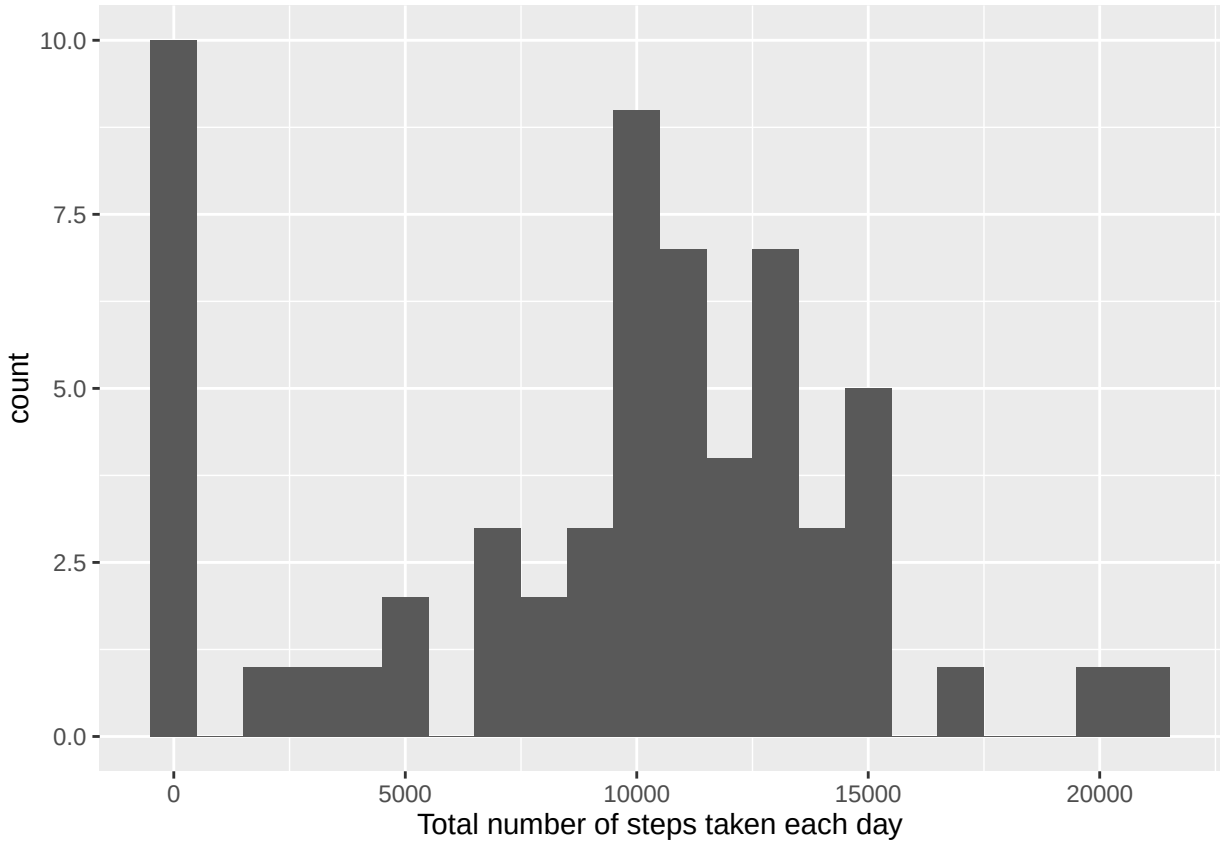
Reproducible Research: Peer Assessment 1

Loading and preprocessing/transforming of the data

```
unzip(zipfile="activity.zip")
data <- read.csv("activity.csv")
```

What is mean total number of steps taken per day?

```
library(ggplot2)
total.steps <- tapply(data$steps, data$date, FUN=sum, na.rm=TRUE)
qplot(total.steps, binwidth=1000, xlab="Total number of steps taken each day")
```




```
mean(total.steps, na.rm=TRUE)
```

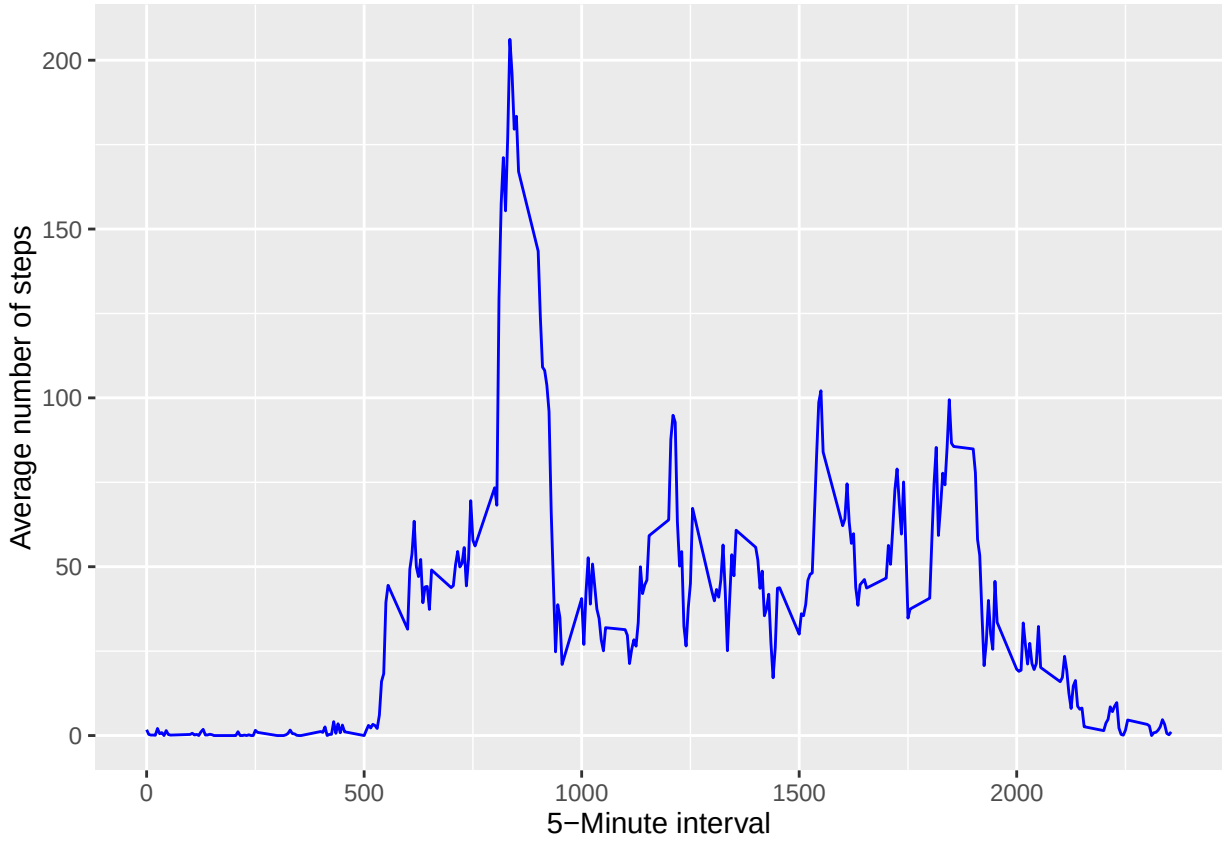
```
## [1] 9354.23
```

```
median(total.steps, na.rm=TRUE)
```

```
## [1] 10395
```

What is the average daily activity pattern?

```
library(ggplot2)
averages <- aggregate(x=list(steps=data$steps), by=list(interval=data$interval),
                      FUN=mean, na.rm=TRUE)
ggplot(data=averages, aes(x=interval, y=steps)) +
    geom_line(colour='blue') +
    xlab("5-Minute interval") +
    ylab("Average number of steps")
```



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

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

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

Imputing missing values

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.

```r
missing <- is.na(data$steps)
# Calculate and report the total number of missing values in the dataset (i.e. the total number of rows
table(missing)
```

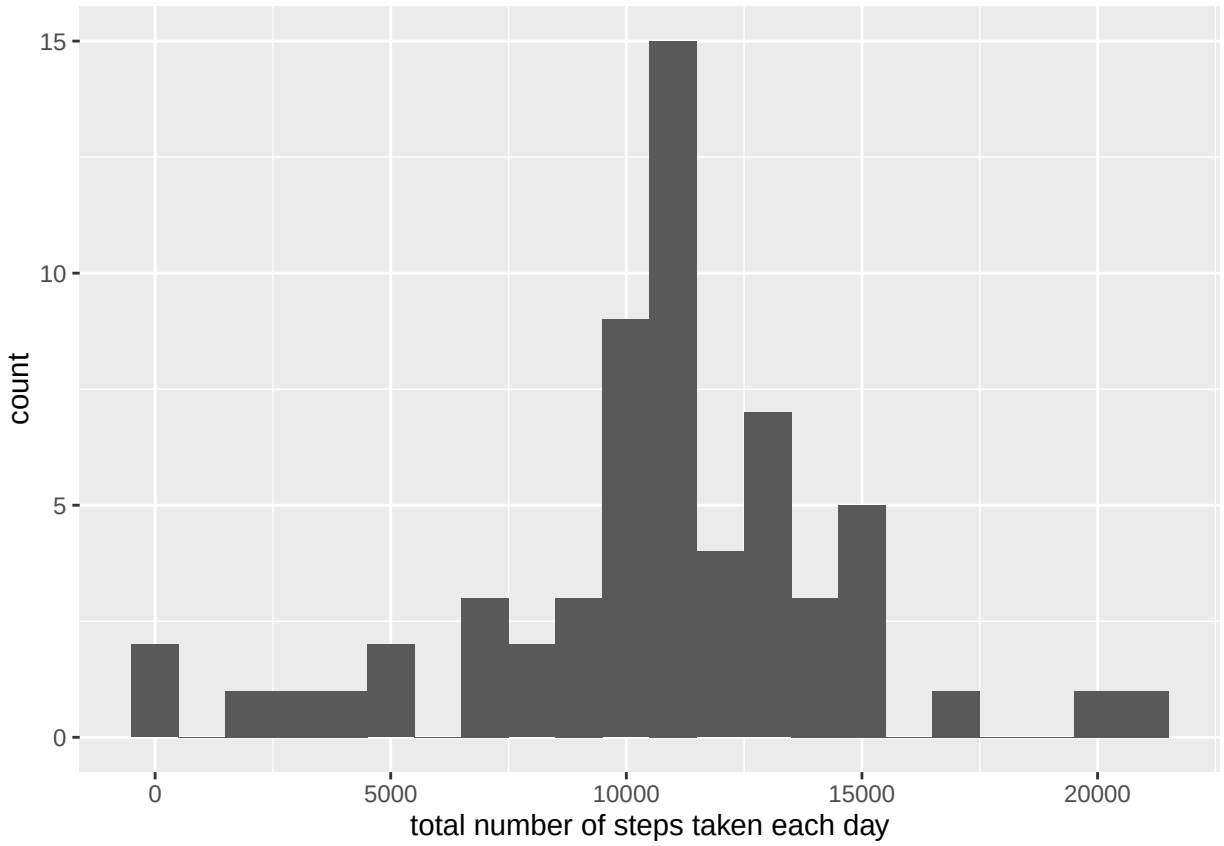
```
## missing
## FALSE  TRUE
## 15264  2304
```

All of the missing values are filled in with mean value for that 5-minute interval.

```r
# Replacing each missing value with the mean value of its 5-minute interval of that day
fill.value <- function(steps, interval) {
    filled <- NA
    if (!is.na(steps))
        filled <- c(steps)
    else
        filled <- (averages[averages$interval==interval, "steps"])
    return(filled)
}
filled.data <- data
filled.data$steps <- mapply(fill.value, filled.data$steps, filled.data$interval)
```

Histogram of the total number of steps taken each day and calculation of the mean and median total number of steps taken per day.

```r
total.steps <- tapply(filled.data$steps, filled.data$date, FUN=sum)
qplot(total.steps, binwidth=1000, xlab="total number of steps taken each day")
```




```r
mean(total.steps)
```

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

```r
median(total.steps)
```

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

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?

Mean and median becomes higher after imputing missing data. This is because in the original data, there are some days with `steps` values `NA` for any `interval`. The impact of imputing the missing values is to have more data, hence to obtain a bigger mean and median value.

Are there differences in activity patterns between weekdays and weekends?

Step - I: Findinf the day of the week for each measurement in the dataset.

```r
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")
}
filled.data$date <- as.Date(filled.data$date)
filled.data$day <- sapply(filled.data$date, FUN=weekday.or.weekend)
```

Step - II: Panel plot containing plots of average number of steps taken on weekdays and weekends.

```
averages <- aggregate(steps ~ interval + day, data=filled.data, mean)
ggplot(averages, aes(interval, steps)) + geom_line(color="red") + facet_grid(day ~ .) +
    xlab("5-minute interval") + ylab("Number of steps")
```