Reproducible_Research_Assignment_1

Pratik Nirhali

February 19, 2017

Intordunction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit (http://www.fitbit.com/), Nike Fuelband (http://www.nike.com/us/en_us/c/nikeplus-fuelband), or Jawbone Up (https://jawbone.com/up). These type of devices are part of the “quantified self” movement - a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

Loading the data

```
activity <- read.csv ("activity.csv", header = T, sep = ",", stringsAsFactors =
F)
```

Process data

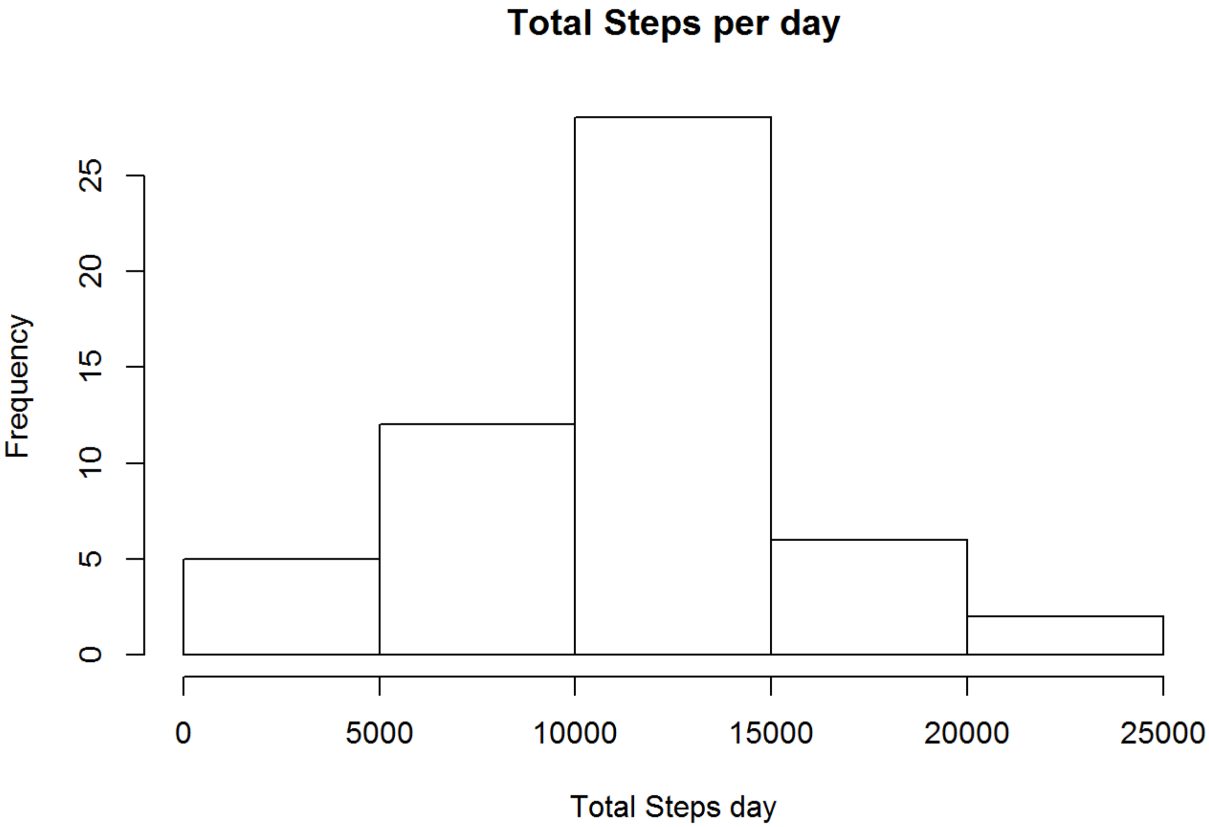
```
totalSteps <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
```

Analysis

What is mean total number of steps taken per day?

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

```
hist(totalSteps$steps,main="Total Steps per day",xlab="Total Steps day",cex.axis=
1,cex.lab = 1)
```



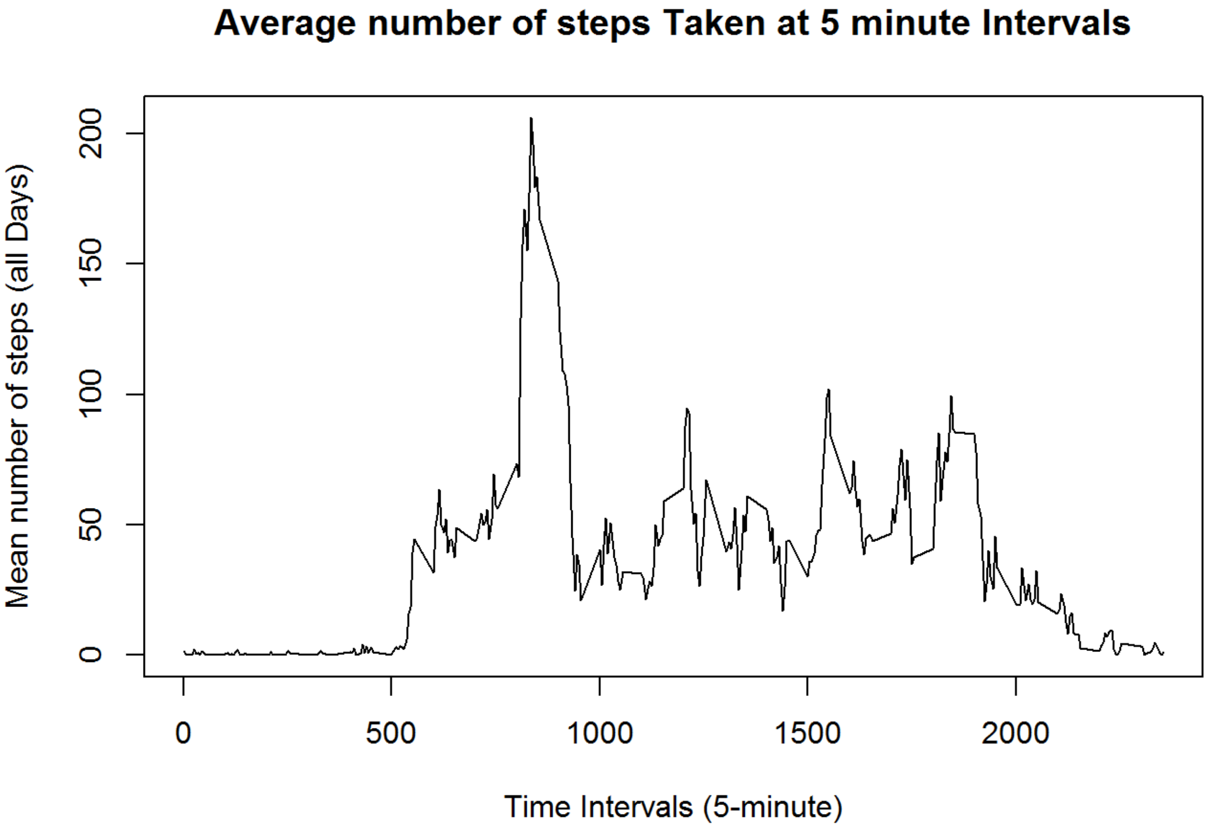
Calculate and report the mean and median total number of steps taken per day

```
mean_steps <- mean(totalSteps$steps)
median_steps <- median(totalSteps$steps)
```

What is the average daily activity pattern?

Make 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)

```
steps_interval <- aggregate(steps ~ interval, data = activity, mean, na.rm = TRU
E)
plot(steps ~ interval, data = steps_interval, type = "l", xlab = "Time Intervals
(5-minute)", ylab = "Mean number of steps (all Days)", main = "Average number of
steps Taken at 5 minute Intervals")
```



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

```
maxStepInterval <- steps_interval[which.max(steps_interval$steps),"interval"]
```

Imputing missing values

Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
missing_rows <- sum(!complete.cases(activity))
```

Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
## This function returns the mean steps for a given interval
getMeanStepsPerInterval <- function(interval){
    steps_interval[steps_interval$interval==interval,"steps"]
}
```

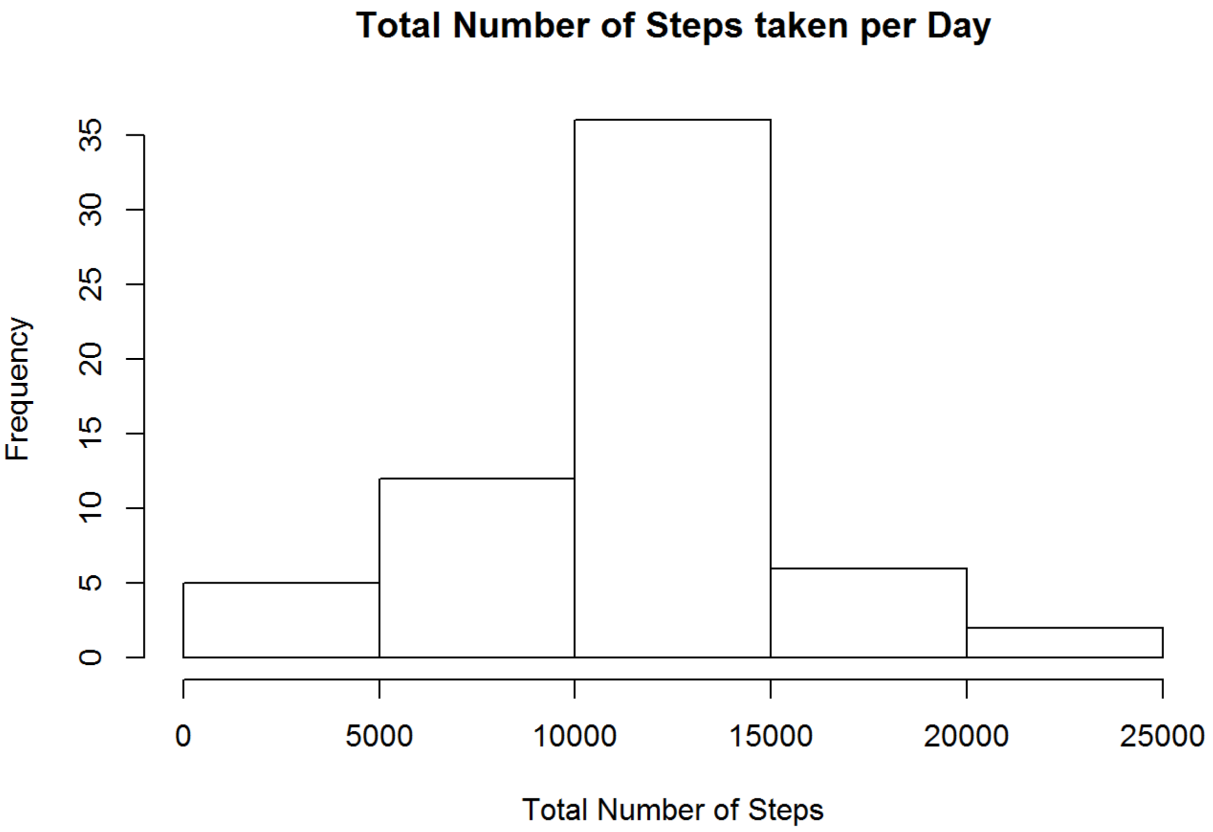
Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
complete.activity <- activity

## Filling the missing values with the mean for that 5-minute interval
flag = 0
for (i in 1:nrow(complete.activity)) {
    if (is.na(complete.activity[i,"steps"])) {
        complete.activity[i,"steps"] <- getMeanStepsPerInterval(complete.activity
[i,"interval"])
        flag = flag + 1
        }
}
```

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

```
total.steps.per.days <- aggregate(steps ~ date, data = complete.activity, sum)
hist(total.steps.per.days$steps, xlab = "Total Number of Steps",
     ylab = "Frequency", main = "Total Number of Steps taken per Day")
```



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

```
showMean <- mean(total.steps.per.days$steps)
showMedian <- median(total.steps.per.days$steps)
```

Do these values differ from the estimates from the first part of the assignment?
Ans: The mean value is the same as the value before imputing missing data, but the median value has changed.

What is the impact of imputing missing data on the estimates of the total daily number of steps?
Ans: The mean value is the same as the value before imputing missing data since the mean value has been used for that particular 5-min interval. The median value is different, since the median index is now being changed after imputing missing values.

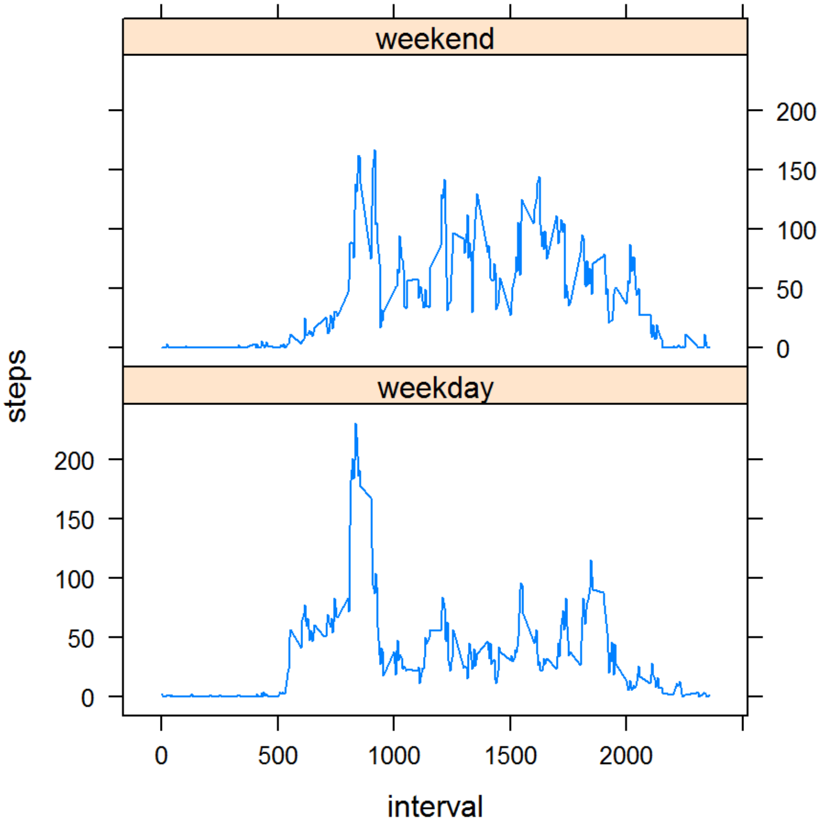
Are there differences in activity patterns between weekdays and weekends?

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.

```
complete.activity$day <- ifelse(as.POSIXlt(as.Date(complete.activity$date))$wday%%6 ==
                                    0, "weekend", "weekday")
complete.activity$day <- factor(complete.activity$day, levels = c("weekday", "weekend"))
```

Make 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).

```
steps.interval= aggregate(steps ~ interval + day, complete.activity, mean)
library(lattice)
xyplot(steps ~ interval | factor(day), data = steps.interval, aspect = 1/2,
       type = "l")
```



We observe that, as expected, the activity profiles between weekdays and weekends greatly differ. During the weekdays, activity peaks in the morning between 7 and 9 and then the activity remains below ~100 steps. In contrast, the weekend data does not show a period with particularly high level of activity, but the activity remains higher than the weekday activity at most times and in several instances it surpases the 100 steps mark and it is overall more evenly distributed throughout the day.