Reproducible Research: Peer Assessment 1

## Loading and preprocessing the data

1. Load the data (i.e. read.csv())

activity<-read.csv("activity.csv",stringsAsFactors=FALSE)

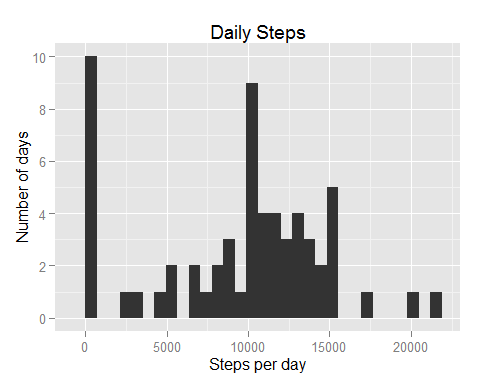
1. Process/transform the data (if necessary) into a format suitable for your analysis

# Convert the data column from character to an appropriate date format  
activity$date<-strptime(activity$date,"%Y-%m-%d")

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

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

library(plyr)  
library(ggplot2)  
  
agg1<-ddply(activity,.(date),summarize,steps=sum(steps,na.rm=T))  
  
ggplot(agg1,aes(x=steps))+scale\_y\_continuous(breaks=seq(0,10,by=2))+  
 ylab("Number of days")+xlab("Steps per day")+ggtitle("Daily Steps")+geom\_histogram()



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

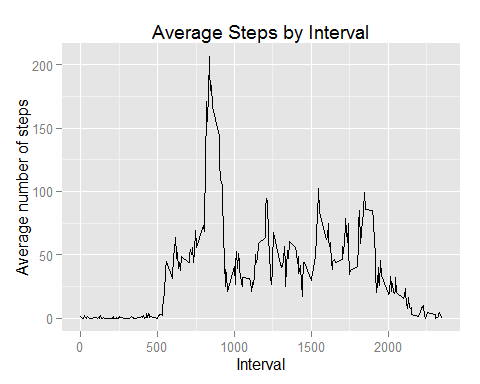
with(agg1,data.frame(mean=mean(steps),median=median(steps)))

## mean median  
## 1 9354.23 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)

agg2<-ddply(activity,.(interval),summarize,mean\_steps=mean(steps,na.rm=TRUE))  
  
ggplot(agg2,aes(x=interval,y=mean\_steps))+ggtitle("Average Steps by Interval")+  
 ylab("Average number of steps")+xlab("Interval")+geom\_line()



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

agg2[agg2$mean\_steps==max(agg2$mean\_steps),]

## interval mean\_steps  
## 104 835 206.1698

## Imputing missing values

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

table(is.na(activity$steps))["TRUE"]

## TRUE   
## 2304

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.

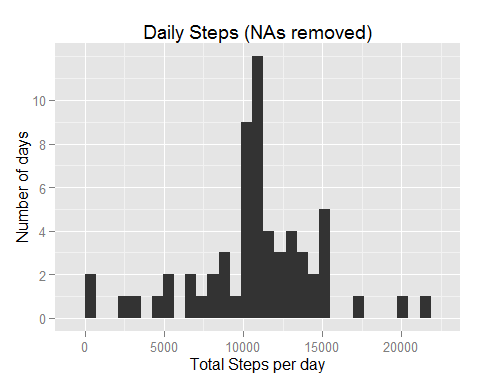
**Strategy:** I'll use mean for 5-minute intervals which have been computed in the previous task.

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

activity\_bis<-activity  
  
# I'll use the data frame agg2 prepared in the previous task  
activity\_bis<-merge(agg2,activity\_bis,by="interval",all=TRUE)  
  
activity\_bis$steps<-with(activity\_bis,ifelse(is.na(steps),mean\_steps,steps))  
  
activity\_bis$mean\_steps<-NULL

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?

agg3<-ddply(activity\_bis,.(date),summarize,steps=sum(steps))  
  
ggplot(agg3,aes(x=steps))+scale\_y\_continuous(breaks=seq(0,10,by=2))+  
 ylab("Number of days")+xlab("Total Steps per day")+ggtitle("Daily Steps (NAs removed)")+geom\_histogram()



with(agg3,data.frame(mean=mean(steps),median=median(steps)))

## mean median  
## 1 10766.19 10766.19

**Answer:** The above numbers differ, they'e now larger as previously ignored values were replaced with certain numbers.

## 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.

# To be sure that the names of days are in English  
Sys.setlocale("LC\_TIME", "English")

## [1] "English\_United States.1252"

activity$dow<-as.factor(ifelse(weekdays(activity$date,abbreviate=T) %in% c("Sat","Sun"),"weekend","weekday"))  
  
# Confirm by displaying first rows  
head(activity)

## steps date interval dow  
## 1 NA 2012-10-01 0 weekday  
## 2 NA 2012-10-01 5 weekday  
## 3 NA 2012-10-01 10 weekday  
## 4 NA 2012-10-01 15 weekday  
## 5 NA 2012-10-01 20 weekday  
## 6 NA 2012-10-01 25 weekday

1. 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). The plot should look something like the following, which was created using **simulated data**:

agg4<-ddply(activity,.(dow,interval),summarize,mean\_steps=mean(steps,na.rm=TRUE))  
  
ggplot(agg4,aes(x=interval,y=mean\_steps))+ggtitle("Average Steps by Interval")+  
 ylab("Average number of steps")+xlab("Interval")+facet\_grid(dow~.)+geom\_line()

