# Case Study: Bellabeat Fitness Data Analysis

The case study follows the six step data analysis process:

Ask

Prepare

Process

Analyze

Share

Act

Since its inception in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-focused wellness company for women. The company has 5 focus products: bellabeat App, Leaf, Time, Spring and bellabeat Membership. Bellabeat is a successful small company, but they have the potential to become a bigger player in the global smart device market. Our team was asked to analyze smart device data to gain insights into how consumers are using their smart devices. The insights gained then help guide the marketing strategy for the company.

1. Ask

   BUSINESS TASK: Analyze Fitbit data to gain insight and help guide marketing strategy for Bellabeat to grow as a global player.

   Primary stakeholders: Urška Sršen and Sando Mur, executive team members.

   Secondary stakeholders: Bellabeat marketing analytics team.

2. Prepare

Data Source: 30 participants FitBit Fitness Tracker Data from Mobius: https://www.kaggle.com/arashnic/fitbit

The dataset has 18 CSV files. The data also follows the ROCCC approach:

- Reliability: The data is from 30 FitBit users who consented to the submission of personal tracker data and generated by from a distributed survey through Amazon Mechanical Turk.
- Original: The data is from 30 FitBit users who have consented to the submission of personal tracker data via Amazon Mechanical Turk.
- Comprehensive: Data minute-level output for physical activity, heart rate, and sleep monitoring. While the data tracks many factors in the user activity and sleep, but the sample size is small and most data is recorded during certain days of the week.
- Current: Data is from March 2016 to May 2016. Data is not current so the users habit may be different now.
- Cited: Unknown.

However, the dataset does have some limitations:

- Most data is recorded from Tuesday to Thursday, which may not be comprehensive enough to form an accurate analysis.
- Only 30 user data is available. The central limit theorem general rule of n≥30 applies and we can use the t test for statistical reference. However, a larger sample size would have been preferred for the analysis.
- For the 8 user data for weight, 5 users manually entered their weight and 3 recorded via a connected wireless device (e.g.: Wifi).
- Upon further investigation with n_distinct() to check for unique user Id, the set has 33 user data from daily activity, 24 from sleep and only 8 from weight. There are 3 extra users and some users opted not to record their data for tracking daily activity and sleep.

3. Process

Examine the data, check for NA, and remove duplicates for three main tables: daily_activity, sleep_day and weight:

```
dim(sleep_day)
sum(is.na(sleep_day))
sum(duplicated(sleep_day))
sleep_day <- sleep_day[!duplicated(sleep_day), ]
```

Add a column for day of the week and convert ActivityDate into date format.

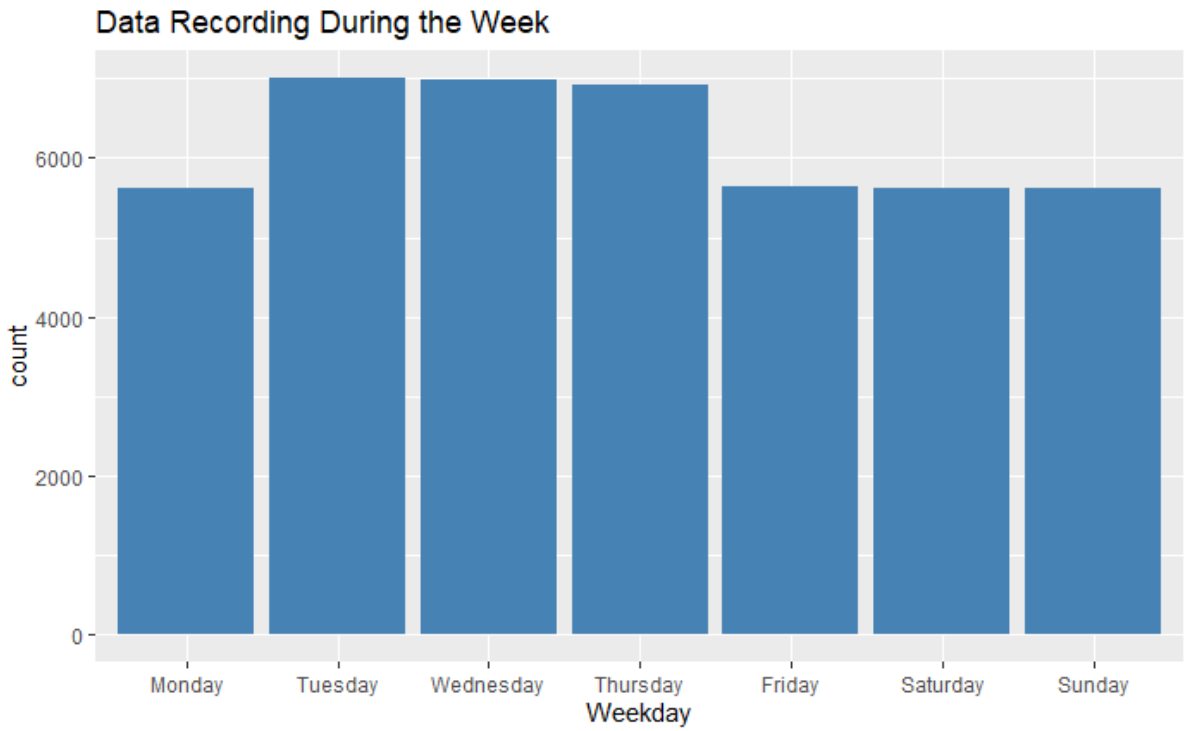
```
daily_activity <- daily_activity %>% mutate( Weekday = weekdays(as.Date(ActivityDate,
"%m/%d/%Y")))
```

Check to see if we have 30 users using n_distinct(). The dataset has 33 users data from daily activity, 24 from sleep and only 8 from weight. If there is a discrepancy such as in the weight table, check to see how the data are recorded. The way the user record the data may give you insight on why there is missing data.

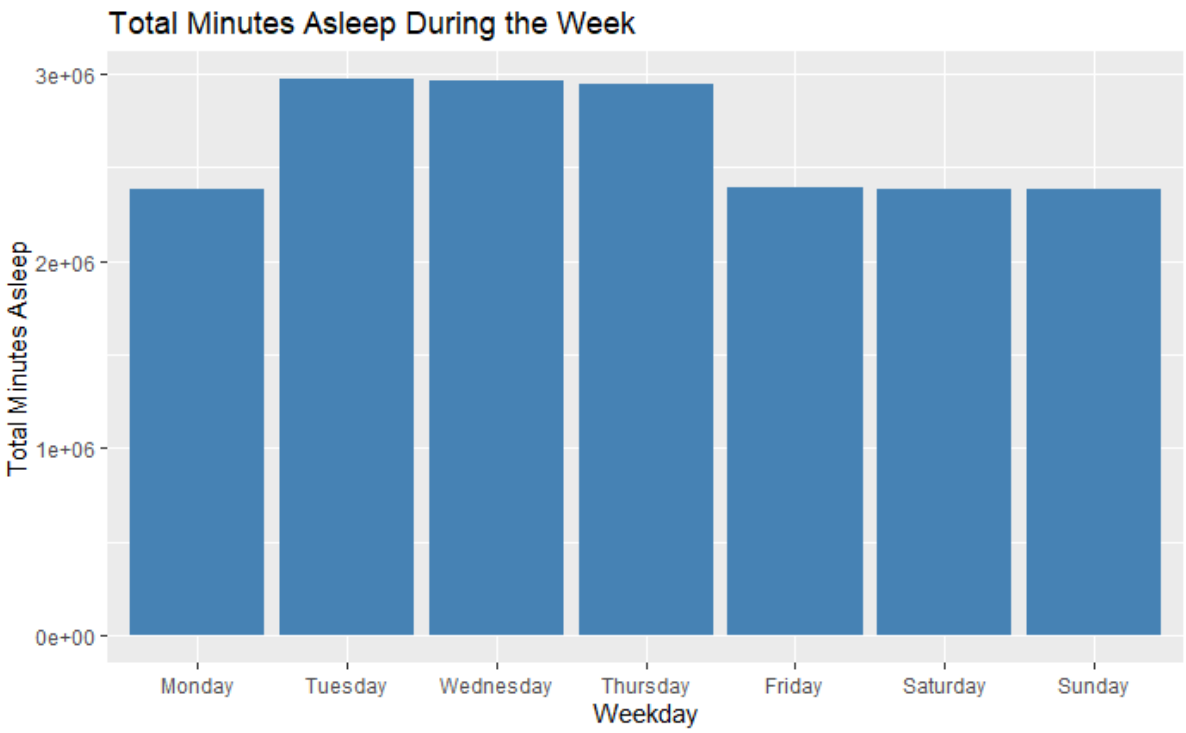
```
weight %>%
  filter(IsManualReport == "True") %>%
  group_by(Id) %>%
  summarise("Manual Weight Report"=n()) %>%
  distinct()
```

Additional insight to be aware of is how often user record their data. We can see from the ggplot() bar graph that the data are greatest from Tuesday to Thursday. We need to investigate the data recording distribution. Monday and Friday are both weekdays, why isn't the data recordings as much as the other weekdays?

```
ggplot(data=merged_data, aes(x=Weekday))+
  geom_bar(fill="steelblue")
```




From weekday's total asleep minutes, we can see the graph look almost same as the graph above! We can confirmed that most sleep data is also recorded during Tuesday to Thursday. This raised a question of comprehensiveness and data accuracy to form a sound analysis.




Merge the three tables:

merged_data <- merge(merged_activity_sleep, weight, by = c("Id"), all=TRUE)
Clean the data to prepare for analysis in the next step: Analyze!

4. Analyze

**Summary:**

Check min, max, mean, median and any outliers. Avg weight is 135 pounds with BMI of 24 and burn 2050 calories. Average steps is 10200, max is almost triple that at 36000 steps. Users spend on average 12 hours a day in sedentary minutes, only half hour in fairly and very active and 4 hours lightly active! Users also get about 7 hours of sleep.

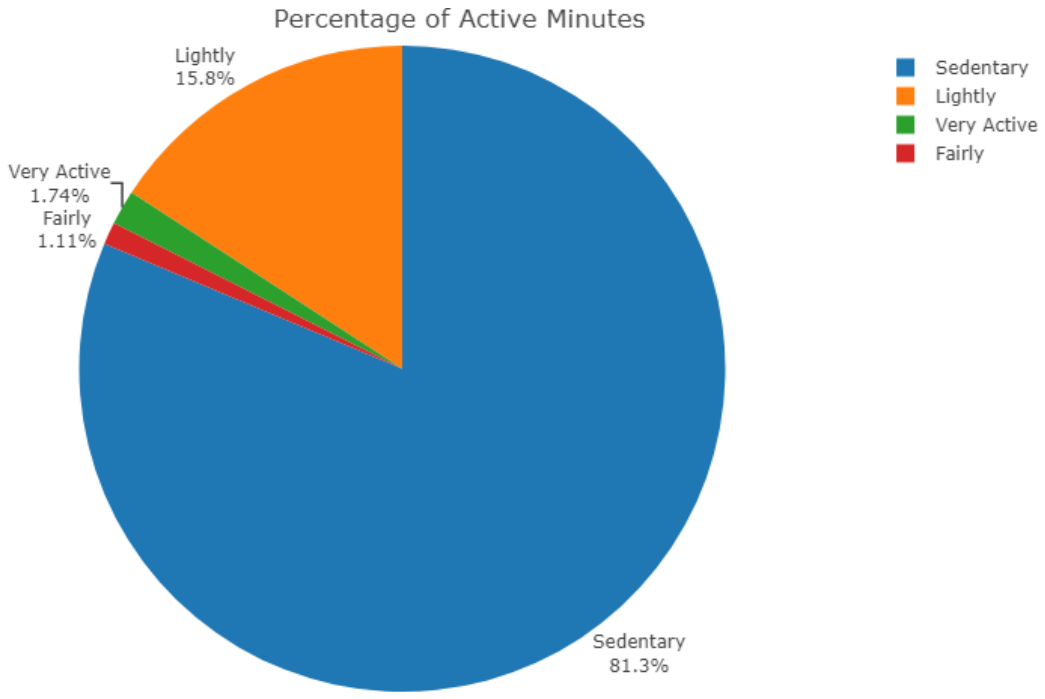
```
merged_data %>%
 dplyr::select(Weekday,
       TotalSteps,
       TotalDistance,
       VeryActiveMinutes,
       FairlyActiveMinutes,
       LightlyActiveMinutes,
       SedentaryMinutes,
       Calories,
       TotalMinutesAsleep,
       TotalTimeInBed,
       WeightPounds,
       BMI
       ) %>%
 summary()
```

```
     Weekday       TotalSteps    TotalDistance    VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes    Calories    TotalMinutesAsleep
Monday   :5609   Min.   :    0   Min.   : 0.000   Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0   Min.   :   0   Min.   : 58.0
Tuesday  :7004   1st Qu.: 5832   1st Qu.: 3.910   1st Qu.:  0.00    1st Qu.:  3.00      1st Qu.:194.0        1st Qu.: 637.0   1st Qu.:1850   1st Qu.:400.0
Wednesday:6988   Median :10199   Median : 6.820   Median : 15.00    Median : 14.00      Median :238.0        Median : 697.0   Median :2046   Median :442.0
Thursday :6930   Mean   : 9373   Mean   : 6.415   Mean   : 23.57    Mean   : 17.82      Mean   :232.2        Mean   : 722.6   Mean   :2103   Mean   :433.8
Friday   :5632   3rd Qu.:12109   3rd Qu.: 8.350   3rd Qu.: 38.00    3rd Qu.: 31.00      3rd Qu.:288.0        3rd Qu.: 745.0   3rd Qu.:2182   3rd Qu.:477.0
Saturday :5616   Max.   :36019   Max.   :28.030   Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0   Max.   :4900   Max.   :796.0
Sunday   :5610                                                                                                                            NA's   :971
TotalTimeInBed   WeightPounds        BMI
Min.   : 61.0   Min.   :116.0   Min.   :21.45
1st Qu.:421.0   1st Qu.:134.9   1st Qu.:23.89
Median :457.0   Median :135.6   Median :24.00
Mean   :458.2   Mean   :139.6   Mean   :24.42
3rd Qu.:510.0   3rd Qu.:136.7   3rd Qu.:24.21
Max.   :961.0   Max.   :294.3   Max.   :47.54
NA's   :971     NA's   :8881    NA's   :8881
```

**Active Minutes:**

Percentage of active minutes in the four categories: very active, fairly active, lightly active and sedentary. From the pie chart, we can see that most users spent 81.3% of their daily activity in sedentary minutes and only 1.74% in very active minutes.

```
percentage <- data.frame(
 level=c("Sedentary", "Lightly", "Fairly", "Very Active"),
 minutes=c(sedentary_percentage,lightly_percentage,fairly_percentage,active_percentage)
)

plot_ly(percentage, labels = ~level, values = ~minutes, type = 'pie',textposition = 'outside',textinfo =
'label+percent') %>%
 layout(title = 'Activity Level Minutes',
       xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
       yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```



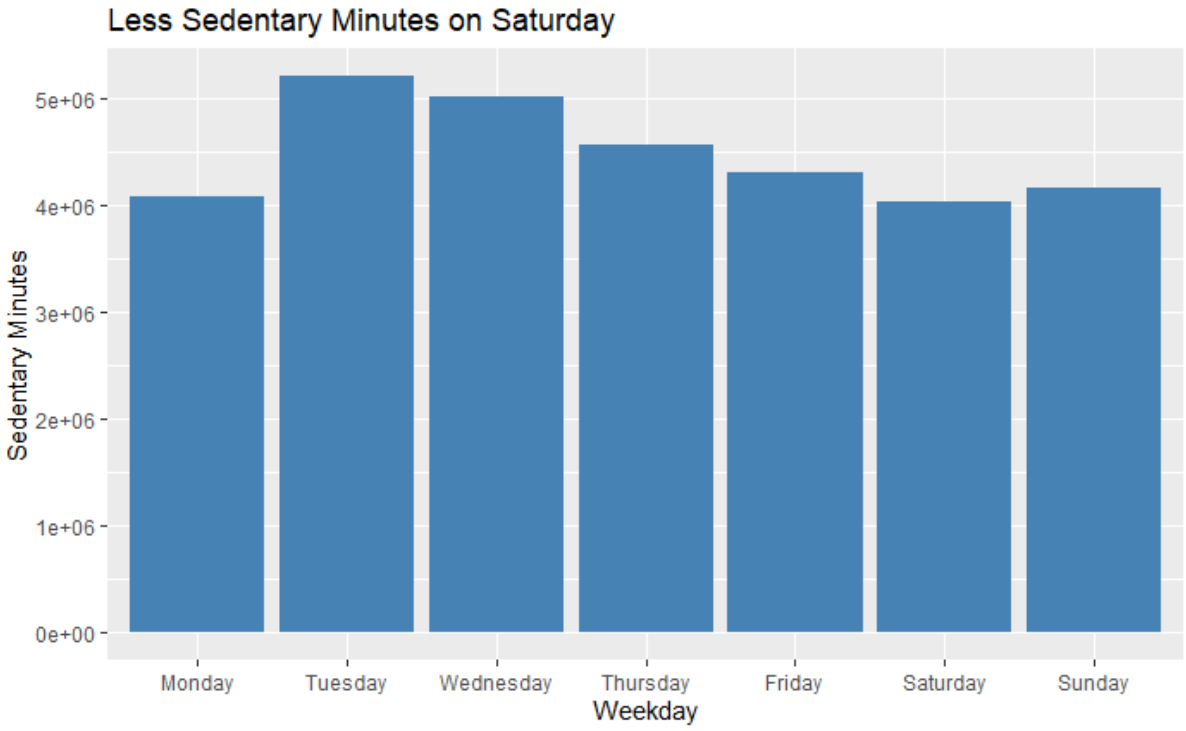

The American Heart Association and World Health Organization recommend at least 150 minutes of moderate-intensity activity or 75 minutes of vigorous activity, or a combination of both, each week. That means one needs an daily goal of 21.4 minutes of FairlyActiveMinutes or 10.7 minutes of VeryActiveMinutes.

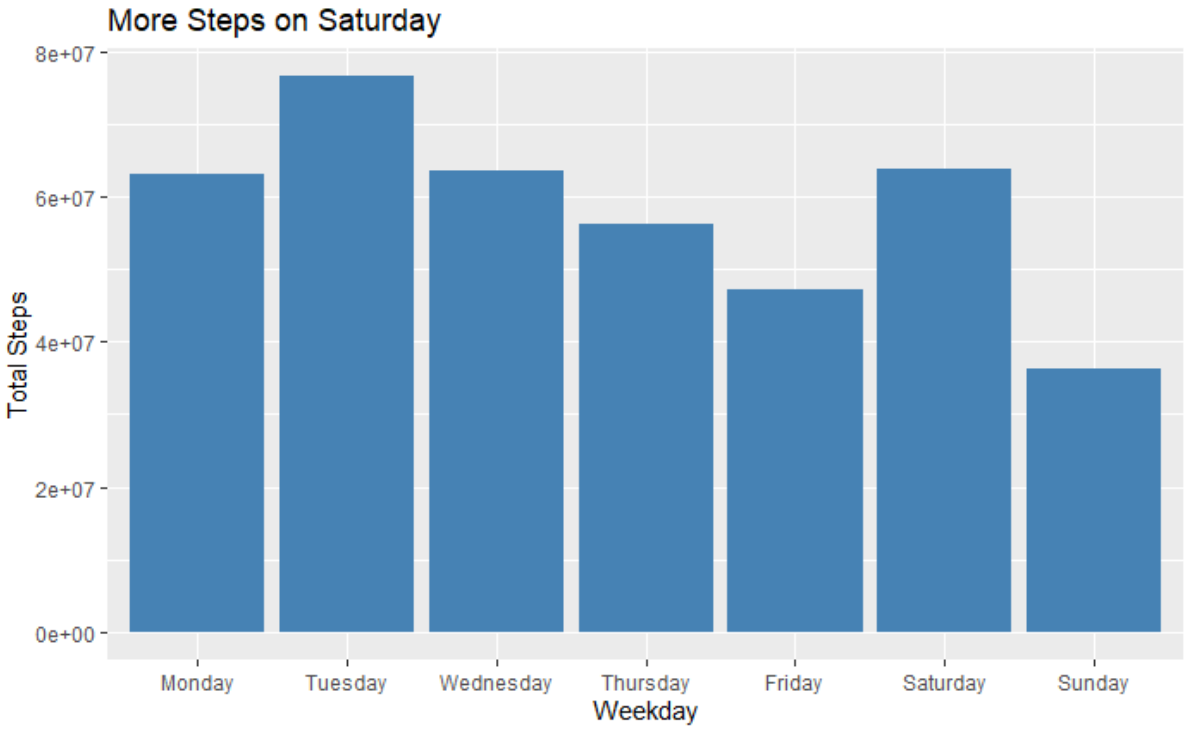
In our dataset, **30 users** met fairly active minutes or very active minutes.

```
active_users <- daily_activity %>%
 filter(FairlyActiveMinutes >= 21.4 | VeryActiveMinutes>=10.7) %>%
 group_by(Id) %>%
 count(Id)
```

**Noticeable Day:**

The bar graph shows that there is a jump on Saturday: user spent LESS time in sedentary minutes and take MORE steps.



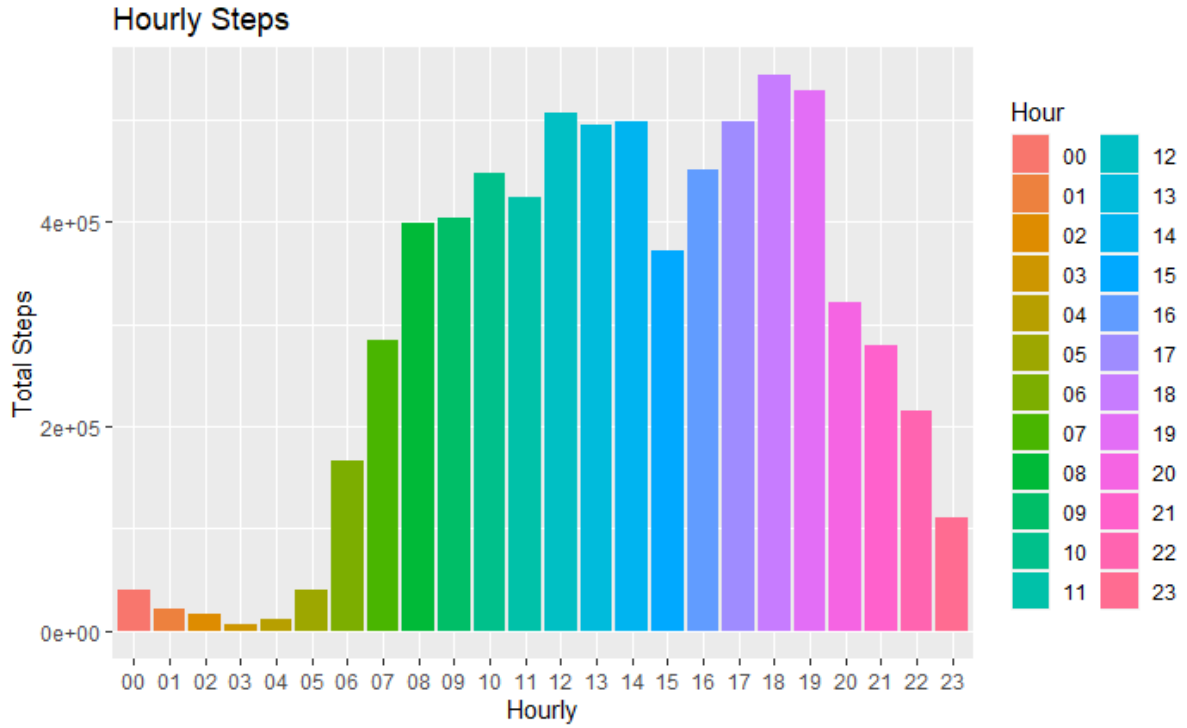

**Total Steps:**

Back to Analyze

Let's look at how active the users are per hourly in total steps. From 5PM to 7PM the users take the most steps.

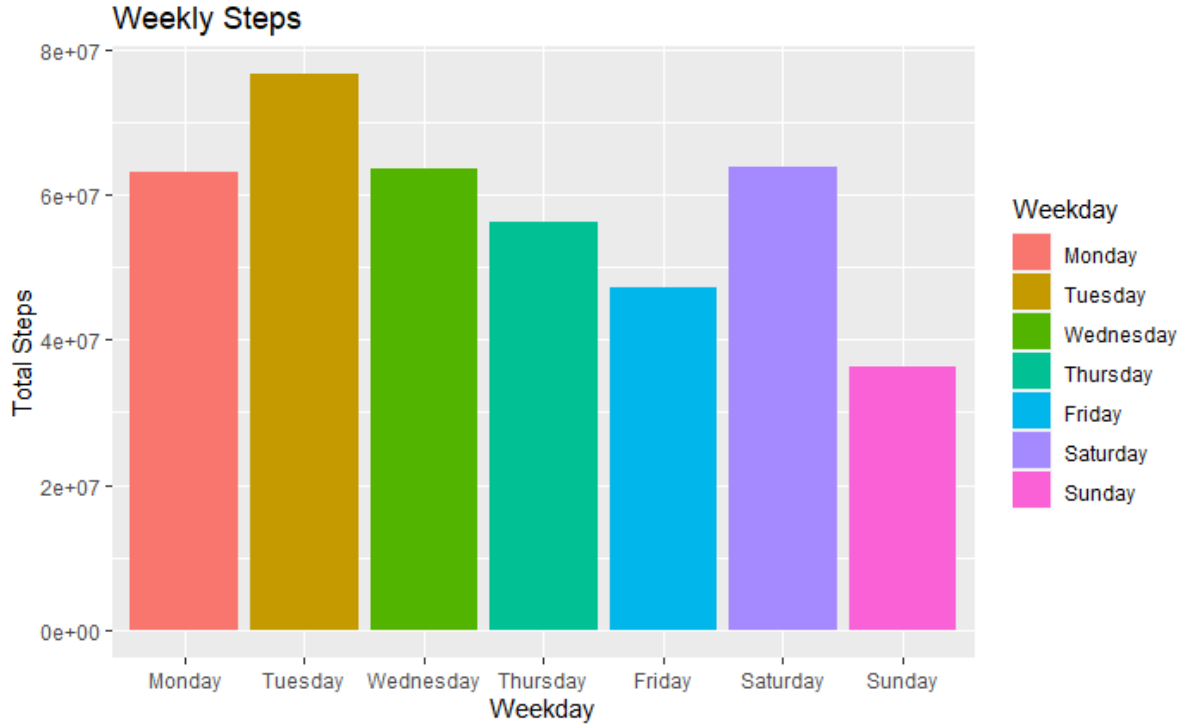
ggplot(data=hourly_step, aes(x=Hour, y=StepTotal, fill=Hour))+

```
geom_bar(stat="identity")+
labs(title="Hourly Steps")
```




This shows how active are the users in the weekly total steps. Tuesday and Saturdays the users take the most steps.

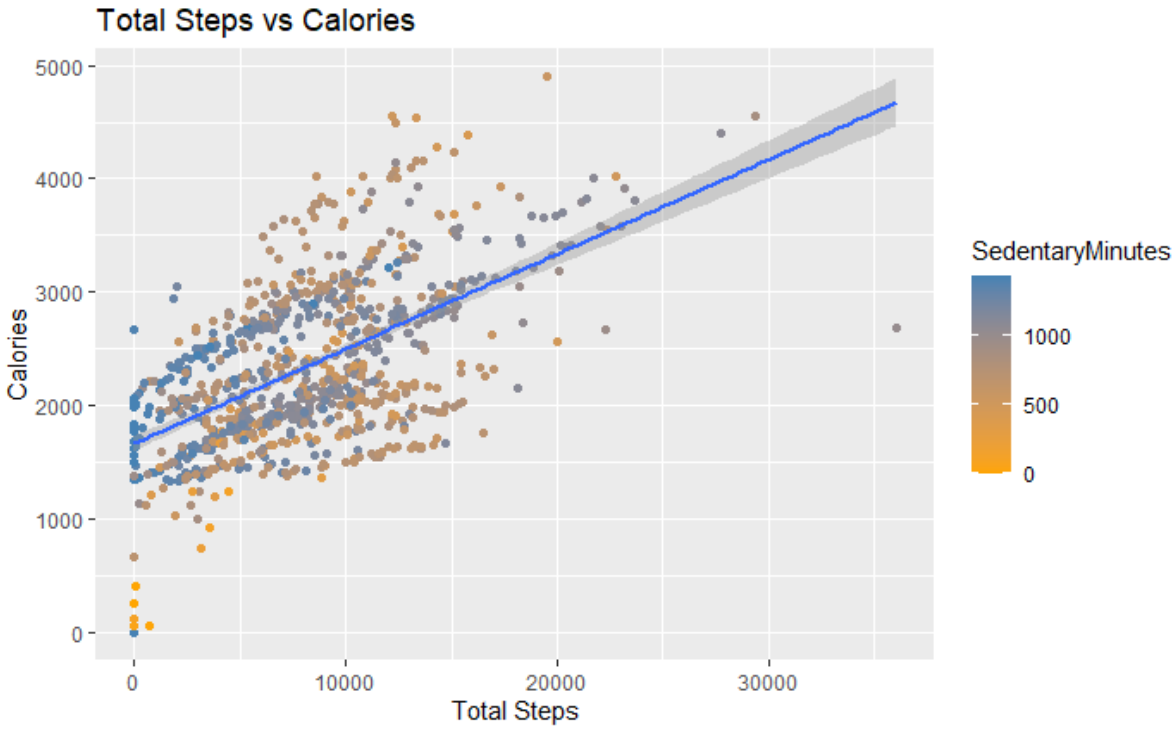
```
ggplot(data=merged_data, aes(x=Weekday, y=TotalSteps, fill=Weekday))+
 geom_bar(stat="identity")+
 ylab("Total Steps")
```



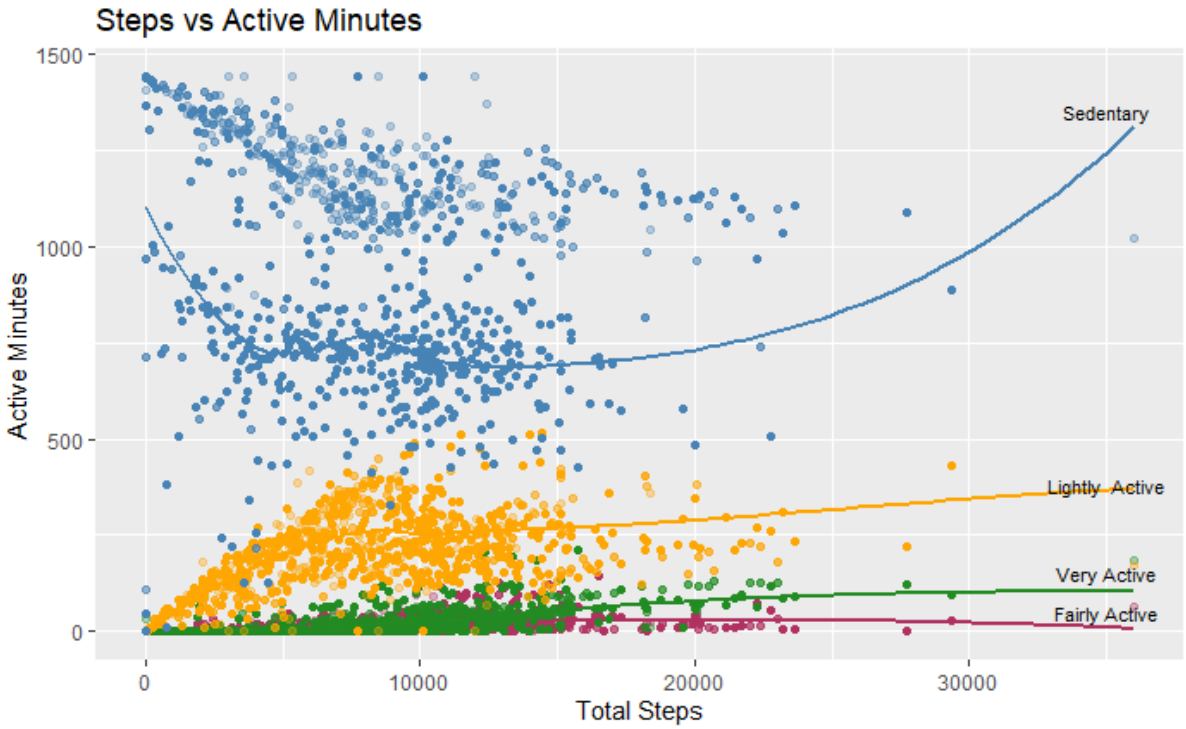
**Interesting findings:**

The more active that you're, the more steps you take, and the more calories you will burn. This is a crystal clear message, but we can still look into the data to find any interesting. Here we see that some users who are sedentary, take minimal steps, but are still able to burn from 1500 to 2500 calories compared to users who are more active, take more steps, but still burn similar calories.

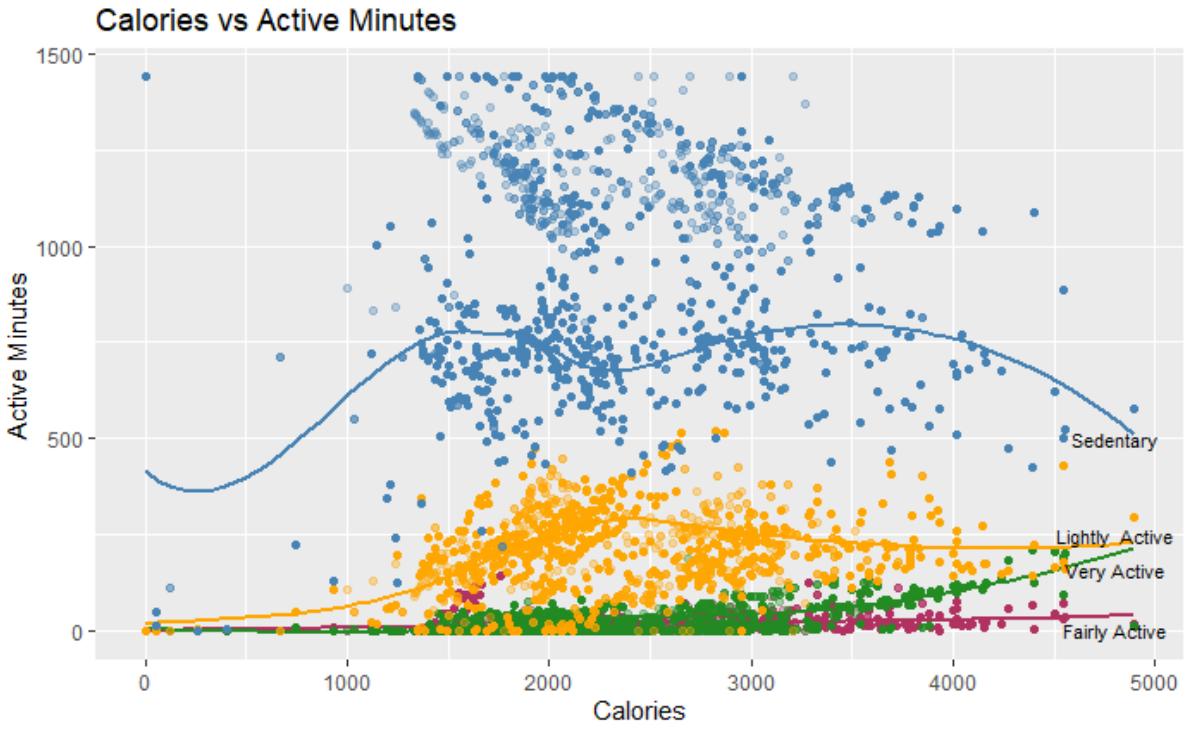
```
ggplot(data=daily_activity, aes(x=TotalSteps, y = Calories, color=SedentaryMinutes))+
  geom_point()+
  stat_smooth(method=lm)+
  scale_color_gradient(low="steelblue", high="orange")
```




Comparing the four active levels to the total steps, we see most data is concentrated on users who take about 5000 to 15000 steps a day. These users spent an average between 8 to 13 hours in sedentary, 5 hours in lightly active, and 1 to 2 hour for fairly and very active.




According to the World Health Organization, a moderately active woman between the ages of 26–50 needs to eat about 2,000 calories per day and a moderately active man between the ages of 26–45 needs 2,600 calories per day to maintain his weight. Comparing the four active levels to the calories, we see most data is concentrated on users who burn 2000 to 3000 calories a day. These users also spent an average between 8 to 13 hours in sedentary, 5 hours in lightly active, and 1 to 2 hour for fairly and very active. Additionally, we see that the sedentary line is leveling off toward the end while fairly + very active line is curing back up. This indicate that the users who burn more calories spend less time in sedentary, more time in fairly + active.



**Sleep:**

According to an article about Fitbit Sleep Study, 55 minutes are spent awake in bed before going to sleep. We have 13 users in our dataset spend 55 minutes awake before alseep.

```
awake_in_bed <- mutate(sleep_day, AwakeTime = TotalTimeInBed - TotalMinutesAsleep)
awake_in_bed <- awake_in_bed %>%
 filter(AwakeTime >= 55) %>%
 group_by(Id) %>%
 arrange(AwakeTime)
```

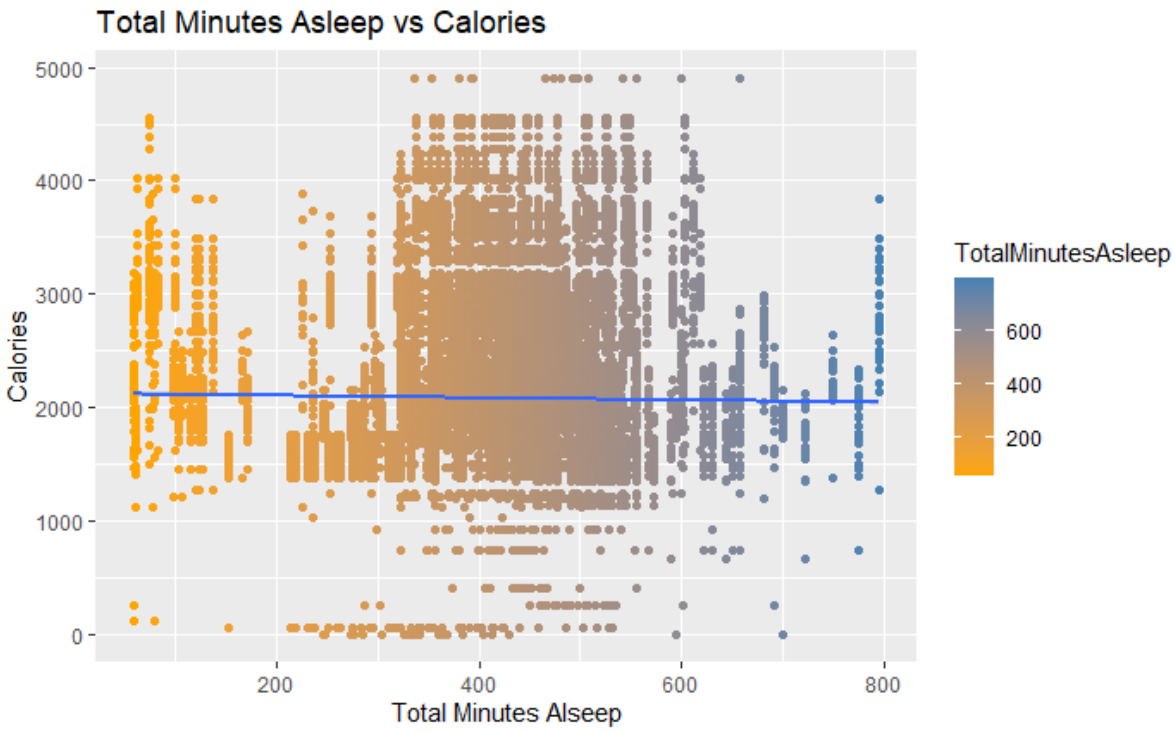
We can use regression analysis look at the variables and correlation. For R-squared, 0% indicates that the model explains none of the variability of the response data around its mean. Higher % indicates that the model explains more of the variability of the response data around its mean. Positive slope means variables increase/decrease with each other, and negative means one variable go up and the other go down. We want to look at if users who spend more time in sedentary minutes spend more time sleeping as well. We can use regression analysis lm() to check for the dependent and independent variables. We also find that how many minutes a user sleeps have very weak correlation with how long they spend in sedentary minutes during the day.

```
sedentary_vs_sleep.mod <- lm(SedentaryMinutes ~ TotalMinutesAsleep, data = merged_data)
summary(sedentary_vs_sleep.mod)
```

```
Call:
lm(formula = VeryActiveMinutes ~ TotalMinutesAsleep, data = merged_data)

Residuals:
    Min      1Q  Median      3Q     Max
-23.500 -22.737  -7.984  14.862 187.401

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        23.595768   0.582829  40.485   <2e-16 ***
TotalMinutesAsleep -0.001652   0.001313  -1.258    0.208
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.26 on 42416 degrees of freedom
  (971 observations deleted due to missingness)
Multiple R-squared:  3.732e-05, Adjusted R-squared:  1.374e-05
F-statistic: 1.583 on 1 and 42416 DF,  p-value: 0.2084
```

How about calories vs asleep? Do people who sleep more burn less calories? Plotting the two variables we can see that there is no such a correlation.

```
ggplot(data=merged_data, aes(x=TotalMinutesAsleep, y = Calories, color=TotalMinutesAsleep))+
 geom_point()+
 labs(title="Total Minutes Asleep vs Calories")+
 xlab("Total Minutes Alseep")+
 stat_smooth(method=lm)+
 scale_color_gradient(low="orange", high="steelblue")
```




5. Share

Then we shared our findings and recommendations with team leaders using a PowerPoint presentation.


6. Act

**Conclusions based on our analysis**

Sedentary make up a significant portion that is 80.5% of users daily active minutes. Users spend on average 11.5 hours a day in sedentary minutes, 4 hours lightly active, and only half-hour in fairly and very active.

We see a significant change on Saturdays: users take more steps, burn more calories, and spend less time sedentary. Sunday is the most "sluggish" day for users.

53.5% of the users who recorded their sleep data spent 55 minutes awake in bed before falling asleep. We could find a way of making good use of this time by offering users a game to play in order to stay occupied.

Users who take the most steps from 5 PM to 7 PM. However, users who are sedentary take minimal steps and burn 1500 to 2500 calories compared to users who are more active, taking more steps, but still burning similar calories.

Educational healthy style campaign encourages users to have short active exercises during the week, longer during the weekends, especially on Sunday where we see the lowest steps and most sedentary minutes.

Educational healthy style campaign can pair with a point-award incentive system. Users completing the whole week's exercise will receive Bellabeat points on products/memberships.

The product, such as Leaf wellness tracker, can beat or vibrate after a prolonged period of sedentary minutes, signaling to the user it's time to get active! Similarly, it can also remind the user it's time to sleep after sensing a prolonged awake time in bed.

Marketing recommendations to expand globally

Obtain a bigger data sample for an accurate analysis while encouraging users to use a wifi-connected scale device instead of making manual weight entries.