Bellabeat Customer Behaviour Analysis

2023-04-12

Scenario

You are a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. You have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy.

Project Task

Sršen asks you to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants you to select one Bellabeat product to apply these insights to in your presentation. Sršen encourages you to use public data that explores smart device users' daily habits. She points you to a specific data set. FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius) This Kaggle data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

Guiding Questions

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

Loading libraries

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0     ✓ purrr   1.0.1
## ✓ tibble  3.1.8     ✓ dplyr   1.1.0
## ✓ tidyr   1.3.0     ✓ stringr 1.5.0
## ✓ readr   2.1.3     ✓ forcats 1.0.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(lubridate)
```

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

```
library(ggplot2)
```

Data importation

I used daily activity and daily sleep tables. This tables have information concerning activity and sleep data which are also collected by bellabeat products.

```
Activity_df <- read.csv("C:\\Users\\gbless7\\Downloads\\fitbit fitness data\\Fitabase Data 4.12.16-5.12.16\\dailyActivity_merged.csv")

Sleep_df <- read.csv("C:\\Users\\gbless7\\Downloads\\fitbit fitness data\\Fitabase Data 4.12.16-5.12.16\\sleepDay_merged.csv")

Weight_df <- read.csv("C:\\Users\\gbless7\\Downloads\\fitbit fitness data\\Fitabase Data 4.12.16-5.12.16\\weightLogInfo_merged.csv")
```

```
head(Activity_df)
```

```
##           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(Sleep_df)
```

```
##           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
```

```
head(Weight_df)
```

```
##           Id                  Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

Data Cleaning

Activity Table

number of columns

```
colnames(Activity_df)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

Data structure

```
str(Activity_df)
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819
...
##  $ TotalDistance            : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

Activity date has character data type instead of date type

Detecting Duplicates

```
duplicates <- Activity_df[duplicated(Activity_df), ]
duplicates
```

```
##  [1] Id                       ActivityDate             TotalSteps
##  [4] TotalDistance            TrackerDistance          LoggedActivitiesDistance
##  [7] VeryActiveDistance       ModeratelyActiveDistance LightActiveDistance
## [10] SedentaryActiveDistance  VeryActiveMinutes        FairlyActiveMinutes
## [13] LightlyActiveMinutes     SedentaryMinutes         Calories
## <0 rows> (or 0-length row.names)
```

There are no duplicates

Distinct users

```
Distinct_users <- Activity_df %>%
  distinct(Id)
count(Distinct_users)
```

```
##    n
## 1 33
```

there are 33 users for the product recording activity data

Sleep Table

number of columns

```
colnames(Sleep_df)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

table structure

```
str(Sleep_df)
```

```
## 'data.frame':    413 obs. of  5 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:
00 AM" "4/16/2016 12:00:00 AM" ...
##  $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

sleep day variable has character data type instead of date type

Detecting duplicates

```
duplicates <- Sleep_df[duplicated(Sleep_df), ]
duplicates
```

```
##             Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 162 4388161847  5/5/2016 12:00:00 AM                 1                471
## 224 4702921684  5/7/2016 12:00:00 AM                 1                520
## 381 8378563200 4/25/2016 12:00:00 AM                 1                388
##     TotalTimeInBed
## 162            495
## 224            543
## 381            402
```

There are 3 duplicates

Weight Data

number of columns

```
colnames(Weight_df)
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

Data structure

```
str(Weight_df)
```

```
## 'data.frame':    67 obs. of  8 variables:
##  $ Id            : num  1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ Date          : chr  "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM"
"4/21/2016 11:59:59 PM" ...
##  $ WeightKg      : num  52.6 52.6 133.5 56.7 57.3 ...
##  $ WeightPounds  : num  116 116 294 125 126 ...
##  $ Fat           : int  22 NA NA NA NA 25 NA NA NA NA ...
##  $ BMI           : num  22.6 22.6 47.5 21.5 21.7 ...
##  $ IsManualReport: chr  "True" "True" "False" "True" ...
##  $ LogId         : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

Date field has character data type instead of date type

Detecting duplicates

```
duplicates_wt <- Weight_df[duplicated(Weight_df),]
duplicates_wt
```

```
## [1] Id             Date           WeightKg       WeightPounds   Fat
## [6] BMI            IsManualReport LogId
## <0 rows> (or 0-length row.names)
```

there are no duplicates

Changing activity date data type to date type and renaming the column

```
Date_1 <- as.Date(Activity_df$ActivityDate, format = "%m/%d/%y")

Activity <- Activity_df %>%
  rename(Date=ActivityDate) %>%
  mutate(Date=Date_1)
head(Activity)
```

```
##           Id       Date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2020-04-12      13162          8.50            8.50
## 2 1503960366 2020-04-13      10735          6.97            6.97
## 3 1503960366 2020-04-14      10460          6.74            6.74
## 4 1503960366 2020-04-15       9762          6.28            6.28
## 5 1503960366 2020-04-16      12669          8.16            8.16
## 6 1503960366 2020-04-17       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
```

Change sleep date data type and renaming sleep day

```
Date_1 <- as.Date(Sleep_df$SleepDay, format = "%m/%d/%y")

Sleep_1 <- Sleep_df %>%
  rename(Date=SleepDay) %>%
  mutate(Date=Date_1)
head(Sleep_1)
```

```
##           Id       Date TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 2020-04-12                 1                327            346
## 2 1503960366 2020-04-13                 2                384            407
## 3 1503960366 2020-04-15                 1                412            442
## 4 1503960366 2020-04-16                 2                340            367
## 5 1503960366 2020-04-17                 1                700            712
## 6 1503960366 2020-04-19                 1                304            320
```

changing weight date type

```
Date_wt <- as.Date(Weight_df$Date, format = "%m/%d/%y")

Wt_df <- Weight_df %>%
  mutate(Date=Date_wt)
head(Wt_df)
```

```
##           Id       Date WeightKg WeightPounds Fat   BMI IsManualReport
## 1 1503960366 2020-05-02     52.6     115.9631  22 22.65           True
## 2 1503960366 2020-05-03     52.6     115.9631  NA 22.65           True
## 3 1927972279 2020-04-13    133.5     294.3171  NA 47.54          False
## 4 2873212765 2020-04-21     56.7     125.0021  NA 21.45           True
## 5 2873212765 2020-05-12     57.3     126.3249  NA 21.69           True
## 6 4319703577 2020-04-17     72.4     159.6147  25 27.45           True
##          LogId
## 1 1.462234e+12
## 2 1.462320e+12
## 3 1.460510e+12
## 4 1.461283e+12
## 5 1.463098e+12
## 6 1.460938e+12
```

Removing duplicates

```
sleep_unique <- distinct(Sleep_1)
head(sleep_unique)
```

```
##           Id       Date TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 2020-04-12                 1                327            346
## 2 1503960366 2020-04-13                 2                384            407
## 3 1503960366 2020-04-15                 1                412            442
## 4 1503960366 2020-04-16                 2                340            367
## 5 1503960366 2020-04-17                 1                700            712
## 6 1503960366 2020-04-19                 1                304            320
```

Data Summarization

```
summary(Activity_df)
```

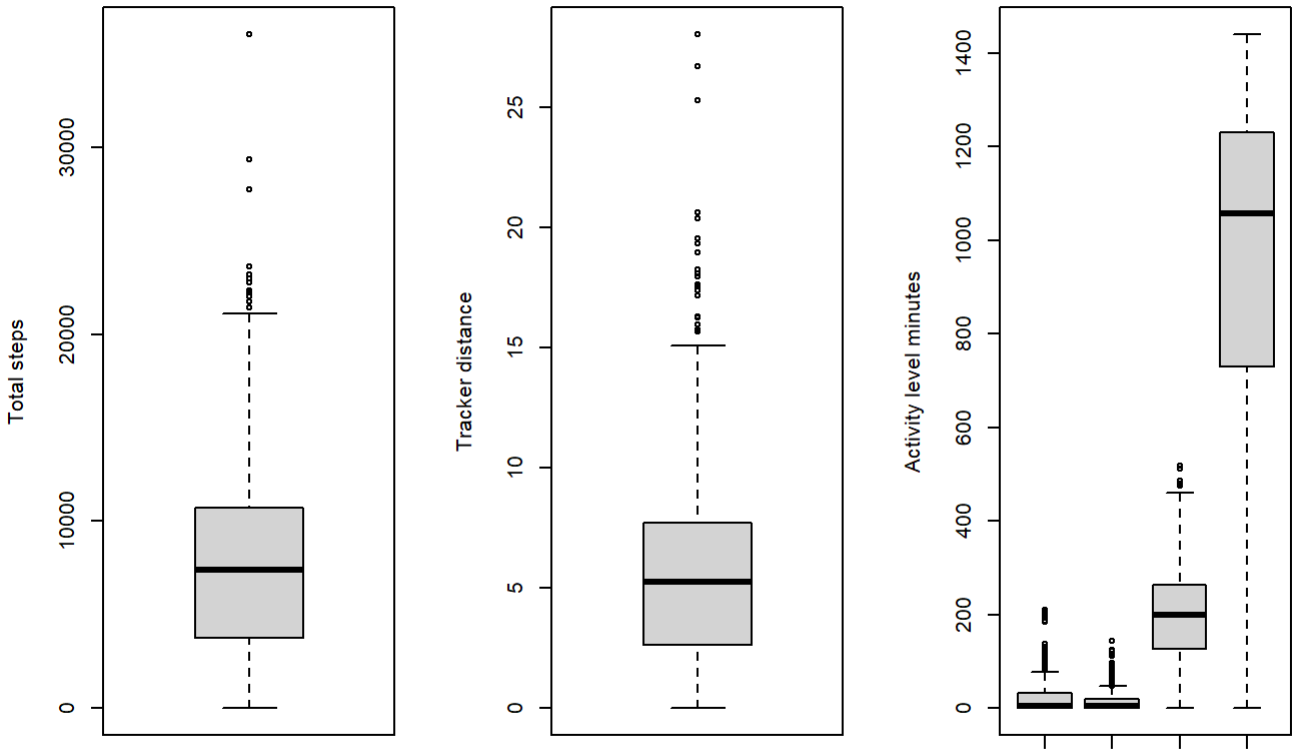
```
##        Id            ActivityDate         TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:940         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Mode  :character   Median : 7406   Median : 5.245
##  Mean   :4.855e+09                      Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09                      3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09                      Max.   :36019   Max.   :28.030
##  TrackerDistance  LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.000   Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 2.620   1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 5.245   Median :0.0000           Median : 0.210
##  Mean   : 5.475   Mean   :0.1082           Mean   : 1.503
##  3rd Qu.: 7.710   3rd Qu.:0.0000           3rd Qu.: 2.053
##  Max.   :28.030   Max.   :4.9421           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
##  Min.   :   0
##  1st Qu.:1828
##  Median :2134
##  Mean   :2304
##  3rd Qu.:2793
##  Max.   :4900
```

box plot to visualize distribution of total steps, tracker distance, and minutes activity levels

```
par(mfrow=c(1,3))
boxplot(Activity_df$TotalSteps, ylab='Total steps')

boxplot(Activity_df$TrackerDistance, ylab='Tracker distance')

boxplot(Activity_df$VeryActiveMinutes,Activity_df$FairlyActiveMinutes,Activity_df$LightlyActiveM
inutes,Activity_df$SedentaryMinutes, ylab='Activity level minutes')
```



Total steps has outliers and data is skewed right, meaning many users take steps below 10,000 per day

Tracker distance has outliers and data is skewed right, meaning many users walk short distances

There are few users that have active distance in very active minutes and fairly active minutes because of small range compared to users in lightly active minutes and sedentary active minutes

Individuals in sedentary active minutes have skewed to left meaning many individuals are less active

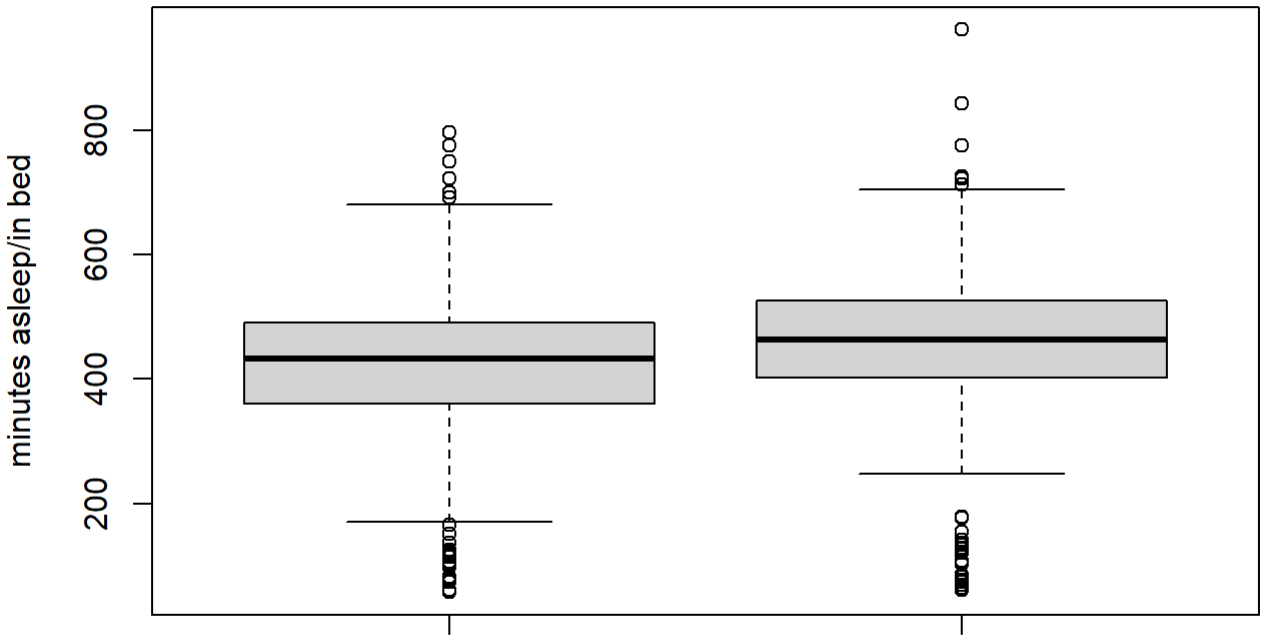
distribution of sleep data

```
summary(Sleep_df)
```

```
##        Id              SleepDay         TotalSleepRecords TotalMinutesAsleep
##  Min.   :1.504e+09   Length:413         Min.   :1.000     Min.   : 58.0
##  1st Qu.:3.977e+09   Class :character   1st Qu.:1.000     1st Qu.:361.0
##  Median :4.703e+09   Mode  :character   Median :1.000     Median :433.0
##  Mean   :5.001e+09                      Mean   :1.119     Mean   :419.5
##  3rd Qu.:6.962e+09                      3rd Qu.:1.000     3rd Qu.:490.0
##  Max.   :8.792e+09                      Max.   :3.000     Max.   :796.0
##  TotalTimeInBed
##  Min.   : 61.0
##  1st Qu.:403.0
##  Median :463.0
##  Mean   :458.6
##  3rd Qu.:526.0
##  Max.   :961.0
```

Box plot to visualize distribution of total minutes asleep and total time in bed

```
par(mfrow=c(1,1))
boxplot(Sleep_df$TotalMinutesAsleep,Sleep_df$TotalTimeInBed,ylab='minutes asleep/in bed')
```



The distribution of time spent asleep is normal but there are outliers

Time spent in bed also has normal distribution but with outliers

The time spent by users in bed is approximately the same with those spending asleep

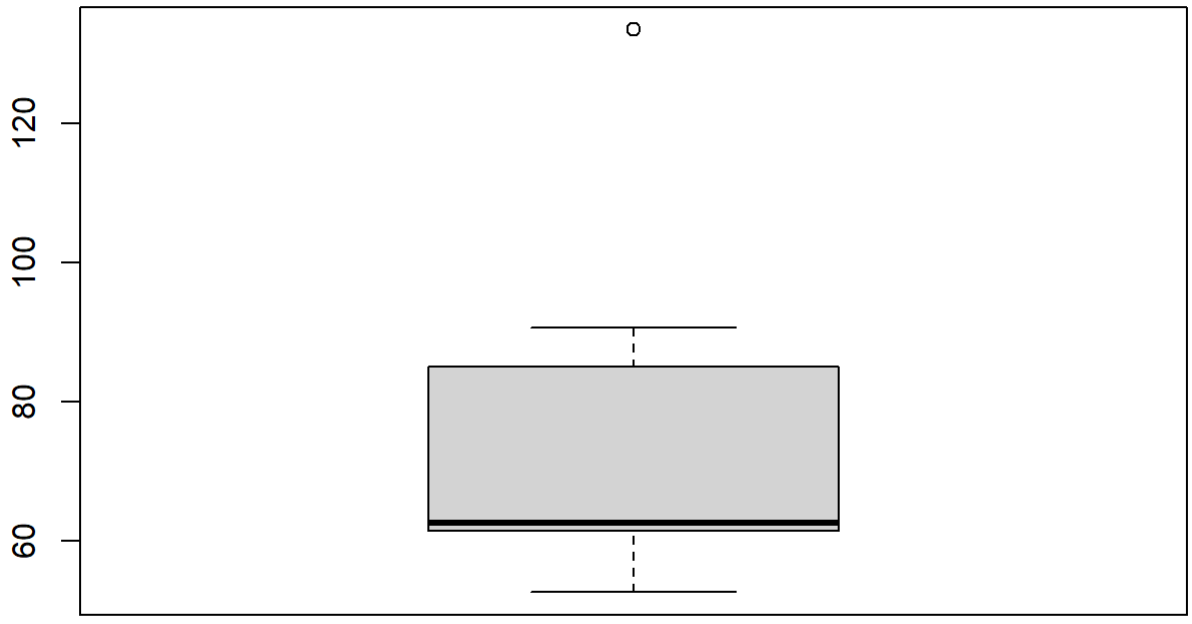
Distribution of selected weight fields

```
Wt_df %>%
  select(WeightKg) %>%
  summary()
```

```
##     WeightKg
##  Min.   : 52.60
##  1st Qu.: 61.40
##  Median : 62.50
##  Mean   : 72.04
##  3rd Qu.: 85.05
##  Max.   :133.50
```

box plot to visualize distribution

```
par(mfrow=c(1,1))
boxplot(Wt_df$WeightKg, ylan="weight")
```



there is an outlier, it will be deleted

```
wt <- subset(Wt_df, Id != "1927972279")
head(wt)
```

```
##           Id       Date WeightKg WeightPounds Fat   BMI IsManualReport
## 1 1503960366 2020-05-02     52.6     115.9631  22 22.65           True
## 2 1503960366 2020-05-03     52.6     115.9631  NA 22.65           True
## 4 2873212765 2020-04-21     56.7     125.0021  NA 21.45           True
## 5 2873212765 2020-05-12     57.3     126.3249  NA 21.69           True
## 6 4319703577 2020-04-17     72.4     159.6147  25 27.45           True
## 7 4319703577 2020-05-04     72.3     159.3942  NA 27.38           True
##          LogId
## 1 1.462234e+12
## 2 1.462320e+12
## 4 1.461283e+12
## 5 1.463098e+12
## 6 1.460938e+12
## 7 1.462406e+12
```

recheck the outlier

```
wt %>%
  select(WeightKg) %>%
  summary()
```

```
##     WeightKg
##  Min.   :52.60
##  1st Qu.:61.40
##  Median :62.45
##  Mean   :71.10
##  3rd Qu.:84.97
##  Max.   :90.70
```

there is large variation in weights of individuals

Data transformation

creating new activity table with summarized variables. for example customer Id '1503960366' took 12207 steps in each day

```
activity_new <- Activity %>%
  group_by(Id) %>%
  summarize(Days=as.numeric(max(Date)-min(Date)),Steps=median(TotalSteps),Distance=median(Tracke
rDistance),ActiveDistance=median(VeryActiveDistance),InactiveDistance=mean(SedentaryActiveDistan
ce),ActiveMinutes=median(VeryActiveMinutes),InactiveMinutes=median(SedentaryMinutes),BurnedCalor
ies=median(Calories)
)
head(activity_new)
```

```
## # A tibble: 6 × 9
##           Id  Days  Steps Distance ActiveDista…¹ Inact…² Activ…³ Inact…⁴ Burne…⁵
##        <dbl> <dbl>  <dbl>    <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1503960366    30 12207     8.03          2.81  0            38     798   1837
## 2 1624580081    30  4026     2.62          0     0.00613       0    1288   1435
## 3 1644430081    29  6684.    4.86          0.300 0.00400       4    1179   2802.
## 4 1844505072    30  2237     1.48          0     0             0    1301   1549
## 5 1927972279    30   152     0.110         0     0             0    1413   2100
## 6 2022484408    30 11548     8.29          2.51  0            36    1112   2529
## # … with abbreviated variable names ¹ActiveDistance, ²InactiveDistance,
## #   ³ActiveMinutes, ⁴InactiveMinutes, ⁵BurnedCalories
```

also creating new sleep table with summarized variables

```
sleep_new <- sleep_unique %>%
  group_by(Id) %>%
  summarize(Days=as.numeric(max(Date)-min(Date)),No_of_Sleeps=median(TotalSleepRecords), Minutes
Asleep=median(TotalMinutesAsleep),MinutesInBed=median(TotalTimeInBed))
head(sleep_new)
```

```
## # A tibble: 6 × 5
##           Id  Days No_of_Sleeps MinutesAsleep MinutesInBed
##        <dbl> <dbl>        <dbl>         <dbl>        <dbl>
## 1 1503960366    29            1           340          367
## 2 1644430081     9            1           130.         148
## 3 1844505072    16            1           644          961
## 4 1927972279    16            1           398          422
## 5 2026352035    30            1           516.         546.
## 6 2320127002     0            1            61           69
```

Joining sleep table and activity table

```
activity_sleep <- inner_join(sleep_new,activity_new,by="Id")
head(activity_sleep)
```

```
## # A tibble: 6 × 13
##           Id Days.x No_of…¹ Minut…² Minut…³ Days.y  Steps Dista…⁴ Activ…⁵ Inact…⁶
##        <dbl>  <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1    1.50e9     29       1     340     367      30 12207    8.03    2.81  0
## 2    1.64e9      9       1     130.    148      29  6684.   4.86    0.300 0.00400
## 3    1.84e9     16       1     644     961      30  2237    1.48    0     0
## 4    1.93e9     16       1     398     422      30   152    0.110   0     0
## 5    2.03e9     30       1     516.    546.     30  5528    3.45    0     0
## 6    2.32e9      0       1      61      69      30  5057    3.41    0     0
## # … with 3 more variables: ActiveMinutes <dbl>, InactiveMinutes <dbl>,
## #   BurnedCalories <dbl>, and abbreviated variable names ¹No_of_Sleeps,
## #   ²MinutesAsleep, ³MinutesInBed, ⁴Distance, ⁵ActiveDistance,
## #   ⁶InactiveDistance
```

joining weight table and activity table

```
wt_ac <- merge(wt, Activity, by=c("Id","Date"))
head(wt_ac)
```

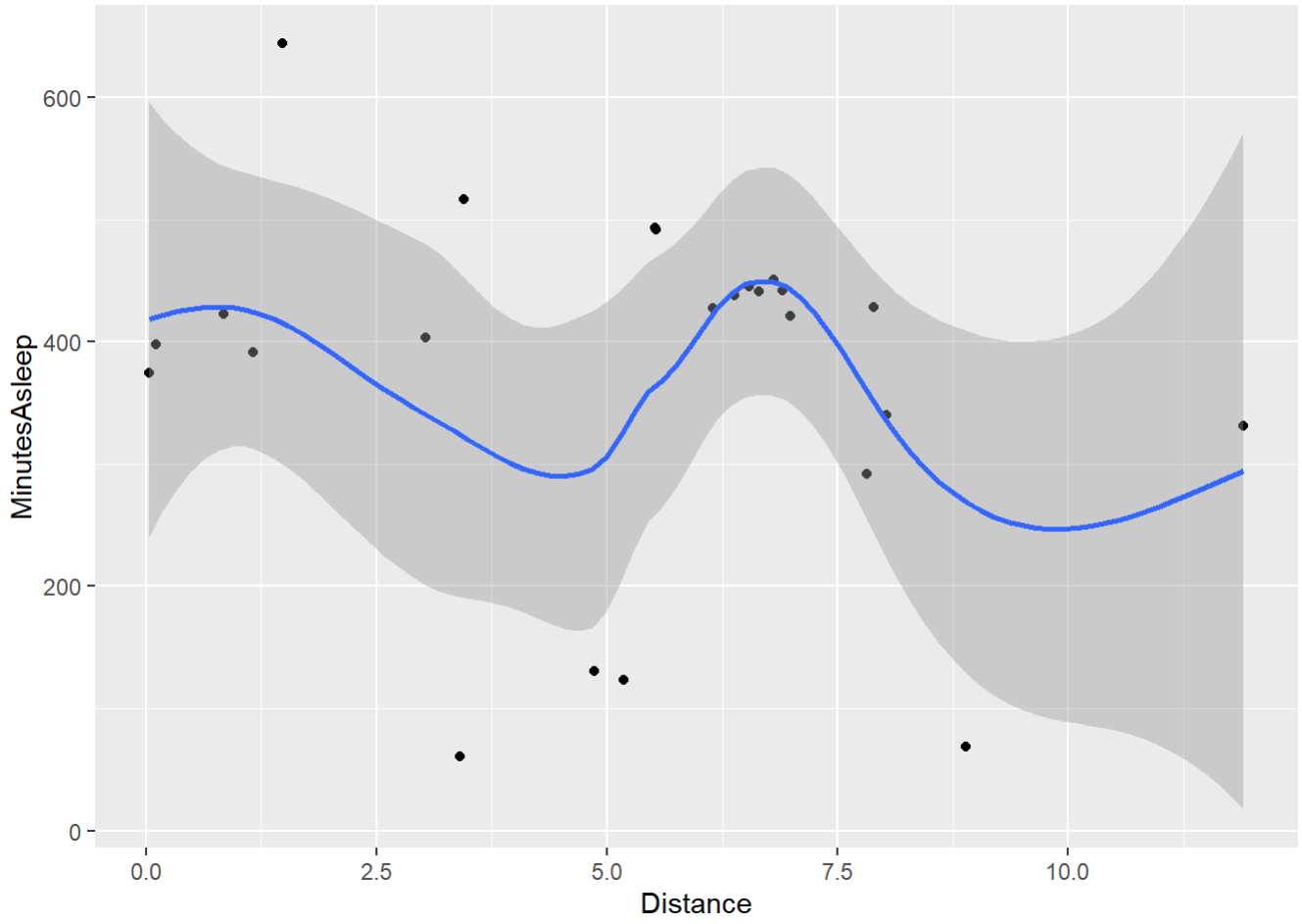
```
##           Id       Date WeightKg WeightPounds Fat   BMI IsManualReport
## 1 1503960366 2020-05-02     52.6     115.9631  22 22.65           True
## 2 1503960366 2020-05-03     52.6     115.9631  NA 22.65           True
## 3 2873212765 2020-04-21     56.7     125.0021  NA 21.45           True
## 4 2873212765 2020-05-12     57.3     126.3249  NA 21.69           True
## 5 4319703577 2020-04-17     72.4     159.6147  25 27.45           True
## 6 4319703577 2020-05-04     72.3     159.3942  NA 27.38           True
##          LogId TotalSteps TotalDistance TrackerDistance
## 1 1.462234e+12      14727          9.71            9.71
## 2 1.462320e+12      15103          9.66            9.66
## 3 1.461283e+12       8859          5.98            5.98
## 4 1.463098e+12       7566          5.11            5.11
## 5 1.460938e+12         29          0.02            0.02
## 6 1.462406e+12      10429          7.02            7.02
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              3.21                     0.57
## 2                        0              3.73                     1.05
## 3                        0              0.13                     0.37
## 4                        0              0.00                     0.00
## 5                        0              0.00                     0.00
## 6                        0              0.59                     0.58
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                5.92                    0.00                41
## 2                4.88                    0.00                50
## 3                5.47                    0.01                 2
## 4                5.11                    0.00                 0
## 5                0.02                    0.00                 0
## 6                5.85                    0.00                 8
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  15                  277              798     2004
## 2                  24                  254              816     1990
## 3                  10                  371             1057     1970
## 4                   0                  268              720     1431
## 5                   0                    3             1363     1464
## 6                  13                  313             1106     2282
```

Data Visualization

Correlation between distance covered and time spent during sleeping

```
ggplot(data = activity_sleep, aes(x=Distance, y=MinutesAsleep))+geom_point()+geom_smooth(na.rm = TRUE)
```

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



correlation coefficient between distance and time spent sleeping

```
coeff <- cor(x=activity_sleep$Distance, y=activity_sleep$MinutesAsleep)
coeff
```

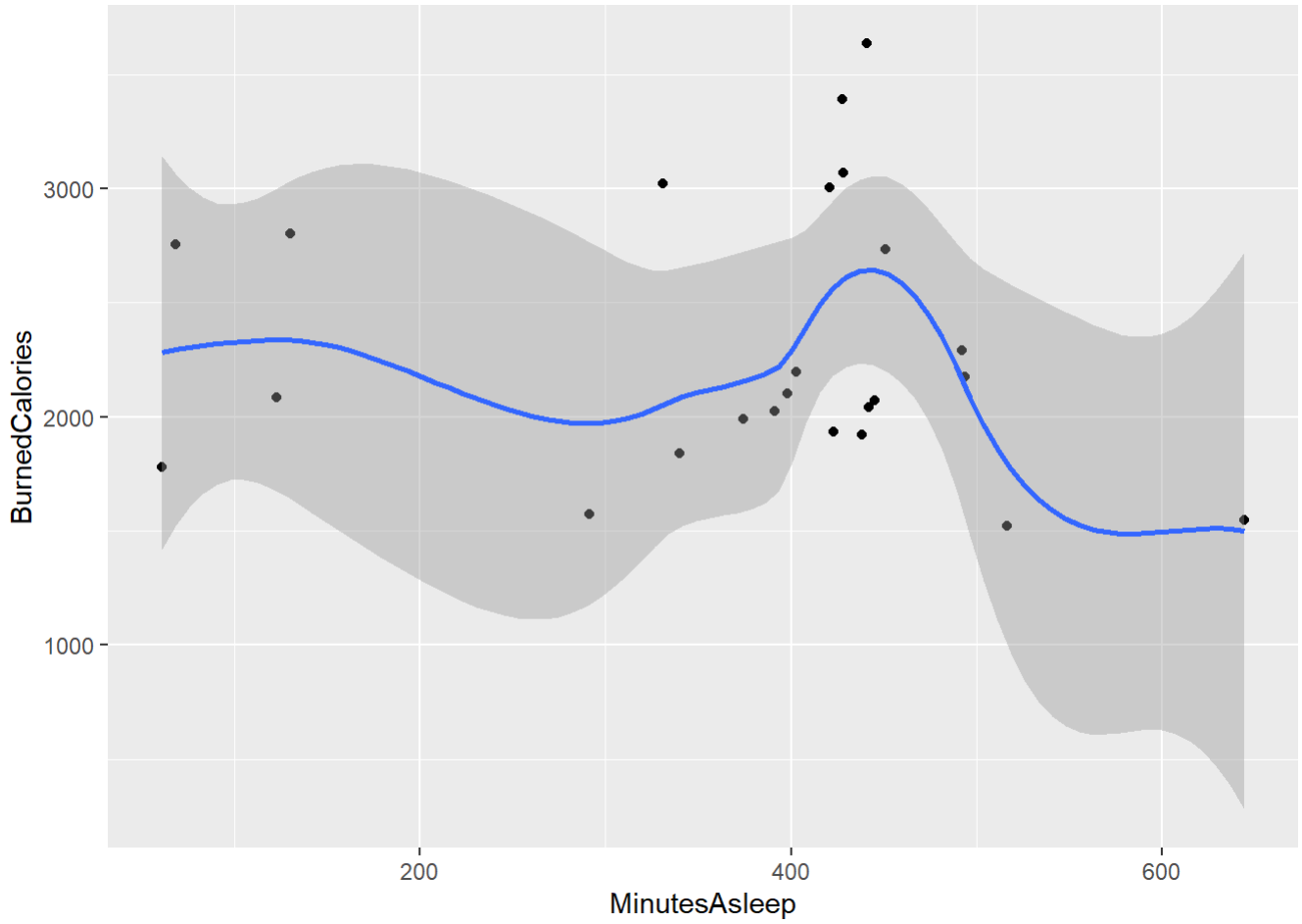
```
## [1] -0.1934977
```

There is negative correlation, that means the less steps are taken the more time is spent sleeping

correlation between burnt calories and time spent sleeping

```
ggplot(data = activity_sleep, aes(x=MinutesAsleep, y=BurnedCalories))+geom_point()+geom_smooth(n
a.rm = TRUE)
```

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



correlation coefficient between burnt calories and time spent sleeping

```
coeff_ <- cor(x=activity_sleep$MinutesAsleep, y=activity_sleep$BurnedCalories)
coeff_
```

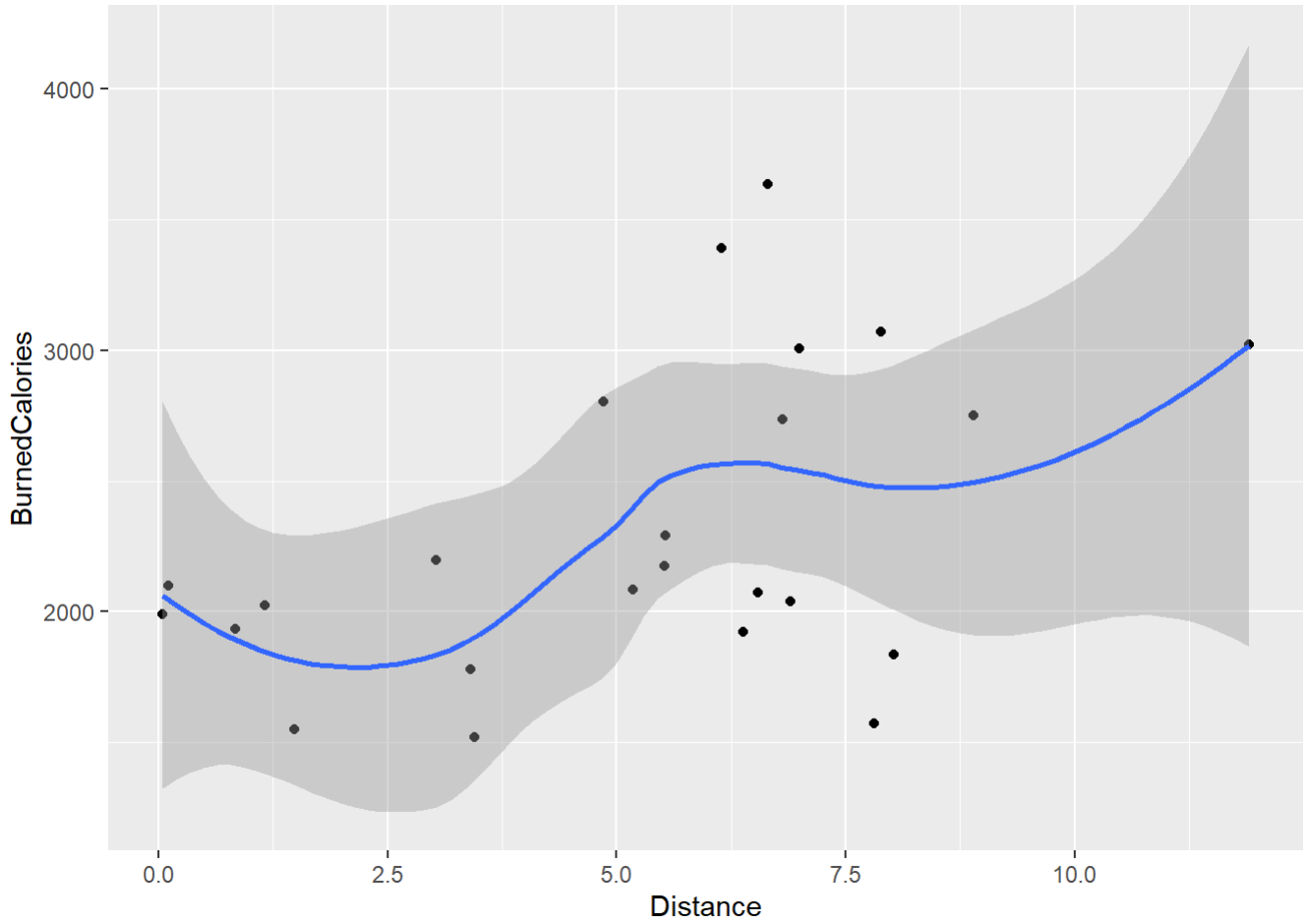
```
## [1] -0.08187504
```

There is negative correlation, meaning the more time is spent sleeping the less calories are burnt

correlation between distance covered and calories burnt

```
ggplot(data = activity_sleep, aes(x=Distance, y=BurnedCalories))+geom_point()+geom_smooth(na.rm
= TRUE)
```

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



correlation coefficient

```
coeff_i <- cor(x=activity_sleep$Distance, y=activity_sleep$BurnedCalories)
coeff_i
```

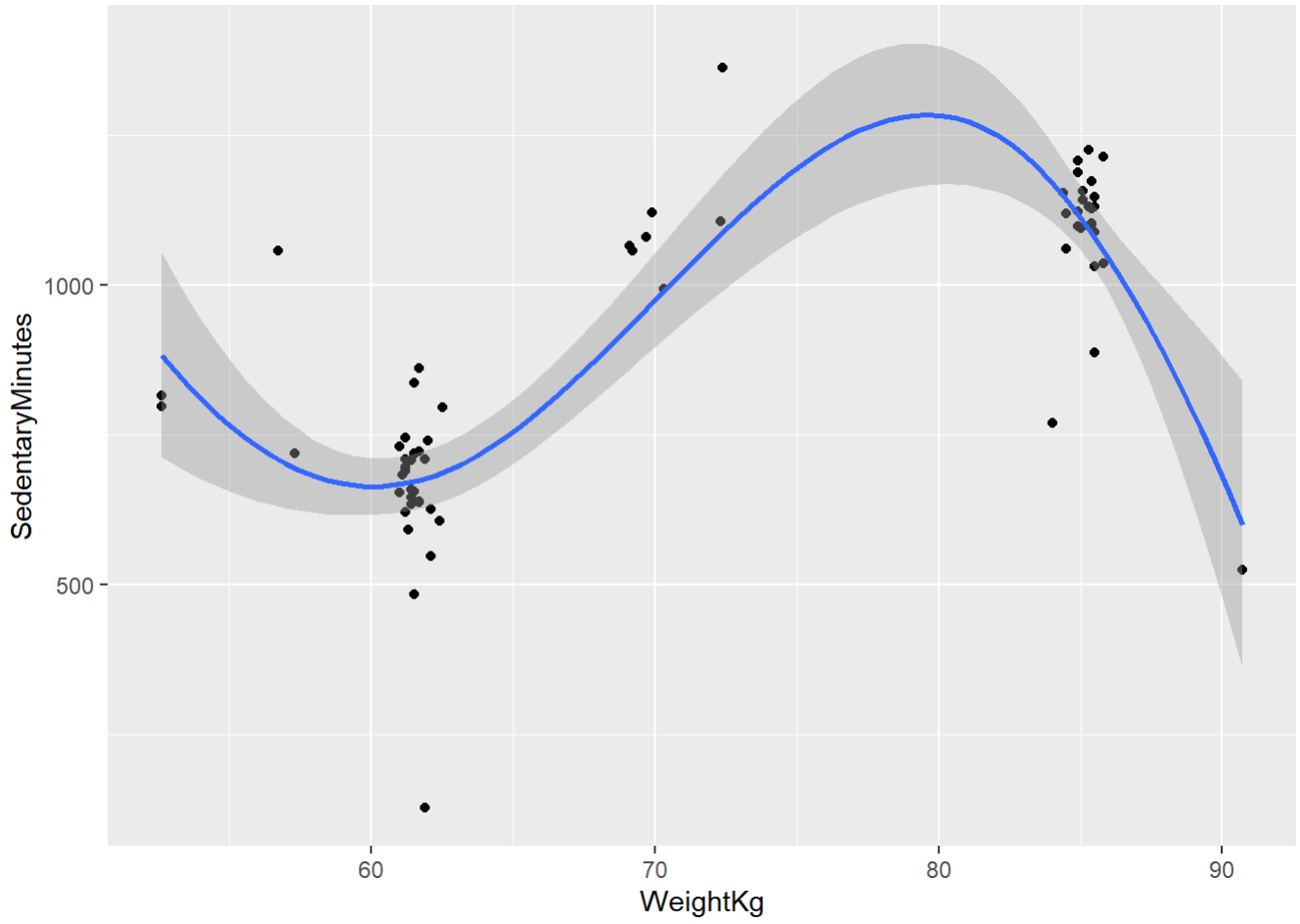
```
## [1] 0.4756133
```

There is positive correlation, that means the more distance is covered, the more calories are burnt

correlation between weight and sedentary minutes

```
ggplot(data = wt_ac, aes(x=WeightKg, y=SedentaryMinutes))+geom_point()+geom_smooth()
```

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



coefficient between weight and sedentary minutes

```
cor_wt <- cor(wt_ac$WeightKg,wt_ac$SedentaryMinutes)
cor_wt
```

```
## [1] 0.7046599
```

there is strong positive correlation between weight and sedentary minutes

Trends and Patterns

- Many users walk short distances that are not intense
- Many users spend most of their time sleeping and in bed
- Customers who spend more time sleeping burn less calories
- Customers covering short distances tend to sleep more
- Customers Covering long distances tend to burn more calories
- Customers who have large weight tend to spend more time sitting/lying while engaging in activities

Recommendation

- The marketing team in bellabeat should come up with a campaign and tell customers the effects being inactive
- In monitoring their body weight they should use bellabeat products