Reproducible Research: Peer Assessment 1

1. Loading and preprocessing the data

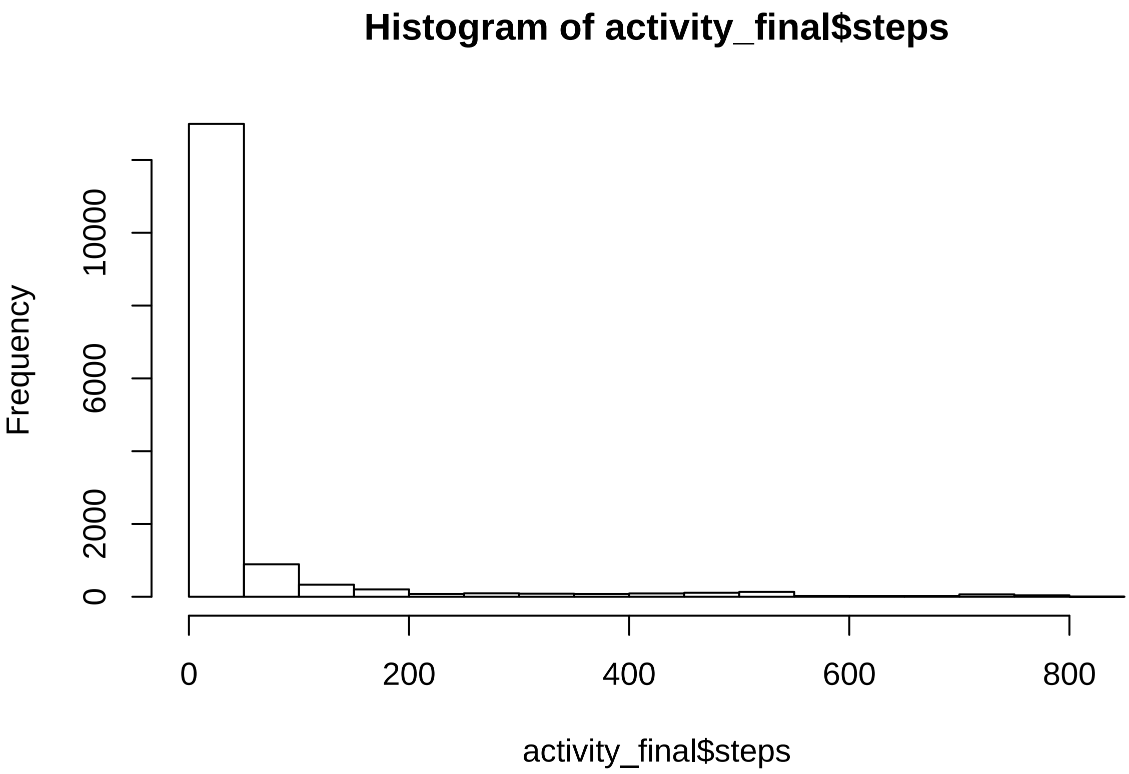
Load the activity.csv dataset from my local path.

```
activity <- read.csv("D:\\Coursera\\Material\\05. Reproducible Research\\CourseProject\\activity.csv")
```

2. What is mean total number of steps taken per day?

First, Remove NA data. And draw the histogram to see the distribution of steps taken per day. Then, claculate the mean and median. We got mean = 10766.19 and median = 10766.

```
activity_final <- activity[!is.na(activity$steps),]
steps_by_date <- aggregate(steps ~ date, data=activity_final, sum)
hist(activity_final$steps)
```



```
mean(steps_by_date$steps)
```

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

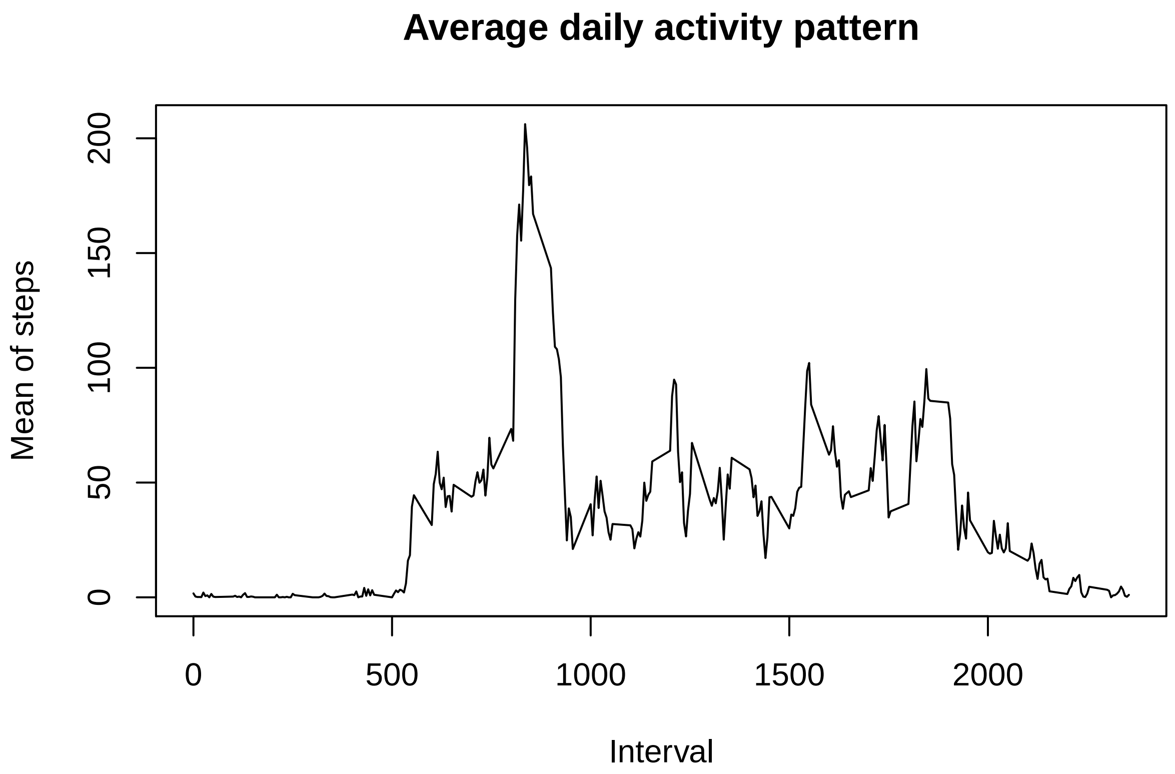
```
median(steps_by_date$steps)
```

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

3. What is the average daily activity pattern?

The average daily activiy pattern is as below plot. You can see the peak at interval 835.

```
steps_by_interval <- tapply(activity_final$steps,activity_final$interval,mean)
plot(row.names(steps_by_interval), steps_by_interval, type="l",
     main="Average daily activity pattern", xlab="Interval", ylab="Mean of steps")
```



```
names(which.max(steps_by_interval))
```

```
## [1] "835"
```

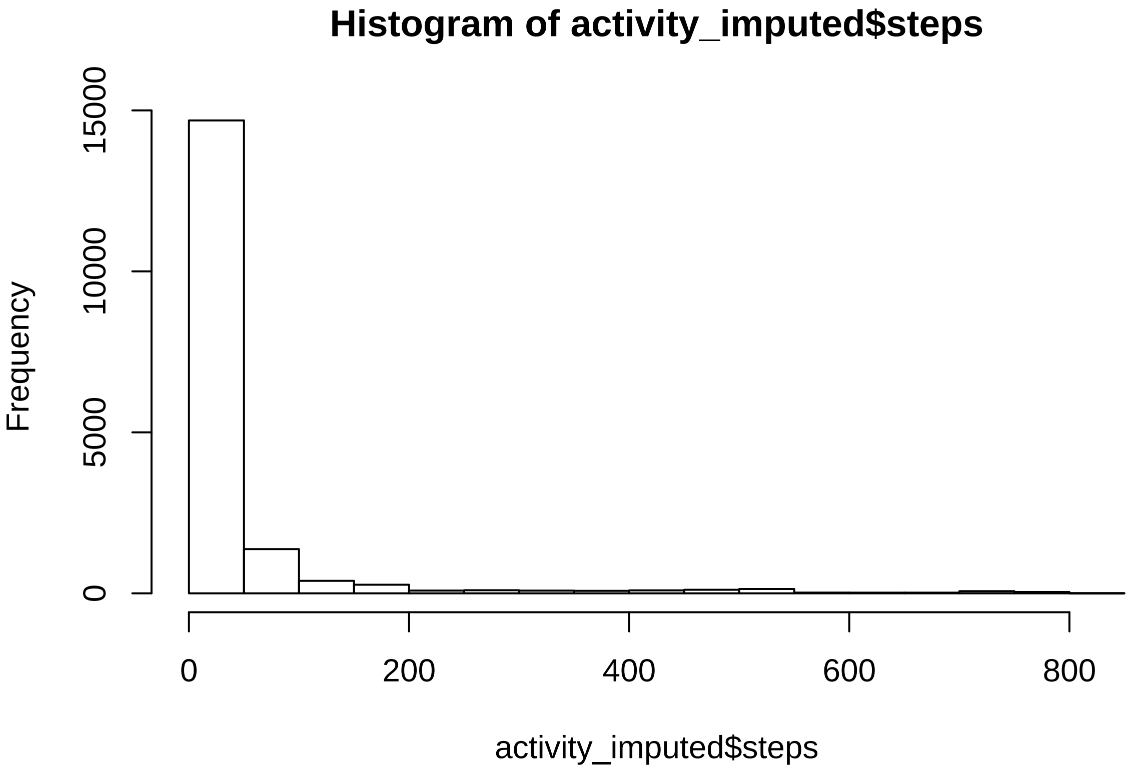
4. Imputing missing values

Try to replace the missing values by the mean of interval.

```
activity_imputed <- activity
for (i in 1:nrow(activity)) {
  if( is.na(activity[i,]$steps) )
  {
     j <- which(rownames(steps_by_interval) == activity[i,]$interval);
     activity_imputed$steps[i] <- as.integer(steps_by_interval[j]);
  }
}
```

Calculate the mean/median again, and draw the histogram & time series plot. We can see that the missing value imputing is pretty make sense because the statistics summary & graph looks almost the same.

```
steps_by_date <- aggregate(steps ~ date, data=activity_imputed, sum)
hist(activity_imputed$steps)
```



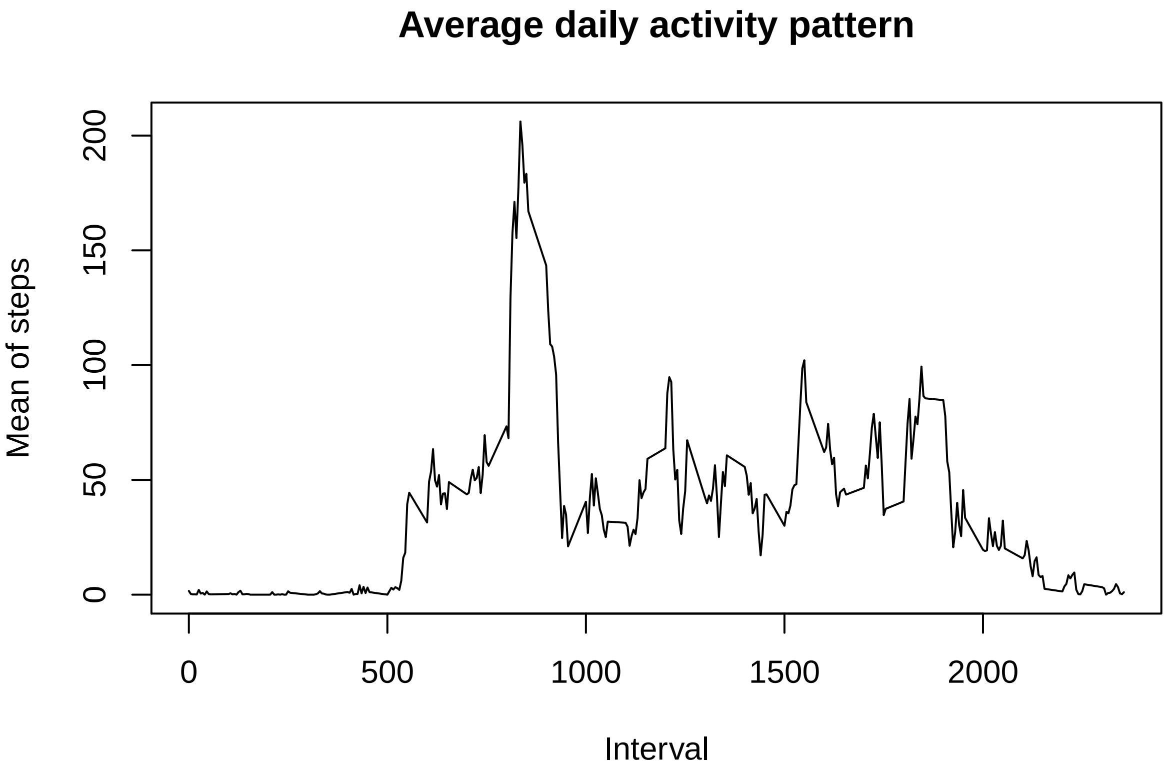
```
mean(steps_by_date$steps)
```

```
## [1] 10749.77
```

```
median(steps_by_date$steps)
```

```
## [1] 10641
```

```
steps_by_interval <- tapply(activity_imputed$steps,activity_imputed$interval,mean)
plot(row.names(steps_by_interval), steps_by_interval, type="l",
     main="Average daily activity pattern", xlab="Interval", ylab="Mean of steps")
```



```
names(which.max(steps_by_interval))
```

```
## [1] "835"
```

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

Yes, it is totally different. From the plot, we can see the weekdays & weekends is totally different steps distribution.