Google Data Analytics Capstone: Bellabeat Case Study

Syeda Zaki

2/21/2022

Ask

Business Task

Bellabeat is a manufacturer of smart health monitoring technology for women. They have five different smart products, Bellabeat App, Leaf - wellness tracker, Time - wellness watch and Spring - smart water bottle. Bellabeat's co-founder and Chief Creative Officer, Urska Srsen wants us to analyze existing data from from other smart device usage that is publicly available and use the insights to to inform business decisions at Bellabeat.

We will be analyzing publicly available Fitbit Fitness Tracker data which contains fitness data from 33 Fitbit users that includes daily physical activity, sleep monitoring and heart rate data. This data is similar to what we would expect to collect from Bellabeat's Time users so it can be used to inform decisions for Bellabeat's Time smart watch.

We will present the insights from the analysis to the co-founder and Chief Executive Officer, Sando Mur, the other co-founder and key member of the Bellabeat executive team and the marketing analytics team that are collecting, analyzing and reporting data to achieve Bellabeats business goal and we are a part of that team. We will use our findings to make recommendations to inform Bellabeats marketing strategy for their Time fitness watch that tracks users activity, sleep and stress data.

Prepare

Load Dependencies

```
# install required package
#install.packages("tidyverse")
library(tidyverse)
#install.packages("here")
library(here)
#install.packages("skimr")
library(skimr)
#install.packages("janitor")
library(janitor)
#install.packages(lubridate)
library(lubridate)
#install.packages("dplyr")
library(dplyr)
#install.packages("gridExtra")
library(gridExtra)
```

```r
#install(ggplplot2)
library(ggplot2)
```

Data Sources

The data we are using is FitBit Fitness Tracker Data https://www.kaggle.com/arashnic/fitbit (CC0: Public Domain, dataset made available through Mobius). This is a Kaggle data set with personal fitness data from thirty three fitbit users.

We will download the data onto onto our computer. We will load the files into RStudio with the help of read.csv().

Read Files

```r
dailyActivity <- read.csv("dailyActivity_merged.csv")
dailyCalories <- read.csv("dailyCalories_merged.csv")
dailyIntensities <- read.csv("dailyIntensities_merged.csv")
dailySteps <- read.csv("dailySteps_merged.csv")
sleepDay <- read.csv("sleepDay_merged.csv")
```

Data Dimensions and Format

- dailyActivity_df dimensions: 940, 15, it is in long format
- dailyCalories_df dimensions: 940, 3, it is in long format
- dailyIntensities_df dimensions: 940, 10, it is in long format
- dailySteps_df dimensions: 940, 3, it is in long format
- sleepDay_df dimensions: 413, 5, it is in long format

Credibility of Data

- The data is from a reliable open source website Kaggle.
- It is original (was reported by 33 Fitbit users) and is comprehensive.
- It has been used by thousands of people.
- This is the data the stakeholders wants us to use.
- The data is similar to what we would have collected from Bellabeat customers to gain insight into smart device usage and hence may be used to gain insights to inform business decisions at Bellabeat.

Limitations of the Data

- The data is from a small sample of people.
- It was collected over a very small time period.
- It is very old (from 2016).
- The gender, height and weight of the participants is not known.

Explore Data

```r
str(dailyActivity)

## 'data.frame':    940 obs. of  15 variables:
##  $ Id                      : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09
```

```
...
##  $ ActivityDate            : chr  "4/12/2016" "4/13/2016" "4/14/2016"
"4/15/2016" ...
##  $ TotalSteps              : int  13162 10735 10460 9762 12669 9705 13019
15506 10544 9819 ...
##  $ TotalDistance           : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance         : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance      : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance     : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes       : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes     : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes    : int  328 217 181 209 221 164 233 264 205 211
...
##  $ SedentaryMinutes        : int  728 776 1218 726 773 539 1149 775 818
838 ...
##  $ Calories                : int  1985 1797 1776 1745 1863 1728 1921 2035
1786 1775 ...

str(dailyCalories)

## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...

str(dailyIntensities)

## 'data.frame':    940 obs. of  10 variables:
##  $ Id                      : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09
...
##  $ ActivityDay             : chr  "4/12/2016" "4/13/2016" "4/14/2016"
"4/15/2016" ...
##  $ SedentaryMinutes        : int  728 776 1218 726 773 539 1149 775 818
838 ...
##  $ LightlyActiveMinutes    : int  328 217 181 209 221 164 233 264 205 211
...
##  $ FairlyActiveMinutes     : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ VeryActiveMinutes       : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LightActiveDistance     : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ VeryActiveDistance      : num  1.88 1.57 2.44 2.14 2.71 ...

str(dailySteps)

## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
```

```
##  $ StepTotal  : int  13162 10735 10460 9762 12669 9705 13019 15506 10544
9819 ...
```

```
str(sleepDay)
```

```
## 'data.frame':    413 obs. of  5 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00
AM" "4/15/2016 12:00:00 AM" "4/16/2016 12:00:00 AM" ...
##  $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

We see that daily activity has all the data from the other daily files. We will use only the dailyActivity and sleepDay datasets for downstream analysis.

Process

Check Data Errors and Inconsistencies

- Convert datasets into tibbles for ease of use.
- Check data for inconsistencies and missing values using skim_without_charts() from skimr package.
- Check number of users for both datasets.

```
# Convert to tibbles
daily_activity <- as_tibble(dailyActivity)
daily_sleep <- as_tibble(sleepDay)

# Check to see if there are any missing values and in-consistencies in the
data
skim_without_charts(daily_activity)
skim_without_charts(daily_sleep)

# Check the number of users in daily activity and daily sleep
activity_users <- n_distinct(daily_activity$Id)
sleep_users <- n_distinct(daily_sleep$Id)
```

- The id column is common between the two files. It has data type numeric.
- The daily_activity and daily_sleep datasets have date columns with different names.
- The activity_day and sleep_day columns contain date time data as character.
- The names of columns are inconsistent between the datasets.
- Daily activity dataset has 33 unique id’s and sleep dataset has 24 unique id’s .
- There are no missing values in either files.

Format Data

We will format and rename columns, change data types of variables to prepare the data for analysis.

```
# Make names consistent in both tibbles
daily_activity <- clean_names(daily_activity)
daily_sleep <- clean_names(daily_sleep)

# Rename column daily_activity from daily_activity to date
daily_activity <- daily_activity %>% rename(date = activity_date)

# Rename column sleep_day to date
daily_sleep <- daily_sleep %>% rename(date = sleep_day)

# Convert id to character
daily_activity$id <- as.character(daily_activity$id)
daily_sleep$id <- as.character(daily_sleep$id)

# Convert daily_activity date from character to date
daily_activity$date <- mdy(daily_activity$date)

# Convert daily_sleep date from character to date
# Seperate the different parts of the date tine string into date, time and NA
daily_sleep <- daily_sleep %>% separate(date,c("date", "time", "NA"), sep = "
")
# Convert the date portion to date format
daily_sleep$date <- mdy(daily_sleep$date)
# Select columns to keep
daily_sleep <- daily_sleep %>% select(id, date,total_sleep_records,
total_minutes_asleep, total_time_in_bed)

# Get summary
str(daily_activity)

## tibble [940 x 15] (S3: tbl_df/tbl/data.frame)
##  $ id                        : chr [1:940] "1503960366" "1503960366"
"1503960366" "1503960366" ...
##  $ date                      : Date[1:940], format: "2016-04-12" "2016-04-
13" ...
##  $ total_steps               : int [1:940] 13162 10735 10460 9762 12669
9705 13019 15506 10544 9819 ...
##  $ total_distance            : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ tracker_distance          : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ logged_activities_distance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_distance      : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  $ moderately_active_distance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ light_active_distance     : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ sedentary_active_distance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_minutes       : int [1:940] 25 21 30 29 36 38 42 50 28 19
...
##  $ fairly_active_minutes     : int [1:940] 13 19 11 34 10 20 16 31 12 8
...
##  $ lightly_active_minutes    : int [1:940] 328 217 181 209 221 164 233 264
```

```
205 211 ...
##  $ sedentary_minutes         : int [1:940] 728 776 1218 726 773 539 1149
775 818 838 ...
##  $ calories                  : int [1:940] 1985 1797 1776 1745 1863 1728
1921 2035 1786 1775 ...
```

```
str(daily_sleep)
```

```
## tibble [413 x 5] (S3: tbl_df/tbl/data.frame)
##  $ id                  : chr [1:413] "1503960366" "1503960366"
"1503960366" "1503960366" ...
##  $ date                : Date[1:413], format: "2016-04-12" "2016-04-13"
...
##  $ total_sleep_records : int [1:413] 1 2 1 2 1 1 1 1 1 1 ...
##  $ total_minutes_asleep: int [1:413] 327 384 412 340 700 304 360 325 361
430 ...
##  $ total_time_in_bed   : int [1:413] 346 407 442 367 712 320 377 364 384
449 ...
```

Data Cleaning and Manipulation

- Cleaned names using clean_names() from janitor package to make them consistent.
- Renamed columns containing date information to date in both datasets using the rename() of dplyr package.
- Converted id column to type character.
- Converted the date column in both datasets to date type using mdy() from lubridate package.

Get Basic Summaries

```
# Daily Activity summary
daily_activity %>% select(total_distance, total_steps, sedentary_minutes) %>%
summary()
```

```
##  total_distance    total_steps    sedentary_minutes
##  Min.   : 0.000   Min.   :    0   Min.   :   0.0
##  1st Qu.: 2.620   1st Qu.: 3790   1st Qu.: 729.8
##  Median : 5.245   Median : 7406   Median :1057.5
##  Mean   : 5.490   Mean   : 7638   Mean   : 991.2
##  3rd Qu.: 7.713   3rd Qu.:10727   3rd Qu.:1229.5
##  Max.   :28.030   Max.   :36019   Max.   :1440.0
```

```
# Sleep data summary
daily_sleep %>% summary()
```

```
##       id                 date            total_sleep_records
##  Length:413         Min.   :2016-04-12   Min.   :1.000
##  Class :character   1st Qu.:2016-04-19   1st Qu.:1.000
##  Mode  :character   Median :2016-04-27   Median :1.000
##                     Mean   :2016-04-26   Mean   :1.119
##                     3rd Qu.:2016-05-04   3rd Qu.:1.000
##                     Max.   :2016-05-12   Max.   :3.000
```

```
##  total_minutes_asleep total_time_in_bed
##  Min.   : 58.0        Min.   : 61.0    
##  1st Qu.:361.0        1st Qu.:403.0    
##  Median :433.0        Median :463.0    
##  Mean   :419.5        Mean   :458.6    
##  3rd Qu.:490.0        3rd Qu.:526.0    
##  Max.   :796.0        Max.   :961.0
```

Now that we have processed the data by formatting it and renaming columns, changing data types of variables and making sure that there are no missing values or inconsistencies the data is clean and ready for analysis.

Analyze

Look for Correlation among Variables

Daily Activity Dataset Correlations

```
# Compute the correlation between the different variables
cor1 <- cor(daily_activity$total_steps, daily_activity$total_distance)
cor2 <- cor(daily_activity$total_steps, daily_activity$sedentary_minutes)
cor3 <- cor(daily_activity$total_steps, daily_activity$calories)
cor4 <- cor(daily_activity$sedentary_minutes, daily_activity$calories)
```

- Total steps vs total distance we can see that the these variables have a linear relationship and they are highly positively correlated with a correlation coefficient of 0.9853688.
- Total steps and sedentary minutes have somewhat a negative correlation with a correlation coefficient of -0.3274835.
- Total steps and calories burned have a correlation coefficient 0.5915681 but they are not highly correlated.
- Sedentary minutes and calories also have a slight negative correlation with a correlation coefficient of -0.106973.

Daily Sleep Dataset Correaltions

```
cor9 <- cor(daily_sleep$total_time_in_bed, daily_sleep$total_minutes_asleep)
```

We see that the total time in bed and total minutes asleep are highly positively correlated with a correlation coefficient of 0.9304575.

Merge daily activity and daily sleep datasets together

We will merge the daily activity and sleep datasets to see if there are any correlations between the variables in the two datasets

```
# Do inner join  to keep only matching rows from both datasets
merged_data <- merge(x=daily_activity, y=daily_sleep, by=c("id","date"))
```

Merged Data Correlations

- Correlation between total distance and total minutes asleep is -0.1721427
- Correlation between total steps and total minutes asleep is -0.1868665
- Correlation between sedentary minutes and total time asleep is -0.599394.
- Correlation between calories and total time asleep is -0.0285257.

From the correlation coefficients we can see that there seems to be no real correlation between total steps and total minutes asleep, total distance and total minutes asleep and calories and total minutes asleep. However, we can see there is a slight negative correlation between sedentary minutes and total minutes asleep which is of interest.

Share

Data Visualizations

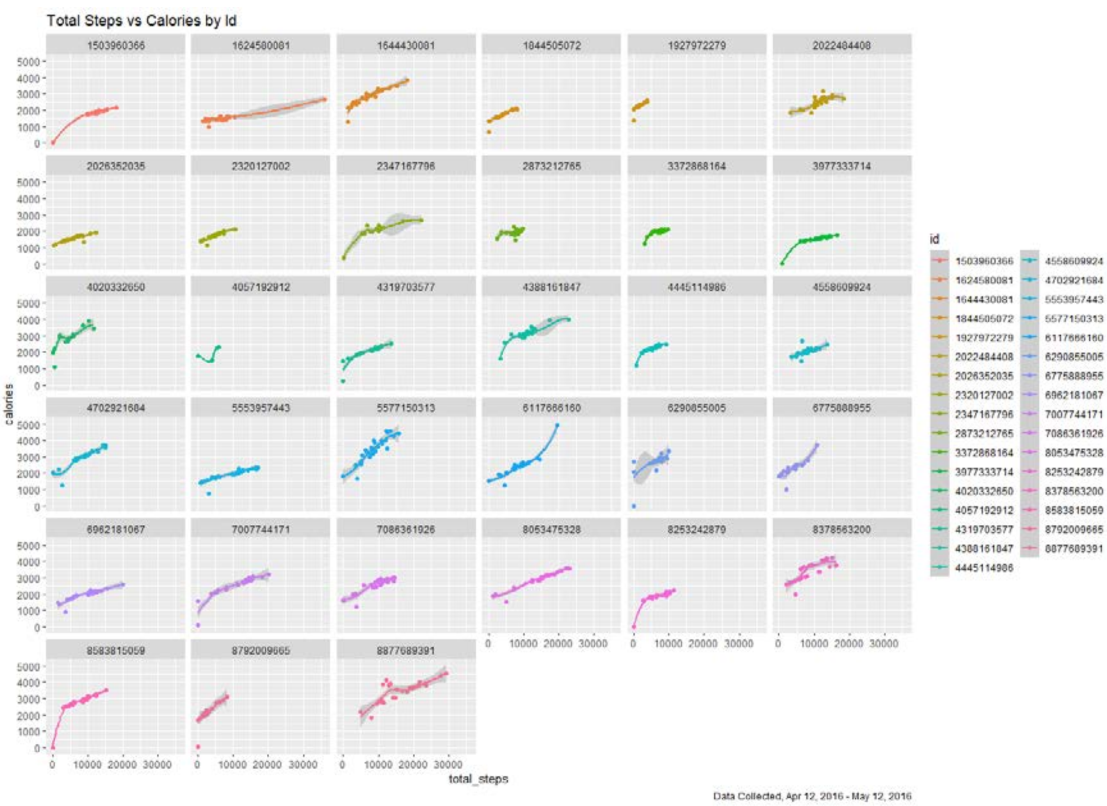
We will create plots to visualize the relationships between the different variables in the datasets to discover trends using the ggplot2 geom_smooth(). This will produce a smoothing line along the data points which will help us discover trends. We will go with the default smooth which is loess smooth and the bands on the plot show the confidence intervals for our predictions.

Plotting Daily Activity Data

```
# Plot total_steps vs total_distance by id
plot1 <-  daily_activity %>% group_by(id) %>% ggplot() + geom_point(mapping =
aes(x=total_steps,y=total_distance,color=id)) +
geom_smooth(mapping=aes(x=total_steps,y=total_distance)) + labs(title =
"Daily Activity: Total Steps vs Total Distance", caption = "Data) Collected,
Apr 12, 2016 - May 12, 2016") + theme(legend.position = "none")

# Plot total_steps vs sedentary_minutes
plot2 <- daily_activity %>% group_by(id) %>% ggplot() + geom_point(mapping =
aes(x=sedentary_minutes,y=total_steps,color=id)) +
geom_smooth(mapping=aes(x=sedentary_minutes,y=total_steps)) + labs(title =
"Daily Activity:Total Steps vs Sedentary Mimnutes", caption = "Data
Collected, Apr 12, 2016 - May 12, 2016") + theme(legend.position = "none")

# Plot total_steps vs calories burned
plot3 <- daily_activity %>% group_by(id) %>% ggplot() + geom_point(mapping =
aes(x=total_steps,y=calories,color=id)) +
geom_smooth(mapping=aes(x=total_steps,y=calories)) + labs(title = "Daily
Activity: Total Steps vs Calories", caption = "Daily Activity:Data Collected,
Apr 12, 2016 - May 12, 2016") + theme(legend.position = "none")

# Plot sedentary minutes vs calories
plot4 <- daily_activity %>% group_by(id) %>% ggplot() + geom_point(mapping =
aes(x=sedentary_minutes,y=calories,color=id)) +
geom_smooth(mapping=aes(x=sedentary_minutes,y=calories)) + labs(title =
```

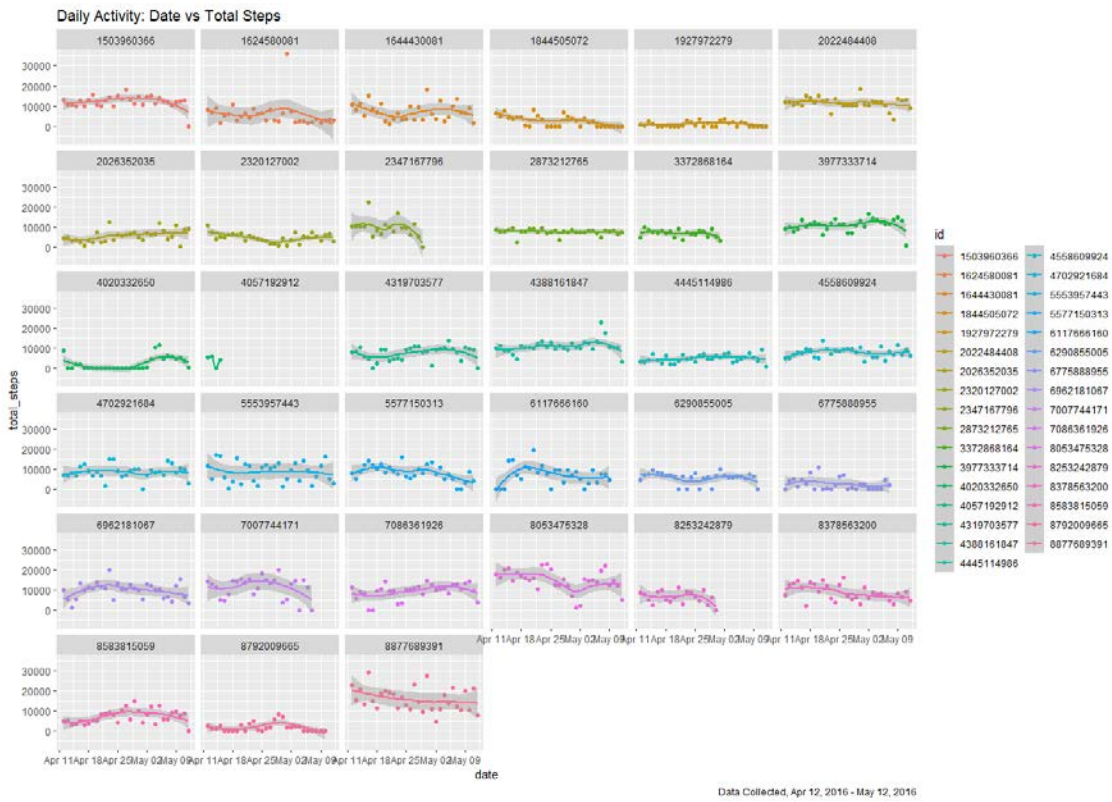
```
"Daily Activity:Sedentary Minutes vs Calories", caption = "Data Collected, Apr 12, 2016 - May 12, 2016")

grid.arrange(plot1,plot2,plot3,plot4)
```

- We can see that total steps and total distance have a highly positive correlation as total steps increase so does the distance which is expected.
- We can see that total steps and sedentary minutes have a very slightly negative correlation, so as total steps increase sedentary minutes decrease which is expected.
- We can see that total steps and calories burned have a somewhat positive correlation but not as much as we would expect. We see that there are some outliers which may be the reason for this and we will investigate this further.
- Sedentary minutes and calories burn have a slightly negative correlation, so more sedentary minutes result in less calories burned which is expected.



We see that the data from user Id 1624580081 is inconsistent and is the reason for the outlier. Since the user has reported very little data we can chose to remove this outlier without affecting the overall result.

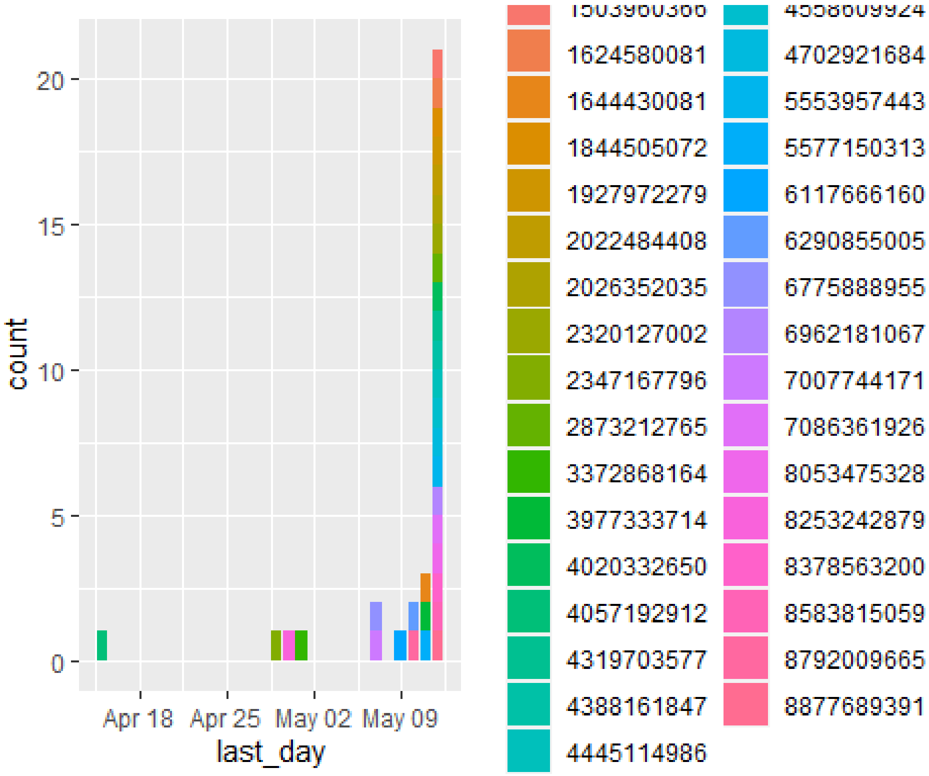
Next, we can look at daily total steps for each user over a month to determine trends in device usage at the user level and see if we can see any trends.



- From the plot we can see that some users stopped using the device before the end of the monitoring period.
- Some of those were consistent in monitoring their activity initially and then suddenly stopped using the device.

We will further investigate to see how many users stopped using the device before the end of the reporting period.

```
daily_activity %>% group_by(id) %>% summarize(last_day = max(date)) %>%
ggplot() + geom_bar(mapping=aes(last_day, fill=id))
```




```
# Number of users who submitted data till May 12, 2016
daily_activity %>% group_by(id) %>% summarize(last_day = max(date)) %>%
filter(last_day == "2016-05-12")

## # A tibble: 21 x 2
##    id         last_day
##    <chr>      <date>
##  1 1503960366 2016-05-12
##  2 1624580081 2016-05-12
##  3 1844505072 2016-05-12
##  4 1927972279 2016-05-12
##  5 2022484408 2016-05-12
##  6 2026352035 2016-05-12
##  7 2320127002 2016-05-12
##  8 2873212765 2016-05-12
##  9 4020332650 2016-05-12
## 10 4319703577 2016-05-12
## # ... with 11 more rows

# Users who stopped reporting data before the end of April 2016
daily_sleep %>% group_by(id) %>% summarize(last_day = max(date)) %>%
filter(last_day < "2016-05-01")

## # A tibble: 4 x 2
##   id         last_day
##   <chr>      <date>
## 1 1927972279 2016-04-28
```

```
## 2 2320127002 2016-04-23
## 3 2347167796 2016-04-29
## 4 6775888955 2016-04-15
```

We see that 21 users used the device to monitor daily activity till 2016-05-12

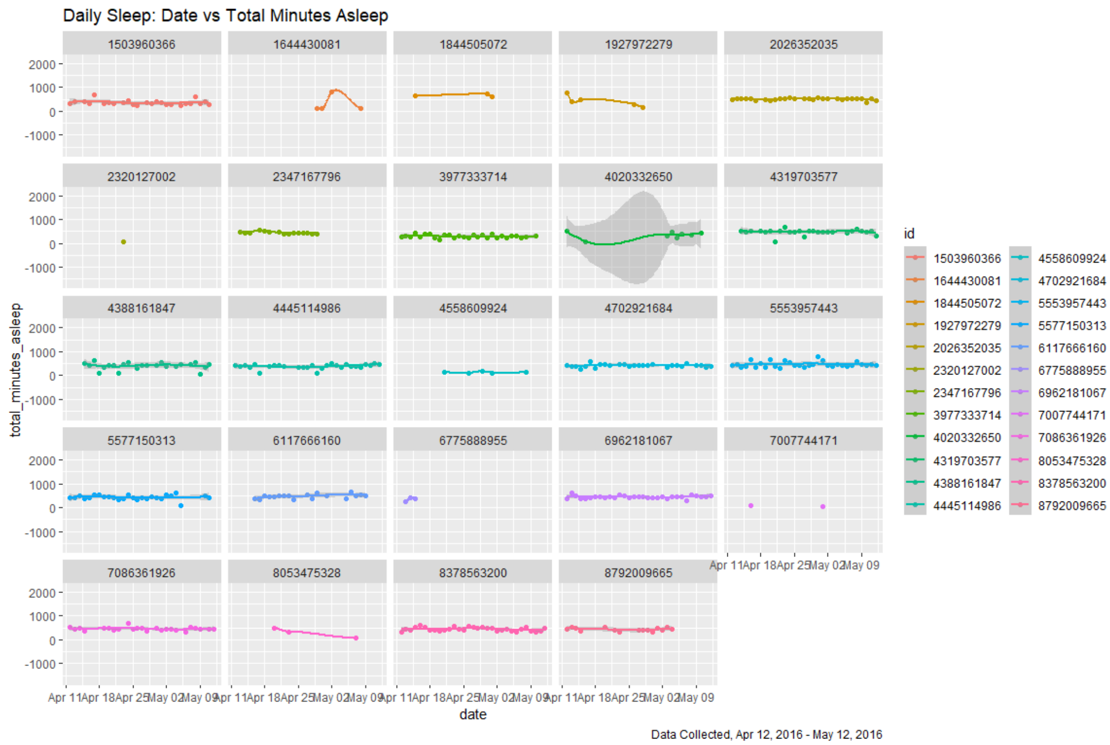
We see that 4 users stopped using the device by 2016-05-01

This information will be important in discovering the trends in device usage over a period of a month.

Plotting Daily Sleep Data

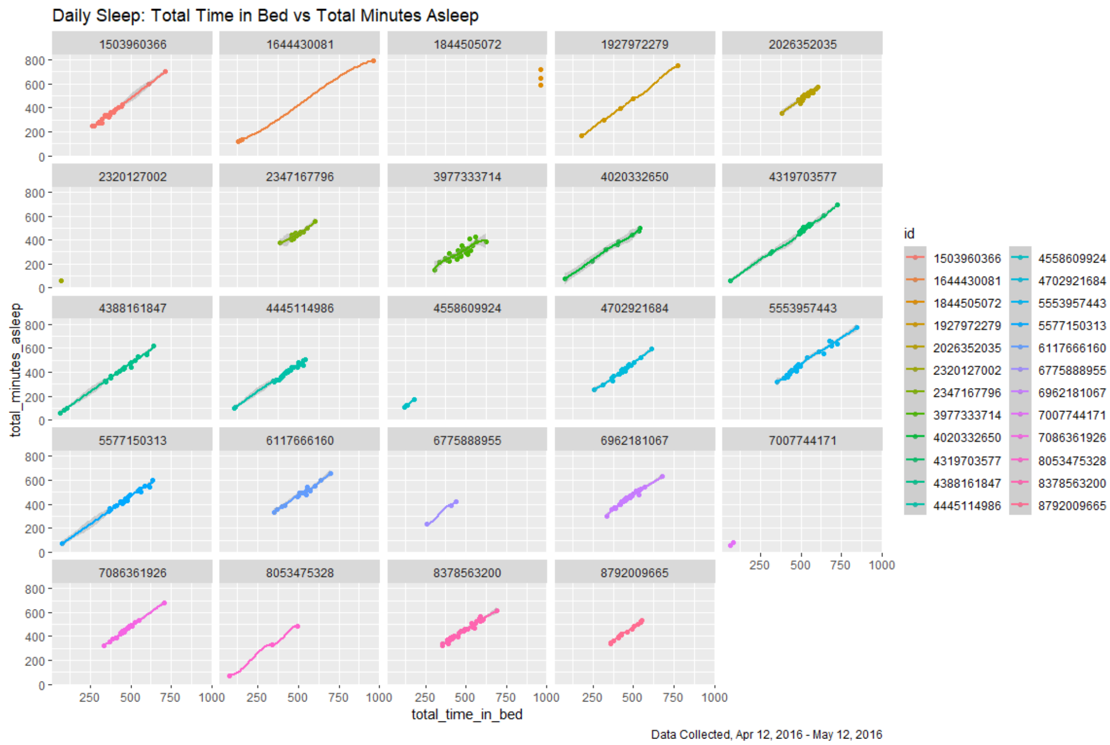
```r
# Plot total time in bed vs total_minutes asleep and color by id
ggplot(data=daily_sleep) + geom_point(mapping = aes(x=total_time_in_bed,
y=total_minutes_asleep, color=id)) + labs(title = "Total Time in Bed vs Total
Minutes Asleep")
```

- Total minutes in bed has an almost linear relationship with total minutes asleep, but there are some outliers. This could be for several reasons. We need to further investigate this to see if there is a trend.
- We can plot the total minutes asleep vs the total minutes in bed by id to determine the trend for each user.Then we can see how this trend develops over the month for each users. We can see for how long and how many times each user used the device. We can facet the plot by id to see the relationships for each user more clearly.
- We will look at total minutes asleep plotted against date to see how the trend develops over the month.



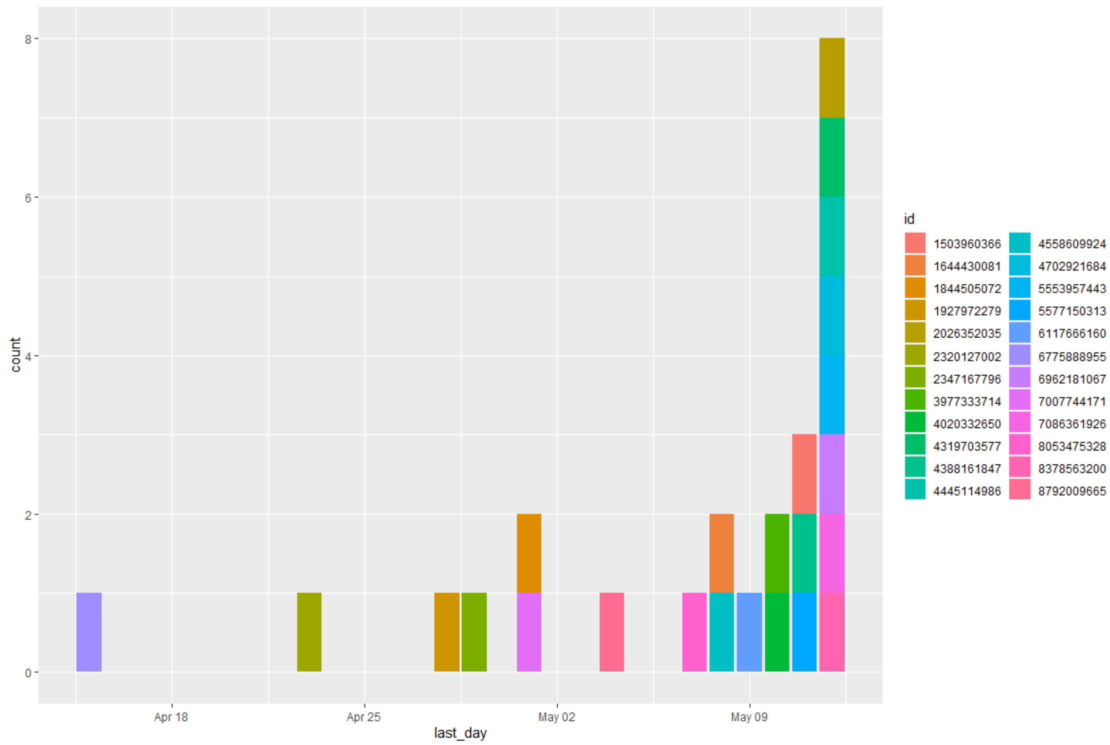
- Out of the 23 participants who turned in sleep data 7 had very few days of sleep data reported.
- Three had only 2 or three data points in total.
- Next we can also look at the relationship between total time in bed vs total minutes asleep.

```
# Plot total time in bed vs total minutes asleep faceted by id
ggplot(data=daily_sleep, aes(x=total_time_in_bed,
y=total_minutes_asleep,color=id)) +
      geom_point() + geom_smooth() +
      facet_wrap(~id) + labs(title = "Daily Sleep: Total Time in Bed vs Total
Minutes Asleep", caption = "Data Collected, Apr 12, 2016 - May 12, 2016")
```



- Out of the 23 participants who turned in sleep data 8 had very few days of sleep data reported.
- This shows that when users did wear the device it did not interfere with their ability to fall asleep.
- Next let us see what was the last day when each user reported sleep data.

```
daily_sleep %>% group_by(id) %>% summarize(last_day = max(date)) %>% ggplot()
+ geom_bar(mapping=aes(last_day, fill=id))
```




```
# Users who stopped reporting data before the end of April 2016
daily_sleep %>% group_by(id) %>% summarize(last_day = max(date)) %>%
filter(last_day < "2016-05-01")
```

```
## # A tibble: 4 x 2
##   id         last_day
##   <chr>      <date>
## 1 1927972279 2016-04-28
## 2 2320127002 2016-04-23
## 3 2347167796 2016-04-29
## 4 6775888955 2016-04-15
```

```
# Number of users who submitted data till May 12, 2016
daily_sleep %>% group_by(id) %>% summarize(last_day = max(date)) %>%
filter(last_day > "2016-05-01")
```

```
## # A tibble: 18 x 2
##    id         last_day
##    <chr>      <date>
##  1 1503960366 2016-05-11
##  2 1644430081 2016-05-08
##  3 2026352035 2016-05-12
##  4 3977333714 2016-05-10
##  5 4020332650 2016-05-10
##  6 4319703577 2016-05-12
##  7 4388161847 2016-05-11
##  8 4445114986 2016-05-12
##  9 4558609924 2016-05-08
```

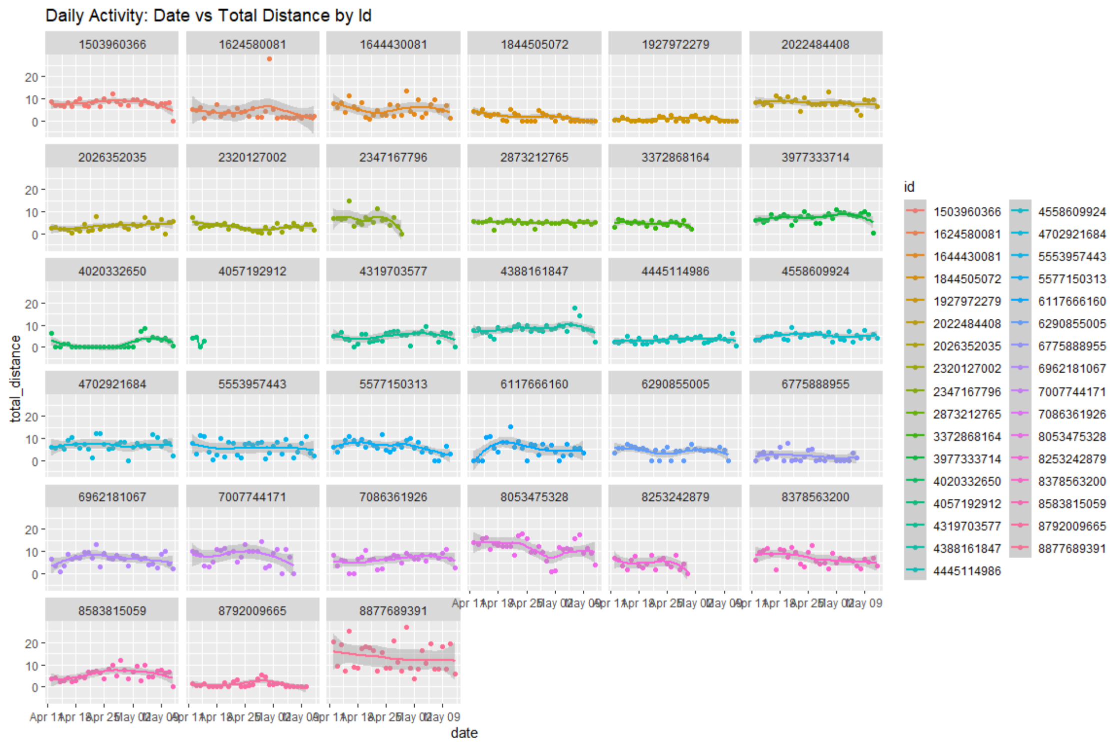
```
## 10 4702921684 2016-05-12
## 11 5553957443 2016-05-12
## 12 5577150313 2016-05-11
## 13 6117666160 2016-05-09
## 14 6962181067 2016-05-12
## 15 7086361926 2016-05-12
## 16 8053475328 2016-05-07
## 17 8378563200 2016-05-12
## 18 8792009665 2016-05-04
```

```
daily_sleep %>% group_by(id) %>% summarize(first_day = min(date))
```

```
## # A tibble: 24 x 2
##    id         first_day
##    <chr>      <date>
##  1 1503960366 2016-04-12
##  2 1644430081 2016-04-29
##  3 1844505072 2016-04-15
##  4 1927972279 2016-04-12
##  5 2026352035 2016-04-12
##  6 2320127002 2016-04-23
##  7 2347167796 2016-04-13
##  8 3977333714 2016-04-12
##  9 4020332650 2016-04-12
## 10 4319703577 2016-04-14
## # ... with 14 more rows
```

- 4 users stopped using the device before 2016-05-01.
- 16 of the users users stopped using the device by 2016-05-11.
- Most users who reported sleep data recorded it till 2016-05-11.
- Next, we will check to see if the same users had more activity data points which will show that they used the device just but not when sleeping.
- Did they use the device during the day but not when sleeping?
- We can check if the users who did not turn in sleep data had turned in activity data.

```
daily_activity %>% group_by(date) %>%
ggplot() + geom_point(mapping = aes(x=date, y=total_distance,color=id)) +
geom_smooth(mapping = aes(x=date, y=total_distance,color=id)) +
facet_wrap(~id) + labs(title = "Daily Activity: Date vs Total Distance by
Id")
```



So we can see that the users who had very little sleep data or missing sleep data had a good amount of activity data.

It is possible that they did not find the device comfortable to wear when sleeping or perhaps they were not interested in collecting sleep data. We can survey the customers to find out.

It would be a good idea to survey customers to ask them what they think about using the device when sleeping vs during the day.

This may be one of opportunities for growth against the competitors that Bellabeat could target.

Explore relationship between activity and sleep data

From the plots and the correlation coefficients we can see that there seems to be no real correlation between total steps and total minutes asleep, total distance and total minutes asleep and calories and total minutes asleep.

We can see there is a slight negative correlation between sedentary minutes and total minutes asleep which is of interest. This data indicates that if users are more active during the day they may get more sleep at night.

Act

Inference and Conclusion

We found the following trends in smart device usage

- All 33 users used the device to monitor daily activity but only 24 used the device to monitor sleep.

- Many users used the device consistently to monitor daily activity for the whole period over which the data was recorded.

- Among the users who did monitor sleep there was a highly positive correlation between total time in bed and total minutes asleep. This shows that the wearing the device to bed did not interfere with the users ability to sleep. However, some users used the device only briefly to monitor sleep and then stopped.

- Some users used the device consistently initially to record activity but not so consistently after the initial 15 day period.

- We found a slightly negative correlation between sedentary minutes and total time asleep.

Recommendation for translating these findings to Bellabeat's customers

Overall, we found that the device use decreases after the initial 15 day period and that some users did not use the device to monitor sleep at all. Of the users that monitor sleep, many use the device only for a short time and then stop.

Since Fitbit fitness tracker is similar to Bellabeat's Time device, which also monitors users activity, sleep and stress data, we suggest the following:

- The stakeholders need to keep the users engaged by offering them some incentive to use the device after the initial 15 days when many user's interest seem to vane. For example, Bellabeat could have competitions or challenges where users compete with each other to score points on a leader board to give them incentive to keep using the device and keep them motivated.

- The stakeholders should survey the users to find out what they liked or disliked about the device, and the reasons for not monitoring sleep data. They should then try to use the findings to address the problems and encourage the users to monitor sleep data. For example, is the device comfortable to wear to bed, is it too large or bulky, and so on?

- The stakeholders could inform users that they found that a decrease in sedentary time during the day (more activity during the day) may improve sleep. So they should use the device to monitor daily activity and aim to be more active during the day to get a good night's sleep, which they can monitor with the device.

- This data is not recent so we need to keep in mind that some of the issues highlighted above might have changed as the smart watch companies have rolled out new products with latest upgrades in technology to address the feedback from users.
- We suggest the stakeholders to collect the latest data from the target customers – women – gathered over a longer period of time to make more informed decisions for users in 2022 and beyond.