Data Science in Medicine - Final Report

Anjali Bari, David Lee,Keerthana Kesavan, Mohal Khandelwal, Rahul Chauhan, Srav

Introduction

Analyzing data to find patterns and trends that could be indicators of future occurrences is the process of predictive analytics. Predictive analytics can be used in the healthcare industry to forecast the likelihood of specific medical disorders or the likelihood that a patient will respond to a specific treatment. Predictive analytics uses methods from data mining, statistics and mathematical modeling to make future predictions about unknowable events. It creates forecasts using historical data. Healthcare practitioners can choose the finest therapies for patients and the most effective ways to customize those treatments to meet their unique needs by using predictive analytics. Additionally, patients who are at risk for complications or relapse can be identified using predictive healthcare analytics, and interventions can be given before issues arise. Predictive analytics has the ability to boost the effectiveness and quality of healthcare services overall.

Wearable technology have been widely employed in the health industry for a variety of purposes, including patient care and personal health.The number of well-known consumer and medical devices that incorporate wearable sensor technologies has gradually increased. In situations involving the elderly, rehabilitation, and people with different disabilities, wearable devices can offer real-time input about a person's health problems. As a result, they can offer an objective alternative to manage and monitor the progression of chronic diseases. The vital indicators such as heart rate, blood pressure, and body temperature are the most often monitored data.

Data Collection

The objective of this project is to determine whether commercial wearable technology can reliably forecast lying, sitting, and various other levels of physical activity. The dataset obtained was from Harvard Dataverse, An experiment was performed where a sample of 46 participants were taken, 26 of these were female. Three different types of devices used for the data are GENEActiv, an Apple Watch, and a Fitbit Charge. Each participant completed a 65-minute regimen that included 25 minutes of relaxing or resting and 40 total minutes on the treadmill. The amount of energy expended was measured using indirect calorimetry.

Source of the Data Set

The given data set has been obtained from the Harvard Dataverse. Let's talk about the attributes of the data set obtained:

1. X1: Serial Number
2. Age: Age of every participant in the sample
3. Gender: Gender of every participant in the sample expressed in terms of "1" and "0" for "Male" and "Female" respectively.

Variables:

- **age**
- **gender :** Female & Male
- **height :** cm
- **weight :** kg
- **steps :** steps/mins
- **calories**
- **distance :** in meters
- **entropy_heart :** Heart rate entropy is used as a commonly used parameter to describe the regularity of the heart rate in the data set.
- **entropy_steps :** The entropy of steps is used as a commonly used parameter to describe the regularity of the steps in the data set.
- **resting_heart :** A normal resting heart rate for adults ranges from 60 to 100 beats per minute. Generally, a lower heart rate at rest implies more efficient heart function and better cardiovascular fitness.
- **corr_heart_steps :** Heart rate/step correlation;This column provides the relation between heart rate and steps for the particular activity.
- **intensity_karvonen :** The Karvonen formula is your heart rate reserve multiplied by the percentage of intensity plus your resting heart rate.
- **sd_norm_heart :** A standard deviation (or) is a measure of how dispersed the data is in relation to the mean.
- **device:**Apple watch & Fitbit

- **activity:** Lying ,Running 3 METS , Running 5 METS ,Running 7 METS,Self Pace walk and Sitting.

Data Cleaning

Let's look at first few rows of our data frame:

```
head(participants_data)
```

```
  X X1 age gender height weight    steps hear_rate   calories    distance
1 1  1  20      1    168   65.4 10.77143  78.53130  0.3445329 0.008326857
2 2  2  20      1    168   65.4 11.47532  78.45339  3.2876255 0.008896346
3 3  3  20      1    168   65.4 12.17922  78.54083  9.4840000 0.009465835
4 4  4  20      1    168   65.4 12.88312  78.62826 10.1545556 0.010035325
5 5  5  20      1    168   65.4 13.58701  78.71569 10.8251111 0.010604814
6 6  6  20      1    168   65.4 14.29091  78.80313 11.4956667 0.011174303
  entropy_heart entropy_setps resting_heart corr_heart_steps norm_heart
1      6.221612      6.116349            59        1.0000000   19.53130
2      6.221612      6.116349            59        1.0000000   19.45339
3      6.221612      6.116349            59        1.0000000   19.54083
4      6.221612      6.116349            59        1.0000000   19.62826
5      6.221612      6.116349            59        0.9828157   19.71569
6      6.221612      6.116349            59        1.0000000   19.80313
  intensity_karvonen sd_norm_heart steps_times_distance      device activity
1          0.1385199      1.000000           0.08969215 apple watch    Lying
2          0.1379673      1.000000           0.10208846 apple watch    Lying
3          0.1385874      1.000000           0.11528650 apple watch    Lying
4          0.1392075      1.000000           0.12928626 apple watch    Lying
5          0.1398276      0.241567           0.14408774 apple watch    Lying
6          0.1404477      0.264722           0.15969095 apple watch  Sitting
```

For the given data set we begin by removing some rows that contain too many (\>10%) NA values for both qualitative and quantitative variables. We also will remove duplicate columns.

Let's first look at the structure of our data frame:

```
str(participants_data)
```

```
'data.frame':   6264 obs. of  20 variables:
 $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
$ X1                  : int  1 2 3 4 5 6 7 8 9 10 ...
$ age                 : int  20 20 20 20 20 20 20 20 20 20 ...
$ gender              : int  1 1 1 1 1 1 1 1 1 1 ...
$ height              : num  168 168 168 168 168 168 168 168 168 168 ...
$ weight              : num  65.4 65.4 65.4 65.4 65.4 65.4 65.4 65.4 65.4 65.4 ...
$ steps               : num  10.8 11.5 12.2 12.9 13.6 ...
$ hear_rate           : num  78.5 78.5 78.5 78.6 78.7 ...
$ calories            : num  0.345 3.288 9.484 10.155 10.825 ...
$ distance            : num  0.00833 0.0089 0.00947 0.01004 0.0106 ...
$ entropy_heart       : num  6.22 6.22 6.22 6.22 6.22 ...
$ entropy_setps       : num  6.12 6.12 6.12 6.12 6.12 ...
$ resting_heart       : num  59 59 59 59 59 59 59 59 59 59 ...
$ corr_heart_steps    : num  1 1 1 1 0.983 ...
$ norm_heart          : num  19.5 19.5 19.5 19.6 19.7 ...
$ intensity_karvonen  : num  0.139 0.138 0.139 0.139 0.14 ...
$ sd_norm_heart       : num  1 1 1 1 0.242 ...
$ steps_times_distance: num  0.0897 0.1021 0.1153 0.1293 0.1441 ...
$ device              : chr  "apple watch" "apple watch" "apple watch" "apple watch" ...
$ activity            : chr  "Lying" "Lying" "Lying" "Lying" ...
```

```r
summary(participants_data)
```

```
       X               X1              age            gender      
 Min.   :   1   Min.   :   1.0   Min.   :18.00   Min.   :0.0000  
 1st Qu.:1567   1st Qu.: 789.8   1st Qu.:23.00   1st Qu.:0.0000  
 Median :3132   Median :1720.0   Median :28.00   Median :0.0000  
 Mean   :3132   Mean   :1771.1   Mean   :29.16   Mean   :0.4765  
 3rd Qu.:4698   3rd Qu.:2759.2   3rd Qu.:33.00   3rd Qu.:1.0000  
 Max.   :6264   Max.   :3670.0   Max.   :56.00   Max.   :1.0000  
     height          weight           steps           hear_rate      
 Min.   :143.0   Min.   : 43.00   Min.   :   1.00   Min.   :  2.222  
 1st Qu.:160.0   1st Qu.: 60.00   1st Qu.:   5.16   1st Qu.: 75.598  
 Median :168.0   Median : 68.00   Median :  10.09   Median : 77.268  
 Mean   :169.7   Mean   : 69.61   Mean   : 109.56   Mean   : 86.142  
 3rd Qu.:180.0   3rd Qu.: 77.30   3rd Qu.: 105.85   3rd Qu.: 95.669  
 Max.   :191.0   Max.   :115.00   Max.   :1714.00   Max.   :194.333  
    calories           distance        entropy_heart   entropy_setps  
 Min.   : 0.05627   Min.   :  0.0004   Min.   :0.000   Min.   :0.000  
 1st Qu.: 0.73587   1st Qu.:  0.0191   1st Qu.:6.109   1st Qu.:5.909  
 Median : 4.00000   Median :  0.1817   Median :6.190   Median :6.157  
 Mean   :19.47182   Mean   : 13.8326   Mean   :6.030   Mean   :5.740  
```

```
 3rd Qu.:20.50000   3rd Qu.: 15.6972   3rd Qu.:6.248   3rd Qu.:6.248
 Max.   :97.50000   Max.   :335.0000   Max.   :6.476   Max.   :6.476
 resting_heart    corr_heart_steps    norm_heart      intensity_karvonen
 Min.   :  3.00   Min.   :-1.0000   Min.   :-76.000   Min.   :-2.714286
 1st Qu.: 58.13   1st Qu.:-0.4673   1st Qu.:  1.149   1st Qu.: 0.009819
 Median : 75.00   Median : 0.6658   Median :  9.820   Median : 0.079529
 Mean   : 65.87   Mean   : 0.3064   Mean   : 20.272   Mean   : 0.155479
 3rd Qu.: 76.14   3rd Qu.: 1.0000   3rd Qu.: 27.077   3rd Qu.: 0.211868
 Max.   :155.00   Max.   : 1.0000   Max.   :156.319   Max.   : 1.297980
 sd_norm_heart     steps_times_distance    device            activity
 Min.   : 0.0000   Min.   :    0.00     Length:6264        Length:6264
 1st Qu.: 0.2647   1st Qu.:    0.66     Class :character   Class :character
 Median : 2.8935   Median :   13.37     Mode  :character   Mode  :character
 Mean   : 8.1108   Mean   :  590.04
 3rd Qu.: 9.6797   3rd Qu.:   93.73
 Max.   :74.4579   Max.   :51520.00
```

Let's look at the names of columns in our data frame and understand if they are in human readable format or not:

```
colnames(participants_data)
```

```
 [1] "X"                    "X1"                   "age"
 [4] "gender"               "height"               "weight"
 [7] "steps"                "hear_rate"            "calories"
[10] "distance"             "entropy_heart"        "entropy_setps"
[13] "resting_heart"        "corr_heart_steps"     "norm_heart"
[16] "intensity_karvonen"   "sd_norm_heart"        "steps_times_distance"
[19] "device"               "activity"
```

As we can see, there are two column names that are X1 and hear_rate that doesn't make any sense, we will proceed to replace X1 with ID and hear_rate with heart_rate.

```
names(participants_data)[2] <- 'ID'
names(participants_data)[8] <- "heart_rate"
names(participants_data)[12] <- "entropy_steps"
```

Let's look at it again:

```
colnames(participants_data)
```

```
 [1] "X"                    "ID"                   "age"
 [4] "gender"               "height"               "weight"
 [7] "steps"                "heart_rate"           "calories"
[10] "distance"             "entropy_heart"        "entropy_steps"
[13] "resting_heart"        "corr_heart_steps"     "norm_heart"
[16] "intensity_karvonen"   "sd_norm_heart"        "steps_times_distance"
[19] "device"               "activity"
```

```
#head(participants_data)
```

Checking for null values in our data frame

```
sum(is.null(participants_data))
```

```
[1] 0
```

Let's take a look at the dimension of our data frame before removing any duplicate values:

```
print(paste(c("Rows: ","Columns: "),dim(participants_data)))
```

```
[1] "Rows:  6264"  "Columns:  20"
```

After removing duplicate rows:

```
new_participants_data<-distinct(participants_data)
head(new_participants_data)
```

```
  X ID age gender height weight    steps heart_rate   calories    distance
1 1  1  20      1    168   65.4 10.77143   78.53130  0.3445329 0.008326857
2 2  2  20      1    168   65.4 11.47532   78.45339  3.2876255 0.008896346
3 3  3  20      1    168   65.4 12.17922   78.54083  9.4840000 0.009465835
4 4  4  20      1    168   65.4 12.88312   78.62826 10.1545556 0.010035325
5 5  5  20      1    168   65.4 13.58701   78.71569 10.8251111 0.010604814
6 6  6  20      1    168   65.4 14.29091   78.80313 11.4956667 0.011174303
  entropy_heart entropy_steps resting_heart corr_heart_steps norm_heart
1      6.221612      6.116349            59        1.0000000   19.53130
2      6.221612      6.116349            59        1.0000000   19.45339
3      6.221612      6.116349            59        1.0000000   19.54083
4      6.221612      6.116349            59        1.0000000   19.62826
```

```
5      6.221612      6.116349            59        0.9828157   19.71569
6      6.221612      6.116349            59        1.0000000   19.80313
  intensity_karvonen sd_norm_heart steps_times_distance      device activity
1          0.1385199      1.000000           0.08969215 apple watch    Lying
2          0.1379673      1.000000           0.10208846 apple watch    Lying
3          0.1385874      1.000000           0.11528650 apple watch    Lying
4          0.1392075      1.000000           0.12928626 apple watch    Lying
5          0.1398276      0.241567           0.14408774 apple watch    Lying
6          0.1404477      0.264722           0.15969095 apple watch  Sitting
```

```
print(paste(c("Rows: ","Columns: "),dim(new_participants_data)))
```

```
[1] "Rows:  6264"  "Columns:  20"
```

Since, we have "1" and "0" for our gender, for our ease we will change it to "Male" and "Female"

```
new_participants_data$gender[new_participants_data$gender == 0] <- "Female"
new_participants_data$gender[new_participants_data$gender == 1] <- "Male"

tail(new_participants_data)
```

```
        X   ID age gender height weight steps heart_rate calories distance
6259 6259 3602  46 Female  157.5   71.4     1         35      1.0        1
6260 6260 3666  46 Female  157.5   71.4     1         35     20.5        1
6261 6261 3667  46 Female  157.5   71.4     1         35     20.5        1
6262 6262 3668  46 Female  157.5   71.4     1         35     20.5        1
6263 6263 3669  46 Female  157.5   71.4     1         35     20.5        1
6264 6264 3670  46 Female  157.5   71.4     1         35     20.5        1
     entropy_heart entropy_steps resting_heart corr_heart_steps norm_heart
6259             0             0            35                1          0
6260             0             0            35                1          0
6261             0             0            35                1          0
6262             0             0            35                1          0
6263             0             0            35                1          0
6264             0             0            35                1          0
     intensity_karvonen sd_norm_heart steps_times_distance device
6259                  0      25.07234                   1 fitbit
6260                  0       0.00000                   1 fitbit
6261                  0       1.00000                   1 fitbit
```

```
6262                 0       1.00000                   1 fitbit
6263                 0       1.00000                   1 fitbit
6264                 0       1.00000                   1 fitbit
          activity
6259          Lying
6260 Running 7 METs
6261 Running 7 METs
6262 Running 7 METs
6263 Running 7 METs
6264 Running 7 METs
```

Segregating the participants who used Apple watch and Fit bit watch into two different data frames:

```
participants_data_apple<-new_participants_data%>%group_by(device)%>%filter(device=="apple
participants_data_fitbit<-new_participants_data%>%group_by(device)%>%filter(device=="fitbi
```

EDA

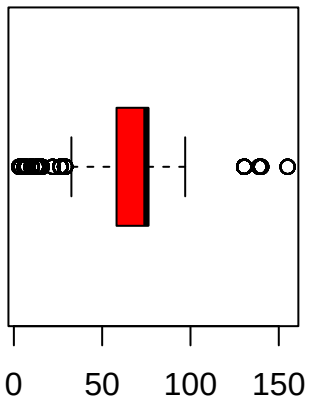
Let's explore a bit more in depth:

Checking for outliers in different columns:

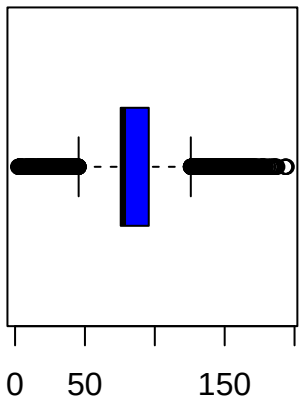
```
par(mfrow=c(1,2))

boxplot(new_participants_data$resting_heart,col="red",
        xlab = "Resting Heart rate",
        horizontal = TRUE)

boxplot(new_participants_data$heart_rate,col="blue",
        xlab = "Heart rate",
        horizontal = TRUE)
```



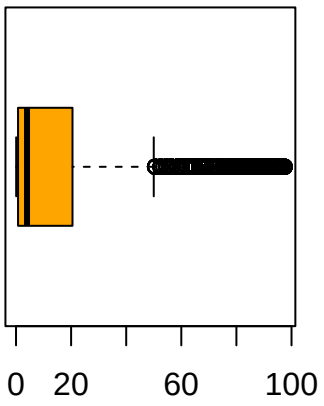
Resting Heart rate

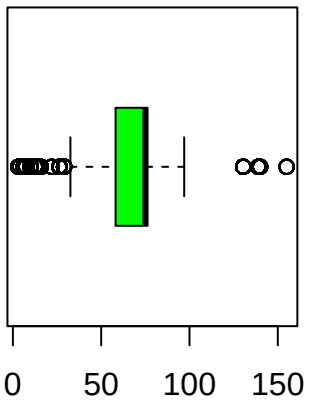


Heart rate

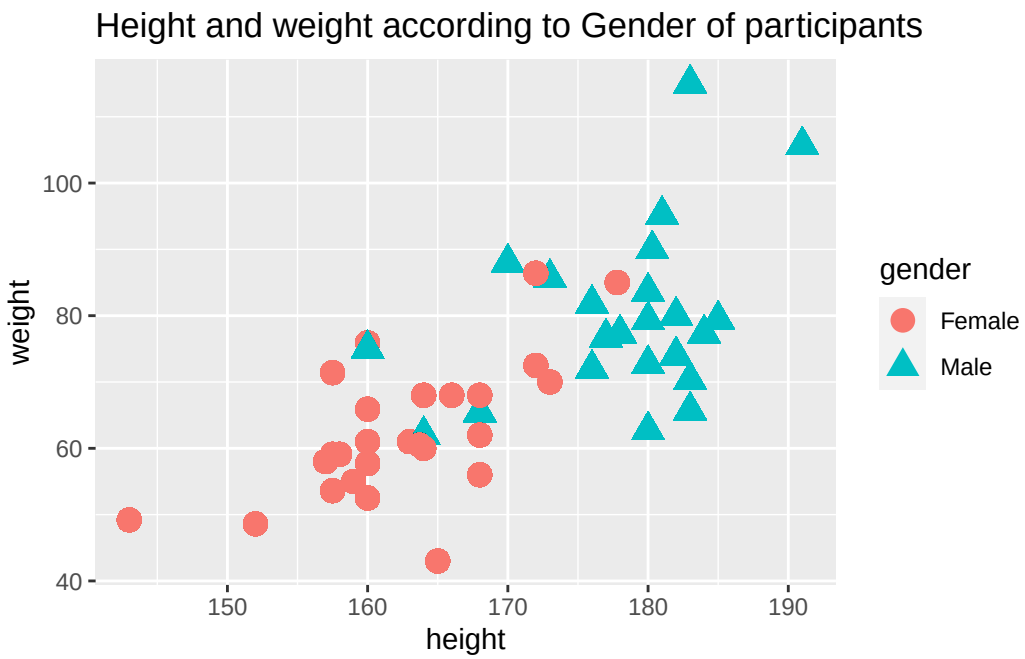
```
boxplot(new_participants_data$calories,col="orange",
        xlab = "Calories",
        horizontal = TRUE)

boxplot(new_participants_data$resting_heart,col="green",
        xlab = "Entropy Heart rate",
        horizontal = TRUE)
```





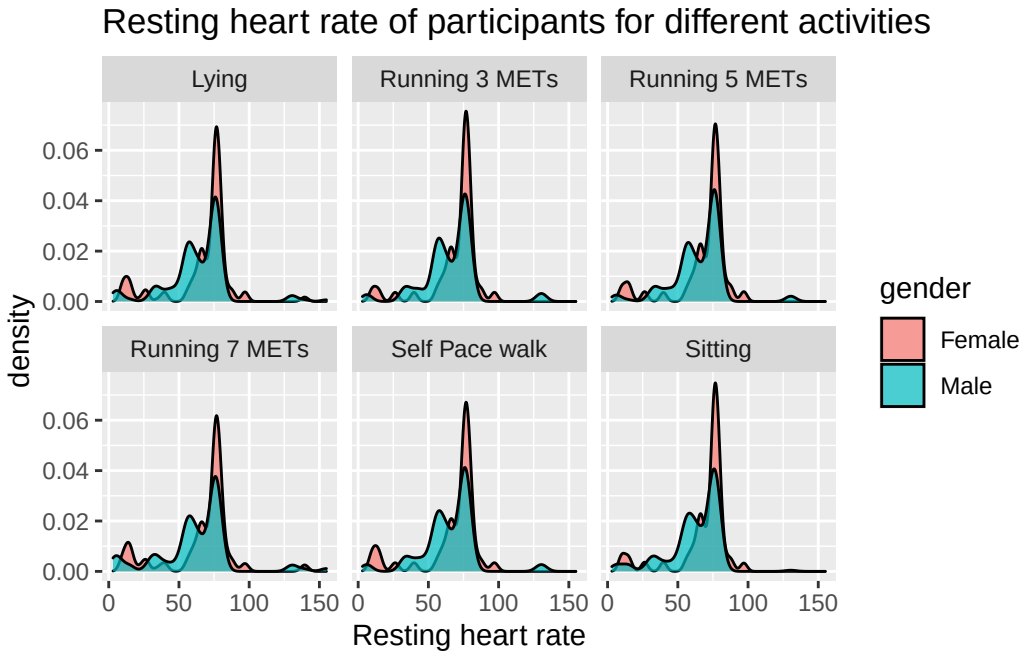

```
ggplot(new_participants_data, aes(x=height, y=weight, color=gender, shape=gender)) +
    geom_point(size=4) +
    labs(title='Height and weight according to Gender of participants')
```



From the above visualization, we can get insights on the height and weight of both Male and Female participants. We can extract the information that says Males in general have greater height and weight as compared to Females. In such case, we can assume a lot of things like, probably they will burn more calories while on treadmill. Yet, another assumption can be something like, they will have a greater heart rate and so on.

No gender based discrimination is intended.

```
ggplot(new_participants_data, aes(x=resting_heart, fill=gender)) + geom_density(alpha=0.7)
    labs(title='Resting heart rate of participants for different activities',
         x='Resting heart rate'
        )
```



The above graphs showcases the heart rate of participants across various activities performed like Lying down, Sitting, running over treadmill for varying speed and so on. We can capture the insight that says female participants have much more heart rate as compared to male participants. Moreover, there are few that have heart rate closer to 0 which is practically not possible, so we can label them as outliers.

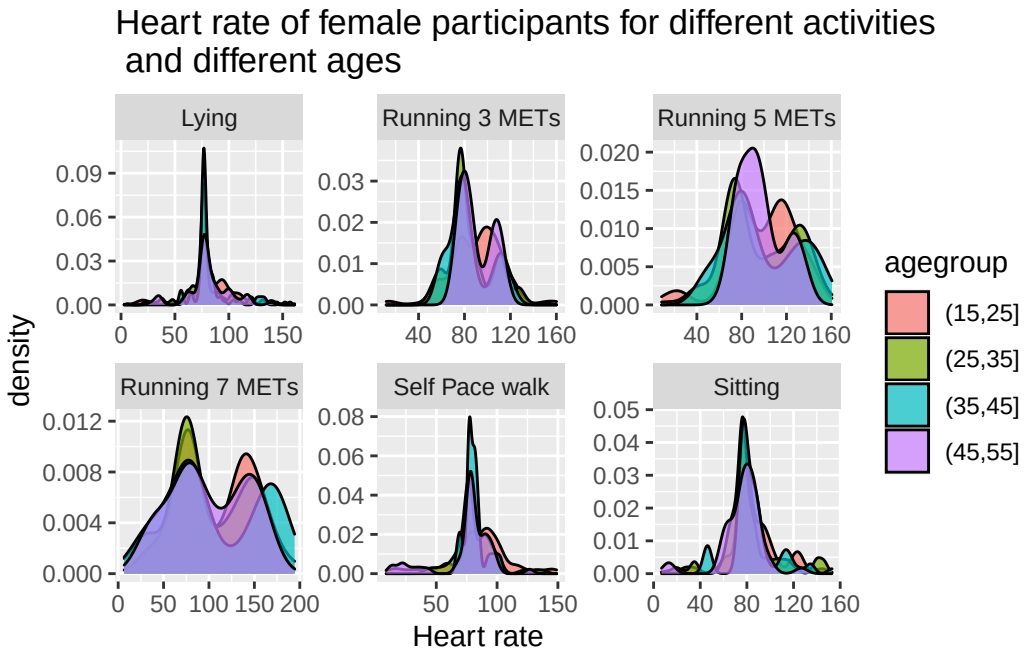
```
ggplot(participants_data_apple, aes(x=heart_rate, y=intensity_karvonen,group_by(device) ,c
    geom_point(size=4) +
    facet_wrap(~activity) +
    labs(title='Heart rate and Intensity vs genre of participants')
```


```
new_participants_data$agegroup = cut(new_participants_data$age,c(15,25,35,45,55,65))
head(new_participants_data)
```

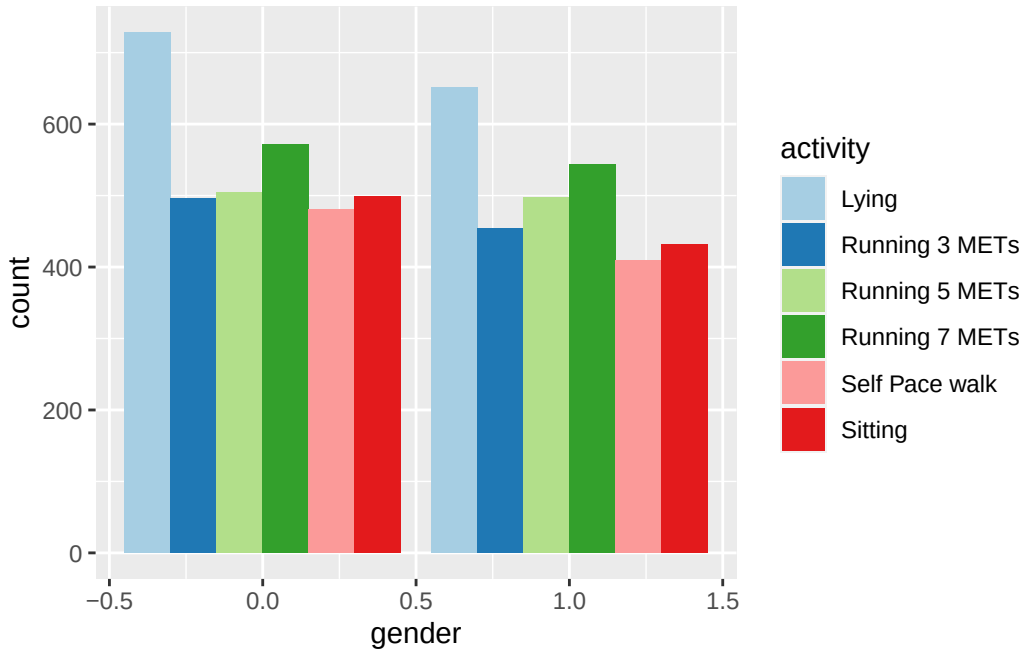
```
  X ID age gender height weight    steps heart_rate   calories    distance
1 1  1  20   Male    168   65.4 10.77143   78.53130  0.3445329 0.008326857
2 2  2  20   Male    168   65.4 11.47532   78.45339  3.2876255 0.008896346
3 3  3  20   Male    168   65.4 12.17922   78.54083  9.4840000 0.009465835
4 4  4  20   Male    168   65.4 12.88312   78.62826 10.1545556 0.010035325
5 5  5  20   Male    168   65.4 13.58701   78.71569 10.8251111 0.010604814
6 6  6  20   Male    168   65.4 14.29091   78.80313 11.4956667 0.011174303
  entropy_heart entropy_steps resting_heart corr_heart_steps norm_heart
1      6.221612      6.116349            59        1.0000000   19.53130
2      6.221612      6.116349            59        1.0000000   19.45339
3      6.221612      6.116349            59        1.0000000   19.54083
4      6.221612      6.116349            59        1.0000000   19.62826
5      6.221612      6.116349            59        0.9828157   19.71569
6      6.221612      6.116349            59        1.0000000   19.80313
  intensity_karvonen sd_norm_heart steps_times_distance      device activity
1          0.1385199      1.000000           0.08969215 apple watch    Lying
2          0.1379673      1.000000           0.10208846 apple watch    Lying
3          0.1385874      1.000000           0.11528650 apple watch    Lying
4          0.1392075      1.000000           0.12928626 apple watch    Lying
```

```
5          0.1398276      0.241567           0.14408774 apple watch    Lying
6          0.1404477      0.264722           0.15969095 apple watch  Sitting
  agegroup
1  (15,25]
2  (15,25]
3  (15,25]
4  (15,25]
5  (15,25]
6  (15,25]
```

```
fem_data <- new_participants_data %>% filter(gender=='Female')
ggplot(fem_data, aes(x=heart_rate, fill=agegroup)) +
    geom_density(alpha=0.7) +facet_wrap(~activity, scale='free') +
    labs(title='Heart rate of female participants for different activities\n and different
         x='Heart rate'
        )
```




```
ggplot(participants_data, aes(x = gender,
           fill = activity)) +
  geom_bar(position = "dodge")+
  scale_fill_brewer(palette = "Paired")
```



We can observe that females wearer chose high MET activities over self-paced walks during the 40-minute treadmill protocol, and they subsequently chose to lay down rather than sit.

Bias

The given data set taken from Harvard data verse contains data related to only 46 participants. Maybe, if we had more than 1000 participants, we would have reached a better conclusion. Moreover, there are various other activities as well that can be recorded in watches, such as rhythm (regular or irregular), ECG, Oxygen level etc. that can be used for further analysis. Moreover, sometimes the watches may not be 100% accurate, due to technical shortcomings giving us undesired results.

Conclusion

To conclude,we can observe that the heart rate is consistent in both apple watch and fitbit for the different physical activities performed.We can also observe that the commercial watches concentrate more on features like heart rate ,steps and calories .We can also observe that there is increase in calories burnt depending on the activity for example running.This observation is crucial because all the parameters used in the smart watch is dependent on the calories parameter.Overall we can observe that females chose high MET activities over self-paced walks

during the 40-minute treadmill protocol, and they subsequently chose to lay down rather than sit.

We utilized different smartwatches in the visualizations to see how the smart watches behave based on our activities to conclude this study. Smartwatches are becoming more and more well-liked because of how convenient and portable they are. Many of them monitor their health using a single smart device to calculate calories, track their workout, and more. The smart watches tracks and alerts users for features like medication reminders, fall detection, and information on your heart rate, sleep, and location around-the-clock.