Reproducible Research - Peer Assesment 1

Igor Siqueira Cortez

14 de outubro de 2015

**Standard Settings**

echo = TRUE # make code visible  
options(scipen = 1) # Turn off scientific notations for numbers

**Loading and processing the Data**

data <- read.csv("activity.csv", colClasses = c("integer", "Date", "factor"))  
data$month <- as.numeric(format(data$date, "%m"))  
notNA<-na.omit(data)  
stepstotal<-aggregate(steps ~ date, data=data, sum, na.rm=TRUE)  
rownames(stepstotal) <- 1:nrow(stepstotal)  
head(stepstotal)

## 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

dim(stepstotal)

## [1] 53 2

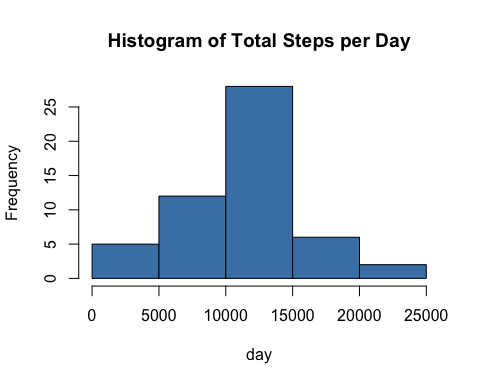
library(ggplot2)  
library(lattice)

**What is the mean of total steps taken per day?**

For this part of the assignment, you can ignore the missing values in the dataset.

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

stepstotal<-aggregate(steps ~ date, data=data, sum, na.rm=TRUE)  
hist(stepstotal$steps, main="Histogram of Total Steps per Day", xlab="day", col="steelblue")



1. Calculate and report the mean and median of steps taken per day

Mean of steps taken per day:

mean(stepstotal$steps)

## [1] 10766.19

Median of steps taken per day:

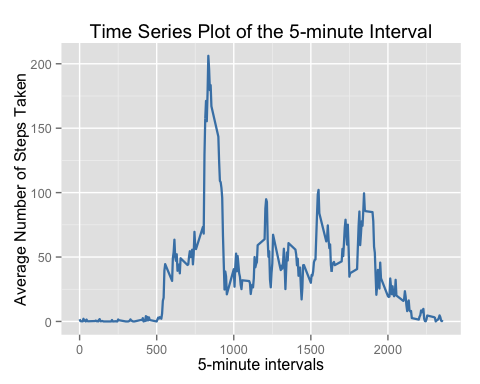
median(stepstotal$steps)

## [1] 10765

**What is the average daily activity pattern ?**

1. 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)

avg\_steps <- aggregate(notNA$steps, list(interval = as.numeric(as.character(notNA$interval))), FUN = "mean")  
names(avg\_steps)[2] <- "meanOfSteps"  
  
ggplot(avg\_steps, aes(interval, meanOfSteps)) + geom\_line(color = "steelblue", size = 0.8) + labs(title = "Time Series Plot of the 5-minute Interval", x = "5-minute intervals", y = "Average Number of Steps Taken")



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

avg\_steps[avg\_steps$meanOfSteps == max(avg\_steps$meanOfSteps), ]

## interval meanOfSteps  
## 104 835 206.1698

**Imputing Missing Values**

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

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] 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.

Strategy used: use the mean.

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

new\_data <- data   
for (i in 1:nrow(new\_data)) {  
 if (is.na(new\_data$steps[i])) {  
 new\_data$steps[i] <- avg\_steps[which(new\_data$interval[i] == avg\_steps$interval), ]$meanOfSteps  
 }  
}  
head(new\_data)

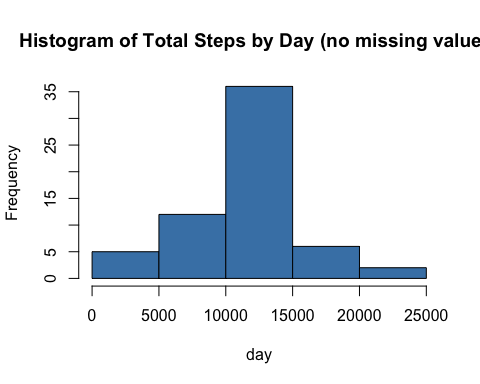
## steps date interval month  
## 1 1.7169811 2012-10-01 0 10  
## 2 0.3396226 2012-10-01 5 10  
## 3 0.1320755 2012-10-01 10 10  
## 4 0.1509434 2012-10-01 15 10  
## 5 0.0754717 2012-10-01 20 10  
## 6 2.0943396 2012-10-01 25 10

sum(is.na(new\_data))

## [1] 0

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.

stepstotal2<-aggregate(steps ~ date, data=new\_data, sum, na.rm=TRUE)  
hist(stepstotal2$steps, main="Histogram of Total Steps by Day (no missing values)", xlab="day", col="steelblue")



4.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?

new\_mean <- mean(stepstotal2$steps)  
new\_mean

## [1] 10766.19

Median total number of steps taken per day:

new\_median <- median(stepstotal2$steps)  
new\_median

## [1] 10766.19

Compare them with the two before imputing missing data:

old\_mean <- mean(stepstotal$steps)  
old\_median <- median(stepstotal$steps)  
new\_mean - old\_mean

## [1] 0

new\_median - old\_median

## [1] 1.188679

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

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.

head(new\_data)

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

new\_data$weekdays <- factor(format(new\_data$date, "%A"))  
levels(new\_data$weekdays)

## [1] "Domingo" "Quarta Feira" "Quinta Feira" "Sábado"   
## [5] "Segunda Feira" "Sexta Feira" "Terça Feira"

levels(new\_data$weekdays) <- list(weekday = c("Segunda Feira", "Terça Feira",  
 "Quarta Feira",   
 "Quinta Feira", "Sexta Feira"),  
 weekend = c("Sábado", "Doming"))  
levels(new\_data$weekdays)

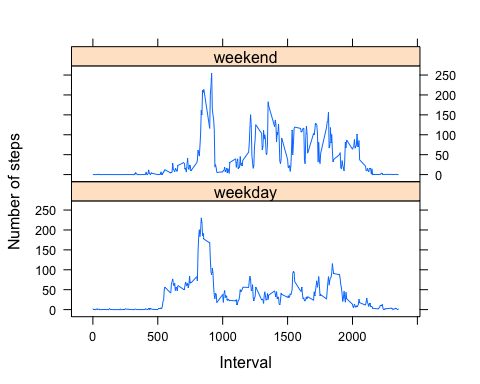
## [1] "weekday" "weekend"

table(new\_data$weekdays)

##   
## weekday weekend   
## 12960 2304

2.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).

avg\_steps <- aggregate(new\_data$steps,   
 list(interval = as.numeric(as.character(new\_data$interval)),   
 weekdays = new\_data$weekdays),  
 FUN = "mean")  
names(avg\_steps)[3] <- "meanOfSteps"  
  
  
xyplot(avg\_steps$meanOfSteps ~ avg\_steps$interval | avg\_steps$weekdays, layout = c(1, 2), type = "l", xlab = "Interval", ylab = "Number of steps")



\*Conclusion The plot indicates that the person moves around more (or more active) during the weekend days.