Bellabeat Analysis

Teddy Mutinge

2023-11-02

# **How Can a Wellness Technology Company Play It Smart?**

## About Company

Bellabeat is a company that develops fitness products for women. Their products include smart water bottles, fashionable fitness watches and jewelry, and yoga mats. Users can access their health data collected through these devices in the Bellabeat app.

Bellabeat’s co-founders would like to analyze data from non-Bellabeat fitness devices to see how consumers are using these products. The company hopes to use these insights to help guide new marketing strategies for the company.

### Key stakeholders

* **Urška Sršen**: Bellabeat’s cofounder and Chief Creative Officer
* **Sando Mu**: Mathematician and Bellabeat’s cofounder
* **The Bellabeat marketing analytics team**: a team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat’s marketing strategy.

### Bellabeat products

* **Bellabeat app**: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
* **Leaf**: Bellabeat’s classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.
* **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.
* **Spring**: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.
* **Bellabeat membership**: Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals

## Business task

To analyze non-Bellabeat smart device data and compare with one Bellabeat product to discover insights to help guide marketing strategies for the company.

### Key Questions

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

## Data

The data used in this analysis was obtained from publicly available dataset on Kaggle, [FitBit Fitness Tracker Data.](https://www.kaggle.com/datasets/arashnic/fitbit)

### Limitations

* Data was of small sample size.
* Data did not include information on sex, age and demographics.
* Data was not current dating back to the year 2016.
* Data was gender biased because it included women only.

For purposes of this analysis two data files were used i.e **dailyActivity\_merged.csv** and **sleepdDay\_merged.csv**.

### Softwares used for analysis

* R
* Spreadsheets(Google Sheets and Excel)

## Process and Analysis

First, we load the packages.Then import the data, transform and analyze

# Load the packages  
  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(readxl)  
library(ggplot2)

# Import the data  
d\_sleep <- read\_excel("sleep.xlsx")  
d\_activity <- read\_excel("activity.xlsx")  
  
head(d\_sleep)

## # A tibble: 6 × 5  
## Id Date TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## <dbl> <dttm> <dbl> <dbl> <dbl>  
## 1 1.50e9 2016-04-12 00:00:00 1 327 346  
## 2 1.50e9 2016-04-13 00:00:00 2 384 407  
## 3 1.50e9 2016-04-15 00:00:00 1 412 442  
## 4 1.50e9 2016-04-16 00:00:00 2 340 367  
## 5 1.50e9 2016-04-17 00:00:00 1 700 712  
## 6 1.50e9 2016-04-19 00:00:00 1 304 320

head(d\_activity)

## # A tibble: 6 × 15  
## Id Date TotalSteps TotalDistance TrackerDistance  
## <dbl> <dttm> <dbl> <dbl> <dbl>  
## 1 1503960366 2016-04-12 00:00:00 13162 8.5 8.5   
## 2 1503960366 2016-04-13 00:00:00 10735 6.97 6.97  
## 3 1503960366 2016-04-14 00:00:00 10460 6.74 6.74  
## 4 1503960366 2016-04-15 00:00:00 9762 6.28 6.28  
## 5 1503960366 2016-04-16 00:00:00 12669 8.16 8.16  
## 6 1503960366 2016-04-17 00:00:00 9705 6.48 6.48  
## # ℹ 10 more variables: LoggedActivitiesDistance <dbl>,  
## # VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,  
## # LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,  
## # VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,  
## # LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>

# Detailed summary of the data  
glimpse(d\_activity)

## Rows: 940  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ Date <dttm> 2016-04-12, 2016-04-13, 2016-04-14, 2016-04-…  
## $ TotalSteps <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019…  
## $ TotalDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveMinutes <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…  
## $ FairlyActiveMinutes <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…  
## $ LightlyActiveMinutes <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, …  
## $ SedentaryMinutes <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…  
## $ Calories <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…

glimpse(d\_sleep)

## Rows: 413  
## Columns: 5  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150…  
## $ Date <dttm> 2016-04-12, 2016-04-13, 2016-04-15, 2016-04-16, 20…  
## $ TotalSleepRecords <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2…  
## $ TotalTimeInBed <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3…

### Data cleaning

We will begin by cleaning the data frames to remove duplicates,check for null values and change the datatype of the date columns from **character** to **date**.

# Clean the column names.  
clean\_names(d\_activity)

## # A tibble: 940 × 15  
## id date total\_steps total\_distance tracker\_distance  
## <dbl> <dttm> <dbl> <dbl> <dbl>  
## 1 1503960366 2016-04-12 00:00:00 13162 8.5 8.5   
## 2 1503960366 2016-04-13 00:00:00 10735 6.97 6.97  
## 3 1503960366 2016-04-14 00:00:00 10460 6.74 6.74  
## 4 1503960366 2016-04-15 00:00:00 9762 6.28 6.28  
## 5 1503960366 2016-04-16 00:00:00 12669 8.16 8.16  
## 6 1503960366 2016-04-17 00:00:00 9705 6.48 6.48  
## 7 1503960366 2016-04-18 00:00:00 13019 8.59 8.59  
## 8 1503960366 2016-04-19 00:00:00 15506 9.88 9.88  
## 9 1503960366 2016-04-20 00:00:00 10544 6.68 6.68  
## 10 1503960366 2016-04-21 00:00:00 9819 6.34 6.34  
## # ℹ 930 more rows  
## # ℹ 10 more variables: logged\_activities\_distance <dbl>,  
## # very\_active\_distance <dbl>, moderately\_active\_distance <dbl>,  
## # light\_active\_distance <dbl>, sedentary\_active\_distance <dbl>,  
## # very\_active\_minutes <dbl>, fairly\_active\_minutes <dbl>,  
## # lightly\_active\_minutes <dbl>, sedentary\_minutes <dbl>, calories <dbl>

clean\_names(d\_sleep)

## # A tibble: 413 × 5  
## id date total\_sleep\_records total\_minutes\_asleep  
## <dbl> <dttm> <dbl> <dbl>  
## 1 1503960366 2016-04-12 00:00:00 1 327  
## 2 1503960366 2016-04-13 00:00:00 2 384  
## 3 1503960366 2016-04-15 00:00:00 1 412  
## 4 1503960366 2016-04-16 00:00:00 2 340  
## 5 1503960366 2016-04-17 00:00:00 1 700  
## 6 1503960366 2016-04-19 00:00:00 1 304  
## 7 1503960366 2016-04-20 00:00:00 1 360  
## 8 1503960366 2016-04-21 00:00:00 1 325  
## 9 1503960366 2016-04-23 00:00:00 1 361  
## 10 1503960366 2016-04-24 00:00:00 1 430  
## # ℹ 403 more rows  
## # ℹ 1 more variable: total\_time\_in\_bed <dbl>

# Check for duplicates.  
sum(duplicated(d\_activity))

## [1] 0

sum(duplicated(d\_sleep))

## [1] 3

# Remove the duplicates found in the d\_sleep data.  
d\_sleep <- d\_sleep[!duplicated(d\_sleep), ]  
# Check for null values if any present in the data.  
sum(is.na(d\_activity))

## [1] 0

sum(is.na(d\_sleep))

## [1] 0

# Convert to date datatype  
d\_activity <- d\_activity %>%   
 mutate\_at(vars(Date), as.Date, format = "%m/%d/%y")%>%   
 rename("date"="Date")  
  
d\_sleep <- d\_sleep %>%   
 mutate\_at(vars(Date), as.Date, format = "%m/%d/%y") %>%  
 rename("date"="Date")

### Merge the data

Here we merge the data using **Id** and **date**.

d\_merged <- merge(d\_activity, d\_sleep, by = c("Id","date"))  
  
# Introduce a new field ad Weekday  
d\_merged<- d\_merged %>%   
 mutate(Weekday = weekdays(as.Date(date, "m/%d/%Y")))  
  
# Check for unique entries  
n\_distinct(d\_merged$Id)

## [1] 24

n\_distinct(d\_sleep$Id)

## [1] 24

n\_distinct(d\_activity$Id)

## [1] 33

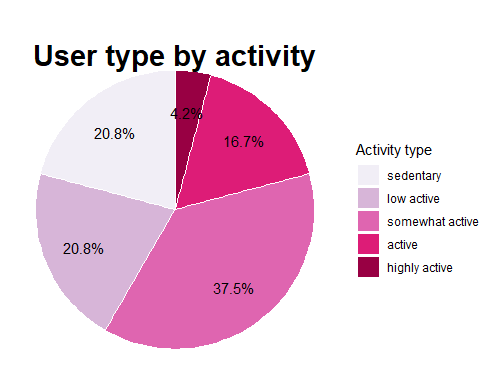
summary(d\_merged)

## Id date TotalSteps TotalDistance   
## Min. :1.504e+09 Min. :2016-04-12 Min. : 17 Min. : 0.010   
## 1st Qu.:3.977e+09 1st Qu.:2016-04-19 1st Qu.: 5189 1st Qu.: 3.592   
## Median :4.703e+09 Median :2016-04-27 Median : 8913 Median : 6.270   
## Mean :4.995e+09 Mean :2016-04-26 Mean : 8515 Mean : 6.012   
## 3rd Qu.:6.962e+09 3rd Qu.:2016-05-04 3rd Qu.:11370 3rd Qu.: 8.005   
## Max. :8.792e+09 Max. :2016-05-12 Max. :22770 Max. :17.540   
## TrackerDistance LoggedActivitiesDistance VeryActiveDistance  
## Min. : 0.010 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 3.592 1st Qu.:0.0000 1st Qu.: 0.000   
## Median : 6.270 Median :0.0000 Median : 0.570   
## Mean : 6.007 Mean :0.1089 Mean : 1.446   
## 3rd Qu.: 7.950 3rd Qu.:0.0000 3rd Qu.: 2.360   
## Max. :17.540 Max. :4.0817 Max. :12.540   
## ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance  
## Min. :0.0000 Min. :0.010 Min. :0.0000000   
## 1st Qu.:0.0000 1st Qu.:2.540 1st Qu.:0.0000000   
## Median :0.4200 Median :3.665 Median :0.0000000   
## Mean :0.7439 Mean :3.791 Mean :0.0009268   
## 3rd Qu.:1.0375 3rd Qu.:4.918 3rd Qu.:0.0000000   
## Max. :6.4800 Max. :9.480 Max. :0.1100000   
## VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes  
## Min. : 0.00 Min. : 0.00 Min. : 2.0 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:158.0 1st Qu.: 631.2   
## Median : 9.00 Median : 11.00 Median :208.0 Median : 717.0   
## Mean : 25.05 Mean : 17.92 Mean :216.5 Mean : 712.1   
## 3rd Qu.: 38.00 3rd Qu.: 26.75 3rd Qu.:263.0 3rd Qu.: 782.8   
## Max. :210.00 Max. :143.00 Max. :518.0 Max. :1265.0   
## Calories TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   
## Min. : 257 Min. :1.00 Min. : 58.0 Min. : 61.0   
## 1st Qu.:1841 1st Qu.:1.00 1st Qu.:361.0 1st Qu.:403.8   
## Median :2207 Median :1.00 Median :432.5 Median :463.0   
## Mean :2389 Mean :1.12 Mean :419.2 Mean :458.5   
## 3rd Qu.:2920 3rd Qu.:1.00 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :4900 Max. :3.00 Max. :796.0 Max. :961.0   
## Weekday   
## Length:410   
## Class :character   
## Mode :character   
##   
##   
##

### Analysis based on activity

We will follow the recommendations developed as a guide on how many steps are sufficient for health benefits among adults based on [healthline article.](https://www.healthline.com/health/how-many-steps-a-day)

# Average Total steps grouped by Id  
d\_avg\_steps <- d\_merged %>%   
 group\_by(Id) %>%   
 summarize(avg\_steps = mean(TotalSteps), avg\_dist = mean(TotalDistance), avg\_cal = mean(Calories))  
  
# Creating user types  
activity\_user\_type <- d\_avg\_steps %>%  
 mutate(activity\_type = case\_when(  
 avg\_steps < 5000 ~ "sedentary",  
 avg\_steps >= 5000 & avg\_steps < 7500 ~ "low active",   
 avg\_steps >= 7500 & avg\_steps < 10000 ~ "somewhat active",   
 avg\_steps >= 10000 & avg\_steps < 12500 ~ "active",  
 avg\_steps >= 12500 ~ "highly active",  
 ))  
  
# Counting the number by user type and finding the percentage   
activity\_user\_type\_percent <- activity\_user\_type %>%   
 group\_by(activity\_type) %>%   
 summarise(total = n()) %>%   
 mutate(totals = sum(total)) %>%   
 group\_by(activity\_type) %>%   
 summarise(total\_percent=total/totals) %>%   
 mutate(percent = scales::percent(total\_percent)) %>%   
 arrange(desc(total\_percent))  
  
  
activity\_user\_type\_percent$activity\_type <- factor(activity\_user\_type\_percent$activity\_type, levels =c("sedentary", "low active", "somewhat active", "active", "highly active"))  
  
  
# Plot a pie chart showing the distribution of user activity   
ggplot(activity\_user\_type\_percent,aes(x="",y = total\_percent, fill=activity\_type)) +  
 geom\_bar(stat="identity", width=1, color="white") +  
 coord\_polar("y", start=0)+  
 scale\_fill\_brewer(palette='PuRd')+  
 theme\_void()+ # remove background, grid, numeric labels  
 theme(plot.title = element\_text(hjust = 0.5,vjust= -5, size = 22, face = "bold")) +  
 geom\_text(aes(label = percent, x=1.2),position = position\_stack(vjust = 0.5))+  
 labs(title="User type by activity")+  
 guides(fill = guide\_legend(title = "Activity type"))



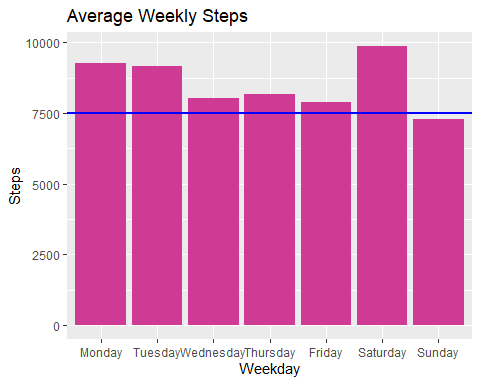
***Conclusions***

1. 58% of the users are active and somewhat active.
2. 21% of the users have active and highly active lifestyles.
3. 21% of the users have sedentary lifestyles.

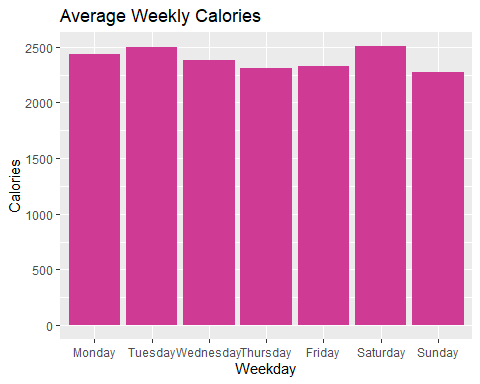
### Weekday distributions

Next we will check weekday distributions based on average steps,calories and sleep.

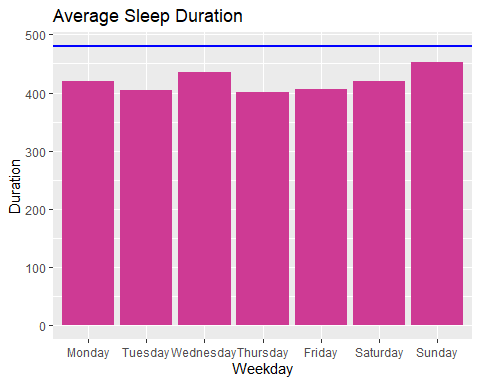
# Averages for Totalsteps,Calories and Sleep duration grouped by Weekday  
activity\_weekdays <- d\_merged %>%   
 group\_by(Weekday) %>%   
 summarise(avg\_steps = mean(TotalSteps),avg\_dist= mean(TotalDistance), avg\_cal = mean(Calories), avg\_sleep= mean(TotalMinutesAsleep))  
   
activity\_weekdays$Weekday <- ordered(activity\_weekdays$Weekday, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))  
  
# Checking weekly average steps  
ggplot(data=activity\_weekdays, aes(x=Weekday, y=avg\_steps)) +   
 geom\_bar(stat="identity", fill="#CE3A94")+  
 geom\_hline(yintercept = 7500, size=1, color = "#0000FF")+  
 labs(title="Average Weekly Steps ",x ="Weekday", y="Steps")



# Checking distribution of calories per weekday  
ggplot(data=activity\_weekdays, aes(x=Weekday, y=avg\_cal)) +   
 geom\_bar(stat="identity", fill="#CE3A94")+  
 labs(title="Average Weekly Calories ",x ="Weekday", y="Calories")



# Distribution of sleep during the week  
ggplot(data=activity\_weekdays, aes(x=Weekday, y=avg\_sleep)) +   
 geom\_bar(stat="identity", fill="#CE3A94")+  
 geom\_hline(yintercept = 480, size=1, color = "#0000FF")+  
 labs(title = "Average Sleep Duration", x= "Weekday", y ="Duration")

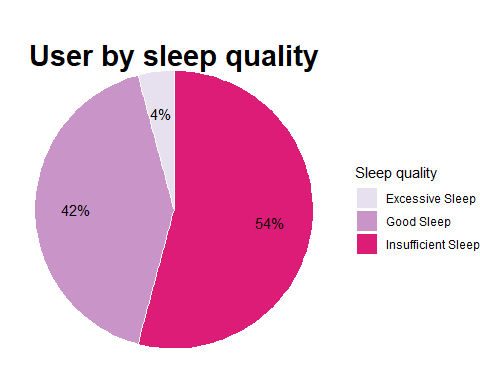
 ***Conclusions***

1. Saturday recorded the highest number of steps on average which corresponds with the high amount of lost calories.
2. Most users did not meet the threshold of average steps on Sunday thus a decline in activity.
3. Sunday recorded highest amount of sleep duration.
4. Below average sleep duration of 8hrs was noted.

### Sleep Quality

Next, we address the issue pertaining sleep quality of the users.

# Average sleep duration grouped by Id  
sleep\_activity <- d\_merged %>%   
 group\_by(Id) %>%   
 summarise(avg\_sleep = mean(TotalMinutesAsleep))  
  
# Creating sleep types  
sleep\_activity <- sleep\_activity %>%   
 mutate(sleep\_quality = case\_when(  
 avg\_sleep < 420 ~ "Insufficient Sleep",  
 avg\_sleep >= 420 & avg\_sleep < 540 ~ "Good Sleep",  
 avg\_sleep > 540 ~ "Excessive Sleep",  
 ))  
  
# Counting the number of sleep type and finding the percentage  
sleep\_activity\_percent <- sleep\_activity %>%   
 group\_by(sleep\_quality) %>%   
 summarise(total = n()) %>%   
 mutate(totals = sum(total)) %>%   
 group\_by(sleep\_quality) %>%   
 summarise(total\_percent = total/totals) %>%   
 mutate(percent = scales::percent(total\_percent)) %>%   
 arrange(desc(total\_percent))  
  
  
# Plot the distribution of sleep types  
ggplot(sleep\_activity\_percent,aes(x="",y = total\_percent, fill=sleep\_quality)) +  
 geom\_bar(stat="identity", width=1, color="white") +  
 coord\_polar("y", start=0)+  
 scale\_fill\_brewer(palette='PuRd')+  
 theme\_void()+ # remove background, grid, numeric labels  
 theme(plot.title = element\_text(hjust = 0.5,vjust= -5, size = 22, face = "bold")) +  
 geom\_text(aes(label = percent, x=1.2),position = position\_stack(vjust = 0.5))+  
 labs(title="User by sleep quality")+  
 guides(fill = guide\_legend(title = "Sleep quality"))

 **Conclusions**

* There is a near equal split between people who do not get enough sleep (less than 7 hours per day) in comparison to those who get the right amount of sleep (7 to 9 hours per day).

### Correlation between Steps and Calories

Does more steps result to high calories burn?

d\_merged %>%   
 group\_by(TotalSteps, Calories) %>%   
 ggplot(aes(x = TotalSteps, y = Calories, color = Calories)) +  
 geom\_point() +  
 geom\_smooth(color = "blue") +   
 labs(title = 'Calories vs. Total Steps',  
 y = 'Calories',  
 x = 'Total Steps')

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



cor(d\_merged$TotalSteps, d\_merged$Calories)

## [1] 0.4063007

A correlation of **0.4** signifies that **somewhat** the calories burn can be attributed to increased activity(number of steps).

### Key findings

Now that the data has been cleaned and analyzed it has provided valuable insights to the marketing team.

* Most users are more active on Saturday and less active on Sunday.
* The quality of sleep among most users is not great.
* Majority of the users are fairly active during the week with at least 7500 steps recorded.
* There is somewhat a correlation between steps taken and calories burnt.

### Recommendations

* Create a feature that would **send notifications to users to remind them of their bedtime hours** to ensure they have adequate sleep.
* Develop **exercise recommendations** based on the lifestyles that contribute positively to their health.
* Add a **music feature** that would keep users motivated during their exercise sessions.