Reproducible Research: Peer Assessment 1

## Introduction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the "quantified self" movement -- a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

## Data

The data for this assignment can be downloaded from the course web site:

Dataset: [Activity monitoring data][datalink] [52K] The variables included in this dataset are:

steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)

date: The date on which the measurement was taken in YYYY-MM-DD format

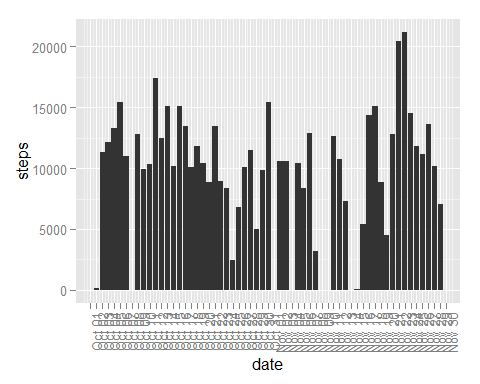
interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

## Loading and preprocessing the data

## Load ggplot2 library  
library(ggplot2)  
## Load scales library  
library(scales)  
  
## Load the data (i.e. read.csv())  
data <- read.csv("activity.csv")  
  
# Sum the total number of steps per day  
total <- aggregate(data$steps, list(date=data$date), sum)  
total$steps <- total$x  
total$date <- as.Date(total$date)  
  
# Calculate the average number of steps per day  
avg <- aggregate(x=data$steps, list(date=data$date), mean)   
avg.steps <- mean(data$steps, na.rm=TRUE)  
  
# Calculate the median number of steps per day  
med <- aggregate(x=data$steps, list(date=data$date), median)  
med.steps <- median(data$steps, na.rm=TRUE)  
  
# Make a histogram of the total number of steps taken each day  
datebreaks <- as.Date(total$date)  
p <- ggplot(total, aes(y=steps, x=date)) +geom\_bar(stat="identity")  
p <- p + theme(axis.text.x = element\_text(angle=90, hjust=1))  
p <- p + scale\_x\_date(breaks=datebreaks, labels=date\_format("%b %d") )  
  
p

## Warning: Removed 8 rows containing missing values (position\_stack).



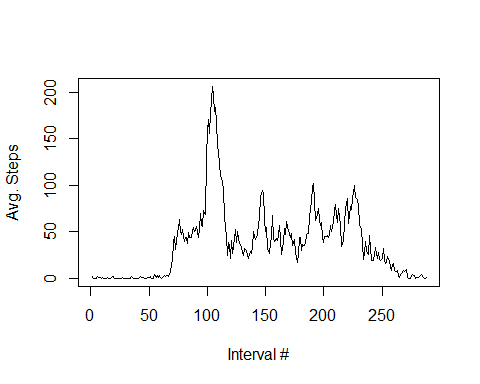
## What is the mean total number of steps taken per day?

The mean number of steps taken per day With missing values is: 37.3826.

The median number of steps taken per day with missing values is: 0.

## What is the average daily activity pattern?

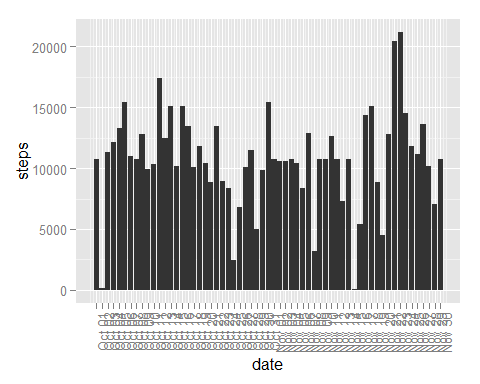
## Create interval bins of 1-288 (intervals per day)  
data$intervalbin <- rep(seq(1,288))  
  
## Calculate the average number of steps per interval  
avginterval <- aggregate(x=data$steps, list(y=data$intervalbin), mean, na.rm=TRUE)  
  
## Interval with the maximum number of steps  
max.step <- avginterval$y[which.max(avginterval$x)]  
  
## Time series plot of interval vs avg number of steps  
plot(avginterval$y, avginterval$x, type="l", xlab="Interval #", ylab="Avg. Steps")



The 5-minute interval that contains the maximum number of steps is: 104

## Imputing missing values

## Calculate number of rows with missing values  
num.rows.missing <- nrow(data[is.na(data$steps)==TRUE,])  
  
## Calculate missing days/intervals  
daysNA <- nrow(array(unique(data$date[is.na(data$steps)])))  
  
## Create a new data set  
## Copy the data frame  
newdata <- data  
  
## Imputing values strategy: assign the missing interval value the mean of the interval of all days  
## Assign mean steps for each missing interval using means calculated above  
newdata[is.na(newdata$steps)==TRUE,1] <- avginterval$x  
  
# Sum the total number of steps per day on the new dataset  
total2 <- aggregate(newdata$steps, list(date=newdata$date), sum)  
total2$steps <- total2$x  
total2$date <- as.Date(total2$date)  
  
# Calculate the average number of steps per day  
avg2 <- aggregate(x=newdata$steps, list(date=newdata$date), mean)   
avg.steps2 <- mean(newdata$steps, na.rm=TRUE)  
  
# Calculate the median number of steps per day  
med2 <- aggregate(x=newdata$steps, list(date=newdata$date), median)  
med.steps2 <- median(newdata$steps, na.rm=TRUE)  
  
# Create a histogram of the number of steps per day  
datebreaks <- as.Date(total2$date)  
p3 <- ggplot(total2, aes(y=steps, x=date)) +geom\_bar(stat="identity")  
p3 <- p3 + theme(axis.text.x = element\_text(angle=90, hjust=1))  
p3 <- p3 + scale\_x\_date(breaks=datebreaks, labels=date\_format("%b %d") )  
  
p3



Total number of rows missing: 2304

Total numnber of days missing: 8

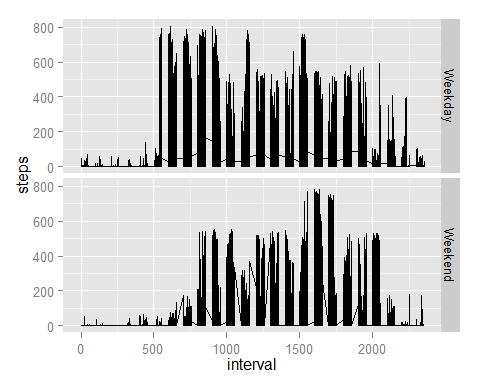
The mean number of steps taken per day Without missing values is: 37.3826.

The median number of steps taken per day without missing values is: 0.

There is no impact to the mean or the median of steps when populating the missing values in the dataset.

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

## Setup a new factor and classify by Weekday and Weekend  
newdata$day <- factor(weekdays(as.Date(newdata$date)))  
levels(newdata$day)[1:2] <- "Weekday"  
levels(newdata$day)[4:6] <- "Weekday"  
levels(newdata$day)[2:3] <- "Weekend"  
  
# Create a time series panel of two plots (weekdays and weekends)  
p4 <- qplot(interval, steps, data=newdata, geom="line", facets=day~.)  
  
p4

 [datalink]:<https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip>