Reproducible Research: Peer Assessment 1

Author: rrwiren

Current date ;) 28 March, 2018

### Loading libraries and some housekeeping before starting

library(ggplot2)  
# library(scales)  
# library(stats)  
# library(tidyr)  
# library(lattice)  
# library(mice)  
library(Hmisc)

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

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

# library(DMwR)  
setwd("~/GitHub/RepData\_PeerAssessment1")

## Loading and preprocessing the data

if(!file.exists('activity.csv')){  
 unzip('activity.zipMiss <- function(x){sum(is.na(x))/length(x)\*100}p')  
}  
activity <- read.csv('activity.csv')

#### Having a quick look at the data

names(activity)

## [1] "steps" "date" "interval"

head(activity, n = 3L)

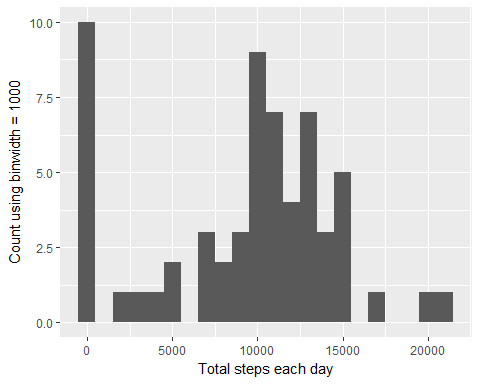
## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10

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

stepsDay <- tapply(activity$steps, activity$date, sum, na.rm=TRUE)

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

qplot(stepsDay, xlab='Total steps each day', ylab='Count using binwidth = 1000', binwidth=1000)



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

stepsDayMean <- mean(stepsDay)  
print(stepsDayMean)

## [1] 9354.23

stepsDayMedian <- median(stepsDay)  
print(stepsDayMedian)

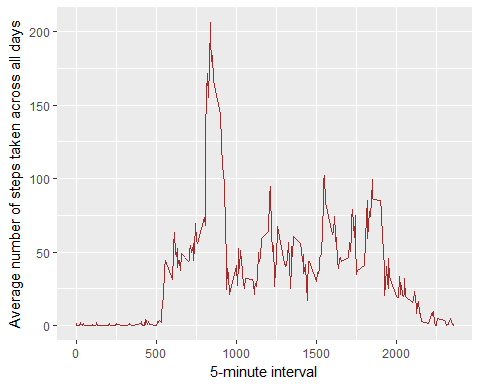
## [1] 10395

## What is the average daily activity pattern?

avgSteps <- aggregate(x=list(meanSteps=activity$steps), by=list(interval=activity$interval), FUN=mean, na.rm=TRUE)

#### 1. Time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

ggplot(data=avgSteps, aes(x=interval, y=meanSteps)) +  
 geom\_line(colour = "brown") +  
 xlab("5-minute interval") +  
 ylab("Average number of steps taken across all days")



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

mostSteps <- which.max(avgSteps$meanSteps)  
timeMostSteps <- gsub("([0-9]{1,2})([0-9]{2})", "\\1:\\2", avgSteps[mostSteps,'interval'])  
print(timeMostSteps)

## [1] "8:35"

## 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)

missingValues <- length(which(is.na(activity$steps)))  
print(missingValues)

## [1] 2304

anyNA(activity$steps)

## [1] TRUE

##### 2. 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.

activityImputedMean <- activity  
activityImputedMean$steps <- impute(activity$steps, fun = mean)

###### Quick check

missingValuesMean <- length(which(is.na(activityImputedMean$steps)))  
print(missingValuesMean)

## [1] 0

anyNA(activityImputedMean$steps)

## [1] FALSE

mean(activity$steps)

## [1] NA

mean(activityImputedMean$steps)

## [1] 37.3826

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

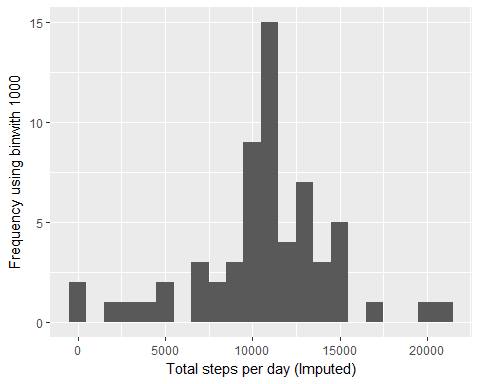
activityImputed <- activityImputedMean  
summary(activityImputed)

##   
## 2304 values imputed to 37.3826

## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 37.38 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## (Other) :15840

#### 4. 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?

stepsDayImputed <- tapply(activityImputed$steps, activityImputed$date, sum)  
qplot(stepsDayImputed, xlab='Total steps per day (Imputed)', ylab='Frequency using binwith 1000', binwidth=1000)



## 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.

#### 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.

#### 2. Panel plot containing a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

## Additional information