Reproducible Research Assignment One

DM Rochman

December 20, 2015

Loading and Preprocessing the Data

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

```
names(activity)
```

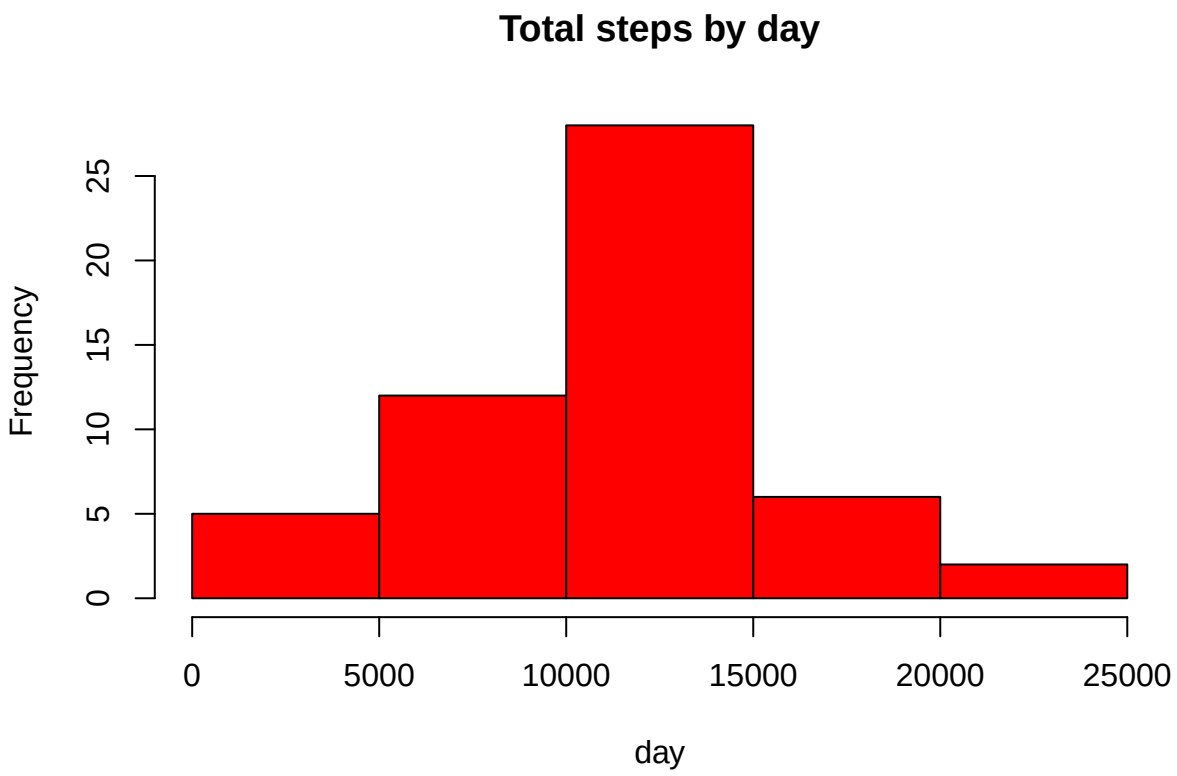
```
## [1] "steps"    "date"     "interval"
```

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

What is the mean total number of steps taken?

I first found the mean and median of steps taken, I then created a dataframe so that I could run plots on the data.

```
StepsTotal <- aggregate(steps ~ date, data = activity, sum, na.rm =
                            TRUE)
hist(StepsTotal$steps, main = "Total steps by day", xlab = "day",
     col = "red")
```




```
mean(StepsTotal$steps)
```

```
## [1] 10766.19
```

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

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

```
steps <- rep(NA, 61)
day <- rep("NA", 61)
stepsday <- tapply(activity$steps, activity$date, sum, na.rm = T)
length(stepsday)
```

```
## [1] 61
```

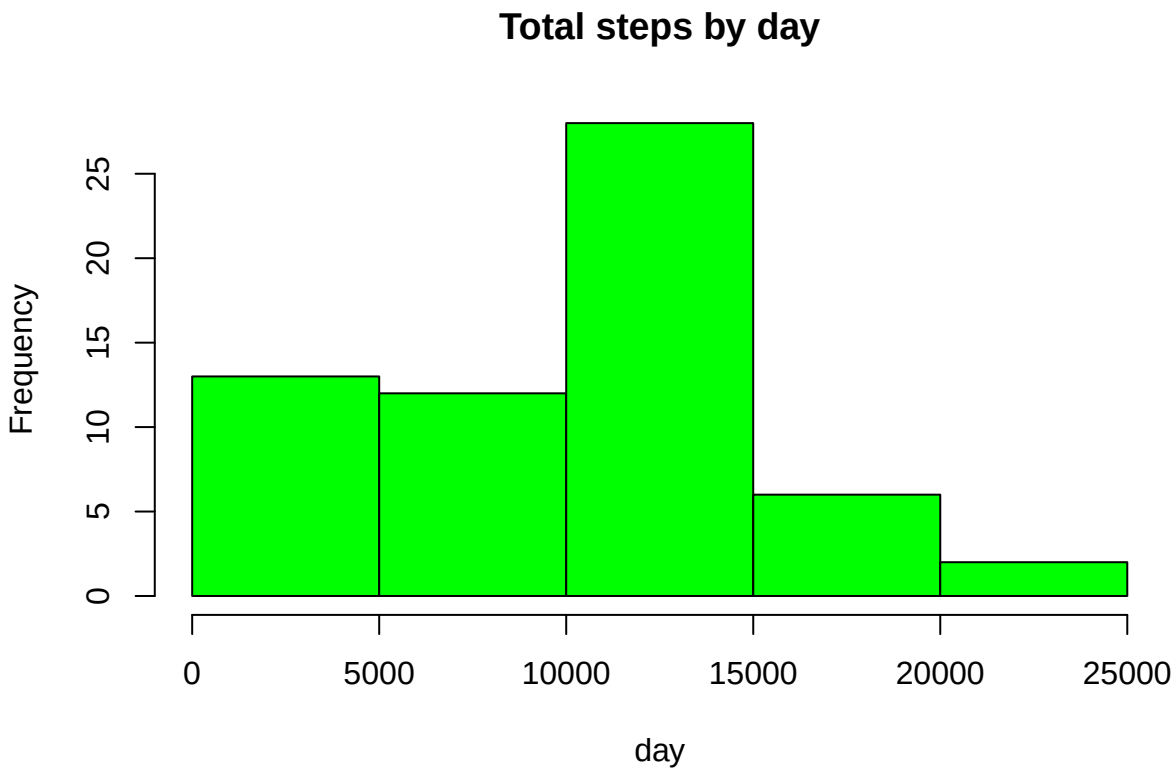
```
for (i in 1:61) {
     steps[i] <- stepsday[[i]]
     day[i] <- names(stepsday)[i]
}

df <- data.frame(day, steps)
head(df)
```

```
##          day steps
## 1 2012-10-01     0
## 2 2012-10-02   126
```

```
## 3 2012-10-03 11352
## 4 2012-10-04 12116
## 5 2012-10-05 13294
## 6 2012-10-06 15420
```

```
hist(df$steps, main = "Total steps by day", xlab = "day", col =
          "green")
```

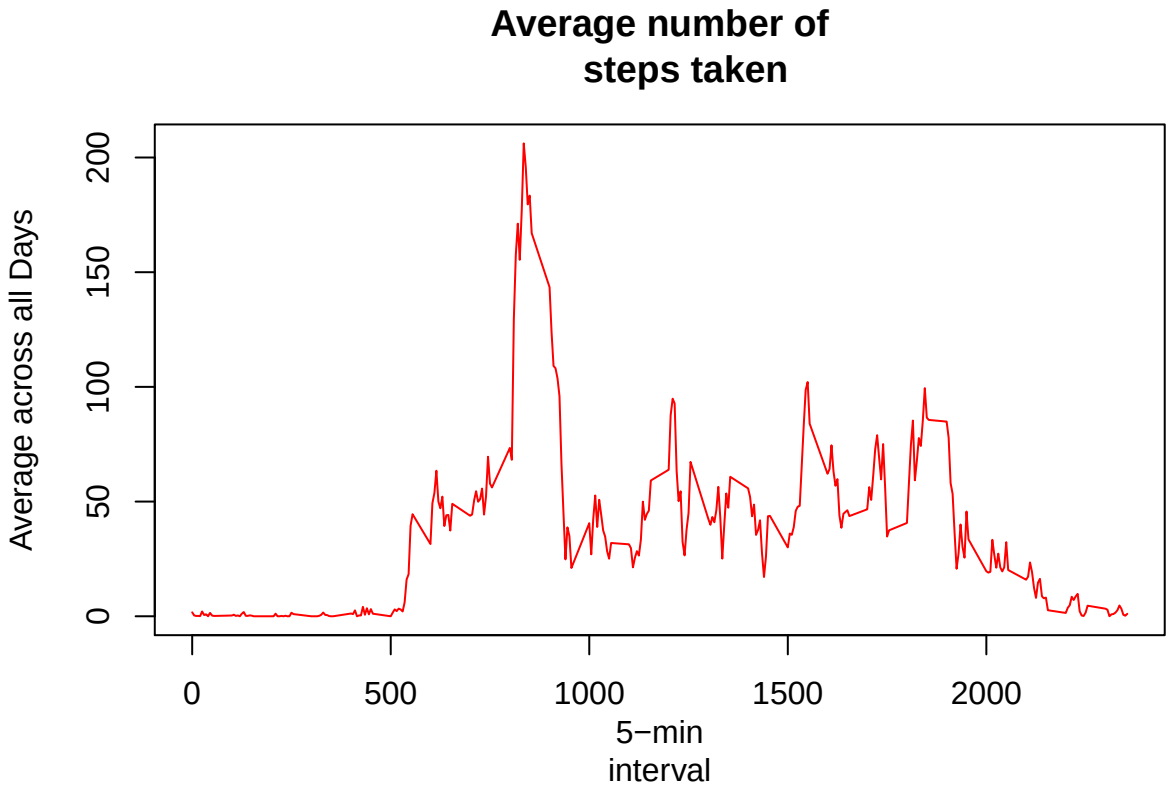


What is the daily average activity pattern?

I created a time series plot and used 5 minute intervals as the X-axis. I then created a mean steps time series plot to find the maximum number of steps

```
time_series <- tapply(activity$steps, activity$interval, mean,
                      na.rm = TRUE)

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 = "red")
```




```
max_interval <- which.max(time_series)
names(max_interval)
```

```
## [1] "835"
```

Imputing Missing Values

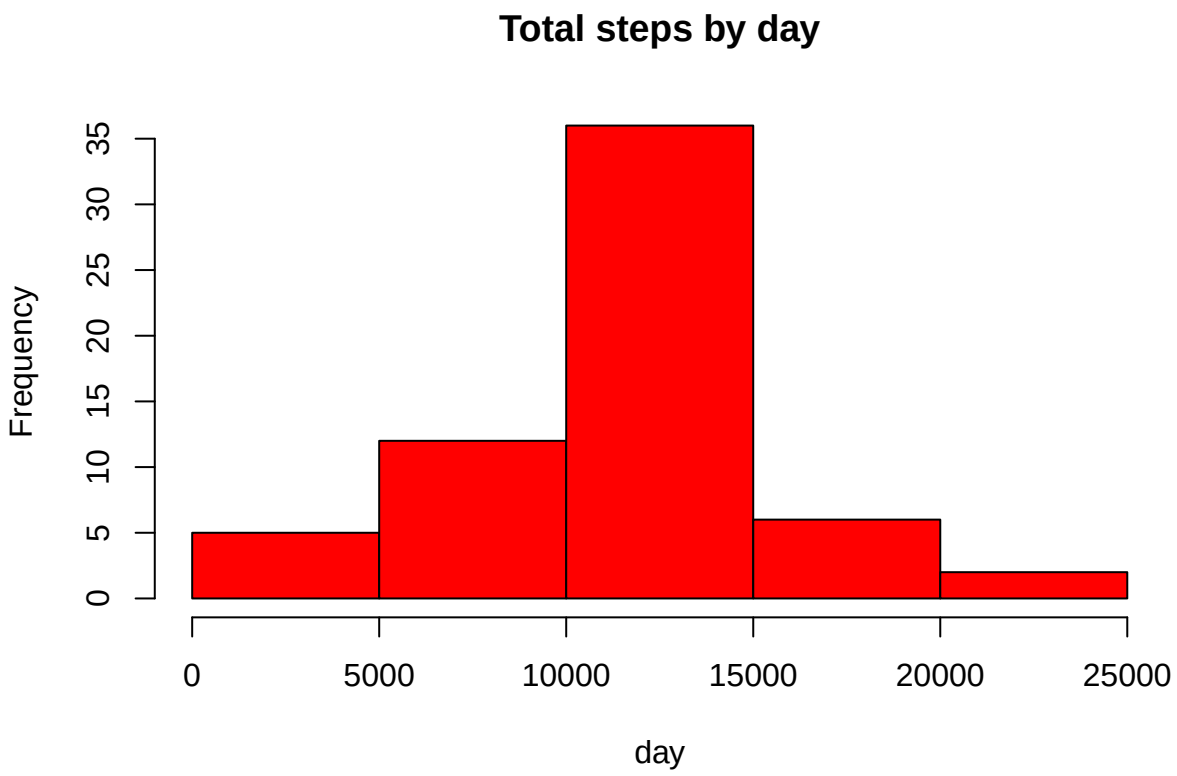
```
activity_NA <- sum(is.na(activity))
activity_NA
```

```
## [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)
}
```

```
new_activity <- activity
new_activity$steps <- fillNA

StepsTotal2 <- aggregate(steps ~ date, data = new_activity, sum,
                         na.rm = TRUE)
hist(StepsTotal2$steps, main = "Total steps by day", xlab = "day",
     col = "red")
```




```
mean(StepsTotal2$steps)
```

```
## [1] 10766.19
```

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

```
## [1] 10766.19
```

Are there differences in activity patterns between weekdays and weekends?

On average there is more overall activity on the 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")
```