Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

```
if (file.exists("activity.csv")){file.remove("activity.csv")}
```

```
## [1] TRUE
```

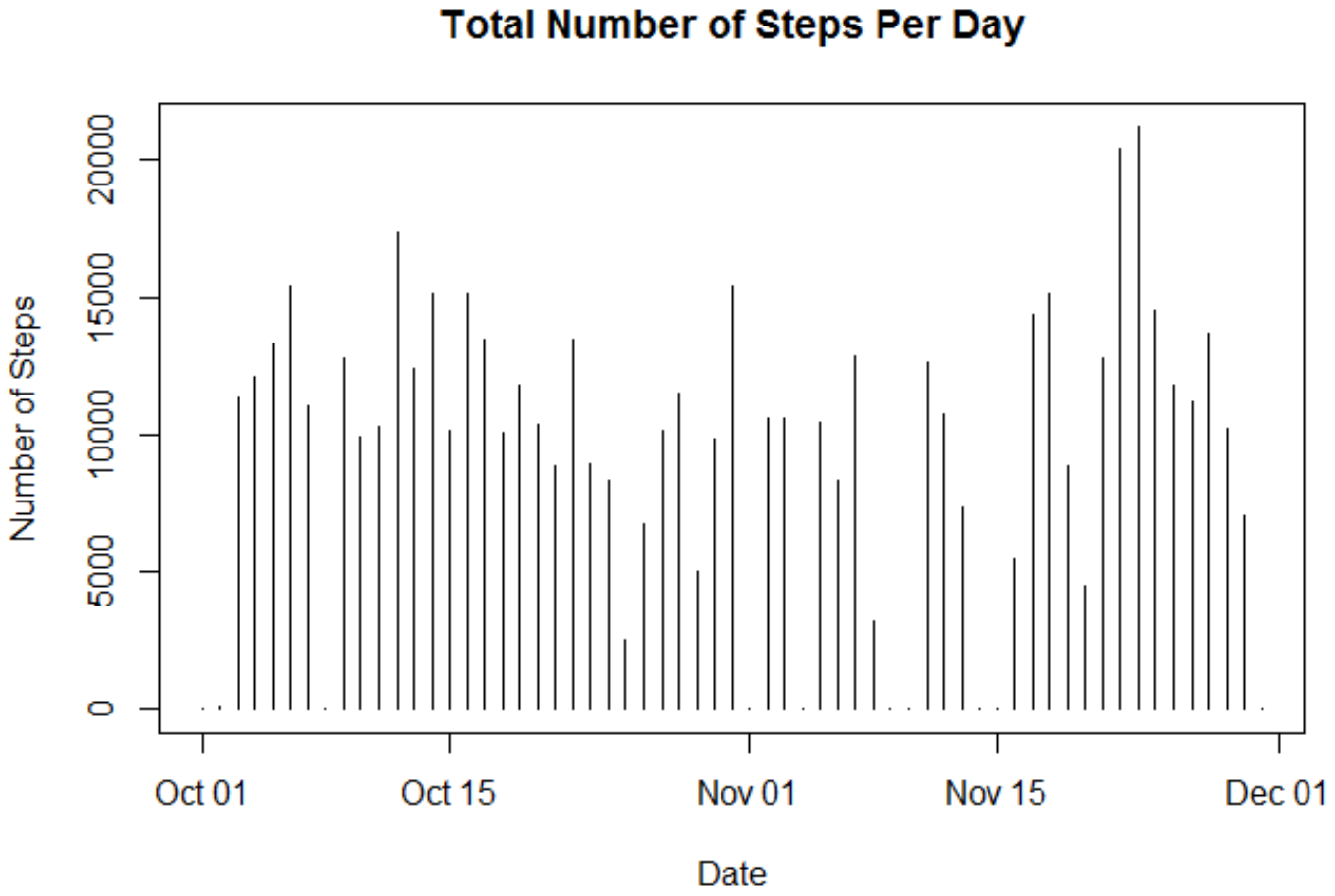
```
unzip('activity.zip')

activity <- read.csv("activity.csv")

activity_sum_day <- as.data.frame.table(with(activity, tapply(steps, date, sum, na.rm=TRU
E)))

names(activity_sum_day) <- c("date", "steps")

activity_mean <- mean(activity_sum_day$steps)
activity_sum_day$date <- as.Date(as.character(activity_sum_day$date))
activity_summary <- summary(activity_sum_day$steps)[c(3,4)]
with(activity_sum_day, plot(date, steps, type = "h", xlab = "Date", ylab = "Number of Ste
ps", main = "Total Number of Steps Per Day"))
```



What is mean total number of steps taken per day?

```
print(activity_summary)
```

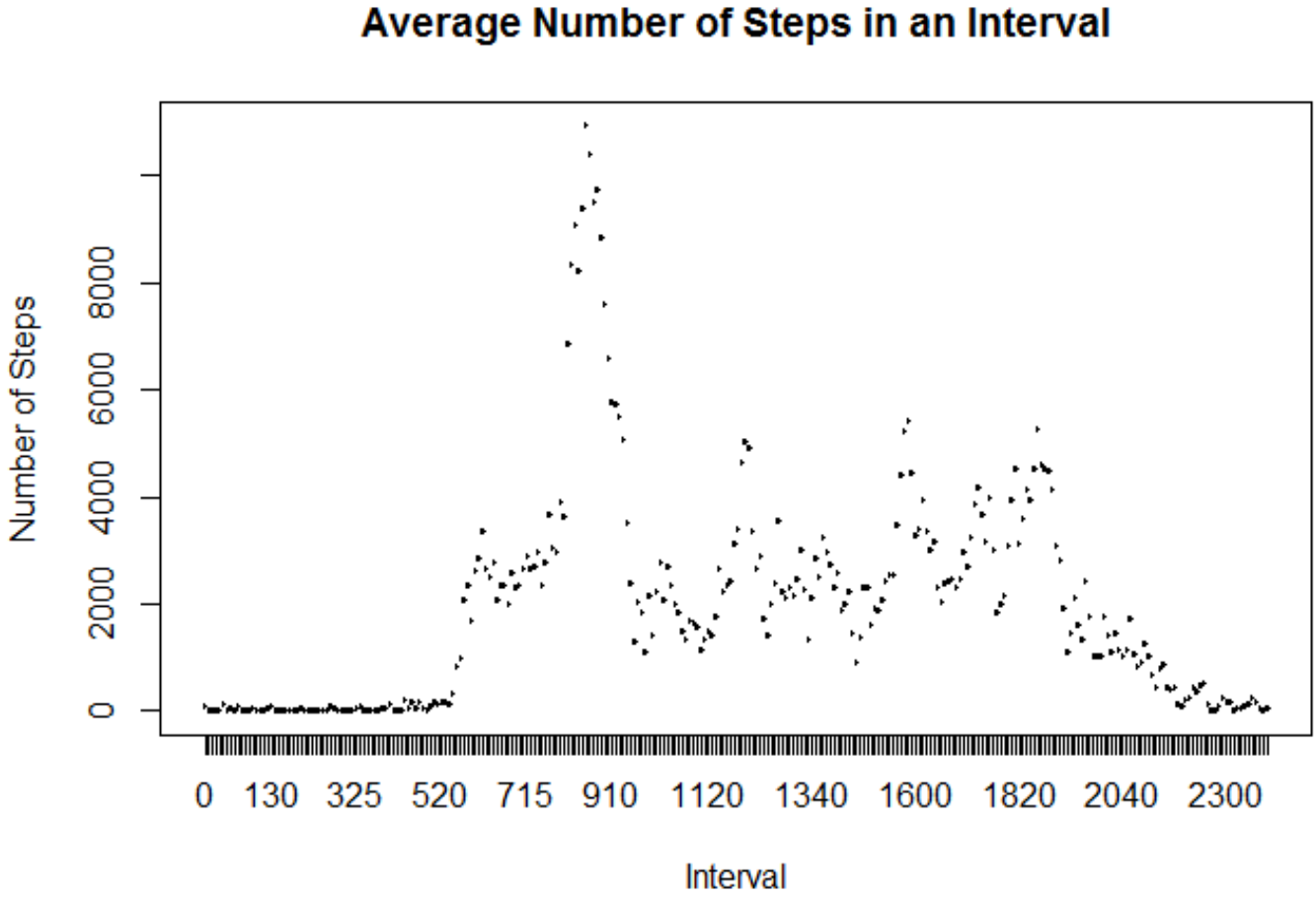
```
## Median   Mean
##  10400   9354
```

What is the average daily activity pattern?

```
activity_average_interval <- as.data.frame.table(with(activity, tapply(steps, as.factor(i
nterval), sum, na.rm=TRUE)))

names(activity_average_interval) <- c("interval", "steps")

with(activity_average_interval, plot(interval, steps, type = "l", xlab = "Interval", ylab
= "Number of Steps", main = "Average Number of Steps in an Interval"))
```




```
max_interval <- which.max( activity_average_interval[,2] )

maximum_step_time <- as.integer(as.character(activity_average_interval[max_interval, 1]))
```

The interval that contains the maximum number of steps

```
print(maximum_step_time)
```

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

Imputing missing values

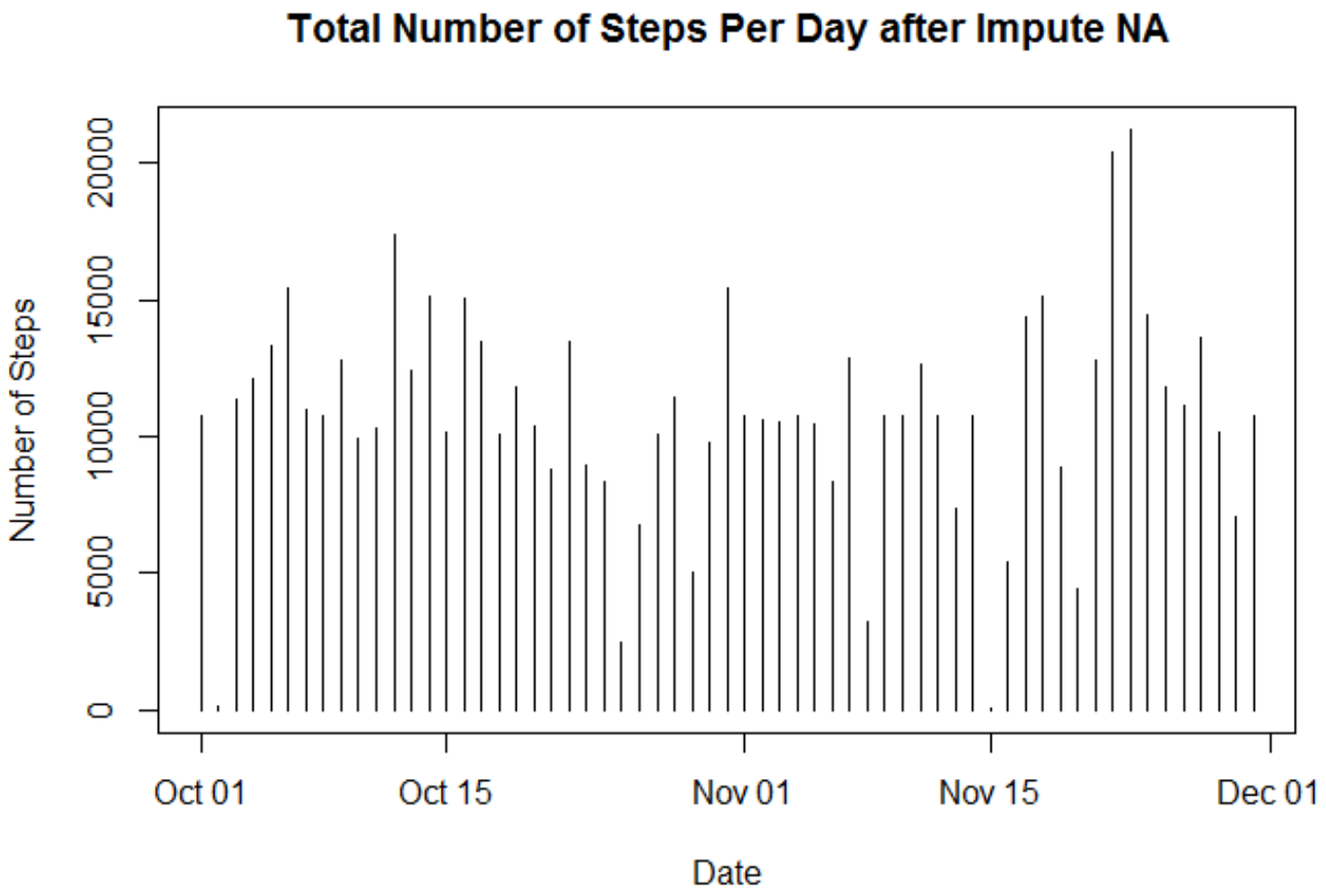
This will take the most amount of time, replace all NA with the mean of the respective interval associated with it.

```
activity_new <- activity
activityna <- sum(is.na(activity$steps))
na_id <- which(is.na(activity$steps))
activity_interval_mean <- as.data.frame.table(with(activity, tapply(steps, as.factor(inte
rval), mean, na.rm=TRUE)))
names(activity_interval_mean) <- c("interval", "steps")
ilength <- length(activity_interval_mean$interval)
for (i in 1:activityna){ for (j in 1:ilength){if (activity[na_id[i],3]==activity_interval
_mean[j,1]){activity_new[na_id[i],1] <- activity_interval_mean[j,2]}else{next}}}

activity_new_sum_day <- as.data.frame.table(with(activity_new, tapply(steps, date, sum)))

names(activity_new_sum_day) <- c("date", "steps")

activity_new_mean <- mean(activity_new_sum_day$steps)
activity_new_sum_day$date <- as.Date(as.character(activity_new_sum_day$date))
activity_new_summary <- summary(activity_new_sum_day$steps)[c(3,4)]
with(activity_new_sum_day, plot(date, steps, type = "h", xlab = "Date", ylab = "Number of
Steps", main = "Total Number of Steps Per Day after Impute NA"))
```



The mean and medium of the steps using the file after imputing

```
print(activity_new_summary)
```

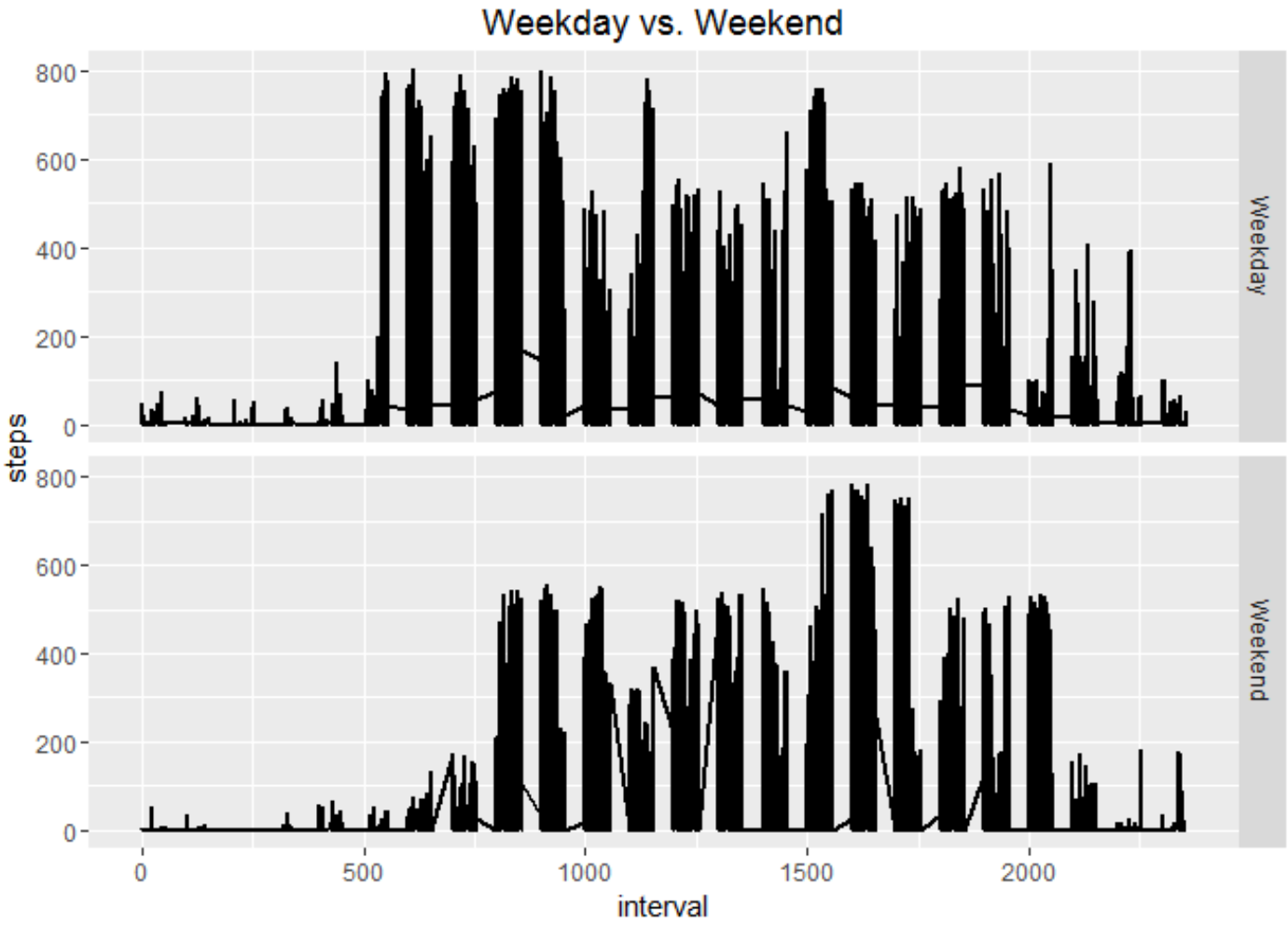
```
## Median   Mean
##  10770  10770
```

Are there differences in activity patterns between weekdays and weekends?

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
activity_new$wk_day <- as.POSIXlt(as.character(activity_new$date))$wday
activity_new_weekday <- subset(activity_new, wk_day>0 & wk_day<6, select = -wk_day)
activity_new_weekend <- subset(activity_new, wk_day<1 | wk_day>5, select = -wk_day)
activity_new_weekday$Wday <- "Weekday"
activity_new_weekend$Wday <- "Weekend"
activity_new2 <- rbind(activity_new_weekday, activity_new_weekend)
par(mfrow = c(2,1))
p <- ggplot(data=activity_new2, aes(x = interval, y = steps))
q <- p+ geom_line(size = 1)
r <- q+ facet_grid(Wday~.)
t <- r+ labs(title = "Weekday vs. Weekend")
print(t)
```



There are subtle differences comparing Weekday and Weekend Activity during most intervals