## Course Project 1

Loading and preprocessing the data

data <- read.csv("E:/S/Coursera/DS-JHU/5Reproducible Research/assignment 2/activity.csv", header = T, sep=",")  
summary(data)

## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 12.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## NA's :2304 (Other) :15840

str(data)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : int NA NA NA NA NA NA NA NA NA NA ...  
## $ date : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

dim(data)

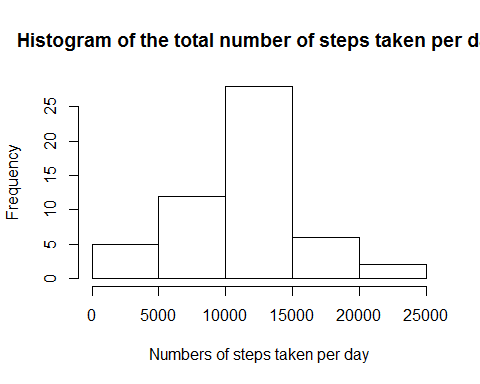
## [1] 17568 3

head(data)

## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

What is mean total number of steps taken per day?

data1 <- tapply(data$steps,data$date,sum)  
hist(data1,xlab="Numbers of steps taken per day",main="Histogram of the total number of steps taken per day")



What is the average daily activity pattern?

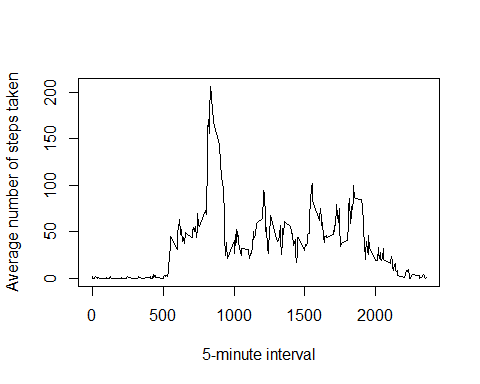
library("dplyr")

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data2 <- filter(data,!is.na(data$steps))  
data22 <- aggregate(data=data2,steps ~ interval,mean)  
plot(data22$interval,data22$steps,type="l",xlab="5-minute interval", ylab="Average number of steps taken")



Imputing missing values 1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

sum(!is.na(data))

## [1] 50400

1. 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.

data3 <- data  
na\_datarows <- is.na(data3$steps)  
data33 <- tapply(data3$steps,data3$interval,mean,na.rm=T)

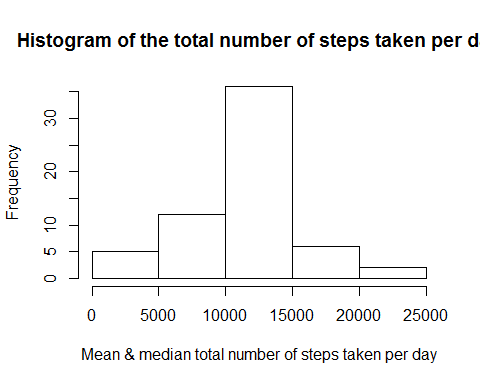
1. Create a new dataset that is equal to the original dataset but with the missing data filled in.

data3$steps[na\_datarows] <- data33[as.character(data3$interval[na\_datarows])]  
head(data3)

## steps date interval  
## 1 1.7169811 2012-10-01 0  
## 2 0.3396226 2012-10-01 5  
## 3 0.1320755 2012-10-01 10  
## 4 0.1509434 2012-10-01 15  
## 5 0.0754717 2012-10-01 20  
## 6 2.0943396 2012-10-01 25

1. Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

data333 <- tapply(data3$steps,data3$date,sum)  
hist(data333,xlab="Mean & median total number of steps taken per day",main="Histogram of the total number of steps taken per day")



Are there differences in activity patterns between weekdays and weekends? For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part. 1. 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.

data4 <- data3  
data4$week\_type<-data4$date  
head(data4)

## steps date interval week\_type  
## 1 1.7169811 2012-10-01 0 2012-10-01  
## 2 0.3396226 2012-10-01 5 2012-10-01  
## 3 0.1320755 2012-10-01 10 2012-10-01  
## 4 0.1509434 2012-10-01 15 2012-10-01  
## 5 0.0754717 2012-10-01 20 2012-10-01  
## 6 2.0943396 2012-10-01 25 2012-10-01

data4$date <- as.Date(data4$date)  
data4$week\_type <- as.Date(data4$week\_type)  
weekend <- function(actdate) {ifelse(weekdays(actdate) == "星期六" | weekdays(actdate) == "星期日","weekend","weekday")}  
data4 <- mutate(data4, week\_type = weekend(data4$date))  
head(data4)

## steps date interval week\_type  
## 1 1.7169811 2012-10-01 0 weekday  
## 2 0.3396226 2012-10-01 5 weekday  
## 3 0.1320755 2012-10-01 10 weekday  
## 4 0.1509434 2012-10-01 15 weekday  
## 5 0.0754717 2012-10-01 20 weekday  
## 6 2.0943396 2012-10-01 25 weekday

1. 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). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

data44 <- aggregate(data4$steps,list(data4$interval, data4$week\_type),sum)  
head(data44)

## Group.1 Group.2 x  
## 1 0 weekday 101.301887  
## 2 5 weekday 20.037736  
## 3 10 weekday 7.792453  
## 4 15 weekday 8.905660  
## 5 20 weekday 4.452830  
## 6 25 weekday 71.566038

data44weekend <- subset(data44,data44$Group.2=="weekend")  
data44weekday <- subset(data44,data44$Group.2=="weekday")  
  
dim(data44)

## [1] 576 3

dim(data44weekday)

## [1] 288 3

dim(data44weekend)

## [1] 288 3

library(lattice)  
xyplot(x~Group.1|Group.2,data=data44,layout=c(1,2),type="l",xlab="5-minute interval",ylab="Steps")

