Case study

Navyadeep

2024-01-08

ASK PHASE:

Stakeholders:

- Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team
- Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Business task:

Analyzing smart device usage data in order to gain insights into how consumers use non-Bellabeat smart devices to gain insights to make use in Bellabeat's marketing strategy.

Prepare Phase

Using Third party data which is containing personal fitness tracker data(structured) from 30 fitbit users including minute-level output for physical activity, heart rate, and sleep monitoring. It includesinformation about daily activity, steps, and heart rate that can be used to explore users' habits.

The data is divided into 18 different csv files that contain the individual metrics as well as a file that that has all the different data matrics merged.

The data is from 03.12.2016-05.12.2016.

Installing Packages:

```
install.packages("tidyverse")
```

```
## 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)
```

Loading the Downloaded packages:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     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(lubridate)
library(janitor)
```

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

Loading Data for analysis:

```
main_df <- read_csv("dailyActivity_merged.csv")
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sleep_df=read_csv("sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
weight_df=read_csv("weightLogInfo_merged.csv")
```

```
## Rows: 67 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
calory_hour=read_csv("hourlyCalories_merged.csv")
```

```
## Rows: 22099 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, Calories
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

To ensure Data integrity lets check internal structure of the data

```
str(main_df)
```

```
## 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] "4/12/2016" "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(sleep_df)
```

```
## spc_tbl_ [413 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:413] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr [1:413] "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:
```

```
##  $ TotalSleepRecords : num [1:413] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:413] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:413] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   SleepDay = col_character(),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(weight_df)
```

```
## spc_tbl_ [67 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id            : num [1:67] 1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ Date          : chr [1:67] "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM" "4
##  $ WeightKg      : num [1:67] 52.6 52.6 133.5 56.7 57.3 ...
##  $ WeightPounds  : num [1:67] 116 116 294 125 126 ...
##  $ Fat           : num [1:67] 22 NA NA NA NA 25 NA NA NA NA ...
##  $ BMI           : num [1:67] 22.6 22.6 47.5 21.5 21.7 ...
##  $ IsManualReport: logi [1:67] TRUE TRUE FALSE TRUE TRUE TRUE ...
##  $ LogId         : num [1:67] 1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   WeightKg = col_double(),
##   ..   WeightPounds = col_double(),
##   ..   Fat = col_double(),
##   ..   BMI = col_double(),
##   ..   IsManualReport = col_logical(),
##   ..   LogId = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(calory_hour)
```

```
## spc_tbl_ [22,099 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id          : num [1:22099] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour: chr [1:22099] "4/12/2016 12:00:00 AM" "4/12/2016 1:00:00 AM" "4/12/2016 2:00:00 AM"
##  $ Calories    : num [1:22099] 81 61 59 47 48 48 48 47 68 141 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityHour = col_character(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Process Phase

After checking the internal structure we got to know that the column named `ActivityDate`in main_df,`SleepDay` in sleep_df has the wrong data type as charactor whereas it should be in Date

format and to change that:

```
main_df <- main_df %>%
  mutate(ActivityDate = as.Date(ActivityDate, format = "%m/%d/%Y"))

sleep_df <- sleep_df %>% mutate(SleepDay = mdy_hms(SleepDay,tz="UTC"))

weight_df <- weight_df %>% mutate(Date = mdy_hms(Date,tz="UTC"))

calory_hour <- calory_hour %>% mutate(ActivityHour = mdy_hms(ActivityHour,tz="UTC"))

# This step is to separate the date and time into individual columns

calory_hour <- mutate(calory_hour,
    date=as.Date(ActivityHour),
    calhours=format(ActivityHour, "%H:%M:%S")
  )
# Deleting the ActivityHour column
calory_hour <- select(calory_hour,-ActivityHour)
```

Now that all the data types are correct lets convert all the column names to lower case :

```
main_df <- rename_with(main_df,tolower)

sleep_df <-rename_with(sleep_df,tolower)

weight_df <- rename_with(weight_df,tolower)

calory_hour <- rename_with(calory_hour,tolower)

main_df <- main_df %>% rename(date = activitydate)

sleep_df <- sleep_df %>% rename( date = sleepday)
```

To check that the column names are consistent and unique:

```
main_df <- clean_names(main_df)

sleep_df <- clean_names(sleep_df)

weight_df <- clean_names(weight_df)

calory_hour <- clean_names(calory_hour)
```

Analyze Phase

```
main_df %>%
  select(totalsteps,veryactivedistance,moderatelyactivedistance,lightactivedistance) %>%
  summary()
```

```
##    totalsteps    veryactivedistance moderatelyactivedistance
##  Min.   :    0   Min.   : 0.000     Min.   :0.0000
##  1st Qu.: 3790   1st Qu.: 0.000     1st Qu.:0.0000
##  Median : 7406   Median : 0.210     Median :0.2400
##  Mean   : 7638   Mean   : 1.503     Mean   :0.5675
```

```
##  3rd Qu.:10727   3rd Qu.: 2.053     3rd Qu.:0.8000
##  Max.   :36019   Max.   :21.920     Max.   :6.4800
##  lightactivedistance
##  Min.   : 0.000
##  1st Qu.: 1.945
##  Median : 3.365
##  Mean   : 3.341
##  3rd Qu.: 4.782
##  Max.   :10.710
```

in the sleep data, An average of 39 minutes are wasted in the bed without sleeping

```
sleep_df <- sleep_df %>%
  mutate(total_time_notslept <- totaltimeinbed - totalminutesasleep)
sleep_df %>%
  select(`total_time_notslept <- totaltimeinbed - totalminutesasleep`) %>%
  summary()
```

```
##  total_time_notslept <- totaltimeinbed - totalminutesasleep
##  Min.   :  0.00
##  1st Qu.: 17.00
##  Median : 25.00
##  Mean   : 39.17
##  3rd Qu.: 40.00
##  Max.   :371.00
```

Checking the average weight and the Body Mass Index

```
weight_df %>%
  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
```

Let us see the calory data per Hour

```
cal_show <- calory_hour %>% group_by(calhours) %>% drop_na() %>% summarise(Total_calories = sum(calorie
```

Joining the Main data with the Weight data:

```
main_weight_df = merge(x=main_df,y=weight_df,by = c("date","id"))
```

Share Phase

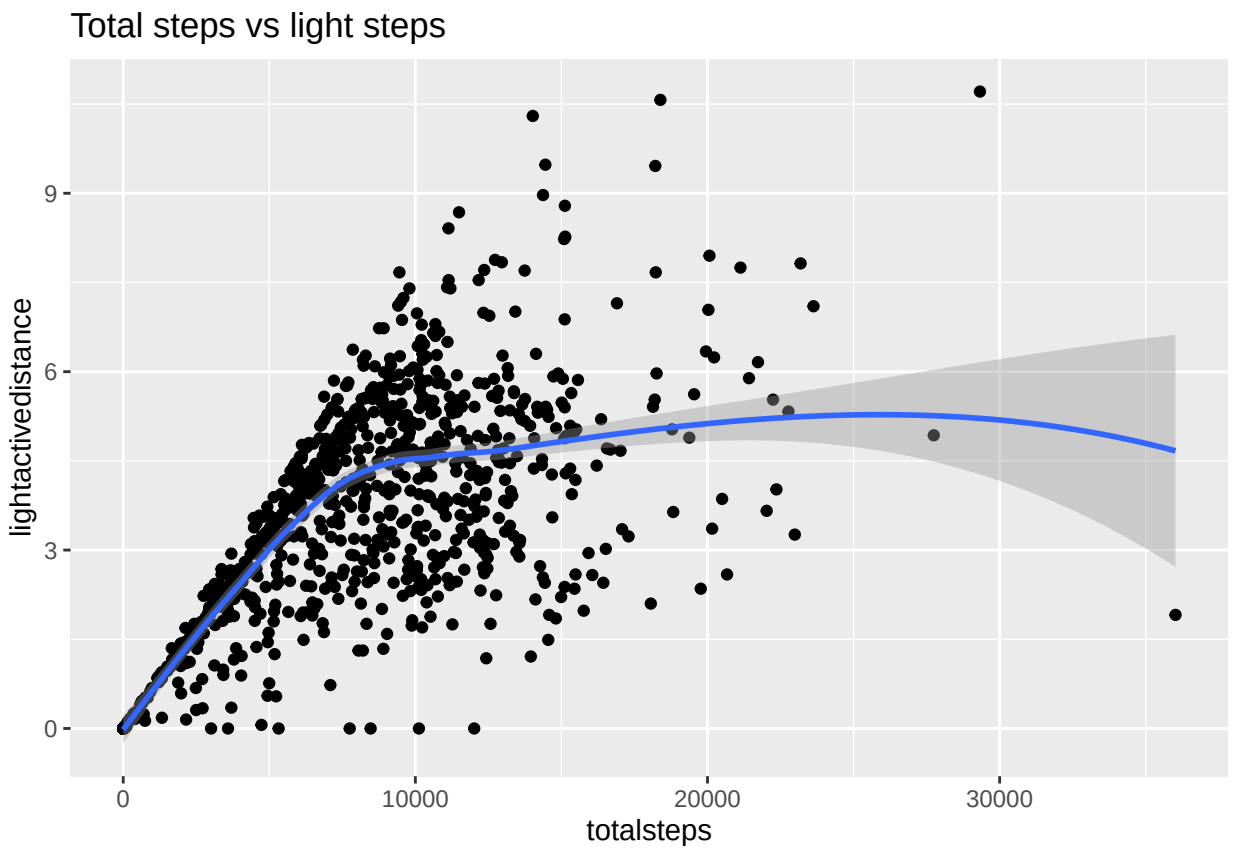
```
str(main_df)
```

```
## tibble [940 x 15] (S3: tbl_df/tbl/data.frame)
##  $ id                       : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ date                     : Date[1:940], format: "2016-04-12" "2016-04-13" ...
##  $ 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 ...
```

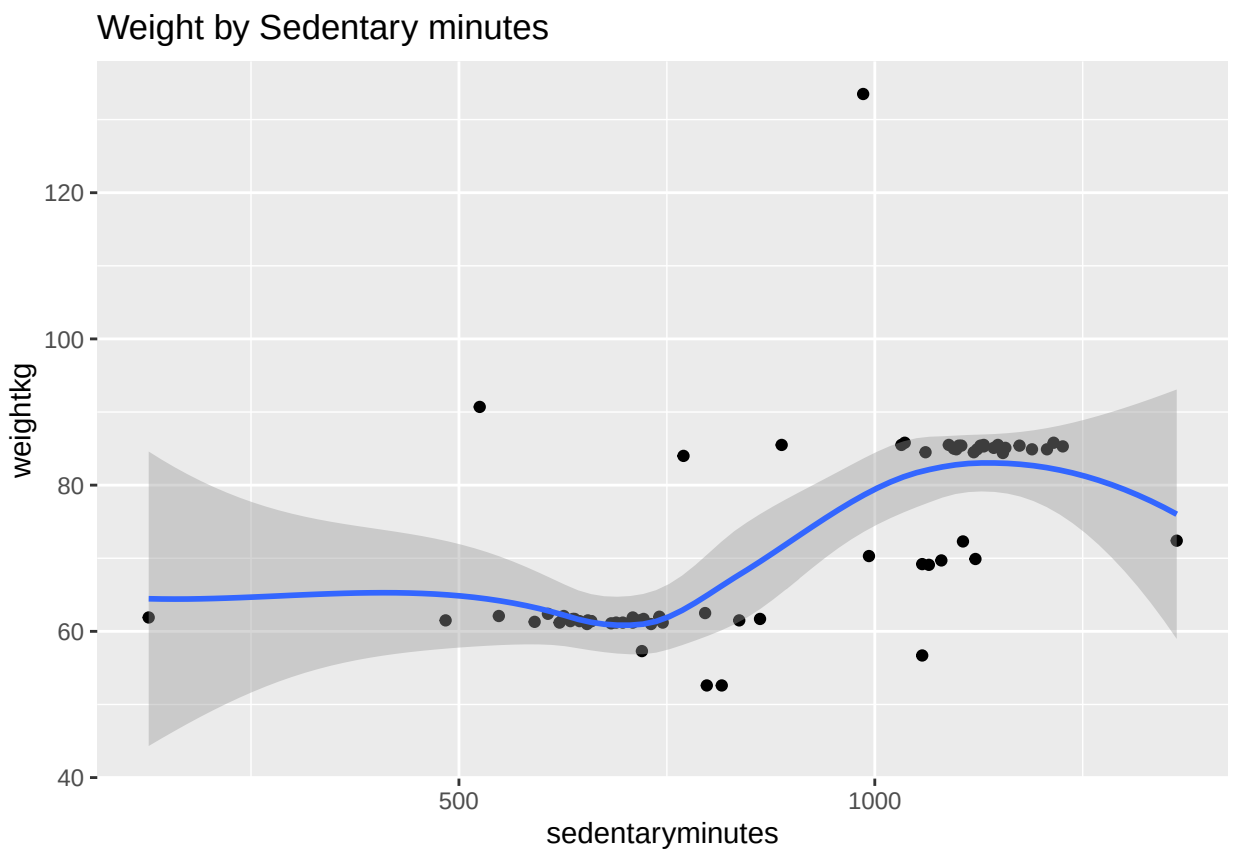
```
ggplot(data=main_df) +
  geom_point(mapping= aes(x=totalsteps,y=lightactivedistance)) +
  geom_smooth(mapping= aes(x=totalsteps,y=lightactivedistance)) +
  labs(title="Total steps vs light steps")
```

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

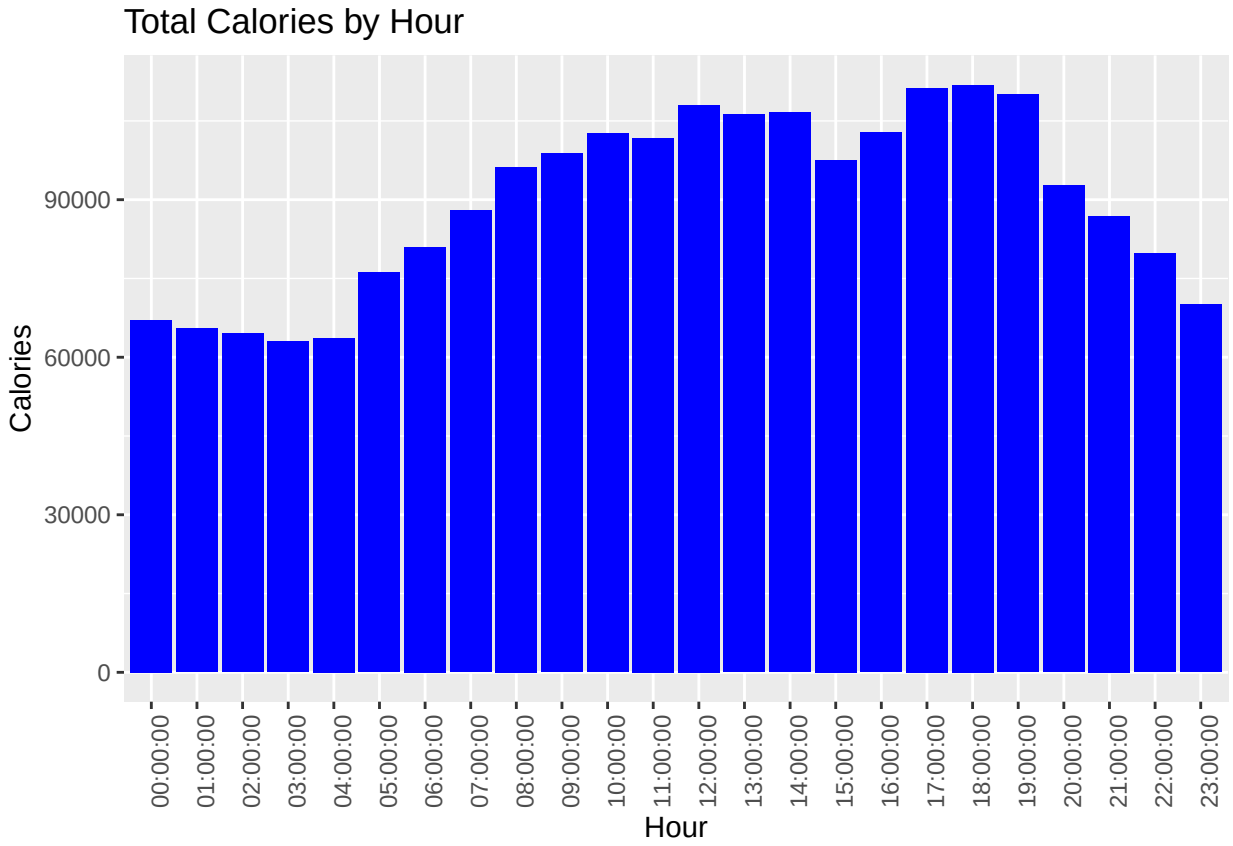



```
ggplot(data=main_weight_df) +
  geom_point(mapping = aes(x = sedentaryminutes ,y = weightkg)) +
  geom_smooth(mapping = aes(x = sedentaryminutes ,y = weightkg)) +
  labs(title = "Weight by Sedentary minutes")
```

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




```
ggplot(cal_show, aes(x = calhours, y = Total_calories)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Total Calories by Hour", x = "Hour", y = "Calories") +
  theme(axis.text.x = element_text(angle = 90))
```



Act Phase

- The graph Total steps vs light steps shows a positive correlation btw the 2 varibles(totalsteps,lightactivedistance) that means people prefer light walking patterns and this can be used to focus marketing strategies promoting light working out with holistic lifestyle rather then promoting intensive workouts.
- The positvie correlation btw sedentary minutes and weight tells that our company's devices should focus more on decreasing the sedentary time of a person to promote reduction in weight if needed and this factor can be used marketing.
- The graph Total Calories by hour clearly shows that the major activity of a person is between 8 AM to 8PM everyday so this data can be used to conclude that marketing adds other then this time would be more beneficial and would have more success