# Reproducible Research - Course Project 1

## Reading the Dataset

library(knitr)  
library(ggplot2)  
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

library(lattice)  
setwd("E:\\Old\_Doc\\sas doc\\Stat\\JHU\\Reproducible Research\\Week2\\Factivity")  
activity <- read.csv("activity.csv")  
summary(activity)

## 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.: 12.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## NA's :2304 (Other) :15840

str(activity)

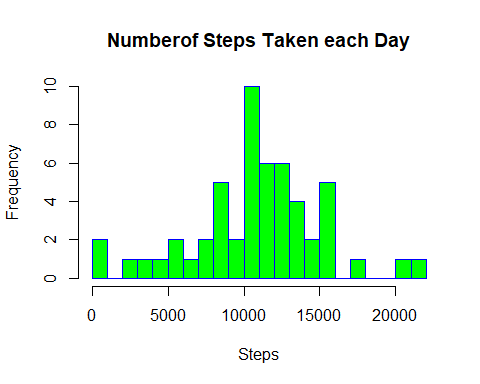
## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : int NA NA NA NA NA NA NA NA NA NA ...  
## $ date : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

activity$date <- as.Date(activity$date, "%Y-%m-%d")  
str(activity)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : int NA NA NA NA NA NA NA NA NA NA ...  
## $ date : Date, format: "2012-10-01" "2012-10-01" ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

## Histogram of the total number of steps taken each day

act <- aggregate(steps ~ date, activity, sum, na.rm = TRUE)  
hist(act$steps, main = "Numberof Steps Taken each Day", xlab = "Steps", border = "blue", col = "green", breaks = 25)



## Mean and median number of steps taken each day

Stepsmean <- mean(act$steps, na.rm = TRUE)  
Stepsmean

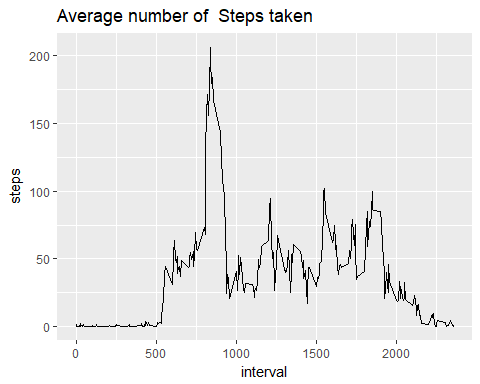
## [1] 10766.19

Stepsmedian <- median(act$steps, na.rm = TRUE)  
Stepsmedian

## [1] 10765

## Time series plot of the average number of steps taken

NotNA <- activity[!is.na(activity$steps),]  
Meantable <- aggregate(steps ~ interval, NotNA, mean)  
ggplot(Meantable, aes(x = interval, y = steps), xlab = "Interval", ylab = "Average Steps") +   
 geom\_line() + ggtitle("Average number of Steps taken")



## The 5-minute interval that, on average, contains the maximum number of step

Maxsteps <- NotNA[which.max(NotNA$steps),]  
Maxsteps

## steps date interval  
## 16492 806 2012-11-27 615

## Code to describe and show a strategy for imputing missing data

sum(is.na(activity$steps))

## [1] 2304

activity2 <- activity  
Missdata <- is.na(activity2$steps)  
avginterval <- tapply(activity2$steps, activity2$interval, mean, na.rm = TRUE, simplify = TRUE)  
activity2$steps[Missdata] <- avginterval[as.character(activity2$interval[Missdata])]  
sum(is.na(activity2$steps))

## [1] 0

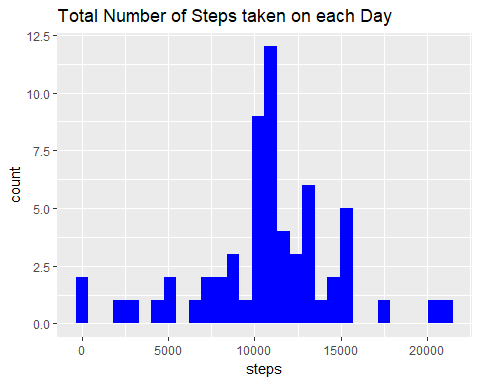
## Histogram of the total number of steps taken each day after missing values are imputed

stepseachday <- aggregate(steps ~ date, activity2, sum)  
stepseachday

## date steps  
## 1 2012-10-01 10766.19  
## 2 2012-10-02 126.00  
## 3 2012-10-03 11352.00  
## 4 2012-10-04 12116.00  
## 5 2012-10-05 13294.00  
## 6 2012-10-06 15420.00  
## 7 2012-10-07 11015.00  
## 8 2012-10-08 10766.19  
## 9 2012-10-09 12811.00  
## 10 2012-10-10 9900.00  
## 11 2012-10-11 10304.00  
## 12 2012-10-12 17382.00  
## 13 2012-10-13 12426.00  
## 14 2012-10-14 15098.00  
## 15 2012-10-15 10139.00  
## 16 2012-10-16 15084.00  
## 17 2012-10-17 13452.00  
## 18 2012-10-18 10056.00  
## 19 2012-10-19 11829.00  
## 20 2012-10-20 10395.00  
## 21 2012-10-21 8821.00  
## 22 2012-10-22 13460.00  
## 23 2012-10-23 8918.00  
## 24 2012-10-24 8355.00  
## 25 2012-10-25 2492.00  
## 26 2012-10-26 6778.00  
## 27 2012-10-27 10119.00  
## 28 2012-10-28 11458.00  
## 29 2012-10-29 5018.00  
## 30 2012-10-30 9819.00  
## 31 2012-10-31 15414.00  
## 32 2012-11-01 10766.19  
## 33 2012-11-02 10600.00  
## 34 2012-11-03 10571.00  
## 35 2012-11-04 10766.19  
## 36 2012-11-05 10439.00  
## 37 2012-11-06 8334.00  
## 38 2012-11-07 12883.00  
## 39 2012-11-08 3219.00  
## 40 2012-11-09 10766.19  
## 41 2012-11-10 10766.19  
## 42 2012-11-11 12608.00  
## 43 2012-11-12 10765.00  
## 44 2012-11-13 7336.00  
## 45 2012-11-14 10766.19  
## 46 2012-11-15 41.00  
## 47 2012-11-16 5441.00  
## 48 2012-11-17 14339.00  
## 49 2012-11-18 15110.00  
## 50 2012-11-19 8841.00  
## 51 2012-11-20 4472.00  
## 52 2012-11-21 12787.00  
## 53 2012-11-22 20427.00  
## 54 2012-11-23 21194.00  
## 55 2012-11-24 14478.00  
## 56 2012-11-25 11834.00  
## 57 2012-11-26 11162.00  
## 58 2012-11-27 13646.00  
## 59 2012-11-28 10183.00  
## 60 2012-11-29 7047.00  
## 61 2012-11-30 10766.19

ggplot(stepseachday, aes(x = steps), xlab = "Date", ylab = "TotalSteps") + geom\_histogram(fill = "blue") +  
 ggtitle("Total Number of Steps taken on each Day")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

activity2 <- mutate(activity2, weektype = ifelse(weekdays(activity2$date) == "Saturday" | weekdays(activity2$date) == "Sunday", "weekend", "weekday"))  
activity2$weektype <- as.factor(activity2$weektype)  
head(activity2)

## steps date interval weektype  
## 1 1.7169811 2012-10-01 0 weekday  
## 2 0.3396226 2012-10-01 5 weekday  
## 3 0.1320755 2012-10-01 10 weekday  
## 4 0.1509434 2012-10-01 15 weekday  
## 5 0.0754717 2012-10-01 20 weekday  
## 6 2.0943396 2012-10-01 25 weekday

tail(activity2)

## steps date interval weektype  
## 17563 2.6037736 2012-11-30 2330 weekday  
## 17564 4.6981132 2012-11-30 2335 weekday  
## 17565 3.3018868 2012-11-30 2340 weekday  
## 17566 0.6415094 2012-11-30 2345 weekday  
## 17567 0.2264151 2012-11-30 2350 weekday  
## 17568 1.0754717 2012-11-30 2355 weekday

intervaldata <- activity2 %>% group\_by(interval, weektype) %>% summarise(steps = mean(steps))  
  
ggplot(intervaldata, aes(x = interval, y = steps, color = weektype)) +   
 geom\_line() + labs(title = "Average number of steps taken across Weekdays and Weekends",   
x = "interval", y = "Steps")

