Reproducible Research: Peer assessment 1

FR

Github repository with RMarkdown source code: https://github.com/froediger/FR_RepData_PeerAssessment1 (https://github.com/froediger/FR_RepData_PeerAssessment1)

Introduction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the “quantified self” movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

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.

Data

The data for this assignment can be downloaded from the course web site:

Dataset: Activity Link: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip)

The variables included in this dataset are:

```
steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)

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

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.

Load required libraries

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.1.3
```

```
library(lattice)
```

Load the required data and preprocess the data

Load the data using `read.csv()`:

```
rdata <- read.csv('activity.csv', header = TRUE, sep = ",",
                  colClasses=c("numeric", "character", "numeric"))
```

Transform the column date to **date** class and **interval** class:

```
rdata$date <- as.Date(rdata$date, format = "%Y-%m-%d")
rdata$interval <- as.factor(rdata$interval)
```

check data using `str()` and `names()`:

```
str(rdata)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: Factor w/ 288 levels "0","5","10","15",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
names(rdata)
```

```
## [1] "steps"    "date"     "interval"
```

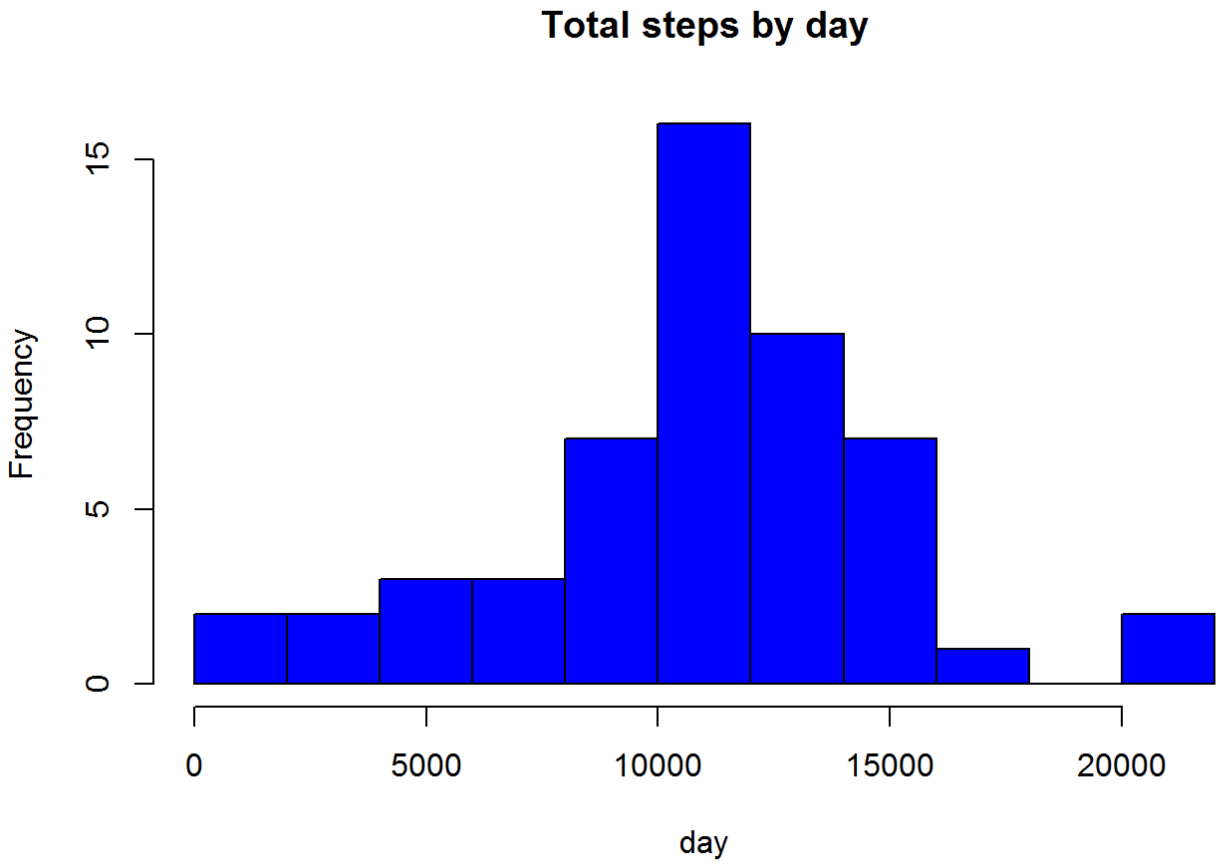
What is mean total number of steps taken per day?

Ignore missing values (*a valid assumption*).

histogram

Building new variable using `aggregate()` to build histogram using `hist()`

```
StepsTotal <- aggregate(steps ~ date, data = rdata, sum, na.rm = TRUE)
hist(StepsTotal$steps, breaks = 8, main = "Total steps by day", xlab = "day", col = "blue")
```



mean and median

Calulating the mean and median using `mean()` and `median()`

```
mean(StepsTotal$steps)
```

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

```
median(StepsTotal$steps)
```

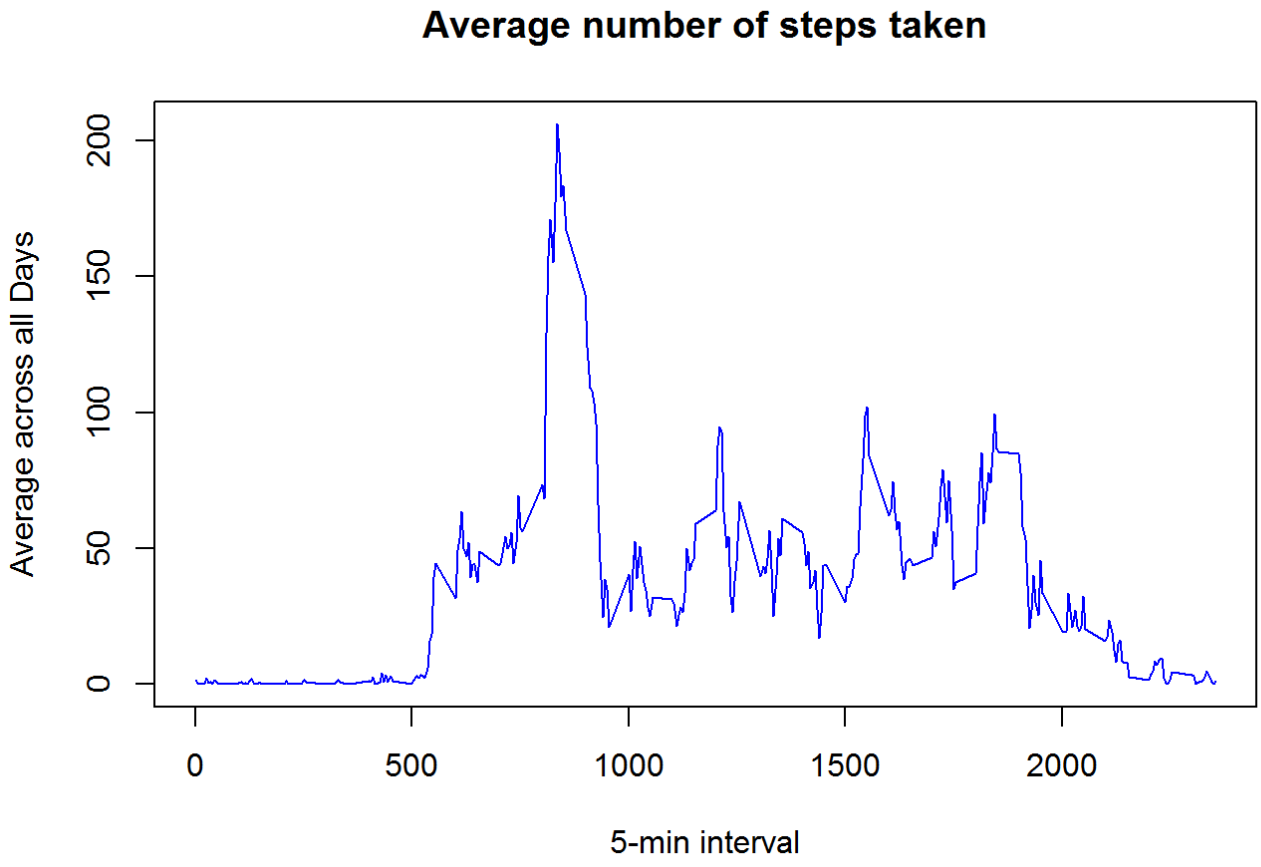
```
## [1] 10765
```

What is the average daily activity pattern?

Building a time series using `tapply()` in preperation to build a plot of the 5-minute interval using `plot()`

```
time_series <- tapply(rdata$steps, rdata$interval, mean, na.rm = TRUE)
```

```
plot(row.names(time_series), time_series, type = "l", xlab = "5-min interval",
    ylab = "Average across all Days", main = "Average number of steps taken",
    col = "blue")
```



Answering the question: **Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?** by using `whicxh.max()` and `names()`

```
max_interval <- which.max(time_series)
names(max_interval)
```

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

Imputing missing values

Calculate the count of missing values using `sum(is.na())`

```
activity_NA <- sum(is.na(rdata))
activity_NA
```

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

filling in all of the missing values in the dataset with a `loop`

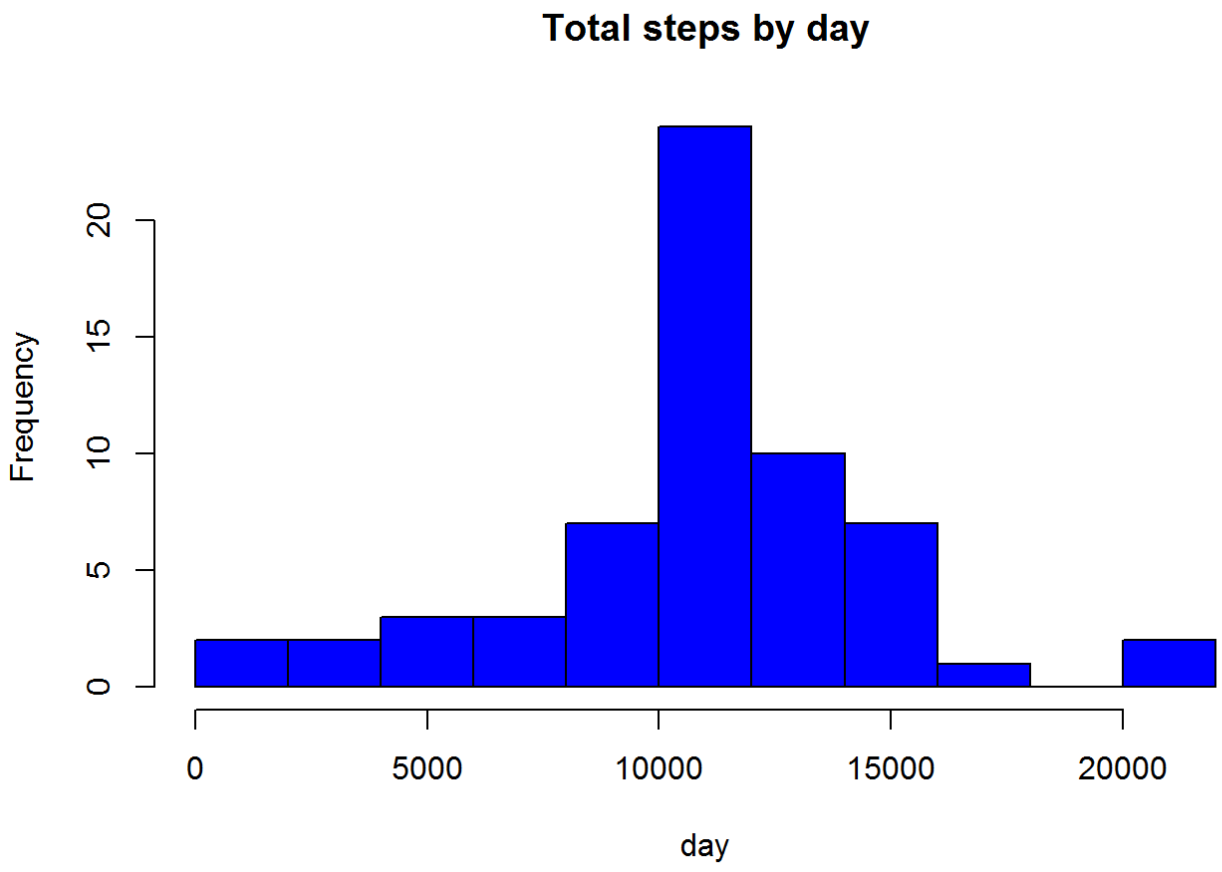
```
StepsAverage <- aggregate(steps ~ interval, data = rdata, FUN = mean)
fillNA <- numeric()
for (i in 1:nrow(rdata)) {
    obs <- rdata[i, ]
    if (is.na(obs$steps)) {
        steps <- subset(StepsAverage, interval == obs$interval)$steps
    } else {
        steps <- obs$steps
    }
    fillNA <- c(fillNA, steps)
}
```

```
new_activity <- rdata
new_activity$steps <- fillNA
```

```
StepsTotal2 <- aggregate(steps ~ date, data = new_activity, sum, na.rm = TRUE)
```

Builing a histogram after filling missing values using `hist()`

```
hist(StepsTotal2$steps,breaks = 8, main = "Total steps by day", xlab = "day", col = "blue")
```



Calulating the mean and median using `mean()` and `median()`

```
mean(StepsTotal2$steps)
```

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

```
median(StepsTotal2$steps)
```

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

Are there differences in activity patterns between weekdays and weekends?

Creating a factor variable to have aplit in **weekday or weekend** using `weekdays()` (*Using the dataset with the filled-in missing values*)

```
day <- weekdays(rdata$date)
daylevel <- vector()
for (i in 1:nrow(rdata)) {
    if (day[i] == "Saturday") {
        daylevel[i] <- "Weekend"
    } else if (day[i] == "Sunday") {
        daylevel[i] <- "Weekend"
    } else {
        daylevel[i] <- "Weekday"
    }
}
rdata$daylevel <- daylevel
rdata$daylevel <- factor(rdata$daylevel)

stepsByDay <- aggregate(steps ~ interval + daylevel, data = rdata, mean)
names(stepsByDay) <- c("interval", "daylevel", "steps")
```

Creating the using `xyplot()`

```
xyplot(steps ~ interval | daylevel, stepsByDay, type = "l", layout = c(1, 2), main = "Weekday v
s. Weekend",
    xlab = "Interval", ylab = "Number of steps")
```

Weekday vs. Weekend