Capstone_Project_2

2022-08-09

Installing Packages

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("magrittr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("fmsb")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
```

```
##     date, intersect, setdiff, union
```

```
library("ggplot2")
library("dplyr")
library("magrittr")
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library("fmsb")
```

Including Plots

```
getwd()
```

```
## [1] "/cloud/project"
```

```
setwd("/cloud/project/Capstone_Project_2")

Daily_Activity <- read_csv("Daily_Activity.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.
```

```
Hourly_Intensities <- read.csv("Hourly_Intensities_merged.csv")
Daily_Calories <- read.csv("Daily_Calories_merged.csv")
Sleep_Day <- read.csv("Sleep_Day_merged.csv")
Hourly_Calories <- read.csv("Hourly_Calories_merged.csv")
```

Exploring a few key tables

```
head(Daily_Activity)
```

```
## # A tibble: 6 x 15
##        Id Activ~1 Total~2 Total~3 Track~4 Logge~5 VeryA~6 Moder~7 Light~8 Seden~9
##     <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 1.50e9 4/12/2~   13162    8.5     8.5         0    1.88   0.550    6.06       0
## 2 1.50e9 4/13/2~   10735    6.97    6.97        0    1.57   0.690    4.71       0
## 3 1.50e9 4/14/2~   10460    6.74    6.74        0    2.44   0.400    3.91       0
## 4 1.50e9 4/15/2~    9762    6.28    6.28        0    2.14   1.26     2.83       0
## 5 1.50e9 4/16/2~   12669    8.16    8.16        0    2.71   0.410    5.04       0
## 6 1.50e9 4/17/2~    9705    6.48    6.48        0    3.19   0.780    2.51       0
## # ... with 5 more variables: VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>, and abbreviated variable names
## #   1: ActivityDate, 2: TotalSteps, 3: TotalDistance, 4: TrackerDistance,
## #   5: LoggedActivitiesDistance, 6: VeryActiveDistance,
## #   7: ModeratelyActiveDistance, 8: LightActiveDistance,
## #   9: SedentaryActiveDistance
## # i Use `colnames()` to see all variable names
```

```
head(Hourly_Intensities)
```

```
##           Id          ActivityHour TotalIntensity AverageIntensity
## 1 1503960366 4/12/2016 12:00:00 AM             20         0.333333
## 2 1503960366  4/12/2016 1:00:00 AM              8         0.133333
## 3 1503960366  4/12/2016 2:00:00 AM              7         0.116667
## 4 1503960366  4/12/2016 3:00:00 AM              0         0.000000
## 5 1503960366  4/12/2016 4:00:00 AM              0         0.000000
## 6 1503960366  4/12/2016 5:00:00 AM              0         0.000000
```

```
head(Daily_Calories)
```

```
##           Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

```
head(Sleep_Day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

```
head(Hourly_Calories)
```

```
##           Id          ActivityHour Calories
## 1 1503960366 4/12/2016 12:00:00 AM       81
## 2 1503960366  4/12/2016 1:00:00 AM       61
## 3 1503960366  4/12/2016 2:00:00 AM       59
## 4 1503960366  4/12/2016 3:00:00 AM       47
## 5 1503960366  4/12/2016 4:00:00 AM       48
## 6 1503960366  4/12/2016 5:00:00 AM       48
```

```r
colnames(Daily_Activity)
```

```
##  [1] "Id"                         "ActivityDate"
##  [3] "TotalSteps"                 "TotalDistance"
##  [5] "TrackerDistance"            "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"         "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"        "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"          "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"       "SedentaryMinutes"
## [15] "Calories"
```

```r
colnames(Hourly_Intensities)
```

```
## [1] "Id"               "ActivityHour"     "TotalIntensity"   "AverageIntensity"
```

```r
colnames(Daily_Calories)
```

```
## [1] "Id"          "ActivityDay" "Calories"
```

```r
colnames(Sleep_Day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

```r
colnames(Hourly_Calories)
```

```
## [1] "Id"           "ActivityHour" "Calories"
```

```r
str(Daily_Activity)
```

```
## spec_tbl_df [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(Hourly_Intensities)
```

```
## 'data.frame':    22099 obs. of  4 variables:
##  $ Id              : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour    : chr  "4/12/2016 12:00:00 AM" "4/12/2016 1:00:00 AM" "4/12/2016 2:00:00 AM" "4/1
##  $ TotalIntensity  : int  20 8 7 0 0 0 0 0 13 30 ...
##  $ AverageIntensity: num  0.333 0.133 0.117 0 0 ...
```

```
str(Daily_Calories)
```

```
## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
str(Sleep_Day)
```

```
## '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(Hourly_Calories)
```

```
## 'data.frame':    22099 obs. of  3 variables:
##  $ Id          : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour: chr  "4/12/2016 12:00:00 AM" "4/12/2016 1:00:00 AM" "4/12/2016 2:00:00 AM" "4/12/20
##  $ Calories    : int  81 61 59 47 48 48 48 47 68 141 ...
```

Converging Data format

```
Daily_Activity$date <- as.POSIXct( Daily_Activity$ActivityDate, format="%m/%d/%Y" )
Daily_Activity$day <- wday(Daily_Activity$date, label=TRUE)

Daily_Calories$date <- as.POSIXct( Daily_Calories$ActivityDay, format="%m/%d/%Y" )
Daily_Calories$day <- wday(Daily_Calories$date, label=TRUE)

Sleep_Day$date <- as.POSIXct( Sleep_Day$SleepDay, format="%m/%d/%Y" )
Sleep_Day$day <- wday(Sleep_Day$date, label=TRUE)
```

Identifiying the number of users

```
n_distinct(Daily_Activity$Id)
```

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

```
n_distinct(Sleep_Day$Id)
```

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

```
n_distinct(Daily_Calories$Id)
```

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

```
n_distinct(Hourly_Calories$Id)
```

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

```
n_distinct(Hourly_Intensities$Id)
```

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

```
nrow(Daily_Activity)
```

```
## [1] 940
```

```
nrow(Sleep_Day)
```

```
## [1] 413
```

```
nrow(Daily_Calories)
```

```
## [1] 940
```

```
nrow(Hourly_Calories)
```

```
## [1] 22099
```

```
nrow(Hourly_Intensities)
```

```
## [1] 22099
```

```
Daily_Activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

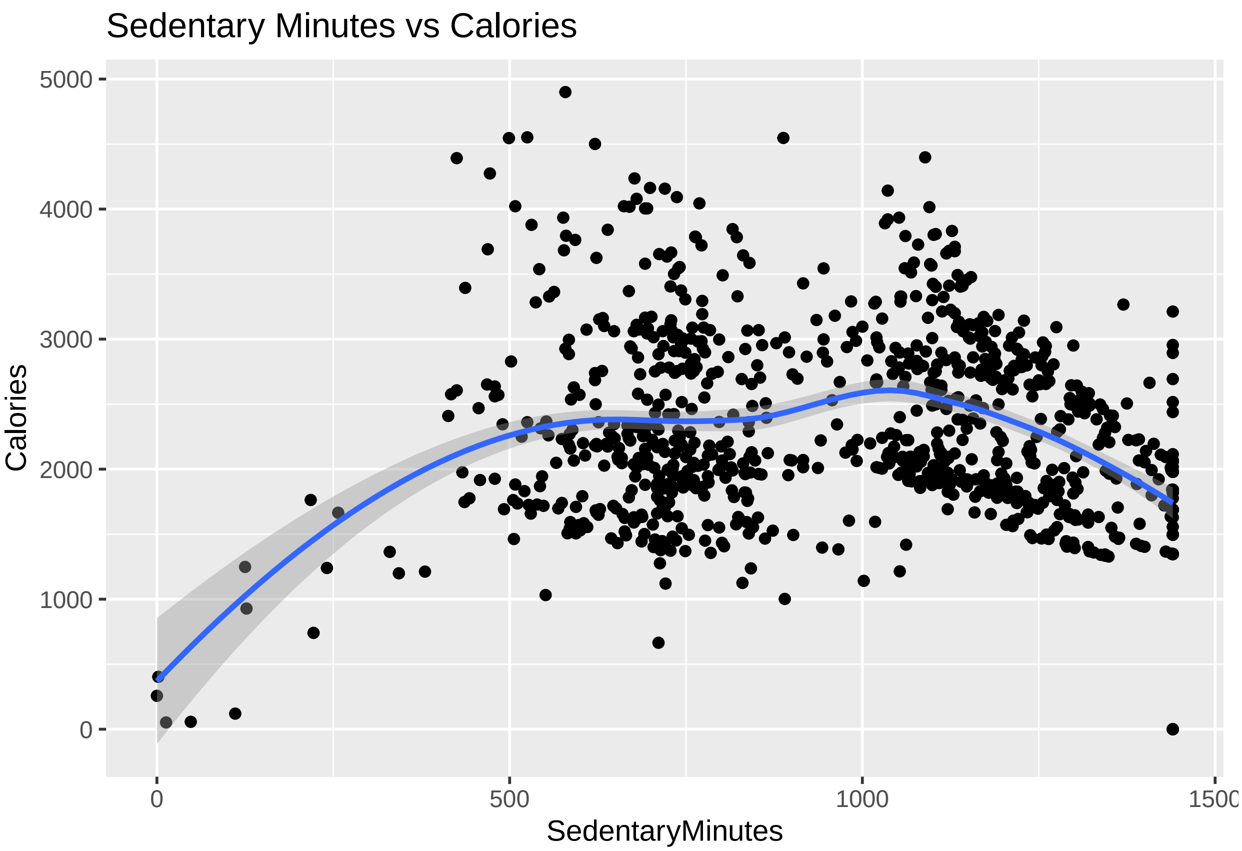
```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8
##  Median : 7406   Median : 5.245   Median :1057.5
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2
##  3rd Qu.:10727   3rd Qu.: 7.713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

Plotting a few explorations

What's the relationship between steps taken in a day and sedentary minutes? How could this help inform the customer segments that we can market to? E.g. position this more as a way to get started in walking more? Or to measure steps that you're already taking?

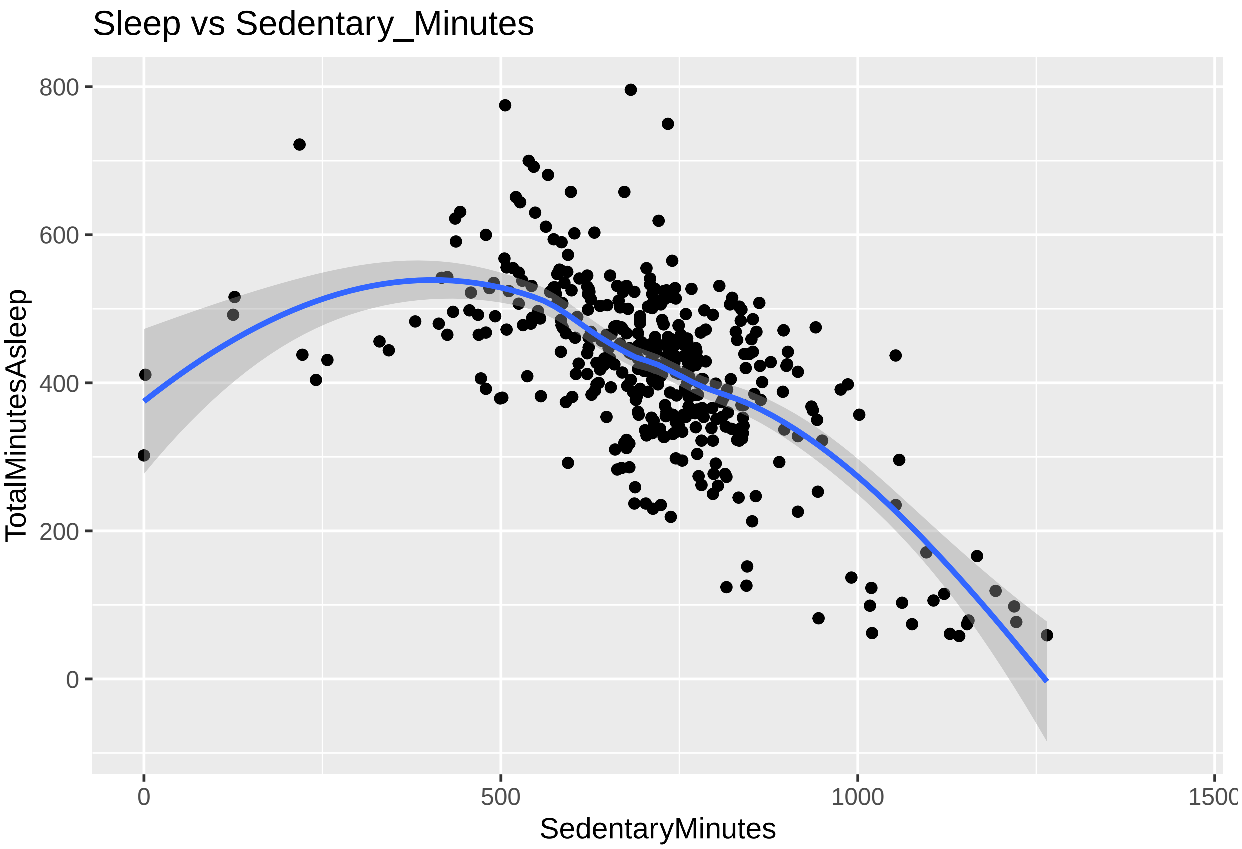
```
ggplot(data=Daily_Activity) +
  geom_point(mapping=aes(x=SedentaryMinutes, y=Calories)) +
  geom_smooth(mapping=aes(x=SedentaryMinutes, y=Calories)) +
  labs(title="Sedentary Minutes vs Calories")
```

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



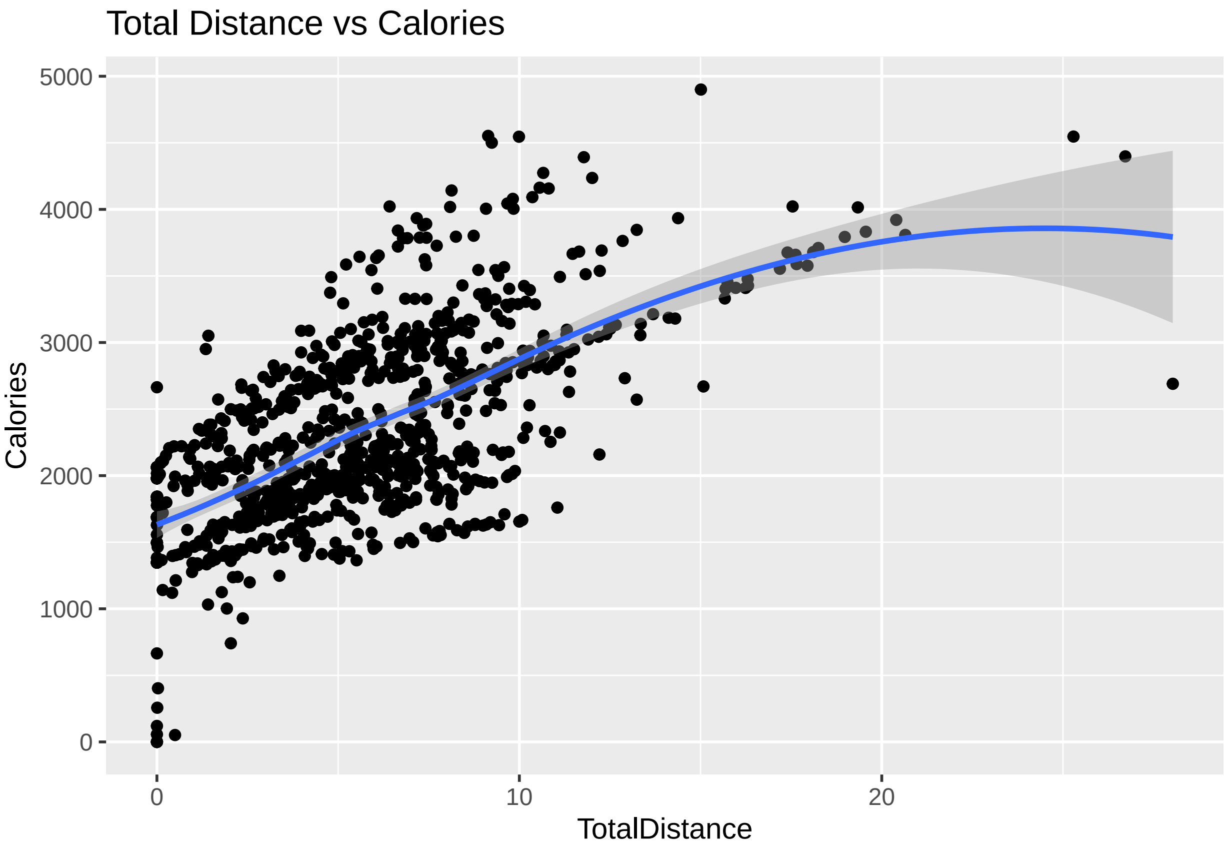
```
merged_data <- Daily_Activity %>%
  left_join(Sleep_Day , by = c('Id', 'date'))

ggplot(data = merged_data) +
  geom_point(mapping=aes(x=SedentaryMinutes,y = TotalMinutesAsleep))+
  geom_smooth(mapping=aes(x=SedentaryMinutes,y = TotalMinutesAsleep))+
  labs(title = "Sleep vs Sedentary_Minutes")
```

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

## Warning: Removed 530 rows containing non-finite values (stat_smooth).

## Warning: Removed 530 rows containing missing values (geom_point).
```



```
ggplot(data=Daily_Activity) +
  geom_point(mapping=aes(x=TotalDistance, y=Calories)) +
  geom_smooth(mapping=aes(x=TotalDistance, y=Calories)) +
  labs(title="Total Distance vs Calories")
```

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



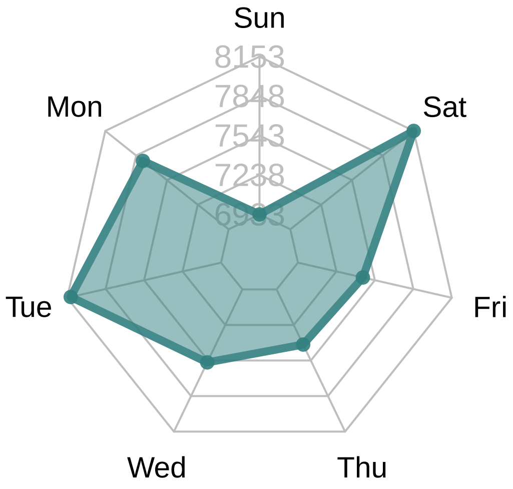
Average number of steps per day of the week

```
steps_results <- Daily_Activity %>%
select(TotalSteps, day) %>%
group_by(day) %>%
summarise(mean_steps = mean(TotalSteps, na.rm = TRUE))

steps_results_t <- t(steps_results)
colnames(steps_results_t) <- steps_results_t[1, ]
steps_results_t = steps_results_t[-1,]
steps_results_t <- t(steps_results_t)
steps_results_t <- as.data.frame(steps_results_t)

steps_results_t$Sun <- as.numeric(steps_results_t$Sun)
steps_results_t$Mon <- as.numeric(steps_results_t$Mon)
steps_results_t$Tue <- as.numeric(steps_results_t$Tue)
steps_results_t$Wed <- as.numeric(steps_results_t$Wed)
steps_results_t$Thu <- as.numeric(steps_results_t$Thu)
steps_results_t$Fri <- as.numeric(steps_results_t$Fri)
steps_results_t$Sat <- as.numeric(steps_results_t$Sat)

min_d <- min(steps_results_t)
max_d <- max(steps_results_t)

steps_results_t[nrow(steps_results_t) + 1,] <- c(max_d, max_d, max_d, max_d, max_d, max_d, max_d)
steps_results_t[nrow(steps_results_t) + 1,] <- c(min_d, min_d, min_d, min_d, min_d, min_d, min_d)
```

```
steps_results_t <- steps_results_t[c(2,3,1),]

steps_results_t <- round(steps_results_t, digits=0)
```

```
radarchart(steps_results_t, axistype=1,
pcol=rgb(0.2,0.5,0.5,0.9) , pfcol=rgb(0.2,0.5,0.5,0.5) , plwd=4 , cglcol="grey", cglty=1, axislabcol="g
```

Average number of steps taken per day

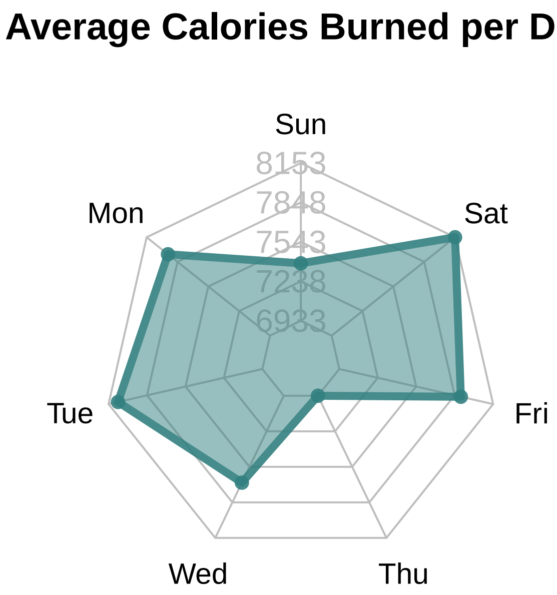


### Average Calories burned per day of the week

```
calories_burned <- merged_data %>%
select(Calories, day.x) %>%
group_by(day.x) %>%
summarise(mean_calories = mean(Calories, na.rm = TRUE))

calories_burned_t <- t(calories_burned)
colnames(calories_burned_t) <- calories_burned_t[1, ]
calories_burned_t = calories_burned_t[-1,]
calories_burned_t <- t(calories_burned_t)
calories_burned_t <- as.data.frame(calories_burned_t)

calories_burned_t$Sun <- as.numeric(calories_burned_t$Sun)
calories_burned_t$Mon <- as.numeric(calories_burned_t$Mon)
calories_burned_t$Tue <- as.numeric(calories_burned_t$Tue)
calories_burned_t$Wed <- as.numeric(calories_burned_t$Wed)
calories_burned_t$Thu <- as.numeric(calories_burned_t$Thu)
calories_burned_t$Fri <- as.numeric(calories_burned_t$Fri)
calories_burned_t$Sat <- as.numeric(calories_burned_t$Sat)

min_d <- min(calories_burned_t)
max_d <- max(calories_burned_t)

calories_burned_t[nrow(calories_burned_t) + 1,] <- c(max_d, max_d, max_d, max_d, max_d, max_d, max_d)
calories_burned_t[nrow(calories_burned_t) + 1,] <- c(min_d, min_d, min_d, min_d, min_d, min_d, min_d)

calories_burned_t <- calories_burned_t[c(2,3,1),]
```

```r
calories_burned_t <- round(calories_burned_t, digits=0)
```

```r
radarchart(calories_burned_t, axistype=1,
pcol=rgb(0.2,0.5,0.5,0.9) , pfcol=rgb(0.2,0.5,0.5,0.5) , plwd=4 , cglcol="grey", cglty=1, axislabcol="g
```



“‘

Background

In the dataset provided I focused on a few key variables that tried to understand if these variables are correlated positively or negatively. I started installing the packages, loading the data sets, reading the data and renaming the columns accordingly, converging data and date format, plotting different comparative graphs to understand correlation.

Improvement points

- Date format is different for each data set, suggestion is to have a standard format of the record.
- Data of sleeping hours are not completed or disclosed by users.

Summary of the analysis

Sedentary Minutes vs Calories - After plotting the chart, we can see that as the amount of calories burnt does not increase when the sedentary minutes increase, meaning that these variables are not strong correlated.

Sleep vs Sedentary_Minutes - Tried to find a correlation between the total minutes sleeping vs the sedentary minutes and we cannot confirm if we have users sleeping more, that will also mean that they have more minutes as sedentaries. In fact, we see a few outliers in the graph, showing users that sleep up to 200 minutes have many more minutes as sedentaries than others (>1000 minutes).

Total Distance vs Calories - We can see a strong correlation between distance walked and calories burned out. It is a good indicative that the data collected correct. The comparison shows that if users walk more, they will burn more calories.

Average number of steps taken per day - The chart shows us what is the average steps taken per day of the week. Thursday, Friday and Sundays are the days of the week with the lowest number of steps taken per day.

Average Calories Burned per Day - This chart shows us the average calories burned per day of the week. We can see that Thursdays we have a a decrease compared to the rest of the week.

Recommendations

- Notify users and provide recommendations to excercise during sedentary minutes. Text message to users to remind them that they have been sedentary for too long.
- The Bellabeat app application could be sending messages and tips of healthy snack type food.
- It would be interesting to have a function that helps the users to build a routine with a fixed sleeping and scheduled exercises.
- The app could also plot comparative of graphs to the users, so they can track their progress throughout the month showing how many stepsn were taken in a weekly basis, calories burned x steps taken and sleeping time during the week.