# Reproducible Course Project 1

*Ophelia*

*June 14, 2016*

```
knitr::opts_chunk$set(echo = TRUE)
```

## part1: read and process data

This is my first knitr document. The actiity monitoring dataset is stored in a csv file and there are 17,568 observation and 3 variables (steps, date, interval) in this dataset.
It is downloaded from https://www.coursera.org/learn/reproducible-research/peer/gYyPt/course-project-1 (https://www.coursera.org/learn/reproducible-research/peer/gYyPt/course-project-1) Here is code chunk reading data from working directory

```
dataset <- read.csv("activity.csv", header=TRUE, sep=",", stringsAsFactors = FALSE)
```

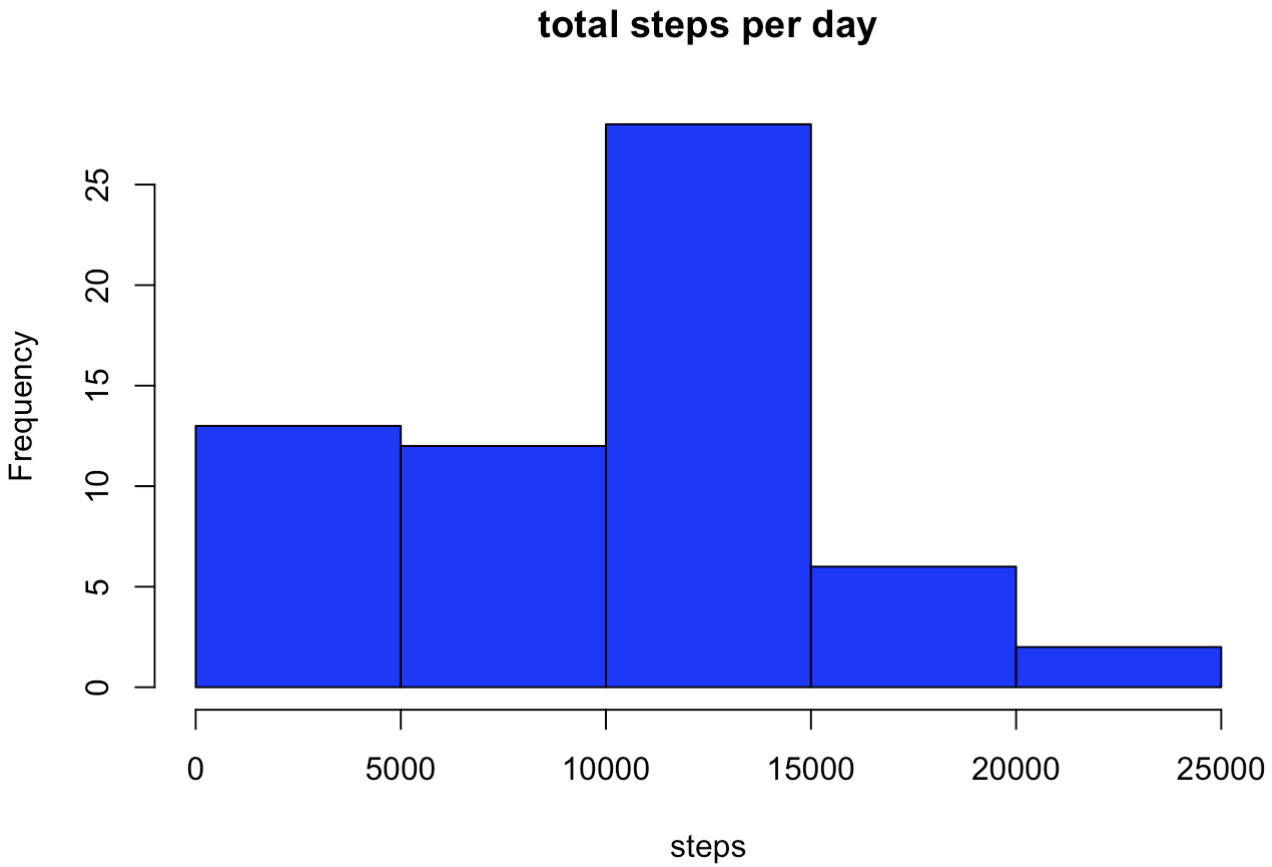
## part2: histogram of total number steps taken per day

Here is code chunk generating plot

```
aggData1 <- aggregate(dataset[,'steps'], by=list(Group.date=dataset$date),FUN=sum,na.rm=TRUE)
aggData1$Group.date <- as.Date(aggData1$Group.date)
head(aggData1)
```

```
##   Group.date     x
## 1 2012-10-01     0
## 2 2012-10-02   126
## 3 2012-10-03 11352
## 4 2012-10-04 12116
## 5 2012-10-05 13294
## 6 2012-10-06 15420
```

```
hist(aggData1$x,col="blue",main="total steps per day",xlab="steps",freq=TRUE)
```




The aggregated data summarizing

total number steps taken each day with missing data removed contains 61 observations, of which 8 of them are zeros steps.

# part3: mean and median values of part2

Here is code chunk of summary of total steps taken each day

```
summary(aggData1)
```

```
##    Group.date               x
##  Min.   :2012-10-01   Min.   :    0
##  1st Qu.:2012-10-16   1st Qu.: 6778
##  Median :2012-10-31   Median :10395
##  Mean   :2012-10-31   Mean   : 9354
##  3rd Qu.:2012-11-15   3rd Qu.:12811
##  Max.   :2012-11-30   Max.   :21194
```

The mean of total numbers of steps taken per day is 9354. The median of total numbers of steps taken per day is 10395.

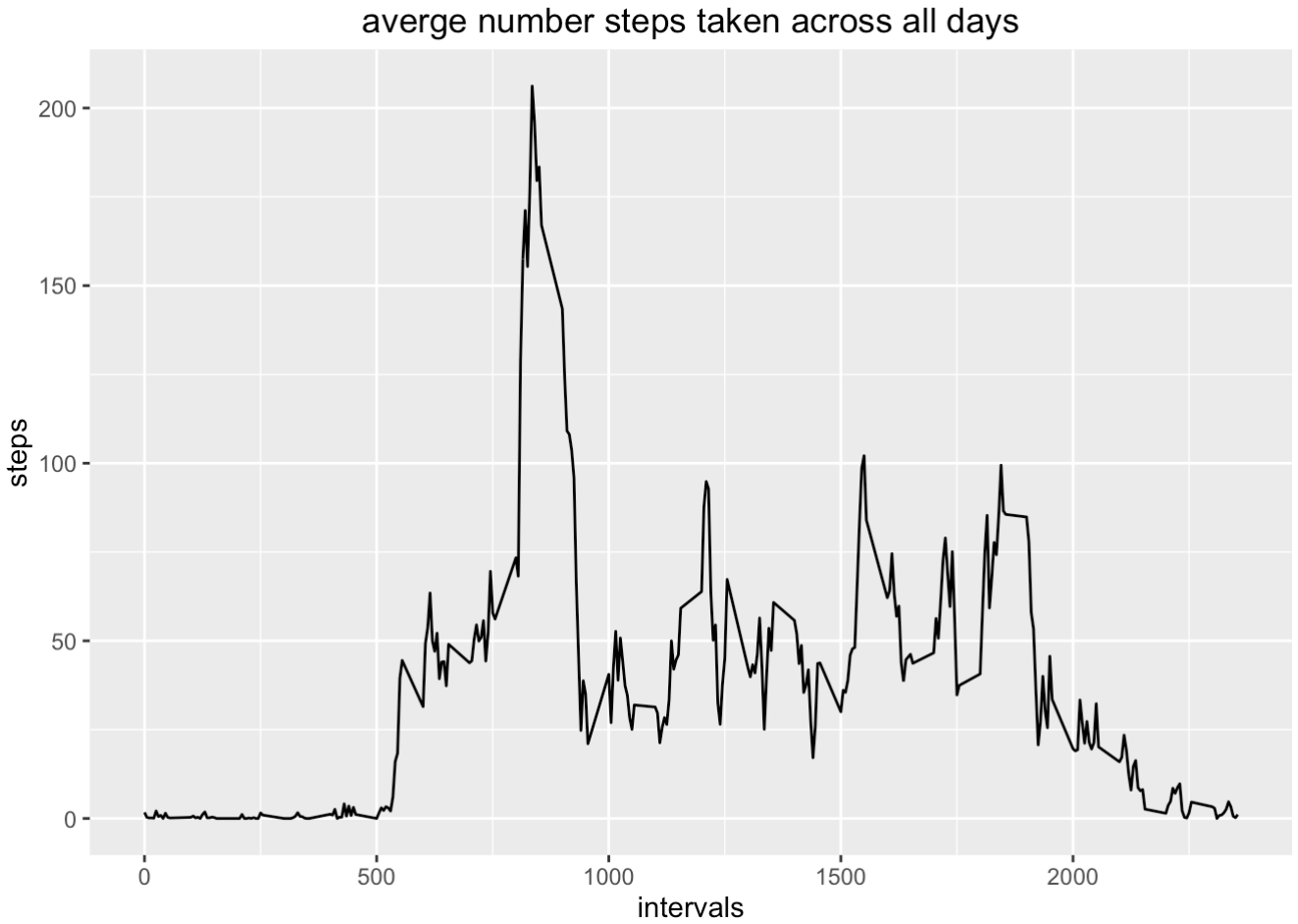
# part4: time series plot of average steps taken

Here is code chunk generating plot

```
library(ggplot2)
aggData2 <- aggregate(dataset[,'steps'], by=list(Group.interval=dataset$interval),FUN=mean,na.rm=TRUE)
head(aggData2)
```

```
##   Group.interval         x
## 1              0 1.7169811
## 2              5 0.3396226
## 3             10 0.1320755
## 4             15 0.1509434
## 5             20 0.0754717
## 6             25 2.0943396
```

```
ggplot(aggData2,aes(Group.interval,x))+geom_line()+xlab("intervals") +ylab("steps")+
ggtitle("averge number steps taken across all days")
```




# part5: 5-minute interval contains maximum number of steps

Here is code chunk to evaluate average number of steps in 5-minute intervals

```
maxStep <- aggData2[which.max(aggData2$x),]
```

Interval 835 contains maximuum steps of 206.

# part6: strategy for imputing missing data

Here is code chunk checking which variables misses more than 5% of data, a safe maximum threshold for dataset

```
pMis <- function(x)(sum(is.na(x)/length(x)*100))
pMiss <- apply(dataset,2,pMis)
pMiss
```

```
##    steps     date interval
## 13.11475  0.00000  0.00000
```

We see that variable steps consists 13% (2304) NAs of total 17568 rows, Here is code chunk using mice package to replace missing data with imputed values in the first of five dataset

```
library(mice)
```

```
## Loading required package: Rcpp
```

```
## mice 2.25 2015-11-09
```

```
tempData1 <- mice(dataset,m=5,maxit=20,meth='pmm',seed=500)
```

```
##
##  iter imp variable
##   1   1  steps
##   1   2  steps
##   1   3  steps
##   1   4  steps
##   1   5  steps
##   2   1  steps
##   2   2  steps
##   2   3  steps
##   2   4  steps
##   2   5  steps
##   3   1  steps
##   3   2  steps
##   3   3  steps
##   3   4  steps
##   3   5  steps
##   4   1  steps
##   4   2  steps
##   4   3  steps
##   4   4  steps
##   4   5  steps
##   5   1  steps
##   5   2  steps
##   5   3  steps
##   5   4  steps
##   5   5  steps
##   6   1  steps
##   6   2  steps
##   6   3  steps
##   6   4  steps
##   6   5  steps
##   7   1  steps
##   7   2  steps
##   7   3  steps
##   7   4  steps
##   7   5  steps
##   8   1  steps
##   8   2  steps
##   8   3  steps
```

```
##    8   4  steps
##    8   5  steps
##    9   1  steps
##    9   2  steps
##    9   3  steps
##    9   4  steps
##    9   5  steps
##    10   1  steps
##    10   2  steps
##    10   3  steps
##    10   4  steps
##    10   5  steps
##    11   1  steps
##    11   2  steps
##    11   3  steps
##    11   4  steps
##    11   5  steps
##    12   1  steps
##    12   2  steps
##    12   3  steps
##    12   4  steps
##    12   5  steps
##    13   1  steps
##    13   2  steps
##    13   3  steps
##    13   4  steps
##    13   5  steps
##    14   1  steps
##    14   2  steps
##    14   3  steps
##    14   4  steps
##    14   5  steps
##    15   1  steps
##    15   2  steps
##    15   3  steps
##    15   4  steps
##    15   5  steps
##    16   1  steps
##    16   2  steps
##    16   3  steps
##    16   4  steps
```

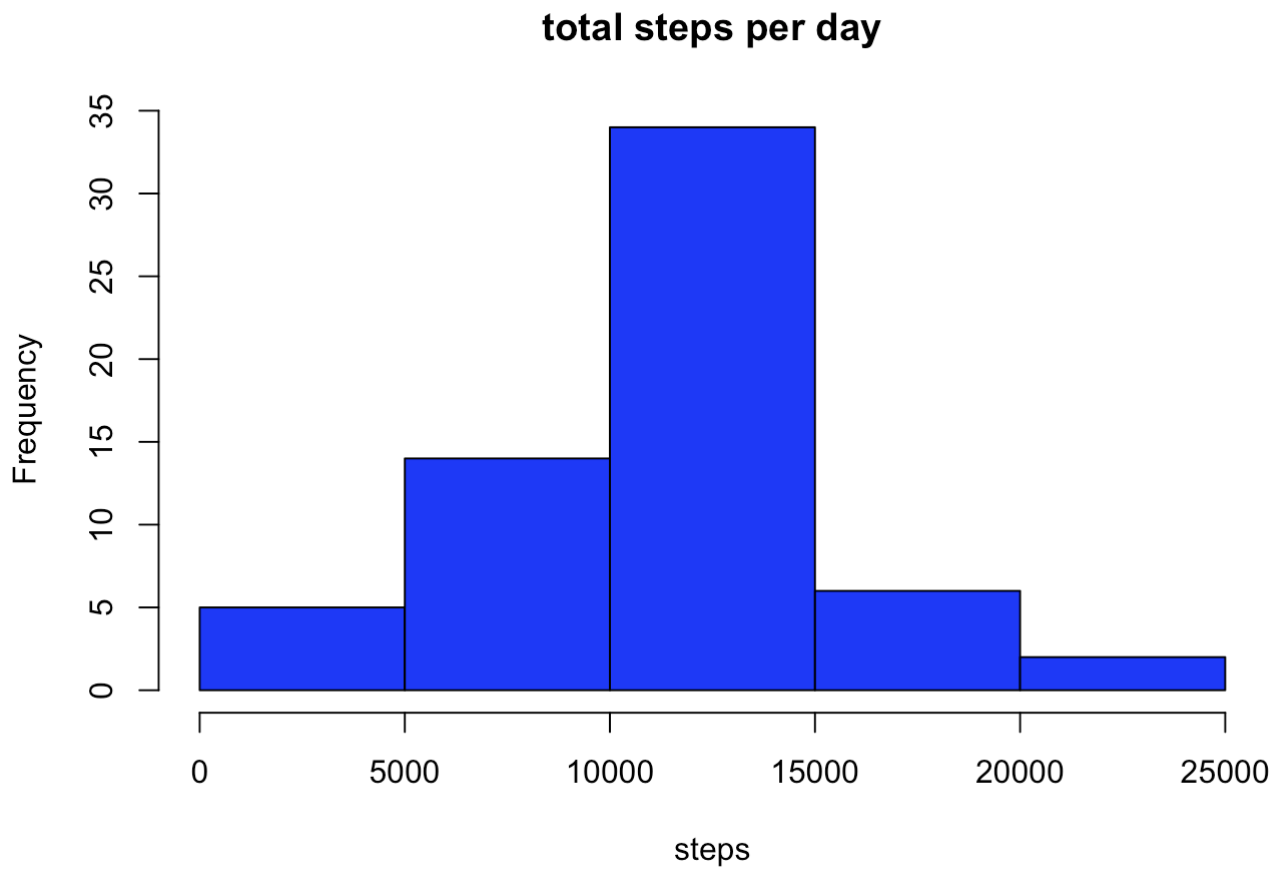
```
##   16   5  steps
##   17   1  steps
##   17   2  steps
##   17   3  steps
##   17   4  steps
##   17   5  steps
##   18   1  steps
##   18   2  steps
##   18   3  steps
##   18   4  steps
##   18   5  steps
##   19   1  steps
##   19   2  steps
##   19   3  steps
##   19   4  steps
##   19   5  steps
##   20   1  steps
##   20   2  steps
##   20   3  steps
##   20   4  steps
##   20   5  steps
```

```
completedData <- complete(tempData1,1,include=TRUE)
```

# part7: histogram of total number steps taken per day after missing values are imputed

Here is code chunk generating plot

```
aggData3 <- aggregate(completedData[,'steps'], by=list(Group.date=dataset$date),FUN=sum,na.rm=TRUE)
aggData3$Group.date <- as.Date(aggData3$Group.date)
hist(aggData3$x,col="blue",main="total steps per day",xlab="steps",freq=TRUE)
```




```
summary(aggData3)
```

```
##    Group.date                x
##  Min.   :2012-10-01   Min.   :   41
##  1st Qu.:2012-10-16   1st Qu.: 9333
##  Median :2012-10-31   Median :10654
##  Mean   :2012-10-31   Mean   :10761
##  3rd Qu.:2012-11-15   3rd Qu.:12811
##  Max.   :2012-11-30   Max.   :21194
```

The mean of total numbers of steps taken per day with imputed values is 10761. The median of total numbers of steps taken per day with imputed values is 10654. The mice function draws the mean and median closer together, making the distribution more symmetrical.

# part8: panel plot average number of steps taken per 5-minute intervals

# comparing weekends and weekdays, using imputed values

Here is code chunk generating plot

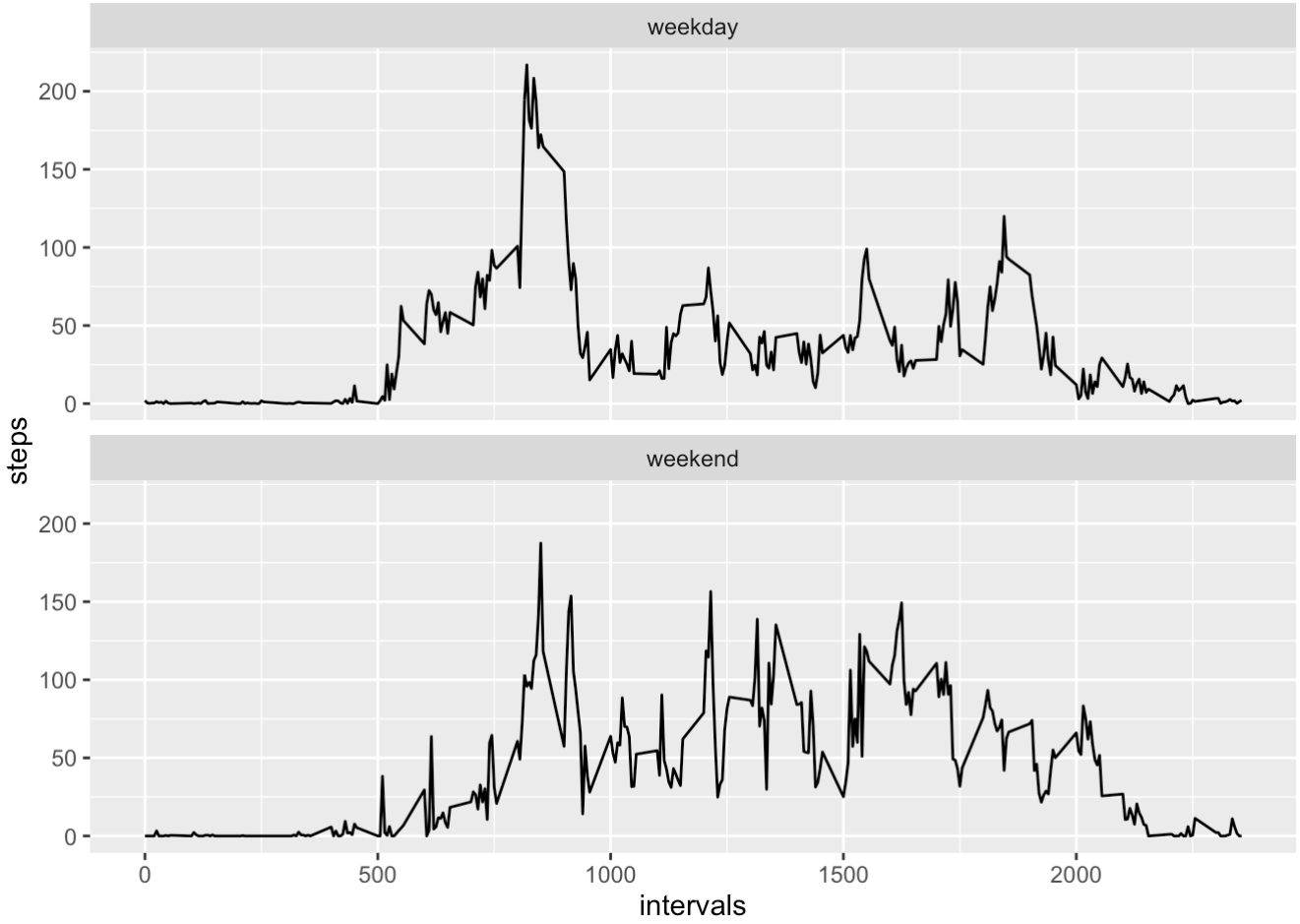
```
temp <- completedData
head(temp)
```

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

```
weekdays1 <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')
temp$date <- as.Date(temp$date)
temp$wDay <- c('weekend','weekday')[(weekdays(temp$date) %in% weekdays1)+1L]
head(temp)
```

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

```
aggData4 <- aggregate(steps ~ interval+wDay, temp, mean)
ggplot(aggData4,aes(interval,steps))+geom_line()+xlab("intervals") +ylab("steps")+
facet_wrap(~wDay,ncol=1)
```


weekday
weekend
steps
intervals
0
50
100
150
200
0
500
1000
1500
2000