Bellabeats_Case_Study

Stephanie Sun

2023-03-03

Phases of Data Analysis

Bellabeats Case Study for Google Data Analytics Certificate Program

1. The Ask

-Business Task: How can the client, Bellabeats, grow their market share in the Smart Device industry?

-Insight into competitor, FitBit, consumer behaviour will inform opportunities for growth by finding patterns and understanding needs.

-Stakeholders: Bellabeat founders - Urska and Sando

2. Prepare The Data

-Client suggests FitBit data from Kaggle by Mobius

-Additional Google Trends data can be used to inform marketing and advertising decisions

-Stored in Google Drive Folder for easy access with Sheets in order to verify data integrity

-Limitations:Sample size is small and is not a strong representation of the population

3. Process The Data

-Spreadsheets used to clean data (looked for nulls, inconsistencies, incomplete columns)

-R will be used for data exploration

-Cleaning documented in notes tab of csv files

4. Data Analysis

-Cleaned datasets downloaded as csv files and imported into R Studio

-Analysis completed in R Studio desktop

-**Most users are a healthy BMI and they tracked their daily activity but not weight or sleep**

5. Share Findings

-Continue scrolling down in this markdown to view findings

6. Recommendations and Next Steps

-Bellabeats can grow their market share by increasing their advertising campaigns during the holidays to take advantage of increased interest for their competitor during that time

-Messaging can be directed towards healthy users, daily activity tracking, or evening/noon workouts

-Opportunity to grow to new users by researching why there are less overweight users and sleep/weight tracking

-A large sample size and greater timeframe for data collected would provide stronger representation of the population

Data Exploration, Visualizations, and Findings

```
install.packages("tidyverse")
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\steph\AppData\Local\Temp\RtmpMJwQJt\downloaded_packages
```

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

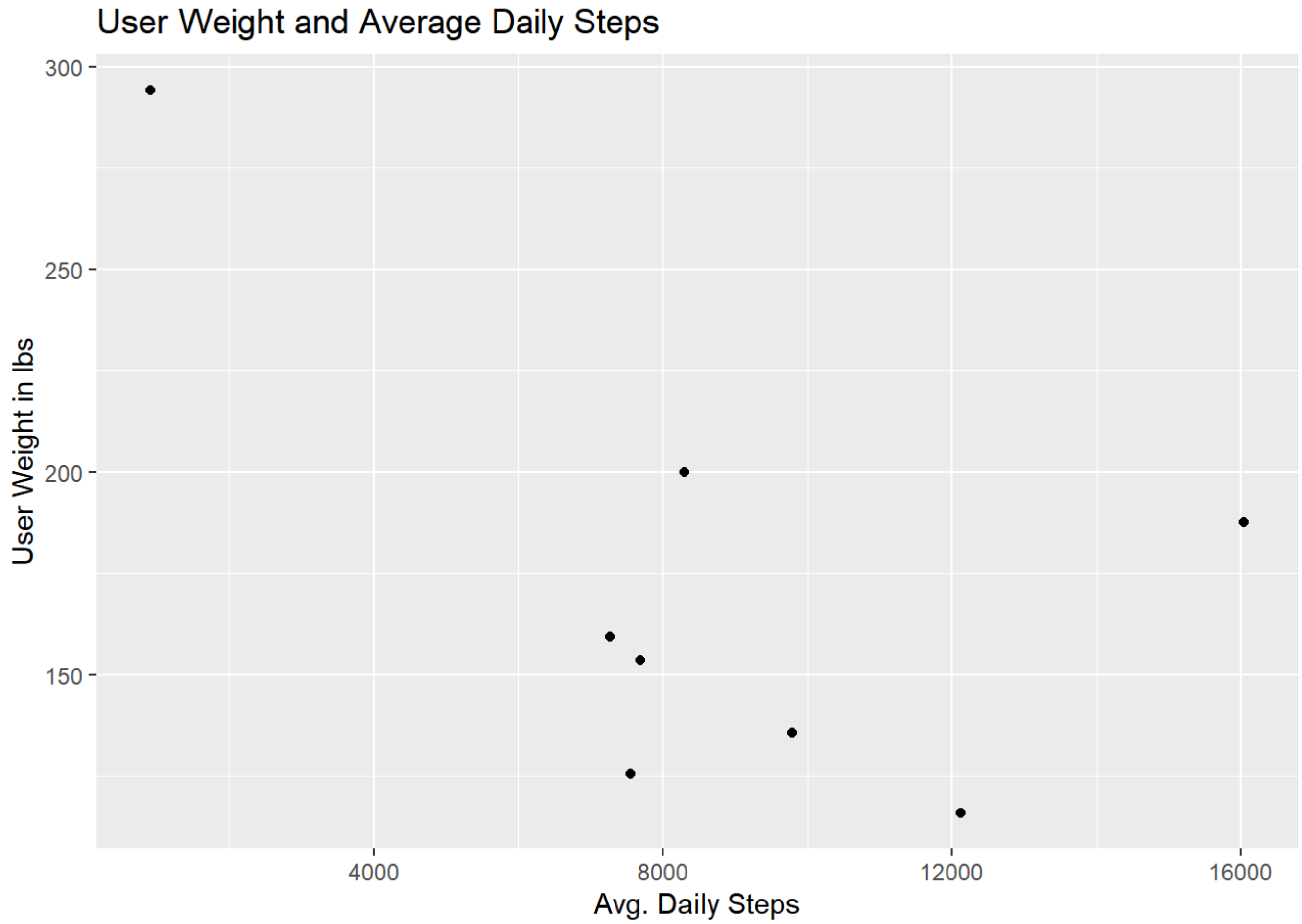
```
fit_bit_gtrend <- read_csv('FitBit google trend - cleaned.csv')

activity <- read_csv('y dailyActivity_merged - dailyActivity_merged.csv')

sleep <- read_csv('y sleepDay_merged - cleaned.csv')

weight <- read_csv('y weightLogInfo_merged - cleaned.csv')

hourly_steps <- read.csv('y hourlySteps_merged - cleaned.csv')

steps <- read.csv('hourly_steps.csv')
```

Only 1 record in the overweight BMI category

```
activity_weight <- inner_join(activity, weight, by = "Id") %>%
  select(Id, TotalSteps, SedentaryMinutes, BMI, WeightPounds) %>%
  group_by(Id) %>%
  summarize(mean_steps=mean(TotalSteps,na.rm = T),
            mean_sedentary_minutes=mean(SedentaryMinutes,na.rm = T),
            mean_bmi=mean(BMI,na.rm = T),
            mean_weightlbs=mean(WeightPounds,na.rm = T))

activity_weight
```

```
## # A tibble: 8 × 5
##           Id mean_steps mean_sedentary_minutes mean_bmi mean_weightlbs
##        <dbl>      <dbl>                  <dbl>    <dbl>          <dbl>
## 1 1503960366     12117.                   848.     22.6            116
## 2 1927972279       916.                  1317.     47.5           294.
## 3 2873212765      7556.                  1097.     21.6           126.
## 4 4319703577      7269.                   736.     27.4           160.
## 5 4558609924      7685.                  1094.     27.2           154.
## 6 5577150313      8304.                   754.     28              200
## 7 6962181067      9795.                   662.     24.0           136.
## 8 8877689391     16040.                  1113.     25.5           188.
```

-Not enough user data to determine if correlation exists between weight and steps



-Dataframe, daily_activity, was created to aggregate average steps by date

```
daily_activity <- activity %>%
  group_by(ActivityDate) %>%
  summarize(mean_steps=mean(TotalSteps))

daily_activity
```

```
## # A tibble: 31 × 2
##    ActivityDate mean_steps
##    <chr>             <dbl>
##  1 4/12/2016         8237.
##  2 4/13/2016         7199.
##  3 4/14/2016         7744.
##  4 4/15/2016         7534.
##  5 4/16/2016         8679.
##  6 4/17/2016         6409.
##  7 4/18/2016         7897.
##  8 4/19/2016         8049.
##  9 4/20/2016         8163.
## 10 4/21/2016         8244.
## # … with 21 more rows
```

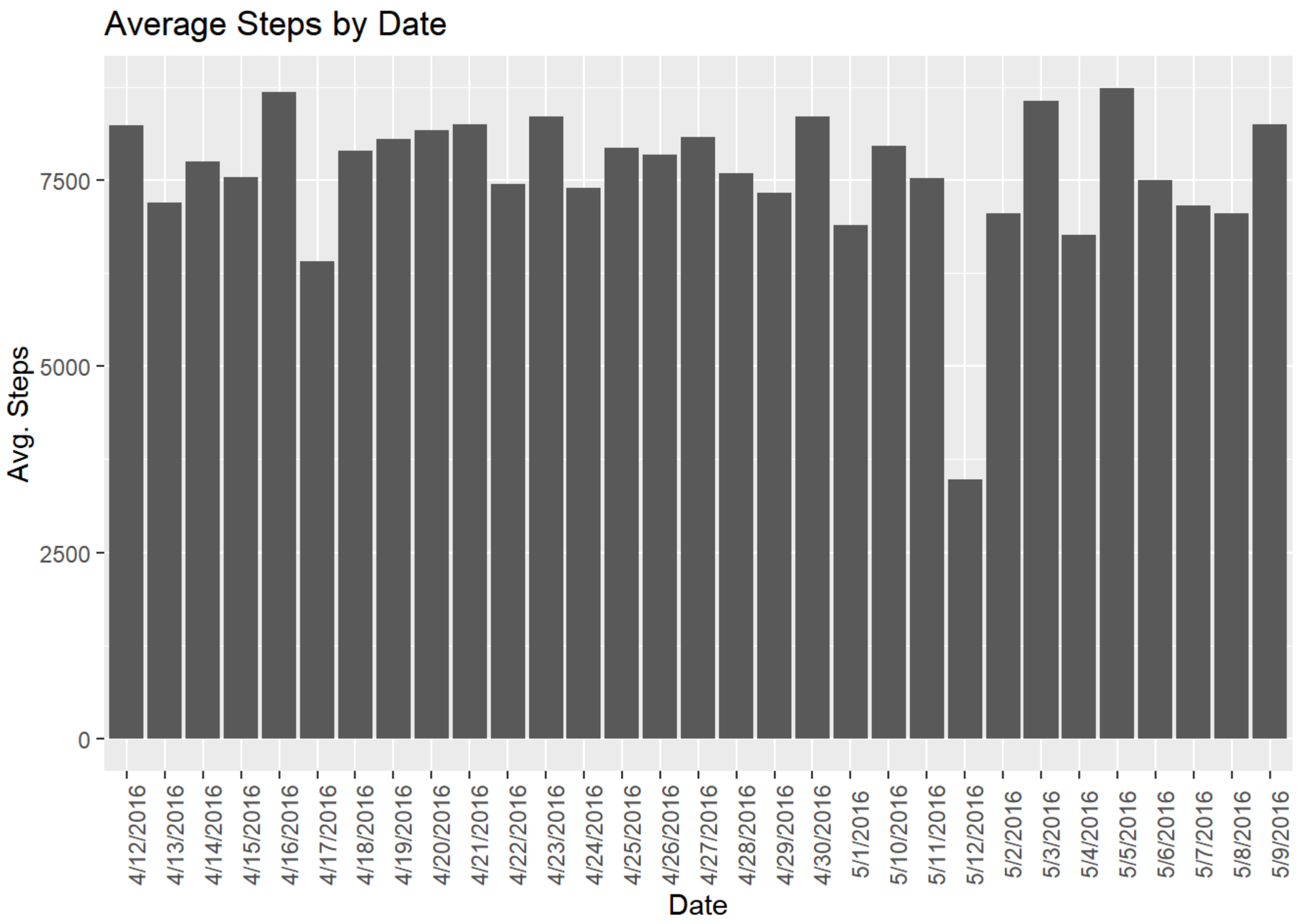
-Sorted daily_activity dataframe by steps to determine if there were more or less activity on certain days of the week

-Daily Activity data also visualized in a graph

-No correlations or patterns found, data is also insufficient

```
daily_activity %>%
  arrange(mean_steps)
```

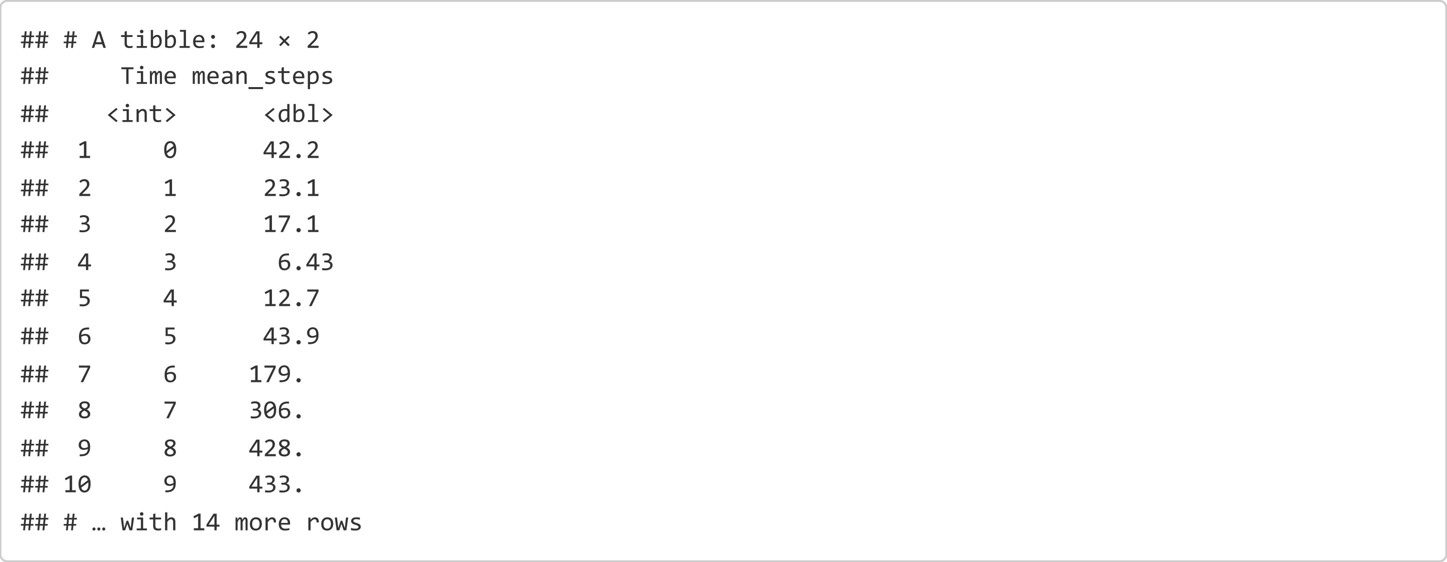
```
## # A tibble: 31 × 2
##    ActivityDate mean_steps
##    <chr>             <dbl>
##  1 5/12/2016         3482.
##  2 4/17/2016         6409.
##  3 5/4/2016          6764.
##  4 5/1/2016          6896.
##  5 5/8/2016          7049.
##  6 5/2/2016          7049.
##  7 5/7/2016          7151.
##  8 4/13/2016         7199.
##  9 4/29/2016         7322.
## 10 4/24/2016         7394.
## # … with 21 more rows
```



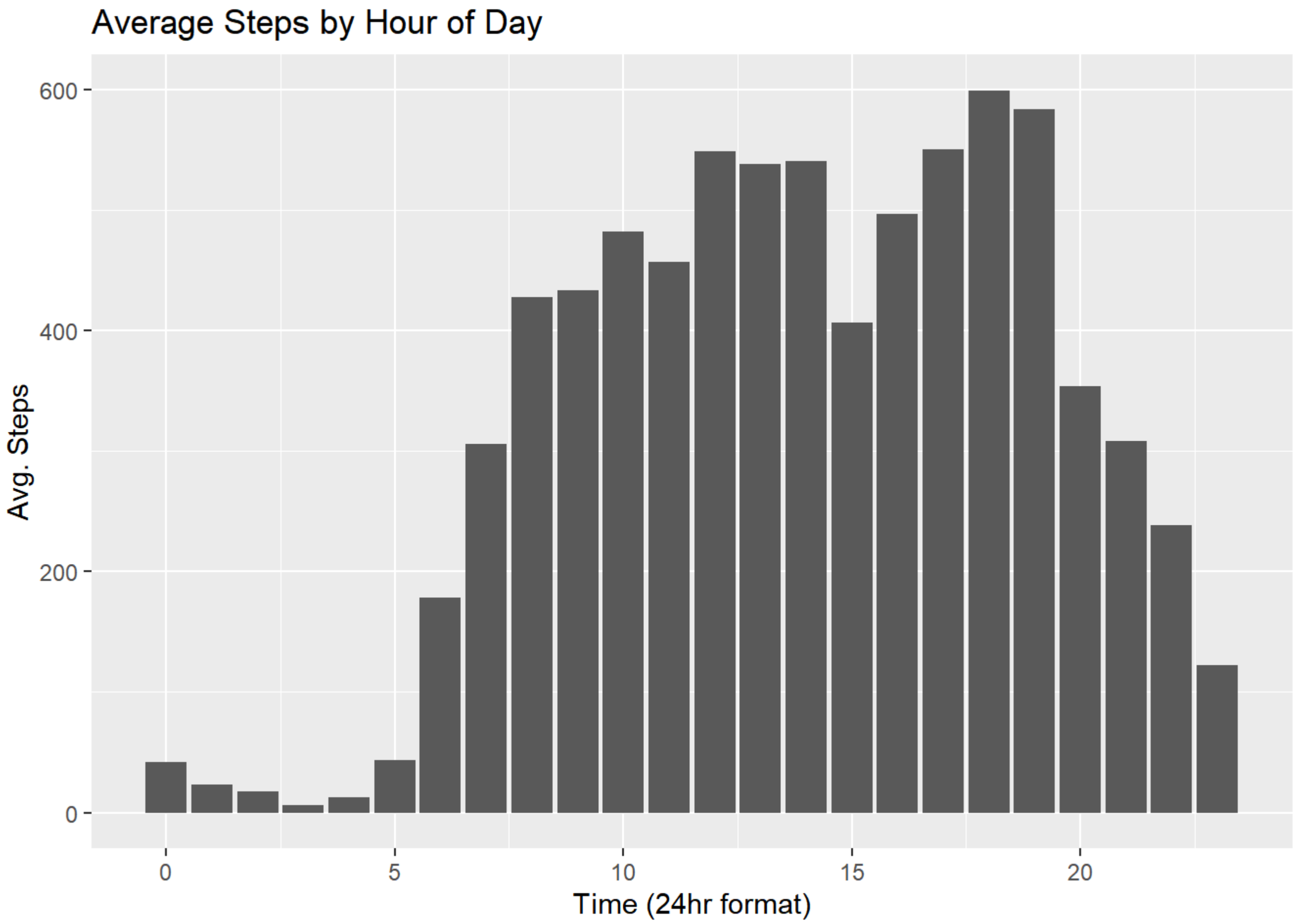
-Next hourly activity was table and graph were created to see if when users were more active

-Found that activity was highest in the evenings and around noon with 6pm being the most active hour

```
summary_steps <- steps %>%
  group_by(Time) %>%
  summarize(mean_steps=mean(StepTotal)) %>%
  arrange(Time)

summary_steps
```



```
## # A tibble: 24 × 2
##     Time mean_steps
##    <int>      <dbl>
##  1     0      42.2
##  2     1      23.1
##  3     2      17.1
##  4     3       6.43
##  5     4      12.7
##  6     5      43.9
##  7     6     179.
##  8     7     306.
##  9     8     428.
## 10     9     433.
## # … with 14 more rows
```



-Dataframes were created and joined to compare the total logs tracked

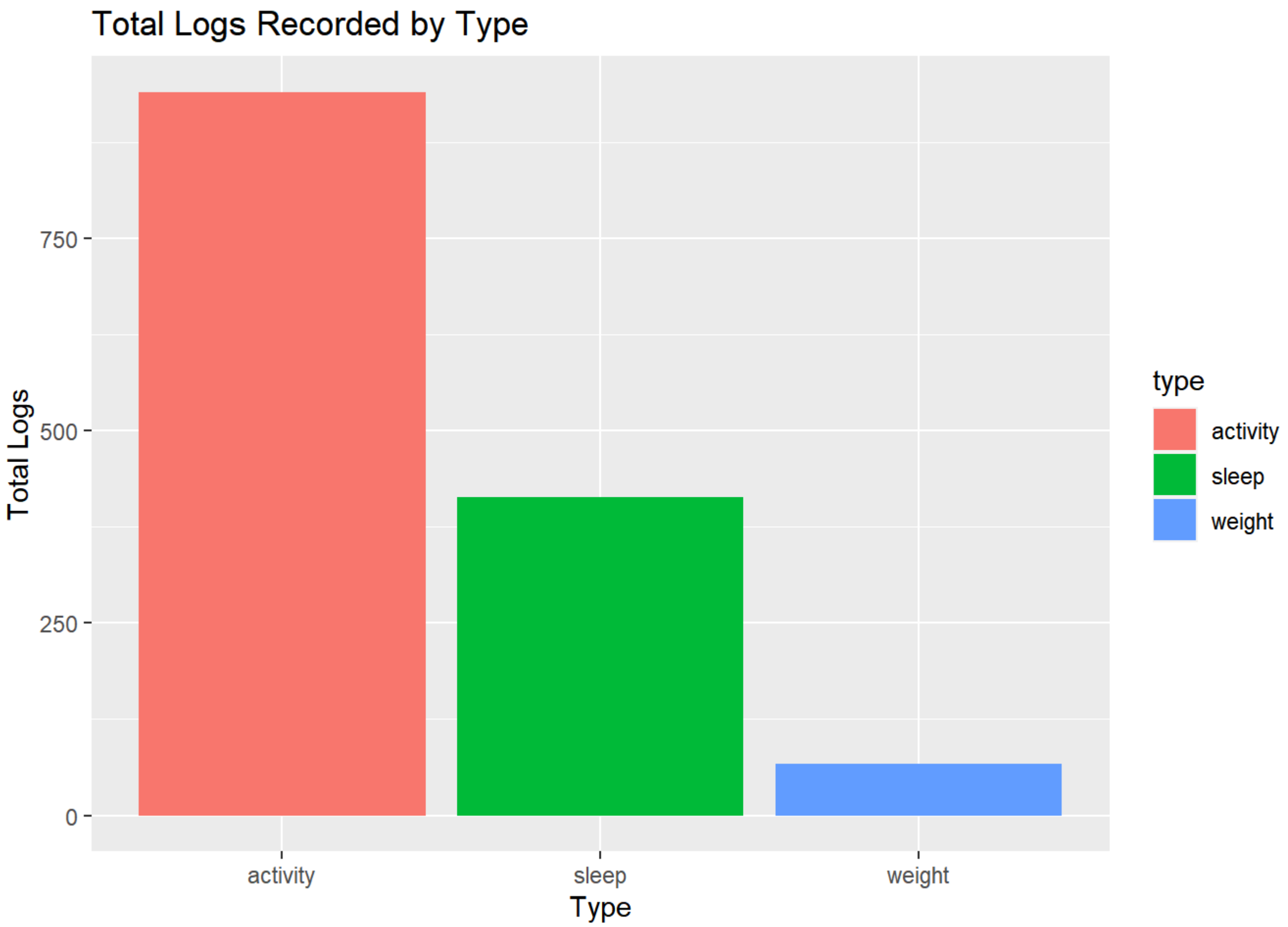
-Pivoted to a longer table to visualize in a bar chart

-Most users tracked their activity, some tracked their sleep, and few tracked their weight

-data tracked over a month, count of 31 indicated daily tracking

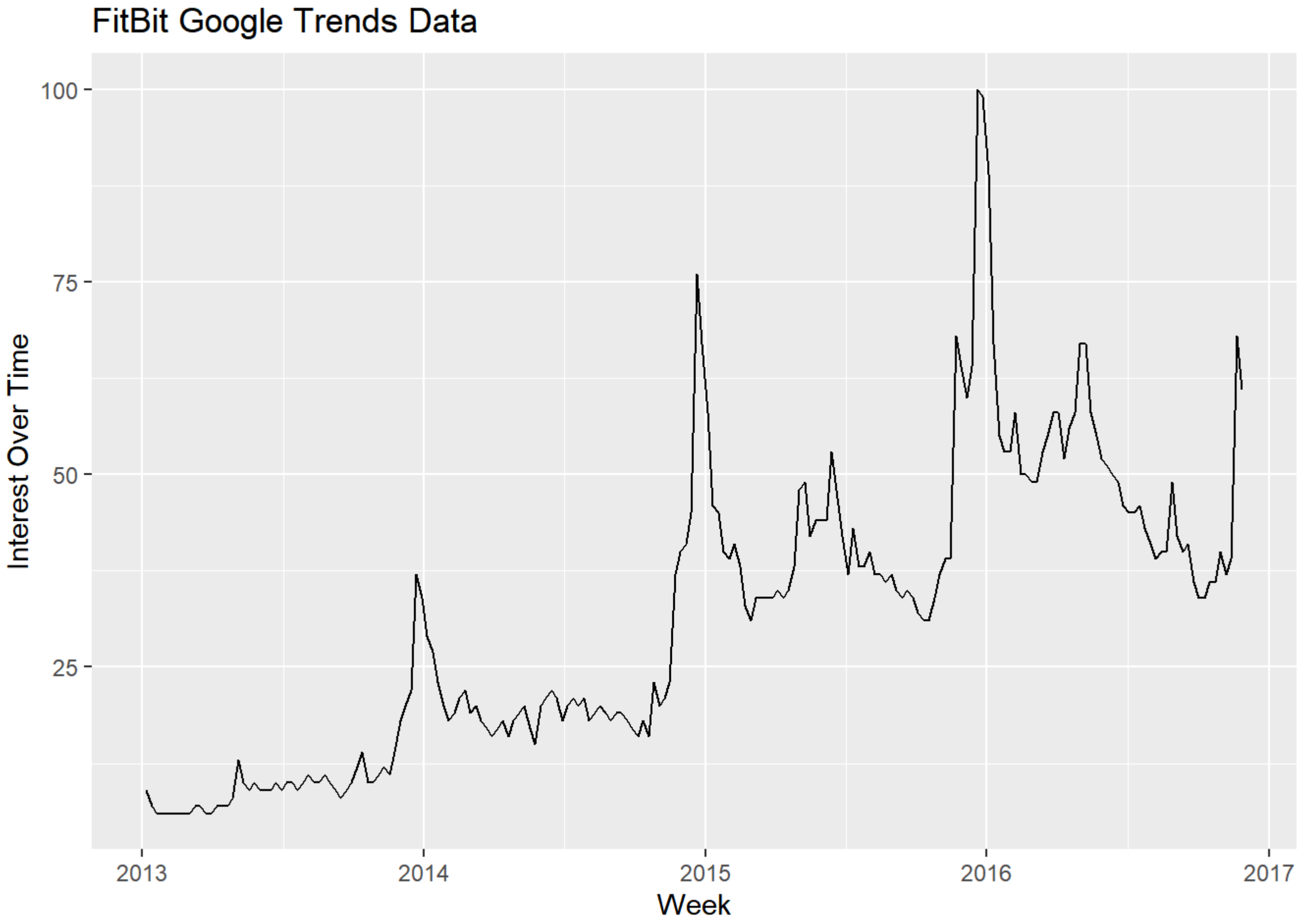
```
activity_logs<- activity %>% #most users track daily
  count(Id)

sleep_logs<- sleep %>% #some users track
  count(Id)


weight_logs<- weight %>% #few users track
  count(Id)


logs <- full_join(activity_logs, sleep_logs, by = "Id") %>%
  full_join(weight_logs, by = "Id") %>%
  rename(activity = n.x, sleep = n.y, weight=n)

logs
```

```
## # A tibble: 33 × 4
##            Id activity sleep weight
##         <dbl>    <int> <int>  <int>
##  1 1503960366       31    25      2
##  2 1624580081       31    NA     NA
##  3 1644430081       30     4     NA
##  4 1844505072       31     3     NA
##  5 1927972279       31     5      1
##  6 2022484408       31    NA     NA
##  7 2026352035       31    28     NA
##  8 2320127002       31     1     NA
##  9 2347167796       18    15     NA
## 10 2873212765       31    NA      2
## # … with 23 more rows
```



-Google Trend graph plotted to show when interest in FitBit is at its peak (December/holidays)

-This may be something Bellabeats marketing and advertising teams may want to take advantage of


FitBit Google Trends Data
100
75
50
25
Interest Over Time
2013
2014
2015
2016
2017
Week