Google Case study: Bellabeat

Sen

2023-05-25

Installing and loading common packages and libraries

```
install.packages('tidyverse')
```

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

```
install.packages("dplyr")
```

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

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

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

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

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

```
install.packages("janitor")
```

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

```
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 errors
```

```
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Loading the CSV files

```
daily_activity <- read.csv("dailyActivity_merged.csv")
sleep_day <- read.csv("sleepDay_merged.csv")
calories_hourly <- read.csv("hourlyCalories_merged.csv")
intensities_hourly <- read.csv("hourlyIntensities_merged.csv")
weight_loginfo <- read.csv("weightLogInfo_merged.csv")
```

Getting the unique participants each dataframe

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

```
n_distinct(calories_hourly$Id)
```

```
## [1] 33
```

```
n_distinct(intensities_hourly$Id)
```

```
## [1] 33
```

```
n_distinct(weight_loginfo$Id)
```

```
## [1] 8
```

summarise data using summary function:

```
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes, Calories) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes    Calories
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :   0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:1828
##  Median : 7406   Median : 5.245   Median :1057.5   Median :2134
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :4900
```

```
daily_activity %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%
  summary()
```

```
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0
##  Median :  4.00    Median :  6.00      Median :199.0
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0
```

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

```
##     Calories
##  Min.   : 42.00
##  1st Qu.: 63.00
##  Median : 83.00
##  Mean   : 97.39
##  3rd Qu.:108.00
##  Max.   :948.00
```

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

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

```
weight_loginfo %>%
  select(WeightKg, BMI) %>%
  summary()
```

```
##     WeightKg           BMI
##  Min.   : 52.60   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:23.96
##  Median : 62.50   Median :24.39
##  Mean   : 72.04   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:25.56
##  Max.   :133.50   Max.   :47.54
```

```
intensities_hourly %>%
  select(ActivityHour, TotalIntensity, AverageIntensity) %>%
  summary()
```

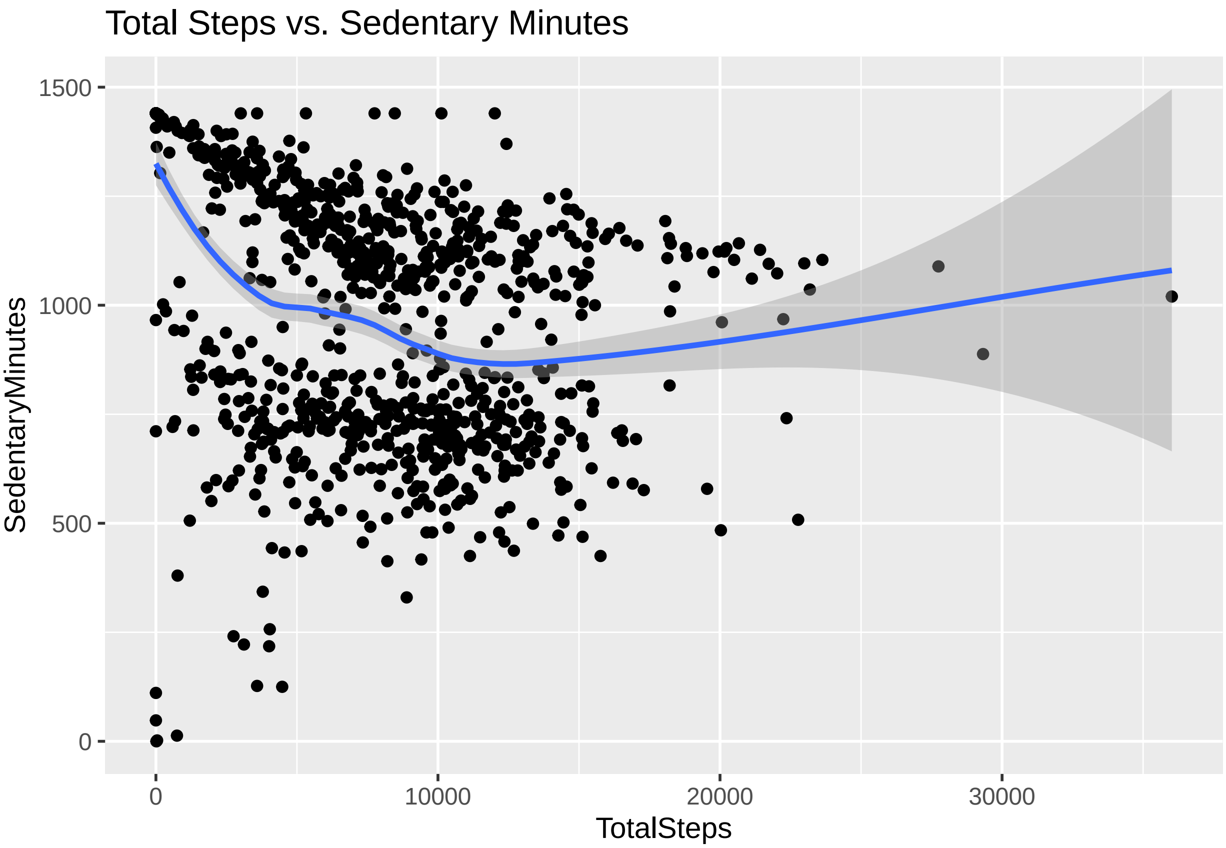
```
##  ActivityHour       TotalIntensity   AverageIntensity
##  Length:22099       Min.   :  0.00   Min.   :0.0000
##  Class :character   1st Qu.:  0.00   1st Qu.:0.0000
##  Mode  :character   Median :  3.00   Median :0.0500
##                     Mean   : 12.04   Mean   :0.2006
##                     3rd Qu.: 16.00   3rd Qu.:0.2667
##                     Max.   :180.00   Max.   :3.0000
```

Plotting a few explorations

Relationship between Total Steps and Sedentary Minutes:

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) +
  geom_point() +
  geom_smooth() +
  labs(title="Total Steps vs. Sedentary Minutes")
```

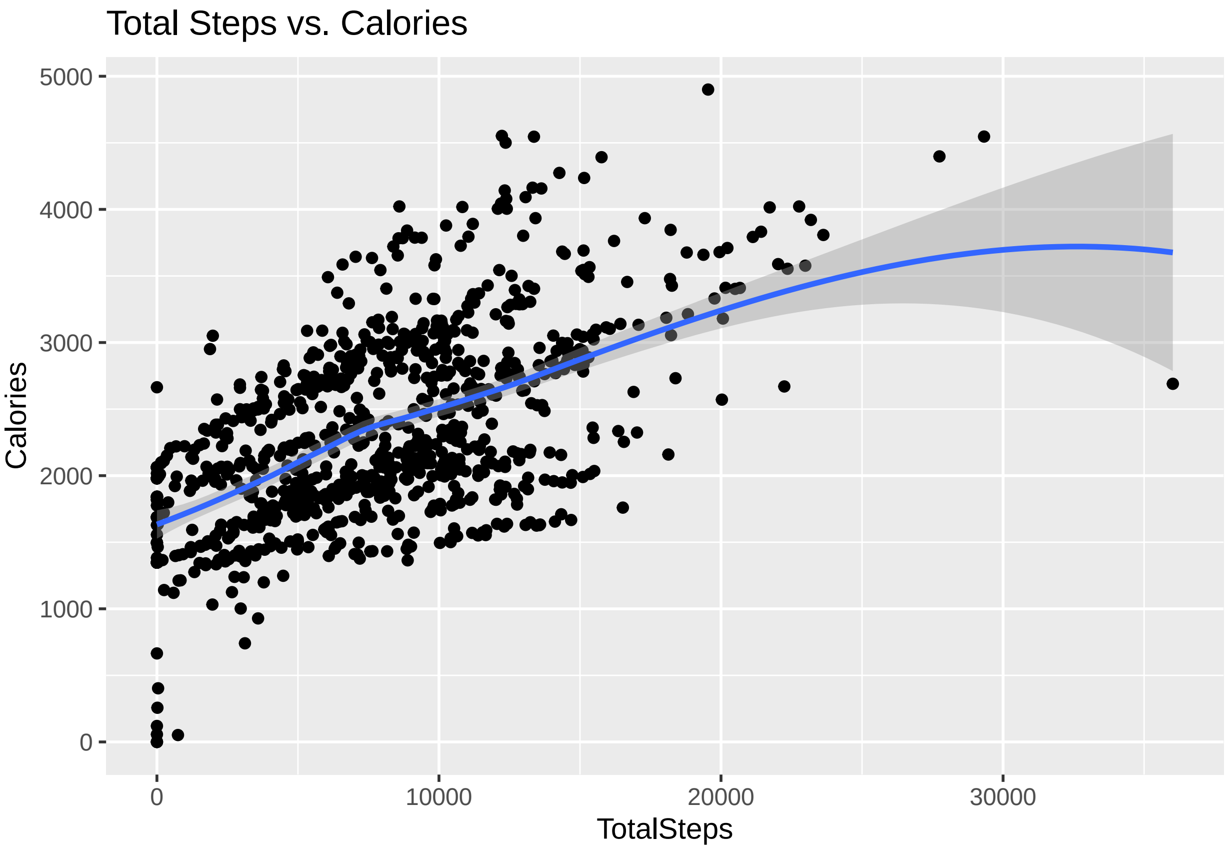
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Relationship between Total Steps and Calories?

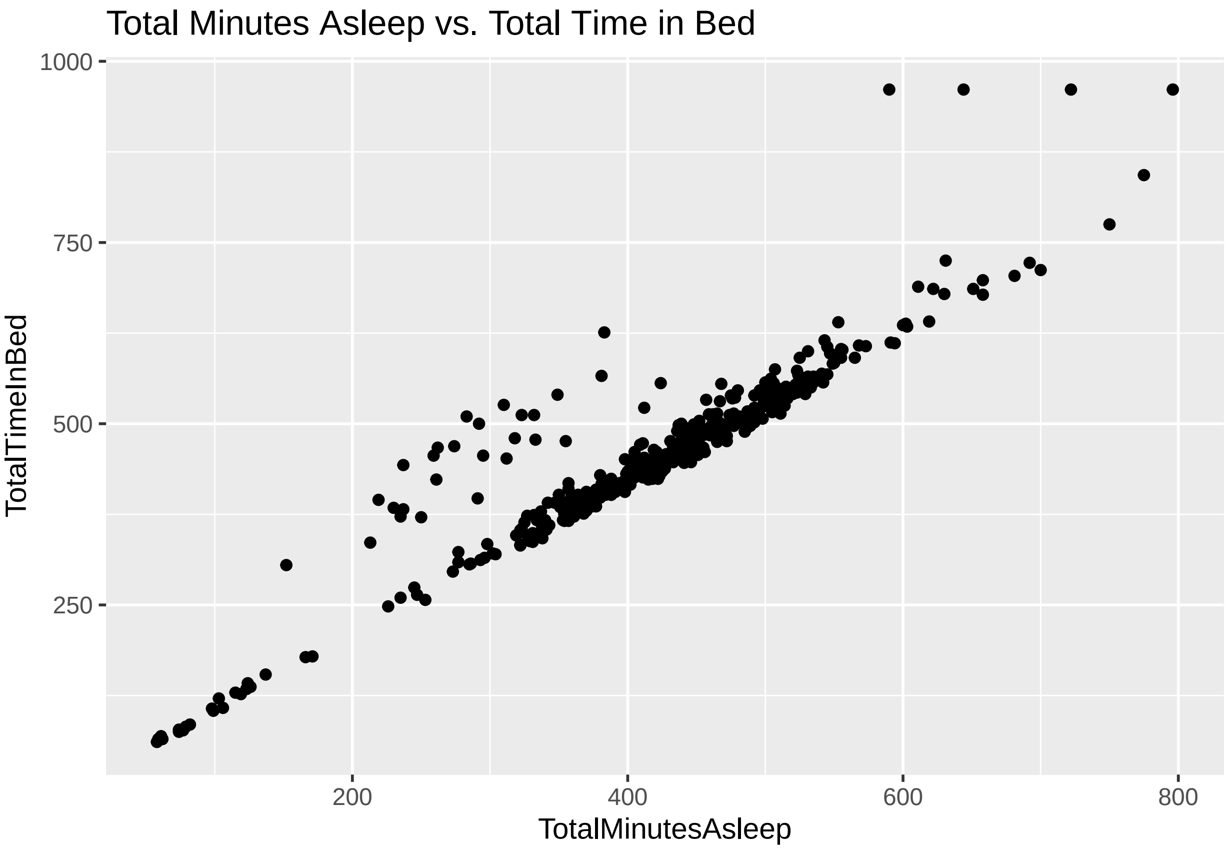
```
ggplot(data=daily_activity, aes(x=TotalSteps, y=Calories)) +
  geom_point() +
  geom_smooth() +
  labs(title="Total Steps vs. Calories")
```

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



Relationship between minutes asleep and time in bed? You might expect it to be almost completely linear - are there any unexpected trends?

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point() +
  labs(title="Total Minutes Asleep vs. Total Time in Bed")
```



Convert date time format.

```
intensities_hourly$ActivityHour=as.POSIXct(intensities_hourly$ActivityHour, format="%m/%d/%Y %I:%M:%S %
intensities_hourly$time <- format(intensities_hourly$ActivityHour, format = "%H:%M:%S")
intensities_hourly$date <- format(intensities_hourly$ActivityHour, format = "%m/%d/%y")
```

```
daily_activity$ActivityDate=as.POSIXct(daily_activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone(
daily_activity$date <- format(daily_activity$ActivityDate, format = "%m/%d/%y")
```

```
sleep_day$SleepDay=as.POSIXct(sleep_day$SleepDay, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
sleep_day$date <- format(sleep_day$SleepDay, format = "%m/%d/%y")
```

Merging these two datasets together

```
combined_data <- merge(sleep_day, daily_activity, by = c('Id', 'date'))
head(combined_data)
```

```
##           Id     date   SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 04/12/16 2016-04-12                 1                327
## 2 1503960366 04/13/16 2016-04-13                 2                384
## 3 1503960366 04/15/16 2016-04-15                 1                412
## 4 1503960366 04/16/16 2016-04-16                 2                340
## 5 1503960366 04/17/16 2016-04-17                 1                700
## 6 1503960366 04/19/16 2016-04-19                 1                304
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346   2016-04-12      13162          8.50            8.50
## 2            407   2016-04-13      10735          6.97            6.97
## 3            442   2016-04-15       9762          6.28            6.28
## 4            367   2016-04-16      12669          8.16            8.16
## 5            712   2016-04-17       9705          6.48            6.48
## 6            320   2016-04-19      15506          9.88            9.88
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.14                     1.26
## 4                        0               2.71                     0.41
## 5                        0               3.19                     0.78
## 6                        0               3.53                     1.32
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                2.83                       0                29
## 4                5.04                       0                36
## 5                2.51                       0                38
## 6                5.03                       0                50
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  34                  209              726     1745
## 4                  10                  221              773     1863
## 5                  20                  164              539     1728
## 6                  31                  264              775     2035
```

Take a look at how many participants are in this data set.

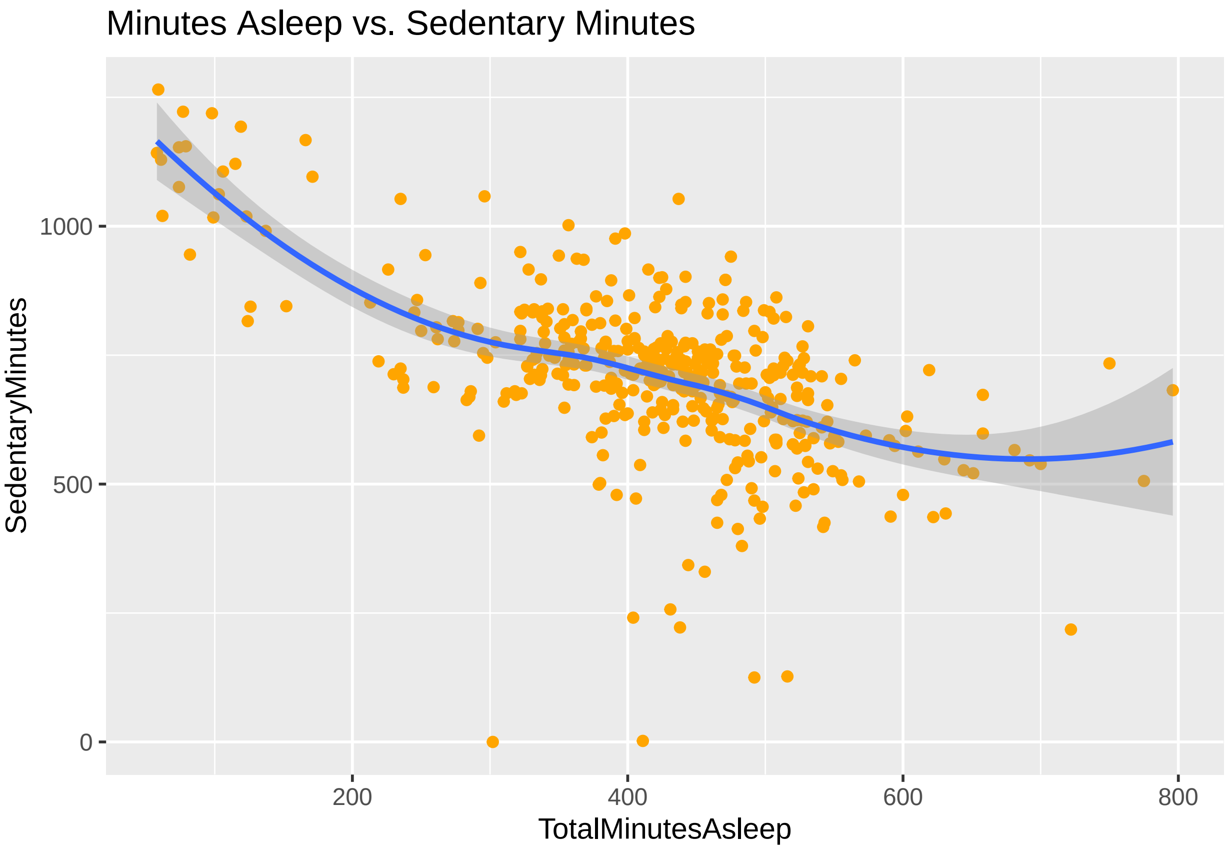
```
n_distinct(combined_data$Id)
```

```
## [1] 24
```

Relationship between Total Minutes Asleep and Sedentry Minutes.

```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=SedentaryMinutes)) +
geom_point(color='orange') +
  geom_smooth() +
  labs(title="Minutes Asleep vs. Sedentary Minutes")
```

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



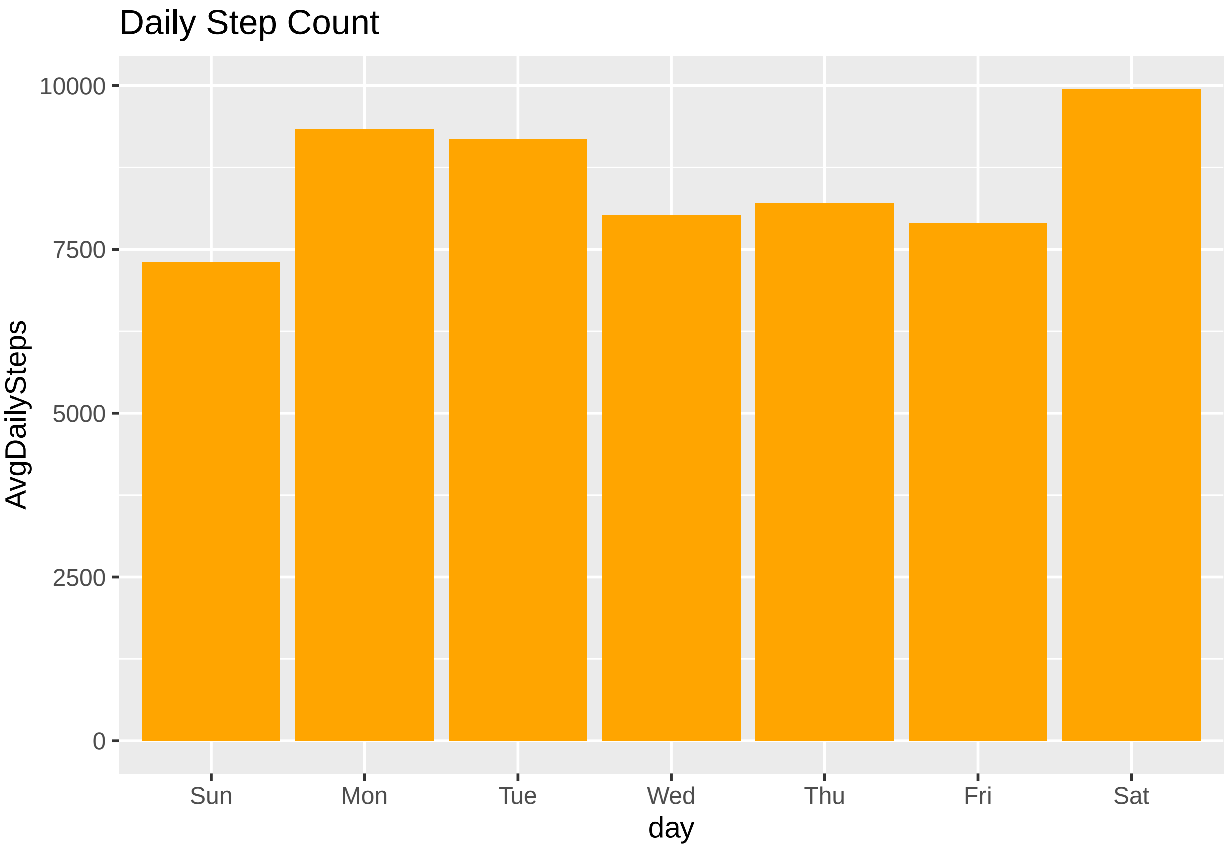
aggregate data by day of week to summarize averages

```
aggregate_data <- mutate(combined_data, day = wday(SleepDay, label = TRUE))
summary_activity_sleep_data <- aggregate_data %>%
  group_by(day) %>%
  summarise(AvgDailySteps = mean(TotalSteps),
            AvgAsleepMinutes = mean(TotalMinutesAsleep),
            AvgAwakeTimeInBed = mean(TotalTimeInBed),
            AvgSedentaryMinutes = mean(SedentaryMinutes),
            AvgLightlyActiveMinutes = mean(LightlyActiveMinutes),
            AvgFairlyActiveMinutes = mean(FairlyActiveMinutes),
            AvgVeryActiveMinutes = mean(VeryActiveMinutes),
            AvgCalories = mean(Calories))
head(summary_activity_sleep_data)
```

```
## # A tibble: 6 x 9
##   day   AvgDailySteps AvgAsleepMinutes AvgAwakeTimeInBed AvgSedentaryMinutes
##   <ord>         <dbl>            <dbl>             <dbl>               <dbl>
## 1 Sun           7298.             453.              504.                688.
## 2 Mon           9340.             419.              456.                718.
## 3 Tue           9183.             405.              443.                740.
## 4 Wed           8023.             435.              470.                714.
## 5 Thu           8205.             402.              436.                701.
## 6 Fri           7901.             405.              445.                743.
## # i 4 more variables: AvgLightlyActiveMinutes <dbl>,
## #   AvgFairlyActiveMinutes <dbl>, AvgVeryActiveMinutes <dbl>, AvgCalories <dbl>
```

Plot data by day of week to summarize averages

```
ggplot(data = summary_activity_sleep_data, mapping = aes(x = day, y = AvgDailySteps)) +
geom_col(fill = "orange") + labs(title = "Daily Step Count")
```


Daily Step Count
AvgDailySteps
Sun
Mon
Tue
Wed
Thu
Fri
Sat
day