Reproducible Research: Peer Assessment 1

## 1. Loading and preprocessing the data

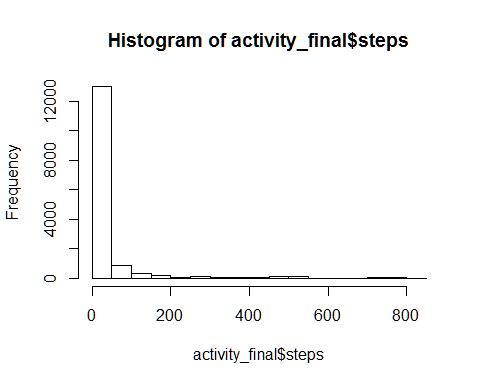
Load the activity.csv dataset from my local path.

activity <- read.csv("D:\\Coursera\\Material\\05. Reproducible Research\\CourseProject\\activity.csv")

## 2. What is Mean/Median total number of steps taken per day?

First, Remove NA data. And draw the histogram to see the distribusion of steps taken per day. Then, claculate the mean and median. We got mean = 10766.19 and median = 10766.

activity\_final <- activity[!is.na(activity$steps),]  
steps\_by\_date <- aggregate(steps ~ date, data=activity\_final, sum)  
hist(activity\_final$steps)



mean(steps\_by\_date$steps)

## [1] 10766.19

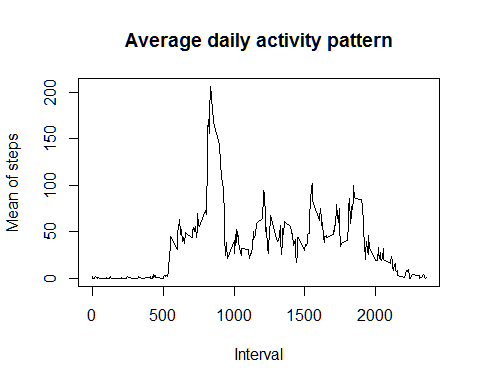
median(steps\_by\_date$steps)

## [1] 10765

## 3. What is the average daily activity pattern?

The average daily activiy pattern is as below plot. You can see the peak at interval 835.

steps\_by\_interval <- tapply(activity\_final$steps,activity\_final$interval,mean)  
plot(row.names(steps\_by\_interval), steps\_by\_interval, type="l",  
 main="Average daily activity pattern", xlab="Interval", ylab="Mean of steps")



names(which.max(steps\_by\_interval))

## [1] "835"

## 4. Imputing missing values

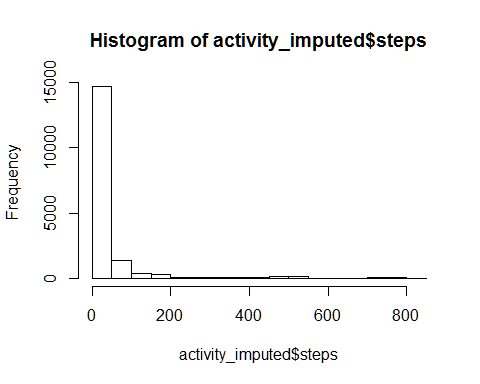
Try to replace the missing values by the mean of interval.

activity\_imputed <- activity  
for (i in 1:nrow(activity)) {  
 if( is.na(activity[i,]$steps) )  
 {  
 j <- which(rownames(steps\_by\_interval) == activity[i,]$interval);  
 activity\_imputed$steps[i] <- as.integer(steps\_by\_interval[j]);  
 }   
}

## 5. Checking if the imputing missing values make sense

Calculate the mean/median again, and draw the histogram & time series plot. We can see that the missing value imputing is pretty make sense because the statistics summary & graph looks almost the same.

steps\_by\_date <- aggregate(steps ~ date, data=activity\_imputed, sum)  
hist(activity\_imputed$steps)



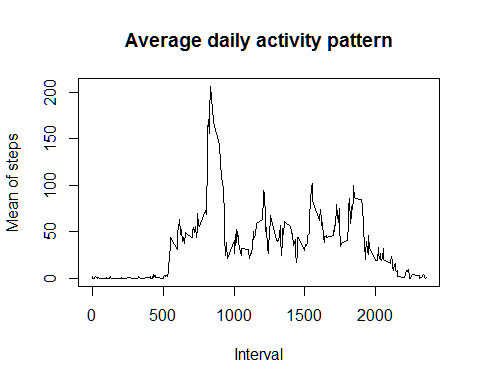
mean(steps\_by\_date$steps)

## [1] 10749.77

median(steps\_by\_date$steps)

## [1] 10641

steps\_by\_interval <- tapply(activity\_imputed$steps,activity\_imputed$interval,mean)  
plot(row.names(steps\_by\_interval), steps\_by\_interval, type="l",  
 main="Average daily activity pattern", xlab="Interval", ylab="Mean of steps")



names(which.max(steps\_by\_interval))

## [1] "835"

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

Yes, it is totally different. From the plot, we can see the weekdays & weekends is totally different steps distribution.