Title My Reproducible Research Project1

```
library(ggplot2)
library(data.table)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

What is mean total number of steps taken per day?

Preprocessing Data

1. load the data (ie. read.csv())

```
Data <- read.csv("activity.csv")
str(Data)
```

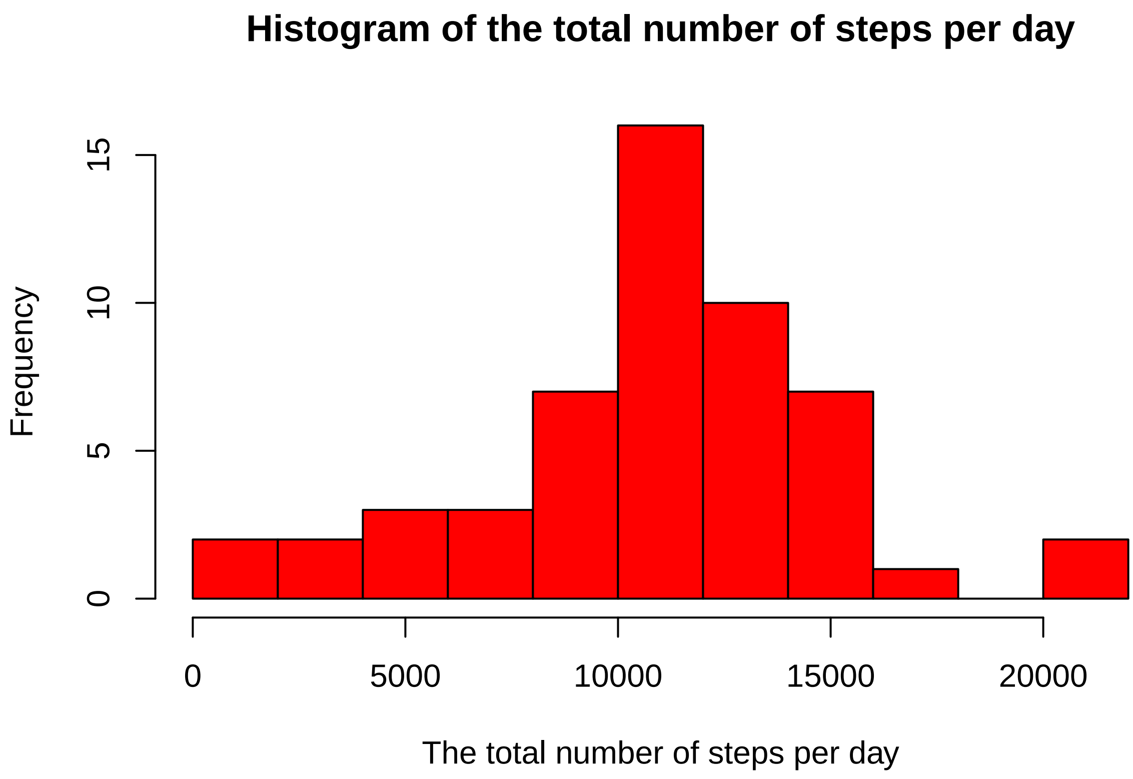
```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

2. Process and transform data if needed

```
Data1 <- Data[complete.cases(Data),]
Tsteps <- aggregate(steps ~ date, Data1, sum)
names(Tsteps)[1] <- "Date"
names(Tsteps)[2] <- "Totalsteps"
```

3. Make a histogram of the total of steps taken per day

```
hist(Tsteps$Totalsteps,
     col="red",
     main="Histogram of the total number of steps per day",
     xlab="The total number of steps per day",
     breaks =10)
```



4. Calculate and report the mean and median number of steps per day

```
mean(Tsteps$Totalsteps)
```

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

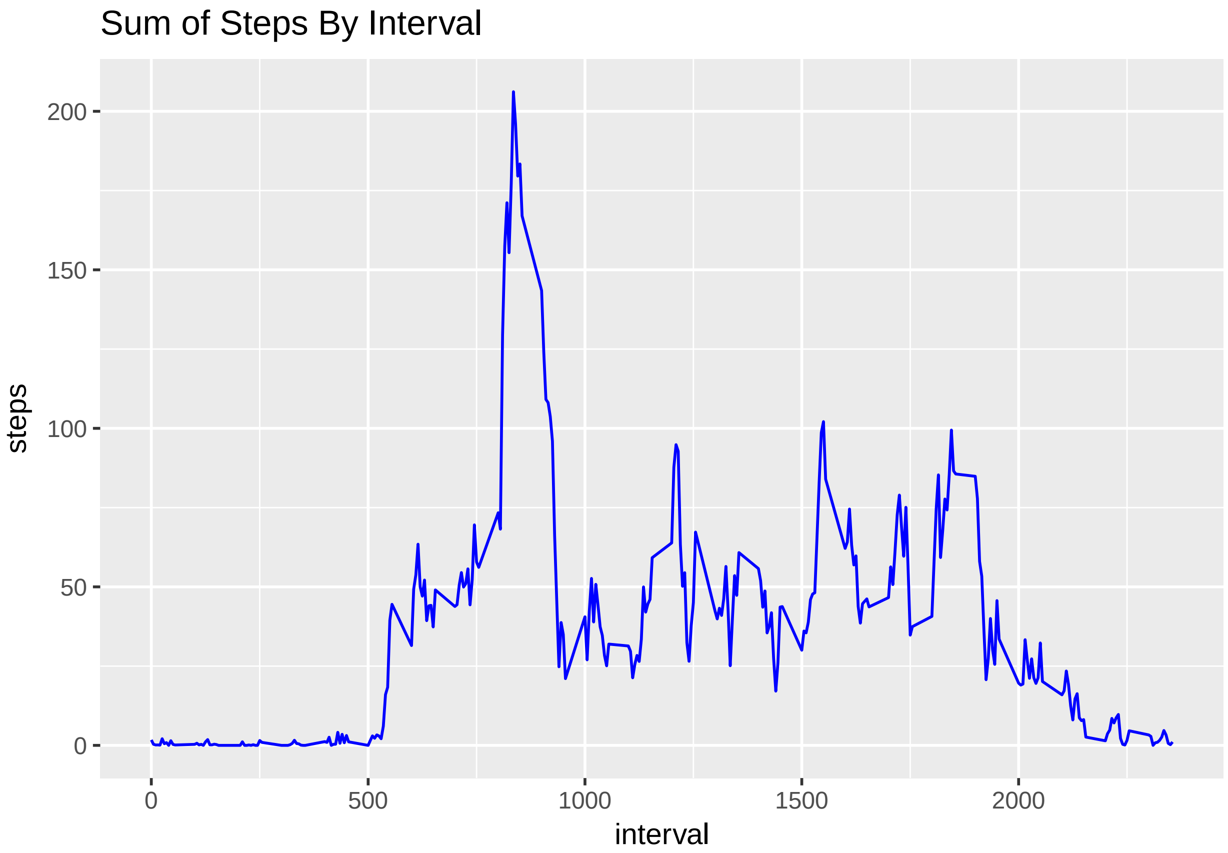
```
median(Tsteps$Totalsteps)
```

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

What is the average daily activity pattern

1. Make a time series plot of 5-min interval and the average number of steps taken, averaged across all days

```
Data2 <- Data[complete.cases(Data),]
MData1 <- aggregate(Data2$steps, by=list(Data2$interval), mean)
names(MData1)[1] <- "interval"
names(MData1)[2] <- "steps"
ggplot(MData1, aes(x=interval, y=steps)) +
        labs(title="Sum of Steps By Interval", x="interval", y="steps")+
        geom_line(color="blue")
```



2.

Which 5-min interval contains max numbers of steps?

```
Interval <- MData1[which.max(MData1$steps),]
Interval
```

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

Imputing Missing Values

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

```
Datamissing <- sum(is.na(Data))
Datamissing
```

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

2. Devise a strategy for filling in all of the missing values in the dataset. My strategy: will use the mean interval steps for the 5-minute interval at a given interval.

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

```
NewData <- Data
NewData <- NewData[complete.cases(NewData$steps),]
MeanByInterval <- aggregate(NewData$steps, by=list(NewData$interval), sum)
names(MeanByInterval)[1]="interval"
names(MeanByInterval)[2]="steps"
```

Impute Method- Attempt 2

```
NewData1 <- Data
Datamissing <- is.na(NewData1$steps)
CleanDatamissing <- NewData1[!is.na(NewData1$steps),]
MeanVals <- tapply(CleanDatamissing$steps, CleanDatamissing$interval, mean, na.rm=TRUE, simplify=TRUE)
NewData1$steps[Datamissing] <- MeanVals[as.character(NewData1$interval[Datamissing])]
sum(Datamissing)
```

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

```
sum(is.na(NewData1$steps))
```

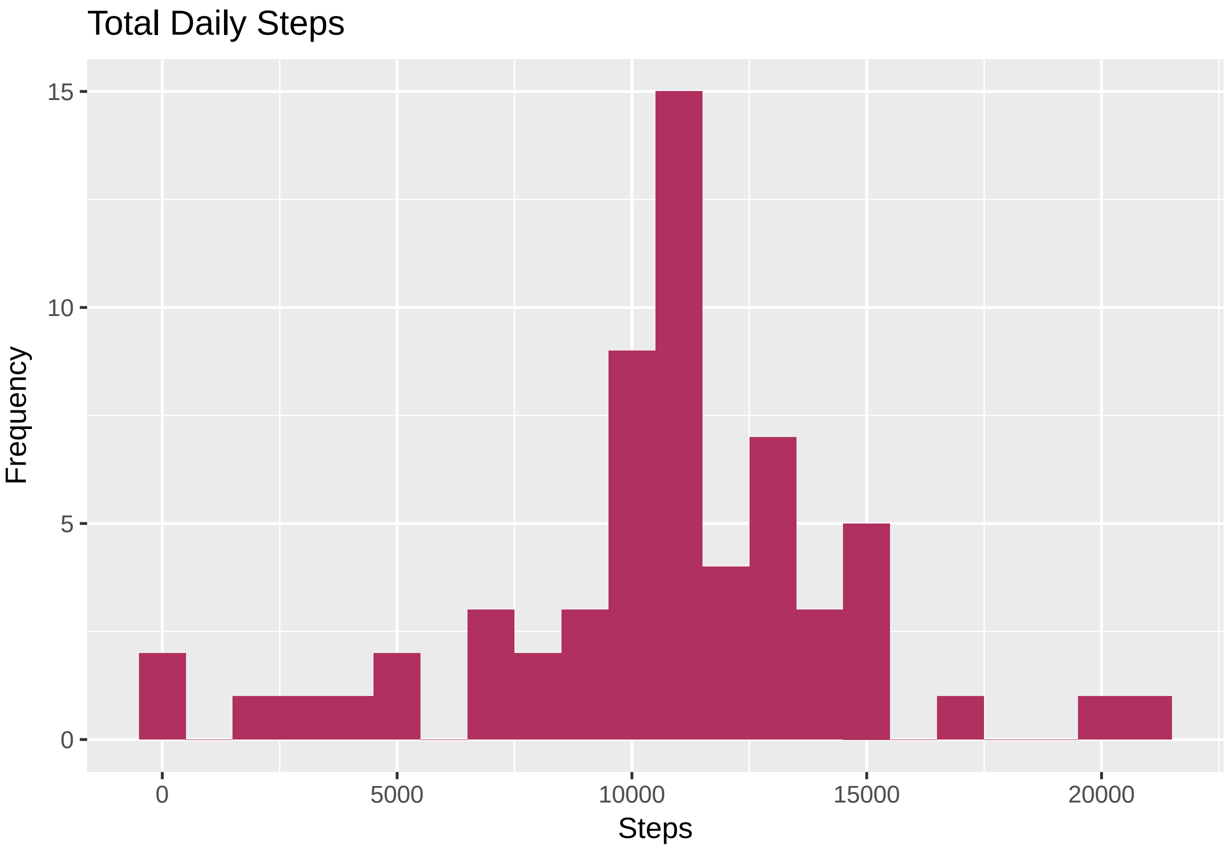
```
## [1] 0
```

4. Make a histogram of the total number of steps taken per day and calculate and report the mean and median total number of steps taken per day.

```
SumDataByDay <- aggregate(NewData1$steps, by=list(NewData1$date), sum)
names(SumDataByDay)[1]="date"
names(SumDataByDay)[2]="totalsteps"
head(SumDataByDay,20)
```

```
##          date totalsteps
## 1  2012-10-01   10766.19
## 2  2012-10-02     126.00
## 3  2012-10-03   11352.00
## 4  2012-10-04   12116.00
## 5  2012-10-05   13294.00
## 6  2012-10-06   15420.00
## 7  2012-10-07   11015.00
## 8  2012-10-08   10766.19
## 9  2012-10-09   12811.00
## 10 2012-10-10    9900.00
## 11 2012-10-11   10304.00
## 12 2012-10-12   17382.00
## 13 2012-10-13   12426.00
## 14 2012-10-14   15098.00
## 15 2012-10-15   10139.00
## 16 2012-10-16   15084.00
## 17 2012-10-17   13452.00
## 18 2012-10-18   10056.00
## 19 2012-10-19   11829.00
## 20 2012-10-20   10395.00
```

```
# Plot using ggplot
ggplot(SumDataByDay, aes(x=totalsteps))+
        geom_histogram(fill="maroon",binwidth=1000)+
        labs(title="Total Daily Steps", x="Steps", y="Frequency")
```



```
# Mean on New Data
mean(SumDataByDay$totalsteps)
```

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

```
#Median on New Data
median(SumDataByDay$totalsteps)
```

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

Yes, they are same with original.

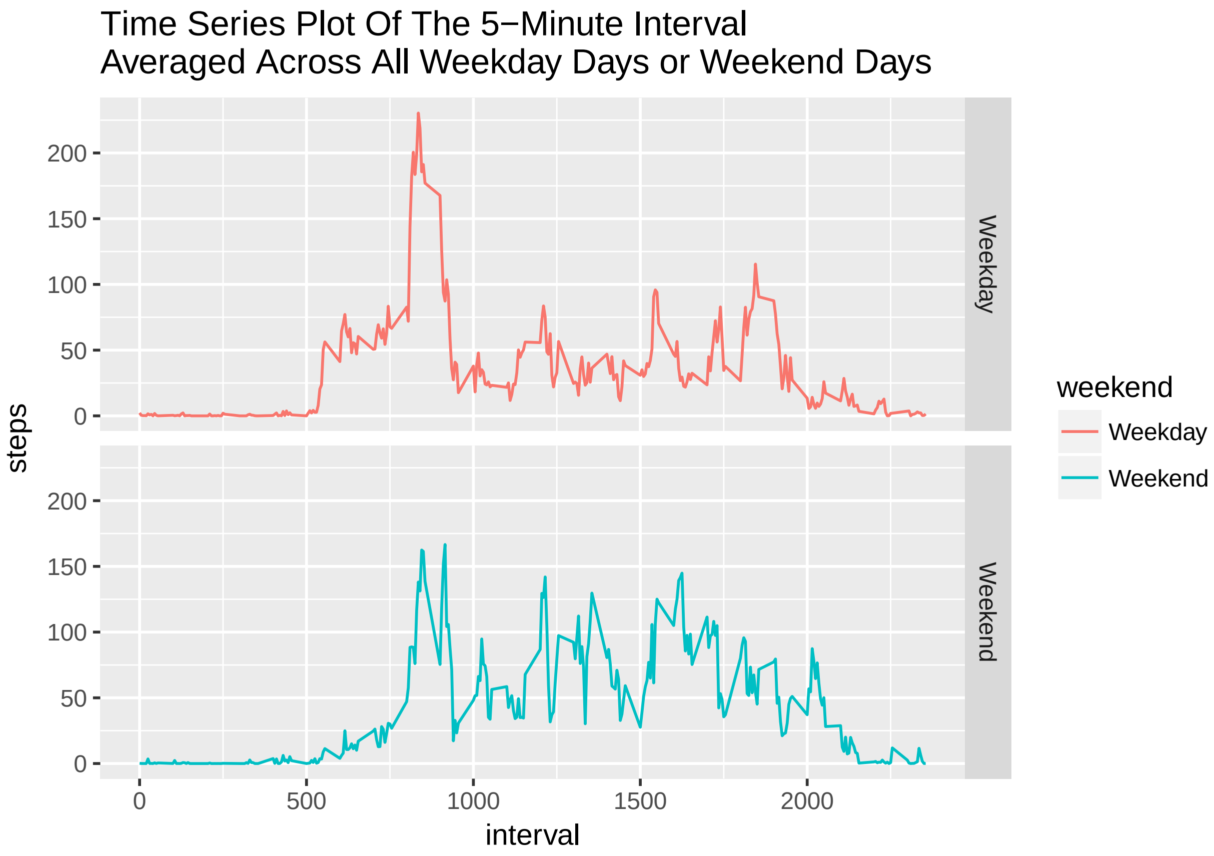
Are there differences in activity patterns between weekdays and weekends?

```
NewData1$weekday <- weekdays(as.Date.character(NewData1$date))
NewData1$weekend <- ifelse(NewData1$weekday=="Saturday"| NewData1$weekday=="Sunday","Weekend","Weekday"
head(NewData1,5)
```

```
##       steps       date interval weekday weekend
## 1 1.7169811 2012-10-01        0  Monday Weekday
## 2 0.3396226 2012-10-01        5  Monday Weekday
## 3 0.1320755 2012-10-01       10  Monday Weekday
## 4 0.1509434 2012-10-01       15  Monday Weekday
## 5 0.0754717 2012-10-01       20  Monday Weekday
```

```
MeanWEWD <- aggregate(NewData1$steps, by=list(NewData1$weekend, NewData1$interval), mean)
names(MeanWEWD)[1]="weekend"
names(MeanWEWD)[2]="interval"
names(MeanWEWD)[3]="steps"
```

```
ggplot(MeanWEWD, aes(x=interval, y=steps, color=weekend))+
        geom_line()+
        facet_grid(weekend ~.)+
        labs(title="Time Series Plot Of The 5-Minute Interval\nAveraged Across All Weekday Days or Weeke
```


Time Series Plot Of The 5−Minute Interval
Averaged Across All Weekday Days or Weekend Days
steps
Weekday
Weekend
interval
weekend
Weekday
Weekend