FitBaseDataSetAnalysis

Sajith

2022-11-18

So,as we said in the intoduction, we are basically analysing the data given to us by the Bellabeat company 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.

DATA FEATURE ENGINEERING

Lets Import the required libraries

```
setwd("D:/GOOGLE(DA)Coursera/Capstone Coursera DA/Case Study/Case Study 2/Fitabase Data 4.12.16-5.12.16")
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

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

```
library(ggplot2)
library(readr)
```

We are only importing only the required files for the analysis as there is so much data that is unneccessary

```
daily_activity <- read.csv("dailyActivity_merged.csv")
daily_calories <- read.csv('dailyCalories_merged.csv')
daily_intensities <- read.csv('dailyIntensities_merged.csv')
daily_steps <- read.csv('dailySteps_merged.csv')
hourly_calories <- read.csv('hourlyCalories_merged.csv')
hourly_intensities <- read.csv('hourlyIntensities_merged.csv')
hourly_steps <- read.csv('hourlySteps_merged.csv') daily_sleep
<- read.csv('sleepDay_merged.csv') weight_log <-
read.csv('weightLogInfo_merged.csv')
```

Reviewing the dataframes

```
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
```

```
head(daily_calories)
```

```
##            Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

```
head(daily_intensities)
```

```
##            Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 1 1503960366   4/12/2016              728                  328
## 2 1503960366   4/13/2016              776                  217
## 3 1503960366   4/14/2016             1218                  181
## 4 1503960366   4/15/2016              726                  209
## 5 1503960366   4/16/2016              773                  221
## 6 1503960366   4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

```
head(daily_sleep)
```

```
##           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(daily_steps)
```

```
##            Id ActivityDay StepTotal
## 1 1503960366   4/12/2016     13162
## 2 1503960366   4/13/2016     10735
## 3 1503960366   4/14/2016     10460
## 4 1503960366   4/15/2016      9762
## 5 1503960366   4/16/2016     12669
## 6 1503960366   4/17/2016      9705
```

```
head(hourly_intensities)
```

```
##            Id          ActivityHour TotalIntensity AverageIntensity
## 1 1503960366 4/12/2016 12:00:00 AM             20         0.333333
## 2 1503960366  4/12/2016 1:00:00 AM              8         0.133333
## 3 1503960366  4/12/2016 2:00:00 AM              7         0.116667
## 4 1503960366  4/12/2016 3:00:00 AM              0         0.000000
## 5 1503960366  4/12/2016 4:00:00 AM              0         0.000000
## 6 1503960366  4/12/2016 5:00:00 AM              0         0.000000
```

```
head(hourly_steps)
```

```
##            Id          ActivityHour StepTotal
## 1 1503960366 4/12/2016 12:00:00 AM       373
## 2 1503960366  4/12/2016 1:00:00 AM       160
## 3 1503960366  4/12/2016 2:00:00 AM       151
## 4 1503960366  4/12/2016 3:00:00 AM         0
## 5 1503960366  4/12/2016 4:00:00 AM         0
## 6 1503960366  4/12/2016 5:00:00 AM         0
```

```
head(hourly_calories)
```

```
##            Id          ActivityHour Calories
## 1 1503960366 4/12/2016 12:00:00 AM       81
## 2 1503960366  4/12/2016 1:00:00 AM       61
## 3 1503960366  4/12/2016 2:00:00 AM       59
## 4 1503960366  4/12/2016 3:00:00 AM       47
## 5 1503960366  4/12/2016 4:00:00 AM       48
## 6 1503960366  4/12/2016 5:00:00 AM       48
```

```
head(weight_log)
```

```
##           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
```

Using the glimpse and summary to shows the data structures and statistical summary

```
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…
```

```
summary(daily_activity)
```

```
##        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
```

```
glimpse(daily_calories)
```

```
## Rows: 940
## Columns: 3
## $ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366…
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/…
## $ Calories    <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775…
```

```
summary(daily_calories)
```

```
##        Id             ActivityDay           Calories
##  Min.   :1.504e+09   Length:940         Min.   :   0
##  1st Qu.:2.320e+09   Class :character   1st Qu.:1828  
##  Median :4.445e+09   Mode  :character   Median :2134  
##  Mean   :4.855e+09                      Mean   :2304  
##  3rd Qu.:6.962e+09                      3rd Qu.:2793  
##  Max.   :8.878e+09                      Max.   :4900  
```

```
glimpse(daily_intensities)
```

```
## Rows: 940
## Columns: 10
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityDay              <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, …
## $ FairlyActiveMinutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…
```

```
summary(daily_intensities)
```

```
##        Id             ActivityDay        SedentaryMinutes LightlyActiveMinutes
##  Min.   :1.504e+09   Length:940         Min.   :   0.0   Min.   :  0.0  
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 729.8   1st Qu.:127.0  
##  Median :4.445e+09   Mode  :character   Median :1057.5   Median :199.0  
##  Mean   :4.855e+09                      Mean   : 991.2   Mean   :192.8  
##  3rd Qu.:6.962e+09                      3rd Qu.:1229.5   3rd Qu.:264.0         ##
Max.   :8.878e+09                      Max.   :1440.0   Max.   :518.0  
##  FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
##  Min.   :  0.00      Min.   :  0.00    Min.   :0.000000       
##  1st Qu.:  0.00      1st Qu.:  0.00    1st Qu.:0.000000       
##  Median :  6.00      Median :  4.00    Median :0.000000       
##  Mean   : 13.56      Mean   : 21.16    Mean   :0.001606       
##  3rd Qu.: 19.00      3rd Qu.: 32.00    3rd Qu.:0.000000       
##  Max.   :143.00      Max.   :210.00    Max.   :0.110000       
##  LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
##  Min.   : 0.000      Min.   :0.0000           Min.   : 0.000    
##  1st Qu.: 1.945      1st Qu.:0.0000           1st Qu.: 0.000    
##  Median : 3.365      Median :0.2400           Median : 0.210    
##  Mean   : 3.341      Mean   :0.5675           Mean   : 1.503    
##  3rd Qu.: 4.782      3rd Qu.:0.8000           3rd Qu.: 2.053    
##  Max.   :10.710      Max.   :6.4800           Max.   :21.920    
```

```
glimpse(daily_sleep)
```

```
## 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, 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…
```

```
summary(daily_sleep)
```

```
##        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
```

```
glimpse(daily_steps)
```

```
## Rows: 940
## Columns: 3
## $ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366…
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/…
## $ StepTotal   <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019, 15506, 1054…
```

```
summary(daily_steps)
```

```
##        Id             ActivityDay          StepTotal    
##  Min.   :1.504e+09   Length:940         Min.   :    0  
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790  
##  Median :4.445e+09   Mode  :character   Median : 7406  
##  Mean   :4.855e+09                      Mean   : 7638  
##  3rd Qu.:6.962e+09                      3rd Qu.:10727  
##  Max.   :8.878e+09                      Max.   :36019  
```

```
glimpse(hourly_calories)
```

```
## Rows: 22,099
## Columns: 3
## $ Id           <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityHour <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/12/20…
## $ Calories     <int> 81, 61, 59, 47, 48, 48, 48, 47, 68, 141, 99, 76, 73, 66, …
```

```
summary(hourly_calories)
```

```
##        Id            ActivityHour          Calories
##  Min.   :1.504e+09   Length:22099       Min.   : 42.00
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 63.00
##  Median :4.445e+09   Mode  :character   Median : 83.00
##  Mean   :4.848e+09                      Mean   : 97.39
##  3rd Qu.:6.962e+09                      3rd Qu.:108.00
##  Max.   :8.878e+09                      Max.   :948.00
```

```
glimpse(hourly_intensities)
```

```
## Rows: 22,099
## Columns: 4
## $ Id               <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 15039…
## $ ActivityHour     <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/1…
## $ TotalIntensity   <int> 20, 8, 7, 0, 0, 0, 0, 0, 13, 30, 29, 12, 11, 6, 36, 5…
## $ AverageIntensity <dbl> 0.333333, 0.133333, 0.116667, 0.000000, 0.000000, 0.0…
```

```
summary(hourly_intensities)
```

```
##        Id             ActivityHour       TotalIntensity   AverageIntensity
##  Min.   :1.504e+09   Length:22099       Min.   :  0.00   Min.   :0.0000  
##  1st Qu.:2.320e+09   Class :character   1st Qu.:  0.00   1st Qu.:0.0000  
##  Median :4.445e+09   Mode  :character   Median :  3.00   Median :0.0500  
##  Mean   :4.848e+09                      Mean   : 12.04   Mean   :0.2006  
##  3rd Qu.:6.962e+09                      3rd Qu.: 16.00   3rd Qu.:0.2667  
##  Max.   :8.878e+09                      Max.   :180.00   Max.   :3.0000  
```

```
glimpse(hourly_steps)
```

```
## Rows: 22,099
## Columns: 3
## $ Id           <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityHour <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/12/20…
## $ StepTotal    <int> 373, 160, 151, 0, 0, 0, 0, 0, 250, 1864, 676, 360, 253, 2…
```

```
summary(hourly_steps)
```

```
##        Id             ActivityHour         StepTotal    
##  Min.   :1.504e+09   Length:22099       Min.   :    0.0  
##  1st Qu.:2.320e+09   Class :character   1st Qu.:    0.0  
##  Median :4.445e+09   Mode  :character   Median :   40.0  
##  Mean   :4.848e+09                      Mean   :  320.2
##  3rd Qu.:6.962e+09                      3rd Qu.:  357.0
##  Max.   :8.878e+09                      Max.   :10554.0
```

```
glimpse(weight_log)
```

```
## Rows: 67
## Columns: 8
## $ Id             <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212…
## $ Date           <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2…
## $ WeightKg       <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, …
## $ WeightPounds   <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6…
## $ Fat            <int> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ BMI            <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,…
## $ IsManualReport <chr> "True", "True", "False", "True", "True", "True", "True"…
## $ LogId          <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,…
```

```
summary(weight_log)
```

```
##        Id                  Date              WeightKg       WeightPounds
##  Min.   :1.504e+09   Length:67          Min.   : 52.60   Min.   :116.0
##  1st Qu.:6.962e+09   Class :character   1st Qu.: 61.40   1st Qu.:135.4
##  Median :6.962e+09   Mode  :character   Median : 62.50   Median :137.8
##  Mean   :7.009e+09                      Mean   : 72.04   Mean   :158.8
##  3rd Qu.:8.878e+09                      3rd Qu.: 85.05   3rd Qu.:187.5  
##  Max.   :8.878e+09                      Max.   :133.50   Max.   :294.3  
##                                                                          
##       Fat             BMI        IsManualReport         LogId          
##  Min.   :22.00   Min.   :21.45   Length:67          Min.   :1.460e+12  
##  1st Qu.:22.75   1st Qu.:23.96   Class :character   1st Qu.:1.461e+12  
##  Median :23.50   Median :24.39   Mode  :character   Median :1.462e+12  
##  Mean   :23.50   Mean   :25.19                      Mean   :1.462e+12  
##  3rd Qu.:24.25   3rd Qu.:25.56                      3rd Qu.:1.462e+12  
##  Max.   :25.00   Max.   :47.54                      Max.   :1.463e+12  
##  NA's   :65
```

Lets clean the dataframes using the janitor and skimr packages

Lets check the null values and remove the null values for better analysis

```
daily_activity %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
daily_calories %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
daily_intensities %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
daily_sleep %>%
  is.na() %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
daily_steps %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
hourly_calories %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
hourly_intensities %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
hourly_steps %>%
  is.na() %>%
  sum()
```

```
## [1] 0
```

```
weight_log %>%
  is.na() %>%
  sum()
```

```
## [1] 65
```

from checking the dataset we can see that weight_log dataset has 65 null values in “FAT” col,lets remove the “FAT” col from weight_log as it is not necessary for our analysis

```
weight_log <- weight_log %>%
  select(-c("Fat"))
```

The next step is to verify the number of unique users as the ID column acts as a foreign key across the whole dataset,therefore I could merge the whole datasets using ID as it is shared by each dataframe

```
n_distinct(daily_activity$Id)
```

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

```
n_distinct(daily_calories$Id)
```

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

```
n_distinct(daily_intensities$Id)
```

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

```
n_distinct(daily_sleep$Id)
```

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

```
n_distinct(daily_steps$Id)
```

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

```
n_distinct(hourly_calories$Id)
```

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

```
n_distinct(hourly_intensities$Id)
```

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

```
n_distinct(hourly_steps$Id)
```

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

```
n_distinct(weight_log$Id)
```

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

Based on the results, there are 24 unique daily users that provided their health metrics info (SleepDay_merged dataframe), 8 unique users provided their daily weight_log_Info health metrics and 33 unique users provided the rest of the health metrics. Hence, the weight_log_info data frame could be dropped as the unique users are too few to give me any insightful information. lets check the duplicate rows in our dataframes

```
sum(duplicated(daily_activity))
```

```
## [1] 0
```

```
sum(duplicated(daily_calories))
```

```
## [1] 0
```

```
sum(duplicated(daily_intensities))
```

```
## [1] 0
```

```
sum(duplicated(daily_sleep))
```

```
## [1] 3
```

```
sum(duplicated(daily_steps))
```

```
## [1] 0
```

```
sum(duplicated(hourly_calories))
```

```
## [1] 0
```

```
sum(duplicated(hourly_intensities))
```

```
## [1] 0
```

```
sum(duplicated(hourly_steps))
```

```
## [1] 0
```

I noticed that daily_sleep has 3 duplicated rows where as none of other has duplicated rows. lets merge the duplicatd rows in dail_sleep

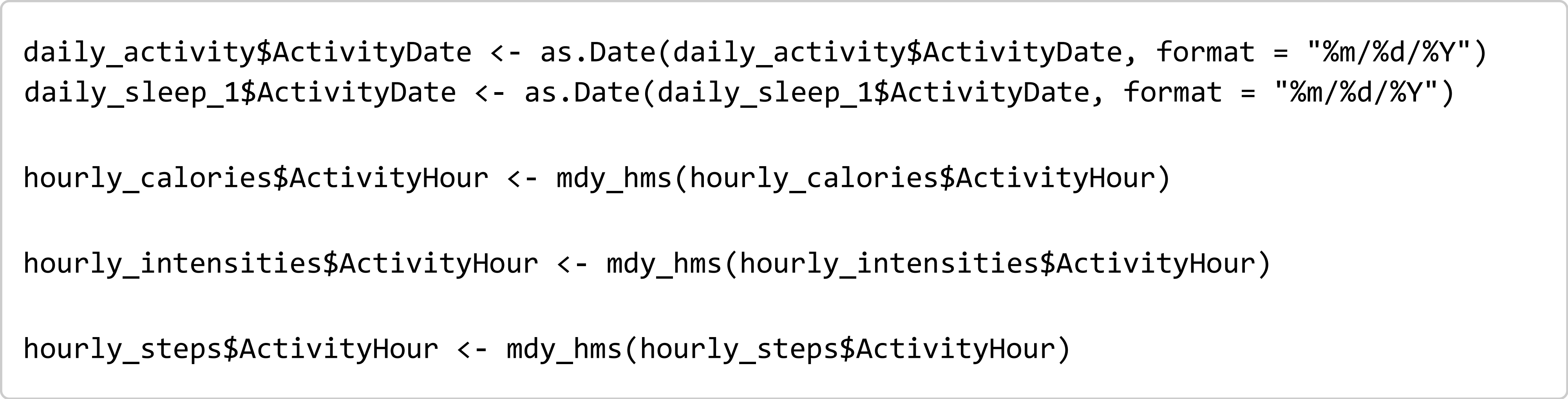
```
daily_sleep_1 <- daily_sleep[!duplicated(daily_sleep),]

sum(duplicated(daily_sleep_1))
```

```
## [1] 0
```

So, I decided to combine the data_sleep_1 and daily_activity but i saw the in daily_activity it is activitydate but in daily_sleep its is sleepdate lets rename it

```
daily_sleep_1 <-daily_sleep_1 %>%
  rename(ActivityDate = SleepDay)
```

There is disparancies in all of the date formats in all of these dataframes lets change into same format using lubridate package In hourly dataframes the Ther is disparencies in the timstamps so lets also correct that



```
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format = "%m/%d/%Y")
daily_sleep_1$ActivityDate <- as.Date(daily_sleep_1$ActivityDate, format = "%m/%d/%Y")

hourly_calories$ActivityHour <- mdy_hms(hourly_calories$ActivityHour)

hourly_intensities$ActivityHour <- mdy_hms(hourly_intensities$ActivityHour)

hourly_steps$ActivityHour <- mdy_hms(hourly_steps$ActivityHour)
```

ANALYZE PHASE

I merge the dailyActivity_merged table and daily_sleep_1 into a new data frame called “daily_activity_and_sleep”

I merge the hourlyCalories, hourlyIntensities these 2 data frames into a single data frame called “hourly_activity”. This is done via “Id” and “ActivityHour”

I merge the hourly activity dataframe and hourly steps in to a new dataframe called “hourly_act”

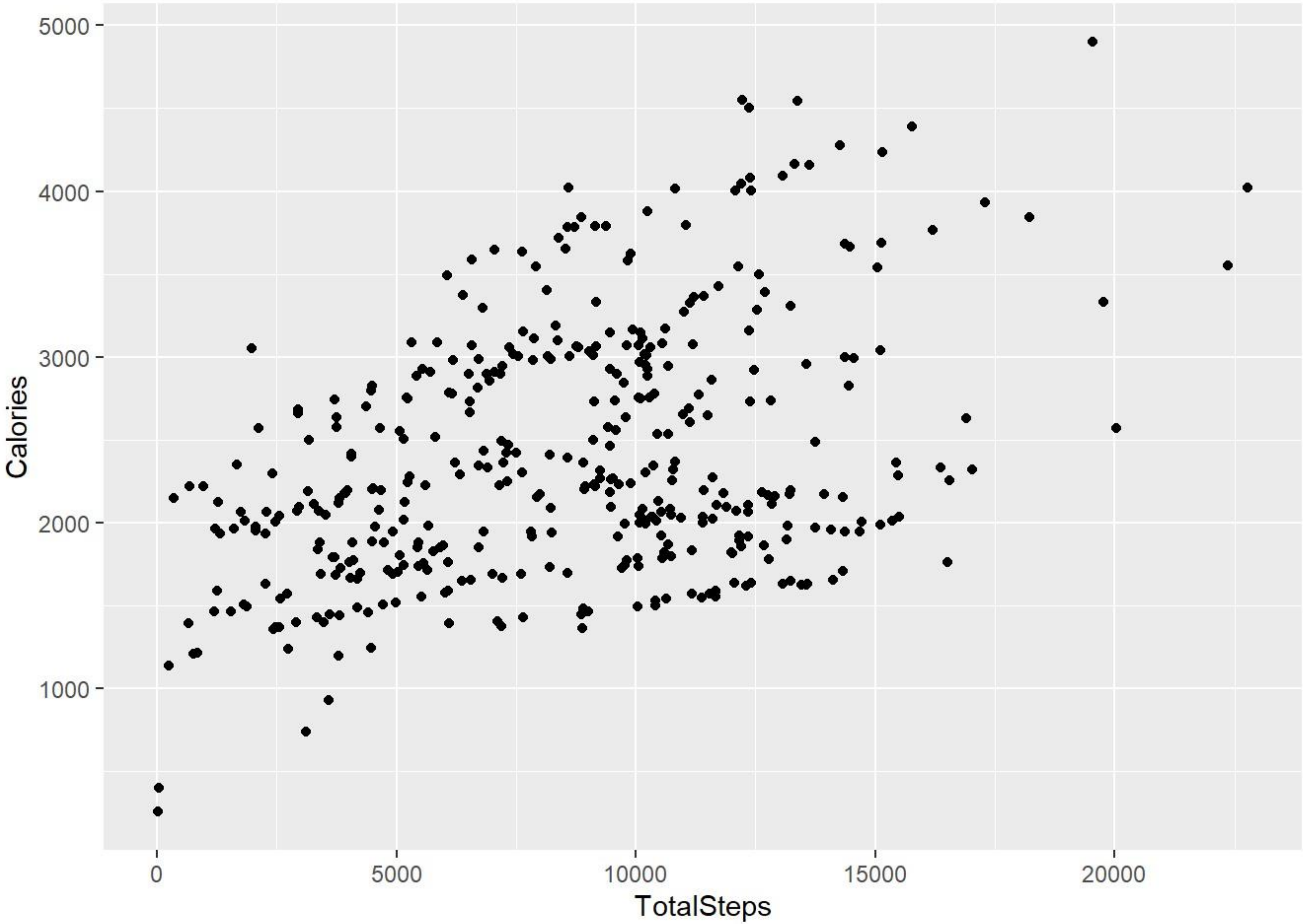
```
daily_activity_and_sleep <-merge(daily_activity,daily_sleep_1,by=c("Id","ActivityDate"))
```

```
 hourly_activity <- merge(hourly_calories,hourly_intensities, by =
c("Id","ActivityHour"))

hourly_act <- merge(hourly_activity,hourly_steps, by = c("Id","ActivityHour"))
```

Data visualization Phase

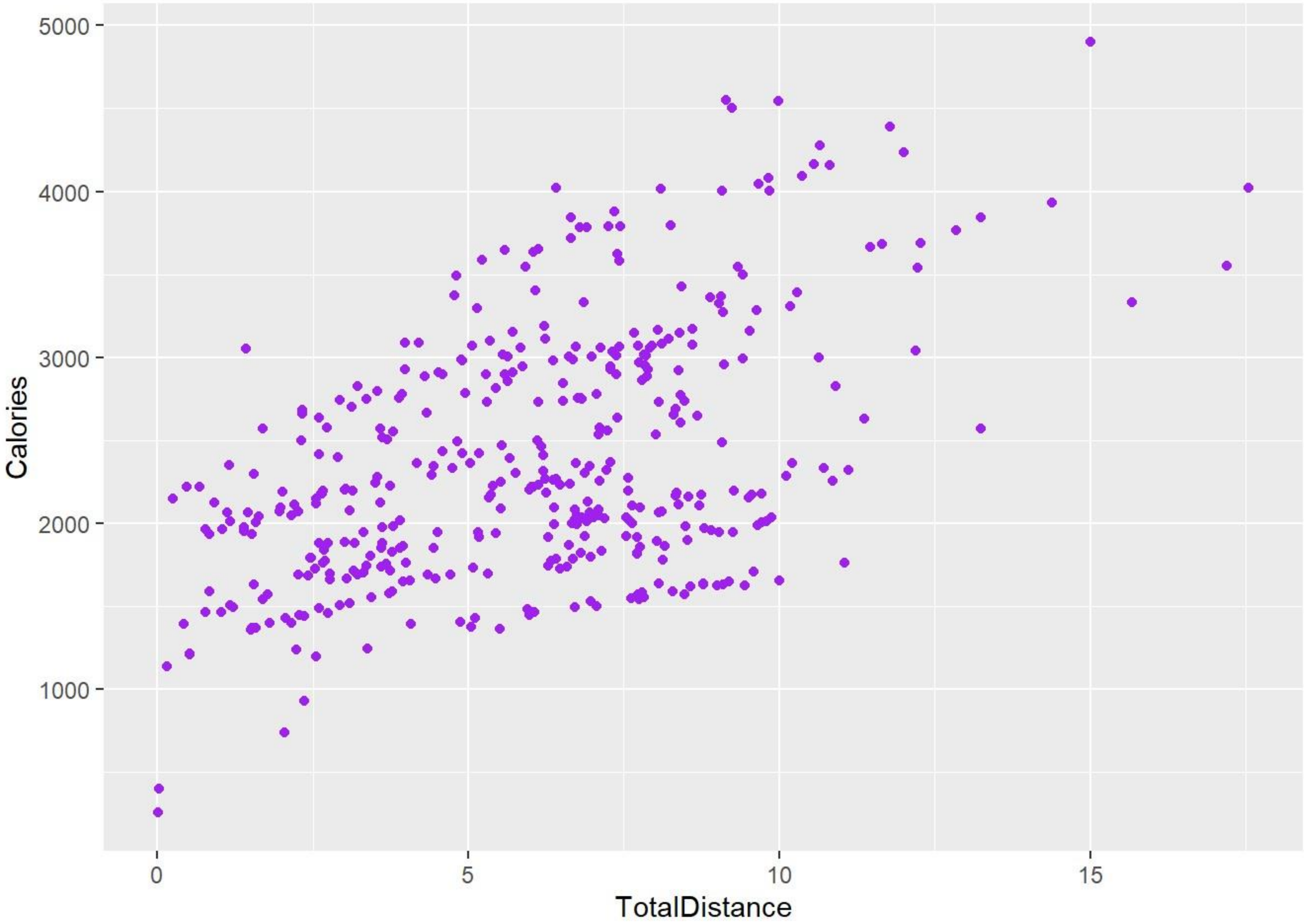
Determination of relationship between TotalSteps and Calories of each users using the daily_activity_with_sleep dataframe

```
ggplot(data=daily_activity_and_sleep)+geom_point(mapping = aes(x=TotalSteps, y=Calories))
```


Calories
5000
4000
3000
2000
1000
0
5000
10000
15000
20000
TotalSteps

Determine the relationship between the distance and the calories burnt

```
ggplot(data=daily_activity_and_sleep)+geom_point(mapping = aes(x=TotalDistance, y=Calories),color="purple")
```


Calories
5000
4000
3000
2000
1000
0
5
10
15
TotalDistance