PA1.html

## install.packages("knitr")  
library(knitr)  
## ("dplyr")  
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

## install.packages("ggplot2")  
library(ggplot2)  
  
  
### read file   
act <- read.csv("activity.csv", header = TRUE, stringsAsFactors = FALSE)  
str(act)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : int NA NA NA NA NA NA NA NA NA NA ...  
## $ date : chr "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

act$date<- as.Date(act$date,'%Y-%m-%d')  
act$steps<-as.numeric(act$steps)  
head(act)

## 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?  
## Calculate the total number of steps taken per day  
## If you do not understand the difference between a histogram and a barplot,  
## research the difference between them. Make a histogram of the total number of   
## steps taken each day  
## Calculate and report the mean and median of the total number of   
## steps taken per day  
  
nrow(act)

## [1] 17568

sum(is.na(act$steps))

## [1] 2304

sum(is.na(as.character(act$Date)) )

## [1] 0

sum(is.na(act$interval))

## [1] 0

act\_nan<- act[which(!is.na(act$steps)),] ### removing null rows  
str(act\_nan)

## 'data.frame': 15264 obs. of 3 variables:  
## $ steps : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ date : Date, format: "2012-10-02" "2012-10-02" ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

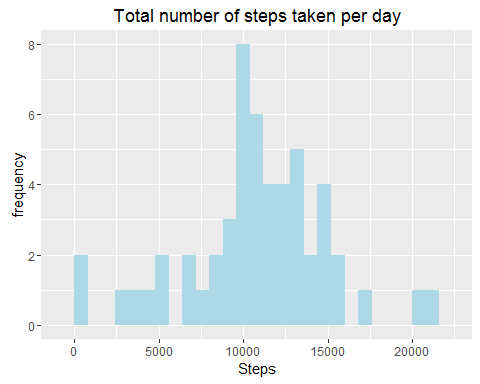
## act\_nan  
head(act\_nan)

## steps date interval  
## 289 0 2012-10-02 0  
## 290 0 2012-10-02 5  
## 291 0 2012-10-02 10  
## 292 0 2012-10-02 15  
## 293 0 2012-10-02 20  
## 294 0 2012-10-02 25

## perday\_stps<-tapply(act\_nan$steps,act\_nan$date)  
  
perday\_stps<-aggregate(steps ~ date, act\_nan , sum)  
  
head(perday\_stps)

## date steps  
## 1 2012-10-02 126  
## 2 2012-10-03 11352  
## 3 2012-10-04 12116  
## 4 2012-10-05 13294  
## 5 2012-10-06 15420  
## 6 2012-10-07 11015

ggplot(perday\_stps, aes(x= steps)) +  
 geom\_histogram(fill="lightblue",binwidth=800) +   
 ggtitle('Total number of steps taken per day') + xlab("Steps") +  
 ylab("frequency")



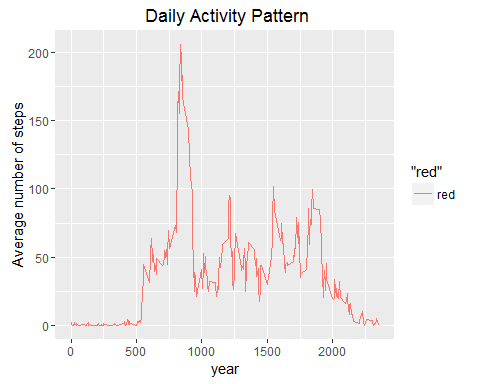
mean(perday\_stps$steps, na.rm = TRUE) ###10766.19

## [1] 10766.19

median(perday\_stps$steps, na.rm = TRUE)### 10765

## [1] 10765

### What is the average daily activity pattern?  
  
avg\_intr\_stpes<-aggregate(steps ~ interval, act\_nan , mean)  
  
   
  
 ggplot(avg\_intr\_stpes , aes(interval, steps, color = "red")) +  
 geom\_line() + xlab("year") + ylab("Average number of steps") +  
 ggtitle(" Daily Activity Pattern")



head(avg\_intr\_stpes)

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

max(avg\_intr\_stpes$steps)#### max steps 206.1698

## [1] 206.1698

### Imputing missing values  
## total NA   
sum(is.na(act$steps)) ##2304

## [1] 2304

sum(is.na(act)) ##2304

## [1] 2304

head(act)

## 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

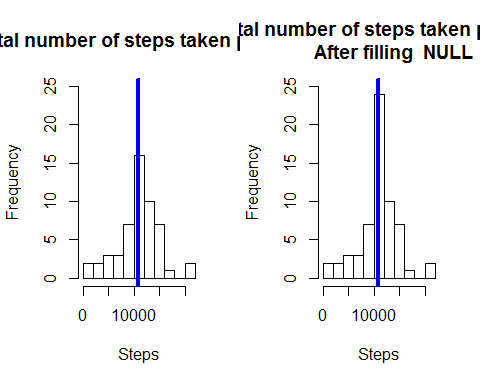
act\_all <- act  
act\_all[which(is.na(act\_all$steps)),1]<- mean(act\_all$steps, na.rm = TRUE)  
sum(is.na(act\_all)) ## no of null 0

## [1] 0

max(act\_all$interval) ##2355

## [1] 2355

## What is the impact of imputing missing data   
## on the estimates of the total daily number of steps?  
  
   
  
par(mfrow=c(1,2))  
  
  
hist(perday\_stps$steps,10,  
 main = "Total number of steps taken per day", xlab = "Steps",ylim =c(0, 25)  
 )   
abline(v=median(perday\_stps$steps),col = 4, lwd = 4)  
  
  
perday\_stps\_nona <-aggregate(steps ~ date, act\_all , sum)  
  
   
hist(perday\_stps\_nona$steps,10,  
 main = "Total number of steps taken per day  
 After filling NULL", xlab = "Steps",ylim =c(0, 25)   
)   
abline(v=median(perday\_stps\_nona$steps),col = 4, lwd = 4)



mean(perday\_stps$steps) ## 10766.19

## [1] 10766.19

median(perday\_stps$steps)## 10765

## [1] 10765

mean(perday\_stps\_nona$steps) ##10766.19

## [1] 10766.19

median(perday\_stps\_nona$steps) ## 10766.19

## [1] 10766.19

mean(perday\_stps$steps)-mean(perday\_stps\_nona$steps) ###0

## [1] 0

median(perday\_stps$steps)-median(perday\_stps\_nona$steps)## small difference

## [1] -1.188679

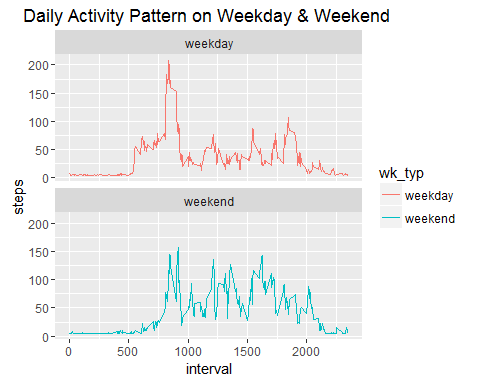
### Are there differences in activity patterns between weekdays and weekends?  
act\_all$date<- as.Date(act\_all$date,'%Y-%m-%d')  
act\_all<- mutate(act\_all, wk\_typ= ifelse(weekdays(act\_all$date)  
 == "Saturday" | weekdays(act\_all$date) =="Sunday", "weekend",  
 "weekday"))  
head(act\_all)

## steps date interval wk\_typ  
## 1 37.3826 2012-10-01 0 weekday  
## 2 37.3826 2012-10-01 5 weekday  
## 3 37.3826 2012-10-01 10 weekday  
## 4 37.3826 2012-10-01 15 weekday  
## 5 37.3826 2012-10-01 20 weekday  
## 6 37.3826 2012-10-01 25 weekday

act\_all$wk\_typ<- as.factor(act\_all$wk\_typ)  
  
head(act\_all)

## steps date interval wk\_typ  
## 1 37.3826 2012-10-01 0 weekday  
## 2 37.3826 2012-10-01 5 weekday  
## 3 37.3826 2012-10-01 10 weekday  
## 4 37.3826 2012-10-01 15 weekday  
## 5 37.3826 2012-10-01 20 weekday  
## 6 37.3826 2012-10-01 25 weekday

act\_weekty<- aggregate(steps ~ interval + wk\_typ , act\_all , mean)  
  
## png("weekly-comapre.png", width=640, height=640)  
ggplot(act\_weekty, aes(x=interval,y=steps ,color =wk\_typ)) +  
 geom\_line() + ggtitle( "Daily Activity Pattern on Weekday & Weekend ") +  
 facet\_wrap(~wk\_typ, ncol = 1, nrow=2)



## png("weekly-comapre\_2.png", width=640, height=480)  
ggplot(act\_weekty, aes(x=interval,y=steps ,color =wk\_typ)) +  
 geom\_line() + ggtitle( "Daily Activity Pattern on Weekday & Weekend ")

