activity project

Loading and preprocessing the data set the working directory and read into the dataset

```
setwd("C:/Users/betty/Desktop/coursera/reproducible research/project1")
activity <- read.csv("activity.csv")
```

remove NA values

```
data_nona <- activity[complete.cases(activity),]
```

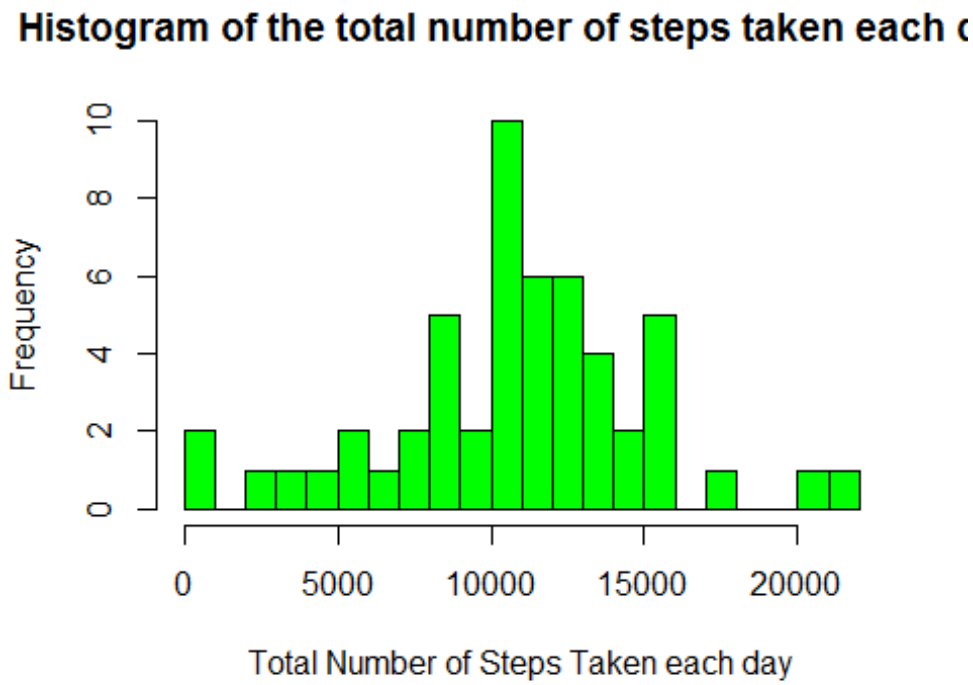
What is mean total number of steps taken per day?

For this part of the assignment, you can ignore the missing values in the dataset. Calculate and report the mean and median of the total number of steps taken per day

```
total_steps <- aggregate(steps~date,data_nona,sum)
```

Histogram of the total number of steps taken each day

```
hist(total_steps$steps, col ="green",
     xlab = "Total Number of Steps Taken each day",
     main = "Histogram of the total number of steps taken each day",
     ylab = "Frequency",breaks = 30)
```



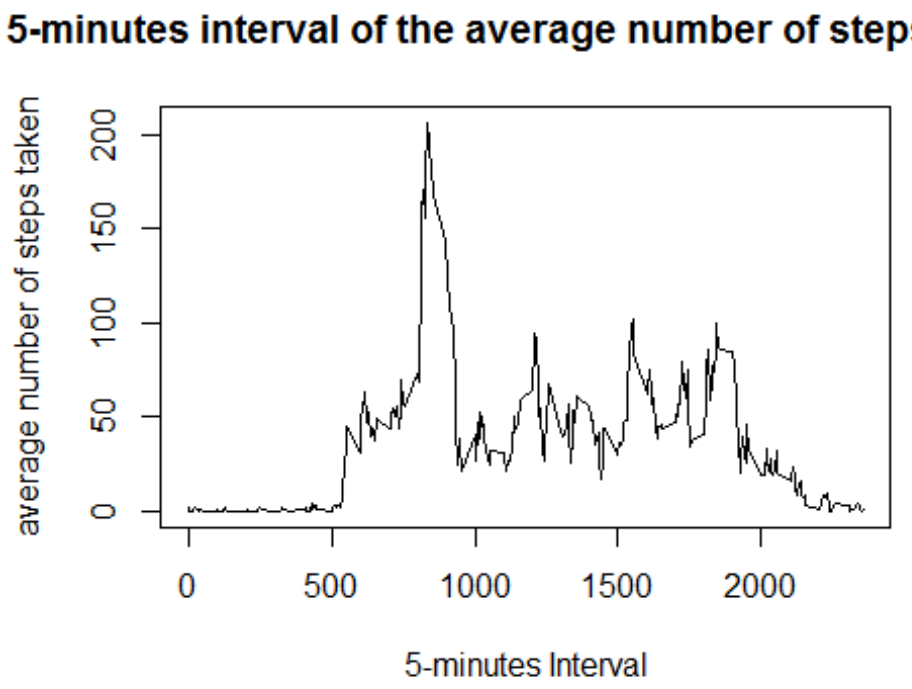
Mean and median number of steps taken each day

```
mean_steps <- mean(total_steps$steps)
median_steps <- median(total_steps$steps)
```

What is the average daily activity pattern?

Time series plot of the average number of steps taken The 5-minute interval that, on average, contains the maximum number of steps

```
aver_steps <- aggregate(steps~interval,data_nona,mean)
plot(x=aver_steps$interval,y=aver_steps$steps,type="l",
     ylab= "average number of steps taken", xlab="5-minutes Interval",
     main = "5-minutes interval of the average number of steps taken")
```



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

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

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

Imputing missing values

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

```
miss <-is.na(activity$steps)
table(miss)
```

```
## miss
## FALSE  TRUE
## 15264  2304
```

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

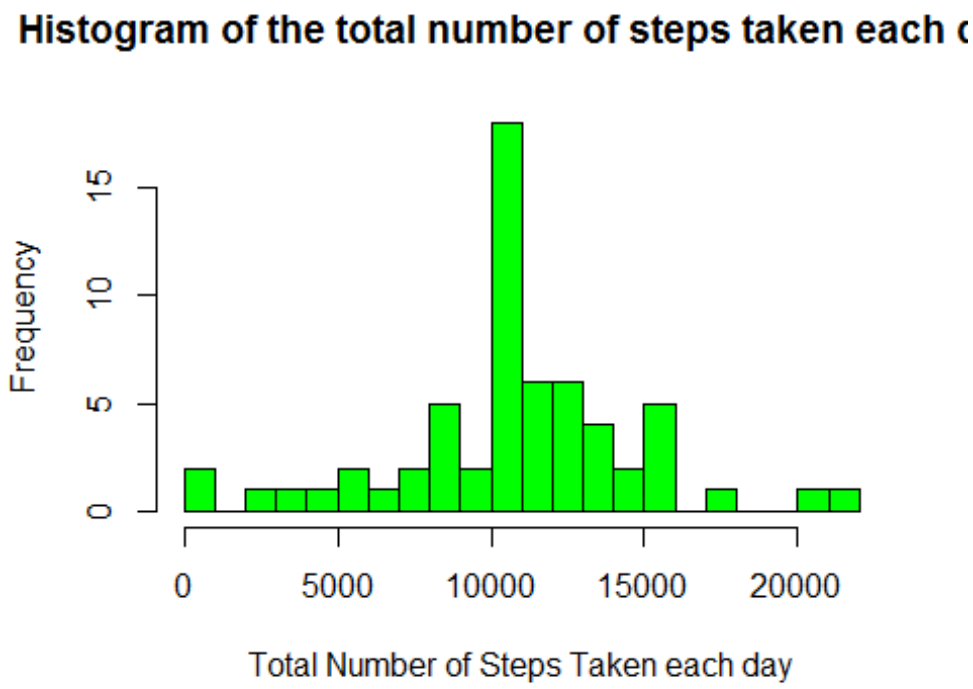
```
newdata <- activity
mis <-is.na(newdata$steps)
avg_interval <- tapply(newdata$steps, newdata$interval, mean, na.rm=TRUE,
simplify=TRUE)
newdata$steps[mis] <- avg_interval[as.character(newdata$interval[mis])]
```

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?

```
newtotal_steps <- aggregate(steps~date,newdata,sum)
```

Histogram of the total number of steps taken each day

```
hist(newtotal_steps$steps, col ="green",
     xlab = "Total Number of Steps Taken each day",
     main = "Histogram of the total number of steps taken each day",
     ylab = "Frequency",breaks = 30)
```



Mean and median number of steps taken each day

```
newmean_steps <- mean(newtotal_steps$steps)
newmedian_steps <- median(newtotal_steps$steps)
```

they do differ but not alot

Are there differences in activity patterns between weekdays and weekends? 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.

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

newdata$date <- as.Date(newdata$date)
newdata <- mutate(newdata, weektype = ifelse(weekdays(newdata$date) ==
"Saturday"
                      | weekdays(newdata$date) == "Sunday", "weekend",
"weekday"))
newdata$weektype <- as.factor(newdata$weektype)
head(newdata)

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

Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

```
newave <- aggregate(steps ~ interval + weektype, newdata, mean)
library(lattice)
xyplot( steps ~ interval|weektype,newave,type="l", layout=c(1,2),
        ylab= "average number of steps taken", xlab="5-minutes Interval",
        main = "5-minutes interval of the average number of steps taken")
```

minutes interval of the average number of steps take