Case Study 1

Yiu

2025-07-18

Introduction and background

Objective:

Gain insight from non-Bellabeat smart device data to produce high-level recommendations for marketing strategy to **one** of their products.

Product Choice:

Time: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.

Business Task:

Identify trends in smart device usage from non-Bellabeat smart devices, and look for any discoveries that can help improve the current marketing strategy of *Time*.

Preparation of Dataset

I am suggested to use the dataset from Kaggle, FitBit Fitness Tracker Data. The dataset originated from zenodo.org, and uploaded by a team RTI International researchers who is thought to be a reliable research organisation, so the dataset has a relatively high credibility. However, the data are gathered through distributed survey via Amazon Mechanical Turk, this may create a bias that people not using that platform cannot be observed. More details on data format can be found in here.

##Let's begin in RStudio.

Install and load the tidyverse

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

```
library(tidyverse)
```

```
dailyActivity <- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/dailyActivity_merged.csv")
heartrate_seconds<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/heartrate_seconds_merged.csv
hourlyCalories<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/hourlyCalories_merged.csv")
hourlyIntensities<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/hourlyIntensities_merged.csv
hourlySteps<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/hourlySteps_merged.csv")
minuteCaloriesNarrow<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/minuteCaloriesNarrow_merg
minuteIntensitiesNarrow<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/minuteIntensitiesNarro
minuteMETsNarrow<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/minuteMETsNarrow_merged.csv"
minuteSleep<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/minuteSleep_merged.csv")
```

```
minuteStepsNarrow<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/minuteStepsNarrow_merged.csv
weightLogInfo<- read.csv("D://mturkfitbit/Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv")
```

Load CSV files

Explore some tables

Now, I will briefly look into each data frames, `str`, `head`, `summary` functions are good options.

```
str(weightLogInfo)
```

Take a look at the weightLogInfo

```
## 'data.frame':    33 obs. of  8 variables:
##  $ Id            : num  1.50e+09 1.93e+09 2.35e+09 2.87e+09 2.87e+09 ...
##  $ Date          : chr  "4/5/2016 11:59:59 PM" "4/10/2016 6:33:26 PM" "4/3/2016 11:59:59 PM" "4/6/20
##  $ WeightKg      : num  53.3 129.6 63.4 56.7 57.2 ...
##  $ WeightPounds  : num  118 286 140 125 126 ...
##  $ Fat           : int  22 NA 10 NA NA NA NA NA NA ...
##  $ BMI           : num  23 46.2 24.8 21.5 21.6 ...
##  $ IsManualReport: chr  "True" "False" "True" "True" ...
##  $ LogId         : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

I discovered some formatting problems in the dataset,the data type of `Date` is recorded as character, but not date; whereas `IsManualReport` column has the same problem, the date type being character rather than logical.

```
weightLogInfo <- weightLogInfo %>%
  mutate(weightLogInfo,NewDate = mdy_hms(Date)) %>%
  mutate(weightLogInfo,IsManualReport_logical = as.logical(IsManualReport))
```

```
str(dailyActivity)
```

Take a look at the dailyActivity

```
## 'data.frame':    457 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr  "3/25/2016" "3/26/2016" "3/27/2016" "3/28/2016" ...
##  $ TotalSteps               : int  11004 17609 12736 13231 12041 10970 12256 12262 11248 10016 ...
##  $ TotalDistance            : num  7.11 11.55 8.53 8.93 7.85 ...
##  $ TrackerDistance          : num  7.11 11.55 8.53 8.93 7.85 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  2.57 6.92 4.66 3.19 2.16 ...
##  $ ModeratelyActiveDistance: num  0.46 0.73 0.16 0.79 1.09 ...
##  $ LightActiveDistance      : num  4.07 3.91 3.71 4.95 4.61 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  33 89 56 39 28 30 33 47 40 15 ...
##  $ FairlyActiveMinutes      : int  12 17 5 20 28 13 12 21 11 30 ...
##  $ LightlyActiveMinutes     : int  205 274 268 224 243 223 239 200 244 314 ...
##  $ SedentaryMinutes         : int  804 588 605 1080 763 1174 820 866 636 655 ...
##  $ Calories                 : int  1819 2154 1944 1932 1886 1820 1889 1868 1843 1850 ...
```

Some information, such `TotalSteps` and `Calories` are already included, so the corresponding sheets, `hourlySteps` and `hourlyCalories`, can be omitted unless more details are needed. The only extra information are intensity, heartrate, METs, sleeping time and weight log.

However, after some research, Fitbit use heartrate and METs to create some high level data, such as intensity and classification of active minutes, I will come back to these data if needed.

Organise the sheets about sleeping time and weight log.

For the sleeping time one, since it is in a long format that check whether the user is sleeping in every minute, so I run this code to find more metadata

```
## Calculate each duration of sleeping session and date of sleep by logId
SleepSessionTime <-
  minuteSleep %>%
  mutate(minuteSleep, NewDate = mdy_hms(date)) %>%
  group_by(Id,logId) %>%
  summarise(sleep_sec = max(NewDate) - min(NewDate), sleep_date = min(NewDate))
```

```
## `summarise()` has grouped output by 'Id'. You can override using the `.groups`
## argument.
```

It turns out that some people have a habit of taking a nap, which may need to switch to calculate sleeping per day instead of per session, and there seems to be inaccurate measurement, such as for `Id` 2022484408 and 7007744171, their mean sleeping time is below 3 hours, the same goes for 1644430081 and 1844505072, having exceptional high sleeping hours, which does not seem to be accurate measurements. I shall discard them when I have to use this data.

On the `weightLogInfo` table, there are several duplicates for some people, because they have multiple measurement across the period of time. I pick the most recent as reference. Also, since almost all the columns inside can be concluded into BMI.

```
## Store in a new data frame
BMIInfo <-
  weightLogInfo %>%
## Change of format
  mutate(weightLogInfo,NewDate = mdy_hms(Date)) %>%
  mutate(weightLogInfo,IsManualReport_logical = as.logical(IsManualReport)) %>%
## Choose the lastest entry for each id
  group_by(Id) %>%
  filter(NewDate == max(NewDate))
```

Understanding some summary statistics

Note that by running the `distinct()` command, it is found that there are only 11 people's BMI and 23 people's sleeping time, which means that not all respondants' data are collected, and, to be clear, in the `dailyActivity` sheet, there are 35 people. This tells us the sample size is 35 instead of 30.

Now, let's move on to the `dailyActivity` sheet. I would like to address that, by definition,

TotalDistance = TrackerDistance + LoggedActivitiesDistance = VeryActiveDistance + ModeratelyActiveDistance + LightActiveDistance + SedentaryActiveDistance

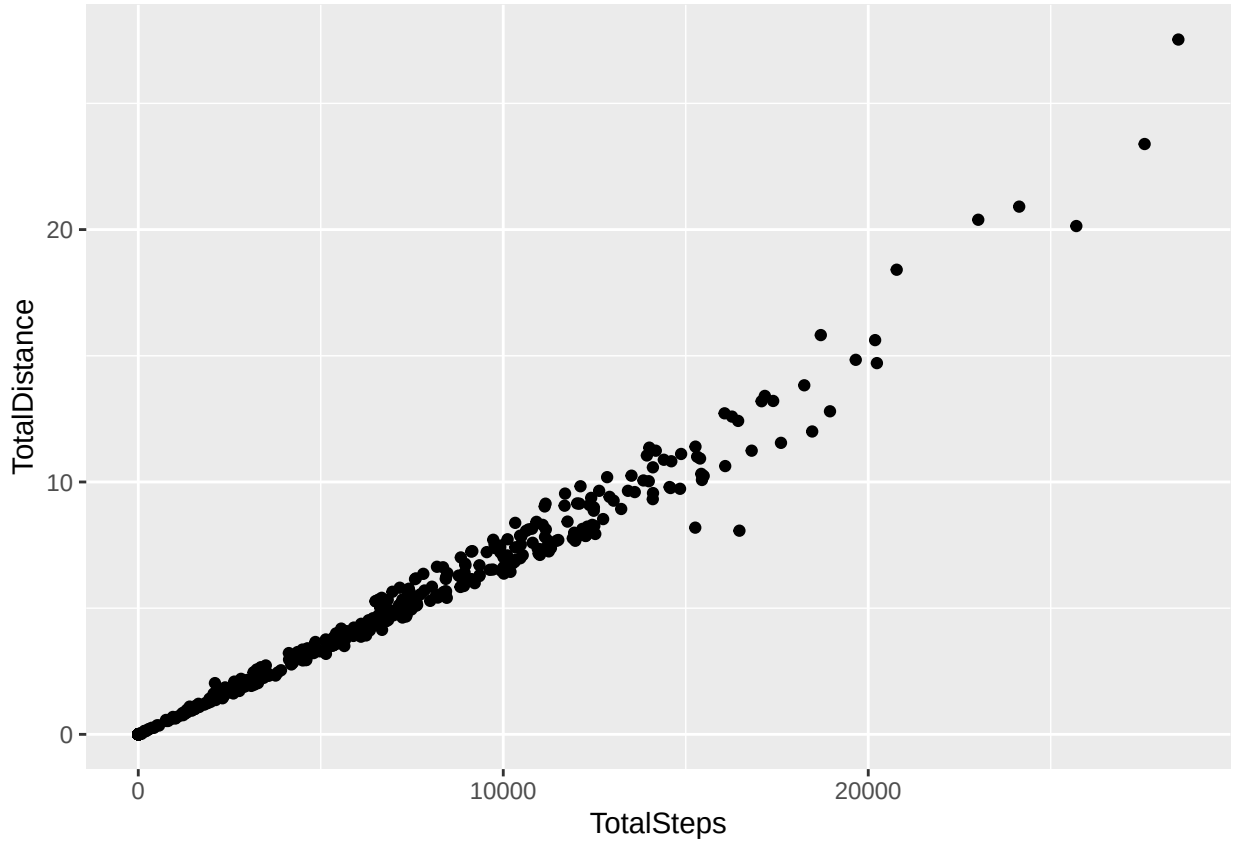
VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes should be 1440 (minutes) which equals to a whole day, but I found out that some entries do not, I shall keep this in mind that this may not be a good indicator. Yet, on the other hand, this is a good metric to view their daily usage.

```
## Insert new columns for total recorded minutes per day.
dailyActivity <- dailyActivity %>%
  mutate(dailyActivity,TotalRecordedMinutes = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMi
  mutate(dailyActivity,Date = mdy(ActivityDate))
```

Plotting a few explorations

By a simple scatter plot, we can observe that TotalSteps is positively variated with TotalDistance. The smartwatches are functioning well.

```
ggplot(data = dailyActivity) +
geom_point(mapping = aes(x=TotalSteps, y=TotalDistance))
```



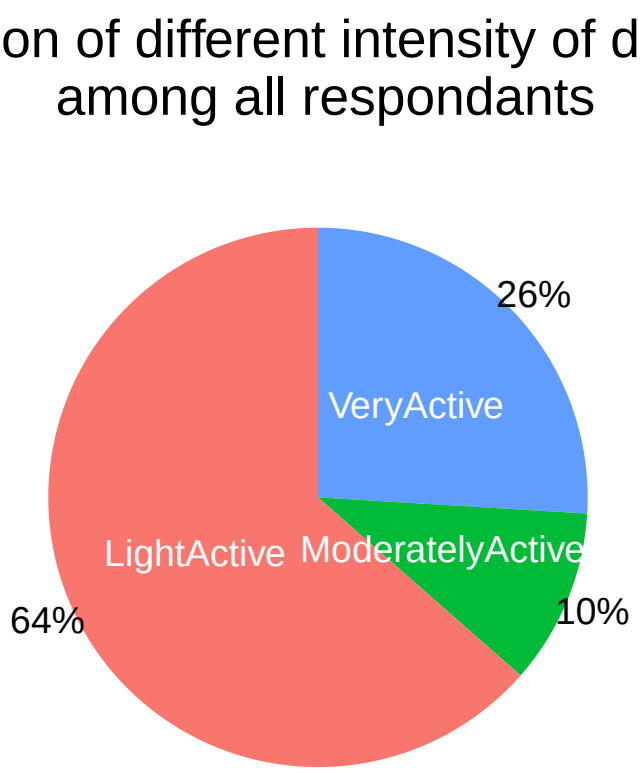
And by a pie chart, we see that higher intensity does not mean making up fewer proportion.

```
## Need to use one of the function in scales
library('scales')
```

```
## Create a new master list and further refine the data
MasterList <- dailyActivity %>%
 group_by(Id) %>%
 summarise(AvgSteps=mean(TotalSteps),NewTotalDistance=sum(TotalDistance), TotalVeryActiveDistance = sum

## Preparation for graphing
distance_intensity <- data_frame(
distance_intensity_type = c("LightActive", "ModeratelyActive", "VeryActive"),
all_respondants_distance_by_intensity = c(sum(MasterList$TotalLightActiveDistance), sum(MasterList$Total
intensity_percent = round((all_respondants_distance_by_intensity / sum(all_respondants_distance_by_inte
)

distance_intensity <- distance_intensity %>%
  arrange(desc(distance_intensity_type)) %>%
  mutate(prop = all_respondants_distance_by_intensity / sum(distance_intensity$all_respondants_distance_
```

```
  mutate(ypos = (cumsum(prop)- 0.5*prop))
```

```
## Pie Chart of proportion among different intersity
ggplot(distance_intensity, aes(x="", y=prop, fill=distance_intensity_type )) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  labs(title="Proportion of different intensity of distance \n among all respondants")+
  theme(legend.position="none",plot.title = element_text(hjust = 0.5, size = 20))+
  geom_text(aes(y = ypos, label = distance_intensity_type), color = "white", size=5) +
  geom_text(aes(x=1.6,y = ypos, label = percent(intensity_percent,scale = 1,suffix = "%") ), color = "bl
```

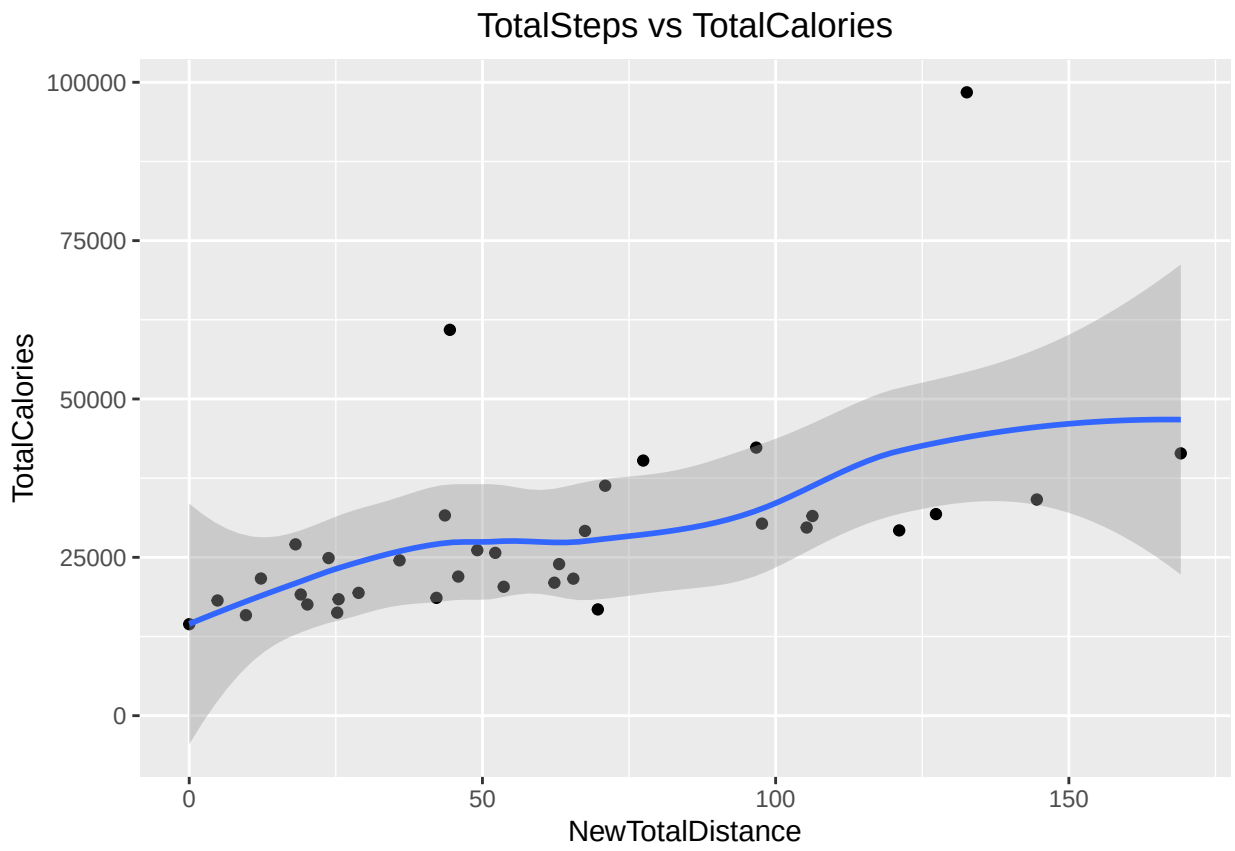


In terms of total distance, it is composed by 64% of light active, 10% of moderately active, and 26% of very active. This may imply there are two extremes, very few people choose to have a moderately active activity.

I further investigate the relationship between number of steps and calories, it is found that they are positively related, but at a mimimal correlation.

```
#Relationship between steps and calories
ggplot(data = MasterList,aes(x=NewTotalDistance, y=TotalCalories)) +
  geom_point()+
  geom_smooth(method="loess")+
  labs(title="TotalSteps vs TotalCalories")+
  theme(plot.title = element_text(hjust = 0.5))
```

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



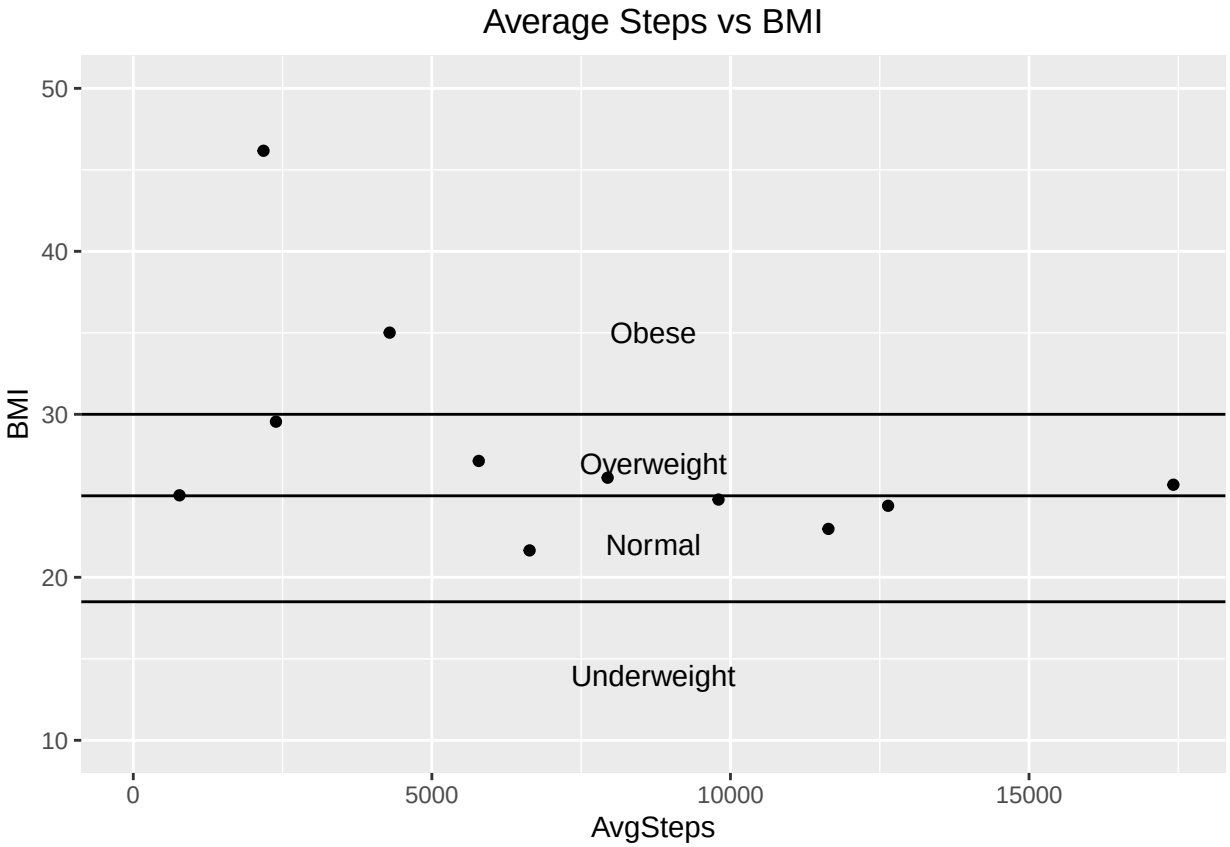
Merging these two datasets together

Next, I moved on to find if there is any relationship between BMI and daily distance. Since they are data from two different sheets, I have to merge them first by creating a new sheet called `MasterList`. I also classify respondants' healthiness based on BMI, depsite being considered as a outdated and inaccurate measure. Yet, no clear correlation is found.

```
## Combining with BMI
MasterList <- left_join(MasterList,BMIInfo)
```

```
## Joining with `by = join_by(Id)`
```

```
    BMIpos <- c(14,22,27,35)
    BMIcategory <- c("Underweight","Normal","Overweight","Obese")
    MidDist <-max((MasterList$AvgSteps)-min(MasterList$AvgSteps)) /2

## Plot with BMI against Average Steps, with BMI classifications
  ggplot(data = MasterList) +
    geom_point(mapping = aes(x=AvgSteps, y=BMI))+
    labs(title="Average Steps vs BMI")+
    theme(plot.title = element_text(hjust = 0.5))+
    ylim(10,50)+
    geom_hline(aes(yintercept = 18.5))+
    geom_hline(aes(yintercept = 25))+
    geom_hline(aes(yintercept = 30))+
    annotate('text', x=MidDist, y=BMIpos, label=BMIcategory, size=4)
```



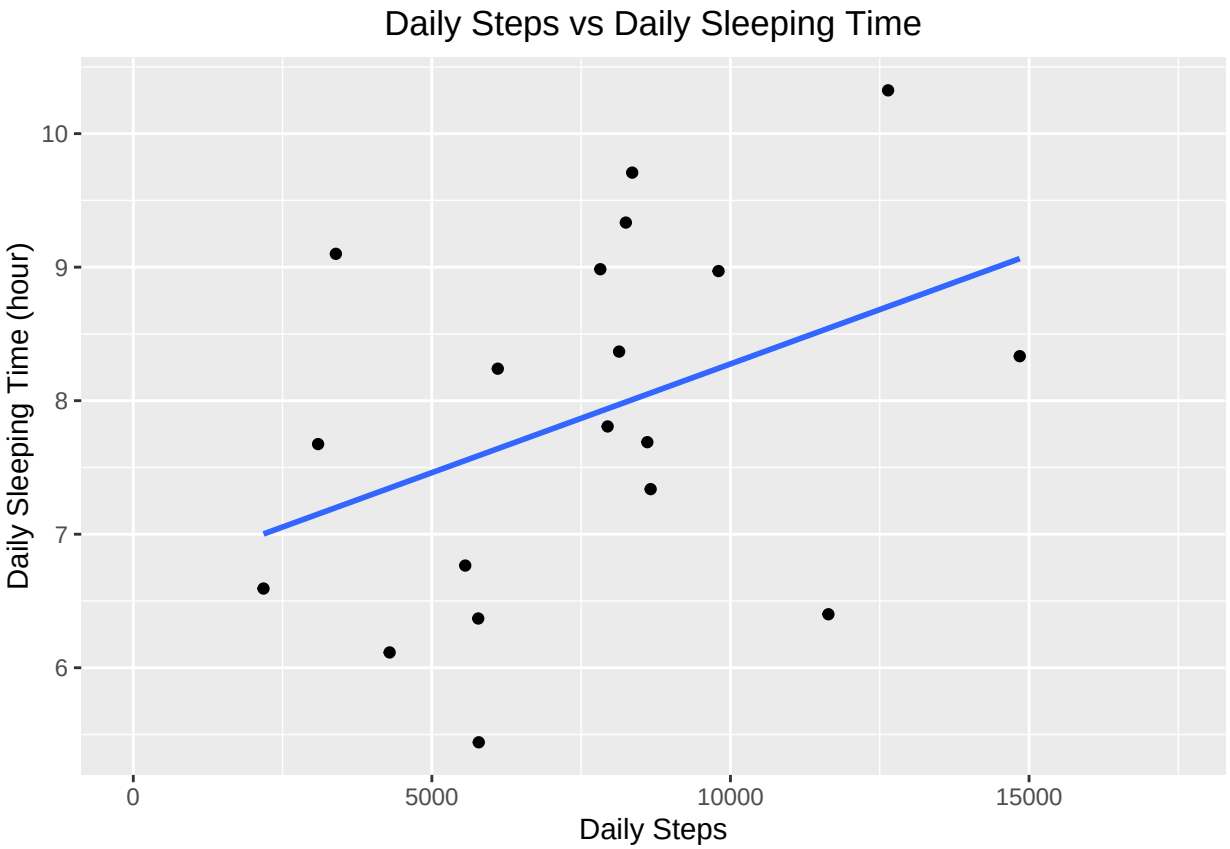
Furthermore, I tried to investigate the correlation sleep time and number of steps. Keeping in mind that some outliers mentioned before should be remove.

```
SleepInfo<-
SleepSessionTime %>%
  mutate(sleep_day = date(sleep_date)) %>%
  group_by(Id) %>%
  summarise(daily_sleep_time = sum(sleep_sec)/n_distinct(sleep_day))

SleepInfo <- SleepInfo[!(SleepInfo$Id %in% c(1644430081,1844505072,2022484408,7007744171)),]

MasterList <- left_join(MasterList, SleepInfo)
```

```
## Joining with `by = join_by(Id)`
```

```
ggplot(data = MasterList,aes(x=AvgSteps, y=daily_sleep_time/3600)) +
  geom_point()+
  labs(title="Daily Steps vs Daily Sleeping Time")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Daily Steps")+
  ylab("Daily Sleeping Time (hour)")+
  geom_smooth(method='lm',se = FALSE)
```

```
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
## `geom_smooth()` using formula = 'y ~ x'
```



Due to the limited data points, the trend is not apparent. However, one can observed that when daily steps increase, the daily sleeping time also increases in general.

Last but not least, we would also like to find out the trend of the device usage.

There is no clear trend against the day of usage.

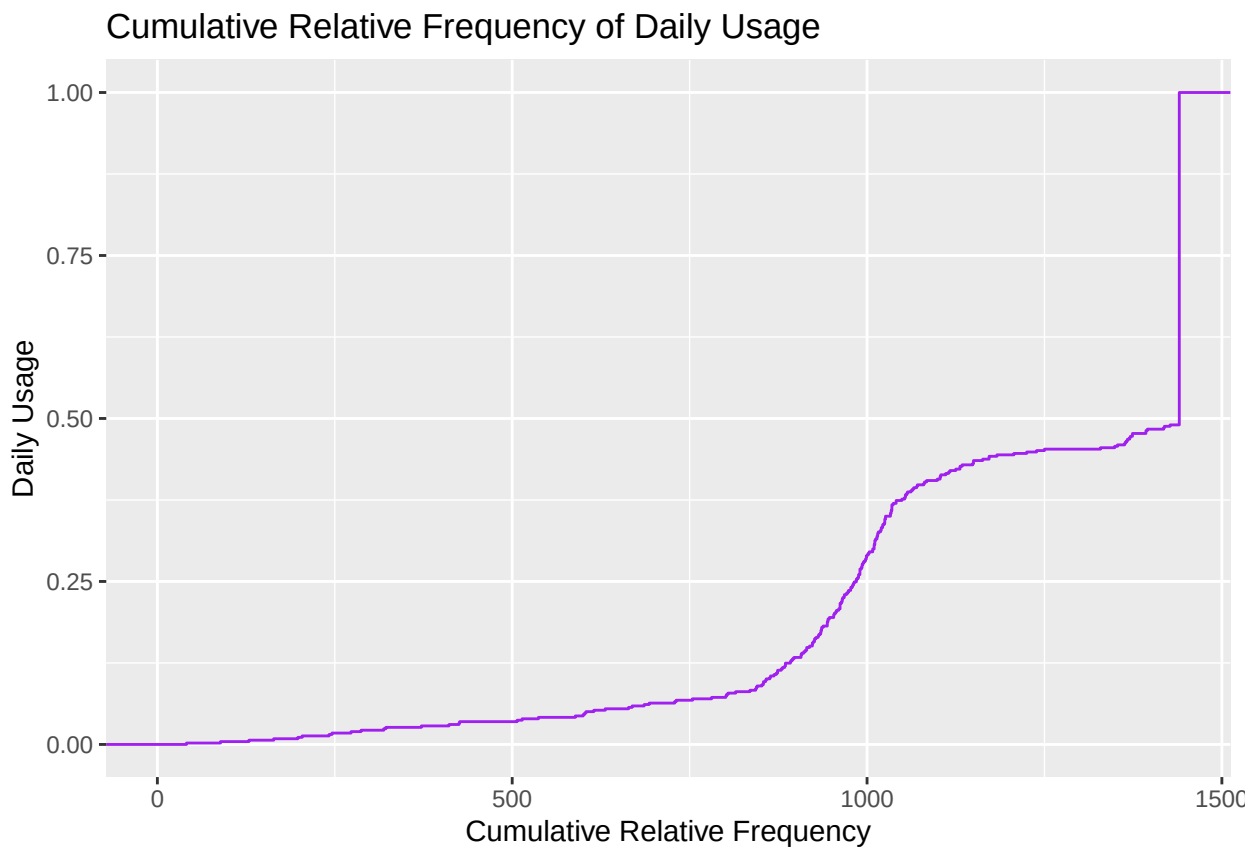
```
dailyActivity %>%
  mutate(dailyActivity,Day = wday(Date, label = TRUE, locale="en")) %>%
  group_by(Day) %>%
  summarise(median_usage = median(TotalRecordedMinutes), mean_usage = mean(TotalRecordedMinutes))
```

```
## # A tibble: 7 x 3
##   Day   median_usage mean_usage
##   <ord>        <dbl>      <dbl>
## 1 Sun           1440      1204.
## 2 Mon           1440      1234.
## 3 Tue           1026      1007.
## 4 Wed           1440      1233.
## 5 Thu           1440      1254.
## 6 Fri           1440      1271.
## 7 Sat           1440      1198.
```

Other schema cannot be used, as this may cause an observer bias. For example, one does not want wear the watch during sports, the time of device usage and active minutes will both decrease, so it is inaccurate to make correlation between them.

```
ggplot(data = dailyActivity, aes(TotalRecordedMinutes))+
  stat_ecdf(geom = "step",color ="purple")+
```

```
  xlim(0,1440)+
  labs(title = "Cumulative Relative Frequency of Daily Usage")+
  xlab("Cumulative Relative Frequency")+
  ylab("Daily Usage")
```

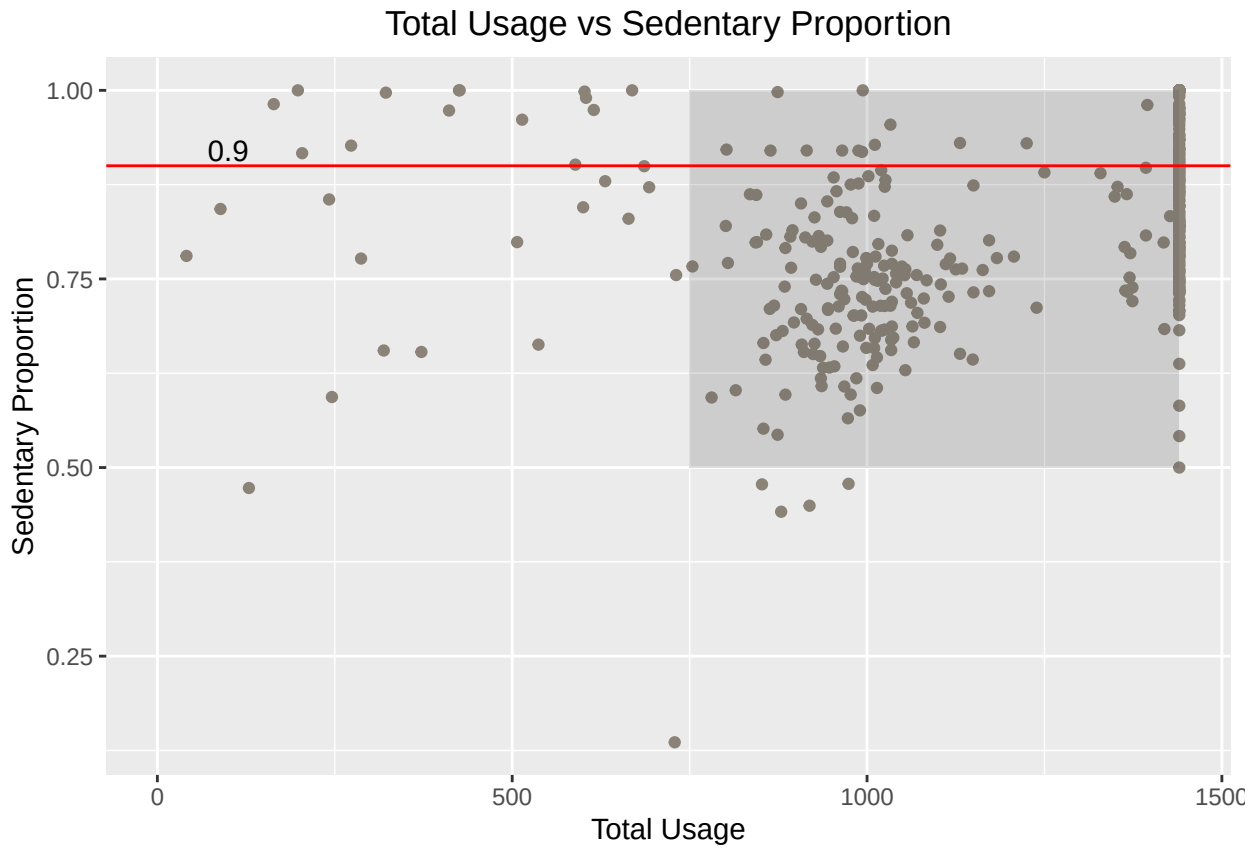


This plot gives a overview how frequent the device is used. It revealed that only a half of the recorded usage showed that the respondant wears it whole day (24 hours). Around one-fourth of the them wear less than 16 hours. It is uncertain whether this is due to "they do not wear them at sleep".

```
dailyActivity <- dailyActivity %>%
  mutate(SedProp =SedentaryMinutes/TotalRecordedMinutes)
```

This calculated the proportion of sedentary minutes over total recorded minutes. I am interested that how is the actual usage from the view of activeness.

```
ggplot(data = dailyActivity)+
  geom_point(mapping = aes(x=TotalRecordedMinutes, y=SedProp),color ="antiquewhite4")+
  labs(title="Total Usage vs Sedentary Proportion")+
  theme(plot.title = element_text(hjust = 0.5))+
  xlim(0,1440)+
  annotate("rect",xmin = 750,xmax = 1440,ymin = 0.5,ymax = 1.0,alpha = 0.2)+
  annotate('text', x=100, y=0.92, label=0.9, size=4)+
  geom_hline(yintercept = 0.9,color ="red")+
  xlab("Total Usage")+
  ylab("Sedentary Proportion")
```



This plot tries to correlate the sedentary proportion of the usage. One can tell from the graph,

(1) Most people are in the first quadrant (usage more than 12 hours, sedentary proportion greater than 0.5)

(2) A clear stack of points on x=1440, which means wear it full day

(3) A few points is above the line y=0.9, meaning some highly inactive usage are recorded. Especially some records are having 100% sedentary proportion, this may imply the respondant is totally idle or suffer from illness. It is very likely to be not wearing it and leaving it on.

Insights and Conclusions

Usage

Some cases of being completely idle, or highly inactive are found.

Not everyone wears it whole day, only a half of them wear them frequently.

Even fewer people wear it during sleeping.

Health

BMI is a outdated measurement, and they may be a lot of factors affect it.

Very little correlation was found.

More steps give more distances and also more calories.

Intensity of distance: Moderately Active (10%) < Very Active (26%) < Light Active (64%), which is unexpected to see the most active category being the second place.

Limitation

The dataset has limited data. More extensive datasets can be used, such as a longer period or across different products to give a more comprehensive analysis.

Suggestion on Marketing Strategy

Let's further relate these insights to what can be done.

(1) Design a better device that is more user-friendly that will yield higher usage time, and so provide more accurate evaluation to users. There are times that users does not wear their smartwatches, especially during sleep (since there are very little sleep records). Hence, this motivates us if there are any ways to improve user experience that make them wear longer, More information can be gathered through surveys.

(2) Develop notification system or any kind of interactions as there are some cases that having high idling time to avoid miscalculation. This can improve users' engagement to their smart devices. A further step can be send reminder to user to encourage health management.

(3) Extend wider metrics to measure healthiness and activity. For example, BMI is not that accurate, and the correlation between steps and calories is weak, this may imply there is still a lot of factor contribute to calories. Based on these outcomes from Fitbit, we shall actively check for better measurement for healthiness to our customers, making sure they have the most accurate understanding on their health.