Reproducible Research: “Peer Assessment 1”

Loading and preprocessing the data

Setting global options to turn warnings off

```
knitr::opts_chunk$set(warning=FALSE)
```

Adding data and loading ggplot2

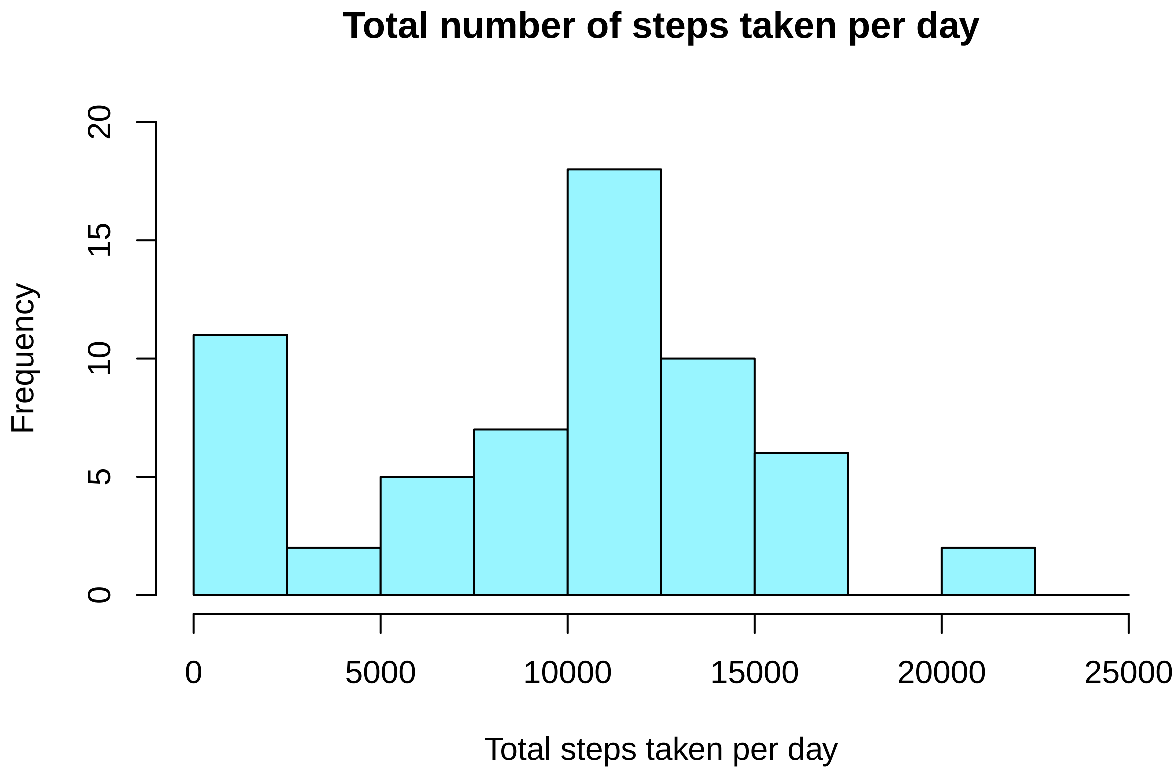
```
library(ggplot2)
activity <-read.csv("activity.csv")
activity$date <-as.POSIXct(activity$date,"%Y-%m-%d")
weekday <- weekdays(activity$date)
activity <- cbind(activity,weekday)

##Verify dataset is preprocessed correctly
summary(activity)
```

```
##      steps             date                 interval        weekday
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0   Length:17568
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8   Class :character
##  Median :  0.00   Median :2012-10-31   Median :1177.5   Mode  :character
##  Mean   : 37.38   Mean   :2012-10-31   Mean   :1177.5
##  3rd Qu.: 12.00   3rd Qu.:2012-11-15   3rd Qu.:1766.2
##  Max.   :806.00   Max.   :2012-11-30   Max.   :2355.0
##  NA's   :2304
```

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

```
activity_total_steps <- with(activity, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))
names(activity_total_steps) <- c("date", "steps")
hist(activity_total_steps$steps, main = "Total number of steps taken per day", xlab = "Total steps taken
```



```
meansteps <- mean(activity_total_steps$steps)
```

Mean total number of steps taken per day is 9354.2295082.

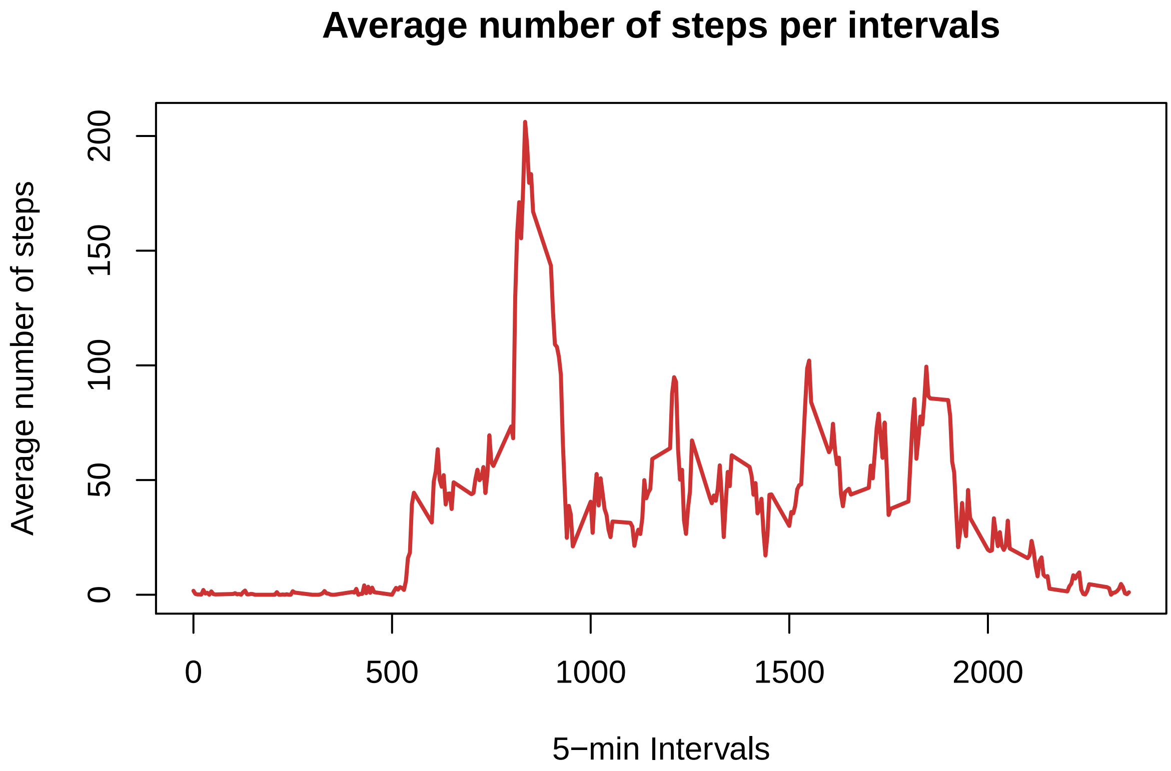
```
mediansteps <- median(activity_total_steps$steps)
```

Median total number of steps taken per day is 10395.

What is the average daily activity pattern?

Time Series plot of the 5-minute interval and average number of steps taken across all days

```
average_daily_activity <- aggregate(activity$steps, by=list(activity$interval), FUN=mean, na.rm=TRUE)
names(average_daily_activity) <- c("interval", "mean")
plot(average_daily_activity$interval, average_daily_activity$mean, type = "l", col="brown3", lwd = 2, xl
```



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

```
maxdaily <-average_daily_activity[which.max(average_daily_activity$mean), ]$interval
```

The 5-min interval that contains the maximum number of steps across all the days in the data set is 835.

Imputing missing values

Total number of missing values in the data set (i.e. the total number of rows with NAs)

```
totalna <-sum(is.na(activity$steps))
```

There are 2304 missing values in the data set.

Fill in all the missing values in the dataset.

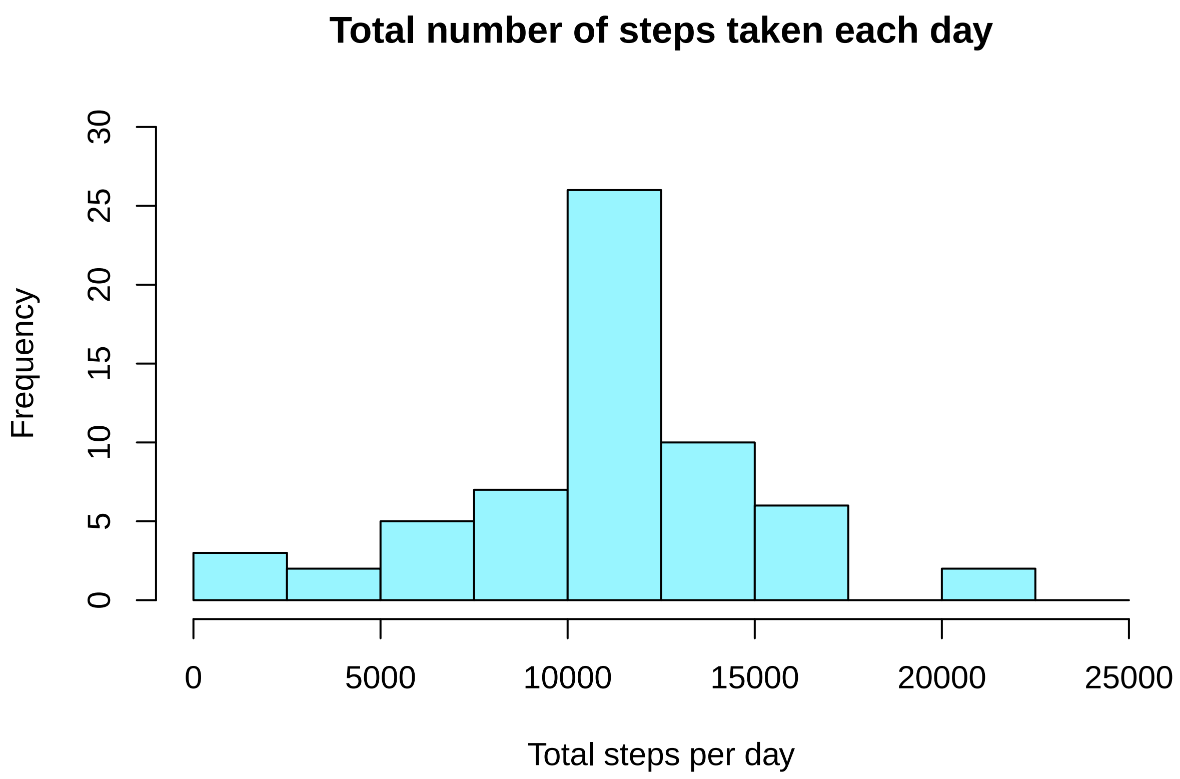
```
imputed_steps <- average_daily_activity$mean[match(activity$interval, average_daily_activity$interval)]
```

Create a new dataset with the missing data filled in

```
activity_imputed <- transform(activity, steps = ifelse(is.na(activity$steps), yes = imputed_steps, no =
total_steps_imputed <- aggregate(steps ~ date, activity_imputed, sum)
names(total_steps_imputed) <- c("date", "daily_steps")
```

Histogram of the total number of steps taken each day

```
hist(total_steps_imputed$daily_steps, col = "cadetblue1", xlab = "Total steps per day", ylim = c(0,30),
```



```
impmeansteps <- mean(total_steps_imputed$daily_steps)
```

Mean total number of steps taken per day is 1.0766189×10^4.

```
impmediansteps <- median(total_steps_imputed$daily_steps)
```

Median total number of steps taken per day is 1.0766189×10^4.

We can see these values are *greater* than the estimates from the first part of the assignments, thus we can conclude that imputing missing data on the estimate of the total daily number of steps increases the mean and median number of steps taken.

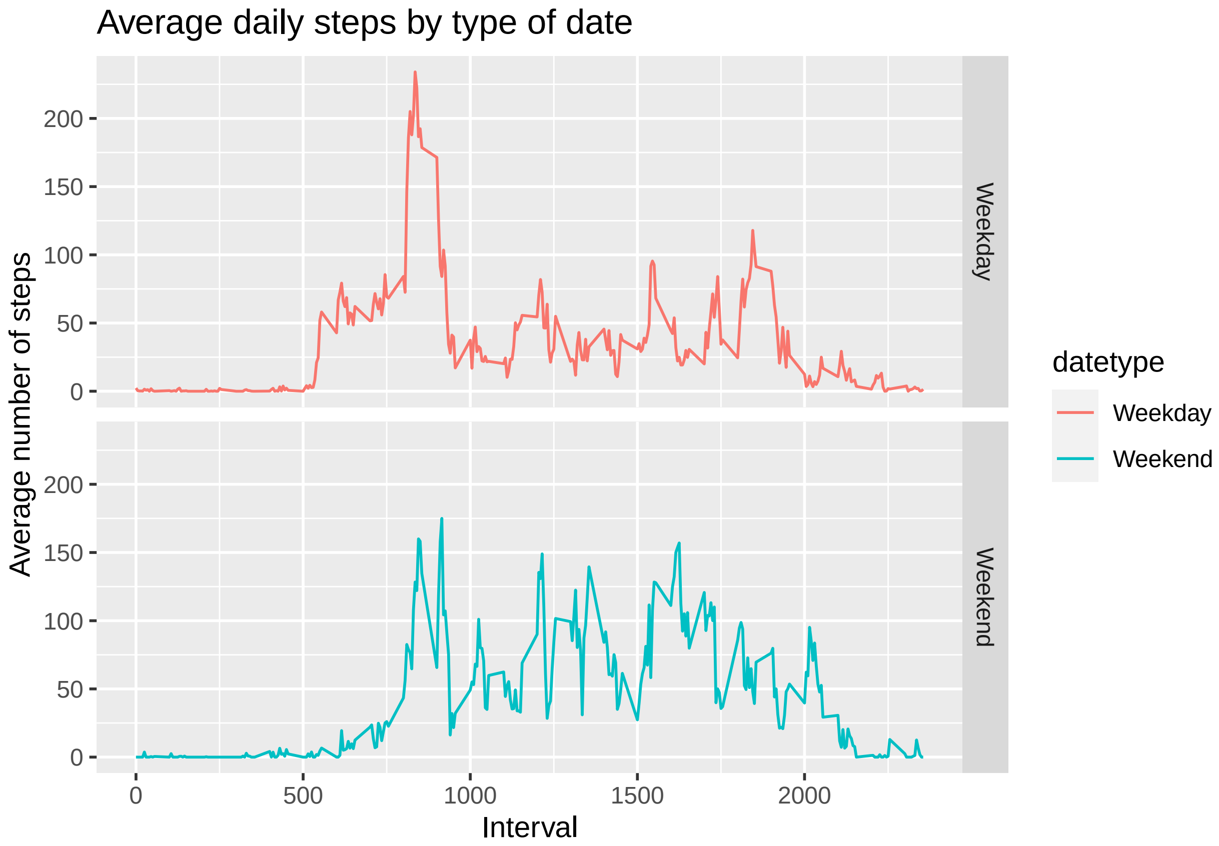
Are there differences in activity patterns between weekdays and weekends?

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

```
library(ggplot2)
activity$date <- as.Date(strptime(activity$date, format="%Y-%m-%d"))
activity$datetype <- sapply(activity$date, function(x)
        {
        if (weekdays(x) == "Saturday" | weekdays(x) =="Sunday")
                {y <- "Weekend"} else
                {y <- "Weekday"}
                y
        })
```

Make a panel plot containing a time series plot of hte 5-minute interval and the average number of steps taken.

```
activity_by_date <- aggregate(steps~interval + datetype, activity, mean, na.rm = TRUE)
plot<- ggplot(activity_by_date, aes(x = interval , y = steps, color = datetype)) + geom_line() + labs(ti
print(plot)
```


Average daily steps by type of date
Average number of steps
Interval
Weekday
Weekend
datetype
Weekday
Weekend