Fitness Tracker data analysis

## Introduction and background

This is a simple analysis performed to discover key trends in fitness tracker market

#### Upload CSV files to R

Files used for analysis uploaded from kaggle: <https://www.kaggle.com/arashnic/fitbit>

#### Installing and loading common packages and libraries

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)  
library(writexl)

#### Loading CSV files

creating a dataframe called “daily\_activity” to upload specific data for analysis from “Fitabase Data 4.12.16-5.12.16”.

daily\_activity <- read.csv("dailyActivity\_merged.csv")

creating another dataframe called “sleep\_day” to upload more data

sleep\_day <- read.csv("sleepDay\_merged.csv")

#### Exploring data

Taking a look at the daily\_activity data.

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

Identifying all the columns in the daily\_activity data.

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"

Taking a look at the sleep\_day data.

head(sleep\_day)

## 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

Identifying all the columns in the daily\_activity data.

colnames(sleep\_day)

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

both datasets have the ‘Id’ field - this can be used to merge the datasets.

#### Understanding some summary statistics

first, identifying the unique participants in each dataframe

n\_distinct(daily\_activity$Id)

## [1] 33

n\_distinct(sleep\_day$Id)

## [1] 24

next how many observations are there in each dataframe?

nrow(daily\_activity)

## [1] 940

nrow(sleep\_day)

## [1] 413

#### More detailed statistics summary

For the daily activity dataframe:

daily\_activity %>%   
 select(TotalSteps,  
 TotalDistance,  
 SedentaryMinutes) %>%  
 summary()

## TotalSteps TotalDistance SedentaryMinutes  
## Min. : 0 Min. : 0.000 Min. : 0.0   
## 1st Qu.: 3790 1st Qu.: 2.620 1st Qu.: 729.8   
## Median : 7406 Median : 5.245 Median :1057.5   
## Mean : 7638 Mean : 5.490 Mean : 991.2   
## 3rd Qu.:10727 3rd Qu.: 7.713 3rd Qu.:1229.5   
## Max. :36019 Max. :28.030 Max. :1440.0

For the sleep dataframe:

sleep\_day %>%   
 select(TotalSleepRecords,  
 TotalMinutesAsleep,  
 TotalTimeInBed) %>%  
 summary()

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

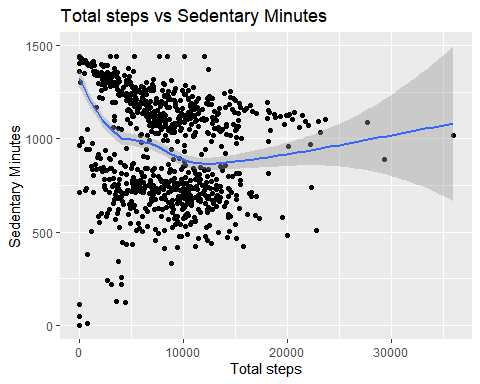
from the summary, the participants walked an average of 7638 steps daily and slept for 7 hours on average

#### Plotting a few explorations

Now to look at the relationship between steps taken in a day and sedentary minutes. This could help inform us on customer market segments to market to.

ggplot(data=daily\_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom\_point()+   
 geom\_smooth()+  
 labs(title = "Total steps vs Sedentary Minutes", x = "Total steps", y = "Sedentary Minutes")

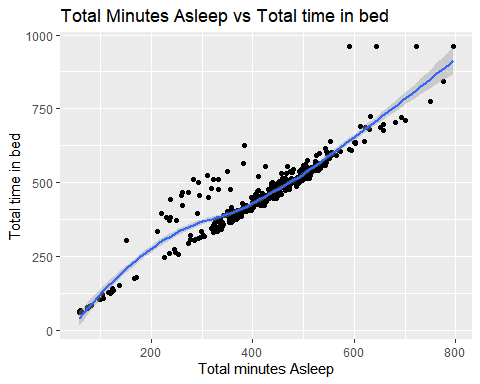
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



Next to look at the relationship between minutes asleep and time in bed.

ggplot(data=sleep\_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_point()+   
 geom\_smooth()+  
 labs(title = "Total Minutes Asleep vs Total time in bed", x = "Total minutes Asleep", y = "Total time in bed")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



#### Merging these two datasets together

combined\_data <- merge(sleep\_day, daily\_activity, by="Id")  
head(combined\_data)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep  
## 1 1503960366 4/12/2016 12:00:00 AM 1 327  
## 2 1503960366 4/12/2016 12:00:00 AM 1 327  
## 3 1503960366 4/12/2016 12:00:00 AM 1 327  
## 4 1503960366 4/12/2016 12:00:00 AM 1 327  
## 5 1503960366 4/12/2016 12:00:00 AM 1 327  
## 6 1503960366 4/12/2016 12:00:00 AM 1 327  
## TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 346 5/7/2016 11992 7.71 7.71  
## 2 346 5/6/2016 12159 8.03 8.03  
## 3 346 5/1/2016 10602 6.81 6.81  
## 4 346 4/30/2016 14673 9.25 9.25  
## 5 346 4/12/2016 13162 8.50 8.50  
## 6 346 4/13/2016 10735 6.97 6.97  
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance  
## 1 0 2.46 2.12  
## 2 0 1.97 0.25  
## 3 0 2.29 1.60  
## 4 0 3.56 1.42  
## 5 0 1.88 0.55  
## 6 0 1.57 0.69  
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes  
## 1 3.13 0 37  
## 2 5.81 0 24  
## 3 2.92 0 33  
## 4 4.27 0 52  
## 5 6.06 0 25  
## 6 4.71 0 21  
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories  
## 1 46 175 833 1821  
## 2 6 289 754 1896  
## 3 35 246 730 1820  
## 4 34 217 712 1947  
## 5 13 328 728 1985  
## 6 19 217 776 1797

#### A summary of combined\_data

head(combined\_data)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep  
## 1 1503960366 4/12/2016 12:00:00 AM 1 327  
## 2 1503960366 4/12/2016 12:00:00 AM 1 327  
## 3 1503960366 4/12/2016 12:00:00 AM 1 327  
## 4 1503960366 4/12/2016 12:00:00 AM 1 327  
## 5 1503960366 4/12/2016 12:00:00 AM 1 327  
## 6 1503960366 4/12/2016 12:00:00 AM 1 327  
## TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 346 5/7/2016 11992 7.71 7.71  
## 2 346 5/6/2016 12159 8.03 8.03  
## 3 346 5/1/2016 10602 6.81 6.81  
## 4 346 4/30/2016 14673 9.25 9.25  
## 5 346 4/12/2016 13162 8.50 8.50  
## 6 346 4/13/2016 10735 6.97 6.97  
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance  
## 1 0 2.46 2.12  
## 2 0 1.97 0.25  
## 3 0 2.29 1.60  
## 4 0 3.56 1.42  
## 5 0 1.88 0.55  
## 6 0 1.57 0.69  
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes  
## 1 3.13 0 37  
## 2 5.81 0 24  
## 3 2.92 0 33  
## 4 4.27 0 52  
## 5 6.06 0 25  
## 6 4.71 0 21  
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories  
## 1 46 175 833 1821  
## 2 6 289 754 1896  
## 3 35 246 730 1820  
## 4 34 217 712 1947  
## 5 13 328 728 1985  
## 6 19 217 776 1797

colnames(combined\_data)

## [1] "Id" "SleepDay"   
## [3] "TotalSleepRecords" "TotalMinutesAsleep"   
## [5] "TotalTimeInBed" "ActivityDate"   
## [7] "TotalSteps" "TotalDistance"   
## [9] "TrackerDistance" "LoggedActivitiesDistance"  
## [11] "VeryActiveDistance" "ModeratelyActiveDistance"  
## [13] "LightActiveDistance" "SedentaryActiveDistance"   
## [15] "VeryActiveMinutes" "FairlyActiveMinutes"   
## [17] "LightlyActiveMinutes" "SedentaryMinutes"   
## [19] "Calories"

combined\_data %>%  
 select(TotalSteps, TotalDistance, Calories, SedentaryMinutes, VeryActiveMinutes, TotalTimeInBed, TotalMinutesAsleep) %>%  
 summary()

## TotalSteps TotalDistance Calories SedentaryMinutes  
## Min. : 0 Min. : 0.000 Min. : 0 Min. : 0.0   
## 1st Qu.: 4660 1st Qu.: 3.180 1st Qu.:1783 1st Qu.: 659.0   
## Median : 8596 Median : 6.120 Median :2162 Median : 734.0   
## Mean : 8117 Mean : 5.735 Mean :2329 Mean : 799.2   
## 3rd Qu.:11317 3rd Qu.: 7.920 3rd Qu.:2865 3rd Qu.: 853.0   
## Max. :22988 Max. :17.950 Max. :4900 Max. :1440.0   
## VeryActiveMinutes TotalTimeInBed TotalMinutesAsleep  
## Min. : 0.00 Min. : 61.0 Min. : 58.0   
## 1st Qu.: 0.00 1st Qu.:402.0 1st Qu.:361.0   
## Median : 8.00 Median :463.0 Median :432.0   
## Mean : 23.97 Mean :458.4 Mean :419.4   
## 3rd Qu.: 36.00 3rd Qu.:526.0 3rd Qu.:492.0   
## Max. :210.00 Max. :961.0 Max. :796.0

Taking a look at how many participants are in this data set.

n\_distinct(combined\_data$Id)

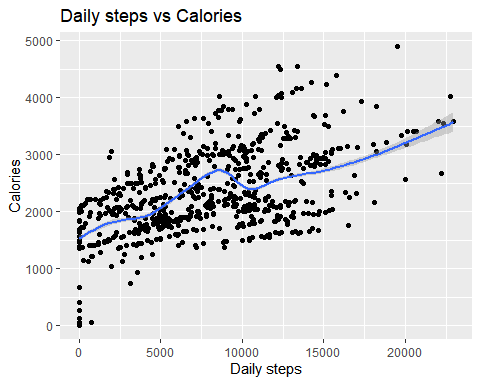
## [1] 24

# Plotting a few more Explorations

Looking at the relationship between daily steps and daily calories

ggplot(combined\_data, aes(x = TotalSteps, y = Calories))+  
 geom\_point()+  
 geom\_smooth()+  
 labs(title = "Daily steps vs Calories", x = "Daily steps", y = "Calories")

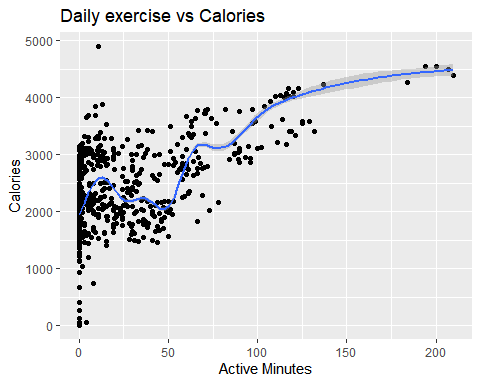
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Now a look at the relationship between daily exercise time(VeryActiveMinutes) and calories

ggplot(combined\_data, aes(x = VeryActiveMinutes, y = Calories))+  
 geom\_point()+  
 geom\_smooth()+  
 labs(title = "Daily exercise vs Calories", x = "Active Minutes", y = "Calories")

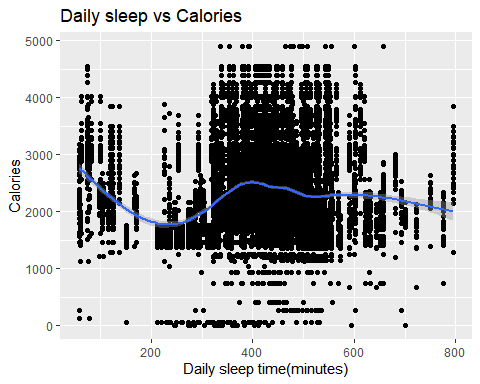
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Now a look at the relationship between daily sleep and calories

ggplot(combined\_data, aes(x = TotalMinutesAsleep, y = Calories))+  
 geom\_point()+  
 geom\_smooth()+  
 labs(title = "Daily sleep vs Calories", x = "Daily sleep time(minutes)", y = "Calories")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Exporting dataframe for further visualization on Tableau

write\_xlsx(x = combined\_data, path = "combined\_data.xlsx", col\_names = TRUE)