```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.4.3
```

Download and save the file acitivity.csv setwd

```
setwd("C:/SM_Projects/Personal/Training/Data Science Orientation/Data_Science_Course_5")
activity <- read.csv("activity.csv", sep = ",")
```

With the str function we can check if the date is in the correct format

```
str(activity)
```

```
## '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 ...
```

Since date shows as a factor, we will need to convert into a Date format

```
activity$date <- as.Date(activity$date)
```

We can use the Str function to check now if the date is converted

```
str(activity)
```

```
## '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 ...
```

Question 1: What is mean total number of steps taken per day?

First we need to remove the NA values from the steps

```
activity<-activity %>% filter(complete.cases(activity))
stepsummary <- tapply(activity$steps, activity$date, FUN = sum, na.rm = TRUE)
```

check stepsummary

```
stepsummary
```

```
## 2012-10-02 2012-10-03 2012-10-04 2012-10-05 2012-10-06 2012-10-07
##        126      11352      12116      13294      15420      11015
## 2012-10-09 2012-10-10 2012-10-11 2012-10-12 2012-10-13 2012-10-14
##      12811       9900      10304      17382      12426      15098
## 2012-10-15 2012-10-16 2012-10-17 2012-10-18 2012-10-19 2012-10-20
##      10139      15084      13452      10056      11829      10395
## 2012-10-21 2012-10-22 2012-10-23 2012-10-24 2012-10-25 2012-10-26
##       8821      13460       8918       8355       2492       6778
## 2012-10-27 2012-10-28 2012-10-29 2012-10-30 2012-10-31 2012-11-02
##      10119      11458       5018       9819      15414      10600
## 2012-11-03 2012-11-05 2012-11-06 2012-11-07 2012-11-08 2012-11-11
##      10571      10439       8334      12883       3219      12608
## 2012-11-12 2012-11-13 2012-11-15 2012-11-16 2012-11-17 2012-11-18
##      10765       7336         41       5441      14339      15110
## 2012-11-19 2012-11-20 2012-11-21 2012-11-22 2012-11-23 2012-11-24
##       8841       4472      12787      20427      21194      14478
## 2012-11-25 2012-11-26 2012-11-27 2012-11-28 2012-11-29
##      11834      11162      13646      10183       7047
```

calculate mean and median

```
mean(stepsummary)
```

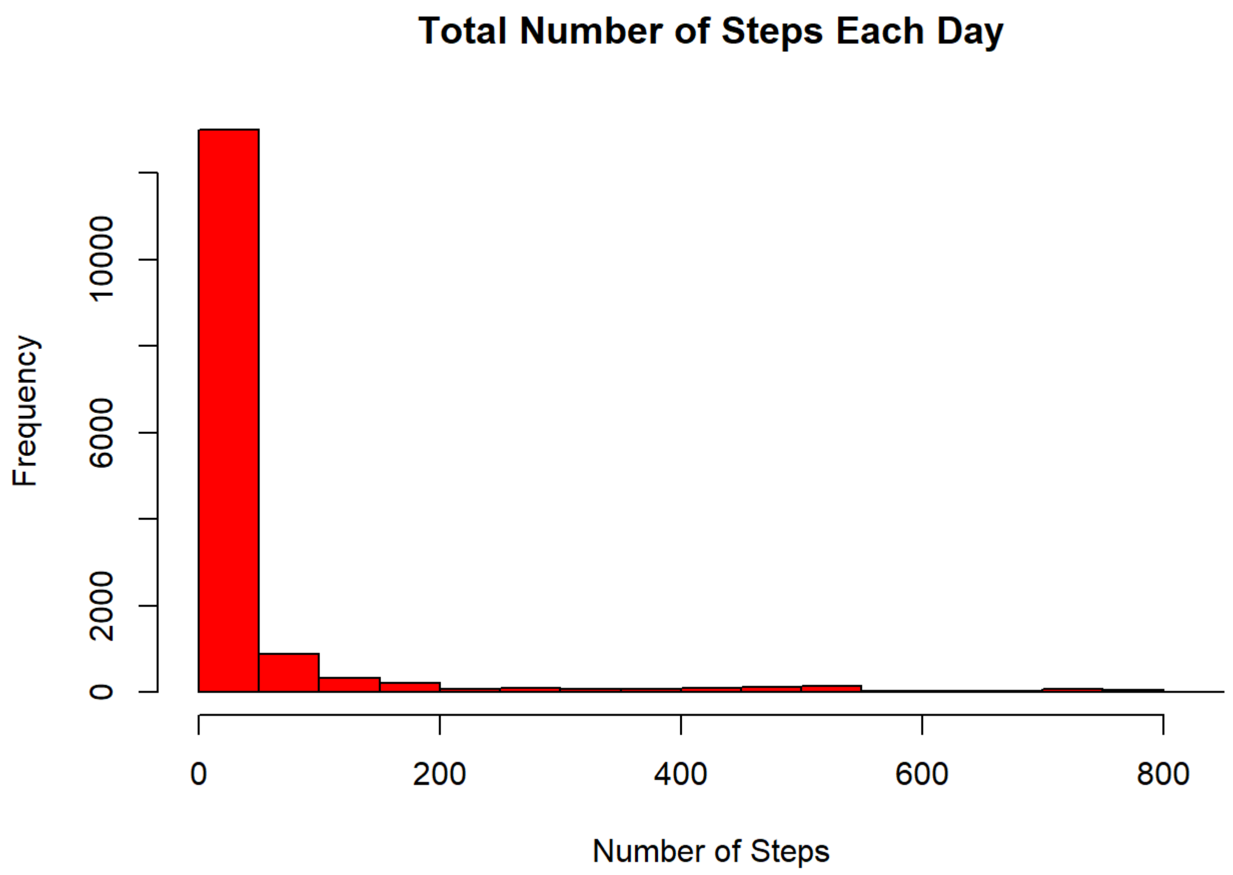
```
## [1] 10766.19
```

```
median(stepsummary)
```

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

plot histogram for total number of steps

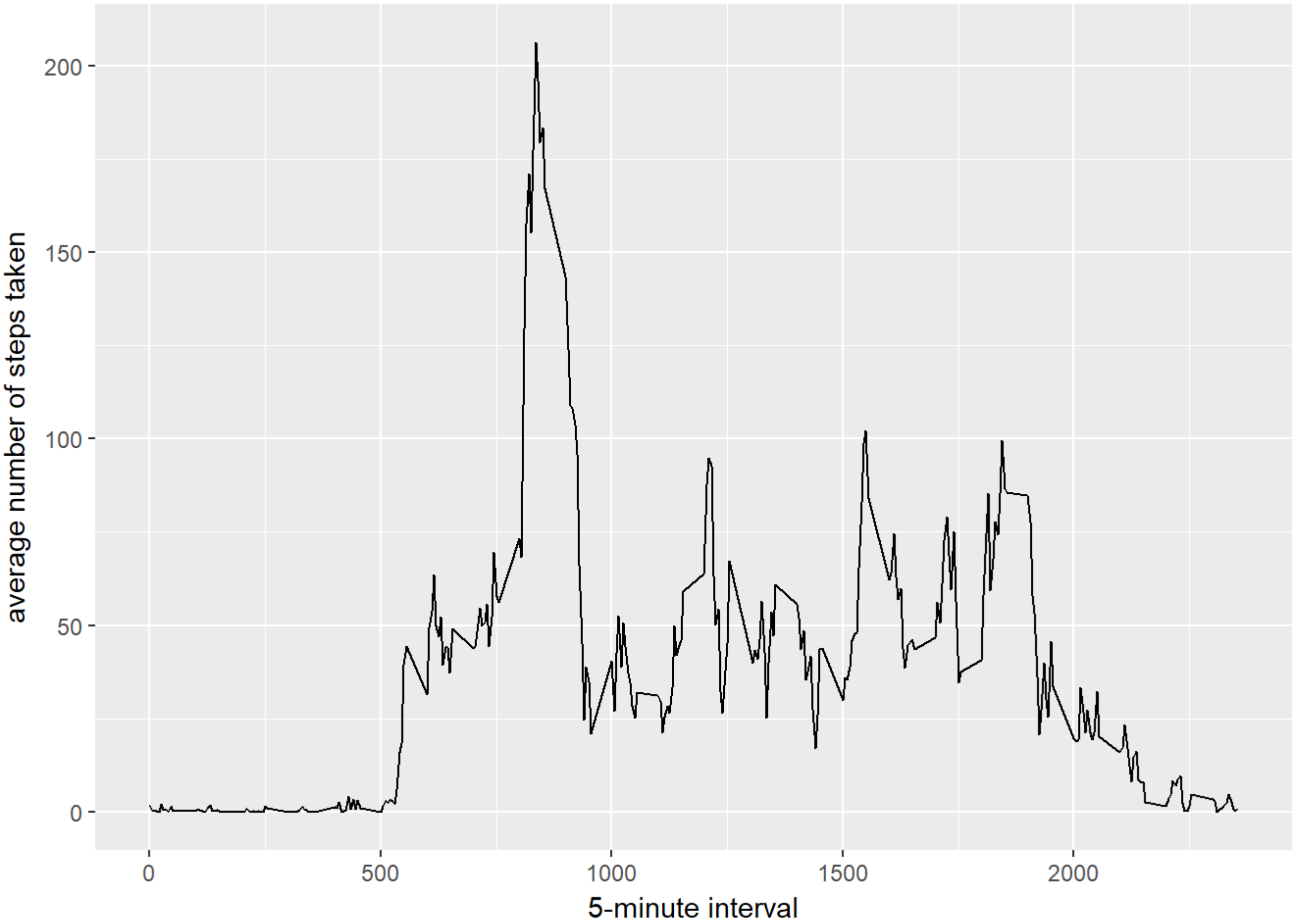
```
hist(activity$steps, xlab = "Number of Steps", main = "Total Number of Steps Each Day", col = 'red')
```



#

Question 1: What is the average daily activity pattern?

```
averagepattern <- aggregate(x = list(steps = activity$steps), by = list(interval = activity$interval), FUN = mean, na.rm = TRUE)
ggplot(data = averagepattern, aes(x = interval, y = steps)) + geom_line() + xlab("5-minute interval") + ylab("average number of steps taken")
```



#Calculating maximum number of steps

```
averagepattern[which.max(averagepattern$steps), ]
```

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

Question 4: Imputing missing values

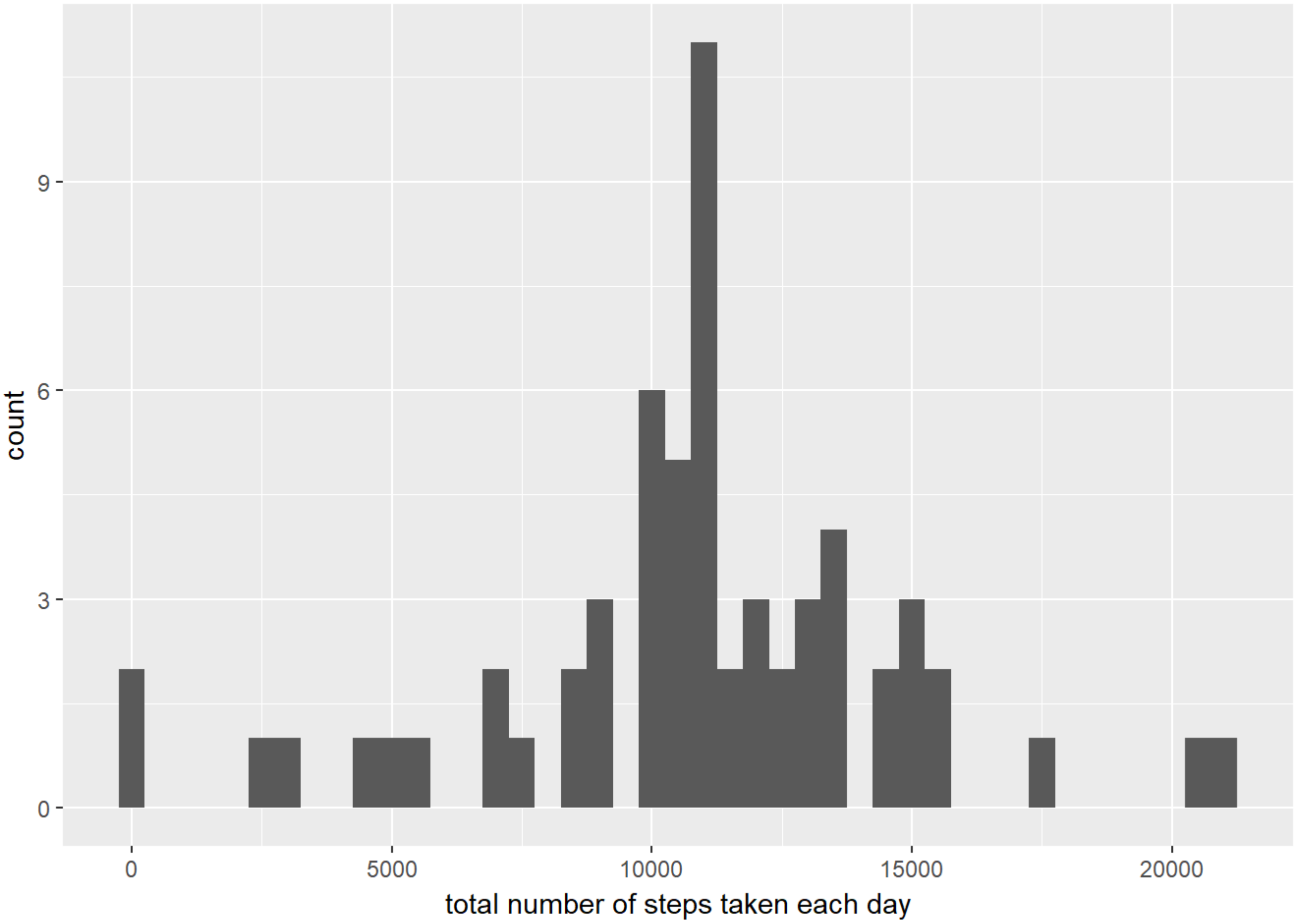
Since I already filtered NA previously. I will create another dataset with original na values

```
activitywithna <- read.csv("activity.csv", sep = ",")
sum(is.na(activitywithna$steps))
```

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

Replacing missing values

```
imputedactivity <- activitywithna %>%
   group_by(interval) %>%
  mutate(steps = replace(steps, is.na(steps), mean(steps, na.rm = TRUE)))
imputedsteps <- tapply(imputedactivity$steps, imputedactivity$date, FUN = sum)
qplot(imputedsteps, binwidth = 500, xlab = "total number of steps taken each day")
```



```
mean(imputedsteps)
```

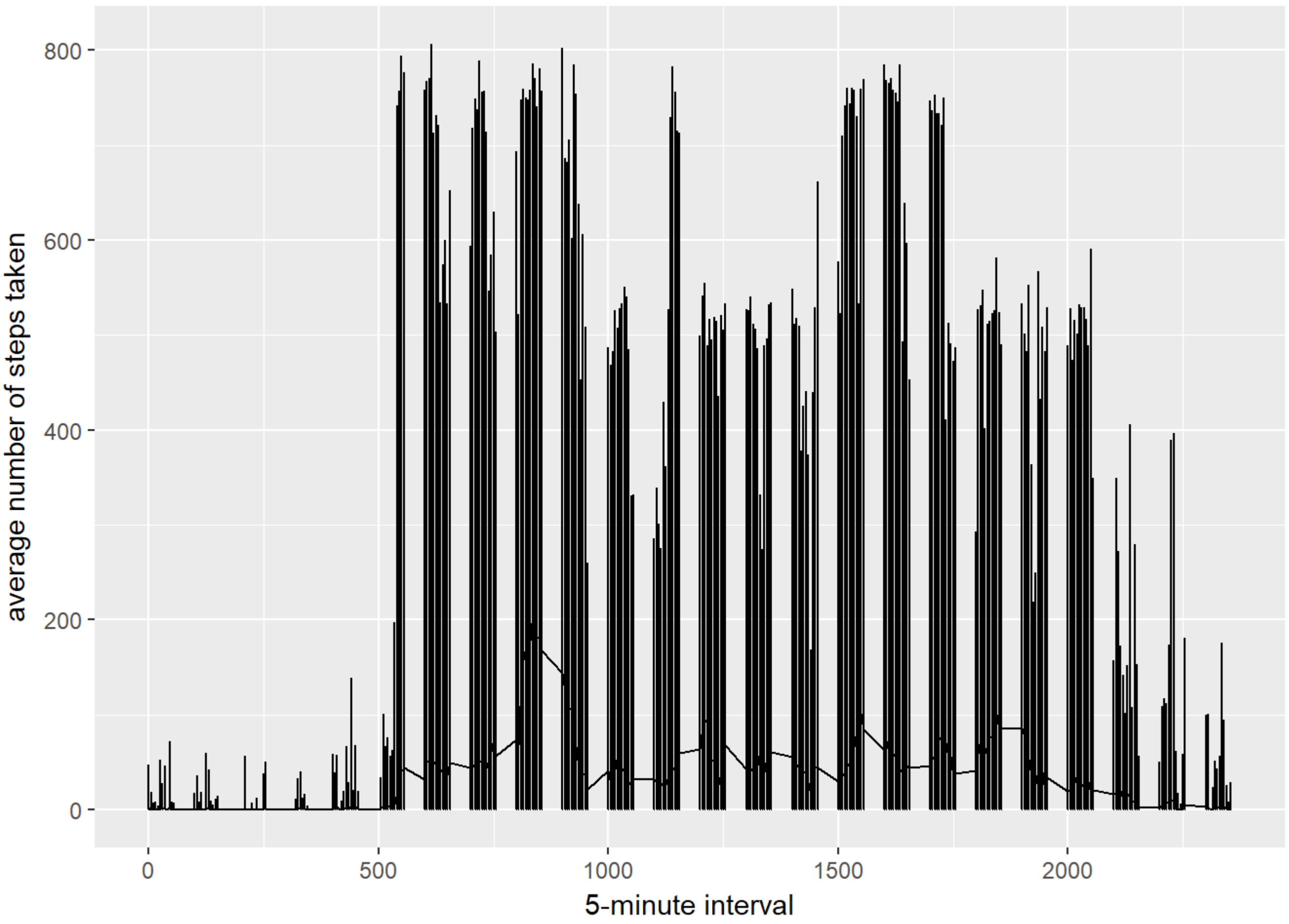
```
## [1] 10766.19
```

```
median(imputedsteps)
```

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

Question 5: Are there differences in activity patterns between weekdays and weekends?

```
 weekday.or.weekend <- function(date) {
    day <- weekdays(date)
  if (day %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
        return("weekday") else if (day %in% c("Saturday", "Sunday"))
                  return("weekend") else stop("invalid date")
  }
ggplot(data = imputedactivity, aes(x = interval, y = steps)) + geom_line() + xlab("5-minute interval") + ylab("average number of steps taken")
```


average number of steps taken
5-minute interval