Google Capstone Project: Bellabeat

Kirk Jimenez

2023-06-24

This project is from the Google Data Analytics course. The analysis conducted herein will follow the six steps of Data Analysis as covered in the course: Ask, Prepare, Process, Analyze, Share, and Act.

Step 1: Ask

This step serves to define the problem and objectives for this case study as well as the desired outcome.

Background For this scenario, I will be acting as 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 in the global fitness 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.

Business Task Analyze FitBit fitness tracker data to discover how consumers utilize the FitBit app and then use data insights to assist in Bellabeat's marketing strategy.

Step 2: Prepare

This phase involves the identification of the data to be used

Information on Data Source This data is gathered from 33 FitBit users who consented to the submission of their personal tracker data. This dataset is publicly available and can be accessed at https://www.kaggle.com/datasets/arashnic/fitbit. For the purposes of this analysis, the dailyActivity_merged.csv was the dataset of focus. The programming language 'R' was used to clean and transform data, as well as to provide data visualizations.

Step 3: Process

This is the step where data is cleaned and transformed for efficient use.

R Environment The tidyverse package was installed, primarily for access to ggplot.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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 error
```

Importing the data set The data was imported into R

```
dailyActivity_merged <- read_csv("dailyActivity_merged.csv")
```

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

```
glimpse(dailyActivity_merged)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ 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~
```

Data Cleaning Renamed lengthy ID numbers to simple, chronological units to clean up upcoming visualization

```
dailyActivity_merged$Id[dailyActivity_merged$Id==(1503960366)] <- 1
dailyActivity_merged$Id[dailyActivity_merged$Id==(1624580081)] <- 2
dailyActivity_merged$Id[dailyActivity_merged$Id==(1644430081)] <- 3
dailyActivity_merged$Id[dailyActivity_merged$Id==(1844505072)] <- 4
dailyActivity_merged$Id[dailyActivity_merged$Id==(1927972279)] <- 5
dailyActivity_merged$Id[dailyActivity_merged$Id==(2022484408)] <- 6
dailyActivity_merged$Id[dailyActivity_merged$Id==(2026352035)] <- 7
dailyActivity_merged$Id[dailyActivity_merged$Id==(2320127002)] <- 8
dailyActivity_merged$Id[dailyActivity_merged$Id==(2347167796)] <- 9
dailyActivity_merged$Id[dailyActivity_merged$Id==(2873212765)] <- 10
dailyActivity_merged$Id[dailyActivity_merged$Id==(3372868164)] <- 11
dailyActivity_merged$Id[dailyActivity_merged$Id==(3977333714)] <- 12
dailyActivity_merged$Id[dailyActivity_merged$Id==(4020332650)] <- 13
dailyActivity_merged$Id[dailyActivity_merged$Id==(4057192912)] <- 14
dailyActivity_merged$Id[dailyActivity_merged$Id==(4319703577)] <- 15
dailyActivity_merged$Id[dailyActivity_merged$Id==(4388161847)] <- 16
dailyActivity_merged$Id[dailyActivity_merged$Id==(4445114986)] <- 17
dailyActivity_merged$Id[dailyActivity_merged$Id==(4558609924)] <- 18
```

```
dailyActivity_merged$Id[dailyActivity_merged$Id==(4702921684)] <- 19
dailyActivity_merged$Id[dailyActivity_merged$Id==(5553957443)] <- 20
dailyActivity_merged$Id[dailyActivity_merged$Id==(5577150313)] <- 21
dailyActivity_merged$Id[dailyActivity_merged$Id==(6117666160)] <- 22
dailyActivity_merged$Id[dailyActivity_merged$Id==(6290855005)] <- 23
dailyActivity_merged$Id[dailyActivity_merged$Id==(6775888955)] <- 24
dailyActivity_merged$Id[dailyActivity_merged$Id==(6962181067)] <- 25
dailyActivity_merged$Id[dailyActivity_merged$Id==(7007744171)] <- 26
dailyActivity_merged$Id[dailyActivity_merged$Id==(7086361926)] <- 27
dailyActivity_merged$Id[dailyActivity_merged$Id==(8053475328)] <- 28
dailyActivity_merged$Id[dailyActivity_merged$Id==(8253242879)] <- 29
dailyActivity_merged$Id[dailyActivity_merged$Id==(8378563200)] <- 30
dailyActivity_merged$Id[dailyActivity_merged$Id==(8583815059)] <- 31
dailyActivity_merged$Id[dailyActivity_merged$Id==(8792009665)] <- 32
dailyActivity_merged$Id[dailyActivity_merged$Id==(8877689391)] <- 33
```

Created new columns in data to ascertain averages for the "VeryActive", "FairlyActive", "LightlyActive", and "Sedentary" values in service of future visualization

```
dailyActivity_merged <- dailyActivity_merged %>%
 mutate(Avg_VeryActiveMinutes=mean(VeryActiveMinutes))
dailyActivity_merged <- dailyActivity_merged %>%
 mutate(Avg_FairlyActiveMinutes=mean(FairlyActiveMinutes))
dailyActivity_merged <- dailyActivity_merged %>%
 mutate(Avg_LightlyActiveMinutes=mean(LightlyActiveMinutes))
dailyActivity_merged <- dailyActivity_merged %>%
 mutate(Avg_SedentaryMinutes=mean(SedentaryMinutes))
```

Step 4: Analyze

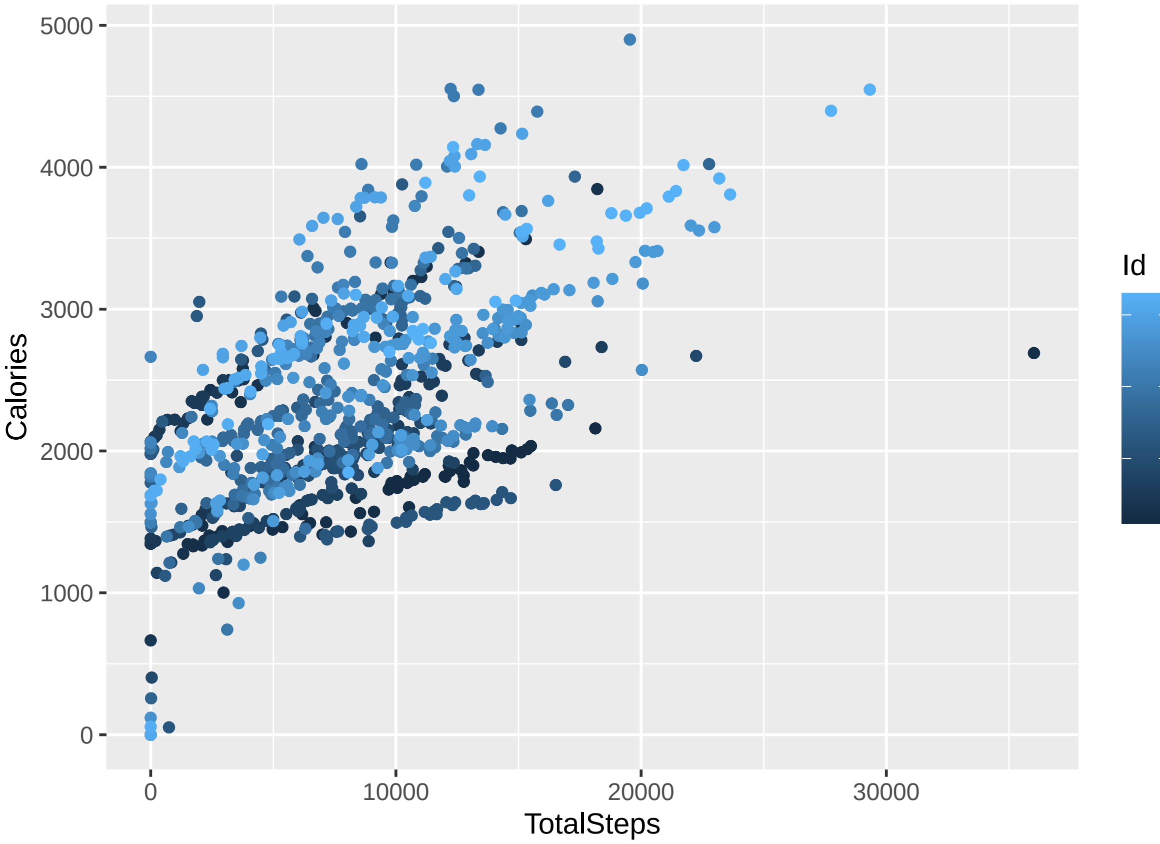
General statistical overview of relevant data points warranting analysis.

Statistical findings Analyzing the relationship between total steps taken and calories burned revealed a positive correlation: on average, the more steps you take, the more calories you burn. Given that this is a defining metric for smart fitness tracking apps, it should be reasonably concluded that having such information readily available would provide an incentive for increased step-making. However, after averaging values for the various levels of activity that were tracked, it was discovered that an overwhelming 81% of logged activity was classified as "Sedentary", disproving this hypothesis.

Step 5: Share

This step includes the visualizations used to communicate analytic findings.

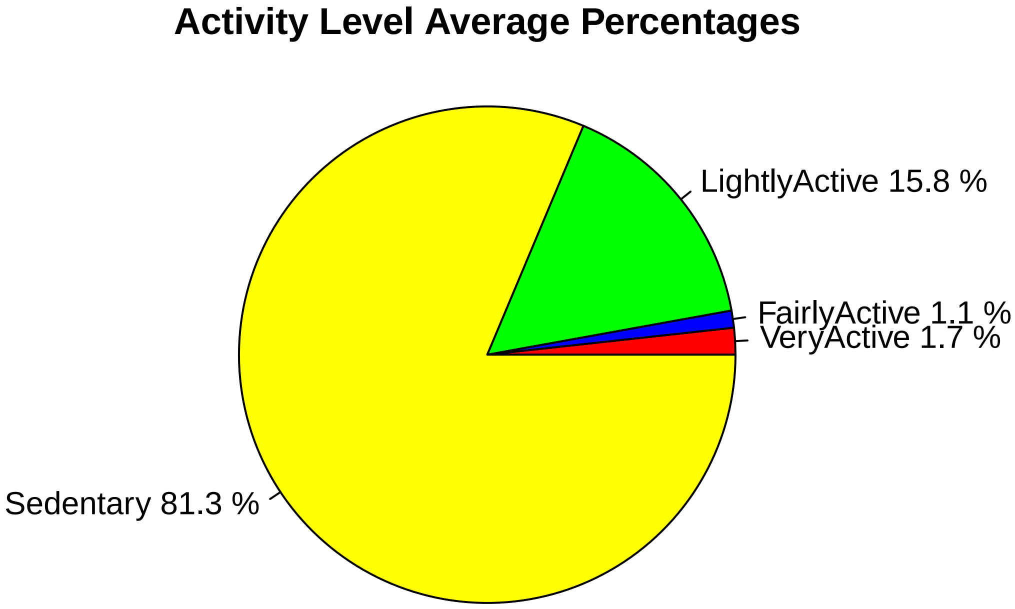
```
ggplot(data=dailyActivity_merged)+geom_point(mapping=aes(x=TotalSteps, y=Calories, color=Id))
```



Data Visualizations

This scatterplot clearly shows the positive correlation discussed earlier between number of steps taken and calories burned. The small amount of outliers can be attributed to potential software glitches, improper use of hardware, or various other uncommon anomalies.

```
activity_level <-c("VeryActive","FairlyActive","LightlyActive","Sedentary")
average_minutes <-c(21.16489,13.56489,192.8128,991.2106)
piepercent <- round(100*average_minutes/sum(average_minutes), 1)
pie(x=average_minutes, labels=paste(activity_level, sep = " ", piepercent,"%"), radius=1.05, main="Acti
```



This pie chart showcases the large, disproportionate amount of sedentary versus active data logged. This showcases that, rather than the software predominately being used to track fitness activities, users are instead tending to log idle, daily activities far more frequently.

Step 6: Act

The final step, wherein insights and recommendations are proffered bases on insights gleaned from the data.

Identified Trend The majority of FitBit use amongst the individuals in this dataset is in activities that the app designates as “Sedentary” instead of using it for its intended function of tracking health habits.

Applying trend to Bellabeat customers Both Bellabeat and FitBit have products that are designed to help users have readily available information regarding their health, habit, and fitness data. This is designed to help them be informed by their metrics to then make healthy lifestyle decisions. This trend seen in the FitBit data highlights a lack of incentive to engage users in maintaining a healthy degree of activity.

Using trend to influence Bellabeat product/marketing strategy In order for Bellabeat to improve on their existing product, it would be recommended that the software be updated to have programs that would provide relevant incentive to users for proper utilization of the product (i.e. actually using it to track fitness activities instead of primarily sedentary activity.) Examples of this could be: a functionality that provides encouraging text whenever a user reaches predetermined fitness milestones, a “currency” that users earn by frequently engaging in very active activities that can be redeemed for cosmetic rewards displayed on the product, automated, encouraging reminders that pop up whenever a user has had a prolonged period of sedentary action.