“Reproducible,Research: Peer Assessment 1”

by 'Gopalkriz'

Date:

[1] “2015-08-16”

Coursera - Reproducible Research (Aug 2015 batch)

Project Introduction

Dataset: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip

ProjectFolder: https://github.com/gopalkriz/RepData_PeerAssessment1

Local Repository:

[1] “Cloned from github D:/VIVEK/DataAnalysis/Coursera/5 RepRes/RepData_PeerAssessment1”

##Setting Global Options
arguments set for code chunks to be included in output

```
library(knitr)
opts_chunk$set(echo = TRUE, results = "asis")
```

##Set Working Directory with required Folder Path

```
WDoriginal<- getwd()
setwd("D:/VIVEK/DataAnalysis/Coursera/5 RepRes/RepData_PeerAssessment1"
```

Project Introduction

About: “quantified self” movement - collect data of personal activity using monitoring devices Data: collected at 5 minute intervals through out the day for two months from an anonymous individual during October and November, 2012. Variable: steps taken in 5 minute intervals each day.

Dataset: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip ProjectFolder: https://github.com/gopalkriz/RepData_PeerAssessment1 Local Repository: [1] “Cloned from github D:/VIVEK/DataAnalysis/Coursera/5 RepRes/RepData_PeerAssessment1”

Loading and preprocessing the data

###Upload the Data

```
unzip("activity.zip")
Data <- read.csv("activity.csv", header = TRUE, sep =",")
```

###Understand the Data dataset: comma-separated-value (CSV) file 17,568 observations Three variables included in this dataset are: steps: taking in a 5-minute interval (missing values are coded as NA) date: on which the measurement was taken in YYYY-MM-DD format interval: continuous counter Identifier of time for the measurement

```
object.size(Data)
```

214232 bytes

```
class(Data)
```

[1] “data.frame”

```
dim(Data); nrow(Data); ncol(Data)
```

[1] 17568 3 [1] 17568 [1] 3

```
names(Data)
```

[1] “steps” “date” “interval”

```
str(Data)
```

'data.frame': 17568 obs. of 3 variables: $ steps : int NA NA NA NA NA NA NA NA NA … $ date : Factor w/ 61 levels “2012-10-01”,“2012-10-02”,..: 1 1 1 1 1 1 1 1 1 … $ interval: int 0 5 10 15 20 25 30 35 40 45 …

```
head(Data)
```

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

```
tail(Data)
```

```
  steps       date interval
```

17563 NA 2012-11-30 2330 17564 NA 2012-11-30 2335 17565 NA 2012-11-30 2340 17566 NA 2012-11-30 2345 17567 NA 2012-11-30 2350 17568 NA 2012-11-30 2355

```
summary(Data)
```

```
 steps                date          interval
```

Min. : 0.00 2012-10-01: 288 Min. : 0.0
1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8
Median : 0.00 2012-10-03: 288 Median :1177.5
Mean : 37.38 2012-10-04: 288 Mean :1177.5
3rd Qu.: 12.00 2012-10-05: 288 3rd Qu.:1766.2
Max. :806.00 2012-10-06: 288 Max. :2355.0
NA's :2304 (Other) :15840

##Question 1.0

What is mean total number of steps taken per day?

note: you can ignore the missing values in the dataset

1.1. Calculate the total number of steps taken per day

Rcode 1.1

```
daysteps <- aggregate(steps ~ date, data = Data, FUN = sum)
```

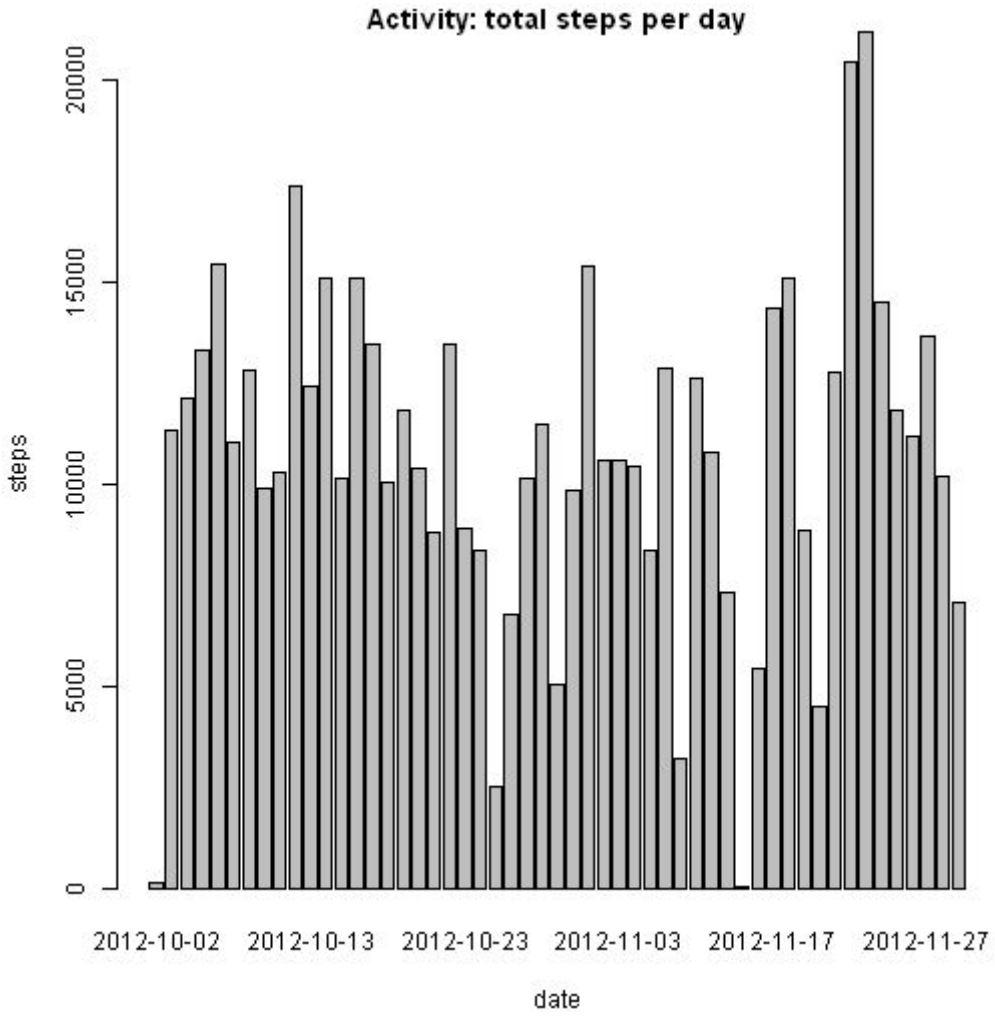
1.2. Plot the total number of steps taken per day

Barplot steps-datewise

Rcode 1.2a

```
par(mfrow = c(1, 1))
par(mar=c(4,4,1,0.5))

plot_001 <- barplot(daysteps$steps, names.arg = daysteps$date, xlab = '
```




```
dev.copy(png,'plot_001.png')
```

png 5

```
invisible(plot_001)
dev.off()
```

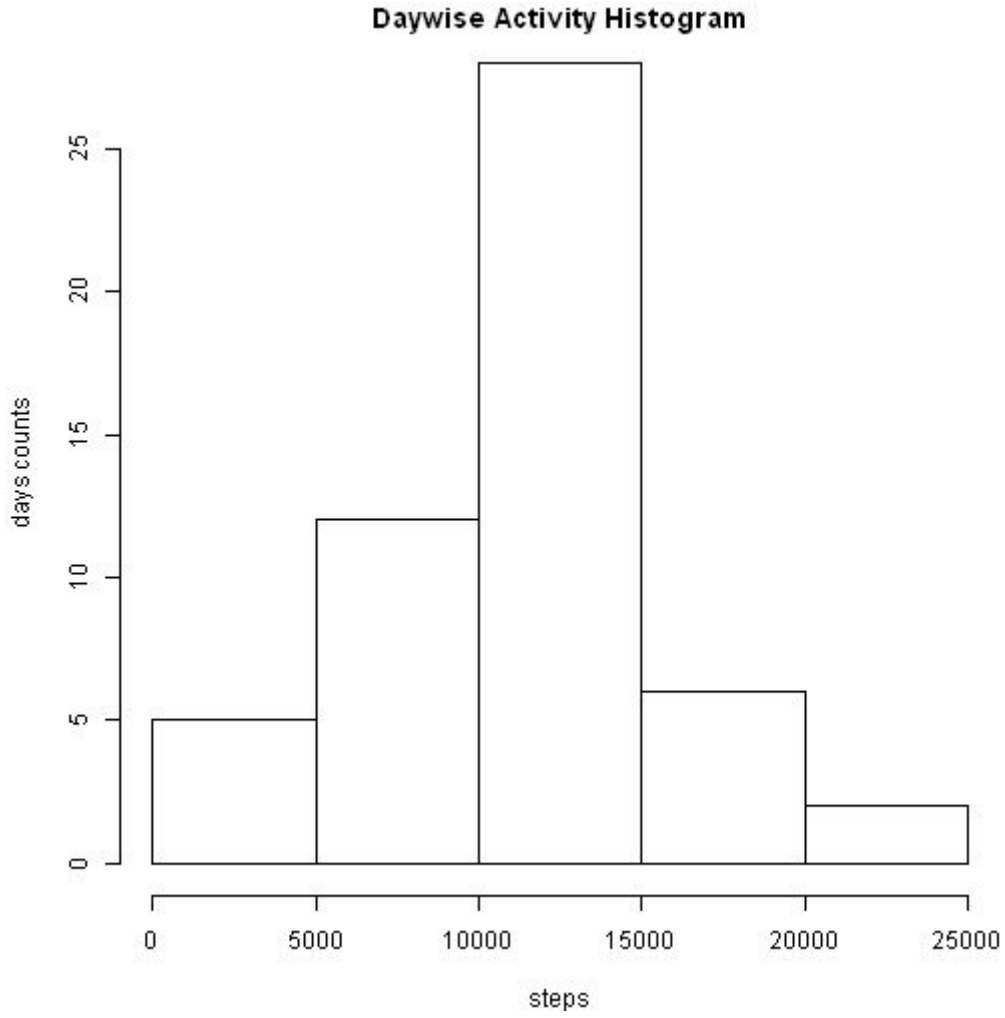
RStudioGD 2

Histogram steps per day

Rcode 1.2b

```
par(mfrow = c(1, 1))
par(mar=c(4,4,1,0.5))

plot_002 <- hist(daysteps$steps, xlab = "steps", ylab = "days counts",
```




```
dev.copy(png,'plot_002.png')
```

png 5

```
invisible(plot_002)
dev.off()
```

RStudioGD 2

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

Rcode 1.3

```
mean(daysteps$steps) #Ans: should be 10766.19
```

[1] 10766.19

```
median(daysteps$steps) #Ans:  should be 10765
```

[1] 10765

##Answer Mean total number of steps taken per day is as per output above.

##Question 2.0

What is the average daily activity pattern?

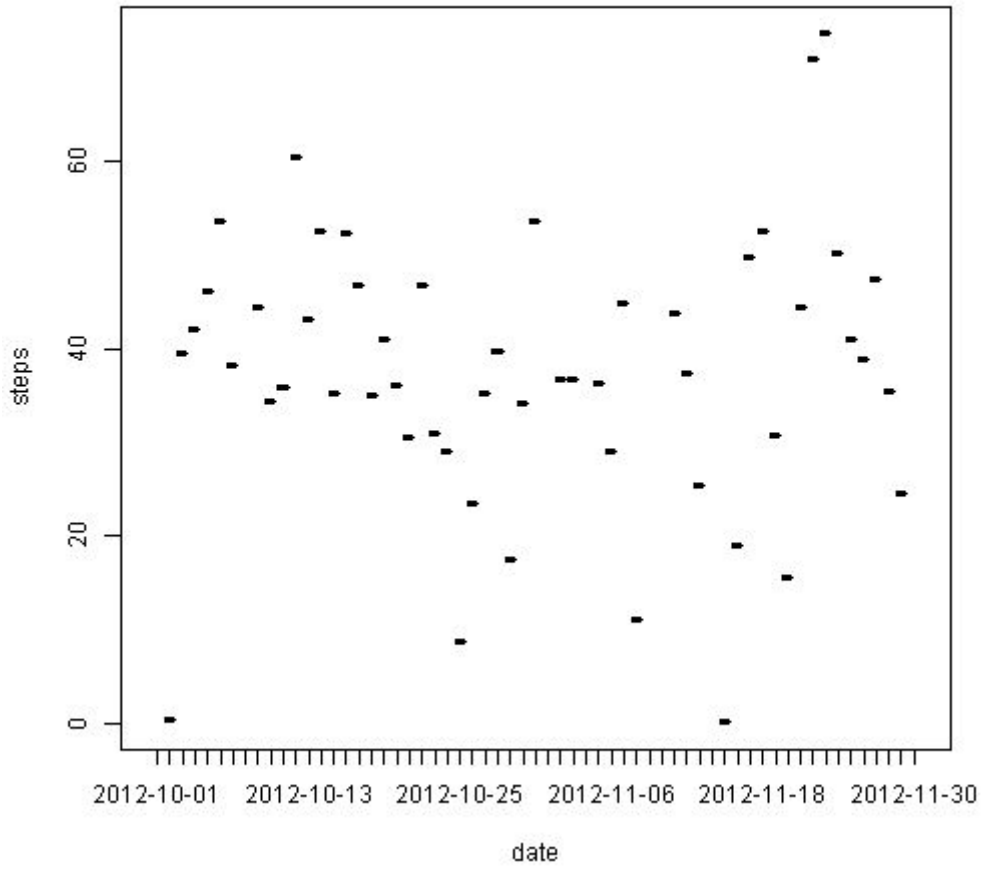
##Answer 2.0

```
dayAvg <- aggregate(steps ~ date, data = Data, FUN = mean)
head(dayAvg)
```

```
    date    steps
```

1 2012-10-02 0.43750 2 2012-10-03 39.41667 3 2012-10-04 42.06944 4 2012-10-05 46.15972 5 2012-10-06 53.54167 6 2012-10-07 38.24653

```
plot(dayAvg)
```



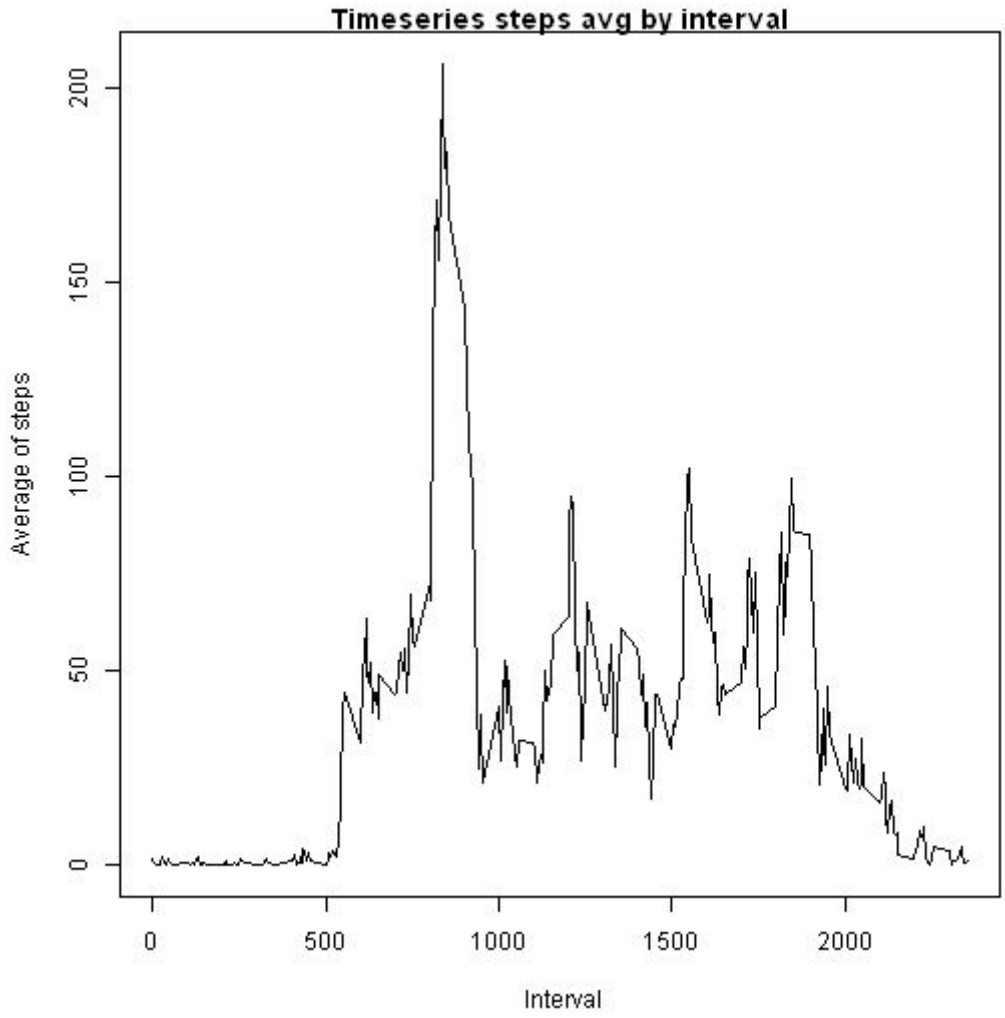
2.1 Make a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

Rcode 2.1

```
StepsGroup <- aggregate(steps ~ interval, data = Data, FUN = mean)
```

```
par(mfrow = c(1, 1))
par(mar=c(4,4,1,0.5))

plot_003 <- plot(StepsGroup, type = "l", xlab = " Interval", ylab= "Ave
```




```
dev.copy(png,'plot_003.png')
```

png 5

```
invisible(plot_003)
dev.off()
```

RStudioGD 2

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

Rcode 2.2

```
StepsGroup$interval[which.max(StepsGroup$steps)] #Ans: 835th Interval
```

[1] 835

```
print("with the value as average:")
```

[1] “with the value as average:”

```
StepsGroup$steps[which.max(StepsGroup$steps)] #Ans: 206.1698
```

[1] 206.1698 ##Answer 2.2 As per above output.

##Question 3.0

Imputing missing values

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

Rcode 3.1

```
sum(is.na(Data)) #Ans: 2304 NAs in entire dataset
```

[1] 2304

```
sum(is.na(Data$steps)) #col1 has all the NAs
```

[1] 2304

```
sum(is.na(Data$date)) #col2 does not have missing values
```

[1] 0

```
sum(is.na(Data$interval)) #col3 does not have missing values
```

[1] 0 ##Answer 3.1 There are 2304 NAs in the data set, all of them in the column with steps measurements

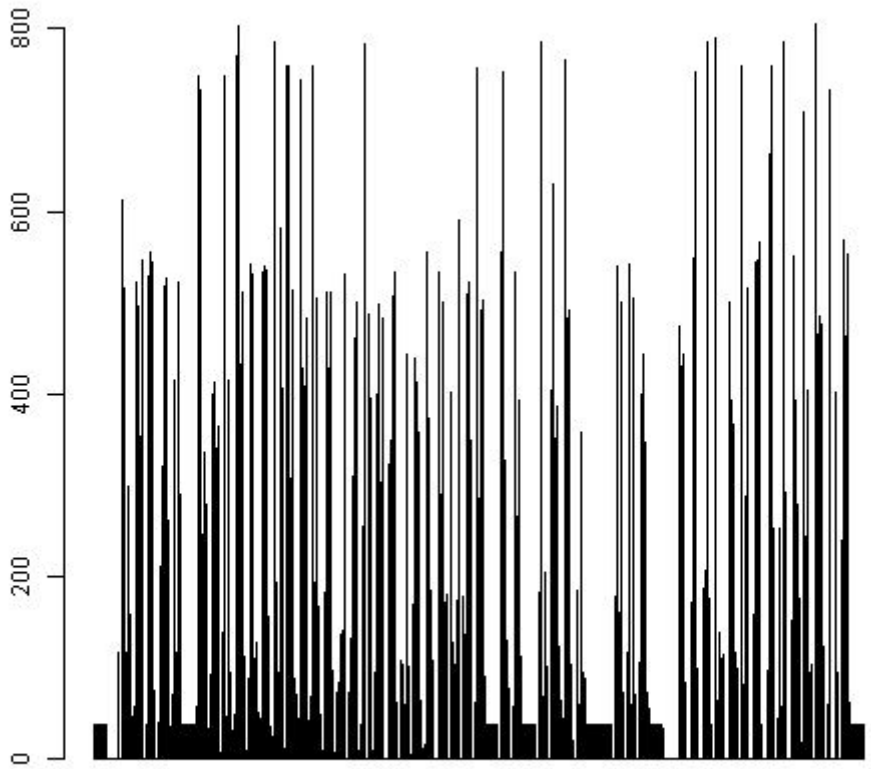
3.2 Devise a strategy for filling in all of the missing values in the dataset. ##Answer 3.2 Method used: means of 5-min intervals as fillers for missing values. Existing values will remain. NA values will be replaced by values from StepsGroup <- aggregate(steps ~ interval, data = Data, FUN = mean)

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

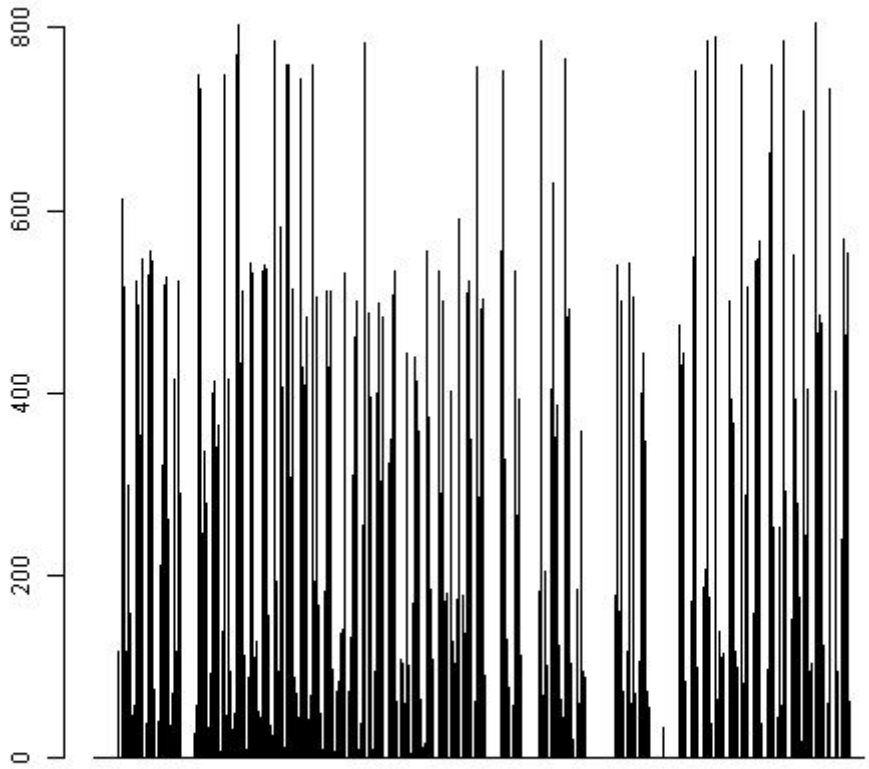
Rcode 3.3

```
library(Hmisc)

cleandata_mean <- Data
cleandata_mean$steps <- impute(Data$steps, fun=mean)
barplot(as.numeric(cleandata_mean$steps))
```




```
cleandata_median <- Data
cleandata_median$steps <- impute(Data$steps, fun=median)
barplot(as.numeric(cleandata_median$steps))
```



3.4a Make a histogram of the total number of steps taken each day.

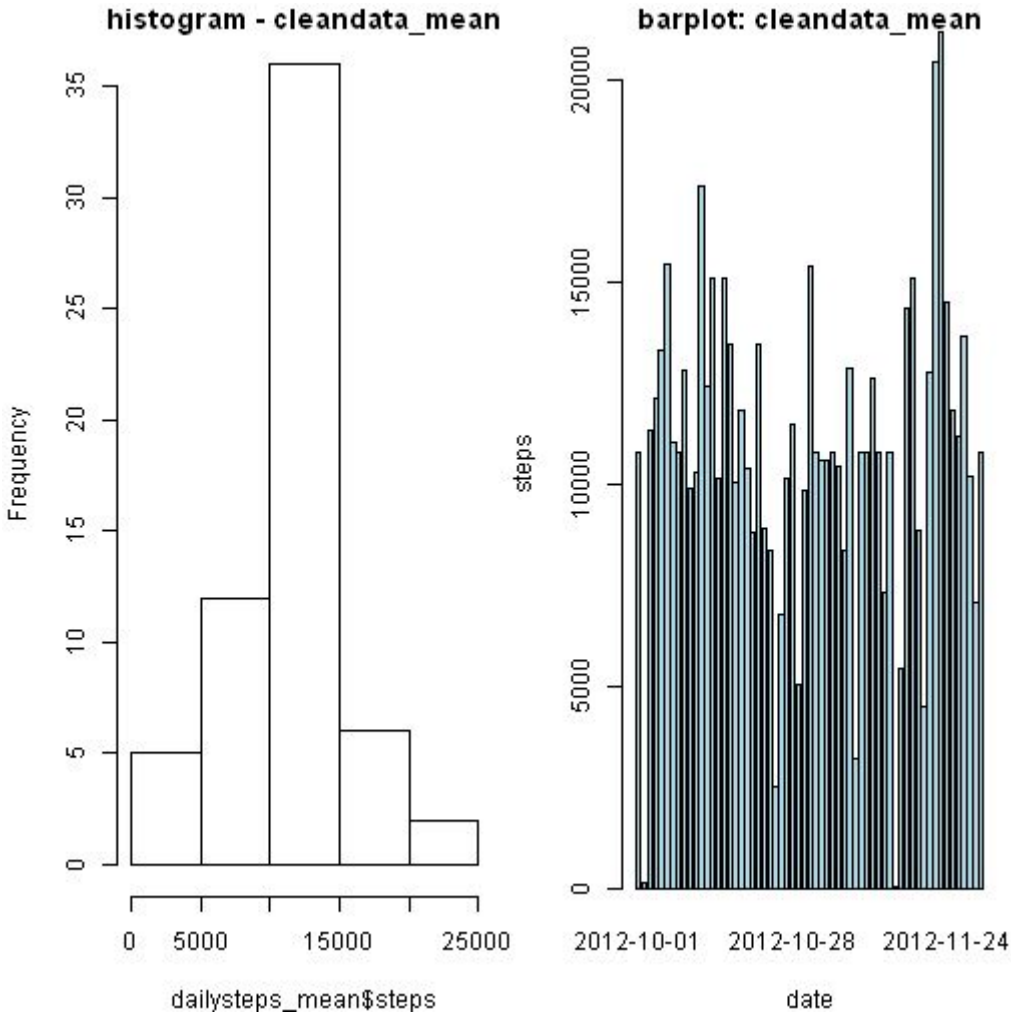
Rcode 3.4a

```
dailysteps_mean <- aggregate(steps ~ date, data = cleandata_mean, FUN =
dailysteps_median <- aggregate(steps ~ date, data = cleandata_median, F
```

```
par(mfrow = c(1, 2))
par(mar=c(4,4,1,0.5))

plot_004a <- hist(dailysteps_mean$steps, main = "histogram - cleandata_

plot_004b <- barplot(dailysteps_mean$steps, names.arg = dailysteps_mean
```




```
dev.copy(png,'plot_004.png')
```

png 5

```
dev.off()
```

RStudioGD 2

3.4b Calculate and report the mean and median total number of steps taken per day.

Rcode 3.4b

```
#Ans: This was 10766.19  from data with NA  (ie from daysteps$steps)
mean(dailysteps_mean$steps)
```

[1] 10766.19

```
#10766.19 mean replace NAs (no difference)
mean(dailysteps_median$steps)
```

[1] 9354.23

```
#9354.23 median replace NAs

#Ans: This was 10765 from data with NA (ie from daysteps$steps)
median(dailysteps_mean$steps)
```

[1] 10766.19

```
#10766.19 mean replace NAs (minimal difference)
median(dailysteps_median$steps)
```

[1] 10395

```
#10395 median replace NAs
```

###Selection of Mean as impute. Cleaned dataset is used for further analysis Data <- cleandata_mean

3.4c Do these values differ from the estimates from the first part of the assignment? ##Answer 3.4c When the impute is Mean is no / minimal difference When the impute is Median there is significant difference Hence, for this dataset, the choice of mean is a better option for impute.

3.4d What is the impact of imputing missing data on the estimates of the total daily number of steps? ##Answer 3.4d Comparing daysteps (data with NAs) vs dailysteps (cleaned dataset), we do not see much

diffrence. The impact in minimal for our stratergy of imputing missing data.

##Question 4.0

Are there differences in activity patterns between weekdays and weekends?

note: dataset with the filled-in missing values

4.1 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.

Create funtion to know daytype (weekday/weekend)

Rcode 4.1a

```
daytype <- function(date) {
    whichday <- weekdays(date)
    if (whichday %in% c("Monday", "Tuesday", "Wednesday", "Thursday",
        return("weekday")
    else (whichday %in% c("Saturday", "Sunday"))
        return("weekend")
}
```

Apply function_daytype on Data

Rcode 4.1b

```
Data$date <- as.Date(Data$date)
Data$daytype <- sapply(Data$date, FUN=daytype)

head(Data); table(Data$daytype)
```

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

weekday weekend 12960 4608

4.2 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).

Rcode 4.2

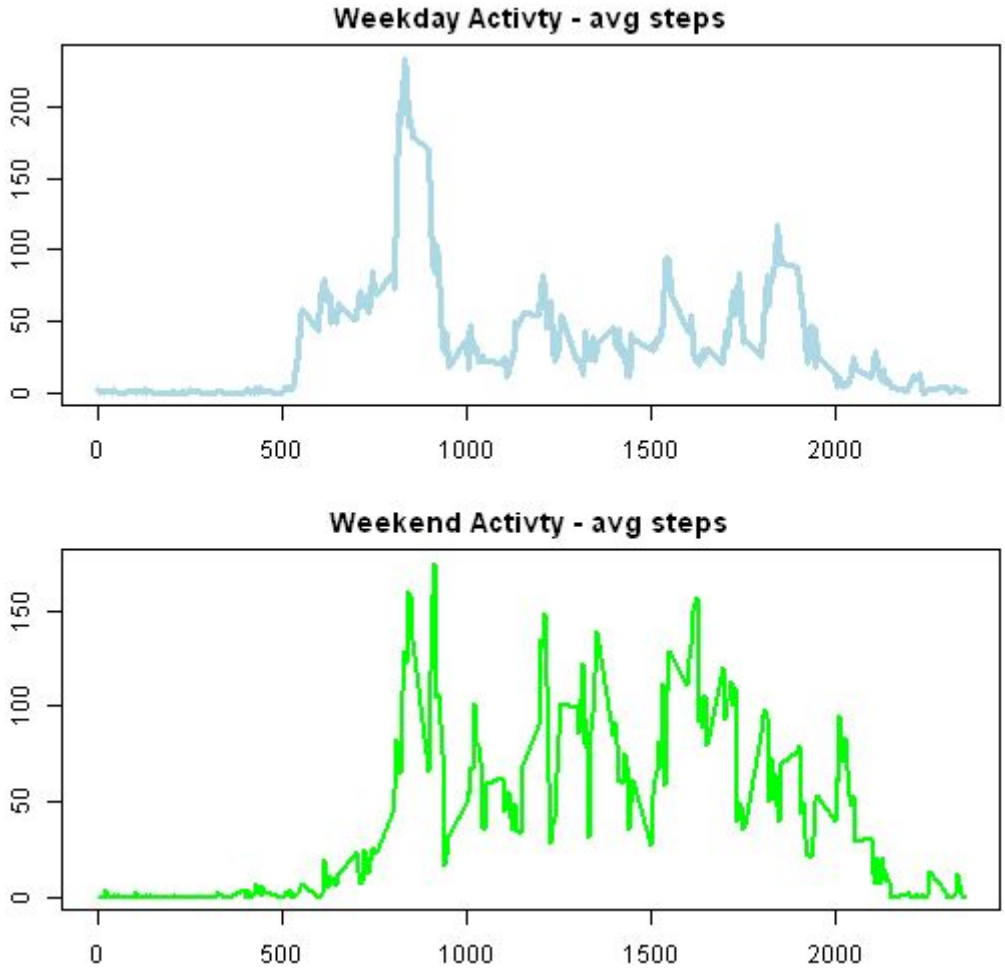
```
weekdayAvg <- aggregate(steps ~ interval, data = Data, subset = Data$da
weekendAvg <- aggregate(steps ~ interval, data = Data, subset = Data$da
```

Plotting the comparison

```
par(mfrow = c(2, 1))
par(mar=c(3,2,2,0.5))

plot_005a <- plot(weekdayAvg , type = "l", main = "Weekday Activty - a

plot_005b <- plot(weekendAvg, type = "l", main = "Weekend Activty - av
```




```
dev.copy(png,'plot005.png')
```

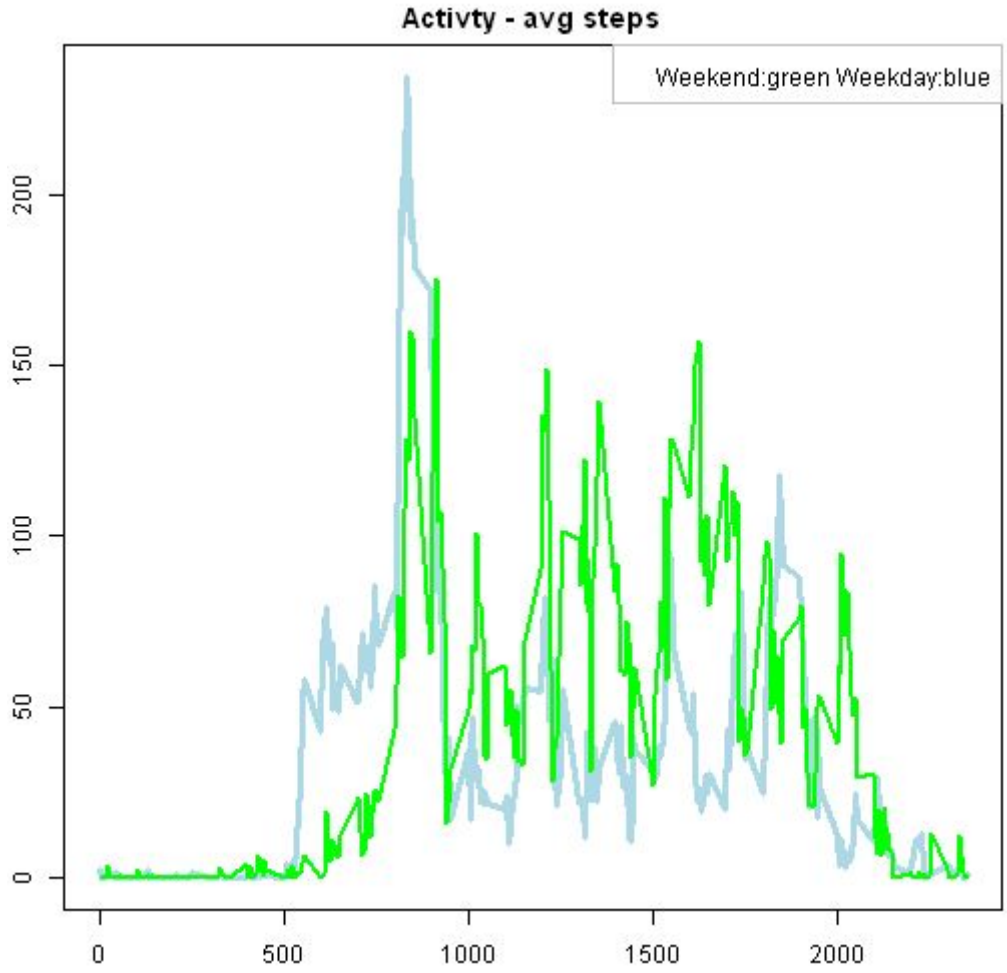
png 5

```
dev.off()
```

RStudioGD 2

```
par(mfrow = c(1, 1))
par(mar=c(3,2,2,0.5))

plot_006 <- plot(weekdayAvg , type = "l", col = "lightblue", lwd =3, x
lines(weekendAvg, type = "l", col = "green", lwd=2)
legend("topright", legend = "Weekend:green Weekday:blue", box.col=8)
```




```
dev.copy(png,'plot006.png')
```

png 5

```
dev.off()
```

RStudioGD 2

##Answer 4.0 Yes, there is a difference in patters of activity between weekdays and weekends?

Weekday activity sees sudden rise start at interval 500. There is lot more activity on weekends for interval 1000-2000

###reset the workign directory to original setwd(WDoriginal);getwd() ###clear the workspace rm(age, age.i, cleandata_mean, cleandata_median, dailysteps, dailysteps_mean, dailysteps_median, Data, dayAvg, daysteps, daytype, MainData, naughts, plot_001, Plot_002, plot_003, Plot_004a, Plot_004b, plot_005a, plot_005b, plot_006, StepsGroup, WDoriginal, weekdayAvg, weekendAvg) ls()

###end