Untitled

Julian Sanchez - [wathnai@hotmail.com](mailto:wathnai@hotmail.com)

22 de marzo de 2018

### Set Working directory and read the DataSet

setwd("~/Data\_Scientist/Reproducible Research/RepData\_PeerAssessment1")  
  
unzip ("activity.zip") # extracts the content of the zip file  
  
activity <- read.csv("activity.csv") # loads the dataset

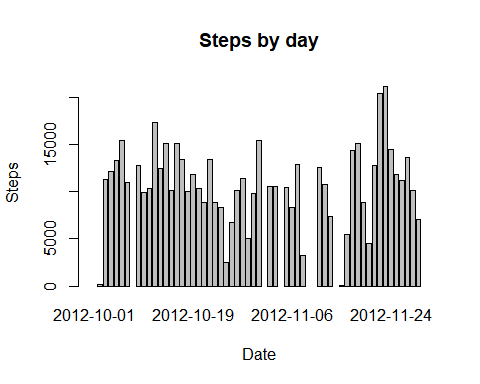
### What is mean total number of steps taken per day?

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

steps\_by\_day <- aggregate(activity$steps, by=list(date=activity$date), sum)  
max(steps\_by\_day$x, na.rm = TRUE)## highest value

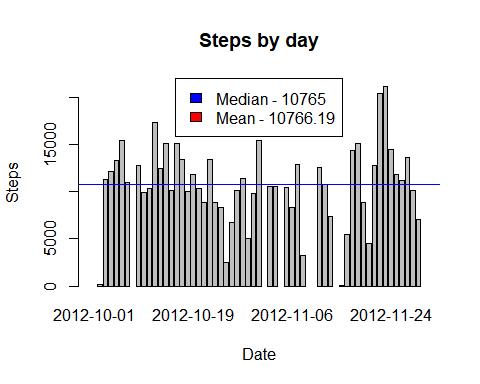
## [1] 21194

barplot(height=steps\_by\_day$x, names.arg=steps\_by\_day$date, xlab="Date", ylim=c(0,22000), ylab="Steps", main = "Steps by day")



2 Calculate and report the mean and median total number of steps taken per day

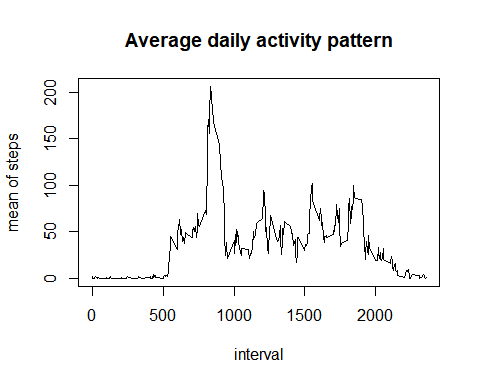
mean\_steps\_by\_day <- mean(steps\_by\_day$x, na.rm = TRUE)  
median\_steps\_by\_day <- median(steps\_by\_day$x, na.rm = TRUE)  
barplot(height=steps\_by\_day$x, names.arg=steps\_by\_day$date, xlab="Date",   
 ylim=c(0,22000), ylab="Steps", main = "Steps by day")  
abline(h = mean\_steps\_by\_day, col = "red")  
abline(h = median\_steps\_by\_day, col = "blue")  
legend("top",   
 legend = c("Median - 10765 ", "Mean - 10766.19"),   
 fill = c("blue", "red"))



### What is the average daily activity pattern?

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)

steps\_by\_interval <- aggregate(steps ~ interval, data = activity, FUN = mean)  
plot(steps\_by\_interval$interval, steps\_by\_interval$steps, type="l", xlab="interval", ylab="mean of steps",   
 main="Average daily activity pattern")



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

maxSteps <- steps\_by\_interval$interval[which.max(steps\_by\_interval$steps)]  
maxSteps

## [1] 835

### 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.

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

sum(is.na(activity))

## [1] 2304

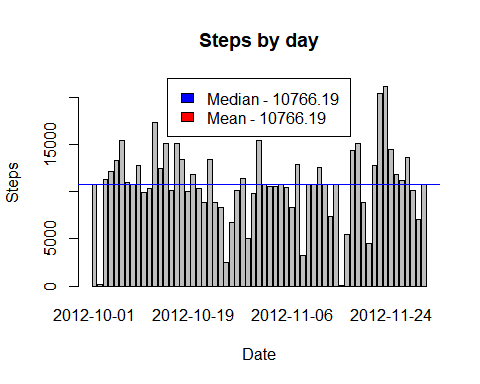
Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

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

#Replacing NAs with the mean of each day  
activity\_na <- activity  
activity\_na[is.na(activity)]=mean(activity$steps, na.rm=TRUE)  
# head(activity\_na)  
# activity\_na  
#which (is.na(activity\_na))  
  
group\_by\_date\_days\_na <- aggregate(activity\_na$steps, by=list(activity\_na$date), FUN=sum)  
group\_by\_date\_days\_mean\_na <- aggregate(activity\_na$steps, by=list(activity\_na$date), FUN=mean)  
group\_by\_date\_days\_median\_na <- aggregate(activity\_na$steps, by=list(activity\_na$date), FUN=median)  
colnames(group\_by\_date\_days\_na) <- c("date", "total\_steps")  
colnames(group\_by\_date\_days\_mean\_na) <- c("date", "steps\_mean")  
colnames(group\_by\_date\_days\_median\_na) <- c("date", "steps\_median")  
# group\_by\_date\_days\_mean\_na  
# group\_by\_date\_days\_na  
# group\_by\_date\_days\_median\_na

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?

#group\_by\_date\_days\_na$date <- as.numeric(group\_by\_date\_days\_na$date)  
  
barplot(group\_by\_date\_days\_na$total\_steps, names.arg=group\_by\_date\_days\_na$date, xlab="Date",   
 ylim=c(0,22000), ylab="Steps", main = "Steps by day")  
abline(h = mean(group\_by\_date\_days\_na$total\_steps), col = "red")  
abline(h = median(group\_by\_date\_days\_na$total\_steps), col = "blue")  
legend("top",   
 legend = c("Median - 10766.19 ", "Mean - 10766.19"),   
 fill = c("blue", "red"))



### 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.

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.

activity\_na$day <- weekdays(as.Date(activity\_na$date))  
activity\_na$weekend <- ifelse(activity\_na$day == "sábado" | activity\_na$day == "domingo","weekend", "weekday")  
  
activity\_na\_weekend <- aggregate(activity\_na$steps, by=list(activity\_na$weekend , activity\_na$interval), FUN=mean)  
colnames(activity\_na\_weekend) <- c("weekday" , "interval", "steps")  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

ggplot(activity\_na\_weekend, aes(x = interval, y = steps)) + ylab("Number of Steps") + geom\_line() + facet\_grid(weekday~.)

