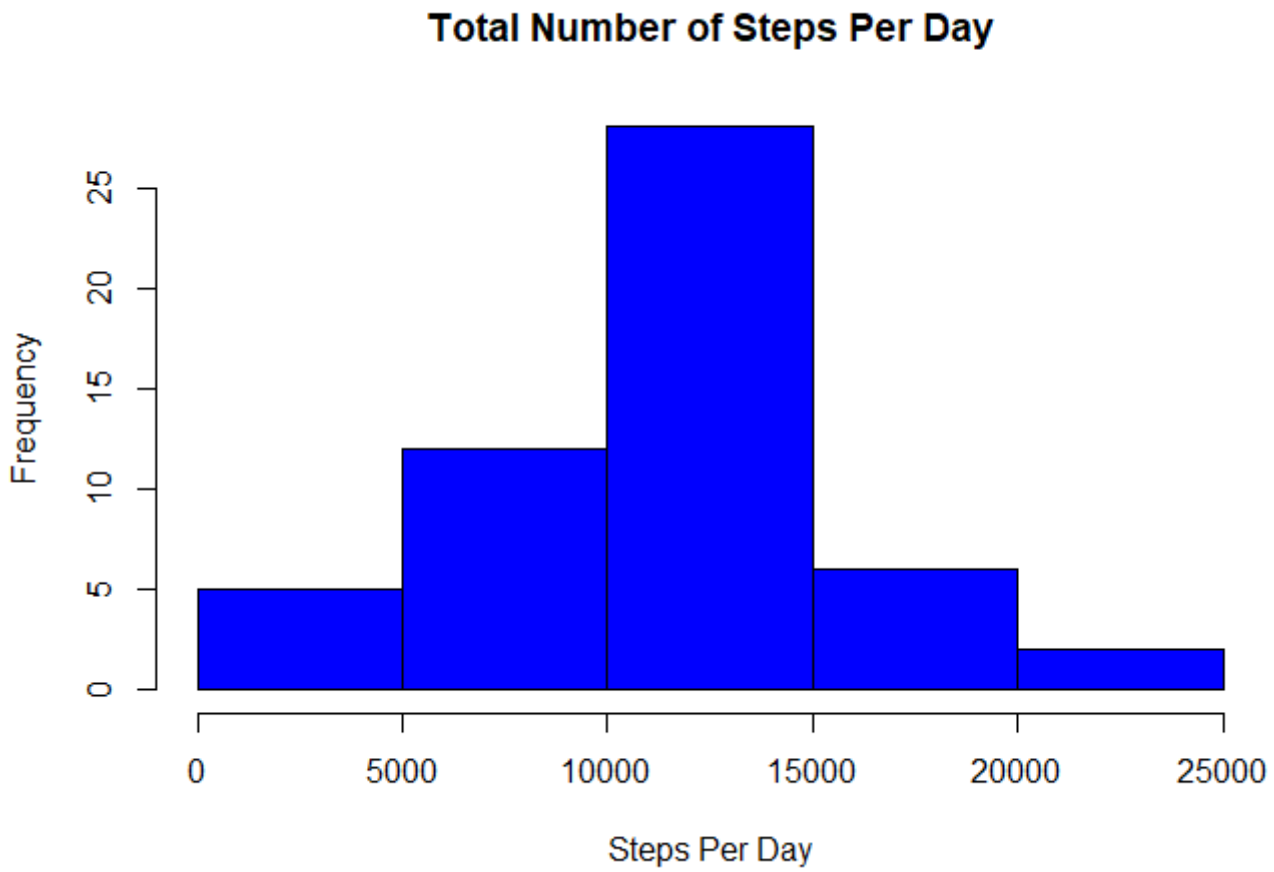
PA1_template

```
setwd("C:/Users/Leslie/Documents/Classes/Data Science/Reproducible Research/Week 2 Assignment")

data <- read.csv("activity.csv")
```

``` Transform the data to make it usable.

```
newdata <- na.omit(data)
```

Calculate the number of steps taken per day.

```
stepsperday <- aggregate(steps ~ date, newdata, sum)
```

Plot a histogram of the total number of steps taken each day.

```
hist(stepsperday$steps, col = "blue", xlab = "Steps Per Day", main = "Total Number of Steps Per Day")
```
```



Calculate and report the mean of the total number of steps taken each day.

```
meansteps <- mean(stepsperday$steps)
meansteps
```

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

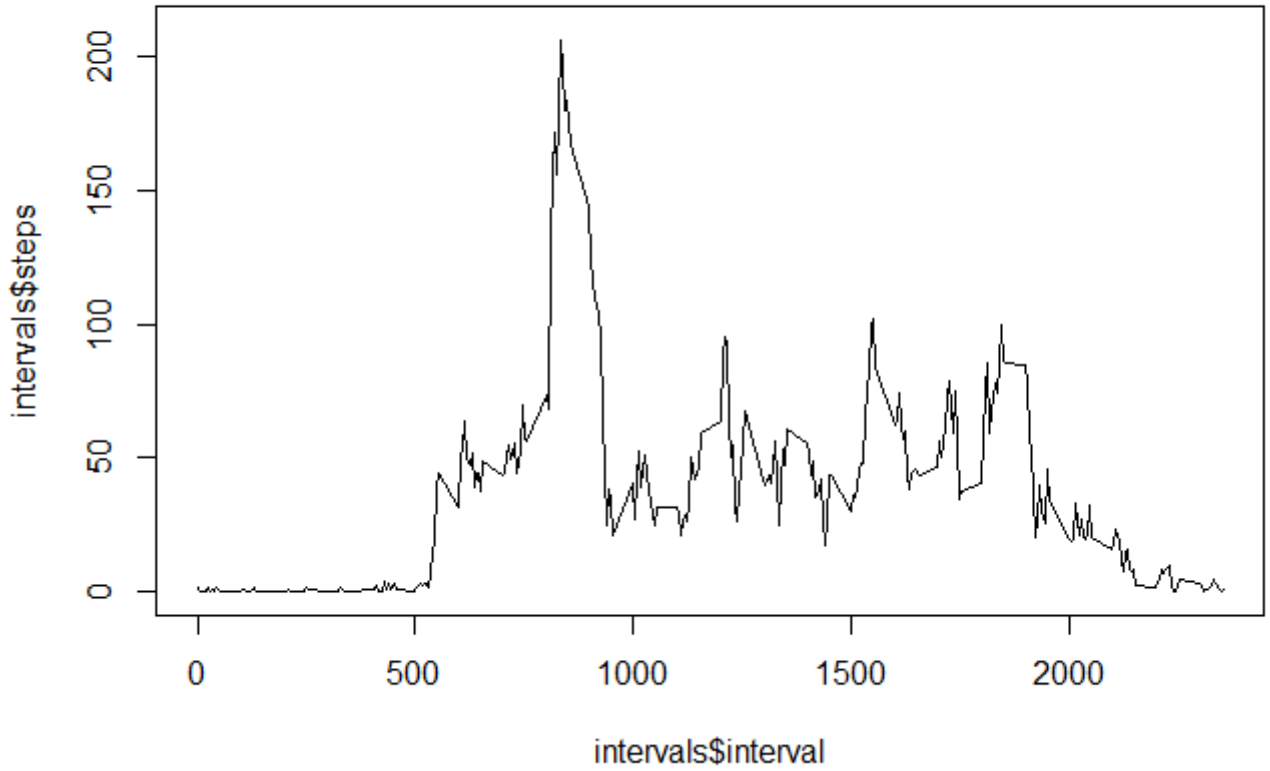
Calculate and report the median of the total number of steps taken each day.

```
mediansteps <- median(stepsperday$steps)
mediansteps
```

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

Time Series Plot

```
library(ggplot2)
intervals <- aggregate(steps ~ interval, newdata, mean)
plot(intervals$interval, intervals$steps, type = "l", ylim = c(0, 210))
```



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

```
which.max( intervals[,1] )
```

```
## [1] 288
```

Calculate the number of missing values in the dataset.

```
sum(is.na(data))
```

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

Devise a strategy for filling in all of the missing values in the dataset.

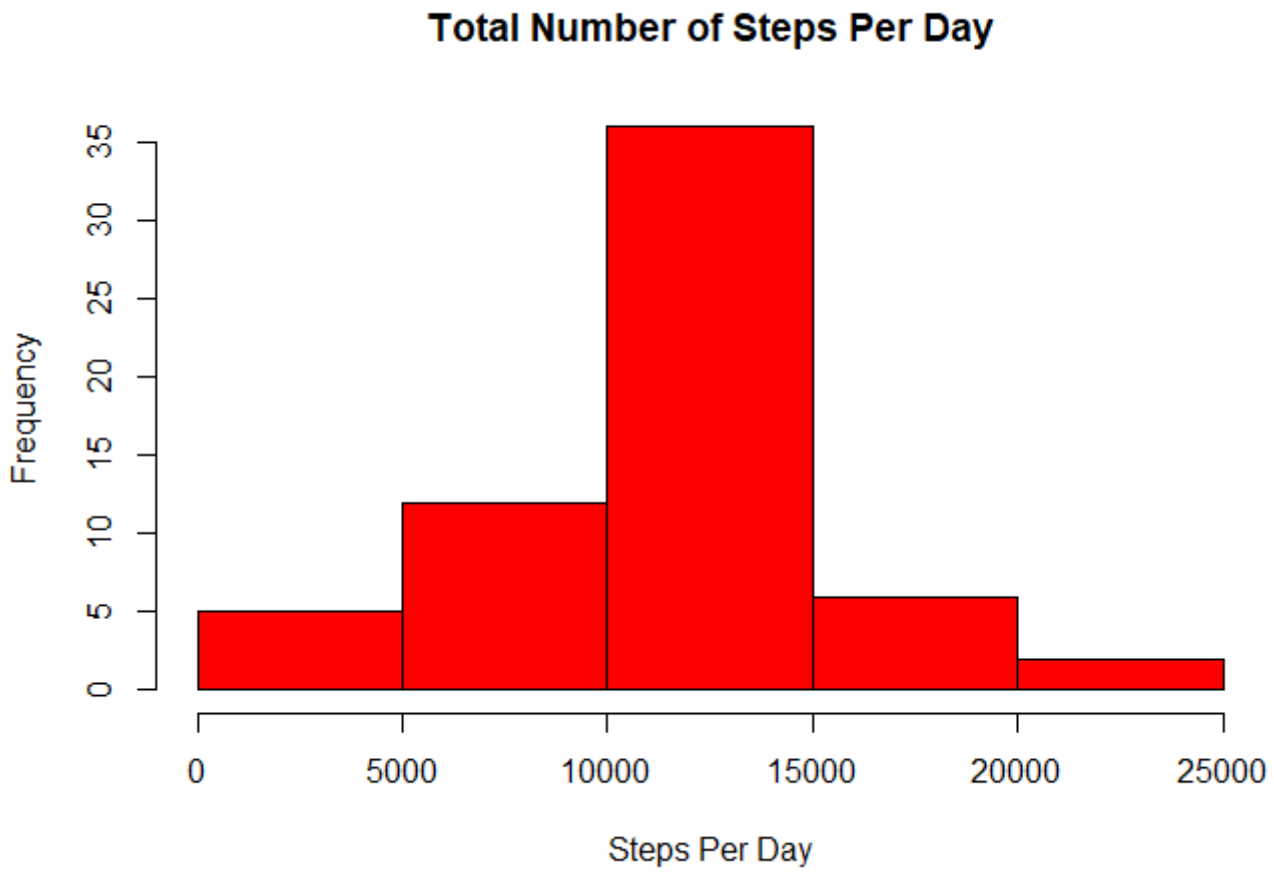
```
nadata <- data
for (i in 1:nrow(nadata)) {
  if (is.na(nadata$steps[i])) {
    intvalue <- nadata$interval[i]
    stepsvalue <- intervals[
      intervals$interval == intvalue,]
    nadata$steps[i] <- stepsvalue$steps
  }
}
```

Create a new dataset that is equal to the original dataset but with the missing data filled in. Missing data imputed with mean of intervals across days.

```
View(nadata)
```

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?

```
stepsperday2 <- aggregate(steps ~ date, nadata, sum)
hist(stepsperday2$steps, col = "red", xlab = "Steps Per Day", main = "Total Number of Steps Per Day")
```



Mean of steps with imputed values:

```
meansteps <- mean(stepsperday2$steps)
meansteps
```

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

Median of steps with imputed values:

```
mediansteps <- median(stepsperday2$steps)
mediansteps
```

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

Original Values: Mean - 10766.19 Median - 10765

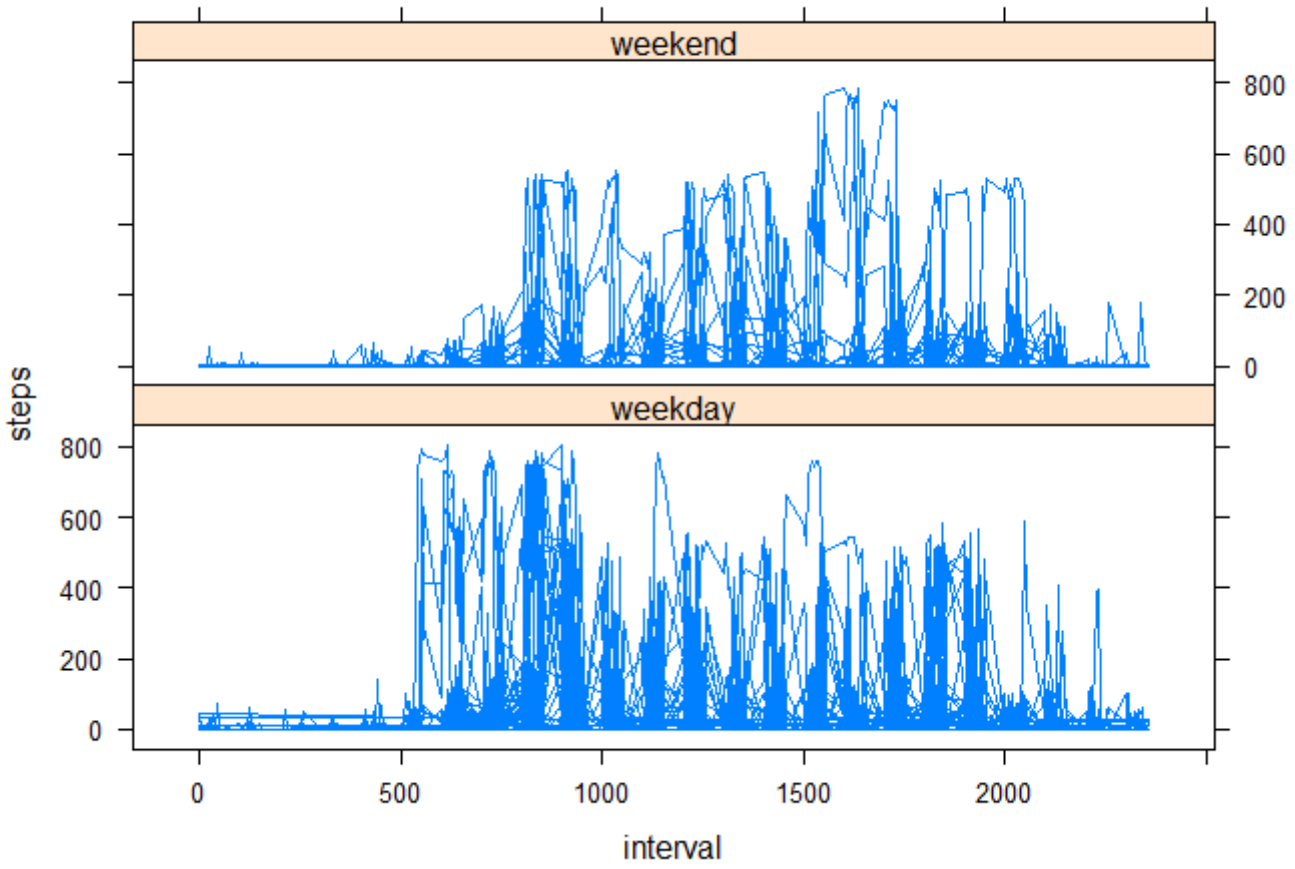
Imputed Values: Mean - 10766.19 Median - 10766.19 After imputing the missing values, the mean remains the same, but the median grows closer to the mean of the data set, which is slightly higher than the median of the original data set.

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.

```
daytype <- function(date) {
    if (weekdays(as.Date(date)) %in% c("Saturday", "Sunday")) {
        "weekend"
    } else {
        "weekday"
    }
}
nadata$daytype <- as.factor(sapply(nadata$date, daytype))
```

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).

```
library(lattice)
xyplot(steps~interval|daytype,
       type="l",
       data = nadata,
       layout=c(1,2))
```


weekend
weekday
steps
interval
0
200
400
600
800
0
500
1000
1500
2000