BellaBeat Women Fitness Analysis

Kunal

2024-08-13

Company Description

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

Analysis Key Points

- 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 directive

Identify potential opportunities for growth and recommendations for the Bellabeat marketing strategy improvement based on trends in smart device usage.

Loading 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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
```

Importing Dataset

```
daily_activities <- read.csv("Daily Activities.csv")
hourly_intensity <- read.csv("Hourly Intensity.csv")
hourly_calories <- read.csv("Hourly Calories.csv")
sleep_data <- read.csv("Sleep Data.csv")
weight_log <- read.csv("Weight Log.csv")
```

Formatting Data Type

```
# Daily Activities
daily_activities$ActivityDate = as.POSIXct(daily_activities$ActivityDate, format = "%m-%d-%Y", tz=Sys.ti
daily_activities$Date = format(daily_activities$ActivityDate, format = "%m/%d/%Y")

# Hourly Intensity
hourly_intensity$ActivityHour = as.POSIXct(hourly_intensity$ActivityHour, format = "%m/%d/%Y %I:%M:%S %
hourly_intensity$Date = format(hourly_intensity$ActivityHour, format = "%m/%d/%Y")
hourly_intensity$Time = format(hourly_intensity$ActivityHour, format = "%H:%M:%S")

# Hourly Calorie
hourly_calories$ActivityHour = as.POSIXct(hourly_calories$ActivityHour, format = "%m/%d/%Y %I:%M:%S %p"
hourly_calories$Date = format(hourly_calories$ActivityHour, format = "%m/%d/%Y")
hourly_calories$Time = format(hourly_calories$ActivityHour, format = "%H:%M:%S")

# Sleep Data
sleep_data$SleepDay = as.POSIXct(sleep_data$SleepDay, format = "%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone()
sleep_data$Date = format(sleep_data$SleepDay, format = "%m/%d/%Y")

# Weight Log
weight_log$Date = as.POSIXct(weight_log$Date, format = "%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
weight_log$Date = format(weight_log$Date, format = "%m/%d/%Y")
```

Number of IDs

```
n_distinct(daily_activities$Id)
```

```
## [1] 36
```

```
n_distinct(hourly_calories$Id)
```

```
## [1] 36
```

```
n_distinct(hourly_intensity$Id)
```

```
## [1] 36
```

```r
n_distinct(sleep_data$Id)
```

```
## [1] 25
```

```r
n_distinct(weight_log$Id)
```

```
## [1] 14
```

Length of Dataset

```r
nrow(daily_activities)
```

```
## [1] 1399
```

```r
nrow(hourly_calories)
```

```
## [1] 46185
```

```r
nrow(hourly_intensity)
```

```
## [1] 46185
```

```r
nrow(sleep_data)
```

```
## [1] 999
```

```r
nrow(weight_log)
```

```
## [1] 999
```

Summarize Dataset

```r
daily_activities %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.: 3146   1st Qu.: 2.170   1st Qu.: 729.0
##  Median : 6999   Median : 4.950   Median :1057.0
##  Mean   : 7281   Mean   : 5.219   Mean   : 992.5
##  3rd Qu.:10544   3rd Qu.: 7.500   3rd Qu.:1244.0
##  Max.   :36019   Max.   :28.030   Max.   :1440.0
##  NA's   :2       NA's   :2        NA's   :2
```

- The mean total steps are 7281 which is not adequate for fitness. The company should try to influence customer to to increase their step count. Many data enters are 0 steps, so the the fitness device was not used on those days, while maximum steps are 36019.
- Average mean Sedentary time is 992.5 minutes or 16.5 hrs which is very high for general fitness.

```
hourly_calories %>%
  select(Calories) %>%
  summary()
```

```
##     Calories
##  Min.   : 42.00
##  1st Qu.: 62.00
##  Median : 80.00
##  Mean   : 95.76
##  3rd Qu.:106.00
##  Max.   :948.00
##  NA's   :2
```

- Hourly Calories Burned are averaged at 96. The calories burned while being sedentary are 42. Most of the people were active only for 1/4 of the day because 3rd Quarter and max values are 106 and 948.

```
hourly_intensity %>%
  select(TotalIntensity) %>%
  summary()
```

```
##  TotalIntensity
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median :  2.0
##  Mean   : 11.4
##  3rd Qu.: 15.0
##  Max.   :180.0
##  NA's   :2
```

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

```
##  TotalMinutesAsleep TotalTimeInBed  TotalSleepRecords
##  Min.   : 58.0      Min.   : 61.0   Min.   :1.000
##  1st Qu.:361.0      1st Qu.:403.0   1st Qu.:1.000
##  Median :433.0      Median :463.0   Median :1.000
##  Mean   :419.5      Mean   :458.6   Mean   :1.119
##  3rd Qu.:490.0      3rd Qu.:526.0   3rd Qu.:1.000
##  Max.   :796.0      Max.   :961.0   Max.   :3.000
##  NA's   :586        NA's   :586     NA's   :586
```

- The average sleep 420 minutes or 7 hours. While Total Time in Bed is almost 40 minutes extra.

```r
weight_log %>%
  select(WeightKg,
         BMI) %>%
  summary()
```

```
##     WeightKg          BMI
##  Min.   : 52.6   Min.   :21.45
##  1st Qu.: 61.5   1st Qu.:24.00
##  Median : 62.5   Median :24.39
##  Mean   : 72.5   Mean   :25.37
##  3rd Qu.: 85.3   3rd Qu.:25.59
##  Max.   :133.5   Max.   :47.54
##  NA's   :899     NA's   :899
```

Insights from Dataset Visualization

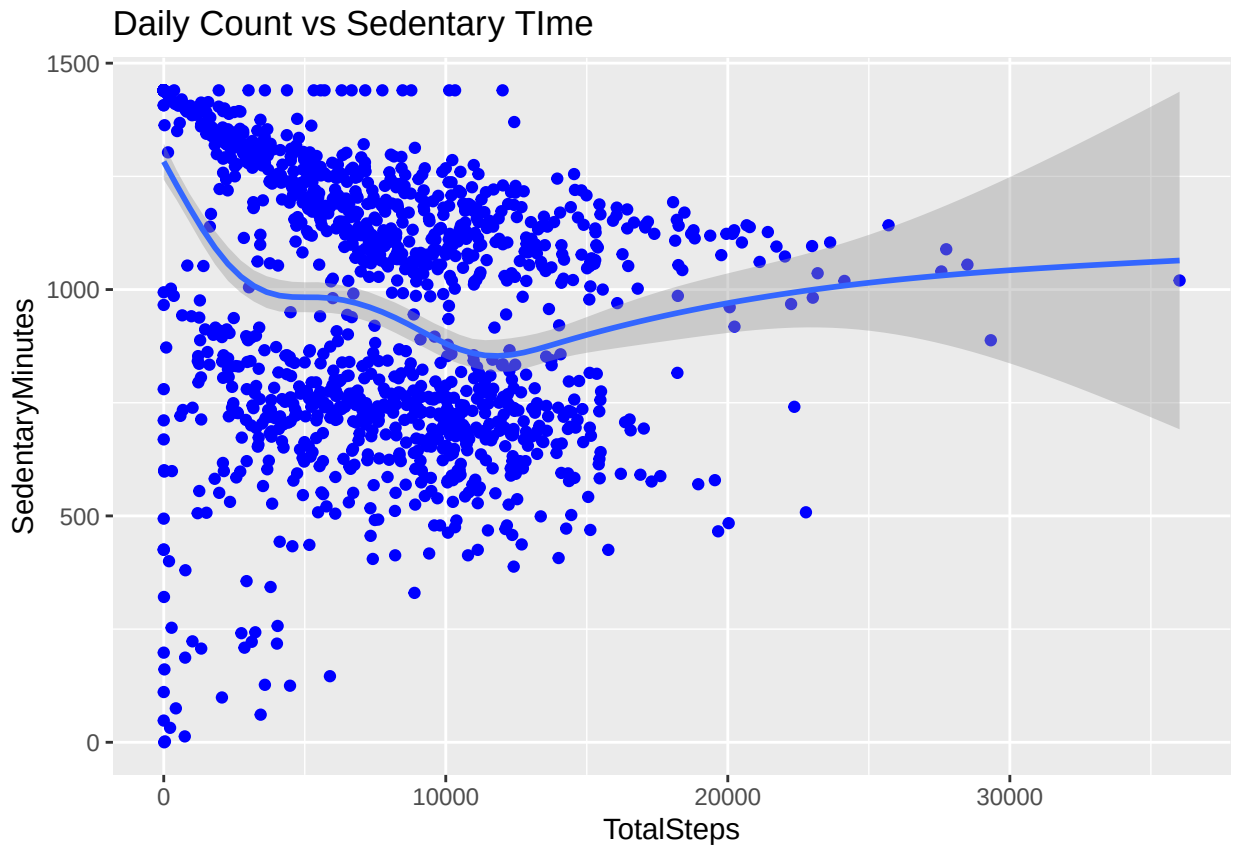
Visualization from Daily Activities

```r
ggplot(data=daily_activities, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point(color="blue") + geom_s
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- There are two Groups of customers have high sedentary time and another one with lower sedentary time. Both Groups are able to perform similiarly while some customers have very high steps count due to high intensity hours.

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

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

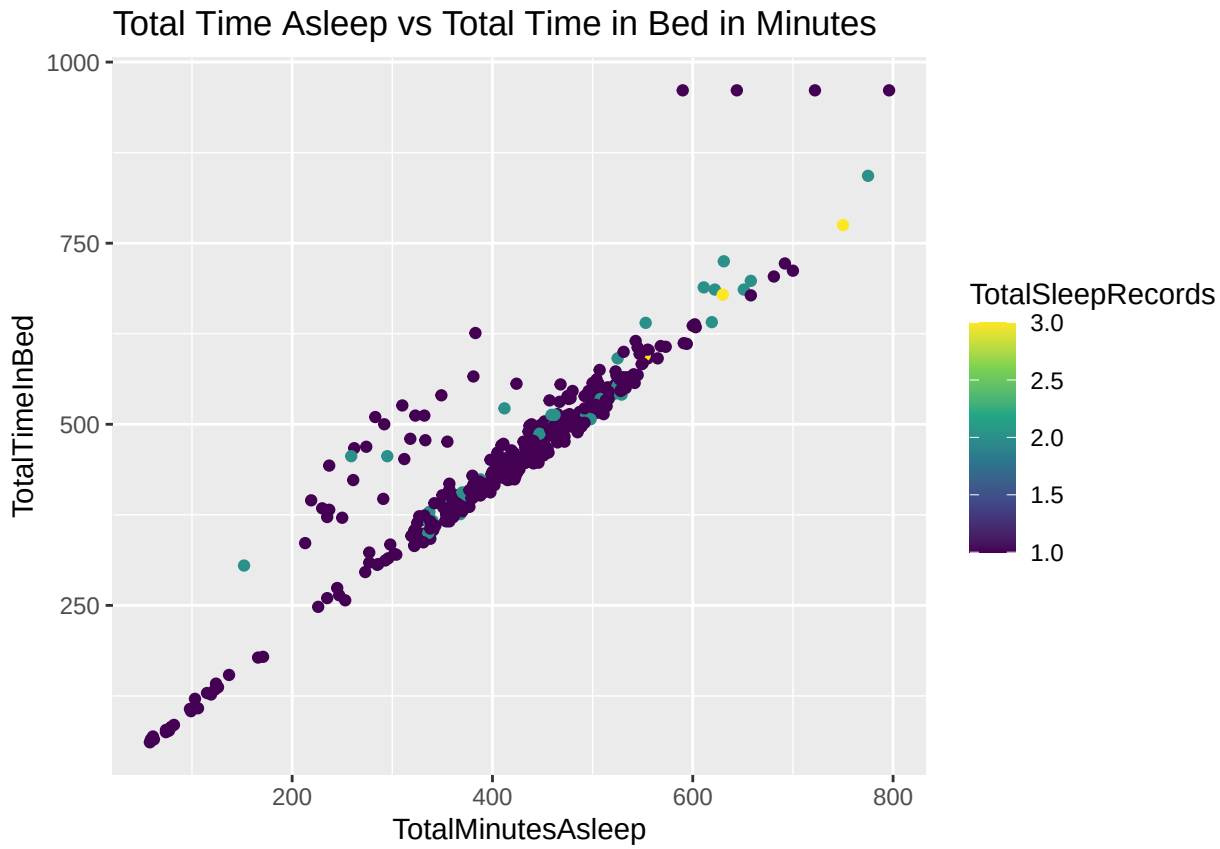
```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

- It is expected that Calories burned would increase with increasing step count. However for the same Step Count, there is variation of Calorie Burned.

```
ggplot(data=sleep_data) +   geom_point(mapping = aes(x=TotalMinutesAsleep, y=TotalTimeInBed, color=Tota
```

```
## Warning: Removed 586 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- Customers who tend to sleep in intervals also tend to sleep more compared to customers who sleep in one intervals.

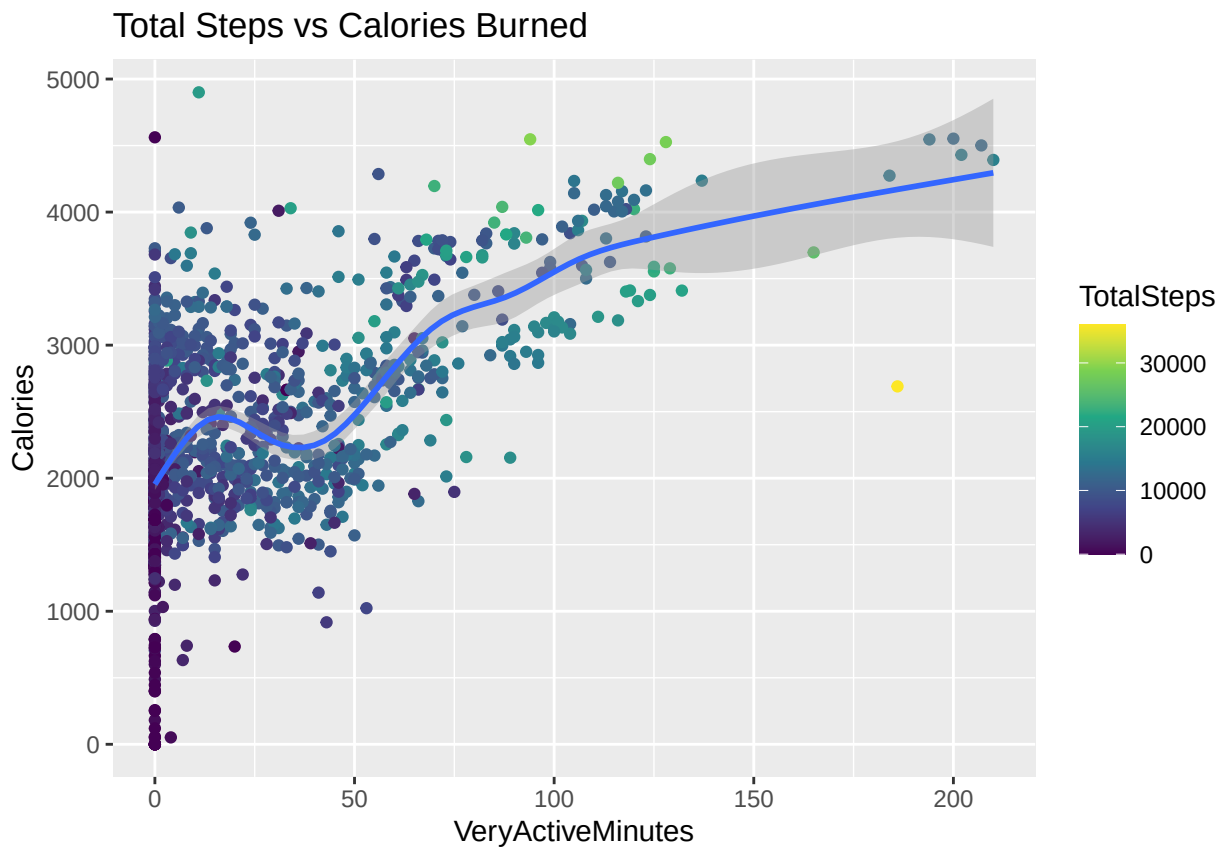
```
ggplot(data=daily_activities, aes(x=VeryActiveMinutes, y=Calories, color=TotalSteps)) + geom_point() + g
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- Customers who have high very active intervals also have high calories burned and tend to have high step count.

```
ggplot(data=daily_activities, aes(x=SedentaryMinutes, y=Calories, color=TotalSteps)) + geom_point() + g
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

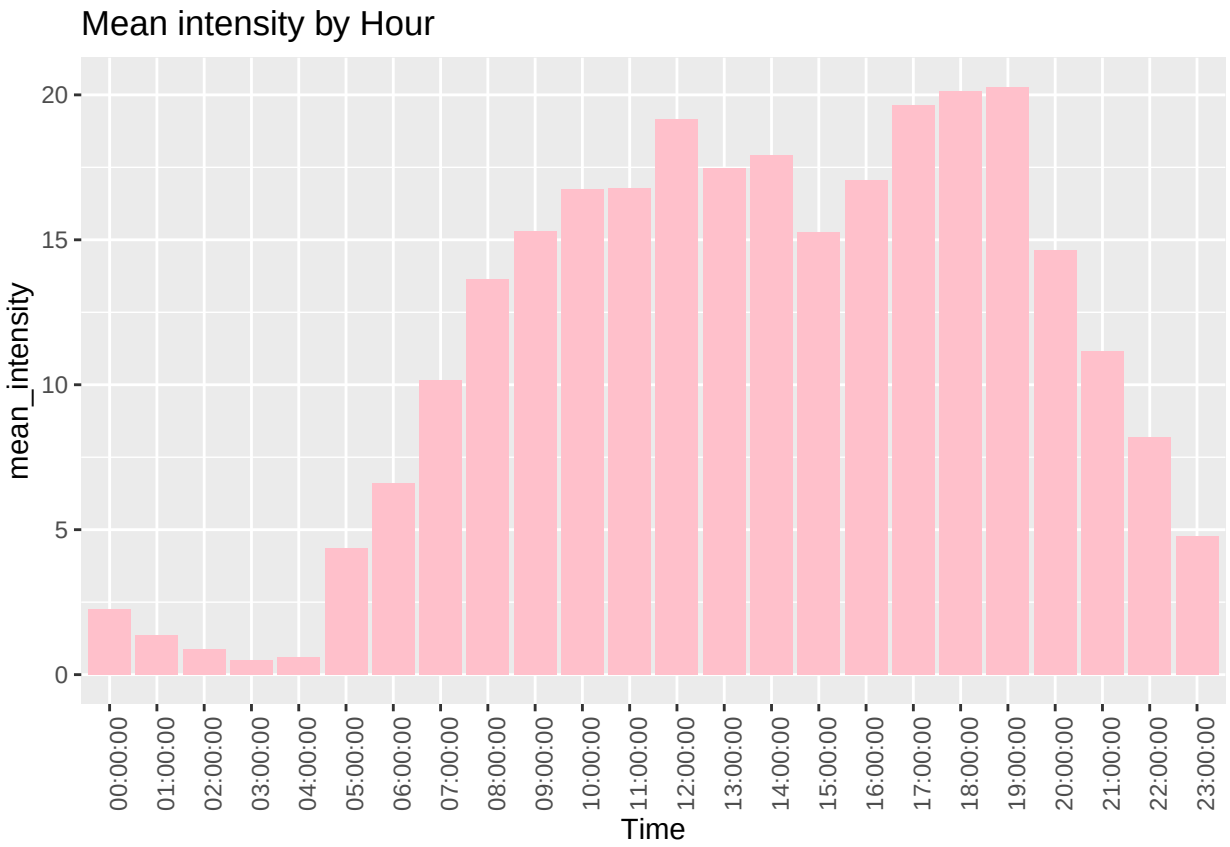
```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

- Customers with 8-20 hrs of sedentary tend to have an active lifestyle.

```
intensity_by_hour <- hourly_intensity %>%
  group_by(Time) %>%
  drop_na() %>%
  summarize(mean_intensity = mean(TotalIntensity))

ggplot(data=intensity_by_hour, aes(x=Time, y=mean_intensity)) + geom_histogram(stat = "identity", fill=
```

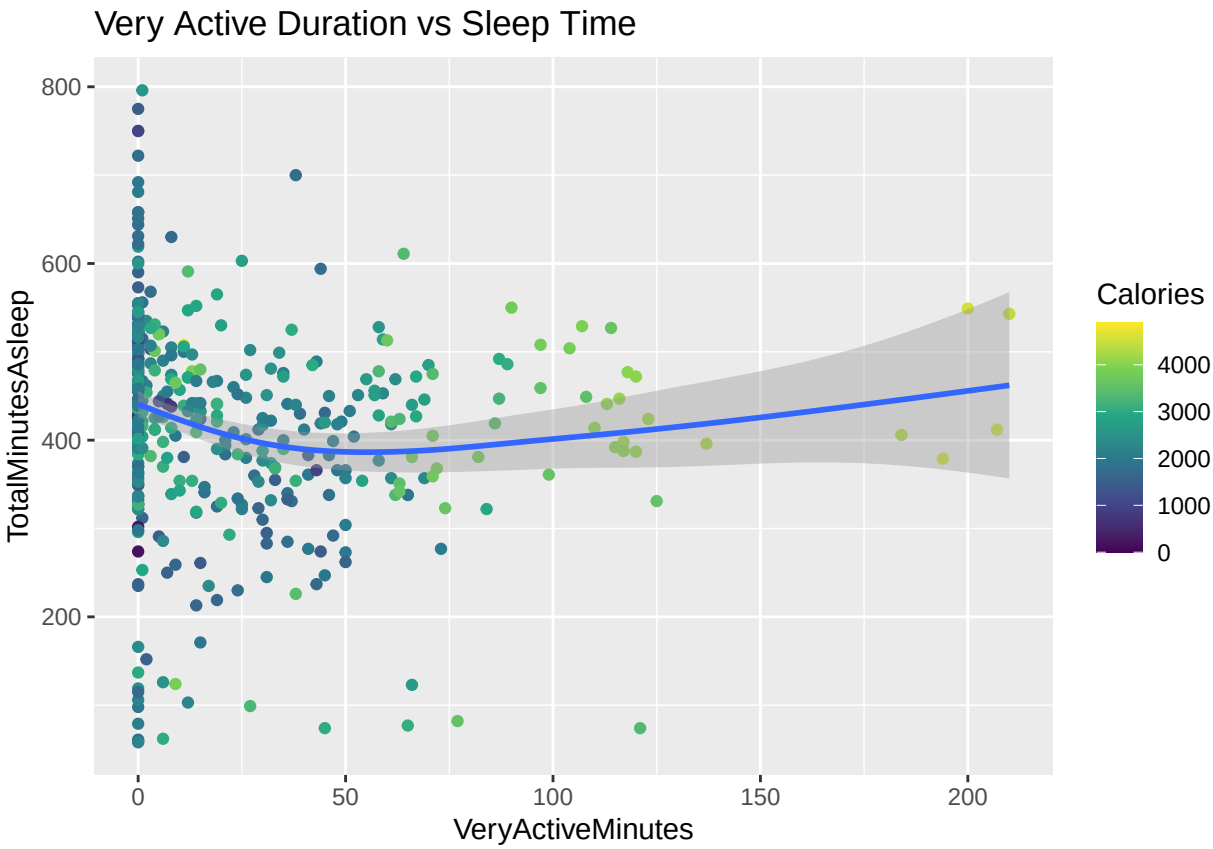
```
## Warning in geom_histogram(stat = "identity", fill = "pink"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



- Customers tends to have high intensity between 9 AM to 7 PM.
- Customer's Active lifestyle starts from 5-8 AM depending on their career and ends around 8-11 PM.

```
combined_activities_sleep <- merge(daily_activities, sleep_data, by= c("Id", "Date"))

ggplot(data=combined_activities_sleep, aes(x=VeryActiveMinutes, y=TotalMinutesAsleep, color=Calories))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 1172 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?

## Warning: Removed 1172 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- Customers with high intensity intervals and calorie count also tend to Sleep more to recover their fatigue.
- Majority of customers have low intervals of high intensity tend to sleep more.

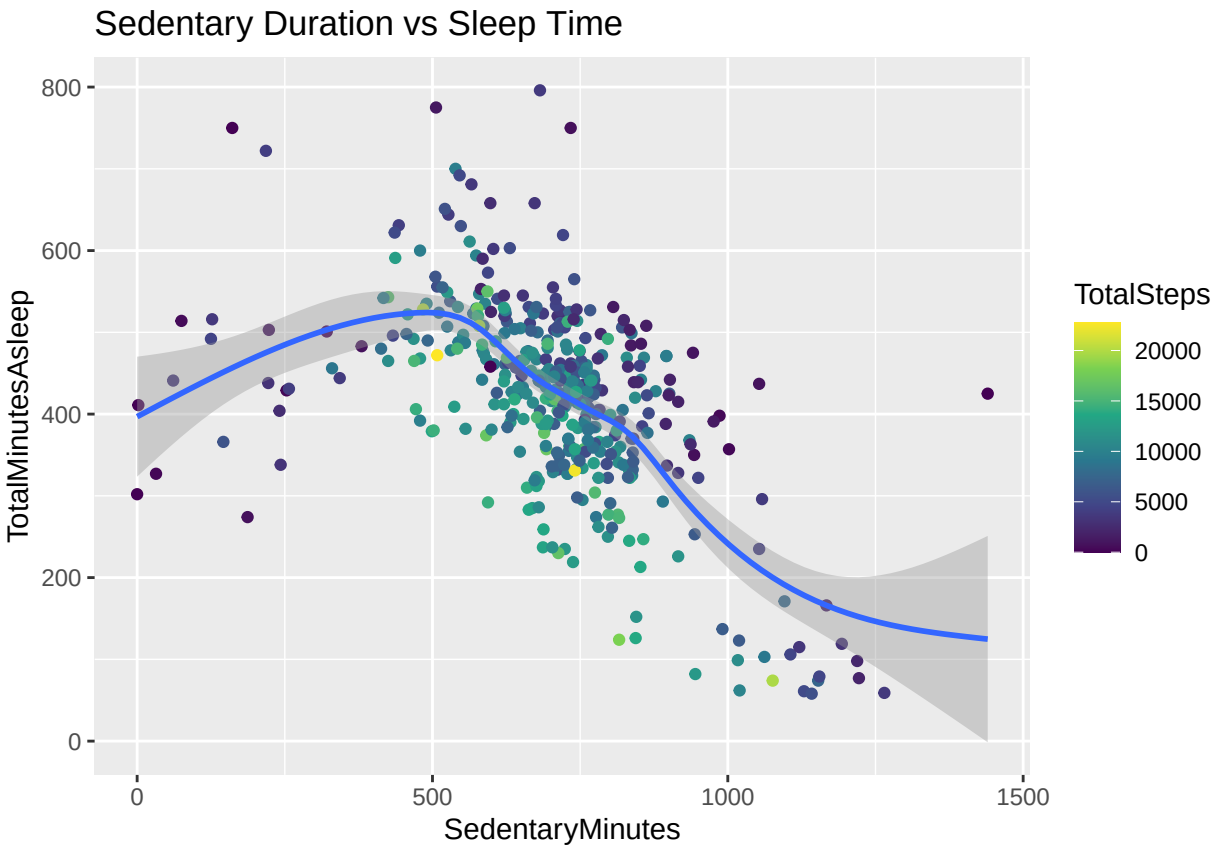
```
ggplot(data=combined_activities_sleep, aes(x=SedentaryMinutes, y=TotalMinutesAsleep, color=TotalSteps))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1172 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 1172 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- Customers who tend to have high sedentary time, also tend to sleep less because there is no cause of fatique.
- Customers with high step count also tend to sleep less compared to customers with low step count.

```
combined_activities_weight <- merge(daily_activities, weight_log, by= "Id")

ggplot(data=combined_activities_weight, aes(x=TotalSteps, y=Calories, color=WeightKg)) + geom_point() +
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1798 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 1798 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

- Customers who tend to have high sedentary time, also tend to sleep less because there is no cause of fatique.
- Customers with high step count also tend to sleep less compared to customers with low step count.