Assignment1

Thomas Cronin

July 16, 2025

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, fig.width = 10, fig.height = 5,
                      fig.keep = 'all' ,fig.path = 'figures\ ', dev = 'png')
```

1. Loading and preprocessing the data:

```
# load the data
activity <- read.csv("activity.csv")
# check the data
head(activity)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

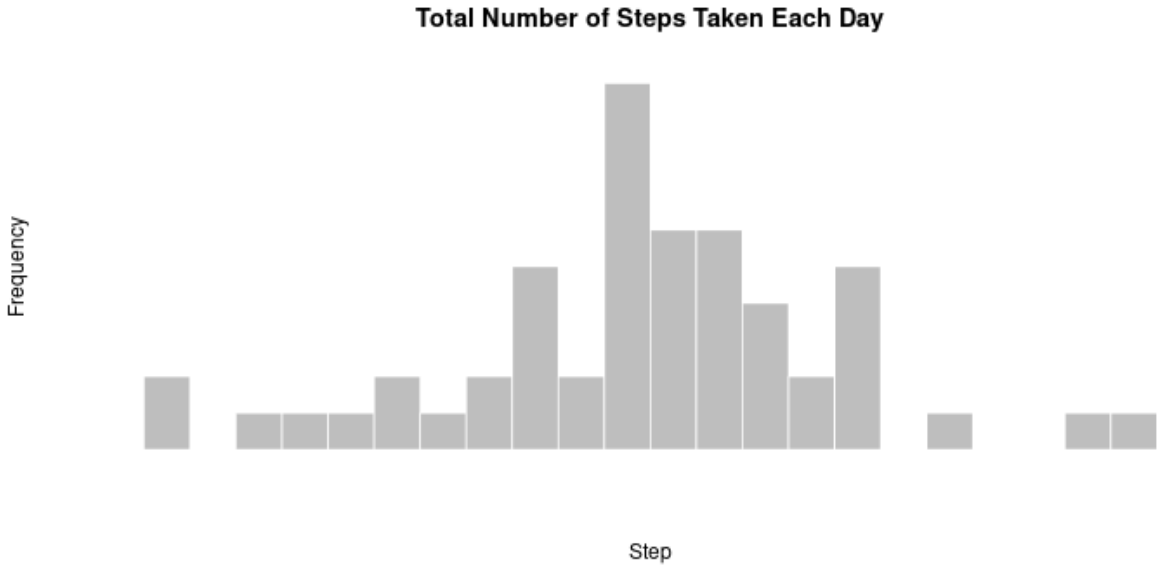
```
# change date to calss factor to class Date
activity$date <- as.Date(activity$date, format = "%Y-%m-%d")
```

2. What is mean total number of steps taken per day?

```
# the total number of steps taken per day is stored in the variable called "total_step"
total_step <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
head(total_step)
```

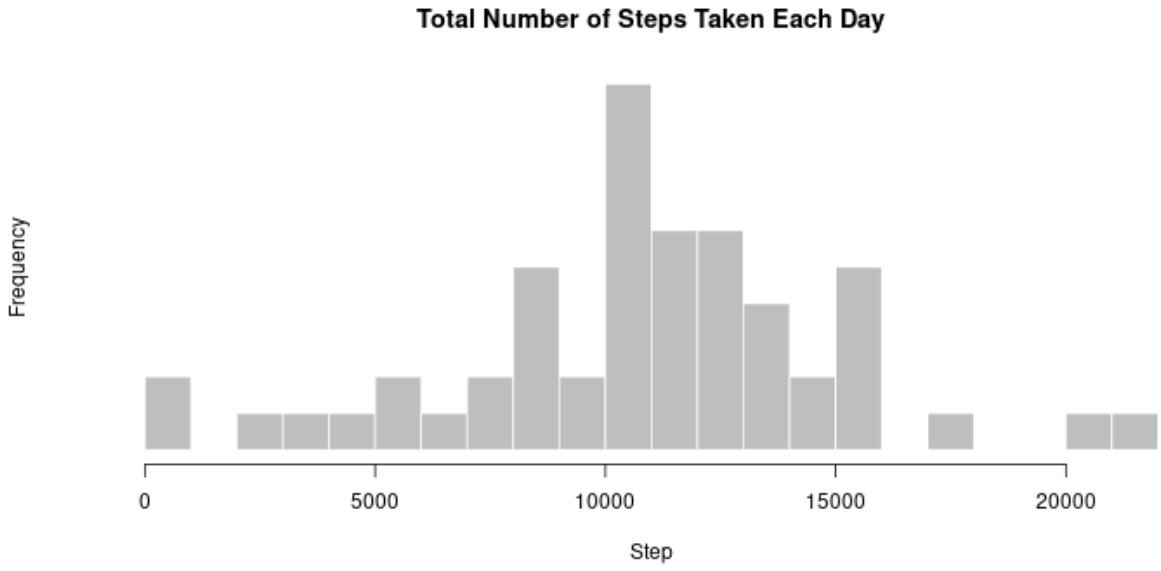
```
##         date steps
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

```
par(mfrow = c(1, 1))
# use base plotting system and more bins than the default setting
hist(total_step$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day",
     col = "grey", border = "white", xlab = "Step", axes = FALSE)
```

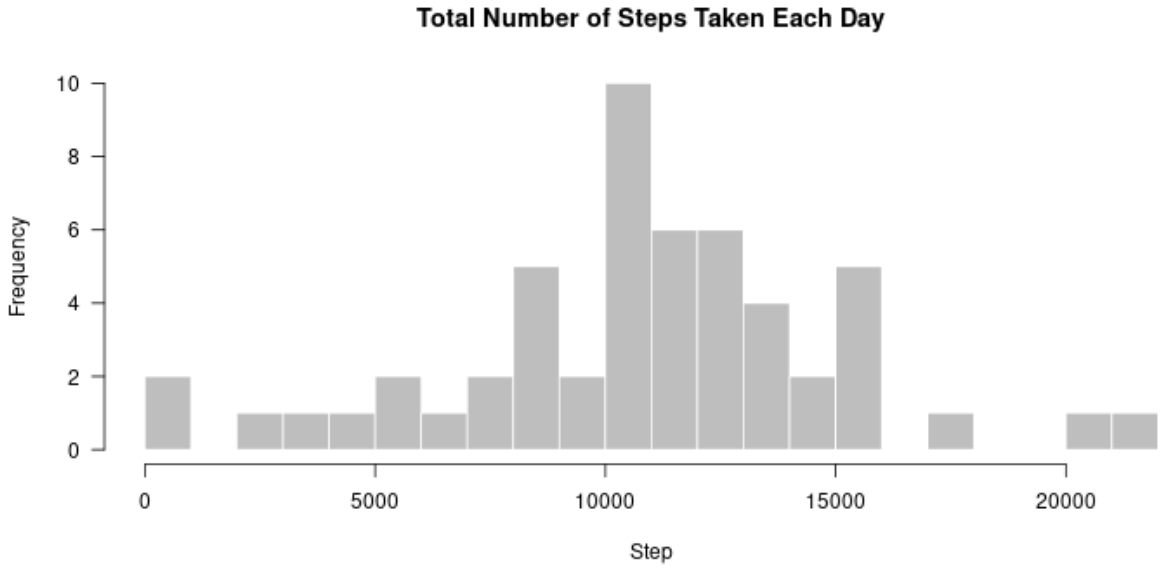


Histogram:

```
axis(1)
```




```
axis(2, las = 1)
```



Mean and Median number of steps taken each day:

```
mean(total_step$steps)
```

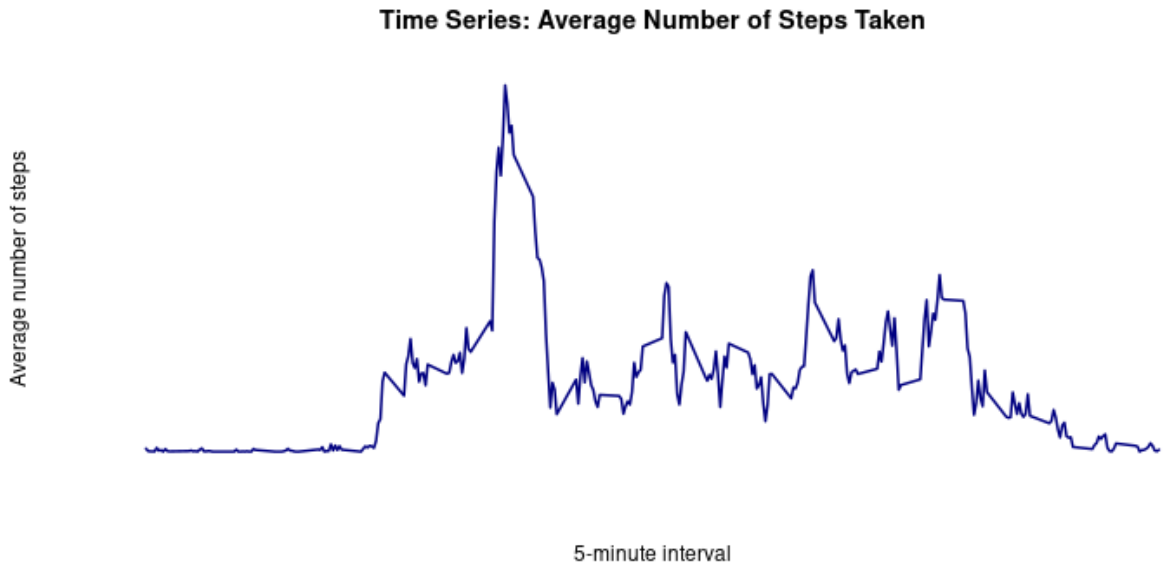
```
## [1] 10766.19
```

```
median(total_step$steps)
```

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

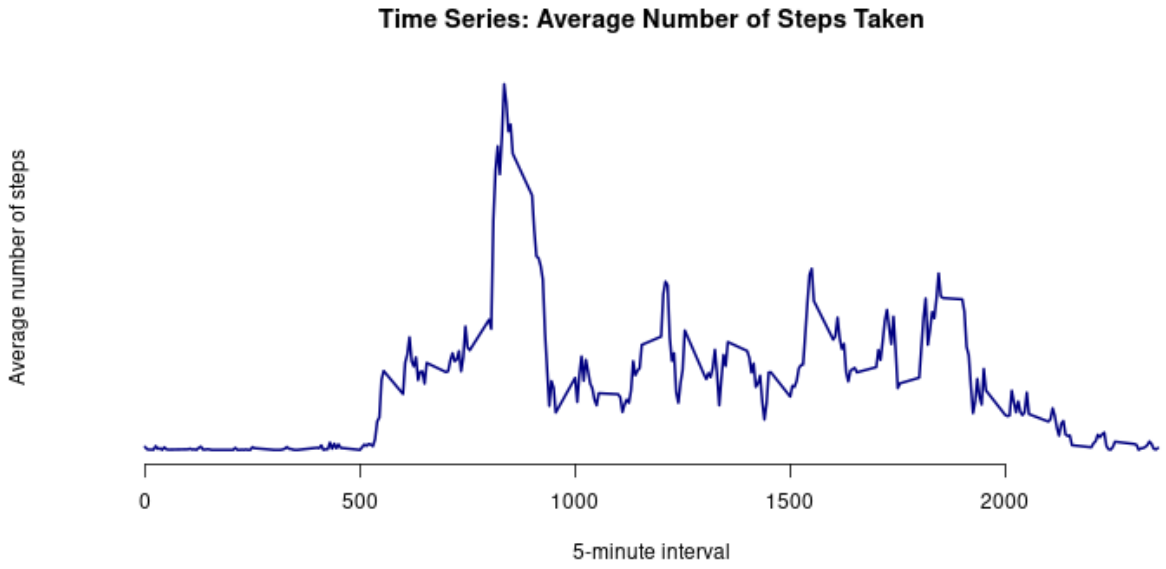
3. What is the average daily activity pattern?

```
avg_step <- aggregate(steps ~ interval, data = activity, mean, na.rm = TRUE)
plot(avg_step$interval, avg_step$steps, type = "l", lwd = 2, col = "navy",
     main = "Time Series: Average Number of Steps Taken", axes = FALSE,
     xlab = "5-minute interval", ylab = "Average number of steps")
```

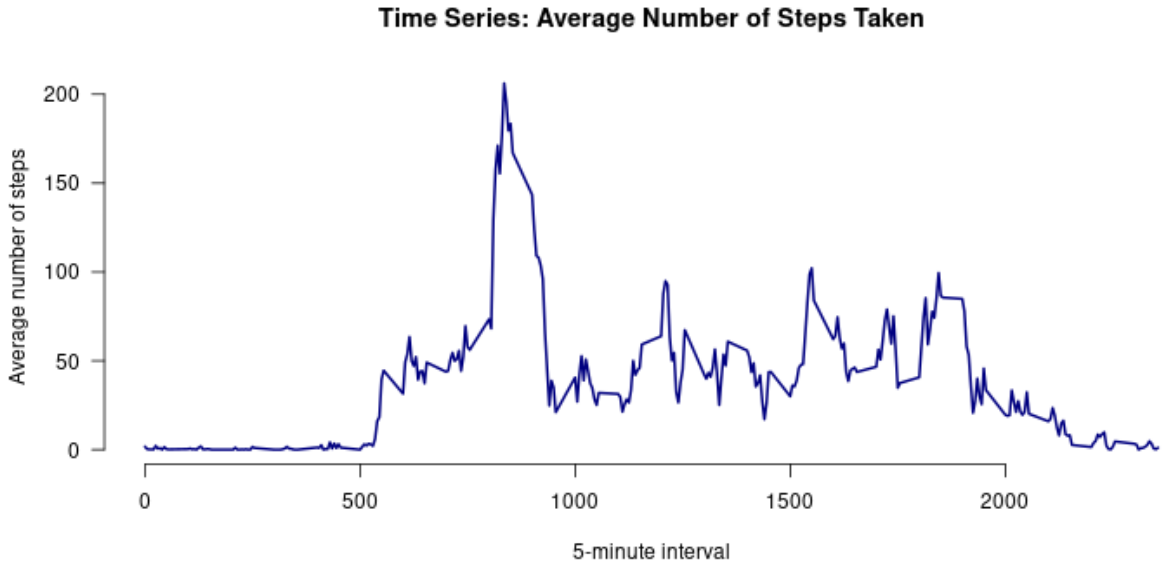


Time series plot:

```
axis(1)
```




```
axis(2, las = 1)
```




```
avg_step$interval[which.max(avg_step$steps)]
```

The 5-minute interval contains the max number of steps:

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

4. Imputing missing values:

The total missing values.

```
sum(is.na(activity)) # or dim(activity[activity$steps == "NA", ])[1]
```

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

assign avg to all NA in a new dataset:

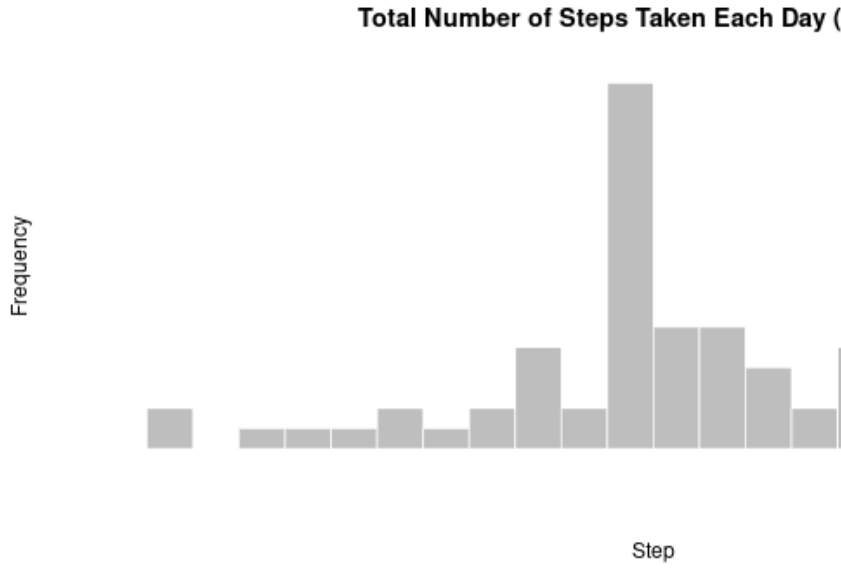
```
imp <- activity # new dataset called imp
for (i in avg_step$interval) {
    imp[imp$interval == i & is.na(imp$steps), ]$steps <-
        avg_step$steps[avg_step$interval == i]
}
head(imp) # no NAs
```

```
##       steps       date interval
## 1 1.7169811 2012-10-01        0
## 2 0.3396226 2012-10-01        5
## 3 0.1320755 2012-10-01       10
## 4 0.1509434 2012-10-01       15
## 5 0.0754717 2012-10-01       20
## 6 2.0943396 2012-10-01       25
```

```
sum(is.na(imp)) # should be 0
```

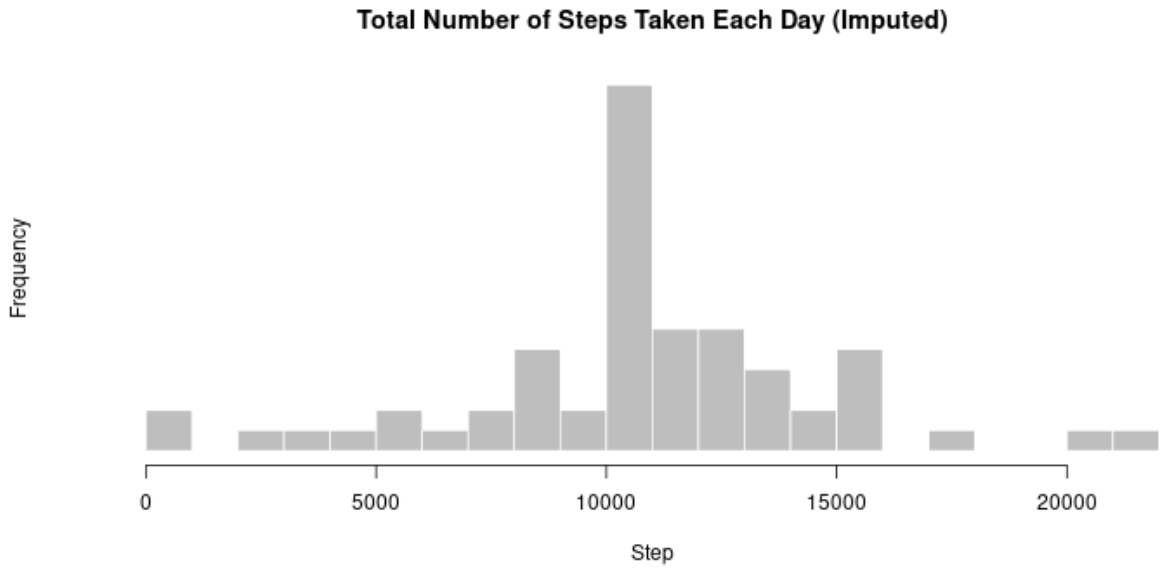
```
## [1] 0
```

```
total_step_imp <- aggregate(steps ~ date, data = imp, sum, na.rm = TRUE)
hist(total_step_imp$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day (Imputed)",
     col = "grey", border = "white", xlab = "Step", axes = FALSE)
```

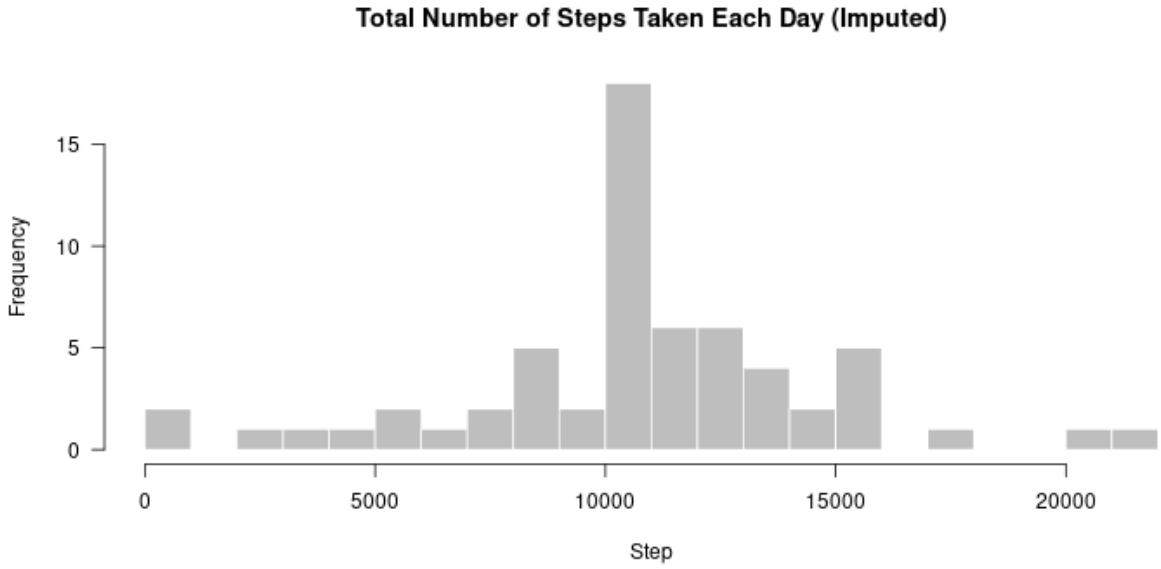


Histogram after missing values are imputed:

```
axis(1)
```




```
axis(2, las = 1)
```




```
mean(total_step_imp$steps)
```

Mean and median after missing values are imputed:

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

```
median(total_step_imp$steps)
```

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

5. Differences in activity patterns between weekdays and weekends:

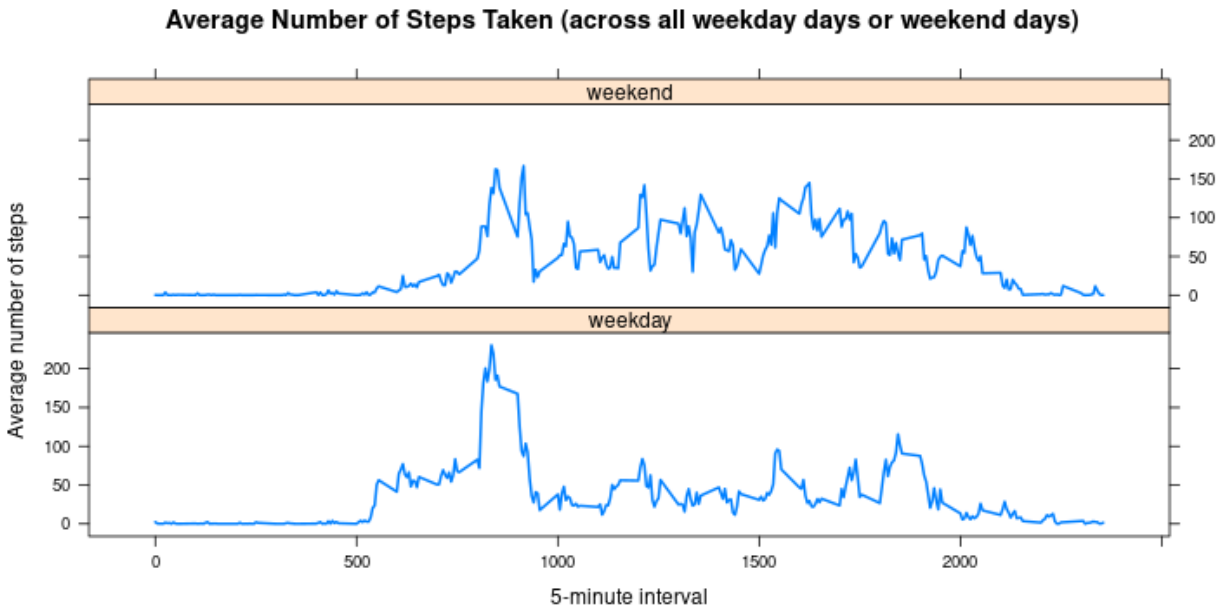
Create new factor variables.:

```
imp$day <- weekdays(imp$date)
imp$week <- ""
imp[imp$day == "Saturday" | imp$day == "Sunday", ]$week <- "weekend"
imp[!(imp$day == "Saturday" | imp$day == "Sunday"), ]$week <- "weekday"
imp$week <- factor(imp$week)
```

```
imp$day <- weekdays(imp$date)
imp$week <- ""
imp[imp$day == "Saturday" | imp$day == "Sunday", ]$week <- "weekend"
imp[!(imp$day == "Saturday" | imp$day == "Sunday"), ]$week <- "weekday"
imp$week <- factor(imp$week)
```

Panel plot:

```
avg_step_imp <- aggregate(steps ~ interval + week, data = imp, mean)
library(lattice)
xyplot(steps ~ interval | week, data = avg_step_imp, type = "l", lwd = 2,
       layout = c(1, 2),
       xlab = "5-minute interval",
       ylab = "Average number of steps",
       main = "Average Number of Steps Taken (across all weekday days or weekend days)")
```


Average Number of Steps Taken (across all weekday days or weekend days)
weekend
weekday
Average number of steps
0
50
100
150
200
0
500
1000
1500
2000
5-minute interval