Activity Monitor Data

by Julie Dragon

For this analysis we will make use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

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

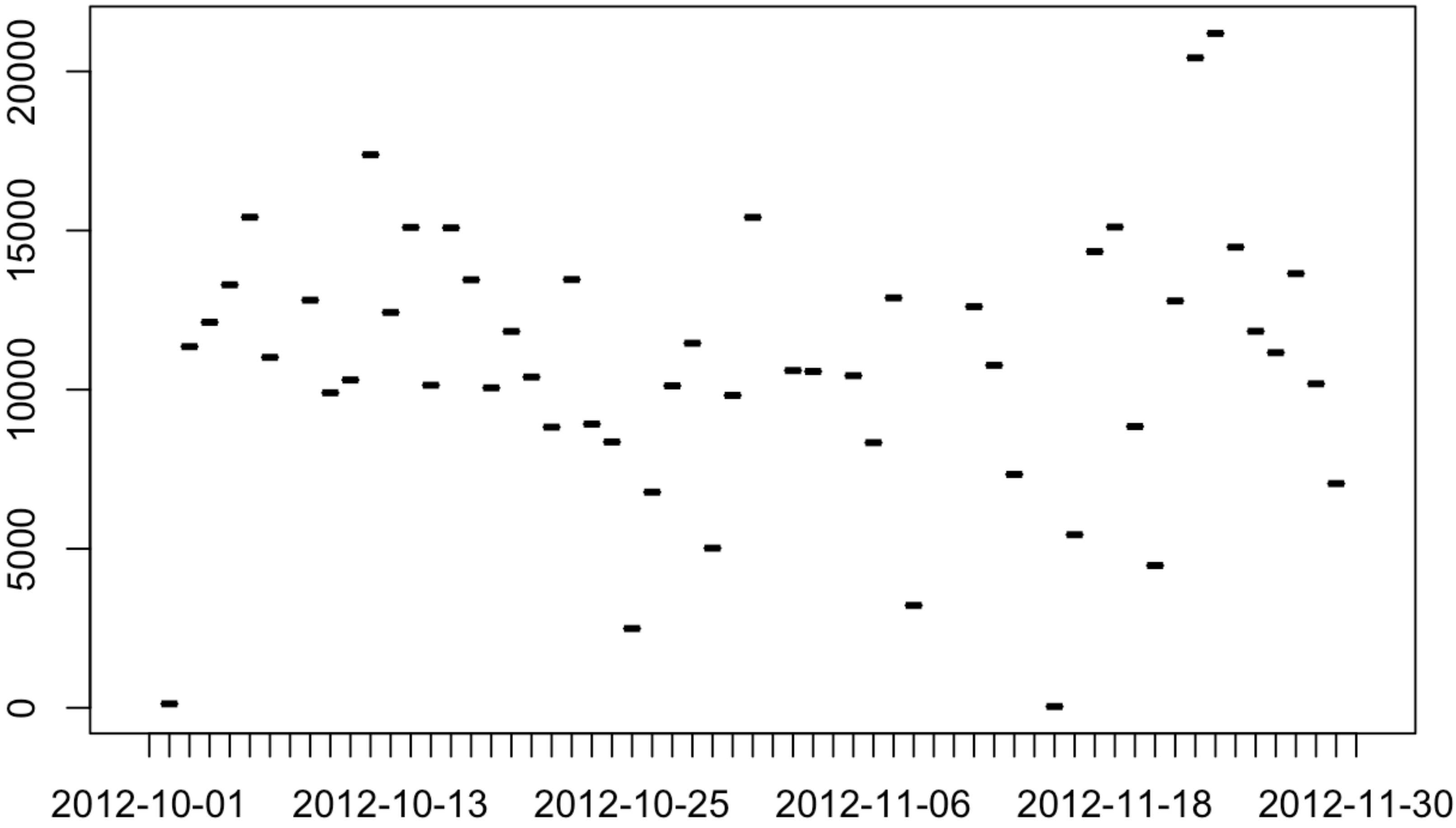
What is mean total number of steps taken per day? For this part of the assignment, you can ignore the missing values in the dataset.

Calculate the total number of steps taken per day

```
timeseries <- activity[!is.na(activity$steps),]
timeseriesTotal <- by (timeseries, timeseries$date,
                    function (i) {data.frame(date=unique(i$date), steps=sum(i$ste
ps))})
timeseriesTotal <- do.call(rbind, timeseriesTotal)
```

Make a histogram of the total number of steps taken each day

```
plot(timeseriesTotal$date, timeseriesTotal$steps, type="h")
```



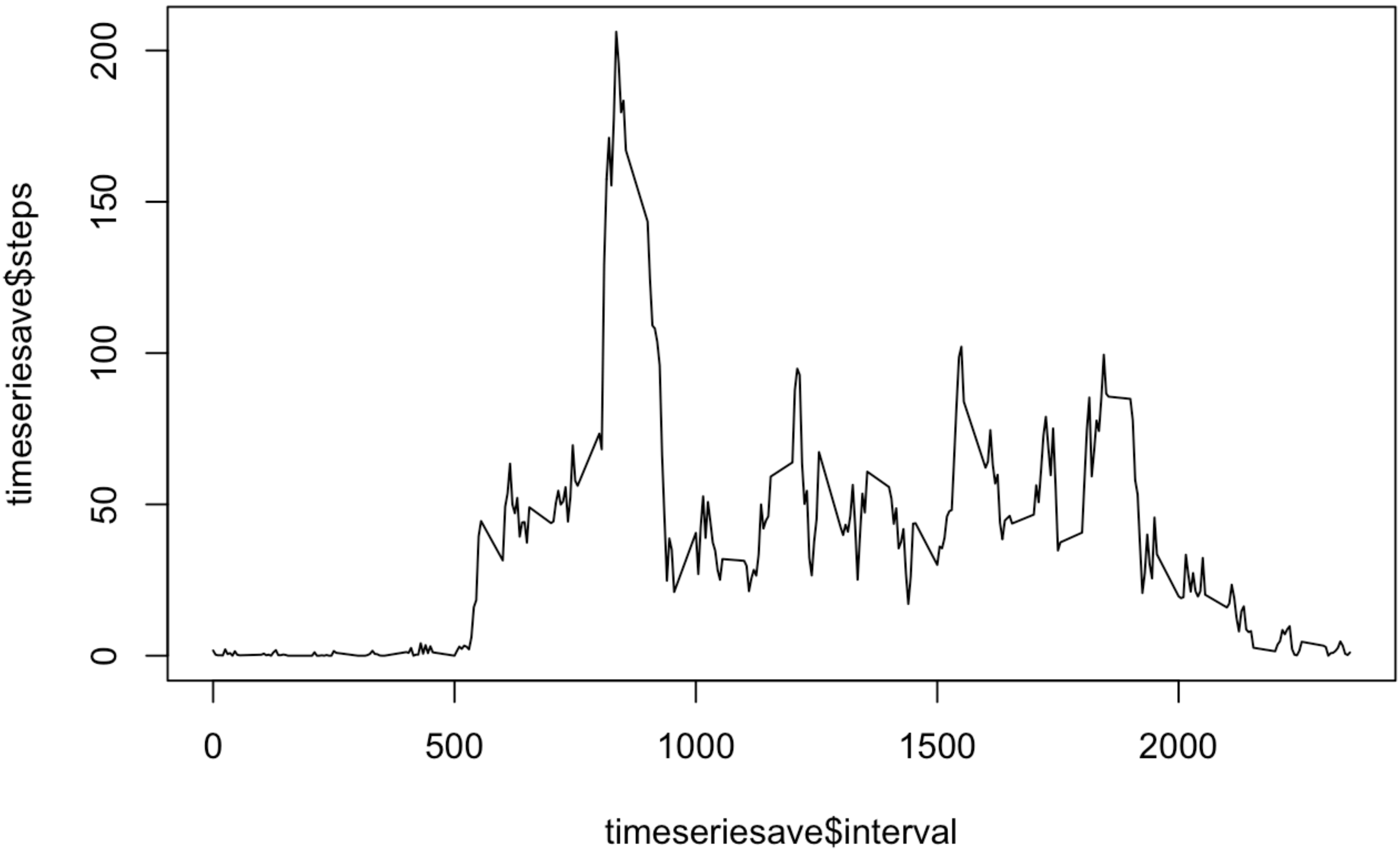
Calculate and report the mean and median of the total number of steps taken per day

```
mean <- mean(timeseriesTotal$steps)
median <-median(timeseriesTotal$steps)
```

The mean number of steps per day is 1.076618910^{4}. The median is 10765.

What is the average daily activity pattern? Make 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 days (y-axis)

```
timeseries <- activity[!is.na(activity$steps),]
timeseriesave <- by (timeseries, timeseries$interval,
                     function (i) {data.frame(interval=unique(i$interval), steps=m
ean(i$steps))})
timeseriesave <- do.call(rbind, timeseriesave)
plot(timeseriesave$interval, timeseriesave$steps, type= "l")
```



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

```
rowMaxSteps <- which(timeseriesave$steps == max(timeseriesave$steps))
intervalmax <- timeseriesave[rowMaxSteps, "interval"]
```

The maximum steps interval is 835.

Imputing missing values Note that there are a number of days/intervals where there are missing values (coded as ????????). The presence of missing days may introduce bias into some calculations or summaries of the data.

Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with ????????s)

```
missing <- sum(is.na(activity$steps))
```

The number of missing values for steps, across days and intervals, is 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.

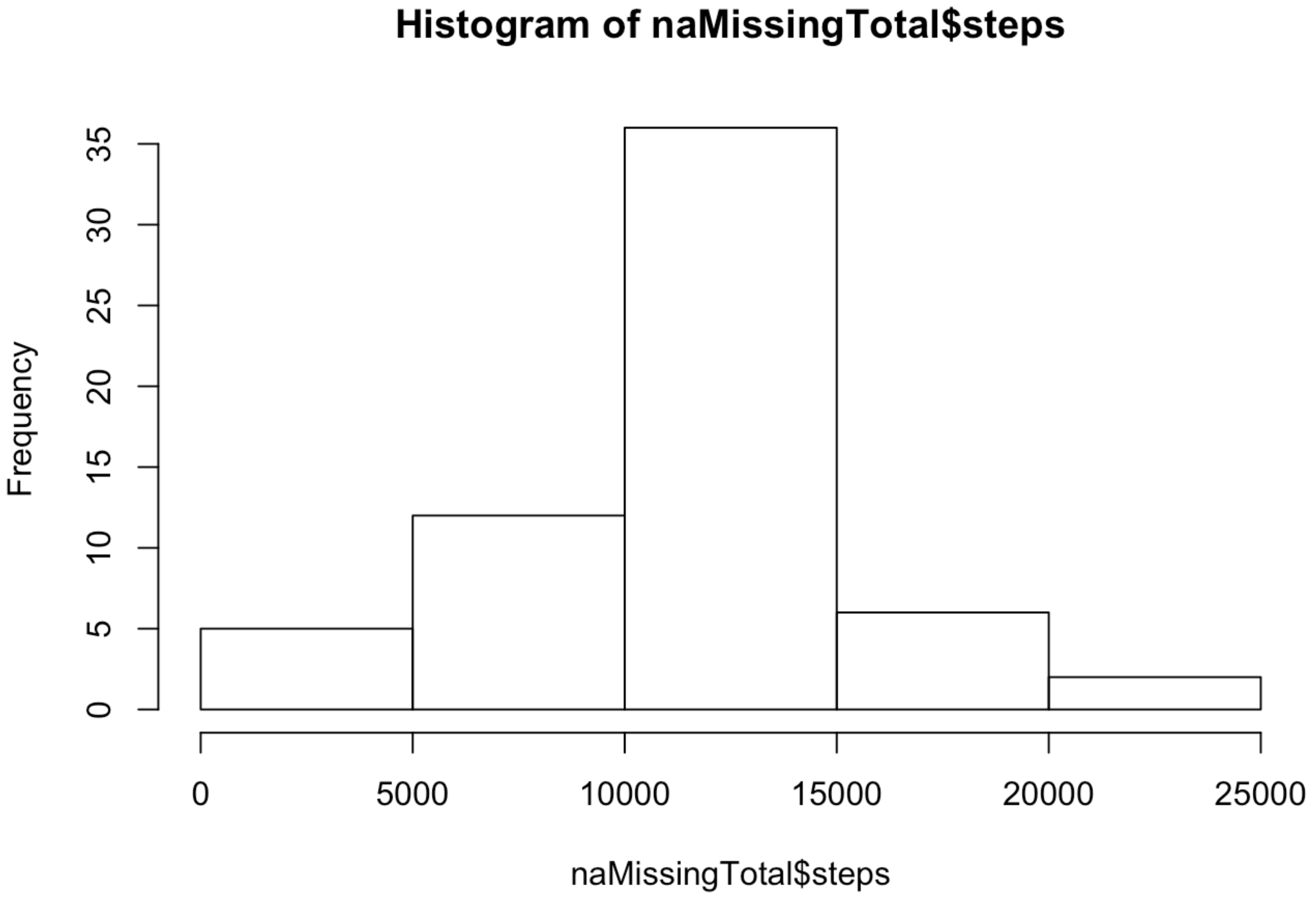
Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
naMissing = activity
for (i in 1:nrow(naMissing)) {
        if (is.na(naMissing[i,"steps"])) {
                interval <- naMissing$interval[i]
                naMissing[i, "steps"] <- timeseriesave[timeseriesave$interval == interval, "steps"]
        }
}
```

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?

```
naMissingTotal <- by (naMissing, naMissing$date,
                     function (i) {data.frame(date=unique(i$date), steps=sum(i$steps))})
naMissingTotal <- do.call(rbind, naMissingTotal)

hist(naMissingTotal$steps)
```



```
naMissingMean <- mean(naMissingTotal$steps)
naMissingMedian <-median(naMissingTotal$steps)
```

The mean number of steps per day is 1.076618910^{4}. The median is 1.076618910^{4}.

Are there differences in activity patterns between weekdays and weekends? For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

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.

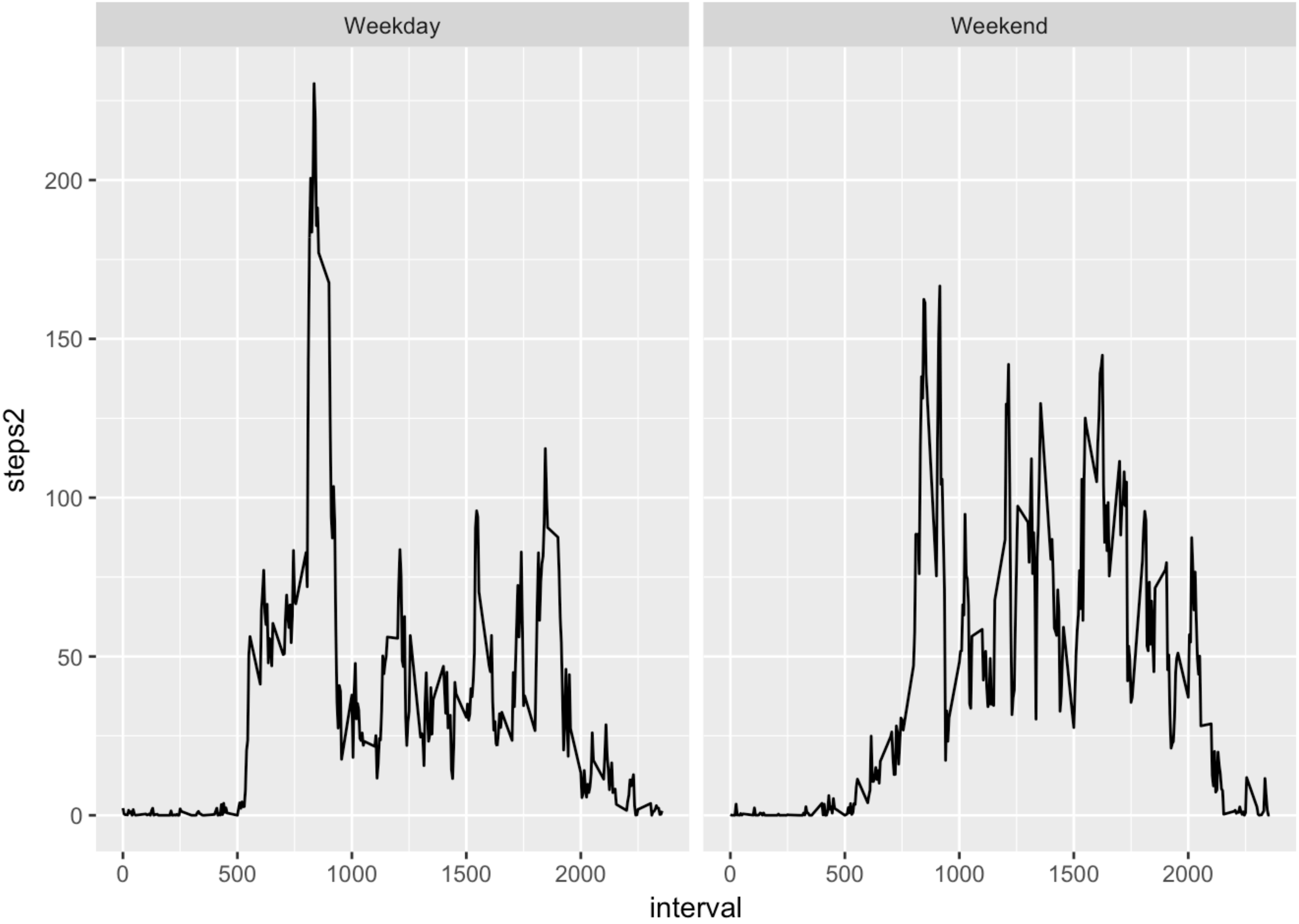
```
dayType <- weekdays(as.POSIXlt(naMissing$date))
naMissing[dayType %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"),
"dayType"] <- "Weekday"
naMissing[dayType %in% c("Saturday", "Sunday"), "dayType"] <- "Weekend"
naMissing$dayType <- factor(naMissing$dayType)
```

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). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
timeseries2Total <- by (naMissing, list(naMissing$interval, naMissing$dayType),
                     function (i) {
                             data.frame(
                                     interval=unique(i$interval),
                                     n=nrow(i),
                                dayType=unique(i$dayType),
                                     steps2=mean(i$steps))
                            }
                     )
timeseries2Total <- do.call(rbind, timeseries2Total)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
ggplot(timeseries2Total, aes(x=interval, y=steps2)) +geom_line() + facet_wrap(~day
Type)
```


Weekday
Weekend
steps2
0
50
100
150
200
interval
0
500
1000
1500
2000
0
500
1000
1500
2000