Course Project 1

author: "Huib"

1. Loading and preprocessing the data

```
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
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union
```

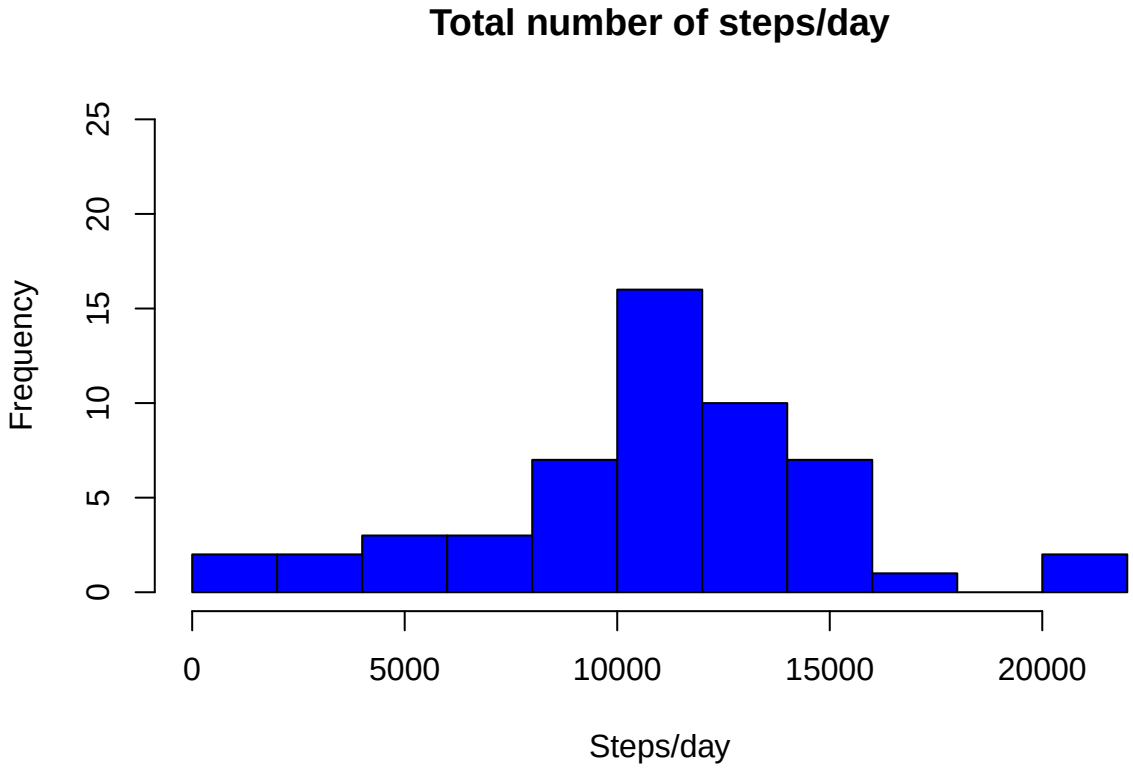
```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
setwd("/Users/huib/Documents/Huib/Coursera/DataScience/5-Reproducible Research/Assignment")
activity <- read.csv("activity.csv")
activity$date <- ymd(activity$date)
str(activity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

2. Histogram of the total number of steps taken each day

```
daystepssum <- with(activity, tapply(steps, date, sum))
hist(daystepssum, xlab = "Steps/day", ylim=c(0,25), breaks=10, main="Total number of steps/day", col="bl
```



3. The mean and median number of steps taken each day

```
mdn <- median(daystepssum, na.rm=TRUE)
mn <- mean(daystepssum, na.rm=TRUE)
```

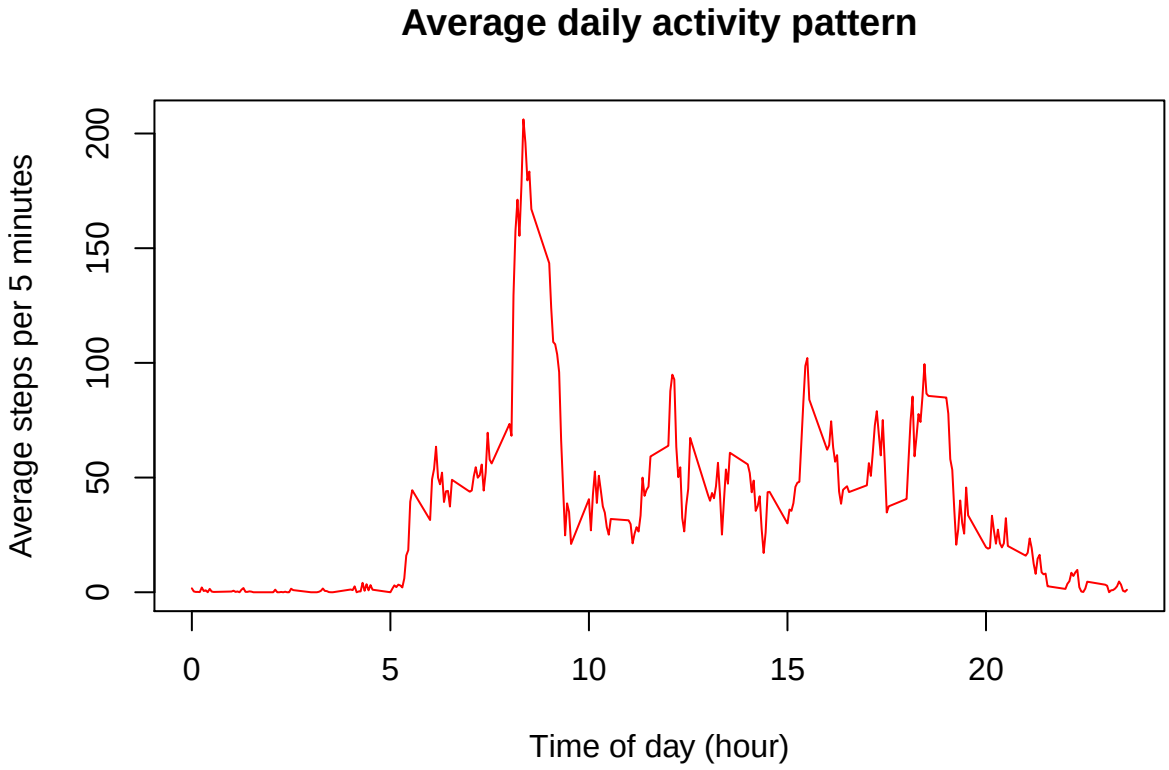
The mean and median number of steps taken each day are:

- mean: 1.0766189×10^4 steps
- median 10765 steps

4. Time series plot of the average number of steps taken

What is the average daily activity pattern?

```
Intervalstepsmean <- with(activity, tapply(steps, interval, mean, na.rm=TRUE))
stepsmean <- data.frame(as.numeric(names(Intervalstepsmean)),Intervalstepsmean)
names(stepsmean) <- c("interval","meansteps")
stepsmean$time <- stepsmean$interval/100
plot(stepsmean$time,stepsmean$meansteps, type="l", xlab = "Time of day (hour)", ylab="Average steps per
```



5. The 5-minute interval that, on average, contains the maximum number of steps

```
mxi <- stepsmean$interval[which.max(stepsmean$meansteps)]
mx <- max(stepsmean$meansteps)
```

The 5-minute interval that, on average, contains the maximum number of steps is:

- 835 with 206.1698113 steps

6. Code to describe and show a strategy for imputing missing data

```
mv <- summary(activity)
```

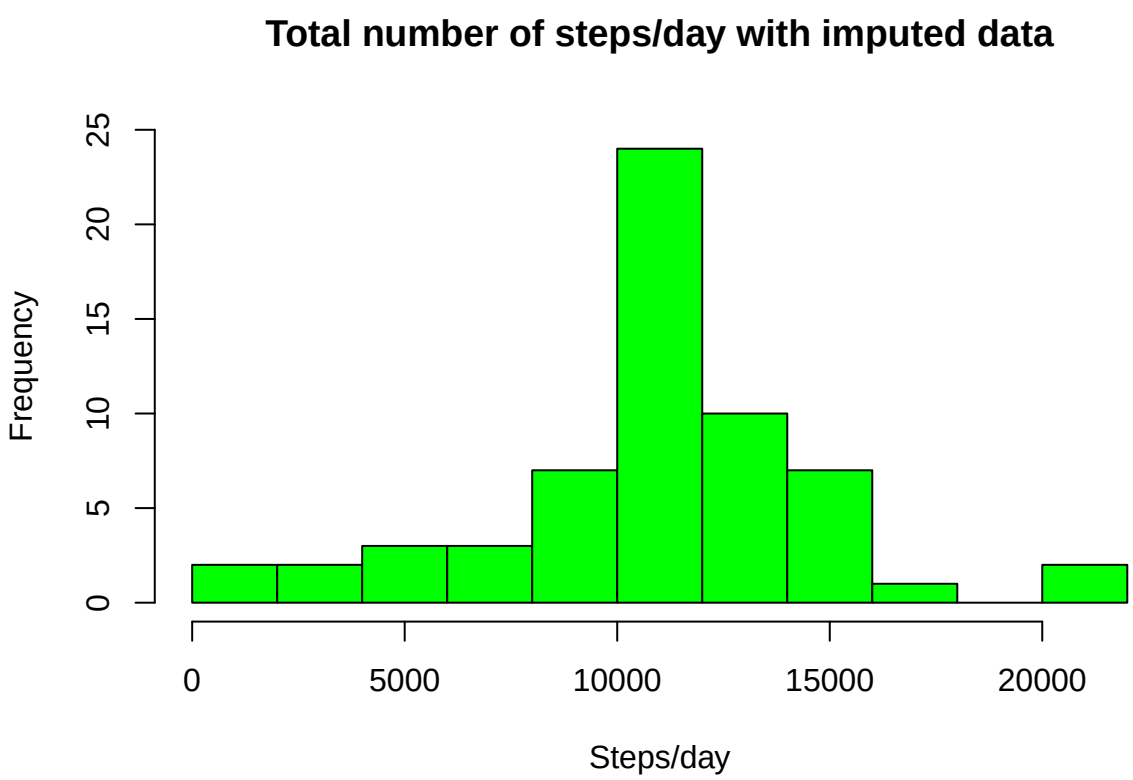
The total number of missing values is: - NA's :2304

```
newactivity <- activity
i<-1
for(i in 1:nrow(activity)){
      if (is.na(activity[i,1]) == TRUE){
        newactivity$steps[i] <- stepsmean[which(stepsmean$interval == activity[i,3]),2]
      }
      i <- i+1
}
str(newactivity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : num  1.717 0.3396 0.1321 0.1509 0.0755 ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

7. Histogram of the total number of steps taken each day after missing values are imputed

```
newdaystepssum <- with(newactivity, tapply(steps, date, sum, na.rm=TRUE))
hist(newdaystepssum, xlab = "Steps/day", ylim=c(0,25), breaks=10, main="Total number of steps/day with
```




```
newmdn <- median(newdaystepssum)
newmn <- mean(newdaystepssum)
```

After imputing NAs, the new mean and median number of steps taken each day are:

- mean: 1.0766189×10^4 steps
- median 1.0766189×10^4 steps

Mean number of steps is the same as with omiting NAs, median is slightly different. Replacing NAs with overall average values for that specific 5 minute time interval increases the median of the total daily number of steps.

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

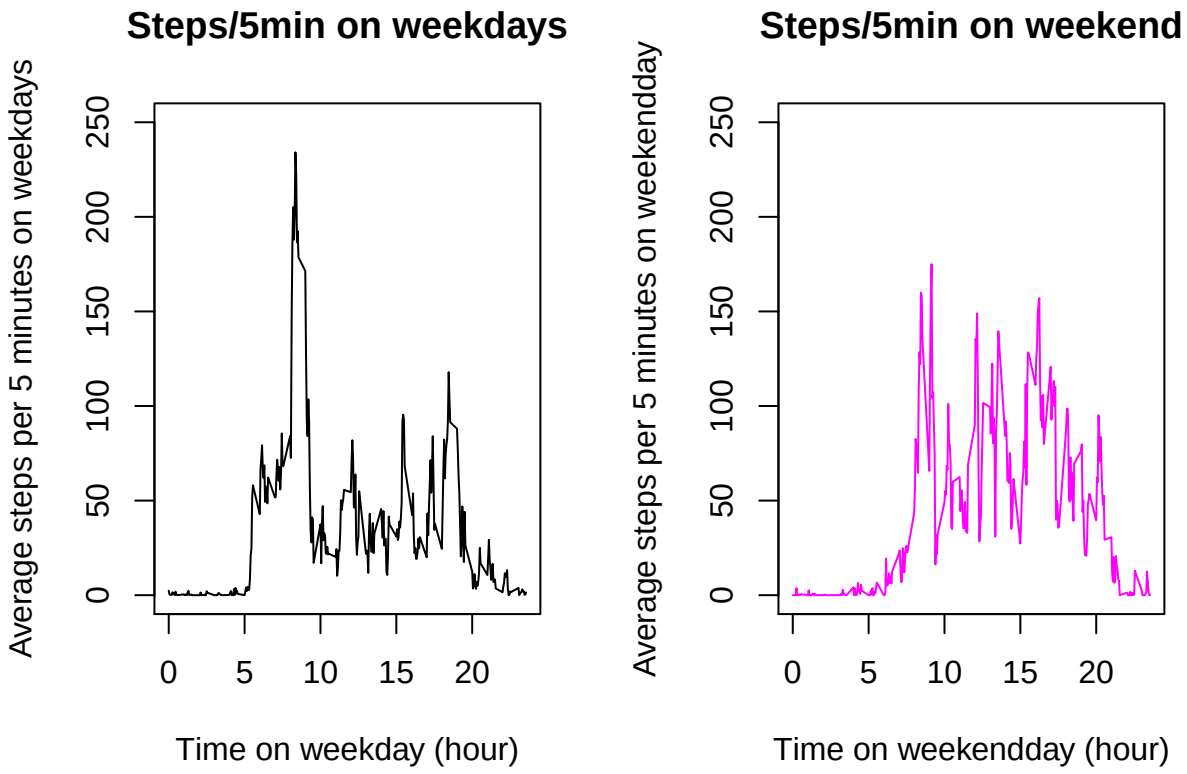
```
activity <- mutate(activity, weekday = wday(date))
actweekend <- filter(activity, weekday == 7 | weekday == 1)
actweekdays <- filter(activity, weekday > 1 & weekday < 7)

Intervalweekdays <- with(actweekdays, tapply(steps, interval, mean, na.rm=TRUE))
weekdaysmean <- data.frame(as.numeric(names(Intervalweekdays)),Intervalweekdays)
names(weekdaysmean) <- c("interval","meansteps")
weekdaysmean$time <- weekdaysmean$interval/100

Intervalweekends <- with(actweekend, tapply(steps, interval, mean, na.rm=TRUE))
weekendmean <- data.frame(as.numeric(names(Intervalweekends)),Intervalweekends)
```

```r
names(weekendmean) <- c("interval","meansteps")
weekendmean$time <- weekendmean$interval/100

par(mfrow = c(1, 2))
plot(weekdaysmean$time,weekdaysmean$meansteps, type="l", ylim=c(0,250), xlab = "Time on weekday (hour)"
plot(weekendmean$time,weekendmean$meansteps, type="l", ylim=c(0,250), xlab = "Time on weekendday (hour)
```



Activity starts later on weekends than on weekdays and seems to be more spread out during the entire day.