Reproducible Research: Peer Assessment 1

Isaac G Veras

05/10/2023

This assignment will be described in multiple parts. You will need to write a report that answers the questions detailed below. Ultimately, you will need to complete the entire assignment in a single R Markdown document that can be processed by knitr and be transformed into an HTML file.

Package installation:

```
if (!require("pacman")) install.packages("pacman")
```

```
## Carregando pacotes exigidos: pacman
```

```
pacman::p_load(pacman,     # Package Manager
               knitr,      # Transform R Markdown documents into various output formats such as HT
ML, PDF, Word and others.
               plyr,       # Data manipulation
               data.table, # Manipulate, process and analyze large data sets
               tidyverse   # Data organization
)
```

Current working directory:

```
R.version.string   # R 4.3.1
```

```
## [1] "R version 4.3.1 (2023-06-16 ucrt)"
```

```
getwd(); cat("\n") # Current working directory
```

```
## [1] "C:/Johns Hopkins - Data Science/Reproducible_Research/RR_Poject1"
```

```
setwd("C:/Johns Hopkins - Data Science/Reproducible_Research/RR_Poject1")
```

Following the results per questions:

Loading and preprocessing the data

To load the data I used read.csv function, considering first file as the headers and all the missing values, as follows:

```
opts_chunk$set(fig_path = "./figure/")
activity_monitoring_data <- read.csv("activity.csv",
                                     header     = TRUE,
                                     na.strings = "NA"
)
head(activity_monitoring_data)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

What is mean total number of steps taken per day?

I will calculate the total average of steps taken per “day”, considering an applied function that adds the total steps per date, then I will calculate the average of StepsPerDay. The result is the next:

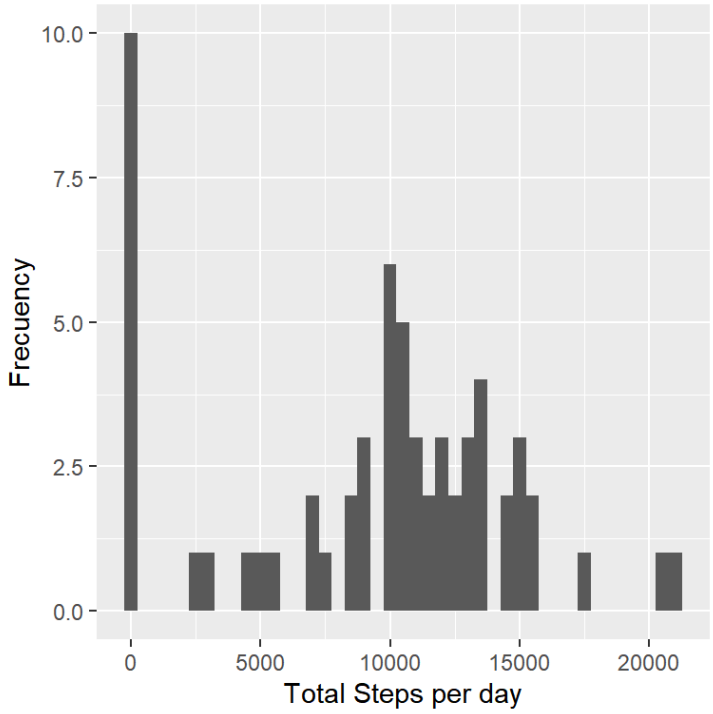
```
opts_chunk$set(fig_path = "./figure/")
steps_per_day <- tapply(activity_monitoring_data$steps,
                        activity_monitoring_data$date,
                        sum,
                        na.rm = TRUE
)
mean_step_per_day <- mean(steps_per_day)
mean_step_per_day
```

```
## [1] 9354.23
```

A histogram of the Total of number of steps by day is calculated and plot by the next code:

```
opts_chunk$set(fig_path = "./figure/")
library(ggplot2)
qplot(steps_per_day,
      xlab     = "Total Steps per day",
      ylab     = "Frecuency",
      binwidth = 500
)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



What is the average daily activity pattern?

To calculate the Mean and the Median number step by day, I used the following code with their results:

```
mean_step_per_day   <- mean(steps_per_day)
mean_step_per_day
```

```
## [1] 9354.23
```

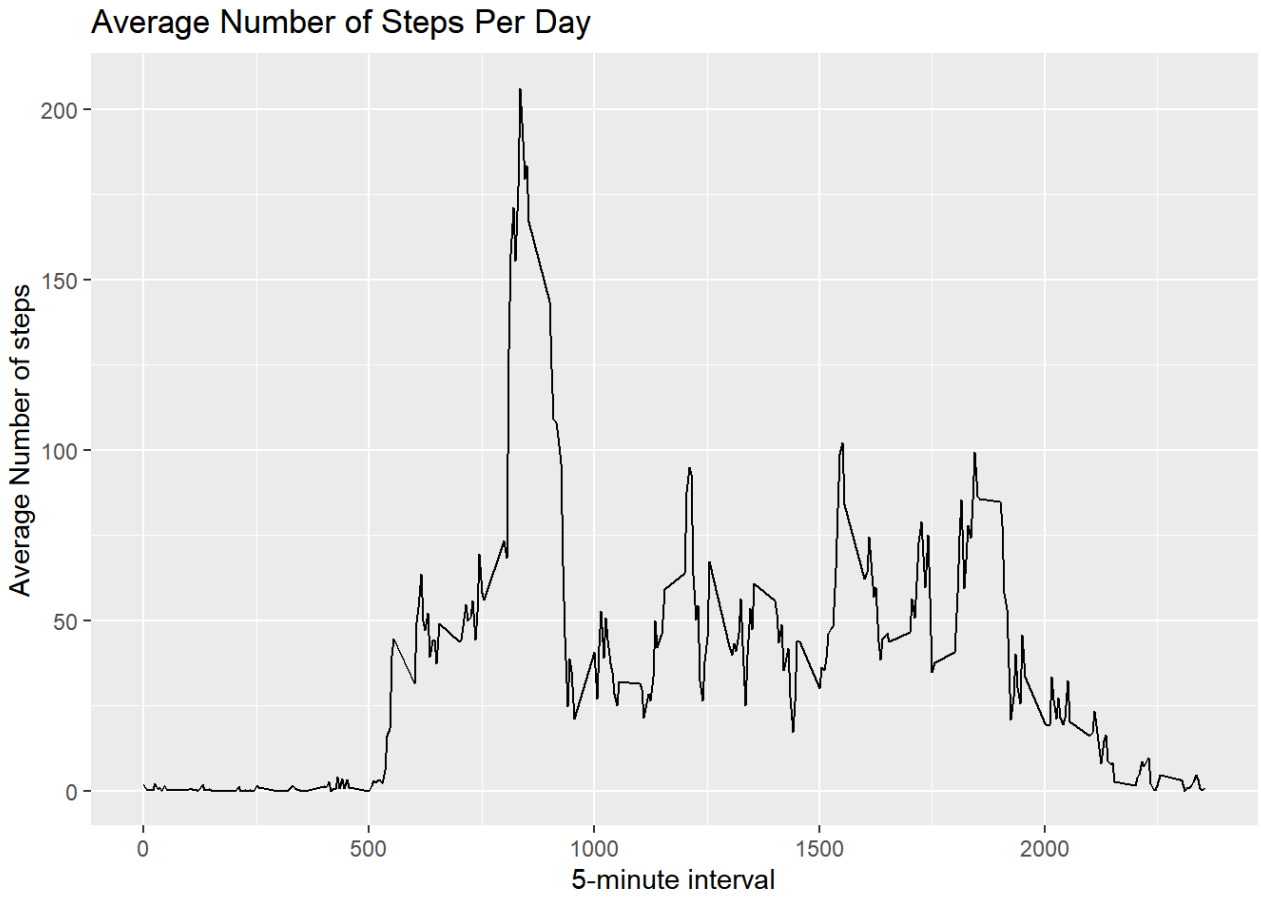
```
median_step_per_day <- median(steps_per_day)
median_step_per_day
```

```
## [1] 10395
```

To create a time series chart that shows the average number of “steps” taken in 5-minute intervals, along with the 5-minute interval that, on average, contains the greatest number of steps, I code the following:

```
opts_chunk$set(fig_path = "./figure/")
ave_day_act_patt <- aggregate(x     = list(meanSteps = activity_monitoring_data$steps),
                              by    = list(interval = activity_monitoring_data$interval),
                              FUN   = mean,
                              na.rm = TRUE
)

ggplot(data = ave_day_act_patt,
       aes(x = interval,
           y = meanSteps)) +
        geom_line() +
        ggtitle("Average Number of Steps Per Day") +
        xlab("5-minute interval") +
        ylab("Average Number of steps"
        )
```



The 5-minutes interval on average per day in the data contains the maximum number of steps?

```
opts_chunk$set(fig_path="./figure/")
max_steps     <- which.max(ave_day_act_patt$meanSteps)
most_of_steps <- gsub("([0-9]{1,2})([0-9]{2})", "\\1:\\2", ave_day_act_patt[max_steps, "interva
l"])
most_of_steps
```

```
## [1] "8:35"
```

This “Interval number” indicates that 8.35 AM is the time when the average person is most active

Code to describe and show a strategy for imputing missing data

The total number of missing values are calculate by the next code

```
MValues<-length(which(is.na(activity_monitoring_data$steps)))
MValues
```

```
## [1] 2304
```

Make a Histogram of the number of total steps taken

by day

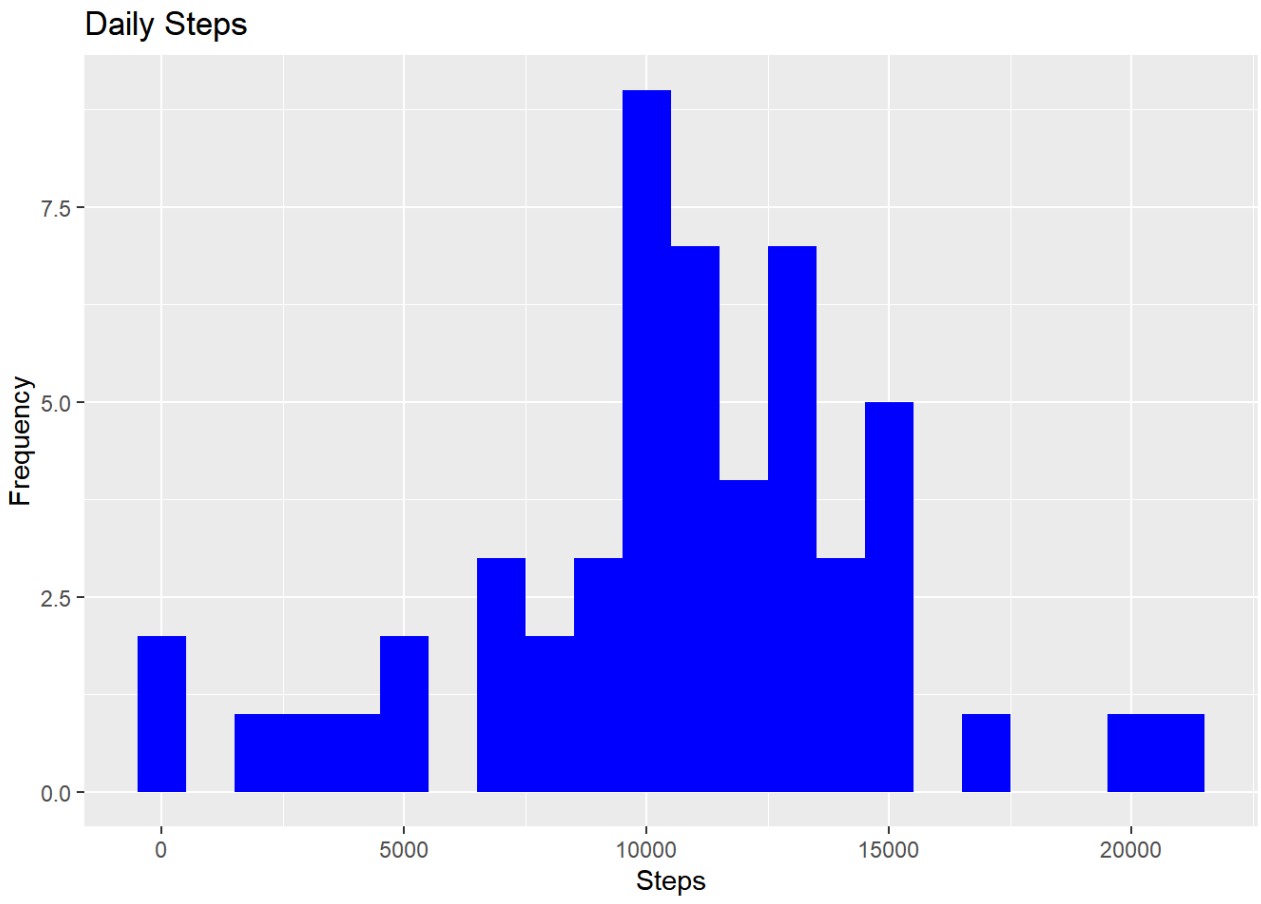
Following the histogram which show the total steps taken by day, in thi section I consider the advantage for the data.table function. Following the code and the histogram.

```
opts_chunk$set(fig_path = "./figure/")
activity   <- data.table::fread(input = "activity.csv")
TotalSteps <- activity[, lapply(.SD, sum), .SDcols = "steps", by = .(date)]
TotalSteps[, .(MeanSteps = mean(steps), MedianSteps = median(steps))]
```

```
##    MeanSteps MedianSteps
## 1:        NA          NA
```

```
ggplot(TotalSteps, aes(x = steps)) +
        geom_histogram(fill     = "blue",
                       binwidth = 1000) +
        labs(title = "Daily Steps",
             x     = "Steps",
             y     = "Frequency"
        )
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_bin()`).
```



Are there differences in activity patterns between weekdays and weekends?

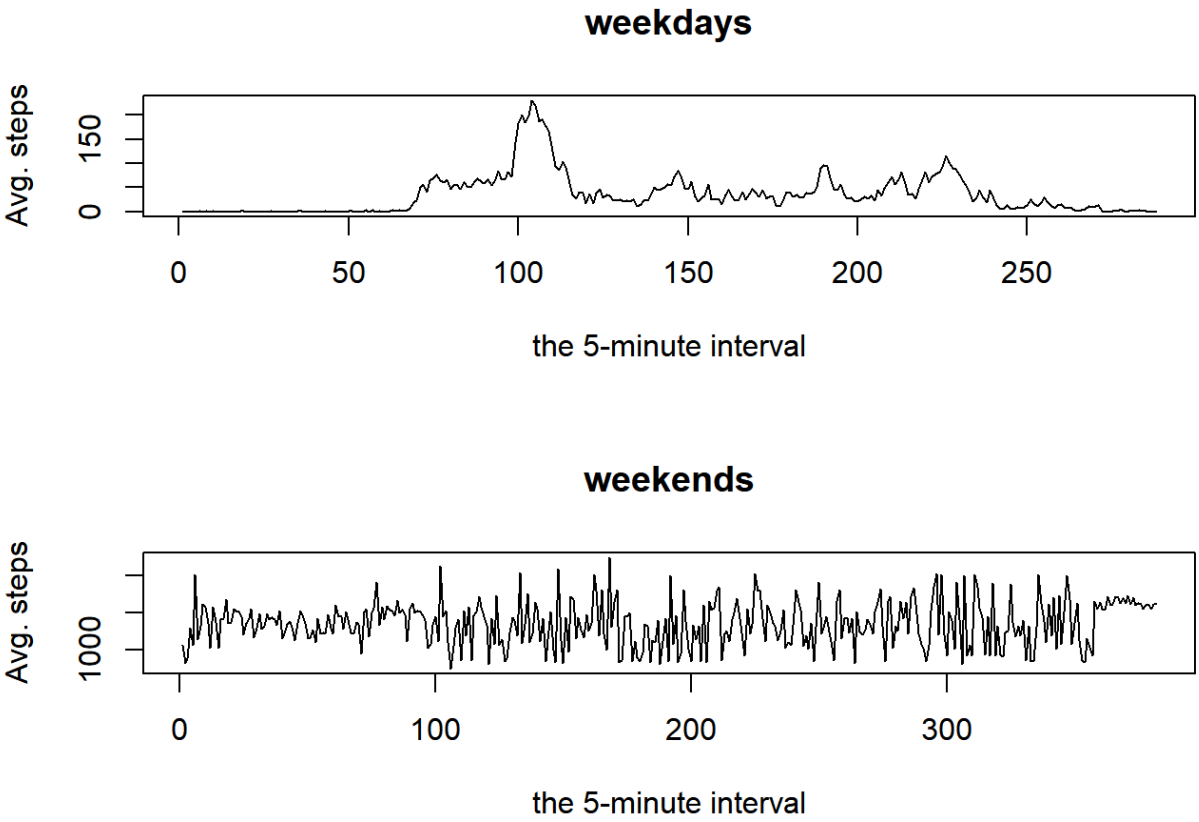
Building a factor variable considering weeks and weekends as follows:

```
opts_chunk$set(fig_path = "./figure/")
activity_monitoring_data$date <- as.POSIXct(activity_monitoring_data$date)
dataFix <- activity_monitoring_data

for (i in unique(dataFix$interval)) {
  dataFix$steps[is.na(dataFix$steps) & dataFix$interval == i] <- round(mean(dataFix$steps[activity
_monitoring_data$interval == i], na.rm = TRUE))
}
dataFix$weekDay <- as.POSIXlt(activity_monitoring_data$date)$wday == 0 | as.POSIXlt(activity_monit
oring_data$date)$wday == 6
dataFix$weekDay <- factor(dataFix$weekDay,
                          levels = c(F, T),
                          labels = c("weekday", "weekend")
)

steps_week_day <- tapply(dataFix$steps[dataFix$weekDay == "weekday"],
                         dataFix$interval[dataFix$weekDay == "weekday"], mean)
steps_week_end <- tapply(dataFix$interval[dataFix$weekDay == "weekend"],
                         dataFix$steps[dataFix$weekDay == "weekend"], mean)

par(mfrow = c(2, 1))
plot(steps_week_day,
     type = "l",
     main = "weekdays",
     xlab = "the 5-minute interval",
     ylab = "Avg. steps"
)
plot(steps_week_end,
     type = "l",
     main = "weekends",
     xlab = "the 5-minute interval",
     ylab = "Avg. steps"
)
```


weekdays
Avg. steps
0
150
0
50
100
150
200
250
the 5-minute interval
weekends
Avg. steps
1000
0
100
200
300
the 5-minute interval