C5 W2 Project 1

Jishan Hossain

April 25, 2017

setwd("C:/My Documents/1 Jishan/2 R (Data Science)/Data Scientists (John Hopkins) Coursera/5 Reproducible Research/Week 2/Project 1")  
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

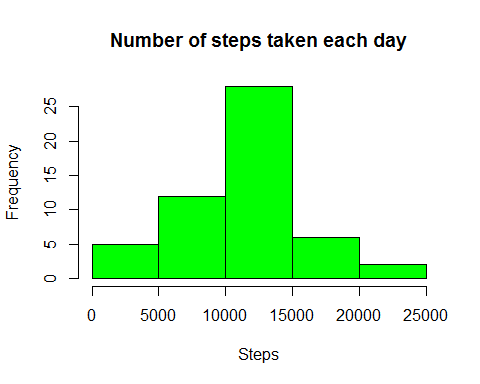
if (!file.exists("./data")) {  
 dir.create("./data")  
}  
f1 <- file.path(getwd(), "./repdata\_data\_activity.zip")  
unzip(f1, exdir = "./data")  
activity <- read.csv("./data/activity.csv")

Q. What is mean total number of steps taken per day? 1. Calculate the total number of steps taken per day

sum\_steps <- aggregate(steps ~ date, activity, sum)

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

hist(sum\_steps$steps,   
 main = "Number of steps taken each day",   
 xlab = "Steps",   
 border = "black",   
 col = "green")



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

mean\_steps <- mean(sum\_steps$steps)  
median\_steps <- sum\_steps[sum\_steps$steps == median(sum\_steps$steps), ]  
mean\_steps

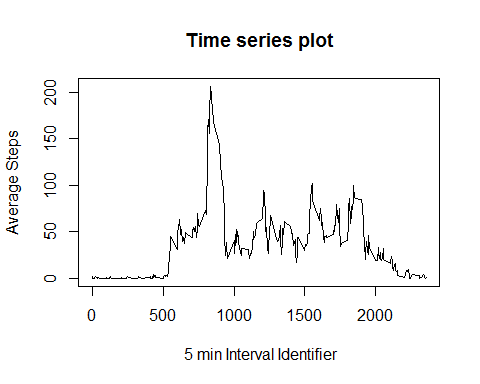
## [1] 10766.19

median\_steps

## date steps  
## 37 2012-11-12 10765

Q. 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)

avg\_steps <- aggregate(steps ~ interval, activity, FUN = mean)  
plot(avg\_steps$interval, avg\_steps$steps, type = "l", main = "Time series plot", xlab = " 5 min Interval Identifier", ylab = "Average Steps")



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

max\_intv <- avg\_steps[(avg\_steps$steps == max(avg\_steps$steps)), ]  
max\_intv

## interval steps  
## 104 835 206.1698

Q. 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)

sum(is.na(activity$steps))

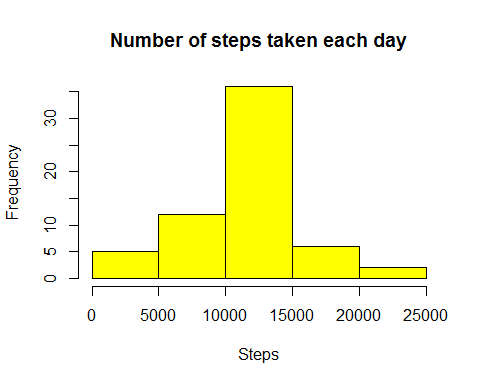
## [1] 2304

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

activity1 <- activity  
activity2 <- activity  
activity1$steps <- ifelse(!is.na(activity1$steps), activity1$steps, avg\_steps$steps)  
activity2$steps <- ifelse(!is.na(activity2$steps), activity2$steps, avg\_steps$steps)

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?

sum\_activity1 <- aggregate(steps ~ date, activity1, sum)  
hist(sum\_activity1$steps,   
 main = "Number of steps taken each day",   
 xlab = "Steps",   
 border = "black",   
 col = "yellow")



The mean and median of the new dataset after imputing missing values have not changed from the original dataset as the missing values had been imputed as the **averages** of 5 minute intervals across all days from the original dataset.

mean\_act1 <- mean(sum\_activity1$steps)  
median\_act1 <- median(sum\_activity1$steps)

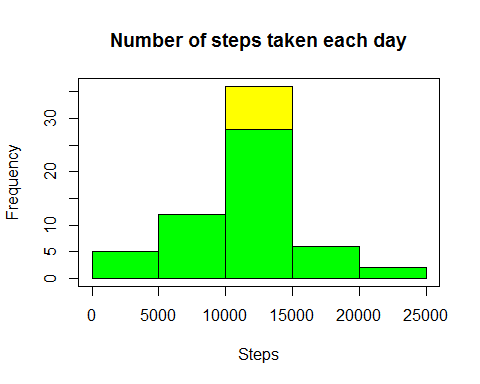
mean\_act1

## [1] 10766.19

median\_act1

## [1] 10766.19

hist(sum\_activity1$steps,   
 main = "Number of steps taken each day",   
 xlab = "Steps",   
 border = "black",   
 col = "yellow")  
hist(sum\_steps$steps,   
 main = "Number of steps taken each day",   
 xlab = "Steps",   
 border = "black",   
 col = "green",   
 add = T)  
box()



Imposing the two histograms to check for differences reveals the difference in frequency between steps brackets of 10000 and 15000.

Q. 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. 2. 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). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

library(lubridate)

##   
## Attaching package: 'lubridate'

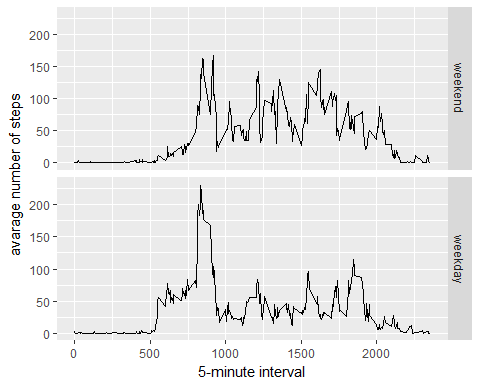
## The following object is masked from 'package:base':  
##   
## date

activity2$date <- as.Date(activity2$date)  
weekdays1 <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')  
activity2$wDay <- factor((weekdays(activity2$date, abbreviate = FALSE) %in% weekdays1), levels = c(FALSE, TRUE), labels = c('weekend', 'weekday'))

avg\_steps2 <- aggregate(steps ~ interval + wDay, data = activity2, FUN = mean)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

ggplot(avg\_steps2, aes(interval, steps)) +   
 geom\_line() +   
 facet\_grid(wDay ~ .) +  
 xlab("5-minute interval") +   
 ylab("avarage number of steps")



Looking at the two time series plots drawn for the weekend and weekdays for average number of steps for 5 minute interval identifier it can be seen that they look similar but not identical.

Weekday and weekend peaks are around similar time intervals but then the average for the rest of the time intervals for weekdays seems to be lesser than that of weekends. Weekdays average falls sharply after the peak and the difference is larger compared to the difference between the peak and the rest in weekend time intervals.