Case Study #1: Bellabeat

Drew Croteau

Introduction:

This case study fulfills the capstone project requirement for the Google Data Analytics Certificate. Although it is optional, this project will bolster my understanding of the topics covered in the course and add to my work portfolio. The project will focus on Track A / Case 2, or more specifically, the ***Bellabeat*** route.

Bellabeat:

A small, but successful, high-tech manufacturer of health-based products for women.

Scenario: I am a junior data analyst in the company's marketing department tasked with analyzing existing product data to unlock new growth opportunities.

Ask:

The ultimate goal, or business task, is to promote company growth through analysis of non-Bellabeat smart device data. In other words, what adjustments can the company make to bring in more customers or make more sales based off the data of a competing company's product. Through this analysis we can recommend adjustments to ***Bellabeat*** products to achieve this goal.

Prepare and Process:

The dataset was pulled from **https://www.kaggle.com/datasets/arashnic/fitbit?resource=download-directory&select=mturkfitbit_export_4.12.16-5.12.16** and was obtained from 30 eligible and consenting FitBit users during the timespan of March 12, 2016 - May 12, 2016.

Data Handling: The dataset is rather big and spread out across two files, 1 pertaining to each month within the timespan mentioned above. Usuable files, notably the minute, sleep, weight, and daily activity files were all uploaded to ***RStudio***. Some files containing the same amount of rows were combined to consolidate column values using the **inner_join()** function.The separated files (by month) were combined through the **full_join()** function. Everything is stored in the *long* format.

Necessary Packages:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidyr)
library(ggplot2)
```

```
overall_averages = read.csv("~/data-analysis-projects-capstone/case_studies/capstone_bellabeat/data_tabl
filtered_steps = read.csv('~/data-analysis-projects-capstone/case_studies/capstone_bellabeat/data_tables
steps_by_date = read.csv("~/data-analysis-projects-capstone/case_studies/capstone_bellabeat/data_tables/
steps_by_day = read.csv("~/data-analysis-projects-capstone/case_studies/capstone_bellabeat/data_tables/s
```

Example:

Here we'll take the sleepDay_merged table and create a new table with the averages of each column.

```
average_sleep_data = sleepDay_merged %>%
  group_by(Id) %>%
  summarize(times_slept = mean(TotalSleepRecords),
            minutes_asleep = mean(TotalMinutesAsleep),
            minutes_in_bed = mean(TotalTimeInBed))
```

Good, now we can combine this new table to an averages table created prior.

```
overall_averages = inner_join(overall_averages, average_sleep_data)
```

Limitations: Variations in specific FitBit product is briefly mentioned in the description on Kaggle but has no obvious indicator within the data itself. Some value(s), like that of the **value** column in the sleep files are non-descriptive. Other names are vague or are missing measurement values such as Calories (is that burned or contained at the time?) or TotalDistance (miles or km?) respectively. Weight logs do not contain all participant IDs.

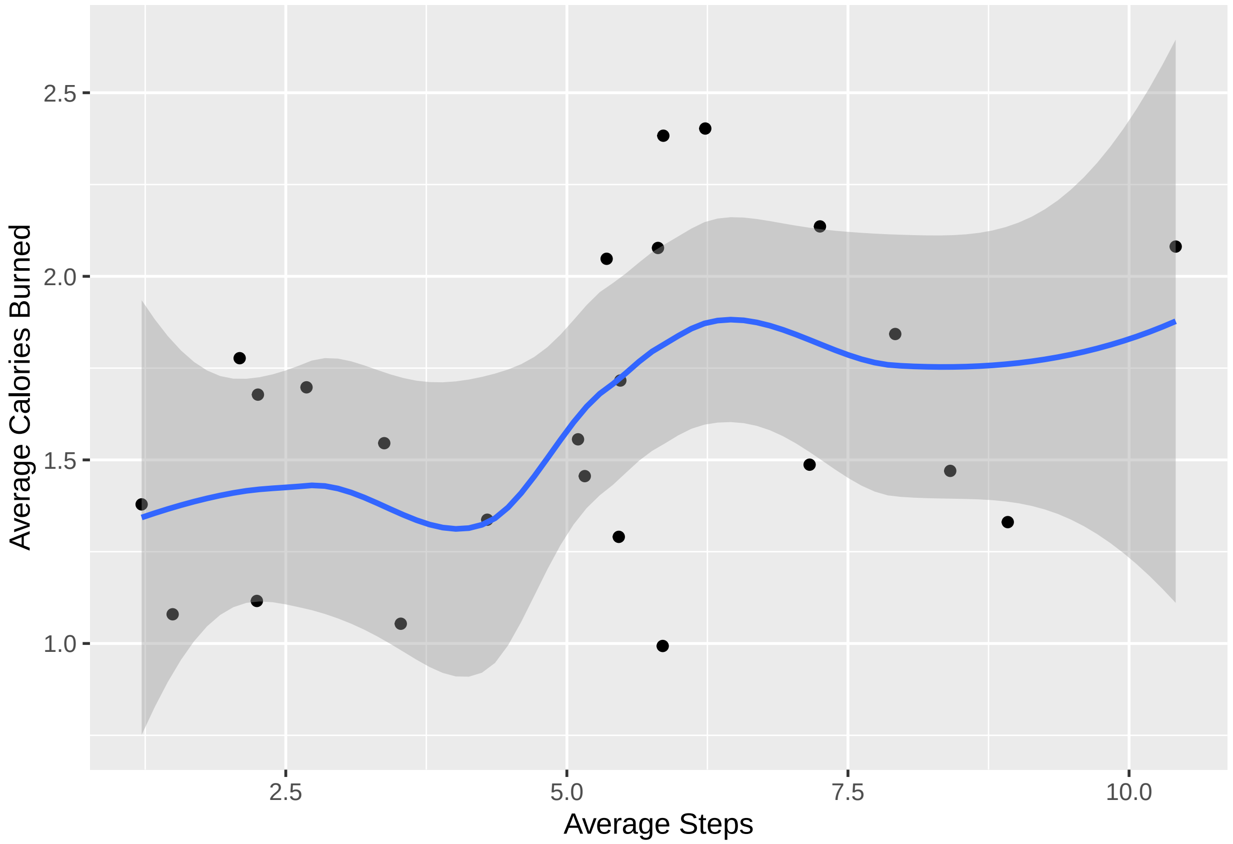
Analyze and Share:

Now that we have consolidated tables, we can start making connections with the data and create some visuals.

```
ggplot(data = overall_averages, aes(x = avg_steps, y = avg_calories_burned)) %>%
  + geom_point() %>%
  + geom_smooth() %>%
  + labs(x = 'Average Steps', y = 'Average Calories Burned')
```

Steps Vs. Calories Burned:

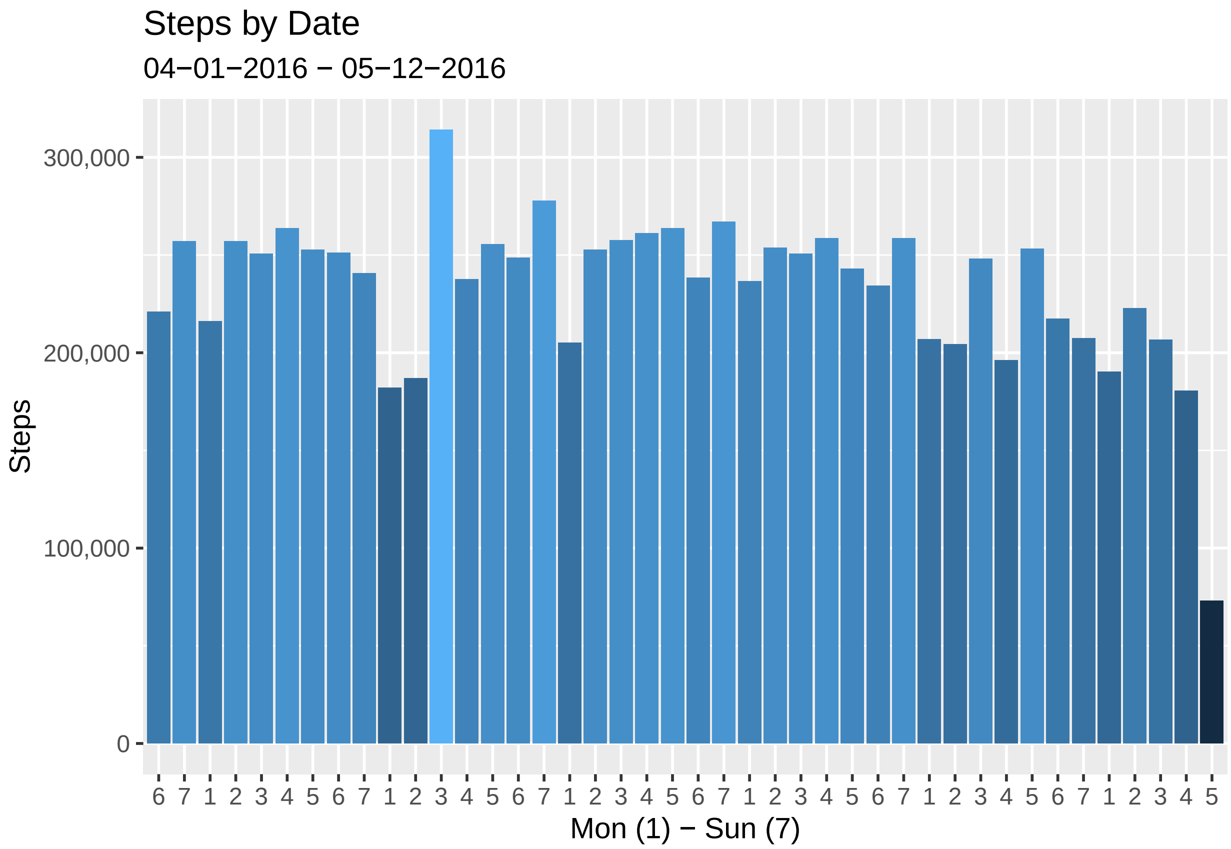
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Analysis: Fairly obvious results, there's a positive correlation between an increase in calories burned and steps taken. There appears to be more to the story, however, as everyone has different averages indicating more variables at play. Perhaps individuals who are heavier will burn more calories with fewer steps. Unfortunately we cannot dive into this notion because specific characteristics of the subjects are unavailable in this dataset.

```
ggplot(data = filtered_steps, aes(x = ActivityDate,
                                  y = total_steps, fill = total_steps)) %>%
  + geom_col() %>%
  + labs(title = 'Steps by Date', subtitle = '04-01-2016 - 05-12-2016',
       x = 'Mon (1) - Sun (7)', y = 'Steps') %>%
  + scale_x_discrete(breaks = filtered_steps$ActivityDate,
                        labels =   wday(filtered_steps$ActivityDate)) %>%
  + scale_y_continuous(labels = scales::comma) %>%
  + theme(legend.position = 'none')
```

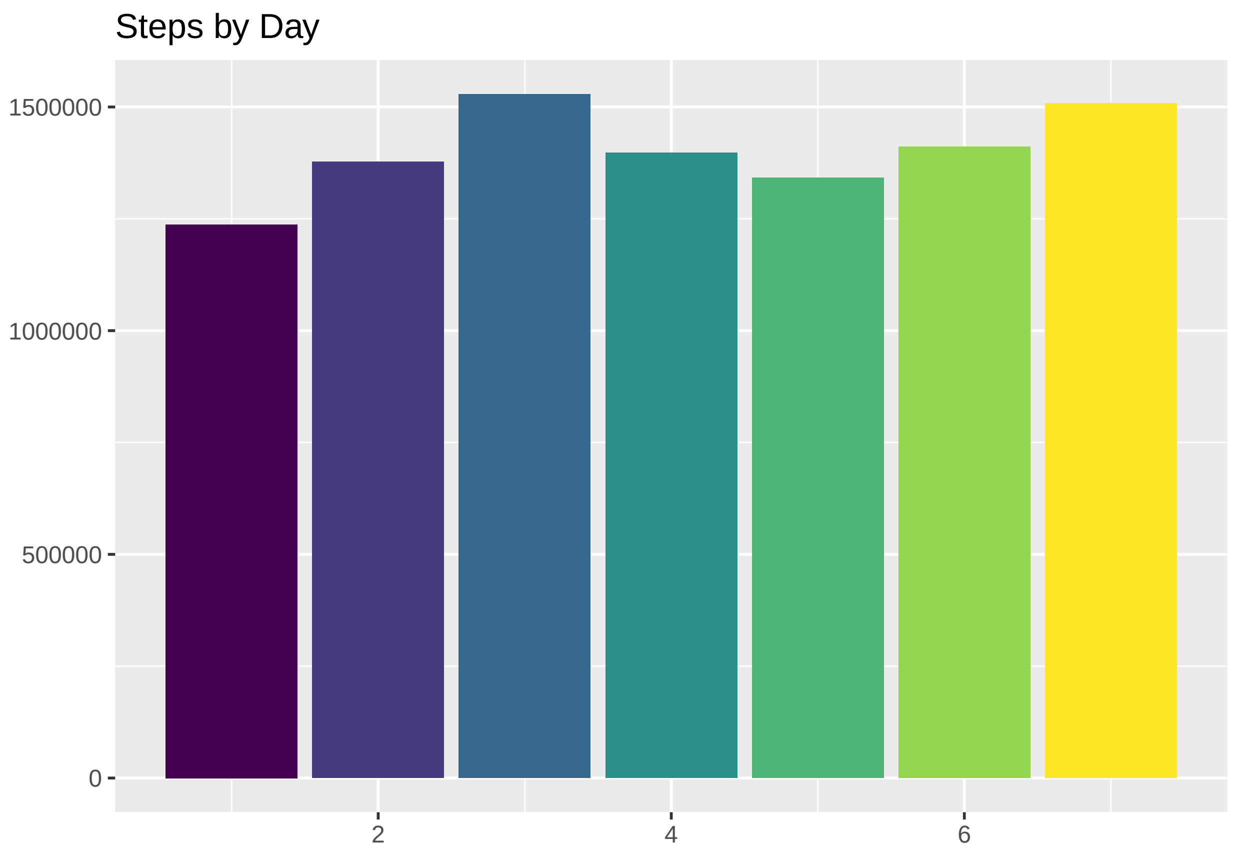
Steps by Date:



Analysis: A little more interesting of a visual, we can see activity on certain days. Monday and Tuesday seem to have the least activity from participants. The weekends seem to hold slightly higher activity. To confirm this, we'll need to create another visual.

```
ggplot(data = steps_by_day, aes(x = day, y = steps, fill = day)) %>%
  + geom_col() %>%
  + scale_fill_viridis_c() %>%
  + labs(title = 'Steps by Day', x = NULL, y = NULL) %>%
  + theme(legend.position = 'none')
```

Steps by Day of the Week



Analysis: Beautiful! Simple but informative. Our observation(s) from the previous visual was correct. Monday and Tuesday are the least active days whereas the weekends hold higher step counts. Interestingly enough, Wednesday appears to be the **MOST** active day of the week.

Act:

Suggestion: Pertaining to the business task of growth, It might be advantageous to focus marketing strategies and advertisements on days where people are more motivated to be active. Namely, Wednesdays and the weekends according to this dataset. No other substantial observations can be expressed using this dataset at this time.