Assignment 1

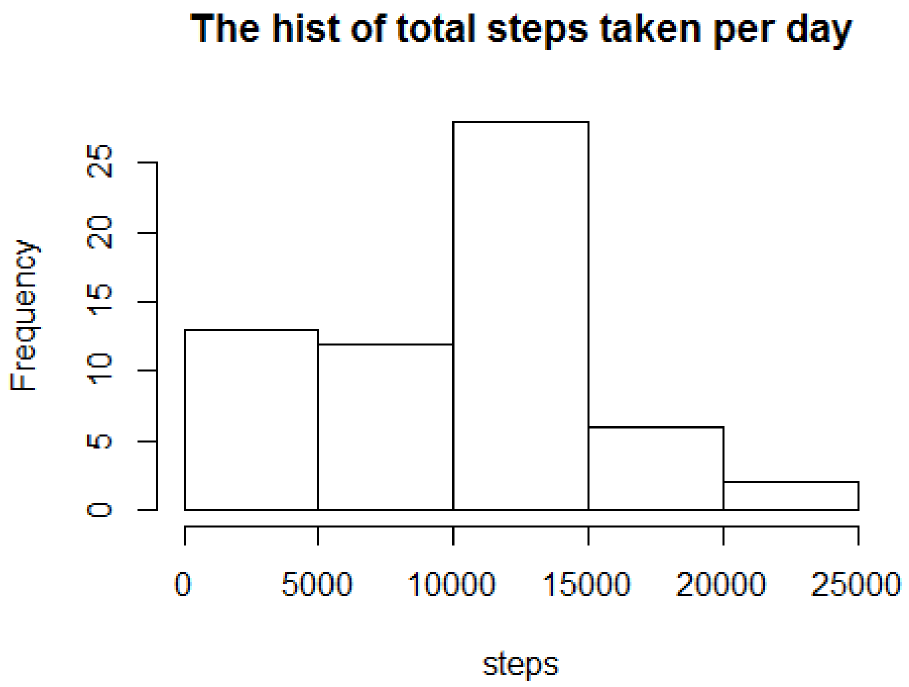
This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

Loading and preprocessing the data

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

The mean total number of steps taken per day

```
day_sum = aggregate(data$steps,by=list(data$date), sum,na.rm=TRUE)
hist(day_sum$x,main = "The hist of total steps taken per day", xlab =
"steps")
```

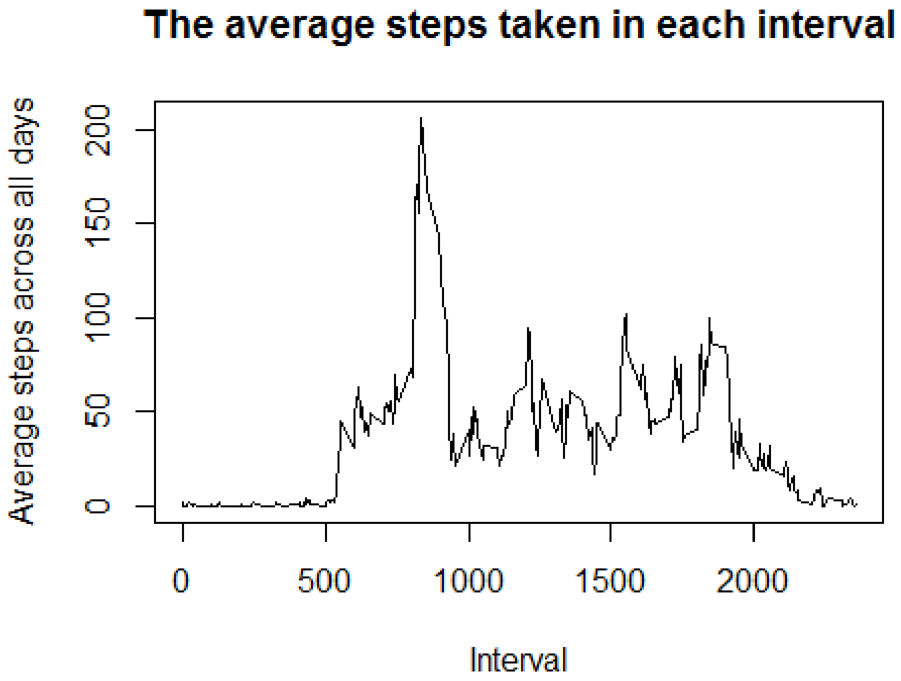


```
summary_steps = summary(day_sum$x)
```

The mean of the total number of steps is 9354, and the median of the total number steps taken per day is 10400.

The average daily activity pattern

```
interval_mean = aggregate(data$steps,by=list(data$interval),
mean,na.rm=TRUE)
plot(interval_mean[,1],interval_mean[,2],main = "The average steps
taken in each interval",xlab = "Interval",ylab = "Average steps across
all days",type = "l")
```



```
interval_max =
interval_mean[which(interval_mean$x==max(interval_mean$x)),]
```

In the [835,840] 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps, which is 206.1698113 steps.

Imputing missing values

```
ind <- complete.cases(data)
data_full = data[ind,]
total_missing = dim(data)[1]-dim(data_full)[1]
```

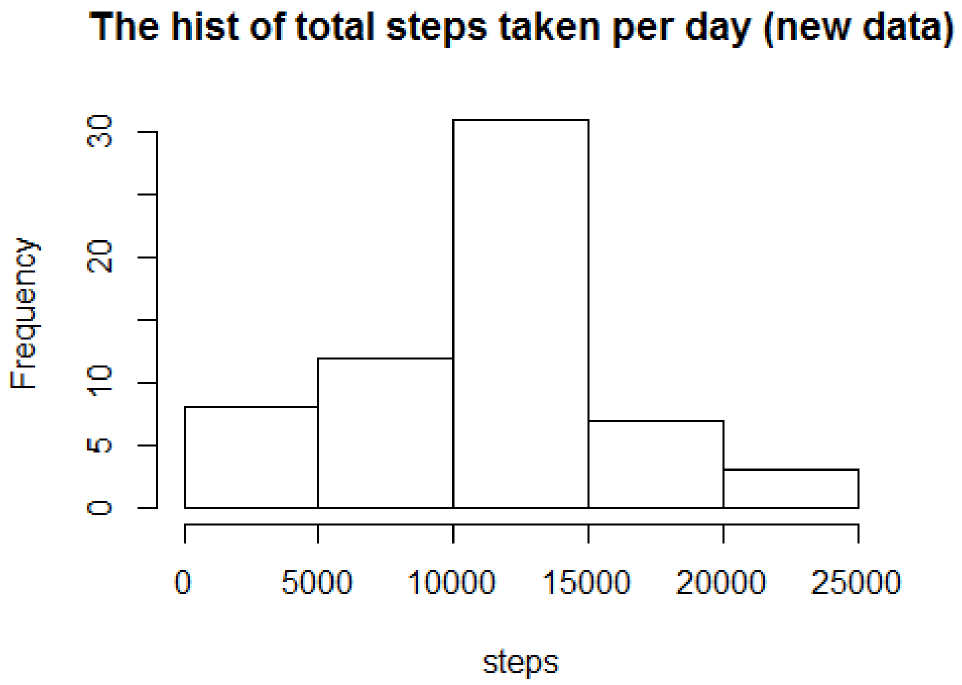
There are 2304 missing values in the dataset.

Next, We are going to fill these missing values, (further investgation suggests that there are missing values in 'steps' colum). We use the mean value of the corresponding interval to fill the mising values

```
data_new = data
fillIndeces = which(!ind)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_filling = merge(data[fillIndeces,], interval_mean,by.x =
"interval", by.y = "Group.1")%>% select(x)
data_new$steps[fillIndeces] = data_filling[,1]
day_sum_new = aggregate(data_new$steps,by=list(data_new$date),
sum,na.rm=TRUE)
hist(day_sum_new$x,main = "The hist of total steps taken per day (new
data)", xlab = "steps")
```

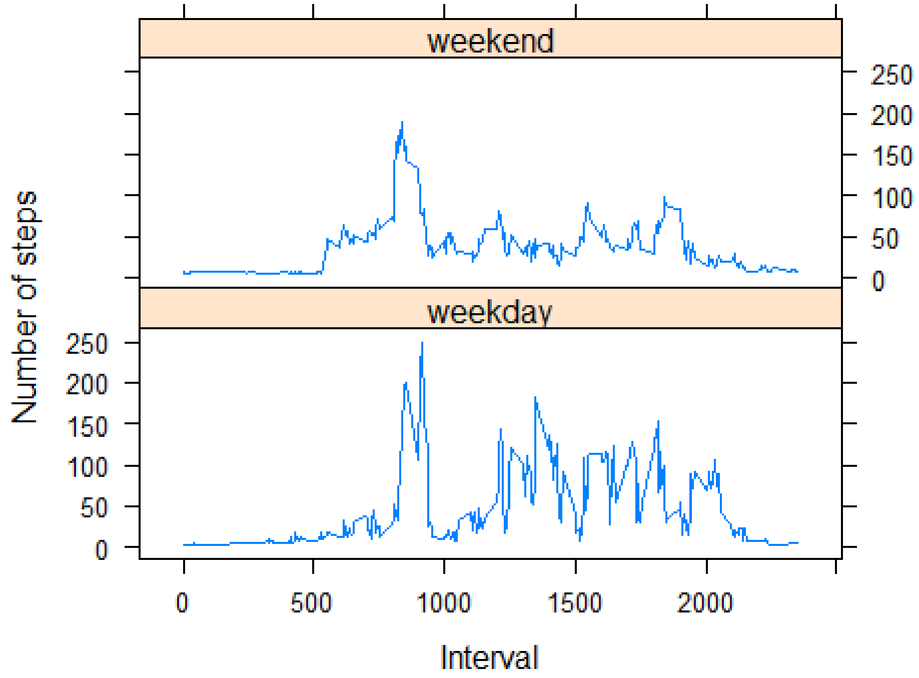


```
summary_steps_new = summary(day_sum_new$x)
```

The mean of the total number of steps is 10770, and the median of the total number steps taken per day is 11020. These values do differs from the estimates from the first part of the assignement. The imputting missing data on the estimates from of the total daily number of steps makes the mean and median bigger.

The difference in activity patters between weekdays and weekends

```
weekdays_data= weekdays(as.Date(data_new$date))
is_weekend<-function(x) {
        if( x == "Saturday" | x == "Saturday") {
                return("weekday")
        }
        else return("weekend")
}
data_new$Weekdays <- sapply(weekdays_data,is_weekend)
interval_mean_week = data_new %>% group_by(Weekdays,interval) %>%
summarise(steps=mean(steps))
library(lattice)
xyplot(interval_mean_week$steps~interval_mean_week$interval |
interval_mean_week$Weekdays, type="l",layout=c(1,2), xlab =
"Interval",ylab = "Number of steps")
```


weekend
weekday
Number of steps
Interval