Analysis of fitness device data

Oghenetega Courage Ayonuwe

2025-05-23

Installation and loading of packages for data wrangling, cleaning and analysis

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library (janitor)
```

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

```
library(skimr)
library(here)
```

```
## here() starts at /cloud/project
```

Importing case study datasets

```
combined <- read_csv("combined_daily_recording_v02.csv")
```

```
## Rows: 457 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.
```

```
heartrate <- read_csv("heartrate_seconds_merged_V02.csv")
```

```
## Rows: 1154681 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
## dbl  (2): Id, Value
## time (1): Time
##
## 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("hourlyCalories_merged_V02.csv")
```

```
## Rows: 24084 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date2
## dbl  (2): Id2, Calories2
## time (1): Time2
##
## 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.
```

```
intensity <- read_csv("hourlyIntensities_merged_V02.csv")
```

```
## Rows: 24084 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Activity_Date
## dbl  (3): Id, Average_Intensity, Total_Intensity
## time (1): Time
##
## 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.
```

```
steps3 <- read_csv("hourlySteps_merged_V03.csv")
```

```
## Rows: 24084 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
```

```
## dbl  (2): Id, StepTotal
## time (1): Hour
##
## 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("minuteSleep_merged_V02.csv")
```

```
## Rows: 198559 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): date
## dbl  (3): Id, value, logId
## time (1): time
##
## 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_V02.csv")
```

```
## Rows: 33 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
## dbl  (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl  (1): IsManualReport
## time (1): Time
##
## 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.
```

Previewing the data

```
# Preview the data
head(combined)
```

```
## # A tibble: 6 x 15
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
##        <dbl> <chr>             <dbl>         <dbl>           <dbl>
## 1 1503960366 3/25/2016         11004          7.11            7.11
## 2 1503960366 3/26/2016         17609         11.6            11.6
## 3 1503960366 3/27/2016         12736          8.53            8.53
## 4 1503960366 3/28/2016         13231          8.93            8.93
## 5 1503960366 3/29/2016         12041          7.85            7.85
## 6 1503960366 3/30/2016         10970          7.16            7.16
## # i 10 more variables: LoggedActivitiesDistance <dbl>,
## #   VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
```

```
head(heartrate)
```

```
## # A tibble: 6 x 4
##   Date                Id Time     Value
##   <chr>            <dbl> <time>   <dbl>
```

```
## 1 3/29/2016 2347167796 22:49:30    65
## 2 3/29/2016 2347167796 22:49:45    65
## 3 3/29/2016 2347167796 22:49:50    63
## 4 3/29/2016 2347167796 22:50:05    63
## 5 3/29/2016 2347167796 22:50:20    63
## 6 3/29/2016 2347167796 22:50:30    62
```

```
head(calories)
```

```
## # A tibble: 6 x 4
##          Id2 Date2     Time2  Calories2
##        <dbl> <chr>     <time>     <dbl>
## 1 1503960366 3/12/2016 00:00         48
## 2 1503960366 3/12/2016 01:00         48
## 3 1503960366 3/12/2016 02:00         48
## 4 1503960366 3/12/2016 03:00         48
## 5 1503960366 3/12/2016 04:00         48
## 6 1503960366 3/12/2016 05:00         48
```

```
head(steps3)
```

```
## # A tibble: 6 x 4
##           Id Date      Hour   StepTotal
##        <dbl> <chr>     <time>     <dbl>
## 1 1503960366 3/12/2016 00:00          0
## 2 1503960366 3/12/2016 01:00          0
## 3 1503960366 3/12/2016 02:00          0
## 4 1503960366 3/12/2016 03:00          0
## 5 1503960366 3/12/2016 04:00          0
## 6 1503960366 3/12/2016 05:00          0
```

Understanding the structure and summary of the dataset

```
str(combined)
```

```
## spc_tbl_ [457 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                       : num [1:457] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr [1:457] "3/25/2016" "3/26/2016" "3/27/2016" "3/28/2016" ...
##  $ TotalSteps               : num [1:457] 11004 17609 12736 13231 12041 ...
##  $ TotalDistance            : num [1:457] 7.11 11.55 8.53 8.93 7.85 ...
##  $ TrackerDistance          : num [1:457] 7.11 11.55 8.53 8.93 7.85 ...
##  $ LoggedActivitiesDistance: num [1:457] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num [1:457] 2.57 6.92 4.66 3.19 2.16 ...
##  $ ModeratelyActiveDistance: num [1:457] 0.46 0.73 0.16 0.79 1.09 ...
##  $ LightActiveDistance      : num [1:457] 4.07 3.91 3.71 4.95 4.61 ...
##  $ SedentaryActiveDistance : num [1:457] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : num [1:457] 33 89 56 39 28 30 33 47 40 15 ...
##  $ FairlyActiveMinutes      : num [1:457] 12 17 5 20 28 13 12 21 11 30 ...
##  $ LightlyActiveMinutes     : num [1:457] 205 274 268 224 243 223 239 200 244 314 ...
##  $ SedentaryMinutes         : num [1:457] 804 588 605 1080 763 ...
##  $ Calories                 : num [1:457] 1819 2154 1944 1932 1886 ...
##  - 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>
```

```
summary(combined)
```

```
##        Id            ActivityDate         TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:457         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.347e+09   Class :character   1st Qu.: 1988   1st Qu.: 1.410
##  Median :4.057e+09   Mode  :character   Median : 5986   Median : 4.090
##  Mean   :4.629e+09                      Mean   : 6547   Mean   : 4.664
##  3rd Qu.:6.392e+09                      3rd Qu.:10198   3rd Qu.: 7.160
##  Max.   :8.878e+09                      Max.   :28497   Max.   :27.530
##  TrackerDistance LoggedActivitiesDistance VeryActiveDistance
##  Min.   : 0.00   Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 1.28   1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 4.09   Median :0.0000           Median : 0.000
##  Mean   : 4.61   Mean   :0.1794           Mean   : 1.181
##  3rd Qu.: 7.11   3rd Qu.:0.0000           3rd Qu.: 1.310
##  Max.   :27.53   Max.   :6.7271           Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.00       Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 0.87       1st Qu.:0.000000
##  Median :0.0200           Median : 2.93       Median :0.000000
##  Mean   :0.4786           Mean   : 2.89       Mean   :0.001904
##  3rd Qu.:0.6700           3rd Qu.: 4.46       3rd Qu.:0.000000
##  Max.   :6.4000           Max.   :12.51       Max.   :0.100000
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :  32.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.: 64.0        1st Qu.: 728.0
##  Median :  0.00    Median :  1.00      Median :181.0        Median :1057.0
##  Mean   : 16.62    Mean   : 13.07      Mean   :170.1        Mean   : 995.3
##  3rd Qu.: 25.00    3rd Qu.: 16.00      3rd Qu.:257.0        3rd Qu.:1285.0
##  Max.   :202.00    Max.   :660.00      Max.   :720.0        Max.   :1440.0
##     Calories
##  Min.   :   0
##  1st Qu.:1776
##  Median :2062
##  Mean   :2189
##  3rd Qu.:2667
##  Max.   :4562
```

Checking for Nulls and duplicates

```
sum(is.null(combined))
```

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

```
sum(duplicated(combined))
```

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

```
sum(is.null(steps3))
```

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

```
sum(duplicated(steps3))
```

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

Converting ActivityDate column from chr to date format

```
steps3$Date <- as.Date(steps3$Date, format = "%m/%d/%y")
```

Checking the changes

```
str(steps3)
```

```
## spc_tbl_ [24,084 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id       : num [1:24084] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Date     : Date[1:24084], format: "2020-03-12" "2020-03-12" ...
##  $ Hour     : 'hms' num [1:24084] 00:00:00 01:00:00 02:00:00 03:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ StepTotal: num [1:24084] 0 0 0 0 0 0 0 0 0 8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   Hour = col_time(format = ""),
##   ..   StepTotal = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Calculating average steps per hour per user for the month of April

```
steps1 <- steps3 %>%
  group_by(Date, Id, Hour) %>%
  summarise(avg_step = mean(StepTotal)) %>%
  filter(Date > '2020-03-31')
```

```
## `summarise()` has grouped output by 'Date', 'Id'. You can override using the
## `.groups` argument.
```

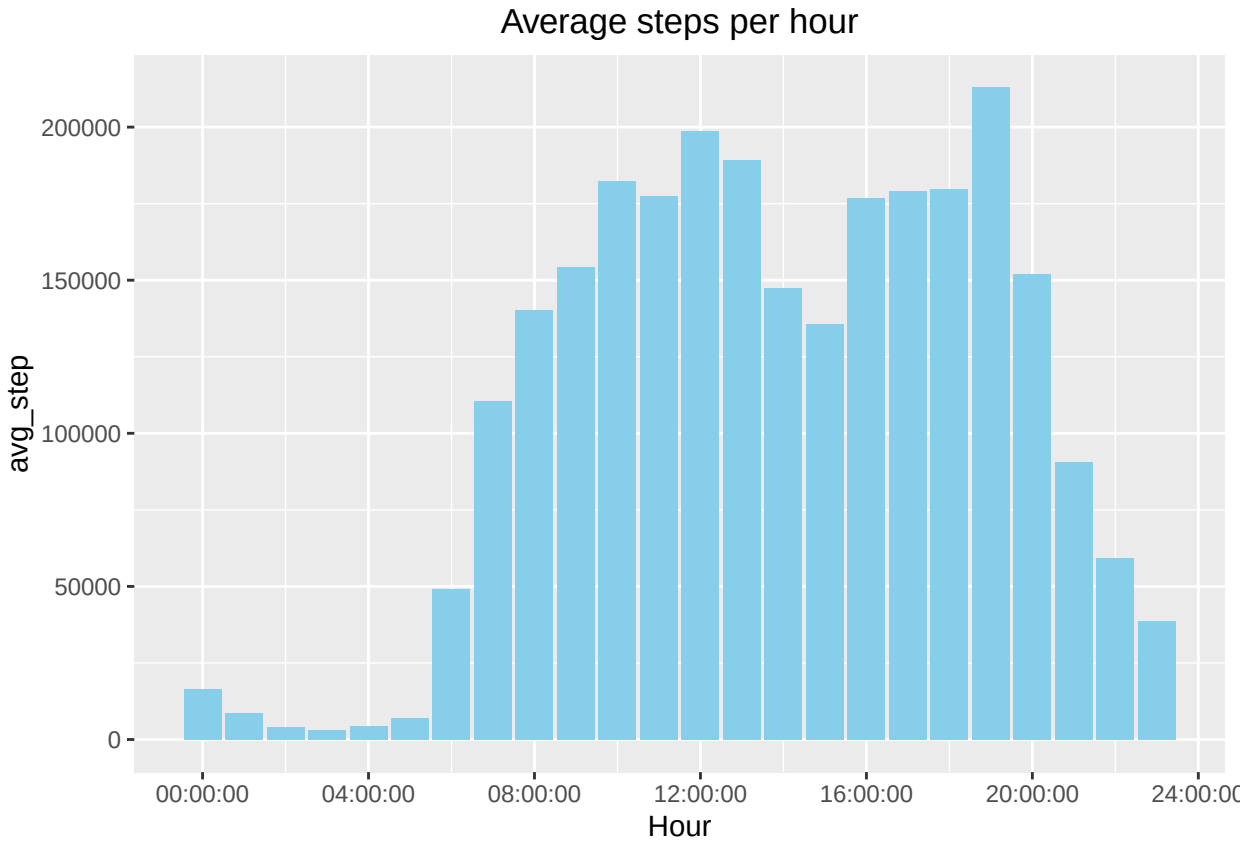
```
steps1
```

```
## # A tibble: 8,244 x 4
## # Groups:   Date, Id [361]
##    Date                Id Hour   avg_step
```

```
##    <date>          <dbl> <time>    <dbl>
##  1 2020-04-01 1503960366 00:00        82
##  2 2020-04-01 1503960366 01:00         7
##  3 2020-04-01 1503960366 02:00        41
##  4 2020-04-01 1503960366 03:00       106
##  5 2020-04-01 1503960366 04:00         0
##  6 2020-04-01 1503960366 05:00         0
##  7 2020-04-01 1503960366 06:00         0
##  8 2020-04-01 1503960366 07:00        43
##  9 2020-04-01 1503960366 08:00       251
## 10 2020-04-01 1503960366 09:00       194
## # i 8,234 more rows
```

```
view(steps1)
```

Visualization of average steps per Hour



Conversion of ActivityDate column from chr to date format

```
calories$Date2 <- as.Date(calories$Date2, format = "%m/%d/%y")
```

Checking the changes

```
str(calories)
```

```
## spc_tbl_ [24,084 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
```

```
##  $ Id2      : num [1:24084] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Date2    : Date[1:24084], format: "2020-03-12" "2020-03-12" ...
##  $ Time2    : 'hms' num [1:24084] 00:00:00 01:00:00 02:00:00 03:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ Calories2: num [1:24084] 48 48 48 48 48 48 48 48 48 49 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id2 = col_double(),
##   ..   Date2 = col_character(),
##   ..   Time2 = col_time(format = ""),
##   ..   Calories2 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

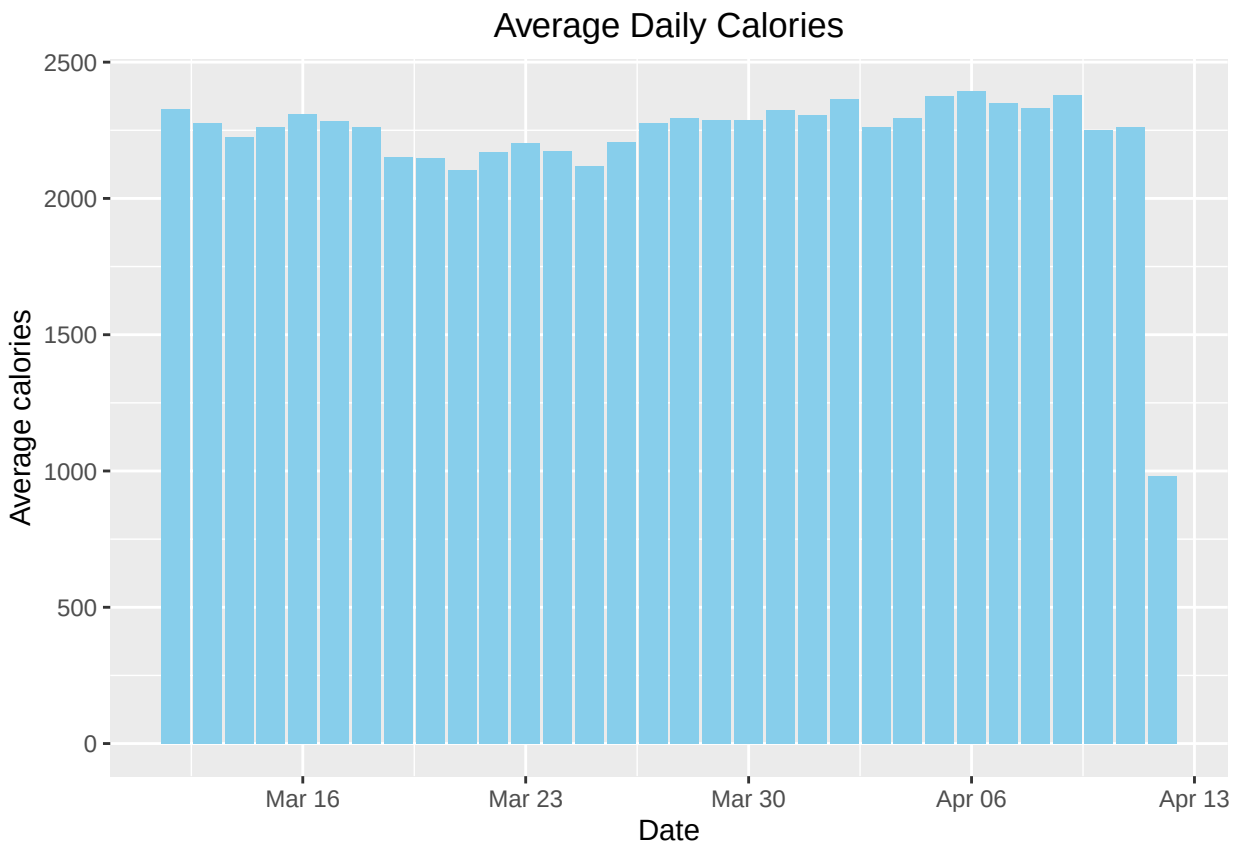
Calculation of average daily calories

```
calories2 <- calories %>%
  group_by(Date2, Time2) %>%
  summarise(avg_calories = mean(Calories2))
```

```
## `summarise()` has grouped output by 'Date2'. You can override using the
## `.groups` argument.
```

```
view(calories2)
```

Visualization of the average daily calories per user



Analysis of sleep patterns

```
# Viewing the structure of the sleep dataset
str(sleep)
```

```
## spc_tbl_ [198,559 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id   : num [1:198559] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ date : chr [1:198559] "3/13/2016" "3/13/2016" "3/13/2016" "3/13/2016" ...
##  $ time : 'hms' num [1:198559] 02:39:30 02:40:30 02:41:30 02:42:30 ...
##   ..- attr(*, "units")= chr "secs"
##  $ value: num [1:198559] 1 1 1 1 1 1 2 2 1 1 ...
##  $ logId: num [1:198559] 1.11e+10 1.11e+10 1.11e+10 1.11e+10 1.11e+10 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   date = col_character(),
##   ..   time = col_time(format = ""),
##   ..   value = col_double(),
##   ..   logId = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Grouping the time into hours from minutes
library(hms)
```

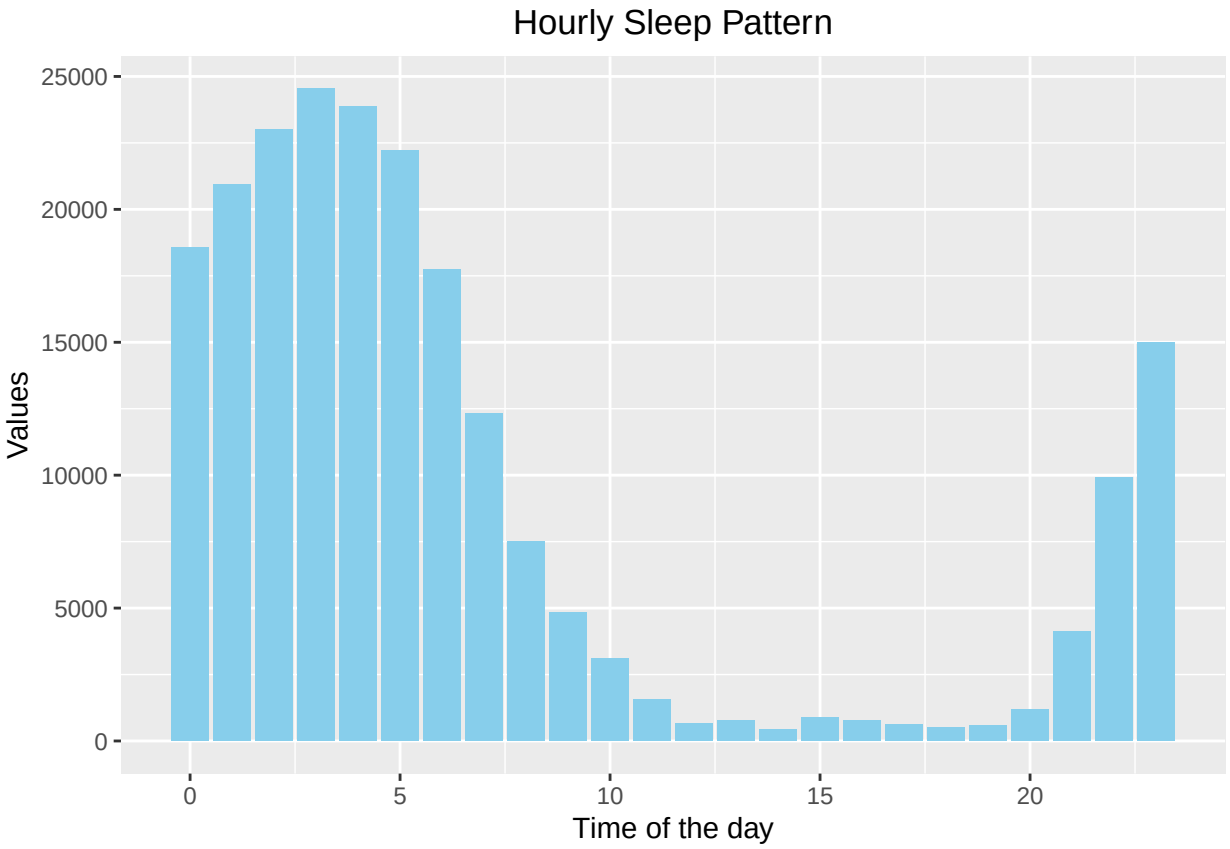
```
##
## Attaching package: 'hms'
```

```
## The following object is masked from 'package:lubridate':
##
##     hms
```

```
library(dplyr)
sleep2 <- sleep %>%
  mutate(hourly = hour(time)) %>%
  group_by(hourly) %>%
  summarise(total_value = sum(value))
sleep2
```

```
## # A tibble: 24 x 2
##    hourly total_value
##     <int>       <dbl>
##  1      0       18566
##  2      1       20953
##  3      2       23012
##  4      3       24529
##  5      4       23885
##  6      5       22212
##  7      6       17740
##  8      7       12308
##  9      8        7502
## 10      9        4851
## # i 14 more rows
```

Visualizing the hourly sleep pattern for the day



Analysis of distance covered vs weight of users

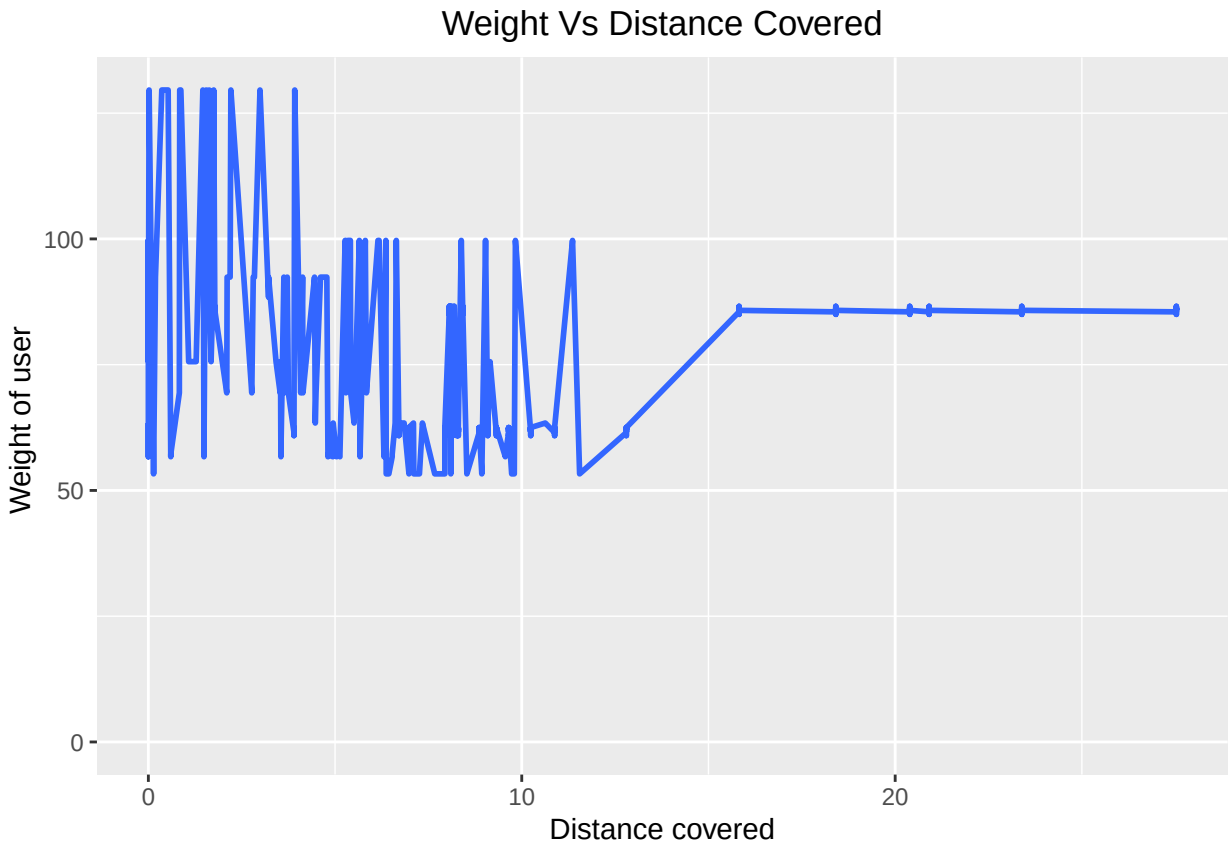
```
#
# Performing a join operation on the 'weight' dataset and 'combined' dataset
merged_weight_combined <- full_join(combined, weight, by = "Id")
```

```
## Warning in full_join(combined, weight, by = "Id"): Detected an unexpected many-to-many relationship b
## i Row 124 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
# Checking the changes
view(merged_weight_combined)
```

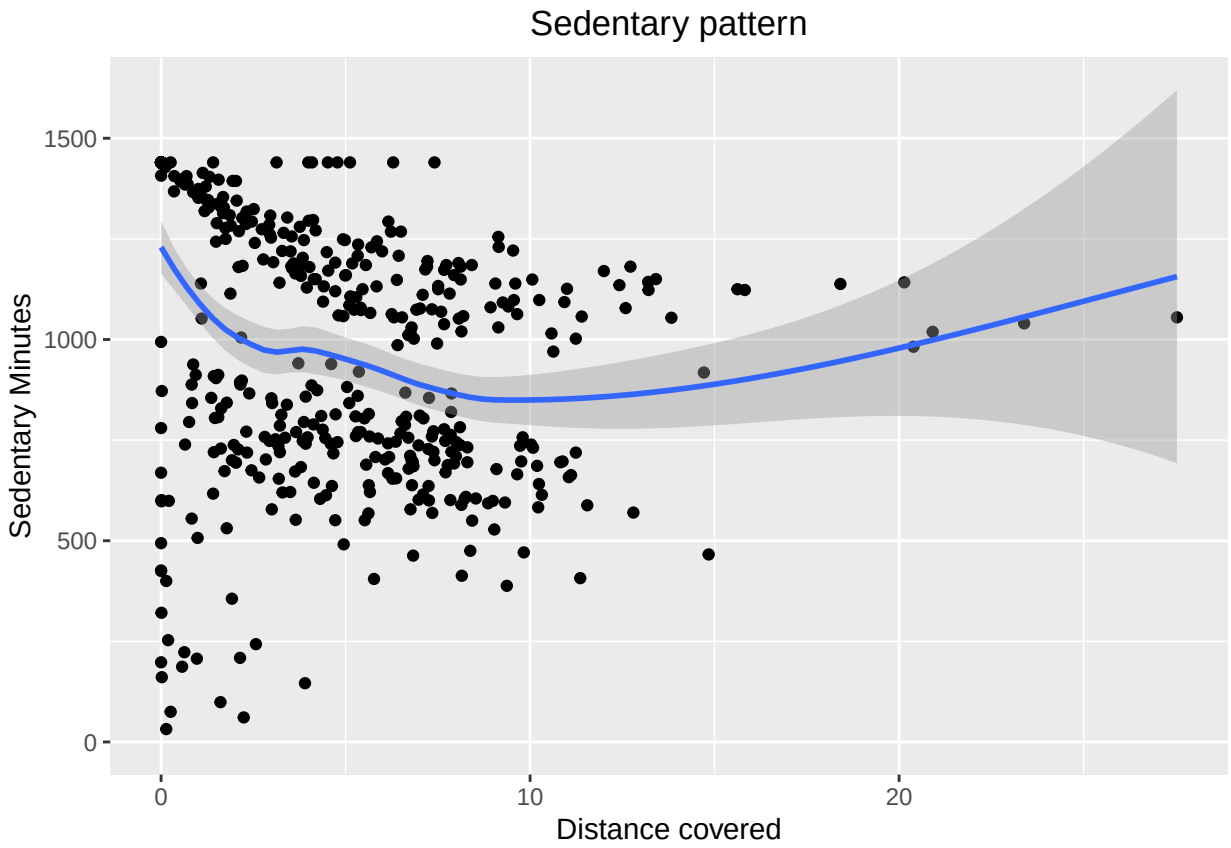
Visualizing weight vs distance covered

```
## Warning: Removed 311 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



Visualizing sedentary patterns

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


Sedentary pattern
Sedentary Minutes
1500
1000
500
0
0
10
20
Distance covered