Google Data Analytics Professional Certificate study case report

Julián David Candela

2023-08-31

This report was created as an assignment for the Google Data Analytics Professional Certificate.

Hypothetical case: The Bellabeat company believes that analyzing physical activity data from smart devices could unlock new business opportunities for the company. I've been asked analyze the FitBit Fitness Tracker Data set from Kaggle to learn how users are using their smart devices and present high-level recommendations for Bellabeat's marketing strategy.

Preliminars

Cleaning:

Google sheets was used to clean and format the data before loading it to R. Most of the process was to make sure that all the columns have a suitable format for the analysis process that we want to carry out.

Installing and loading packages and libraries:

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

Loading the data:

```
daily_activity <- read.csv("dailyActivity_merged.csv")
sleep_day <- read.csv("cleaned.csv")
weight <- read.csv("weightLogInfo_merged.csv")
calories<- read.csv("dailyCalories_merged.csv")
```

Basic Information about TotalMinutesAsleep and TotalTimeInBed

We calculate the average of the variables TotalTimeInBed and TotalMinutesAsleep as an estimator of the amount of time a user is in bed and sleeps. We also calculate the minimum and maximum of these variables.

```
sleep_day %>%   drop_na() %>%  summarize(average_time_bed=mean(TotalTimeInBed), min_time_bed=min(TotalT
```

```
##   average_time_bed min_time_bed max_time_bed
## 1         458.6392           61          961
```

```
sleep_day  %>%  drop_na() %>%  summarize(average_time_asleep=mean(TotalMinutesAsleep), min_time_asleep=
```

```
##   average_time_asleep min_time_asleep max_time_asleep
## 1            419.4673             58            796
```

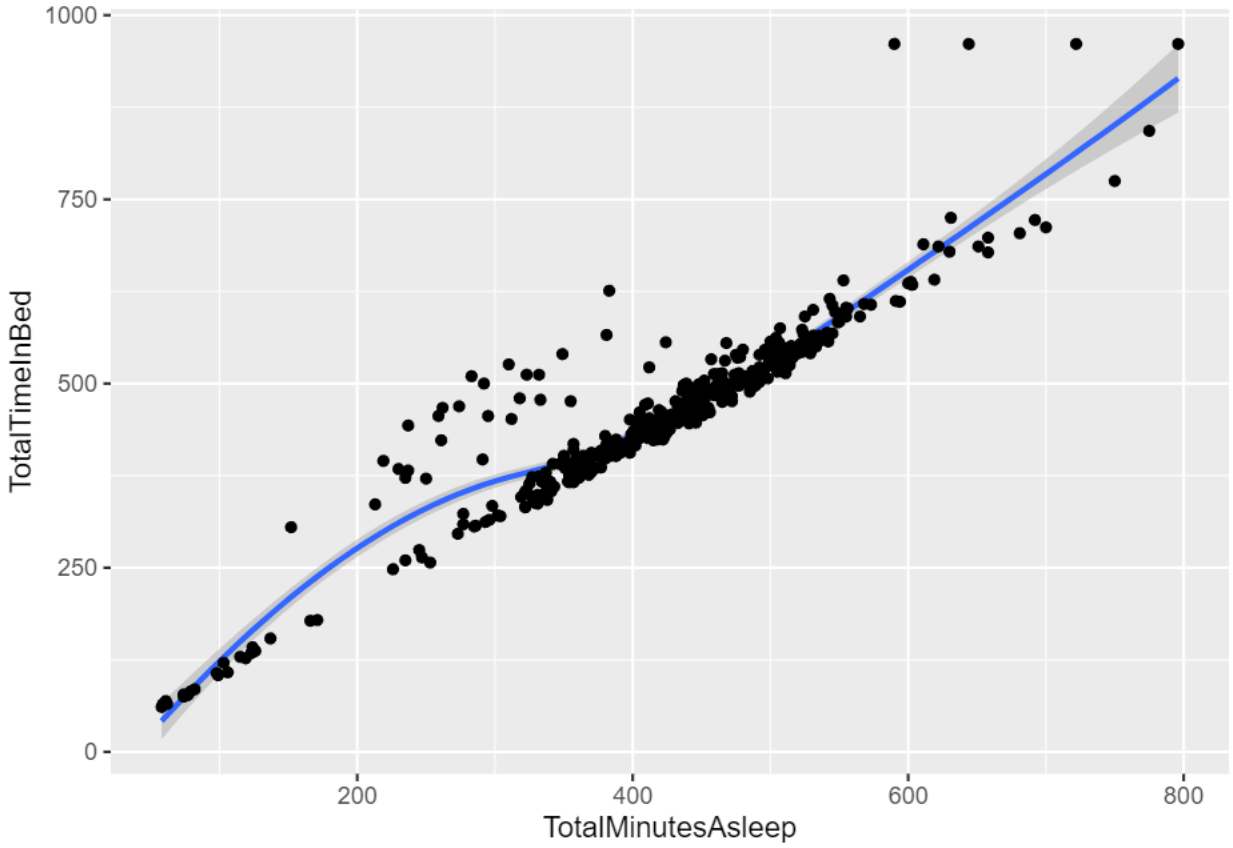
Comparing the variables TotalMinutesAsleep vs TotalTimeInBed

We want to compare these two variables to find any possible relation between them.

Plotting:

For that, we begin plotting the variables in order to study the behavior of the data. We also add a trend line.

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +geom_smooth()+ geom_point()
```



Correlation of the variables TotalMinutesAsleep and TotalTimeInBed:

The plotting of the variables TotalMinutesAsleep vs TotalTimeInBed suggests they are correlated, we verify this by calculating the correlation coefficient.

```
cor(x=sleep_day$TotalMinutesAsleep, y=sleep_day$TotalTimeInBed)
```

```
## [1] 0.9304575
```

Diference betwwen TotalMinutesAsleep and TotalTimeInBed:

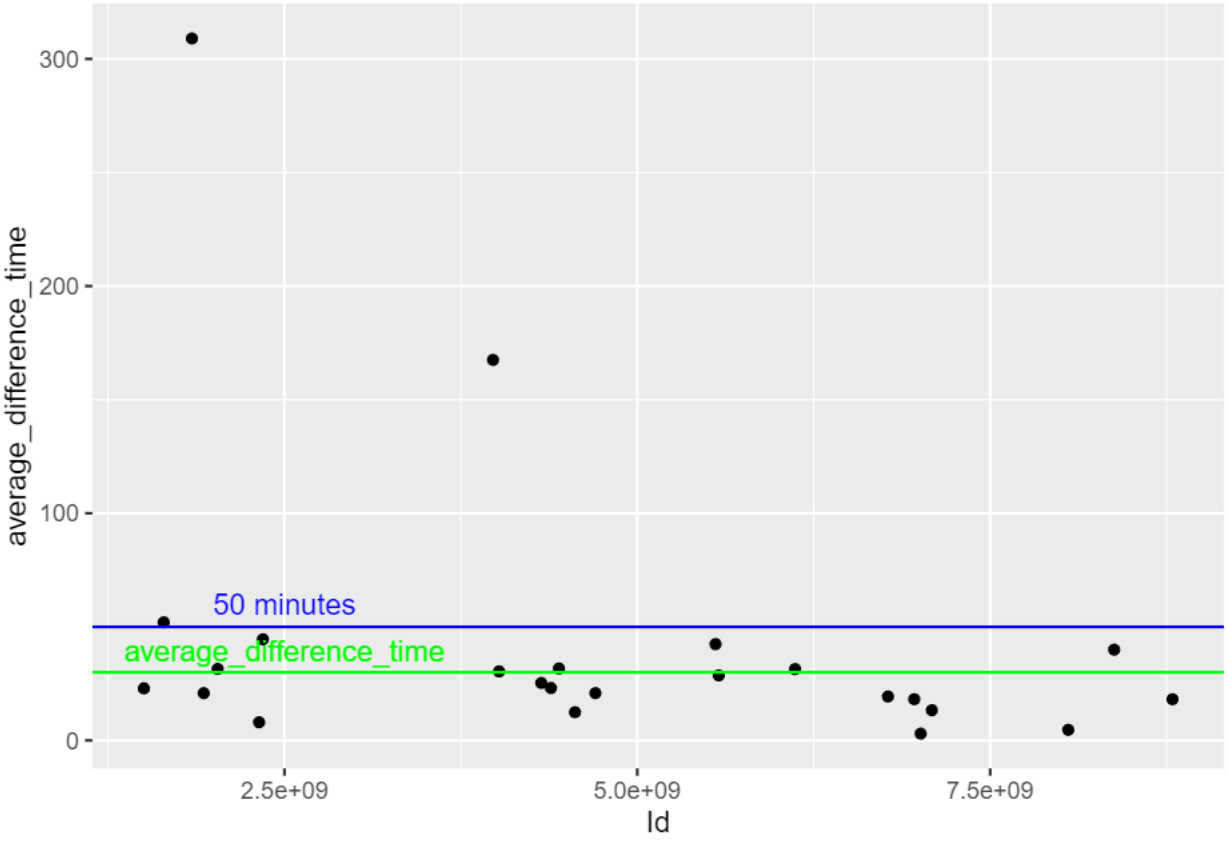
Now, we want compare the two variables TotalMinutesAsleep and TotalTimeInBed in order to to obtain an estimate of their difference. First we calculate the average difference of these variables

```
sleep_day %>% drop_na() %>%  summarize(average_difference_time=mean(TotalMinutesDif))
```

```
##   average_difference_time
## 1                39.17191
```

also, we decide to plot the average difference by user and add two lines to the graphic allow the reader to compare with two different values the average_difference_time and the value 50 minutes.

```
sleep_day  %>% group_by(Id) %>% drop_na() %>%  summarize(average_difference_time=mean(TotalMinutesDif))
```



Analysis by user

Now, we want to make some analysis by user in order to have a better understanding of the behavior of the data.

Basic inforation TotalMinutesAsleep and otalTimeInBed by user:

```
sleep_day %>% group_by(Id) %>% drop_na() %>%  summarize(average_time_bed=mean(TotalTimeInBed),min_time_
```

```
## # A tibble: 24 x 4
##            Id average_time_bed min_time_bed max_time_bed
##         <dbl>            <dbl>        <int>        <int>
##  1 1503960366             383.          264          712
##  2 1644430081             346           127          961
##  3 1844505072             961           961          961
##  4 1927972279             438.          178          775
##  5 2026352035             538.          380          607
##  6 2320127002              69            69           69
##  7 2347167796             491.          386          602
##  8 3977333714             461.          305          626
##  9 4020332650             380.           77          541
## 10 4319703577             502.           65          722
```

```
## # i 14 more rows
```

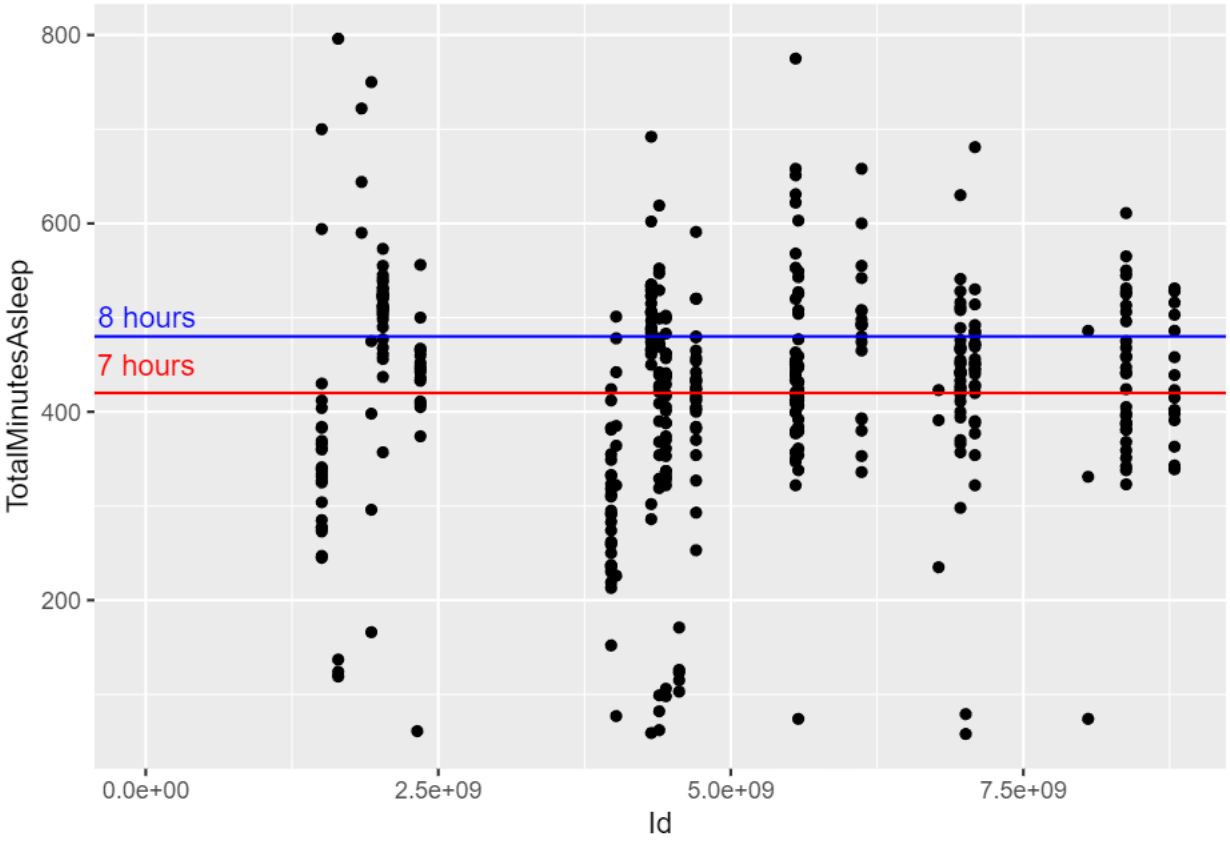
```
sleep_day  %>% group_by(Id) %>% drop_na() %>%  summarize(mean_time_asleep=mean(TotalMinutesAsleep), min
```

```
## # A tibble: 24 x 4
##            Id mean_time_asleep min_time_bed max_time_bed
##         <dbl>            <dbl>        <int>        <int>
##  1 1503960366             360.          245          700
##  2 1644430081             294           119          796
##  3 1844505072             652           590          722
##  4 1927972279             417           166          750
##  5 2026352035             506.          357          573
##  6 2320127002              61            61           61
##  7 2347167796             447.          374          556
##  8 3977333714             294.          152          424
##  9 4020332650             349.           77          501
## 10 4319703577             477.           59          692
## # i 14 more rows
```

Comparing TotalMinutesAsleep vs the recomended amount of sleep

Also, we want to compare how often users sleep an amount of hours below the recommended amount of sleep.

```
ggplot(data=sleep_day, aes(x=Id, y=TotalMinutesAsleep)) + geom_point()+  geom_hline(yintercept = 420, c
```



Analysis by weekday

We are also interested to know if the weekday is related to the amount of hours that the user rests.

Basic inforation TotalMinutesAsleep and otalTimeInBed by weekday:

```
sleep_day %>% group_by(DayWeek) %>% drop_na() %>%  summarize(average_time_bed=mean(TotalTimeInBed),min_
```

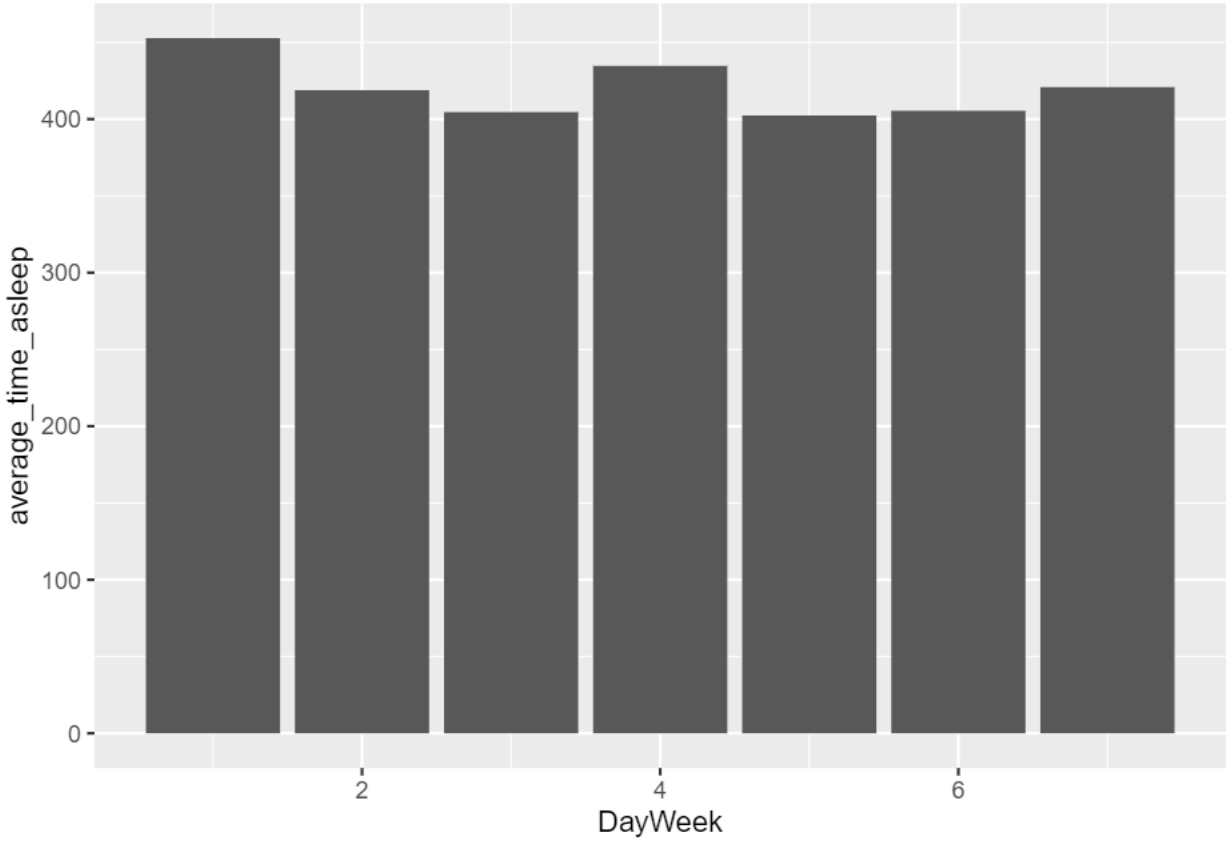
```
## # A tibble: 7 x 4
##   DayWeek average_time_bed min_time_bed max_time_bed
##     <int>            <dbl>        <int>        <int>
## 1       1             504.           61          961
## 2       2             456.           65          961
## 3       3             443.          121          775
## 4       4             470.          260          679
## 5       5             436.           65          568
## 6       6             445.           85          961
## 7       7             461.           69          961
```

```
sleep_day  %>% group_by(DayWeek) %>% drop_na() %>%  summarize(mean_time_asleep=mean(TotalMinutesAsleep)
```

```
## # A tibble: 7 x 4
##   DayWeek mean_time_asleep min_time_asleep max_time_asleep
##     <int>            <dbl>           <int>           <int>
## 1       1             453.              58             700
## 2       2             419.              62             796
## 3       3             405.             103             750
## 4       4             435.             152             658
## 5       5             402.              59             545
## 6       6             405.              82             658
## 7       7             421.              61             775
```

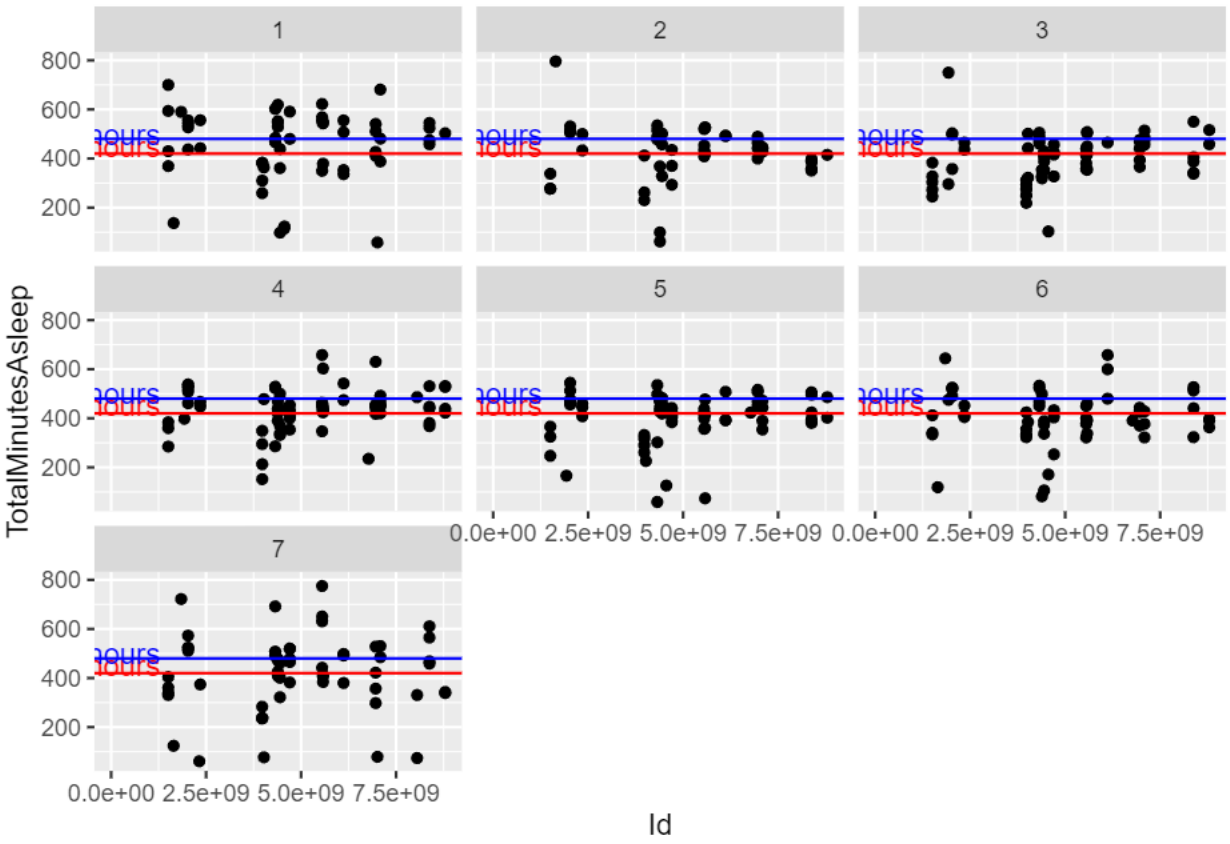
We also add a graphic to have a better understanding of the of the average_time_asleep behavior per day.

```
sleep_day  %>% group_by(DayWeek) %>% drop_na() %>%  summarize(average_time_asleep=mean(TotalMinutesAsle
arrange(DayWeek)  %>%
ggplot(aes(x=DayWeek, y=average_time_asleep)) + geom_col(position = "dodge")
```



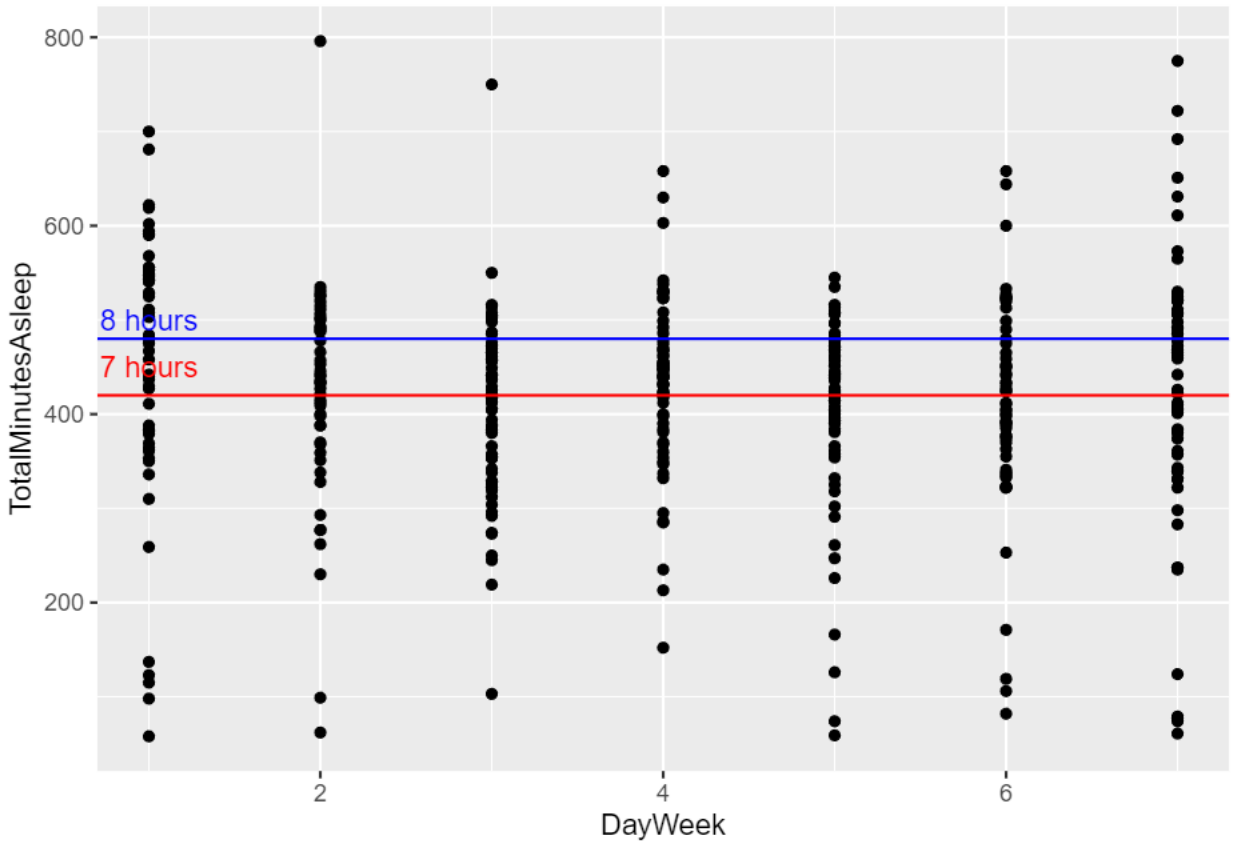
Furthermore, we want to compare the Total time asleep vs the recommended amount of sleep by weekday to identify possible trends.

```
ggplot(data=sleep_day, aes(x=Id, y=TotalMinutesAsleep)) + geom_point()+  geom_hline(yintercept = 420, c
```



now, we classify the data provided to make a comparison of sleep hours by weekday

```
ggplot(data=sleep_day, aes(x=DayWeek, y=TotalMinutesAsleep)) + geom_point()+  geom_hline(yintercept = 4
```



Merging 2 data sets together:

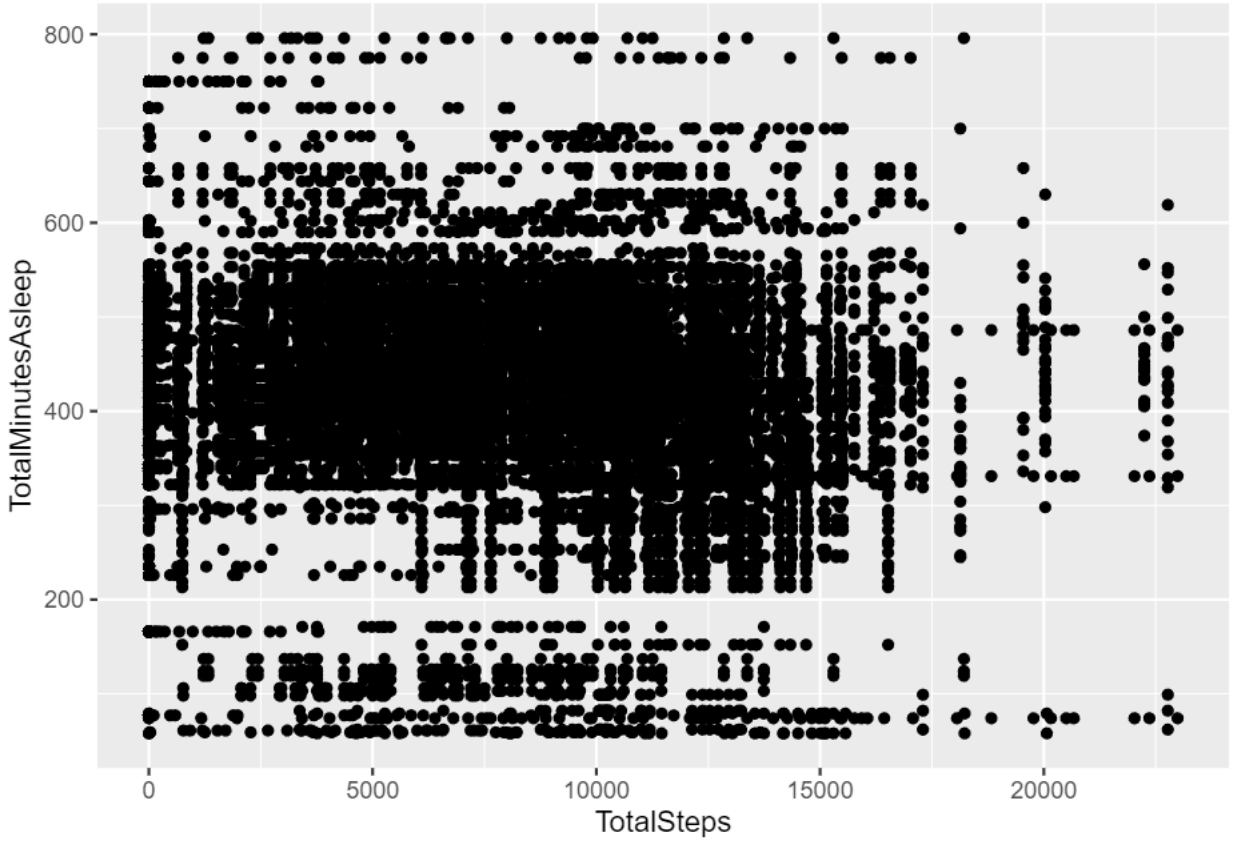
In order to deduce other possible trends, we merge the two data sets "sleep_day" and "daily_activity".

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

Plotting the merged data:

now, we plot the variables TotalSteps vs TotalMinutesAsleep

```
ggplot(data=combined_data, aes(x=TotalSteps, y=TotalMinutesAsleep))+ geom_point()
```



Correlation:

We calculate the correlation between the variables **TotalSteps** vs **TotalMinutesAsleep**

```
cor(x=combined_data$TotalSteps, y=combined_data$TotalMinutesAsleep)
```

```
## [1] -0.09854146
```

Results

Summary:

- The average TotalMinutesAsleep is 419.46 min, that is approximately 7 hours. However the graphics show that most user's sleep times are below the healthy recommended amount of time sleep.
- The correlation coefficient r=0.93 guarantee that the variables TotalTimeInBed and TotalMinutesAsleep are highly correlated.
- The average Total Time Difference between the variables TotalTimeInBed and TotalMinutesAsleep is 39.17 minutes. The graphics showed that this Difference is below average for most of the users.
- In average, the day in which the least number of hours of sleep is recorded is Thursday.
- In average, the day in which the most number of hours of sleep is recorded is Sunday.
- The correlation coefficient r=-0.09 guarantee that the variables TotalSteps and TotalMinutesAsleep are not correlated.

Observations:

- Although, in average, the amount of total time for a user to be asleep is reasonable good, making the analysis by user we found out that actually some of the users don't sleep enough.
- For most users, the average Difference between the variables TotalTimeInBed and TotalMinutesAsleep is below 39.17 minutes. Thus, approximately, we expect a user to sleep the same amount of time that they are in bed except for 40 minutes.
- Despite the fact that, on average, Sundays and Thursdays are the days when users have more and fewer hours of sleep, respectively; When examining the plotting made, we conclude that there is not a clear tendency for users to sleep better one day than another.
- The data suggest that the variables TotalSteps and TotalMinutesAsleep are not related.

Recommendations:

- The data suggests that almost all users sleeps most of the time that they are in bed; however, the data also shows that in many cases this last amount of time is not enough for a user to rest well. We recommend to add an alert to the device that suggest the user to go to bed early and that it indicates how much time will the user sleep if going to bed right now, based on his usual time to wake up.
- We suggest to recollect more data and make a new analysis in order to deduce new trends.