Bellabeat case study

2022-11-30

##data analysis and visualization using R programming

#installing common packages and libraries

```
install.packages("tidyverse")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.1 -- Conflicts ------------------------------------------ tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(knitr)
```

#importing the sleepDay_merged into sleep_data

```
sleep_data <- read.csv("sleepDay_merged.csv")
```

#reading the table sleep_data

```
head(sleep_data)
```

```
##           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
```

#column names of the table sleep_data

```
colnames(sleep_data)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
```

```
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

#summary of columns

```
sleep_data%>%
  select(TotalTimeInBed,TotalMinutesAsleep,TotalSleepRecords)%>%
  summary()
```

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

#changing the column names into lower case

```
names(sleep_data)<-tolower(names(sleep_data))
```

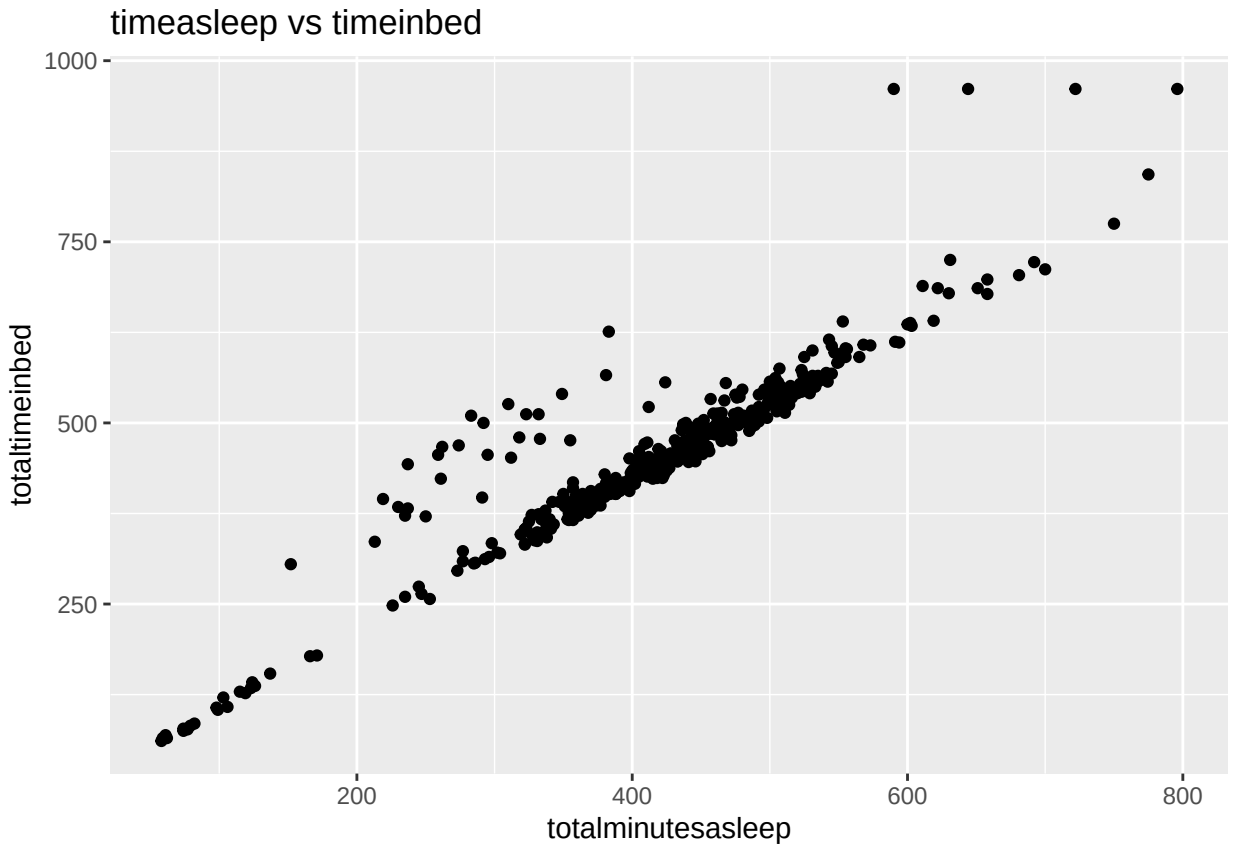
#checking the columns names

```
colnames(sleep_data)
```

```
## [1] "id"                 "sleepday"           "totalsleeprecords"
## [4] "totalminutesasleep" "totaltimeinbed"
```

#relation between totalminutesasleep and totaltimeinbed

```
ggplot(data=sleep_data,aes(x=totalminutesasleep, y=totaltimeinbed))+geom_point()+ggtitle("timeasleep vs
```



#importing another dataset into daily_activity

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

#reading the data daily_activity

```
head(daily_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
```

#converting the column names into lower case

```
names(daily_activity)<-tolower(names(daily_activity))
```

```
colnames(daily_activity)
```

```
##  [1] "id"                       "activitydate"
##  [3] "totalsteps"               "totaldistance"
##  [5] "trackerdistance"          "loggedactivitiesdistance"
##  [7] "veryactivedistance"       "moderatelyactivedistance"
##  [9] "lightactivedistance"      "sedentaryactivedistance"
## [11] "veryactiveminutes"        "fairlyactiveminutes"
## [13] "lightlyactiveminutes"     "sedentaryminutes"
## [15] "calories"
```

#changing the activitydate format as y-m-d

```
daily_activity$activitydate <- format(as.Date(daily_activity$activitydate, format = "%m/%d/%Y"), "%Y-%m
```

```
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
```

#loading date related library

```
library(lubridate)
```

```
## Loading required package: timechange
```

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

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

#calculating the day of week from the activitydate

```
daily_activity$weekday <- wday(daily_activity$activitydate, label=TRUE)
```

```
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 weekday
## 1                  13                  328              728     1985     Tue
## 2                  19                  217              776     1797     Wed
## 3                  11                  181             1218     1776     Thu
## 4                  34                  209              726     1745     Fri
## 5                  10                  221              773     1863     Sat
## 6                  20                  164              539     1728     Sun
```

#extracting and formating the date from sleepday

```
sleep_data$sleepdate <- format(as.Date(sleep_data$sleepday, format = "%m/%d/%Y"), "%Y-%m-%d")
head(sleep_data)
```

```
##           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  sleepdate
## 1            346 2016-04-12
## 2            407 2016-04-13
## 3            442 2016-04-15
## 4            367 2016-04-16
## 5            712 2016-04-17
## 6            320 2016-04-19
```

#merging the two data sets

```
combined_data <- merge(daily_activity, sleep_data, by.x=c("id","activitydate"),by.y = c("id","sleepdate
```

```
head(combined_data)
```

```
##           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   2016-04-15       9762          6.28            6.28
## 4 1503960366   2016-04-16      12669          8.16            8.16
## 5 1503960366   2016-04-17       9705          6.48            6.48
## 6 1503960366   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 weekday
## 1                  13                  328              728     1985     Tue
## 2                  19                  217              776     1797     Wed
## 3                  34                  209              726     1745     Fri
## 4                  10                  221              773     1863     Sat
## 5                  20                  164              539     1728     Sun
## 6                  31                  264              775     2035     Tue
##                  sleepday totalsleeprecords totalminutesasleep totaltimeinbed
## 1 4/12/2016 12:00:00 AM                 1                327            346
## 2 4/13/2016 12:00:00 AM                 2                384            407
## 3 4/15/2016 12:00:00 AM                 1                412            442
## 4 4/16/2016 12:00:00 AM                 2                340            367
## 5 4/17/2016 12:00:00 AM                 1                700            712
## 6 4/19/2016 12:00:00 AM                 1                304            320
```

#converting the minutes into hours

```
combined_data$totalhoursasleep<-round(combined_data$totalminutesasleep/60,2)
head(combined_data)
```

```
##           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   2016-04-15       9762          6.28            6.28
## 4 1503960366   2016-04-16      12669          8.16            8.16
## 5 1503960366   2016-04-17       9705          6.48            6.48
## 6 1503960366   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 weekday
## 1                  13                  328              728     1985     Tue
## 2                  19                  217              776     1797     Wed
## 3                  34                  209              726     1745     Fri
## 4                  10                  221              773     1863     Sat
```

```
## 5                  20                 164              539     1728      Sun
## 6                  31                 264              775     2035      Tue
##                sleepday totalsleeprecords totalminutesasleep totaltimeinbed
## 1 4/12/2016 12:00:00 AM                 1                327            346
## 2 4/13/2016 12:00:00 AM                 2                384            407
## 3 4/15/2016 12:00:00 AM                 1                412            442
## 4 4/16/2016 12:00:00 AM                 2                340            367
## 5 4/17/2016 12:00:00 AM                 1                700            712
## 6 4/19/2016 12:00:00 AM                 1                304            320
##   totalhoursasleep
## 1             5.45
## 2             6.40
## 3             6.87
## 4             5.67
## 5            11.67
## 6             5.07
```

#formating the totalhourasleep

```
combined_data$formathoursasleep <- strftime(as.POSIXct(combined_data$totalhoursasleep*3600,origin='1900
```

```
head(combined_data)
```

```
##           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   2016-04-15       9762          6.28            6.28
## 4 1503960366   2016-04-16      12669          8.16            8.16
## 5 1503960366   2016-04-17       9705          6.48            6.48
## 6 1503960366   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 weekday
## 1                  13                 328              728     1985      Tue
## 2                  19                 217              776     1797      Wed
## 3                  34                 209              726     1745      Fri
## 4                  10                 221              773     1863      Sat
## 5                  20                 164              539     1728      Sun
## 6                  31                 264              775     2035      Tue
##                sleepday totalsleeprecords totalminutesasleep totaltimeinbed
## 1 4/12/2016 12:00:00 AM                 1                327            346
## 2 4/13/2016 12:00:00 AM                 2                384            407
## 3 4/15/2016 12:00:00 AM                 1                412            442
```

```
## 4 4/16/2016 12:00:00 AM                2                340            367
## 5 4/17/2016 12:00:00 AM                1                700            712
## 6 4/19/2016 12:00:00 AM                1                304            320
##   totalhoursasleep formathoursasleep
## 1             5.45          05:27:00
## 2             6.40          06:24:00
## 3             6.87          06:52:12
## 4             5.67          05:40:12
## 5            11.67          11:40:12
## 6             5.07          05:04:12
```

#importing correlation package

```
library(corrr)
```

#finding the correlation between the columns

```
combined_data %>%
  correlate()
```

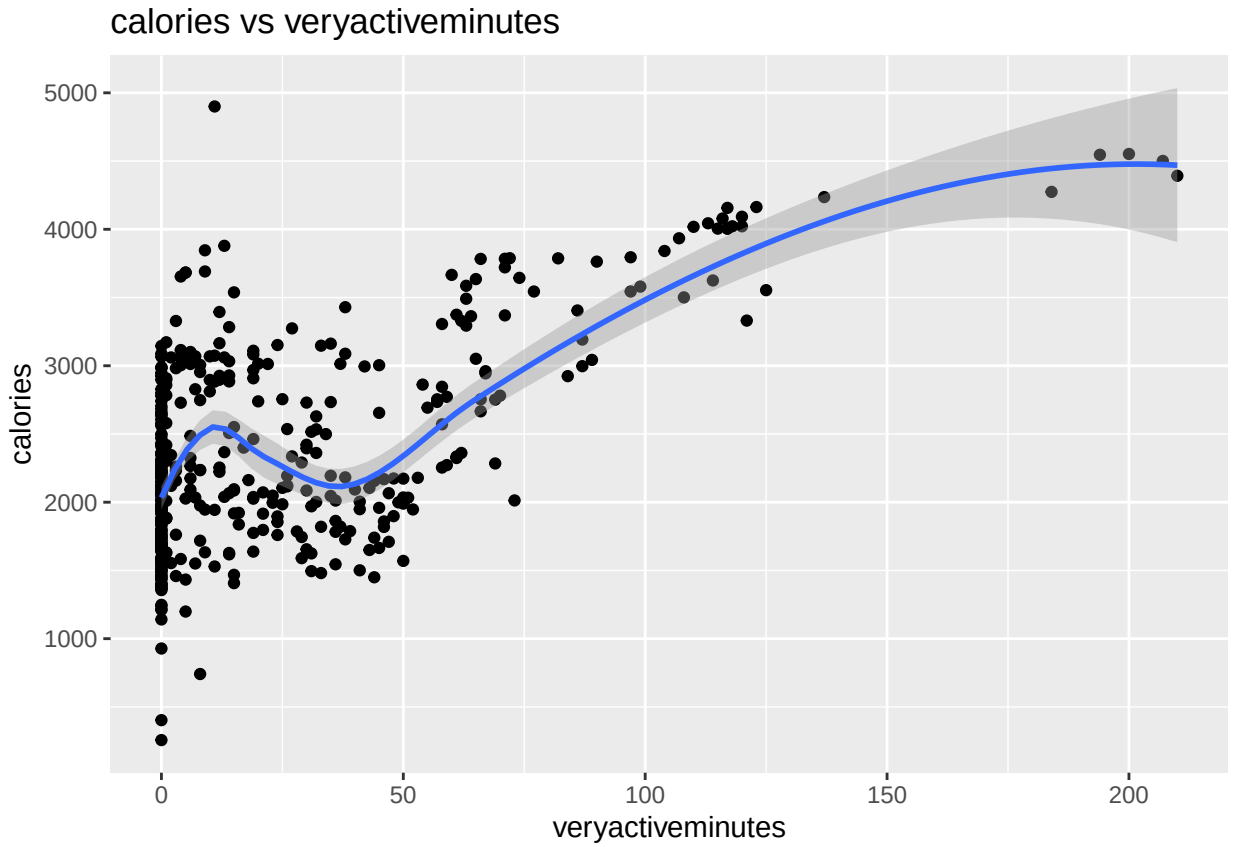
```
## Non-numeric variables removed from input: `activitydate`, `weekday`, `sleepday`, and `formathoursasl
## * Method: 'pearson'
## * Missing treated using: 'pairwise.complete.obs'
```

```
## # A tibble: 18 x 19
##    term             id total~1 total~2 track~3 logge~4 veryac~5 moder~6 lighta~7
##    <chr>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>   <dbl>    <dbl>
##  1 id         NA        0.0113  0.0784  0.0770  0.330   0.207   -0.0515 -0.0755
##  2 totalsteps  0.0113  NA       0.982   0.982   0.120   0.677    0.581   0.587
##  3 totaldist~  0.0784   0.982  NA       1.00    0.174   0.703    0.561   0.598
##  4 trackerdi~  0.0770   0.982   1.00   NA       0.162   0.702    0.560   0.599
##  5 loggedact~  0.330    0.120   0.174   0.162  NA       0.201    0.0269  0.0610
##  6 veryactiv~  0.207    0.677   0.703   0.702   0.201  NA        0.225  -0.0387
##  7 moderatel~ -0.0515   0.581   0.561   0.560   0.0269  0.225   NA       0.157
##  8 lightacti~ -0.0755   0.587   0.598   0.599   0.0610 -0.0387   0.157  NA
##  9 sedentary~  0.0334   0.0467  0.0590  0.0390  0.374   0.0525   0.0649  0.00186
## 10 veryactiv~  0.270    0.541   0.574   0.574   0.351   0.850    0.183  -0.0672
## 11 fairlyact~ -0.0214   0.570   0.550   0.550  -0.0238  0.286    0.945   0.0947
## 12 lightlyac~ -0.181    0.419   0.387   0.388  -0.0912 -0.169    0.0343  0.853
## 13 sedentary~ -0.00742 -0.132  -0.128  -0.128   0.0234  0.00655 -0.0422 -0.211
## 14 calories    0.403    0.412   0.528   0.529   0.323   0.440    0.0791  0.342
## 15 totalslee~ -0.0126  -0.162  -0.144  -0.144  -0.0473 -0.0948  -0.0641 -0.102
## 16 totalminu~  0.0801  -0.187  -0.172  -0.173  -0.0413 -0.105   -0.239  -0.0426
## 17 totaltime~  0.00211 -0.164  -0.158  -0.158  -0.0617 -0.113   -0.0950 -0.0915
## 18 totalhour~  0.0801  -0.187  -0.172  -0.173  -0.0413 -0.105   -0.239  -0.0425
## # ... with 10 more variables: sedentaryactivedistance <dbl>,
## #   veryactiveminutes <dbl>, fairlyactiveminutes <dbl>,
## #   lightlyactiveminutes <dbl>, sedentaryminutes <dbl>, calories <dbl>,
## #   totalsleeprecords <dbl>, totalminutesasleep <dbl>, totaltimeinbed <dbl>,
## #   totalhoursasleep <dbl>, and abbreviated variable names 1: totalsteps,
## #   2: totaldistance, 3: trackerdistance, 4: loggedactivitiesdistance,
## #   5: veryactivedistance, 6: moderatelyactivedistance, ...
```

#correlation between veryactiveminutes and calories

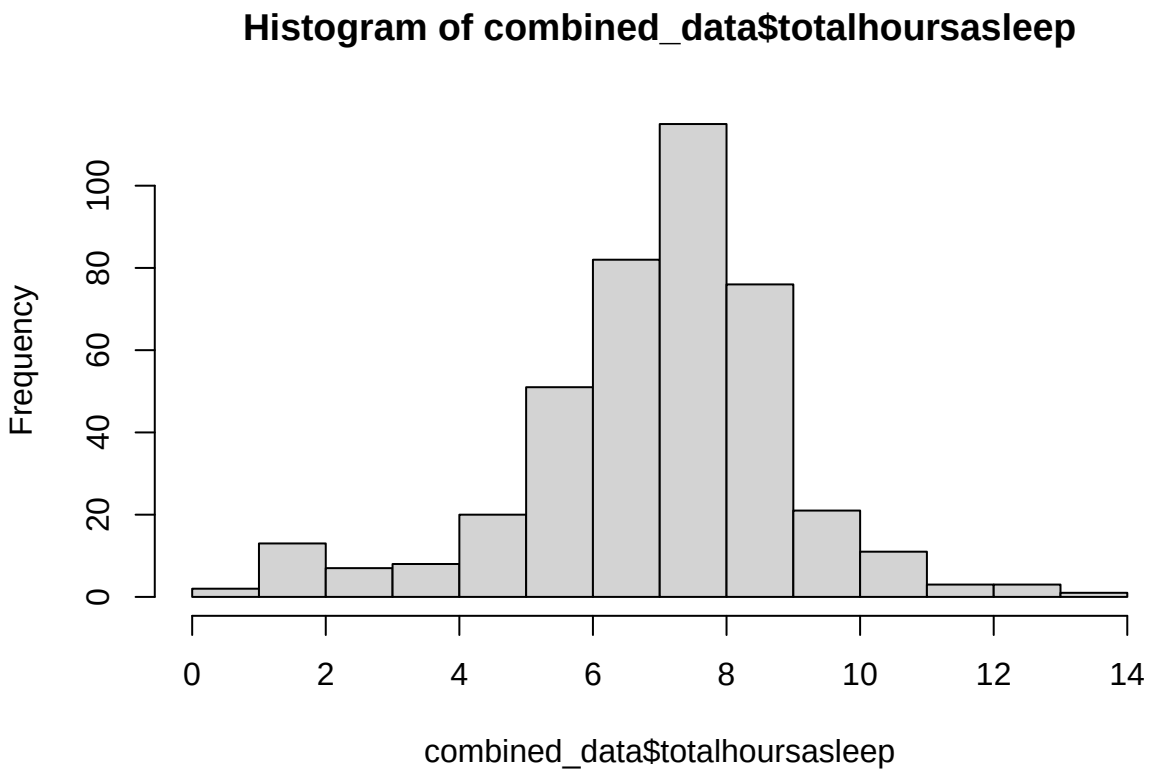
```
ggplot(data=combined_data,aes(x=veryactiveminutes,y=calories))+geom_point()+
  geom_smooth()+ggtitle("calories vs veryactiveminutes")
```

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



#totalhoursasleep

```
hist(x=combined_data$totalhoursasleep,breaks=10)
```



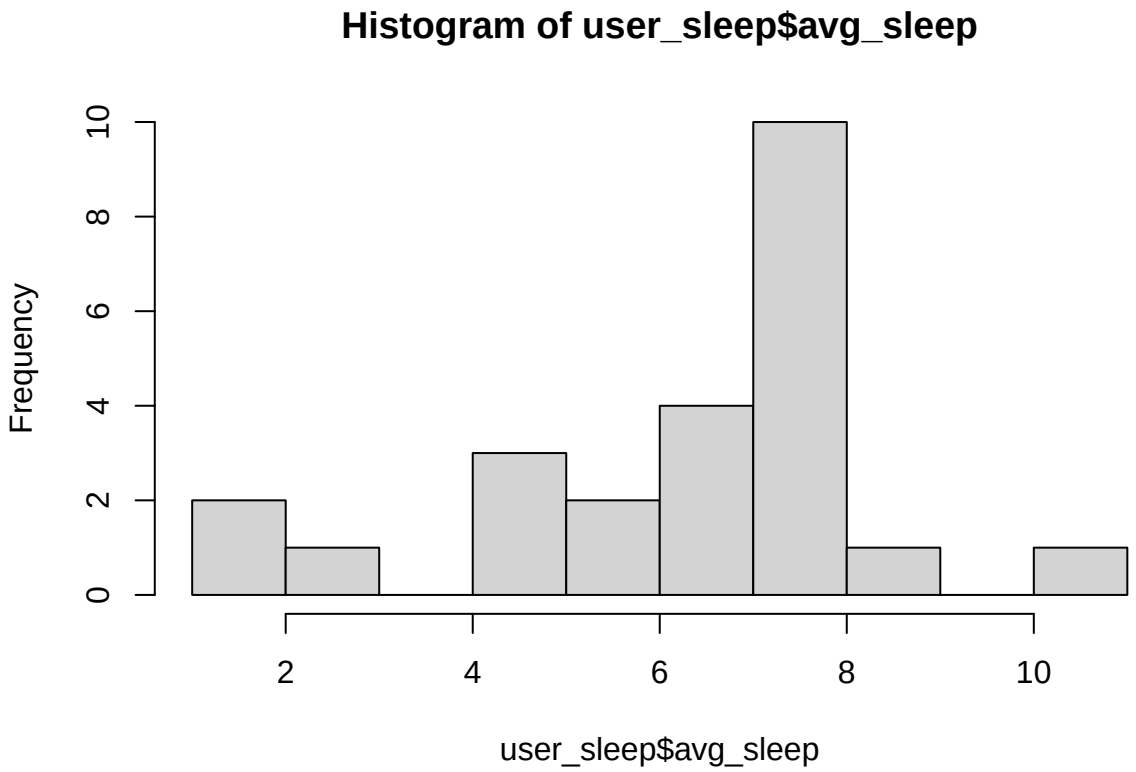
#finding average sleep of users

```
user_sleep<-combined_data%>%
  select(id,totalhoursasleep)%>%
  group_by(id)%>%
  summarize(avg_sleep=mean(totalhoursasleep))
head(user_sleep)
```

```
## # A tibble: 6 x 2
##           id avg_sleep
##        <dbl>     <dbl>
## 1 1503960366      6.01
## 2 1644430081      4.9
## 3 1844505072     10.9
## 4 1927972279      6.95
## 5 2026352035      8.44
## 6 2320127002      1.02
```

#average sleep data

```
hist(x=user_sleep$avg_sleep,breaks=10)
```



#average distance calculation

```
sleep_activity<-combined_data%>%
  select(id,totaldistance)%>%
  group_by(id)%>%
  summarize(avg_distance=mean(totaldistance))
head(sleep_activity)
```

```
## # A tibble: 6 x 2
##           id avg_distance
##        <dbl>        <dbl>
## 1 1503960366         7.97
## 2 1644430081         5.79
## 3 1844505072         2.30
## 4 1927972279         1.03
## 5 2026352035         3.49
## 6 2320127002         3.42
```

#merging avg_sleep and avg_distance by id

```
sleep_distance<-merge(user_sleep,sleep_activity,by="id")
head(sleep_distance)
```

```
##           id avg_sleep avg_distance
## 1 1503960366  6.005600     7.971200
## 2 1644430081  4.900000     5.792500
## 3 1844505072 10.863333     2.303333
## 4 1927972279  6.950000     1.032000
## 5 2026352035  8.436429     3.487143
## 6 2320127002  1.020000     3.420000
```

#finding correlation between avg_sleep and avg_distance

```
ggplot(sleep_distance,aes(avg_sleep,avg_distance))+geom_point()+geom_smooth()+
  ggtitle("avg_sleep vs avg_distance")
```

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