Reproducible research course project 1

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Loading and processing the data:

```
setwd("~/Documents/Rlearning/RepData_PeerAssessment1")
data <- read.csv(unz("activity.zip", "activity.csv"), header = TRUE,
                 sep = ",")
```

```
data$date <- as.Date(data$date)
data <- transform(data, as.factor(date))
summary(data)
```

```
##      steps             date               interval
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8
##  Median :  0.00   Median :2012-10-31   Median :1177.5
##  Mean   : 37.38   Mean   :2012-10-31   Mean   :1177.5
##  3rd Qu.: 12.00   3rd Qu.:2012-11-15   3rd Qu.:1766.2
##  Max.   :806.00   Max.   :2012-11-30   Max.   :2355.0
##  NA's   :2304
```

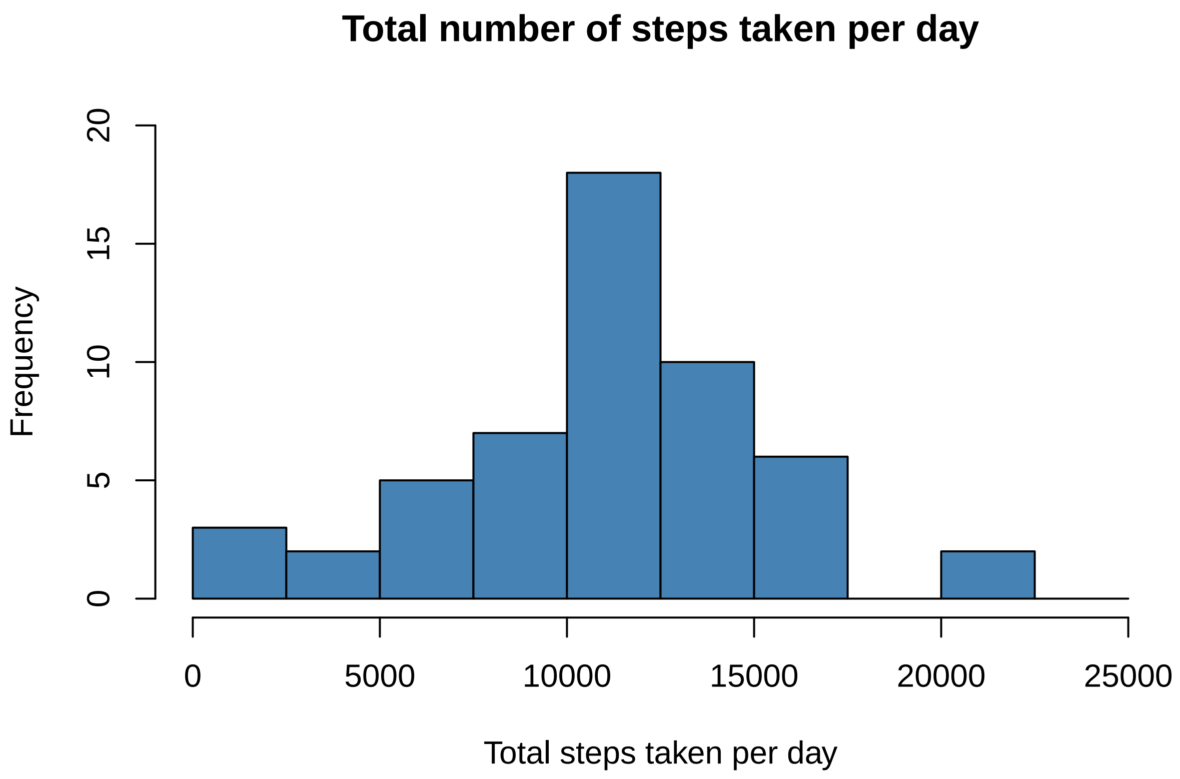
What is mean total number of steps taken per day?

1. Calculate the total number of steps taken per day

```
stepbydate <- aggregate(steps ~ date, data, sum)
```

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

```
hist(stepbydate$steps, main = "Total number of steps taken per day",
     xlab = "Total steps taken per day", col = "steelblue",
     ylim = c(0,20), breaks = seq(0,25000, by=2500))
```



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

```
mean(stepbydate$steps)
```

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

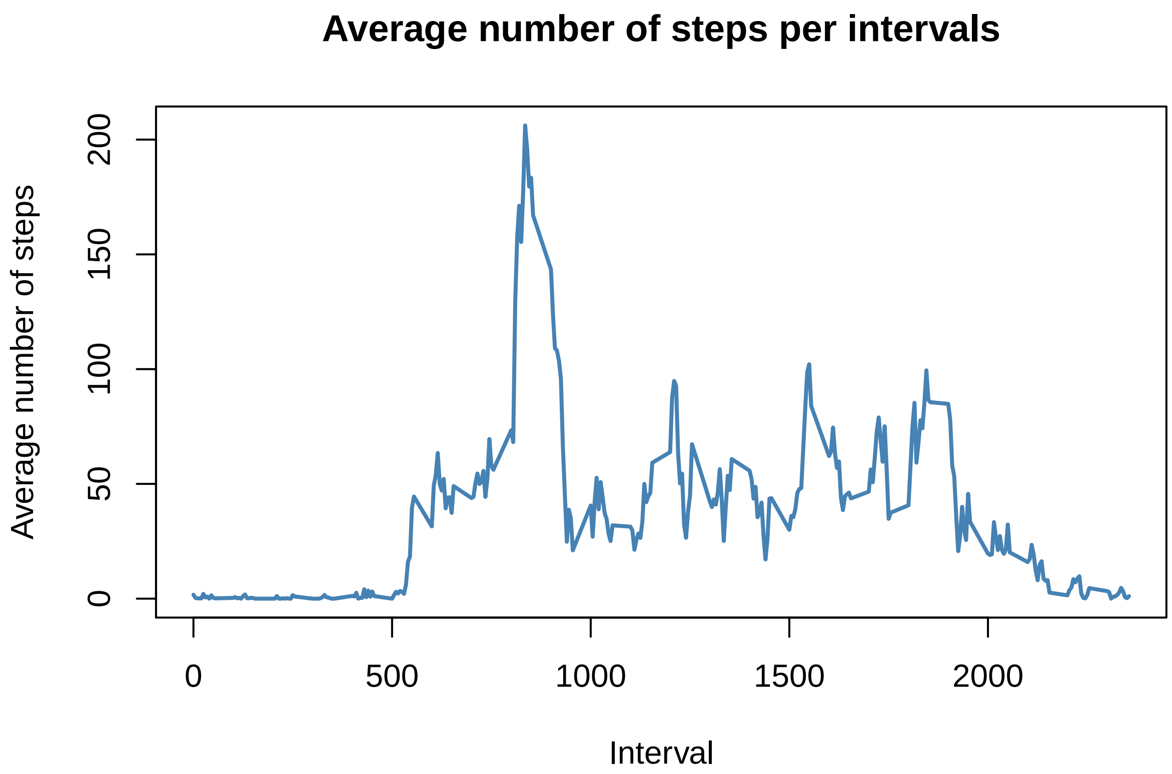
```
median(stepbydate$steps)
```

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

What is the average daily activity pattern?

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

```
stepsInterval <- aggregate(steps~interval, data=data, mean, na.rm=T)
plot(stepsInterval$interval,stepsInterval$steps, type="l",
     col="steelblue", lwd = 2,
     xlab="Interval", ylab="Average number of steps",
     main="Average number of steps per intervals")
```



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

```
stepsInterval[which.max(stepsInterval$steps), ]$interval
```

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

Imputing missing values

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

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

```
## [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.

```
imputed_steps <- stepsInterval$steps[match(data$interval, stepsInterval$interval)]
```

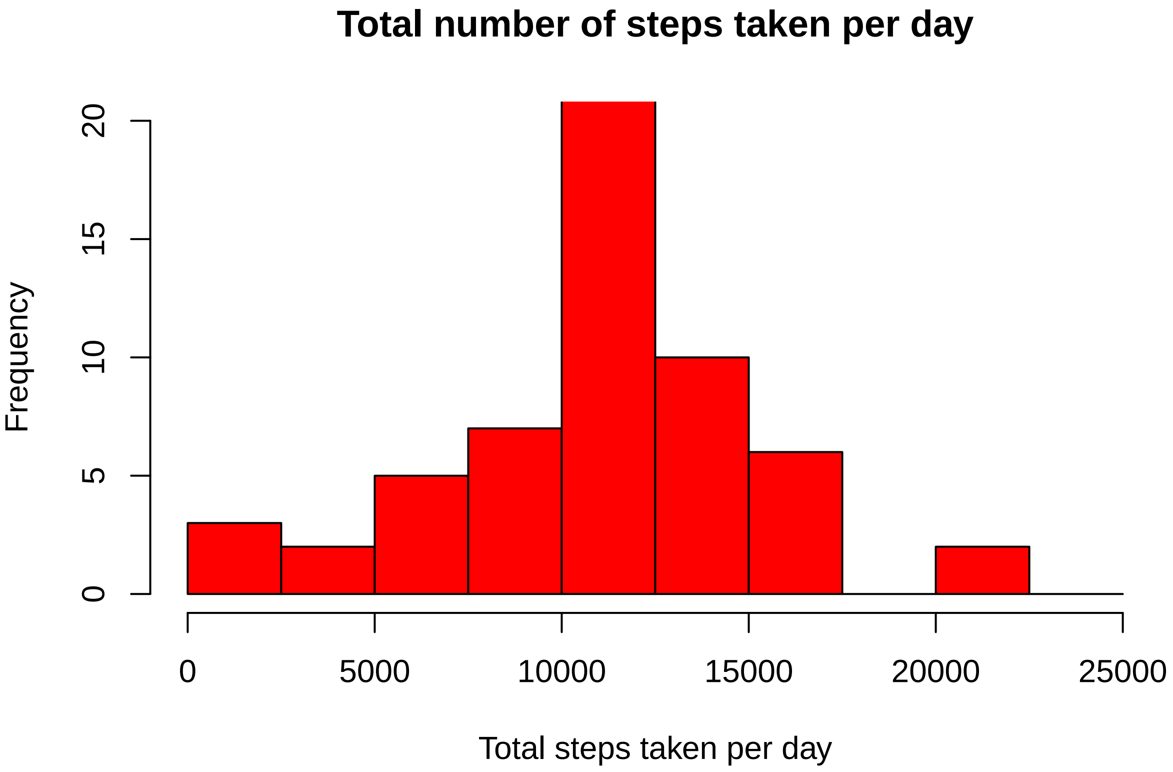
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
data_imputed <- transform(data, steps = ifelse(is.na(data$steps), yes = imputed_steps, no = data$steps)
```

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?

```
stepbydate_imputed <- aggregate(steps ~ date, data_imputed, sum)

hist(stepbydate_imputed$steps, main = "Total number of steps taken per day",
     xlab = "Total steps taken per day", col = "red",
     ylim = c(0,20), breaks = seq(0,25000, by=2500))
```



```
mean(stepbydate_imputed$steps)
```

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

```
median(stepbydate_imputed$steps)
```

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

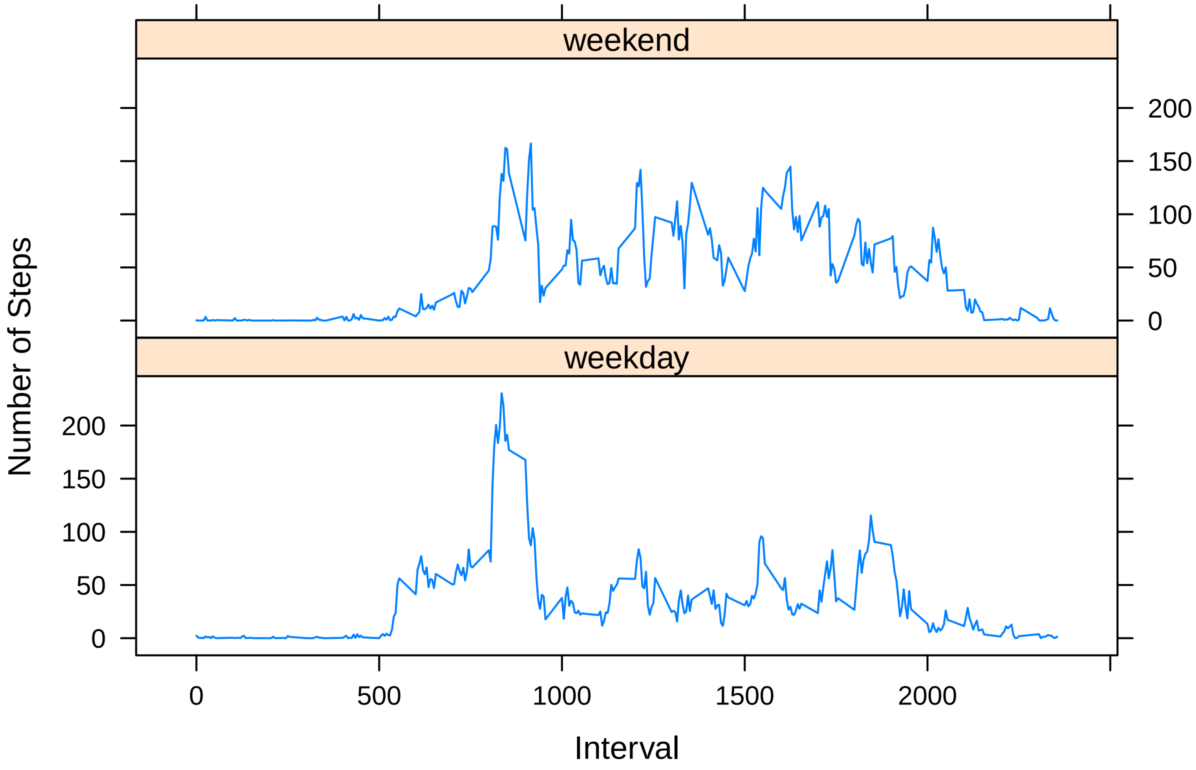
Are there differences in activity patterns between weekdays and weekends?

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.

```
data_imputed$day <-NA
for (i in 1:nrow(data_imputed)){
        if(weekdays(as.Date(data_imputed[i,]$date)) == "Saturday"|
           weekdays(as.Date(data_imputed[i,]$date)) == "Sunday"){
                data_imputed[i,]$day<-"weekend"
        }else{
                   data_imputed[i,]$day<-"weekday"
           }
}
```

2.Make a panel plot containing a time series plot (i.e. `type = "l"`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.

```
data_imputed$day<-factor(data_imputed$day)
stepbyday_imputed <- aggregate(steps~interval + day,
                               data_imputed, mean, na.rm = TRUE)
library(lattice)
xyplot(steps ~ interval | factor(day),
       data=stepbyday_imputed,
       type="l",
       layout = c(1,2),
       xlab = "Interval",
       ylab = "Number of Steps")
```


weekend
weekday
Number of Steps
Interval