google data analytics capstone project

Najith

1/23/2022

Fitness Tracer data Analysis

Purpose of this analysis

To analyze smart device usage data in order to gain insight into how consumers use these products to track their activity.

This dataset used in this analysis is generated by respondents to a distributed survey via Amazon Mechanical Turk. Thirty three eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.

Importing Packages and data

importing all the needed libraries for data wrangling and data analysis.

Reading the dataset (which are csv files) and converting the data into tibbles.

```
activity <- 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.
```

```
calories <- read_csv("dailyCalories_merged.csv")
```

```
## Rows: 940 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## 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.
```

```
intensities <- read_csv("dailyIntensities_merged.csv")
```

```
## Rows: 940 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (9): Id, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, Ve...

##
## 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.
```

```
steps <- read_csv("dailySteps_merged.csv")
```

```
## Rows: 940 Columns: 3

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (2): Id, StepTotal

##
## 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 <- 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 <- 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.
```

Understanding the data

Lets start with exploring the data. We take a look at the data that we have.We are checking if there is any duplicates or null in the data.

```
head(activity)
```

```
## # A tibble: 6 x 15
##        Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~
##     <dbl> <chr>             <dbl>         <dbl>           <dbl>            <dbl>
## 1  1.50e9 4/12/2016         13162          8.5             8.5                 0
```

```
## 2  1.50e9 4/13/2016         10735          6.97           6.97                0
## 3  1.50e9 4/14/2016         10460          6.74           6.74                0
## 4  1.50e9 4/15/2016          9762          6.28           6.28                0
## 5  1.50e9 4/16/2016         12669          8.16           8.16                0
## 6  1.50e9 4/17/2016          9705          6.48           6.48                0
## # ... with 9 more variables: VeryActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>
```

```
head(calories)
```

```
## # A tibble: 6 x 3
##           Id ActivityDay Calories
##        <dbl> <chr>          <dbl>
## 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(intensities)
```

```
## # A tibble: 6 x 10
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes FairlyActiveMinu~
##        <dbl> <chr>                  <dbl>                 <dbl>              <dbl>
## 1 1503960366 4/12/2016                728                   328                 13
## 2 1503960366 4/13/2016                776                   217                 19
## 3 1503960366 4/14/2016               1218                   181                 11
## 4 1503960366 4/15/2016                726                   209                 34
## 5 1503960366 4/16/2016                773                   221                 10
## 6 1503960366 4/17/2016                539                   164                 20
## # ... with 5 more variables: VeryActiveMinutes <dbl>,
## #   SedentaryActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, VeryActiveDistance <dbl>
```

```
head(steps)
```

```
## # A tibble: 6 x 3
##           Id ActivityDay StepTotal
##        <dbl> <chr>           <dbl>
## 1 1503960366 4/12/2016       13162
## 2 1503960366 4/13/2016       10735
## 3 1503960366 4/14/2016       10460
## 4 1503960366 4/15/2016        9762
## 5 1503960366 4/16/2016       12669
## 6 1503960366 4/17/2016        9705
```

```
head(sleep)
```

```
## # A tibble: 6 x 5
##           Id SleepDay          TotalSleepRecor~ TotalMinutesAsle~ TotalTimeInBed
##        <dbl> <chr>                        <dbl>             <dbl>          <dbl>
## 1 1503960366 4/12/2016 12:00:~                1               327            346
## 2 1503960366 4/13/2016 12:00:~                2               384            407
```

```
## 3 1503960366 4/15/2016 12:00:~                1              412            442
## 4 1503960366 4/16/2016 12:00:~                2              340            367
## 5 1503960366 4/17/2016 12:00:~                1              700            712
## 6 1503960366 4/19/2016 12:00:~                1              304            320
```

```
head(weight)
```

```
## # A tibble: 6 x 8
##           Id Date      WeightKg WeightPounds   Fat   BMI IsManualReport    LogId
##        <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl> <lgl>              <dbl>
## 1 1503960366 5/2/2016~     52.6         116.    22  22.6 TRUE            1.46e12
## 2 1503960366 5/3/2016~     52.6         116.    NA  22.6 TRUE            1.46e12
## 3 1927972279 4/13/201~    134.          294.    NA  47.5 FALSE           1.46e12
## 4 2873212765 4/21/201~     56.7         125.    NA  21.5 TRUE            1.46e12
## 5 2873212765 5/12/201~     57.3         126.    NA  21.7 TRUE            1.46e12
## 6 4319703577 4/17/201~     72.4         160.    25  27.5 TRUE            1.46e12
```

```
str(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(calories)
```

```
## spec_tbl_df [940 x 3] (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 ...
##  $ ActivityDay: chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDay = col_character(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(intensities)
```

```
## spec_tbl_df [940 x 10] (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 ...
##  $ ActivityDay              : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ SedentaryMinutes         : num [1:940] 728 776 1218 726 773 ...
##  $ LightlyActiveMinutes     : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ FairlyActiveMinutes      : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ VeryActiveMinutes        : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ LightActiveDistance      : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ VeryActiveDistance       : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDay = col_character(),
##   ..   SedentaryMinutes = col_double(),
##   ..   LightlyActiveMinutes = col_double(),
##   ..   FairlyActiveMinutes = col_double(),
##   ..   VeryActiveMinutes = col_double(),
##   ..   SedentaryActiveDistance = col_double(),
##   ..   LightActiveDistance = col_double(),
##   ..   ModeratelyActiveDistance = col_double(),
##   ..   VeryActiveDistance = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(steps)
```

```
## spec_tbl_df [940 x 3] (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 ...
##  $ ActivityDay: chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ StepTotal  : num [1:940] 13162 10735 10460 9762 12669 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDay = col_character(),
##   ..   StepTotal = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(sleep)
```

```
## spec_tbl_df [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 12:00:0
##  $ 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)
```

```
## spec_tbl_df [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>
```

```
colnames(activity)
```

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

```
colnames(calories)
```

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

```
colnames(intensities)
```

```
##  [1] "Id"                       "ActivityDay"
##  [3] "SedentaryMinutes"         "LightlyActiveMinutes"
##  [5] "FairlyActiveMinutes"      "VeryActiveMinutes"
##  [7] "SedentaryActiveDistance"  "LightActiveDistance"
##  [9] "ModeratelyActiveDistance" "VeryActiveDistance"
```

```
colnames(steps)
```

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

```
colnames(sleep)
```

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

```
colnames(weight)
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

```
skim_without_charts(activity)
```

Table 1: Data summary

Name	activity
Number of rows	940
Number of columns	15
Column type frequency:	
character	1
numeric	14
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ActivityDate	0	1	8	9	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.855407e+09	2.924805e+09	1503960366	2320127e+09	4.445115e+09	6.962181e+09	8.877689e+09
TotalSteps	0	1	7.637910e+03	5.087150e+03	0	3.789750e+03	7.405500e+03	1.0727700e+04	3.6401900e+04
TotalDistance	0	1	5.490000e+00	3.920000e+00	0	2.620000e+00	5.240000e+00	7.710000e+00	2.8030000e+01
TrackerDistance	0	1	5.480000e+00	3.910000e+00	0	2.620000e+00	5.240000e+00	7.710000e+00	2.8030000e+01
LoggedActivitiesDistance	0	1	1.100000e-01	6.200000e-01	0	0.000000e+00	0.000000e+00	0.000000e+00	4.940000e+00
VeryActiveDistance	0	1	1.500000e+00	2.060000e+00	0	0.000000e+00	0.000000e-01	2.050000e+00	2.092000e+01
ModeratelyActiveDistance	0	1	5.700000e-01	8.800000e-01	0	0.000000e+00	0.000000e-01	8.000000e-01	6.480000e+00

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
LightActiveDistance	0	1	3.340000e+00	2.040000e+00	0	1.950000e+00	3.280000e+00	4.780000e+00	1.071000e+01
SedentaryActiveDistance	0	1	0.000000e+00	1.000000e-02	0	0.000000e+00	0.000000e+00	0.000000e+00	1.100000e-01
VeryActiveMinutes	0	1	2.116000e+01	2.284000e+01	0	0.000000e+00	4.000000e+00	3.200000e+01	2.100000e+02
FairlyActiveMinutes	0	1	1.356000e+01	1.999000e+01	0	0.000000e+00	6.000000e+00	1.900000e+01	1.430000e+02
LightlyActiveMinutes	0	1	1.928100e+02	1.091700e+02	0	1.270000e+02	1.990000e+02	2.640000e+02	5.180000e+02
SedentaryMinutes	0	1	9.912100e+02	3.012700e+02	0	7.297500e+02	1.057500e+03	1.329500e+03	1.440000e+03
Calories	0	1	2.303610e+03	7.181700e+02	0	1.828500e+03	2.134000e+03	2.793250e+03	4.900000e+03

```
skim_without_charts(calories)
```

Table 4: Data summary

Name	calories
Number of rows	940
Number of columns	3
Column type frequency:	
character	1
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ActivityDay	0	1	8	9	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.855407e+09	2.424805e+09	1503960366	2320127002	4445114986	6.962181e+09	8877689391
Calories	0	1	2.303610e+03	7.181700e+02	0	1828.5	2134	2.793250e+03	4900

```
skim_without_charts(intensities)
```

Table 7: Data summary

Name	intensities
Number of rows	940
Number of columns	10
Column type frequency:	
character	1
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ActivityDay	0	1	8	9	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.855407e+09	2.424805e+09	1503960366	2.320127e+09	4.445115e+09	6.962181e+09	8.877689e+09
SedentaryMinutes	0	1	9.912100e+02	3.012700e+02	0	7.297500e+02	1.057500e+03	1.229500e+03	1.440000e+03
LightlyActiveMinutes	0	1	1.928100e+02	1.091700e+02	0	1.270000e+02	1.990000e+02	2.640000e+02	5.128000e+02
FairlyActiveMinutes	0	1	1.356000e+01	9.999000e+01	0	0.000000e+00	0.000000e+00	1.000000e+01	1.143000e+02
VeryActiveMinutes	0	1	2.116000e+02	2.284000e+01	0	0.000000e+00	4.000000e+00	3.200000e+02	2.110000e+02
SedentaryActiveDistance	0	1	0.000000e+00	1.000000e-02	0	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e-01
LightActiveDistance	0	1	3.340000e+00	2.040000e+00	0	1.950000e+00	3.360000e+00	4.780000e+00	1.071000e+01
ModeratelyActiveDistance	0	1	5.700000e-01	8.800000e-01	0	0.000000e+00	2.000000e-01	8.000000e-01	6.480000e+00
VeryActiveDistance	0	1	1.500000e+00	2.060000e+00	0	0.000000e+00	2.000000e-01	2.050000e+00	2.092000e+01

```
skim_without_charts(steps)
```

Table 10: Data summary

Name	steps
Number of rows	940
Number of columns	3
Column type frequency:	
character	1
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ActivityDay	0	1	8	9	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.855407e+09	2.424805e+09	1503960366	2.320127e+09	4.445115e+09	6.962181e+09	8.877689e+09
StepTotal	0	1	7.637910e+03	5.087150e+03	0	3.789750e+03	7405.5	10727	36019

```
skim_without_charts(sleep)
```

Table 13: Data summary

Name	sleep
Number of rows	413
Number of columns	5
Column type frequency:	
character	1
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SleepDay	0	1	20	21	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	5.000979e+09	2.906036e+09	1503960366	2977333714	4702921684	6962181067	8792009665
TotalSleepRecords	0	1	1.120000e+00	3.050000e-01	1	1	1	1	3
TotalMinutesAsleep	0	1	4.194700e+02	1.218340e+02	58	361	433	490	796
TotalTimeInBed	0	1	4.586400e+02	1.227100e+02	61	403	463	526	961

```
skim_without_charts(weight)
```

Table 16: Data summary

Name	weight
Number of rows	67
Number of columns	8
Column type frequency:	
character	1
logical	1
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Date	0	1	19	21	0	56	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
IsManualReport	0	1	0.61	TRU: 41, FAL: 26

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1.00	7.009282e+09	2.950322e+09	1.503960e+09	4.962181e+09	4.962181e+09	8.877689e+09	8.877689e+09
WeightKg	0	1.00	7.204000e+01	3.392000e+01	5.260000e+01	6.140000e+01	6.250000e+01	8.505000e+01	1.335000e+02
WeightPounds	0	1.00	1.588100e+02	7.477000e+01	1.159600e+02	1.353600e+02	1.377900e+02	1.875000e+02	2.943200e+02
Fat	65	0.03	2.350000e+01	1.120000e+01	2.200000e+01	2.275000e+01	2.350000e+01	2.425000e+01	2.500000e+01
BMI	0	1.00	2.519000e+01	3.107000e+00	2.145000e+01	2.396000e+01	2.439000e+01	2.556000e+01	4.754000e+01
LogId	0	1.00	1.461772e+12	2.829948e+08	1.460444e+12	1.461079e+12	1.461802e+12	1.462375e+12	1.463098e+12

Data Cleaning

We remove the duplicates in the dataset, so we use n_distinct to check distinct number of ids and we check for nulls and duplicate in the dataset.

```
n_distinct(activity$Id)
```

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

```
n_distinct(calories$Id)
```

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

```
n_distinct(intensities$Id)
```

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

```
n_distinct(steps$Id)
```

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

```
n_distinct(sleep$Id)
```

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

```
n_distinct(weight$Id)
```

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

```
sum(duplicated(activity))
```

```
## [1] 0
```

```
sum(is.na(activity))
```

```
## [1] 0
```

```
sum(duplicated(calories))
```

```
## [1] 0
```

```
sum(is.na(calories))
```

```
## [1] 0
```

```
sum(duplicated(intensities))
```

```
## [1] 0
```

```
sum(is.na(intensities))
```

```
## [1] 0
```

```
sum(duplicated(steps))
```

```
## [1] 0
```

```
sum(is.na(steps))
```

```
## [1] 0
```

```
sum(duplicated(sleep))
```

```
## [1] 3
```

```
sum(is.na(sleep))
```

```
## [1] 0
```

```
sum(duplicated(weight))
```

```
## [1] 0
```

```
sum(is.na(weight))
```

```
## [1] 65
```

We see that there is duplicates in the sleep dataset so we should remove the duplicate

```
sleep <- distinct(sleep)
sum(duplicated(sleep))
```

```
## [1] 0
```

From dataset we see that the date and time in the dataset is of different format or sometime in the same column. Lets change the date in the dataset to the same format for all the tibbles and create a new column for time and have a uniform format.

```
activity_v2 <- activity %>%
  mutate(ActivityDate = mdy(ActivityDate)) %>%
  rename(date=ActivityDate)
steps_v2 <- steps %>%
  mutate(ActivityDay = mdy(ActivityDay)) %>%
  rename(date=ActivityDay)
intensities_v2 <- intensities %>%
  mutate(ActivityDay = mdy(ActivityDay)) %>%
  rename(date=ActivityDay)
calories_v2 <- calories %>%
  mutate(ActivityDate = mdy(ActivityDay)) %>%
  rename(date=ActivityDay)
sleep_v2 <- sleep %>%
  mutate(SleepDay = mdy_hms(SleepDay),date = as.Date(SleepDay),
  time = format(SleepDay, "%T"))
weight_v2 <- weight %>%
```

```
  mutate(Date = mdy_hms(Date), date= as.Date(Date),
         time = format(Date, "%T"))
```

Delete the columns which contains the old date format and time. Check the number of rows after the duplicates are deleted.

```
sleep_v3 <- sleep_v2
weight_v3 <- weight_v2

sleep_v3[,c("SleepDay")] <- list(NULL)
weight_v3[,c("Date")] <- list(NULL)

head(sleep_v3)
```

```
## # A tibble: 6 x 6
##           Id TotalSleepRecords TotalMinutesAsle~ TotalTimeInBed date       time
##        <dbl>             <dbl>             <dbl>          <dbl> <date>     <chr>
## 1 1503960366                 1               327            346 2016-04-12 00:0~
## 2 1503960366                 2               384            407 2016-04-13 00:0~
## 3 1503960366                 1               412            442 2016-04-15 00:0~
## 4 1503960366                 2               340            367 2016-04-16 00:0~
## 5 1503960366                 1               700            712 2016-04-17 00:0~
## 6 1503960366                 1               304            320 2016-04-19 00:0~
```

```
head(weight_v3)
```

```
## # A tibble: 6 x 9
##           Id WeightKg WeightPounds   Fat   BMI IsManualReport   LogId date
##        <dbl>    <dbl>        <dbl> <dbl> <dbl> <lgl>            <dbl> <date>
## 1 1503960366     52.6         116.    22  22.6 TRUE           1.46e12 2016-05-02
## 2 1503960366     52.6         116.    NA  22.6 TRUE           1.46e12 2016-05-03
## 3 1927972279    134.          294.    NA  47.5 FALSE          1.46e12 2016-04-13
## 4 2873212765     56.7         125.    NA  21.5 TRUE           1.46e12 2016-04-21
## 5 2873212765     57.3         126.    NA  21.7 TRUE           1.46e12 2016-05-12
## 6 4319703577     72.4         160.    25  27.5 TRUE           1.46e12 2016-04-17
## # ... with 1 more variable: time <chr>
```

```
nrow(sleep_v3)
```

```
## [1] 410
```

```
nrow(weight_v3)
```

```
## [1] 67
```

```
nrow(activity_v2)
```

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

Merging the activity and sleep dataset into a single dataset called activity_sleep for easy use.

```
activity_sleep <- merge(activity_v2, sleep_v3, by= c("Id","date"), all=TRUE)
nrow(activity_sleep)
```

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

```
activity_sleep_wt <- merge(activity_sleep, weight_v3, by = c("Id", "date"), all=TRUE)
nrow(activity_sleep_wt)
```

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

```
head(activity_sleep_wt)
```

```
##           Id       date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2016-04-12      13162          8.50            8.50
## 2 1503960366 2016-04-13      10735          6.97            6.97
## 3 1503960366 2016-04-14      10460          6.74            6.74
## 4 1503960366 2016-04-15       9762          6.28            6.28
## 5 1503960366 2016-04-16      12669          8.16            8.16
## 6 1503960366 2016-04-17       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
##   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   time.x WeightKg
## 1                 1                327            346 00:00:00       NA
## 2                 2                384            407 00:00:00       NA
## 3                NA                 NA             NA     <NA>       NA
## 4                 1                412            442 00:00:00       NA
## 5                 2                340            367 00:00:00       NA
## 6                 1                700            712 00:00:00       NA
##   WeightPounds Fat BMI IsManualReport LogId time.y
## 1           NA  NA  NA             NA    NA   <NA>
## 2           NA  NA  NA             NA    NA   <NA>
## 3           NA  NA  NA             NA    NA   <NA>
## 4           NA  NA  NA             NA    NA   <NA>
## 5           NA  NA  NA             NA    NA   <NA>
## 6           NA  NA  NA             NA    NA   <NA>
```

Add the weekdays column to the newly created dataset using mutate and creating a completely new dataset as activity_final.

```
activity_final <- activity_sleep_wt %>%
  mutate(day = weekdays(date))
head(activity_final)
```

```
##           Id       date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2016-04-12      13162          8.50            8.50
```

```
## 2 1503960366 2016-04-13      10735          6.97           6.97
## 3 1503960366 2016-04-14      10460          6.74           6.74
## 4 1503960366 2016-04-15       9762          6.28           6.28
## 5 1503960366 2016-04-16      12669          8.16           8.16
## 6 1503960366 2016-04-17       9705          6.48           6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
##   TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   time.x WeightKg
## 1                 1                327            346 00:00:00       NA
## 2                 2                384            407 00:00:00       NA
## 3                NA                 NA             NA     <NA>       NA
## 4                 1                412            442 00:00:00       NA
## 5                 2                340            367 00:00:00       NA
## 6                 1                700            712 00:00:00       NA
##   WeightPounds Fat BMI IsManualReport LogId time.y       day
## 1           NA  NA  NA             NA    NA   <NA>   Tuesday
## 2           NA  NA  NA             NA    NA   <NA> Wednesday
## 3           NA  NA  NA             NA    NA   <NA>  Thursday
## 4           NA  NA  NA             NA    NA   <NA>    Friday
## 5           NA  NA  NA             NA    NA   <NA>  Saturday
## 6           NA  NA  NA             NA    NA   <NA>    Sunday
```

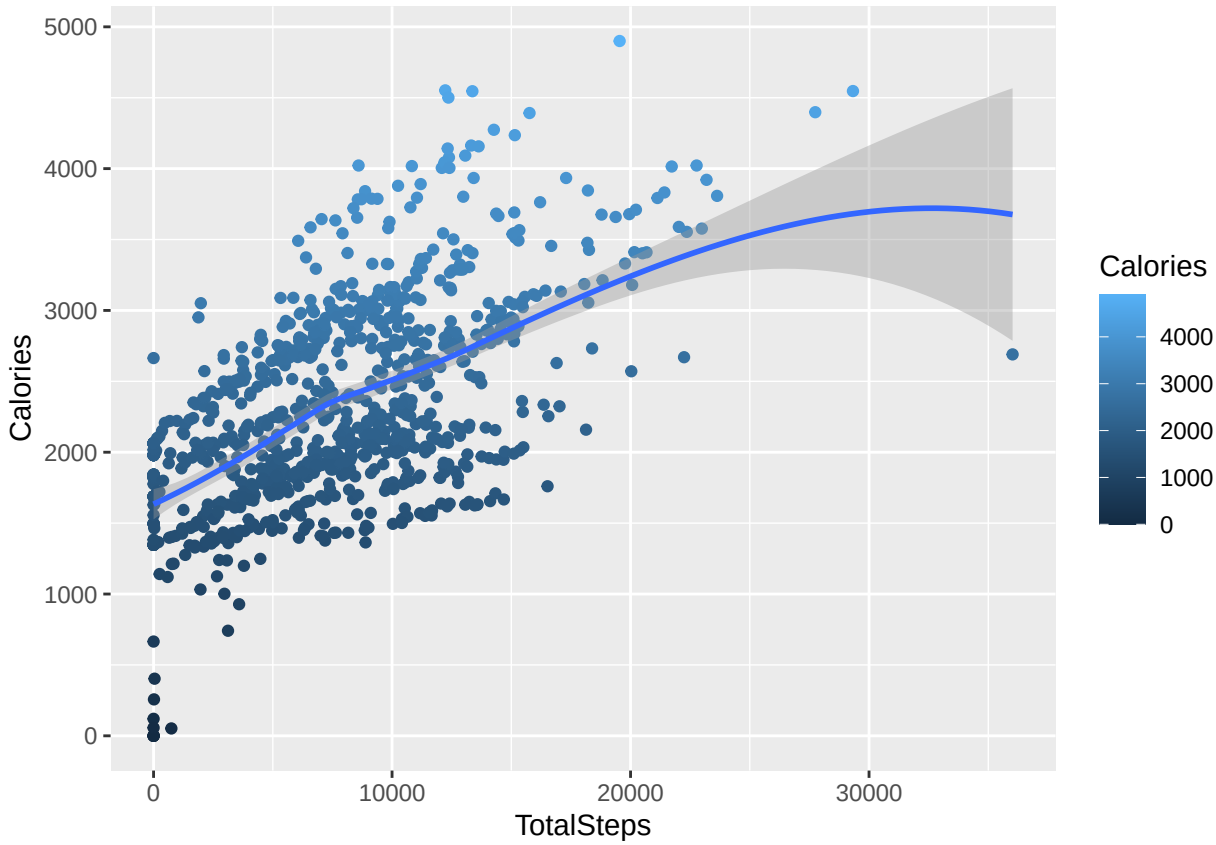
Data Analysis

Total steps vs Calories burnt

The hypothesis here is that the more steps is taken a day the more calories are burnt.

```
ggplot(data=activity_final, aes(x=TotalSteps,
                y= Calories,
                color=Calories)) +
  geom_point() +
  geom_smooth()
```

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



Conclusion: We see that the hypothesis is true and the Calories are burnt more the more total steps taken by the people is.

Average steps VS Average sleep VS Average Calories burnt on weekdays

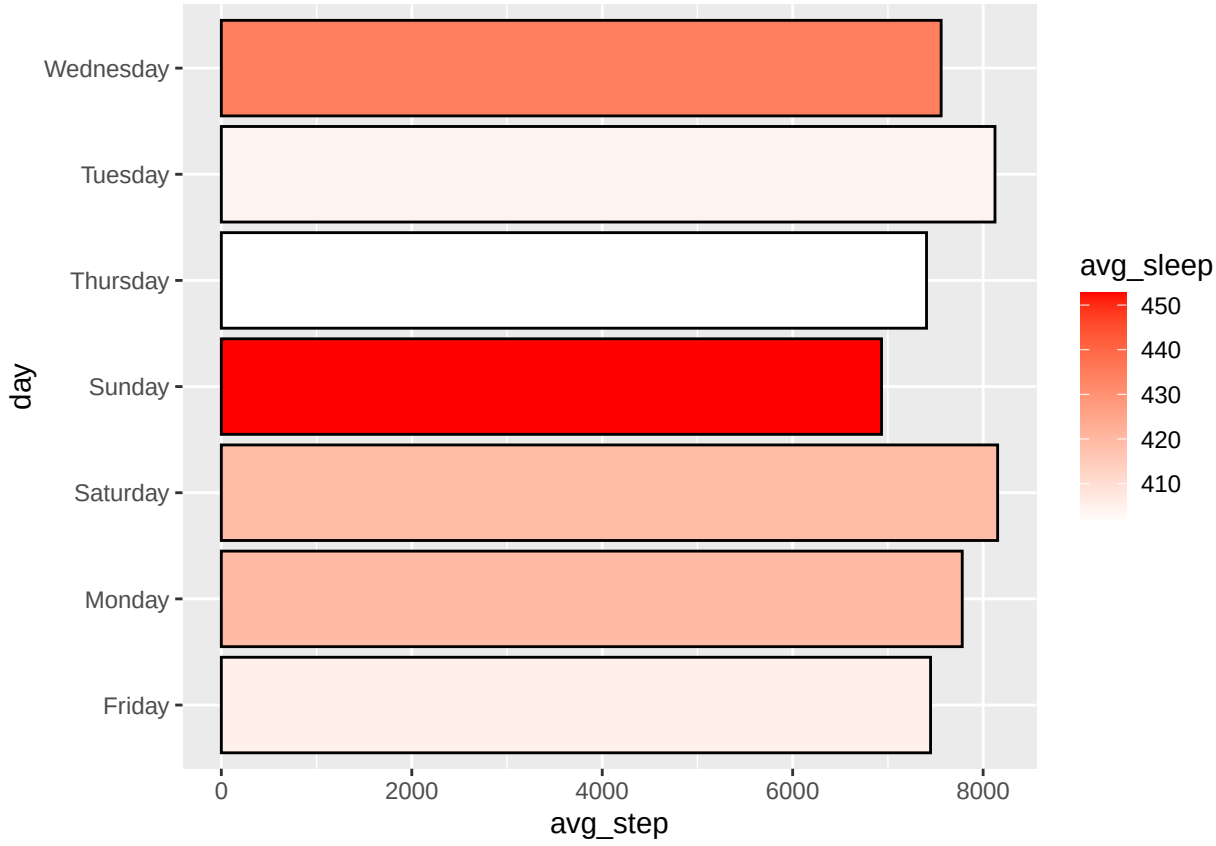
The more the people workout the more they should be tired so the more they should sleep. It might change depending on the data.Lets create a new dataset which has total steps and total minutes as sleep and average it up based on weekdays.

```
avg_step_sleep <- activity_final %>%
  select(day,TotalSteps, TotalMinutesAsleep) %>%
  group_by(day) %>%
  summarise(avg_step= mean(TotalSteps), avg_sleep= mean(TotalMinutesAsleep,na.rm = TRUE))

head(avg_step_sleep)
```

```
## # A tibble: 6 x 3
##   day       avg_step avg_sleep
##   <chr>        <dbl>     <dbl>
## 1 Friday       7448.      405.
## 2 Monday       7781.      420.
## 3 Saturday     8153.      419.
## 4 Sunday       6933.      453.
## 5 Thursday     7406.      401.
## 6 Tuesday      8125.      405.
```

Lets check how the steps taken and the total minutes slept changes vary on each weekday.

```
ggplot(data=avg_step_sleep)+
  geom_col(mapping= aes(x=day,y=avg_step,fill=avg_sleep),colour="black")+
  coord_flip()+
  scale_fill_gradient(low="white",high="red")
```



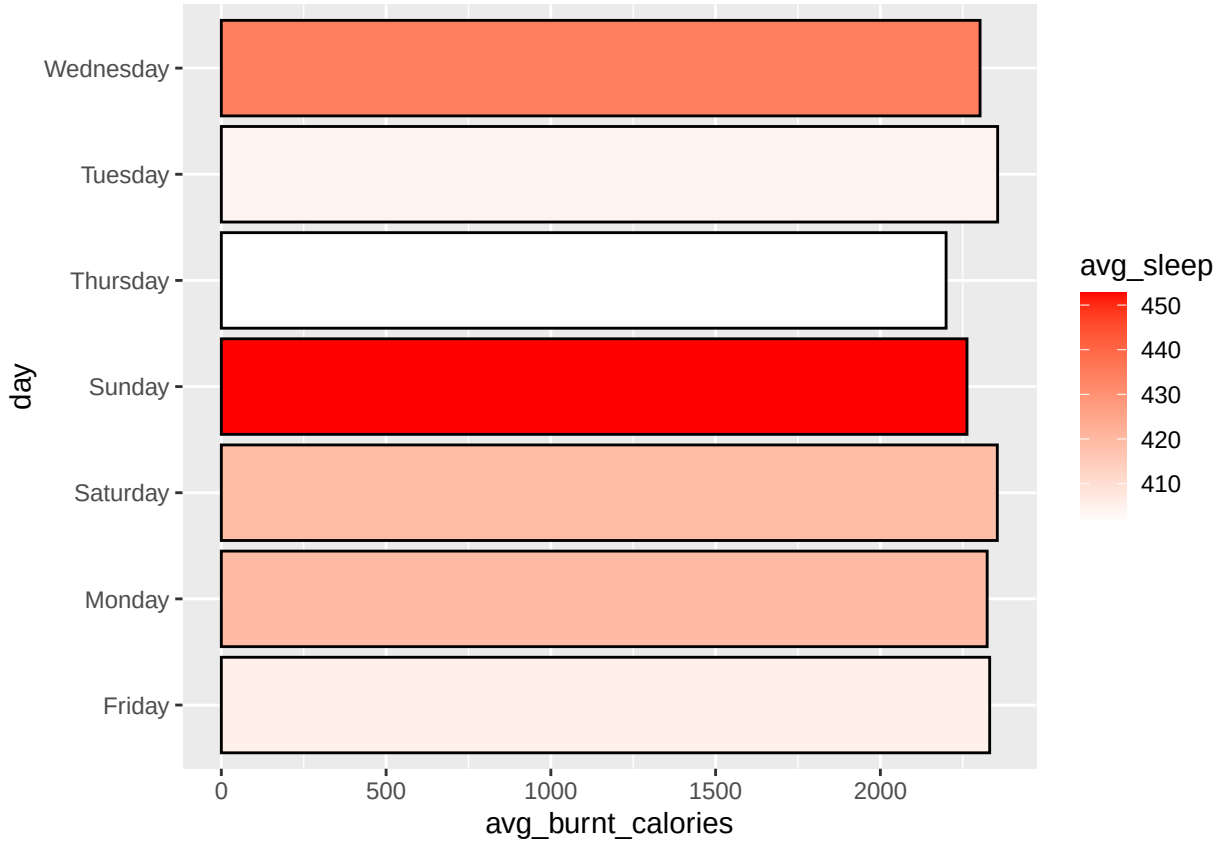
We see that they do not have a proportional relationship. But we also get to see that Thursday is the day where both the steps taken and the amount of sleep that people get are less on average.Lets We also check if the calories burnt and sleep has any relationship.

```
avg_Calories_burnt_v1 <- activity_final %>%
  select(day,Calories) %>%
  group_by(day) %>%
  summarise(avg_burnt_calories= mean(Calories))
head(avg_Calories_burnt_v1)
```

```
## # A tibble: 6 x 2
##   day       avg_burnt_calories
##   <chr>                  <dbl>
## 1 Friday                 2332.
## 2 Monday                 2324.
## 3 Saturday               2355.
## 4 Sunday                 2263
## 5 Thursday               2200.
## 6 Tuesday                2356.
```

```
avg_Calories_burnt <- merge(avg_Calories_burnt_v1, avg_step_sleep,
                            by= c("day"), all=TRUE)
```

```
ggplot(data=avg_Calories_burnt)+
  geom_col(mapping= aes(x=day,y=avg_burnt_calories,fill=avg_sleep),colour="black")+
  coord_flip()+
  scale_fill_gradient(low="white",high="red")
```



Conclusion: We see that on Thursdays on an average people sleep less as well as take lesser steps and so burn less calories than any weekdays.

Device usage

Lets find the number of hours the devices is being used by people. We are splitting them into 3 types and they are as follows: 1. Light Use: Usage anywhere below 11 hours a day 2. Moderate Use: Usage anywhere above 11 hours and below or equal to 18 hours a day. 3. High Use: Usage above 18 hours and below or equal to 24 hours a day

This will give a breakdown of how much number of people use the devices each day.

```
#Create total active minutes and total bed minutes of each person every day
device_usage <- activity_final %>%
  mutate(total_Activity_usage=(VeryActiveMinutes+FairlyActiveMinutes+
              LightlyActiveMinutes+SedentaryMinutes)/60,
         total_sleep_usage=TotalTimeInBed/60) %>%
  select(Id,date,day,total_Activity_usage,total_sleep_usage)
head(device_usage)
```

```
##           Id       date       day total_Activity_usage total_sleep_usage
## 1 1503960366 2016-04-12   Tuesday              18.23333          5.766667
## 2 1503960366 2016-04-13 Wednesday              17.21667          6.783333
## 3 1503960366 2016-04-14  Thursday              24.00000                NA
```

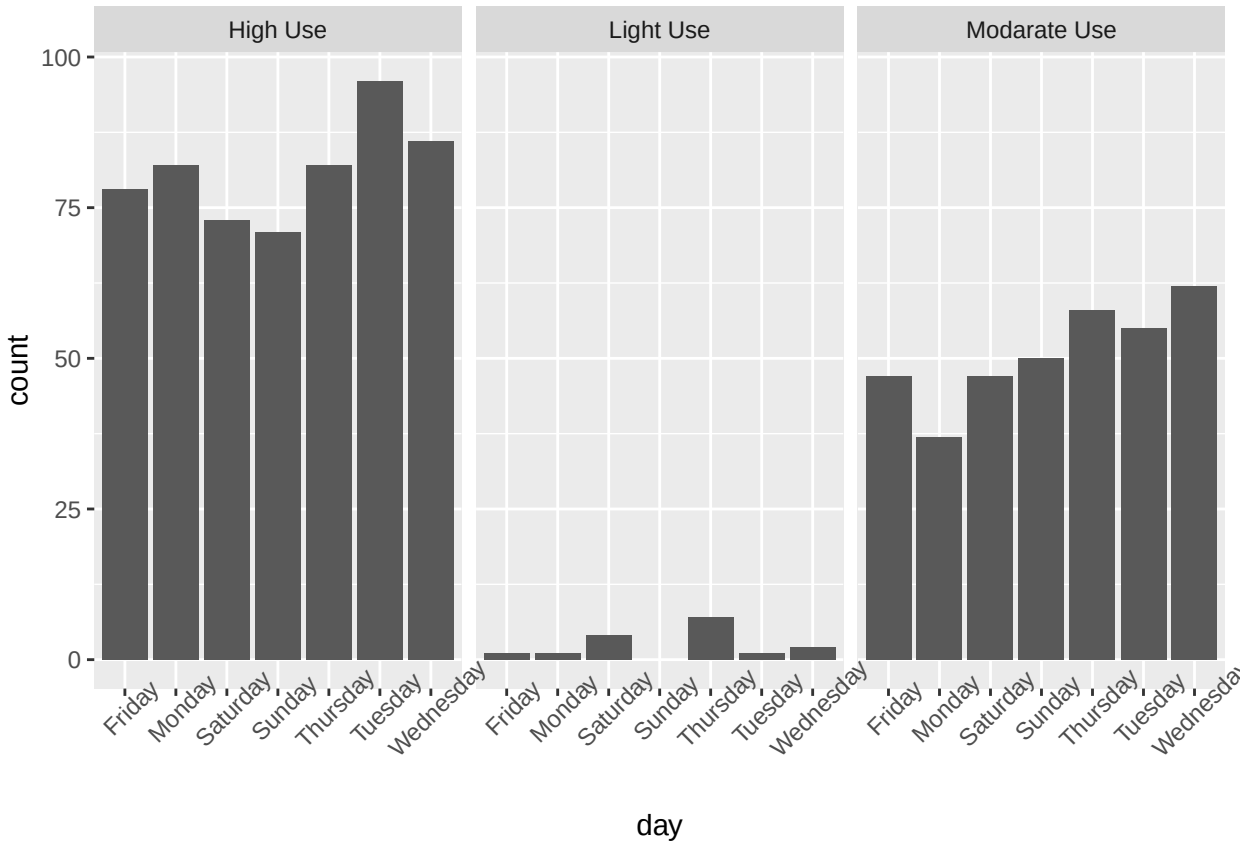
```
## 4 1503960366 2016-04-15    Friday             16.63333          7.366667
## 5 1503960366 2016-04-16  Saturday             17.33333          6.116667
## 6 1503960366 2016-04-17    Sunday             12.68333         11.866667
```

```
#type of usage people do each day

device_usage_type <- device_usage %>%
  mutate(activity_type = case_when(
    total_Activity_usage<11 ~ "Light Use",
    total_Activity_usage>11 & total_Activity_usage<= 18 ~ "Modarate Use",
    total_Activity_usage<=24 ~ "High Use"))

head(device_usage_type)
```

```
##           Id       date       day total_Activity_usage total_sleep_usage
## 1 1503960366 2016-04-12   Tuesday             18.23333          5.766667
## 2 1503960366 2016-04-13 Wednesday             17.21667          6.783333
## 3 1503960366 2016-04-14  Thursday             24.00000                NA
## 4 1503960366 2016-04-15    Friday             16.63333          7.366667
## 5 1503960366 2016-04-16  Saturday             17.33333          6.116667
## 6 1503960366 2016-04-17    Sunday             12.68333         11.866667
##   activity_type
## 1      High Use
## 2  Modarate Use
## 3      High Use
## 4  Modarate Use
## 5  Modarate Use
## 6  Modarate Use
```

Lets visualise and see how the usage vary in the dataset, by using a bar chart visualization.

```
ggplot(data=device_usage_type)+
  geom_bar(mapping=aes(x=day))+
  facet_wrap(~activity_type)+
  theme(axis.text.x = element_text(angle = 45))
```

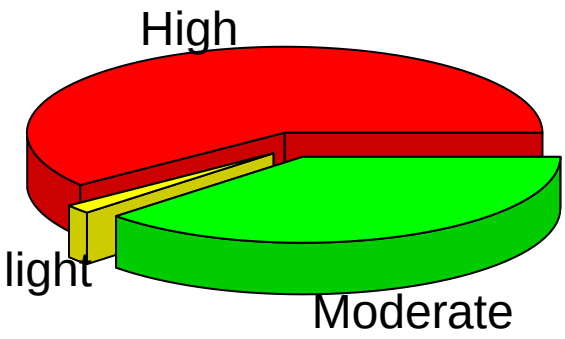


People seem to use the device more between high use and moderate use each day. Now lets find the percentage of each type or the dataset.Lets visualize the percentage in a pie chart.

```
usage_perc <- device_usage_type %>%
  group_by(activity_type) %>%
  summarise(total=n()) %>%
  mutate(count=sum(total)) %>%
  group_by(activity_type) %>%
  summarise(activity_type_perc= (total/count)*100)

head(usage_perc)
```

```
## # A tibble: 3 x 2
##   activity_type activity_type_perc
##   <chr>                      <dbl>
## 1 High Use                   60.4
## 2 Light Use                   1.70
## 3 Modarate Use               37.9
```

```
pie3D(usage_perc$activity_type_perc,,labels= c("High","light", "Moderate"),
    border="Black", col= c("red", "yellow", "green"), radius = 1, explode=0.1,
    main="Device Usage")
```



Conclusion: We see that the people in the dataset use the devices on weekdays on an average more than 18 hours or more than 11 hours, which is high use and moderate use.

Average Minutes VS Intensities

The last analysis is to calculate the intensity level based on the amount of minutes spent during the morning run or walk.

Intensity level - Intensity level is the measure of energy exerted during the exercise or workout done by an individual.

In this analysis the intensity level is measured using four states

1.Very active 2.Fairly active 3.Lightly active 4.Sedentarily active

```
activity_intensities  <- intensities_v2 %>%
  group_by(Id) %>%
  summarise(Very_Active_Minutes = mean(VeryActiveMinutes),
            Fairly_Active_Minutes = mean(FairlyActiveMinutes),
            Light_Active_Minutes = mean(LightlyActiveMinutes),
            Sedentary_Active_Minutes = mean(SedentaryMinutes))
head(activity_intensities)
```

```
## # A tibble: 6 x 5
##           Id Very_Active_Min~ Fairly_Active_M~ Light_Active_Mi~ Sedentary_Activ~
##        <dbl>            <dbl>            <dbl>            <dbl>            <dbl>
## 1 1503960366           38.7             19.2             220.              848.
## 2 1624580081            8.68             5.81            153.             1258.
## 3 1644430081            9.57            21.4             178.             1162.
## 4 1844505072            0.129            1.29            115.             1207.
## 5 1927972279            1.32             0.774            38.6            1317.
## 6 2022484408           36.3             19.4             257.             1113.
```

```
intensities_row <- c("Very_Active_Minutes", "Fairly_Active_Minutes",
                     "Light_Active_Minutes", "Sedentary_Active_Minutes")
avg_activity_intensities <- activity_intensities %>%
  summarise(Mean_Very_Active_Minutes = mean(Very_Active_Minutes),
            Mean_Fairly_Active_Minutes=mean(Fairly_Active_Minutes),
            Mean_Light_Active_Minutes=mean(Light_Active_Minutes),
            Mean_Sedentary_Active_Minutes=mean(Sedentary_Active_Minutes))
```

```
head(avg_activity_intensities)
```

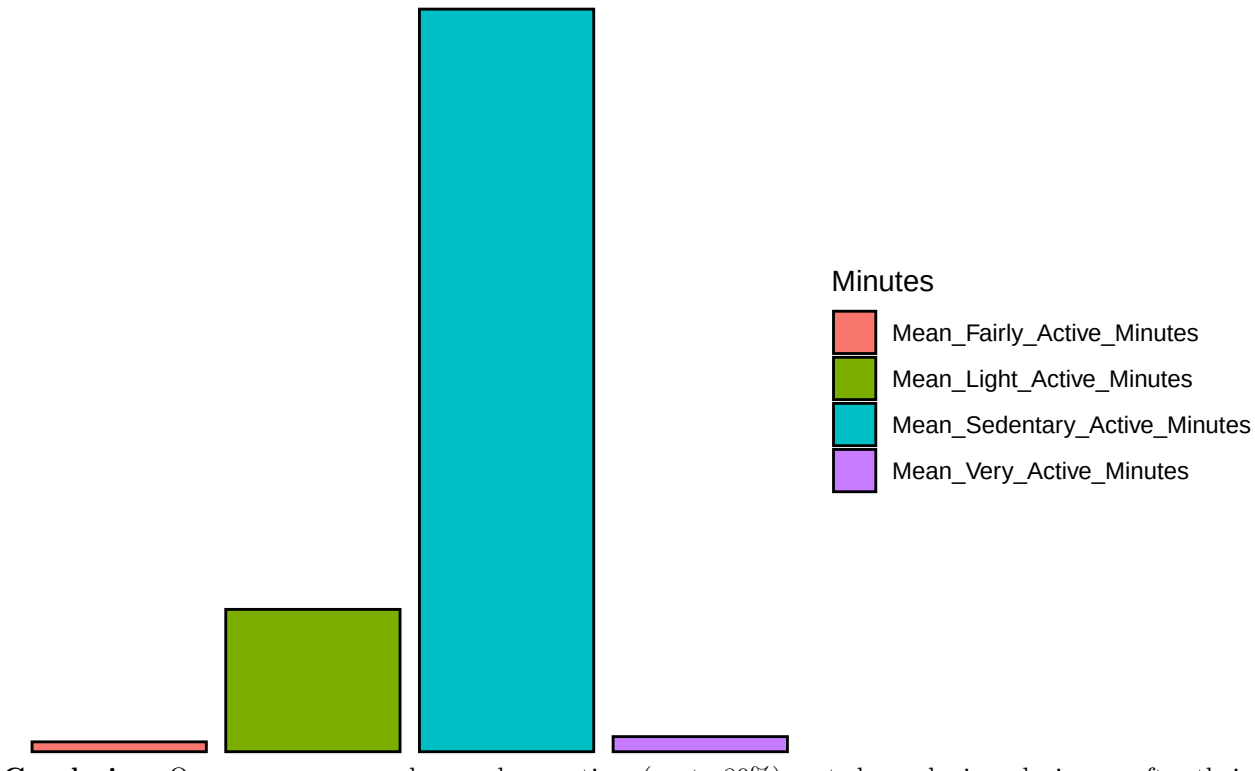
```
## # A tibble: 1 x 4
##   Mean_Very_Active_Minutes Mean_Fairly_Acti~ Mean_Light_Activ~ Mean_Sedentary_A~
##                      <dbl>             <dbl>             <dbl>             <dbl>
## 1                     20.3              13.3              192.              999.
```

We now try to visualize the table using ggplot function.

```
avg_intensities_pie <- avg_activity_intensities %>%
  pivot_longer(everything()) %>%
  rename(Minutes=name)
head(avg_intensities_pie)
```

```
## # A tibble: 4 x 2
##   Minutes                       value
##   <chr>                         <dbl>
## 1 Mean_Very_Active_Minutes       20.3
## 2 Mean_Fairly_Active_Minutes     13.3
## 3 Mean_Light_Active_Minutes     192.
## 4 Mean_Sedentary_Active_Minutes 999.
```

```
ggplot(data = avg_intensities_pie) +
  geom_col(mapping= aes(x = Minutes, y = value, fill=Minutes),color="black") +
  labs(title = "Minutes Spent vs Intensity Level") +
  theme_void()
```



Conclusion: On an average, people spend more time (up to 80%) seated or relaxing, during or after their morning run or stroll.

Inference

1.Calories are burnt more the more total steps taken by the people is. But there is no data for calories intake. 2.On Thursdays on an average people sleep less as well as take lesser steps and so burn less calories than any weekdays. 3.People in the dataset use the devices on weekdays on an average more than 18 hours or more than 11 hours, which is high use and moderate use. Its split between the two more or less into equal halves. 4.On an average, people spend more time (up to 80%) seated or relaxing, during or after their morning run or stroll.

Recommendations

1.Create a feature in the device which also understands calories intake. 2.Offer some special achievements of promotions on thursdays to make people active and get enough sleep. 3.Working on an AI voice assistant that could be an added benefit that could motivate consumers to complete their daily goals to improve device usage and may also help avoid sedentary minutes.