Course Project 1 - Reporducible Research

Loading and Preprocessing the Data

The code below completes the following tasks:
1. Load the data
2. Process/transfor the data into a format suitable for analysis

```
data<-read.csv("activity.csv",sep=",",head=T,colClasses=c("integer","Date","integer"))
```

What is mean total number of steps taken per day?

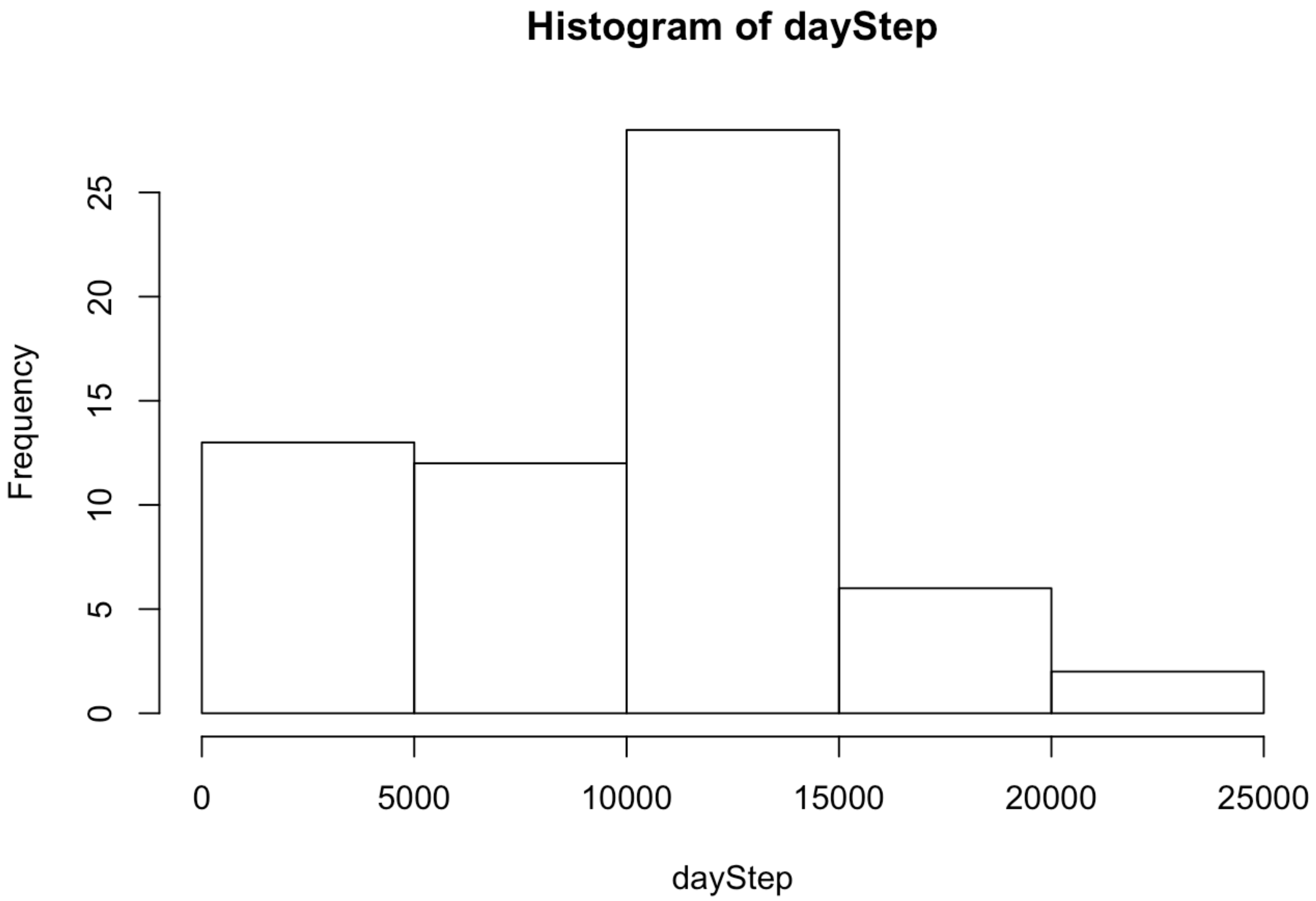
For this part of the assignment, you can ignore the missing values in the dataset.

1. Calculate the total number of steps taken per day

```
dayStep <- sapply(split(data$steps,data$date),sum,na.rm=T)
```

2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day

```
hist(dayStep)
```



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

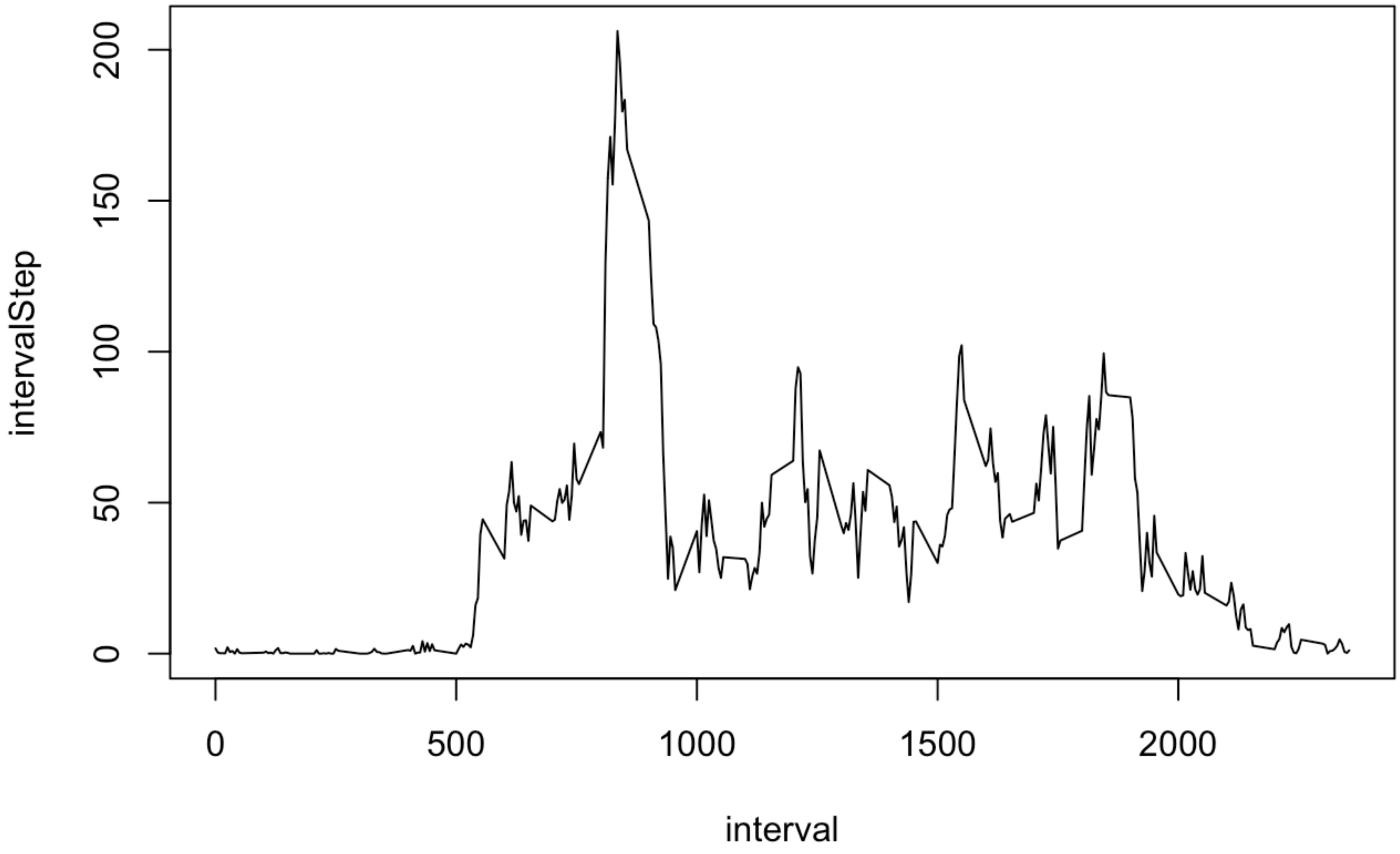
```
median_dayStep <-median(dayStep)
mean_dayStep <- mean(dayStep)
```

The mean number of daily steps is **9354.23** and the median is **1.039510^{4}** steps.

What is the average daily activity pattern?

1. 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)

```
intervalStep <- sapply(split(data$steps,data$interval),mean,na.rm=T)
interval<-unique(data$interval)
plot(x=interval, y=intervalStep,type ="l")
```



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

```
max_intervalStep<-which.max(intervalStep)
```

The 5 min interval at **835** contained the most average steps.
On average, there were **206.1698113** in this period.

Imputing missing values

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

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

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

There are **2304** records with missing step data.

2. 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.

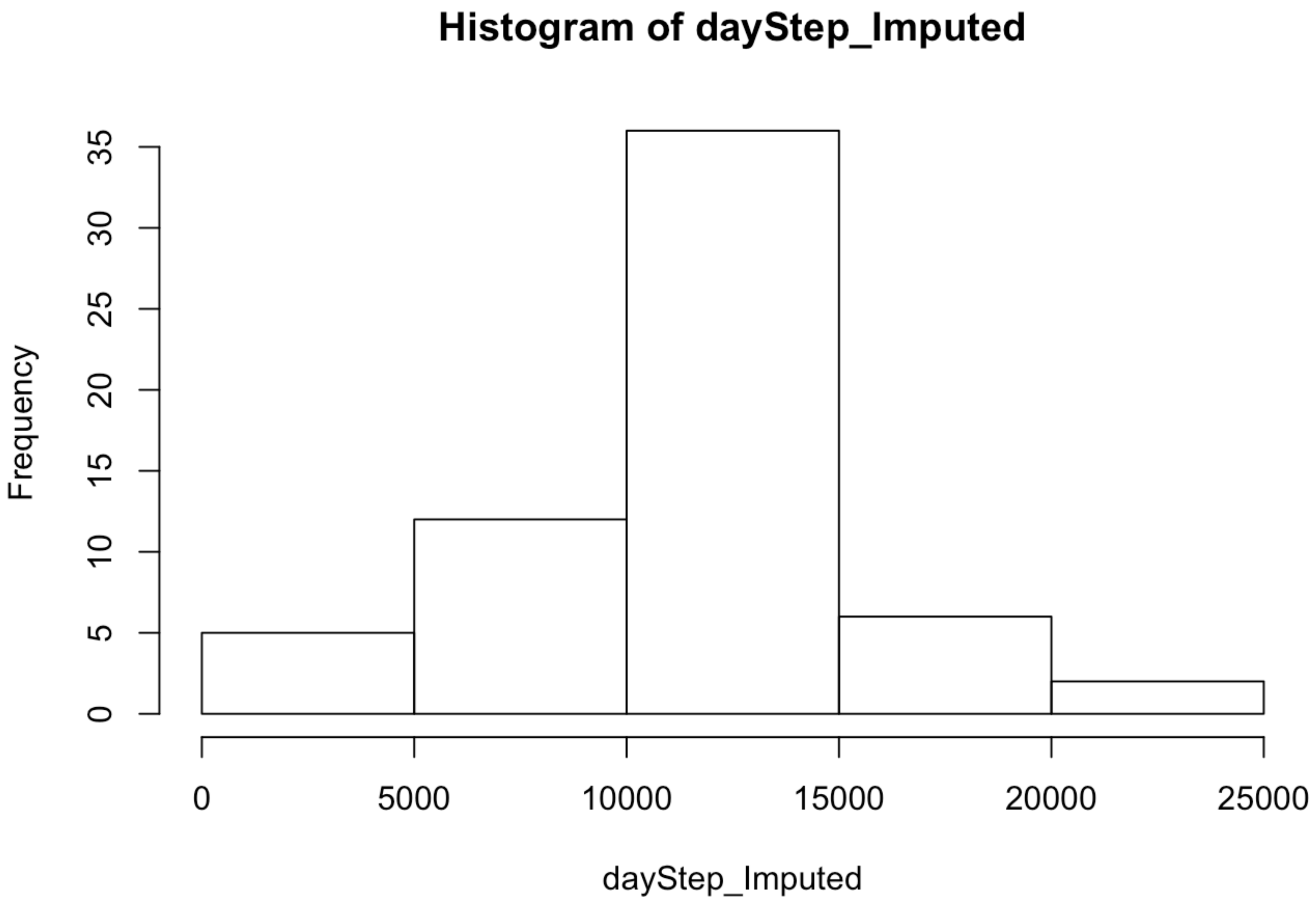
```
data$missing<-is.na(data$steps)
len<-length(data$steps)
data$stepsImputed <- data$steps
for(i in 1:len){
  if(data$missing[i]==TRUE){
    data$stepsImputed[i]<-intervalStep[match(data$interval[i],interval)]}
}
```

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

```
dayStep_Imputed <- sapply(split(data$stepsImputed,data$date),sum)
```

4. 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?

```
hist(dayStep_Imputed)
```



```
mean_dayStep_Imputed <- mean(dayStep_Imputed)
median_dayStep_Imputed <- median(dayStep_Imputed)
```

WITHOUT Imputing Missing Values

The mean number of daily steps is **9354.23** and the median is **10395**.

WITH Imputing Missing Values

The mean number of daily steps is **10766** and the median is **10766**.

Are there differences in activity patterns between weekdays and weekends?

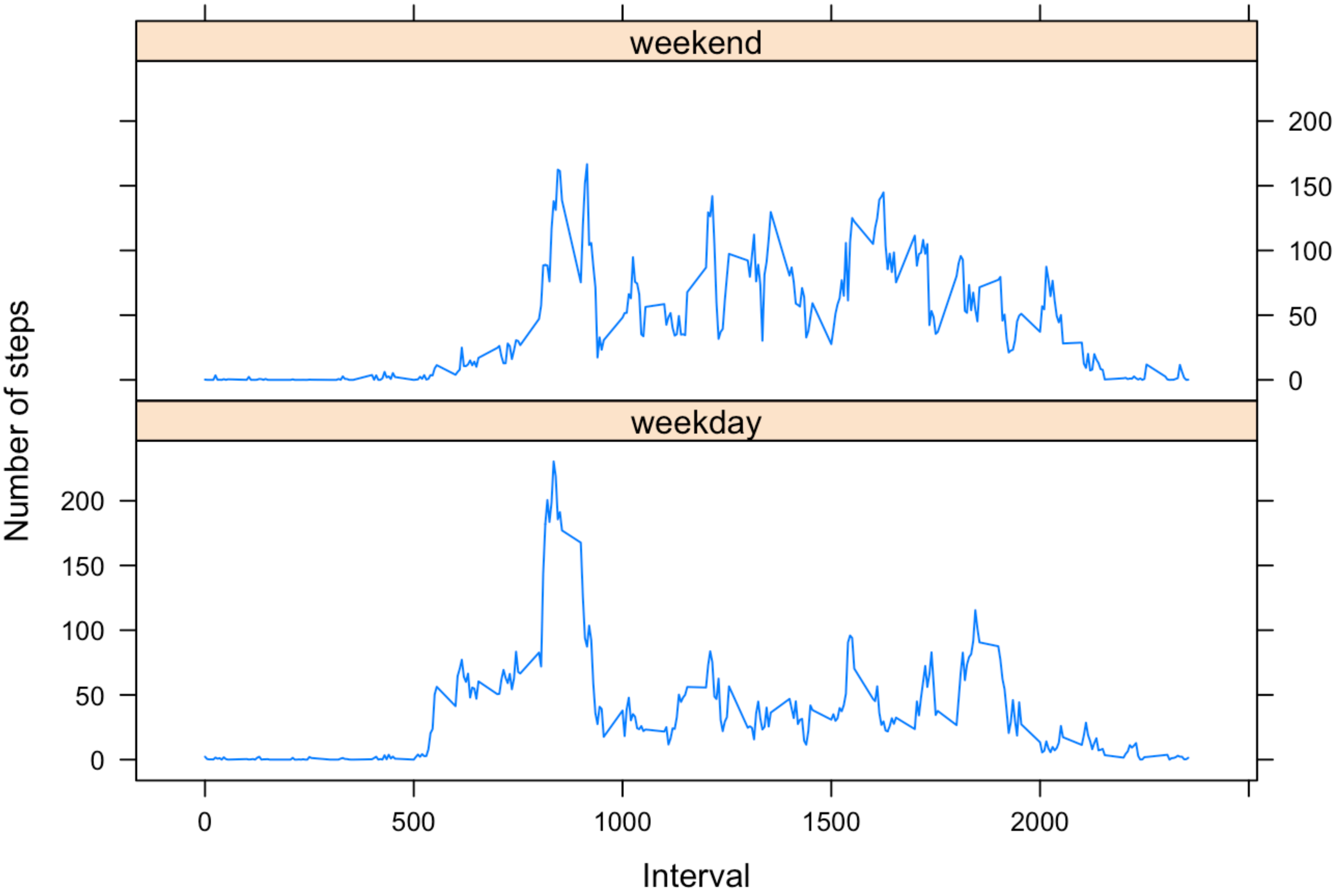
For this part the `weekdays`() function may be of some help here. Use the dataset with the filled-in missing values for this part.

1. 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.

```
data$day <- weekdays(data$date)
len2 <- length(data$date)

for (j in 1:len2){
  if(data$day[j]=="Sunday"){
    data$wkend[j] <- "weekend"
  } else if (data$day[j]=="Saturday"){
    data$wkend[j] <- "weekend"
  }else {
    data$wkend[j] <- "weekday"}
}
```

2. 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.

```
intervalStep_wk <- aggregate(stepsImputed ~ interval+wkend, data=data, FUN="mean", na
.rm=TRUE)

library(lattice)

xyplot(stepsImputed ~ interval | wkend, data=intervalStep_wk, type = "l", layout= c(1
, 2), xlab="Interval",ylab="Number of steps")
```


weekend
weekday
Number of steps
Interval
0
50
100
150
200
0
500
1000
1500
2000