Reproducable Research : Project 1

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.

Loading and preprocessing the data

Process/transform the data (if necessary) into a format suitable for your analysis

```
mData <- read.csv("activity.csv", header = TRUE, stringsAsFactors=FALSE)
mData$date <- as.Date(mData$date)
str(mData)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

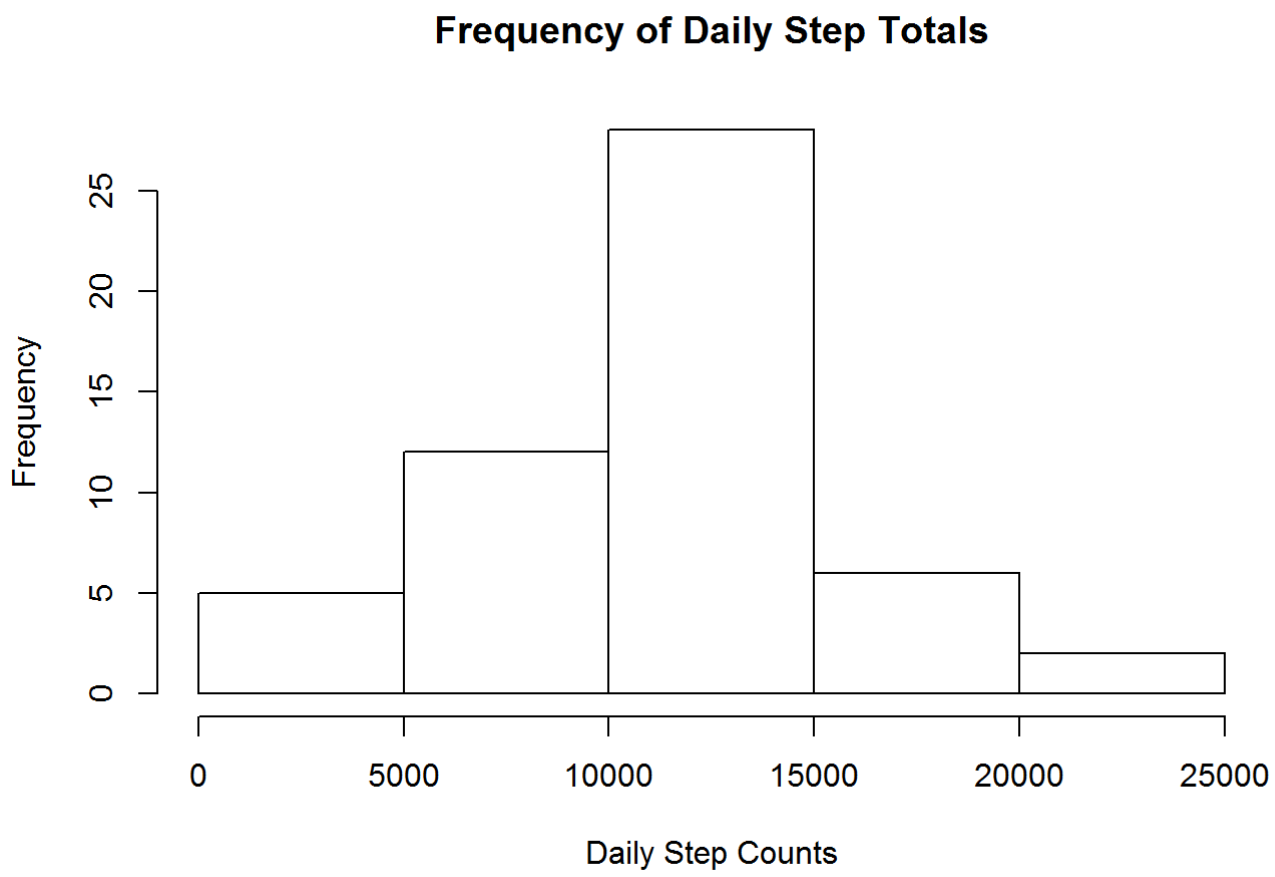
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
by_Date <- group_by(mData, date)

s_Steps <- summarise(by_Date, sum(steps))

head(s_Steps, 3)
```

```
## # A tibble: 3 × 2
##         date `sum(steps)`
##       <date>        <int>
## 1 2012-10-01           NA
## 2 2012-10-02          126
## 3 2012-10-03        11352
```

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

```
hist(s_Steps$`sum(steps)`, breaks=6, main = "Frequency of Daily Step Totals", xlab="Daily Step C
ounts")
```



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

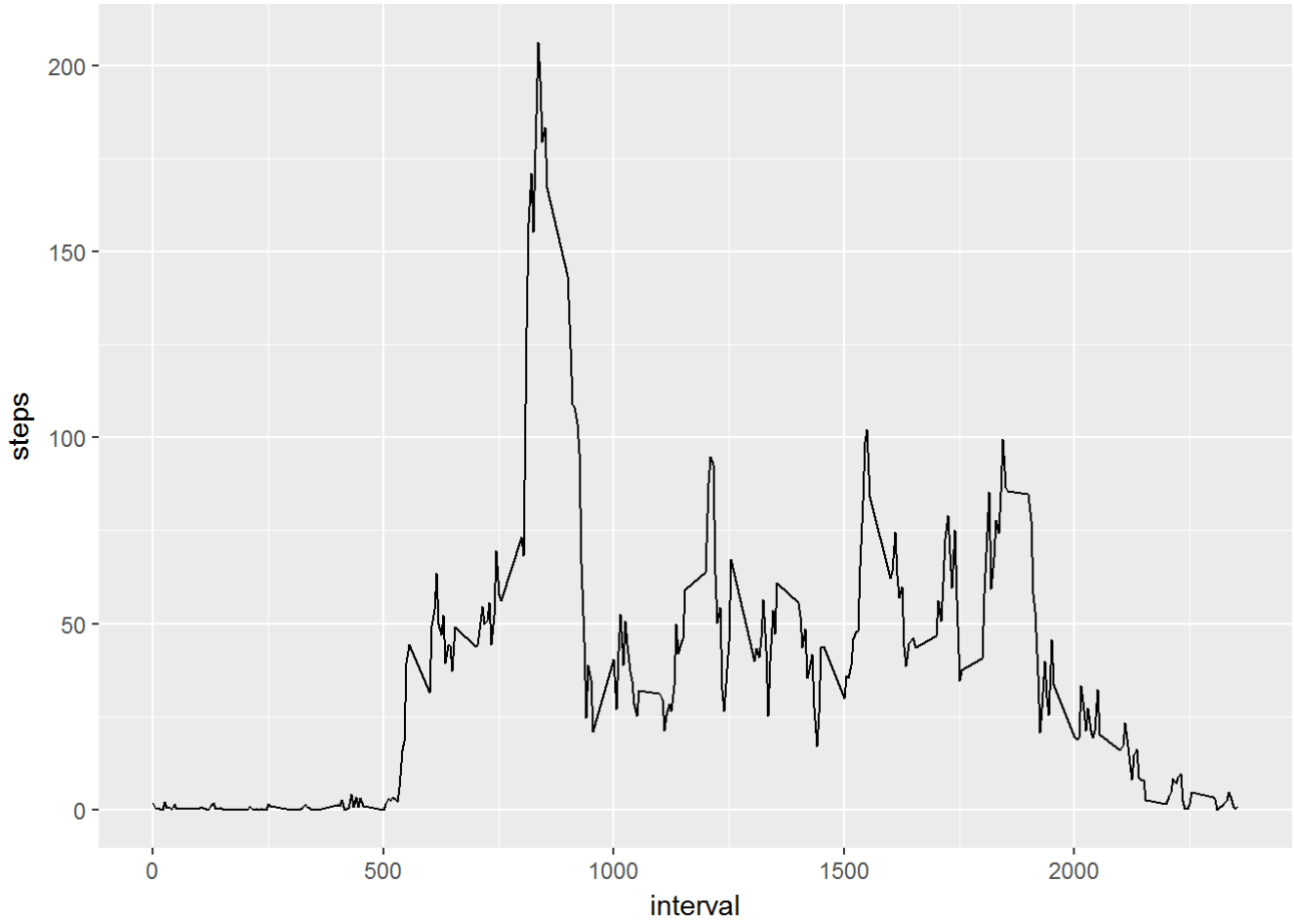
```
summary(s_Steps$`sum(steps)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      41    8841   10760   10770   13290   21190       8
```

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)

```
library(ggplot2)

by_5min <- aggregate(steps ~ interval, data = mData, FUN = mean, na.rm = TRUE)

qplot(interval, steps, data=by_5min, geom="line" )
```



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

```
by_5min$interval[which.max(by_5min$steps)]
```

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

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)

```
sum(is.na(mData$steps))
```

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

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.

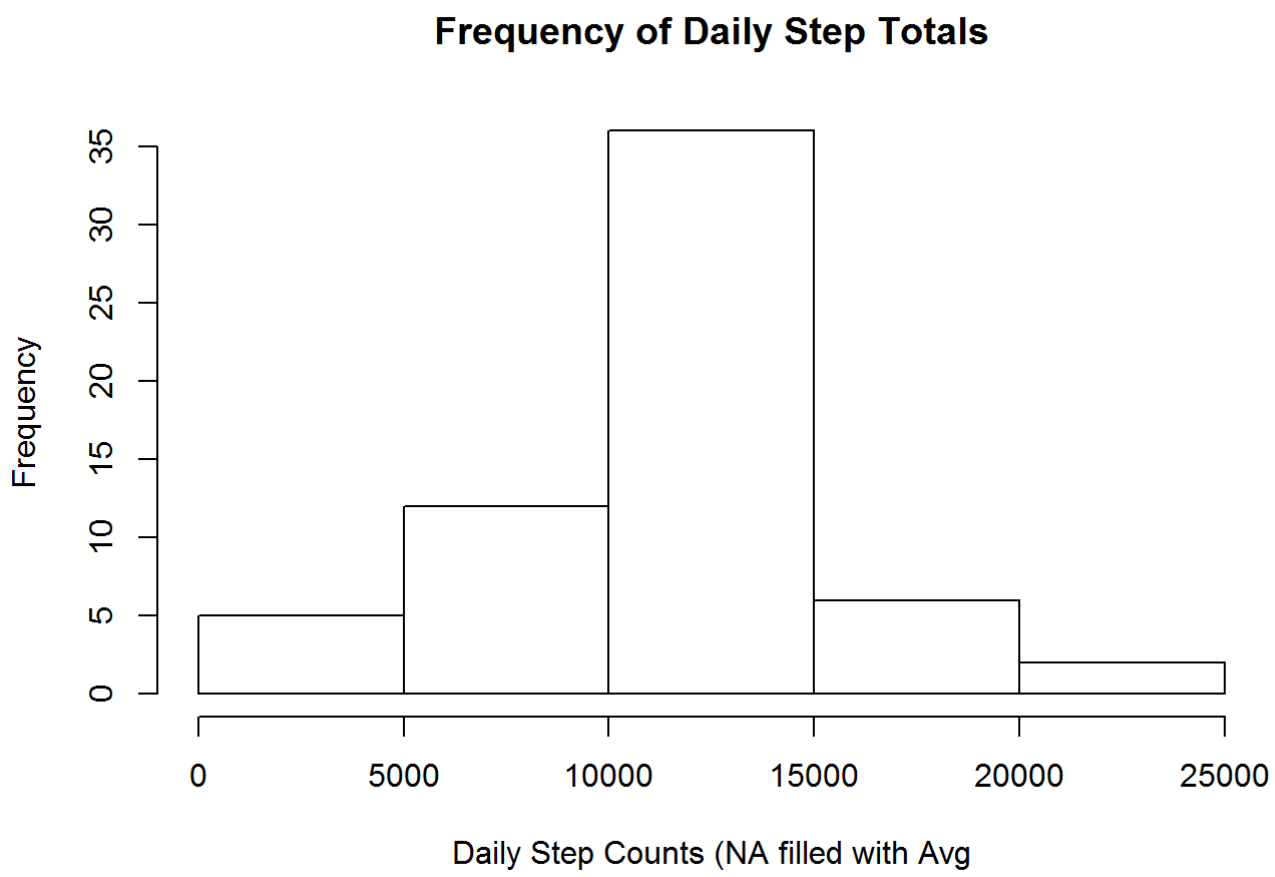
```
# create logical vector for steps values that are NA
nData <- mData

rowsNA <- is.na(nData$steps)
```

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

```
# the mean of daily steps is 10770 from Q#3 above. There are 288 5 min intervals per day.

nData$steps[rowsNA] = 10770/288
```

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.

```
by_Datn <- group_by(nData, date)

t_Steps <- summarise(by_Datn, sum(steps))

hist(t_Steps$`sum(steps)`, breaks=6, main = "Frequency of Daily Step Totals", xlab="Daily Step C
ounts (NA filled with Avg")
```




```
summary(t_Steps$`sum(steps)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      41    9819   10770   10770   12810   21190
```

5. 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?

Answer: The histrogram frequencies are higher, but the MEAN did not change.

Are there differences in activity patterns between weekdays and weekends?

1. Create a new factor variable in the dataset with two levels â“ âœweekdayâ and âœweekendâ indicating whether a given date is a weekday or weekend day. (Use the dataset with the filled-in missing values for this part.)

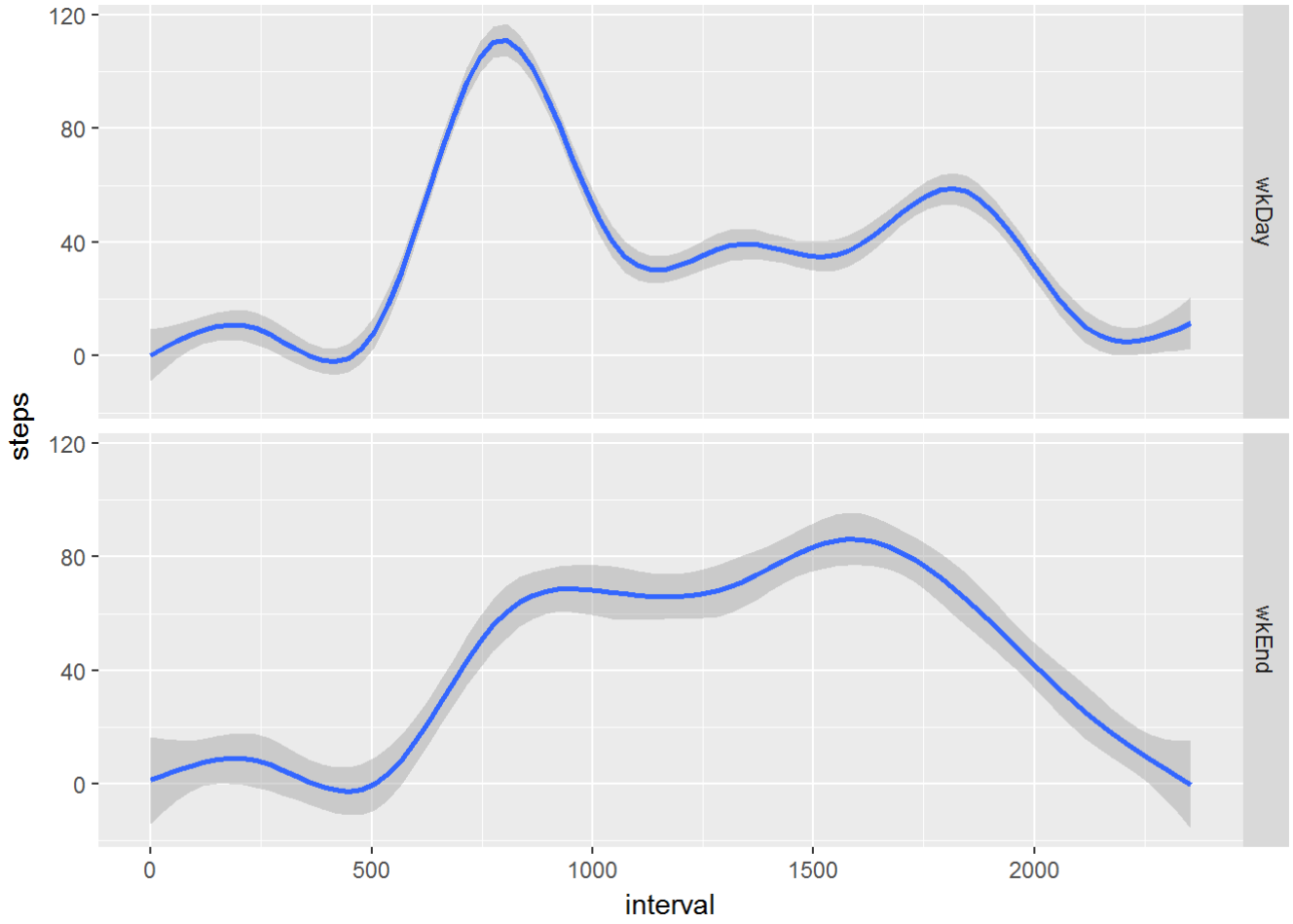
```
nData$day <- weekdays(nData$date)

nData$week <- NULL

nrows <- nrow(nData)

for(i in 1:nrows){

        if(nData$day[i] == "Saturday" | nData$day[i] == "Sunday") {

         nData$week[i] <- "wkEnd"

        }

        else
        {
         nData$week[i] <- "wkDay"

        }

}

nData$week <- as.factor(nData$week)

str(nData)
```

```
## 'data.frame':    17568 obs. of  5 variables:
##  $ steps   : num  37.4 37.4 37.4 37.4 37.4 ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
##  $ day     : chr  "Monday" "Monday" "Monday" "Monday" ...
##  $ week    : Factor w/ 2 levels "wkDay","wkEnd": 1 1 1 1 1 1 1 1 1 1 ...
```

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).

```
qplot(interval, steps, data=nData, geom="smooth", facets =week~.)
```

```
## `geom_smooth()` using method = 'gam'
```




```
# strangely, if I use geom="line", the result is incorrect
```

end