Bella Beat Case Study Capstone by: Amanda Waldron

Loading up data and packages

```
library(jtools)
library(xtable)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'broom':
##   method            from
##   tidy.glht         jtools
##   tidy.summary.glht jtools

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v stringr 1.5.0
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
daily_activity <- read.csv("dailyActivity_merged.csv")
```

We'll create another dataframe for the sleep data.

```
sleep_day <- read.csv("sleepDay_merged.csv")
```

Exploring a few key tables

Take a look at the daily_activity data.

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

Identify all the columsn in the daily_activity data.

```
colnames(daily_activity)
```

```
##  [1] "Id"                       "ActivityDate"            
##  [3] "TotalSteps"               "TotalDistance"           
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance" 
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"     
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"        
## [15] "Calories"
```

Take a look at the sleep_day data.

```
head(sleep_day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

Identify all the columns in the daily_activity data.

```
colnames(sleep_day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

Note that both datasets have the 'Id' field - this can be used to merge the datasets.

Understanding some summary statistics

How many unique participants are there in each dataframe? It looks like there may be more participants in the daily activity dataset than the sleep dataset.

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

How many observations are there in each dataframe?

```
nrow(daily_activity)
```

```
## [1] 940
```

```
nrow(sleep_day)
```

```
## [1] 413
```

Statistical Summary - For the daily activity dataframe:

```
daily_activity %>%
  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
```

Statistical Summary - For the sleep dataframe:

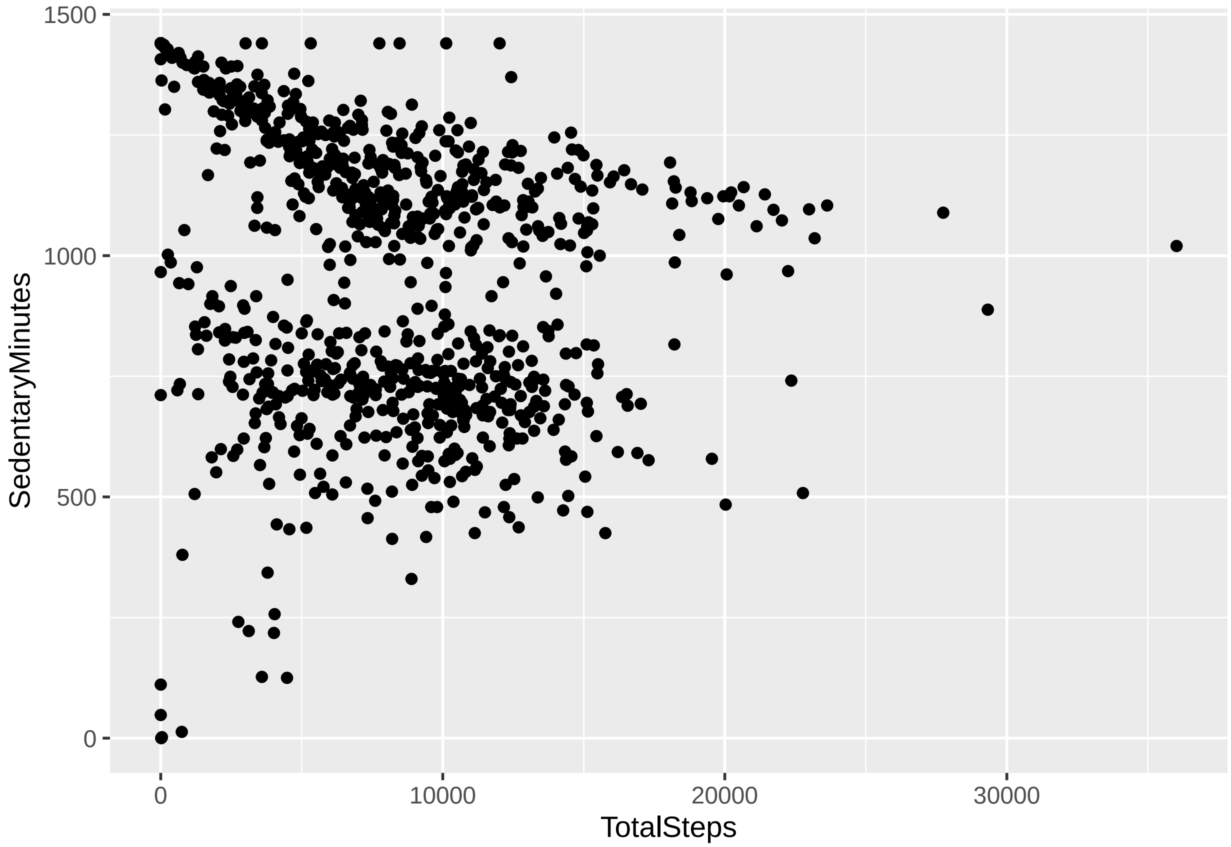
```
sleep_day %>%
  select(TotalSleepRecords,
  TotalMinutesAsleep,
  TotalTimeInBed) %>%
  summary()
```

```
##  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
```

What does this tell us about how this sample of people's activities?

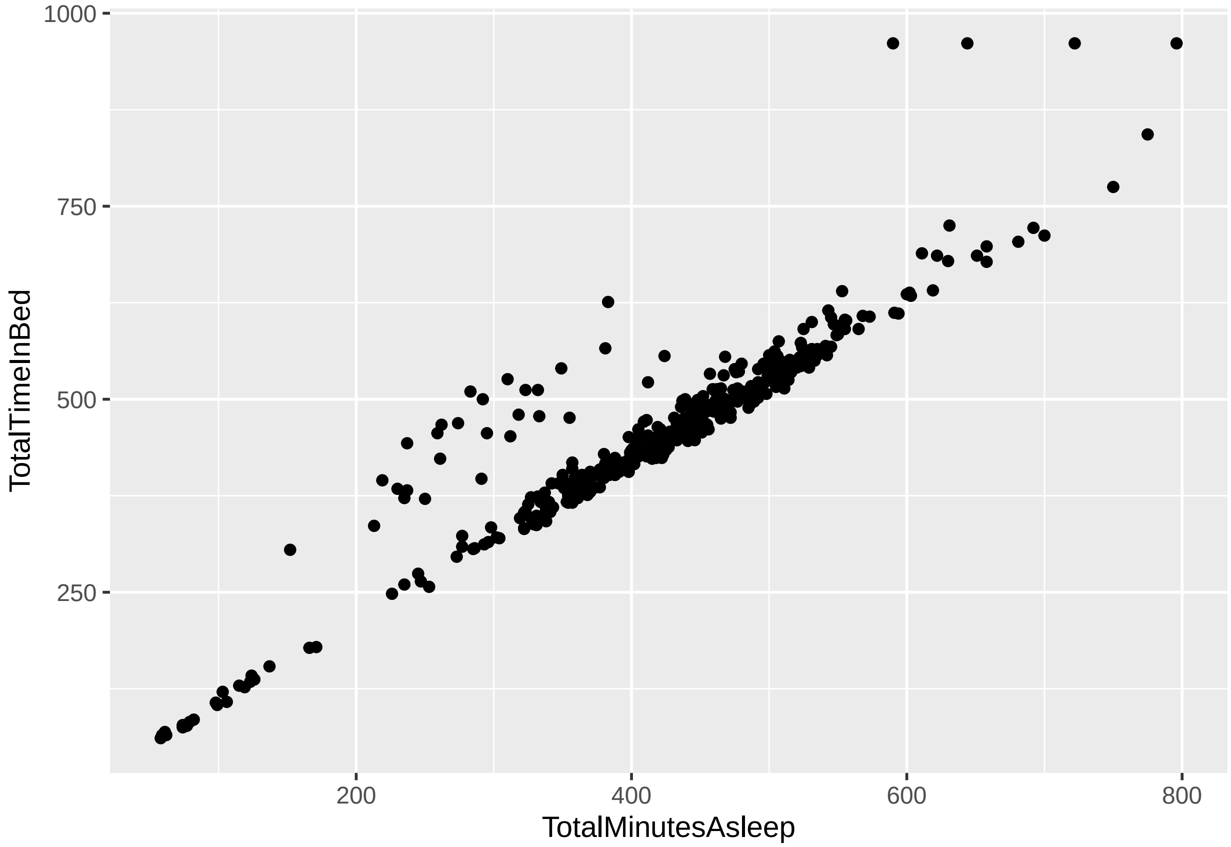
Plotting a few explorations

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point()
```



There is a negative correlation, thus a negateive relationship between steps taken in a day and sedentary minutes. It looks like the higher sedentary minutes a person, lesser steps are taken. This can help inform the customer segments we can market to by marketing to people as a way to get more steps in. The new device can encourage them similar to an apple watch where it reminds you to stand up but take it a step further and remind the person that it is time to move as you have been sitting for “x” amount of time hence the sedentary minutes that are being calculated.

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



There is a positive correlation with some outliers. It is mostly linear with somepeople appearing to spend a lot of time in bed but not so much sleeping.

These trends indicate that the product might need to put more energy into marketing sleep hygiene and its relationship to the users health.

Merging these two datasets together

```
combined_data <- merge(sleep_day, daily_activity, by="Id")
```

Take a look at how many participants are in this data set.

```
n_distinct(combined_data$Id)
```

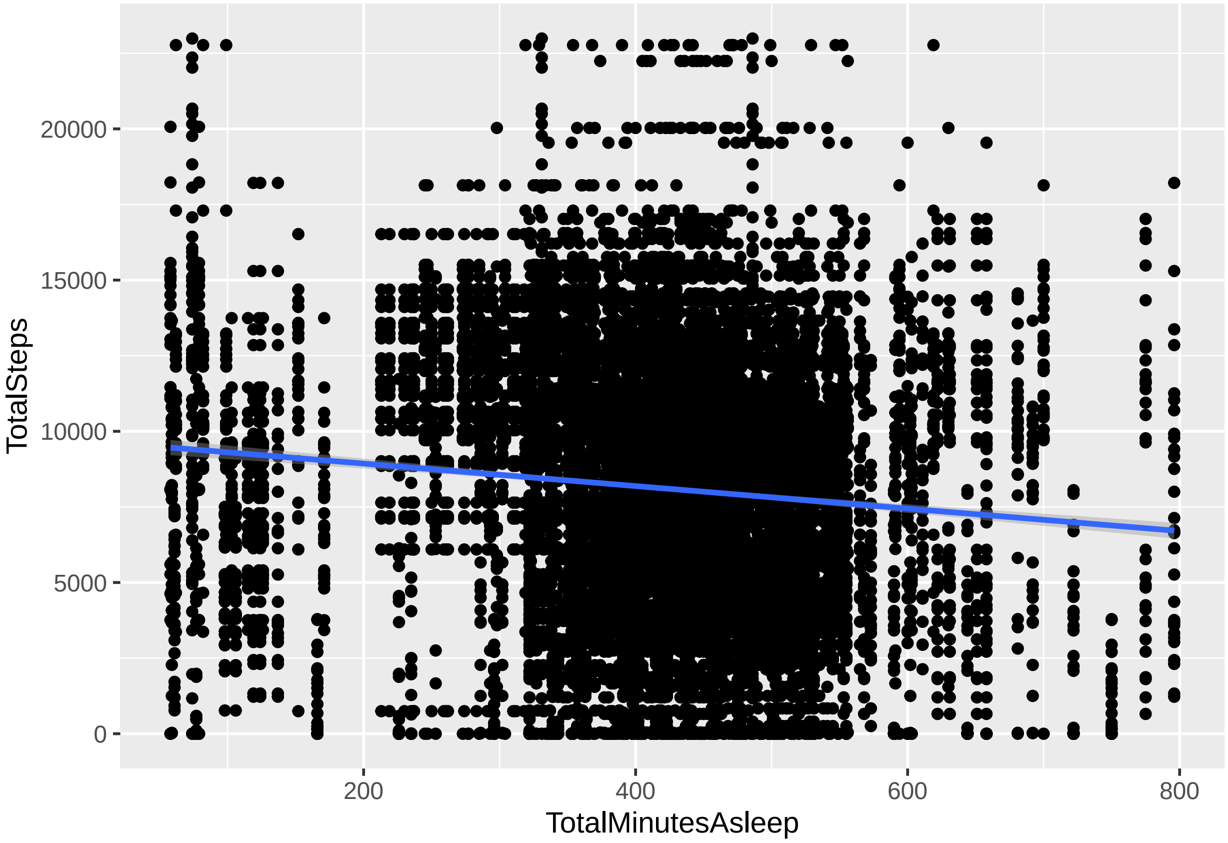
```
## [1] 24
```

Note that there were more participant Ids in the daily activity dataset that have been filtered out using merge.

Activity and Sleep

```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=TotalSteps)) + geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



I was hoping to find a clear and strong relationship that indicates that participants who sleep more also take more steps or fewer steps per day. I am not finding a clear relationship between the two. This indicates to me that Bella Beat should really zero in on sleeping data and activity during the day. If we can get more information collected on this then I think a relationship between the two can be found. Sleep is important for health and that includes having energy to sustain you throughout the day. Also activity throughout the day usually helps a person sleep better. My point is to say the data here is not supporting this but I am sure if Bella Beat has better tracking on this information then that relationship will be supported and shown in their data. Therefore, marketing should focus on a product that has advanced tracking of sleep and acitivity so that we can see the relationship between the two in the average person and in turn help the person see these things themselves and track their own behaviors.

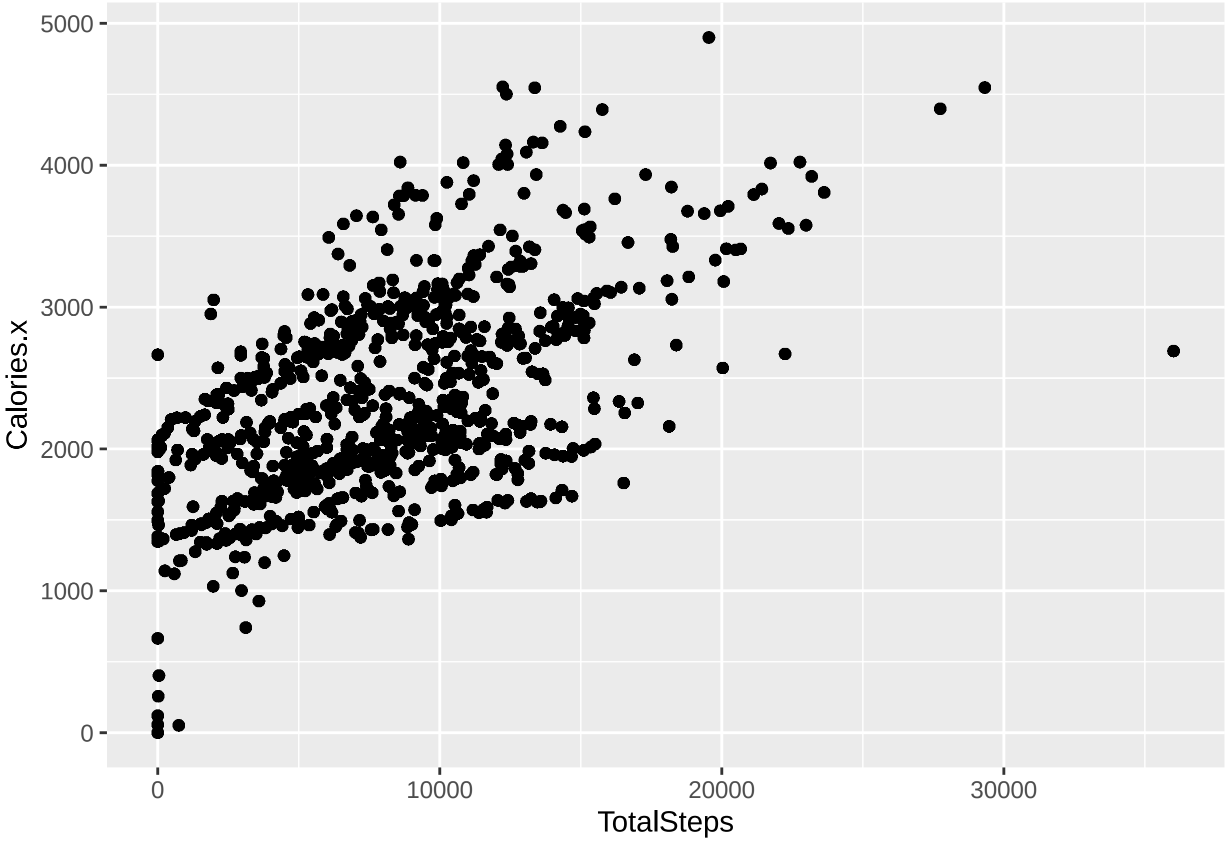
```
calories_day <- read.csv("dailyCalories_merged.csv")

calories_day %>% head()
```

```
##           Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

```
combined_data2 <- merge(daily_activity, calories_day, by= "Id")
```

```
ggplot(data=combined_data2, aes(x=TotalSteps, y=Calories.x)) +
        geom_point()
```



There seems to be positive relationship between total steps taken and calories but it is not strong. This indicates that they can probably build upon this such as calories burned, intake, and activity throughout the day to support a healthy lifestyle. I am sure with further exploration that the data can tell us more but so far it is clear to me that marketing can focus on digging deeper into total steps a person takes per day and calories taken in and bruned along with sleep and activity data. This will help optimize their usefulness to the average person who wants to track these things, lead a healthier lifestyle and be able to see the relationships in their day to day to know what they themselves can work on as an individual which means we have to provide the ability to get more detailed data that allows this.