Reproducible Research - Peer Assessment 1

Michel Janos

15 de julho de 2015

Executive Summary

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.

Loading and preprocessing the data

```
DF <- read.csv("activity.csv",header=TRUE,na.strings="NA")
head(DF)
```

```
##   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
```

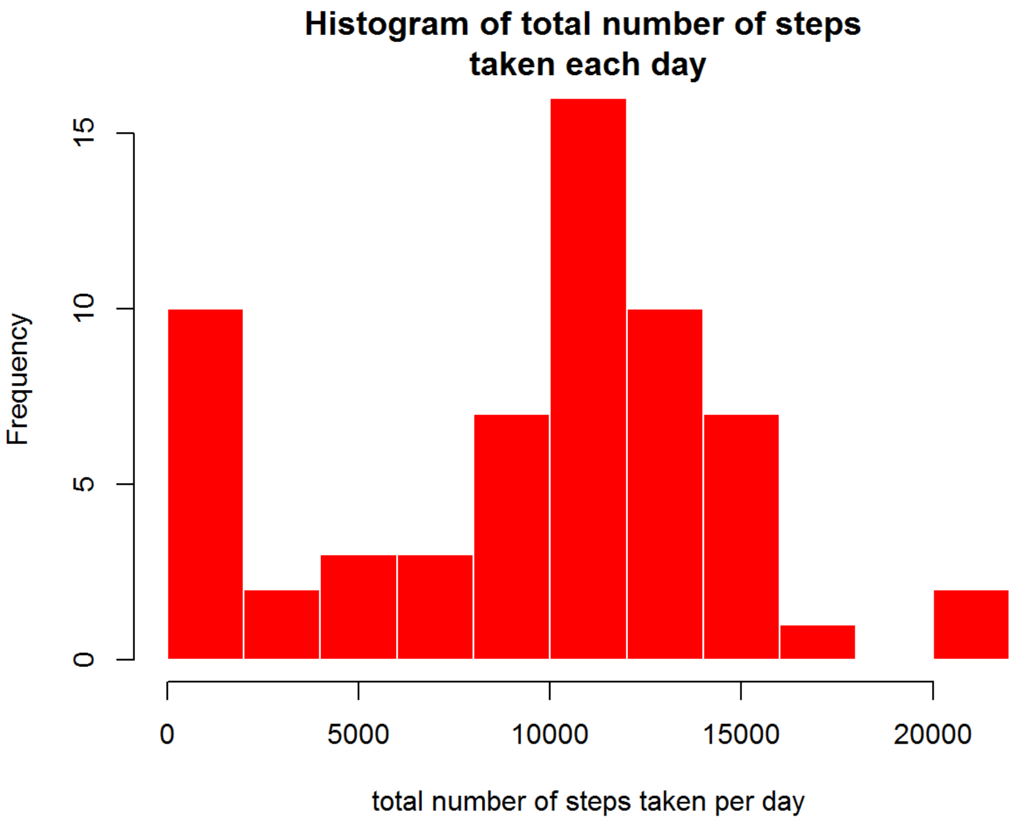
Changing the class of date

```
class(DF$date)
```

```
## [1] "factor"
```

```
DF$date <- as.Date(DF$date)
class(DF$date)
```

```
## [1] "Date"
```



What is mean total number of steps taken per day?

Histogram of the total number of steps taken each day

As we don’t have the total number of steps recorded each day in our dataset we need to sum the data for each day.

```
steps_T <- tapply(DF$steps, DF$date, FUN=sum, na.rm=TRUE)

par(mar=c(5,5,2,3))

hist(steps_T,breaks=11,freq=TRUE,border=F,col= "red",main="Histogram of total number of s
teps \ntaken each day",xlab="total number of steps taken per day")
```

Mean and Median total number of steps taken per day

The mean and median are computed and the missing values are removed.

```
mean1 <- mean(steps_T, na.rm=TRUE)

median1 <- median(steps_T, na.rm=TRUE)

mean1

## [1] 9354.23

median1

## [1] 10395
```

On average, 9354 steps are taken per day. The median is 10395.

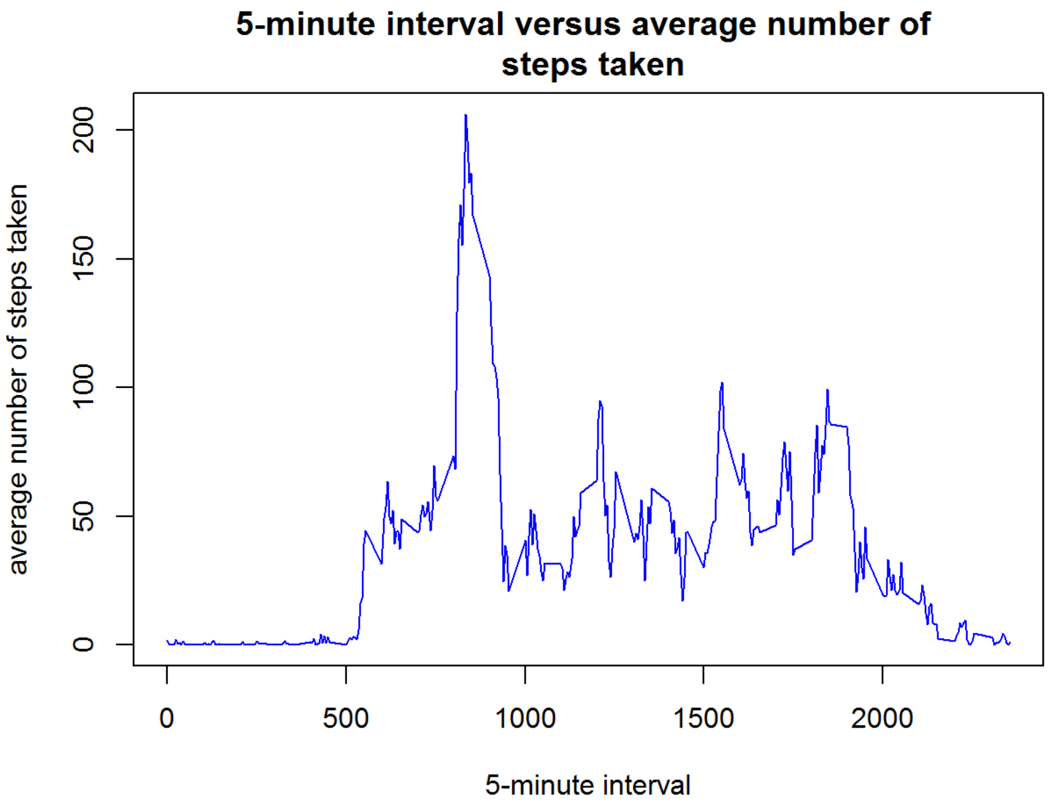
What is the average daily activity pattern?

We plot a time series of the 5-minute interval and the average number of steps taken, averaged across all days.

```
par(mar=c(5,5,3,3))

av <- aggregate(x=list(steps=DF$steps),by=list(interval=DF$interval),FUN=mean,na.rm=TRUE)

plot(av$steps~av$interval,type="l",col="blue",xlab="5-minute interval",pch=2,main= "5-min
ute interval versus average number of \n steps taken",ylab="average number of steps taken
")
```



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

```
av[which.max(av$steps),]

##     interval    steps
## 104      835 206.1698
```

The maximum number of steps is 206.1698 in interval 835.

Imputting Missing Values

Total number of 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.
With the table and is.na() functions we find the NA’a number.

```
table(is.na(DF))
```

```
##
## FALSE  TRUE
## 50400  2304
```

There are 2304 missing values in this data set.

Filling in all the missing values of the data set

We replace the missing values by the mean for the corresponding 5-minute interval.

```
fil <- function(steps, interval) {
filled <- NA
if (!is.na(steps))
filled <- steps
else
filled <- av[av$interval==interval, "steps"]
return(filled)
}
```

New dataset with missing data filled

Show that all the missing values are gone using the table and is.na() functions.

```
newDF <- DF
newDF$steps <- mapply(fil, newDF$steps, newDF$interval)
```

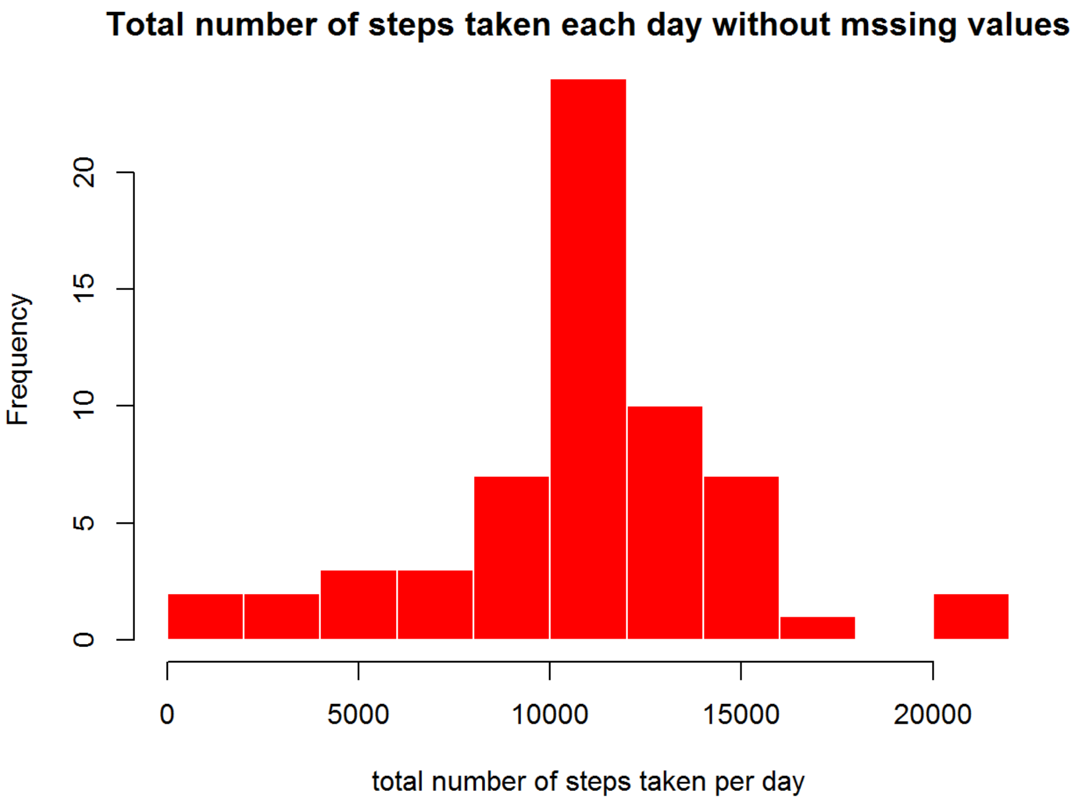
show that the missing values are gone

```
table(is.na(newDF))
```

```
##
## FALSE
## 52704
```

4.Histogram of the total number of steps taken each day using the new data set

```
steps_T2 <- tapply(newDF$steps, newDF$date, FUN=sum, na.rm=TRUE)
par(mar=c(5,5,2,3))
hist(steps_T2,
breaks=11,
freq=TRUE,
border=FALSE,
col= "red",
main="Total number of steps taken each day without mssing values",xlab="total number of steps taken per day")
```



Are there differences in activity patterns between weekdays and weekends?

Create new factor variable with two levels - “weekday” and “weekend”

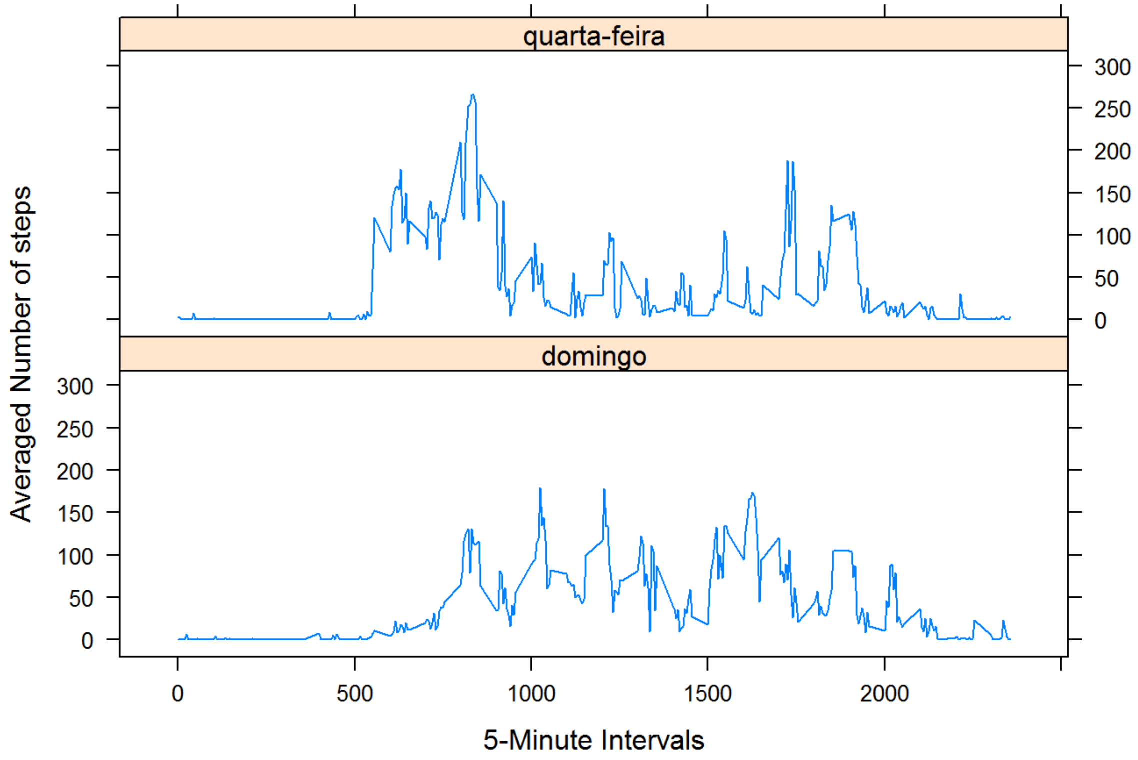
```
newDF$date <- as.Date(newDF$date)
newDF$weekdays <- format(newDF$date, "%A")
levels(newDF$weekdays) <- list(weekday = c("segunda-feira", "terça-feira","quarta-feira", "quinta-feira", "sexta-feira"),weekend = c("sábado", "domingo"))
head(newDF)
```

```
##        steps       date interval      weekdays
## 1 1.7169811 2012-10-01        0 segunda-feira
## 2 0.3396226 2012-10-01        5 segunda-feira
## 3 0.1320755 2012-10-01       10 segunda-feira
## 4 0.1509434 2012-10-01       15 segunda-feira
## 5 0.0754717 2012-10-01       20 segunda-feira
## 6 2.0943396 2012-10-01       25 segunda-feira
```

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

```
library(lattice)
av <- aggregate(steps ~ interval + weekdays, data=newDF, mean)
xyplot(av$steps ~ av$interval | av$weekdays,layout = c(1, 2),type = "l",xlab = "5-Minute
Intervals",ylab = "Averaged Number of steps")
```


quarta-feira
domingo
Averaged Number of steps
5-Minute Intervals