Case Study 2: How Can a Wellness Technology Company Play It Smart?

Busisiwe C Ringane - Google Data Analytics Capstone - Portfolio-Ready Case Study 31/11/2021

Executive Summary

The aim of the exercise is to give insight into how users of Bellbear smart watch use the smart watch. The data reflect a good positive correlation between Calories burned, and distance walked/run. As users run or walk longer distances, the calories burned increases. This translated to a strong correlation with total Steps taken and Calories (r^2 =0.97) . Clearly amount of activity has a direct impact on calories burned. The data was split into three groups: less than 5km, mid distance 5 – 10, and long distance which is

greater than 10 km. There is a difference in calories burned between the groups, with calories burned increasing with distance.

We can also see some overall weekly trends, such as high activity on Tuesday and Saturday and low activity on Sunday and Wednesday.

Sleep data shows that average sleeping range is 6 hours – 9 hours.

1. Introduction and Background

This project is part of Coursesa Google Data Analytics course. The case study is based on Bellabeat data. Bellabeat, is a high-tech company that manufactures health-focused smart products. The product aims at informing and inspiring women around the world to lead a healthy lifestyle. Data collected includes steps taken, hear rate, sleeping patterns, stress, and reproductive health thus empowering women with knowledge about their own health and habits.

The aim of the project is to analyse the data collected from smart device data to gain insight into how consumers are using their smart devices. The insights discovered will then help guide marketing strategy for the company.

Key Stakeholders: The key stakeholders are (1)Urška Sršen: Bellabeat's cofounder and Chief Creative Officer Sando Mur: Mathematician, Bellabeat's cofounder and key member of the Bellabeat executive team Bellabeat marketing analytics team and (2) A team of data analysts guiding Bellabeat's marketing strategy.

Background Scenario: Bellabeat is a successful small company, but they have the potential to become a larger player in the market of health smart devices. Urška Sršenbelieves that analysing smart device fitness data could help unlock new growth opportunities for the company.

Using the Case Study Roadmap as a guide, this analysis will follow the steps of the data analysis process: **Ask Prepare, Process, Analyse, Share, and Act.** A brief summary is given below (**Table 1**).

Step	Requirements
Ask	Understanding of the problem being resolved
Prepare	what data do I need
Process	Cleaning of the data
Analyse	Sort and format your data to make it easier to: Perform calculations, combine data from multiple sources, Create tables with your results
Share	Communicate data effectively
Act	Provide recommendation and data driven decision

Table 1. The Step for data analysis

Step 1: Ask

In this step define the problem and objectives of our case study and its desired outcome.

The **Business questions** and objectives are:

What are some trends in smart device usage?

How could these trends help influence Bellabeat marketing strategy?

Step 2: Prepare

In the preparation phase, the source of data is being identified and its limitations. We are using Python for data cleaning, transformation and visualisation.

2.1 Data Source:

The data was extracted from https://www.kaggle.com/arashnic/fitbit as csv files. Data is publicly available and stored in 18 csv files. This raw dataset will be stored locally for cleaning and processing via RStudio and Microsoft Excel .

2.2 Limitations of Data Set:

The following limitations were identified:

- Data may not be timely or relevant Data was collected 5 years ago in 2016. Users' daily activity, fitness and sleeping habits, diet and food consumption may have changed since then.
- Sample size not representative of the entire fitness population.
- As data is collected in a survey, we are unable to ascertain its integrity or accuracy.
- Data collection is not controlled or monitored; hence some variables do not have sufficient data. Data such as weight and BMI was inputted manually which is why this data was not recorded consistently for allusers. Sleep was also not consistently recorded which might mean that users removed their device during the night.
- The data did not disclose gender of the users.

2.3. Data ROCC

A good data source is ROCCC which stands for **R**eliable, **O**riginal, **C**omprehensive, **C**urrent, and **C**ited. The summary of the finding are

	Level of Confidence	Comments
Reliable	Low	Not reliable as less than 40 respondents
Original	Low.	Third party provider. Overall, this dataset is considered *"bad quality data"* and it **is not recommended** to produce business recommendations based on this data"
Comprehensive	Med	Parameters match most of Bellabeat products' parameters
Current	Low	Data is from 2016
Cited	Low	Data collected from third party, hence unknown

Table 2. ROCC for the dataset

2.4 Data Selection

The following file is selected and copied for analysis:

dailyActivity_merged.csv

dailyCalories_merged.csv

sleep_Day Meged.csv

Weight.csv

We'll follow typical naming conventions based on the csv file names. However each file will be prefixed with the **date when amendments were made**.

Step 3: Process

In order to process and view these datasets, the following packages were installed:

- tidyverse packages
- `ggplot2`
- `tibble`
- `tidyr`
- `readr`

Many of the datasets were very large files and could only be opened in R Studio or Postgres (SQL). The data will be cleaned and ensuring that it is correct, relevant, complete and free of error and outlier by performing Data cleaning. The data cleaning will be done in Python, SQL and Excel. The choice is dependent on the author.

For each file

- Check for missing or null values
- Transform data — format data type
- Perform preliminary statistical analysis.
- Date format concerns: There is lack of consistency in the data format. In some instances it included both the date includes the time (e.g. 4/13/2016 12:00:00 AM). Using the following line of code, we omit the time by converting to the %m/%d/%y format. As a matter of fact, we also ensure that the "daily_activity" data frame follows an identical format and is of the same data type (date or date-time) as well.

3.1. File name: DailyCalories_merged.csv

Original file had 940 records. The following amendments were done to the file used:

1. 16 records had a variation in tracker distance and total distance. These records were excluded from the calories file as it is uncertain if there is a correction factor within the application. This yielded a similar mean of 5.408
2. One observation had 36019 steps and 2690 steps. This correlate with the high distance of 28 units, though it is an outlier it is acceptable.
3. Few records had calories of 0, no activity is associated with these records, and this is expected.

4. The date format varied within the file, for normality and consistency, In SQL, the extract function was used to extract the dates and ensure the data has same format. The final dataset has 925 records; summary is shown below (**Table 3**):

```
       Id            ActivityDate        DayofWeek          TotalSteps    TotalDistance
 Min.   :1.504e+09   Length:925         Length:925         Min.   :    0   Min.   : 0.000
 1st Qu.:2.320e+09   Class :character   Class :character   1st Qu.: 3758   1st Qu.: 2.600
 Median :4.445e+09   Mode  :character   Mode  :character   Median : 7328   Median : 5.180
 Mean   :4.821e+09                                         Mean   : 7530   Mean   : 5.408
 3rd Qu.:6.962e+09                                         3rd Qu.:10602   3rd Qu.: 7.570
 Max.   :8.878e+09                                         Max.   :36019   Max.   :28.030
 TrackerDistance     X           LoggedActivitiesDistance VeryActiveDistance
 Min.   : 0.000   Mode:logical   Min.   :0.00000          Min.   : 0.00
 1st Qu.: 2.600   TRUE:925       1st Qu.:0.00000          1st Qu.: 0.00
 Median : 5.180                  Median :0.00000          Median : 0.19
 Mean   : 5.408                  Mean   :0.03952          Mean   : 1.46
 3rd Qu.: 7.570                  3rd Qu.:0.00000          3rd Qu.: 1.94
 Max.   :28.030                  Max.   :2.25308          Max.   :21.92
 ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
 Min.   :0.0000           Min.   : 0.00       Min.   :0.000000        Min.   :  0.00
 1st Qu.:0.0000           1st Qu.: 1.91       1st Qu.:0.000000        1st Qu.:  0.00
 Median :0.2300           Median : 3.31       Median :0.000000        Median :  3.00
 Mean   :0.5585           Mean   : 3.31       Mean   :0.001384        Mean   : 20.68
 3rd Qu.:0.7900           3rd Qu.: 4.73       3rd Qu.:0.000000        3rd Qu.: 30.00
 Max.   :6.4800           Max.   :10.71       Max.   :0.070000        Max.   :210.00
 FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes    Calories
 Min.   :  0.00      Min.   :  0          Min.   :   0.0   Min.   :   0
 1st Qu.:  0.00      1st Qu.:125          1st Qu.: 729.0   1st Qu.:1821
 Median :  6.00      Median :196          Median :1061.0   Median :2124
 Mean   : 13.42      Mean   :191          Mean   : 991.6   Mean   :2296
 3rd Qu.: 19.00      3rd Qu.:262          3rd Qu.:1234.0   3rd Qu.:2782
 Max.   :143.00      Max.   :518          Max.   :1440.0   Max.   :4900
```

Table 3. The summary of Daily Calories file

3.2. File name: dailyActivity_merged.csv

Original file had 416 records. The following amendments were done to the file used:

1. Three records had duplicated dates and Hours. These were deleted
2. No record was found to have sleep time being greater than in bed.
3. In SQL, the hours slept and hours in bed were calculated from Total Minures Asleep and Total Minutes in bed.
4. Both data set were loaded in R Studio for analysis.
5. Following these cleaning steps, our data frames are ready for the Analyze step.

```
> summary(sleep)
       Id              SleepDay              Day          TotalSleepRecords TotalMinutesAsleep
 Min.   :1.504e+09   Length:413         Mode:logical   Min.   :1.000     Min.   : 58.0
 1st Qu.:3.977e+09   Class :character   NA's:413       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
  Hours_Slept     TotalTimeInBed     Hrs_Bed
 Min.   : 0.970   Min.   : 61.0   Min.   : 1.020
 1st Qu.: 6.020   1st Qu.:403.0   1st Qu.: 6.720
 Median : 7.220   Median :463.0   Median : 7.720
 Mean   : 6.992   Mean   :458.6   Mean   : 7.644
 3rd Qu.: 8.170   3rd Qu.:526.0   3rd Qu.: 8.770
 Max.   :13.270   Max.   :961.0   Max.   :16.020
```

Table 4. Summary of daily Activitiy

Step 4: Analyze

Now that the data is stored appropriately and has been prepared for analysis we can start putting it to work. The following libraries were included:

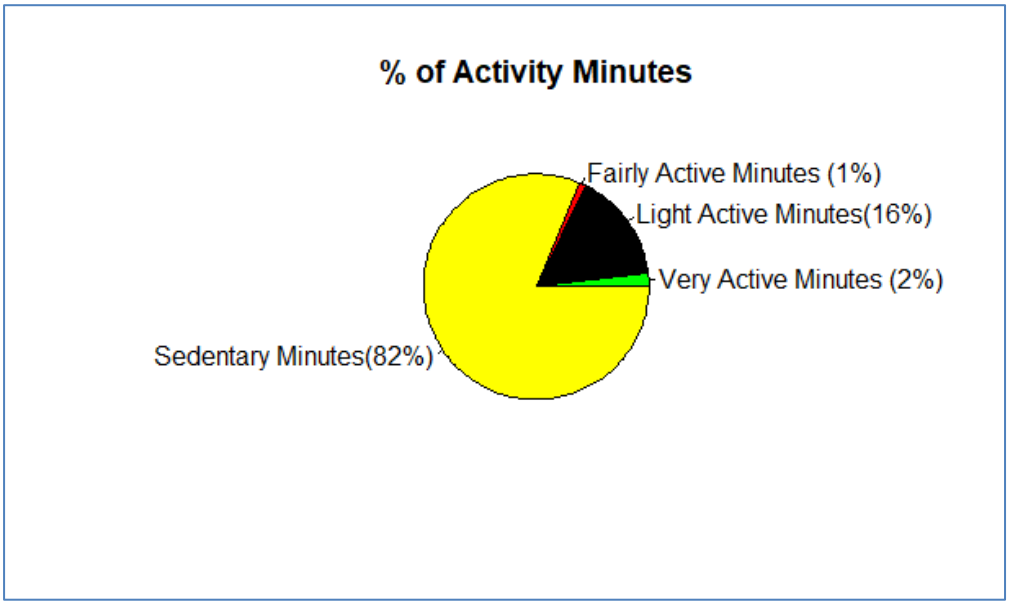
library(dplyr)

library(tidyr)

library(ggplot2)

4.1. Calories measurement

A wellness technology measures the number of calories you burn in a day this based on the steps you took through the day, plus any exercise you logged using the application. Less than 20% of the time is for very active, fairly active and lightly active. 82% of the time is when members are not active. Only 3 % is dedicated to the very active and fairly active. This is highly motivational the app was developed to encourage fitness. Users can determine how active they want to be.




```
 Pie chart in R
new_pies <- c(2,16,1,82)

# labels for the pie chart
new_labels <- c("Very Active Minutes (2%)","Light Active Minutes(16%)", "Fairly Active Minutes (1%)",  "Sedentary Minutes

# colors for pies
new_colors <- c("green", "black", "red", "yellow")
# print the pie chart
pie(new_pies, label = new_labels, main = "% of Activity Minutes",
    col = new_colors)

# explanation box below the chart
legend("", new_labels, fill = new_colors)
```

Figure 1. The activity minutes breakdown

Calories are burned when one is active even when one is an sedentary state. The estimates calories burned by first calculating your basal metabolic rate (BMR). This is the rate at which you naturally burn calories while you're perfectly sedentary, whether or not you exercise. It increases with level of activity.

The average calories burned is low on Thursday. This correlates with the low steps taken. Most calories were burned on Tuesday and Saturday. The histogram, shows the average calories burned of Bell app usage in terms of days of the week (see Error! Reference source not found.). They are high during working days as they include other activities such as daily commute and low over the weekend on Sunday. However it could imply that the users keep their tracker on during the week and forget to track their activity on the app Sunday.

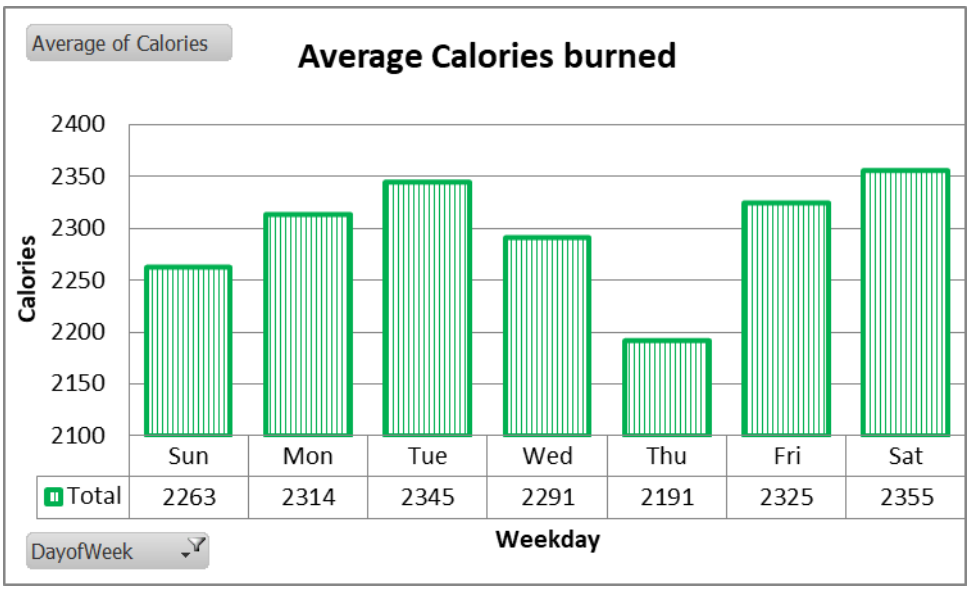


Figure 2. Average Calories burned

For analysis two variables are used daily Total distance and daily Calories. These two variables can be used to identify the relationship between total calories burnt by a user by walking specific miles on an average on daily basis.

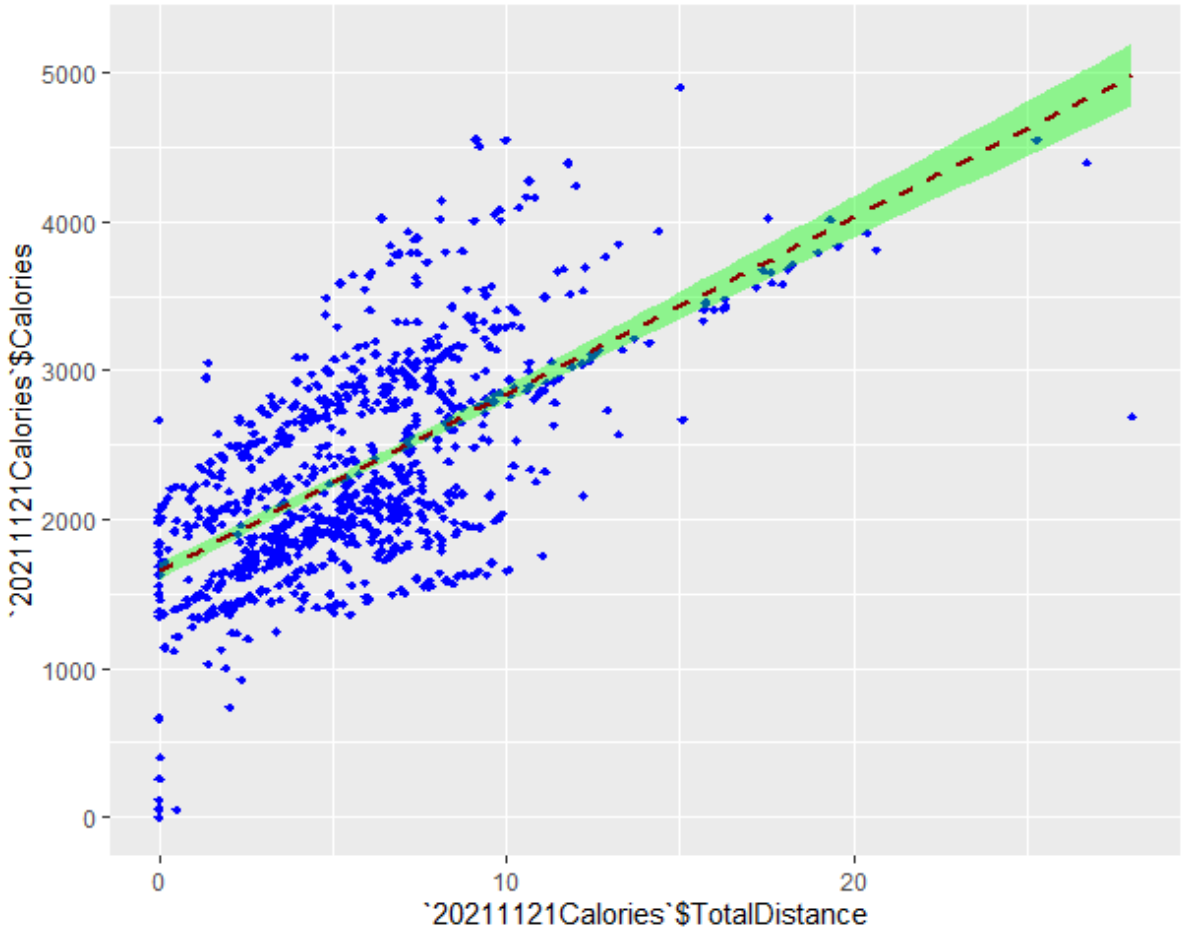


Figure 3. The Calories and Total distance

After plotting the daily total distance against the daily calories burnt, it was clear that there is a positive correlation between calories and total distance some users are walking or running shorter distances and burning higher calories. While some are walking or running long distances but burning fewer calories. Users with high-calorie loss are running faster and usually covering longer distances and users in the less cal-loss category are walking less distance, and users with mid-calorie loss are walking or slow running. To specifically get the idea of how much usually the users' burn calories while walking or traveling how much distance, let's categorize the data based on the amount of distance. Three groups were specified (**Figure 4**):

Distance (Group 1): < 5
Distance (Group 3) : 5 – 10
Distance (Group 2) : >10

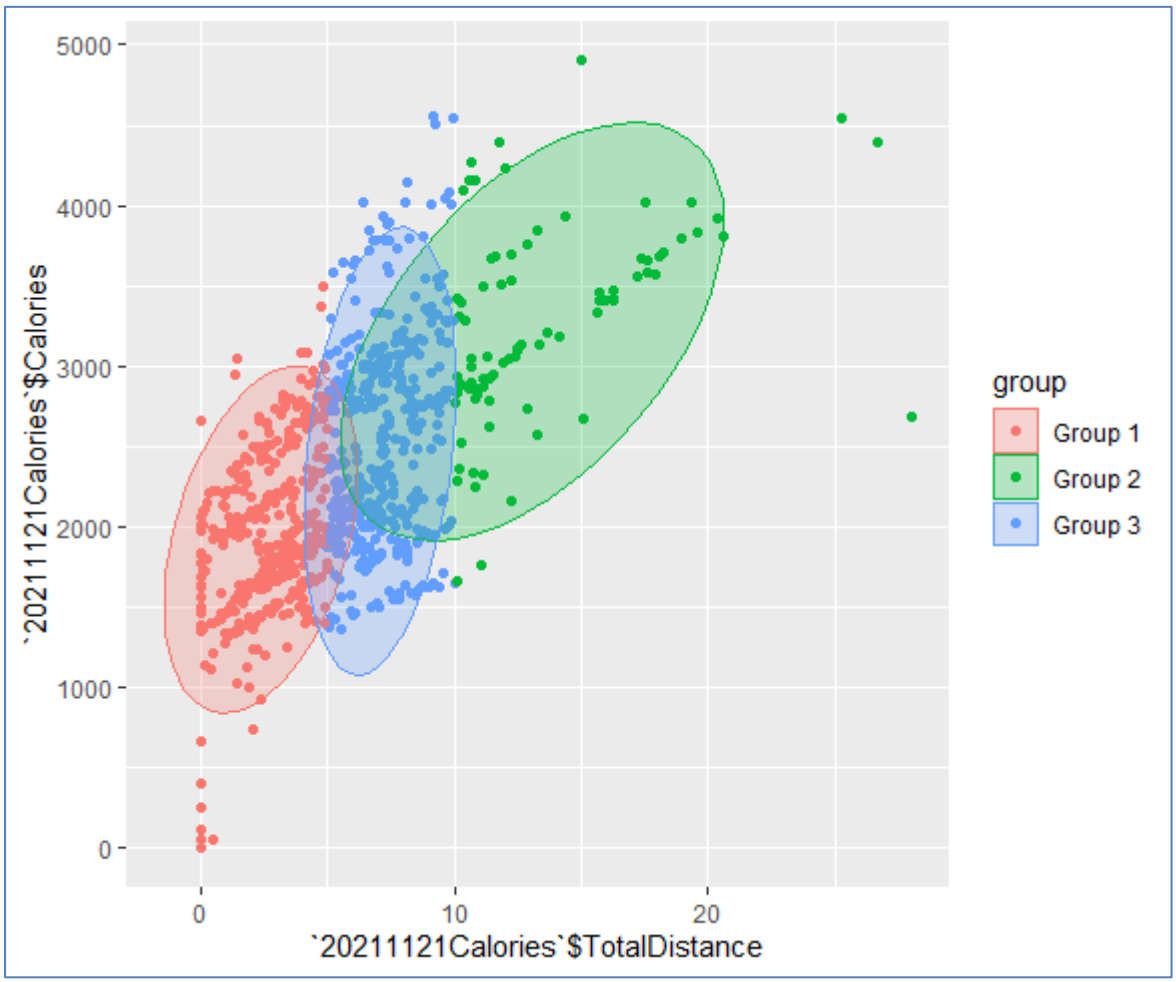
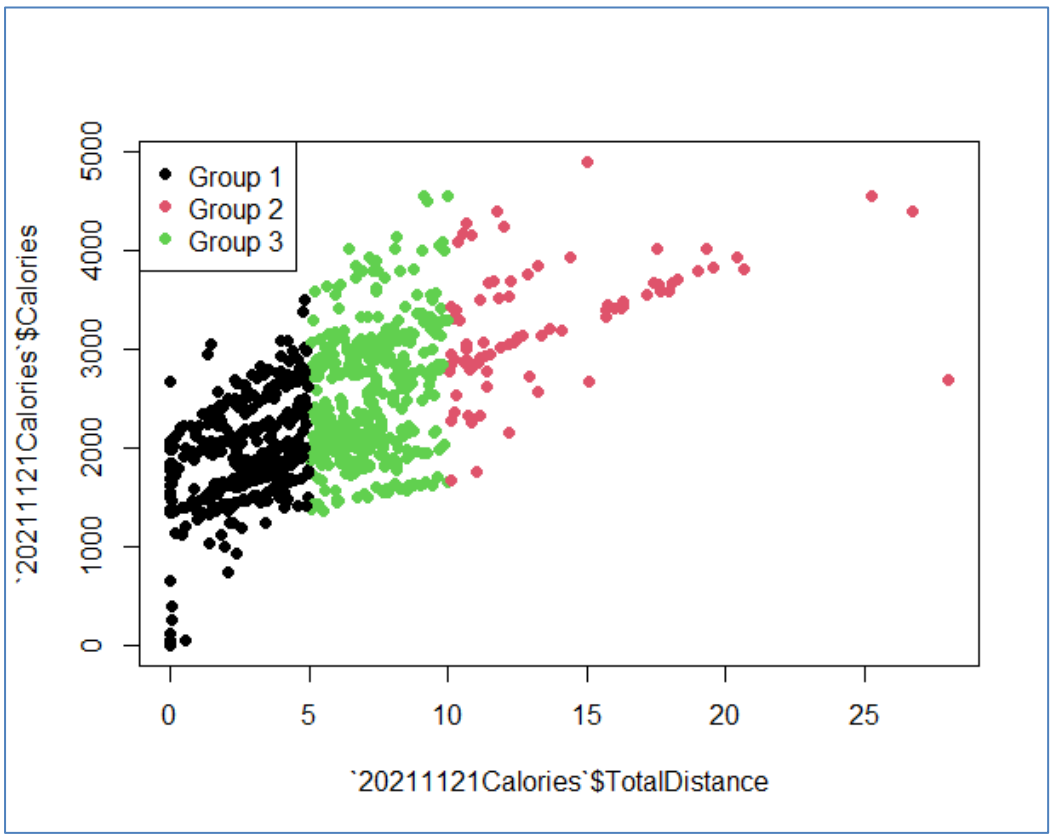


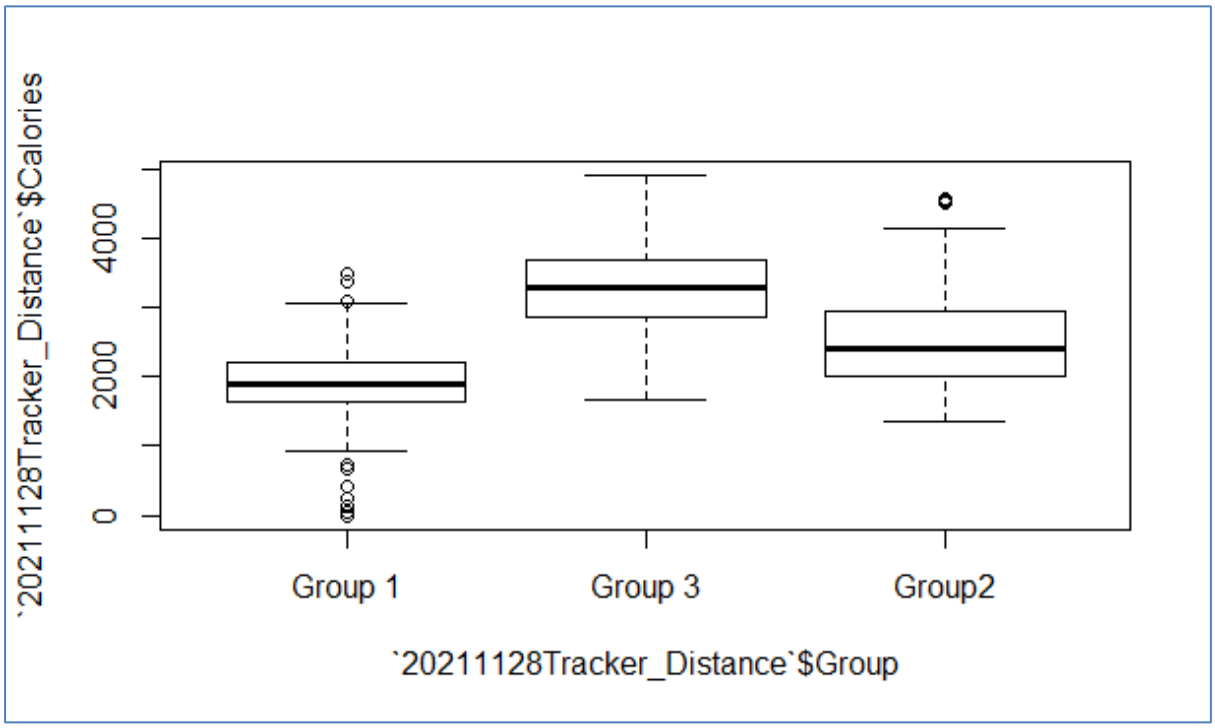
Figure 4. Grouped Calories and Total Distance



```
> ggplot(`20211121Calories`, aes(x = `20211121Calories`$TotalDistance, y = `20211121Calories`
$Calories, color = group)) +
+     geom_point() +
+     stat_ellipse(geom = "polygon",
+                  aes(fill = group),
+                  alpha = 0.25)
```

Figure 5. Scatter plot Grouped Calories and Total Distance

Based on the scatterplot graphs (**Figure 5**), it can be seen how and why some users are burning fewer calories than others. It can also be assumed based on the visuals that, users with high-calorie loss are running/walking and usually covering longer distances and users in the with low calories category are walking/running less distance, and users with mid-calorie loss are walking or running. The box plot below does highlight that there are different cluster based o distance run. The average calories burned are 1919, 2516 and 3276 per respective group. The longer the distance run, the higher the calories burned (**Figure 6**).



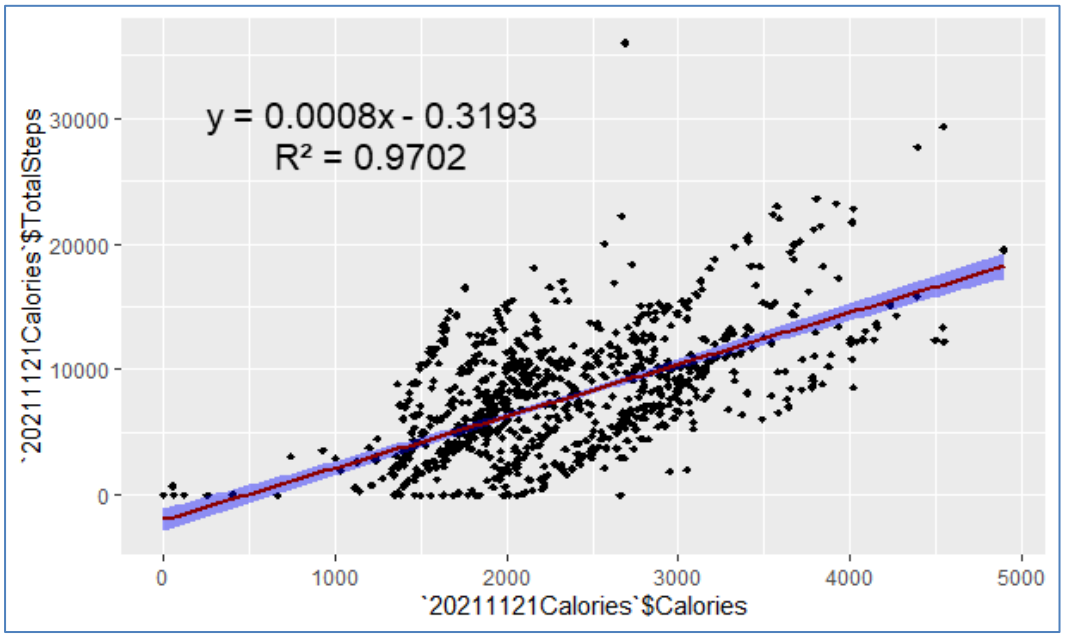

```
> boxplot(Tracker_Distance$Calories ~ Tracker_Distance$Group, data = Tracker_Distance, col = "white")
>
> # Points
> stripchart(x ~ group,
+            data = df,
+            method = "jitter",
+            pch = 19,
+            col = 2:4,
+            vertical = TRUE,
+            add = TRUE)
```

Figure 6. Box plot of Distance between the three groups

4.2. Daily Activity: Steps, Calories

The file used for the analysis is the dailyActivity_merged.csv.

There is a strong positive correlation of r = 0.97 between Steps and Calories. As more steps are taken. These steps are mostly taken on Tuesday.




```
> ggplot(`20211121Calories`, aes(x=`20211121Calories`$Calories, y=`20211121Calories`$Tota
lSteps)) +
+     geom_point(shape=18, color="black")+
+     geom_smooth(method=lm,  linetype="solid",
+                 color="darkred", fill="blue")
```

Figure 7. Calories and total step Scatter plot

4.3. Heart Rate monitor

The heart rate monitor was analysed in Microsoft Excel. Only 7 users had their hear rate monitored. Note that the recording for the 7 records were not done consistently among the 7 users.

The basic statistics extracted is shown below.

Row Labels	Min	Max	Average
User 1	38	203	80.24
User 2	63	125	93.78
User 3	49	195	76.72
User 4	46	191	82.30
User 5	39	180	66.13
User 6	44	199	81.67
User 7	47	106	62.77
Total/Average	**38**	**203**	**77.02**

Table 5 Heart rate summary

The average heartrate for different users is between 60 – 90, however there is a wider range of 30 up to 200 as shown in the histogram, below (Figure 8).

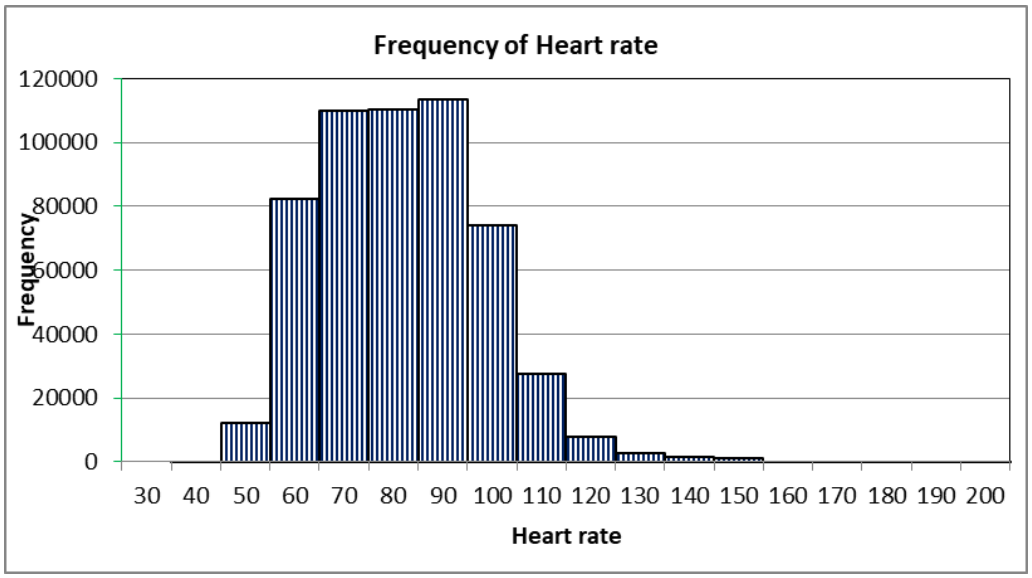


Figure 8. Heartrate histogram

The average line plots of the heart rate shows a less variable rate in the 15 days period where measurements were done.

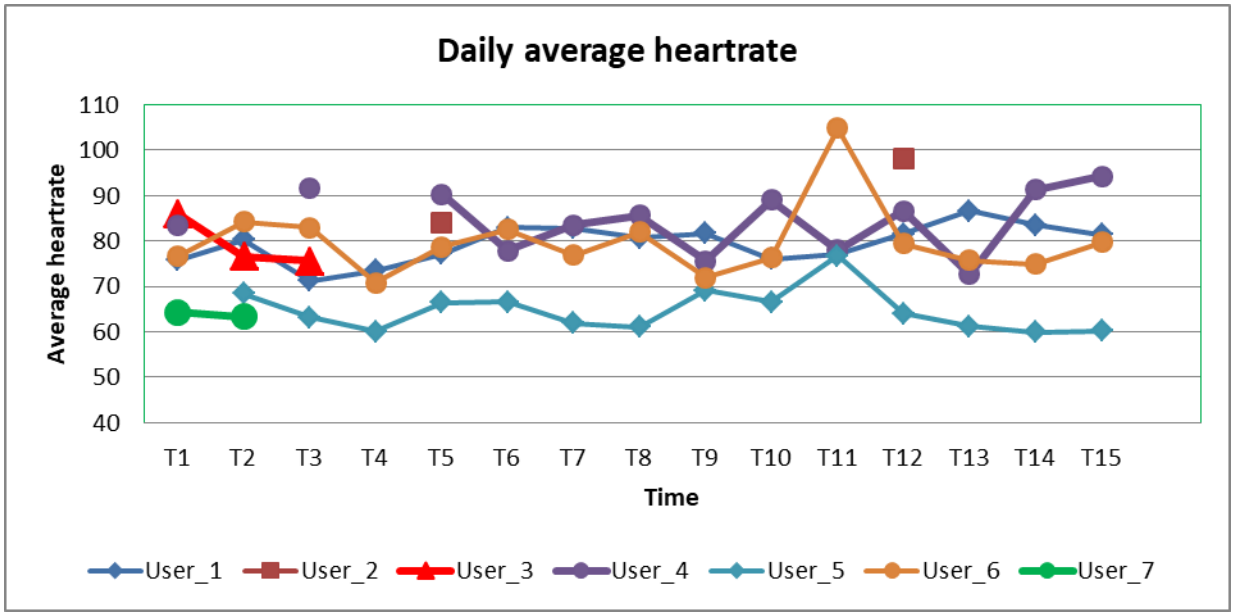


Figure 9. Line plot of the individual heart rate

A stream chart of the 7 individuals is shown below (**Figure 10),** recorded over 15 days labelled as T1 – T15. The plot reflects daily average. The inconsistent monitoring of the heart rate is clear with User 2, User 3 and User 7, where certain days were not recorded. Furthermore there is little variation in the average heart rate as shown by the thickness of the stream chart.

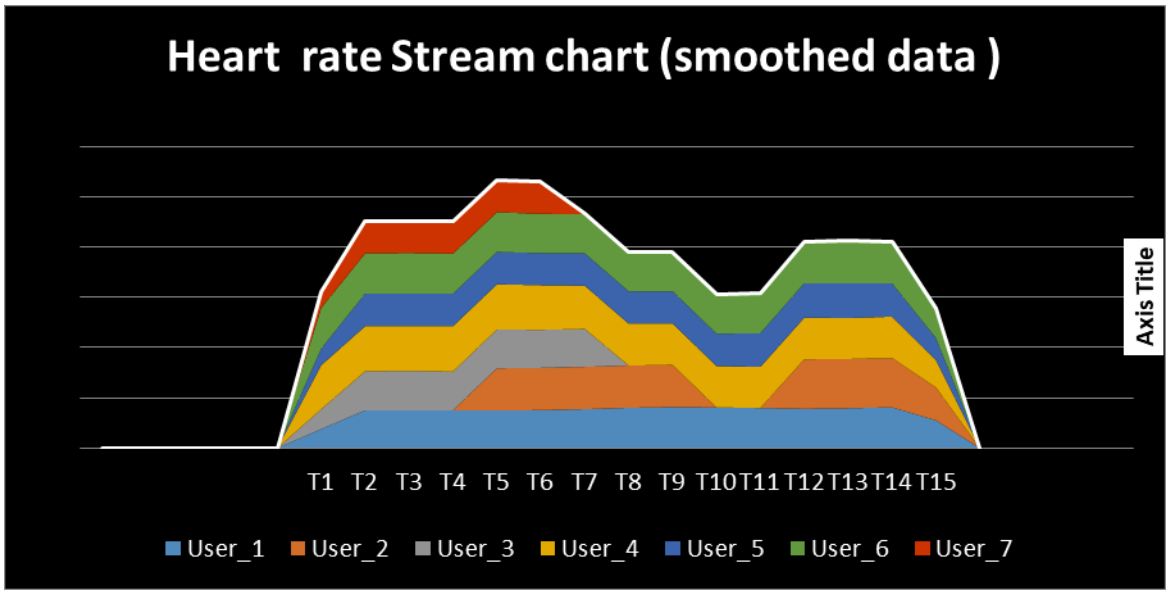


Figure 10. The heart rate streamline plot

4.4. Weight

Only 8 users recorded their weight, yielding 67 records. The summary of the weight is shown in **Figure 11** and **Table 6**. For comparison one has to see the change in weight with time for candidates, however the sample size is not sufficient to determine if there is a significant change in weight with time. The maximum change in weight is 2%, with most of the records indicating a zero change in weight. This is mainly due to the fact that the weight needs to be manually entered, hence majority of the users are not doing it.

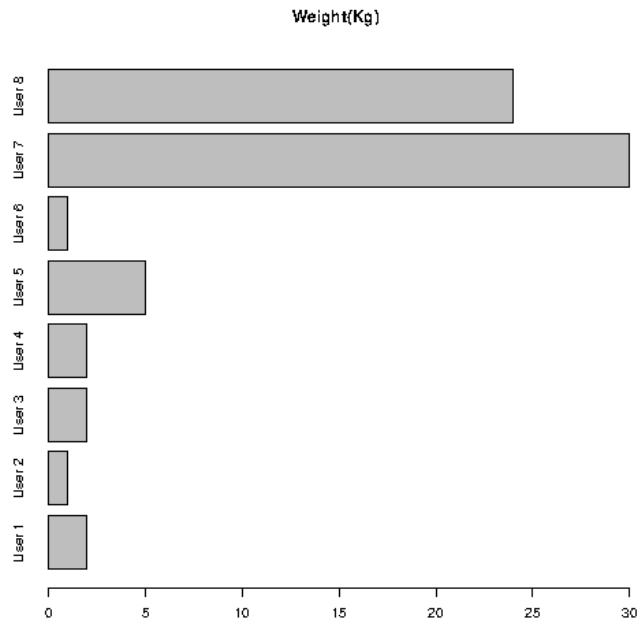


Figure 11. Weight for the recorded users

```
plot(`20211203Merged`$Calories, `20211203Merged`$HoursSlept, main="C
alories and Hours Slept",
+      xlab="Calories ", ylab="Hours Slept ", pch=19)
```

Users	Count WeightKg	Average WeightKg	Min WeightKg	Max WeightKg	Variation
User 1	2	52.6	52.6	52.6	0%
User 2	1	133.5	133.5	133.5	0%
User 3	2	57	56.7	57.3	1%
User 4	2	72.35	72.3	72.4	0%
User 5	5	69.64	69.1	70.3	2%
User 6	1	90.7	90.7	90.7	0%
User 7	30	61.55	61	62.5	2%
User 8	24	85.15	84	85.8	2%
Total	67	72.04			

Table 6. Summary table of the recorded weight

4.5. Sleep

Data Manipulation:

- Three records had duplicated dates and hours. These were deleted
- No record was found to have sleep time being greater than in bed
- The minutes were then converted to hours by dividing the minutes by 60
- The final dataset was reduced from 413 to 410. The basic statistics of the hours slept and hours in bed is shown below, with the histogram of the hours slept shown in **Figure 12**:

```
   HoursSlept       Hours_.Bed
Min.    : 0.970   Min.    : 1.020
1st Qu.: 6.020   1st Qu.: 6.732
Median : 7.210   Median : 7.720
Mean   : 6.987   Mean   : 7.641
3rd Qu.: 8.170   3rd Qu.: 8.770
Max.   :13.270   Max.   :16.020
```

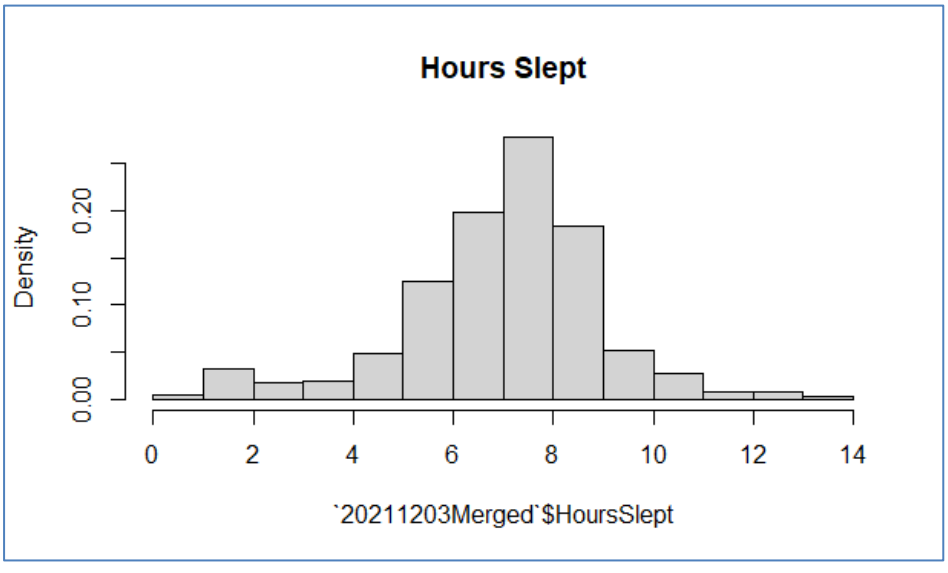


Figure 12. Hours Slept Histogram

There is little variation in the hours spent in bed and hours spent in bed as shown by the

blox plot in **Figure 13.**

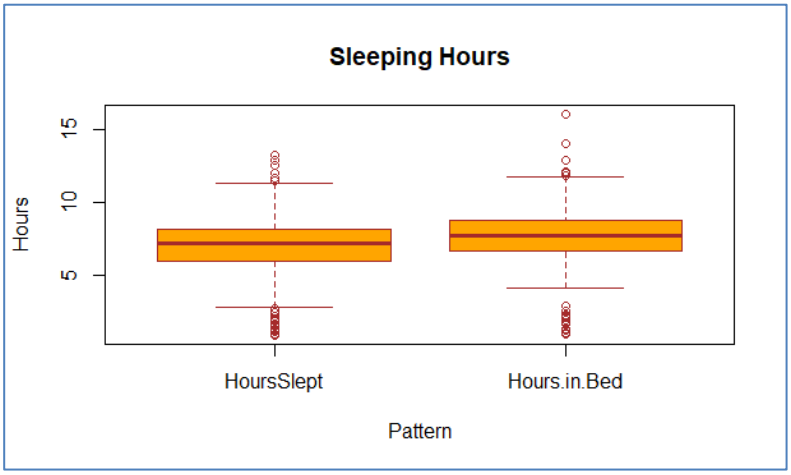
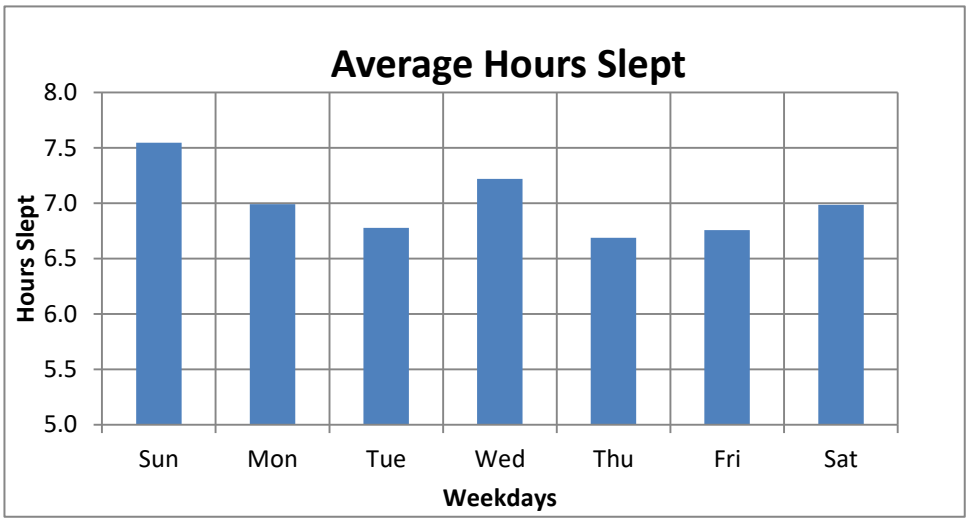


Figure 13. Box plot of hours in bed and hours slept

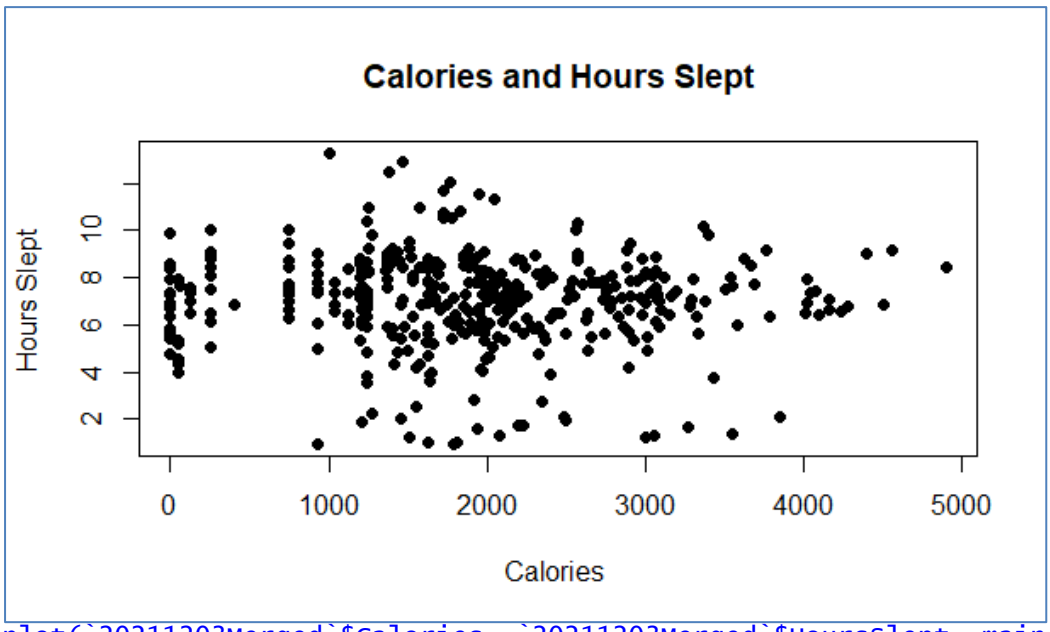
Hours slept is generally above 6.5 hrs a day, with the most hour's slept being on a Sunday and Wednesday. The least numbers of hours slept on average is probably on Thursday. This might coincide with the normal social life of Thursdays being a time where most people will go out clubbing as it is ladies day.




```
~/Capstone/Capstone/
boxplot(`20211122sleep_hrs`$Hourssleept,`20211122sleep_hrs`$Hours.in.Bed,
        data= `20211122sleep_hrs`,
        names = c("Hourssleept", "Hours.in.Bed"),
        main="Sleeping Hours",
        xlab="Pattern",
        ylab="Hours",
        col="orange",
        border="brown")
```

Figure 14

No correlation between hours slept and calories burned.




```
plot(`20211203Merged`$Calories, `20211203Merged`$HoursSlept, main="C
alories and Hours Slept",
+      xlab="Calories ", ylab="Hours Slept ", pch=19)
```

STEP 6: ACT

Conclusion and Recommendations

In the final step, insights will be given and providing recommendations based on our analysis. Overall, the data was not comprehensive enough to make a reliable analysis. But it did provide some valuable insights into how users interact with their wearable fitness device.

Here, we revisit our business questions and share with you our high-level businessrecommendations.

1. **What are the trends identified?**

- The amount of calories burned is directly related to the activity and intensity. Those who with high step count or were active had a high number of calories burned.
- The is generally a
-

2**. How could these trends help influence Bellabeat marketing strategy?**

- Bellabeat marketing team can encourage users by educating and equipping themwith knowledge about fitness benefits, suggest different types of exercise (ie.simple 10

minutes exercise on weekday and a more intense exercise on weekends) and calories intake and burnt rate information on the Bellabeat app.

- On weekends, Bellabeat app can also prompt notification to encourage users to exercise. His can include giving challenges to the users in order to motivate them to exercise. The other option is to get famous people to motivate users, similar to what Nike app did by enlisting Kevin Hart to motivate users to run/walk more
- The data clearly shows he amount of activity and the intensity of activity has a direct effect on calories burned and heart rate. We can also see some overall weekly trends, such as high activity on Tuesday andSaturday and low activity on Sunday and Wednesday
- Sleep data shows that no users slept more than 8 hours and this data is also unreliable as many users did not wear their smart watc at night

Overall, the data was not large or comprehensive enough to make a reliable analysis. But looking at the small sample size, it does give insight insights into how users interact with their wearble fitness device.

Few basic information such as might assist in the interaction with the devises:

- No water intake data has been collected. Collectnh hydration data can assist Bellabeat to be competitive.
- Disclosing gender or age of user is helpful for determining basic health checks such as BMI. This will be useful for analysis

The company is well positioned to market the Leaf device as easy-to-wear with a battery-life of six months device.