# Reproducible Research Course Project 1

### Introduction

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.

The variables included in this dataset are:

- steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)
- date: The date on which the measurement was taken in YYYY-MM-DD format
- interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

## Process data

```
rm(list = ls())
library(ggplot2)

#fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
#filename<-"./Reproducible Research/week2/repdata%2Fdata%2Factivity.zip"
#download.file(fileURL, filename)
#unzip(filename)

activity<- read.csv("./Reproducible Research/week2/activity.csv")
summary(activity)
```

```
##      steps            date              interval
##  Min.   :  0.00   Length:17568       Min.   :   0.0
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8
##  Median :  0.00   Mode  :character   Median :1177.5
##  Mean   : 37.38                      Mean   :1177.5
##  3rd Qu.: 12.00                      3rd Qu.:1766.2
##  Max.   :806.00                      Max.   :2355.0
##  NA's   :2304
```

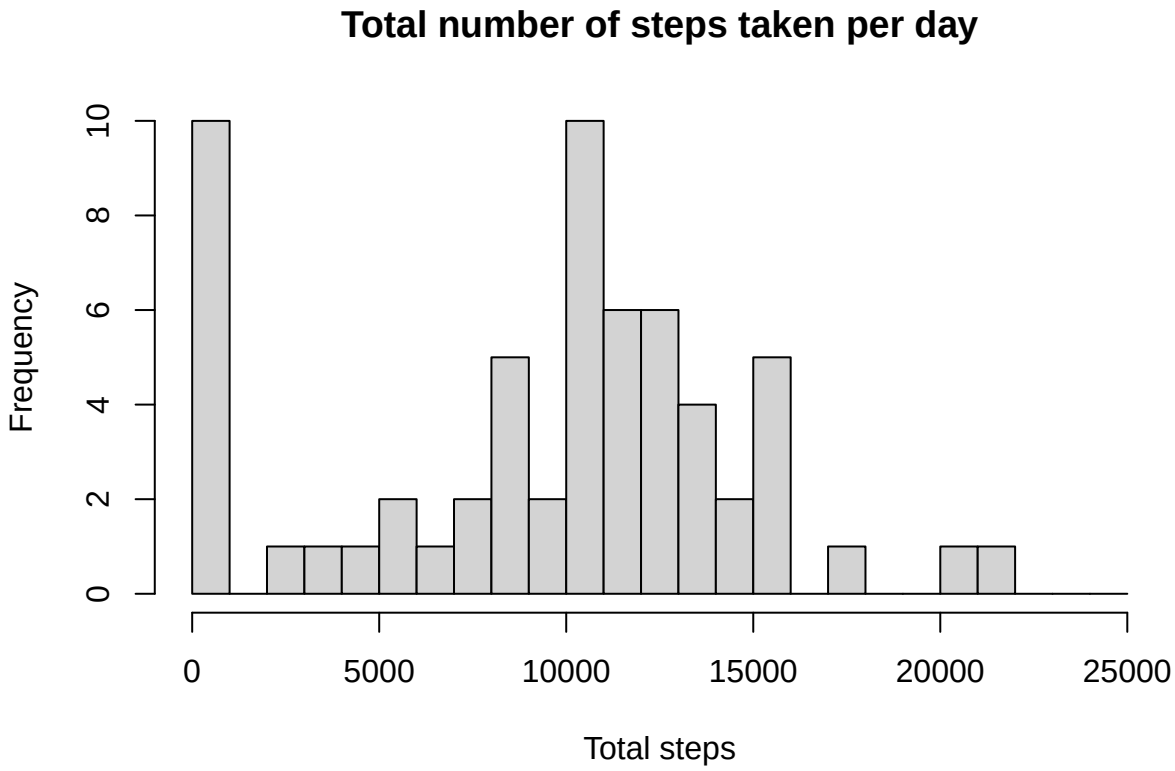
### What is mean total number of steps taken per day?

1. Calculate the total number of steps taken per day

```
totalSteps <- with(activity, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))
names(totalSteps)<- c("dates", "steps")
```

2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day

```
hist(totalSteps$steps, main = "Total number of steps taken per day", xlab = "Total steps", breaks = seq
```




3. Calculate and report the mean and median of the total number of steps taken per day

```
mean(totalSteps$steps, na.rm = TRUE)
```

```
## [1] 9354.23
```

```
median(totalSteps$steps, na.rm = TRUE)
```

```
## [1] 10395
```

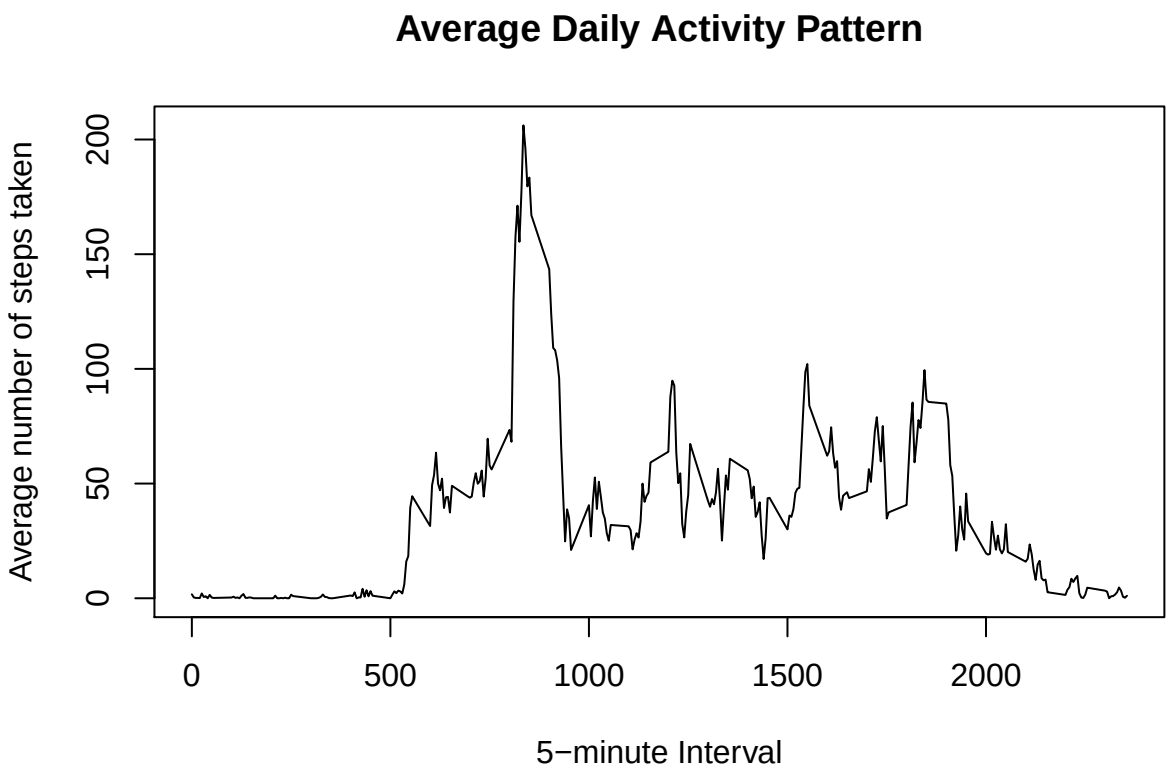
## What is the average daily activity pattern?

1. Make a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
adl<- aggregate(activity$steps, by= list(activity$interval), FUN = mean , na.rm = TRUE)

names(adl)<-c("interval", "average")

plot(adl$interval, adl$average, type = "l", main = "Average Daily Activity Pattern", xlab = "5-minute In
```




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

```
adl[which.max(adl$average),]$interval
```

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

## Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

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

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

2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
step_clean<- adl$average[match(activity$interval, adl$interval)]
activity_clean <- transform(activity, steps = ifelse(is.na(activity$steps), yes = step_clean, no = acti
```

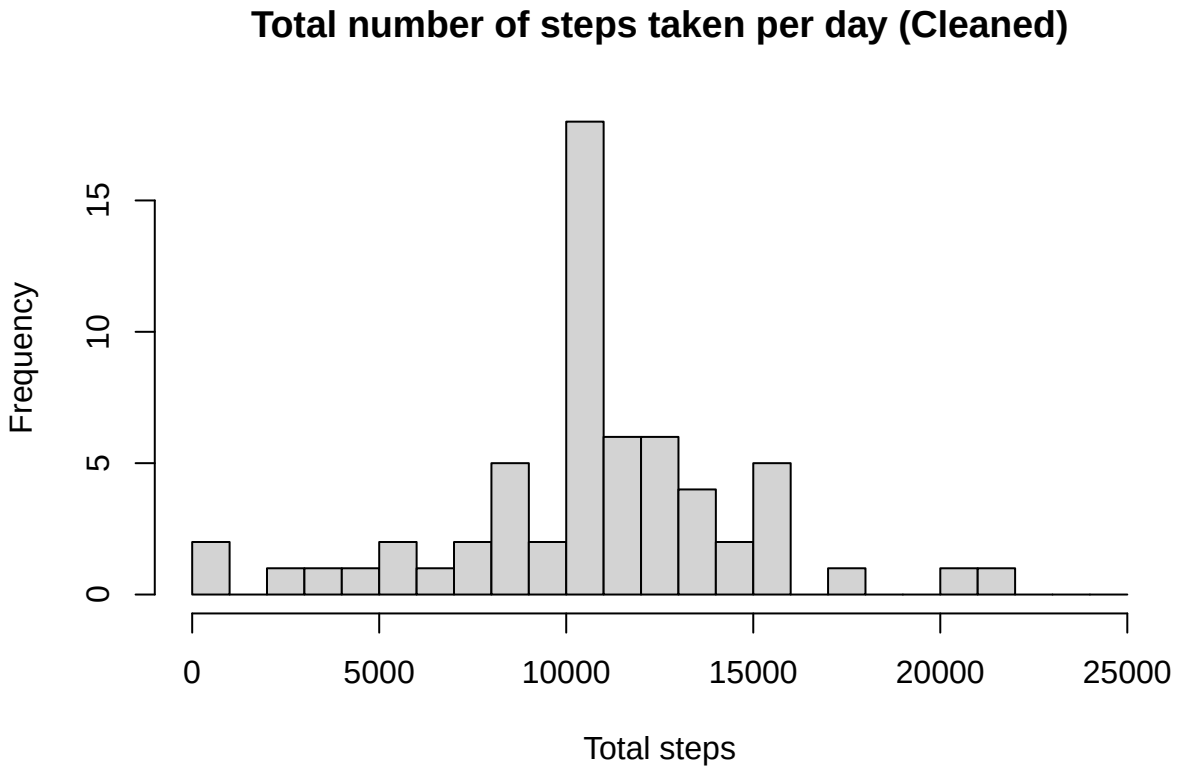
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
totalCleanSteps <- with(activity_clean, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))
names(totalCleanSteps)<- c("dates", "steps")
```

4. Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first

part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

```
hist(totalCleanSteps$steps, main = "Total number of steps taken per day (Cleaned)", xlab = "Total steps
```




```
mean(totalCleanSteps$steps)
```

```
## [1] 10766.19
```

```
median(totalCleanSteps$steps)
```

```
## [1] 10766.19
```

The new mean and median are the same at 10766.19. The original mean was 9354.23 and the original median was 10395. Hence, the process of imputing missing values improve the mean and median of the dataset.

## Are there differences in activity patterns between weekdays and weekends?

For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

1. Create a new factor variable in the dataset with two levels – "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
activity_clean$date<- as.POSIXct(activity_clean$date,  "%Y-%m-%d" )
activity_clean$weektype <- ifelse(weekdays(activity_clean$date) %in% c('Saturday', 'Sunday'), 'Weekend'
```

2. Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
activity_clean.weektype<- aggregate(steps~interval+weektype, activity_clean, mean, na.rm =TRUE)
ggplot(activity_clean.weektype, aes(x = interval, y = steps, color = weektype))+ geom_line() + labs(tit
```