Reproducible Research: Peer Assessment 1

Below we analyze the data for Activity monitoring. The data is obtained from: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip

Loading and preprocessing the data

Load data and transform data:

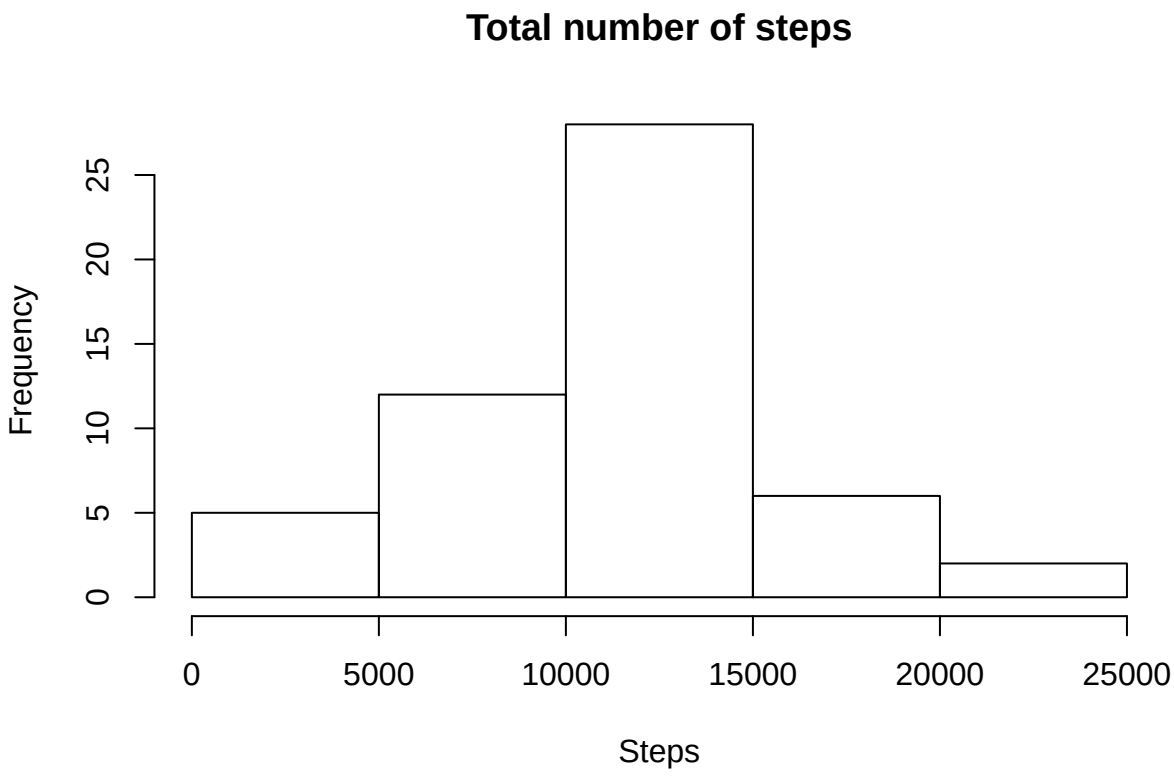
```
data_orig <- read.csv("activity.csv")          # loading
data <- data_orig[complete.cases(data_orig),] # removing rows with empty entry
```

What is mean total number of steps taken per day?

We first look into the simple statistics of the number of steps taken per day:

```
dates <-unique(data$date)
total_steps <- c()                             # total number of steps
for (date in dates){
    data_sub <- data[data$date == date,]
    total_step <- sum(data_sub$steps)
    total_steps <- c(total_steps, total_step)
}

# Below is plot a histogram of total steps
hist(total_steps, xlab = "Steps", main = "Total number of steps")
```




```
# The mean and median
mean(total_steps)
```

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

```
median(total_steps)
```

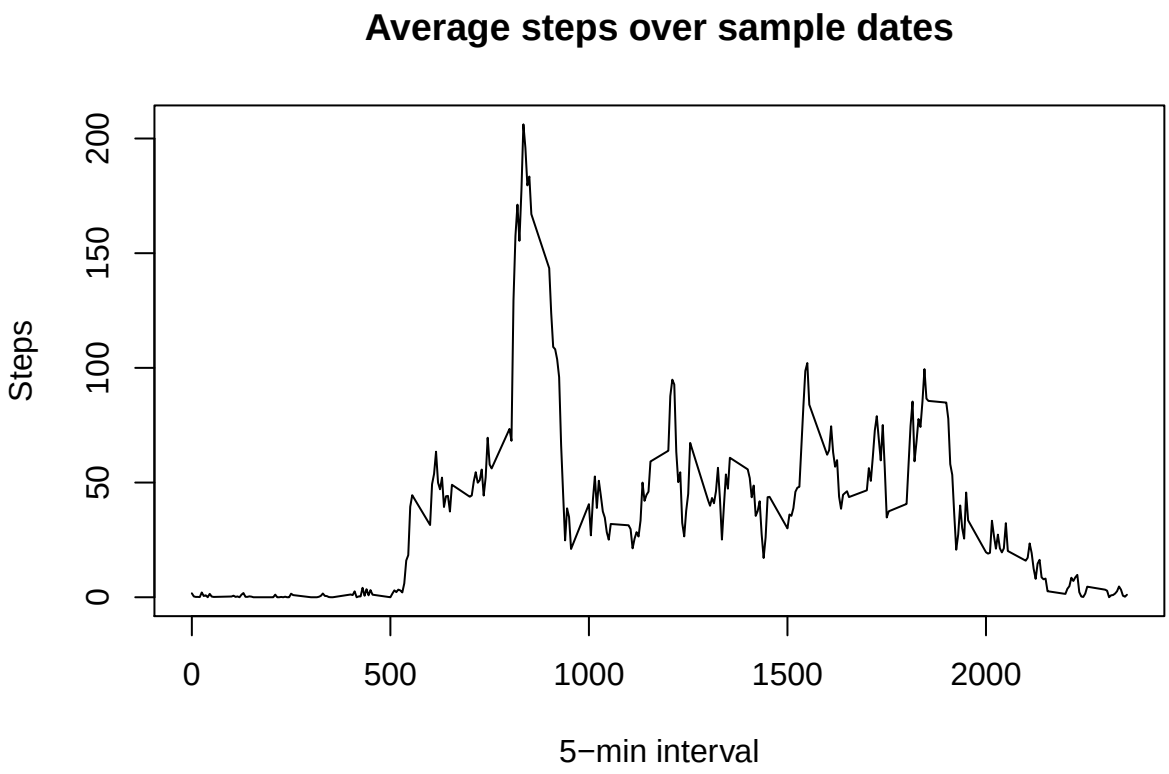
```
## [1] 10765
```

What is the average daily activity pattern?

Secondly, we look into the average daily activity pattern:

```
intervals <- unique(data$interval)
mean_steps <- c()                   # average steps in an interval
for (interval in intervals){
    data_sub <- data[data$interval == interval,]
    mean_step <- mean(data_sub$steps)
    mean_steps <- c(mean_steps, mean_step)
}

plot(intervals,mean_steps, type = "l", xlab = "5-min interval", ylab = "Steps", main = "Average steps o
```




```
# The maximum steps taken during a day on average is found at interval:
data$interval[which(mean_steps == max(mean_steps))]
```

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

Imputing missing values

With the data obtained from the previous step, we now fill in the missing value (i.e. NA) in the original data with the mean_step value for the corresponding 5-min interval, and re-evaluate the statistics in step 1:

```
# The total number of lines with missing values are:
nrow(data_orig) - nrow(data)
```

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

```
# Now fill in the missing value (i.e. NA) in the original data with the mean_step value for the corresp
data_mod <- data_orig
for (interval in intervals){
    data_mod[data_mod$interval == interval & is.na(data_mod$steps), 1] <- mean_steps[which(intervals ==
}

# Now re-evaluate the statistics in step 1
dates <-unique(data_mod$date)
```

```
total_steps <- c()                              # total number of steps
for (date in dates){
    data_sub <- data_mod[data_mod$date == date,]
    total_step <- sum(data_sub$steps)
    total_steps <- c(total_steps, total_step)
}

# The mean and median of the data with fixed original data
mean(total_steps)
```

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

```
median(total_steps)
```

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

```
# We note that the mean is not changed, but the median is shifted towards the mean
```

We note that from the previous step, the mean is not changed, but the median is shifted towards the original mean, which is calculated by removing the missing data points from the original data set.

Are there differences in activity patterns between weekdays and weekends?

Finally, we investigate whether the activity level can be different for weekends vs weekdays from this data set:

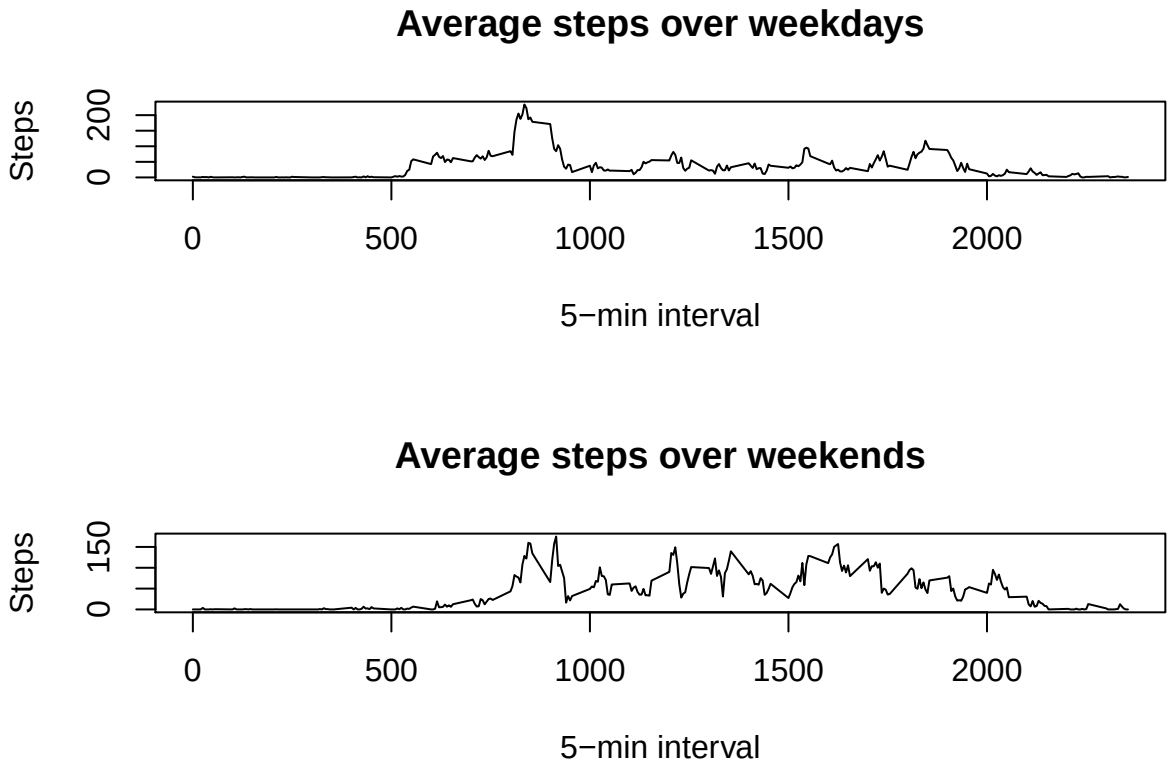
```
# Make a new data set for which the days are classified as weekday or weekend
data_comp <- transform(data, date = as.character(date))
dates <-unique(data_comp$date)
total_steps <- c()                              # total number of steps
for (date in dates){
    if (weekdays(as.Date(date)) == "Saturday" | weekdays(as.Date(date)) == "Sunday") {
        data_comp[data_comp$date == date, 2] <- "weekend"
    }
    else {
        data_comp[data_comp$date == date, 2] <- "weekday"
    }
}

# Now average the activity levels for weekdays and weekends separately:
```

```
intervals <- unique(data_comp$interval)
mean_steps_wkday <- c()
mean_steps_wkend <- c()
for (interval in intervals){
    data_sub_wkday <- data[data_comp$interval == interval & data_comp$date == "weekday",]
    data_sub_wkend <- data[data_comp$interval == interval & data_comp$date == "weekend",]
    mean_step_wkday <- mean(data_sub_wkday$steps)
    mean_step_wkend <- mean(data_sub_wkend$steps)
    mean_steps_wkday <- c(mean_steps_wkday, mean_step_wkday)
    mean_steps_wkend <- c(mean_steps_wkend, mean_step_wkend)
```

```
}

# now plot the data
par(mfrow = c(2, 1))
plot(intervals,mean_steps_wkday, type = "l", xlab = "5-min interval", ylab = "Steps", main = "Average st
plot(intervals,mean_steps_wkend, type = "l", xlab = "5-min interval", ylab = "Steps", main = "Average st
```



We see that the weekend activity levels are indeed different from those from the weekends, with fewer steps taken on average during the day, and the peak intervals flattened out.