# Reproducible Research: Peer Assessment 1

## Loading and preprocessing the data

I read in the data from th .csv file:

```
data <- read.csv("activity.csv", stringsAsFactors=FALSE)
```

For processing, I will use the following packages:

```
library(dplyr)
library(lubridate)
library(ggplot2)
```

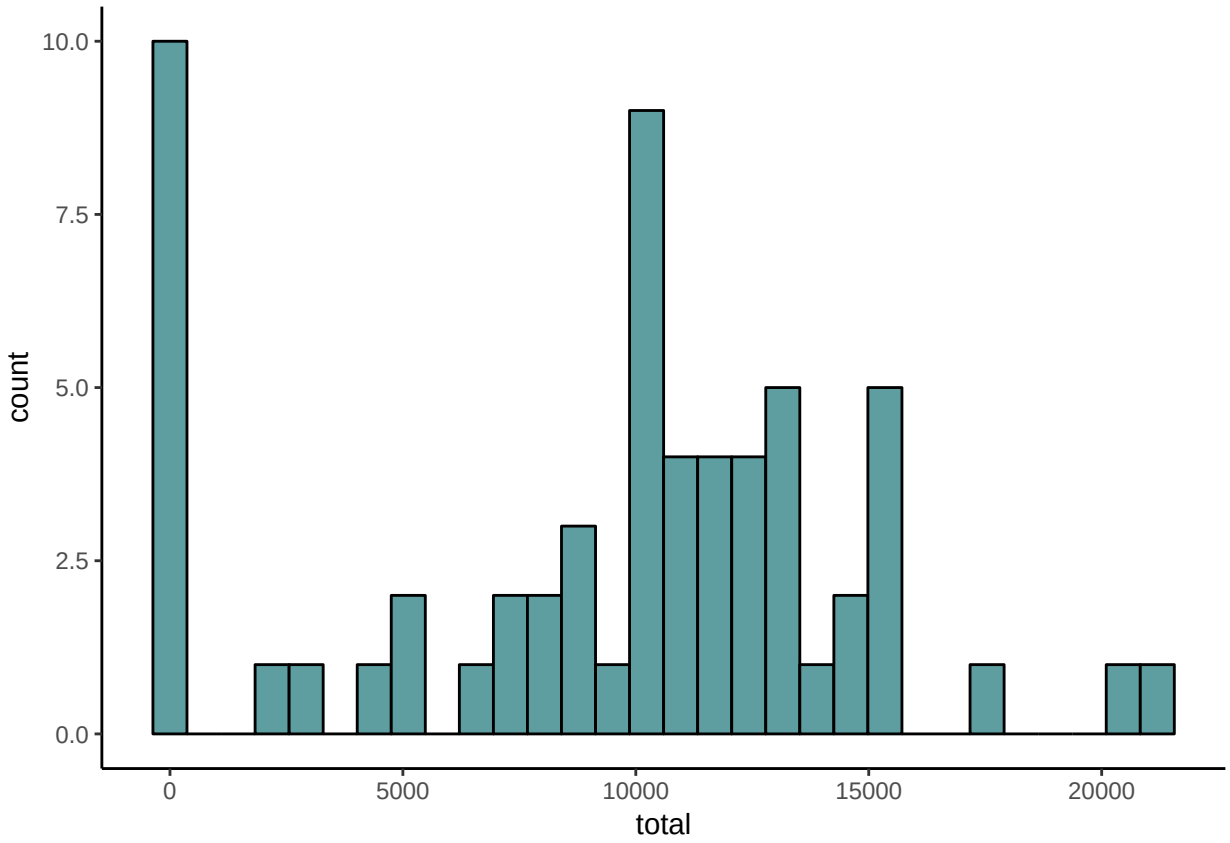
## What is mean total number of steps taken per day?

### 1. Calculate the total number of steps taken per day:

```
group_by(data, as.factor(date)) %>%
  summarize( . , total=sum(steps, na.rm=TRUE)) %>%
  rename( . , date=1) -> sums
```

### 2. Make a histogram of the total number of steps taken each day:

```
p <- ggplot(sums, aes(x=total))
p +
    theme_classic() +
    geom_histogram(
            color="black",
            fill="cadetblue",
            bins=30)
```




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

```
mean(sums$total)
```

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

```
median(sums$total)
```

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

## What is the average daily activity pattern?

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

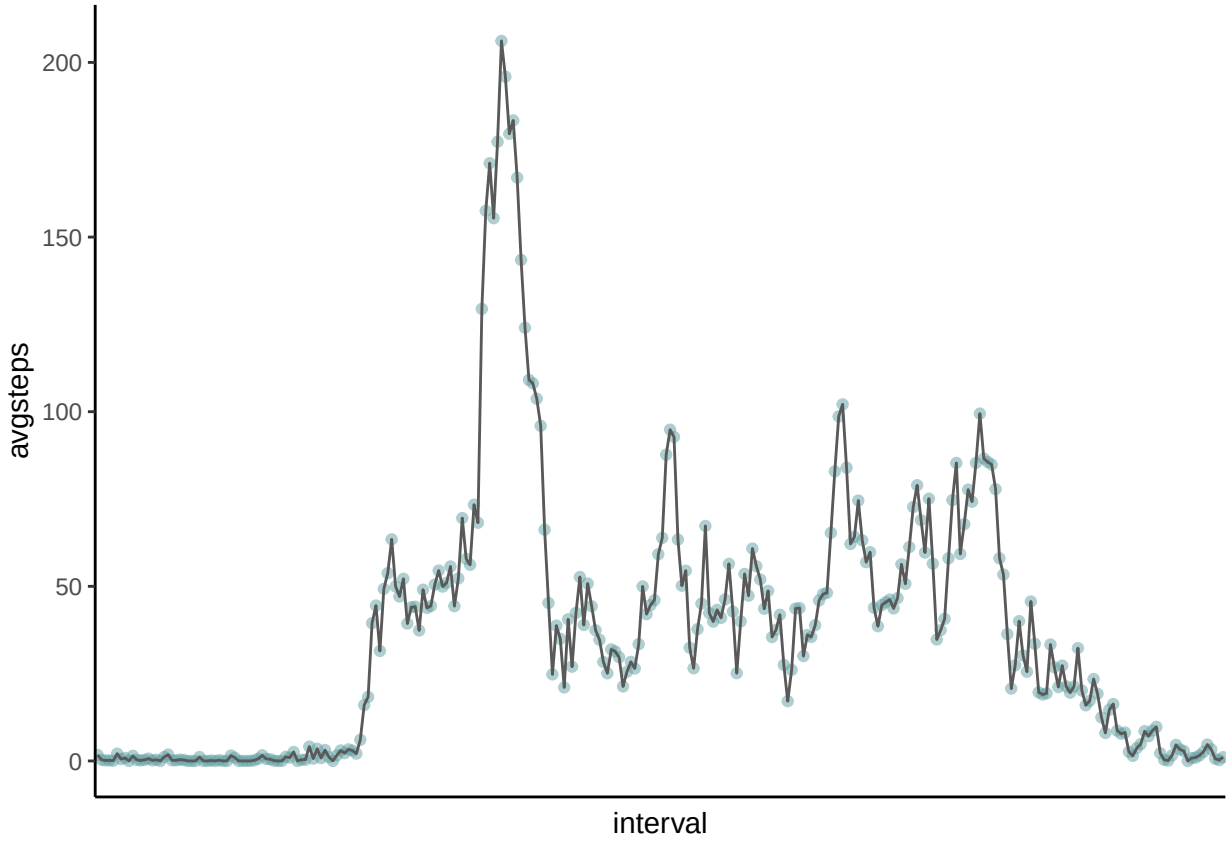
First I calculate the averages for each interval:

```
group_by(data, as.factor(interval)) %>%
  summarize( . , avgsteps=mean(steps, na.rm=TRUE)) %>%
  rename( . , interval=1) -> avgs
```

Then I plot:

```
g <- ggplot(avgs, aes(x=interval, y=avgsteps, group=1))
g +
    theme_classic() +
    theme(
            axis.text.x=element_blank(),
```

```
          axis.ticks.x=element_blank()) +
    geom_point(color="cadetblue", alpha=1/2) +
    geom_line(color="gray35")
```




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

```
avgs$interval[which(avgs$avgsteps==max(avgs$avgsteps))]
```

```
## [1] 835
## 288 Levels: 0 5 10 15 20 25 30 35 40 45 50 55 100 105 110 115 120 ... 2355
```

## Imputing missing values

**1. Calculate and report the total number of missing values in the dataset:**

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

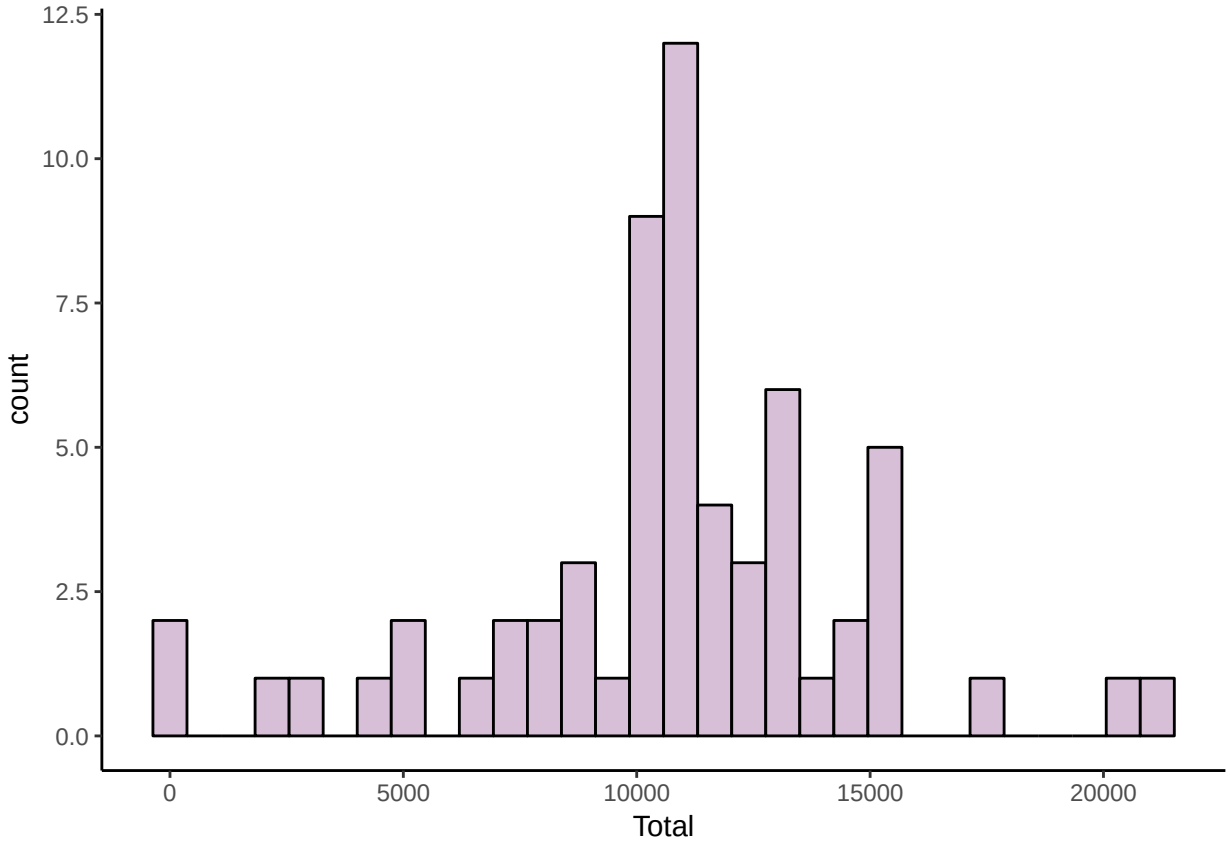
```
## [1] 2304
```

**2 & 3 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 AND Create a new dataset that is equal to the original dataset but with the missing data filled in.**

```
dataIm <- data
for (i in 1: length(dataIm$steps)){
    if (is.na(dataIm$steps[i])){
        int <- dataIm$interval[i]
        dataIm$steps[i] <- avgs$avgsteps[which(avgs$interval==int)]
    }
}

group_by(dataIm, as.factor(date)) %>%
  summarize( . , total=sum(steps, na.rm=TRUE)) %>%
  rename( . , date=1) -> sumsIm

group_by(dataIm, as.factor(interval)) %>%
  summarize( . , avgsteps=mean(steps, na.rm=TRUE)) %>%
  rename( . ,interval=1) -> avgsIm

dataIm %>%
  group_by( . , date) %>%
  summarize( . , total=sum(steps, na.rm=TRUE)) -> sumsIm
```

**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?**

```
pIm <- ggplot(sumsIm, aes(x=total))
p1 +
    theme_classic() +
    geom_histogram(
            color="black",
            fill="thistle",
            bins=30)
```




The mean and median with the imputed values are:

```
mean(sumsIm$total)
```

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

```
median(sumsIm$total)
```

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

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

**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:**

```
dataIm$date <- as.Date(dataIm$date)
days <- character(0)
for (i in 1: length(dataIm$date)){
    if (day(dataIm$date[i])==6 || day(dataIm$date[i])==7){
        days[i]<-"Weekend"
    }else{
        days[i]<-"Weekday"
    }
}

day <- factor(days, levels=c("Weekday", "Weekend"))

dataIm %>%
```

```
  mutate( . , day = day)                        %>%
  mutate( . ,finterval=as.factor(interval)) %>%
  select( . , c(1,2,5,4))                       %>%
  rename( . ,interval=finterval)                %>%
  group_by( . ,day,interval)                    -> dataIm

dataIm %>%
  summarize( . ,avgsteps=mean(steps)) -> avgsIm
```

**2. Make a panel plot containing a time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days. See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.**

```
gIm <- ggplot(avgs1,aes(x=interval,y=avgsteps, group=1))

gIm +
    theme_classic() +
    theme(
            axis.text.x=element_blank(),
            axis.ticks.x=element_blank()
        ) +
    geom_point(color="aquamarine4", alpha=1/2) +
    geom_line(color="gray35") +
    facet_grid(.~day)
```