Reproducible Research week2 Project 1

Xiaoqing XU (Sabrina)

July 9, 2017

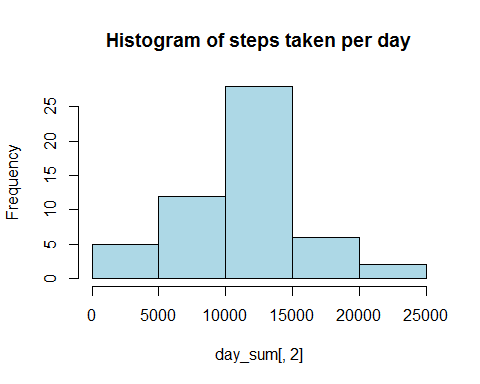
This project is to analyze data collected from FitBit.

## Loading and preprocessing the data

setwd <- ("C:\\0Rdata\\Coursera")  
d <- read.csv("activity.csv")

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

day\_sum <- aggregate(d[,1], list(d$date),FUN=sum)  
#histogram of steps taken per day  
hist(x = day\_sum[,2],main = "Histogram of steps taken per day", col = "lightblue")



#calculate and report mean and median  
mean\_day <- mean(day\_sum$x,na.rm = TRUE)  
mean\_day

## [1] 10766.19

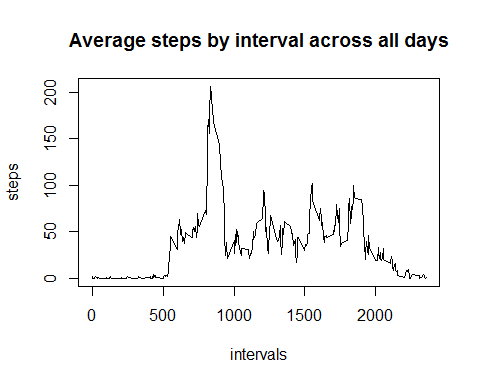
median\_day <- median(day\_sum$x, na.rm = TRUE)  
median\_day

## [1] 10765

So the mean value of steps every day is 10766 and the median value of steps per day is 10765.

## What is the average daily activity pattern?

#plot time-series data by interval  
mean\_interval <- aggregate(d[,1], list(d$interval), mean, na.rm = TRUE)  
plot(x = mean\_interval[,1], y = mean\_interval[,2],type = "l", xlab = "intervals", ylab="steps", main = "Average steps by interval across all days")



#plot(y = d$steps, x=d$interval, type = "l", xlab = "5 min interval", ylab = "Steps", xaxt = "n"); axis(1, at = seq(0, 2500, by = 10))

## Imputing missing values

Missing data needs to be imputed. A simple imputation was applied here by replacing NAs with average steps in its relavent interval across all days. For eamples, for interval 5 of 2012-10-02, the average step for "interval 5" of all days was used to replace the NA.

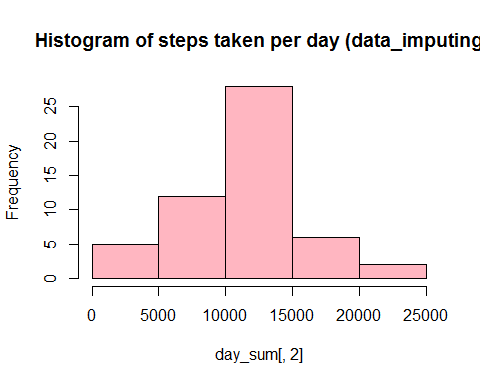
#count number of rows with NAs  
sum(is.na(d$steps))

## [1] 2304

d\_im <- d  
d\_im <- transform(d, steps = ifelse(is.na(d$steps), mean\_interval$x[match(d$interval, mean\_interval$Group.1)],d$steps))

Then the new histogram on imputed data is created and mean and median of imputed data is calculated.

hist(x = day\_sum[,2],main = "Histogram of steps taken per day (data\_imputing", col = "lightpink")



day\_sum\_im <- aggregate(d\_im[,1], list(d\_im$date),FUN=sum)  
mean\_day\_im <- mean(day\_sum\_im$x,na.rm = TRUE)  
mean\_day\_im

## [1] 10766.19

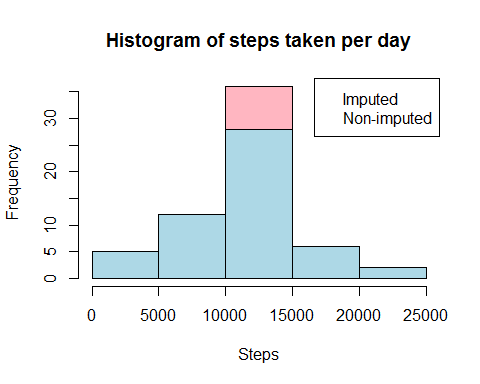
median\_day\_im <- median(day\_sum\_im$x, na.rm = TRUE)  
median\_day\_im

## [1] 10766.19

mean\_interval\_im <- aggregate(d\_im[,1], list(d\_im$interval), mean, na.rm = TRUE)

So there're difference in mean and median values between raw data and imputed data.A histogram is shown below.

hist(x = day\_sum\_im[,2],main = "Histogram of steps taken per day", col = "lightpink", xlab = "Steps")  
hist(day\_sum$x, main = "Total Steps every Day", col="lightblue", xlab="Number of Steps", add=T)  
legend("topright", c("Imputed", "Non-imputed"), col=c("lightpink", "lightblue"))



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

The difference is displayed in the line plot below.

weekdays <- c("Monday", "Tuesday", "Wedesday", "Thursday",   
 "Friday")  
d\_im$dow = as.factor(ifelse(is.element(weekdays(as.Date(d\_im$date)),weekdays), "Weekday", "Weekend"))  
  
steps\_by\_interval\_i <- aggregate(steps ~ interval + dow, d\_im, mean)  
  
  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

ggplot(data=steps\_by\_interval\_i,aes(x=interval, y=steps, group = dow))+  
 geom\_line(position = position\_dodge(), aes(colour=dow))

## Warning: Width not defined. Set with `position\_dodge(width = ?)`

