Reproducible Research-Peer Assessment 1

Charles Njelita

Wednesday, October 15, 2014

Loading and preprocessing the data

```
activity <- read.csv("activity.csv", colClasses = c("numeric", "character",
    "numeric"))
names(activity)
```

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

```
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
```

```
summary(activity)
```

```
##      steps           date              interval
##  Min.   :  0.0   Length:17568       Min.   :   0
##  1st Qu.:  0.0   Class :character   1st Qu.: 589
##  Median :  0.0   Mode  :character   Median :1178
##  Mean   : 37.4                      Mean   :1178
##  3rd Qu.: 12.0                      3rd Qu.:1766
##  Max.   :806.0                      Max.   :2355
##  NA's   :2304
```

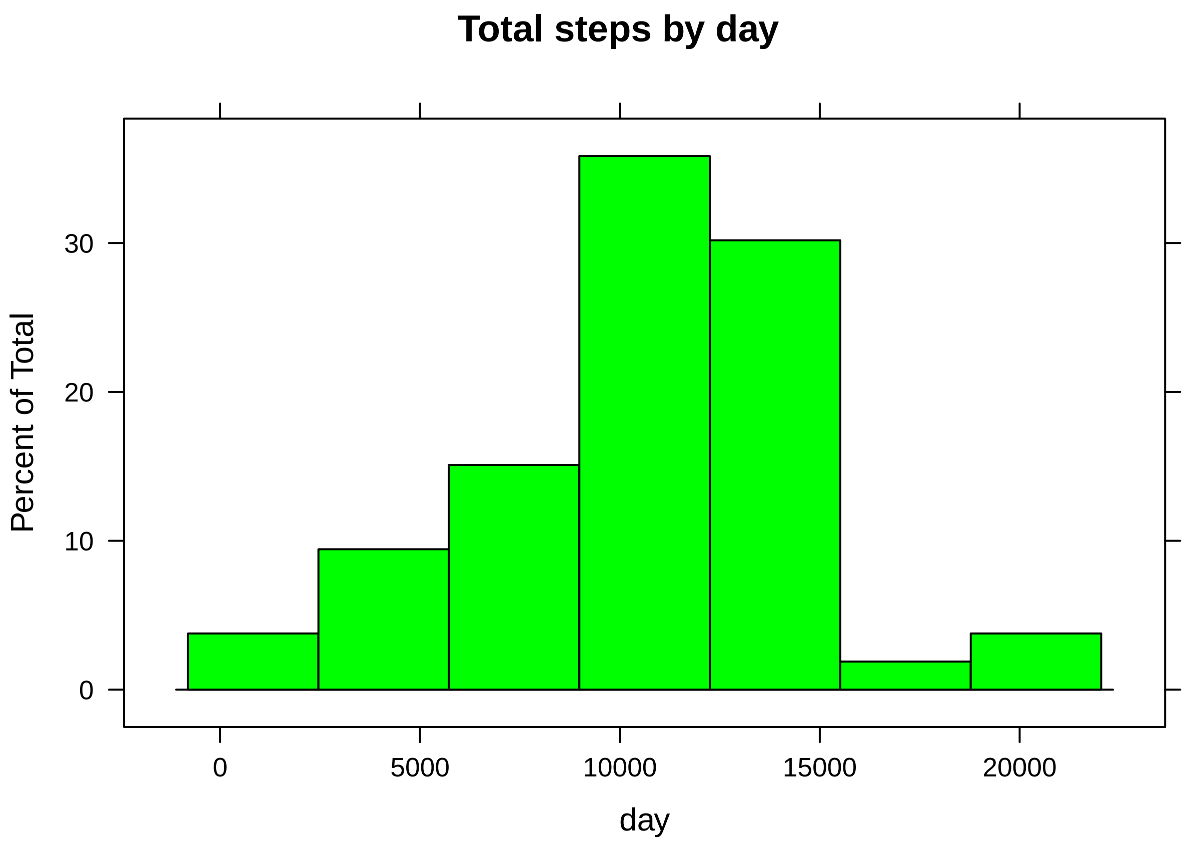
plots the activities:

What is mean total number of steps taken per day?

```
library(ggplot2)
#First is using aggregate function
StepsTotal <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
print(StepsTotal)
```

```
##          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
## 7  2012-10-09 12811
## 8  2012-10-10  9900
## 9  2012-10-11 10304
## 10 2012-10-12 17382
## 11 2012-10-13 12426
## 12 2012-10-14 15098
## 13 2012-10-15 10139
## 14 2012-10-16 15084
## 15 2012-10-17 13452
## 16 2012-10-18 10056
## 17 2012-10-19 11829
## 18 2012-10-20 10395
## 19 2012-10-21  8821
## 20 2012-10-22 13460
## 21 2012-10-23  8918
## 22 2012-10-24  8355
## 23 2012-10-25  2492
## 24 2012-10-26  6778
## 25 2012-10-27 10119
## 26 2012-10-28 11458
## 27 2012-10-29  5018
## 28 2012-10-30  9819
## 29 2012-10-31 15414
## 30 2012-11-02 10600
## 31 2012-11-03 10571
## 32 2012-11-05 10439
## 33 2012-11-06  8334
## 34 2012-11-07 12883
## 35 2012-11-08  3219
## 36 2012-11-11 12608
## 37 2012-11-12 10765
## 38 2012-11-13  7336
## 39 2012-11-15    41
## 40 2012-11-16  5441
## 41 2012-11-17 14339
## 42 2012-11-18 15110
## 43 2012-11-19  8841
## 44 2012-11-20  4472
## 45 2012-11-21 12787
## 46 2012-11-22 20427
## 47 2012-11-23 21194
## 48 2012-11-24 14478
## 49 2012-11-25 11834
## 50 2012-11-26 11162
## 51 2012-11-27 13646
## 52 2012-11-28 10183
## 53 2012-11-29  7047
```

```r
#Second we use histogram
histogram(StepsTotal$steps, main = "Total steps by day", xlab = "day", col = "green")
```



```
# Mean and Median are as follows:
mean(StepsTotal$steps)
```

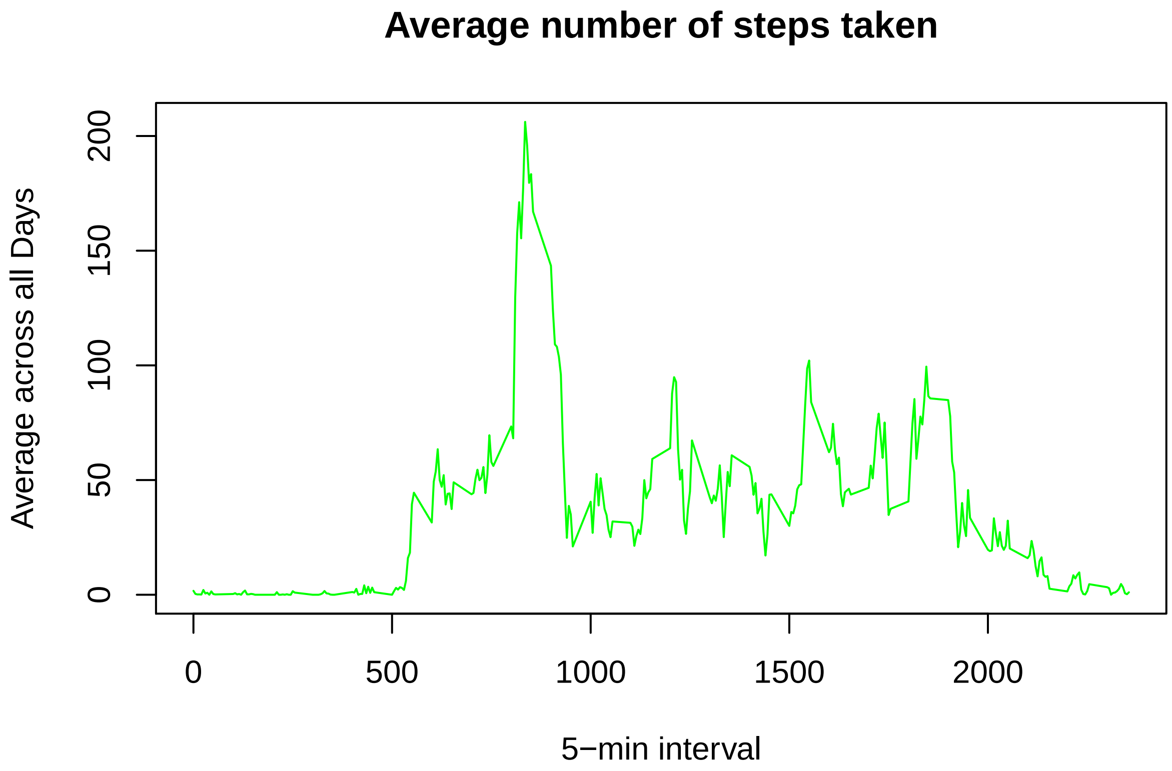
```
## [1] 10766
```

```
median(StepsTotal$steps)
```

```
## [1] 10765
```

What is the average daily activity pattern?

```
time_series <- tapply(activity$steps, activity$interval, mean, na.rm = TRUE)
## We make plot
plot(row.names(time_series), time_series, type = "l", xlab = "5-min interval",
         ylab = "Average across all Days", main = "Average number of steps taken",
    col = "green")
```



Imputing missing values

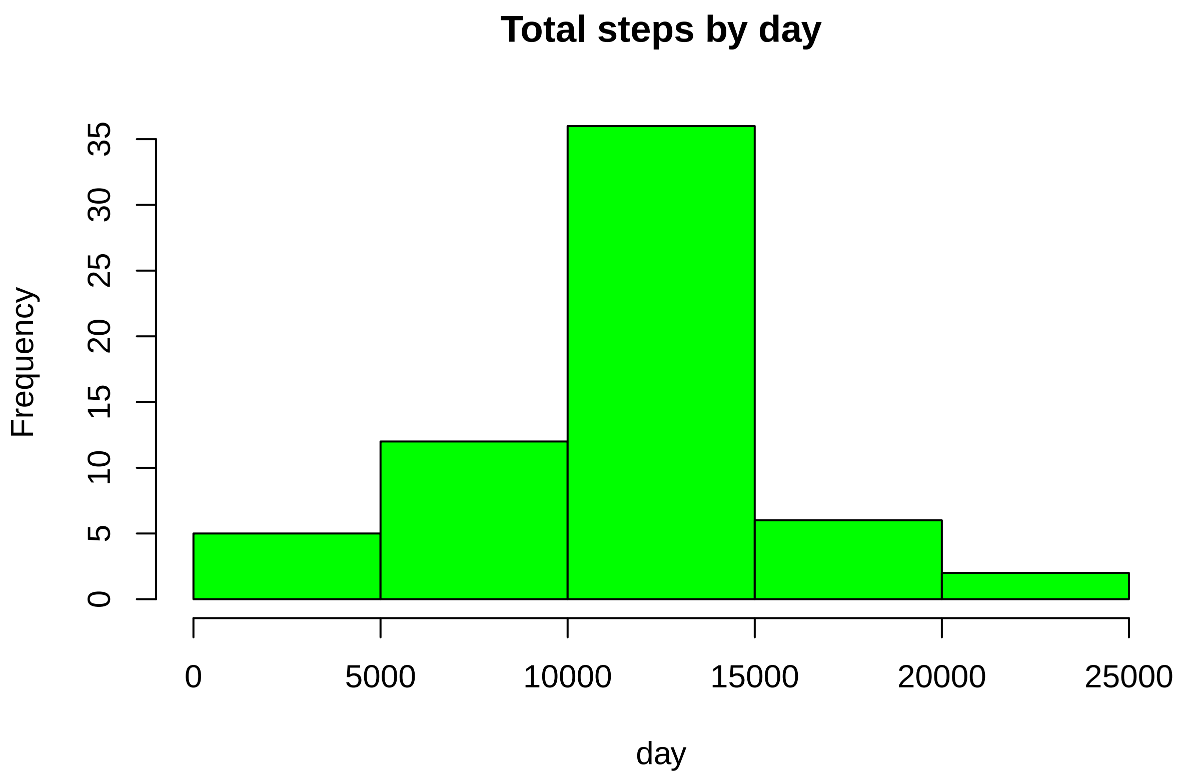
```
activity_NA <- sum(is.na(activity))
print(activity_NA)
```

```
## [1] 2304
```

```
StepsAverage <- aggregate(steps ~ interval, data = activity, FUN = mean)
fillNA <- numeric()
for (i in 1:nrow(activity)) {
    obs <- activity[i, ]
    if (is.na(obs$steps)) {
        steps <- subset(StepsAverage, interval == obs$interval)$steps
    } else {
        steps <- obs$steps
    }
    fillNA <- c(fillNA, steps)
}

# We create a new dataset that is equal to the original dataset but with the missing data filled in.
new_activity <- activity
new_activity$steps <- fillNA
```

```
# Make a histogram of the total number of steps taken each day and Calculate and report.
StepsTotal2 <- aggregate(steps ~ date, data = new_activity, sum, na.rm = TRUE)
hist(StepsTotal2$steps, main = "Total steps by day", xlab = "day", col = "green")
```



```
#the mean and median are as follows:
mean(StepsTotal2$steps)
```

```
## [1] 10766
```

```
median(StepsTotal2$steps)
```

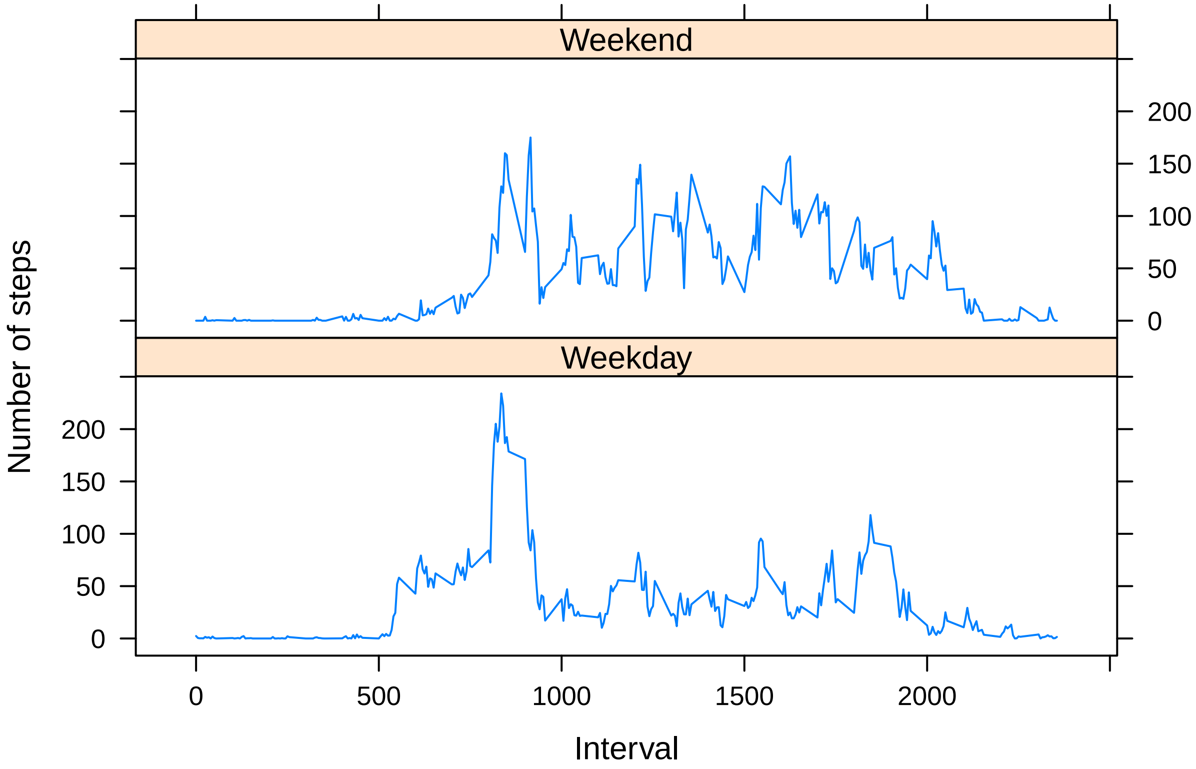
```
## [1] 10766
```

Are there differences in activity patterns between weekdays and weekends?

```
day <- weekdays(activity$date)
daylevel <- vector()
for (i in 1:nrow(activity)) {
    if (day[i] == "Saturday") {
        daylevel[i] <- "Weekend"
    } else if (day[i] == "Sunday") {
        daylevel[i] <- "Weekend"
    } else {
        daylevel[i] <- "Weekday"
    }
}
activity$daylevel <- daylevel
activity$daylevel <- factor(activity$daylevel)

stepsByDay <- aggregate(steps ~ interval + daylevel, data = activity, mean)
names(stepsByDay) <- c("interval", "daylevel", "steps")
```

```
xyplot(steps ~ interval | daylevel, stepsByDay, type = "l", layout = c(1, 2),
    xlab = "Interval", ylab = "Number of steps")
```


Weekend
Weekday
Number of steps
Interval