Reproducible Research: Peer Assessment 1

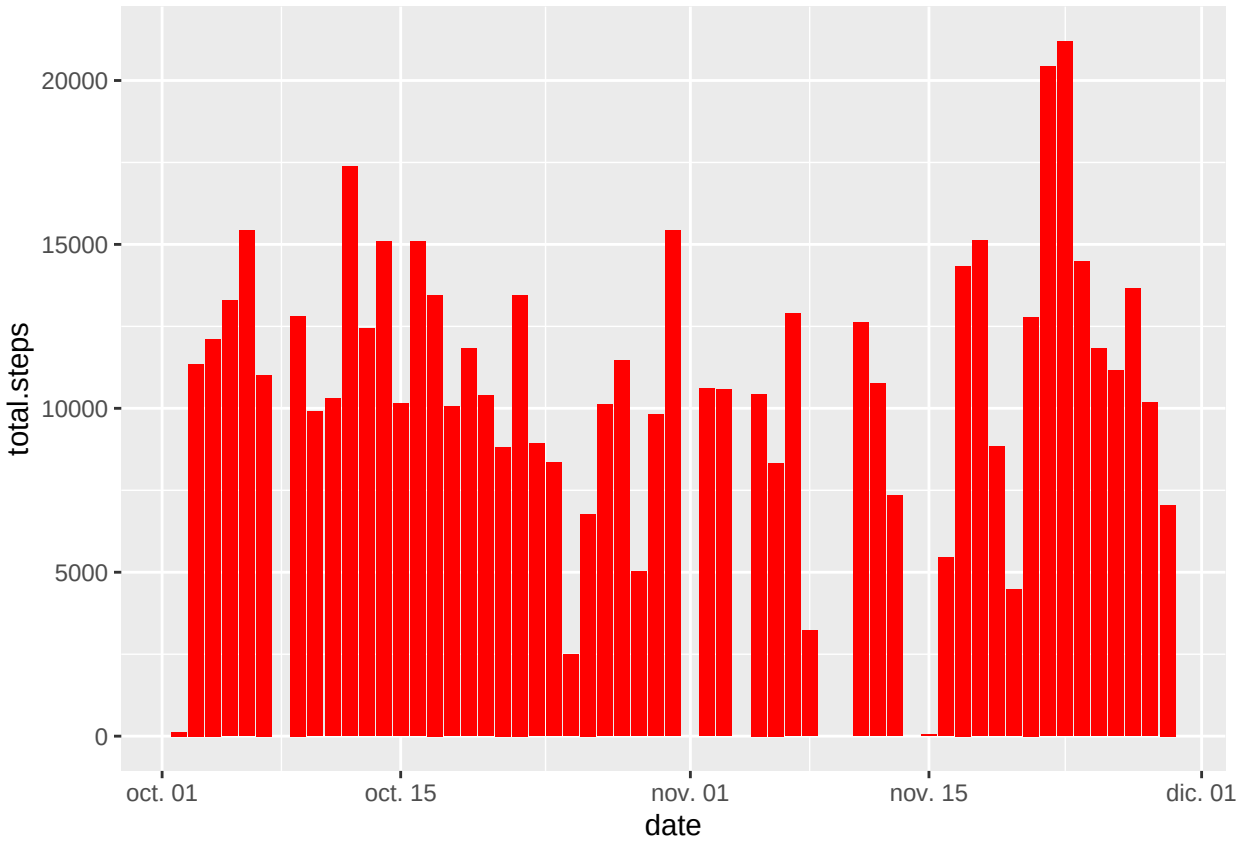
Loading and preprocessing the data

```
library(dplyr)
library(ggplot2)
library(lubridate)
unzip(zipfile = "activity.zip", exdir = "./data")
activity <- read.csv("./data/activity.csv", header = TRUE)
activity$date <- as.Date(activity$date)
```

What is mean total number of steps taken per day?

1) histogram of the total number of steps taken each day

```
act1 <- activity %>% group_by(date) %>% summarise(total.steps=sum(steps))
ggplot(act1, aes(x=date, y=total.steps)) + geom_histogram(stat="identity", fill="red")
```



2) Mean and Median for of the total number of steps taken per day

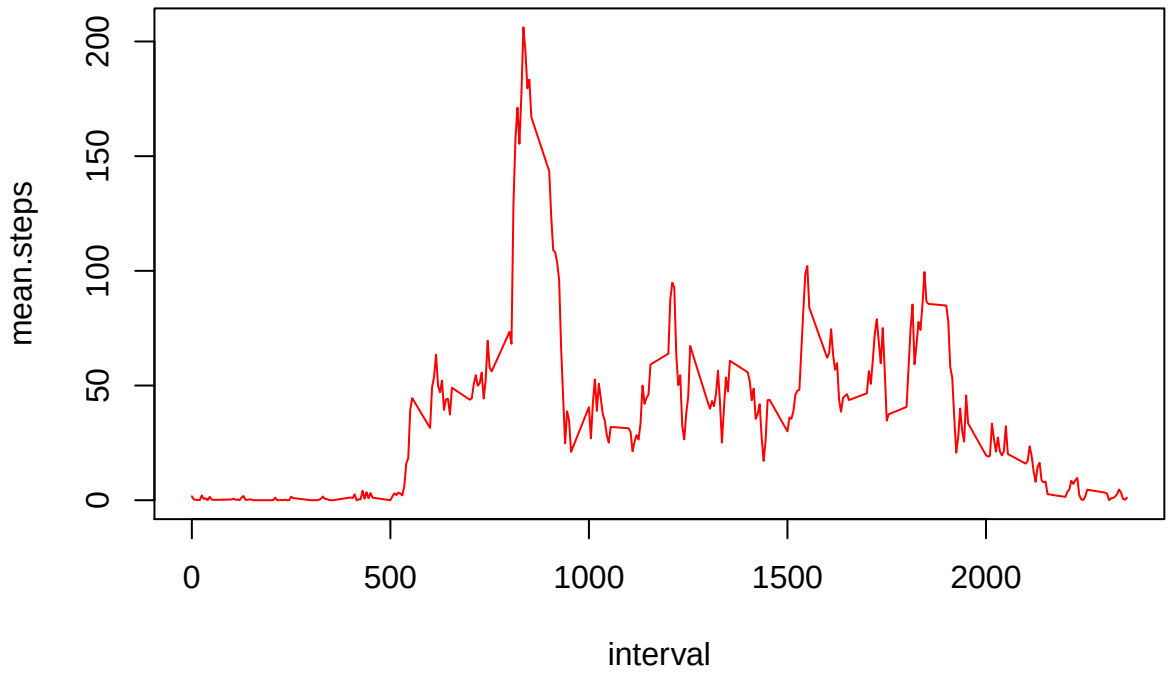
```
tmp <- act1 %>% summarise(mean.steps = mean(total.steps, na.rm = TRUE), median.steps = median(total.ste
tmp
```

```
## # A tibble: 1 x 2
##   mean.steps median.steps
##        <dbl>        <int>
## 1     10766.        10765
```

What is the average daily activity pattern?

1) a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days

```
act2 <- activity %>% group_by(interval) %>% summarise(mean.steps=mean(steps, na.rm=TRUE))
with(act2, plot(interval, mean.steps, type = "l", col = "red"))
```



2) five-minute interval, contains the maximum number of steps

```
head(act2 %>% arrange(desc(mean.steps)), 1)
```

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

Imputing missing values

1) total number of missing values in the dataset

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

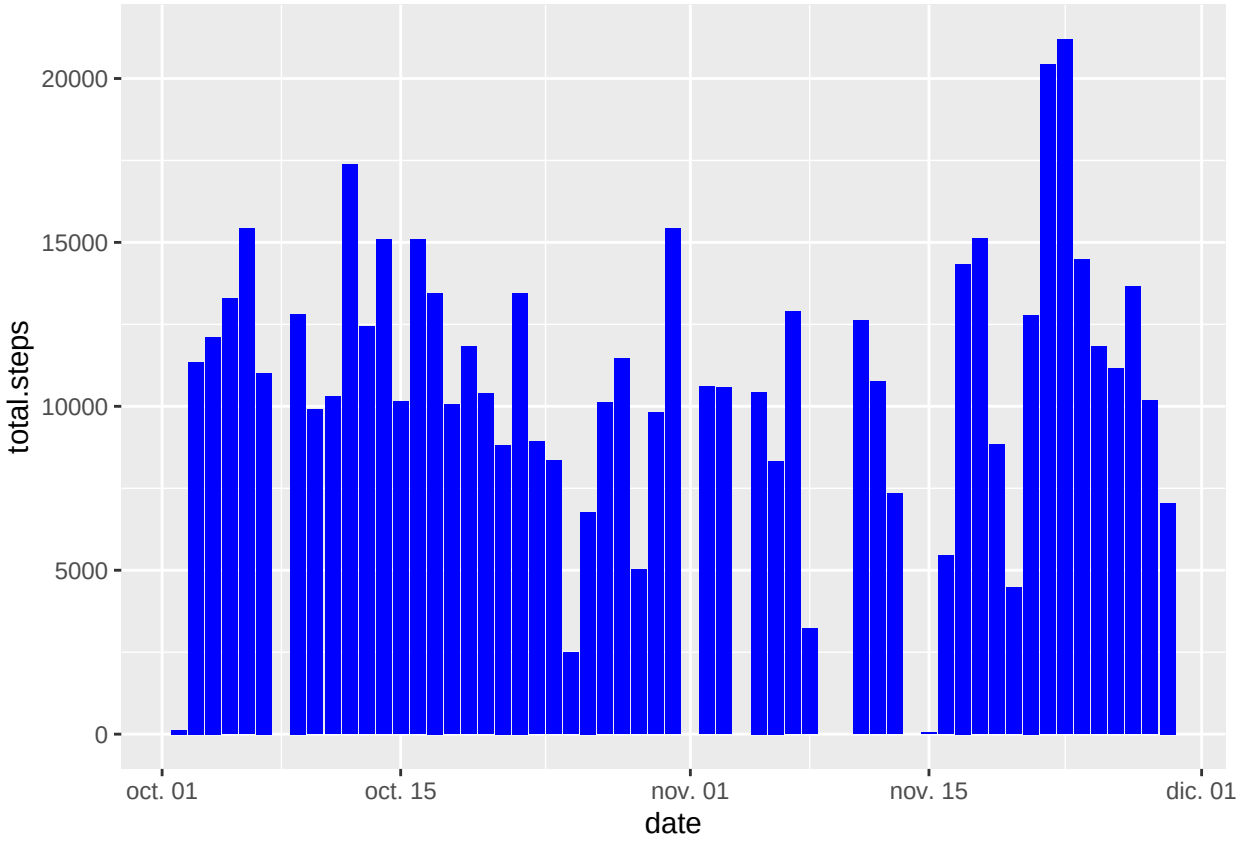
```
## [1] 2304
```

2) filling in all of the missing values in the dataset by the mean for that day and Create a new dataset with not missing values

```
act.imp <- activity
act3 <- activity %>% group_by(wday(date)) %>% summarise(mean.steps=mean(steps, na.rm=TRUE))
for(i in 1:7) {
    act.imp[is.na(act.imp$steps) & wday(act.imp$date) == i, 1] <- act3[i,2]
}
act.imp[is.na(act.imp$steps), 1] <- mean(act.imp$steps, na.rm=T)
```

3) histogram of the total number of steps taken each day

```
act.imp1 <- act.imp %>% group_by(date) %>% summarise(total.steps=sum(steps, na.rm=TRUE))
ggplot(act1, aes(x=date, y=total.steps)) + geom_histogram(stat="identity", fill="blue")
```



4) Mean and Median for of the total number of steps taken per day with imputes missing values

```
tmp2 <- act.imp1 %>% summarise(mean.steps = mean(total.steps, na.rm = TRUE), median.steps = median(tota
tmp2
```

```
## # A tibble: 1 x 2
##   mean.steps median.steps
##        <dbl>        <dbl>
## 1     10821.        11015
```

5) impact of imputing missing data on the estimates of the total daily number of steps, difference in the mean and median

```
tmp2[1,1] - tmp[1,1]
```

```
##   mean.steps
## 1   55.02092
```

```
tmp2[1,2] - tmp[1,2]
```

```
##   median.steps
## 1          250
```

Are there differences in activity patterns between weekdays and weekends?

1) Create a new factor variable in the database

```
act.imp <- act.imp %>% mutate(type_day = as.factor(ifelse(wday(act.imp$date) %in% c(2,3,4,5,6),"weekday
```

2) time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```
act4 <- act.imp %>% group_by(interval) %>% filter(type_day == "weekday") %>% summarise(mean.steps.weekda
tmp3 <- act.imp %>% group_by(interval) %>% filter(type_day == "weekend") %>% summarise(mean.steps.weeke
act4 <- cbind(act4,tmp3[,2])
with(act4, plot(interval, mean.steps.weekdays, type = "l", col = "red"))
with(act4, points(interval, mean.steps.weekend, type = "l", col = "blue"))
legend("topright", legend = c("mean.steps.weekdays", "mean.steps.weekend"), col = c("red", "blue"), cex
```