Reproducible Research: Peer Assessment 1

Frank DU

9/4/2020

Loading and preprocessing the data

```
library(ggplot2)
activity <- read.csv('activity.csv')
activity$date <- as.POSIXct(activity$date, '%Y-%m-%d')
weekday <- weekdays(activity$date)
activity <- cbind(activity, weekday)

summary(activity)
```

```
##      steps             date                interval         weekday
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0   Length:17568
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8   Class :character
##  Median :  0.00   Median :2012-10-31   Median :1177.5   Mode  :character
##  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
```

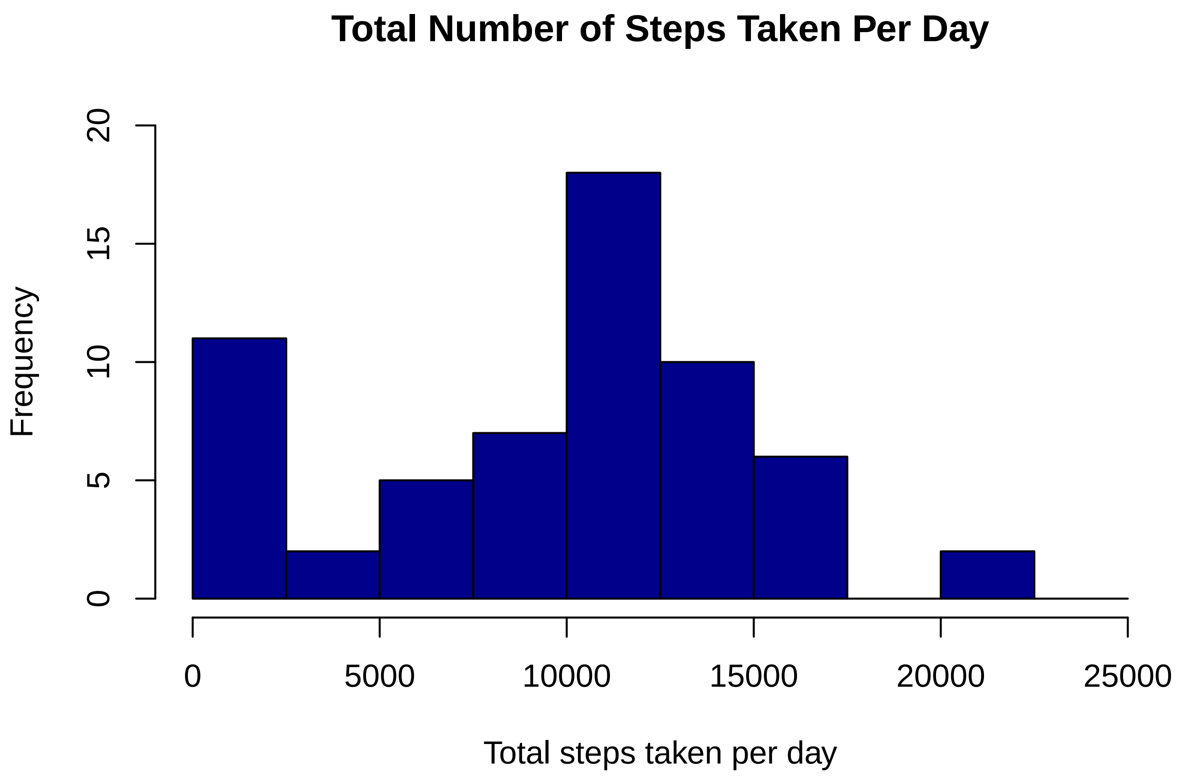
What is mean total number of steps taken per day?

Processing the data.

```
total_steps <- with(activity, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))
names(total_steps) <- c('date', 'steps')
```

Here is the histogram of the total number of steps taken each day.

```
hist(total_steps$steps, main = 'Total Number of Steps Taken Per Day', xlab = 'Total steps taken per day
```



Here is the mean of total number of steps taken per day.

```
mean(total_steps$steps)
```

```
## [1] 9354.23
```

Here is the median of the total number of steps taken per day.

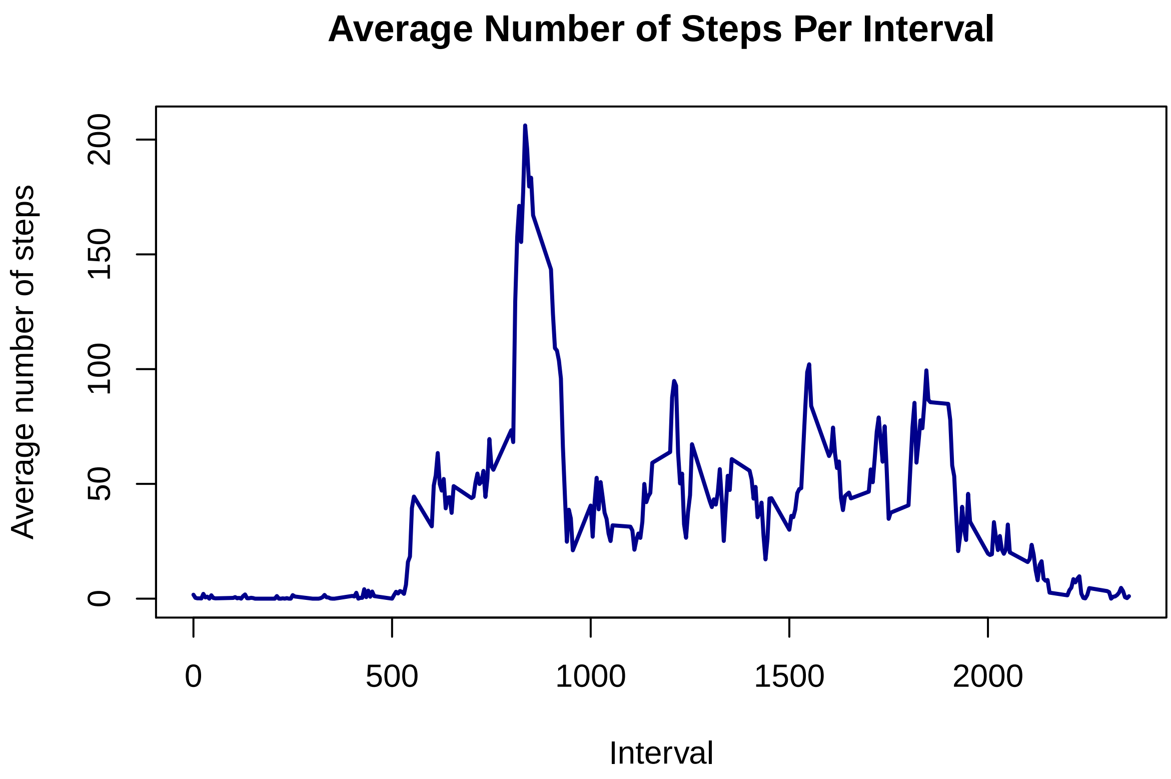
```
median(total_steps$steps)
```

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

What is the average daily activity pattern?

Make a time series plot (i.e. `type = "l"`type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
daily_average <- aggregate(activity$steps, by = list(activity$interval), FUN = mean, na.rm = TRUE)
names(daily_average) <- c('interval', 'mean')
plot(daily_average$interval, daily_average$mean, type = 'l', col = 'darkblue', lwd = 2, xlab = 'Interva
```



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

```
daily_average[which.max(daily_average$mean),]$interval
```

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

Imputing missing values

Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with `NA`NAs)

```
sum(is.na(activity$steps))
```

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

Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

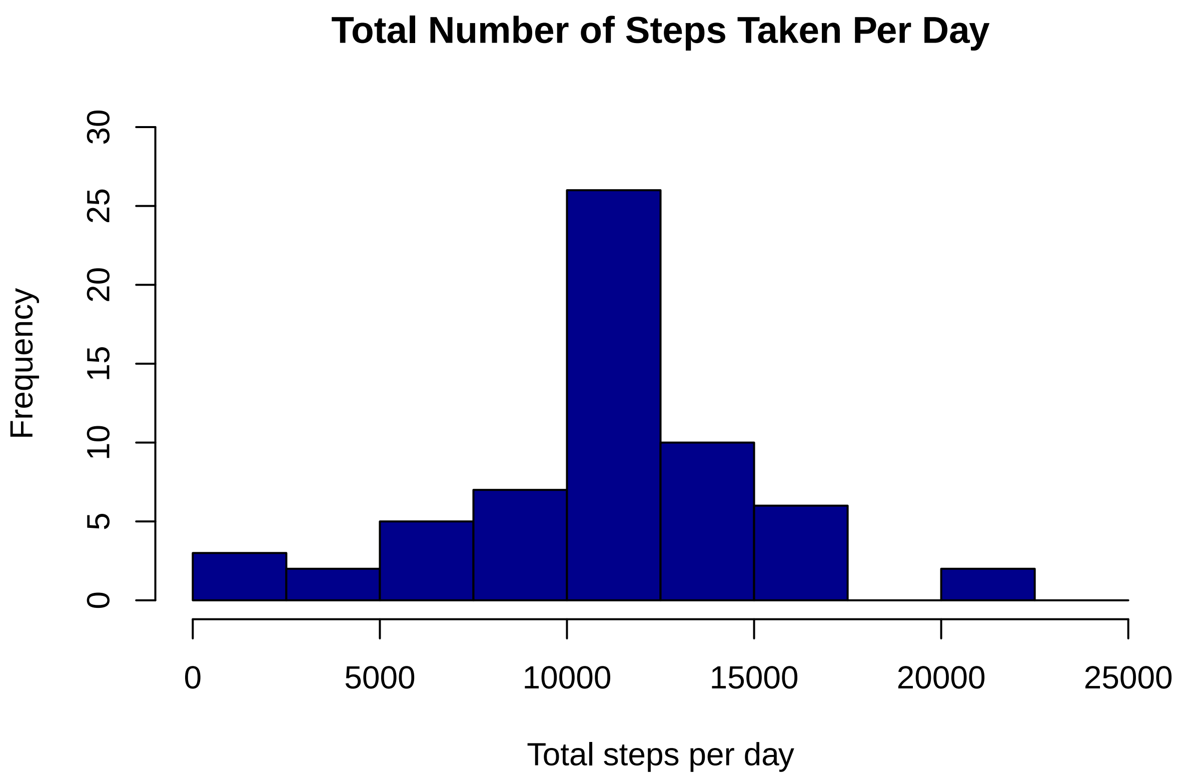
```
imputed_data <- daily_average$mean[match(daily_average$interval, activity$interval)]
```

Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
activity_imputed <- transform(activity, steps = ifelse(is.na(activity$steps), yes = imputed_data, no =
```

Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. 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?

```
total_steps_imputed <- aggregate(steps ~ date, activity_imputed, sum)
names(total_steps_imputed) <- c('date', 'daily_steps')
hist(total_steps_imputed$daily_steps, col = "darkblue", xlab = "Total steps per day", ylim = c(0,30), ma
```



Here is the mean of the total number of steps taken per day.

```
mean(total_steps_imputed$daily_steps)
```

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

Here is the median of the total number of steps taken per day.

```
median(total_steps_imputed$daily_steps)
```

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

The values of mean and median are relatively closed to the estimation in the first part of this assignment. The imputation of this values would increase the mean value.

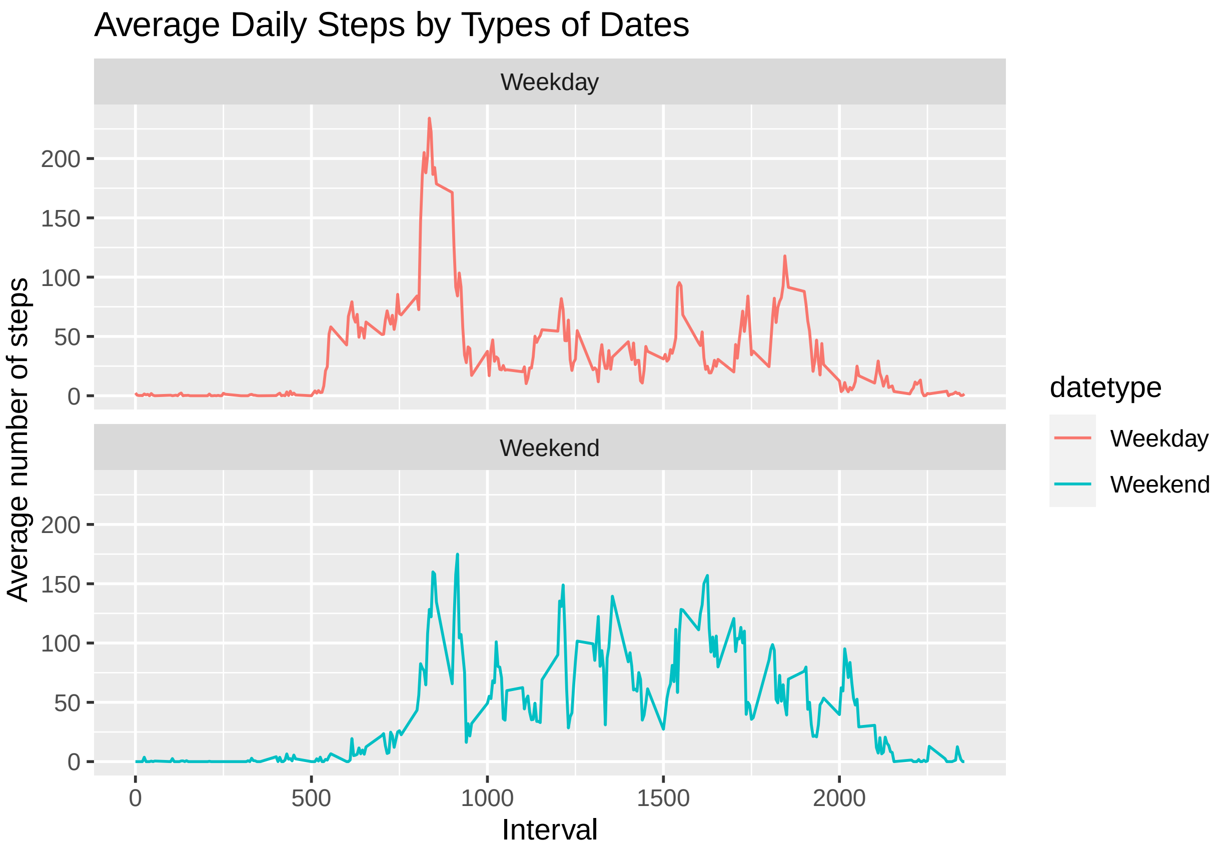
Are there differences in activity patterns between weekdays and weekends?

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.

```
activity$date <- as.Date(strptime(activity$date, format="%Y-%m-%d"))
activity$datetype <- sapply(activity$date, function(x) {
        if (weekdays(x) == "Saturday" | weekdays(x) =="Sunday")
                {y <- "Weekend"} else
                {y <- "Weekday"}
                y
        })
```

Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
activity_by_date <- aggregate(steps~interval + datetype, activity, mean, na.rm = TRUE)
plot<- ggplot(activity_by_date, aes(x = interval , y = steps, color = datetype)) +
       geom_line() +
       labs(title = "Average Daily Steps by Types of Dates", x = "Interval", y = "Average number of step
       facet_wrap(~ datetype, ncol = 1, nrow=2)
print(plot)
```


Average Daily Steps by Types of Dates
Weekday
Weekend
Average number of steps
Interval
datetype
Weekday
Weekend