Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

Extract the data from the zipped file and load it as a dataframe, and view data.

```
unzip("activity.zip", exdir="./data")
df <- read.csv("./data/activity.csv")
head(df)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

Create a new column of the dataframe which is a date-and-time in a native R format.

```
df$minutes <- df$interval %% 100
df$hours <- (df$interval - df$minutes)/100
df$date_and_time <- strptime(paste(df$date, df$hours, df$minutes),
                             format="%Y-%m-%d %H %M", tz="GMT")
```

Finally we can check that we can check the start and finish date for the logging, and that we have continuous records (even though some are recorded as NA):

```
summary(df$date_and_time)
```

```
##                  Min.               1st Qu.                Median
## "2012-10-01 00:00:00" "2012-10-16 05:58:45" "2012-10-31 11:57:30"
##                  Mean               3rd Qu.                  Max.
## "2012-10-31 11:57:30" "2012-11-15 17:56:15" "2012-11-30 23:55:00"
```

```
summary(as.numeric(diff(df$date_and_time)))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5       5       5       5       5       5
```

What is mean total number of steps taken per day?

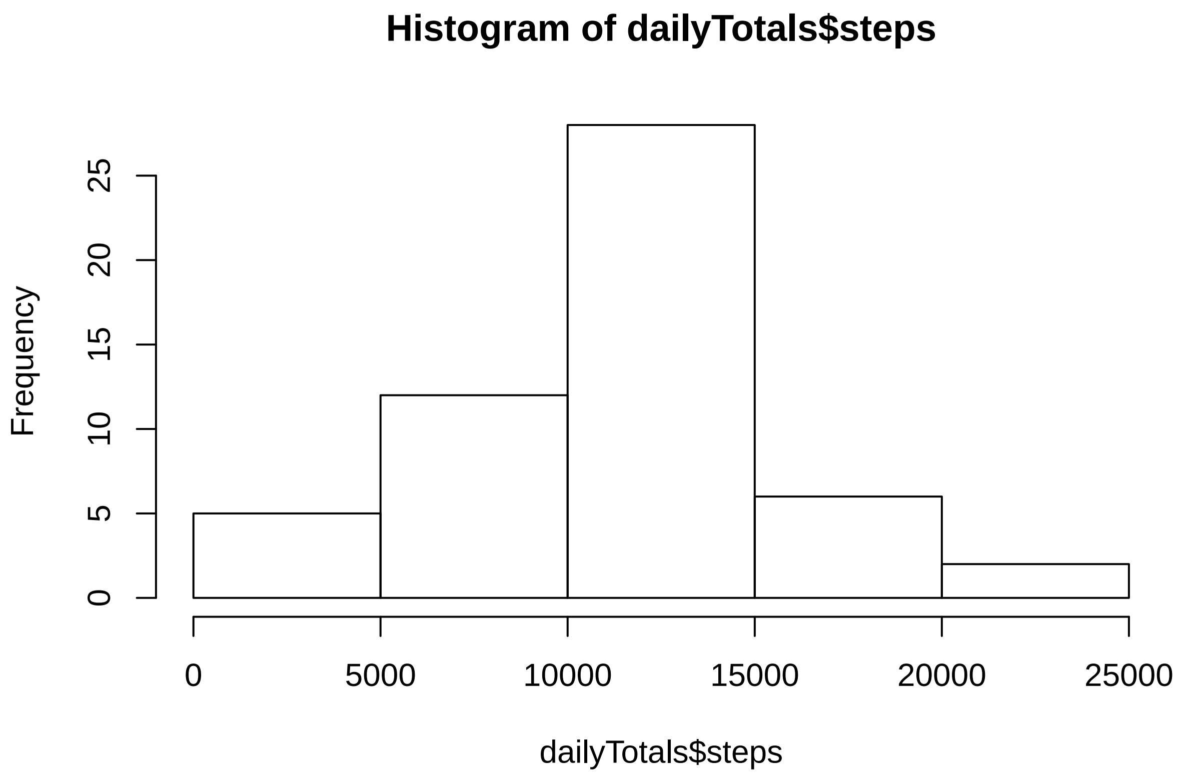
Aggregate the daily totals of steps. Note some days are not represented in this total as some entire days have NA steps.

```
dailyTotals <- aggregate(steps ~ date, df, FUN = sum)
head(dailyTotals)
```

```
##          date steps
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

A histogram of the total number of steps taken each day:

```
hist(dailyTotals$steps)
```



The mean and median of the number of steps per day are:

```
summary(dailyTotals$steps)[c(4,3)]
```

Mean Median 10800 10800

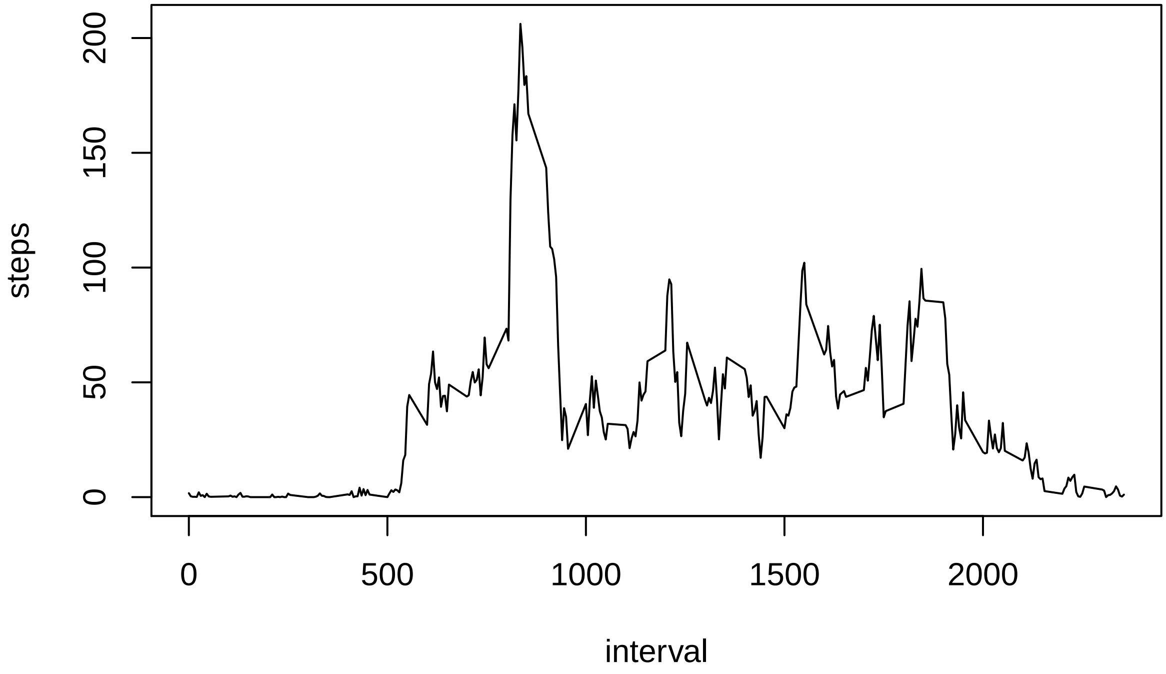
What is the average daily activity pattern?

Aggregate each of the time-steps into an average value for that time-step across all days, and plot the result.

```
intervalMeans <- aggregate(steps ~ interval, df, FUN = mean)
head(intervalMeans)
```

```
##   interval   steps
## 1        0 1.71698
## 2        5 0.33962
## 3       10 0.13208
## 4       15 0.15094
## 5       20 0.07547
## 6       25 2.09434
```

```
plot(intervalMeans, type="l")
```



The interval number which has on average the most steps is:

```
intervalMeans[intervalMeans$steps == max(intervalMeans$steps), 'interval']
```

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

i.e. from 08:25 till 08:30 in the morning, assuming intervals are labelled with their end time.

Imputing missing values

The total number of intervals with missing data is:

```
sum(is.na(df$steps))
```

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

Representing the following fraction of all of the data:

```
sum(is.na(df$steps))/length(df$steps)
```

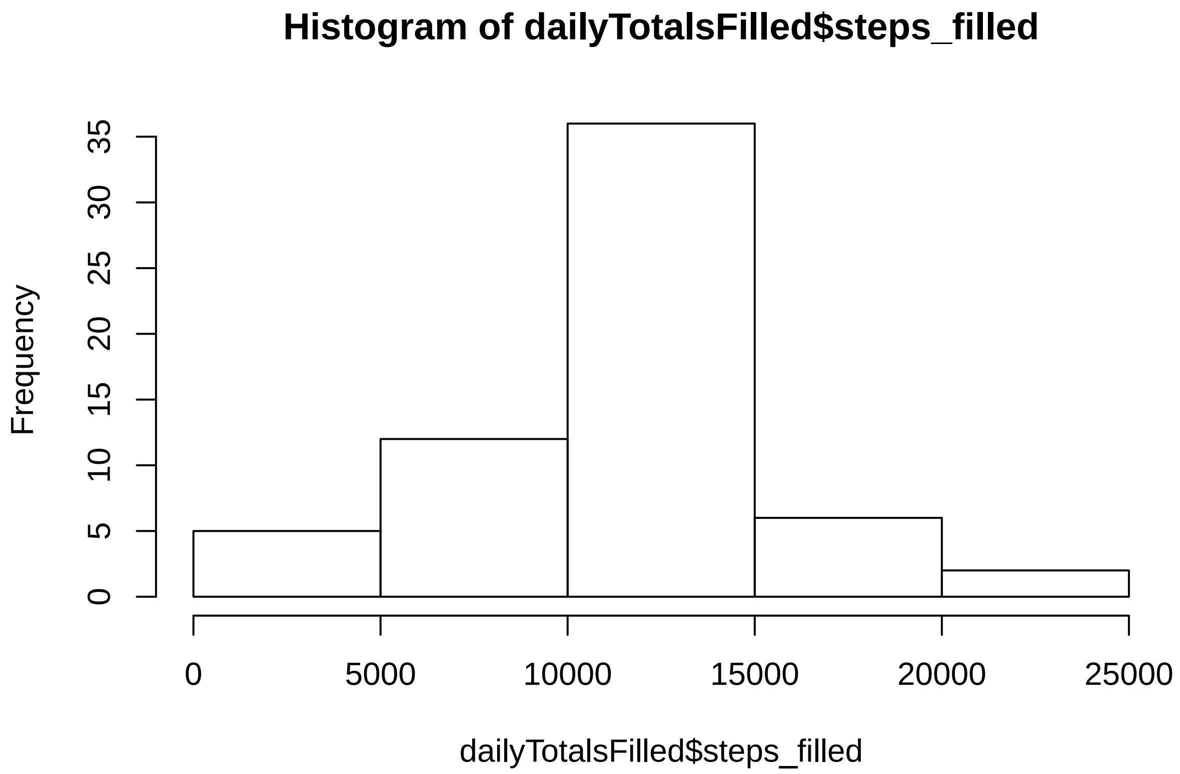
```
## [1] 0.1311
```

I am imputing missing values based on the mean for that 5-minute interval (this was chosen in preference to the average for the day as some days have no recorded data). Save this in a new column which includes the original steps data and the filled in data.

```
df$steps_filled <- apply(df[,c('steps', 'interval')], 1, function(x) { ifelse(is.na(x[1]),
                        intervalMeans[intervalMeans$interval == x[2], 'steps'], x[1]) } )
```

Here is a histogram of the daily totals used including the imputed values (NB: all days are now represented, unlike the original histogram).

```
dailyTotalsFilled <- aggregate(steps_filled ~ date, df, FUN = sum)
hist(dailyTotalsFilled$steps_filled)
```



```
summary(dailyTotalsFilled$steps_filled)[c(4,3)]
```

Mean Median 10800 10800

Are there differences in activity patterns between weekdays and weekends?