# Reproducible Research: Peer Assessment 1

## Loading and preprocessing the data

setwd("~/GitHub/RepData\_PeerAssessment1")  
activity <- read.csv("activity.csv", colClasses = c("integer","Date","integer"))

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

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

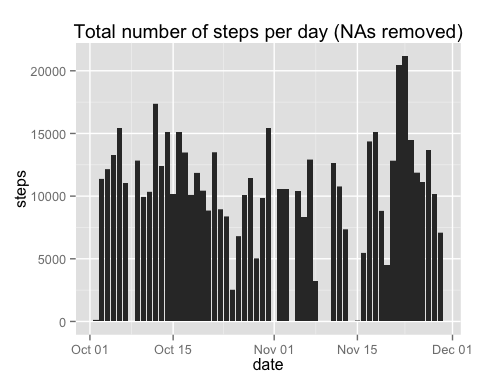
**1.** Make a histogram of the total number of steps taken each day

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

library(ggplot2)  
qplot(date, steps, data = activity %>% group\_by(date), geom = "bar", stat = "identity", main = "Total number of steps per day (NAs removed)")

## Warning: Removed 2304 rows containing missing values (position\_stack).



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

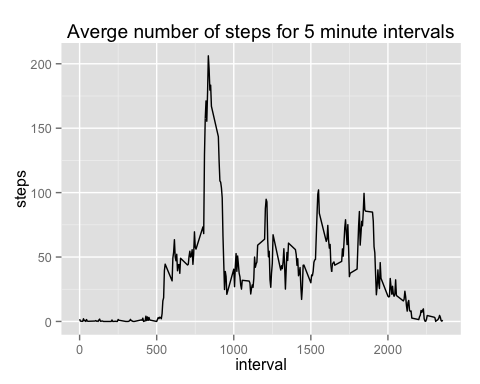
tot\_steps <- activity %>% group\_by(date) %>% summarise(tot\_steps = sum(steps, na.rm=TRUE))  
avg\_steps <- mean(tot\_steps$tot\_steps, na.rm=TRUE)  
med\_steps <- median(tot\_steps$tot\_steps, na.rm=TRUE)

Mean number of steps per day: 9354.2295 Median number of steps perday: 10395

## 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)

qplot(interval, steps, data = activity %>% group\_by(interval) %>% summarise(steps= mean(steps, na.rm=TRUE)), geom = "line", main = "Averge number of steps for 5 minute intervals")



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

d1 <- activity %>% group\_by(interval) %>% summarise(avg\_steps = mean(steps, na.rm=TRUE))  
with(d1,d1[order(-avg\_steps),])

## Source: local data frame [288 x 2]  
##   
## interval avg\_steps  
## 104 835 206.2  
## 105 840 195.9  
## 107 850 183.4  
## 106 845 179.6  
## 103 830 177.3  
## 101 820 171.2  
## 108 855 167.0  
## 100 815 157.5  
## 102 825 155.4  
## 109 900 143.5  
## .. ... ...

The interval with the maximum average number of steps is **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.

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

summary(activity)

## steps date interval   
## Min. : 0.0 Min. :2012-10-01 Min. : 0   
## 1st Qu.: 0.0 1st Qu.:2012-10-16 1st Qu.: 589   
## Median : 0.0 Median :2012-10-31 Median :1178   
## Mean : 37.4 Mean :2012-10-31 Mean :1178   
## 3rd Qu.: 12.0 3rd Qu.:2012-11-15 3rd Qu.:1766   
## Max. :806.0 Max. :2012-11-30 Max. :2355   
## NA's :2304

There are **2304 NAs**

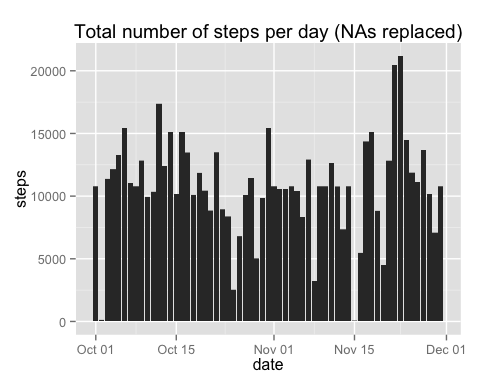
1. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

#### Replace missing values with the daily mean

act\_nareplace <- activity #Duplicate original data frame  
#Replace values with mean number of steps over all observations  
act\_nareplace$steps[is.na(act\_nareplace$steps)] <- mean(act\_nareplace$steps, na.rm = TRUE)

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?

qplot(date, steps, data = act\_nareplace %>% group\_by(date), geom = "bar", stat = "identity", main = "Total number of steps per day (NAs replaced)")



summary(act\_nareplace %>% group\_by(date) %>% summarise(tot\_steps = sum(steps)))

## date tot\_steps   
## Min. :2012-10-01 Min. : 41   
## 1st Qu.:2012-10-16 1st Qu.: 9819   
## Median :2012-10-31 Median :10766   
## Mean :2012-10-31 Mean :10766   
## 3rd Qu.:2012-11-15 3rd Qu.:12811   
## Max. :2012-11-30 Max. :21194

New mean = 10766 New median = 10766

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

act\_nareplace$weekday <- weekdays(act\_nareplace$date)  
act\_nareplace$weekend <- act\_nareplace$weekday=="Saturday" | act\_nareplace$weekday=="Sunday"  
act\_nareplace$weekend <- ifelse(act\_nareplace$weekend==TRUE,"Weekend","Weekday")  
  
qplot(interval, steps, data = act\_nareplace %>% group\_by(interval,weekend) %>% summarise(steps = mean(steps)), facets = weekend~., geom = "line", ylab = "Number of steps")

