Bellabeat Project

William Kerney

03-26-2024

Introduction

The following is the capstone project for the Google Data Analytics professional certificate. Examining a dataset from a fictional company with data that is researched an sourced from a public Kaggle dataset.

About the company

Bellabeat is a high-tech company that manufactures health-focused smart products that track sleep, water intake, steps as well as other personal health statistics. One of the owners used her background as an artist to develop technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Finding and Loading Data

```r
#Find and Load Packages
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("tidyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("lubridate")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

Questions to Guide our Analysis

- 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?

Business Task

Analyze smart device usage data in order to gain insight into how people are already using their smart devices. Then, using this information, deliver high-level recommendations to the marketing team for strategies.

```r
#Loading packages to work with data.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
#Importing and converting table names.
daily_activity <-read_csv("dailyActivity_merged - dailyActivity_merged.csv")
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (13): Id, TotalDistance, TrackerDistance, LoggedActivitiesDistance, Very...
## num  (1): TotalSteps
##
## 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.
```

```r
heart_rate <- read_csv("heartrate_seconds_merged.csv")
```

```
## Rows: 1154681 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Time
## dbl (2): Id, Value
##
## 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.
```

```r
daily_calories <- read_csv("dailyCalories_merged - dailyCalories_merged.csv")
```

```
## Rows: 940 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (2): Id, Calories
##
## 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.
```

```r
weight_data <- read_csv("weightLogInfo_merged -  weightLogInfo_merged Clean.csv")
```

```
## Rows: 67 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
##
## 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.
```

```r
sleep_data <- read_csv("sleepDay_merged - sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## 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.
```

By importing and analyzing these tables, we can gain comprehensive insights into how users are utilizing their smart devices for health and wellness. These insights can then be used to formulate high-level recommendations for the marketing team, such as:

- Tailoring marketing campaigns to target specific demographics of users based on their activity levels, sleep patterns, and health goals.
- Develop personalized recommendations to encourage users to engage more with their smart devices and achieve their health and wellness goals.
- Partnering with other health and wellness brands to offer integrated solutions that address users' holistic needs, such as fitness equipment, nutrition supplements, or wellness programs. Bellabeat specifically offers products such as an app, a watch, a smart water bottle and a necklace.

```r
#Using Head() to get a better view of the data sources
head(daily_activity)
```

```
## # A tibble: 6 x 15
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
##        <dbl> <chr>             <dbl>         <dbl>           <dbl>
## 1 1503960366 4/12/2016         13162          8.5             8.5
## 2 1503960366 4/13/2016         10735          6.97            6.97
## 3 1503960366 4/14/2016         10460          6.74            6.74
## 4 1503960366 4/15/2016          9762          6.28            6.28
## 5 1503960366 4/16/2016         12669          8.16            8.16
## 6 1503960366 4/17/2016          9705          6.48            6.48
## # i 10 more variables: LoggedActivitiesDistance <dbl>,
## #   VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
```

```
head(heart_rate)
```

```
## # A tibble: 6 x 3
##           Id Time                 Value
##        <dbl> <chr>                <dbl>
## 1 2022484408 4/1/2016 7:54:00 AM    93
## 2 2022484408 4/1/2016 7:54:05 AM    91
## 3 2022484408 4/1/2016 7:54:10 AM    96
## 4 2022484408 4/1/2016 7:54:15 AM    98
## 5 2022484408 4/1/2016 7:54:20 AM   100
## 6 2022484408 4/1/2016 7:54:25 AM   101
```

```
head(daily_calories)
```

```
## # A tibble: 6 x 3
##           Id ActivityDay Calories
##        <dbl> <chr>          <dbl>
## 1 1503960366 4/12/2016       1985
## 2 1503960366 4/13/2016       1797
## 3 1503960366 4/14/2016       1776
## 4 1503960366 4/15/2016       1745
## 5 1503960366 4/16/2016       1863
## 6 1503960366 4/17/2016       1728
```

```
head(weight_data)
```

```
## # A tibble: 6 x 8
##           Id Date      WeightKg WeightPounds   Fat   BMI IsManualReport    LogId
##        <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl> <lgl>             <dbl>
## 1 1503960366 5/2/2016      52.6         116.    22  22.6 TRUE            1.46e12
## 2 1503960366 5/3/2016      52.6         116.    NA  22.6 TRUE            1.46e12
## 3 1927972279 4/13/2016    134.          294.    NA  47.5 FALSE           1.46e12
## 4 2873212765 4/21/2016     56.7         125.    NA  21.5 TRUE            1.46e12
## 5 2873212765 5/12/2016     57.3         126.    NA  21.7 TRUE            1.46e12
## 6 4319703577 4/17/2016     72.4         160.    25  27.5 TRUE            1.46e12
```

```
head(sleep_data)
```

```
## # A tibble: 6 x 5
##           Id SleepDay  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##        <dbl> <chr>                 <dbl>              <dbl>          <dbl>
## 1 1503960366 4/12/2016                 1                327            346
## 2 1503960366 4/13/2016                 2                384            407
## 3 1503960366 4/15/2016                 1                412            442
## 4 1503960366 4/16/2016                 2                340            367
## 5 1503960366 4/17/2016                 1                700            712
## 6 1503960366 4/19/2016                 1                304            320
```

```r
#Using str() to get a view of the structures of the dataframes
str(daily_activity)
```

```
## spc_tbl_ [940 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                      : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate            : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps              : num [1:940] 13162 10735 10460 9762 12669 ...
##  $ TotalDistance           : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance         : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance      : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance     : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes       : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes     : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes    : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes        : num [1:940] 728 776 1218 726 773 ...
##  $ Calories                : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDate = col_character(),
##   ..   TotalSteps = col_number(),
##   ..   TotalDistance = col_double(),
##   ..   TrackerDistance = col_double(),
##   ..   LoggedActivitiesDistance = col_double(),
##   ..   VeryActiveDistance = col_double(),
##   ..   ModeratelyActiveDistance = col_double(),
##   ..   LightActiveDistance = col_double(),
##   ..   SedentaryActiveDistance = col_double(),
##   ..   VeryActiveMinutes = col_double(),
##   ..   FairlyActiveMinutes = col_double(),
##   ..   LightlyActiveMinutes = col_double(),
##   ..   SedentaryMinutes = col_double(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(heart_rate)
```

```
## spc_tbl_ [1,154,681 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id   : num [1:1154681] 2.02e+09 2.02e+09 2.02e+09 2.02e+09 2.02e+09 ...
##  $ Time : chr [1:1154681] "4/1/2016 7:54:00 AM" "4/1/2016 7:54:05 AM" "4/1/2016 7:54:10 AM" "4/1/201
##  $ Value: num [1:1154681] 93 91 96 98 100 101 104 105 102 106 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Time = col_character(),
##   ..   Value = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(daily_calories)
```

```
## spc_tbl_ [940 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id         : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDay = col_character(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(weight_data)
```

```
## spc_tbl_ [67 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id            : num [1:67] 1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ Date          : chr [1:67] "5/2/2016" "5/3/2016" "4/13/2016" "4/21/2016" ...
##  $ WeightKg      : num [1:67] 52.6 52.6 133.5 56.7 57.3 ...
##  $ WeightPounds  : num [1:67] 116 116 294 125 126 ...
##  $ Fat           : num [1:67] 22 NA NA NA NA 25 NA NA NA NA ...
##  $ BMI           : num [1:67] 22.6 22.6 47.5 21.5 21.7 ...
##  $ IsManualReport: logi [1:67] TRUE TRUE FALSE TRUE TRUE TRUE ...
##  $ LogId         : num [1:67] 1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   WeightKg = col_double(),
##   ..   WeightPounds = col_double(),
##   ..   Fat = col_double(),
##   ..   BMI = col_double(),
##   ..   IsManualReport = col_logical(),
##   ..   LogId = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
str(sleep_data)
```

```
## spc_tbl_ [413 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:413] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr [1:413] "4/12/2016" "4/13/2016" "4/15/2016" "4/16/2016" ...
##  $ TotalSleepRecords : num [1:413] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:413] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:413] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   SleepDay = col_character(),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Converting Inconsistencies in Data Types

```
#Cleaning and formatting dates to be consistent
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format = "%m/%d/%Y")
sleep_data$SleepDay <- as.Date(sleep_data$SleepDay, format = "%m/%d/%Y")
heart_rate$Time <- as.Date(heart_rate$Time, format = "%m/%d/%Y")
daily_calories$ActivityDay <- as.Date(daily_calories$ActivityDay, format = "%m/%d/%Y")
weight_data$Date <- as.Date(weight_data$Date, format = "%m/%d/%Y")
```

Data exploration through summarizing different columns of dataframes.

```
#Daily Activity
daily_activity %>%
  select(TotalSteps, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Cal
  summary()
```

```
##    TotalSteps    VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
##  Min.   :    0   Min.   :  0.00    Min.   :  0.00      Min.   :  0.0
##  1st Qu.: 3790   1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0
##  Median : 7406   Median :  4.00    Median :  6.00      Median :199.0
##  Mean   : 7638   Mean   : 21.16    Mean   : 13.56      Mean   :192.8
##  3rd Qu.:10727   3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0
##  Max.   :36019   Max.   :210.00    Max.   :143.00      Max.   :518.0
##  SedentaryMinutes    Calories
##  Min.   :   0.0   Min.   :   0
##  1st Qu.: 729.8   1st Qu.:1828
##  Median :1057.5   Median :2134
##  Mean   : 991.2   Mean   :2304
##  3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :1440.0   Max.   :4900
```

```
#Sleep Data
sleep_data %>%
  select(TotalMinutesAsleep, TotalTimeInBed) %>%
  summary()
```

```
##  TotalMinutesAsleep TotalTimeInBed
##  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
```

```
#Heart Rate
heart_rate %>%
  select(Value) %>%
  summary()
```

```
##      Value
##  Min.   : 36.00
##  1st Qu.: 66.00
##  Median : 77.00
##  Mean   : 79.76
##  3rd Qu.: 90.00
##  Max.   :185.00
```

```
#Weight Data
weight_data %>%
  select(WeightKg, WeightPounds, Fat, BMI) %>%
  summary()
```

```
##     WeightKg       WeightPounds        Fat             BMI
##  Min.   : 52.60   Min.   :116.0   Min.   :22.00   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:135.4   1st Qu.:22.75   1st Qu.:23.96
##  Median : 62.50   Median :137.8   Median :23.50   Median :24.39
##  Mean   : 72.04   Mean   :158.8   Mean   :23.50   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:187.5   3rd Qu.:24.25   3rd Qu.:25.56
##  Max.   :133.50   Max.   :294.3   Max.   :25.00   Max.   :47.54
##                                   NA's   :65
```

```
#Daily Calories
daily_calories %>%
  select(Calories) %>%
  summary()
```

```
##     Calories
##  Min.   :   0
##  1st Qu.:1828
##  Median :2134
##  Mean   :2304
##  3rd Qu.:2793
##  Max.   :4900
```

Summary

Physical Activity: Users are moderately active, with a focus on light activities and occasional vigorous exercise.

Sedentary Behavior: Users spend a significant amount of time sedentary, although some have more active lifestyles.

Sleep Patterns: Most users sleep around 7 hours per night, but there's variability in sleep duration.

Body Measurements: On average, users maintain healthy body weights, but there's missing data for body fat percentage.

Caloric Intake: Daily caloric intake varies among users, suggesting diverse dietary habits.

```
#Merging data
merged_data <- merge(daily_activity, sleep_data, by=c('Id'))
head(merged_data)
```

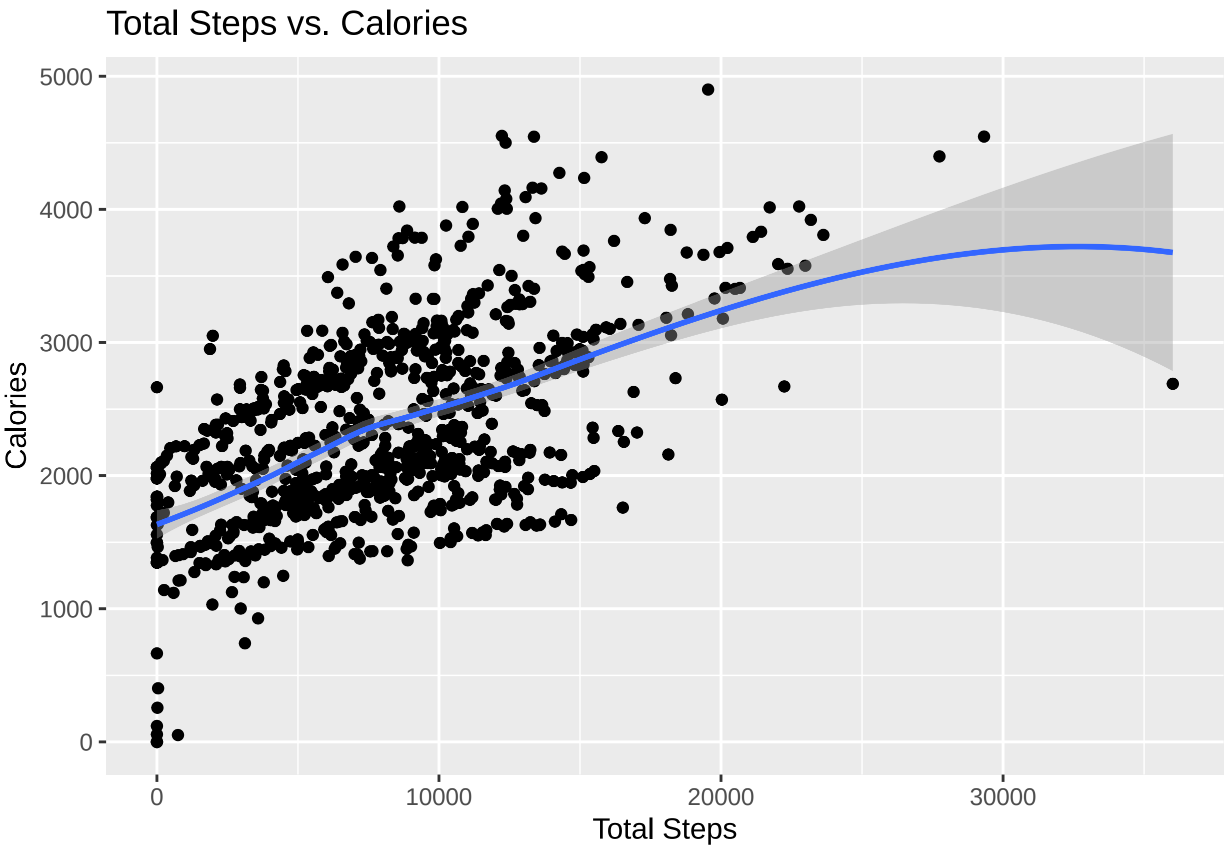
```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   2016-05-07      11992          7.71            7.71
## 2 1503960366   2016-05-07      11992          7.71            7.71
## 3 1503960366   2016-05-07      11992          7.71            7.71
## 4 1503960366   2016-05-07      11992          7.71            7.71
## 5 1503960366   2016-05-07      11992          7.71            7.71
## 6 1503960366   2016-05-07      11992          7.71            7.71
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               2.46                     2.12
## 2                        0               2.46                     2.12
## 3                        0               2.46                     2.12
## 4                        0               2.46                     2.12
## 5                        0               2.46                     2.12
## 6                        0               2.46                     2.12
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                3.13                       0                37
## 2                3.13                       0                37
## 3                3.13                       0                37
## 4                3.13                       0                37
## 5                3.13                       0                37
## 6                3.13                       0                37
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories   SleepDay
## 1                  46                  175              833     1821 2016-04-12
## 2                  46                  175              833     1821 2016-04-13
## 3                  46                  175              833     1821 2016-04-15
## 4                  46                  175              833     1821 2016-04-16
## 5                  46                  175              833     1821 2016-04-17
## 6                  46                  175              833     1821 2016-04-19
##   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1                 1                327            346
## 2                 2                384            407
## 3                 1                412            442
## 4                 2                340            367
## 5                 1                700            712
## 6                 1                304            320
```

Now we can visualize our findings.

Total Steps vs. Calories

```
#Total steps vs. Calories
ggplot(data=daily_activity, aes(x=TotalSteps, y=Calories)) +
  geom_point() + geom_smooth() + labs(title="Total Steps vs. Calories", x= "Total Steps", y="Calories")
```

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

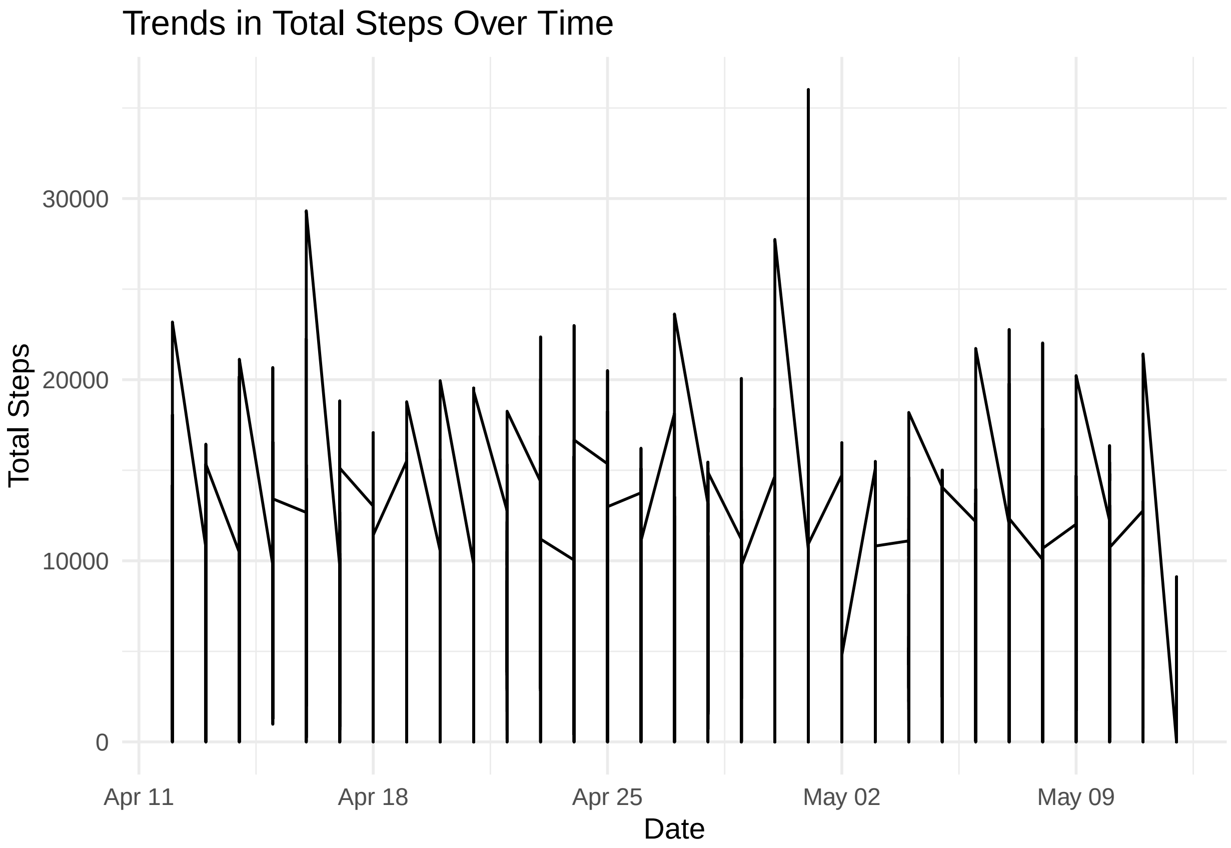


This visualization helps us better understand the relationship between total steps taken and calories burned. Understanding this relationship can help customers gauge the effectiveness of their physical activity in terms of caloric expenditure,hopefully motivating them to achieve their fitness goals.

Total Steps Over Time

```
#Total steps taken over time

ggplot(data = daily_activity, aes(x = ActivityDate, y = TotalSteps)) +
  geom_line() +
  labs(title = "Trends in Total Steps Over Time",
       x = "Date",
       y = "Total Steps") +
  theme_minimal()
```

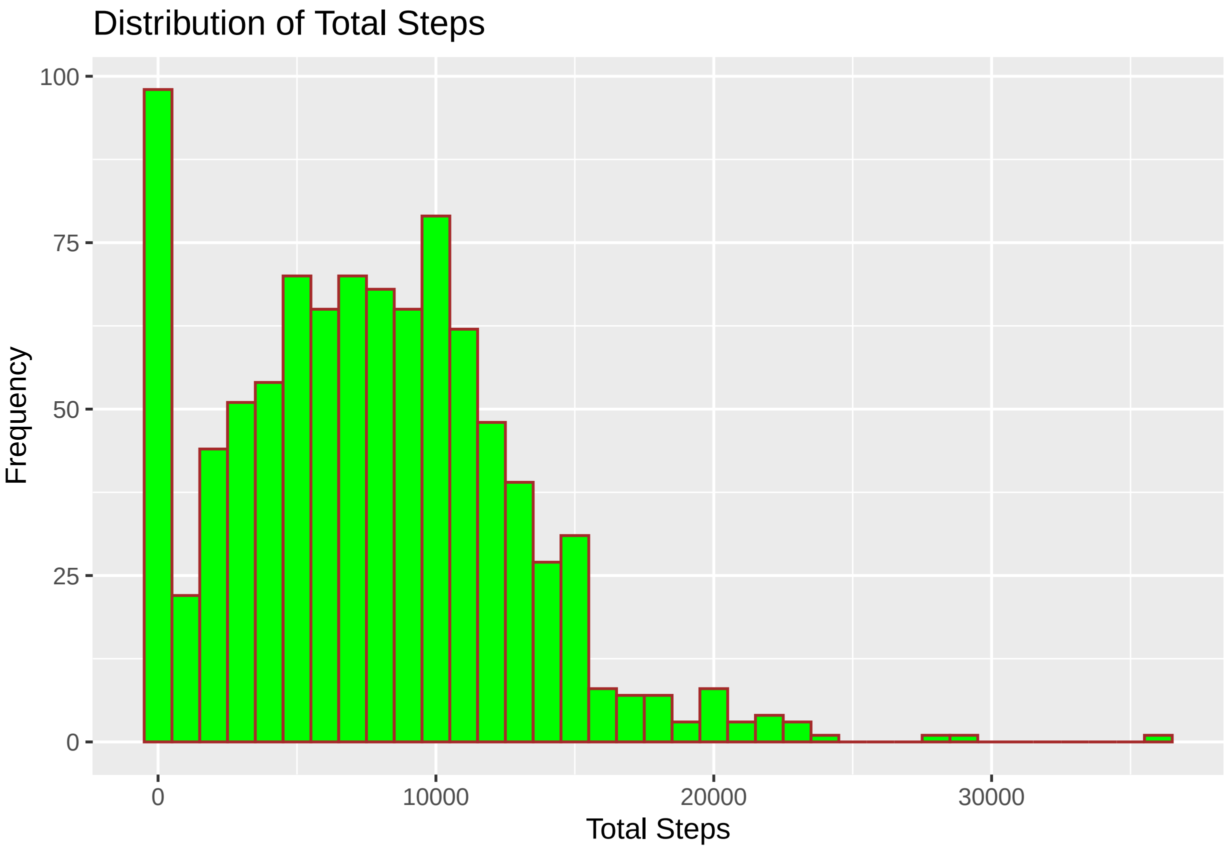


This visualization illustrates the trend in total steps taken over time, which reflects smart device usage for tracking physical activity. It also helps identify patterns, fluctuations, and overall trends in activity levels among users.

Total Steps Distribution

```r
#Total Steps Distribution

ggplot(data = daily_activity, aes(x = TotalSteps)) +
  geom_histogram(binwidth = 1000, fill = "green", color = "brown") +
  labs(title = "Distribution of Total Steps",
       x = "Total Steps",
       y = "Frequency")
```



This histogram of total steps provides insights into user activity levels. We are looking for common activity levels, peaks indicating popular activity levels, and outliers suggesting extreme behaviors. Understanding these patterns can inform marketing strategies and product development to better help serve user needs.