Coursera - Reproducible Research Course Project 1

Charl Arthur Henry Cowley

May 18, 2017

## Loading and preprocessing the data

1.Load the data using read.csv

activity <- read.csv("activity.csv",header = TRUE)  
class(activity$date)

## [1] "factor"

head(activity)

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

1. We see that "date" is of class "factor". We will now change it to "POSIXct".

activity$date <- as.POSIXct(activity$date)  
class(activity$date)

## [1] "POSIXct" "POSIXt"

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

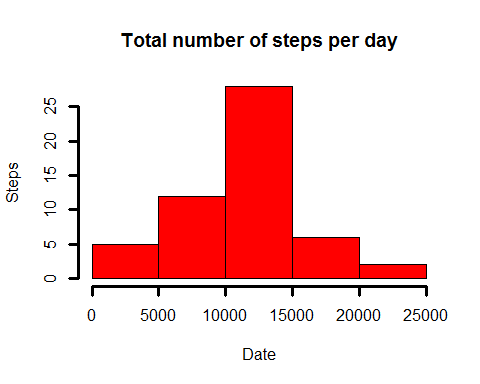
1. Calculate the total number of steps taken per day

daymean <- aggregate(steps ~ date, data = activity, sum)  
head(daymean)

## date steps  
## 1 2012-10-02 126  
## 2 2012-10-03 11352  
## 3 2012-10-04 12116  
## 4 2012-10-05 13294  
## 5 2012-10-06 15420  
## 6 2012-10-07 11015

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

par(mfrow=c(1,1))  
hist(daymean$steps,lwd=3,col="red",xlab="Date",ylab="Steps",main="Total number of steps per day")



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

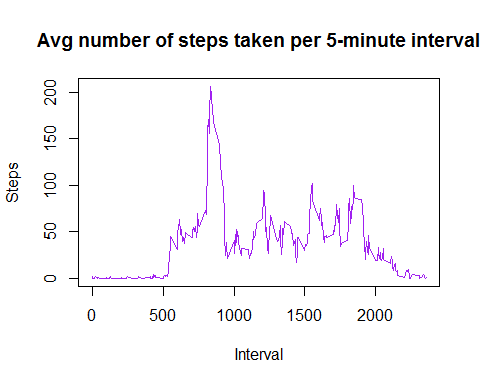
h <- summary(daymean$steps)  
h[c("Median","Mean")]

## Median Mean   
## 10760 10770

## What is the average daily activity pattern?

1. Make a time series plot of the 5-minute interval and the average number of steps taken, averaged across all days

intmean <- aggregate(steps~interval,data=activity,mean)  
plot(intmean$interval,intmean$steps,type="l",xlab="Interval",col="purple",ylab="Steps",main="Avg number of steps taken per 5-minute interval")



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

intmean[which(intmean$steps == max(intmean$steps)),]

## interval steps  
## 104 835 206.1698

## Imputing missing values

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

summary(activity)

## steps date interval   
## Min. : 0.00 Min. :2012-10-01 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.:2012-10-16 1st Qu.: 588.8   
## Median : 0.00 Median :2012-10-31 Median :1177.5   
## Mean : 37.38 Mean :2012-10-31 Mean :1177.5   
## 3rd Qu.: 12.00 3rd Qu.:2012-11-15 3rd Qu.:1766.2   
## Max. :806.00 Max. :2012-11-30 Max. :2355.0   
## NA's :2304

We see that there are only missing values in the "Steps" variable (2304 missings)

1. Devise a strategy for filling in all of the missing values in the data set. I imputed the mean for that 5-minute interval.
2. Create a new data set that is equal to the original data set but with the missing data filled in.

library(dplyr)

##   
## Attaching package: 'dplyr'

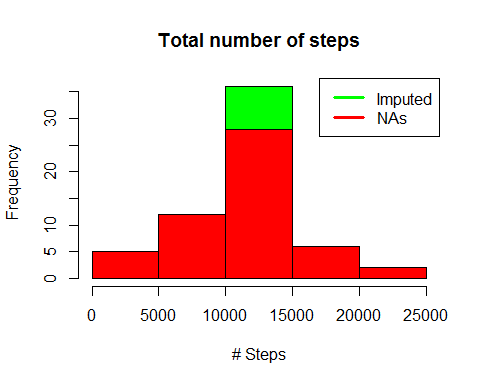
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df1 <- activity %>%  
 group\_by(interval) %>%  
 summarise(avg=mean(steps,na.rm=TRUE)) %>%  
 merge(activity, ., all.x=TRUE) %>%  
 mutate(steps=ifelse(is.na(steps)==TRUE,avg, steps))%>%  
 select(-avg)

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

daymean2 <- aggregate(steps ~ date, data = df1, sum)  
par(mfrow=c(1,1))  
hist(daymean2$steps,col="green",xlab="# Steps",ylab="Frequency",main="Total number of steps")  
hist(daymean$steps,col="red",xlab="# Steps",ylab="Frequency",add=TRUE)  
legend("topright",c("Imputed","NAs"),col=c("green","red"),lwd=3)



Do these values differ from the estimates from the first part of the assignment? Yes.

What is the impact of imputing missing data on the estimates of the total daily number of steps? Imputing missing data increases the estimates of the total daily number of steps.

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

h <- summary(daymean2$steps)  
h[c("Median","Mean")]

## Median Mean   
## 10770 10770

## 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.

weekend <- c("Friday","Saturday","Sunday")  
activity$typeday <- factor((weekdays(activity$date) %in% weekend),levels=c(FALSE,TRUE),labels=c("Weekday","Weekend"))

1. Make a panel plot containing a time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days.

library(lattice)  
step\_int <- aggregate(steps ~ interval + typeday,data = activity, mean)  
  
xyplot(step\_int$steps ~ step\_int$interval | step\_int$typeday, main = "Mean Steps (daily by interval)",xlab="Interval",ylab="Steps",layout=c(1,2),type="l")

