BellaBeat Analysis

Nidhi

2023-10-06

Setting up my environment

Installing the packages

Loading the packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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(janitor)
```

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

```
library(lubridate)
library(skimr)
```

Load the data

After this, we would need to import the datasets into RStudio

```
daily_activity <- read_csv("/cloud/project/CaseStudy2_Project/dailyActivity_merged.csv")
daily_sleep <- read.csv("/cloud/project/CaseStudy2_Project/sleepDay_merged.csv")
weight_log <- read.csv("/cloud/project/CaseStudy2_Project/weightLogInfo_merged.csv")
```

Inspect data

Here, we will check the data types and columns of each data tables.

```
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] "04/12/2016" "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_double(),
##   ..   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>
```

```
str(daily_sleep)
```

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

```
str(weight_log)
```

```
## '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" "4/21/2
##  $ WeightKg      : num  52.6 52.6 133.5 56.7 57.3 ...
##  $ WeightPounds  : 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 ...
##  $ IsManualReport: chr  "True" "True" "False" "True" ...
##  $ LogId         : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

After reviewing the output, we found several issues:

- The naming of the column names is in camelCase
- daily_activity$ActivityDate — Is formatted as CHR not as a date format
- daily_sleep$SleepDay — Is formatted as CHR not as a date format
- weight_log$Date — Is formatted as CHR not as a date format
- weight_log$IsManualReport is formated as CHR not boolean

Clean and format columns

```
daily_activity <- clean_names(daily_activity)
daily_sleep <- clean_names(daily_sleep)
weight_log <- clean_names(weight_log)
daily_activity$activity_date <- as.Date(daily_activity$activity_date,'%m/%d/%y')
daily_sleep$sleep_day <- as.Date(daily_sleep$sleep_day, '%m/%d/%y')
weight_log$date <- parse_date_time(weight_log$date, '%m/%d/%y %H:%M:%S %p')
weight_log$is_manual_report <- as.logical(weight_log$is_manual_report)
```

Analysing and share the data

It will tell total activity minutes

```
daily_activity$day_of_week <- wday(daily_activity$activity_date, label = T, abbr = T)
daily_activity$total_active_hours = round((daily_activity$very_active_minutes + daily_activity$fairly_a
daily_activity$sedentary_hours = round((daily_activity$sedentary_minutes)/60, digits = 2)
```

It will tell us time taken to get sleep :

```
daily_sleep$hours_in_bed = round((daily_sleep$total_time_in_bed)/60, digits = 2)
daily_sleep$hours_asleep = round((daily_sleep$total_minutes_asleep)/60, digits = 2)
daily_sleep$time_taken_to_sleep = (daily_sleep$total_time_in_bed - daily_sleep$total_minutes_asleep)
```

We will add a column in weight_log table 'bmi2' to check if the person is healthy, overweight or underweight.

```
weight_log <- weight_log %>%
  mutate(bmi2 = case_when(
    bmi > 24.9 ~ 'Overweight',
    bmi < 18.5 ~ 'Underweight',
    TRUE ~ 'Healthy'
  ))
```

we will add a new column in daily sleep and check who is having good, poor, over sleeping

```
daily_sleep <- daily_sleep %>%
  mutate(sleep_calculation = case_when(
    hours_asleep > 9 ~ 'Over Sleep',
    hours_asleep < 7 ~ 'Poor Sleep' ,
    TRUE ~ 'Good Sleep'
  ))
```

We will remove the zero rows for calories and total active hours.

```
daily_activity_cl <- daily_activity[!(daily_activity$calories<=0),]
```

```
daily_activity_cl <- daily_activity_cl[!(daily_activity_cl$total_active_hours<=0.00),]
```

We are now combining the tables for further analysis. We are merging (daily_sleep, daily_activity_cl) and (daily_activity_cl, weight_log) and then (activity_weight, daily_sleep) for further analysis.

```
merged_sleep_activity <- merge(daily_sleep, daily_activity_cl, by.x = "id", by.y = "id", all.x=TRUE, al
head(merged_sleep_activity)
```

```
##           id  sleep_day total_sleep_records total_minutes_asleep
## 1 1503960366 2020-04-12                   1                  327
## 2 1503960366 2020-04-12                   1                  327
## 3 1503960366 2020-04-12                   1                  327
## 4 1503960366 2020-04-12                   1                  327
## 5 1503960366 2020-04-12                   1                  327
## 6 1503960366 2020-04-12                   1                  327
##   total_time_in_bed hours_in_bed hours_asleep time_taken_to_sleep
## 1               346         5.77         5.45                  19
## 2               346         5.77         5.45                  19
## 3               346         5.77         5.45                  19
## 4               346         5.77         5.45                  19
## 5               346         5.77         5.45                  19
## 6               346         5.77         5.45                  19
##   sleep_calculation activity_date total_steps total_distance tracker_distance
## 1        Poor Sleep    2020-04-14       10460           6.74             6.74
## 2        Poor Sleep    2020-04-24       10039           6.41             6.41
## 3        Poor Sleep    2020-04-13       10735           6.97             6.97
## 4        Poor Sleep    2020-04-23       14371           9.04             9.04
## 5        Poor Sleep    2020-04-25       15355           9.80             9.80
## 6        Poor Sleep    2020-04-15        9762           6.28             6.28
##   logged_activities_distance very_active_distance moderately_active_distance
## 1                          0                2.44                       0.40
## 2                          0                2.92                       0.21
## 3                          0                1.57                       0.69
## 4                          0                2.81                       0.87
## 5                          0                5.29                       0.57
## 6                          0                2.14                       1.26
##   light_active_distance sedentary_active_distance very_active_minutes
## 1                  3.91                         0                  30
## 2                  3.28                         0                  39
## 3                  4.71                         0                  21
## 4                  5.36                         0                  41
## 5                  3.94                         0                  73
## 6                  2.83                         0                  29
##   fairly_active_minutes lightly_active_minutes sedentary_minutes calories
## 1                    11                   181              1218     1776
## 2                     5                   238               709     1788
## 3                    19                   217               776     1797
## 4                    21                   262               732     1949
## 5                    14                   216               814     2013
## 6                    34                   209               726     1745
##   day_of_week total_active_hours sedentary_hours
## 1         Tue               3.70           20.30
## 2         Fri               4.70           11.82
## 3         Mon               4.28           12.93
## 4         Thu               5.40           12.20
## 5         Sat               5.05           13.57
## 6         Wed               4.53           12.10
```

```
activity_weight <- merge(daily_activity_cl, weight_log, by=c('id'))
```

```
activity_weight_sleep <- merge(activity_weight, daily_sleep, by=c('id'))
```

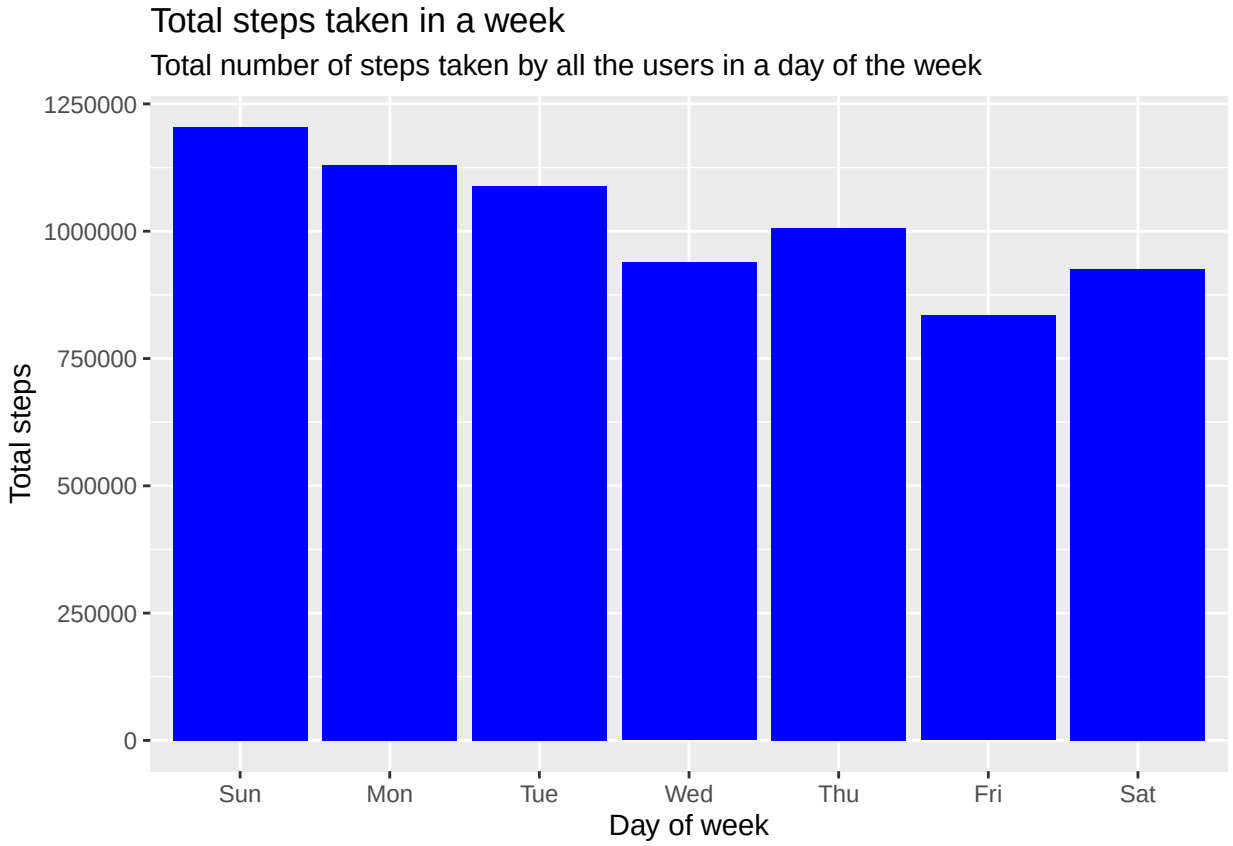
Summarize the data and creating Plots

We can find the general trends in the data to get insights and answers to our business problem. Here, are some of the insights we get from the data:

Which days the users are most active

Here we will check the relationship between totalsteps, very active minutes, calories.

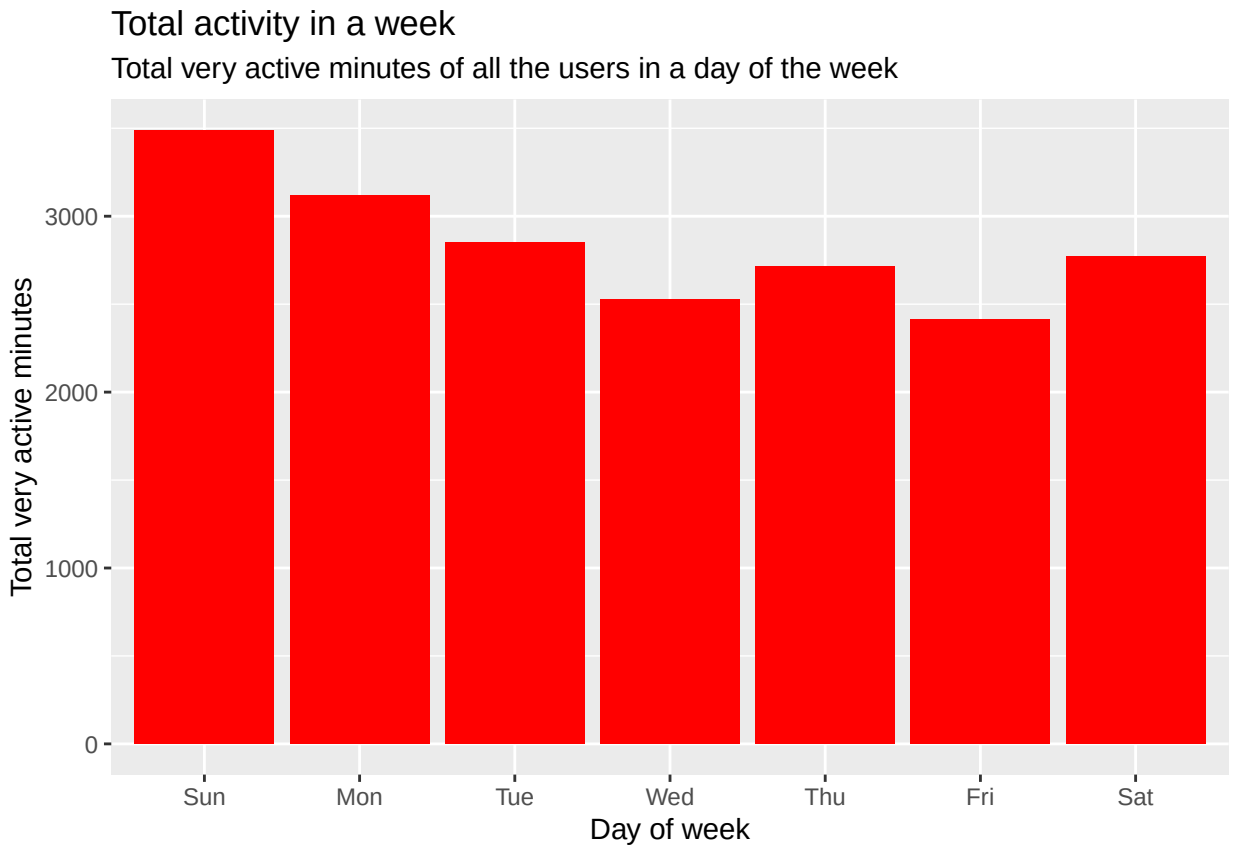
```
ggplot(data = daily_activity_cl) +
  aes(x = day_of_week, y = total_steps) +
  geom_col(fill =  'blue') +
  labs(x = 'Day of week', y = 'Total steps', title = 'Total steps taken in a week', subtitle = 'Total nu
```




```
ggsave('total_steps.png')
```

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

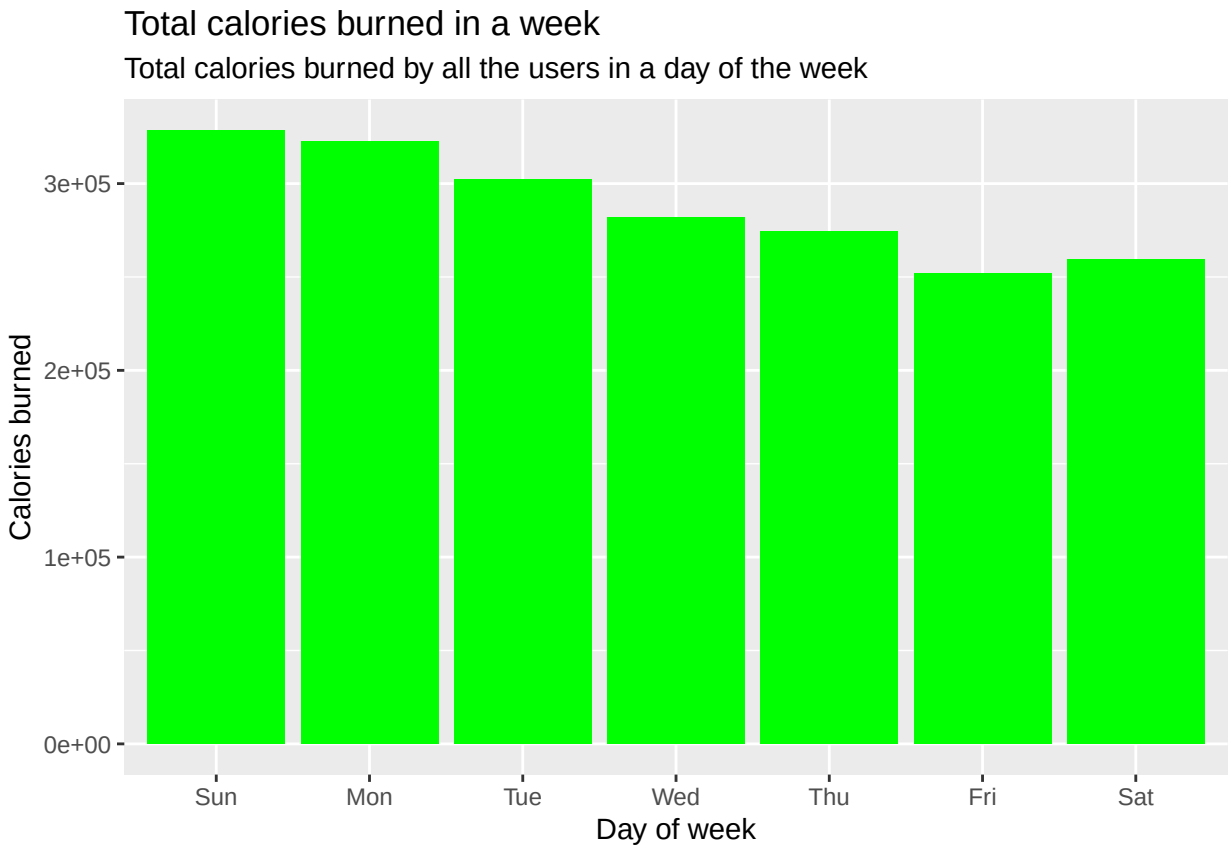
```
ggplot(data = daily_activity_cl) +
  aes(x = day_of_week, y = very_active_minutes) +
  geom_col(fill =  'red') +
  labs(x = 'Day of week', y = 'Total very active minutes', title = 'Total activity in a week', subtitle
```




```
ggsave('total_activity.png')
```

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

```
  ggplot(data = daily_activity_cl) +
  aes(x = day_of_week, y = calories) +
  geom_col(fill =  'Green') +
  labs(x = 'Day of week', y = 'Calories burned', title = 'Total calories burned in a week', subtitle =
```



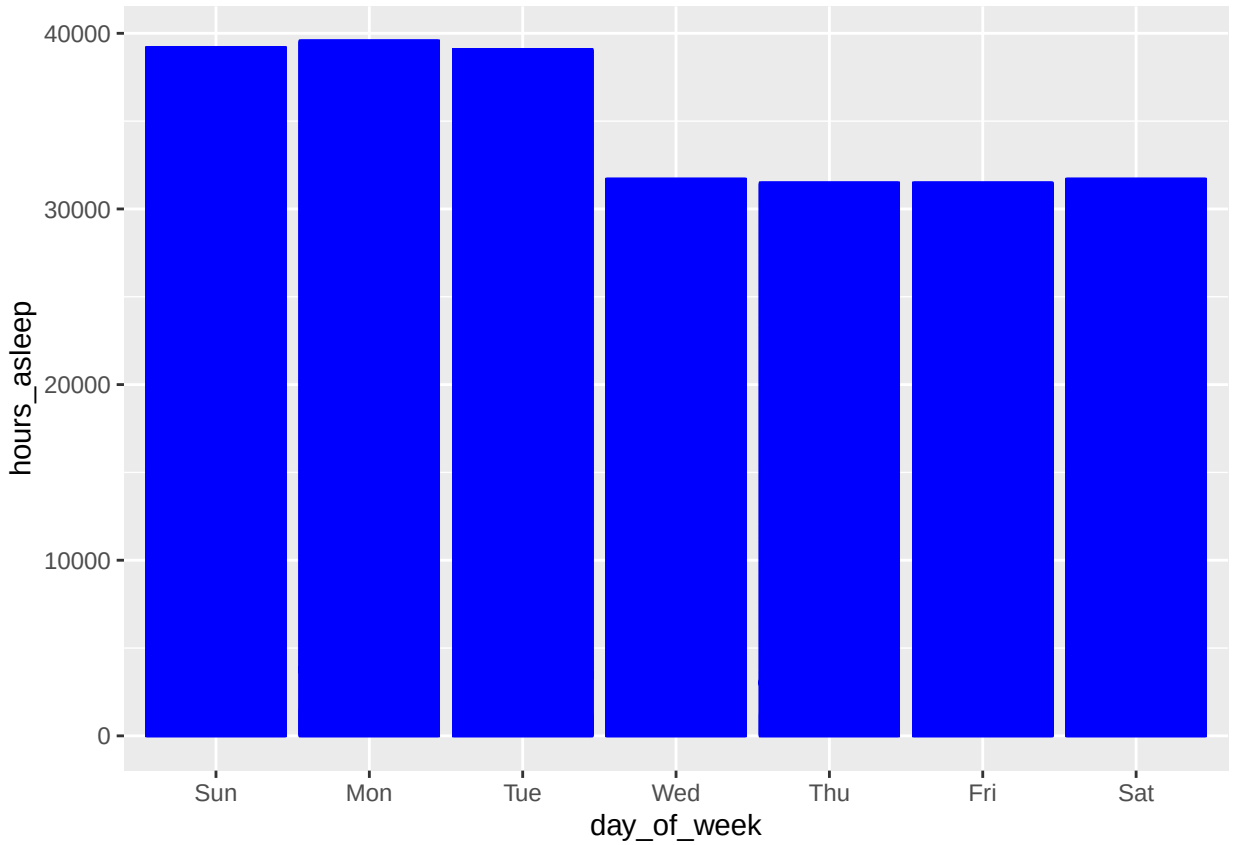

```
ggsave('total_calories.png')
```

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

Here, we can see the most active days for the Fitbit users were on Sunday, with a slow decline throughout the week.

Which days users are having good sleep and which days the users are most inactive

```
ggplot(data = activity_weight_sleep, aes(day_of_week,hours_asleep)) +
    geom_col(color="Blue")
```



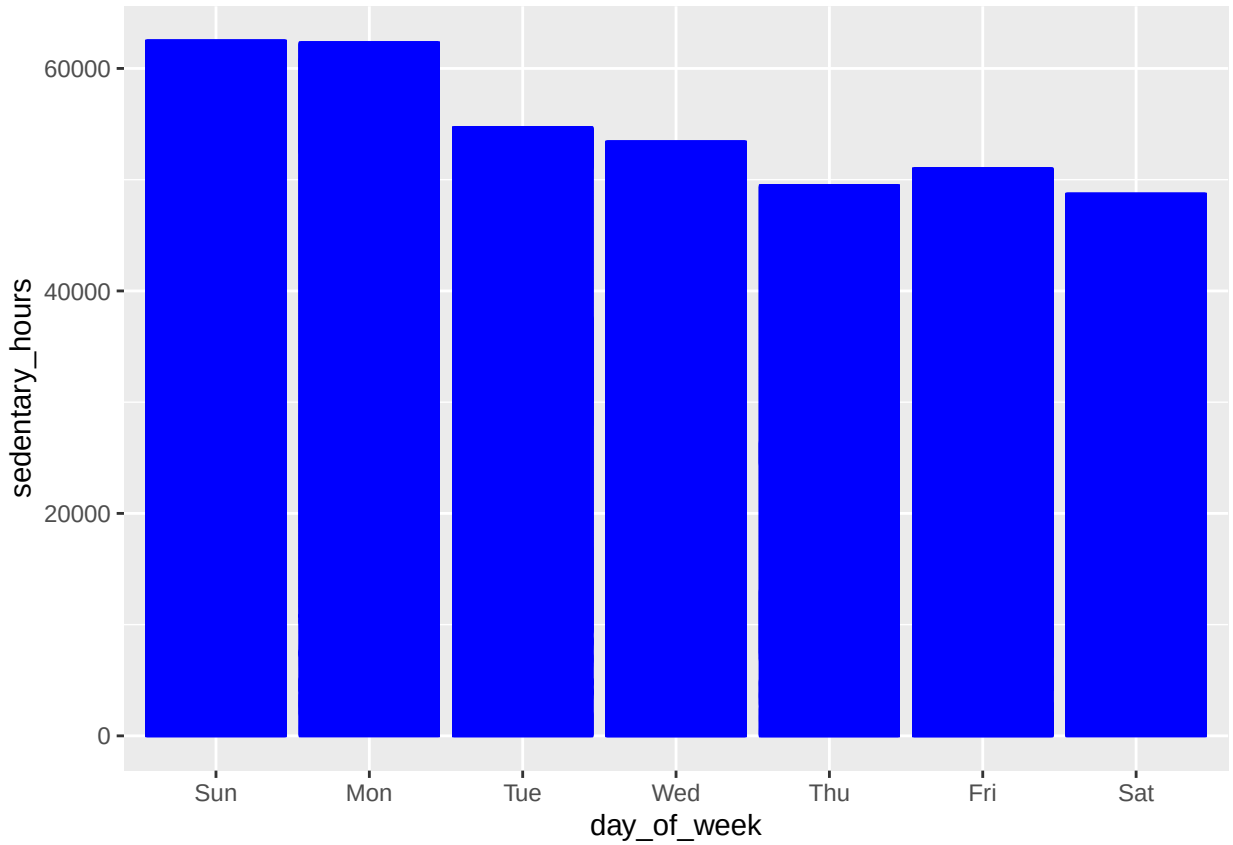

```
  labs(x = 'Day of week', y = 'hours_asleeo', title = 'Total sleep hours in a day of a week', subtitle =
```

```
## $x
## [1] "Day of week"
##
## $y
## [1] "hours_asleeo"
##
## $title
## [1] "Total sleep hours in a day of a week"
##
## $subtitle
## [1] "Which days users have the best sleep in a day of the week"
##
## attr(,"class")
## [1] "labels"
```

```
ggsave('sleep_day.png')
```

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

```
ggplot(data = activity_weight_sleep, aes(day_of_week,sedentary_hours)) +
    geom_col(color="Blue")
```




```
labs(x = 'Day of week', y = 'Sedentary hours', title = 'Total sedantary hours in a day of a week', sub
```

```
## $x
## [1] "Day of week"
##
## $y
## [1] "Sedentary hours"
##
## $title
## [1] "Total sedantary hours in a day of a week"
##
## $subtitle
## [1] "Which days users are most inactive in a day of the week"
##
## attr(,"class")
## [1] "labels"
```

```
ggsave('Sedentary_day.png')
```

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

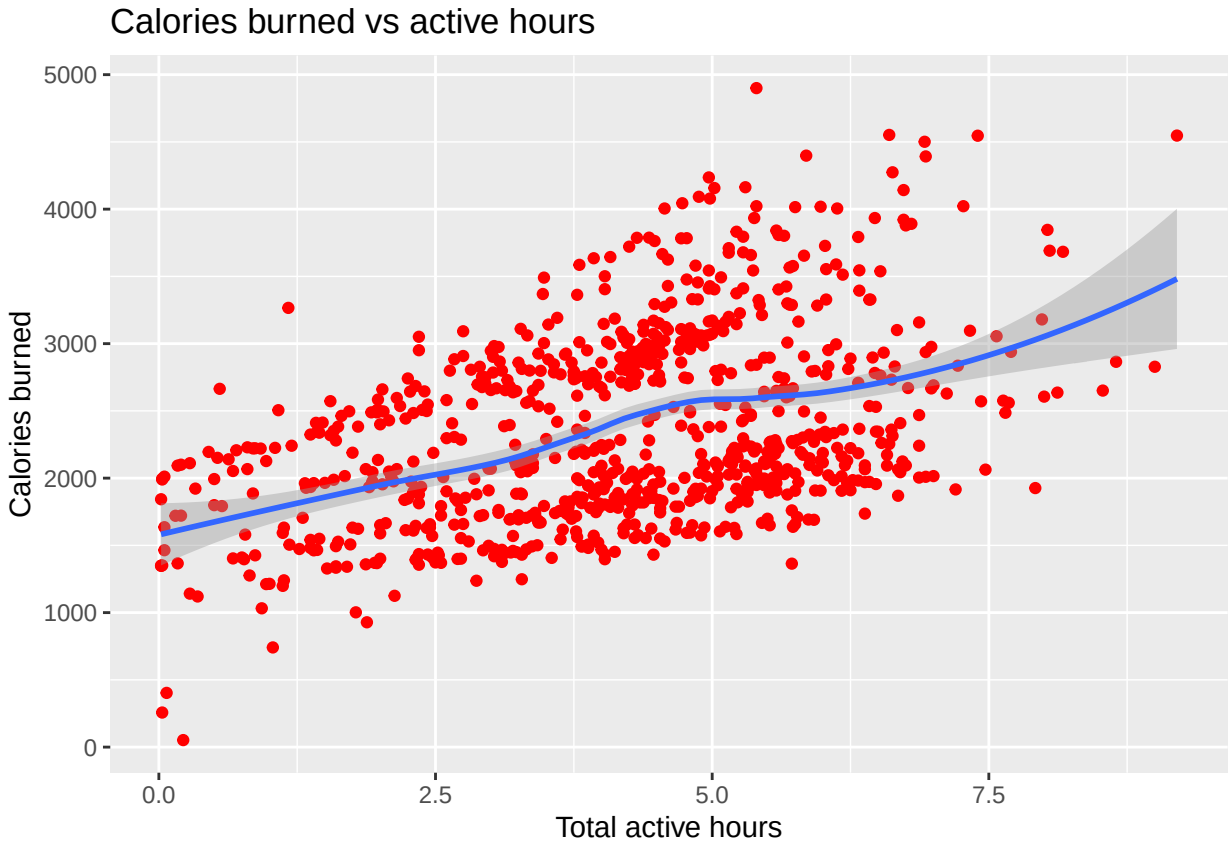
From this we can infer that at start of the week users are sleep best and are more inactive.

The relationship between total active hours, total steps taken, sleep and sedentary hours against calories burned

```
ggplot(data = daily_activity_cl) +
  aes(x= total_active_hours, y = calories) +
  geom_point(color = 'red') +
```

```
  geom_smooth() +
  labs(x = 'Total active hours', y = 'Calories burned', title = 'Calories burned vs active hours')
```

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

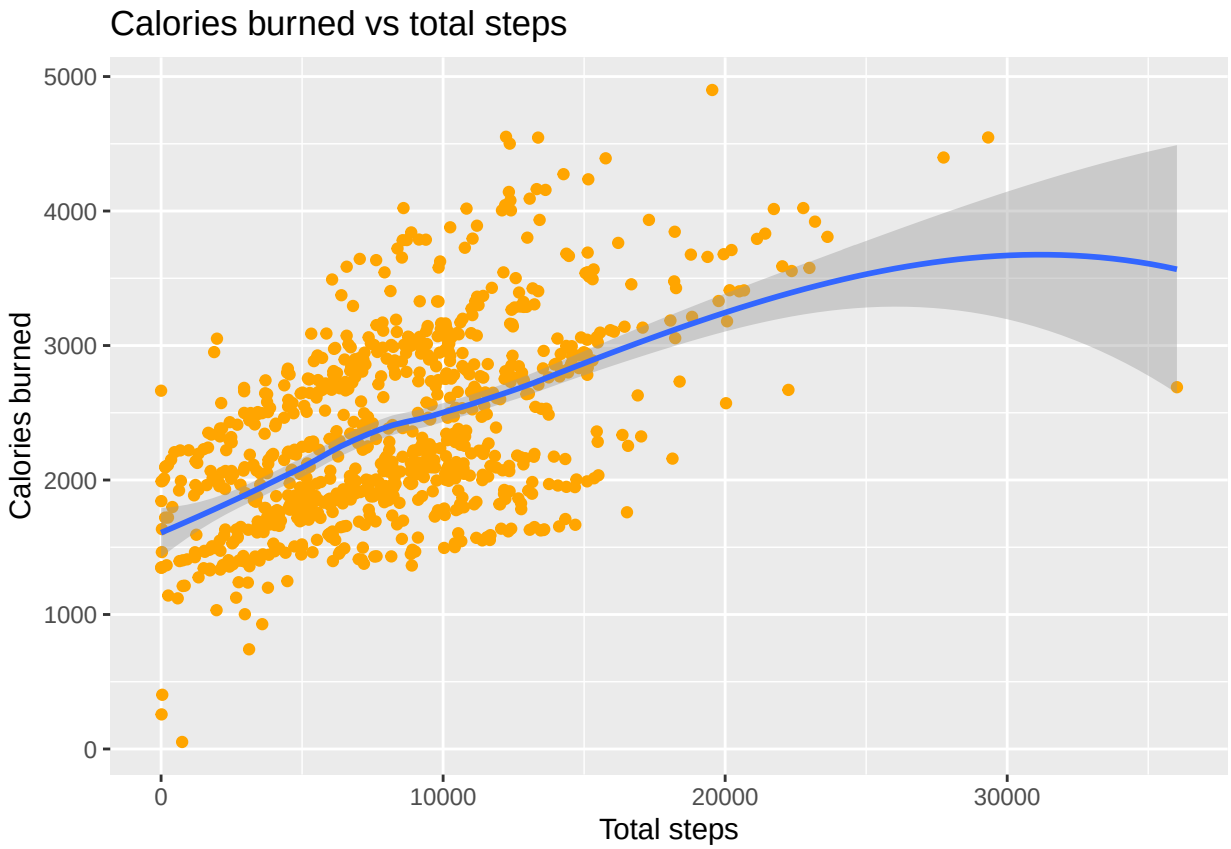



```
ggsave('calories_burned_vs_active_hours.png')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
ggplot(data = daily_activity_cl) +
  aes(x= total_steps, y = calories) +
  geom_point(color = 'orange') +
  geom_smooth() +
  labs(x = 'Total steps', y = 'Calories burned', title = 'Calories burned vs total steps')
```

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

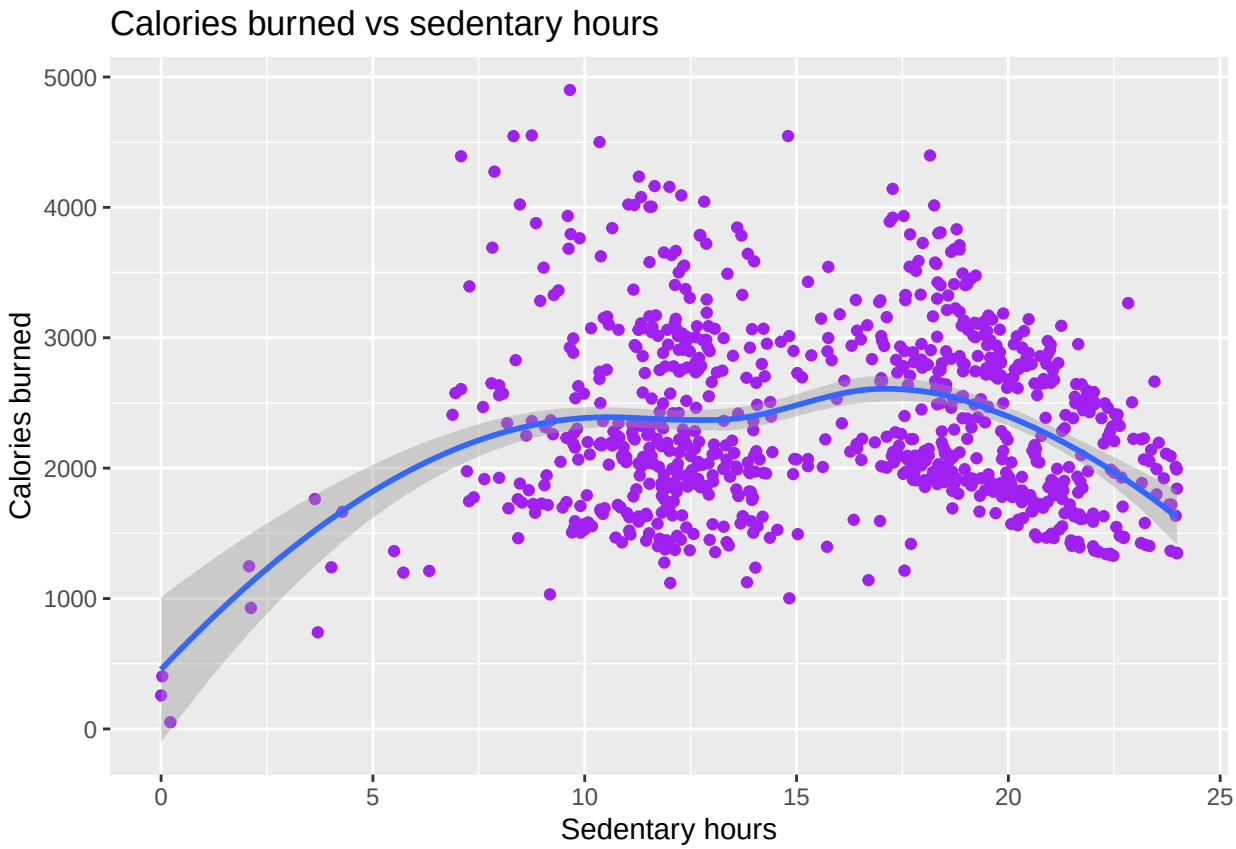



```
ggsave('calories_burned_vs_total_steps.png')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
ggplot(data = daily_activity_cl) +
  aes(x= sedentary_hours, y = calories) +
  geom_point(color = 'purple') +
  geom_smooth() +
  labs(x = 'Sedentary hours', y = 'Calories burned', title = 'Calories burned vs sedentary hours')
```

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

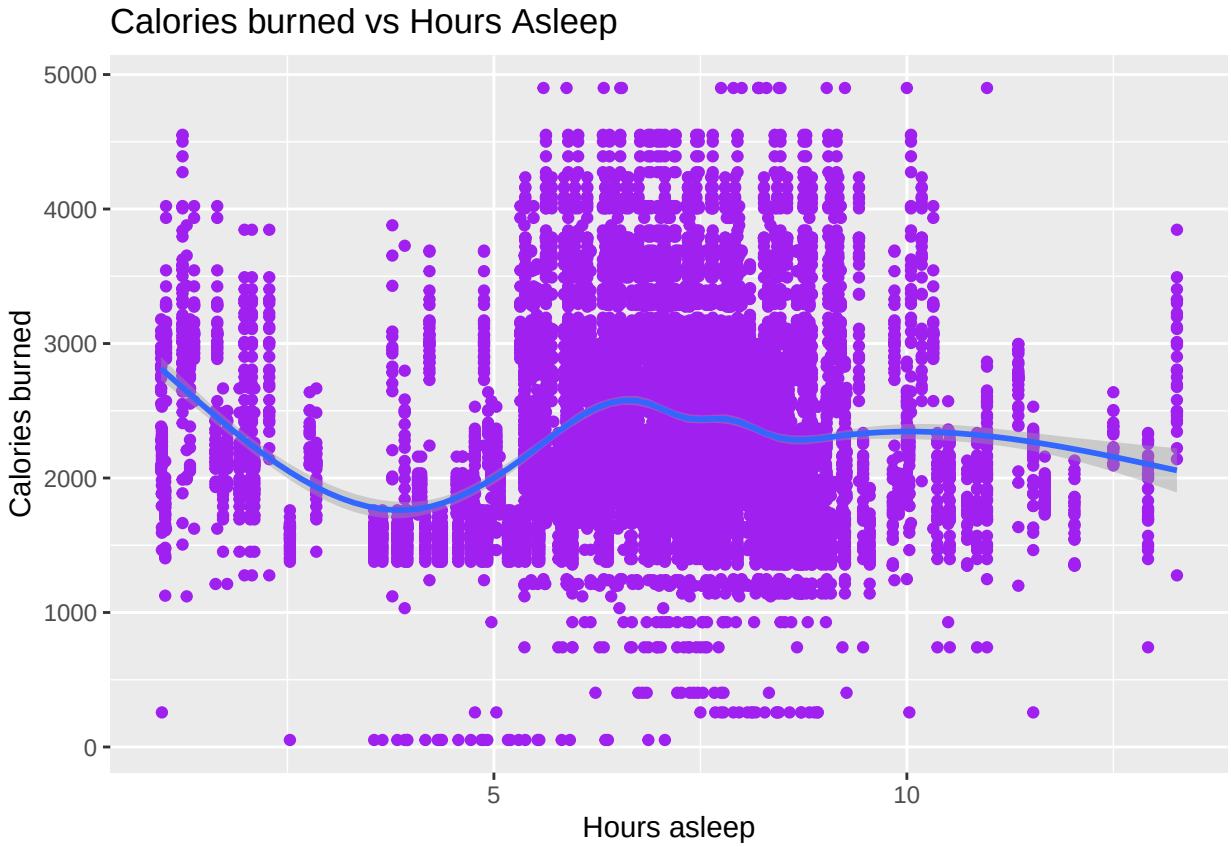



```
ggsave('sedentary_hours_vs_calories_burned.png')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
ggplot(data = merged_sleep_activity) +
   geom_point(mapping=aes(x= hours_asleep, y = calories), color = 'Purple')+
   geom_smooth(mapping=aes(x= hours_asleep, y = calories)) +
    labs(x = 'Hours asleep', y = 'Calories burned', title = 'Calories burned vs Hours Asleep')
```

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




```
ggsave('Hours_Asleep_vs_calories_burned.png')
```

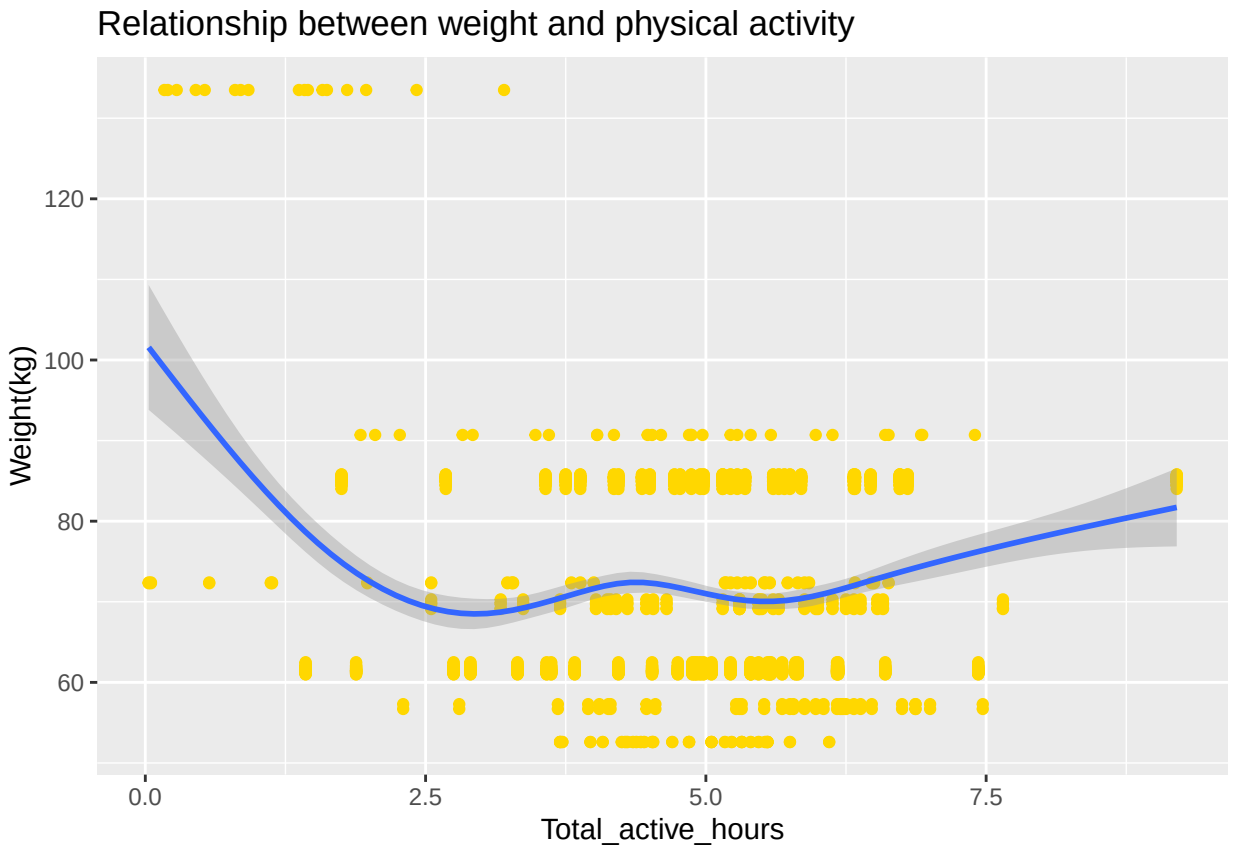
```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

We can tell that there is a positive correlation between calories burned and total steps taken/total active hours. However, in the last chart, we can see that the relationship between sedentary hours and calories burned was fairly positive up till about the 17-hour mark and when a person is not active for more than 17 hrs then the calorie burned is decreasing. The graph between sleep hours and calorie burned also tells also that when people is not sleep for atleast 5 hours the calories burned is decreasing and from 5 hrs to 7hrs it is the maximum and till 10 hrs it is constant and when people is sleeping more than 10 hrs it again starts to fall.

The relationship between weight, total active hours

```
ggplot(data = activity_weight,aes(x= total_active_hours, y = weight_kg)) +
  geom_point(color = 'Gold') +
  geom_smooth(orientation = "x")+
  labs(x = 'Total_active_hours', y = 'Weight(kg)', title = 'Relationship between weight and physical ac
```

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




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

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

We can infer that users weighing around 60kg & 85kg are the most active

The number of overweight users

```
daily_activity_cl %>% distinct(id)
```

```
## # A tibble: 33 x 1
##            id
##         <dbl>
##  1 1503960366
##  2 1624580081
##  3 1644430081
##  4 1844505072
##  5 1927972279
##  6 2022484408
##  7 2026352035
##  8 2320127002
##  9 2347167796
## 10 2873212765
## # i 23 more rows
```

```
activity_weight %>% distinct(id)
```

```
##           id
## 1 1503960366
```

```
## 2 1927972279
## 3 2873212765
## 4 4319703577
## 5 4558609924
## 6 5577150313
## 7 6962181067
## 8 8877689391
```

```
activity_weight %>% count(bmi2,id)
```

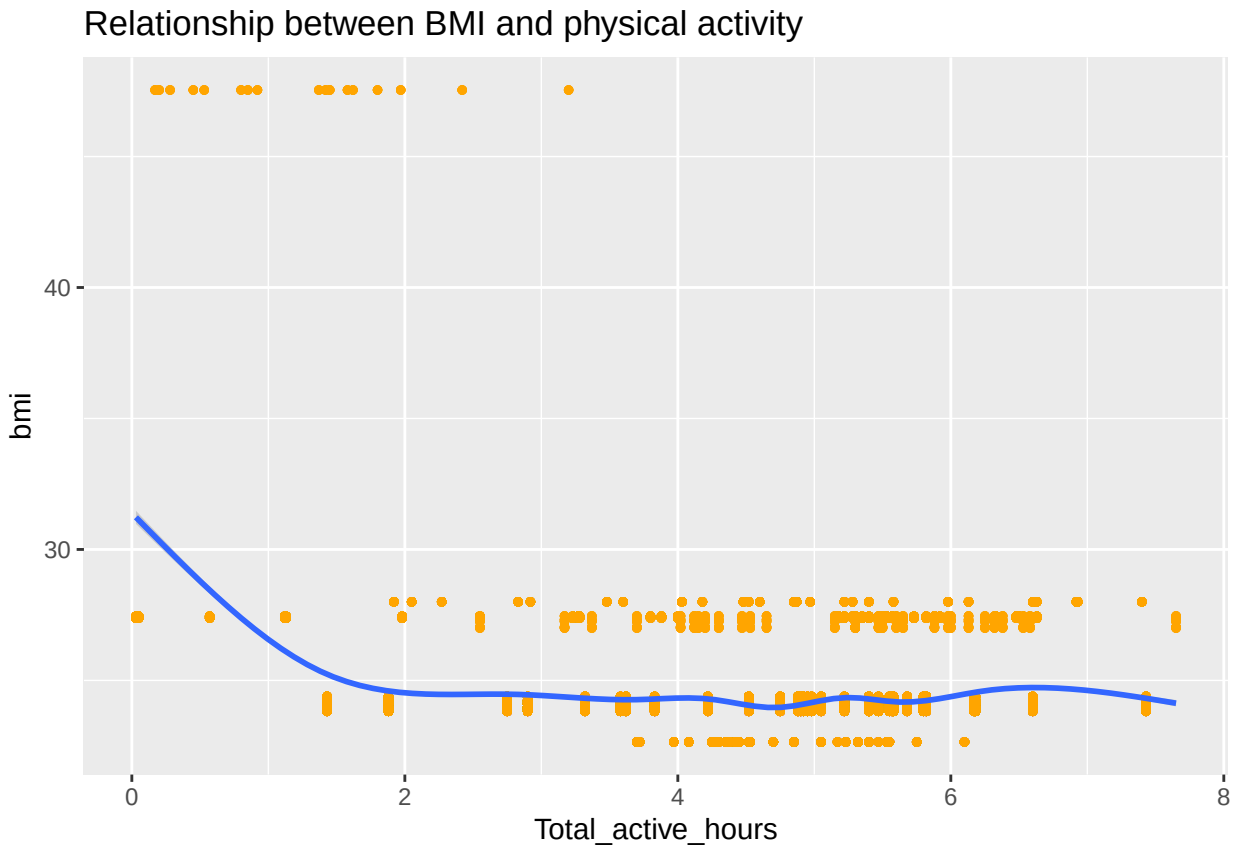
```
##         bmi2         id   n
## 1    Healthy 1503960366  60
## 2    Healthy 2873212765  62
## 3    Healthy 6962181067 930
## 4 Overweight 1927972279  17
## 5 Overweight 4319703577  60
## 6 Overweight 4558609924 155
## 7 Overweight 5577150313  28
## 8 Overweight 8877689391 744
```

Here we can conclude that out of the 33 users, only 8 submitted their responses regarding weight. 5 users are overweight and only 3 are within the healthy BMI range of 18.5–24.9

The relationship between good sleep, total activity and Health(BMI)

```
ggplot(data = activity_weight_sleep, aes(total_active_hours,bmi)) +
  geom_point(color = "Orange",size = 1) +
  geom_smooth() +
   labs(x = 'Total_active_hours', y = 'bmi', title = 'Relationship between BMI and physical activity')
```

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

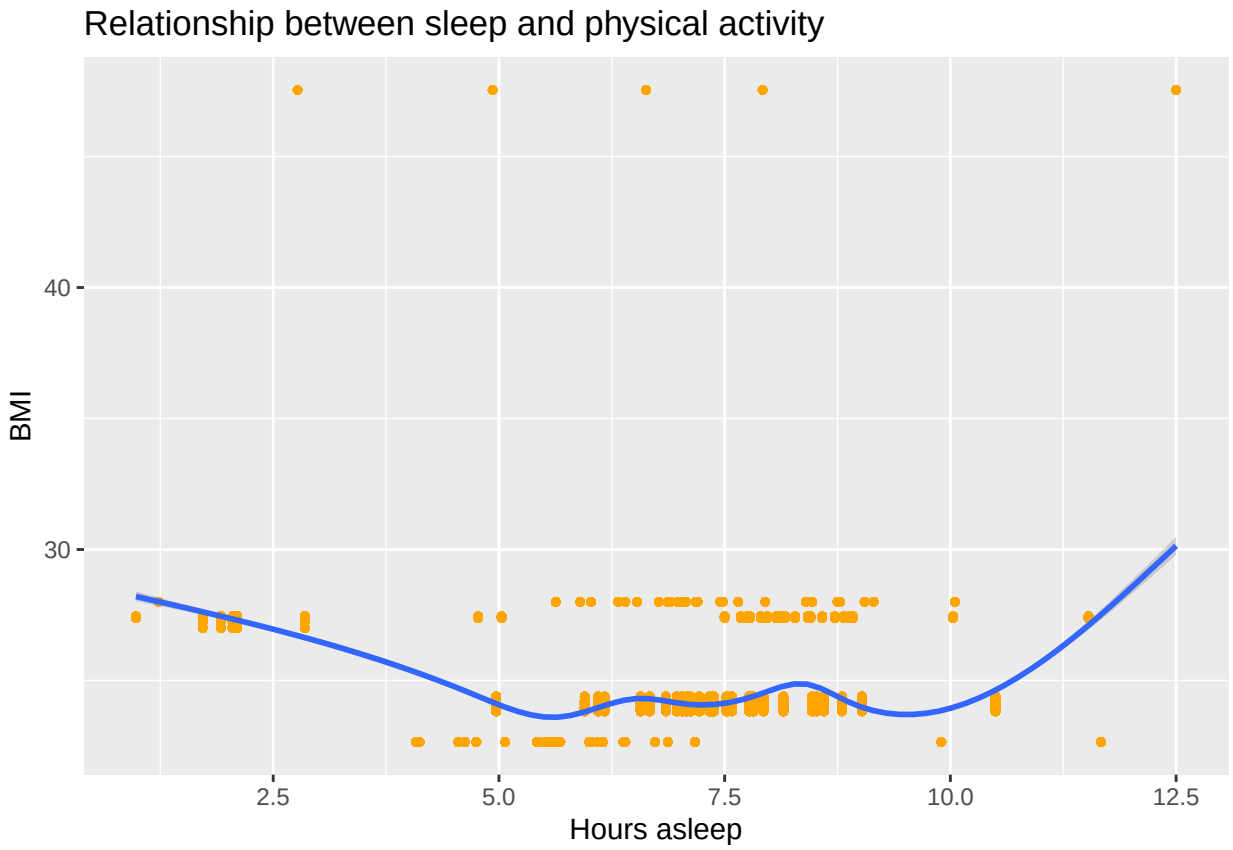



```
ggsave('bmi_physical_activity.png')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
ggplot(data = activity_weight_sleep, aes(hours_asleep,bmi)) +
  geom_point(color = "Orange",size = 1) +
  geom_smooth()+
   labs(x = 'Hours asleep', y = 'BMI', title = 'Relationship between sleep and physical activity')
```

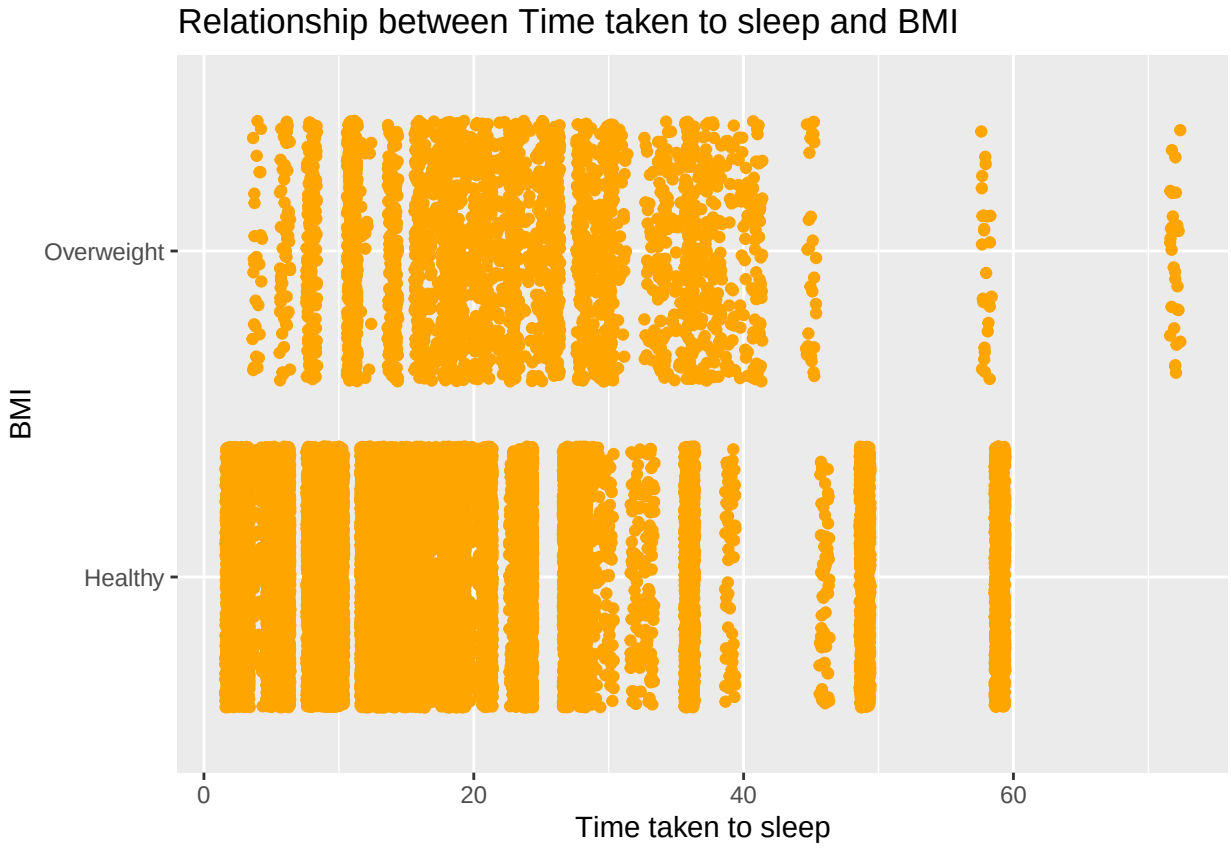
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
ggsave('sleep_BMI.png')
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
ggplot(data = activity_weight_sleep, aes(time_taken_to_sleep,bmi2)) +
  geom_jitter(color = "Orange")+
labs(x = 'Time taken to sleep', y = 'BMI', title = 'Relationship between Time taken to sleep and BMI')
```




```
ggsave('time_taken_to_sleep_bmi.png')
```

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

From below we get find general trends in the data:

```
summary(daily_activity_cl$total_steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    4920    8053    8319   11100   36019
```

```
summary(daily_activity_cl$very_active_minutes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    7.00   23.21   36.00  210.00
```

```
summary(daily_sleep$hours_asleep)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.970   6.020   7.220   6.992   8.170  13.270
```

From here we can say, the average steps taken by active users are 8053 who are very active for around 7 hours and their average sleep 7.22.

#Act

In the previous section of Analyze & Share, we have covered the 1st and 2nd business task which are:

- What are some trends in smart device usage
- How could these trends apply to Bellabeat customers

We have analysed and shared several trends in the smart device usage and which could be followed by bellabeat customers.

Based on my findings, I would like to share my views on this matter.

Users spend more time engaged in physical activity specifically on Sundays, which then proceeds to decrease throughout the week with a slight peak on Thursdays which then sees a slow climb on Saturdays.

I suspect that:

Motivation levels & free time are higher on the weekends, which would provide an opportunity for users to sneak in a workout. As work load decreases, a window of opportunity to exercise would present itself in the midweek (Thursdays) We see an alltime low of recorded activity on Friday's and some on Saturdays due to the possibility of social engagement with friends/coworkers.

- Now to answer the final business task, I would like to share my recommendations based on my findings to help influence Bellabeat's marketing strategy.

-Bellabeat could host events limited to those that are enrolled in their Bellabeat memberships which would reward users who engage in a healthy lifestyle(IE 8k steps a day, less than 7 hours sedentary etc.) with points. With enough points, users could then use points to purchase products that help supplement a healthy lifestyle.

-Bellabeat could partner with healthcare or sports brands to reward users who consistently engage in a healthy lifestyle with coupons/store discounts.

-Bellabeat could introduce some 5mins or 10 mins videos on easy but impactful workout that could help its inactive users to motivate them in doing some activities.

-Bellabeat could send notifications if inactivity is very less or sleep is not well

some general recommendations to further improve Bellabeat's products:

-Bellabeat could implement personalized milestones, to encourage users to slowly engage in a more healthy lifestyle.

-Bellabeat could introduce some 5mins or 10 mins videos on easy but impactful workout that could help its inactive users to motivate them in doing some activities.

-Bellabeat could send notifications if inactivity is very less or sleep is not well

Additional remarks:

Bellabeat should require users to input their height, weight and their activity levels so that BMR calculations and a more accurate.

Bellabeat should create devices that would track sleep more sophisticatedly (REM sleep tracking, deep sleep tracking) to provide more insights into sleep health.