# Data Analysis Case study: How Can a Wellness Technology Company Play It Smart?

## Loading libraries

```
library(tidyverse)
library(readr)
```

## Importing needed CSV files

For First analysis here i have chosen two csv files from the data set which are for activity data and sleep data.

```
Activity <- read.csv('C:/Users/LENOVO/Desktop/Case Study_How a wellness company play it smart(Bellabeat)/Fitabase
Data 4.12.16-5.12.16/Selected_For_case_study_V.1/dailyActivity_merged_v.2.csv')
SleepDay <- read.csv("C:/Users/LENOVO/Desktop/Case Study_How a wellness company play it smart(Bellabeat)/Fitabase
Data 4.12.16-5.12.16/Selected_For_case_study_V.1/sleepDay_merged_v.2.csv")
```

## Exploring files

Now, that we have imported necessary csv files let’s get summary of data via multiple functions that R provides.

```
head(Activity)
```

```
##           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(SleepDay)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366   04/12/2016 00:00:00                 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
```

```
n_distinct(Activity)
```

```
## [1] 862
```

```
summary(Activity)
```

```
##        Id             ActivityDate        TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:862         Min.   :    8   Min.   : 0.010
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 4927   1st Qu.: 3.373
##  Median :4.445e+09   Mode  :character   Median : 8054   Median : 5.590
##  Mean   :4.861e+09                      Mean   : 8329   Mean   : 5.986
##  3rd Qu.:6.962e+09                      3rd Qu.:11096   3rd Qu.: 7.905
##  Max.   :8.878e+09                      Max.   :36019   Max.   :28.030
##  TrackerDistance  LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.010   Min.   :0.000            Min.   : 0.000
##  1st Qu.: 3.373   1st Qu.:0.000            1st Qu.: 0.000
##  Median : 5.590   Median :0.000            Median : 0.410
##  Mean   : 5.971   Mean   :0.118            Mean   : 1.639
##  3rd Qu.: 7.880   3rd Qu.:0.000            3rd Qu.: 2.277
##  Max.   :28.030   Max.   :4.942            Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 2.350      1st Qu.:0.000000
##  Median :0.3100           Median : 3.580      Median :0.000000
##  Mean   :0.6189           Mean   : 3.643      Mean   :0.001752
##  3rd Qu.:0.8675           3rd Qu.: 4.897      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:147.0        1st Qu.: 721.2
##  Median :  7.00    Median :  8.00      Median :208.5        Median :1020.5
##  Mean   : 23.04    Mean   : 14.79      Mean   :210.3        Mean   : 955.2
##  3rd Qu.: 35.00    3rd Qu.: 21.00      3rd Qu.:272.0        3rd Qu.:1189.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
##     Calories
##  Min.   :  52
##  1st Qu.:1857
##  Median :2220
##  Mean   :2362
##  3rd Qu.:2832
##  Max.   :4900
```

```
n_distinct(SleepDay)
```

```
## [1] 410
```

```
glimpse(SleepDay)
```

```
## Rows: 410
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~
## $ SleepDay           <chr> "04/12/2016 00:00:00", "4/13/2016 12:00:00 AM", "4/~
## $ TotalSleepRecords  <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~
## $ TotalTimeInBed     <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~
```

```
nrow(SleepDay)
```

```
## [1] 410
```

```
nrow(Activity)
```

```
## [1] 862
```

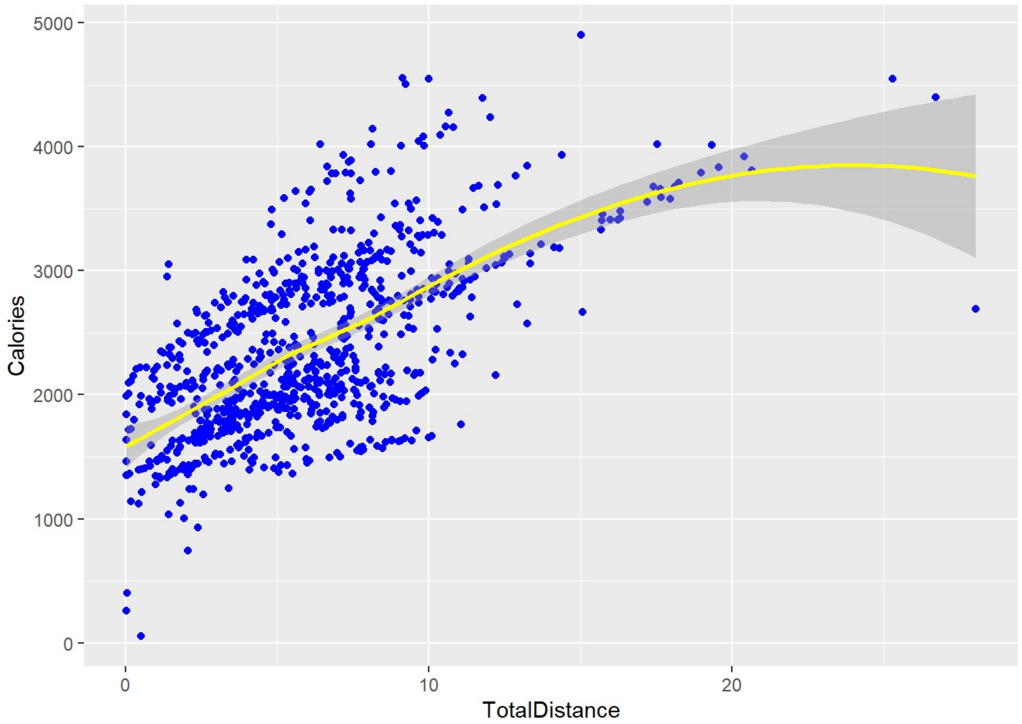
So, As we see there are 410 distinct Records in SleepDay data frame and 862 distinct records in Activity data frame.This observation is very essential to spot any duplicated record from skewing our analysis results.

## Categorizing Participates

Categorizing participator’s TotalMinutesAsleep into a data frame named status which shows the sleep time which indicate their Sleep Status

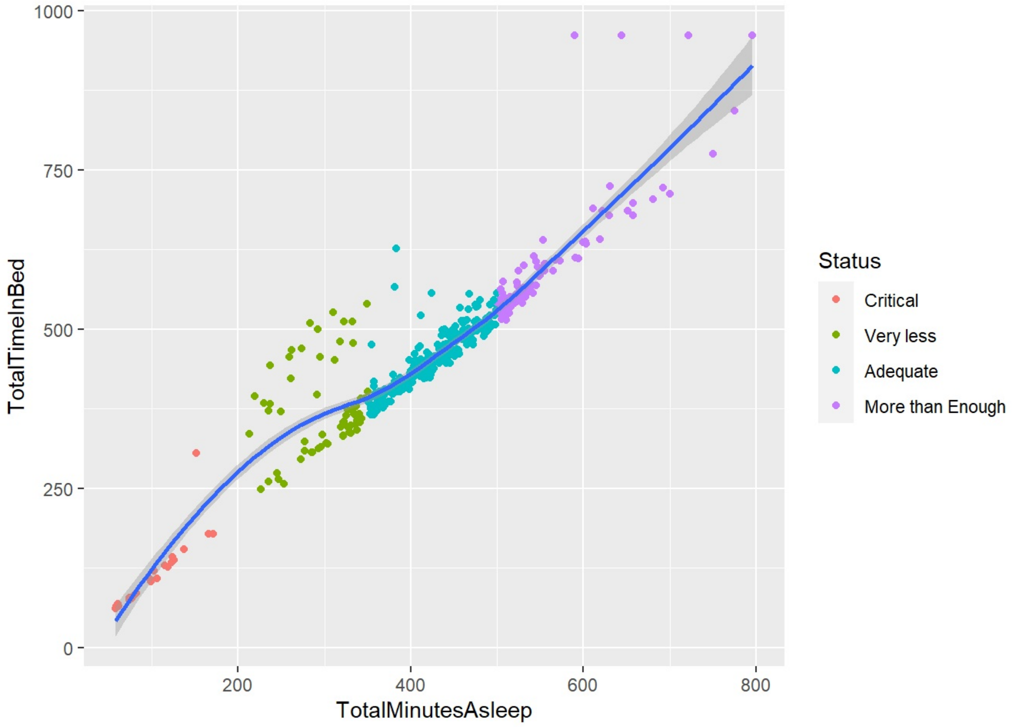
```
attach(SleepDay)

Status <- cut(TotalMinutesAsleep,breaks=c(0,200,350,500,800),labels = c("Critical","Very less","Adequate","More t
han Enough"))
ggplot(data = Activity,mapping = aes(x=TotalDistance,y=Calories))+geom_point(colour = "Blue")+geom_smooth(colour
= " Yellow ")
```



There is direct Corrolation between Total Distance a preson walks to calories he will burn during his Session

```
ggplot(data = SleepDay)+geom_point(mapping = aes(x=TotalMinutesAsleep,y = TotalTimeInBed,colour = Status))+geom_s
mooth(mapping = aes(x=TotalMinutesAsleep,y = TotalTimeInBed))
```



It is evident that people who are more in bed tend to be sleeping less proportionally and they get distracted from sleep doing other activities

## Reverting the Search-FLow

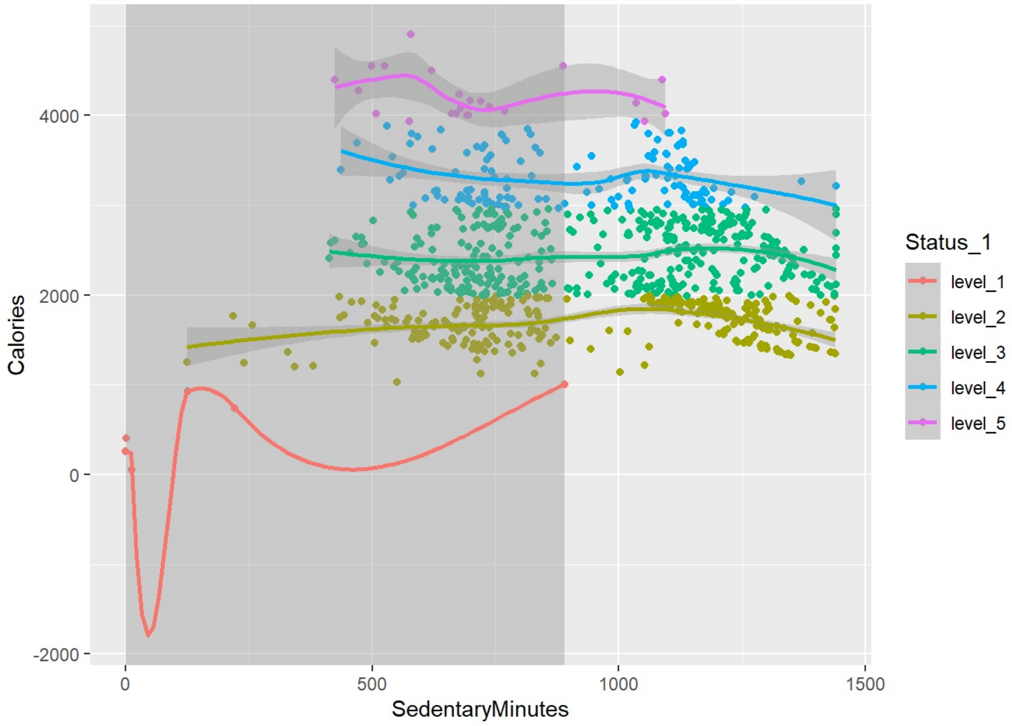
Detaching the Data frame as the search path by default was changes when we used attach() function but that change should not be permitted to persist,thus for that we use detach(“”)

```
detach("SleepDay")
SleepDay <- mutate(SleepDay,Hours = TotalMinutesAsleep / 60)
```

# Analysizing if there is relation between calories burned ,Total steps and sendentary minutes

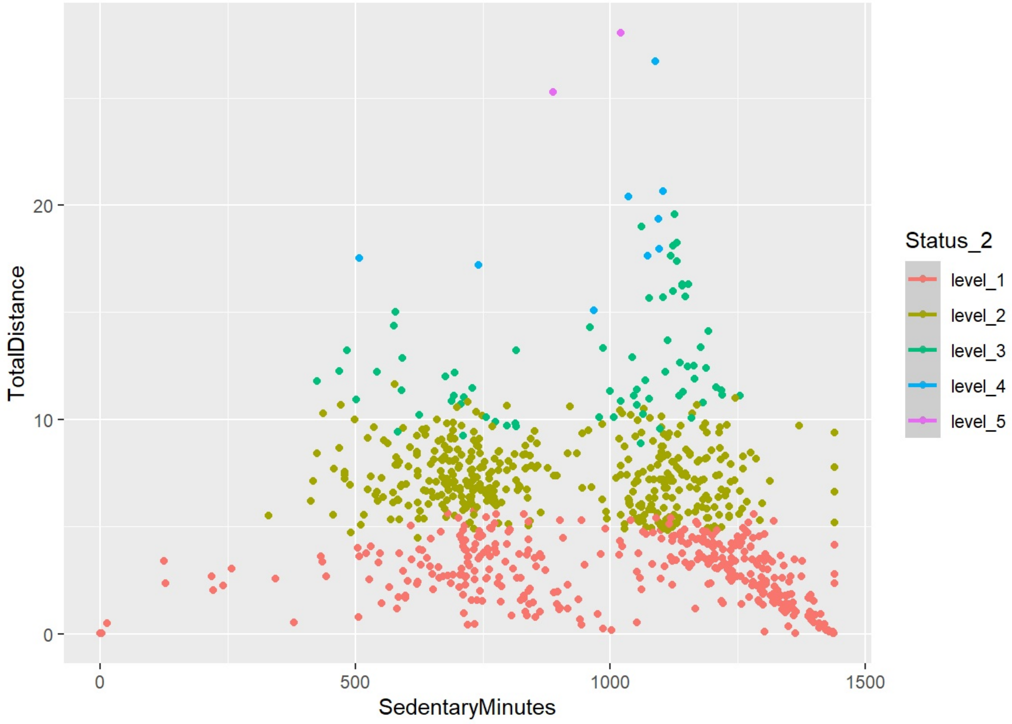
```
attach(Activity)
attach(SleepDay)

Status_1 <- cut(Calories,breaks=5,labels = c("level_1","level_2","level_3","level_4","level_5"))
ggplot(data = Activity,mapping = aes(x=SedentaryMinutes,y=Calories,colour = Status_1))+geom_point()+geom_smooth()
## there is a relationship
```



This visualization shows four different types of people, based on their body type, which can be described in this situation as burning calories respect to sedentary lifestyle

```
Status_2 <- cut(TotalSteps,breaks=5,labels = c("level_1","level_2","level_3","level_4","level_5"))
ggplot(data = Activity,mapping = aes(x=SedentaryMinutes,y=TotalDistance,colour = Status_2))+geom_point() + geom_s
mooth()##There is no relationship to explore here
```



We can gather from above scatter plot that people are more likely to be sedentary for more than 500 minutes per day and also that there is no direct relation between sedentary time to Total distance they cover daily.for majority of people everyday the maximum distance was up-to 10 kilometers and and levels suggest different group of people divided based on their Sedentary Minutes and Total Distance they covered.

# Combining SleepDay and dailyActivity files

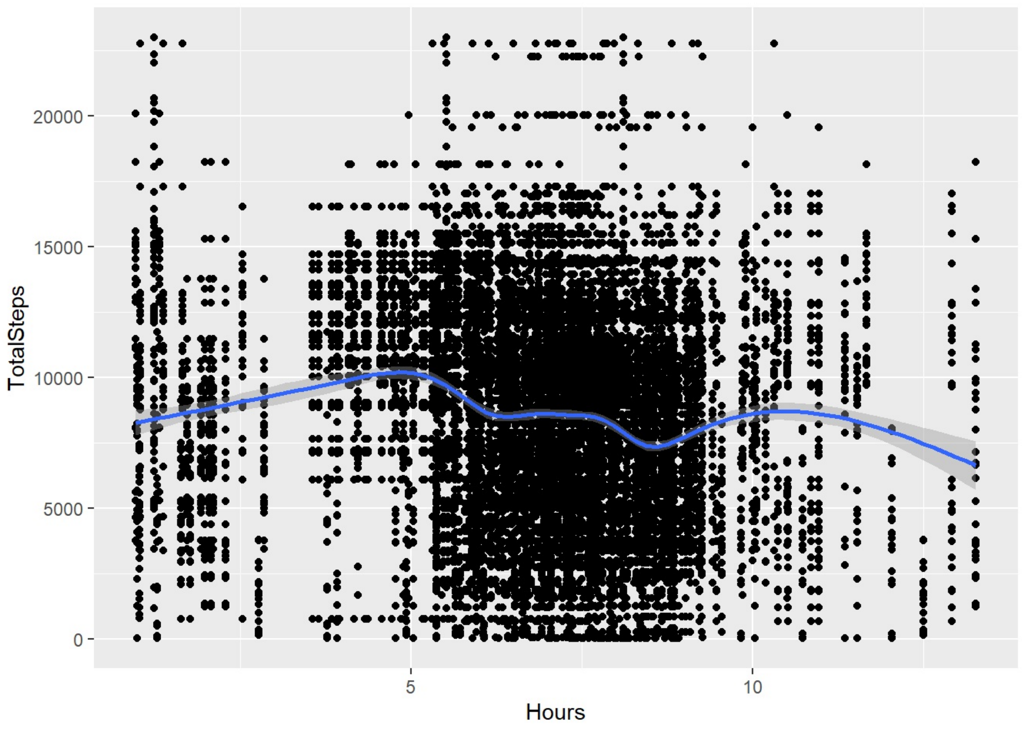
```
combined_data <- merge(SleepDay, Activity, by="Id")
n_distinct(combined_data$Id)
```

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

There was use of inner join in the merge function by default so there are fewer Ids left than it was in Activity data frame.This has happened because both csv files didn't had equal number of participate IDs which should be resolved as there can be many NULL values in merged data frame if the records of Ids from Activity data frame that are not present in SleepDay data frame, permitted to exist

# Analyzing if there is relationship between sleep time and total steps.

```
ggplot(data = combined_data,mapping = aes(x=Hours,y=TotalSteps))+geom_point() + geom_smooth() ##There is relation that there are logs suggesting that at times when sleep time was between 5 to 9 there was goal task achievement of 10000 steps.
```



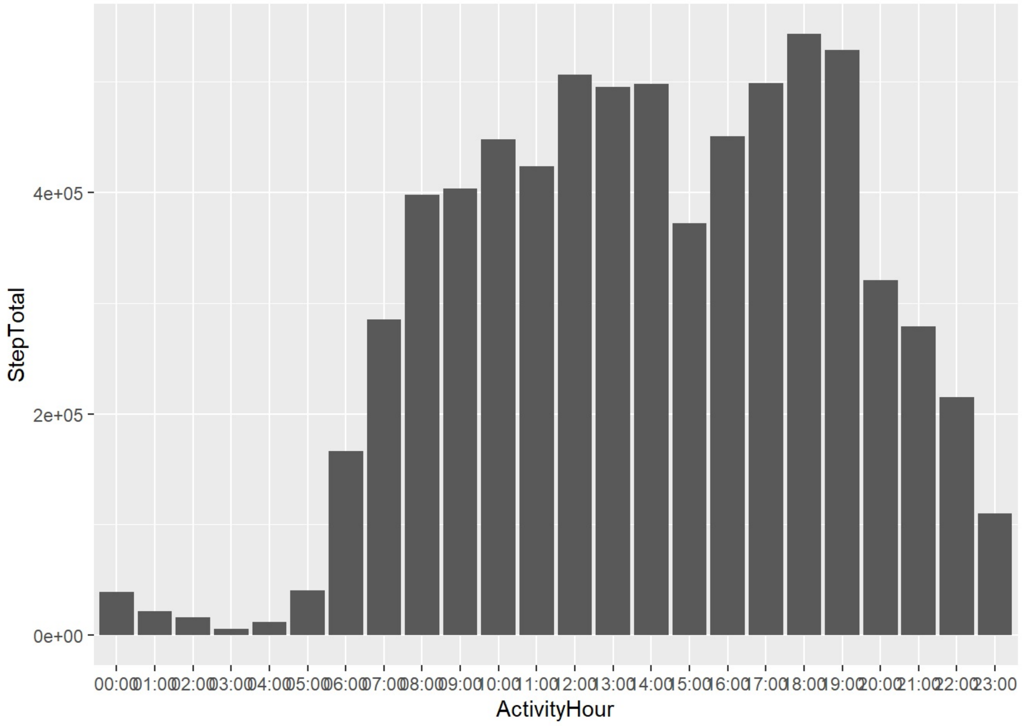
```
detach(Activity)
detach(SleepDay)
```

We can conclude that large chunk of people who sleeps 5 to 10 hours a day are getting To the threshold of 10000 steps recommended by many Health Experts.Also, as sleep time inceases people are getting more and more in Active in their Activities which can be seen by Trend line suggesting relative downward trend from 6 hour mark.But as sample size is too small we can't generalize this finding but this is to be true only for these 24 participates.

# Combining Hourly Data of steps intensities and calories

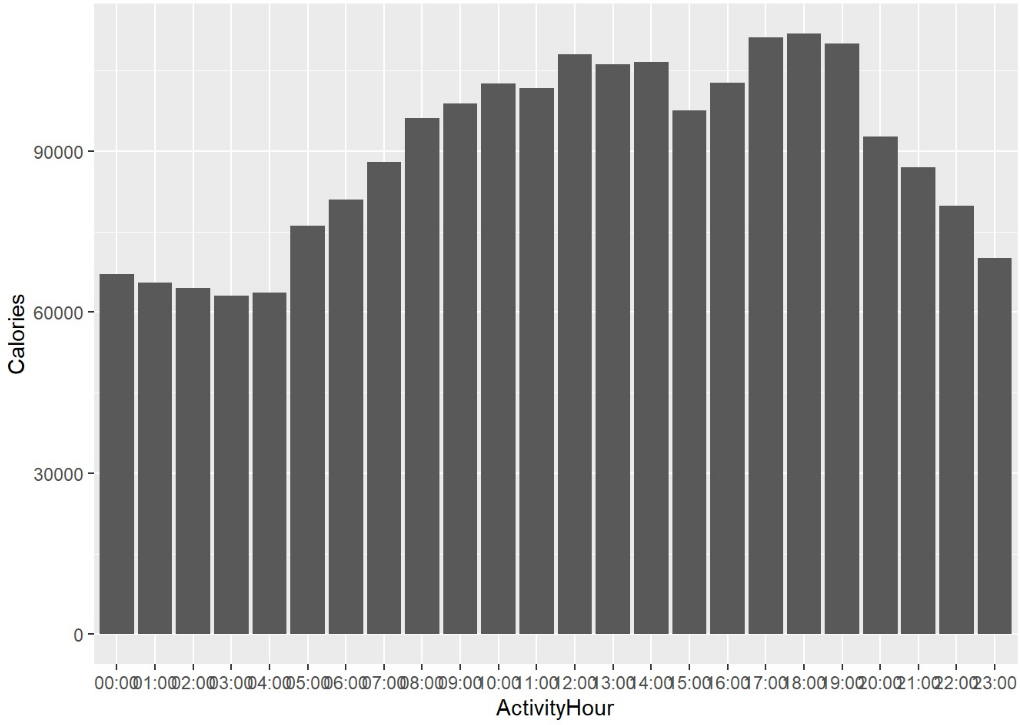
```
hourlySteps <- read.csv('C:/Users/LENOVO/Desktop/Case Study_How a wellness company play it smart(Bellabeat)/Fitab
ase Data 4.12.16-5.12.16/Selected_For_case_study_V.1/hourlySteps_v.2.csv')
hourlyCalories <- read.csv('C:/Users/LENOVO/Desktop/Case Study_How a wellness company play it smart(Bellabeat)/Fi
tabase Data 4.12.16-5.12.16/Selected_For_case_study_V.1/hourlyCalories_v.2.csv')
hourlyIntensities <- read.csv('C:/Users/LENOVO/Desktop/Case Study_How a wellness company play it smart(Bellabeat)
/Fitabase Data 4.12.16-5.12.16/Selected_For_case_study_V.1/hourlyIntensities_v.2.csv')
```

```
ggplot(data = hourlySteps) + geom_col(mapping = aes(y=StepTotal,x=ActivityHour))
```



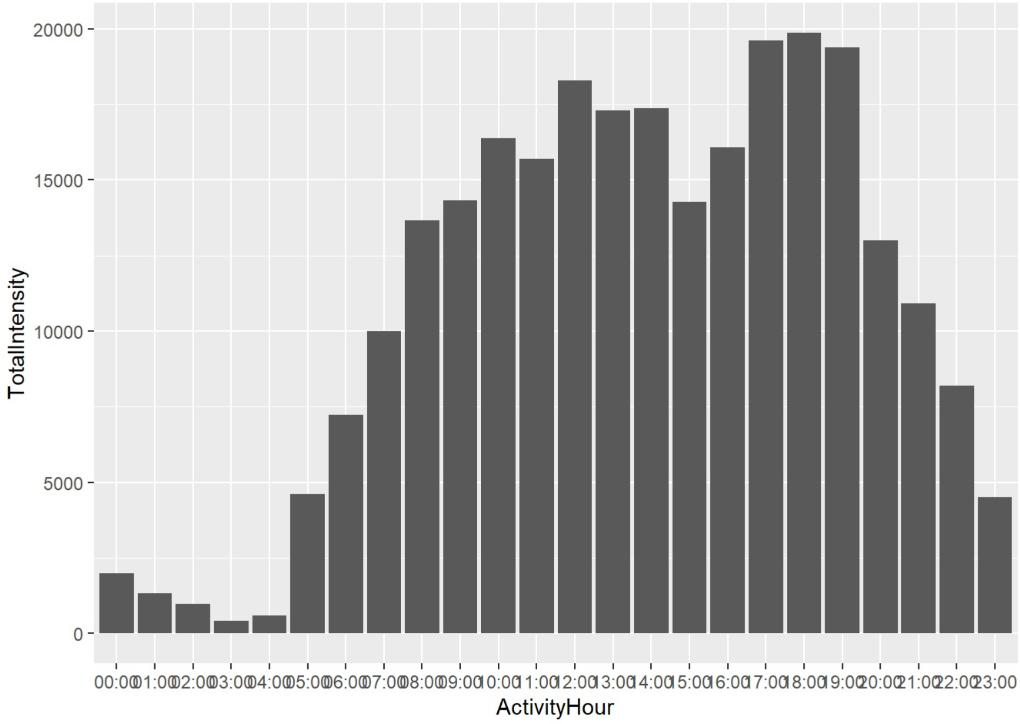
There is relation that this group of participants walked more steps during period of 8:00 to 19:00 and After 19:00 there is steep fall in walked steps which suggest this group is in rest stage of the day.

```
ggplot(data = hourlyCalories) + geom_col(mapping = aes(y=Calories,x=ActivityHour))
```



highest calorific burn can be seen at 18:00 which suggests that max steps walked in evening walks

```
ggplot(data = hourlyIntensities) + geom_col(mapping = aes(y=TotalIntensity,x=ActivityHour))
```



this shows that people were able to apply more intensity from 5 to 7 in the evening

These are some other relations that i liked to show, which is for Activity Log According to Time of the day.

# Recommendations for the Stakeholders

***1.Based on the activity levels and amount of calories burned, users appear to burn more calories with more exercise which in turn will help them become leaner and lead a healthy life. so For that, Bellabeat can motivate users to exercise more through reminders and tips or insights to staying motivated. They could also offer app dashboard for Activity***

***they have logged and a leader-board in which there will be option for users to compete with their connections which will be fueled by ranking system.Also, company can give points according to steps the user have walked which can then used by user to get discounted items from the company’s website.***

***2.The data suggests majority of user live either a light or sedentary lifestyle, which may be due to lack of motivation Thus there can be a feature of push notifications which daily sends motivational quotes to the user.As there is lacking motivation user should be positively reinforced when they complete certain threshold of activity that will encourage user psychologically.***

***3.To promote better sleep habits, Bellabeat may combine reminders with an app that informs users the best time to go to bed and wake up so they can feel refreshed in the morning and get enough sleep which will lead to better performance throughout the day and they will not feel burned out when trying to complete their Daily step count.***

***4.There is dire need of larger sample size in order to improve Credibility of the analysis.***

***5.There is need to acquire current data in order to better reflect current consumer behavior or trends in smart device usage.Collect data from internal sources to increase credibility and reliability of the data sets which will produce accurate analysis regularly and also give insights into the changing behavior patterns of the customers and that will answer the question of In what ways that particular group of people can be attracted towards our product/service by incorporating Cluster analysis.***