Course Project 1

Kieran Higgins

21/04/2020

Loading and preprocessing the data

```
activityData <- read.csv("activity.csv") #load the data
activityData$date <- as.Date(activityData$date, format = "%Y-%m-%d") #formats as date
```

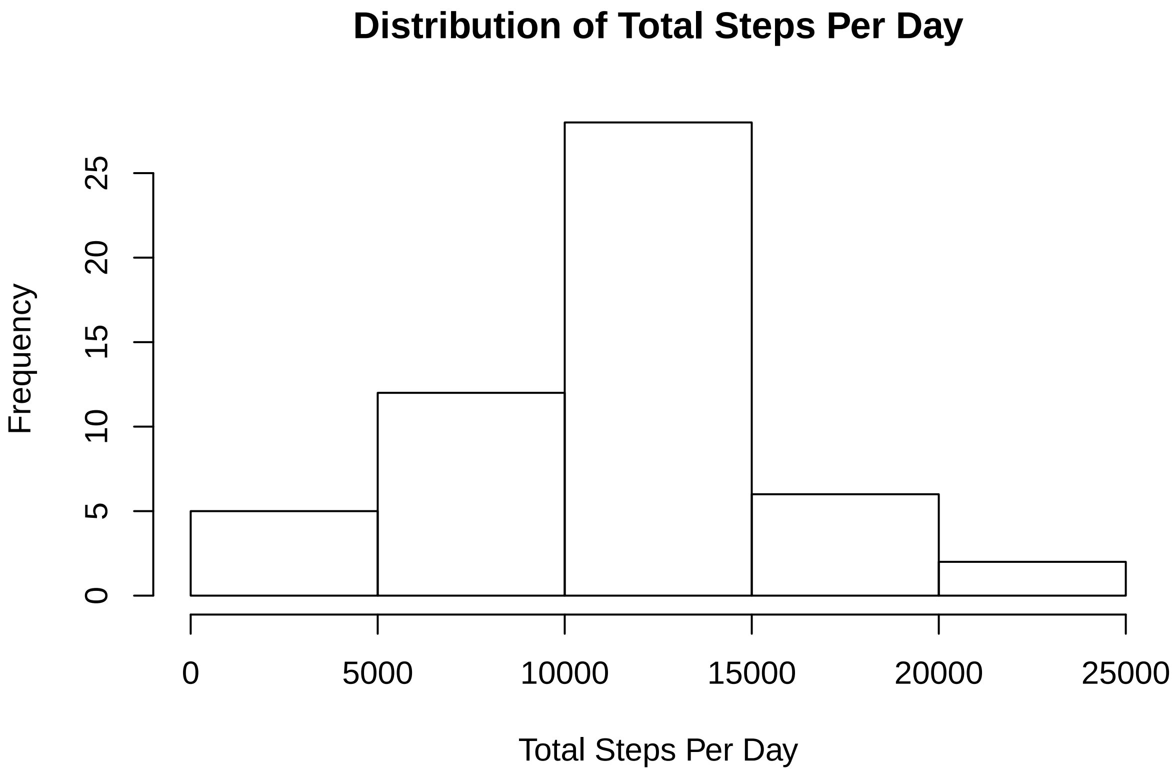
What is mean total number of steps taken per day?

Calculate total number of steps per day:

```
TotalStepsPerDay <- aggregate(steps ~ date, activityData, FUN = "sum")
```

Plot a histogram of the total number of steps taken each day

```
hist(TotalStepsPerDay$steps, xlab = "Total Steps Per Day",
     main = "Distribution of Total Steps Per Day")
```



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

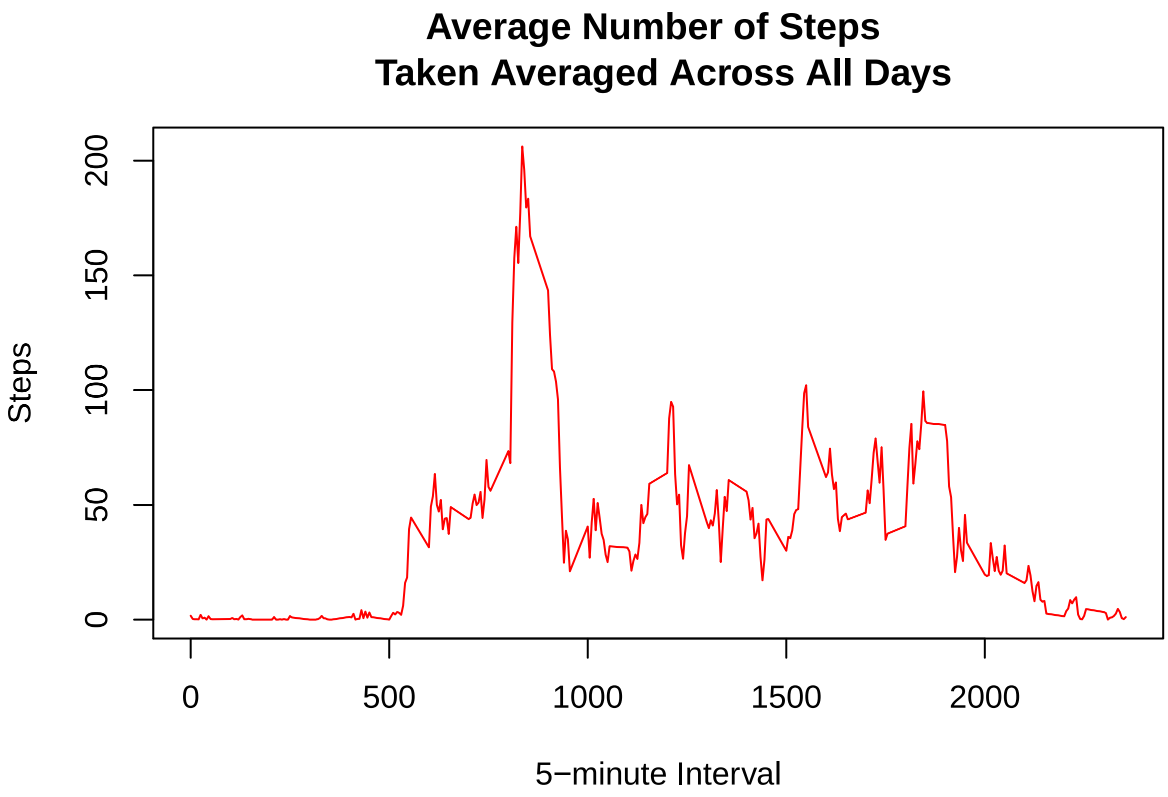
```r
summary(TotalStepsPerDay$steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      41    8841   10765   10766   13294   21194
```

What is the average daily activity pattern?

Make a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```r
MeanStepsPerInterval <- aggregate(steps ~ interval, activityData, FUN = "mean")
plot(MeanStepsPerInterval, type = "l", xlab = "5-minute Interval", ylab = "Steps",
     main = "Average Number of Steps \n Taken Averaged Across All Days", col = "red")
```



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

```
print(MeanStepsPerInterval$interval[which.max(MeanStepsPerInterval$steps)])
```

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

Imputting missing values

Calculate and report the total number of missing values in the dataset

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

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

Devise a strategy for filling in all of the missing values in the dataset and create a new dataset that is equal to the original dataset but with the missing data filled in.

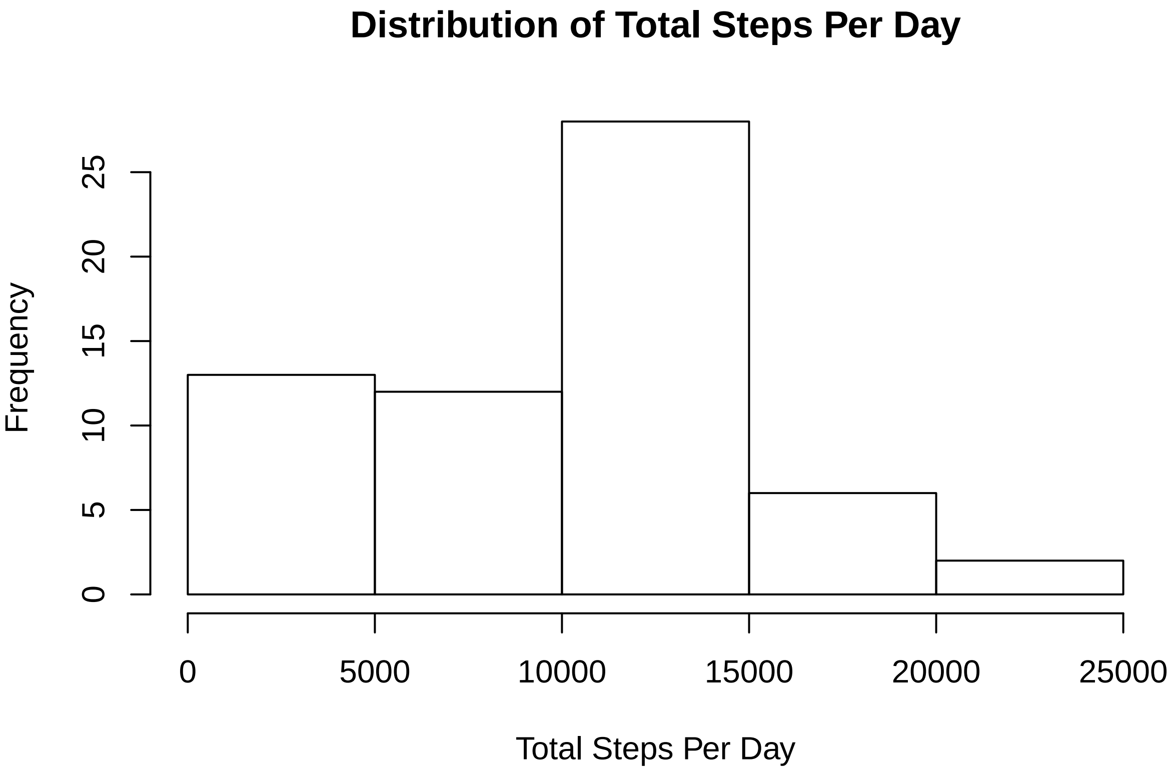
```
activityDataImputed <- activityData #make new dataset

#replace missing values with 0
activityDataImputed$steps[is.na(activityDataImputed$steps)] <- 0
sum(is.na(activityDataImputed$steps)) #check no more NAs
```

```
## [1] 0
```

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

```
TotalStepsPerDayImputed <- aggregate(steps ~ date, activityDataImputed, FUN = "sum")
hist(TotalStepsPerDayImputed$steps, xlab = "Total Steps Per Day",
     main = "Distribution of Total Steps Per Day")
```



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

```
summary(TotalStepsPerDayImputed$steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0    6778   10395    9354   12811   21194
```

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?

> Yes, this increases the number of values in the bin 0-5000. Both the median and the mean lower, from 10765 to 10395 and from 10766 to 9354 respectively.

Are there differences in activity patterns between weekdays and weekends?

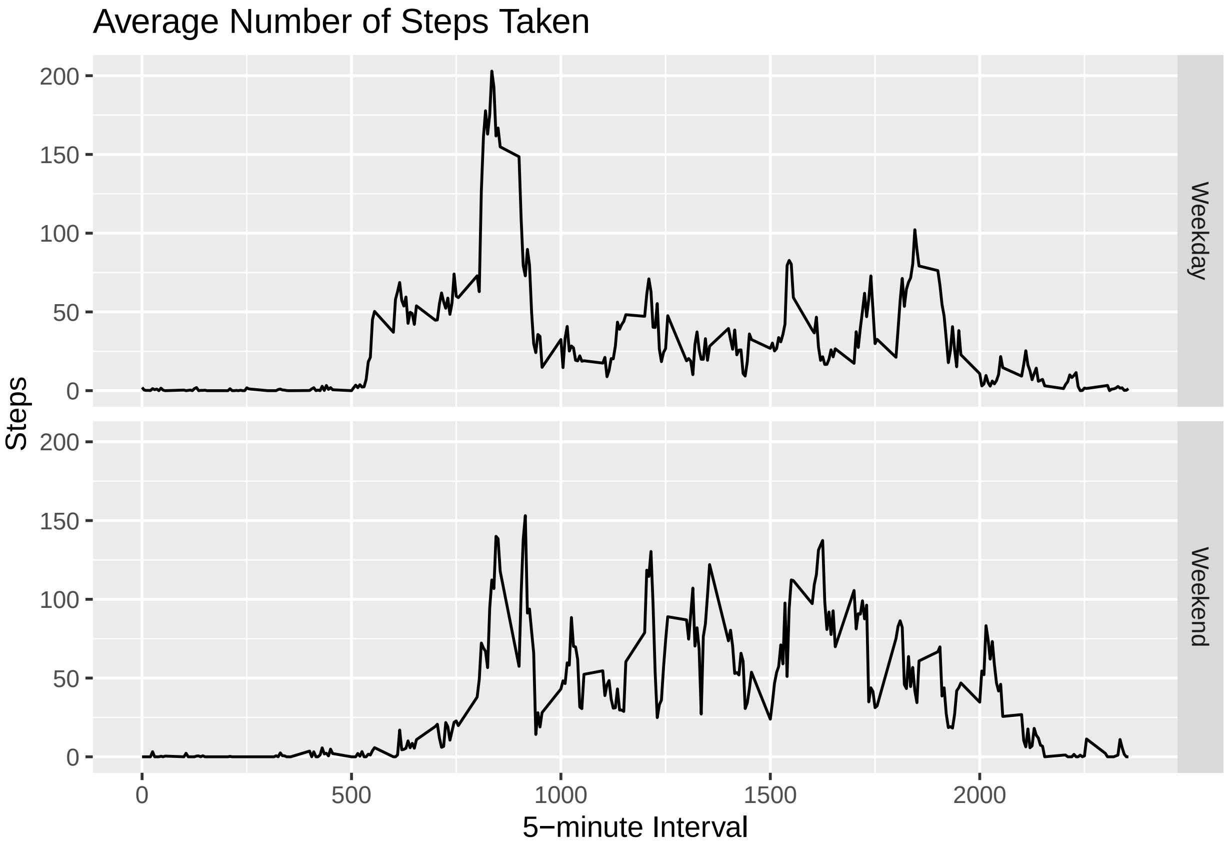
Create a new factor variable in the dataset with two levels – "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
activityDataImputed$weekday <- weekdays(activityDataImputed$date)
activityDataImputed$weekday[
  activityDataImputed$weekday %in% c('Saturday','Sunday')] <- "Weekend"
```

```
activityDataImputed$weekday[activityDataImputed$weekday !="Weekend"] <- "Weekday"
activityDataImputed$weekday <- as.factor(activityDataImputed$weekday)
```

Make a panel plot containing a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```
library(ggplot2)
byDayType <- aggregate(steps ~ interval + weekday, activityDataImputed, FUN = "mean")
qplot(interval, steps, facets = weekday~., geom= "line", data = byDayType,
      xlab = "5-minute Interval", ylab = "Steps", main = "Average Number of Steps Taken")
```


Average Number of Steps Taken
Steps
Weekday
Weekend
5−minute Interval