = data.frame(Participant =  c(data_merged$char_ID, data_merged$char_ID), Purpose = rep(c('Acti
head(long_data)
```

```
##   Participant  Purpose NumOfDays
## 1  1503960366 Activity        30
## 2  1624580081 Activity        31
## 3  1644430081 Activity        30
## 4  1844505072 Activity        20
## 5  1927972279 Activity        17
## 6  2022484408 Activity        31
```

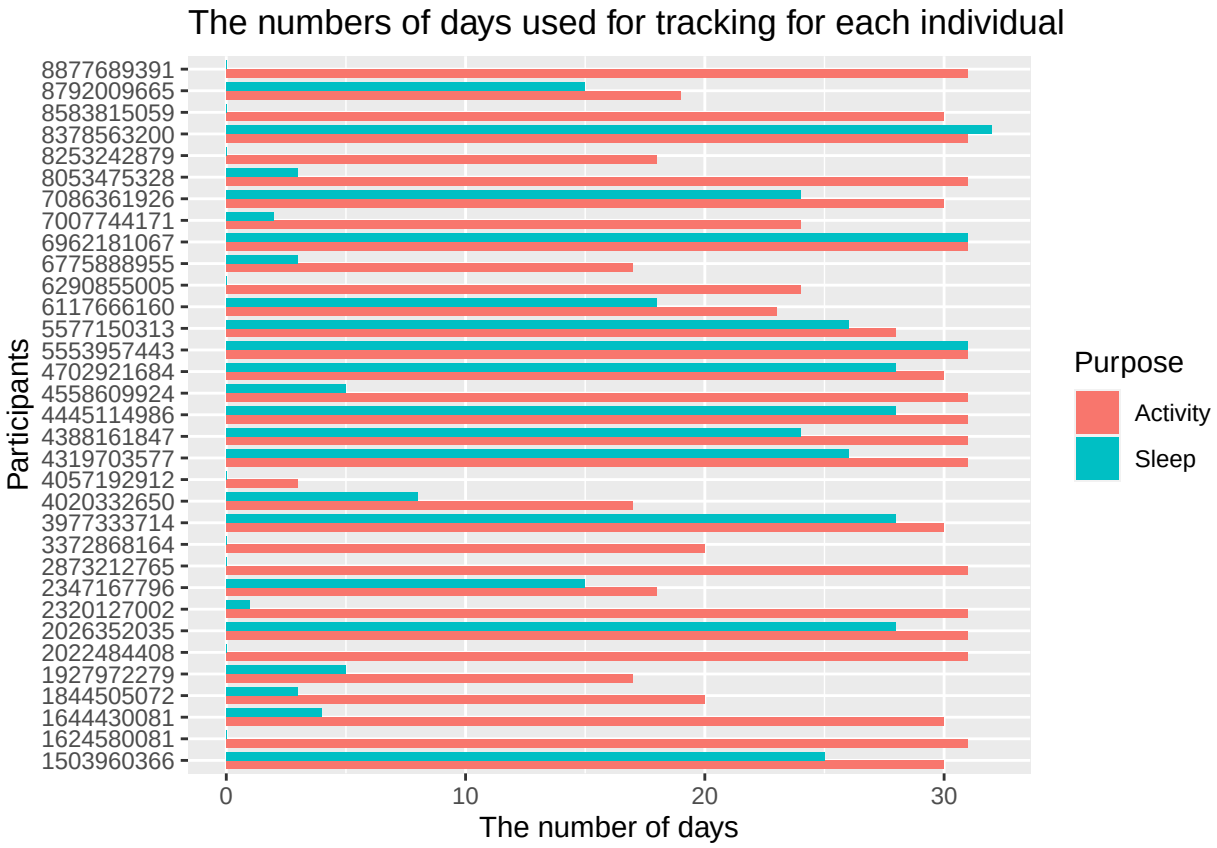
We also computed the averaged number of days for each purpose across participants and its standard errors.

```
sum_data = data.frame(Purpose = c("Activity","Sleep"), Average = c(mean(data_merged$NumOfActiveDays), m
sum_data
```

```
##    Purpose  Average       SE
## 1 Activity 26.12121 1.186858
## 2    Sleep 12.51515 2.171231
```

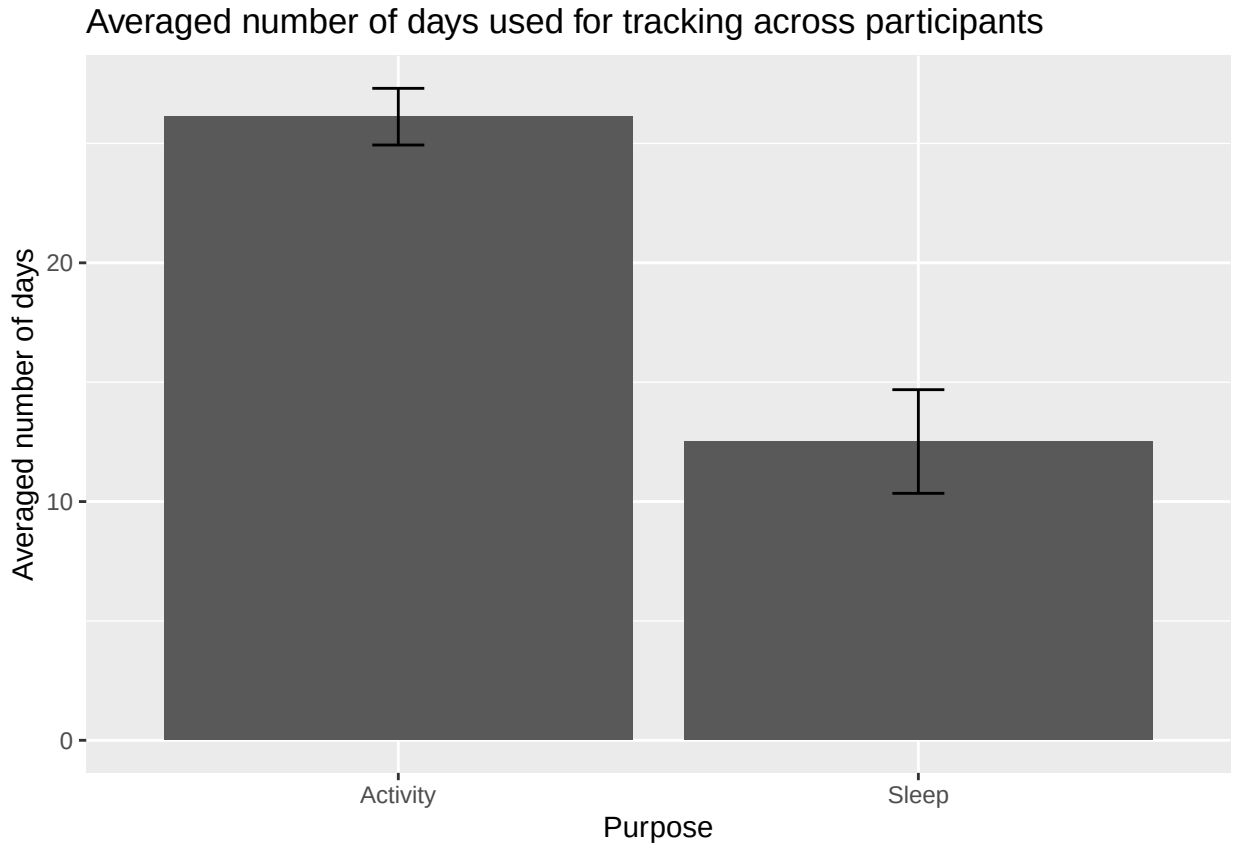
Visualization The following bar plot indicates the number of days each participant used the device for tracking activity and sleep.

```
ggplot(long_data, aes(y = Participant, x = NumOfDays, fill=Purpose)) +
  geom_bar(position = 'dodge', stat = "identity", width = 0.8) +
  labs(title="The numbers of days used for tracking for each individual", y = "Participants", x = "The
```



The following bar plot indicates their average across participants. The error bars indicate the standard error.

```
ggplot(sum_data, aes(x=Purpose, y=Average)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_errorbar(aes(ymin=Average-SE, ymax=Average+SE), width=.1, position=position_dodge(.9)) +
  labs(title="Averaged number of days used for tracking across participants", x="Purpose", y="Averaged
```



As the bar chart indicated, people tended to use their devices for tracking activity rather than sleep. A paired t-test revealed that there was a statistically signiticant difference between the activity and sleep ($p < 0.001$).

```
t.test(data_merged$NumOfActiveDays, data_merged$NumOfSleepDays, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  data_merged$NumOfActiveDays and data_merged$NumOfSleepDays
## t = 6.6733, df = 32, p-value = 1.569e-07
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   9.452973 17.759149
## sample estimates:
## mean difference
##        13.60606
```

```
#r = cor(data_merged$NumOfActiveDays, data_merged$NumOfSleepDays)
#ggplot(data_merged, mapping=aes(x=NumOfActiveDays, y=NumOfSleepDays)) + geom_point() +
#  geom_smooth(method="lm") +
#  labs(title="Association between the numbers of days used for tracking activity and sleep", x="The nu
#  annotate(geom = "text", x = 4, y = 31, label=sprintf("R = %.2f", r))
```

Summary

The results showed that people tended to use FitBit device for tracking their activity rather than their sleep, which supported our hypothesis. Note that the participants tracked their activity on 26.12 ± 1.19 days out of 31 days on average. On the other hand, they tracked their sleep only on 12.52 ± 2.17 days on average. This indicates that people are much more intersted in monitoring their activity than their sleep.

Study 2

We found that most of the FitBit users use their devices for tracking their activities. Now we are interested in investigating what kind of activities (e.g., light, moderate, and intensive) they perform while using the devices.

Hypothesis

- People use their smart devices for tracking light activity rather than moderate and intensive activities.

Since light activity such as walking is a part of daily activity, people use their smart devices mostly while performing light activity.

Materials and Methods

Again, we used FitBit Fitness Tracker Data (https://www.kaggle.com/datasets/arashnic/fitbit) to test the hypothesis.

Using the data, we compared the duration of using their FitBit for tracking different types of activity. The metrics we used were: * Total lightly, fairly, and very active duration in minutes. * Ratio of duration of each types of activities to the total active duration.

We submitted the two metrics into a one-way repeated measures analysis of variance (ANOVA) to test a null-hypothesis: there is no difference in the mean active durations among three types of activities.

Analysis

Data cleaning in SQL We removed invalided observations where all the columns are 0 (indicating he/she did not use his/her device on that day) and computed the metrics (total lightly, fairly, and very active duration in minutes, and Ratio of duration of each types of activities to the total active duration) using the raw daily usage information in Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv using the following queries.

WITH data_converted AS (SELECT ID, sum(LightlyActiveMinutes) AS LightlyActiveMinutesTotal, sum(FairlyActiveMinutes) AS FairlyActiveMinutesTotal, sum(VeryActiveMinutes) AS VeryActiveMinutesTotal, (sum(LightlyActiveMinutes)+sum(FairlyActiveMinutes)+sum(VeryActiveMinutes)) AS ActiveMinutesTotal FROM `bellabeat-case-study-361204.FitBit_Fitness_Tracker_Data.Daily_Activity` WHERE TotalDistance != 0 GROUP BY ID ORDER BY ID)

SELECT ID, LightlyActiveMinutesTotal, FairlyActiveMinutesTotal, VeryActiveMinutesTotal, ActiveMinutesTotal, LightlyActiveMinutesTotal/ActiveMinutesTotal*100 AS LightlyActivePerc, FairlyActiveMinutesTotal/ActiveMinutesTotal*100 AS FairlyActivePerc, VeryActiveMinutesTotal/ActiveMinutesTotal*100 AS VeryActivePerc FROM data_converted

We then saved the table as activityLogPercentage.csv for the following analysis in R. In R, the data table looks like as follows:

```
activity_merged = read_csv("Fitabase_Data_4_12_16-5_12_16/activityLogPercentage.csv")
```

```
## Rows: 33 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): ID, LightlyActiveMinutesTotal, FairlyActiveMinutesTotal, VeryActive...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(activity_merged)
```

```
## # A tibble: 6 x 8
##           ID LightlyActiveMinu~1 Fairl~2 VeryA~3 Activ~4 Light~5 Fairl~6 VeryA~7
##        <dbl>               <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1503960366                6818     594    1200    8612    79.2    6.90  13.9
## 2 1624580081                4758     180     269    5207    91.4    3.46   5.17
## 3 1644430081                5354     641     287    6282    85.2   10.2    4.57
## 4 1844505072                3578      40       4    3622    98.8    1.10   0.110
## 5 1927972279                1196      24      41    1261    94.8    1.90   3.25
## 6 2022484408                7981     600    1125    9706    82.2    6.18  11.6
## # ... with abbreviated variable names 1: LightlyActiveMinutesTotal,
## #   2: FairlyActiveMinutesTotal, 3: VeryActiveMinutesTotal,
## #   4: ActiveMinutesTotal, 5: LightlyActivePerc, 6: FairlyActivePerc,
## #   7: VeryActivePerc
```

We converted the data table to a long format from a wide format and added a new column that contained the participant ID in a character type was added as follows:

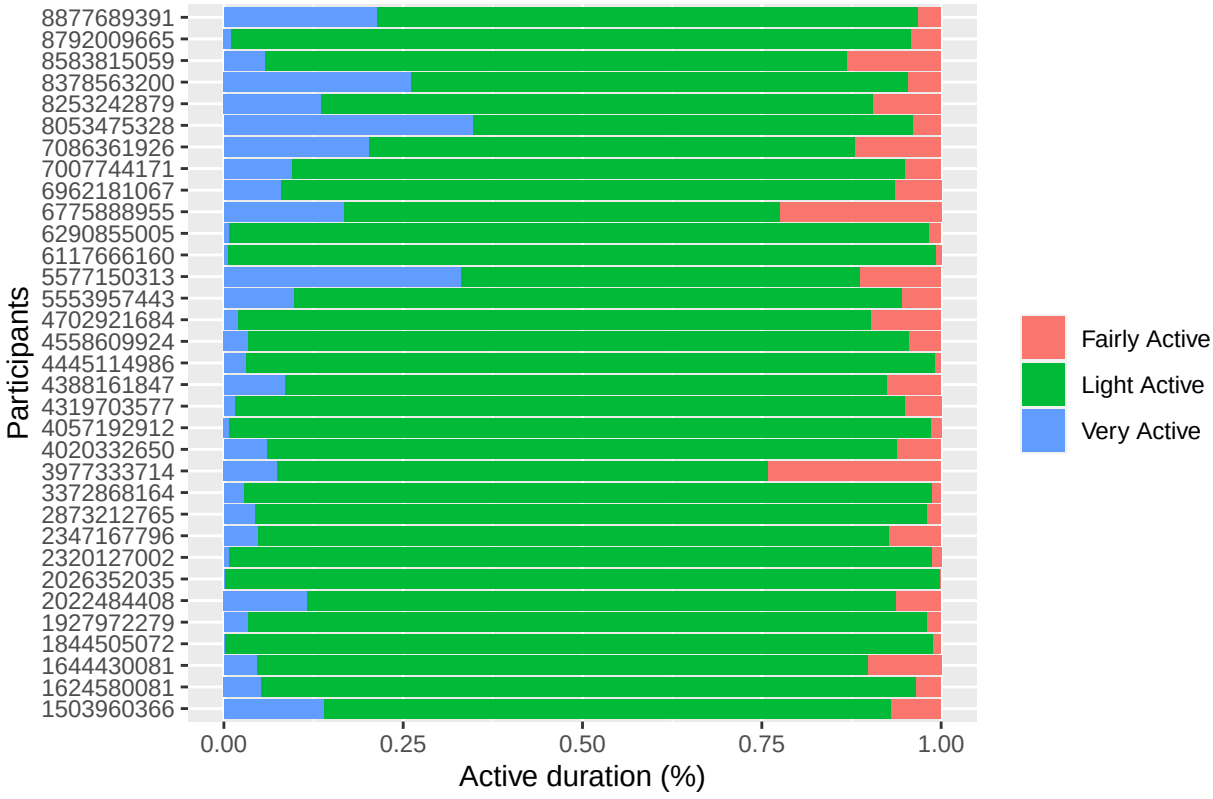
```
activity_wide = summarize(activity_merged, ID, LightlyActivePerc, FairlyActivePerc, VeryActivePerc)
activity_long = gather(activity_wide, Activity, Percentage, -ID)
activity_long = mutate(activity_long, char_ID = as.character(activity_long$ID))
head(activity_long)
```

```
## # A tibble: 6 x 4
##           ID Activity          Percentage char_ID
##        <dbl> <chr>                  <dbl> <chr>
## 1 1503960366 LightlyActivePerc       79.2 1503960366
## 2 1624580081 LightlyActivePerc       91.4 1624580081
## 3 1644430081 LightlyActivePerc       85.2 1644430081
## 4 1844505072 LightlyActivePerc       98.8 1844505072
## 5 1927972279 LightlyActivePerc       94.8 1927972279
## 6 2022484408 LightlyActivePerc       82.2 2022484408
```

The followin chart indicates how long each participants have used their devices for tracking light, moderate, and intensive activity in percentage. As shown in the chart, all the participants mostly used their devices for monitoring light activity.

```
ggplot(activity_long, aes(y = char_ID, x = Percentage, fill=Activity)) +
  geom_col(position="fill", stat="identity") +
  #theme(axis.text.x = element_text(angle=90)) +
  labs(title = "", y = "Participants", x = "Active duration (%)") +
  scale_fill_discrete("", labels = c('Fairly Active', 'Light Active', 'Very Active'))
```

```
## Warning: Ignoring unknown parameters: stat
```




```
  #scale_fill_manual("", values = c('FairlyActivePerc'='blue', 'LightlyActivePerc'='red', 'VeryActivePe
```

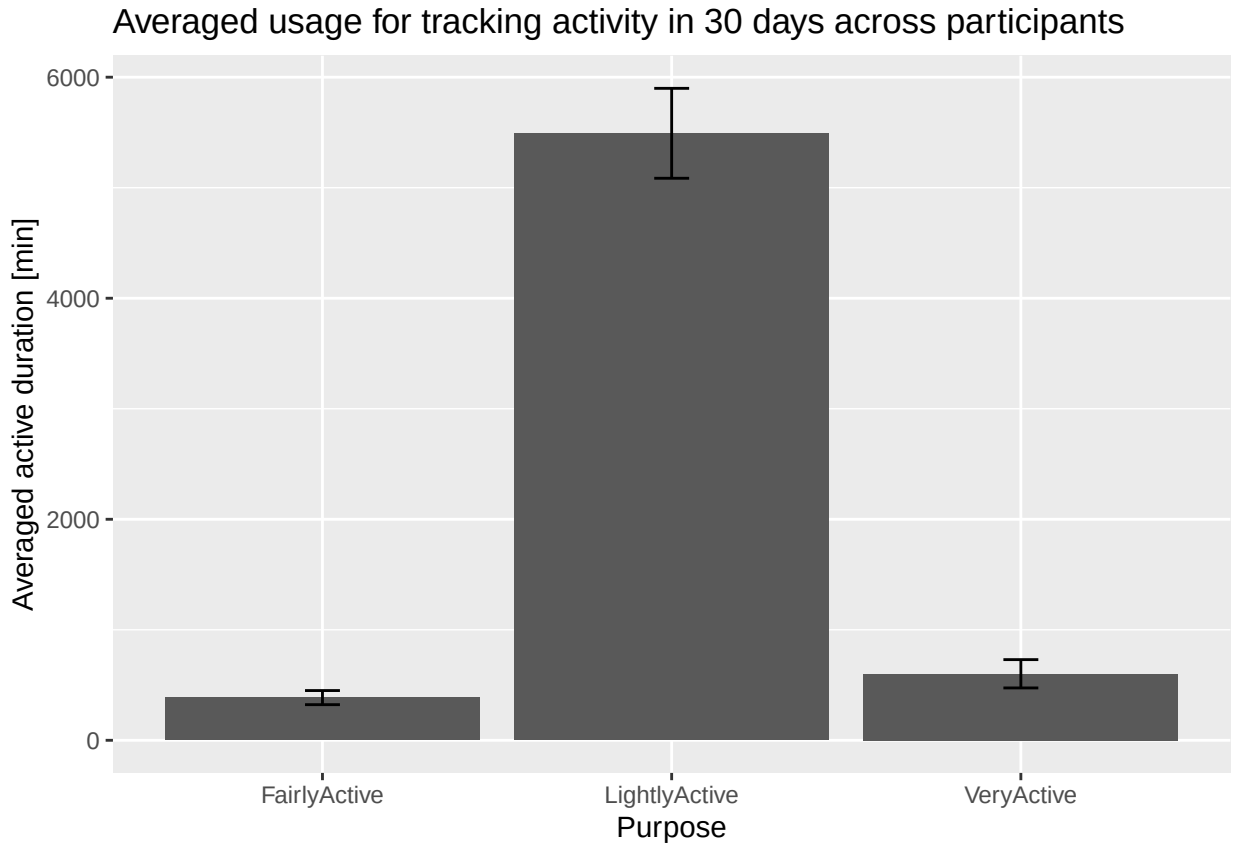
We also computed the averaged duration for each activity across participants and its standard errors as follows:

```
activity_sum = data.frame(Purpose = c("LightlyActive","FairlyActive", "VeryActive"), Average = c(mean(a
activity_sum
```

```
##         Purpose   Average        SE
## 1 LightlyActive 5492.2121 406.89756
## 2  FairlyActive  386.3939  63.81722
## 3    VeryActive  601.8788 128.08886
```

The following bar plot indicates the averaged duration the participants used their devices to monitor each activity in minutes.

```
ggplot(activity_sum, aes(x=Purpose, y=Average)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_errorbar(aes(ymin=Average-SE, ymax=Average+SE), width=.1, position=position_dodge(.9)) +
  labs(title="Averaged usage for tracking activity in 30 days across participants", x="Purpose", y="Ave
```



To perform a statistical test, we made another long data table that contained total duration for each activity for each individual as follows:

```
active_min_wide = summarize(activity_merged, ID, LightlyActiveMinutesTotal, FairlyActiveMinutesTotal, V
active_min_long = gather(active_min_wide, Activity, Duration, -ID)
active_min_long = mutate(active_min_long, char_ID = as.character(active_min_long$ID))
head(active_min_long)
```

```
## # A tibble: 6 x 4
##           ID Activity                   Duration char_ID
##        <dbl> <chr>                         <dbl> <chr>
## 1 1503960366 LightlyActiveMinutesTotal      6818 1503960366
## 2 1624580081 LightlyActiveMinutesTotal      4758 1624580081
## 3 1644430081 LightlyActiveMinutesTotal      5354 1644430081
## 4 1844505072 LightlyActiveMinutesTotal      3578 1844505072
## 5 1927972279 LightlyActiveMinutesTotal      1196 1927972279
## 6 2022484408 LightlyActiveMinutesTotal      7981 2022484408
```

As the bar plot indicated, all the participants tended to use their devices for tracking light activity. An one-way repeated measures ANOVA revealed that there was a statistically singificant main effect of the type of activity ($p < 0.001$) as follows:

```
one.way <- aov(Duration ~ Activity, data = active_min_long)
summary(one.way)
```

```
##             Df    Sum Sq   Mean Sq F value Pr(>F)
## Activity     2 550342905 275171452   134.5 <2e-16 ***
## Residuals   96 196463543   2046495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

The results showed that people are nearly only interested in tracking light activity rather than moderate and intensive activities using their FitBit, which supported our hypothesis. The averaged duration for tracking light activity across participants was 5492.21 $\pm$ 405.90 min, which was about 85% of the total activity tracking duration, and those for tracking moderate and intensive activities were 386.29 $\pm$ 63.82 min and 601.88 $\pm$ 128.09 min, respectively.

General discussions and recommendation

In this project, we analyzed the FitBit usage data to investigate how FitBit users have been using their devices to inform Bellabeat's marketing stragety. The analysis revealed that:

- FitBit users are more interested in monitoring their active duration.
- FitBit users are in general lightly active.

Assuming that potential users of FitBit and Bellabeat's devices are greatly overlapped, this analytics project suggested that the potential users of Bellabeat's products would be more interested in monitoring casual activities (e.g., walking). Therefore, Bellabeat should put in more efforts on implementing products, functionality and services related to monitoring and enhancing activity and advertise them. For example:

- Produce many different types of devices that can be worn casually and looks nice on casual clothes. The existing necklace and bracelet devices would be a Bellabeat's strenth as FitBit does not have any devices like them. Improving the design of the devices, e.g., by collaborating with famous fashion brands, would attract more customers.
- Add functions to keep up users' motivation to be active. For example, when a device detect that an user is not active, the device reminds him/her to stand up, go for a walk, and so on.

The limitation of this analysis is that we did not have data related to the users' stress tracking. Therefore, there might be some users who are more interested in tracking their stress levels than their activities.