Reproducible Research Peer Assessment 1 - Analyzing Activity Monitoring Device Data

Introduction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the “quantified self” movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But this data remains under-utilized because the raw data is hard to obtain and there are limited tools and statistical methods available for interpreting the data. This assignment makes 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.

Data

The data for this assignment can be downloaded from the course web site: Dataset: Activity monitoring data [52K] The variables included in this dataset are: steps: Number of steps taking in a 5-minute interval (missing values are coded as NA) date: The date on which the measurement was taken in YYYY-MM-DD format interval: Identifier for the 5-minute interval in which measurement was taken The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

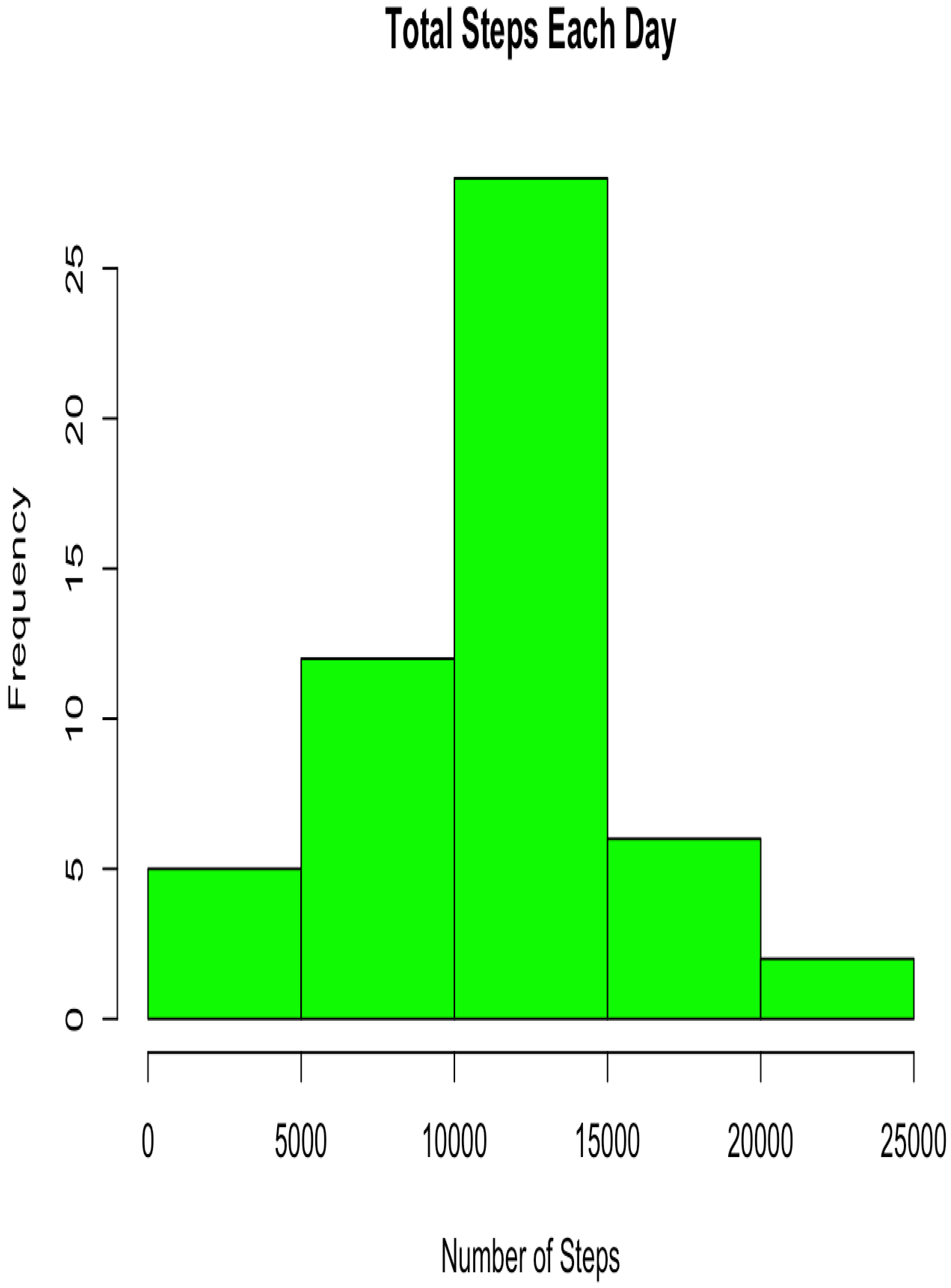
```
# Set Global Echo = On


# Load data
if (!file.exists("activity.csv") )
    {
     dlurl <- 'http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip'
     download.file(dlurl,destfile='repdata%2Fdata%2Factivity.zip',mode='wb')
     unzip('repdata%2Fdata%2Factivity.zip')
    }


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

3a.What is mean total number of steps taken per day?

```
steps_by_day <- aggregate(steps ~ date, data, sum)
hist(steps_by_day$steps, main = paste("Total Steps Each Day"), col="green",xlab="Number o
f Steps")
```


Total Steps Each Day
Frequency
0
5
10
15
20
25
0
5000
10000
15000
20000
25000
Number of Steps

```
rmean <- mean(steps_by_day$steps)
rmean
```

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

```
rmedian <- median(steps_by_day$steps)
rmedian
```

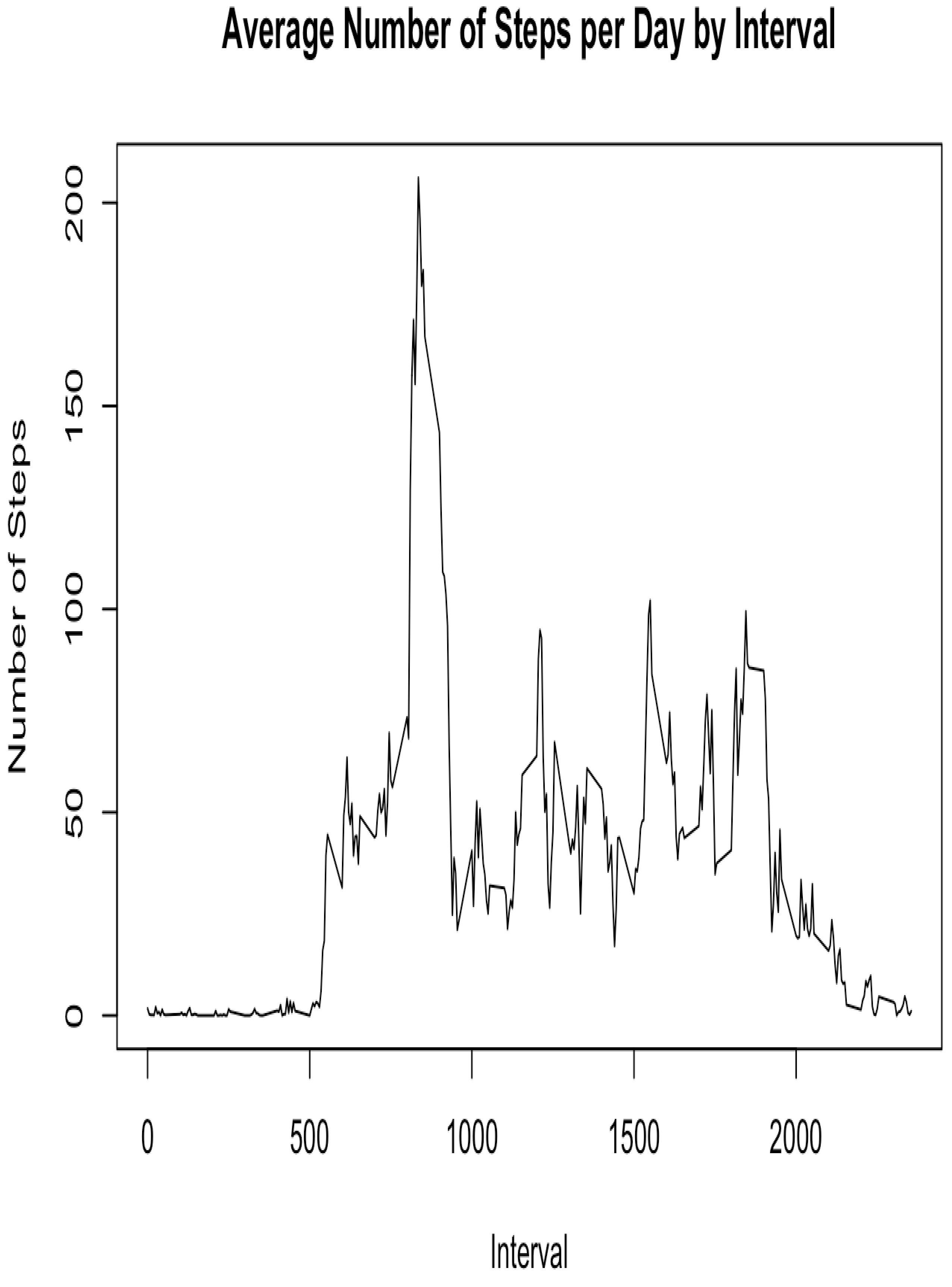
```
## [1] 10765
```

The mean is 1.076618910^{4} and the median is 10765

3b.What is the average daily activity pattern?

1. Calculate average steps for each interval for all days
2. Plot the Average Number Steps per Day by Interval
3. Find interval with most average steps

```
steps_by_interval <- aggregate(steps ~ interval, data, mean)
plot(steps_by_interval$interval,steps_by_interval$steps, type="l", xlab="Interval", ylab=
"Number of Steps",main="Average Number of Steps per Day by Interval")
```


Average Number of Steps per Day by Interval
Number of Steps
0
50
100
150
200
0
500
1000
1500
2000
Interval

```
max_interval <- steps_by_interval[which.max(steps_by_interval$steps),1]
max_interval
## [1] 835
```

The interval with most steps is 835

3c.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 NAs) 2. Using Mean for the day compute missing values 3. Create a new dataset that is equal to the original dataset but with the missing data filled in. 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.

1.Calculate and report the total number of missing values in the dataset

```
NATotal <- sum(!complete.cases(data))
NATotal
## [1] 2304
```

Total Number of Missing values are 2304

2.Using Mean for the day compute missing values

```
StepsAverage <- aggregate(steps ~ interval, data = data, FUN = mean)
fillNA <- numeric()
for (i in 1:nrow(data)) {
    obs <- data[i, ]
    if (is.na(obs$steps)) {
        steps <- subset(StepsAverage, interval == obs$interval)$steps
    } else {
        steps <- obs$steps
    }
    fillNA <- c(fillNA, steps)
}
```

3. Create a new dataset including the imputed missing values

```
new_activity <- data
new_activity$steps <- fillNA
```

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.

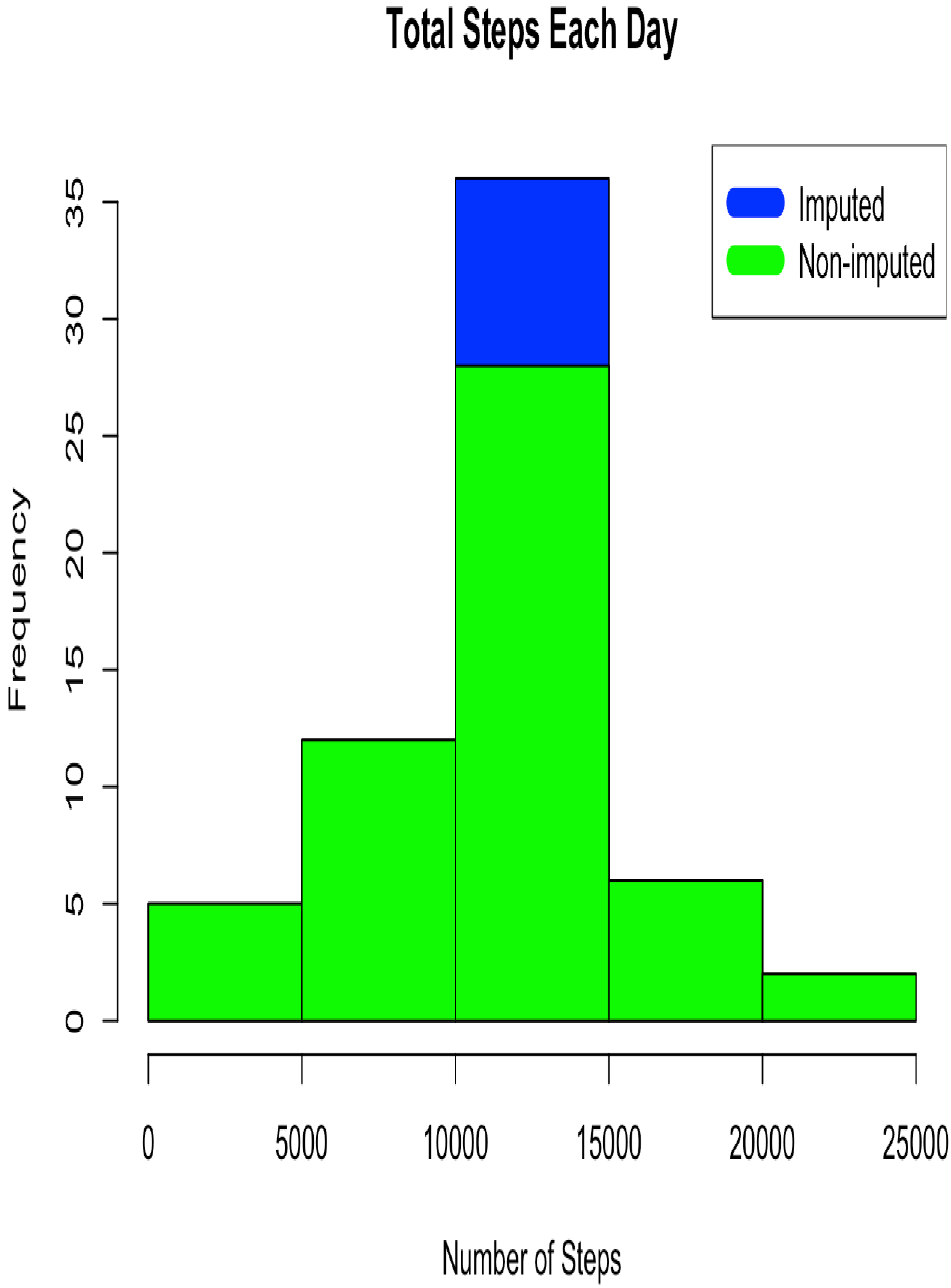
```
StepsTotalUnion <- aggregate(steps ~ date, data = new_activity, sum, na.rm = TRUE)

hist(StepsTotalUnion$steps, main = paste("Total Steps Each Day"), col="blue", xlab="Number of Steps")

#Create Histogram to show difference.

hist(steps_by_day$steps, main = paste("Total Steps Each Day"), col="green", xlab="Number of Steps", add=T)

legend("topright", c("Imputed", "Non-imputed"), col=c("blue", "green"), lwd=10)
```


Total Steps Each Day
Imputed
Non-imputed
Frequency
0
5
10
15
20
25
30
35
0
5000
10000
15000
20000
25000
Number of Steps

Calculate Mean

```
rmeantotal <- mean(StepsTotalUnion$steps)

rmeantotal

## [1] 10766.19
```

Calculate Median

```
rmediantotal <- median(StepsTotalUnion$steps)

rmediantotal

## [1] 10766.19
```

Do these values differ from the estimates from the first part of the assignment?

```
rmediandiff <- rmediantotal - rmedian

rmediandiff

## [1] 1.188679

rmeandiff <- rmeantotal - rmean

rmeandiff

## [1] 0
```

Ans. The mean(Mean Var: 0) is the same however the median does have a small variance(Median Var:1.1886792). between the total which includes the missing values to the base
What is the impact of imputing missing data on the estimates of the total daily number of steps?
On observation the impact of the missing data has the biggest effect on the 10000 - 150000 step interval and changes frequency from 27.5 to 35 a variance of 7.5

3d.Are there differences in activity patterns between weekdays and weekends?

Created a plot to compare and contrast number of steps between the week and weekend. There is a higher peak earlier on weekdays, and more overall activity on weekends.

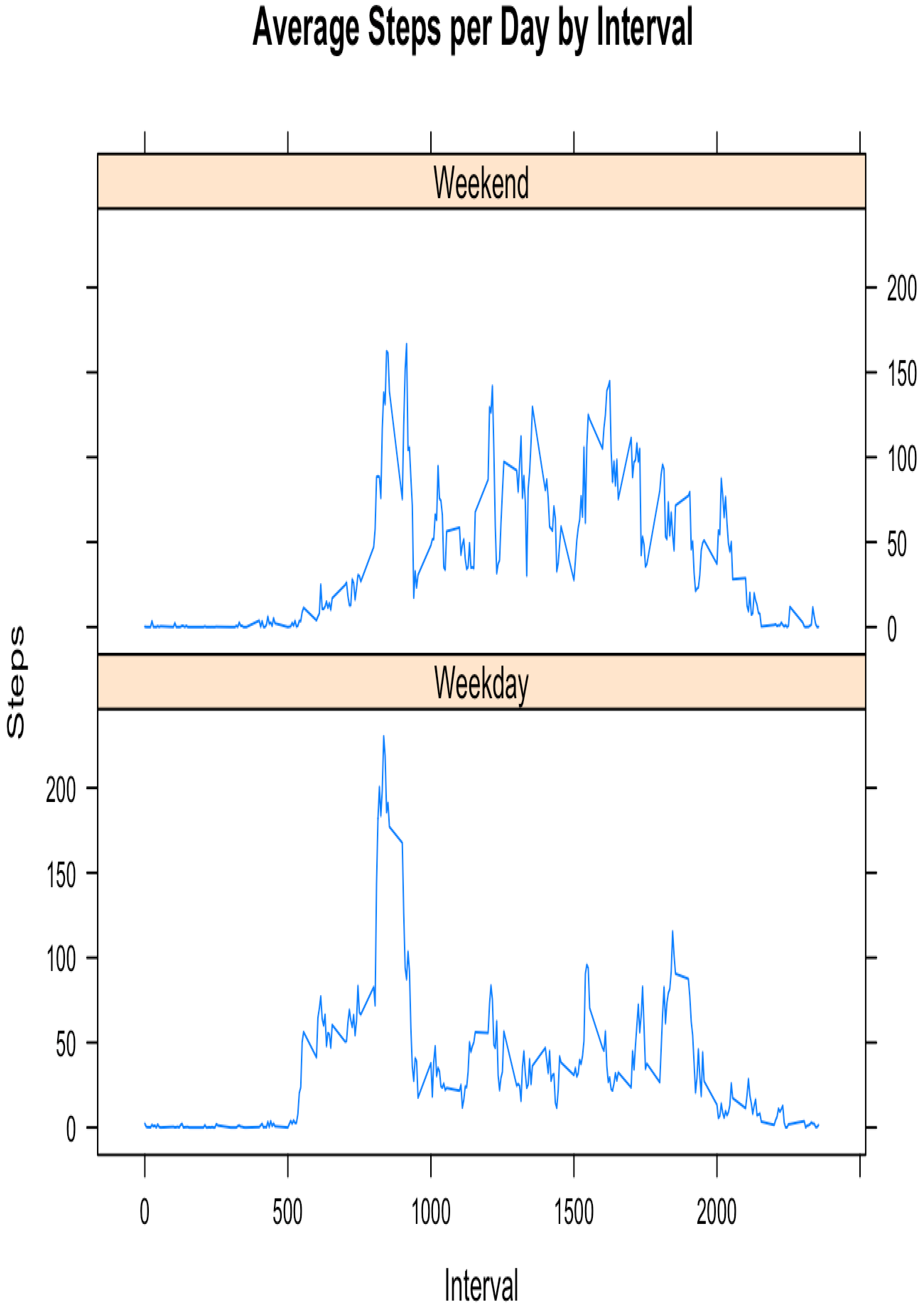
```
weekdays <- c("Monday", "Tuesday", "Wednesday", "Thursday",
              "Friday")

new_activity$dow = as.factor(ifelse(is.element(weekdays(as.Date(new_activity$date)),weekdays), "Weekday", "Weekend"))

StepsTotalUnion <- aggregate(steps ~ interval + dow, new_activity, mean)

library(lattice)

xyplot(StepsTotalUnion$steps ~ StepsTotalUnion$interval|StepsTotalUnion$dow, main="Average Steps per Day by Interval",xlab="Interval", ylab="Steps",layout=c(1,2), type="l")
```


Average Steps per Day by Interval
Weekend
Weekday
Steps
Interval
0
50
100
150
200
0
500
1000
1500
2000