Peer Assessment 1

```
library(ggplot2)
library(scales)
library(lattice)
```

Loading and preprocessing the data

Load data

```
activitydata <- read.csv("activity.csv", stringsAsFactors=FALSE)
```

Process data

```
#Change Date format
activitydata$date <- as.POSIXct(activitydata$date, format="%Y-%m-%d")

activitydata <- data.frame(date=activitydata$date,
                           weekday=tolower(weekdays(activitydata$date)),
                           steps=activitydata$steps,
                           interval=activitydata$interval)

activitydata <- cbind(activitydata,
                      daytype=ifelse(activitydata$weekday == "saturday" |
                                     activitydata$weekday == "sunday", "weekend",
                                     "weekday"))

activity2 <- data.frame(date=activitydata$date,
                        weekday=activitydata$weekday,
                        daytype=activitydata$daytype,
                        interval=activitydata$interval,
                        steps=activitydata$steps)
```

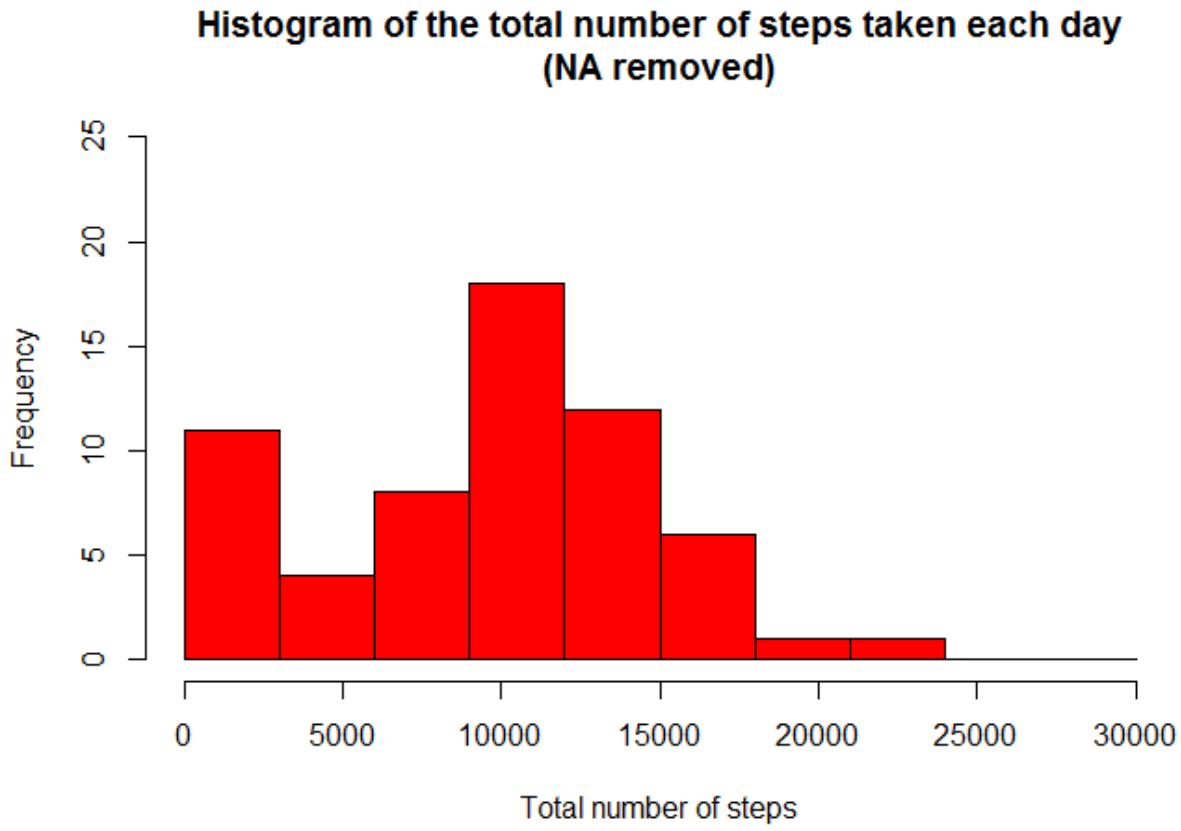
What is the mean total number of steps taken per day?

```
activitysum <- aggregate(activity2$steps, by=list(activity2$date), FUN=sum, na.rm=TRUE)

names(activitysum) <- c("date", "total")
```

1. Histogram of total steps taken

```
hist(activitysum$total,
     breaks=seq(from=0, to=30000, by=3000),
     col="red",
     xlab="Total number of steps",
     ylim=c(0, 25),
     main="Histogram of the total number of steps taken each day\n(NA removed)")
```



2. Mean

```
mean(activitysum$total)
```

```
## [1] 9354.23
```

2. Median

```
median(activitysum$total)
```

```
## [1] 10395
```

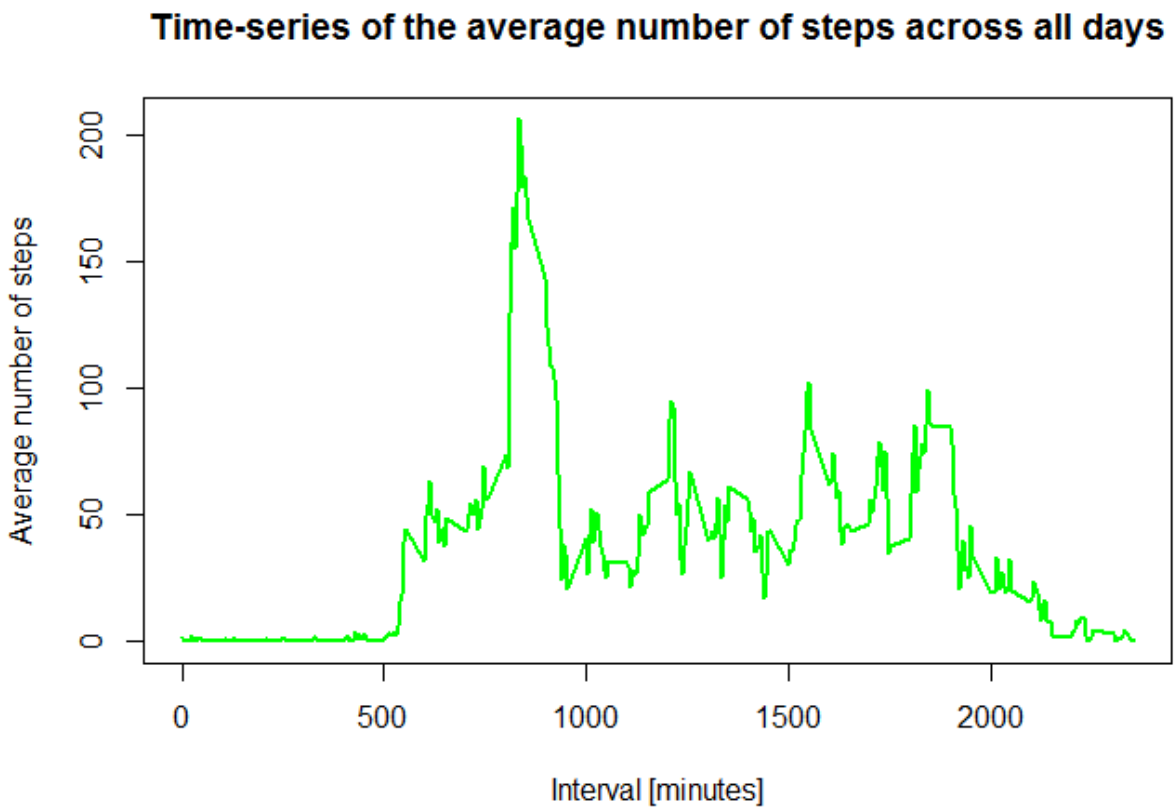
What is the average daily activity pattern?

1. Time series plot

```
activitymean <- aggregate(activity2$steps,
                          by=list(activity2$interval),
                          FUN=mean,
                          na.rm=TRUE)

names(activitymean) <- c("interval", "mean")
```

```
plot(activitymean$interval,
     activitymean$mean,
     type="l",
     col="green",
     lwd=2,
     xlab="Interval [minutes]",
     ylab="Average number of steps",
     main="Time-series of the average number of steps across all days")
```



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

```
activitymax <- activitymean[which.max(activitymean$steps),]

activitymax
```

```
## [1] interval mean
## <0 rows> (or 0-length row.names)
```

Input missing values

1. Number of NA's in dataset

```
NAcount <- sum(is.na(activity2$steps))

NAcount
```

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

The number of NA's is 2304

2. Inout values for NA's

```
NAloc <- which(is.na(activity2$steps))

NAvalue <- rep(mean(activity2$steps, na.rm=TRUE), times=length(NAloc))
```

3. Create new dataset with NA's filled in

```
activity2[NAloc, "steps"] <- NAvalue
```

Display results

```
head(activity2)
```

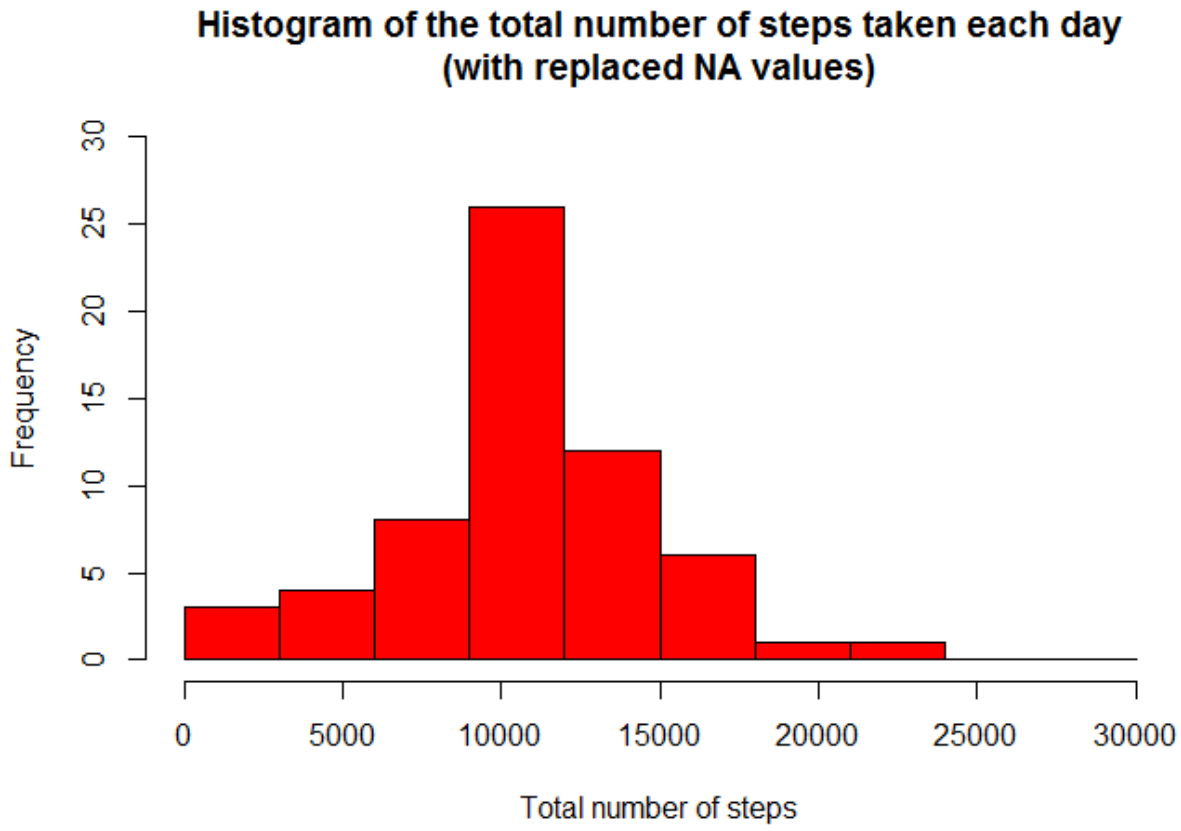
```
##         date weekday daytype interval   steps
## 1 2012-10-01  monday weekday        0 37.3826
## 2 2012-10-01  monday weekday        5 37.3826
## 3 2012-10-01  monday weekday       10 37.3826
## 4 2012-10-01  monday weekday       15 37.3826
## 5 2012-10-01  monday weekday       20 37.3826
## 6 2012-10-01  monday weekday       25 37.3826
```

4.

```
activitysum2 <- aggregate(activity2$steps, by=list(activity2$date), FUN=sum)

# Rename the attributes
names(activitysum2) <- c("date", "total")

# Compute the histogram of the total number of steps each day
hist(activitysum2$total,
     breaks=seq(from=0, to=30000, by=3000),
     col="red",
     xlab="Total number of steps",
     ylim=c(0, 30),
     main="Histogram of the total number of steps taken each day\n(with replaced NA values)")
```



New mean and median

```
mean(activitysum2$total)
```

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

```
median(activitysum2$total)
```

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

The values of the mean and median are higher, because we replaced the values of NA with the mean. Before, the records with NA's were removed and were not accounted for.

Are there differences in activity patterns between weekdays and weekends?

1. Create a new dataset with weekday and weekend

```
#This was done during the preprocessing and transforming step of the data

head(activity2)
```

```
##         date weekday daytype interval   steps
```

```
## 1 2012-10-01  monday weekday         0 37.3826
## 2 2012-10-01  monday weekday         5 37.3826
## 3 2012-10-01  monday weekday        10 37.3826
## 4 2012-10-01  monday weekday        15 37.3826
## 5 2012-10-01  monday weekday        20 37.3826
## 6 2012-10-01  monday weekday        25 37.3826
```

2. 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 weekend days (y-axis)

```
activitymean2 <- aggregate(activity2$steps,
                        by=list(activity2$daytype,
                                activity2$weekday, activity2$interval), mean)

names(activitymean2) <- c("daytype", "weekday", "interval", "mean")
```

Time series plot

```
xyplot(mean ~ interval | daytype, activitymean2,
       type="l",
       lwd=1,
       xlab="Interval",
       ylab="Number of steps",
       layout=c(1,2))
```