# Exploratory data science\_ Week 2\_ project 1

## ———..//..—————-

### Loading lib. needed for running this project

library(ggplot2)  
library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

# P1: Loading and preprocessing data

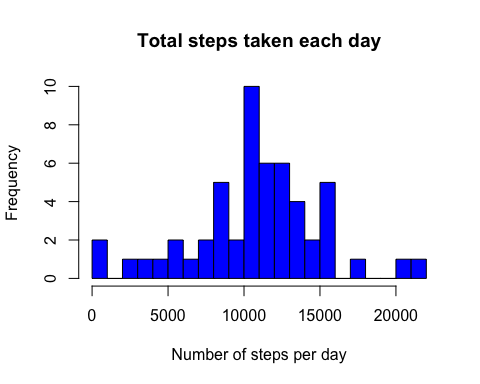
### Setting working directory and read data

### after download data in Repdata folder in Desktop

setwd("~/Desktop/Repdata")  
  
if (!file.exists("activity.csv") ) {  
 unzip("activity.zip")  
}  
data <- read.csv("~/Desktop/Repdata/activity.csv", header = TRUE)  
mydata <- na.omit(data)  
## data has 17568 obs. of 3 vars.(steps, date and intercal);   
## mydata has 15264 obs. instead ; the NA\_num = 2304\_ the diff.obs. bet data and mydata

# P2: What is the mean of total steps taken everyday? Histogram

stepsByday<- aggregate(mydata$steps, by=list(steps.date=mydata$date), sum)  
hist(stepsByday$x, col = "blue", breaks = 25,  
 main = "Total steps taken each day",  
 xlab="Number of steps per day")



## stepsByday has 53 obs. and 2 vars.: 1 col\_ steps.date (date) and 2\_col\_ x (value)  
  
### Calculate the mean and media  
mean\_steps<-mean(stepsByday[,2])  
print(mean\_steps)

## [1] 10766.19

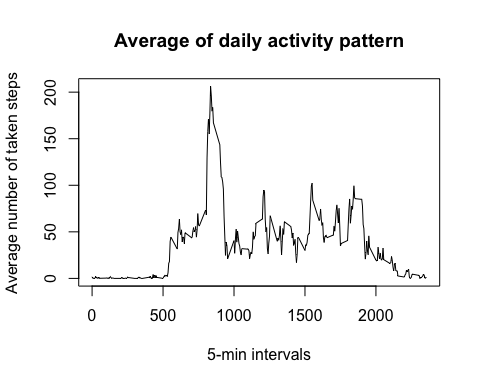
### Result: [1] 10766.19  
median\_steps<-median(stepsByday[,2])  
print(median\_steps)

## [1] 10765

### Result: [1] 10765

# P3:The avg of daily activity pattern along 5-min intervals?

avg\_day<-aggregate(mydata$steps, by=list(Intervals=mydata$interval), mean)  
## agv\_day has 288 obs. of 2 vars.: 1\_col\_ intervals and 2\_col\_ x value   
  
plot(avg\_day$Intervals, avg\_day$x, type = "l",  
 main = "Average of daily activity pattern",  
 ylab = "Average number of taken steps", xlab= "5-min intervals")



## Intervals with max. of number of steps  
max\_val <-which.max(avg\_day$x)  
max\_interval<-avg\_day[max\_val,1]  
print(max\_interval)

## [1] 835

### Result: [1] 835

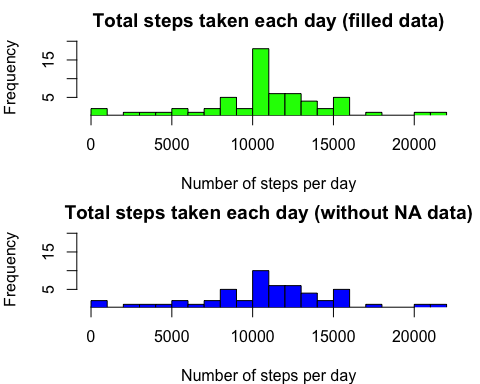
# P4: Imputing with missing values with mean value

###( since the values of the mean and median are similar)

NA\_num<-length (which(is.na(data$steps)))  
print(NA\_num)

## [1] 2304

### Result: [1] 2304 same as Q1 diff. obs.   
  
# filled missing data with the mean val and make histogram  
mydatafilled<-data  
mydatafilled$steps<-impute(data$steps, mean)  
  
stepsByday1<-aggregate(mydatafilled$steps, by=list(Steps.Date = mydatafilled$date),sum)  
par(mfrow = c(2,1), mar=c(4,4,2,1))  
hist(stepsByday1$x, col = "green", breaks = 25,  
 main = "Total steps taken each day (filled data)",  
 xlab = "Number of steps per day", ylim = c(1, 20))  
hist(stepsByday$x, col = "blue", breaks = 25,  
 main = "Total steps taken each day (without NA data)",  
 xlab="Number of steps per day", ylim = c(1, 20))



### Calculate the mean and media  
mean\_steps1<-mean(stepsByday1[,2])  
print(mean\_steps1)

## [1] 10766.19

### Result: [1] 10766.19  
  
median\_steps1<-median(stepsByday1[,2])  
print(median\_steps1)

## [1] 10766.19

### Result: [1] 10766.19

# P5: Difference activity in pattern bet. weekdays and weekend

### Define the weekday and weekend from stepsByday1 data  
weekday.or.weekend <- function(date) {  
 day <- weekdays(date)  
 if (day %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))  
 return("weekday")  
 else if (day %in% c("Saturday", "Sunday"))  
 return("weekend")  
 else  
 stop("invalid date")  
 }  
mydatafilled$date<-as.Date(mydatafilled$date)  
mydatafilled$day <-sapply(mydatafilled$date, weekday.or.weekend)  
  
## Make ggplot  
avg\_date <- aggregate(steps~ interval + day, data=mydatafilled, mean)  
ggplot(avg\_date, aes(interval, steps))+ geom\_line() +  
 facet\_grid(day ~.)+ labs(x="5-min interval", y="Number of steps")+  
 ggtitle ("Weekdays and Weekend activity patterns")

