An analysis on personal activity monitoring device data

Mohammad Arifur Rahman Bhuiyan

October 25, 2018

## Loading and preprocessing the data

Load the data

setwd("C:/Coursera Data Science/Reproducible Research/Week 2")  
activity <- read.table("activity.csv", sep=",", header=T)

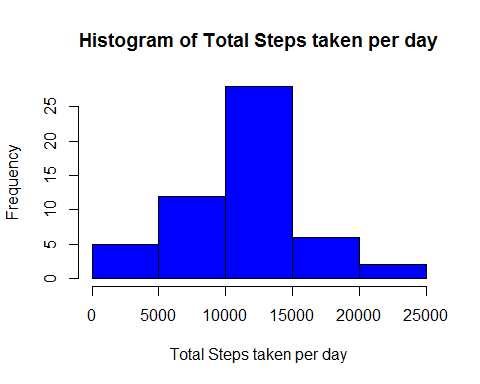
Process/transform the data (if necessary) into a format suitable for your analysis

totalSteps <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)

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

Make a histogram of the total number of steps taken each day

hist(totalSteps$steps,col="blue",main="Histogram of Total Steps taken per day",xlab="Total Steps taken per day",cex.axis=1,cex.lab = 1)



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

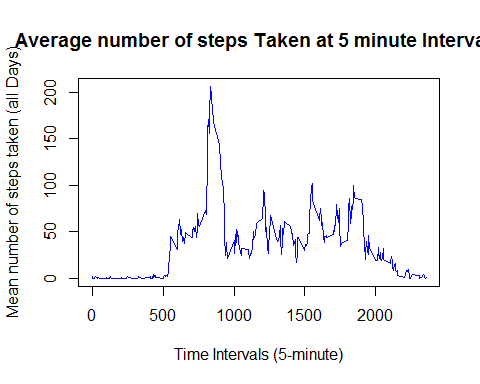
mean\_steps <- mean(totalSteps$steps)  
median\_steps <- median(totalSteps$steps)

-The mean total number of steps taken per day is 1.0766 × 104 steps -The median total number of steps taken per day is 10765 steps

# What is the average daily activity pattern?

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)

steps\_interval <- aggregate(steps ~ interval, data = activity, mean, na.rm = TRUE)  
plot(steps ~ interval, data = steps\_interval, type = "l", xlab = "Time Intervals (5-minute)", ylab = "Mean number of steps taken (all Days)", main = "Average number of steps Taken at 5 minute Intervals", col = "blue")



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

maxStepInterval <- steps\_interval[which.max(steps\_interval$steps),"interval"]

-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)

missing\_rows <- sum(!complete.cases(activity))

-The total number of missing rows is 2304

# 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.

## This function returns the mean steps for a given interval  
getMeanStepsPerInterval <- function(interval){  
 steps\_interval[steps\_interval$interval==interval,"steps"]  
 }

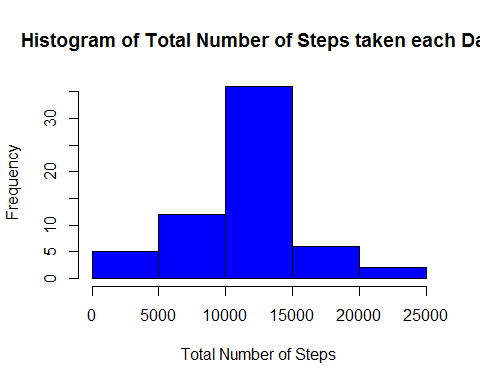
# Create a new dataset that is equal to the original dataset but with the missing data filled in.

complete.activity <- activity  
  
## Filling the missing values with the mean for that 5-minute interval  
flag = 0  
for (i in 1:nrow(complete.activity)) {  
 if (is.na(complete.activity[i,"steps"])) {  
 complete.activity[i,"steps"] <- getMeanStepsPerInterval(complete.activity[i,"interval"])  
 flag = flag + 1  
 }  
 }

-Total of 2304 missing values were filled.

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

total.steps.per.days <- aggregate(steps ~ date, data = complete.activity, sum)  
hist(total.steps.per.days$steps, col = "blue", xlab = "Total Number of Steps",   
 ylab = "Frequency", main = "Histogram of Total Number of Steps taken each Day")



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

showMean <- mean(total.steps.per.days$steps)  
showMedian <- median(total.steps.per.days$steps)

-Mean total number of steps taken per day is 1.0766 × 104 -Median total number of steps taken per day is 1.0766 × 104 #Do these values differ from the estimates from the first part of the assignment?

-The mean value is the same as the value before imputing missing data, but the median value has changed. #What is the impact of imputing missing data on the estimates of the total daily number of steps?

-The mean value is the same as the value before imputing missing data since the mean value has been used for that particular 5-min interval. The median value is different, since the median index is now being changed after imputing missing values.

# Are there 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.

complete.activity$day <- ifelse(as.POSIXlt(as.Date(complete.activity$date))$wday%%6 ==   
 0, "weekend", "weekday")  
complete.activity$day <- factor(complete.activity$day, levels = c("weekday", "weekend"))

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).

steps.interval= aggregate(steps ~ interval + day, complete.activity, mean)  
library(lattice)  
xyplot(steps ~ interval | factor(day), data = steps.interval, aspect = 1/2,   
 type = "l")

