Reproducible Research: Peer Assessment 1

Deborah Passey

6/15/2019

Peer Assignment 1 - RMarkdown File

Loading and preprocessing the data

### Loading Data

```
        library("data.table")
        path <- getwd()
        download.file(url = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zi
p", destfile = paste(path, "dataFiles.zip",sep = "/"))
        unzip(zipfile = "dataFiles.zip")
        activity <- read.csv("activity.csv")
```

### Preprocessing the Data

```
    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
```

```
    summary(activity$steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    0.00   37.38   12.00  806.00    2304
```

```
    summary(activity$interval)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   588.8  1177.5  1177.5  1766.2  2355.0
```

```
    library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

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

```
    day <- function(x) format(as.Date(x), "%A", na.rm=TRUE)
    activity$day <- day(activity$date)
```

What is mean total number of steps taken per day?

```
    library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

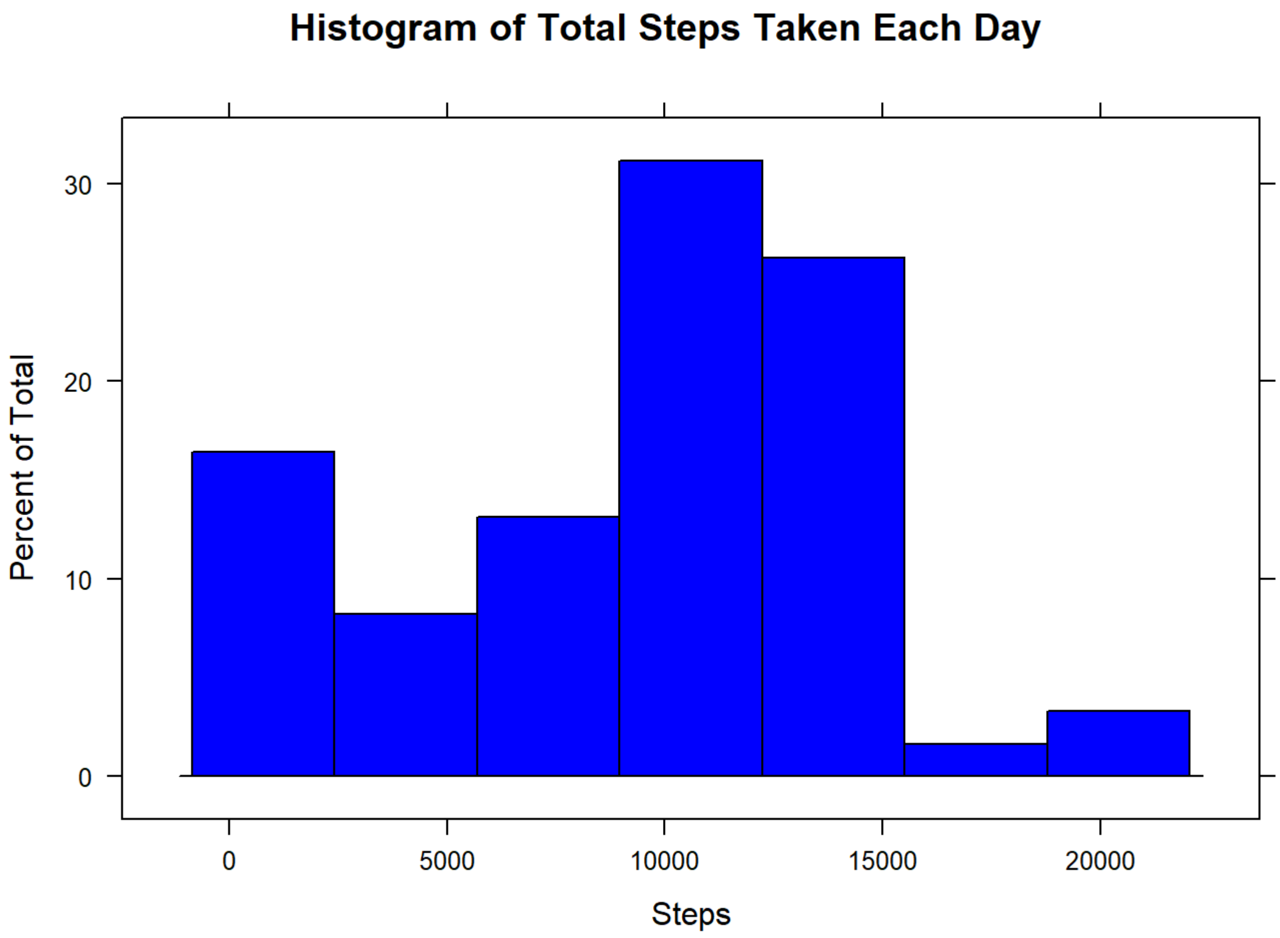
```
    totalsteps <- summarise(group_by(activity, date), totalsteps = sum(steps, na.rm=TRUE))
    print(totalsteps)
```

```
## # A tibble: 61 x 2
##    date       totalsteps
##    <fct>           <int>
##  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
##  7 2012-10-07      11015
##  8 2012-10-08          0
##  9 2012-10-09      12811
## 10 2012-10-10       9900
## # ... with 51 more rows
```

```
    meansteps <- as.integer(mean(totalsteps$totalsteps), na.rm=TRUE)
    mediansteps <- as.integer(median(totalsteps$totalsteps))
```

### Histogram of The Total Steps Taken Each Day

```
    library(lattice)
    histogram(~totalsteps, data=totalsteps, main="Histogram of Total Steps Taken Each Day", col=
"blue", xlab="Steps")
```



### Mean Steps Taken Each Day

```
    print(meansteps)
```

```
## [1] 9354
```

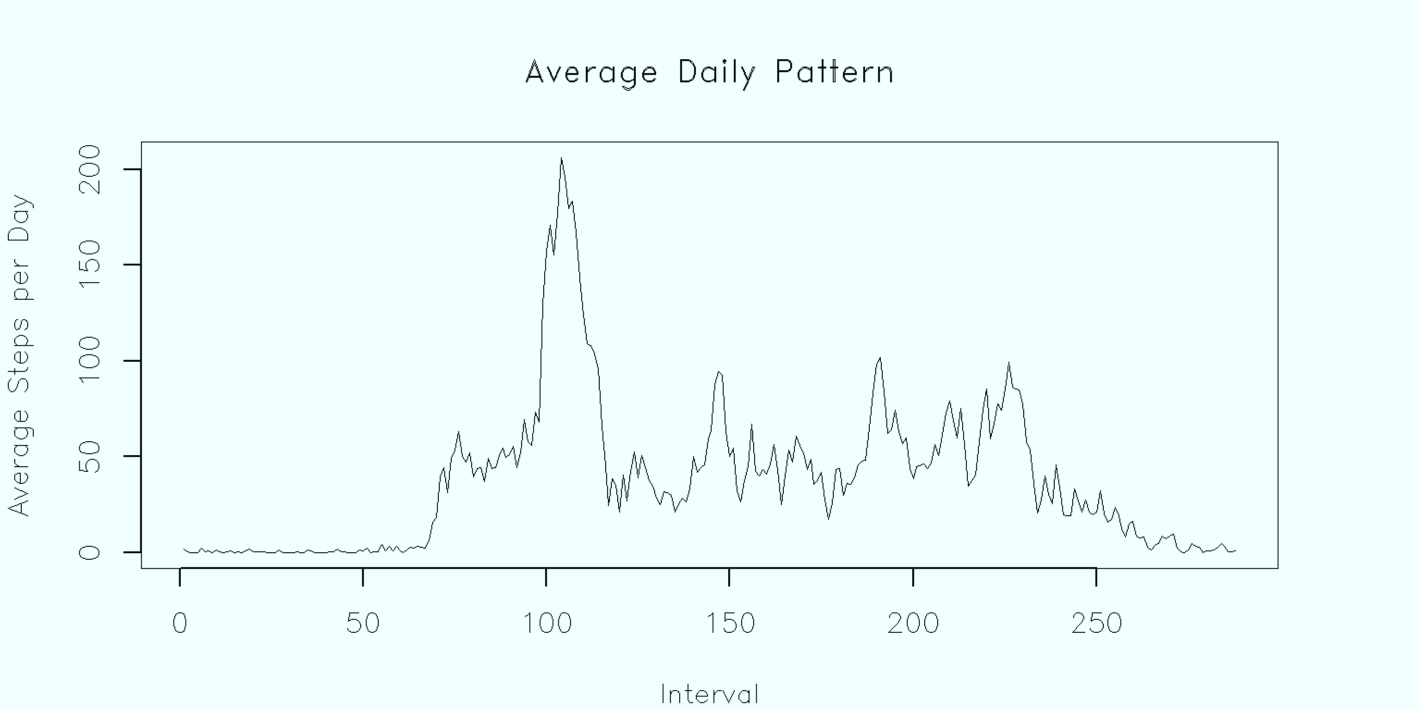
### Median Steps Taken Each Day

```
    print(mediansteps)
```

```
## [1] 10395
```

What is the average daily activity pattern?

```
library(dplyr)
    intervals <- summarise(group_by(activity, interval), meansteps = mean(steps, na.rm=TRUE))
      par(mar=c(4,4,4,4), bg="azure", family="HersheySans", lwd=0.25)
    with(intervals, plot(meansteps, type="l", xlab="Interval", ylab="Average Steps per Day", mai
n="Average Daily Pattern"))
```



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

```
      library(dplyr)
      top_n(intervals, 1, meansteps)
```

```
## # A tibble: 1 x 2
##   interval meansteps
##      <int>     <dbl>
## 1      835      206.
```

The 835th interval has the maximum number of steps.

Imputing Missing Values

### Total Number of Missing Variables

```
    library(dplyr)
    missing <- activity %>%
                filter(is.na(steps))
    table(missing)
```

```
## < table of extent 0 x 61 x 288 x 6 >
```

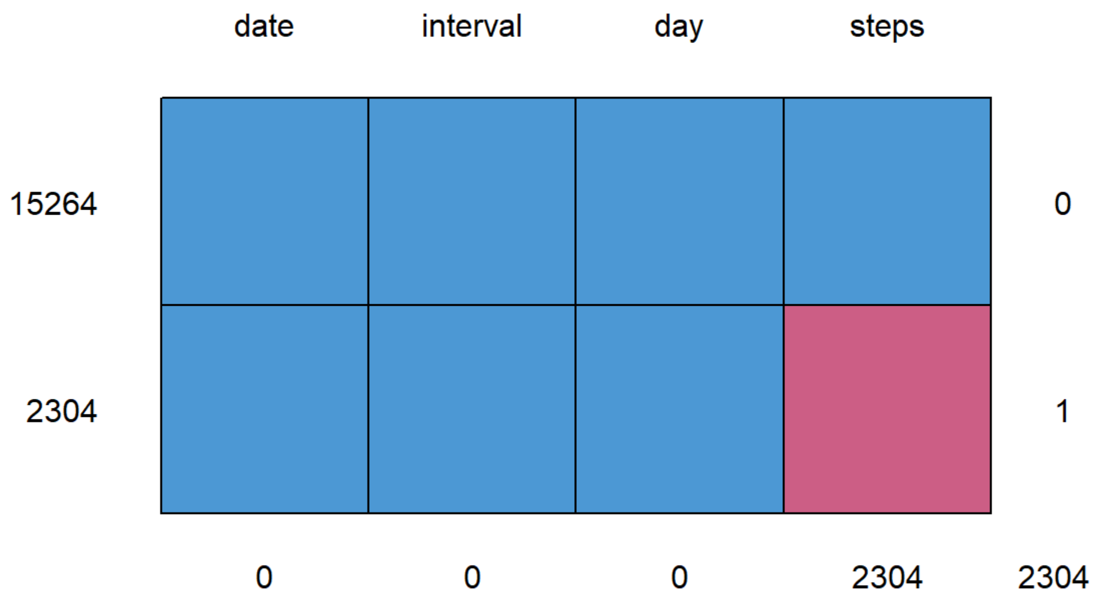
### Looking at Missingness Pattern

```
    library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
    md.pattern(activity)
```



```
##       date interval day steps
## 15264    1        1   1     1    0
## 2304     1        1   1     0    1
##          0        0   0  2304 2304
```

### Imputing Missing Values Using Predictive Mean Matching

```
    imputed_data <- mice(activity, method="mean")
```

```
## 
##  iter imp variable
##   1   1  steps
##   1   2  steps
##   1   3  steps
##   1   4  steps
##   1   5  steps
##   2   1  steps
##   2   2  steps
##   2   3  steps
##   2   4  steps
##   2   5  steps
##   3   1  steps
##   3   2  steps
##   3   3  steps
##   3   4  steps
##   3   5  steps
##   4   1  steps
##   4   2  steps
##   4   3  steps
##   4   4  steps
##   4   5  steps
##   5   1  steps
##   5   2  steps
##   5   3  steps
##   5   4  steps
##   5   5  steps
```

```
## Warning: Number of logged events: 26
```

```
    completedData <- complete(imputed_data,1)
    summary(completedData)
```

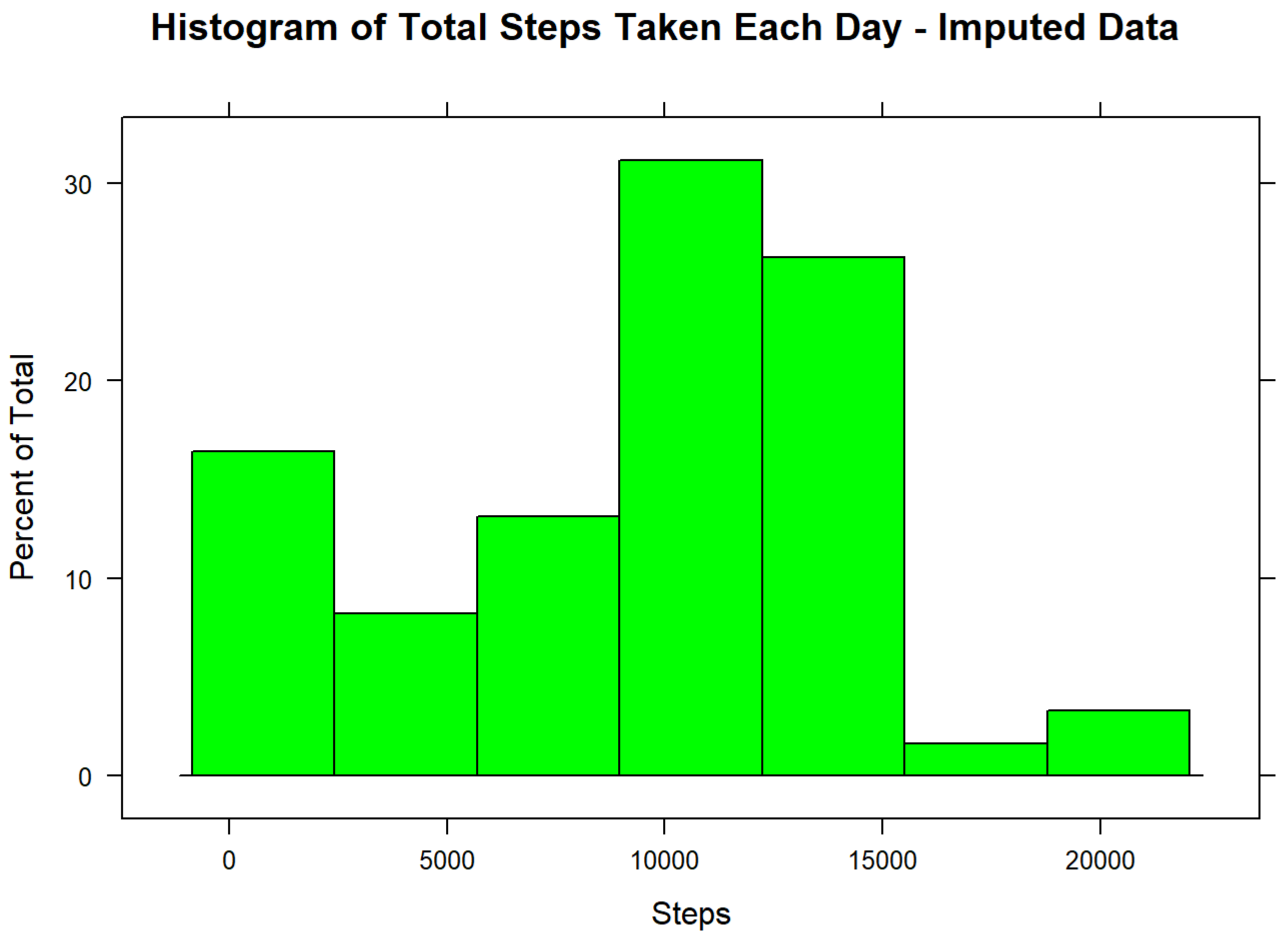
```
##      steps                 date          interval          day           
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0   Length:17568      
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8   Class :character  
##  Median :  0.00   2012-10-03:  288   Median :1177.5   Mode  :character  
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5                     
##  3rd Qu.: 37.38   2012-10-05:  288   3rd Qu.:1766.2                     
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0                     
##                   (Other)   :15840
```

```
    summary(activity)
```

```
##      steps                  date          interval          day           
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0   Length:17568      
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8   Class :character  
##  Median :  0.00   2012-10-03:  288   Median :1177.5   Mode  :character  
##  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                                      
```

### Histogram of Imputed Data

```
    library(dplyr)
    intervals <- summarise(group_by(completedData, interval), meansteps = mean(steps, na.rm=TRUE
))
       library(lattice)
    histogram(~totalsteps, data=totalsteps, main="Histogram of Total Steps Taken Each Day - Impu
ted Data", col="green", xlab="Steps")
```



### Mean and Median Steps of Imputed Data

```
     library(dplyr)
    totalsteps2 <- summarise(group_by(completedData, date), totalsteps = sum(steps, na.rm=TRUE))
    print(totalsteps)
```

```
## # A tibble: 61 x 2
##    date       totalsteps
##    <fct>           <int>
##  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
##  7 2012-10-07      11015
##  8 2012-10-08          0
##  9 2012-10-09      12811
## 10 2012-10-10       9900
## # ... with 51 more rows
```

```
    meansteps2 <- as.integer(mean(totalsteps2$totalsteps), na.rm=TRUE)
    mediansteps2 <- as.integer(median(totalsteps2$totalsteps))

    print(meansteps2)
```

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

```
    print(mediansteps2)
```

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

### 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?

Overall, the imputed data made the mean and median equal. in the original data set the mean was lower than the median.

Are there differences in activity patterns between weekdays and weekends?

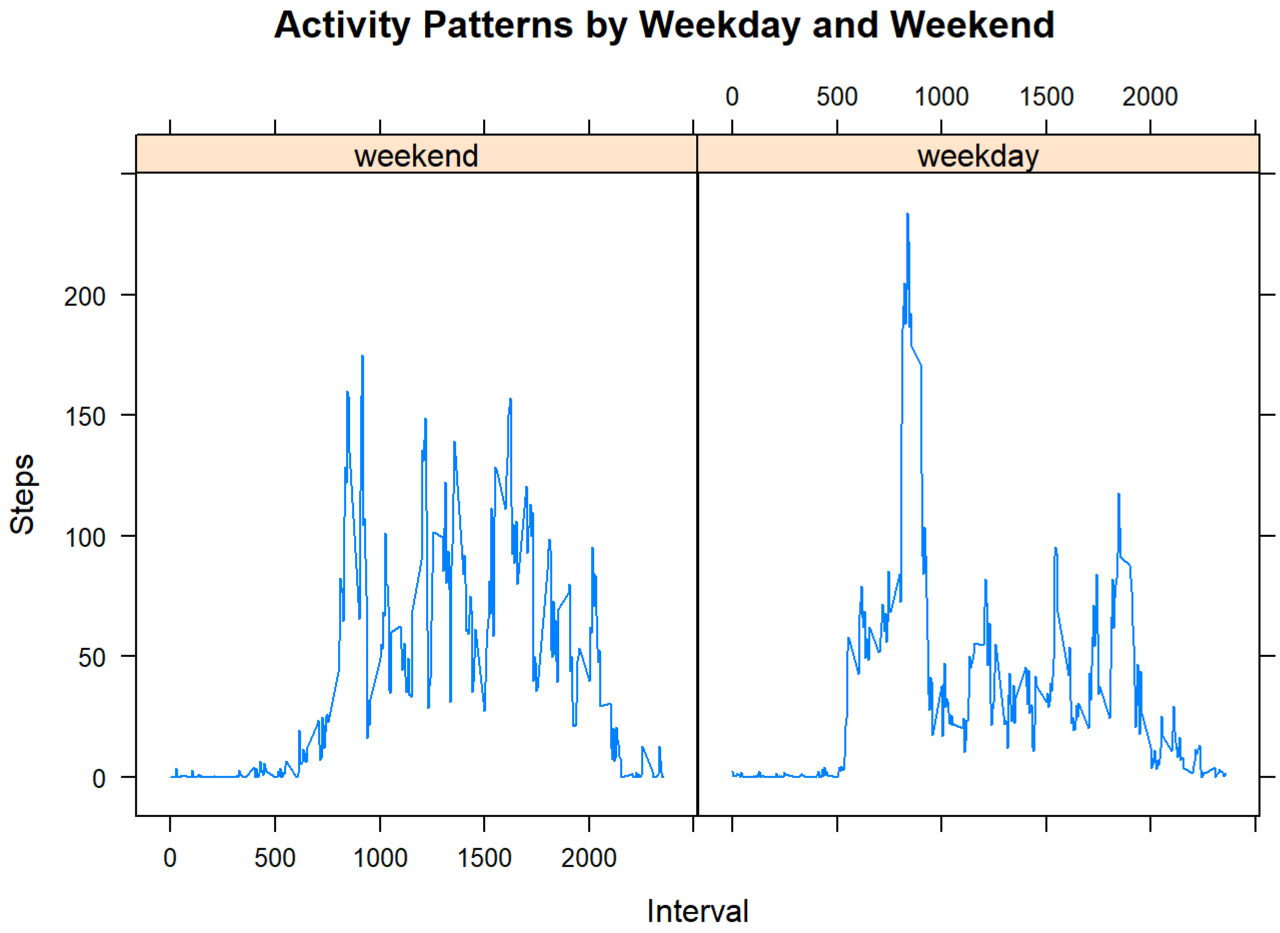
```
### Create "weekday" and "weekend" variables
```

```
    activity$date <- as.Date(activity$date)
    wkdays <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')
    activity$weekday <- factor((weekdays(activity$date) %in% wkdays),
                       levels=c(FALSE, TRUE), labels=c('weekend', 'weekday'))
    summary(activity$weekday)
```

```
## weekend weekday
##    4608   12960
```

```
### Create Panel Plot of Activity Patterns by Weekday and Weekend
```

```
  library(dplyr)
    intervals2 <- summarise(group_by(activity, interval, weekday), meansteps = mean(steps, na.rm
=TRUE))
    library(lattice)
    xyplot(meansteps ~ interval | weekday, data = intervals2, type="l", xlab="Interval", ylab="S
teps", main="Activity Patterns by Weekday and Weekend")
```



There are differences in the activity patterns when looking at the weekday and weekend plots.