Reproducible Research: Peer Assessment 1

Mark Gao

Monday, April 18, 2015

Loading and preprocessing the data

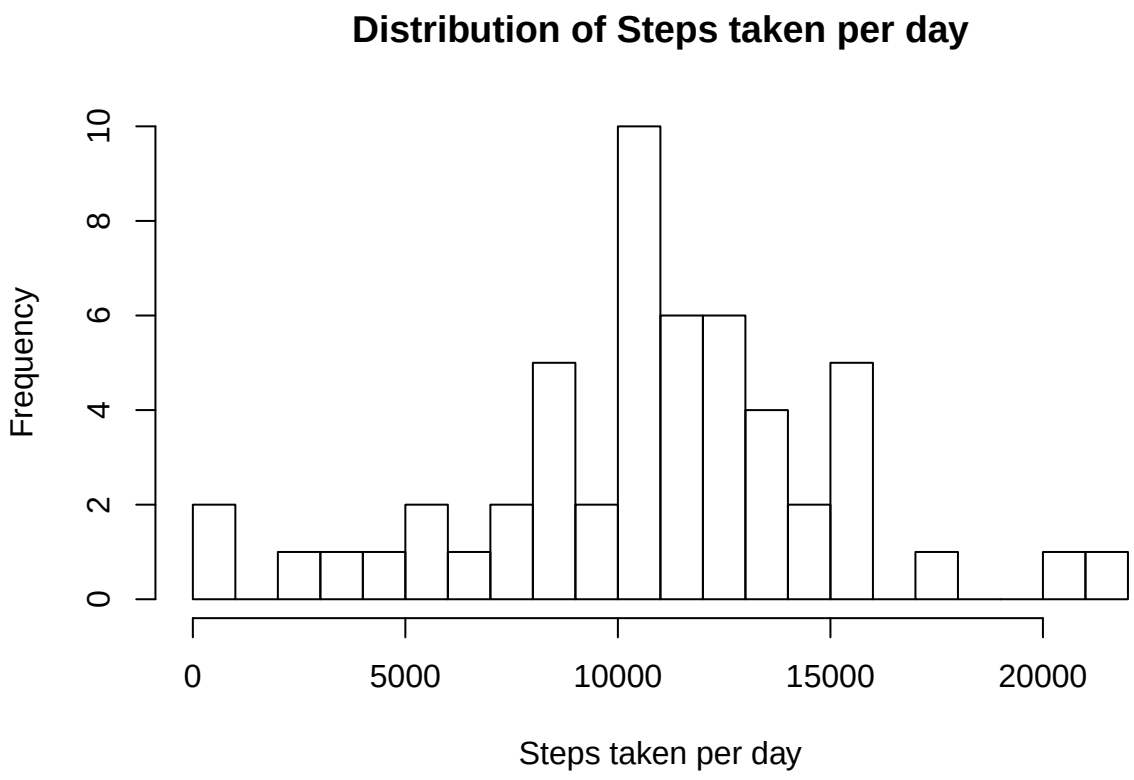
```
library(dplyr)
library(lattice)
```

```
## Warning: package 'lattice' was built under R version 3.1.3
```

```
sdata <- read.csv('activity.csv')  ##original source data
sdata$date <- as.Date(sdata$date, format = "%Y-%m-%d")
pdata <-as.data.frame(sdata[is.na(sdata[,1])==FALSE,])  ## processed data, filter out NA rows
```

What is mean total number of steps taken per day?

```
total_per_day <-pdata %>% group_by(date) %>% summarise(total = sum(steps))
hist(total_per_day$total, breaks = max(total_per_day$total)/1200, main = "Distribution of Steps taken p
```



```
mean1 <- as.integer(mean(total_per_day$total))
median1 <- median(total_per_day$total)
```

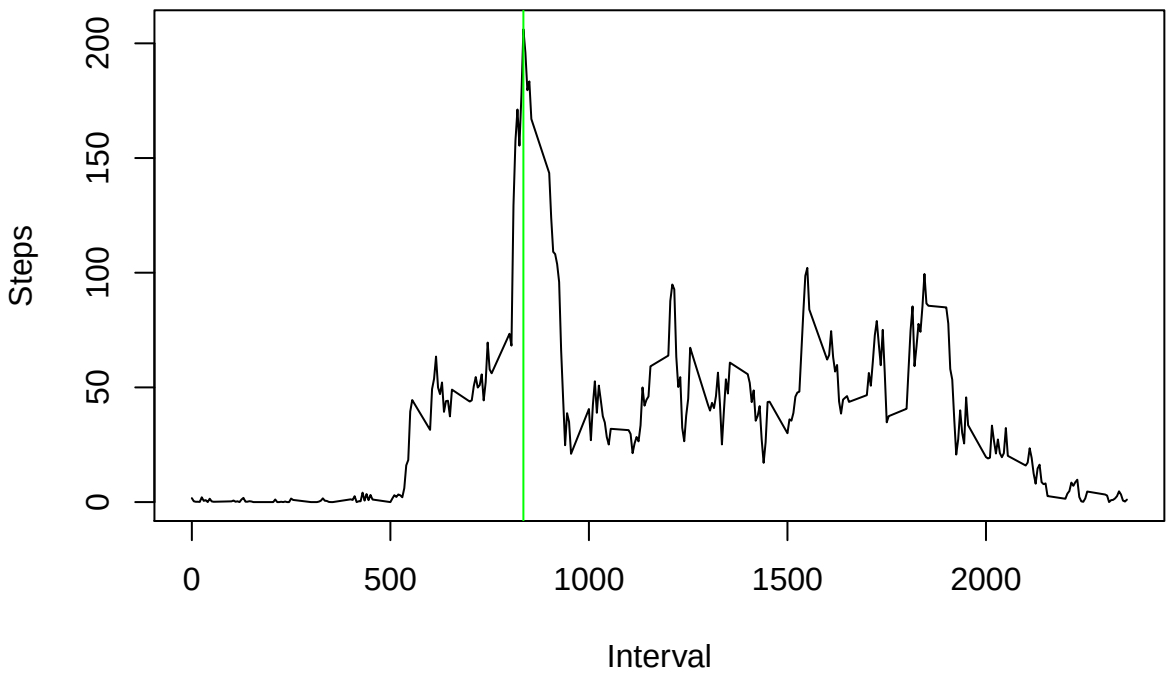
mean of the total number of steps taken per day: 10766;

median of the total number of steps taken per day: 10765

What is the average daily activity pattern?

```
total_per_int <-pdata %>% group_by(interval) %>% summarise(int_total = mean(steps))
max_steps <- max(total_per_int$int_total)
max_interval <-  total_per_int %>% filter(int_total==max_steps)

plot(total_per_int$interval, total_per_int$int_total, type="l", xlab="Interval", ylab= "Steps")
abline(v = max_interval$interval, col="green")
```



The 5-minute interval=835 contains the maximum number of steps 206.1698113

Imputing missing values

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

```
count(sdata) - count(pdata)
```

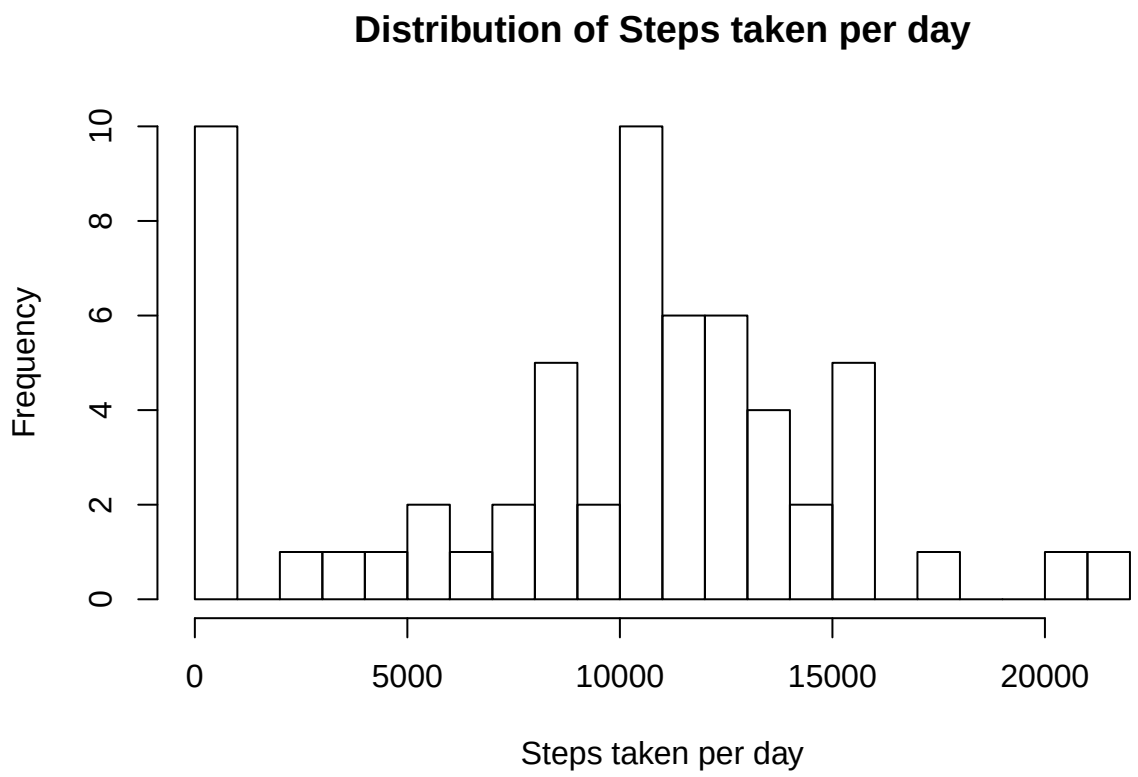
```
##      n
## 1 2304
```

Fill in all of the missing values in the dataset by 0 as a strategy of simplifying imputing missing values (may not the best. We will analysis the difference next)

```
sdata$steps[is.na(sdata$steps)]<-0
```

Then calculate and report the mean and median total number of steps taken per day

```
total_per_day2 <-sdata %>% group_by(date) %>% summarise(total = sum(steps))
hist(total_per_day2$total, breaks = max(total_per_day2$total)/1200, main = "Distribution of Steps taken
```




```
mean2 <- as.integer(mean(total_per_day2$total))
median2 <- as.integer(median(total_per_day2$total))
```

mean of the total number of steps taken per day: 9354;

median of the total number of steps taken per day: 10395

We can see above imputing missing data stragety makes mean and median of the total number of steps taken per day lower. Mean value drops more than median value does. This is another evidance that median value is impacted less by missing values than average value (mean) does.

Are there differences in activity patterns between weekdays and weekends?

First, add two new columns, weekday and weekendFlag

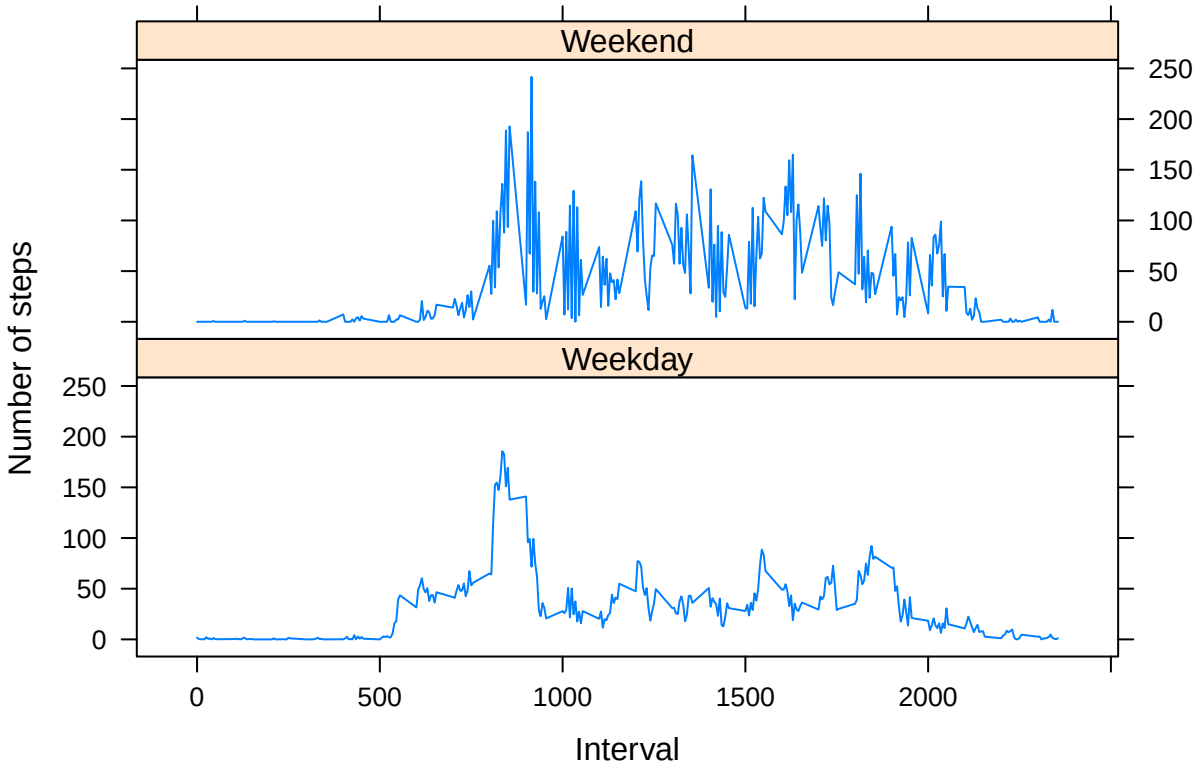
```
data_wks <- cbind(sdata,weekday=weekdays(sdata$date), weekendFlag=weekdays(sdata$date)==c("Sunday","Satu
data_wks$weekendFlag[data_wks$weekendFlag==FALSE] <- "Weekday"
data_wks$weekendFlag[data_wks$weekendFlag==TRUE] <- "Weekend"
head(data_wks)
```

```
##   steps       date interval weekday weekendFlag
## 1     0 2012-10-01        0  Monday     Weekday
## 2     0 2012-10-01        5  Monday     Weekday
## 3     0 2012-10-01       10  Monday     Weekday
## 4     0 2012-10-01       15  Monday     Weekday
## 5     0 2012-10-01       20  Monday     Weekday
## 6     0 2012-10-01       25  Monday     Weekday
```

Then calculate 5-minute average number of steps taken, averaged across all weekday days or weekend days

```
avr_per_int <-data_wks %>% group_by(interval,weekendFlag) %>% summarise( muSteps = mean(steps))

xyplot(muSteps ~ interval| weekendFlag,
           data = avr_per_int,
           type = "l",
           xlab = "Interval",
           ylab = "Number of steps",
           layout=c(1,2))
```



From above diagram, we can see some differences in activity patterns between weekdays and weekends. For example:

1. Wake-up and go to bed earlier in weekdays than in weekends
2. More activities after 10am in weekends than in weekdays