title: Peer assessment 1 Reproducible Research Week 2 author: Madelon den Boeft date: March 13, 2018 output: md_document
==========

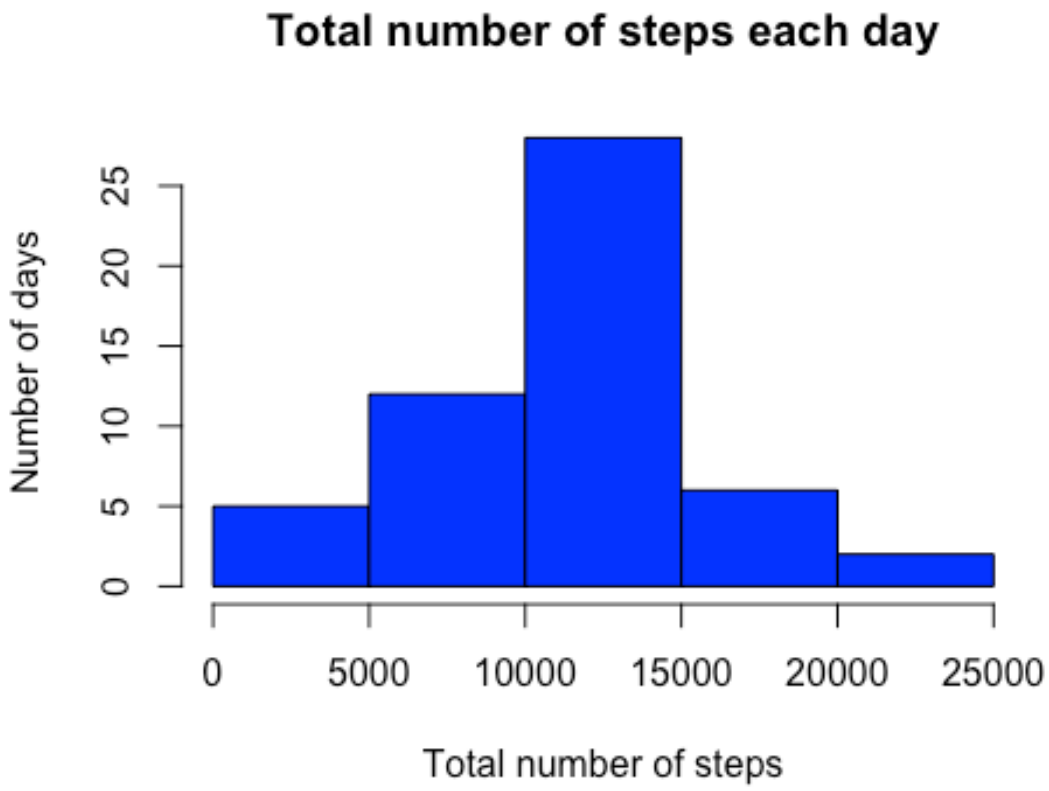
1. Load the data activity

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

2. What is the mean total number of steps taken per day? ignore missing values?

Make a histogram of the total number of steps taken each day

```
total_steps_bydate<- aggregate(steps ~ date, activity_dataset, sum)
hist(total_steps_bydate$steps, main = paste("Total number of steps each
day"), col="blue", xlab="Total number of steps", ylab= "Number of days")
```



Calculate and report the mean and median total number of steps taken each day

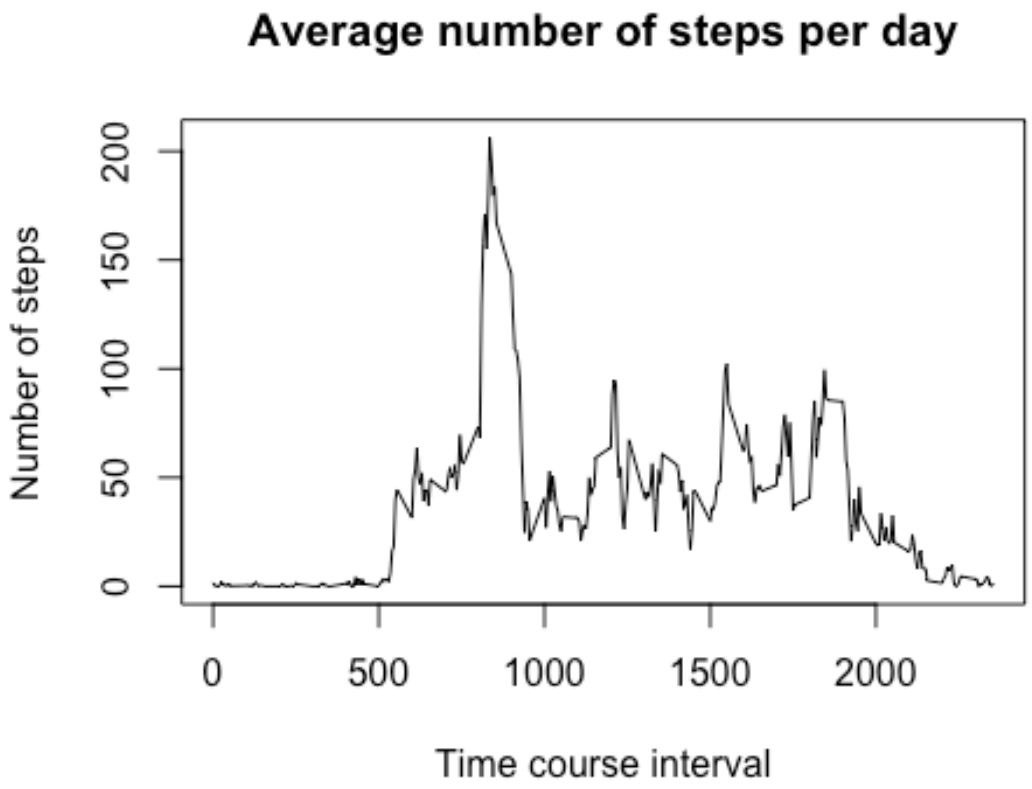
```
mean_steps <- mean(total_steps_bydate$steps)
median_steps <- median(total_steps_bydate$steps)
sprintf("Mean total number of steps taken per day = %.2f", mean_steps)

## [1] "Mean total number of steps taken per day = 10766.19"

sprintf("Median total number of steps taken per day = %.2f", median_steps)
```

```
## [1] "Median total number of steps taken per day = 10765.00"
```

3. What is the average daily activity pattern?

Make a time series plot of the 5-minute interval and average number of steps taken averaged across all days

```
steps_by_interval <- aggregate(steps ~ interval, activity_dataset, mean)
plot(steps_by_interval$interval, steps_by_interval$steps, type="l",
xlab="Time course interval", ylab="Number of steps",main="Average number of
steps per day")
```



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

```
max_interval <- steps_by_interval[which.max(steps_by_interval$steps),1]
sprintf("5-minute interval with maximum number of steps = %.2f",
max_interval)

## [1] "5-minute interval with maximum number of steps = 835.00"
```

4. Imputing missing values

Calculate and report the total number of missing values in the dataset

```
total_missing_data <- sum(!complete.cases(activity_dataset))
sprintf("Total number of missing values  = %.2f", total_missing_data)

## [1] "Total number of missing values  = 2304.00"
```

Devise a strategy for filling in all the missing values in the dataset Create a new dataset that is equal to the original one but with the missing data filled in

```
imputed_data <- transform(activity_dataset, steps =
ifelse(is.na(activity_dataset$steps),
steps_by_interval$steps[match(activity_dataset$interval,
steps_by_interval$interval)], activity_dataset$steps))
```

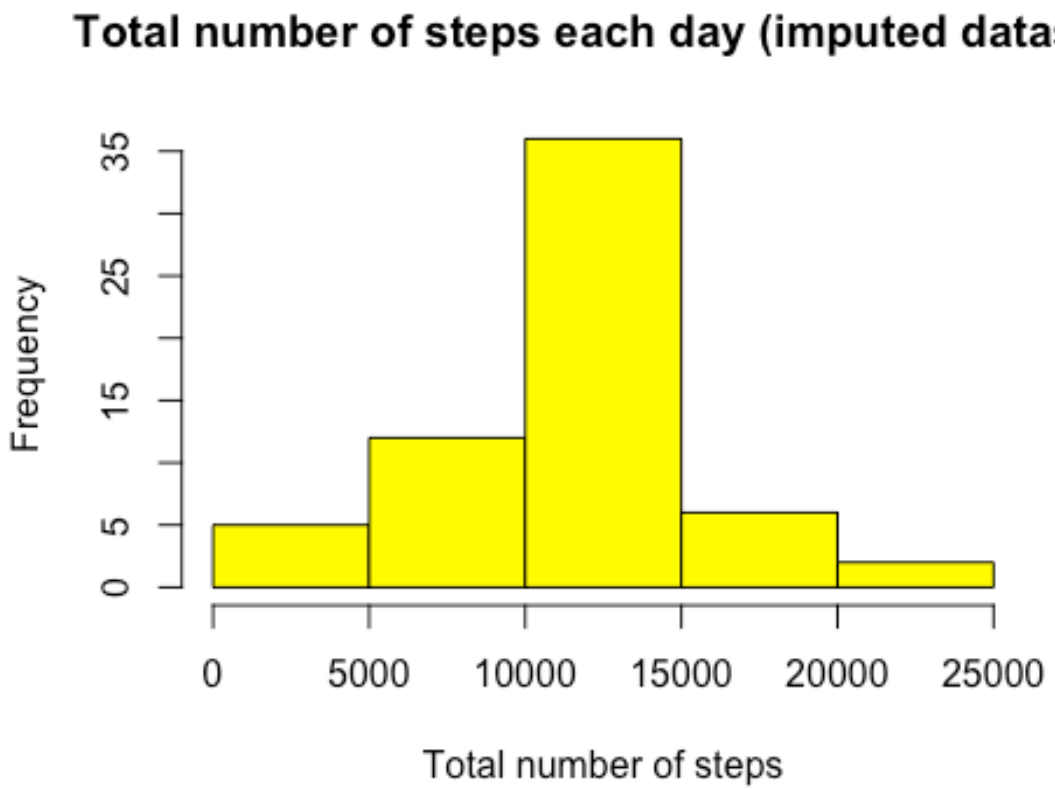
Test variable missing data to see if all the missings are gone / indeed imputed

```
test_missing_data <- sum(!complete.cases(imputed_data))
```

Test result positive! Continue with the imputed_data dataset...

Make a histogram of the total number of steps each day

```
total_steps_eachday_imputed<- aggregate(steps ~ date, imputed_data, sum)
hist(total_steps_eachday_imputed$steps, main = paste("Total number of steps
each day (imputed dataset)"), col="yellow", xlab="Total number of steps")
```



Calculate and report the mean and median total number of steps each day

```
mean_steps_imputed <- mean(total_steps_eachday_imputed$steps)
median_steps_imputed <- median(total_steps_eachday_imputed$steps)
sprintf("Mean total number of steps taken per day (imputed)= %.2f",
mean_steps_imputed)

## [1] "Mean total number of steps taken per day (imputed)= 10766.19"

sprintf("Median total number of steps taken per day (imputed) = %.2f",
median_steps_imputed)

## [1] "Median total number of steps taken per day (imputed) = 10766.19"
```

Do these values differ?

```
sprintf("The difference in mean between the original and imputed dataset is
%.2f ",mean_steps_imputed-mean_steps)

## [1] "The difference in mean between the original and imputed dataset is
0.00 "

sprintf("The difference in median between the original and imputed dataset is
%.2f ",median_steps_imputed-median_steps)

## [1] "The difference in median between the original and imputed dataset is
1.19 "
```

What is the impact of imputing the missing values on the estimates of the total daily number of steps?

```
total <- sum(total_steps_bydate$steps)
total_imputed <- sum(total_steps_eachday_imputed$steps)
sprintf("The difference in total number of steps between original and imputed
dataset = %.2f", total_imputed-total)

## [1] "The difference in total number of steps between original and imputed
dataset = 86129.51"
```

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

Create a factor variable with two levels “weekday” and “weekend”

```
weekend_days <- c("zaterdag", "zondag")
week_days <- c("maandag", "dinsdag", "woensdag", "donderdag", "vrijdag")

imputed_data$dow =
as.factor(ifelse(is.element(weekdays(as.Date(imputed_data$date)),weekend_days
), "Weekend", "Weekday"))
imputed_steps_by_interval <- aggregate(steps ~ interval + dow, imputed_data,
mean)
```

Make a panel plot containinga time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days

```
library(lattice)
xyplot(imputed_steps_by_interval$steps ~
imputed_steps_by_interval$interval|imputed_steps_by_interval$dow,
main="Average Steps per Day by Interval",xlab="Interval",
ylab="Steps",layout=c(1,2), type="l")
```