Reproducible Research - Course Project 1

José Maria Fernandes Marlet

2022-07-13

Introduction

This is the Project 1 of the course Reproducible Research for Peer Assessment.

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

The variables included in this dataset are:

```
1. steps: Number of steps taking in a 5-minute interval (missing values are coded as\color{red}{\verb|NA|}NA)

2. date: The date on which the measurement was taken in YYYY-MM-DD format

3. interval: Identifier for the 5-minute interval in which measurement was taken
```

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

More details can be appreciated at https://www.coursera.org/learn/reproducible-research/peer/gYyPt/course-project-1

Loading libraries

```r
library(ggplot2)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Loading and processing the data

```
activityData <- read.csv("F:/Desktop/Coursera - Ciência de Dados/Johns Hopkins/Reproducible Research/Reproducible Research/activity.csv")

summary(activityData)
```

```
##      steps            date              interval     
##  Min.   :  0.00   Length:17568       Min.   :   0.0  
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8  
##  Median :  0.00   Mode  :character   Median :1177.5  
##  Mean   : 37.38                      Mean   :1177.5  
##  3rd Qu.: 12.00                      3rd Qu.:1766.2  
##  Max.   :806.00                      Max.   :2355.0  
##  NA's   :2304
```

What is mean total number of steps taken per day?

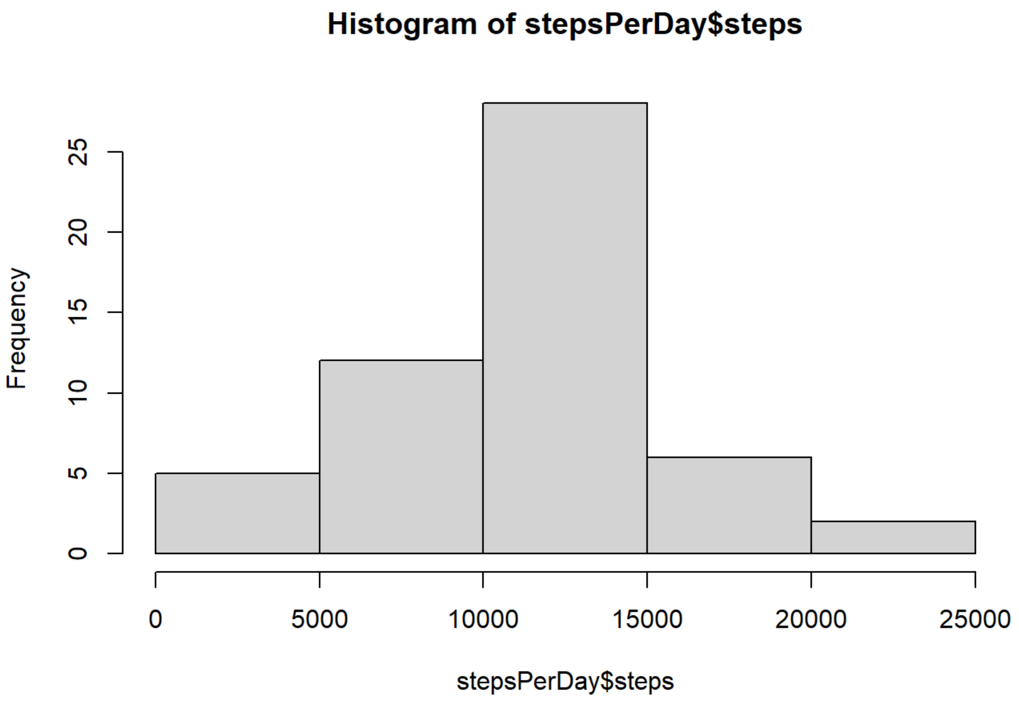
For this part of the assignment, you can ignore the missing values in the dataset.

1. Calculate the total number of steps taken per day without NA values.

```
stepsPerDay <- aggregate(steps ~ date, activityData, sum, na.rm=TRUE)
```

2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day, without NA.

```
hist(stepsPerDay$steps)
```



3. Calculate and report the mean and median of the total number of steps taken per day.

```
meanStepsPerDay <- mean(stepsPerDay$steps)
meanStepsPerDay
```

```
## [1] 10766.19
```

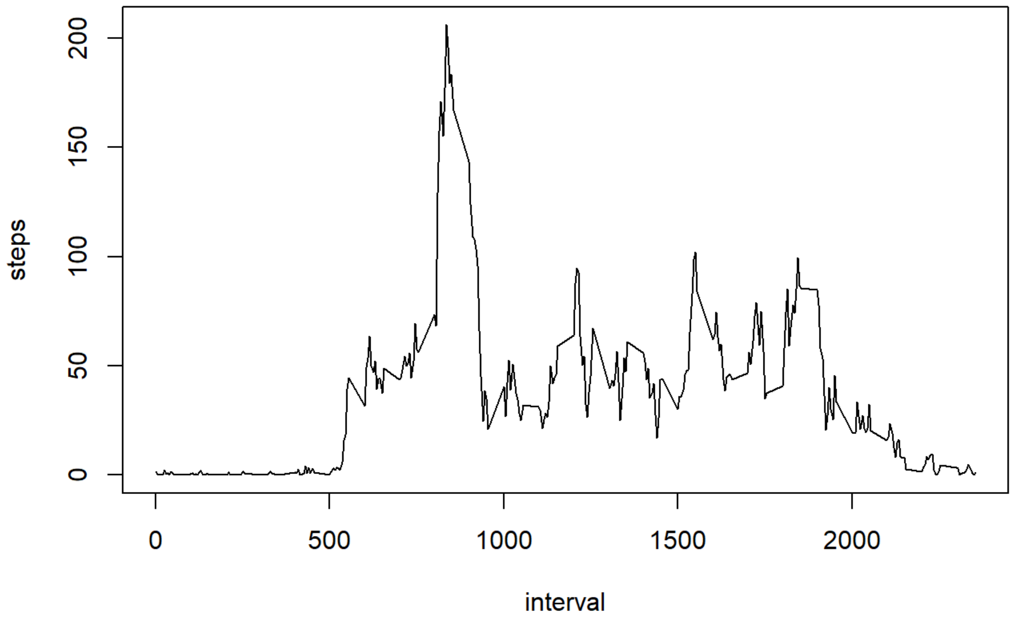
```
medianStepsPerDay <- median(stepsPerDay$steps)
medianStepsPerDay
```

```
## [1] 10765
```

What is the average daily activity pattern?

1. Make a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
stepsPerInterval<-aggregate(steps~interval, data=activityData, mean, na.rm=TRUE)
plot(steps~interval, data=stepsPerInterval, type="l")
```



2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
intervalWithMaxNbSteps <- stepsPerInterval[which.max(stepsPerInterval$steps),]$interval
intervalWithMaxNbSteps
```

```
## [1] 835
```

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
totalValuesMissings <- sum(is.na(activityData$steps))
totalValuesMissings
```

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

2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

Estrategy: fill missing values in the dataset with the mean per interval. Here’s the function that will return, for a particular interval, the mean value:

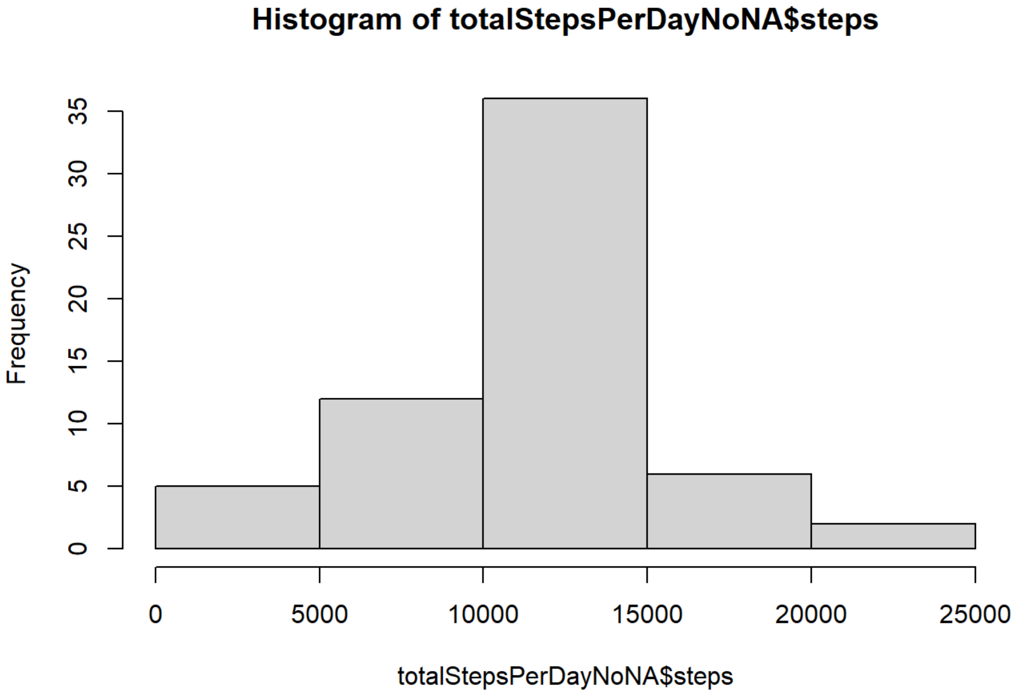
```
getMeanStepsPerInterval<-function(interval){
  stepsPerInterval[stepsPerInterval$interval==interval,]$steps

}
```

3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
activityDataNoNA<-activityData
for(i in 1:nrow(activityDataNoNA)){
  if(is.na(activityDataNoNA[i,]$steps)){
    activityDataNoNA[i,]$steps <- getMeanStepsPerInterval(activityDataNoNA[i,]$interval)
  }
}
```

4. Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

```
totalStepsPerDayNoNA <- aggregate(steps ~ date, data=activityDataNoNA, sum)
hist(totalStepsPerDayNoNA$steps)
```



```
meanStepsPerDayNoNA <- mean(totalStepsPerDayNoNA$steps)
meanStepsPerDayNoNA
```

```
## [1] 10766.19
```

```
medianStepsPerDayNoNA <- median(totalStepsPerDayNoNA$steps)
medianStepsPerDayNoNA
```

```
## [1] 10766.19
```

There is no difference between the means (10766.19), but there is a small difference in the median (10765 vs 10766.19)

Are there differences in activity patterns between weekdays and weekends?

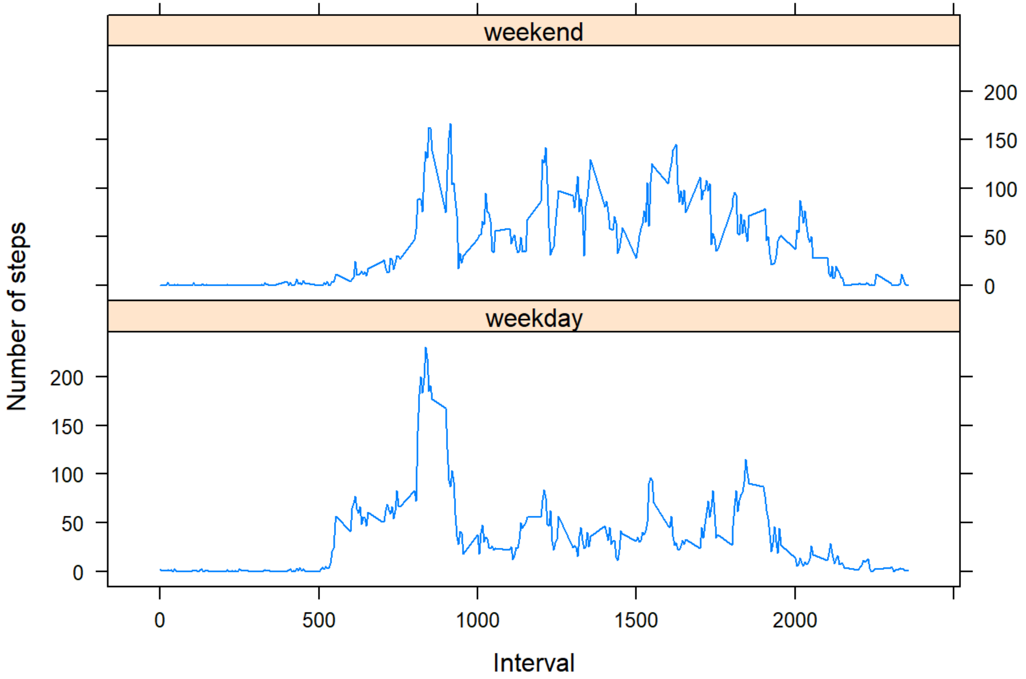
For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

1. Create a new factor variable in the dataset with two levels – “weekday” and “weekend” indicating whether a given date is a weekday or weekend day.

```
activityDataNoNA$date <- as.Date(strptime(activityDataNoNA$date, format="%Y-%m-%d"))
activityDataNoNA$day <- weekdays(activityDataNoNA$date)
for (i in 1:nrow(activityDataNoNA)) {
  if (activityDataNoNA[i,]$day %in% c("sábado","domingo")) {
    activityDataNoNA[i,]$day<-"weekend"
  }
  else{
    activityDataNoNA[i,]$day<-"weekday"
  }
}
stepsByDay <- aggregate(activityDataNoNA$steps ~ activityDataNoNA$interval + activityDataNoNA$day, activityDataNoNA, mean)
```

2. Make a panel plot containing a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
names(stepsByDay) <- c("interval", "day", "steps")
library(lattice)
xyplot(steps ~ interval | day, stepsByDay, type = "l", layout = c(1, 2),
       xlab = "Interval", ylab = "Number of steps")
```



References:

1. https://github.com/schen57/Reproducible-Research-Course-Project-1/blob/master/Reproducible%20Research%20Project%201.Rmd
2. https://rpubs.com/julienbalmont/Coursera-Reproducible-research-Course-project-1
3. https://josephtabadero.neocities.org/PA1_template.html