

Reproducible Research: Peer Assessment 1

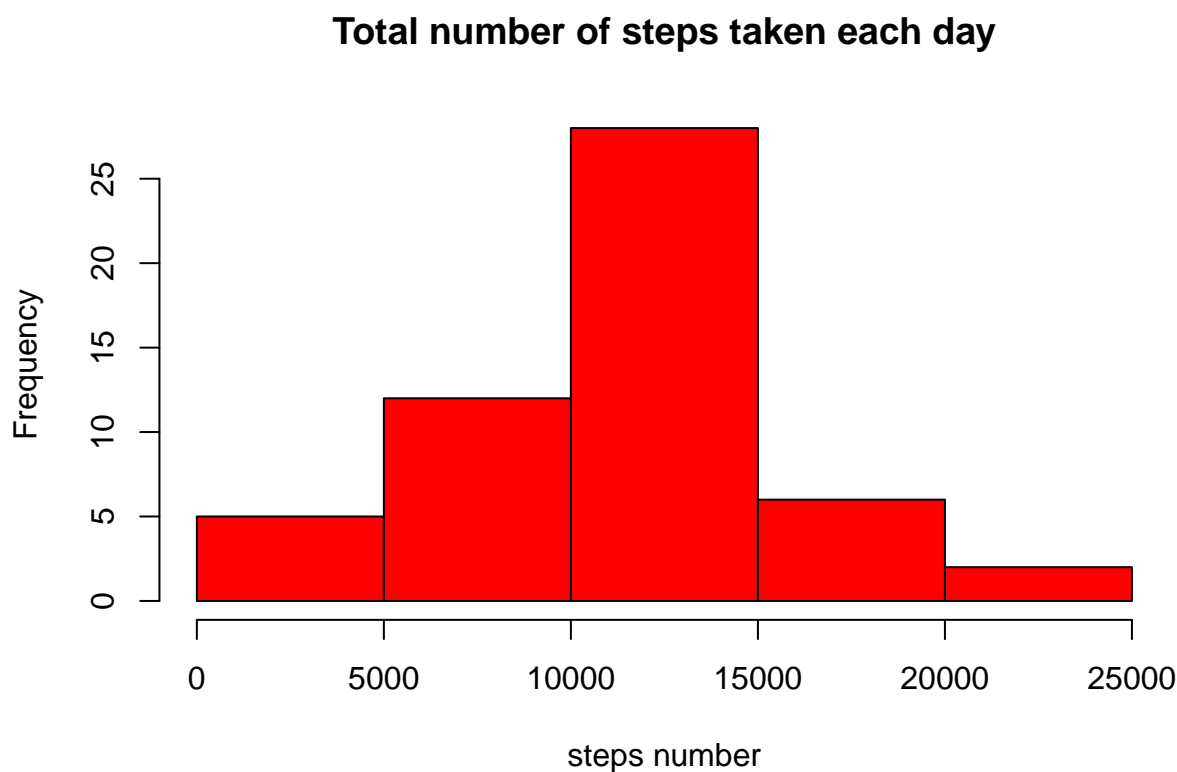
Loading and preprocessing the data

```
#read the original dataset into table
activity <- read.table("activity.csv", sep = ",", header = T)

#sum the steps by date and excluding the missing values
steps_day <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
```

What is mean total number of steps taken per day?

```
#draw the histogram of the total number of steps taken each day
hist(steps_day$steps, col = 'red', main = "Total number of steps taken each day", xlab = "steps number",
```



```
#report the mean and median number of steps taken each day
steps_mean <- mean(steps_day$steps)
steps_mean
```

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

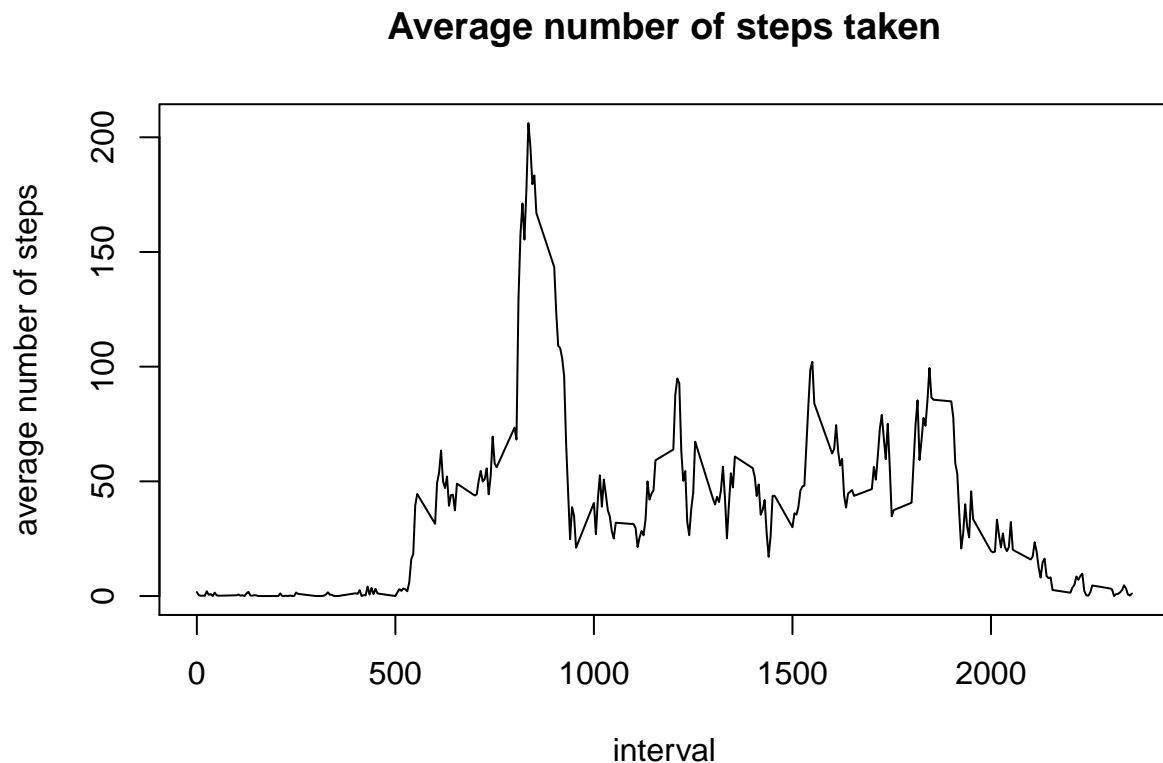
```
steps_median <- median(steps_day$steps)
steps_median
```

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

What is the average daily activity pattern?

```
#Average the number of steps taken across all dates
steps_interval <- aggregate(steps~interval, data = activity, mean, na.rm = TRUE)

#make the time series plot
plot(steps~interval, data = steps_interval, type = "l", xlab = "interval", ylab = "average number of steps taken")
```



```
#the 5-minute interval that, on average, contains the maximum number of steps
steps_interval[which.max(steps_interval$steps), "interval"]
```

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

Imputing missing values

Our strategy use the mean for that 5-minute interval to impute the missing values.

```
#new data set with filled values call activity_filled
activity_filled <- activity

#fill the missing data
for (i in 1:nrow(activity_filled)){
  if(is.na(activity_filled[i, "steps"])){
```

```

    activity_filled[i,"steps"] <- steps_interval[steps_interval$interval==activity_filled[i,"interval"]
  }

```

```

#histogram of the total number of steps taken each day after missing values are imputed
#sum the steps by date and excluding the missing values

```

```

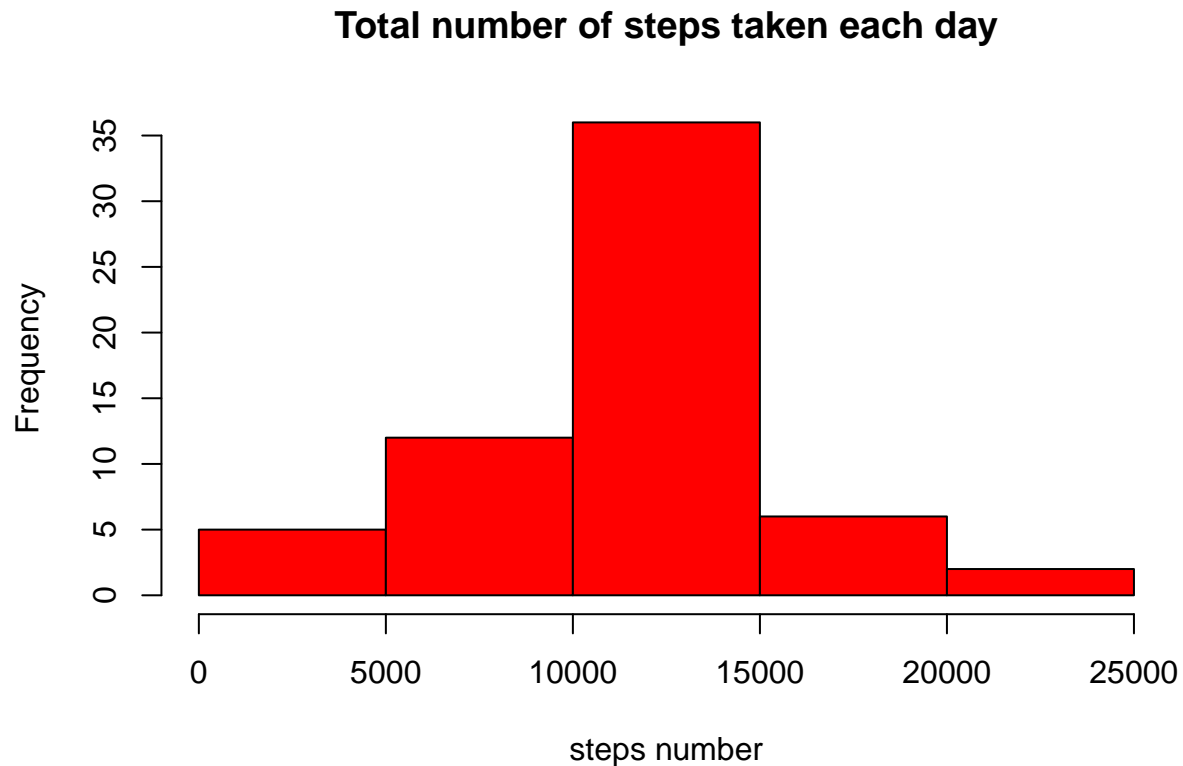
steps_day_filled <- aggregate(steps ~ date, data = activity_filled, sum)

```

```

hist(steps_day_filled$steps, col = 'red', main = "Total number of steps taken each day", xlab = "steps number")

```



```

#report the mean and median number of steps taken each day

```

```

steps_mean_filled <- mean(steps_day_filled$steps)

```

```

steps_mean_filled

```

```

## [1] 10766.19

```

```

steps_median_filled <- median(steps_day_filled$steps)

```

```

steps_median_filled

```

```

## [1] 10766.19

```

we can conclude from our calculations that the mean value is the same before and after imputing missing values but median changed.

As for the total daily number of steps, the mean value should be the same but the median should change.

Are there differences in activity patterns between weekdays and weekends?

```
library(lattice)
#utilize POSIXlt's wday (0~6 represents days) to create a new factor weekdays & weekends
activity_filled$day <- ifelse(as.POSIXlt(as.Date(activity_filled$date))$wday%%6 == 0, "weekend", "weekday")
activity_filled$day <- factor(activity_filled$day, levels = c("weekday", "weekend"))

#Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends
panel_plot_data= aggregate(steps ~ interval + day, activity_filled, mean)
xyplot(steps ~ interval | factor(day), data = panel_plot_data, type = "l", aspect = 1/2)
```

