# Report of an Activity Monitoring Data Analysis

## 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 data for this assignment can be downloaded from the course web site. 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

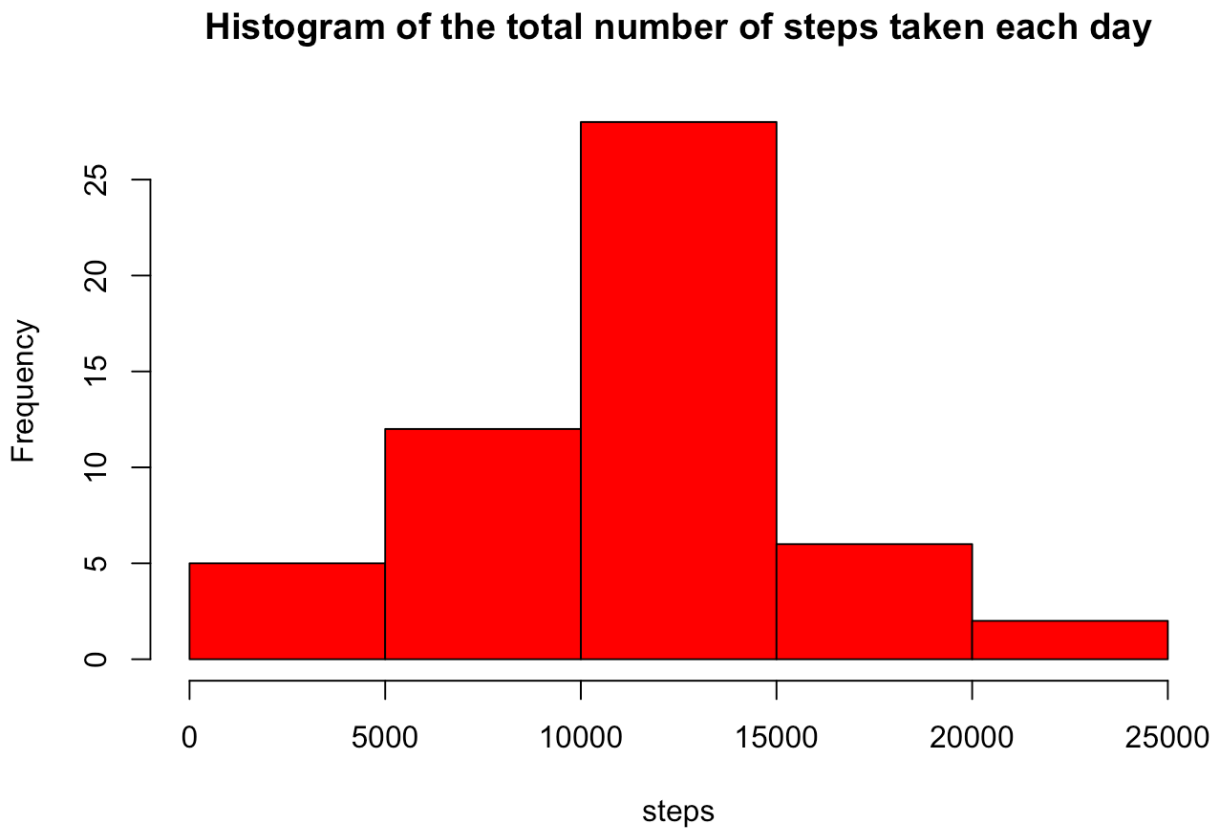
The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

The code for reading in the dataset from the CSV file is as follows.

```
setwd("~/Desktop/Data Science Courses/Reproducible Research")
activity <- read.csv("activity.csv")
```

First we will make a histogram for the total number of steps taken each day. The R code and the output plot for the histogram are as shown below.

```
total_steps <- aggregate(steps ~ date, activity, sum)
with(total_steps,hist(steps, col="red", main = "Histogram of the total number o
f steps taken each day"))
```




Next we will calculate the mean and median number of steps taken each day (ignoring missing values in the data set) as follows.

```
mean(total_steps$steps)
```

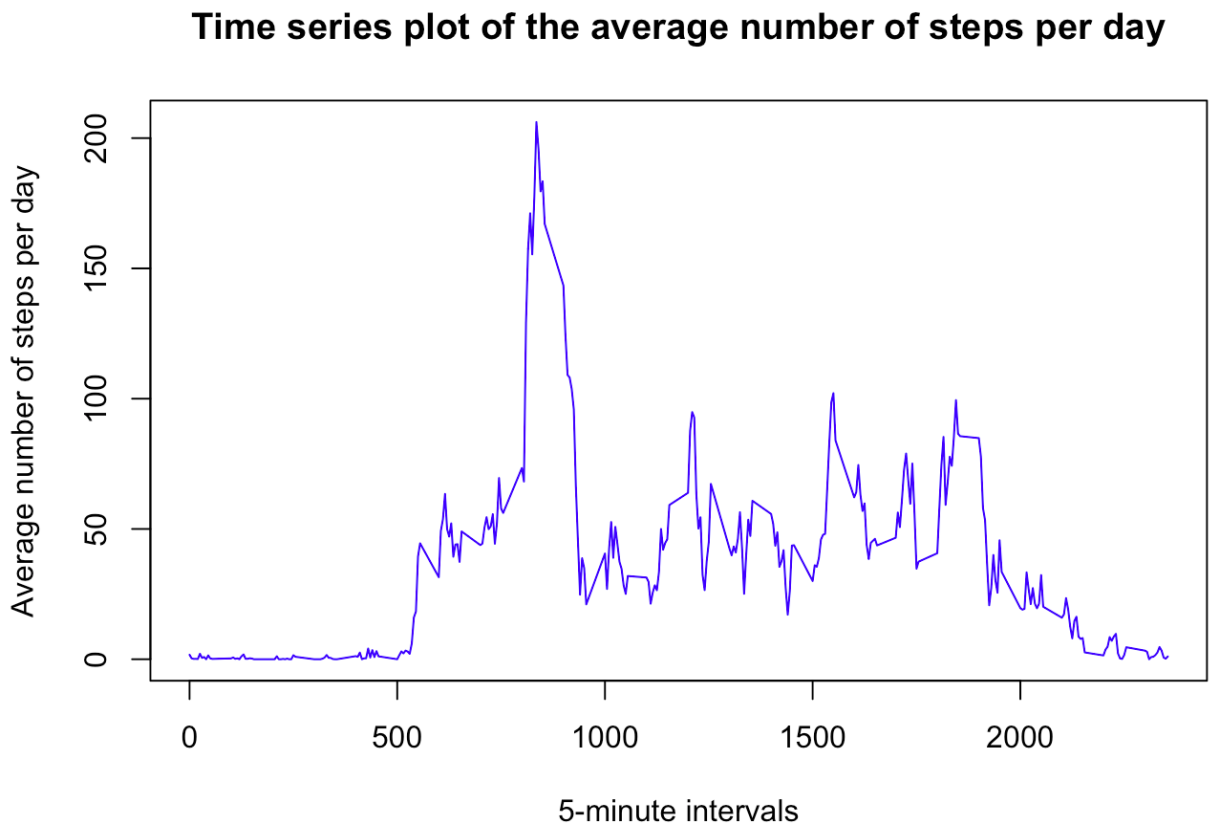
```
## [1] 10766.19
```

```
median(total_steps$steps)
```

```
## [1] 10765
```

After that we will make a time series plot for the average number of steps taken. We will plot the 5-minute intervals on the x-axis and the average number of steps taken, averaged across all days on the y-axis.

```
daily_average <- aggregate(steps ~ interval, activity, mean)
with(daily_average, plot(interval, steps, col = "blue", type = "l", xlab = "5-m
inute intervals",ylab ="Average number of steps per day", main = "Time series p
lot of the average number of steps per day"))
```




Let us now calculate the 5-minute interval that, on average, contains the maximum number of steps.

```
daily_average$interval[which(daily_average$steps == max(daily_average$steps))]
```

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

So the 835th 5-minute interval contains the maximum number of steps. Now we are ready to deal with the missing values in the data set. Let us first see how many rows of the data set have missing values.

```
table(is.na(activity))
```

```
##
## FALSE  TRUE
## 50400  2304
```

So there are 2304 rows with missing values. To keep things simple, we will impute the missing values for the number of steps by the average number of steps for the entire data set. First we have to load the 'Hmisc' library which contains the impute function.

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
activity$imputed_steps <- with(activity, impute(steps, mean))
```

Let us quickly check how the mean and median for the average number of steps per day has changed after imputing the missing values in the data set.

```
total_steps_imputed <- aggregate(imputed_steps ~ date, activity, sum)
mean(total_steps_imputed$imputed_steps)
```

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

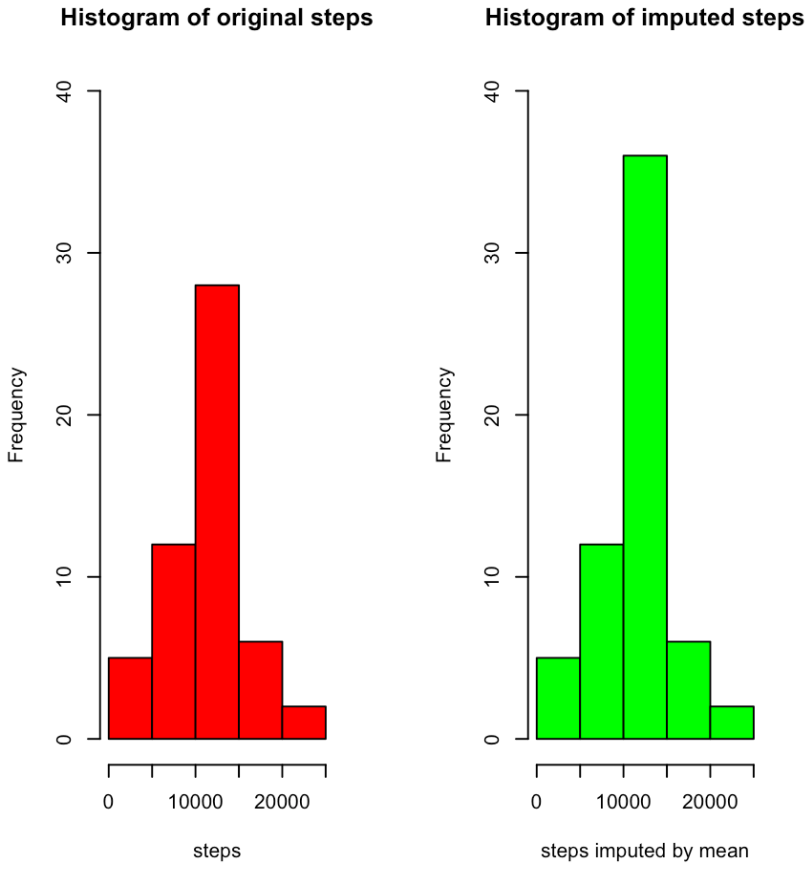
```
median(total_steps_imputed$imputed_steps)
```

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

Note that the new mean is the same as the old mean of 10766.19. However the new median has changed from 10765 to 10766.19. Next we will plot a histogram for the total number of steps taken each day using the imputed data set. We will make a panel plot showing the total number of steps taken each day for both the old data with the missing values and the new data with the imputed values

to see how they compare.

```
par(mfrow = c(1,3), mar = c(5,5,4,3))
with(total_steps,hist(steps, col="red", ylim = c(0,40),main = "Histogram of ori
ginal steps"))
with(total_steps_imputed,hist(imputed_steps, col="green", ylim = c(0,40), main
= "Histogram of imputed steps", xlab = "steps imputed by mean"))
```

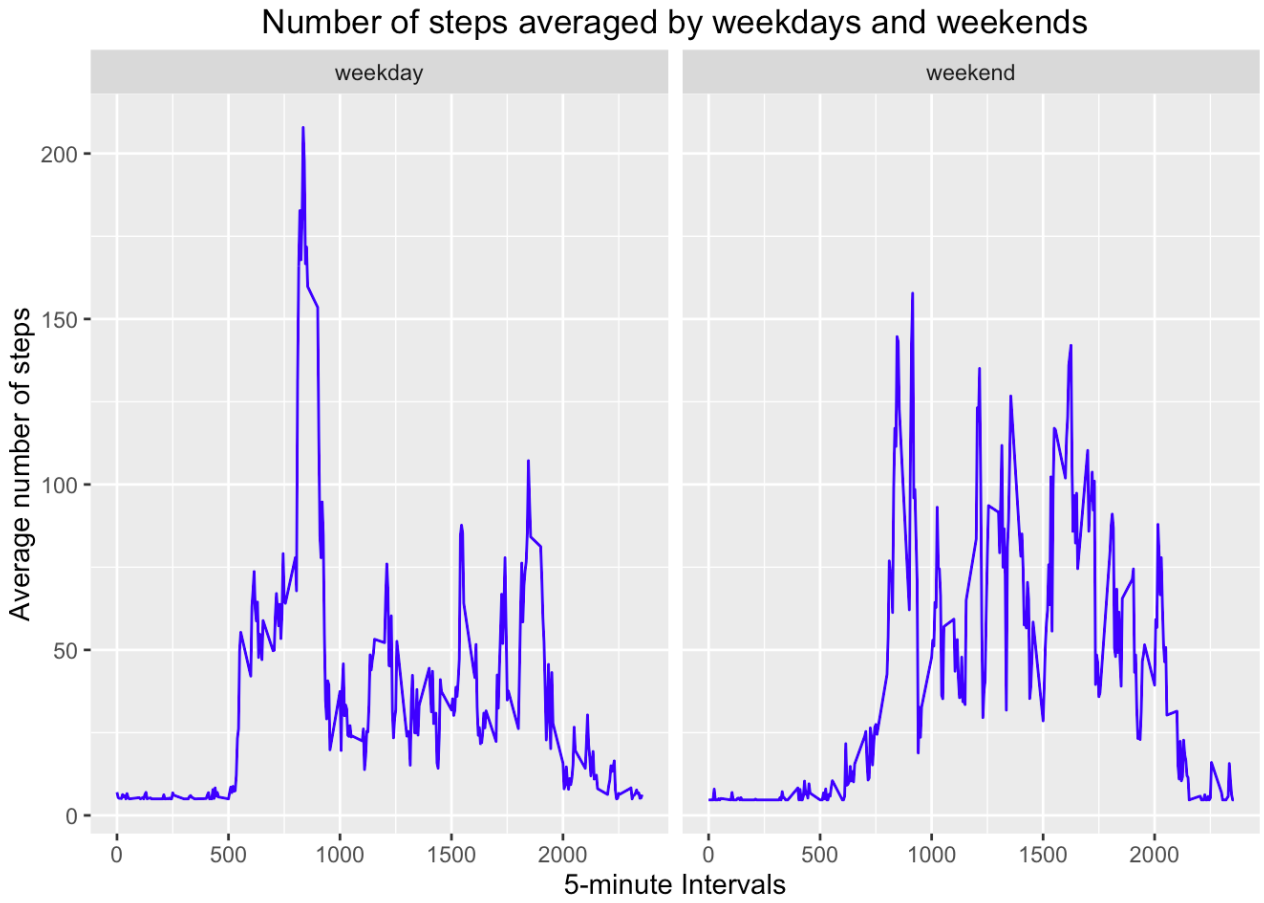



Finally we will make a panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends. To do so, we will first create a new factor variable in the activity dataset called ‘day’ with two levels – “weekday” and “weekend” indicating whether a given date is a weekday or weekend day.

```
activity$date <-as.Date(activity$date)
activity$day <- as.character(weekdays(activity$date))
activity$day[activity$day == "Saturday"] <- "weekend"
activity$day[activity$day == "Sunday"] <- "weekend"
activity$day[activity$day == "Monday"] <- "weekday"
activity$day[activity$day == "Tuesday"] <- "weekday"
activity$day[activity$day == "Wednesday"] <- "weekday"
activity$day[activity$day == "Thursday"] <- "weekday"
activity$day[activity$day == "Friday"] <- "weekday"
```

Next we will 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)
steps_by_day_of_week <- aggregate(imputed_steps ~ day + interval, activity, mea
n)
g <- ggplot(steps_by_day_of_week, aes(x = interval, y = imputed_steps, fill = d
ay))
g + geom_line(stat = "identity", col="blue") + facet_grid(.~ day) + xlab("5-min
ute Intervals") + ylab("Average number of steps") + ggtitle("Number of steps av
eraged by weekdays and weekends")
```




The plots seem to suggest that on an average, people tend to move around more during the early part of the day on weekdays after which the movement tends to fall off as the day progresses. There is a little spike in movement at the end of the day when people are probably getting ready to leave work and go home. However on weekends people tend to move around more throughout the day as compared to weekdays which is probably a result of their indulging in a variety of weekend activities throughout the day.