Bellabeat Analysis and Report

2025-02-3

### Business Task

Bellabeat has requested an analysis regarding the usage of non-bellabeat smart devices in order to gain insights about the consumer base.

### Data Used

dailyActivity\_merged <- read.csv("dailyActivity\_merged.csv")  
sleepDay\_merged <- read.csv("sleepDay\_merged.csv")  
dailyIntensities\_merged <-read.csv("dailyIntensities\_merged.csv")  
hourlyIntensities\_merged <- read.csv("hourlyIntensities\_merged.csv")

The data used in this analysis comprised of the FitBit Fitness Tracker Data provided by Mobius. Tables used from the original dataset included the dailyActivity\_merged, sleepDay\_merged and the hourlyIntensities\_merged. Copies with added columns were made in order to make deriving insights easier.

### Data Cleaning and Manipulation

Exploratory analysis of the data as well as reviews of the data on Kaggle provided an understanding that the data was clean. However, for safety sake, I used drop\_na() frequently to account for any missing values.

Other manipulation of the data included merging the sleepDay\_Day\_merged table with the daily\_activity\_merged table. This allowed us to check for any relationships between sleep and activity

I also mutated ActivityDate, SleepDay and ActivityHour columns from the dailyActivity\_merged, sleepDay\_merged and the hourlyIntensities\_merged tables respectively, in order to perform the aforementioned merge as well as for aggregation operations that required different information from those columns such as just the day (to groupby the day) or just the hour (to groupby the hour).

### Summary of Analysis

Much of my analysis was spent exploring the data for any possible relationships. I will lean into the insights generated from the clear relationships shown. I began my analysis with some simple summary statistics to have an understanding of the data and the activity trends of consumers in the dataset.I then generated some graphs to see how some relationships that would seem obvious actually was represented in the data.

I then went deeper, adding columns and merging the sleepdata and activities data to search for the presence of relationships between sleep and activity. I ended my analysis with 4 questions that I hoped would drive some final insights for Bellabeat.

### Analysis Report with Recommendations

#### I. EXPLORATION

##### i. Exploratory Summary Statistics

n\_distinct(dailyActivity\_merged$Id)

## [1] 33

n\_distinct(sleepDay\_merged$Id)

## [1] 24

nrow(dailyActivity\_merged)

## [1] 940

nrow(sleepDay\_merged)

## [1] 413

dailyActivity\_merged %>% 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

sleepDay\_merged %>% select(TotalSleepRecords,TotalMinutesAsleep,TotalTimeInBed) %>% summary()

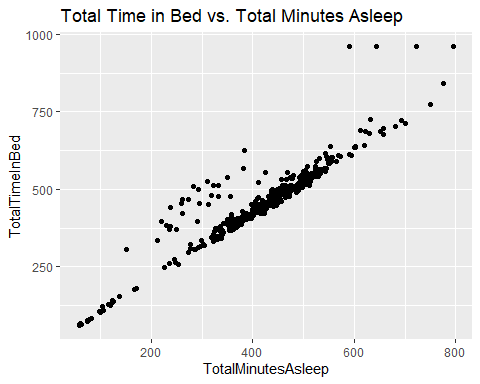
## TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   
## Min. :1.000 Min. : 58.0 Min. : 61.0   
## 1st Qu.:1.000 1st Qu.:361.0 1st Qu.:403.0   
## Median :1.000 Median :433.0 Median :463.0   
## Mean :1.119 Mean :419.5 Mean :458.6   
## 3rd Qu.:1.000 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :3.000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

The mean totalsteps for this dataset is 7638 per day. This is quite low as the CDC recommends 8000 to 10000 a day for adults under 60. The average time sedentary is 16.52 hours. Just these few statistics show the low average amount of activity taking place in the dataset.

##### ii. Exploratory Data Visualizations

1. TotalMinutesAsleep and TotalMinutesinBed from the sleepDay data.

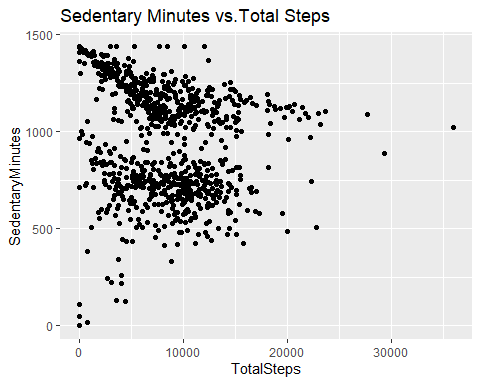
ggplot(data=sleepDay\_merged, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_point() + labs(title= "Total Time in Bed vs. Total Minutes Asleep")



This Data made practical sense and highlights the many users who spend a significant amount of time in bed without sleeping.

1. SedentaryMinutes and TotalSteps from the dailyActivity data.

ggplot(data=dailyActivity\_merged, aes(x=TotalSteps,y=SedentaryMinutes)) + geom\_point() + labs(title= "Sedentary Minutes vs.Total Steps")



Surprisingly, there was not a clear trend between Sedentary Minutes and Total Steps.

#### II. DEEPER ANALYSIS

##### i. Total Minutes Active

To go further in my analysis, I thought it would be helpful to add to the dailyActivity table by combining the LightlyActive,FairlyActive and VeryActive columns into a single total\_minutes\_active column. This would allow for an easy measurement of activity to analyse.

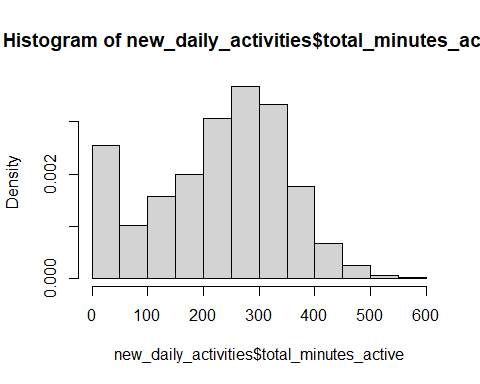
new\_daily\_activities <- dailyActivity\_merged %>% mutate(total\_minutes\_active = LightlyActiveMinutes+FairlyActiveMinutes+VeryActiveMinutes)

Here is a summary and histogram of total\_minutes\_active

summary(new\_daily\_activities$total\_minutes\_active)

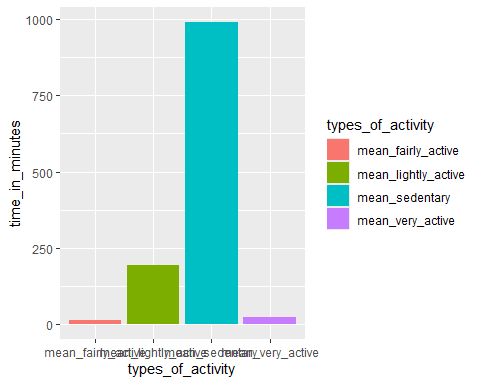
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 146.8 247.0 227.5 317.2 552.0

hist(new\_daily\_activities$total\_minutes\_active, probability = TRUE)

 The mean amount of time being active is 227.5 minutes which is 3.79 hours. These seems like an encouraging number, however it is not clear how that data is spread, as total minutes included, light, fair and very active minutes.

The next plot will show how activity is distributed between these categories and compare it to sedentary activity as well. I took the means of each of the activity categories and plotted them to see the average amount of time spent on each category across the dataset.

activity\_spread\_mean <- data.table::copy(new\_daily\_activities) %>% select('SedentaryMinutes', 'LightlyActiveMinutes', 'FairlyActiveMinutes', 'VeryActiveMinutes','VeryActiveMinutes') %>% summarize(mean\_sedentary = mean(SedentaryMinutes), mean\_lightly\_active = mean(LightlyActiveMinutes), mean\_fairly\_active = mean(FairlyActiveMinutes), mean\_very\_active = mean(VeryActiveMinutes))  
  
activity\_spread\_mean <- pivot\_longer(activity\_spread\_mean, cols=c('mean\_sedentary', 'mean\_lightly\_active', 'mean\_fairly\_active', 'mean\_very\_active'), names\_to = c('types\_of\_activity'), values\_to = 'time\_in\_minutes')  
  
ggplot(data=activity\_spread\_mean, aes(x=types\_of\_activity, y= time\_in\_minutes, fill = types\_of\_activity)) + geom\_col()



print(activity\_spread\_mean)

## # A tibble: 4 × 2  
## types\_of\_activity time\_in\_minutes  
## <chr> <dbl>  
## 1 mean\_sedentary 991.   
## 2 mean\_lightly\_active 193.   
## 3 mean\_fairly\_active 13.6  
## 4 mean\_very\_active 21.2

Most of the activity spent with reference to the total\_minutes\_active, would be spent on light activity, then very intense, then fairly intense activity. Sedentary activity takes up most of the time on average.

##### ii. Sleep Data with Activity Data

The next step of the analysis was to merge sleep data with the new\_daily\_activities table in order to find any relationships between sleep and activity.

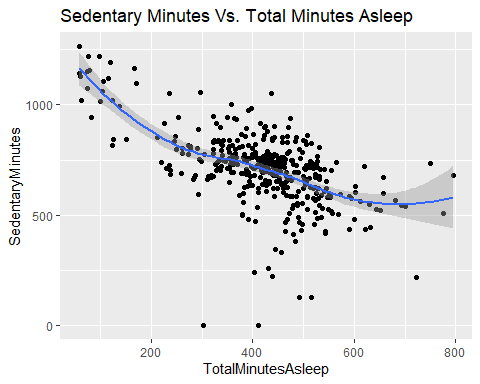
sleepDay\_date <- data.table::copy(sleepDay\_merged) %>% mutate(new\_date = mdy\_hms(SleepDay))  
  
new\_daily\_activities\_date <- data.table::copy(new\_daily\_activities) %>% mutate(new\_date = mdy(ActivityDate))  
  
  
combined\_data <- merge(sleepDay\_date, new\_daily\_activities\_date, by= c("Id","new\_date")) %>% select(-SleepDay,-ActivityDate)

After various checks for relationships between TotalTimeAsleep and the following: TotalSteps, total\_minutes\_active, LightlyActiveMinutes, FairlyActiveMinutes and VeryActiveMinutes, there was found to be no relationship between sleep and activity, except for the final attempt:

TotalMinutesAsleep and SedentaryActivity.

ggplot(data=combined\_data, aes(x=TotalMinutesAsleep,y=SedentaryMinutes)) + geom\_point(color = "black") + geom\_smooth() + labs(title = "Sedentary Minutes Vs. Total Minutes Asleep")

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

 The downward trend of the data suggests that more sleep correlates with less sedentary activity.

#### III. Final Driving Questions

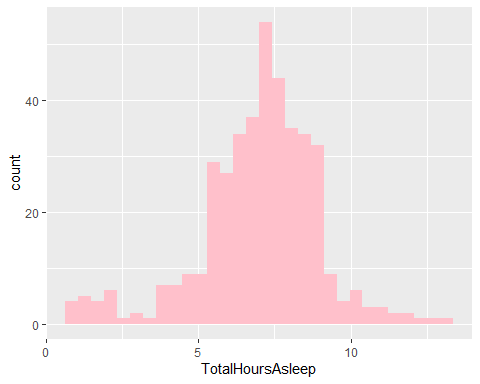
To complete my analysis, I posed some final questions in search for trends and provide recommendations:

1. What does the distribution of sleep look like?
2. What days are the most/least active on average, for most people?
3. What times of day are most/least active?
4. What days have the most/least sleep?

##### i) What does the distribution of sleep look like?

hours\_sleep <- select(combined\_data, 'Id','TotalMinutesAsleep') %>% mutate(TotalHoursAsleep = TotalMinutesAsleep/60)   
ggplot(data=hours\_sleep, aes(x=TotalHoursAsleep)) + geom\_histogram(fill = "pink")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



summary(hours\_sleep$TotalHoursAsleep)

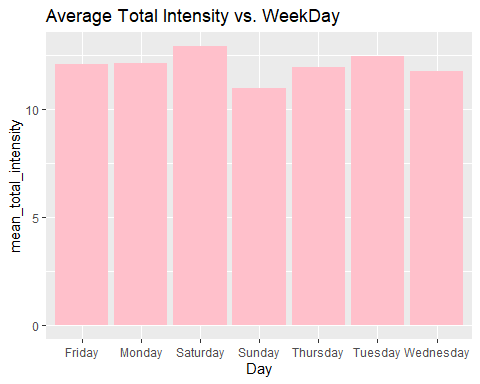
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.9667 6.0167 7.2167 6.9911 8.1667 13.2667

The average amount of sleep is 6.99. Most of the data is falling between the 6.017 and 8.17 hours of sleep.

##### ii) What days are the most/least active on average, for most people?

mutated\_int\_days <- data.table::copy(hourlyIntensities\_merged) %>% mutate(Day = mdy\_hms(ActivityHour) %>% weekdays()) %>% group\_by(Day) %>% drop\_na() %>% summarise(mean\_total\_intensity = mean(TotalIntensity))  
  
ggplot(data = mutated\_int\_days, aes(x=Day, y=mean\_total\_intensity)) + geom\_histogram(stat = "identity", fill = "pink") + theme(axis.text.x = element\_text(angle = 0)) + labs(title = "Average Total Intensity vs. WeekDay" )

## Warning in geom\_histogram(stat = "identity", fill = "pink"): Ignoring unknown  
## parameters: `binwidth`, `bins`, and `pad`

 Tuesday and Saturday were on average the most active while Sundays and Wednesdays were the least active.

##### ii) What times of day are most/least active?

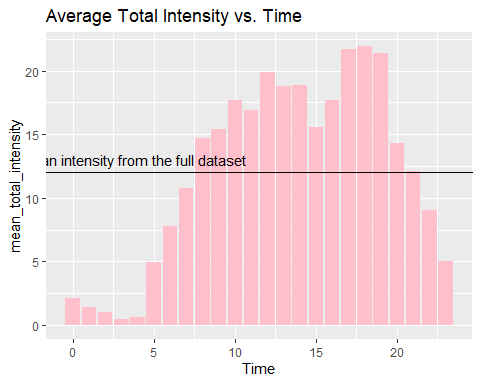
I grouped the intensity data by times of day, finding the means from each hour. Then I compared each mean to the total mean of the data for intensity.

mutated\_int\_hours <- data.table::copy(hourlyIntensities\_merged) %>% mutate(Time = mdy\_hms(ActivityHour) %>% hour()) %>% group\_by(Time) %>% drop\_na() %>% summarise(mean\_total\_intensity = mean(TotalIntensity))  
  
summary(hourlyIntensities\_merged$TotalIntensity)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 3.00 12.04 16.00 180.00

ggplot(data = mutated\_int\_hours, aes(x=Time, y=mean\_total\_intensity)) + geom\_histogram(stat = "identity", fill = "pink") + theme(axis.text.x = element\_text(angle = 0)) + labs(title = "Average Total Intensity vs. Time" ) + geom\_abline(slope = 0, intercept = 12.04) + annotate("text", x= 3.7, y= 13.04, label = "mean intensity from the full dataset", size= 4)

## Warning in geom\_histogram(stat = "identity", fill = "pink"): Ignoring unknown  
## parameters: `binwidth`, `bins`, and `pad`

 Opportunities for increased physical activity can be early in the morning such as from 4-8, or at night from 7-9.

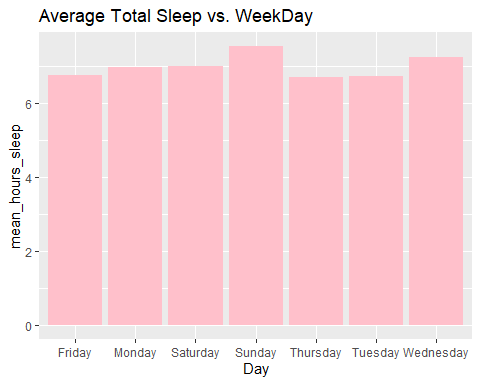
##### iv) What days have the most/least sleep?

mutated\_int\_sleep <- data.table::copy(combined\_data) %>% mutate(Day = weekdays(new\_date), TotalHoursAsleep = TotalMinutesAsleep/60) %>% group\_by(Day) %>% drop\_na() %>% summarise(mean\_hours\_sleep = mean(TotalHoursAsleep))  
  
print(mutated\_int\_sleep)

## # A tibble: 7 × 2  
## Day mean\_hours\_sleep  
## <chr> <dbl>  
## 1 Friday 6.76  
## 2 Monday 6.98  
## 3 Saturday 7.01  
## 4 Sunday 7.55  
## 5 Thursday 6.71  
## 6 Tuesday 6.74  
## 7 Wednesday 7.24

ggplot(data = mutated\_int\_sleep, aes(x=Day, y=mean\_hours\_sleep)) + geom\_histogram(stat = "identity", fill = "pink") + theme(axis.text.x = element\_text(angle = 0)) + labs(title = "Average Total Sleep vs. WeekDay" )

## Warning in geom\_histogram(stat = "identity", fill = "pink"): Ignoring unknown  
## parameters: `binwidth`, `bins`, and `pad`



Monday, Tuesday, Thursday and Friday stick out with the lowest amounts of sleep on average. The average amounts of sleep on these days are all under 7 hours, the recommended minimum amount of sleep by the CDC.

Wednesday and Sunday bave the highest amounts of sleep. This may be because Wednesdays are the middle of the week, hence consumers are tired from the first 3 days of work and Sundays are usually spent watching sports or resting after activities like church and other family gatherings.

#### IV. Final Conclusion and Recommendations

##### Conclusion

The low average of footsteps below the 3rd quartile of totalsteps indicated more activity would be needed by most people on the dataset. After looking at the spread of activity, most of time spent in the data is on sedentary activity, then light, then very, then fairly intense activity. There is a lack of meaningful exercise. In fact a lack of any exercise across the dataset compared to sedentary activity. The people in the upper 25% of the data are quite active but the lower 75% need some help.

The following recommends can be implemented to the Bellabeat membership in conjunction with the Lead or the Time in order to drive results towards users goals.

##### Recommendations:

* Since less sedentary activity is correlated with more sleep, according to our graph, Bellabeat can implement daily reminders for more sleep throught the app or through the Spring/Time. It can tailor the reminders to the days where the users are on average, under the recommended amount of 7 hours.
* There can be specific reminders to work out on Wednesdays and Sundays where it seems like most people have less activity.
* Finally, Bellabeat can provide insights to users about their weekly activity and offer suggestions for meeting weekly, monthy and yearly goals. Since the hours of 4-8, and 7-9 seem to be low on activity, Bellabeat can offer potential times in those ranges for users to implement an hour long intense workout. If users are older, having a fairly active or even a light walk/jog could be recommended instead. This would increase the amount of active minutes for each user.

#### Thank you for reviewing my report