**Reproducible Research Peer Assessment1**

**Loading/preprocessing data**

url = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"

download.file(url, "Activity monitoring data.zip")

unzip("Activity monitoring data.zip")

mydata = read.csv("activity.csv")

names(mydata)

## [1] "steps" "date" "interval"

**Transform to suitable format**

mydata$date <- as.Date(mydata$date, "%Y-%m-%d")

**Mean and median of total # of steps?**

**plot histogram with mean and median lines added**

nsteps = aggregate(steps ~ date, mydata, FUN=sum, na.rm=T)

mean(nsteps$steps)

## [1] 10766.19

median(nsteps$steps)

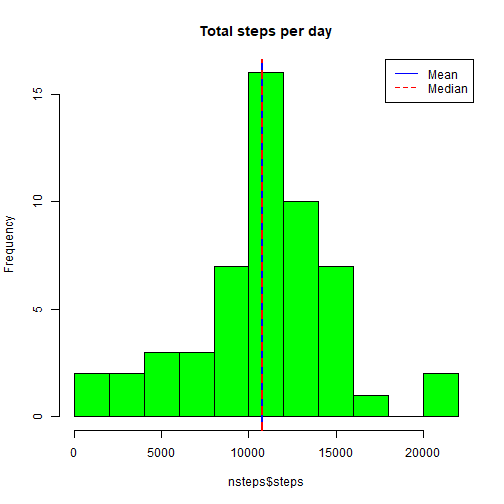
## [1] 10765

hist(nsteps$steps, breaks = 10, col="green", main="Total steps per day")

abline(v=mean(nsteps$steps), lty=1,lwd=2, col="blue")

abline(v=median(nsteps$steps), lty=2,lwd=2, col="red")

legend("topright",lty=c(1,2),col=c("blue","red"),legend=c("Mean","Median"))

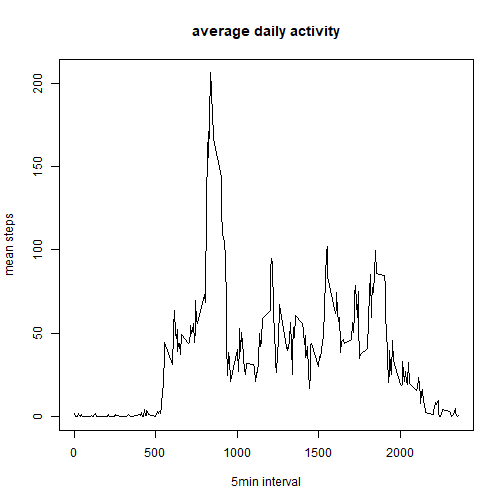
[](https://github.com/liuliuc/RepData_PeerAssessment1/blob/master/figure/unnamed-chunk-3-1.png) Mean and median of the total number of steps taken per day 10766.19 and 10765.

**Average daily activity pattern**

**line plot with interval and steps**

meanInt = aggregate(steps ~ interval, mydata, FUN = mean, na.rm=T)

plot(meanInt$interval,meanInt$steps,type="l",xlab="5min interval",ylab="mean steps",main="average daily activity")

[](https://github.com/liuliuc/RepData_PeerAssessment1/blob/master/figure/unnamed-chunk-4-1.png)

meanInt[which.max(meanInt$steps),]

## interval steps

## 104 835 206.1698

On average across all the days in the dataset, the 835 interval contains the maximum number of steps.

**Imputing missing values**

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

nrow(mydata[is.na(mydata$steps),])

## [1] 2304

**Devise a strategy for filling in all of the missing values in the dataset.**

NAdata = mydata[is.na(mydata$steps),]

NAdata$steps[NAdata$interval==meanInt$interval]= meanInt$steps

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

gooddata = mydata[!is.na(mydata$steps),]

newdata = rbind(NAdata,gooddata)

**Total number of steps taken and mean/median calculation**

nsteps = aggregate(steps ~ date, newdata, FUN=sum, na.rm=T)

mean(nsteps$steps)

## [1] 10766.19

median(nsteps$steps)

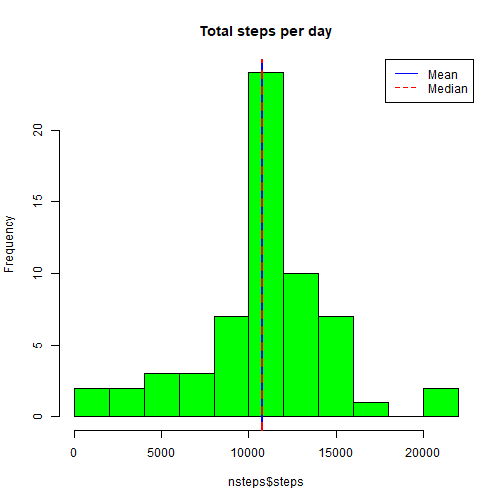
## [1] 10766.19

hist(nsteps$steps, breaks = 10, col="green", main="Total steps per day")

abline(v=mean(nsteps$steps), lty=1,lwd=2, col="blue")

abline(v=median(nsteps$steps), lty=2,lwd=2, col="red")

legend("topright",lty=c(1,2),col=c("blue","red"),legend=c("Mean","Median"))

[](https://github.com/liuliuc/RepData_PeerAssessment1/blob/master/figure/unnamed-chunk-8-1.png) The mean and median total number of steps taken per day changed to be the same after imputing missing data on the estimates of the total daily number of steps.

**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.**

newdata$dayf <- ifelse(weekdays(newdata$date) %in% c("Saturday","Sunday"),"weekend","weekday")

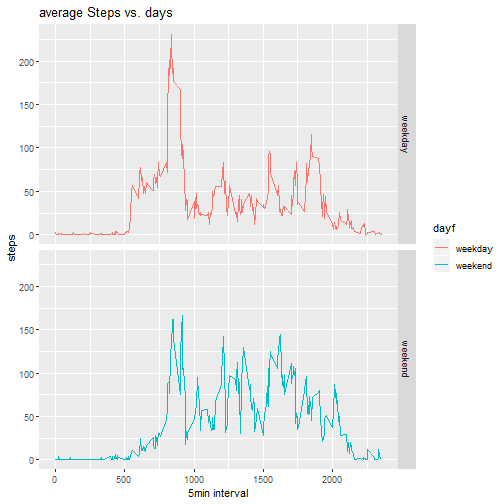
**Make a panel plot containing a time series plot (i.e. type="l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).**

library(ggplot2)

meanInt = aggregate(steps ~ interval+dayf, newdata, FUN = mean)

g <- ggplot(meanInt, aes(interval, steps, color = dayf))

g + geom\_line() + facet\_grid(dayf~.) + labs(title="average Steps vs. days", y="steps", x= "5min interval")

[](https://github.com/liuliuc/RepData_PeerAssessment1/blob/master/figure/unnamed-chunk-10-1.png)