Google Data Analytics Capstone- Bellabet Case Study

Menbere Hailu

2022-08-04

1. Introduction of business task

Bellabeat is a high-tech design and manufacturing company that produces health-focused smart devices for women. The goal of this report is to provide an analysis of how consumers are using their smart devices. The data is collected from a set of 30 Fitbit Fitness users over 31 days here.

Bellabet is looking to expanding their business and would like to receive data-driven recommendations to improve marketing strategy. In other to perform this task, a comprehensive analysis on Smart device usage data is required.

Question for Analysis

I was tasked to analyze the smart device usage data in order to gain insights on how customers use their smart devices. These questions would serve as a guide during our 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

- Urška Sršen - Bellabeat cofounder and Chief Creative Officer
- Sando Mur - Bellabeat cofounder and key member of Bellabeat executive team
- Bellabeat Marketing Analytics team

2. Prepare

Data source

For doing this analysis the data source is from kaggle, the archived dataset contains 18 different excel files for thier perspective information among thos data 3 of them are organized in wide format the rest are organized long format

Data credibility and integrity Due to the limitation of size (30 users) and not having any demographic information we could encounter a sampling bias. We are not sure if the sample is representative of the population as a whole. Another problem we would encounter is that the dataset is not current and also the time limitation of the survey (2 months long). That is why we will give our case study an operational approach.

3. Processs

For doing this analysis I have choose R(programming language)

import packages used for the analysis

```r
# Install packages
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
install.packages("lubridate")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
install.packages("reshape2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
install.packages("scales")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
# load installed library
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()
```

```r
library(lubridate)
```

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

```r
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
```

```
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Loading dataset

```
day_activity<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/dailyActivity_merge
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...
##
## 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.
```

```
hour_Intensities<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/hourlyIntensit
```

```
## Rows: 22099 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (3): Id, TotalIntensity, AverageIntensity
##
## 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.
```

```
hour_calories<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/hourlyCalories_me
```

```
## Rows: 22099 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, Calories
##
## 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.
```

```
hour_step<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/hourlySteps_merged.csv
```

```
## Rows: 22099 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, StepTotal
##
## 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.
```

```
day_sleep<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SleepDay
```

```
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## 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.
```

```
weight_info<-read_csv("/cloud/project/Google_Capstone/Fitabase Data 4.12.16-5.12.16/weightLogInfo_merge
```

```
## Rows: 67 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
##
## 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.
```

lets look one of data set

```
head(day_activity)
```

```
## # A tibble: 6 x 15
##       Id Activ~1 Total~2 Total~3 Track~4 Logge~5 VeryA~6 Moder~7 Light~8 Seden~9
##    <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1.50e9 4/12/2~   13162    8.5     8.5        0    1.88   0.550    6.06       0
## 2 1.50e9 4/13/2~   10735    6.97    6.97       0    1.57   0.690    4.71       0
## 3 1.50e9 4/14/2~   10460    6.74    6.74       0    2.44   0.400    3.91       0
## 4 1.50e9 4/15/2~    9762    6.28    6.28       0    2.14   1.26     2.83       0
## 5 1.50e9 4/16/2~   12669    8.16    8.16       0    2.71   0.410    5.04       0
## 6 1.50e9 4/17/2~    9705    6.48    6.48       0    3.19   0.780    2.51       0
## # ... with 5 more variables: VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>, and abbreviated variable names
## #   1: ActivityDate, 2: TotalSteps, 3: TotalDistance, 4: TrackerDistance,
## #   5: LoggedActivitiesDistance, 6: VeryActiveDistance,
## #   7: ModeratelyActiveDistance, 8: LightActiveDistance,
## #   9: SedentaryActiveDistance
## # i Use `colnames()` to see all variable names
```

Merge Among the imported data most of Hourly_intensity,Hourly_ calories and hourly_steps have common column so it is better to merge those data in to ne data set remove those document for file managment

```
hour_daily<- merge(x=hour_Intensities,y=hour_calories,by=c("Id","ActivityHour"))
hour_activity<-merge(x=hour_daily,y=hour_step, by=c("Id","ActivityHour"))
```

```
### remove unwanted dataset ######
rm(hour_Intensities,hour_calories,hour_step)
```

I spotted some problems with the timestamp data. So before analysis, I need to convert it to **date time** format and split to date and time.

```
day_activity$ActivityDate<-as.POSIXct(day_activity$ActivityDate, format= "%m/%d/%Y")
day_activity$Date<-format(day_activity$ActivityDate, format = "%m/%d/%y")
hour_activity$ActivityHour<-as.POSIXct(hour_activity$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys
hour_activity$ActivityTime<-format(hour_activity$ActivityHour, format="%H:%M:%S")
```

```
hour_activity$ActivityDate<-format(hour_activity$ActivityHour, format= "%m/%d/%y")

day_sleep$SleepDay<-as.POSIXct(day_sleep$SleepDay, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
day_sleep$Date<-format(day_sleep$SleepDay, format = "%m/%d/%y")
```

```
#for further analysis we need to merge
sleep_activity<- merge(x=day_activity,y=day_sleep, by= c("Id","Date"))
```

Fixing formatting

Assigning Days of the week to the datasets I would be assigning a Days column to help with my analysis. I do believe it would provide a great insight as it would be helpful to analyse activities by each day of the week to help see patterns or trends.

```
day_activity<-transform(day_activity,Day_of_week=weekdays(ActivityDate))
hour_activity<-transform(hour_activity,Day_of_week=weekdays(ActivityHour))
day_sleep<-transform(day_sleep,Day_of_week=weekdays(SleepDay))
```

4. Analyse

```
n_distinct(day_activity$Id)
```

Explore and summarize data

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

```
n_distinct(hour_activity$Id)
```

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

```
n_distinct(day_sleep$Id)
```

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

```
n_distinct(weight_info$Id)
```

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

This information tells us about number participants in each data sets.

There is 33 participants in the day_activity and hourly_activity , 24 in the day_sleep and only 8 in the weight data set. 8 participants is not significant to make any recommendations and conclusions based on this data.

Average of Multiple Activities

```
# day_activity
Basic_day_activity_summary<-day_activity %>%
  summarise_at(c( Average_step="TotalSteps",
                      Average_distance="TotalDistance",
                  Average_Sed_min="SedentaryMinutes", Average_calories="Calories"),mean)
Basic_day_activity_summary
```

```
##   Average_step Average_distance Average_Sed_min Average_calories
## 1     7637.911         5.489702        991.2106          2303.61
```

Average per week

```
Basic_Day_of_week_activity_summary<-day_activity %>%
  group_by(Day_of_week)%>%
  summarise_at(c( Average_step="TotalSteps",
                  Average_distance="TotalDistance",
                  Average_Sed_min="SedentaryMinutes", Average_calories="Calories"),mean)

  Basic_Day_of_week_activity_summary
```

```
## # A tibble: 7 x 5
##   Day_of_week Average_step Average_distance Average_Sed_min Average_calories
##   <chr>              <dbl>            <dbl>           <dbl>            <dbl>
## 1 Friday             7448.             5.31           1000.            2332.
## 2 Monday             7781.             5.55           1028.            2324.
## 3 Saturday           8153.             5.85            964.            2355.
## 4 Sunday             6933.             5.03            990.            2263
## 5 Thursday           7406.             5.31            962.            2200.
## 6 Tuesday            8125.             5.83           1007.            2356.
## 7 Wednesday          7559.             5.49            989.            2303.
```

- Users walk more on Saturdays 8,152 steps.This could be due to the fact that most users are workers or students and have Saturdays for exercises.
- On Tuesdays, users take about 8,125 steps and expend more calories than on Saturdays! More data would be required to produce better insights.
- From the results, the average steps for all users is 7,637 steps which is below the recommended steps of 10,000 according to CDC(Centers for Disease Control and Prevention).
- Also, despite Sunday having the lowest average steps taken in a day with 6,933 steps the amount of calories expended is higher than Thursday with 7,405 steps. This shows us that users should not only increase their steps but their very active minutes daily if they plan on increasing the amount of calories burnt per day.

Average sleep minutes and time in bed per week

I would like to find out users sleep patterns for each day of the week

```
Average_sleep_week<-day_sleep %>%
    group_by(Day_of_week)%>%
    summarise_at(c(Average_TotalMinutesAsleep="TotalMinutesAsleep",Average_TotalTimeInBed="TotalTimeInB
  Average_sleep_week
```

```
## # A tibble: 7 x 3
##   Day_of_week Average_TotalMinutesAsleep Average_TotalTimeInBed
##   <chr>                            <dbl>                  <dbl>
## 1 Friday                            405.                   445.
## 2 Monday                            419.                   456.
## 3 Saturday                          421.                   461.
## 4 Sunday                            453.                   504.
## 5 Thursday                          402.                   436.
## 6 Tuesday                           405.                   443.
## 7 Wednesday                         435.                   470.
```

```
 Average_sleep<-day_sleep %>%
    summarise_at(c(Average_TotalMinutesAsleep="TotalMinutesAsleep",Average_TotalTimeInBed="TotalTimeInB
  Average_sleep
```

```
##   Average_TotalMinutesAsleep Average_TotalTimeInBed
```

```
## 1                   419.4673                458.6392
```

- Users tend to sleep more on weekends particularly on Sunday(7.5 hours). This could be as a result of users working on weekdays and having early/late shifts.
- Users have an average sleep time of about 7 hours and according to the CDC, adults of ages 18-60 years are required to sleep for 7 or more hours per night.

Correlation of Day activities

```
  # correlation
  correlated_variables<-select(day_activity, Calories, TotalSteps:SedentaryMinutes, -Day_of_week)
  correlation<-cor(correlated_variables)
   correlation
```

```
##                             Calories  TotalSteps TotalDistance TrackerDistance
## Calories                  1.00000000  0.59156809    0.64496187      0.64531330
## TotalSteps                0.59156809  1.00000000    0.98536884      0.98482223
## TotalDistance             0.64496187  0.98536884    1.00000000      0.99950473
## TrackerDistance           0.64531330  0.98482223    0.99950473      1.00000000
## LoggedActivitiesDistance  0.20759511  0.18184869    0.18833178      0.16258530
## VeryActiveDistance        0.49195856  0.74011458    0.79458162      0.79433807
## ModeratelyActiveDistance  0.21678987  0.50710545    0.47075827      0.47027739
## LightActiveDistance       0.46691676  0.69220820    0.66200154      0.66136481
## SedentaryActiveDistance   0.04365187  0.07050474    0.08238905      0.07459089
## VeryActiveMinutes         0.61583827  0.66707870    0.68129743      0.68081599
## FairlyActiveMinutes       0.29762347  0.49869337    0.46289889      0.46315415
## LightlyActiveMinutes      0.28671753  0.56960021    0.51630049      0.51471308
## SedentaryMinutes         -0.10697305 -0.32748355   -0.28809436     -0.28934322
##                           LoggedActivitiesDistance VeryActiveDistance
## Calories                                0.20759511         0.49195856
## TotalSteps                              0.18184869         0.74011458
## TotalDistance                           0.18833178         0.79458162
## TrackerDistance                         0.16258530         0.79433807
## LoggedActivitiesDistance                1.00000000         0.15085226
## VeryActiveDistance                      0.15085226         1.00000000
## ModeratelyActiveDistance                0.07652693         0.19298587
## LightActiveDistance                     0.13830151         0.15766926
## SedentaryActiveDistance                 0.15499618         0.04611675
## VeryActiveMinutes                       0.23444286         0.82668146
## FairlyActiveMinutes                     0.05385996         0.21173011
## LightlyActiveMinutes                    0.10213494         0.05984538
## SedentaryMinutes                       -0.04699945        -0.06175419
##                           ModeratelyActiveDistance LightActiveDistance
## Calories                                0.216789872           0.4669168
## TotalSteps                              0.507105449           0.6922082
## TotalDistance                           0.470758273           0.6620015
## TrackerDistance                         0.470277391           0.6613648
## LoggedActivitiesDistance                0.076526932           0.1383015
## VeryActiveDistance                      0.192985874           0.1576693
## ModeratelyActiveDistance                1.000000000           0.2378474
## LightActiveDistance                     0.237847447           1.0000000
## SedentaryActiveDistance                 0.005793403           0.0995032
## VeryActiveMinutes                       0.225464009           0.1549665
## FairlyActiveMinutes                     0.946934035           0.2201291
## LightlyActiveMinutes                    0.162091885           0.8856971
```

```
## SedentaryMinutes                    -0.221436057          -0.4135517
##                           SedentaryActiveDistance VeryActiveMinutes
## Calories                              0.043651875       0.615838268
## TotalSteps                            0.070504742       0.667078697
## TotalDistance                         0.082389046       0.681297434
## TrackerDistance                       0.074590885       0.680815987
## LoggedActivitiesDistance              0.154996178       0.234442856
## VeryActiveDistance                    0.046116748       0.826681461
## ModeratelyActiveDistance              0.005793403       0.225464009
## LightActiveDistance                   0.099503204       0.154966479
## SedentaryActiveDistance               1.000000000       0.008258149
## VeryActiveMinutes                     0.008258149       1.000000000
## FairlyActiveMinutes                  -0.022360869       0.312420353
## LightlyActiveMinutes                  0.124184729       0.051925909
## SedentaryMinutes                      0.035474606      -0.164670992
##                           FairlyActiveMinutes LightlyActiveMinutes
## Calories                          0.29762347           0.28671753
## TotalSteps                        0.49869337           0.56960021
## TotalDistance                     0.46289889           0.51630049
## TrackerDistance                   0.46315415           0.51471308
## LoggedActivitiesDistance          0.05385996           0.10213494
## VeryActiveDistance                0.21173011           0.05984538
## ModeratelyActiveDistance          0.94693404           0.16209189
## LightActiveDistance               0.22012907           0.88569707
## SedentaryActiveDistance          -0.02236087           0.12418473
## VeryActiveMinutes                 0.31242035           0.05192591
## FairlyActiveMinutes               1.00000000           0.14881991
## LightlyActiveMinutes              0.14881991           1.00000000
## SedentaryMinutes                 -0.23744642          -0.43710390
##                           SedentaryMinutes
## Calories                      -0.10697305
## TotalSteps                    -0.32748355
## TotalDistance                 -0.28809436
## TrackerDistance               -0.28934322
## LoggedActivitiesDistance      -0.04699945
## VeryActiveDistance            -0.06175419
## ModeratelyActiveDistance      -0.22143606
## LightActiveDistance           -0.41355171
## SedentaryActiveDistance        0.03547461
## VeryActiveMinutes             -0.16467099
## FairlyActiveMinutes           -0.23744642
## LightlyActiveMinutes          -0.43710390
## SedentaryMinutes               1.00000000
```

- From the correlation result, we can see multiple variables(such as Total Steps, Total Distance, Very Active Minutes, e.t.c) all have positive correlation with calories.

The sedentary minutes has no correlation negative correlation with calories expended.

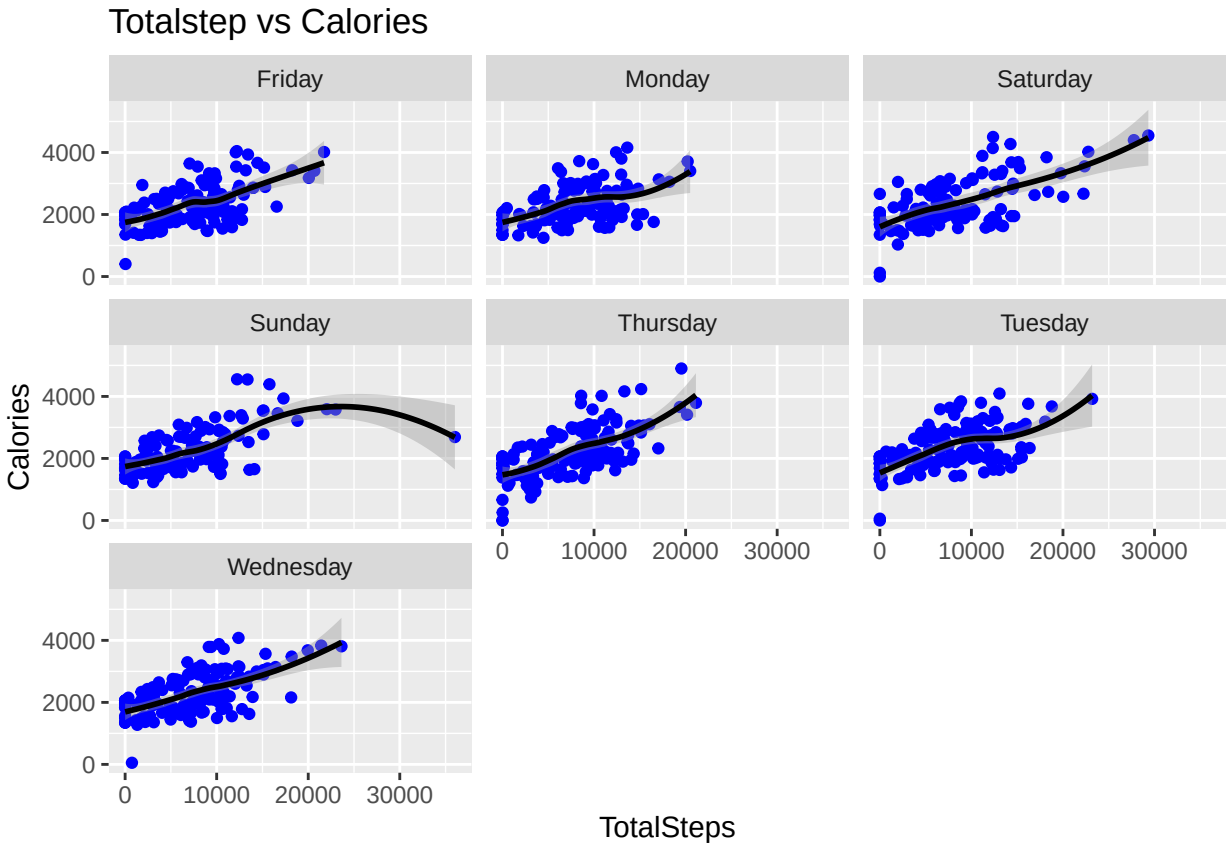
5.Visualization

Correlation between Total steps and Calories Each week

```
##Total step Vs Calories and each week
  ggplot(data=day_activity,aes(x=TotalSteps,y=Calories))+
    geom_point(color="blue")+geom_smooth(color= "black")+
```

```
    facet_wrap(~ Day_of_week)+
    labs(title="Totalstep vs Calories")
```

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



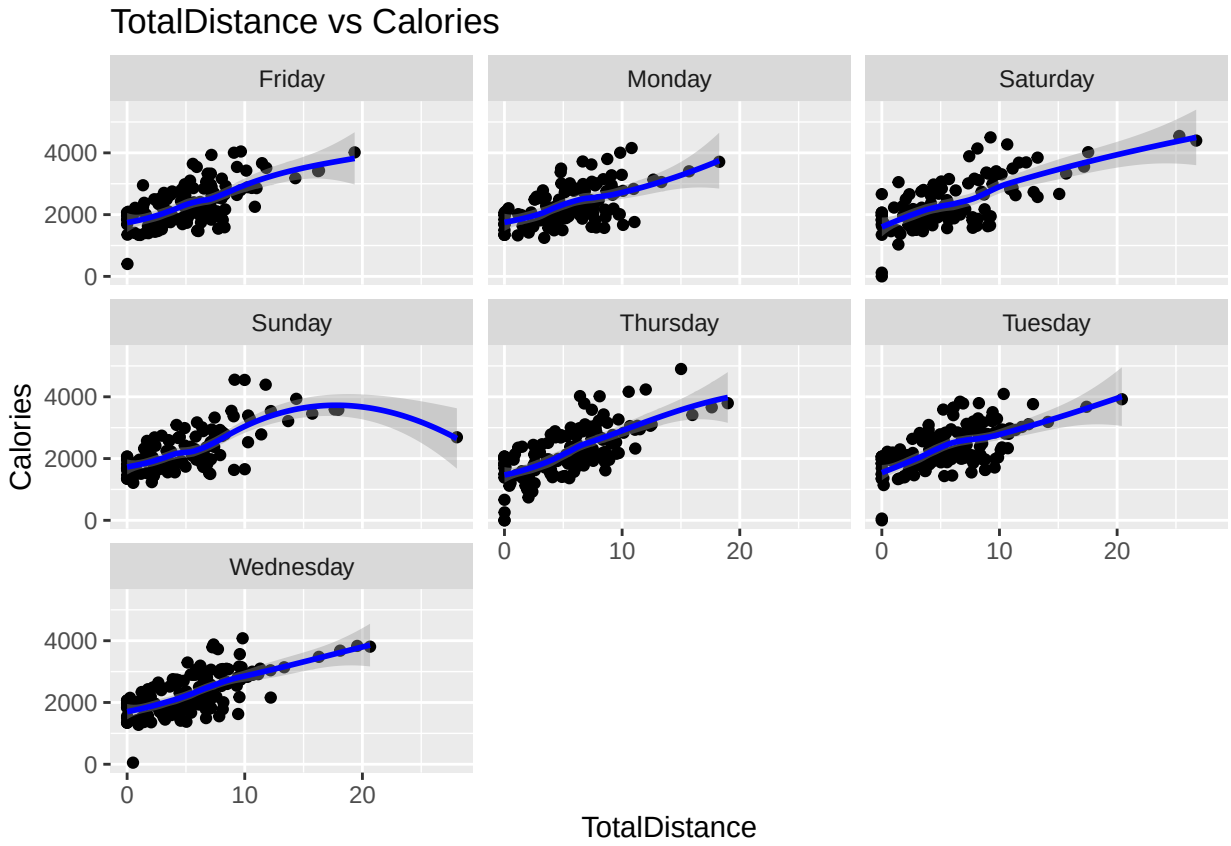
I see positive correlation here between Total Steps and Calories each week, which is obvious - the more active we are, the more calories we burn.

Total distance Vs Calories per week

```
###TotalDistance Vs calories

    ggplot(data=day_activity,aes(x=TotalDistance,y=Calories))+
      geom_point()+geom_smooth(color= "blue")+
      facet_wrap(~ Day_of_week)+
    labs(title="TotalDistance vs Calories")
```

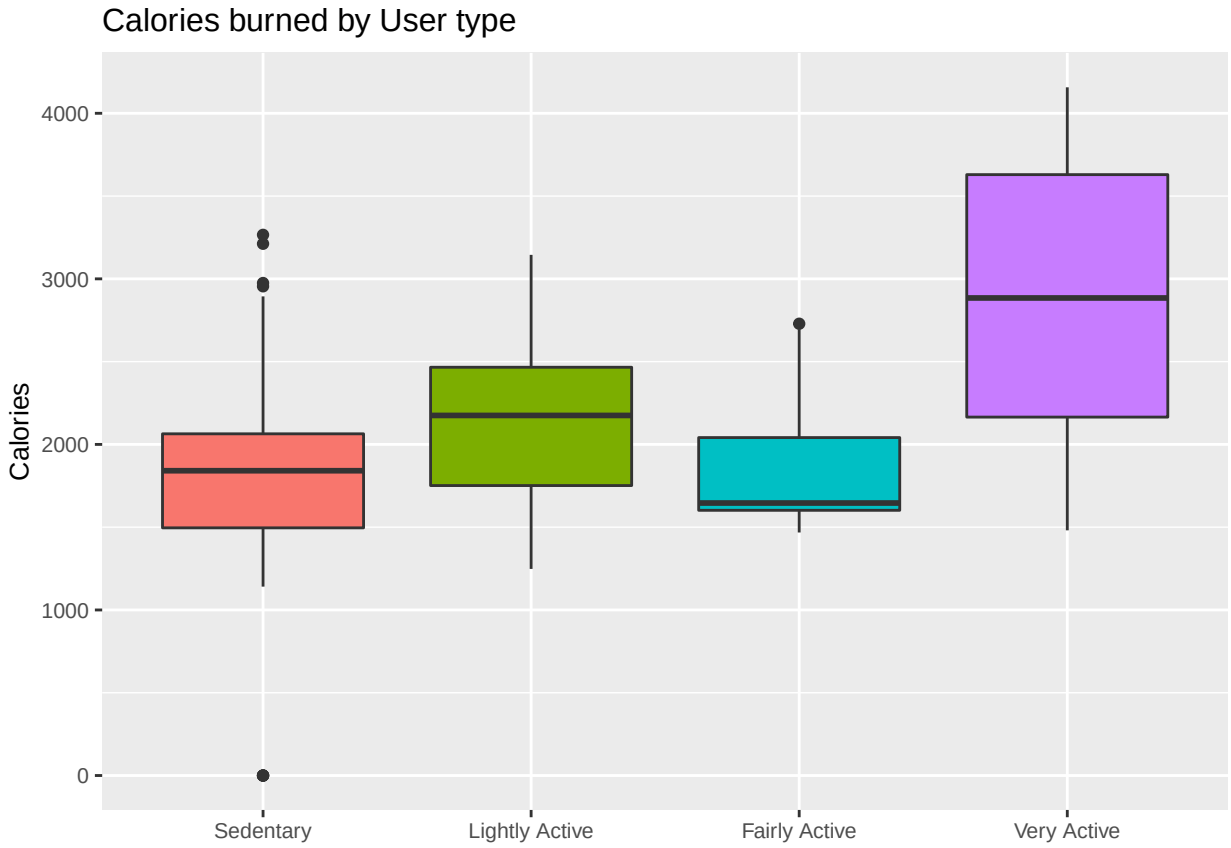
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



I see positive correlation here between Total Distance and Calories each week, which is obvious - the more active we are, the more calories we burn But it depends of week day activities.

User Type per Calories

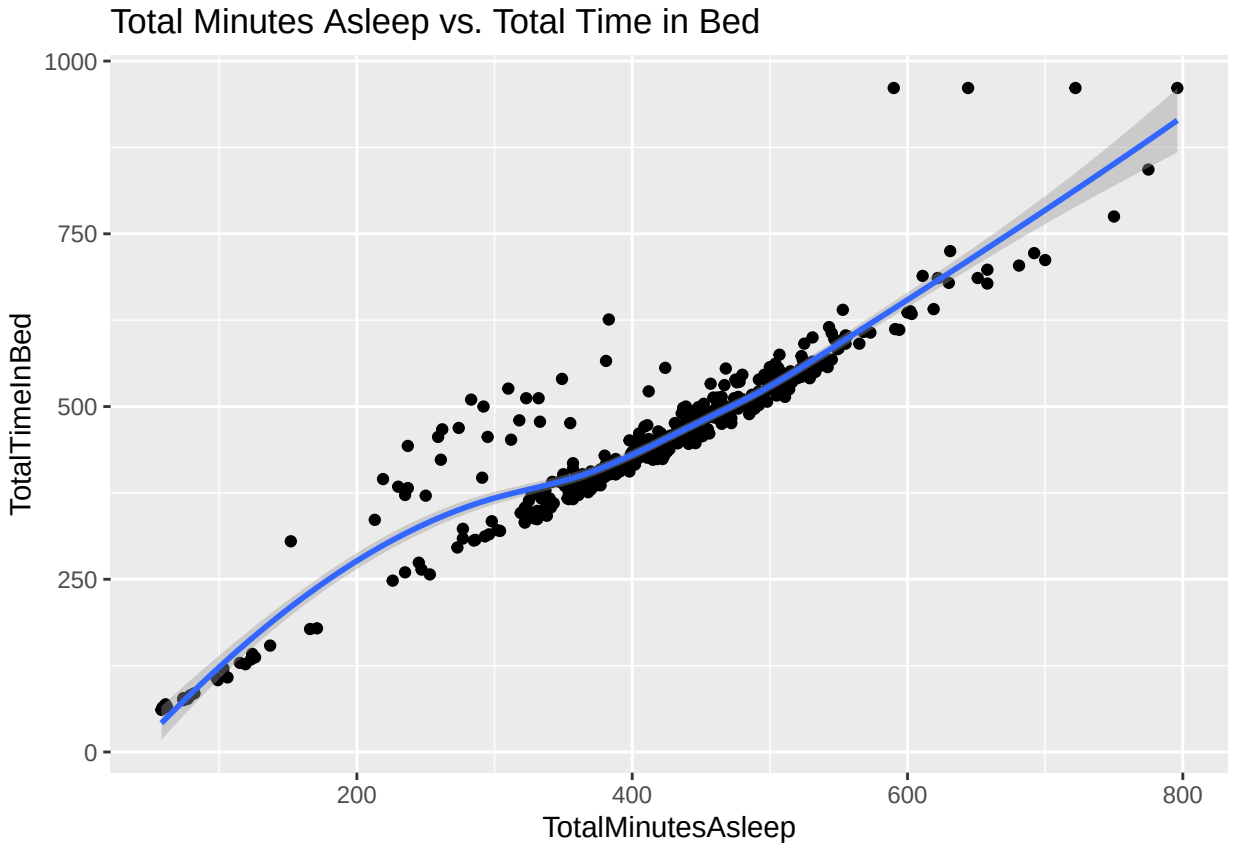
```
## usertype distribution vs calories
    data_by_usertype <- day_activity %>%
      summarise(
        user_type = factor(case_when(
          SedentaryMinutes > mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes)
          SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes > mean(LightlyActiveMinutes)
          SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes)
          SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes)
        ),levels=c("Sedentary", "Lightly Active", "Fairly Active", "Very Active")), Calories, .group=Id
      group_by(user_type)%>%
      drop_na()
    ggplot(data_by_usertype, aes(user_type, Calories, fill=user_type)) +
      geom_boxplot() +
      theme(legend.position="none") +
      labs(title="Calories burned by User type", x=NULL) +
      theme(legend.position="none", text = element_text(size = 10),plot.title = element_text(hjust = 0.
```



Very active users burn a lot of calories other than others **Total Minutes Asleep vs. Total Time in Bed**

```
    ggplot(data=day_sleep, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
      geom_point()+ geom_smooth()+
      labs(title="Total Minutes Asleep vs. Total Time in Bed")
```

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



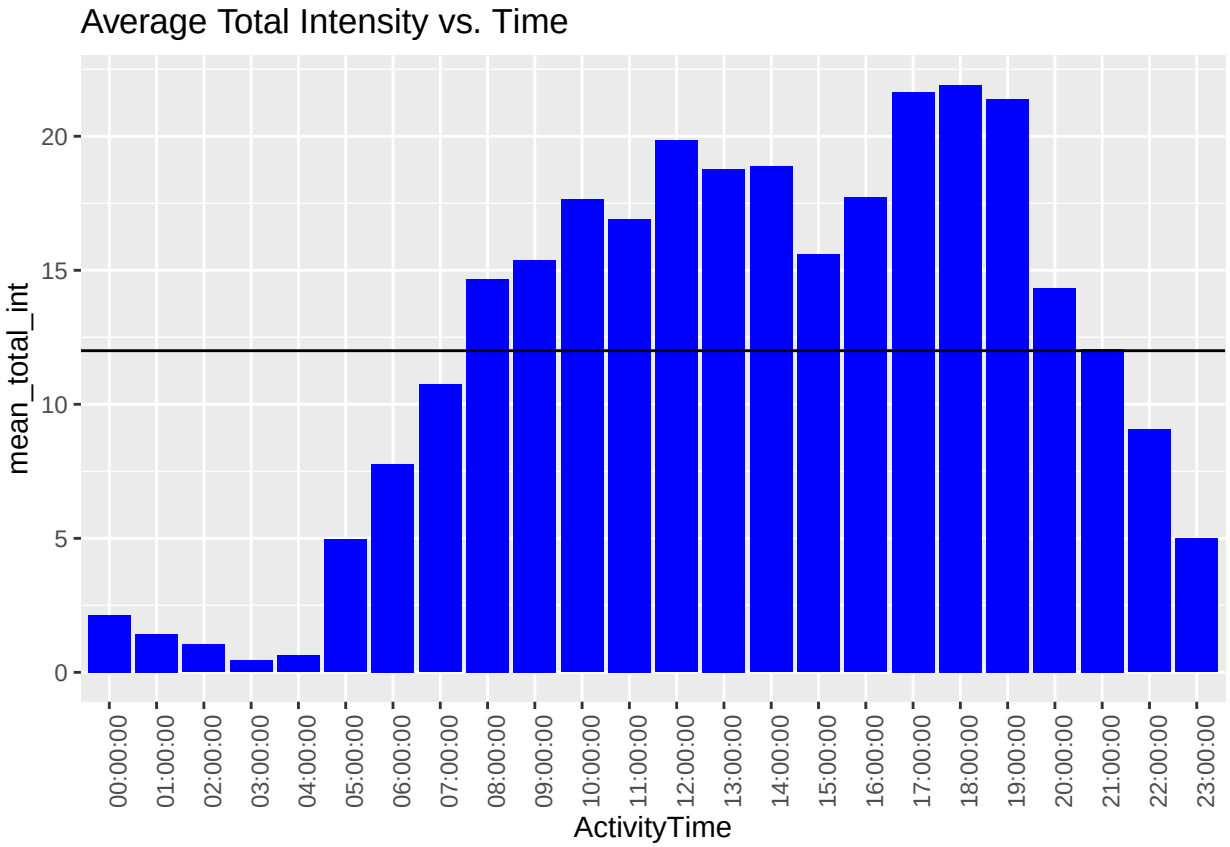
The relationship between Total Minutes Asleep and Total Time in Bed looks linear. So if the Bellabeat users want to improve their sleep, we should consider using notification to go to sleep.

Average Total Intensity vs. Time

```
#####Total average intensity per time
    intensity <- hour_activity %>%
      group_by(ActivityTime) %>%
      drop_na() %>%
      summarise(mean_total_int = mean(TotalIntensity))

    ggplot(data=intensity, aes(x=ActivityTime, y=mean_total_int)) + geom_histogram(stat = "identity", fi
      geom_hline(yintercept = 12)+
      theme(axis.text.x = element_text(angle = 90)) +
      labs(title="Average Total Intensity vs. Time")
```

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



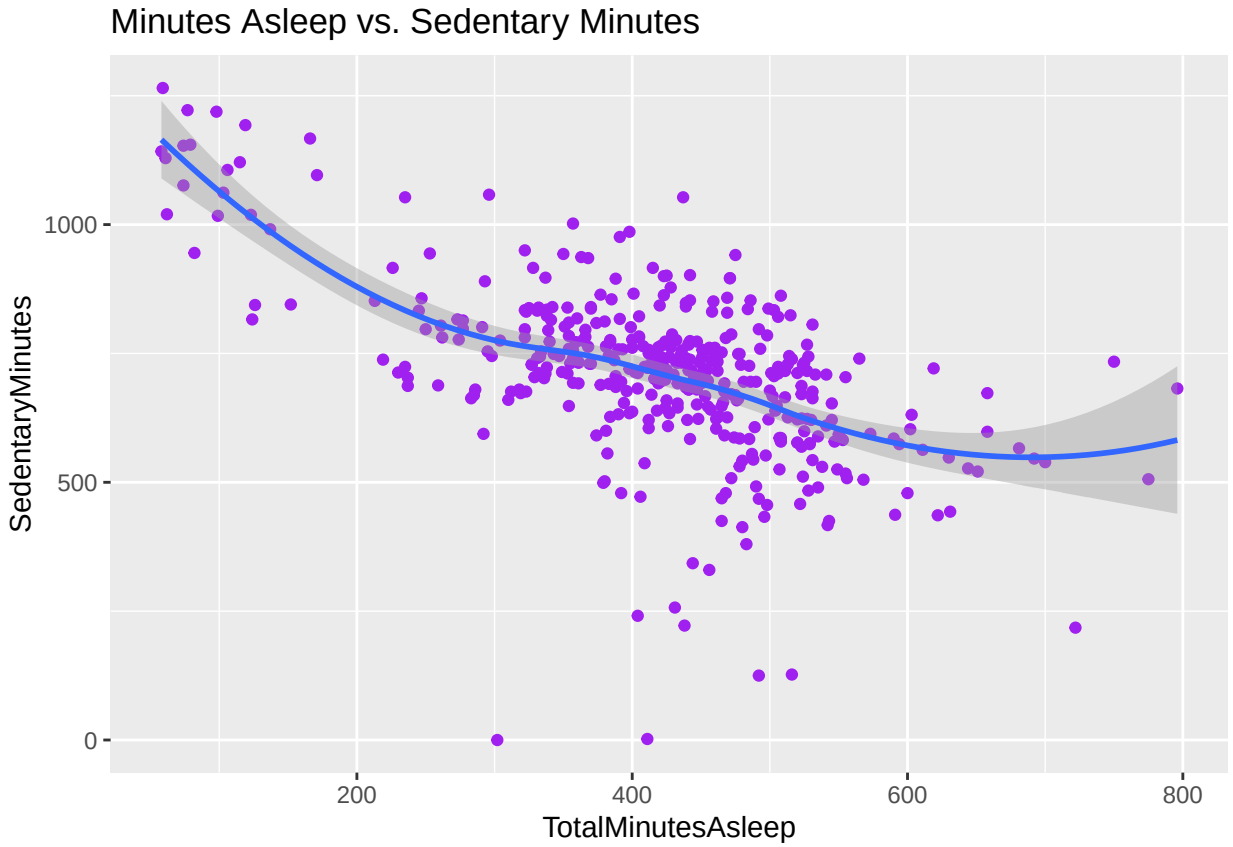
After visualizing Total Intensity hourly, I found out that people are more active between 5 am and 10pm.

Most activity happens between 5 pm and 7 pm - I suppose, that people go to a gym or for a walk after finishing work. **We can use this time in the Bellabeat app to remind and motivate users to go for a run or walk.**

Minute sleep and Sandary minutes

```
    ggplot(data=sleep_activity, aes(x=TotalMinutesAsleep, y=SedentaryMinutes)) +
      geom_point(color='purple') + geom_smooth() +
      labs(title="Minutes Asleep vs. Sedentary Minutes")
```

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



-Here we can clearly see the negative relationship between Sedentary Minutes and Sleep time.

-As an idea: if Bellabeat users want to improve their sleep, Bellabeat app can recommend reducing sedentary time.

-Keep in mind that we need to support this insights with more data, because correlation between some data doesn't mean

6.Recommendation for Bellabate Campign

- Average steps per day are 4738, which is quite lower than the healthy count of 8000 - 10000 steps given by CDC, thus the app can motivate users to achieve the daily target of 10000 steps.
- The app can include a weight loss program where users are made aware of their calorie burn and active time.
- The data shows users with high sedentary time have lower sleep time which affects quality sleep that in turn has negative health effects, thus the app can remind users to take a walk or do movement at regular intervals.
- More time in bed shows more sleep time, thus the app can notify users of the bed time on daily basis which can also improve their sleep cycle and overall mental and physical health.
- Users can be motivated to do high intensity exercise between 6 to 8 PM as data shows they are most active in that time frame of the day.

Thank you for Your time to read this !!!!