Bellabeat Case Study Report

Anjana

2025-06-18

Business Task

The goal of this project is to analyze smart device usage data to identify how customers are using their smart devices, enabling informed and strategic marketing decisions.

Data Sources Used

The primary data source used for this project was the Fitbit Fitness Tracker Data. It contains minute-level output for physical activity, heart rate, sleep monitoring, as well as information about daily activity and steps, useful for exploring user habits.

Installing and Loading necessary packages

The following packages were installed using install.packages() and loaded using library():

- tidyverse
- lubridate
- tidyr
- here
- dplyr
- janitor
- skimr

Cleaning and Manipulation of Data

The cleaning and manipulation of data was done with the programming language R, using RStudio Cloud. The files taken into consideration were:

- Daily Activity
- Daily Calories
- Sleep Day

Importing Datasets

The datasets were imported as follows:

```
daily_activity <- read.csv("~/Bellabeat_Capstone_GDAC/Bellabeat_Files/Bellabeat_Report_Files/dailyActivi
# head(daily_activity)

daily_calories <- read.csv("~/Bellabeat_Capstone_GDAC/Bellabeat_Files/Bellabeat_Report_Files/dailyCalori
# head(daily_calories)

sleep_day <- read.csv("~/Bellabeat_Capstone_GDAC/Bellabeat_Files/Bellabeat_Report_Files/sleepDay_merged
# head(sleep_day)
```

Checking for duplicates Upon executing the code segment below:

```
sum(duplicated(daily_activity))
sum(duplicated(daily_calories))
sum(duplicated(sleep_day))
```

```
[1] 0
```

```
[1] 0
```

```
[1] 3
```

It can be observed that there are three row-wise duplicates in the "sleepDay" dataset. This is resolved by removing them:

```
sleep_day_dist <- distinct(sleep_day)
sum(duplicated(sleep_day_dist))
```

```
[1] 0
```

```
glimpse(daily_activity)
```

Date Formatting

```
Rows: 457
Columns: 15
$ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
$ ActivityDate             <chr> "3/25/2016", "3/26/2016", "3/27/2016", "3/28/~
$ TotalSteps               <int> 11004, 17609, 12736, 13231, 12041, 10970, 122~
$ TotalDistance            <dbl> 7.11, 11.55, 8.53, 8.93, 7.85, 7.16, 7.86, 7.~
$ TrackerDistance          <dbl> 7.11, 11.55, 8.53, 8.93, 7.85, 7.16, 7.86, 7.~
$ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ VeryActiveDistance       <dbl> 2.57, 6.92, 4.66, 3.19, 2.16, 2.36, 2.29, 3.3~
$ ModeratelyActiveDistance <dbl> 0.46, 0.73, 0.16, 0.79, 1.09, 0.51, 0.49, 0.8~
$ LightActiveDistance      <dbl> 4.07, 3.91, 3.71, 4.95, 4.61, 4.29, 5.04, 3.6~
$ SedentaryActiveDistance  <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
$ VeryActiveMinutes        <int> 33, 89, 56, 39, 28, 30, 33, 47, 40, 15, 43, 3~
$ FairlyActiveMinutes      <int> 12, 17, 5, 20, 28, 13, 12, 21, 11, 30, 18, 18~
$ LightlyActiveMinutes     <int> 205, 274, 268, 224, 243, 223, 239, 200, 244, ~
$ SedentaryMinutes         <int> 804, 588, 605, 1080, 763, 1174, 820, 866, 636~
$ Calories                 <int> 1819, 2154, 1944, 1932, 1886, 1820, 1889, 186~
```

```r
glimpse(daily_calories)
```

```
Rows: 940
Columns: 3
$ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366~
$ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/~
$ Calories    <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775~
```

```r
glimpse(sleep_day_dist)
```

```
Rows: 410
Columns: 5
$ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~
$ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "~
$ TotalSleepRecords  <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~
$ TotalTimeInBed     <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~
```

The dates/days in all three datasets are in character format. We will be converting this to date and datetime formats to ensure consistency.

The dates listed in the "daily_calories" and "sleep_day" can be used to combine the datasets. However, the dates are in date+time format in the "sleep_day" dataset. They can be separated the following way:

```r
  # Conversion from Char to Date/DateTime

daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format = "%m/%d/%Y")
# head(daily_activity$ActivityDate)

# head(daily_calories$ActivityDay)

daily_calories$ActivityDay <- as.Date(daily_calories$ActivityDay, format = "%m/%d/%Y")
# head(daily_calories$ActivityDay)

sleep_day_dist$day_and_time <- mdy_hms(sleep_day_dist$SleepDay)
# glimpse(sleep_day_dist)
# View(sleep_day_dist)

  # Separating datetime into date and time

sleep_day_dist$date <- as.Date(sleep_day_dist$day_and_time)
sleep_day_dist$time <- format(sleep_day_dist$day_and_time, format = "%I:%M:%S %p")

# glimpse(sleep_day_dist)
# View(sleep_day_dist)

  # Checking if date values match in "daily_calories" and "sleep_day_dist"

all(sleep_day_dist$date %in% daily_calories$ActivityDay)
```

```
[1] TRUE
```

```r
all(daily_calories$ActivityDay %in% sleep_day_dist$date)
```

```
[1] TRUE
```

```r
  # Combining dates columns in "daily_calories" and "sleep_day"

daily_cals_sleep_merged <- merge(daily_calories, sleep_day_dist, by.x = c("Id", "ActivityDay"), by.y = c

# View(daily_cals_sleep_merged)

  # Accounting for NA values

clean_cals_sleep <- daily_cals_sleep_merged %>% distinct() %>% drop_na()
# View(clean_cals_sleep)
```

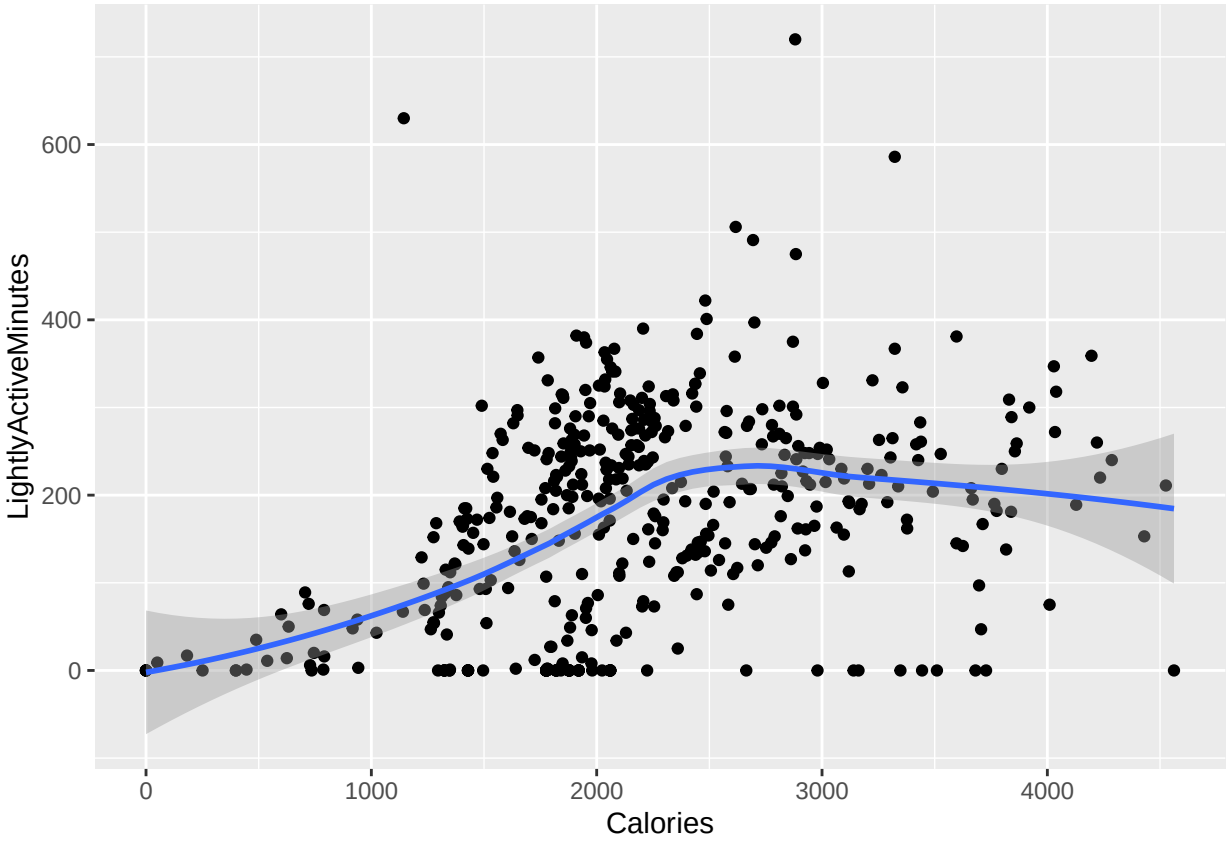
Now we move on to the summary of the analysis in the next section.

Analyzing the Data

What are some trends in smart device usage? Lightly Active Minutes vs Calories

```r
ggplot(data = daily_activity) + geom_point(mapping = aes(x=Calories, y=LightlyActiveMinutes)) + geom_sm
```

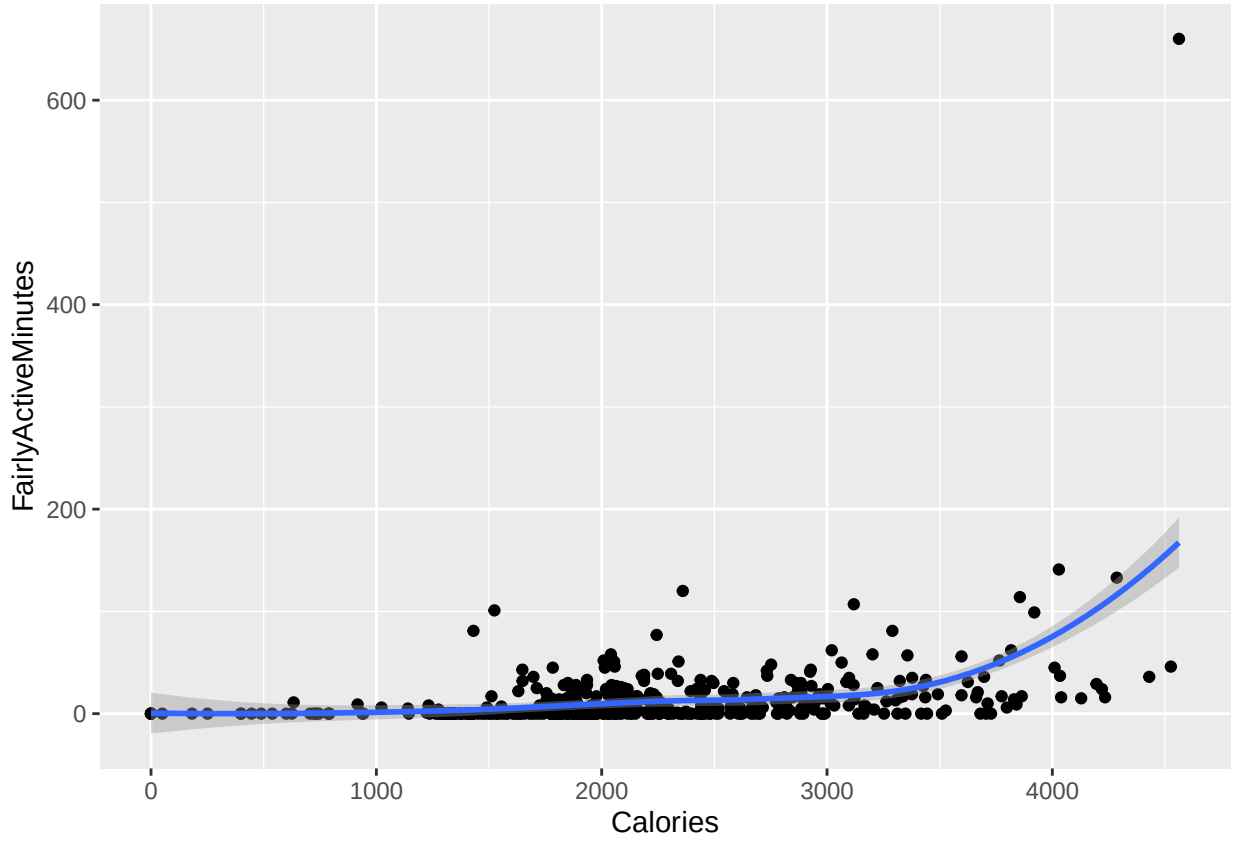
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Fairly Active Minutes vs Calories

```
ggplot(data = daily_activity) + geom_point(mapping = aes(x=Calories, y=FairlyActiveMinutes)) + geom_smo
```

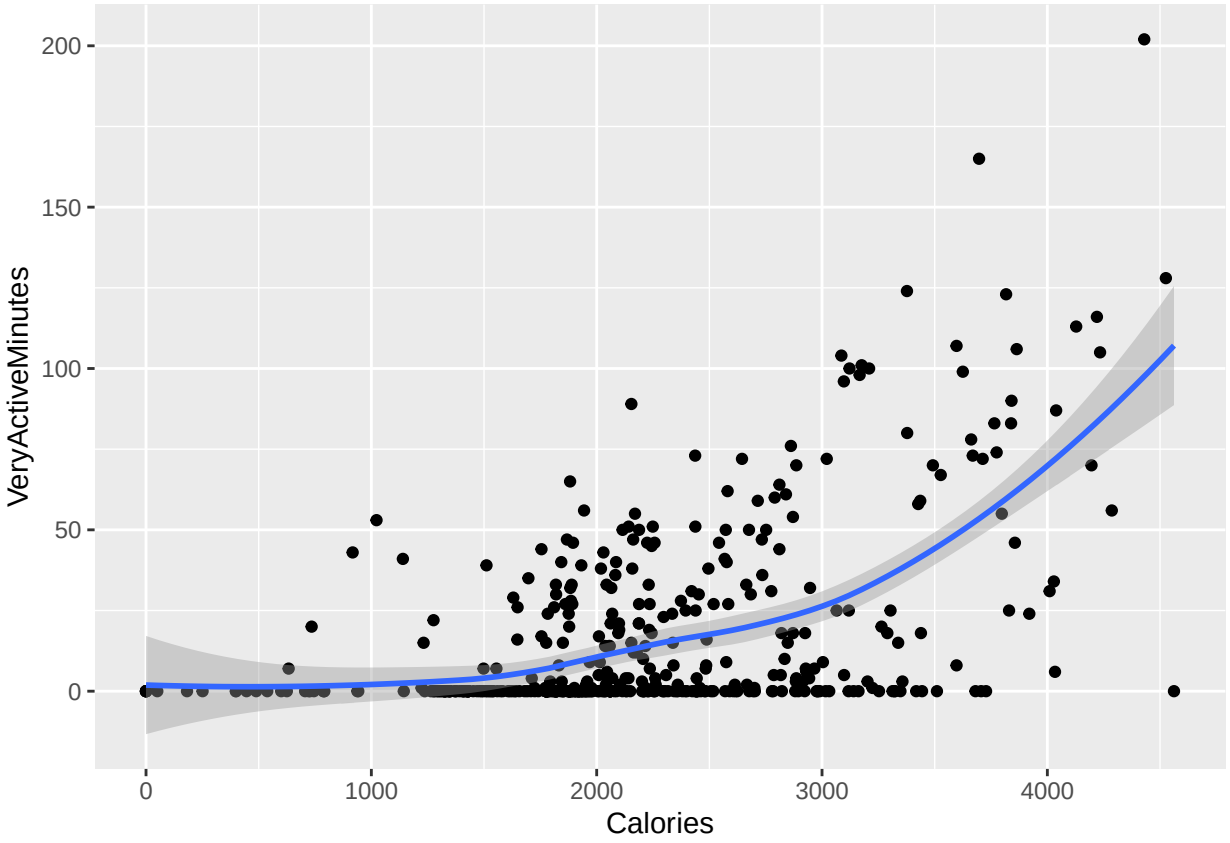
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Very Active Minutes vs Calories

```
ggplot(data = daily_activity) + geom_point(mapping = aes(x=Calories, y=VeryActiveMinutes)) + geom_smoot
```

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

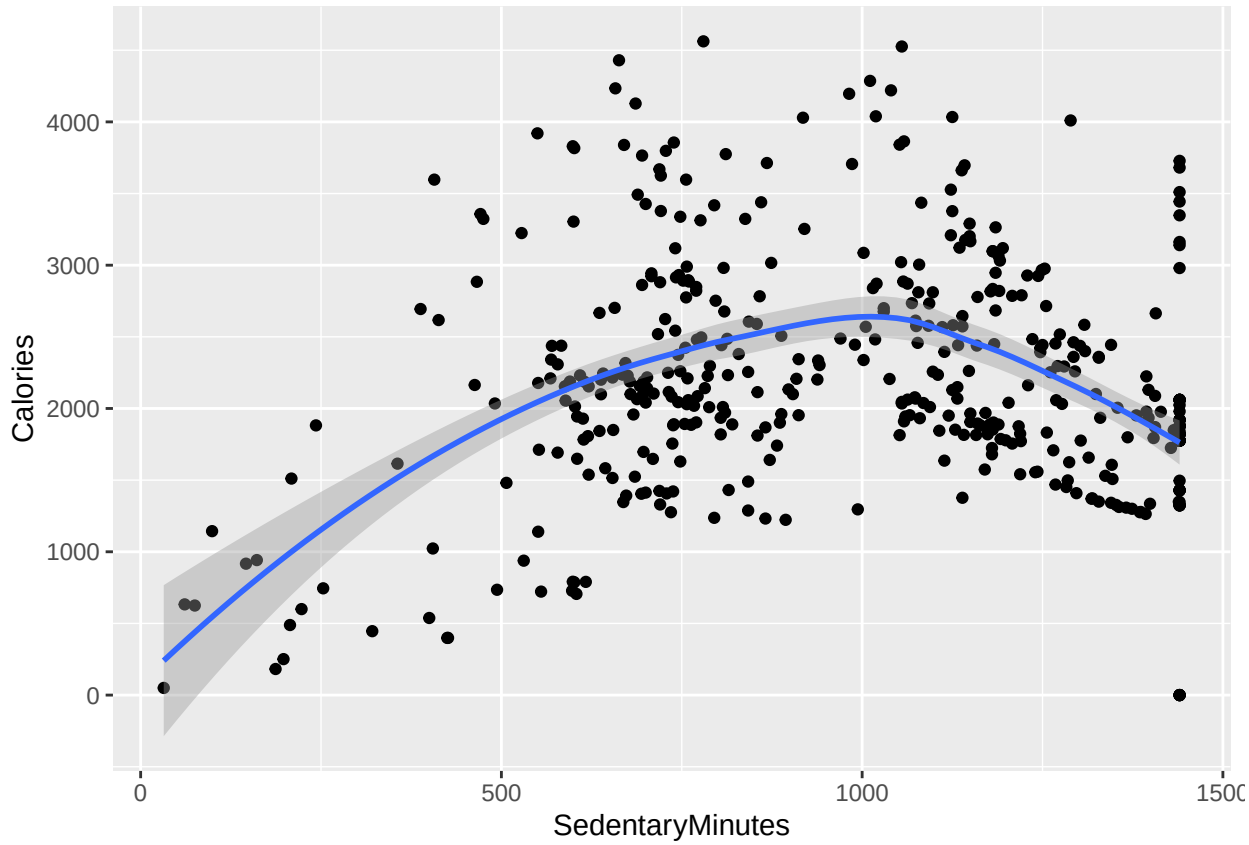



```
# ggplot(data = daily_activity, aes(x=Calories, y=VeryActiveMinutes)) + geom_col(fill='blue') + labs(ti
```

Sedentary Minutes vs Calories

```
ggplot(data = daily_activity) + geom_point(mapping = aes(x=SedentaryMinutes, y=Calories)) + geom_smooth
```

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

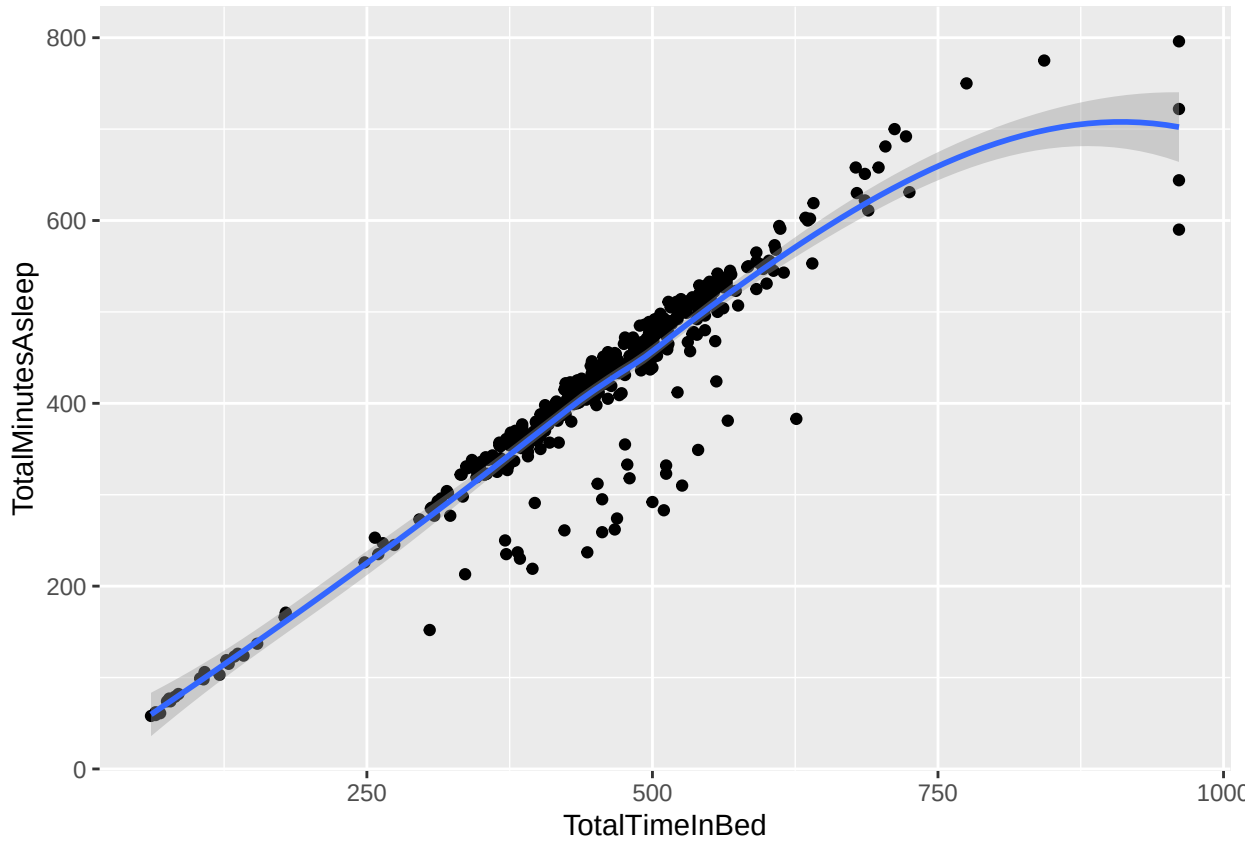


There appear to be *essentially* linear relationships for lightly active, very active, and fairly active users. That is, with an increase in the time of usage for very active, lightly active, and fairly active users, the number of calories burnt also increases. However, this is not the case for the sedentary users – the calories burnt initially increase, then decrease with even more sedentary time. While being more distributed, the overall results seem to indicate that the more sedentary people are, the lesser calories they burn.

Before examining the relationship between total sleep and calories, let us look at the total minutes asleep and total time in bed:

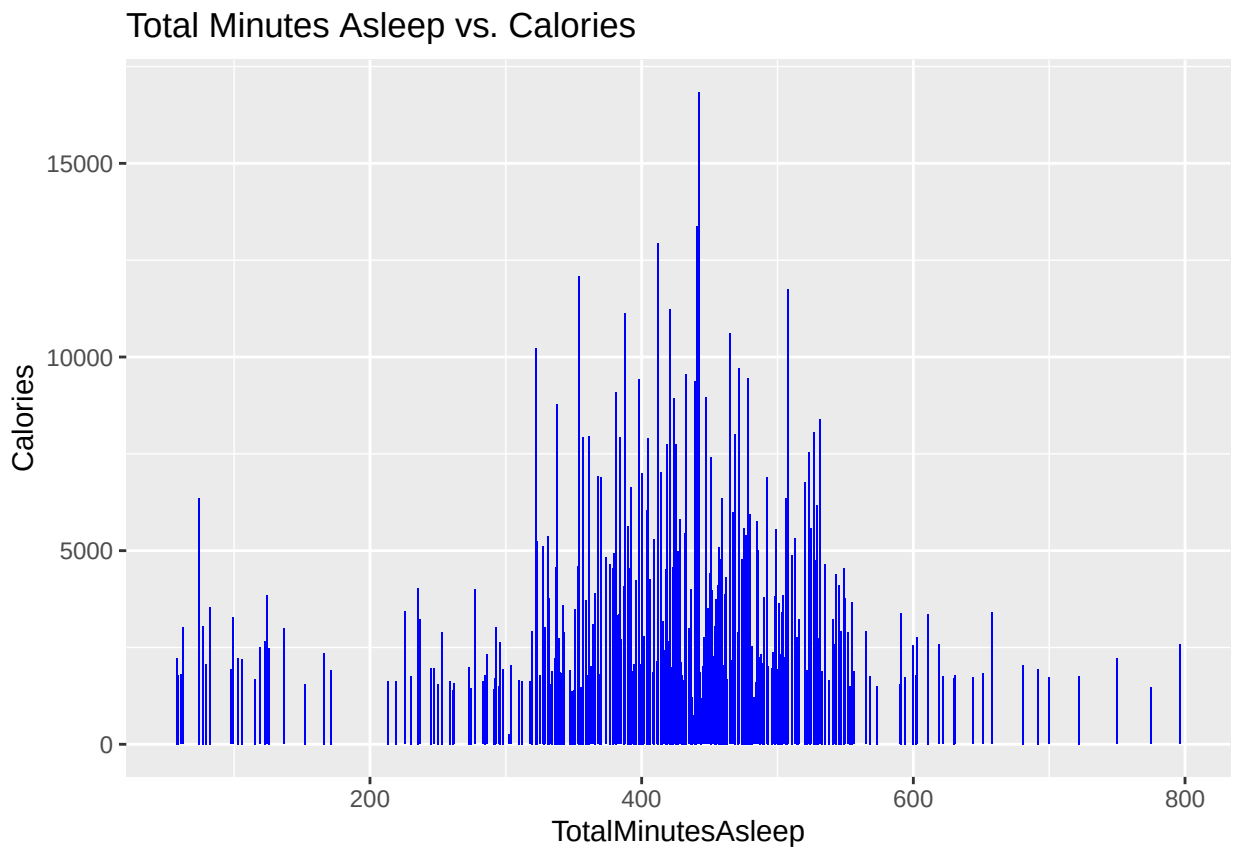
```
ggplot(data = clean_cals_sleep) + geom_point(mapping = aes(x=TotalTimeInBed, y=TotalMinutesAsleep)) + g
```

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



Plotting Sleep against Calories:

```
ggplot(data = clean_cals_sleep, aes(x=TotalMinutesAsleep, y=Calories)) + geom_col(fill='blue') + labs(t
```



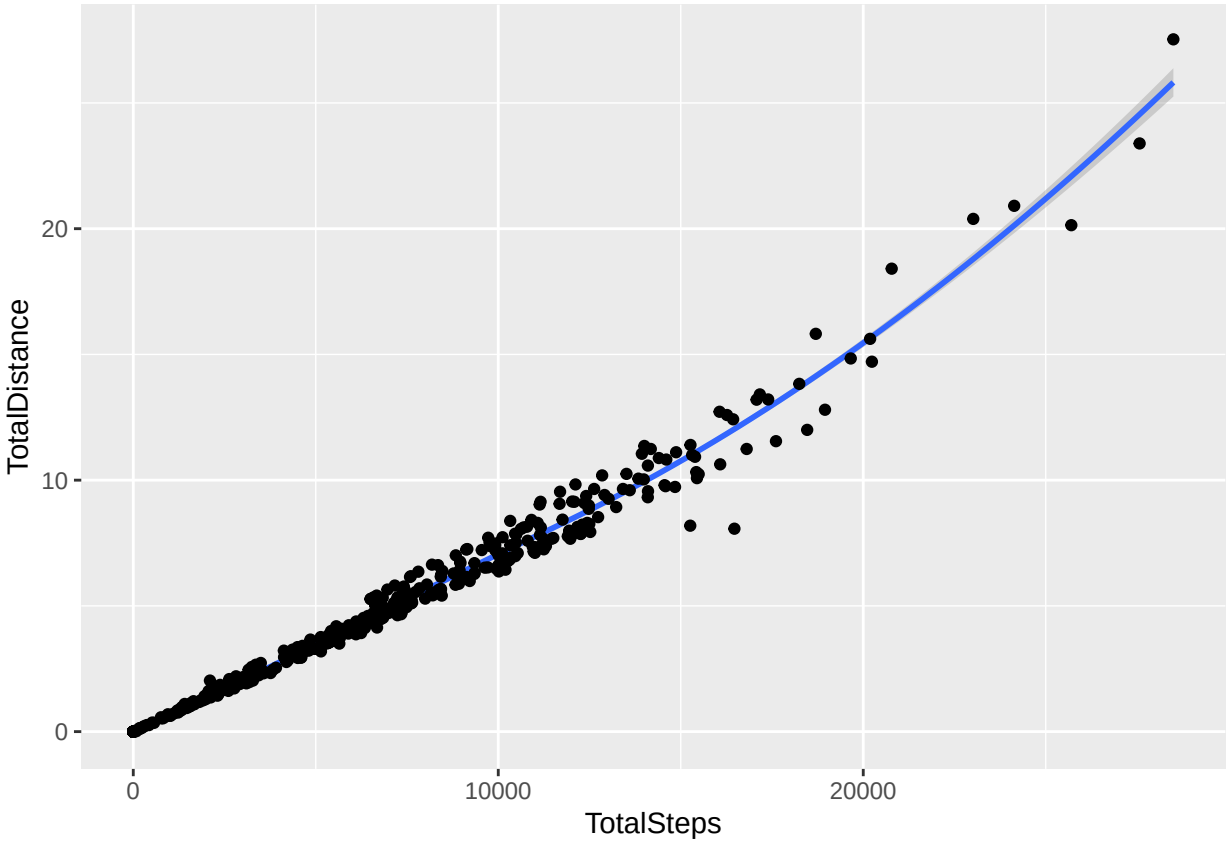
In the above graph, we can see that the distribution resembles a bell-shaped curve. This could be an indication that too little or too much sleep affects the energy levels of the people, who burn lesser calories, compared to the people who sleep in moderation – the values in the middle of the range; they burn more calories effectively.

But this information alone might not be sufficient.

We can also look at the relationship between Total Steps and Calories. Let us cross check the relationship between total steps and total distance before we start.

```
ggplot(data = daily_activity) + geom_smooth(mapping = aes(x=TotalSteps, y=TotalDistance)) + geom_point(
```

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

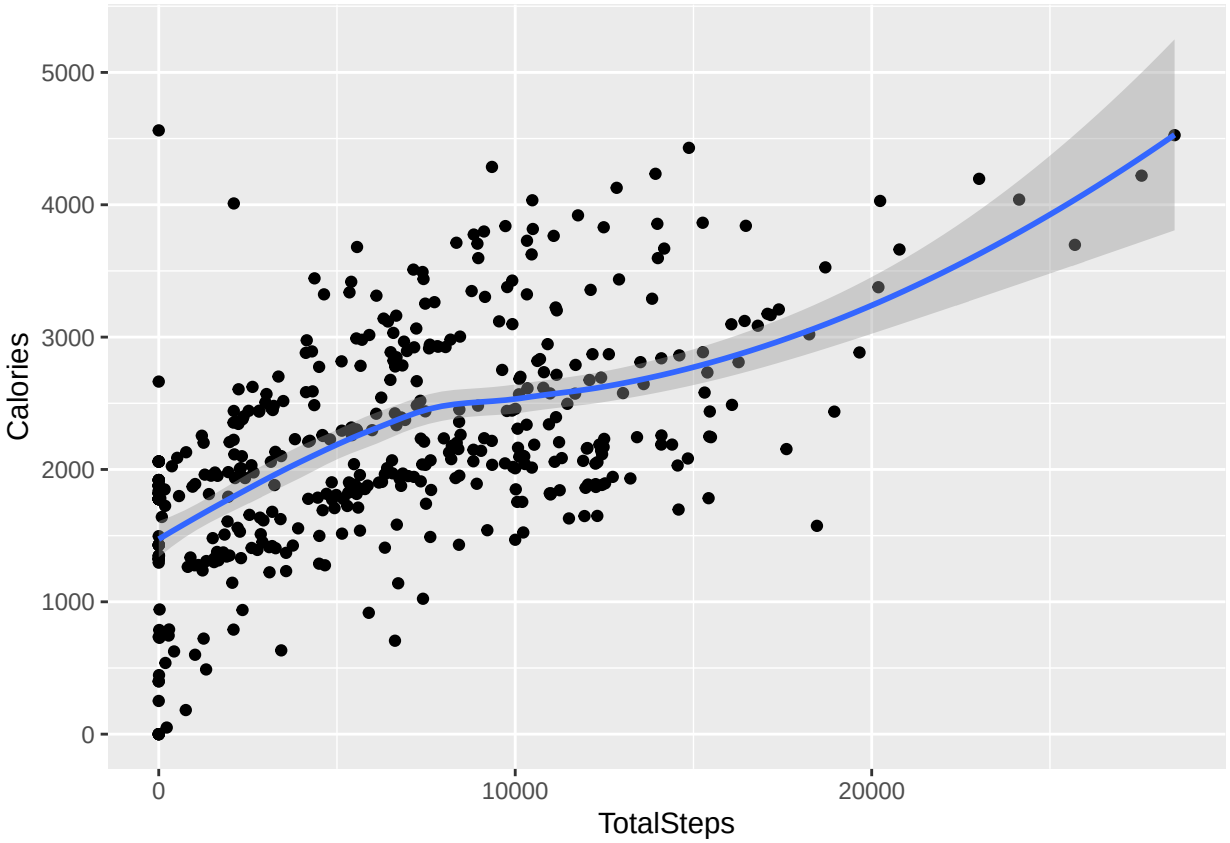


As expected. the relationship is linear – the more the steps, the more the distance covered. This is important to cover, as the distance covered could be a result of driving, or travelling in vehicles where people don't actually cover steps.

Plotting Total Steps and Calories:

```
ggplot(data = daily_activity) + geom_point(mapping = aes(x=TotalSteps, y=Calories)) + geom_smooth(mappi
```

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



The relationship is essentially linear in this case as well, implying that more steps correspond to more calories burnt.

How could these trends apply to Bellabeat customers? Through analyzing the data, we arrived at the following findings:

- **More Activity:** Lightly active, fairly active, and very active users burn more calories with an increase in their activity duration, while sedentary users burn lesser calories the longer they remain sedentary.
- **Sleep Duration:** When people sleep too little or too less, they burn lesser calories, than when people sleep in moderation.
- **Daily Steps:** When people have a higher record of total steps, they burn more calories.

Bellabeat users may be trying to find a balance, in terms of how active they should be, how much they should sleep, and how many steps they should walk in a day, in order to burn more calories.

How could these trends help influence Bellabeat marketing strategy? Some recommendations based on the above findings are as follows:

- Users can be provided with the option to set goals for how many calories they want to burn over a time period, say, in a month, and have a customized plan set out.
- By balancing a lack of one factor by compensating in another, while also ensuring that users are comfortable, and maintain their overall health in all aspects of:
  activity levels, sleep duration and daily steps, users can burn more calories to reach their goal efficiently.

- Bellabeat can send push notifications to keep users on track with their set objectives, and send gentle warnings to help them regulate their activity. For instance, if a user has been sedentary for too long, they can be alerted with a message that would make them add to their activity levels for the day.

By exploring the above strategies, Bellabeat can provide services that satisfy their users, and sustain the company's purpose while continuing to grow in the market.