

# GoogleCapstoneMarkdown

Dillon O'Rourke

2022-03-15

---

## Disclaimer

This case study is a project for my online portfolio as part of the Google Data Analytics professional certificate and is entirely fictional.

---

## Introduction

I am a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company.

### *The purpose of this project;*

Is to use data to gain insights into customer usage of the non-company products and identify new growth opportunities for the company.

An analysis of smart device usage could identify trends which could be investigated further to explore if they can be applied to Bellabeat customers. Applying these trends and behaviors to the company's marketing strategy could improve future sales revenue.

### *The key business tasks;*

- Analyse smart device data and gain insight into how consumers are using their devices.
  - 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?
- Create a presentation of my analysis and high-level recommendations for how these trends can inform Bellabeat marketing strategy to the executive team.

## *Assumptions or theories*

### *Products*

- **Bellabeat app:** The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
- **Leaf:** Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.
- **Time:** This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.
- **Spring:** This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.
- **Bellabeat membership:** Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

### *The Data*

1. FitBit Fitness Tracker Data.

Personal fitness tracker data collected from 30 consenting fitbit users in 2016. It includes minute level output for exercise, heart rate and daily activity.

- **Source:** External 3rd party data.
- **Type:** Structured continuous & discrete mostly nominal quantitative data.
- **Structure:** Time-series data for each individual user in several tables. Data ranges from daily to minute timestamps. Heart rate, calories, steps, sleep and weight information are provided by the data.

The daily data will be analysed first as it will provide the most high level view look for trends in daily data and then hourly or minute data can be delved into to provide a higher resolution view of a trend.

### **Importing Packages**

```
library(lubridate)
library(tidyverse)
library(ggplot2)
library(dplyr)
```

### **Importing the data**

```
daily_activity <- read.csv('C:/Users/Dillon/Documents/Coding Learning/Google Data Analytics Course/Google Data Analytics Course/daily_activity.csv')
sleepday <- read.csv('C:/Users/Dillon/Documents/Coding Learning/Google Data Analytics Course/Google Data Analytics Course/sleepday.csv')
```

### **Taking a look at the data;**

```
tbl_day_act <- as_tibble(daily_activity)
tbl_day_act
```

```
## # A tibble: 940 x 15
##       Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~
##       <dbl> <chr>          <int>          <dbl>          <dbl>          <dbl>
##  1 1.50e9 4/12/2016          13162           8.5            8.5            0
##  2 1.50e9 4/13/2016          10735           6.97           6.97           0
##  3 1.50e9 4/14/2016          10460           6.74           6.74           0
##  4 1.50e9 4/15/2016           9762           6.28           6.28           0
##  5 1.50e9 4/16/2016          12669           8.16           8.16           0
##  6 1.50e9 4/17/2016           9705           6.48           6.48           0
##  7 1.50e9 4/18/2016          13019           8.59           8.59           0
##  8 1.50e9 4/19/2016          15506           9.88           9.88           0
##  9 1.50e9 4/20/2016          10544           6.68           6.68           0
## 10 1.50e9 4/21/2016           9819           6.34           6.34           0
## # ... with 930 more rows, and 9 more variables: VeryActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   SedentaryActiveDistance <dbl>, VeryActiveMinutes <int>,
## #   FairlyActiveMinutes <int>, LightlyActiveMinutes <int>,
## #   SedentaryMinutes <int>, Calories <int>
```

```
tbl_sleep <- as_tibble(sleepday)
tbl_sleep
```

```
## # A tibble: 413 x 5
##       Id SleepDay          TotalSleepReco~ TotalMinutesAsl~ TotalTimeInBed
##       <dbl> <chr>          <int>          <int>          <int>
##  1 1503960366 4/12/2016 12:00:~           1           327           346
##  2 1503960366 4/13/2016 12:00:~           2           384           407
##  3 1503960366 4/15/2016 12:00:~           1           412           442
##  4 1503960366 4/16/2016 12:00:~           2           340           367
##  5 1503960366 4/17/2016 12:00:~           1           700           712
##  6 1503960366 4/19/2016 12:00:~           1           304           320
##  7 1503960366 4/20/2016 12:00:~           1           360           377
##  8 1503960366 4/21/2016 12:00:~           1           325           364
##  9 1503960366 4/23/2016 12:00:~           1           361           384
## 10 1503960366 4/24/2016 12:00:~           1           430           449
## # ... with 403 more rows
```

## Data Cleaning & Filtering

Renaming column names for readability.

```
tbl_day_act <- rename(tbl_day_act, ID = Id, Date = ActivityDate)
tbl_sleep <- rename(tbl_sleep, ID = Id, Datetime = SleepDay)
```

**Change Date & ID format** The format of column “ActivityDate” needs to be changed to a date format. Same with the sleep table

```
tbl_day_act$Date <- mdy(tbl_day_act$Date)
tbl_day_act$ID <- as.character(tbl_day_act$ID)

tbl_sleep$Datetime<-gsub(" AM","",as.character(tbl_sleep$Datetime))
tbl_sleep$Datetime <- mdy_hms(tbl_sleep$Datetime)
tbl_sleep$ID <- as.character(tbl_sleep$ID)
```

**Check for duplicates** A return of 1 from the sum value after removal means there are none left not that there is 1. In the case of the sleep table, there are 4 rows that are duplicates, removing 3 leaves one 1.

```
sum(duplicated(tbl_day_act), vars = TRUE)
```

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

```
tbl_day_act <- tbl_day_act[!duplicated(tbl_day_act), ]
```

```
sum(duplicated(tbl_sleep), vars = TRUE)
```

```
## [1] 4
```

```
tbl_sleep <- tbl_sleep[!duplicated(tbl_sleep), ]
sum(duplicated(tbl_sleep), vars = TRUE)
```

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

---

## Exploratory Analysis

**Calculating the number of unique (distinct) user IDs in both tables.** The data source stated that 30 users supplied Fitbit tracker data. Therefore, Expectation = maximum value of 30

```
n_distinct(tbl_day_act$ID)
```

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

```
n_distinct(tbl_sleep$ID)
```

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

There are 33 user IDs in this table. *Cross-checking in excel:* using Data -> Remove Duplicates gave 33 for dailyActivity\_merged.csv and 24 for the sleepDay\_merged.csv table. The value of 33 could be users with multiple accounts, a second account created by a user during the timeframe that data was collected etc.

**Calculating the number of unique dates in the daily activity table** The data source stated that the data was collected between the 12th of April and the 12th of May. If the end date is included there should be 31 distinct dates. Creating a table of amount of days logged by each person(number of distinct dates for each ID)

```
n_distinct(tbl_day_act$Date)
```

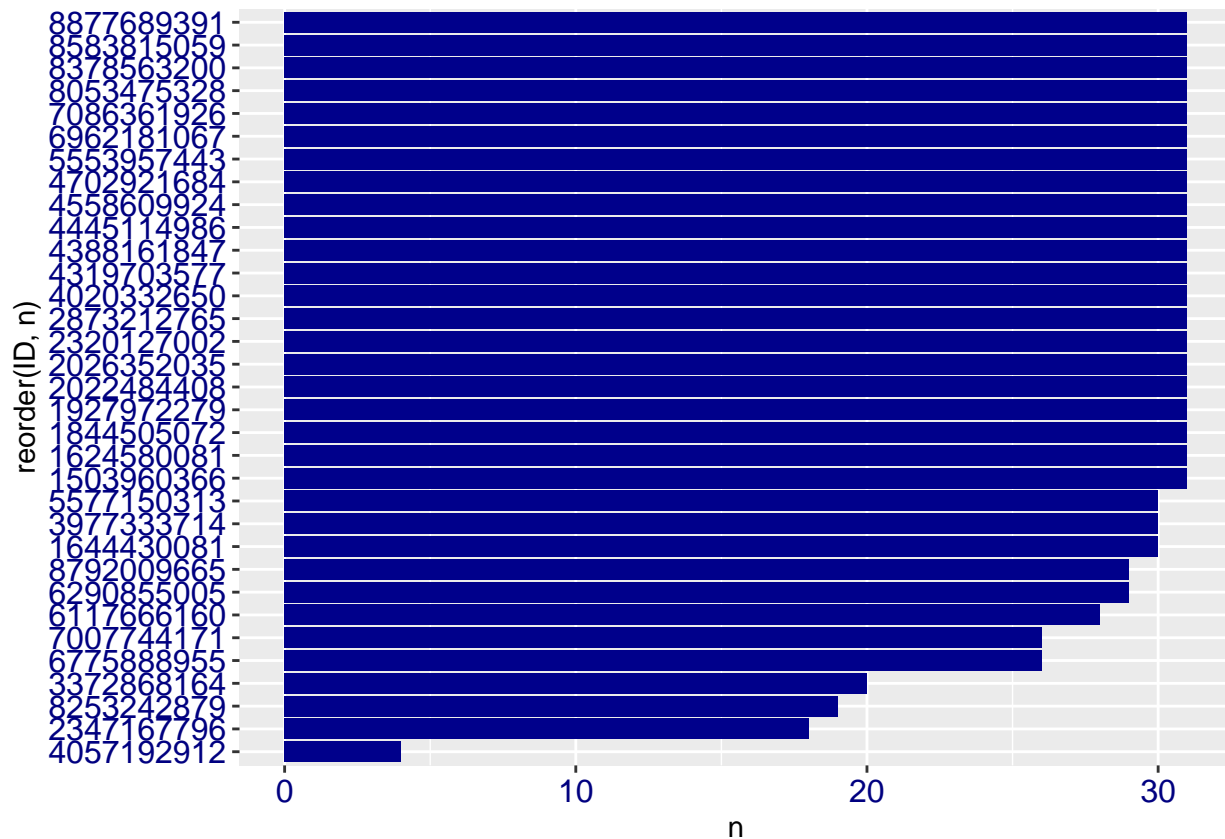
```
## [1] 31
```

```
#number of distinct dates for each ID
```

```
data_days <- as_tibble(count_(unique(tbl_day_act), vars = 'ID'))  
data_days
```

```
## # A tibble: 33 x 2  
##       ID           n  
##   <chr>     <int>  
## 1 1503960366    31  
## 2 1624580081    31  
## 3 1644430081    30  
## 4 1844505072    31  
## 5 1927972279    31  
## 6 2022484408    31  
## 7 2026352035    31  
## 8 2320127002    31  
## 9 2347167796    18  
## 10 2873212765    31  
## # ... with 23 more rows
```

```
ggplot(data = data_days, aes(x = reorder(ID, n), y = n)) +  
  geom_bar(stat = 'identity', fill='blue4') +  
  theme(  
    axis.text.x = element_text(color = "navyblue", size = 12, angle = 0),  
    axis.text.y = element_text(color = "navyblue", size = 12, angle = 0)) +  
  coord_flip()
```



Summary statistics about each data frame For the daily activity dataframe:

```
data_days %>%
  select(n)%>%
  summary()
```

```
##           n
##  Min.    : 4.00
## 1st Qu.:29.00
##  Median :31.00
##   Mean  :28.48
## 3rd Qu.:31.00
##   Max.  :31.00
```

```
tbl_day_act %>%
  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
```

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

```
## TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## Min. :1.00 Min. : 58.0 Min. : 61.0
## 1st Qu.:1.00 1st Qu.:361.0 1st Qu.:403.8
## Median :1.00 Median :432.5 Median :463.0
## Mean :1.12 Mean :419.2 Mean :458.5
## 3rd Qu.:1.00 3rd Qu.:490.0 3rd Qu.:526.0
## Max. :3.00 Max. :796.0 Max. :961.0
```

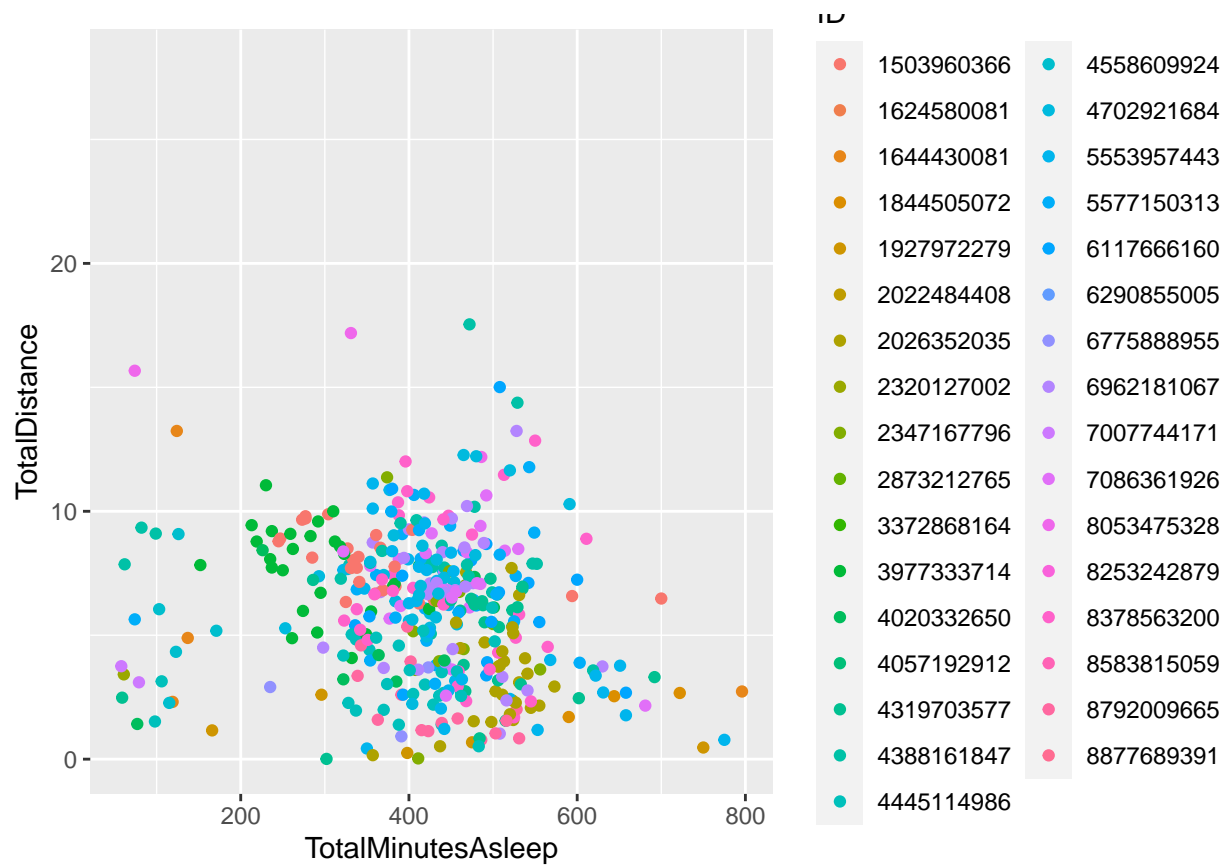
From the summaries we can see that this group of Fitbit users has averaged roughly 7500 steps per day.

**Outer joining the sleep & daily activity tables** The tables could be inner joined but with an outer join, the option to filter out missing sleep data is there while still having access all of the data in the same dataframe.

```
tbl_sleep$Datetime <- date(tbl_sleep$Datetime)
tbl_sleep <- rename(tbl_sleep, Date = Datetime)
act_sleep <- as_tibble(merge(x=tbl_day_act, y=tbl_sleep, by= c('Date', 'ID'), all=TRUE))
```

**Investigating the effect of sleep on activity** 1. Filter out N/As in the sleep data. 2. Create a scatter plot of everyone's total time asleep vs total distance. 3. Fit a line and correlate both variables. 4. What type of exercise does sleep correlate best to?

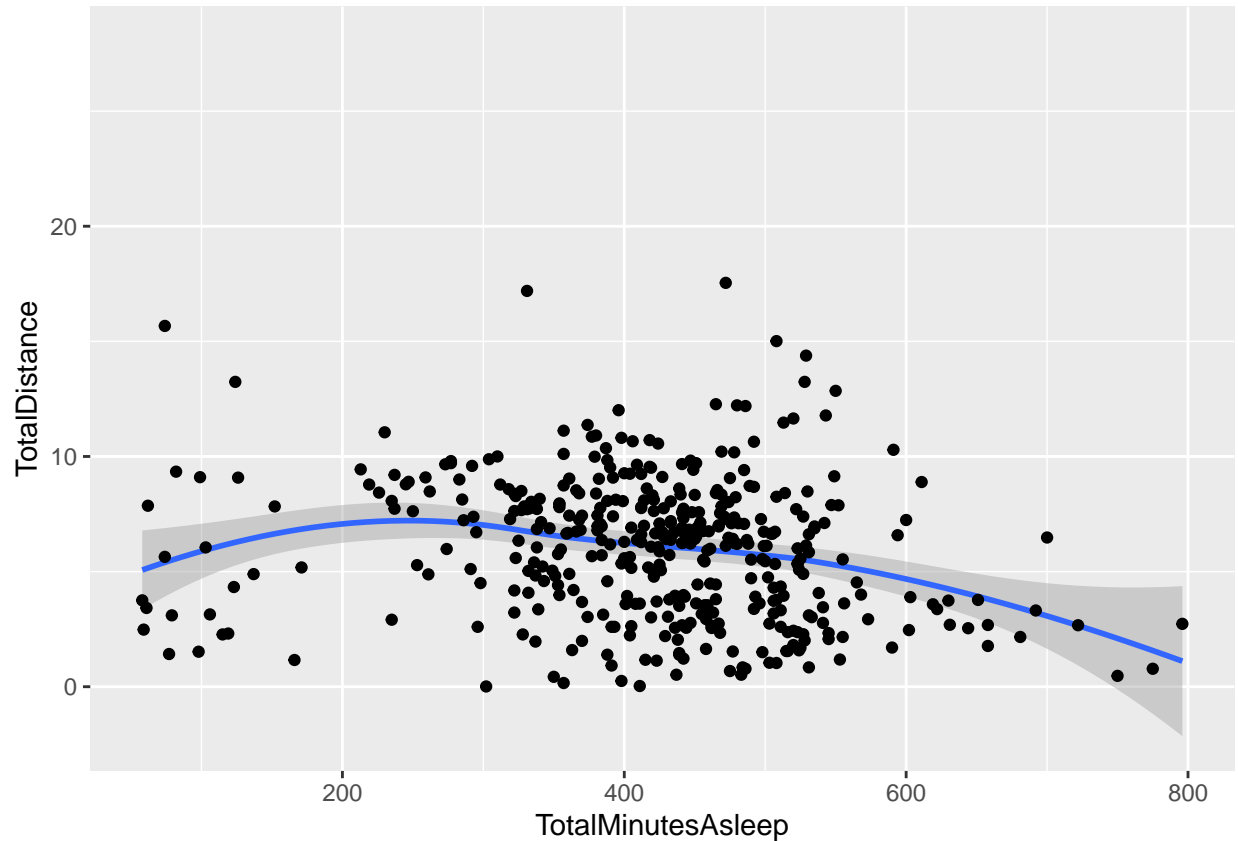
```
ggplot(data = act_sleep, aes(x = TotalMinutesAsleep, y = TotalDistance, color = ID)) +
  geom_point()
```



```
ggplot(data = act_sleep) +
  geom_smooth(mapping = aes(x = TotalMinutesAsleep, y = TotalDistance)) +
  geom_point(mapping = aes(x = TotalMinutesAsleep, y = TotalDistance))
```

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



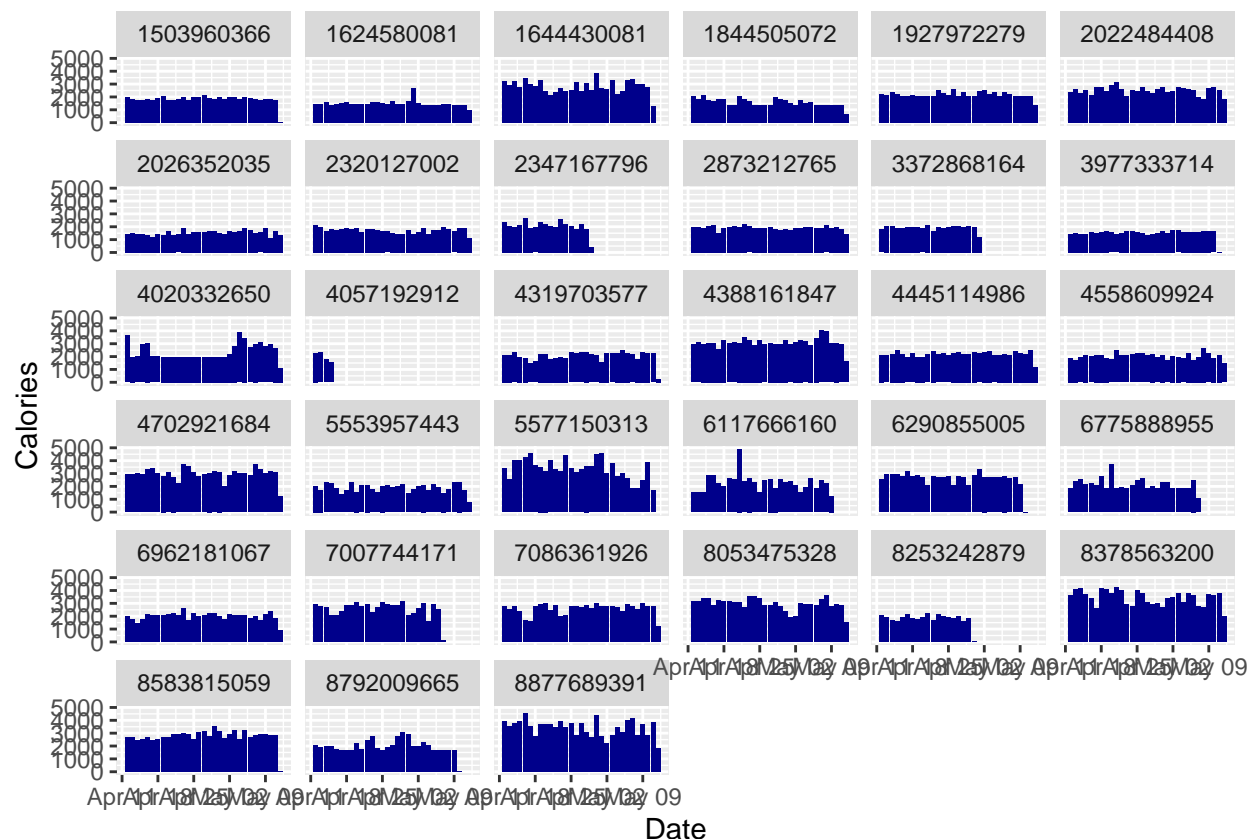


There is an indication of a negative correlation. More sleep equals less distance traveled.

## Problems

- The Fitbit data used was collected from only 30 users. This is a very small sample size when considering that there are around 30 million users. Our sample size therefore represents roughly 1 millionth or 0.0001% of the population.
- In addition to the above, the data was collected in 2016 so it is outdated. Were this study carried out a year ago during a pandemic and lockdowns, the data might even be misleading as people's exercise and eating habits are likely not the same during a pandemic.
- During the exploration of the dailyactivity table, 33 unique IDs were found. This is unexpected because it was stated that Fitbit data was collected from 30 users and yet there are 33 unique user IDs in the daily activity table.

```
ggplot(data = tbl_day_act, aes(x = Date, y = Calories)) +
  geom_bar(stat = 'identity', fill='blue4') +
  facet_wrap(~ID)
```



## Findings

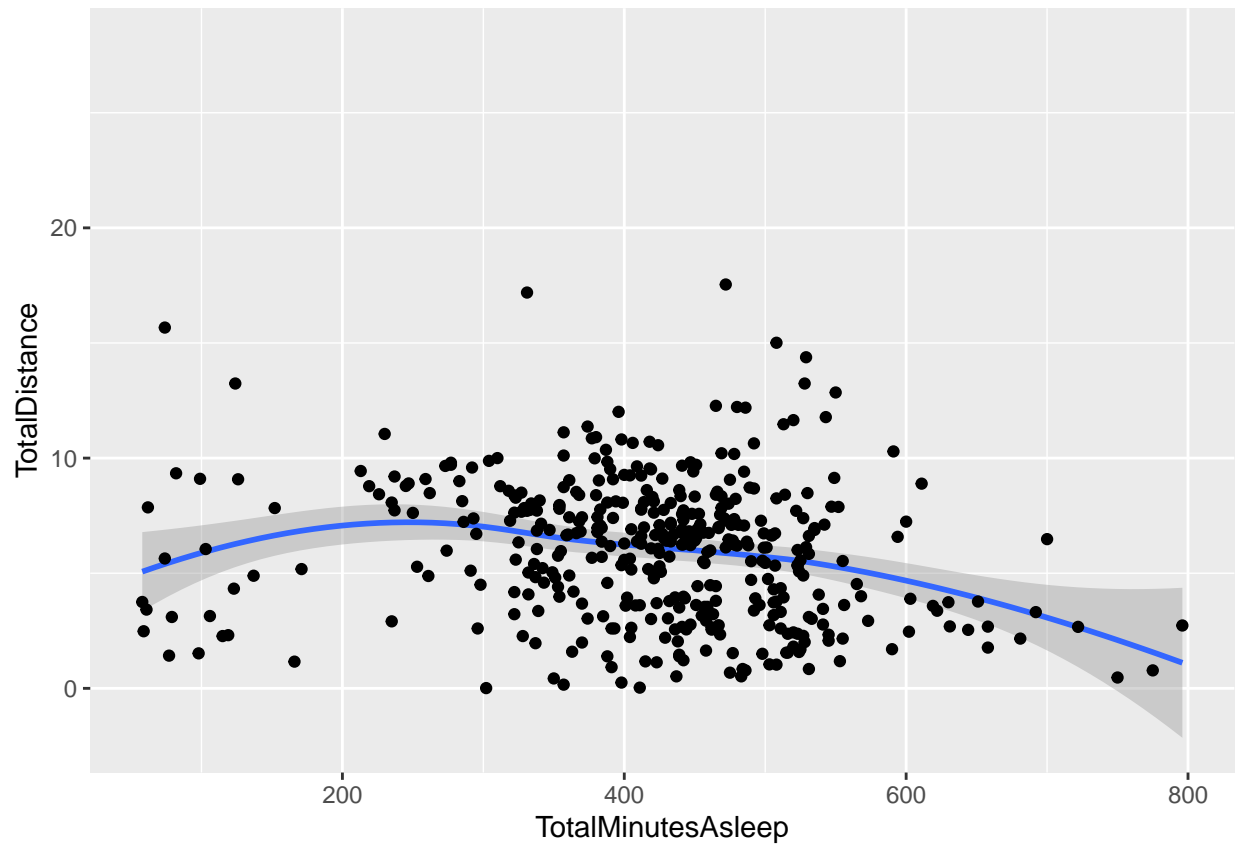
1. The average daily steps in this sample group was roughly 7,500.
2. There is a negative correlation between total daily distance and time spent asleep.

## Discussion

1. **Daily Step Average** The average daily step count for our sample was roughly 7,500. This is below the recommended CDC goal for a daily average of 10,000 steps/day according to this document. This means our users are below the recommended daily count of 10,000 by 2,500 steps. However they exceed the US average of 3,000 - 4,000 by 3,500 - 4,500 steps.
2. **Time asleep vs total distance**

```
ggplot(data = act_sleep) +
  geom_smooth(mapping = aes(x = TotalMinutesAsleep, y = TotalDistance)) +
  geom_point(mapping = aes(x = TotalMinutesAsleep, y = TotalDistance))
```

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



There seems to be no positive correlation here, in fact there is an indication of a negative correlation. Which sounds counter intuitive because one might expect that more well rested person would be less sedentary. But when given further thought, it seems sensible. The simple fact that the more time spent asleep means less time spent moving and since total distance simply tracks movement and not exercise, means more sleep equals less distance traveled.

---

## Recommendations

### 1. Positive reinforcement and encouragement prompt;

- Implement a prompt in the app that uses positive reinforcement to increase a users step count by comparing their step count to the national average. (The cdc document on physical activity linked in the discussion section can be used for America). e.g. User 'X' had a daily step count of 8,000. The app would use a notification or in app message to congratulate a user that it surpassed the avg Americans step count by 4,000 steps today.
- Using user 'X' as our example again, the notification could also advise the user that the CDC recommends setting a goal of 10,000 steps per day and that they “only have 2,000” to go.