Bellabeat project

Hao Tran

2022-05-28

Case Study 2: How can a wellness company play it smart?

In this case study, I will perform data analysis for Bellabeat, a high-tech manufacturer of health-focused products for women. I will analyze smart device data to gain insight into how consumers are using their smart devices. This analysis will help guide future marketing strategies for Bellabeat's team.

Insights: The high relation between TotalSteps and TotalDistance exposes that the users of the smart health trackers are usually running or jogging much more than using bicycles or cars. This means the customer segment that Bellabeat should focus on is the people who use running/walking as a way of commutting. This is also to say, the potential customers can be those who run and walk as a way of working out.

After analyzing the Fitbit dataset, I have discovered the general insights towards the use of smart device of Fitbit's customers:

- We expect that people who use the smart devices partially have concerned about their health. And they have, through inspecting their own health ratios. However, from the data analysis, haft of them, on average in a month, had not had enough time for activities per day. When it comes to sleeping, the significant portion of the valid users (whose information recorded is greater than 14 days) sleep for at least 7 hours a day on average, which means there is a small number of customers did not sleep enough.
- The customers prefer running/walking as a way to travel than using by bikes, cars, or other vehicles.

Installing and loading the essential packages

```
install.packages('tidyverse')
```

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

```
install.packages('reshape2')
```

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

```
library(reshape2)
library(tidyverse)
```

```
## ── Attaching packages
## ───────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.7     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

Loading CSV files

The project is conducted on two dataset: dailyActivity and sleepDay

```
daily_activity <- read.csv("dailyActivity_merged.csv")
sleep_day <- read.csv("sleepDay_merged.csv")
```

Exploring a few key tables

Take a look at the daily_activity data.

```
glimpse(daily_activity)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…
## $ TotalSteps               <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019…
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…
## $ FairlyActiveMinutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, …
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…
## $ Calories                 <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…
```

- Remove the duplicates

```
daily_activity <- unique(daily_activity)
```

- Formatting data We will transform ActivityDate from Character to Date

```
daily_activity$ActivityDate <- as.Date(mdy(daily_activity$ActivityDate))
is.Date(daily_activity$ActivityDate)
```

```
## [1] TRUE
```

Now, ActivityDate is in Date. Create a column which is the total time for VeryActiveMinutes and FairyActiveMinutes

```
daily_activity$TotalActiveMinutes <- daily_activity$VeryActiveMinutes + daily_activity$FairlyActiveMinutes
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   2016-04-12      13162          8.50            8.50
## 2 1503960366   2016-04-13      10735          6.97            6.97
## 3 1503960366   2016-04-14      10460          6.74            6.74
## 4 1503960366   2016-04-15       9762          6.28            6.28
## 5 1503960366   2016-04-16      12669          8.16            8.16
## 6 1503960366   2016-04-17       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
##   TotalActiveMinutes
## 1                 38
## 2                 40
## 3                 41
## 4                 63
## 5                 46
## 6                 58
```

- Take a look at the statistical summary of daily_activity.

```
daily_activity %>%
  select(-c(Id, ActivityDate, LoggedActivitiesDistance)) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    TrackerDistance  VeryActiveDistance
##  Min.   :    0   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 2.620   1st Qu.: 0.000
##  Median : 7406   Median : 5.245   Median : 5.245   Median : 0.210
##  Mean   : 7638   Mean   : 5.490   Mean   : 5.475   Mean   : 1.503
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.: 7.710   3rd Qu.: 2.053
##  Max.   :36019   Max.   :28.030   Max.   :28.030   Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5675           Mean   : 3.341      Mean   :0.001606
##  3rd Qu.:0.8000           3rd Qu.: 4.782      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.:127.0        1st Qu.: 729.8
##  Median :  4.00    Median :  6.00      Median :199.0        Median :1057.5
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8        Mean   : 991.2
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0        3rd Qu.:1229.5
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
##     Calories    TotalActiveMinutes
##  Min.   :   0   Min.   :  0.00
##  1st Qu.:1828   1st Qu.:  0.00
##  Median :2134   Median : 21.00
##  Mean   :2304   Mean   : 34.73
##  3rd Qu.:2793   3rd Qu.: 57.00
##  Max.   :4900   Max.   :275.00
```

- Let see how users spend their time for physical activities In the table, we can see that the numbers of date recorded for each user are not similar, many of them are recorded in a very short time. I recommend to only consider the users that have information for around 1 month which are greater that 27 days. At first, we create the data frame includes information of date and the average total time

```
Sub_daily_activity <- daily_activity %>%
  group_by(Id) %>%
  select(c(Id, ActivityDate, TotalActiveMinutes)) %>%
  mutate(Avg_activity_minutes = mean(TotalActiveMinutes))
```

We are going to figure out the users that have enough data recorded. Then, for these users, let see how many of them are fairly and very active in at least 30 mins per day on average.

```
Number_date_recorded <- as.data.frame(table(Sub_daily_activity$Id)) %>%
  filter(Freq > 27)
length(Number_date_recorded$Var1)
```

```
## [1] 27
```

There are 27 users that use the smart devices for about a month.

```
Longer_than_30mins <- Sub_daily_activity %>%
  filter(Id %in% Number_date_recorded$Var1) %>%
  filter(Avg_activity_minutes > 30)
length(unique(Longer_than_30mins$Id))
```

```
## [1] 14
```

There are 14 (out of) users are the faily and very active at least 30 mins a day on average.

- Data normalization TotalSteps column has the large range of values than other variables. So we need to center it. At this point, I use the Min-Max normalization method.

```
daily_activity_normalized <- daily_activity  %>%
 select(c(Id, ActivityDate, TotalDistance, VeryActiveDistance,  ModeratelyActiveDistance, LightActiveDistance, Se
dentaryActiveDistance,TotalActiveMinutes, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, Sedentary
Minutes, Calories))
daily_activity_normalized$TotalSteps <- (daily_activity$TotalSteps - min(daily_activity$TotalSteps)) / (max(daily
_activity$TotalSteps)- min(daily_activity$TotalSteps))
```

Take a look at the Sleep_day date.

```
glimpse(sleep_day)
```

```
## Rows: 413
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150…
## $ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "…
## $ TotalSleepRecords  <int> 1, 2, 1, 2, 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…
```

- Remove duplicates

```
sleep_day <- unique(sleep_day)
```

- Formatting data We transform SleepDate from Character to Date. Then, change TotalSleepRecords from integer to factor

```
sleep_day$SleepDay <- as.Date(mdy_hms(sleep_day$SleepDay))
sleep_day$TotalSleepRecords <- as.factor(sleep_day$TotalSleepRecords)
is.Date(sleep_day$SleepDay)
```

```
## [1] TRUE
```

- Take a look at the statistical summary of sleep_day.

```
sleep_day %>%
  select(-c(Id, SleepDay)) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  1:364             Min.   : 58.0      Min.   : 61.0
##  2: 43             1st Qu.:361.0      1st Qu.:403.8
##  3:  3             Median :432.5      Median :463.0
##                    Mean   :419.2      Mean   :458.5
##                    3rd Qu.:490.0      3rd Qu.:526.0
##                    Max.   :796.0      Max.   :961.0
```

FitBit's Customers usualy sleep for 419 mins (equivalent to approximately 7 hours a day)

Similar to Daily_actitity, we inspect the time for sleeping of the users using the Fitbit. However, because of the lack of observations, we will use the information of customer whose sleep recorded in at least 2 weeks.

```
Sub_sleep_day <- sleep_day %>%
  group_by(Id) %>%
  select(c(Id, SleepDay,TotalMinutesAsleep)) %>%
  mutate(Avg_Asleep_minutes = mean(TotalMinutesAsleep))
```

Filter out the users did not commit for a month

```
Number_date_recorded1 <- as.data.frame(table(Sub_sleep_day$Id)) %>%
  filter(Freq >= 14)
length(unique(Number_date_recorded1$Var1))
```

```
## [1] 15
```

Let see how many users slept for at least 7 hours on average

```
At_least_7hours <- Sub_sleep_day %>%
  filter(Id %in% Number_date_recorded1$Var1) %>%
  filter(Avg_Asleep_minutes >= 420)
length(unique(At_least_7hours$Id))
```

```
## [1] 10
```

There are 11 out of 15 people that slept at least 7 hours per day on average in during 2 weeks.

Merge daily_activity and sleep_day to inspect the relationship between variables from to tables

```
combined_data <- merge(daily_activity_normalized, sleep_day, by = "Id")
head(combined_data)
```

```
##           Id ActivityDate TotalDistance VeryActiveDistance
## 1 1503960366   2016-05-07          7.71               2.46
## 2 1503960366   2016-05-07          7.71               2.46
## 3 1503960366   2016-05-07          7.71               2.46
## 4 1503960366   2016-05-07          7.71               2.46
## 5 1503960366   2016-05-07          7.71               2.46
## 6 1503960366   2016-05-07          7.71               2.46
##   ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
## 1                     2.12                3.13                       0
## 2                     2.12                3.13                       0
## 3                     2.12                3.13                       0
## 4                     2.12                3.13                       0
## 5                     2.12                3.13                       0
## 6                     2.12                3.13                       0
##   TotalActiveMinutes VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
## 1                 83                37                  46                  175
## 2                 83                37                  46                  175
## 3                 83                37                  46                  175
## 4                 83                37                  46                  175
## 5                 83                37                  46                  175
## 6                 83                37                  46                  175
##   SedentaryMinutes Calories TotalSteps   SleepDay TotalSleepRecords
## 1              833     1821  0.3329354 2016-04-12                 1
## 2              833     1821  0.3329354 2016-04-13                 2
## 3              833     1821  0.3329354 2016-04-15                 1
## 4              833     1821  0.3329354 2016-04-16                 2
## 5              833     1821  0.3329354 2016-04-17                 1
## 6              833     1821  0.3329354 2016-04-19                 1
##   TotalMinutesAsleep TotalTimeInBed
## 1                327            346
## 2                384            407
## 3                412            442
## 4                340            367
## 5                700            712
## 6                304            320
```

Across the combined_data, **TotalDistance** is the sum of **VeryActiveDistance**, **ModeratelyActiveDistance, LightActiveDistance**, and **TotalActiveMinites** is the sum of **VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes**, we will not consider the relationship of the elements and their sum silmutaneously. Instead, we split the merged data into 2 plots.

Plotting 1 includes the sum and other variables

```
combined_data1 <- combined_data %>%
  select(-c(Id,ActivityDate, SleepDay, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, VeryAct
iveMinutes, FairlyActiveMinutes,TotalSleepRecords))
head(combined_data1)
```

```
##   TotalDistance SedentaryActiveDistance TotalActiveMinutes LightlyActiveMinutes
## 1          7.71                       0                 83                  175
## 2          7.71                       0                 83                  175
## 3          7.71                       0                 83                  175
## 4          7.71                       0                 83                  175
## 5          7.71                       0                 83                  175
## 6          7.71                       0                 83                  175
##   SedentaryMinutes Calories TotalSteps TotalMinutesAsleep TotalTimeInBed
## 1              833     1821  0.3329354                327            346
## 2              833     1821  0.3329354                384            407
## 3              833     1821  0.3329354                412            442
## 4              833     1821  0.3329354                340            367
## 5              833     1821  0.3329354                700            712
## 6              833     1821  0.3329354                304            320
```

We calculate the correlation coefficients

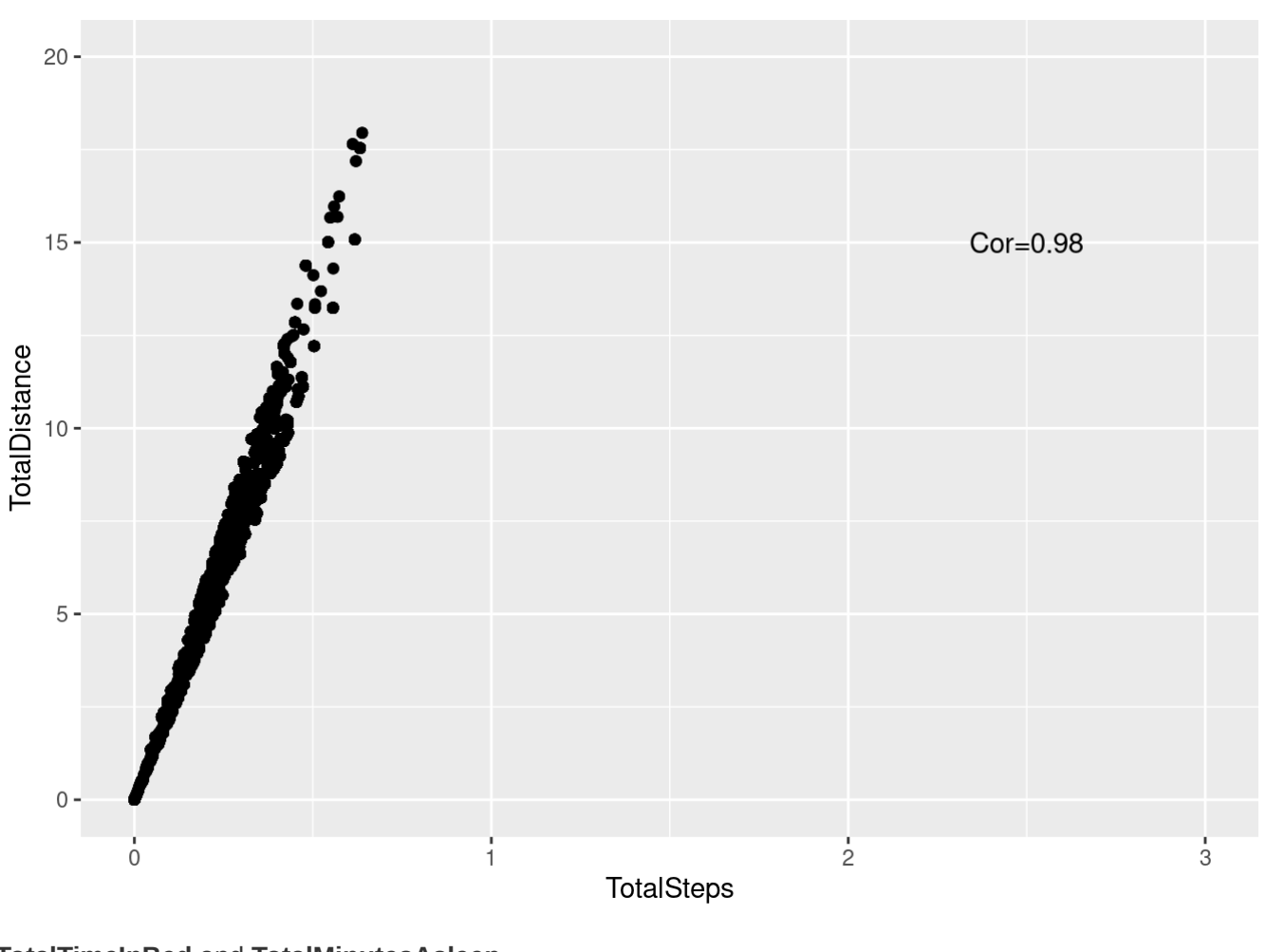
```
correlation1<- cor(combined_data1)
```

Visualize with a heatmap

```
heatmap1 <- melt(correlation1)
head(heatmap1)
```

```
##                      Var1          Var2       value
## 1           TotalDistance TotalDistance  1.00000000
## 2 SedentaryActiveDistance TotalDistance  0.07562118
## 3      TotalActiveMinutes TotalDistance  0.70824462
## 4    LightlyActiveMinutes TotalDistance  0.51256520
## 5        SedentaryMinutes TotalDistance -0.31407273
## 6                Calories TotalDistance  0.55069644
```

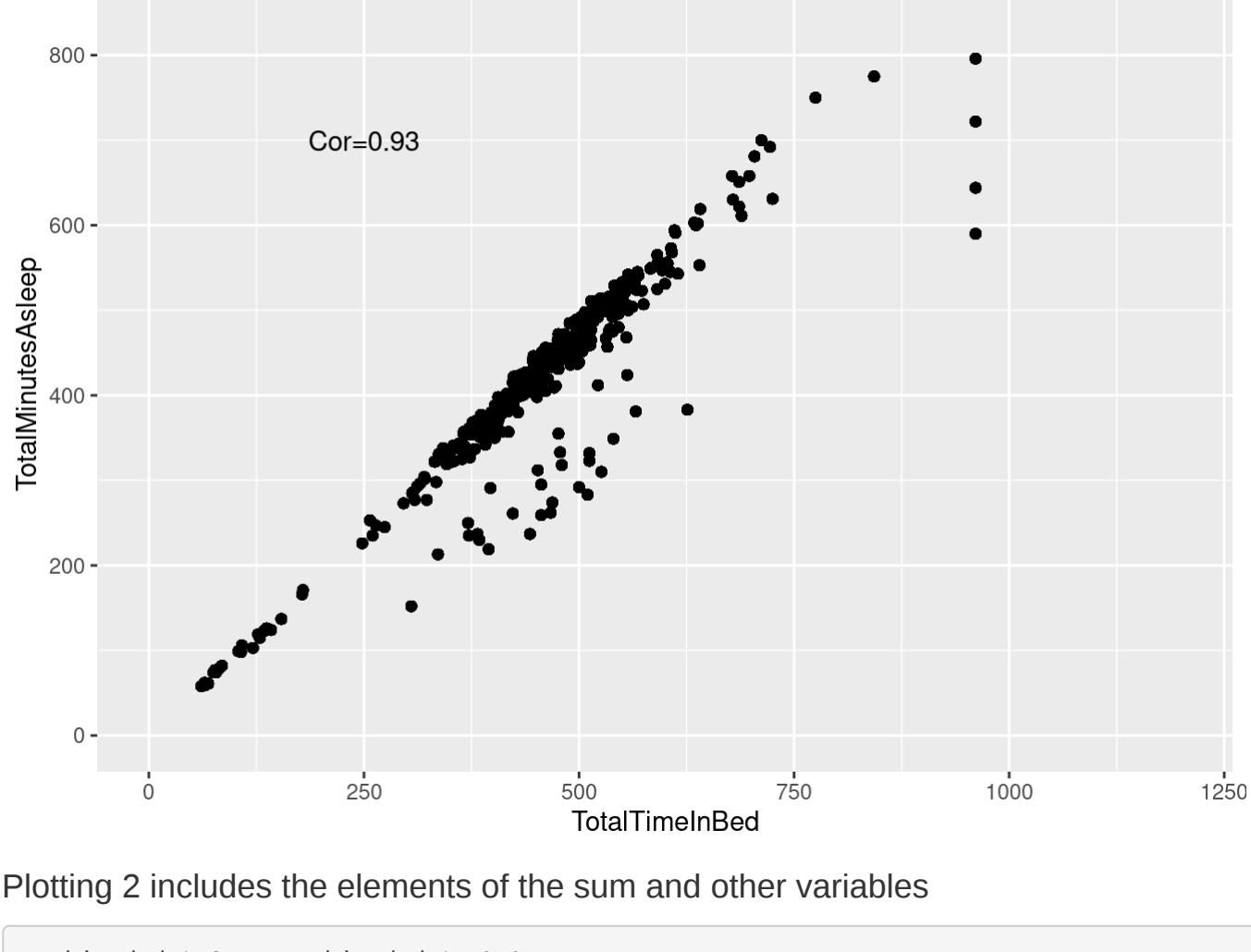
```
ggplot(data = heatmap1, aes(x= Var1, y=Var2, fill = value)) +
  geom_tile(color="white") +
  scale_fill_gradient2(low = "blue",high= "red", mid = "white", midpoint = 0) +
  theme(axis.text.x = element_text(angle = 45,vjust = 1, hjust = 1),
        axis.title.x = element_blank(), axis.title.y = element_blank())+
  labs(fill = 'Correlation values') + geom_text(aes(label= round(value,2)),
                                                 color = 'black', size= 3)
```



From the heatmap, some relations should be considered significant (higher or approximately to 0.8)

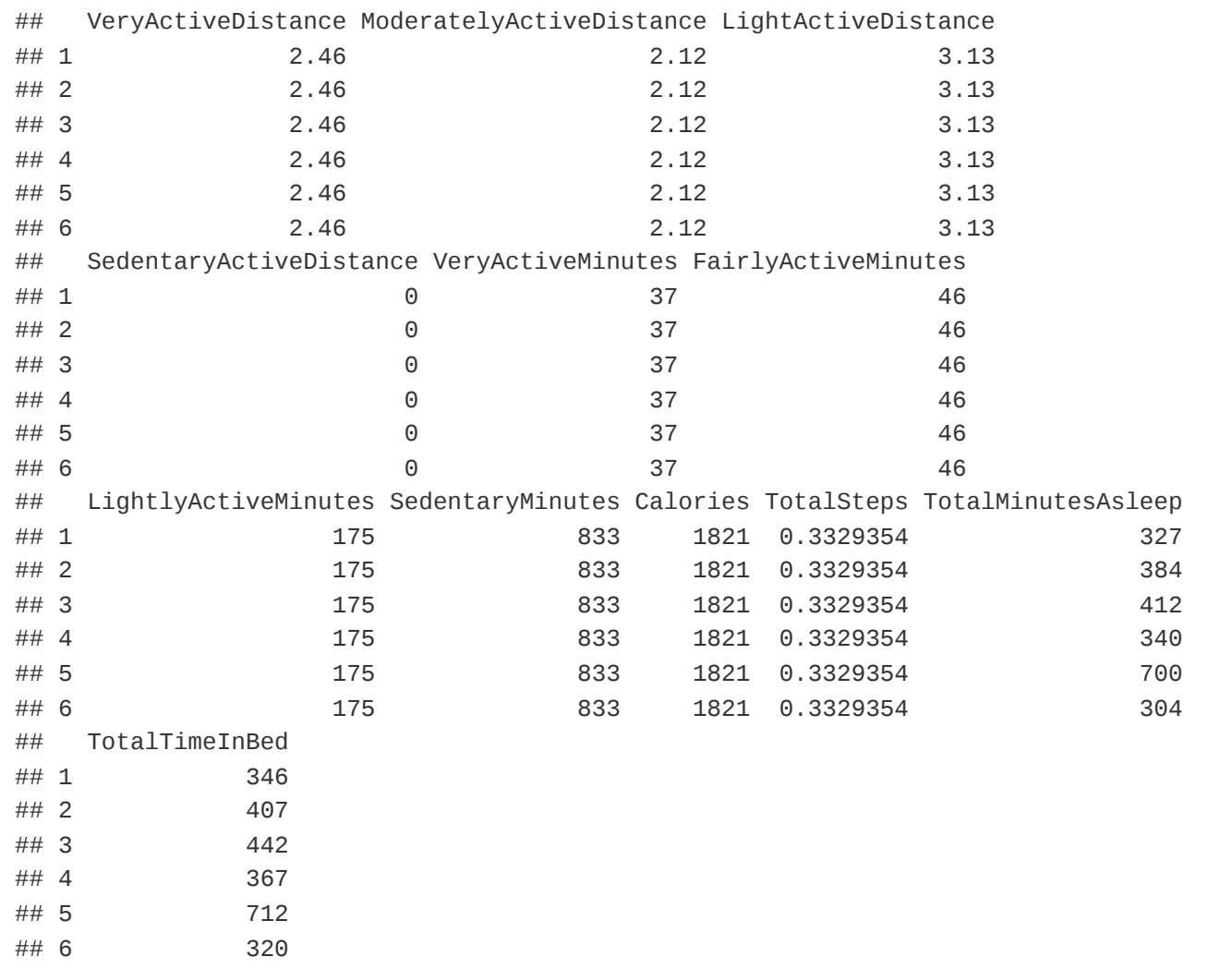
TotalSteps and **TotalDistance**

```
ggplot(data=combined_data1) + geom_point(mapping = aes(x=TotalSteps,y= TotalDistance))+   annotate("text", 2.5,15
,label=paste0("Cor=",round(cor(combined_data1$TotalSteps,combined_data1$TotalDistance),2))) + ylim(0,20) + xlim(0
,3)
```



TotalTimeInBed and **TotalMinutesAsleep**

```
ggplot(data=combined_data1) + geom_point(mapping = aes(x=TotalTimeInBed,y= TotalMinutesAsleep)) +
  annotate("text", 250,700,label=paste0("Cor=",round(cor(combined_data1$TotalTimeInBed,combined_data1$TotalMinute
sAsleep),2))) + ylim(0,850) + xlim(0,1200)
```



Plotting 2 includes the elements of the sum and other variables

```
combined_data2 <- combined_data %>%
  select(-c(Id,ActivityDate, SleepDay,TotalSleepRecords, TotalDistance,TotalActiveMinutes))
head(combined_data2)
```

```
##   VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
## 1               2.46                     2.12                3.13
## 2               2.46                     2.12                3.13
## 3               2.46                     2.12                3.13
## 4               2.46                     2.12                3.13
## 5               2.46                     2.12                3.13
## 6               2.46                     2.12                3.13
##   SedentaryActiveDistance VeryActiveMinutes FairlyActiveMinutes
## 1                       0                37                  46
## 2                       0                37                  46
## 3                       0                37                  46
## 4                       0                37                  46
## 5                       0                37                  46
## 6                       0                37                  46
##   LightlyActiveMinutes SedentaryMinutes Calories TotalSteps TotalMinutesAsleep
## 1                  175              833     1821  0.3329354                327
## 2                  175              833     1821  0.3329354                384
## 3                  175              833     1821  0.3329354                412
## 4                  175              833     1821  0.3329354                340
## 5                  175              833     1821  0.3329354                700
## 6                  175              833     1821  0.3329354                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

We calculate the correlation coefficients for this table

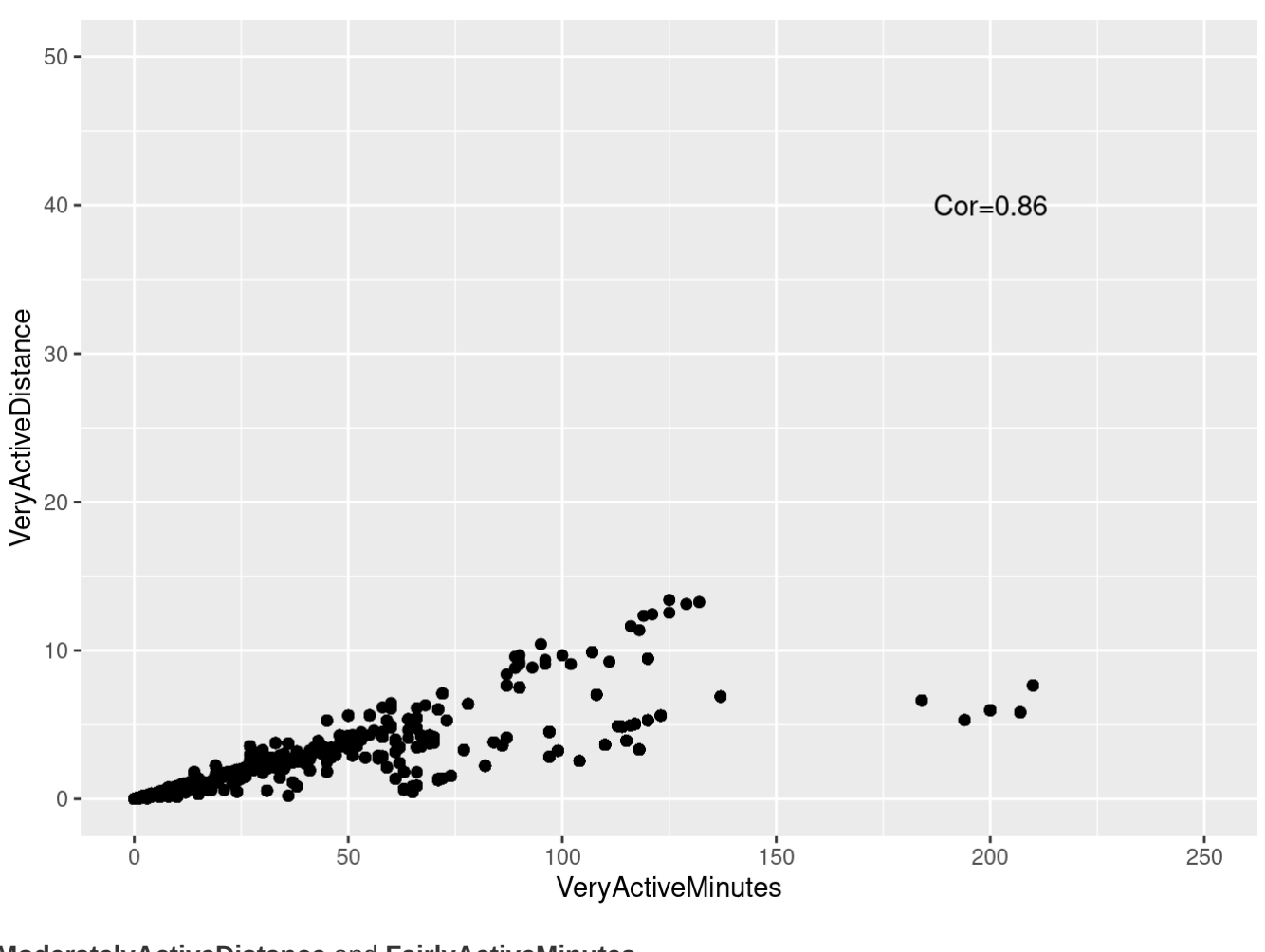
```
correlation2 <- cor(combined_data2)
```

Visualize with heatmap

```
heatmap2 <- melt(correlation2)
head(heatmap2)
```

```
##                       Var1               Var2      value
## 1       VeryActiveDistance VeryActiveDistance 1.00000000
## 2 ModeratelyActiveDistance VeryActiveDistance 0.24037875
## 3      LightActiveDistance VeryActiveDistance 0.07295066
## 4  SedentaryActiveDistance VeryActiveDistance 0.07139521
## 5        VeryActiveMinutes VeryActiveDistance 0.85796442
## 6      FairlyActiveMinutes VeryActiveDistance 0.29406749
```

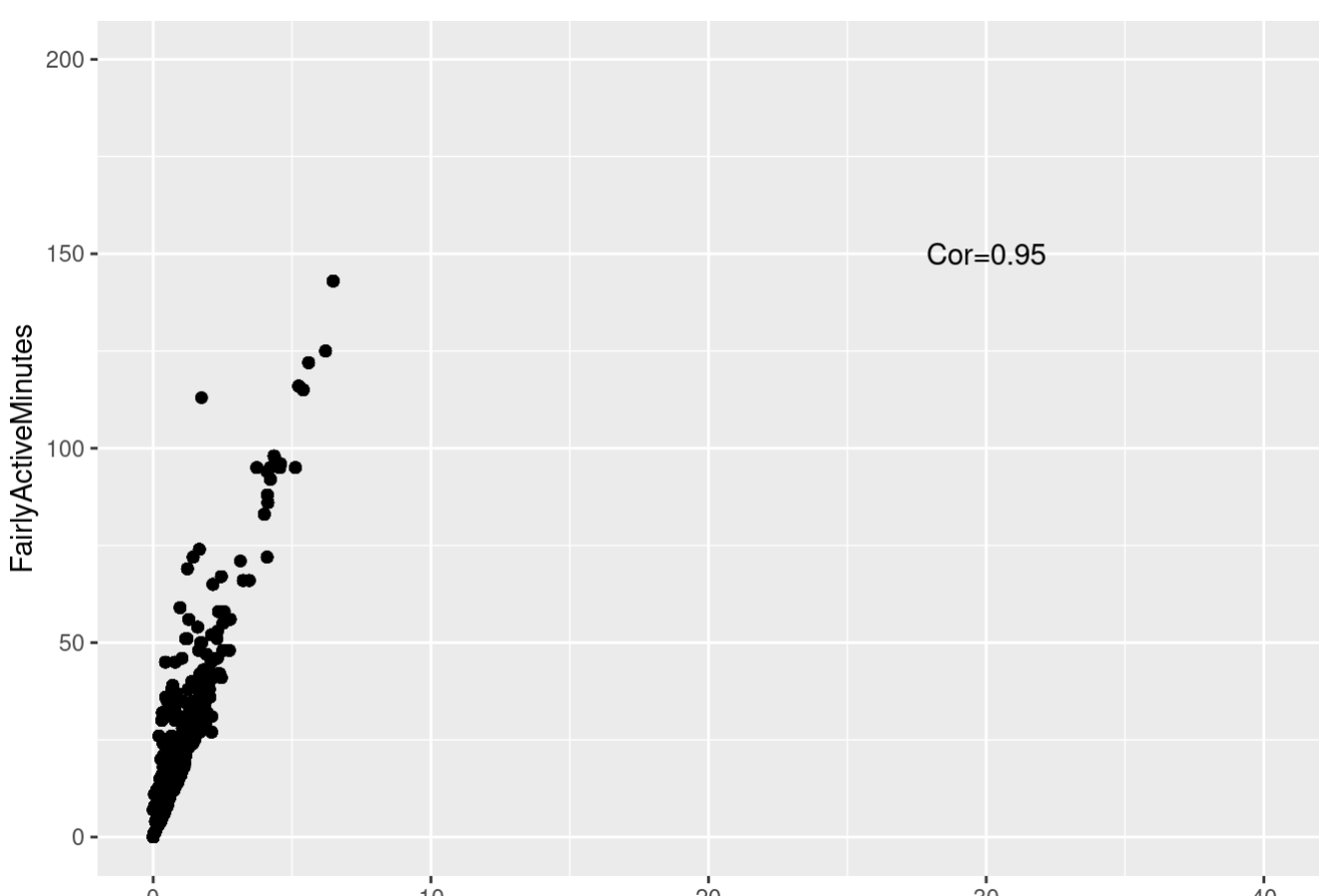
```
ggplot(data = heatmap2, aes(x= Var1, y=Var2, fill = value)) + geom_tile(color="white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) + theme(axis.text.x = element_tex
t(angle = 45,vjust = 1, hjust = 1),axis.title.x = element_blank(), axis.title.y = element_blank()) + labs(fill =
'Correlation values') + geom_text(aes(label= round(value,2)),color = 'black', size= 3)
```



Some relations should be considered significant (higher or approximately to 0.8)

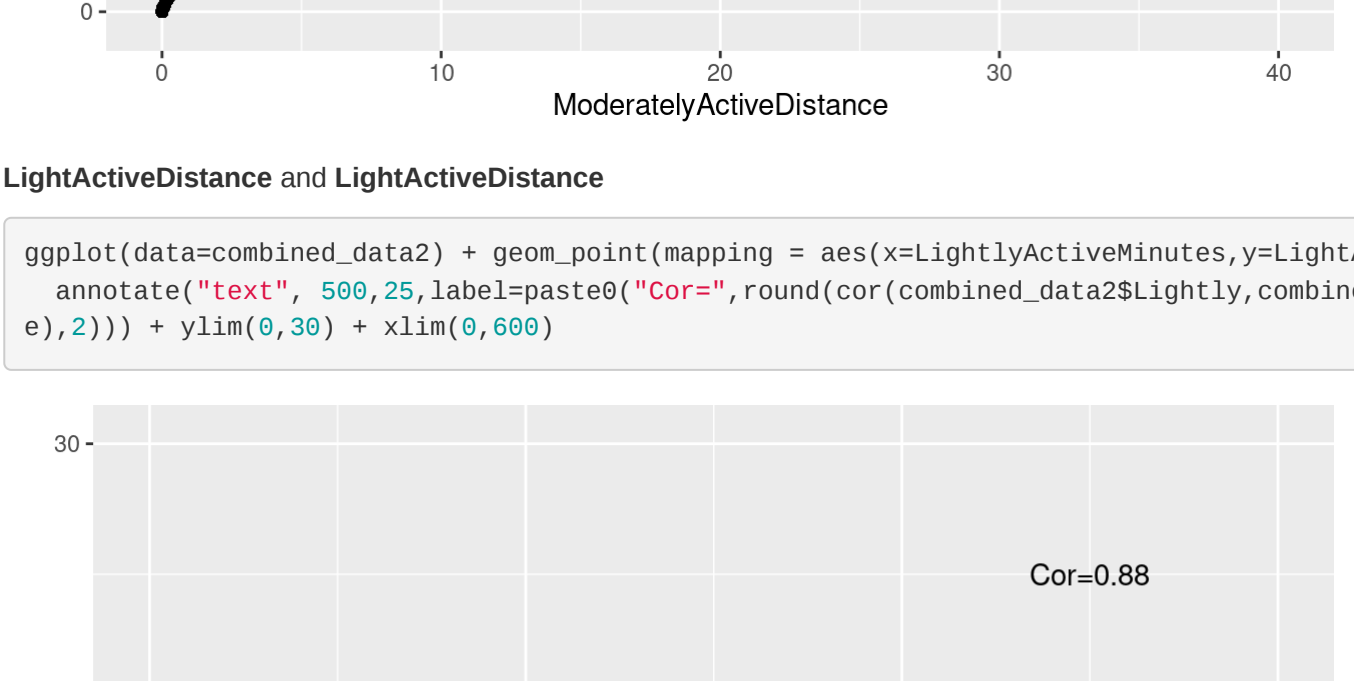
VeryActiveMinutes and **VeryActiveDistance**

```
ggplot(data=combined_data2) + geom_point(mapping = aes(x=VeryActiveMinutes,y=VeryActiveDistance)) +
  annotate("text", 200,40,label=paste0("Cor=",round(cor(combined_data2$VeryActiveMinutes,combined_data2$VeryActiv
eDistance),2))) + ylim(0,50) + xlim(0,250)
```



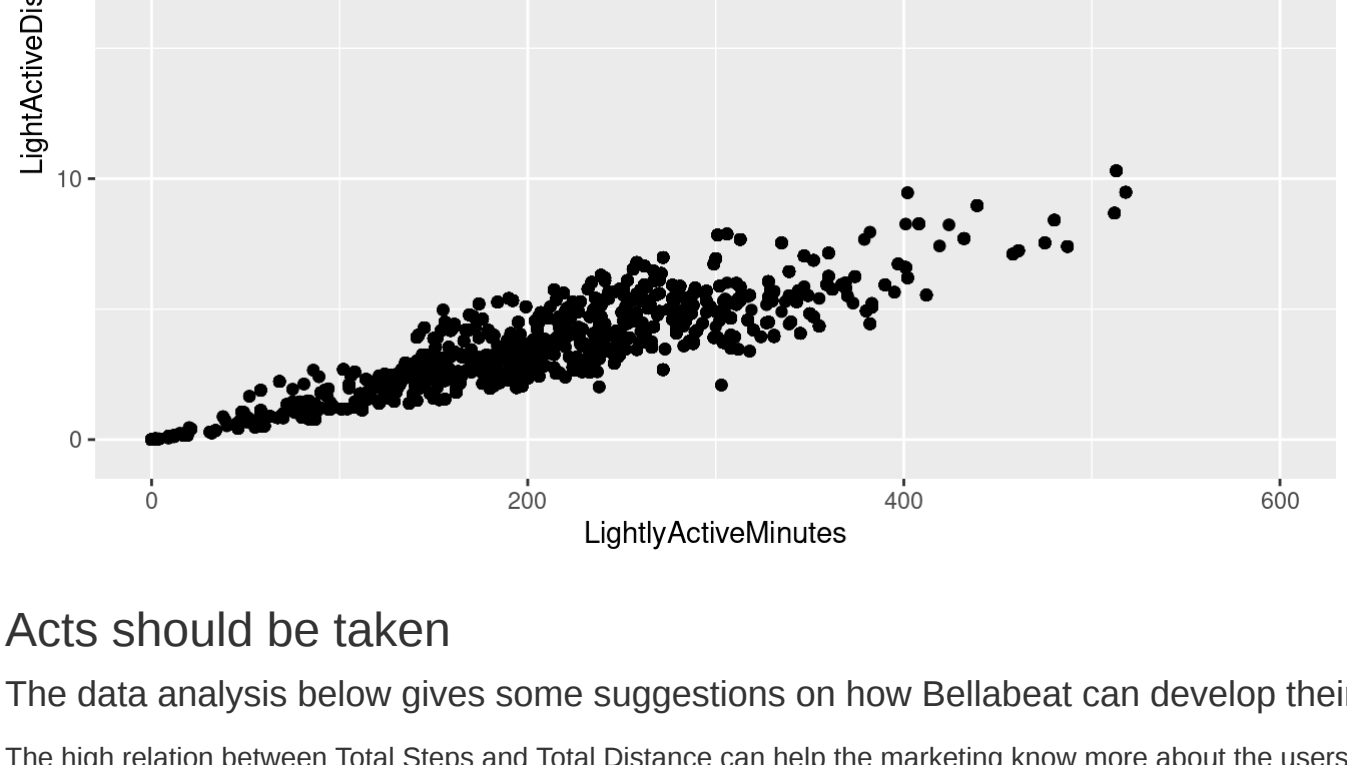
ModeratelyActiveDistance and **FairlyActiveMinutes**

```
ggplot(data=combined_data2) + geom_point(mapping = aes(x=ModeratelyActiveDistance,y=FairlyActiveMinutes)) +
  annotate("text", 30,150,label=paste0("Cor=",round(cor(combined_data2$ModeratelyActiveDistance,combined_data2$Fa
irlyActiveMinutes),2))) + ylim(0,200) + xlim(0,40)
```



LightActiveDistance and **LightActiveDistance**

```
ggplot(data=combined_data2) + geom_point(mapping = aes(x=LightlyActiveMinutes,y=LightActiveDistance)) +
  annotate("text", 500,25,label=paste0("Cor=",round(cor(combined_data2$Lightly,combined_data2$LightActiveDistanc
e),2))) + ylim(0,30) + xlim(0,600)
```



Acts should be taken

The data analysis below gives some suggestions on how Bellabeat can develop their products.

The high relation between Total Steps and Total Distance can help the marketing know more about the users of smart devices. this finding lead to a new opportunity of a customer segment. They should focus on the customer that usually travel by foot or other ones who often run or jog. Also, the people who are supposed to be live in a healthy way actually have some problems with the total time for physical activities and sleeping. The teams backing for Bellabeat app should pay their effort on solving these problems for the future users. Finally, more research should be taken further to explore more knowledge of this data and the answer the questions around the high degree of relation between levels of minute and distance, between total time asleep and total time in bed.