BellaBeat Capstone Case Study

Adeola

2022-10-25

Introduction

This is a Capstone Project for the Google Data Analytics Certification. The Case Study involves Bellabeat, a company that provides smart devices with a focus on women. They are looking to improve their business and as a junior data analyst, I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights I discover will then help guide marketing strategy for the company. In order to answer the key business questions, I will follow the steps of the data analysis process: ask, prepare, process, analyze, share, and act.

Ask Phase

The key stakeholders?

- **Urska Srsen:** cofounder and Chief Creative Officer
- **Sando Mur:** cofounder, mathematician, key member of executive team
- **Marketing and Analytics Team**

The business task is to analyze how customers use smart device products. We will attempt to find answers to these questions as a guide to our analysis:

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

Prepare

- **Where is your data stored?** The data is gathered from a public data that explores smart device users's daily habits and shared by the user with their consent. The dataset used for this case study is FitBit Fitness Tracker Data. This dataset was uploaded to Kaggle by the user Möbius.
- **How is the data organized? Is it in long or wide format?** This data is organized in both long and wide format depending on the tool used to download and open the file, it is also organized in various formats for time and date.
- **Are there issues with bias or credibility in this data? Does your data ROCCC?** The data is presumed reliable due to it being provided by the company. But unfortunately we cannot provide an up to date analysis due to this data being from 3/2/2016 - 5/12/2016.
- **How are you addressing licensing, privacy, security, and accessibility?** The data gathered has been anonymized with no personal information included.
- **How does it help you answer your question?** With the data provided we will be able to analyze it and look for any trends that the stakeholders could use to improve upon their product.
- **Are there any problems with the data?** As noted above the data is outdated because it was uploaded 6 years ago so current data would be needed to make an accurate analysis for the current year. While the description of the dataset on Kaggle says there are 30 users in the data, after getting familiar with the data i discovered there are 33 unique users. Also 33 is a small sample size to get an unbiased result.

Process Data Phase

From the complete dataset which consists of 18 csv files, I used 2 sets of data, which were: * bella_beat - dailyActivity_merged.csv * bella_beat - sleep_day_merged.csv

The activity dataset has a lot of it combined so that will be the primary dataset that I use.

Data Cleaning

For this case study I will be using R for my analysis but I primarily used Excel to clean and organize the datasets. Then used RStudio for analysis and visualizations. Cleaning the data consisted of:

- checked for duplicates- deleting any that were found. (there were 3 found in sleep dataset)
- formatted the date to short date format to be reflected as yyy/mm/dd
- formatted the distances to be a Number and uniform to 2 decimal places
- deleted TrackerDistance column as it was the same as the TotalDistance column
- deleted the LoggedActivitiesDistance column as it had mostly zero's in it.
- checked both datasets for missing values

With the data now clean I began my analysis in R which began with:

- uploading csv file to Rstudio: I uploaded the bella_beat - dailyActivity_merged.csv and bella_beat - sleep_day_merged.csv
- installing and loading the necessary packages:

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(readr)
library(tidyr)
library(dplyr)
```

Loading CSV files

Here I am creating a dataframe named 'daily_activity and sleep_day' respectively i will read in one of the CSV files from the dataset and view them.

```
daily_activity <- read_csv("bellabeat_project/bella_beat - dailyActivity_merged.csv")
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesD...
## date  (1): ActivityDate
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
view(daily_activity)
sleep_day <- read_csv("bellabeat_project/bella_beat - sleep_day.csv")
```

```
## Rows: 410 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
## date (1): SleepDay
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
view(sleep_day)
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"
```

```
head(daily_activity)
```

```
## # A tibble: 6 x 15
##           Id ActivityD~1 Total~2 Total~3 Track~4 Logge~5 VeryA~6 Moder~7 Light~8
##        <dbl> <date>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1503960366 2016-04-12    13162    8.5     8.5        0    1.88    0.55    6.06
## 2 1503960366 2016-04-13    10735    6.97    6.97       0    1.57    0.69    4.71
## 3 1503960366 2016-04-14    10460    6.74    6.74       0    2.44    0.4     3.91
## 4 1503960366 2016-04-15     9762    6.28    6.28       0    2.14    1.26    2.83
## 5 1503960366 2016-04-16    12669    8.16    8.16       0    2.71    0.41    5.04
## 6 1503960366 2016-04-17     9705    6.48    6.48       0    3.19    0.78    2.51
## # ... with 6 more variables: SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>, and
## #   abbreviated variable names 1: ActivityDate, 2: TotalSteps,
## #   3: TotalDistance, 4: TrackerDistance, 5: LoggedActivitiesDistance,
## #   6: VeryActiveDistance, 7: ModeratelyActiveDistance, 8: LightActiveDistance
```

```
colnames(sleep_day)
```

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

```
head(sleep_day)
```

```
## # A tibble: 6 x 5
##           Id SleepDay   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##        <dbl> <date>                 <dbl>              <dbl>          <dbl>
## 1 1503960366 2016-04-12                 1                327            346
## 2 1503960366 2016-04-13                 2                384            407
## 3 1503960366 2016-04-15                 1                412            442
## 4 1503960366 2016-04-16                 2                340            367
## 5 1503960366 2016-04-17                 1                700            712
## 6 1503960366 2016-04-19                 1                304            320
```

```
str(daily_activity)
```

```
## spc_tbl_ [940 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                      : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate            : Date[1:940], format: "2016-04-12" "2016-04-13" ...
##  $ TotalSteps              : num [1:940] 13162 10735 10460 9762 12669 ...
##  $ TotalDistance           : num [1:940] 8.5 6.97 6.74 6.28 8.16 6.48 8.59 9.88 6.68 6.34 ...
##  $ TrackerDistance         : num [1:940] 8.5 6.97 6.74 6.28 8.16 6.48 8.59 9.88 6.68 6.34 ...
##  $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance      : num [1:940] 1.88 1.57 2.44 2.14 2.71 3.19 3.25 3.53 1.96 1.34 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 0.78 0.64 1.32 0.48 0.35 ...
##  $ LightActiveDistance     : num [1:940] 6.06 4.71 3.91 2.83 5.04 2.51 4.71 5.03 4.24 4.65 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes       : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes     : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes    : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes        : num [1:940] 728 776 1218 726 773 ...
##  $ Calories                : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDate = col_date(format = ""),
##   ..   TotalSteps = col_double(),
##   ..   TotalDistance = col_double(),
##   ..   TrackerDistance = col_double(),
##   ..   LoggedActivitiesDistance = col_double(),
##   ..   VeryActiveDistance = col_double(),
##   ..   ModeratelyActiveDistance = col_double(),
##   ..   LightActiveDistance = col_double(),
##   ..   SedentaryActiveDistance = col_double(),
##   ..   VeryActiveMinutes = col_double(),
##   ..   FairlyActiveMinutes = col_double(),
##   ..   LightlyActiveMinutes = col_double(),
##   ..   SedentaryMinutes = col_double(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(sleep_day)
```

```
## spc_tbl_ [410 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : Date[1:410], format: "2016-04-12" "2016-04-13" ...
##  $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   SleepDay = col_date(format = ""),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

To get the number of unique users we run the below code;

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

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

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

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

I'll rename date columns in the two datasets to "date" so that I can merge the data based on their date and Id which are the two variables both datasets have in common.

```
sleep_day <- rename (sleep_day, date = SleepDay)
view(sleep_day)
daily_activity <- rename(daily_activity, date = ActivityDate)
view(daily_activity)
```

then i will merge the two datasets and check to see if the data is merged.

```
dailyactivity_sleepday_merged <- merge(x=daily_activity, y=sleep_day, by=c("Id","date"), all=TRUE)
view(dailyactivity_sleepday_merged)
head(dailyactivity_sleepday_merged)
```

```
##           Id       date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2016-04-12      13162          8.50            8.50
## 2 1503960366 2016-04-13      10735          6.97            6.97
## 3 1503960366 2016-04-14      10460          6.74            6.74
## 4 1503960366 2016-04-15       9762          6.28            6.28
## 5 1503960366 2016-04-16      12669          8.16            8.16
## 6 1503960366 2016-04-17       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
##   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1                 1                327            346
## 2                 2                384            407
## 3                NA                 NA             NA
## 4                 1                412            442
## 5                 2                340            367
## 6                 1                700            712
```

Adding a weekday column

```
dailyactivity_sleepday_merged <- transform(dailyactivity_sleepday_merged, weekday=weekdays(date))
glimpse(dailyactivity_sleepday_merged)
```

```
## Rows: 940
## Columns: 19
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ date                     <date> 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~
## $ TotalSleepRecords        <dbl> 1, 2, NA, 1, 2, 1, NA, 1, 1, 1, NA, 1, 1, 1, ~
## $ TotalMinutesAsleep       <dbl> 327, 384, NA, 412, 340, 700, NA, 304, 360, 32~
## $ TotalTimeInBed           <dbl> 346, 407, NA, 442, 367, 712, NA, 320, 377, 36~
## $ weekday                  <chr> "Tuesday", "Wednesday", "Thursday", "Friday",~
```

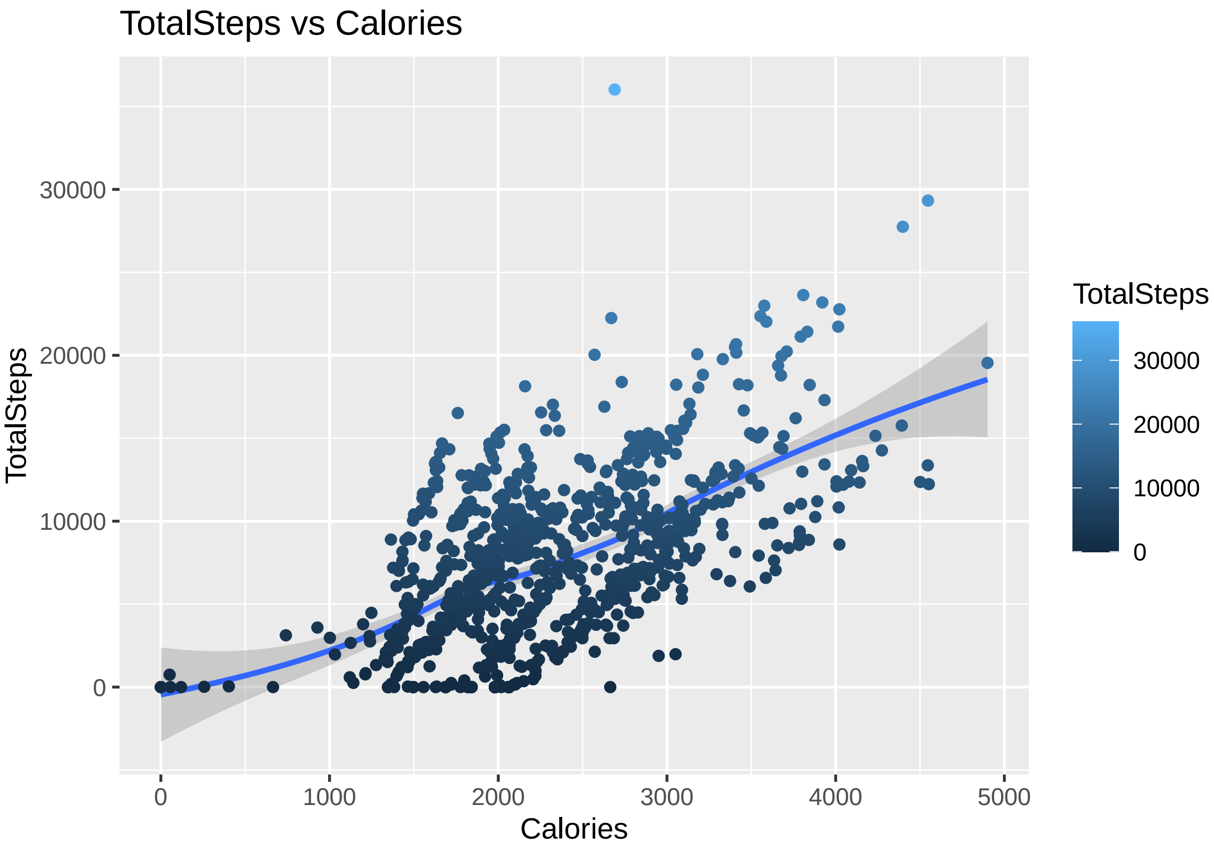
```
view(dailyactivity_sleepday_merged)
```

Analyze & Share

Now that the data has been merged, I created a visualization to check the correlation between TotalSteps and calories

```
ggplot(data = dailyactivity_sleepday_merged) +
  geom_smooth(mapping = aes(x=Calories, y = TotalSteps, colour = TotalSteps)) +
  geom_point(mapping = aes(x=Calories, y = TotalSteps, colour = TotalSteps)) +
  labs(title = "TotalSteps vs Calories")
```

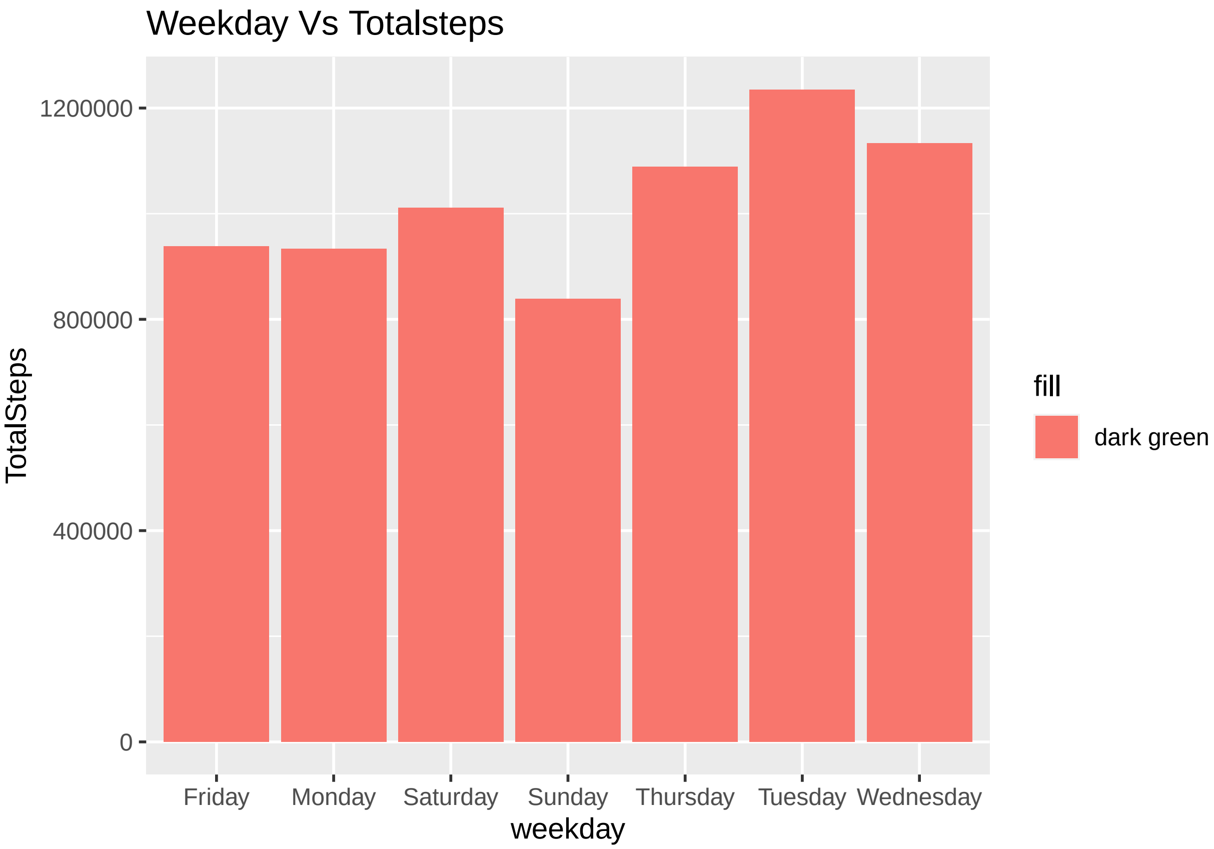
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The plot above shows a positive correlation between the amount of steps one takes and calories burned. the higher the steps the higher amount of calories burned.

Range of steps per weekday

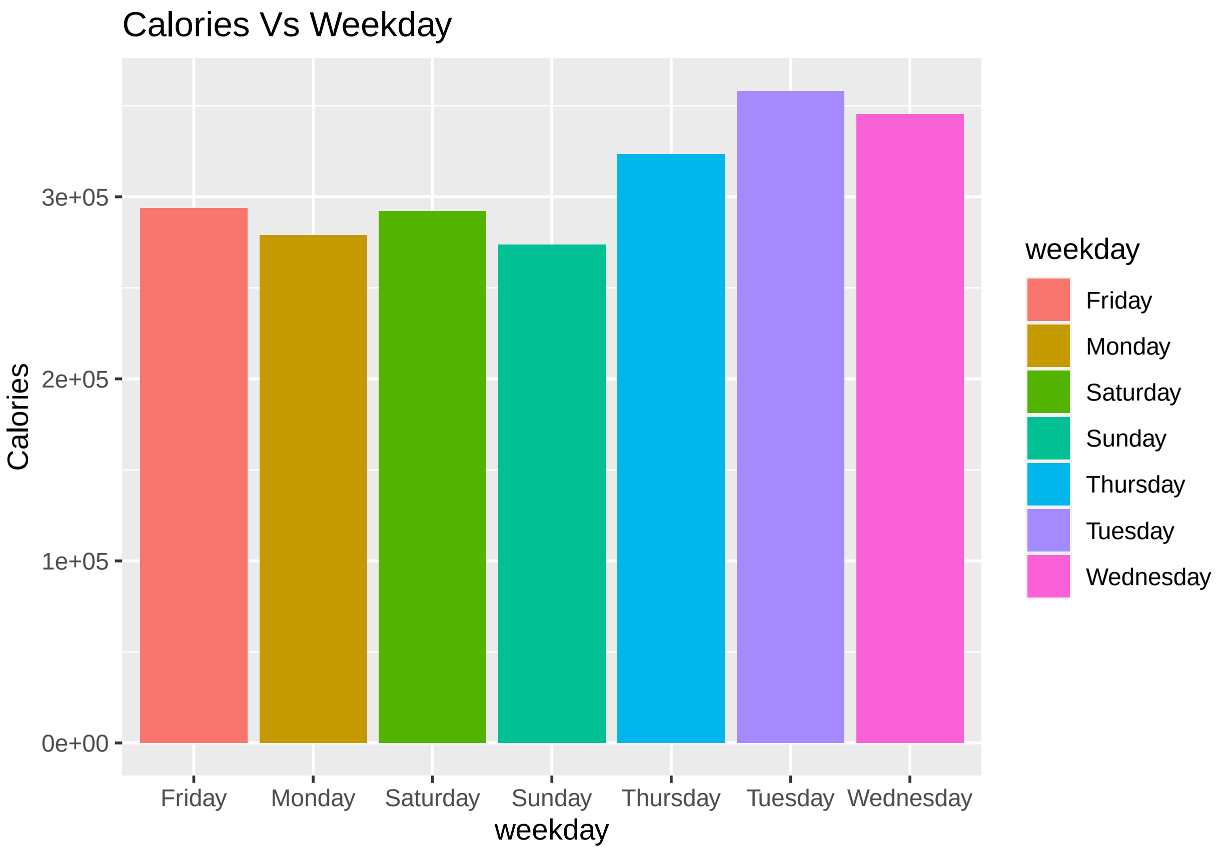
```
ggplot(data = dailyactivity_sleepday_merged) +
  geom_col (mapping = aes(x = weekday, y = TotalSteps, fill = "dark green")) +
  labs(title = "Weekday Vs Totalsteps")
```



The chat above shows that Tuesday has the highest TotalSteps taken.

Total minutes asleep across weekdays

```
ggplot(data = dailyactivity_sleepday_merged) +
  geom_col(mapping = aes(x=weekday, y=Calories, fill = weekday)) +
  labs(title = "Calories Vs Weekday")
```

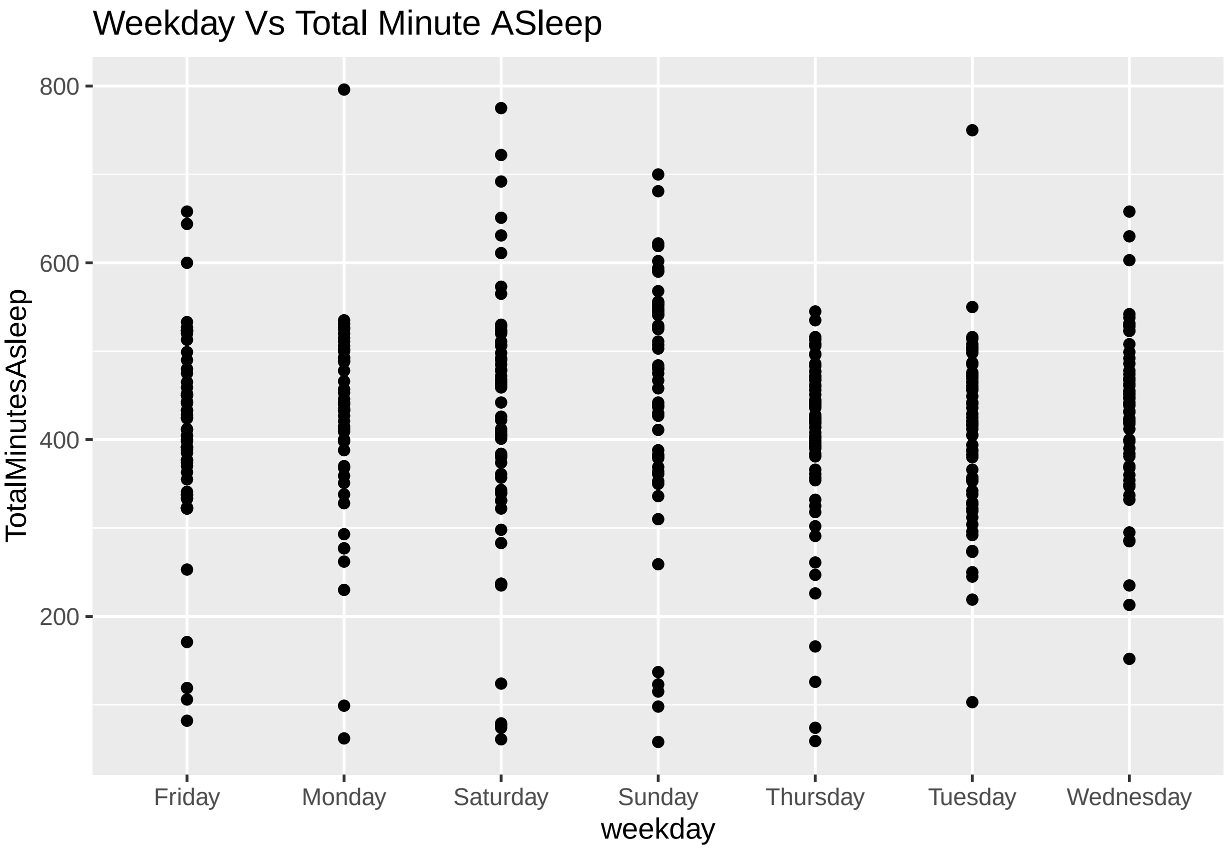


This plot is showing us a grand total of how many calories each participant burned each weekday. I wanted to get an idea of what days of the week people are burning the most and it turns out they are most active on tuesday

Total minute asleep across weekday

```
ggplot(data = dailyactivity_sleepday_merged) +
  geom_point(mapping = aes(x = weekday, y = TotalMinutesAsleep,)) +
  labs(title = "Weekday Vs Total Minute ASleep")
```

```
## Warning: Removed 530 rows containing missing values (geom_point).
```



We are looking at the total minutes alseep across weekdays. From earlier plots we can see a trend of participants reporting a greater amount of activity Tuesday thru Thursday and Saturdays. But, when looking at the amount of sleep people are reporting approximately the same amount across the board, no matter how much exercise/ activity they reported those days. For the most part it looks like people are sleeping anywhere from 6.6 to 8.5 hours on a nightly basis. Couple things to note are: the concentration between 25%-75% of participants is smallest in the middle of the week (Wed/Thur/Fri), with more variation on the weekends.

Act

Based on the insights found above we have the following recommendations.

- Our first recommendation would be to collect more data internally for us to analyze. It is difficult to make strong recommendations due to the dataset being outdated and a small sample size. So we would ask to be provided with a recent and larger sample of data in order to provide the best analysis possible.
- With the data provided there is a strong correlation between distance walked and calories burned. Something to aid users would be to create notifications when the user has been sedentary for too long as a way to encourage them to become active.
- As we have found that certain days of the week are more problematic on average when it comes to both activity and sleep, a recommendation would be to track this data in the app and give notifications on these problem days to remind the user so they can act to improve upon them.
- Based on this analysis I would recommend Bellabeat continue to market to woman focused on overall health and wellness and on-the-go women.