Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

The data given is in a zip file. The data is in the file activity.csv.

- It is read and stored in a dataframe using the read.csv() function.
- The data is then aggregated by date.

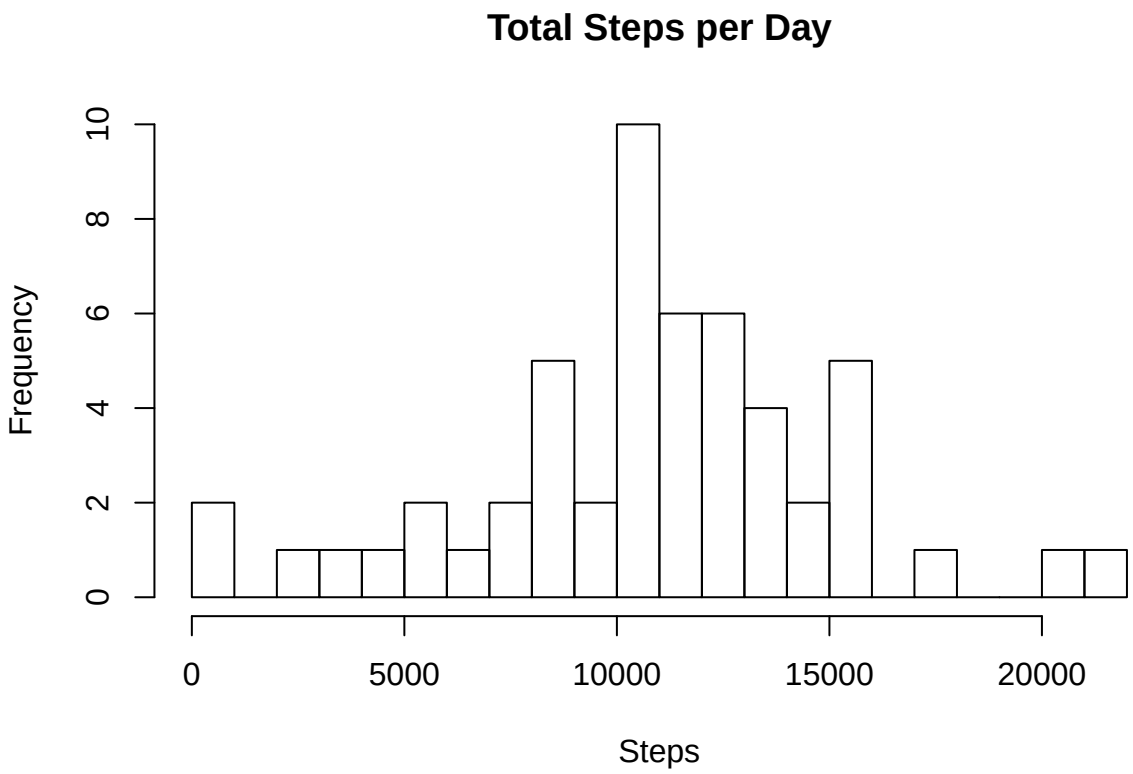
```
## Goal 1: Code for reading in the dataset and/or processing the data

if (file.exists("activity.csv")){
  originaldata <- read.csv("activity.csv")
}
aggregateddata <- aggregate(steps ~ date, data=originaldata, sum, na.rm = TRUE)
```

What is mean total number of steps taken per day?

Ignoring the missing values in the dataset.

- Calculating the total number of steps taken per day
- A histogram is produced for the total number of steps per day.

```
## Goal 2: Histogram of the total number of steps taken each day
hist(aggregateddata$steps, breaks=20, main="Total Steps per Day", xlab="Steps", ylab="Frequency")
```



* Calculating the mean and median of the number of steps

```
## Goal 3: Mean and median number of steps taken each day
originaldata_mean <- mean(originaldata$steps, na.rm=TRUE)
originaldata_median <- median(originaldata$steps, na.rm=TRUE)
print(paste("The mean steps per day is: ", originaldata_mean))
```

```
## [1] "The mean steps per day is:  37.3825995807128"
```

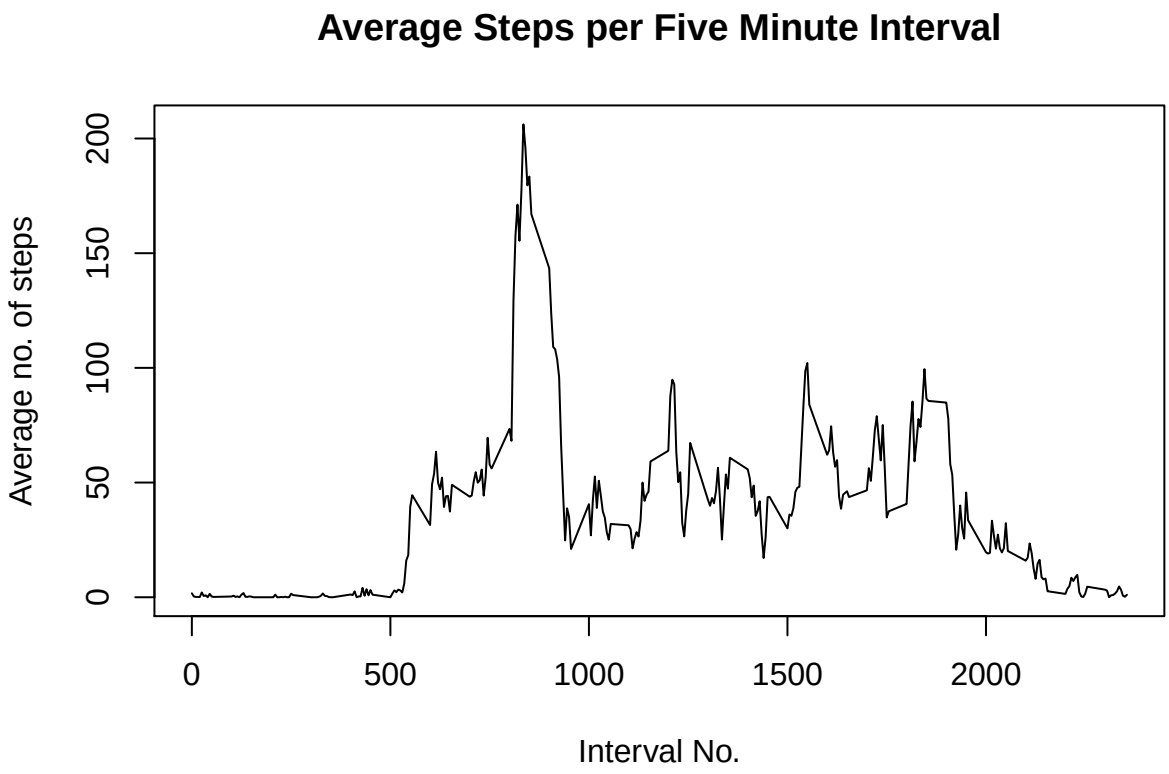
```
print(paste("The median steps per day is: ", originaldata_median))
```

```
## [1] "The median steps per day is:  0"
```

What is the average daily activity pattern?

- 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) is plotted

```
## Goal 4: Time series plot of the average number of steps taken
aggregateddatabysteps <- aggregate(steps ~ interval,
                                   data=originaldata, mean, na.rm=TRUE)
plot(aggregateddatabysteps$interval, aggregateddatabysteps$steps, type="l",
     main="Average Steps per Five Minute Interval",
     xlab="Interval No.", ylab="Average no. of steps")
```



* The 5-minute interval, on average across all the days in the dataset, containing the maximum number of steps is claculated

```
## Goal 5: The 5-minute interval that, on average, contains the maximum number of steps
maxstepsforinterval <- max(aggregateddatabysteps$steps)
print(paste("The maximum number of steps in a five minute interval was: ",
            maxstepsforinterval))
```

```
## [1] "The maximum number of steps in a five minute interval was:  206.169811320755"
```

Imputing missing values

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.

- The total number of missing values in the dataset calculated and reported (i.e. the total number of rows with NAs)

```
## Goal 6: Code to describe and show a strategy for imputing missing data
missingdata <- sum(is.na(originaldata$steps))
print(paste("There are", missingdata, "missing values in the steps variable."))
```

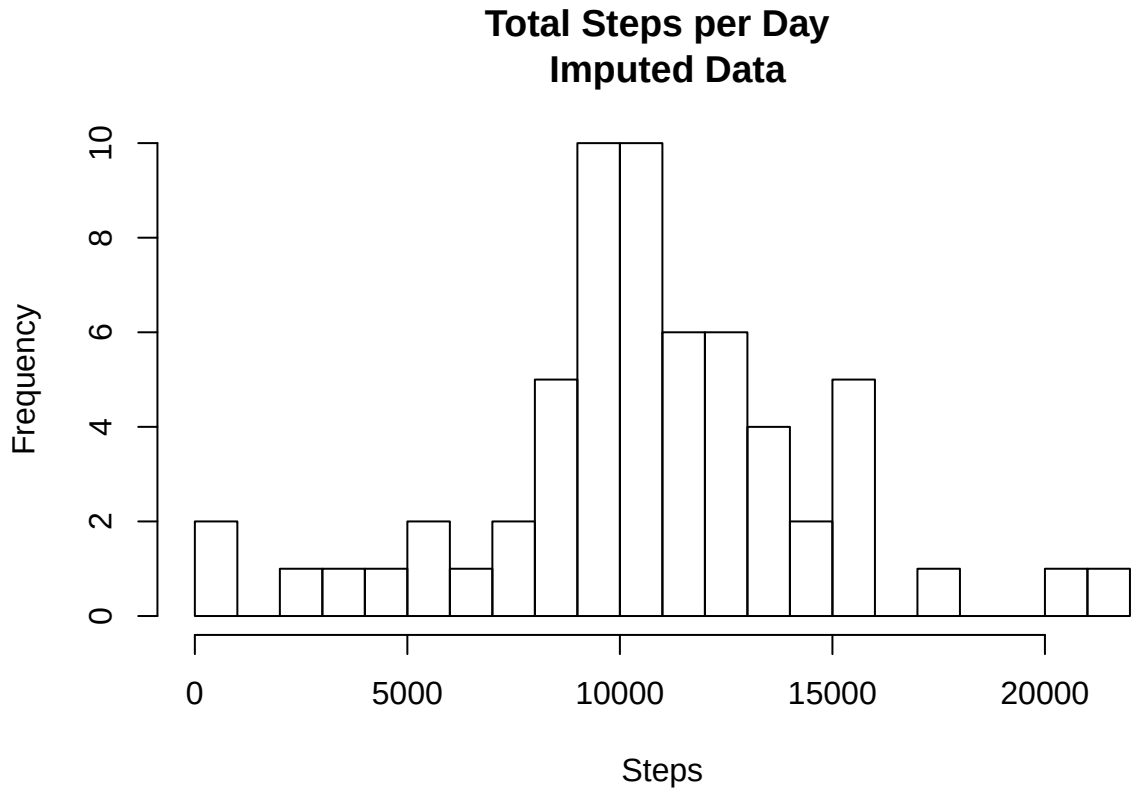
```
## [1] "There are 2304 missing values in the steps variable."
```

- A strategy is deviced for filling in all of the missing values in the dataset. The mean of the steps for the 5-minute interval is used to fill in the na values.
- A new dataset is created that is equal to the original dataset but with the missing data filled in.

```
## Replacing the NA values with the mean of the steps
imputeddata <- originaldata
imputeddata$steps[is.na(imputeddata$steps)] <- median(aggregateddatabysteps$steps, na.rm=TRUE)
imputeddataday <- aggregate(steps ~ date, data=imputeddata, sum, na.rm=TRUE)
```

- 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?

```
## Goal 7: Histogram of the total number of steps taken each day after missing values are imputed
hist(imputeddataday$steps, breaks=20, main="Total Steps per Day \n Imputed Data",
     xlab="Steps", ylab="Frequency")
```



* The mean and median total number of steps taken per day for the imputed dataset is calculated and reported

```
## Goal 8: Calculating the mean and median of the new dataset
imputeddata_mean <- mean(imputeddata$steps)
imputeddata_median <- median(imputeddata$steps)
print(paste("Mean for imputed data is:", imputeddata_mean,
            "whereas original mean was:", originaldata_mean))
```

```
## [1] "Mean for imputed data is: 36.9538268550022 whereas original mean was: 37.3825995807128"
```

```
print(paste("Median for imputed data is:", imputeddata_median,
            "whereas original median was:", originaldata_median))
```

```
## [1] "Median for imputed data is: 0 whereas original median was: 0"
```

- This clear shows how these values differ from the original data values. The impact of imputing missing data on the estimates of the total daily number of steps shows that there are higher frquency counts in the histogram at the center region but it was missing in the original data.

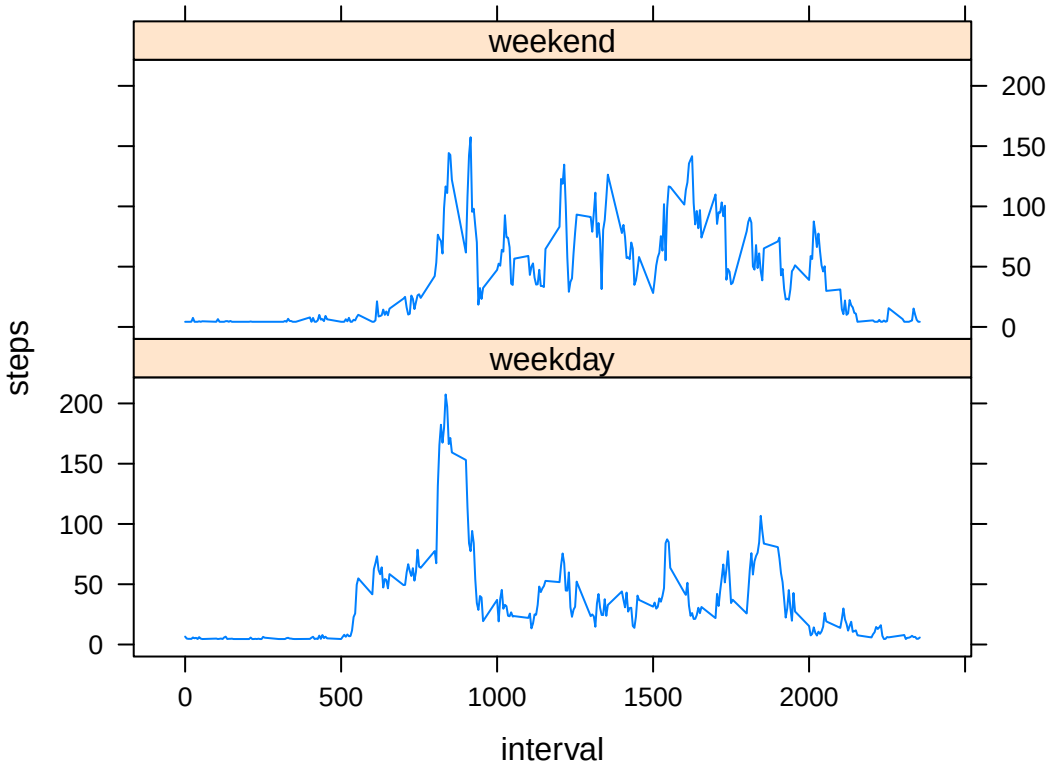
Are there differences in activity patterns between weekdays and weekends?

- A new factor variable is created in the dataset with two levels - "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.
- For this the date variable is converted to date format first
- The weekday() function is used to calculate the day of the week for each date and they are stored in the dayname vafriable newly created
- If the value is either "saturday" or sunday then weekend is assigned to the weekend variable that is newly created.

```
##  Goal 9:Panel plot comparing the average number of steps taken per 5-minute
## interval across weekdays and weekends
imputeddata$date <- as.Date(imputeddata$date)
imputeddata$dayname <- weekdays(imputeddata$date)
imputeddata$weekend <- as.factor(ifelse(imputeddata$dayname == "Saturday" |
                                        imputeddata$dayname == "Sunday", "weekend", "weekday"))
library(lattice)
```

- 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) is made to show the differences in activity patterns between weekdays and weekends.

```
plotdata <- aggregate(steps ~ interval + weekend, imputeddata, mean)
xyplot(steps ~ interval | factor(weekend), data=plotdata, aspect=1/3, type="l")
```


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