Reproducible Research: Peer Assessment 1

## Loading and preprocessing the data

**1. Load the data**

activity<-read.csv("activity.csv")

**2. Process/transform the data (if necessary) into a format suitable for your analysis**

We need to transform the "date" from factor type to date type

activity$date<-as.Date(activity$date)  
head(activity)

## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

tail(activity)

## steps date interval  
## 17563 NA 2012-11-30 2330  
## 17564 NA 2012-11-30 2335  
## 17565 NA 2012-11-30 2340  
## 17566 NA 2012-11-30 2345  
## 17567 NA 2012-11-30 2350  
## 17568 NA 2012-11-30 2355

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

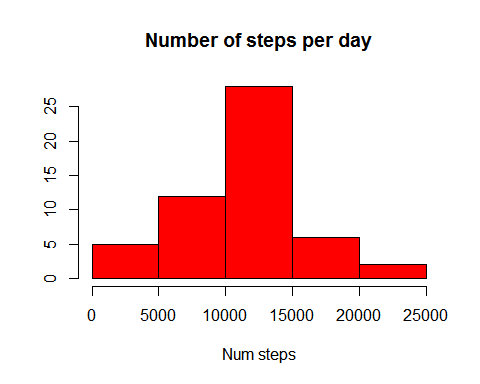
**1. Calculate the total number of steps taken per day**

sum\_steps\_days <- aggregate(steps ~ date, activity, FUN="sum")  
head(sum\_steps\_days)

## date steps  
## 1 2012-10-02 126  
## 2 2012-10-03 11352  
## 3 2012-10-04 12116  
## 4 2012-10-05 13294  
## 5 2012-10-06 15420  
## 6 2012-10-07 11015

**2. Make a histogram of the total number of steps taken each day**

hist(sum\_steps\_days$steps,main="Number of steps per day", xlab="Num steps",ylab="",col="red")



**3. Calculate and report the mean and median total number of steps taken per day**

mean\_total\_steps<-mean(sum\_steps\_days$steps,na.rm=TRUE)  
median\_total\_steps<-median(sum\_steps\_days$steps,na.rm=TRUE)

The mean total number of steps per taken day is:

## [1] 10766.19

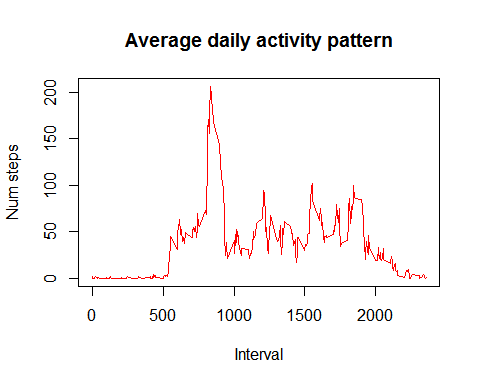
The median total number of steps per taken day is:

## [1] 10765

## What is the average daily activity pattern?

**1. Make 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 days (y-axis)**

mean\_steps\_interval <- aggregate(steps ~ interval, activity, FUN="mean")  
plot(mean\_steps\_interval$interval,mean\_steps\_interval$steps,main="Average daily activity pattern", ylab="Num steps",xlab="Interval",type="n")  
lines(mean\_steps\_interval$interval,mean\_steps\_interval$steps,col="red")



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

interval\_max <- mean\_steps\_interval$interval[which.max(mean\_steps\_interval$steps)]  
interval\_max

## [1] 835

## Imputing missing values

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

sum(is.na(activity$steps))

## [1] 2304

**2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.**  
I will fill the data with the mean of the 5-minutes interval calculated previusly

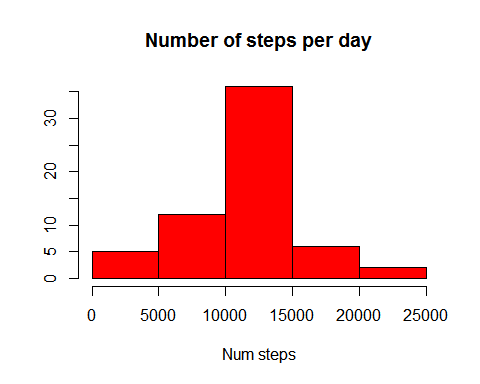
**3. Create a new dataset that is equal to the original dataset but with the missing data filled in.**  
First I create a function "impute" in case the number of steps is NA then it is replaced with the mean of the 5 minutes interval.

impute <- function(steps, interval) {  
 if (is.na(steps)) {  
 mean\_steps\_interval[mean\_steps\_interval$interval==interval,]$steps  
 } else {  
 steps  
 }  
}  
activity\_imputed<-activity  
activity\_imputed$steps <- mapply(impute, activity$steps,activity$interval)  
head(activity\_imputed)

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

**4. 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?**

sum\_steps\_days\_imputed <- aggregate(steps ~ date, activity\_imputed, FUN="sum")  
hist(sum\_steps\_days\_imputed$steps,main="Number of steps per day", xlab="Num steps",ylab="",col="red")



mean(sum\_steps\_days$steps,na.rm=TRUE)

## [1] 10766.19

mean(sum\_steps\_days\_imputed$steps,na.rm=TRUE)

## [1] 10766.19

median(sum\_steps\_days$steps,na.rm=TRUE)

## [1] 10765

median(sum\_steps\_days\_imputed$steps,na.rm=TRUE)

## [1] 10766.19

The mean is the same but the median has increased.

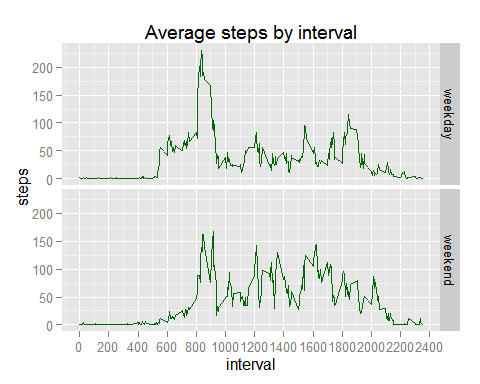
## Are there differences in activity patterns between weekdays and weekends?

**1.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.**

daytype <- function(day) {  
 wd<-weekdays(day)  
 if ((wd=="domingo")|(wd=="sábado")) {"weekend"}  
 else {"weekday"}  
}  
activity\_imputed$daytype <-mapply(daytype, activity\_imputed$date)

**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). The plot should look something like the following, which was created using simulated data:**

mean\_steps\_interval\_imputed <- aggregate( steps ~ interval+daytype, activity\_imputed, FUN="mean")  
  
library(ggplot2)  
g<-ggplot(mean\_steps\_interval\_imputed,aes(x=interval,y=steps))  
# Adding a title to the graph  
g<-g+labs(title="Average steps by interval")  
# Plotting the data ussing bar type   
g<-g+geom\_freqpoly(colour="darkgreen",fill="light gray",stat="identity")  
  
# Laying out panels in a grid by type  
g<-g+facet\_grid(daytype~.)  
g<-g+scale\_x\_continuous(breaks=seq(0,2400, by=200))  
print(g)



I want to see the comparation of the acumulated steps during weekday and weekend

wd<- subset(mean\_steps\_interval\_imputed,daytype=="weekend")  
we<- subset(mean\_steps\_interval\_imputed,daytype=="weekday")  
wd$steps.acu=cumsum(wd$steps)  
we$steps.acu=cumsum(we$steps)  
w<-rbind(wd,we)

And we plot the results

g<-ggplot(w,aes(x=interval,y=steps.acu, group=daytype))  
g<-g + geom\_line()  
g<-g+labs(title="Acumulated average steps by interval")+ylab("Steps")  
g<-g + geom\_line(aes(colour = daytype))  
g<-g+scale\_x\_continuous(breaks=seq(0,2400, by=200))  
print(g)

