Reproducible Research: Peer Assessment 1

December 26, 2016

## Loading and preprocessing the data

library(ggplot2)  
library(lattice)  
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.4.4

setwd("~/R/coursera")  
mydf <- read.csv("activity.csv", stringsAsFactors=FALSE)  
  
dim(mydf)

## [1] 17568 3

dim(mydf[complete.cases(mydf),])

## [1] 15264 3

names(mydf)

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

sum(is.na(mydf$steps)) ## There are NAs.

## [1] 2304

sum(is.na(mydf$date)) ## There isn't NA.

## [1] 0

sum(is.na(mydf$interval)) ## There isn't NA.

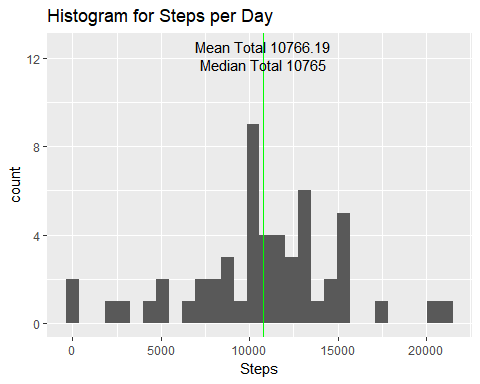
## [1] 0

mydf2 <- mydf[complete.cases(mydf),] ## Remove rows contain NA.

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

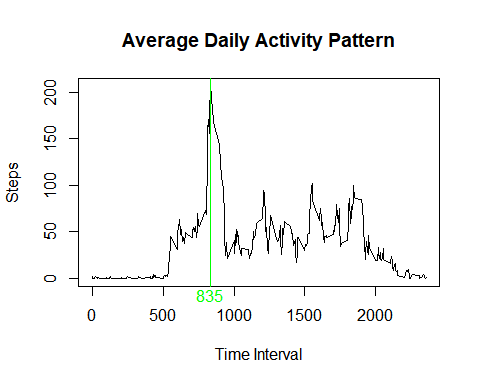
myfun <- function(mydf) {  
 mydf <- aggregate(steps ~ date, mydf, FUN = sum)  
 steps\_mean <- mean(mydf$steps)  
 steps\_median <- median(mydf$steps)  
 q <- qplot(mydf$steps,  
 main = "Histogram for Steps per Day",  
 xlab = "Steps") +  
 geom\_vline(xintercept = steps\_mean, color = "red") +  
 geom\_vline(xintercept = steps\_median, color = "green") +  
 annotate("text", x = steps\_mean, y = 12.5,  
 label = paste("Mean Total", format(steps\_mean))) +  
 annotate("text", x = steps\_median, y = 11.7,  
 label = paste("Median Total", format(steps\_median)))  
}  
q1 <- myfun(mydf2); q1

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



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

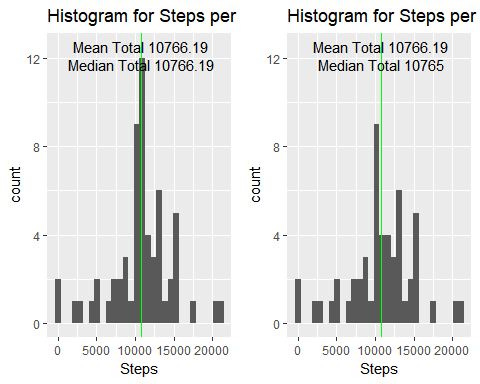
mydf4 <- aggregate(steps ~ interval, mydf2, FUN = mean)  
max\_steps <- max(mydf4$steps)  
max\_interval <- mydf4[mydf4$steps == max\_steps, ]$interval  
plot(mydf4$interval, mydf4$steps, type = "l",  
 main = "Average Daily Activity Pattern",  
 xlab = "Time Interval",  
 ylab = "Steps")  
abline(v=max\_interval, col = "green")  
mtext(max\_interval, side = 1,   
 line = 0, at = max\_interval,  
 col = "green")



## 3. Imputing missing values

row\_na <- nrow(mydf) - nrow(mydf2) ## 2304 rows with NA  
mydf5 <- mydf  
for (i in 1:nrow(mydf5)) {  
 if (is.na(mydf5[i, ]$steps)) {  
 mydf5[i, ]$steps <-   
 mydf4[mydf4$interval == mydf5[i, ]$interval, ]$steps  
 }  
}  
q2 <- myfun(mydf5)  
grid.arrange(q2, q1, ncol=2)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



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

mydf5$day <- NA  
for (i in 1:nrow(mydf5)) {  
 if (weekdays(as.Date(mydf5[i, ]$date)) == "Saturday" |   
 weekdays(as.Date(mydf5[i, ]$date)) == "Sunday") {  
 mydf5[i, ]$day <- "weekend"  
 } else {  
 mydf5[i, ]$day <- "weekday"  
 }  
}  
mydf5$day <- factor(mydf5$day)  
mydf6 <- aggregate(steps ~ interval + day, mydf5, FUN = mean)  
xyplot(steps ~ interval | factor(day),   
 data = mydf6,  
 type = "l",  
 layout = c(1,2),  
 xlab = "Interval",  
 ylab = "Number of Steps")

