Google Data Analytics: Case Study 2 Documentation

2023-10-14

Load necessary packages

```
install.packages('janitor')
```

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

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

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

```
install.packages('skimr')
```

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

```
library(skimr)
install.packages('lubridate')
```

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

```
library(lubridate)
```

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

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

```
install.packages('ggplot2')
```

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

```
library(ggplot2)
install.packages('tidyverse')
```

```
## 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   1.1.3     v stringr 1.5.0
## v forcats 1.0.0     v tibble  3.2.1
```

```
## v purrr   1.0.2     v tidyr   1.3.0
## v readr   2.1.4
```

```
## -- 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
```

```
install.packages('formatR')
```

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

Upload csv

```
weightLog <- read.csv('weightLogInfo_merged.csv')
dailyActivity <- read.csv('dailyActivity_merged.csv')
dailySleep <- read.csv('sleepDay_merged.csv')
```

Check Formatting & Clean names

```
dailyActivity <- clean_names(dailyActivity)
dailySleep <- clean_names(dailySleep)
weightLog <- clean_names(weightLog)
```

Check structure of each data frame

```
str(dailyActivity) #date needs to be formatted as date not chr
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ activity_date            : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ total_steps              : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ total_distance           : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ tracker_distance         : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ logged_activities_distance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_distance     : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ moderately_active_distance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ light_active_distance    : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ sedentary_active_distance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_minutes      : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ fairly_active_minutes    : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ lightly_active_minutes   : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ sedentary_minutes        : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
str(dailySleep) #date needs to be formatted as date not chr
```

```
## 'data.frame':    413 obs. of  5 variables:
##  $ id                 : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ sleep_day          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM
##  $ total_sleep_records : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ total_minutes_asleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ total_time_in_bed   : int  346 407 442 367 712 320 377 364 384 449 ...
```

```
str(weightLog) #date needs to be formatted as date not chr
```

```
## 'data.frame':    67 obs. of  8 variables:
```

```
##  $ id              : num  1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ date            : chr  "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM" "4/21/
##  $ weight_kg       : num  52.6 52.6 133.5 56.7 57.3 ...
##  $ weight_pounds   : num  116 116 294 125 126 ...
##  $ fat             : int  22 NA NA NA NA 25 NA NA NA NA ...
##  $ bmi             : num  22.6 22.6 47.5 21.5 21.7 ...
##  $ is_manual_report: chr  "True" "True" "False" "True" ...
##  $ log_id          : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

```
# ismanuelreport needs to be logical not chr
```

Fix the format

```
dailyActivity$activity_date <- as.Date(dailyActivity$activity_date, '%m/%d/%y' )
dailySleep$sleep_day <- as.Date(dailySleep$sleep_day, '%m/%d/%y' )
weightLog$date <- parse_date_time(weightLog$date,'%m/%d/%y %H:%M:%S %p' )
weightLog$is_manual_report <- as.logical(weightLog$is_manual_report)
```

remove fat column

```
#remove fat column in weightLog
weightLog <- weightLog %>% select(-c(fat))
```

Add columns for day of week, total active hours & sedentary hours & BMI label.

```
dailyActivity$day_of_week <- wday(dailyActivity$activity_date, label = T, abbr = T) # labels day as a s
dailyActivity$total_active_hrs = round((dailyActivity$very_active_minutes+dailyActivity$fairly_active_mi
+dailyActivity$lightly_active_minutes)/60, digits=2)

dailySleep$hrs_asleep = round((dailySleep$total_minutes_asleep)/60, digits = 2)

dailySleep$time_taken_to_sleep = (dailySleep$total_time_in_bed - dailySleep$total_minutes_asleep)

#add correspond label for numerical bmi in a new column
weightLog <- weightLog %>% mutate(bmi2 = case_when(
  bmi > 24.9 ~'overweight',
  bmi < 18.5 ~'underweight',
  TRUE ~ 'healthy'
))
```

Remove 0s for calories and total active hrs

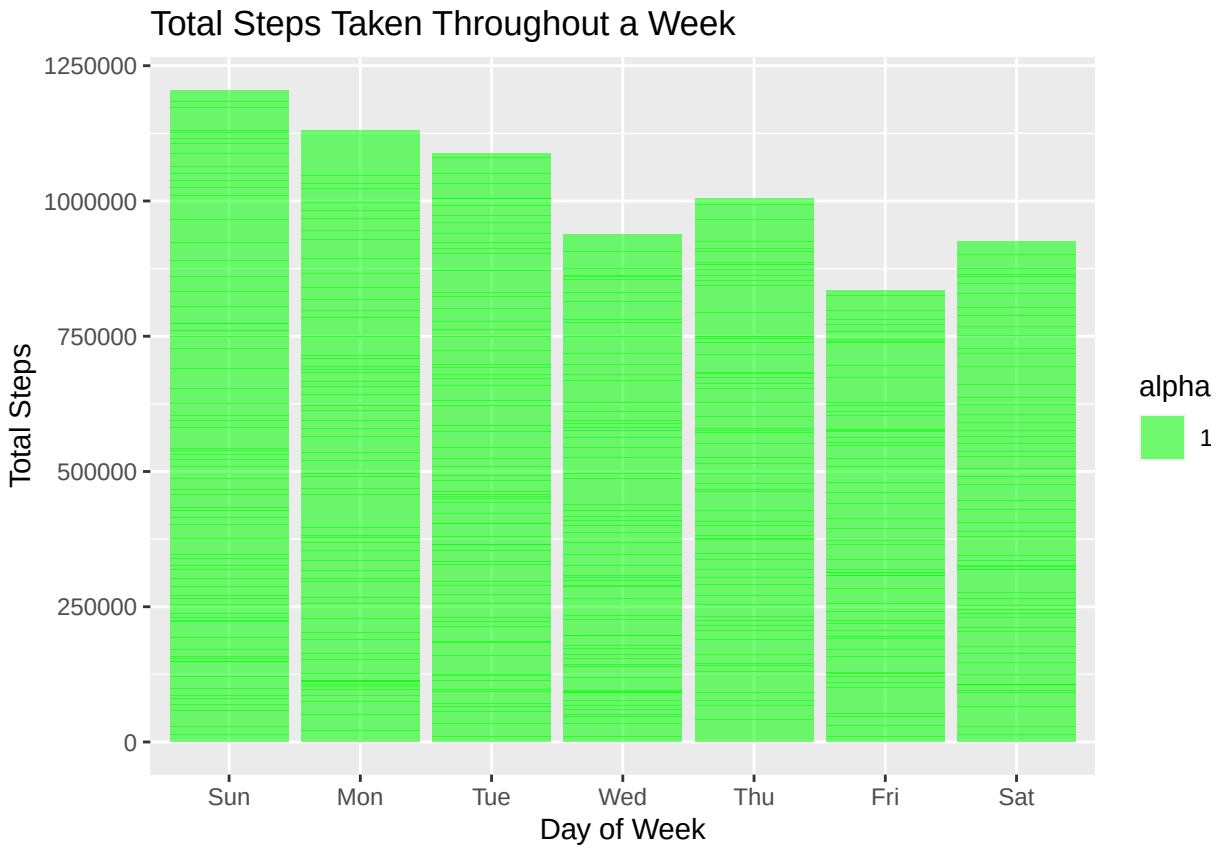
```
daily_activity_cleaned <- dailyActivity[!(dailyActivity$calories<=0),]
daily_activity_cleaned <- daily_activity_cleaned[!(daily_activity_cleaned$total_active_hrs<=0.00),]
```

Visualize

Days Active

Days active throughout the week

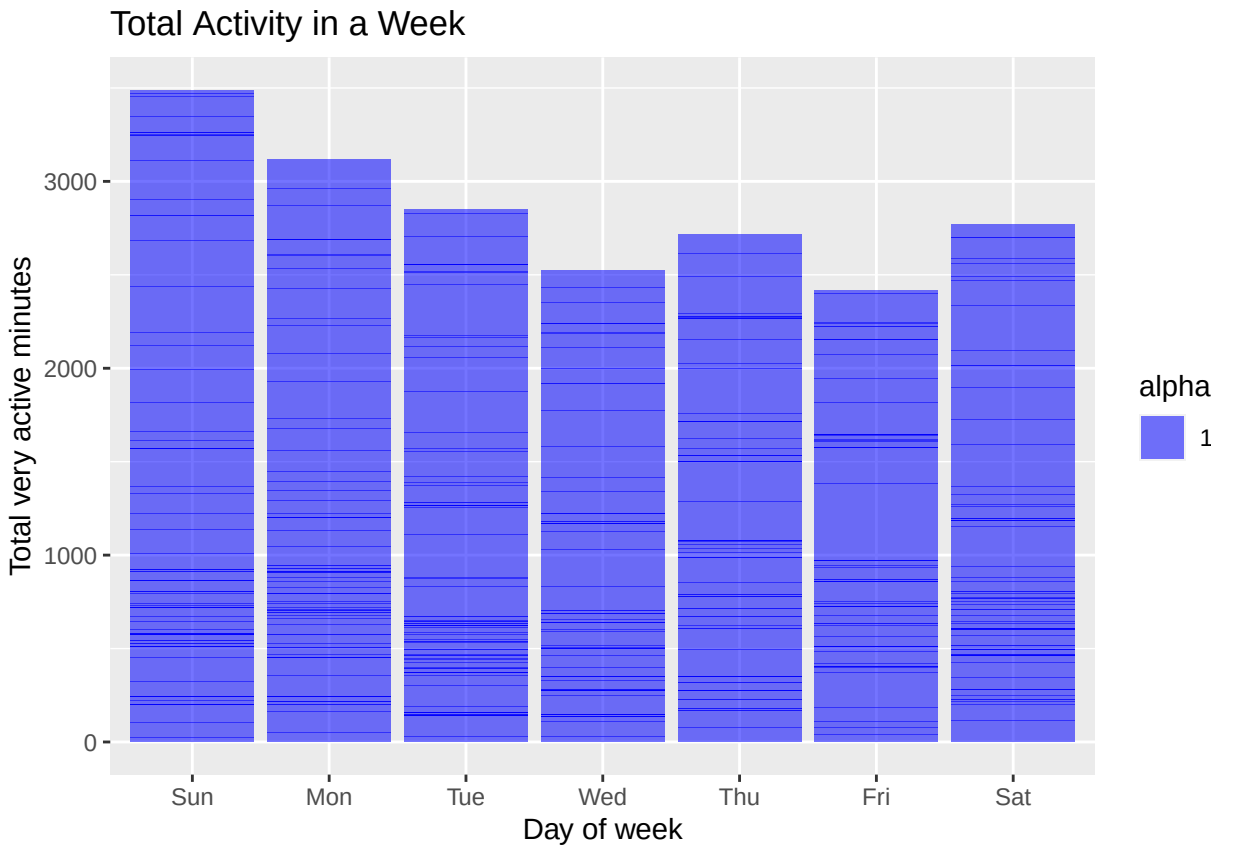
```
options(scipen = 999) #removes scientific notation
ggplot(data = daily_activity_cleaned) + aes(x=day_of_week,y= total_steps, alpha=1) +
  geom_col(fill = 'green') +
  labs(x= 'Day of Week', y= 'Total Steps', title = 'Total Steps Taken Throughout a Week')
```




```
ggsave('total_steps_week.png')
```

```
## Saving 6.5 x 4.5 in image
```

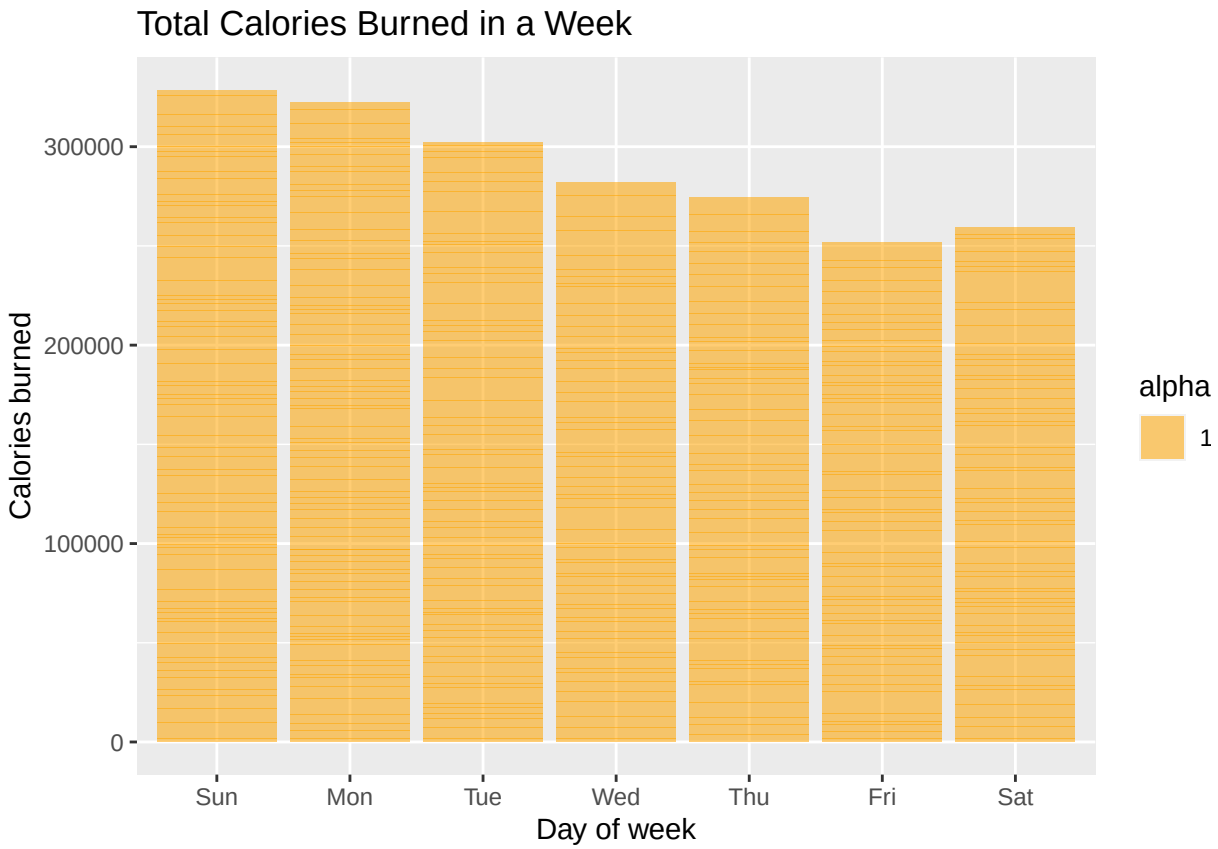
```
ggplot(data = daily_activity_cleaned) +
  aes(x = day_of_week, y = very_active_minutes, alpha=1) +
  geom_col(fill =  'blue') +
  labs(x = 'Day of week', y = 'Total very active minutes', title = 'Total Activity in a Week')
```




```
ggsave('total_activity_week.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(data = daily_activity_cleaned) +
  aes(x = day_of_week, y = calories, alpha=1) +
  geom_col(fill =  'orange') +
  labs(x = 'Day of week', y = 'Calories burned', title = 'Total Calories Burned in a Week')
```




```
ggsave('total_calories_week.png')
```

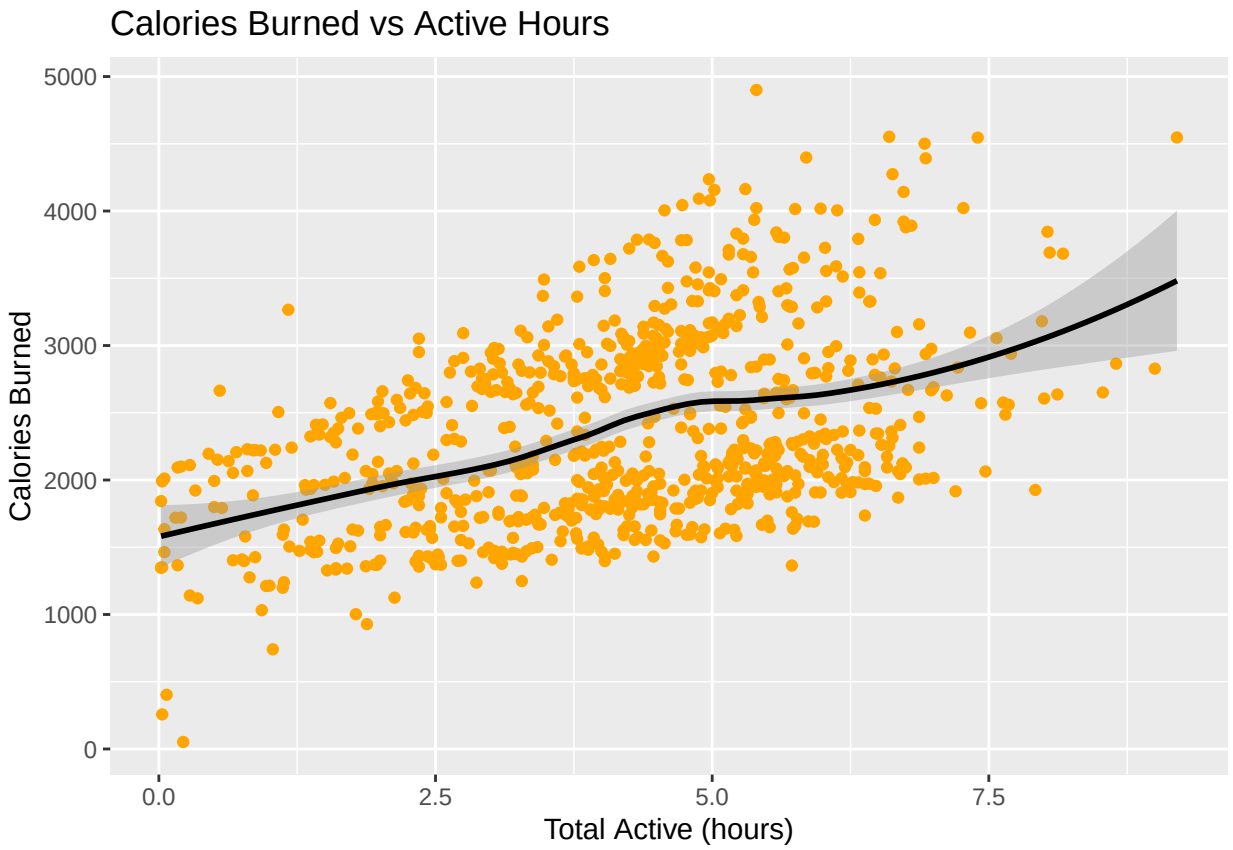
```
## Saving 6.5 x 4.5 in image
```

Trends: Daily activity levels are highest on Sundays and gradually decrease through the week.

Calories burned compared to activity: sedentary, active hours, total steps

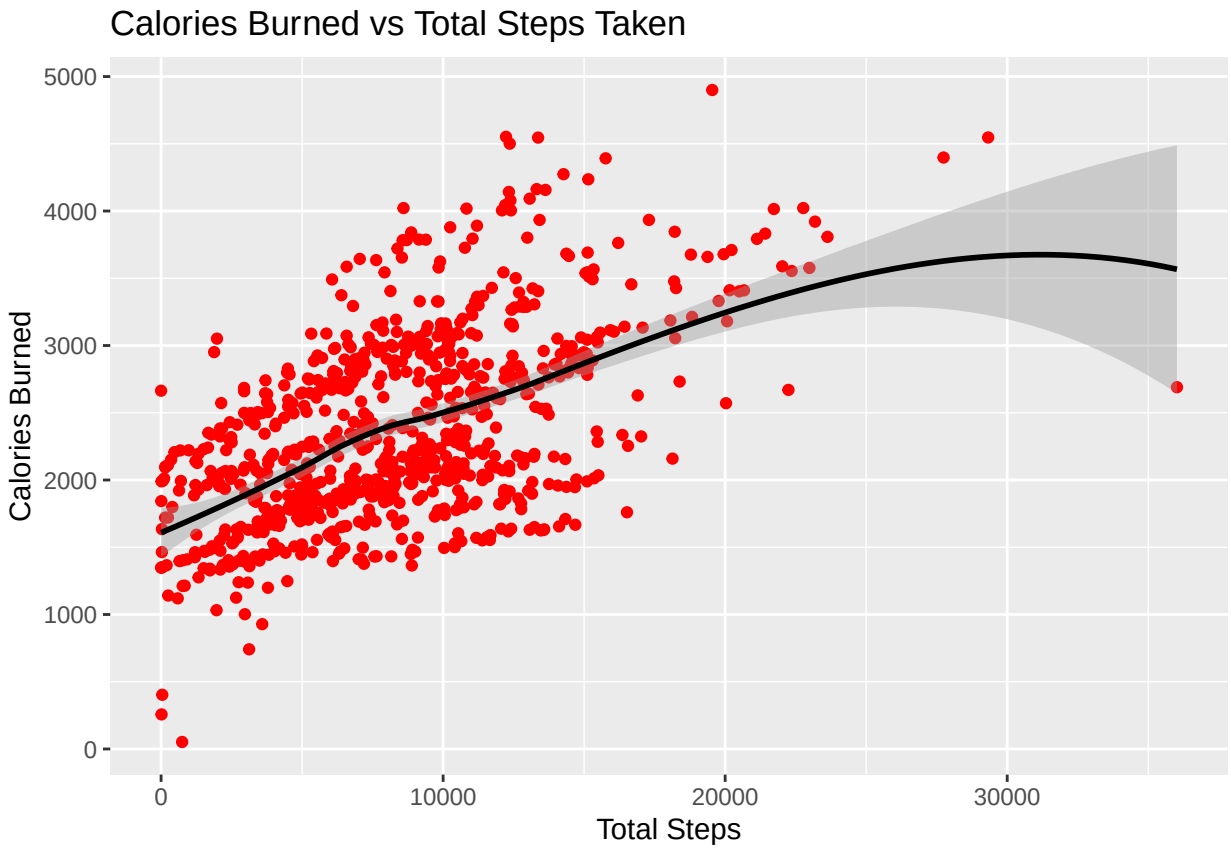
```
#active hours vs calories burned
ggplot(data= daily_activity_cleaned ) +
  aes(x=total_active_hrs,y=calories) +
  geom_point(color='orange') +
  geom_smooth(color='black') +
  labs(x='Total Active (hours)', y='Calories Burned',
       title = 'Calories Burned vs Active Hours')
```

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



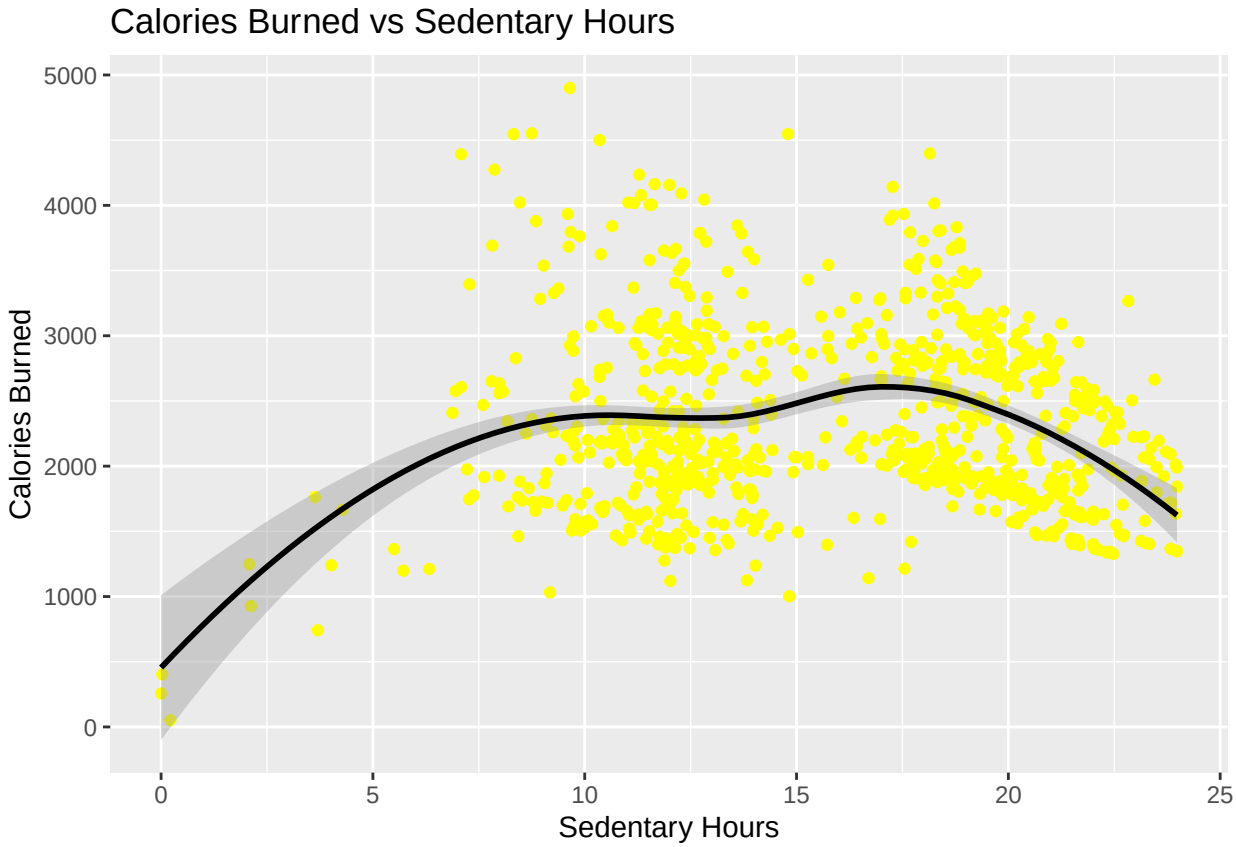

```
#total steps vs calories burned
ggplot(data= daily_activity_cleaned ) +
  aes(x=total_steps,y=calories) +
  geom_point(color='red') +
  geom_smooth(color='black') +
  labs(x='Total Steps', y='Calories Burned',
       title = 'Calories Burned vs Total Steps Taken')
```

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




```
#sedentary hours vs calories burned
daily_activity_cleaned$sedentary_hours =
  round((daily_activity_cleaned$sedentary_minutes)/60, digits = 2)

ggplot(data= daily_activity_cleaned ) +
  aes(x=sedentary_hours,y=calories) +
  geom_point(color='yellow') +
  geom_smooth(color='black') +
  labs(x='Sedentary Hours', y='Calories Burned', title = 'Calories Burned vs Sedentary Hours')
```

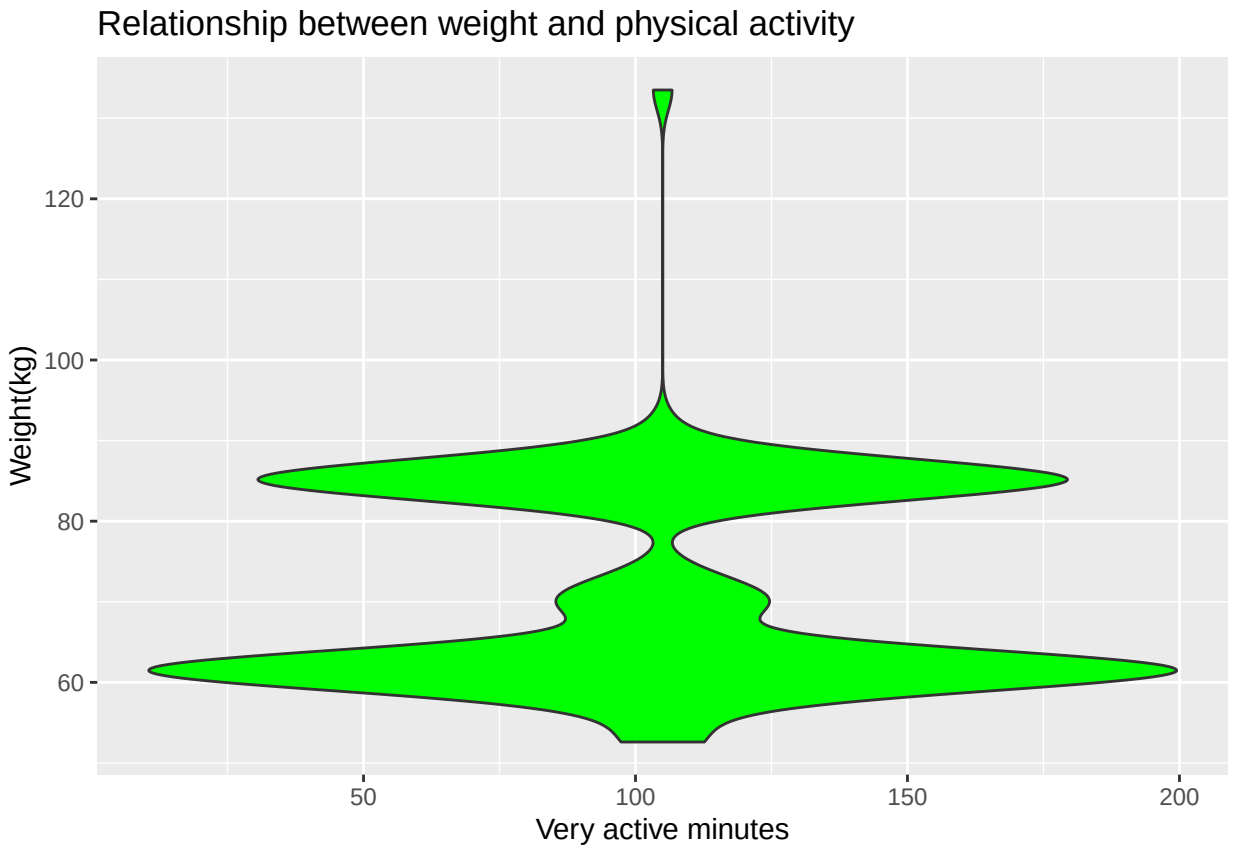
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Trends: More calories are burned as more hours are spent active and more steps are taken. There is no clear relationship between hours spent sedentary and calories burnt; there is a slight positive relationship until the 16 hour mark.

Weight vs Activity Levels

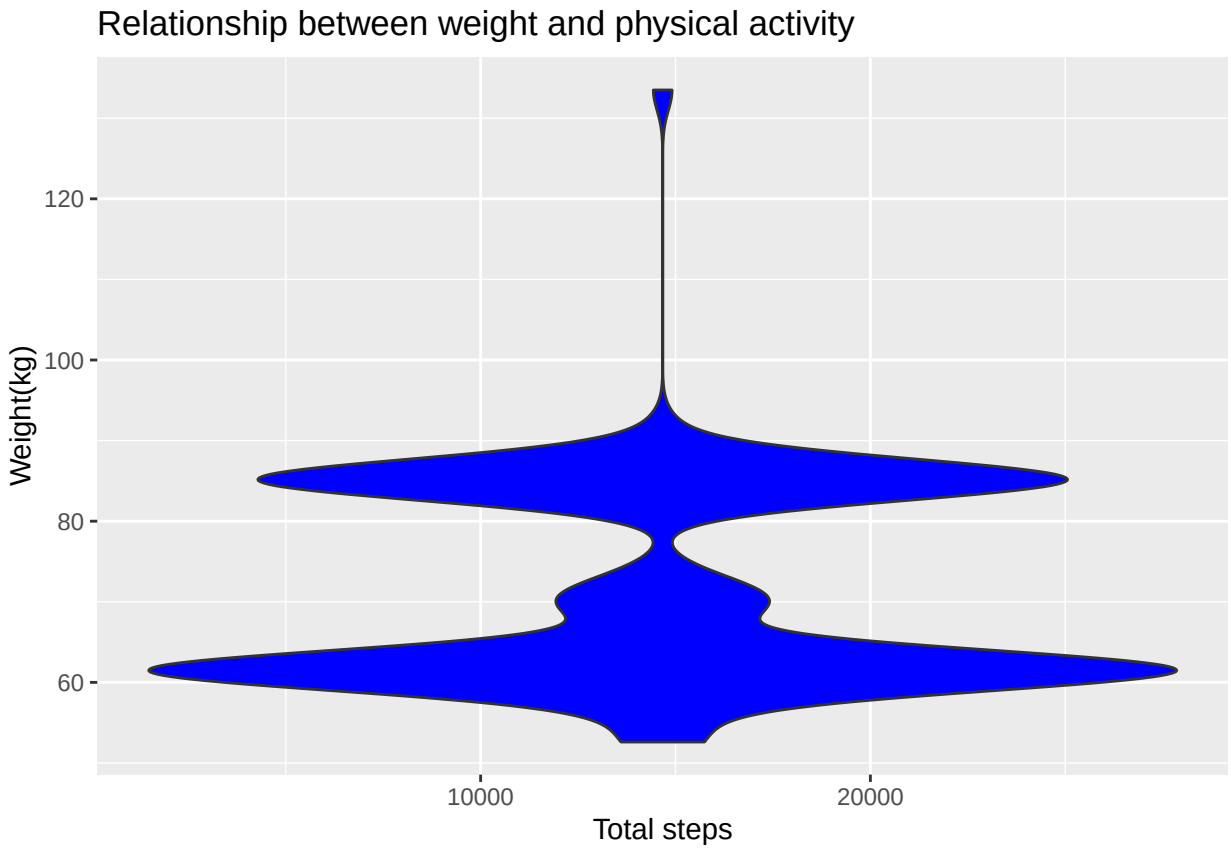
```
activity_weight <- merge(daily_activity_cleaned, weightLog, by=c('id'))
# uses id to add weightlog columns into daily activity

ggplot(data = activity_weight) +
  aes(x = very_active_minutes, y = weight_kg) +
  geom_violin(fill = 'green') +
  labs(x = 'Very active minutes', y = 'Weight(kg)',
       title = 'Relationship between weight and physical activity')
```




```
ggsave('weight_physical_activity.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(data = activity_weight) +
  aes(x = total_steps, y = weight_kg) +
  geom_violin(fill = 'blue') +
  labs(x = 'Total steps', y = 'Weight(kg)',
       title = 'Relationship between weight and physical activity')
```




```
ggsave('weight_physical_activity.png')
```

```
## Saving 6.5 x 4.5 in image
```

Trends: As weight increases, there is less activity and less total steps. Those who weigh between 60-90 kg are most active.