Reproductible Research: Peer Assignment 1

Yue Liu

1/4/2017

Overview

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 Load the data and transform data when necessary. Remove NAs to begin with since they might cause biased results.

```
#set working directory
setwd("/Users/Eva/Documents/Coursera Courses/ReproductibleResearch")
#Load the data
rawdata <- read.csv("activity.csv", stringsAsFactors = FALSE)
head(rawdata)
```

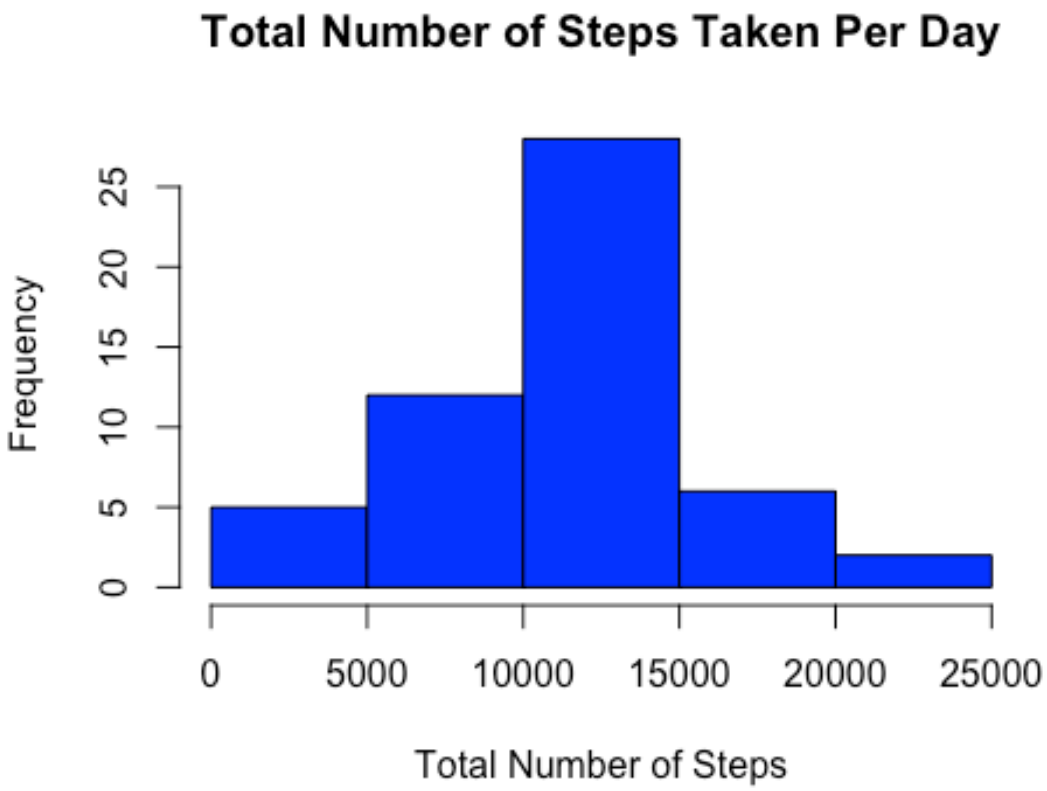
```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```
#transform the date data
rawdata$date <- as.POSIXct(rawdata$date, format="%Y-%m-%d")
```

What is mean total number of steps taken per day?

To begin with, let's look at the histogram of the total number of steps taken each day (missing values are removed) .

```
#remove NAs
rawdata.noNA <- na.omit(rawdata)
StepsPerDay <- aggregate(steps ~ date, rawdata.noNA, FUN = sum)
#histogram
hist(StepsPerDay$steps, col="blue", main="Total Number of Steps Taken P
er Day",xlab="Total Number of Steps")
```




```
#calculate the mean and median
```

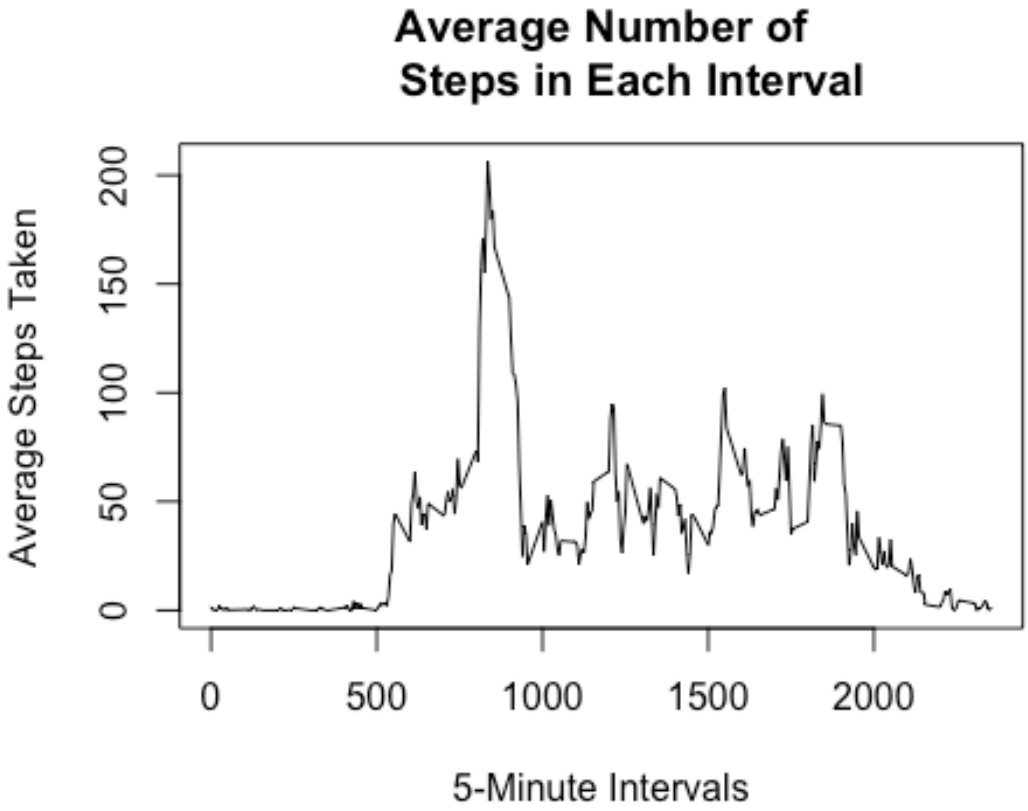
The mean and median of the total steps taken each day are:

```
TotalStepsMean <- mean(StepsPerDay$steps)
TotalStepsMean

## [1] 10766.19

TotalStepsMed <- median(StepsPerDay$steps)
TotalStepsMed

## [1] 10765
```

What is the average daily activity pattern?

We are going to make a time series plot of the average steps taken across all days in each 5-minute intervals during a day.

```
StepsPerInterval <- aggregate(steps ~ interval, rawdata.noNA, mean)
plot(StepsPerInterval$interval,type = "l", StepsPerInterval$steps, main
="Average Number of
     Steps in Each Interval",xlab="5-Minute Intervals",ylab="Average St
eps Taken")
```



Next, we will find out which 5-minute interval has the maximum average steps.

```
#find out the interval for the maximum average steps
index <- which.max(StepsPerInterval$steps)
StepsPerInterval[index,]

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

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as ⍰⍰). The presence of missing days may introduce bias into some calculations or summaries of the data. First, let's find out how many missing values we have.

```
sum(is.na(rawdata$steps))

## [1] 2304
```

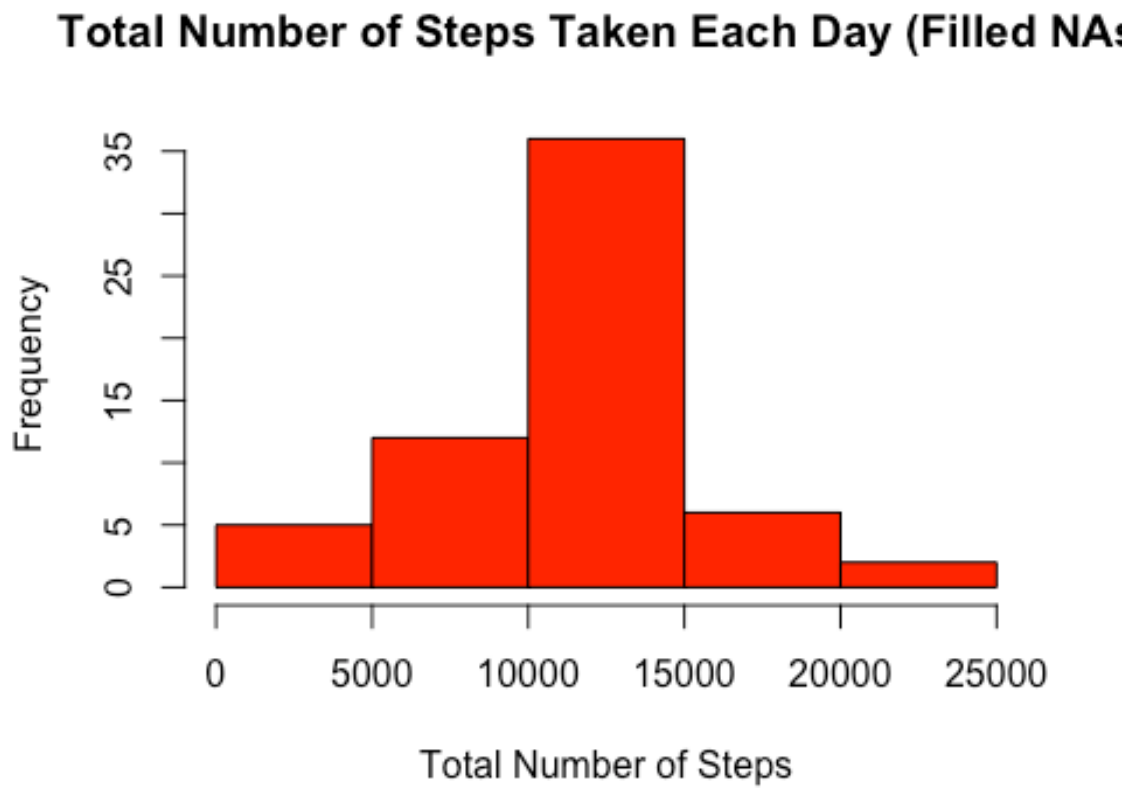
Now we are going to replace the NAs with the average steps for that 5-minute interval. And then we create a new dataset.

```
rawdata2 <- rawdata
for (i in which(is.na(rawdata2$steps))){
```

```
        #get the 5-minute interval corresponding to the NA, and then th
e average steps for that particular interval
        int <- rawdata2$interval[i]
        avgStepsInt <- StepsPerInterval[which(StepsPerInterval$interval
 == int),]$steps
        #assign the mean steps for NA
        rawdata2$steps[i] <- avgStepsInt
}
#compute total number of steps in each day with filled NA values
StepsPerDayFilled <- aggregate(steps ~ date, rawdata2,sum)
```

In addition, let's look at the histogram of the total number of steps taken with NA being replaced with the average steps for the corresponding 5-minute interval. Then we compute the new mean and median of the total steps taken each day. And we can compare our results with the numbers we got earlier.

```
hist(StepsPerDayFilled$steps, col="red",main="Total Number of Steps Tak
en Each Day (Filled NAs)",
     xlab="Total Number of Steps")
```




```
TotalStepsMean2 <- mean(StepsPerDayFilled$steps)
TotalStepsMean2

## [1] 10766.19
```

```
TotalStepsMed2 <- median(StepsPerDayFilled$steps)
TotalStepsMed2

## [1] 10766.19
```

Are there differences in activity patterns between weekdays and weekends?

We will use the new dataset we created before with NAs being filled in. To be begin with, we will create a new factor variable indicating whether the given date is a weekday or weekend day.

```
rawdata2$daytype <- factor(ifelse(weekdays(rawdata2$date) %in% c("Saturday","Sunday"),
                        "weekend","weekday"))
```

Next, we draw time series plots showing the average number of steps in different 5-minute intervals and different day type.

```
#time series plots of the average steps taken in each intervals which averaged across all weekdays and weekends
avgStepsPerIntervalBydaytype <- aggregate(steps ~ interval + daytype, rawdata2, mean)
library(lattice)
xyplot(steps ~ interval|daytype, data=avgStepsPerIntervalBydaytype
       ,type="l",layout=c(1,2),main="Average Steps in 5-Minute Intervals During Weekends vs Weekdays",
       xlab="5-Minute Intervals",ylab="Average Steps Taken")
```

e Steps in 5-Minute Intervals During Weekends vs We