Introduction

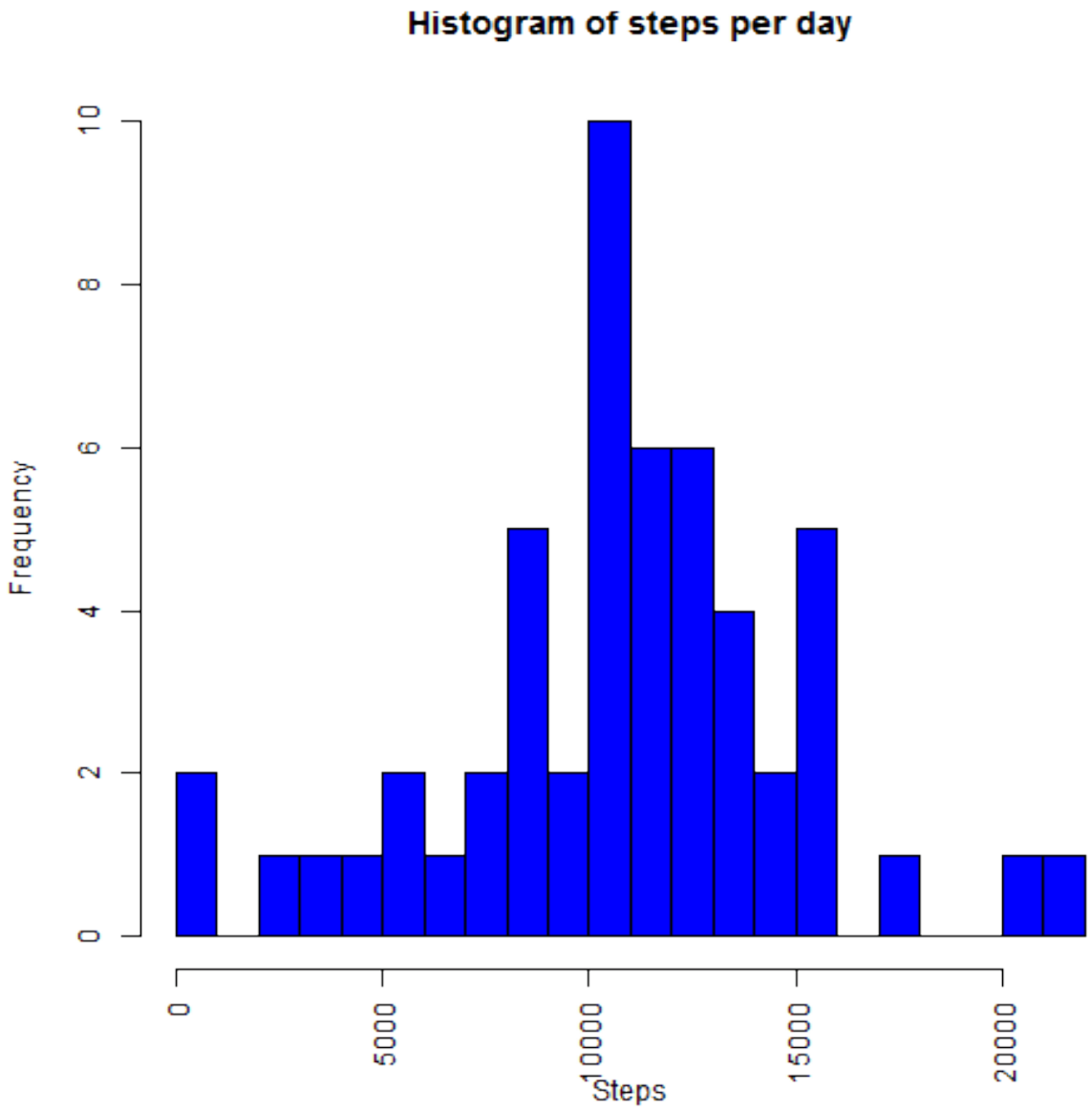
Introduction

This is the first assignment for the course reproductible research. The first step is to load the data. I also create another set eliminating the rows in which there is no value for steps.

```
data <- read.csv("activity.csv", header=T)
dat<- data[!is.na(data$steps),]
```

First question: What is mean total number of steps taken per day?

```
stepsday <-  aggregate(dat$steps, list(dat$date), sum)
colnames(stepsday) <- c('date', 'sumsteps')
plot <- hist(stepsday$sumsteps, breaks =20, las=3, xlab="Steps", main="Histogram of steps per day", col=
```




```
mean(stepsday$sumsteps)
```

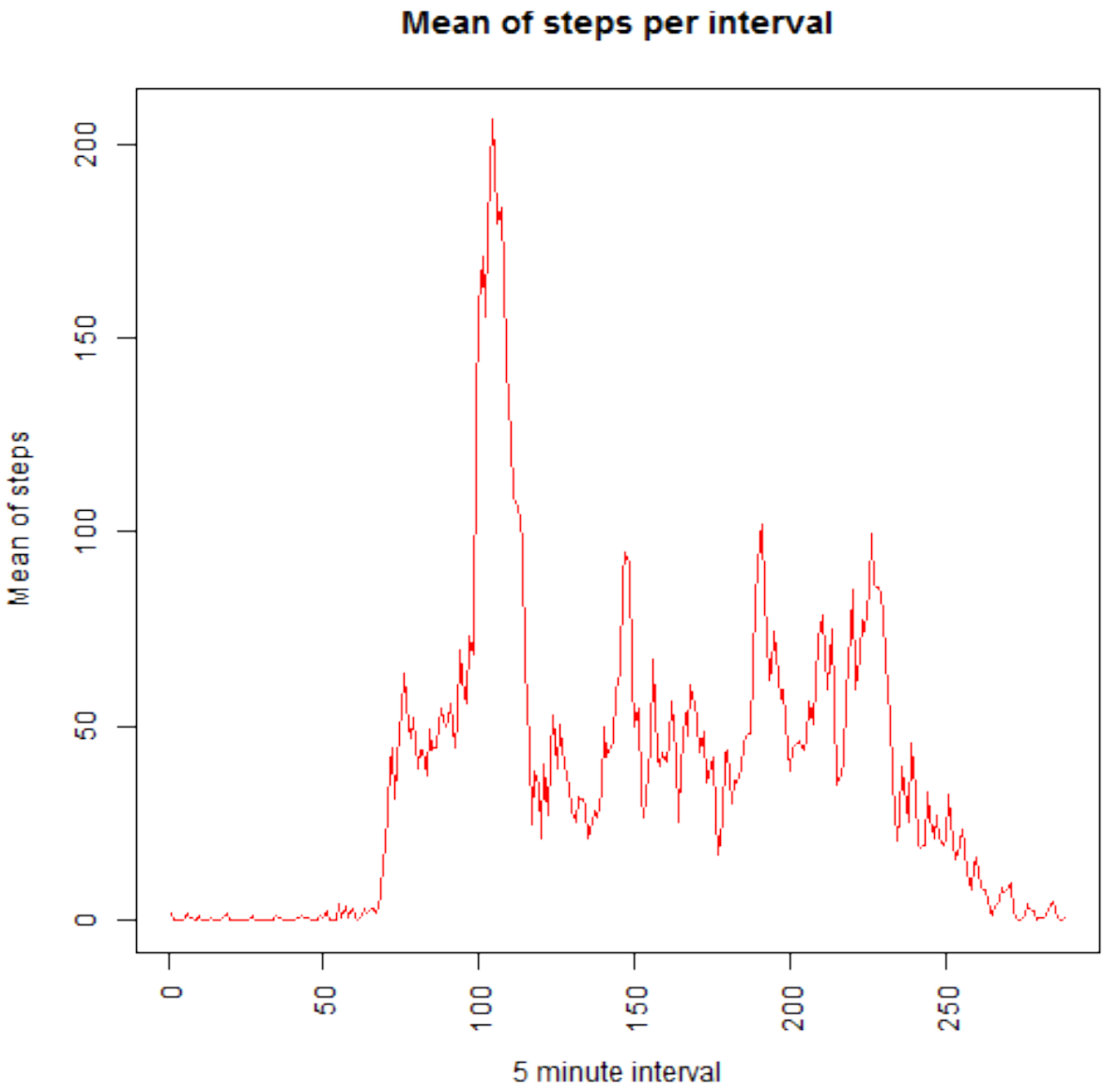
```
## [1] 10766
```

```
median(stepsday$sumsteps)
```

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

Second question: What is the average daily activity pattern?

```
stepsinterval <-  aggregate(dat$steps, list(dat$interval), mean)
colnames(stepsinterval) <- c('interval', 'meansteps')
plot2 <- plot(stepsinterval$meansteps, type = "l", las=3, main="Mean of steps per interval", xlab="5 mi
```



As observed in the graph, 104 is the 5 minute interval with the largest mean (206.17 steps).

Imputing missing values

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

```
nrow(data[is.na(data$steps),])
```

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

The next step is to fill the NA values. The strategy chosen has been fill them in with the mean of the 5-minute interval. I cerate another data frame data2 for clarity reasons.

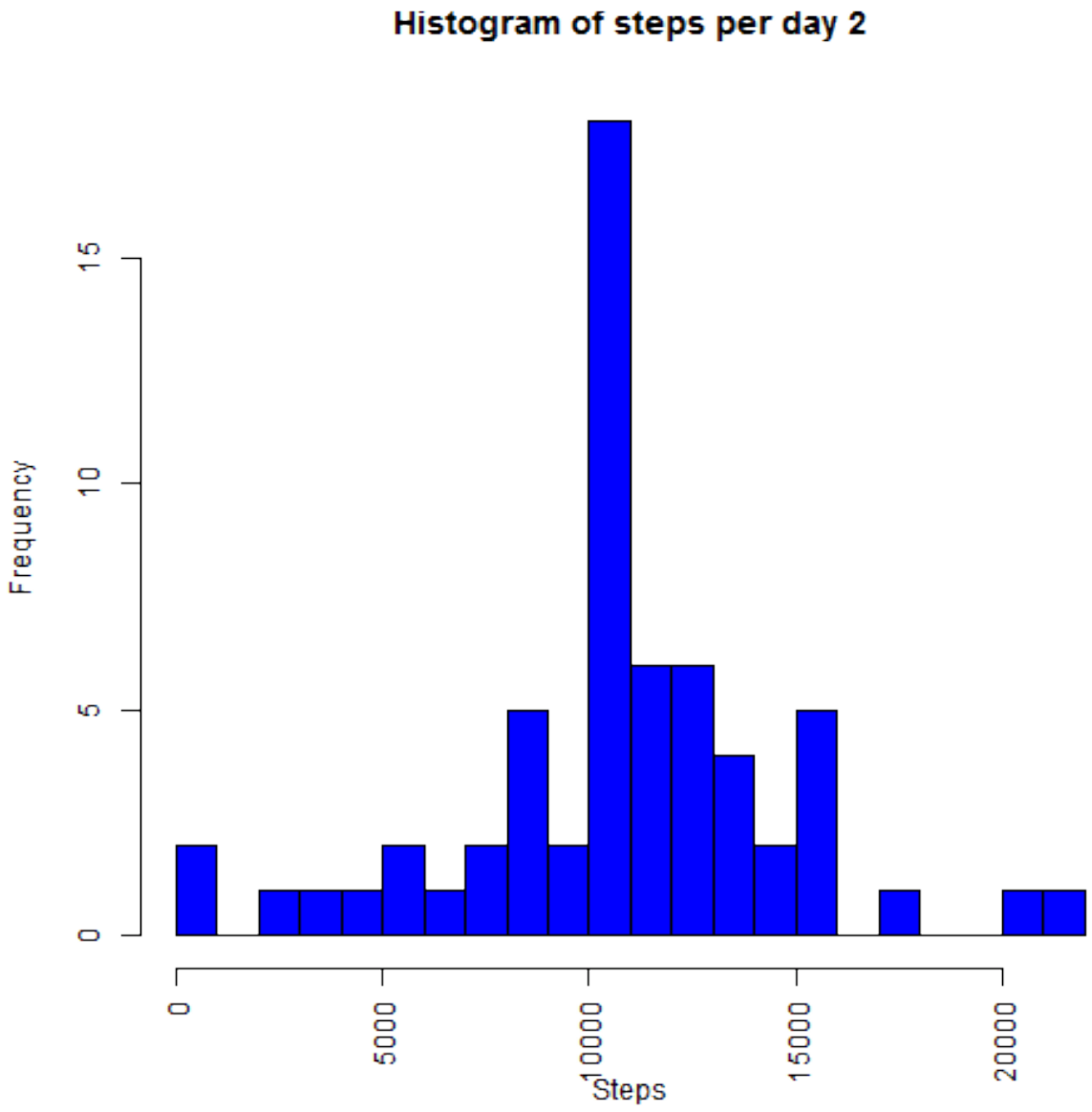
```
data2 <- merge(data, stepsinterval, by = "interval")
data2 <- within(data2 , steps[is.na(data2$steps)] <- meansteps[is.na(data2$steps)])
```

And now I create another data set that is equal to the original dataset but with the missing data filled in.

```
data3 <- data.frame(data2$steps, data2$date, data2$interval)
colnames(data3) <- c('steps', 'date', 'interval')
```

With the new data, I do a histogram again and I report the mean and the median.

```
stepsday <-  aggregate(data3$steps, list(data3$date), sum)
colnames(stepsday) <- c('date', 'sumsteps')
plot <- hist(stepsday$sumsteps, breaks =20, las=3, xlab="Steps", main="Histogram of steps per day 2", c
```



```
mean(stepsday$sumsteps)
```

```
## [1] 10766
```

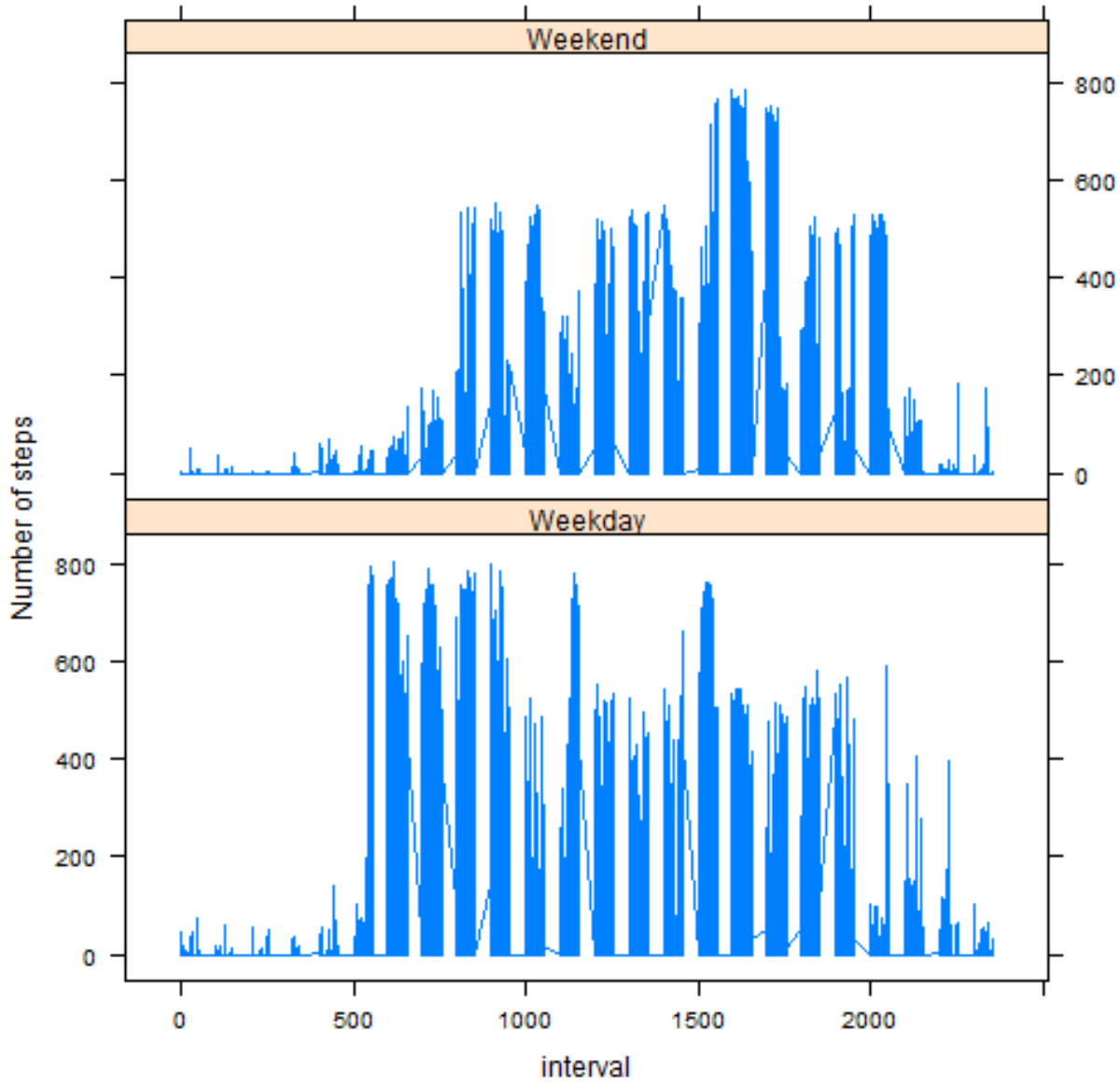
```
median(stepsday$sumsteps)
```

```
## [1] 10766
```

We can see in the histogram that the shape is the same as in the first histogram but the frequency has increased. This is due to the larger number of data points as in the first histogram, values NA were not included anywhere in the graph. The mean is the same. Since we have used the mean per interval, it makes sense that this value is not very affected. The median has sightly increased when including the mean per interval which makes us think that the original data was sightly skewed. The total daily number of steps will be a more centered distribution when imputing missing data

Are there differences in activity patterns between weekdays and weekends?

```
data3$weekday <-ifelse((  (weekdays(as.Date(data3$date))) %in% c('Monday','Tuesday','Wednesday','Thursda
xyplot(steps ~ interval | weekday, data = data3, layout = c(1, 2), type = "l" , ylab="Number of steps")
```


Weekend
Weekday
Number of steps
interval
0
200
400
600
800
0
500
1000
1500
2000