Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

```
originData<-read.csv("activity.csv")
summary(originData)
```

```
##      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
```

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.

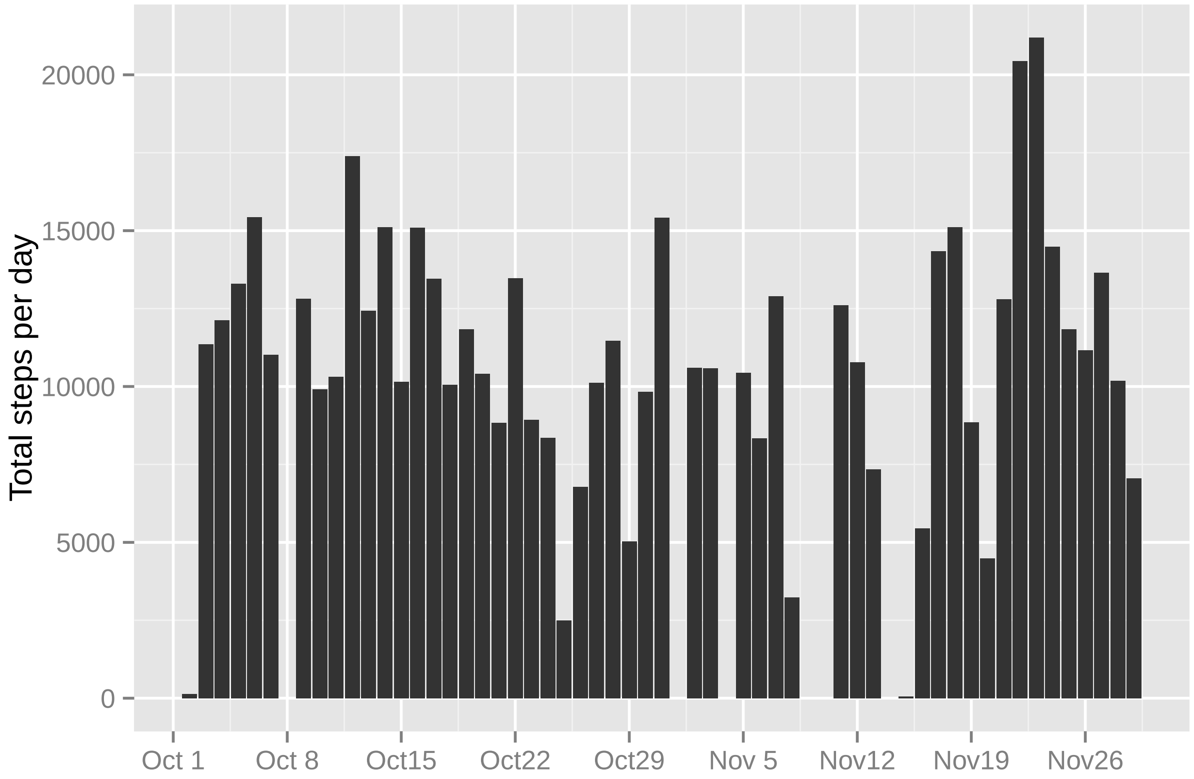
1. Make a histogram of the total number of steps taken per day

```
# Remove row with NA in the dataset
data<-originData[complete.cases(originData),]
summary(data)
```

```
##      steps                date          interval     
##  Min.   :  0.00   2012-10-02:  288   Min.   :   0.0  
##  1st Qu.:  0.00   2012-10-03:  288   1st Qu.: 588.8  
##  Median :  0.00   2012-10-04:  288   Median :1177.5  
##  Mean   : 37.38   2012-10-05:  288   Mean   :1177.5  
##  3rd Qu.: 12.00   2012-10-06:  288   3rd Qu.:1766.2  
##  Max.   :806.00   2012-10-07:  288   Max.   :2355.0  
##                   (Other)   :13536
```

```
#change date character to date format
data$date<-as.Date(data$date,"%Y-%m-%d")

#Make histogram
library(ggplot2)
library(scales)
g<-ggplot(data,aes(date,steps))
g+geom_histogram(stat="identity")+
        scale_x_date( labels=date_format("%b%e"),breaks = date_breaks("week"))+
        labs(x="",y="Total steps per day")
```



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

```
library(plyr)
StepPerDay<-ddply(data,c("date"),summarise,steps=sum(steps))
summary(StepPerDay)
```

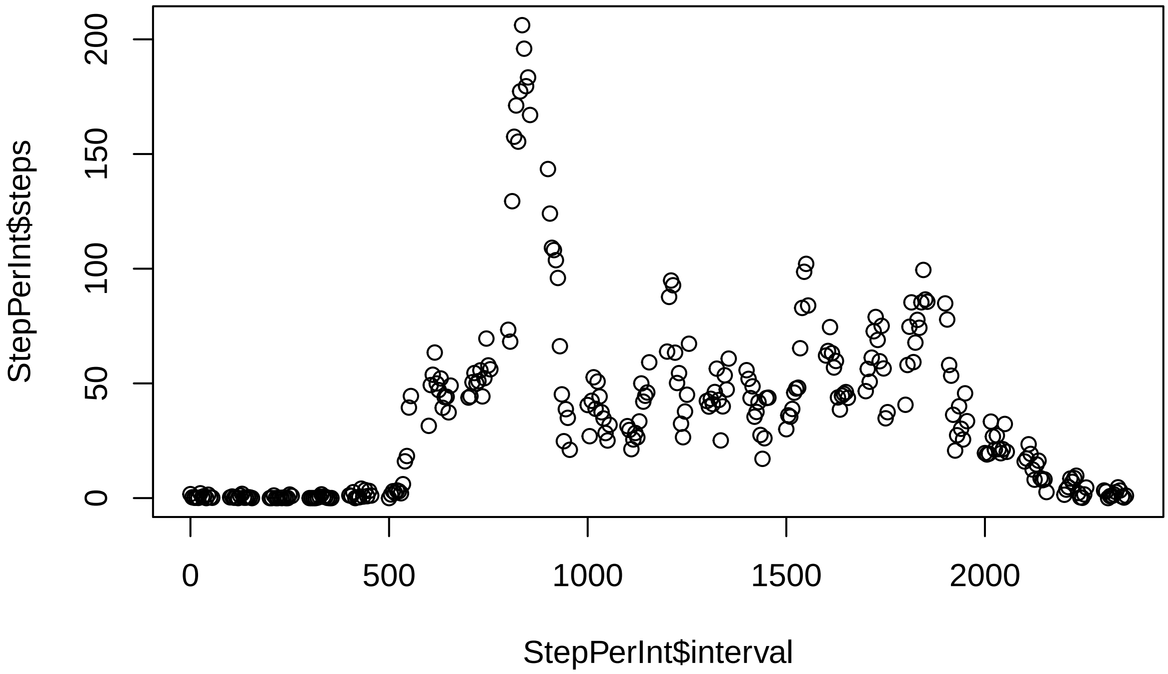
```
##       date                 steps
##  Min.   :2012-10-02   Min.   :   41
##  1st Qu.:2012-10-16   1st Qu.: 8841
##  Median :2012-10-29   Median :10765
##  Mean   :2012-10-30   Mean   :10766
##  3rd Qu.:2012-11-16   3rd Qu.:13294
##  Max.   :2012-11-29   Max.   :21194
```

The mean total number of step per day was 10766 (steps) the median was 10765 steps.

What is the average daily activity pattern?

```
StepPerInt<-ddply(data,c("interval"),summarise,steps=mean(steps))

plot(StepPerInt$interval,StepPerInt$steps)
```



The interval has maximum average of number of steps

```
StepPerInt[StepPerInt$steps==max(StepPerInt$steps),]
```

```
##     interval    steps
## 104      835 206.1698
```

835 is the interval time that the average of number of steps is maximum (~206 steps)

Imputing missing values

1. I will impute the NA by average value of that interval

```
ImpData<-originData # create a new imputed data frame
i<-1

while (i<=nrow(ImpData)){
        if (is.na(ImpData$steps[i])){
                j<-ImpData$interval[i]
                ImpData[i,1]<-StepPerInt[StepPerInt$interval==j,]$steps
        }
        i<-i+1
}
#View summary of imputed data
summary(ImpData)
```

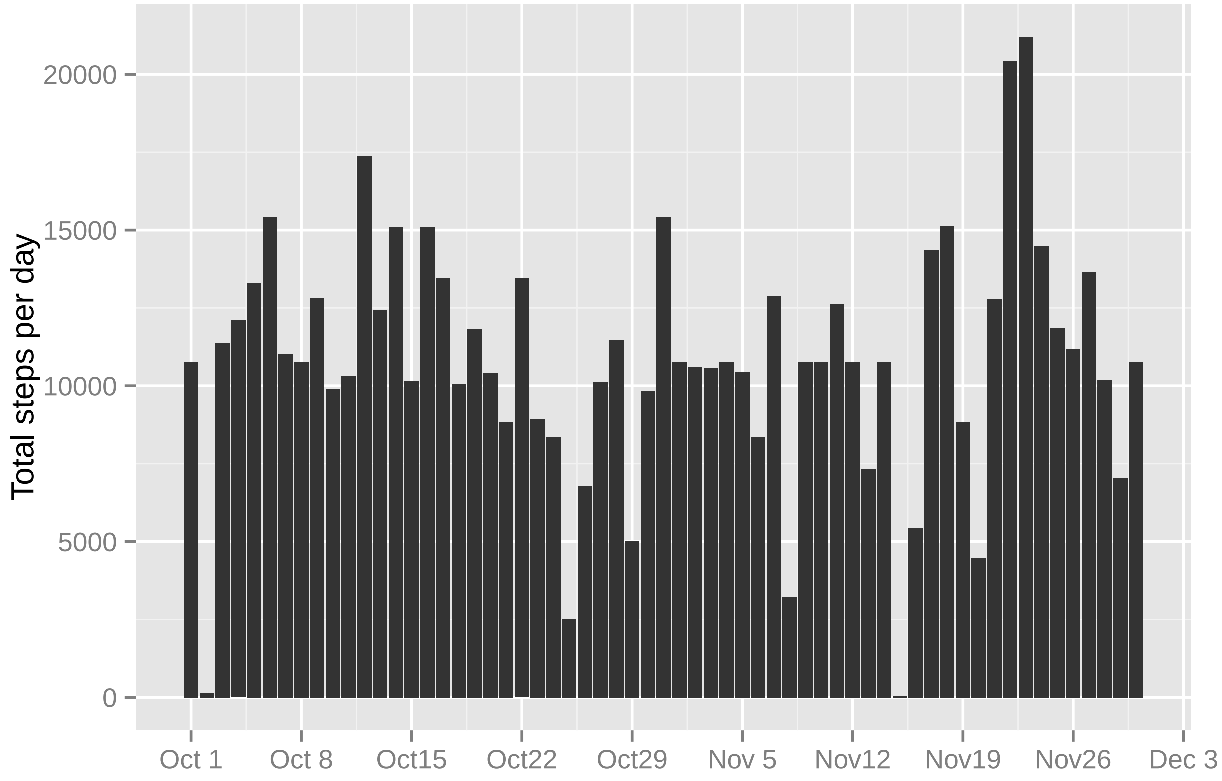
```
##      steps                date          interval
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8
##  Median :  0.00   2012-10-03:  288   Median :1177.5
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5
##  3rd Qu.: 27.00   2012-10-05:  288   3rd Qu.:1766.2
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0
##                   (Other)   :15840
```

There was no NA now.

2. Make histogram of mean daily steps

```
#change format of character into date
ImpData$date<-as.Date(ImpData$date,format="%Y-%m-%d")

#Make histogram by using ggplot2 and scales packages
library(ggplot2)
library(scales)
his<-ggplot(ImpData,aes(date,steps))
his+geom_histogram(stat="identity")+
        scale_x_date( labels=date_format("%b%e"),breaks = date_breaks("week"))+
        labs(x="",y="Total steps per day")
```



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

```
# Make table data
StepPerDay1<-ddply(ImpData,c("date"),summarise,steps=sum(steps))

#Caculate mean and median
summary(StepPerDay1)
```

```
##        date                 steps
##  Min.   :2012-10-01   Min.   :   41
##  1st Qu.:2012-10-16   1st Qu.: 9819
##  Median :2012-10-31   Median :10766
##  Mean   :2012-10-31   Mean   :10766
##  3rd Qu.:2012-11-15   3rd Qu.:12811
##  Max.   :2012-11-30   Max.   :21194
```

We can now see that mean of total steps per day was not changed. However, median of total steps per day in imputed data has changed toward the mean value.

Are there differences in activity patterns between weekdays and weekends?

1.Format the day to become weedays and weekends

```
#Change date format into day
ImpData$date<-format(ImpData$date,"%a",trim=T)
table(ImpData$date) #View number of date now
```

```
##
##  Fri  Mon  Sat  Sun  Thu  Tue  Wed
## 2592 2592 2304 2304 2592 2592 2592
```

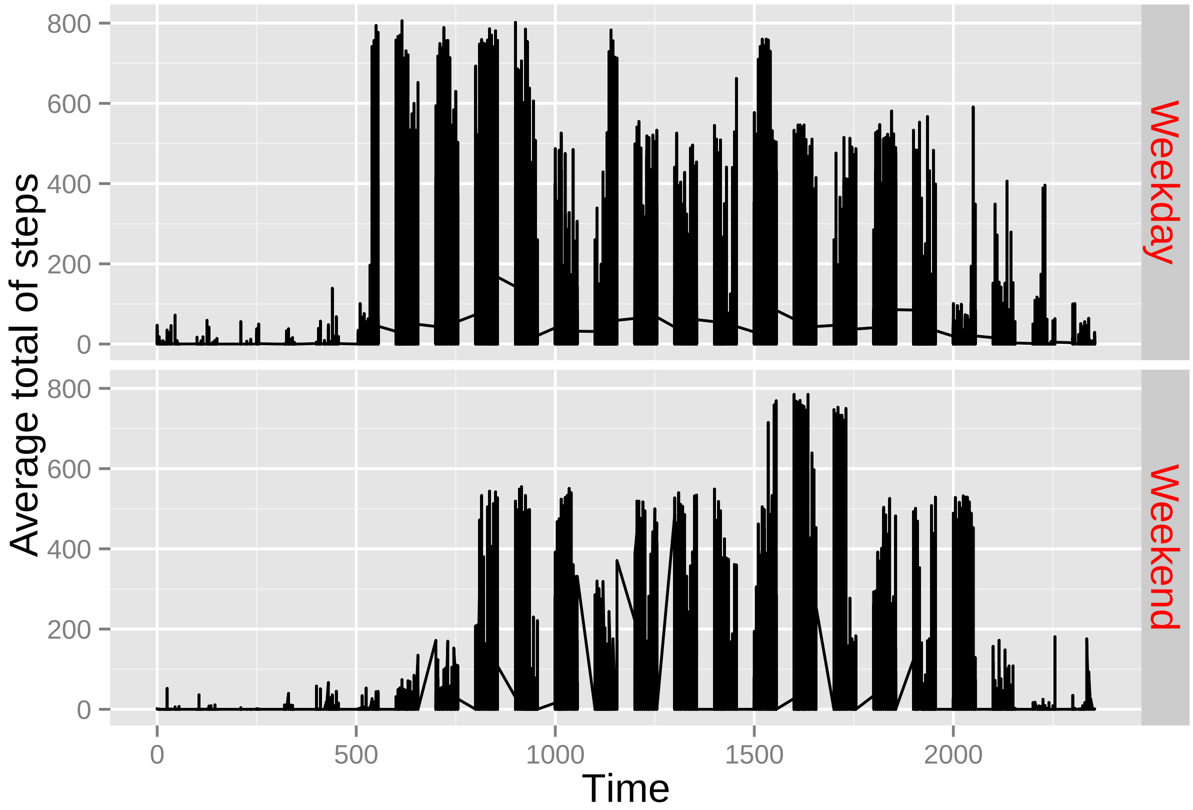
```
#Change date to weekday and weekend
for (i in 1:nrow(ImpData)){
        ifelse(ImpData$date[i]=="Sat" | ImpData$date[i]=="Sun",
               ImpData[i,2]<-"Weekend",
              ImpData[i,2]<-"Weekday" )}

str(ImpData)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : num  1.717 0.3396 0.1321 0.1509 0.0755 ...
##  $ date    : chr  "Weekday" "Weekday" "Weekday" "Weekday" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

2. Make graph to see steps by interval in weekday and weekend

```
library(ggplot2)
we<-ggplot(ImpData,aes(interval,steps))
we+geom_line()+facet_grid(date~.)+
        labs(x="Time",y="Average total of steps")+
        theme(strip.text.y=element_text(size=15,color="red"))+
        theme(axis.title.x=element_text(size=15))+
        theme(axis.title.y=element_text(size=15))
```



We can see that in the weekday, the high activity start earlier from 5 AM, while for the weekend, the acitivity start late at around 8 AM. after 8 AM.