BellaBeats CaseStudy

Malav

11/07/2021

About the company

Bellabeat is a high-tech company that manufactures health-focused smart products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in theglobal smart device market. Urška Sršen and Sando Mur founded Bellabeat.

Business task

You are a junior data analyst working on the marketing analyst team at Bellabeat. You have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy.

Ask

Questions for the analysis *What are some trends in smart device usage?* How could these trends apply to Bellabeat customers? *How could these trends help influence Bellabeat marketing strategy?

Stakeholders

1) **Primary**: Bellabeat's cofounder and Chief Creative Officer
2) **Secondary**: Bellabeat marketing analytics team and executive team

Prepare

Data Source *FitBit Fitness Tracker Data* (CC0: Public Domain, dataset made available through Mobius): This Kaggle 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 users' habits.

Data Limitations

1) The 30 user sample size may not fully represent the population
2) Bellabeats only makes product for women, but the data seems to contain no gender. That is, We may not know whether the user is male or female.

Process

```
library(here)
```

Loading the packages

```
## here() starts at D:/Data Analytics/Project/Case Studies/BellaBeats Case study - 2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

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

```
activity <- read.csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
calories <- read.csv("Fitabase Data 4.12.16-5.12.16/hourlyCalories_merged.csv")
intensity <- read.csv("Fitabase Data 4.12.16-5.12.16/hourlyIntensities_merged.csv")
sleep <- read.csv("Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")
weight <- read.csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv")
```

```
head(activity)
```

Importing datasets

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       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(calories)
```

```
##           Id         ActivityHour Calories
## 1 1503960366 4/12/2016 12:00:00 AM       81
## 2 1503960366  4/12/2016 1:00:00 AM       61
## 3 1503960366  4/12/2016 2:00:00 AM       59
## 4 1503960366  4/12/2016 3:00:00 AM       47
## 5 1503960366  4/12/2016 4:00:00 AM       48
## 6 1503960366  4/12/2016 5:00:00 AM       48
```

```
head(intensity)
```

```
##           Id         ActivityHour TotalIntensity AverageIntensity
## 1 1503960366 4/12/2016 12:00:00 AM             20         0.333333
## 2 1503960366  4/12/2016 1:00:00 AM              8         0.133333
## 3 1503960366  4/12/2016 2:00:00 AM              7         0.116667
## 4 1503960366  4/12/2016 3:00:00 AM              0         0.000000
## 5 1503960366  4/12/2016 4:00:00 AM              0         0.000000
## 6 1503960366  4/12/2016 5:00:00 AM              0         0.000000
```

```
head(sleep)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

```
head(weight)
```

```
##           Id                  Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

All these datsets have Id Column in common. This Information maybe useful if we want to merge the datasets.

Formatting the data Date is not as per our requirment, so it must be formatted.

```
# intensities
intensity$ActivityHour=as.POSIXct(intensity$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone
intensity$time <- format(intensity$ActivityHour, format = "%H:%M:%S")
intensity$date <- format(intensity$ActivityHour, format = "%m/%d/%y")
# calories
calories$ActivityHour=as.POSIXct(calories$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone(
calories$time <- format(calories$ActivityHour, format = "%H:%M:%S")
calories$date <- format(calories$ActivityHour, format = "%m/%d/%y")
# activity
activity$ActivityDate=as.POSIXct(activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone())
activity$date <- format(activity$ActivityDate, format = "%m/%d/%y")
# sleep
sleep$SleepDay=as.POSIXct(sleep$SleepDay, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
sleep$date <- format(sleep$SleepDay, format = "%m/%d/%y")
```

```
n_distinct(activity$Id)
```

Checking the number of unique IDs in datasets

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

```
n_distinct(calories$Id)
```

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

```
n_distinct(intensity$Id)
```

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

```
n_distinct(sleep$Id)
```

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

```
n_distinct(weight$Id)
```

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

We have 33 unique IDs in activity, calories and intensity datasets. 24 in sleep and 8 in weight dataset. Having only 8 parcipants will not be able to contribute towards any kind conclusions or reccommendations.

Analyze

Quick statistical summary of the datsaets

1) Activity

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

```
##    TotalSteps    TotalDistance    SedentaryMinutes    Calories   
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :   0  
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:1828  
##  Median : 7406   Median : 5.245   Median :1057.5   Median :2134  
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   :2304  
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.:2793  
##  Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :4900  
```

2) Distance

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

```
##  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
##  SedentaryActiveDistance
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.001606
##  3rd Qu.:0.000000
##  Max.   :0.110000
```

3) Minutes Active

```
# explore num of active minutes per category
activity %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%
  summary()
```

```
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0
##  Median :  4.00    Median :  6.00      Median :199.0
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0
```

4) Calories Burnt

```
# calories
calories %>%
  select(Calories) %>%
  summary()
```

```
##     Calories
##  Min.   : 42.00
##  1st Qu.: 63.00
##  Median : 83.00
##  Mean   : 97.39
##  3rd Qu.:108.00
##  Max.   :948.00
```

5) Sleep Record

```
# sleep
sleep %>%
  select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

6) Weight and BMI

```
# weight
weight %>%
  select(WeightKg, BMI) %>%
  summary()
```

```
##     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
```

Some findings from the summary above

1) 7638 is the average number of steps taken by participants, which is a little less. According to healthline, 10,000 steps/day is a reasonable target for healthy adults.
2) 991 minutes or around 16.5 hours is the average sedentary time. It must be reduced in order to be active.
3) A person sleep an average of 7 hours in a day, which seems reasonable.

Merging data We will be merging the two datasets, sleep and activity, by IDs and date in order to visualize. Also, We will be using inner join.

```
merged <- merge(sleep, activity, by=c('Id', 'date'))
head(merged)
```

```
##           Id     date   SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 04/12/16 2016-04-12                 1                327
## 2 1503960366 04/13/16 2016-04-13                 2                384
## 3 1503960366 04/15/16 2016-04-15                 1                412
## 4 1503960366 04/16/16 2016-04-16                 2                340
## 5 1503960366 04/17/16 2016-04-17                 1                700
## 6 1503960366 04/19/16 2016-04-19                 1                304
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346   2016-04-12      13162          8.50            8.50
## 2            407   2016-04-13      10735          6.97            6.97
## 3            442   2016-04-15       9762          6.28            6.28
## 4            367   2016-04-16      12669          8.16            8.16
## 5            712   2016-04-17       9705          6.48            6.48
## 6            320   2016-04-19      15506          9.88            9.88
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.14                     1.26
## 4                        0               2.71                     0.41
## 5                        0               3.19                     0.78
## 6                        0               3.53                     1.32
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                2.83                       0                29
## 4                5.04                       0                36
## 5                2.51                       0                38
## 6                5.03                       0                50
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  34                  209              726     1745
## 4                  10                  221              773     1863
## 5                  20                  164              539     1728
## 6                  31                  264              775     2035
```

```
weight_activity_merged <- merge(activity, weight, by="Id")
colnames(weight_activity_merged)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"                 "date"
## [17] "Date"                     "WeightKg"
## [19] "WeightPounds"             "Fat"
## [21] "BMI"                      "IsManualReport"
## [23] "LogId"
```

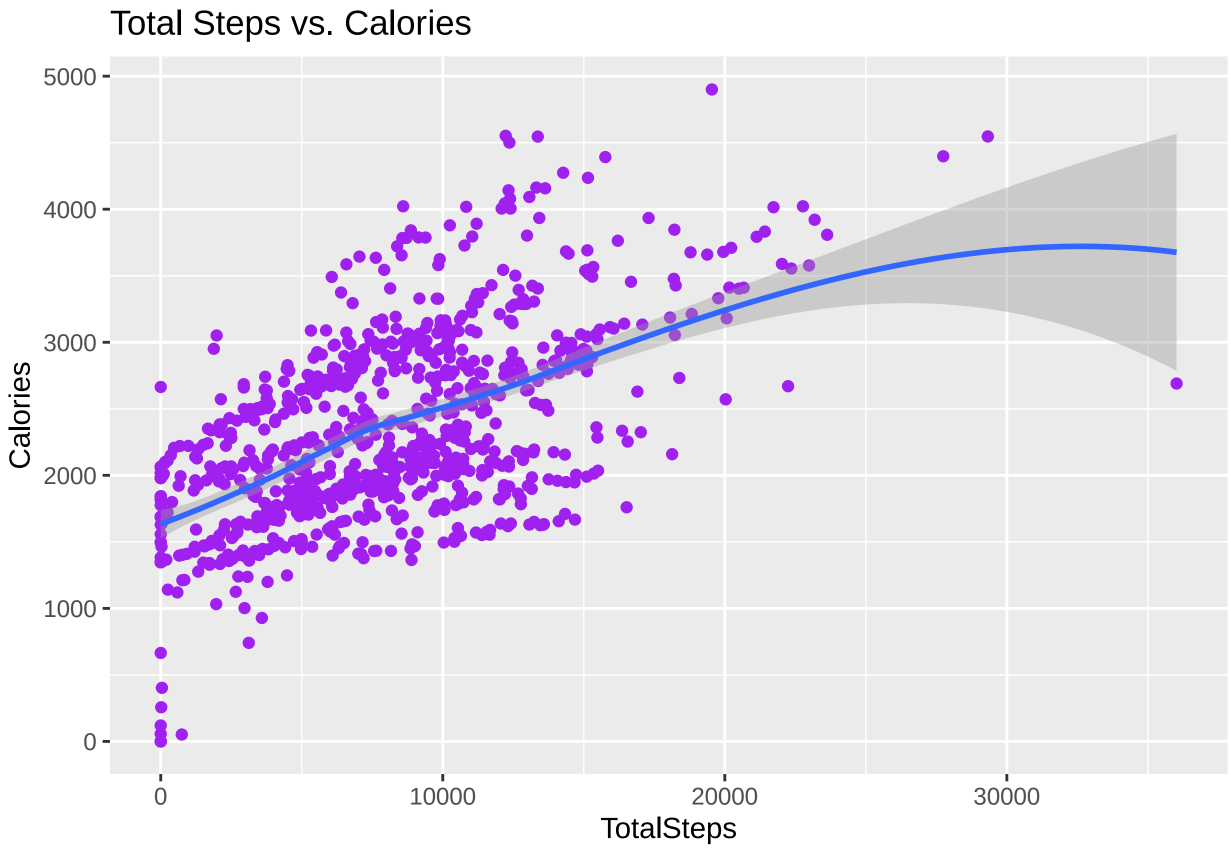
Share

Creating visualizations

1) Total steps vs Calories

```
ggplot(data=activity) +
  geom_point(mapping = aes(x=TotalSteps, y=Calories), color = 'purple') +
  geom_smooth(mapping = aes(x=TotalSteps, y=Calories)) + labs(title="Total Steps vs. Calories")
```

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

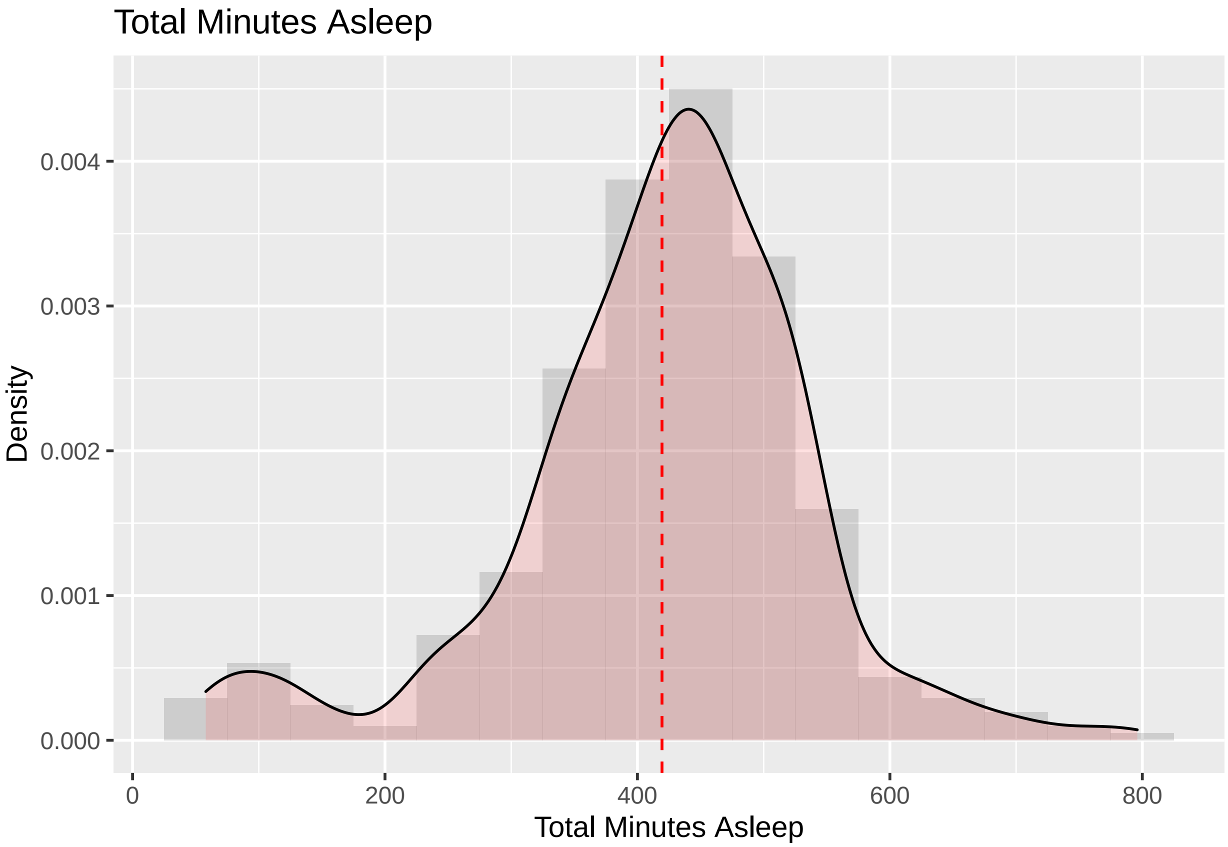


We can safely assume the positive correlation between calories and total steps.It is justifies by the graph. It is also logical as more we walk more, calories we will burn

2) Distribution of total minutes of sleep

```
ggplot(data = sleep, aes(x = sleep$TotalMinutesAsleep)) +
  geom_histogram(aes(y=..density..), binwidth=50,alpha=0.2)+
  geom_density(alpha=0.2, fill="#FF6666") +
  geom_vline(aes(xintercept=mean(TotalMinutesAsleep, na.rm=T)), color="red", linetype="dashed")+
  labs(title="Total Minutes Asleep", x= "Total Minutes Asleep", y="Density")
```

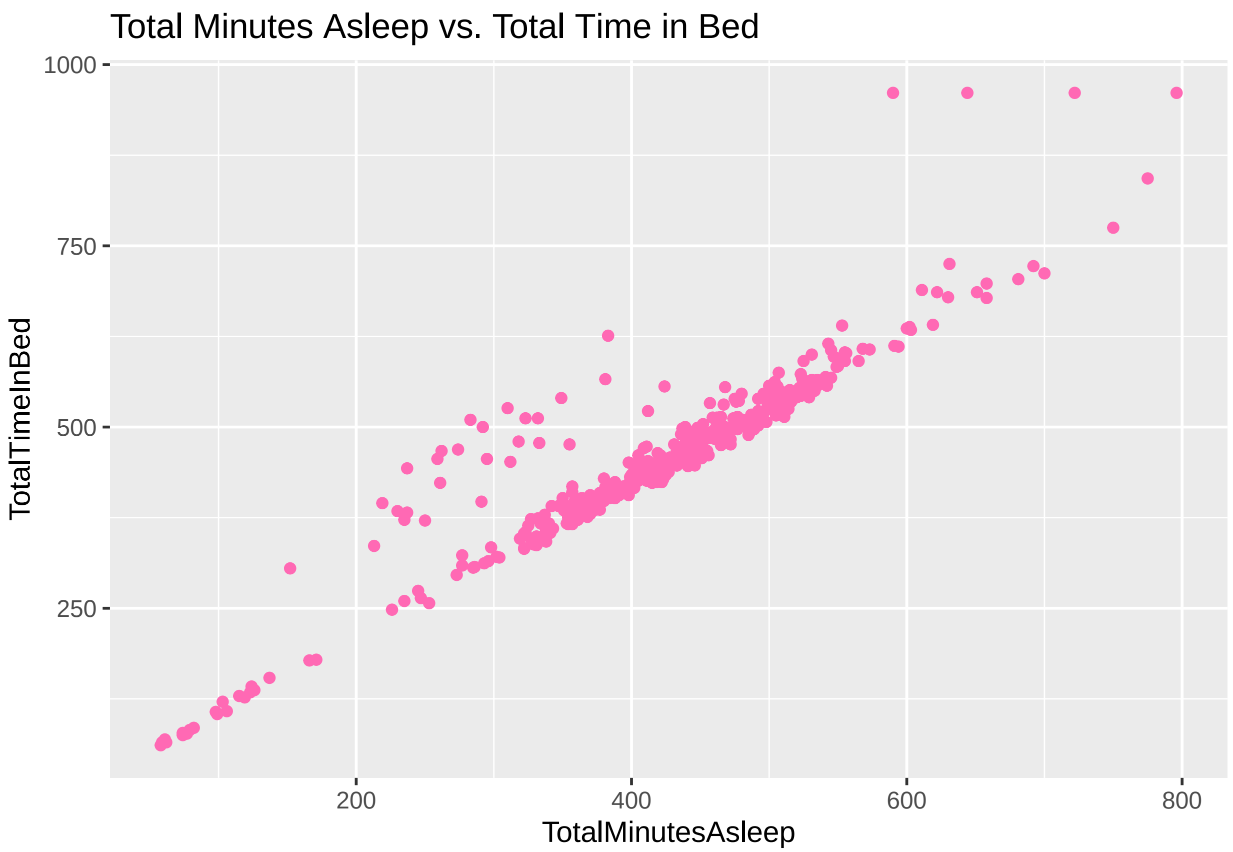
```
## Warning: Use of ‘sleep$TotalMinutesAsleep‘ is discouraged. Use
## ‘TotalMinutesAsleep‘ instead.

## Warning: Use of ‘sleep$TotalMinutesAsleep‘ is discouraged. Use
## ‘TotalMinutesAsleep‘ instead.
```



Sleep time is normally distrbutes among particaipants

3) Toal minutes asleep VS Total time in bed

```
ggplot(data=sleep) +
  geom_point(mapping = aes(x=TotalMinutesAsleep, y=TotalTimeInBed) , color = 'hotpink')+ labs(title="To
```



We can clearly see that the relationship between total minutes asleep and total time in bed is linear

4) Weight Vs Category

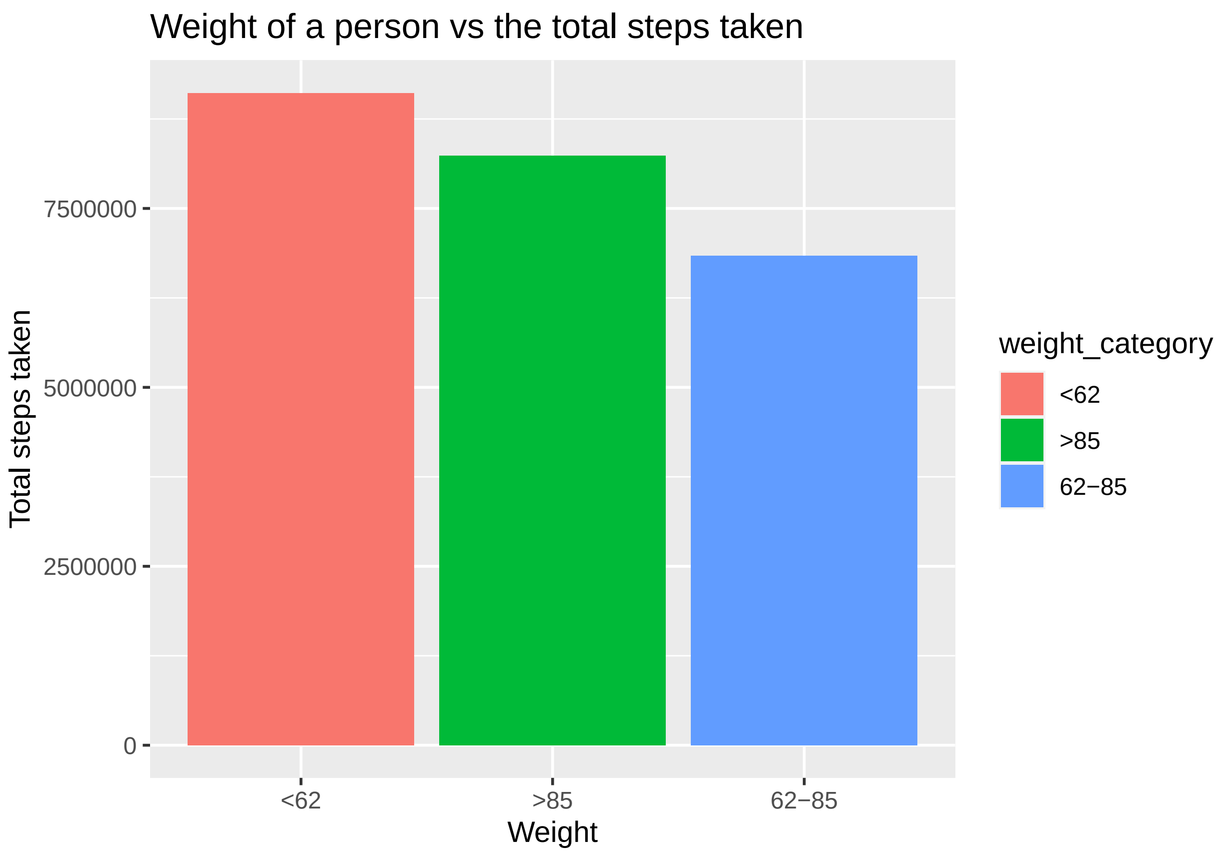
Firstly we want to categorize people into weight category so that it is easier for us to visulaize and draw conclusions.

```
weight_weight_activity_merged_v2 <- weight_activity_merged %>%
  mutate(weight_category = case_when(WeightKg>62 & WeightKg <85  ~ "62-85", WeightKg >=85 ~ ">85", TRUE
```

```
ggplot(data = weight_weight_activity_merged_v2) +
  geom_col(mapping = aes(x = weight_weight_activity_merged_v2$weight_category, y = weight_weight_activi
  labs(title="Weight of a person vs the total steps taken",
x = "Weight", y = "Total steps taken")
```

```
## Warning: Use of ‘weight_weight_activity_merged_v2$weight_category‘ is
## discouraged. Use ‘weight_category‘ instead.
```

```
## Warning: Use of ‘weight_weight_activity_merged_v2$TotalSteps‘ is discouraged.
## Use ‘TotalSteps‘ instead.
```



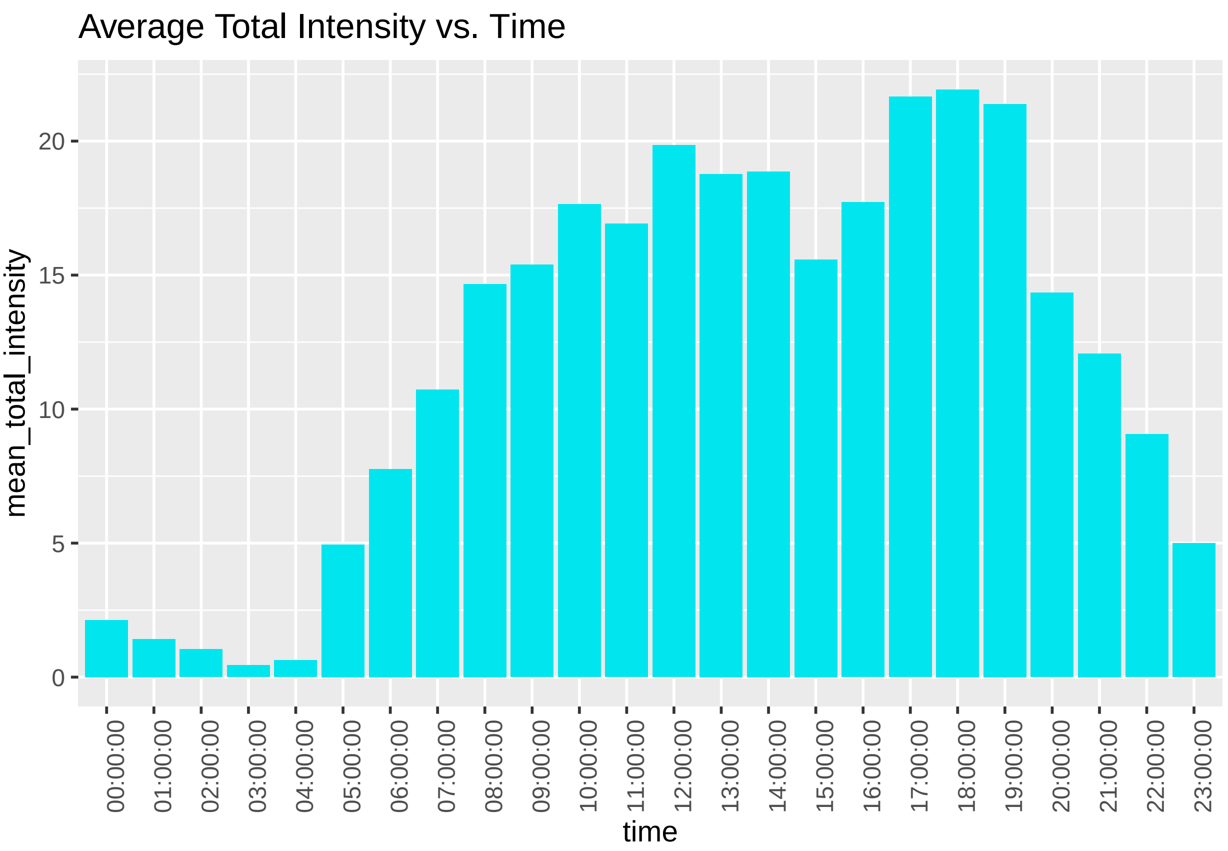
We can clearly see that people who weight less than 62 tend to walk more. So fitbit could recommend people with with greater than 62 to walk more in order to stay fit

5) Average Total intensity hourly

```
intensity_new <- intensity %>%
  group_by(time) %>%
  drop_na() %>%
  summarise(mean_total_intensity = mean(TotalIntensity))
```

```
ggplot(data=intensity_new, aes(x=time, y=mean_total_intensity)) + geom_histogram(stat = "identity", fil
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Average Total Intensity vs. Time")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Here we can clearly see that people tend to be more active between 5 Am and 7 Pm. Maybe after 7 Pm we can we people to take a walk or hit the gym after office hours

6) Active Minutes by weekdays

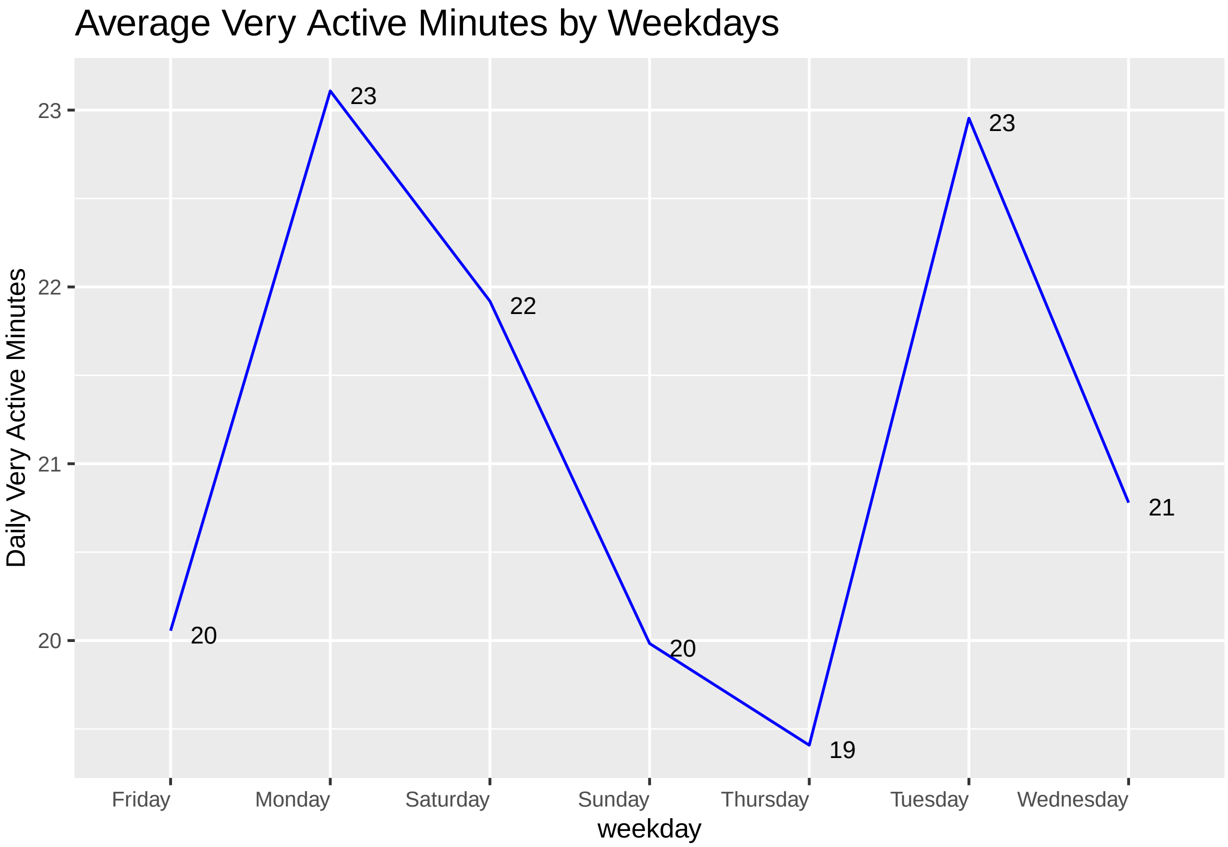
Changing the format of the date so that we can extract the weekday from it.

```
activity$ActivityDate=as.POSIXct(activity$ActivityDate, format="%Y-%m-%d", tz=Sys.timezone())
activity$date <- format(activity$ActivityDate, format = "%Y-%m-%d")
```

Extracting the weekday from date.

```
activity$weekday <- weekdays(as.POSIXct(activity$date), abbreviate = F)
```

```
average_very_active_minutes <-activity %>%
  group_by(weekday) %>%
  summarise_at(vars(VeryActiveMinutes),
             list(VeryActiveMinutes = mean))

ggplot(average_very_active_minutes, aes(x=weekday,y=VeryActiveMinutes, group=1))+
  geom_line(color = "blue")+
  labs(title="Average Very Active Minutes by Weekdays", x= "weekday", y="Daily Very Active Minutes")+
  geom_text(aes(label=round(VeryActiveMinutes, digits=0), hjust=-0.75, vjust=0.75),size=3)+
  theme(plot.title = element_text(size=14), text = element_text(size=10), axis.text.x = element_text(an
```



Lowest activty level are on thursdays, fridays and sundays. People start the week being motivated, but get demotivated by mid-week maybe due to work stress.

Act

Recommendation:-

- Bellabeat app must remind people with weight between 62 - 80 to walk more frequently.
- App to remind people to workout post office hours.
- App must motivate people to workout near the weekends.
- App needs to remind people to reduce sedentary time.
- App need to set a target of 9000 steps / day which needs to be completed by person. Or remind the person if the task is not completed.