#15.05.2014 Assignment 1 done for the Coursera course “Reproducible Research” of “Data Science” Specialization track.

Loading and preprocessing the data

```
data <- read.csv("activity.csv")

# create the subset and remove NAs
cleandata <- subset(data, is.na(data$steps) == F)
```

What is mean total number of steps taken per day?

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

Loading the needed libraries for plyr and lattice

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.0.3
```

```
library(lattice)
```

```
## Warning: package 'lattice' was built under R version 3.0.3
```

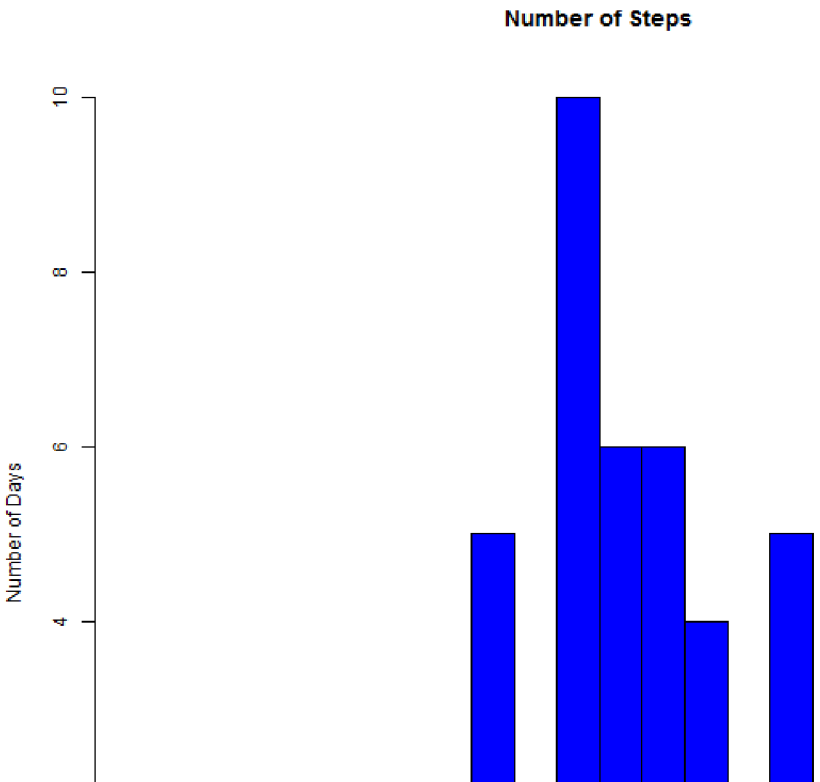
Calculating the total number of steps taken each day

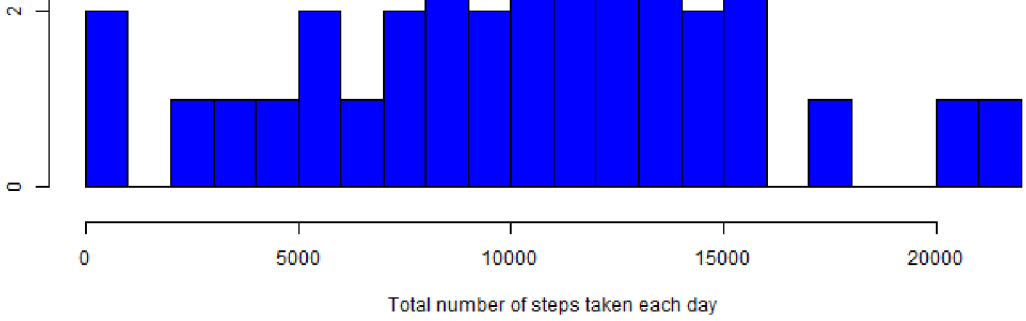
```

tpd <- ddply(cleandata, .(date), summarise, steps = sum(steps))
```

Creating the corresponding plot

```
hist(tpd$steps, breaks = 20, main = "Number of Steps", xlab = "Total number of steps taken
each day",
    ylab = "Number of Days", col = "blue")
```





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

Mean:

```
mean(tpd$steps)
```

```
## [1] 10766
```

Median:

```
median(tpd$steps)
```

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

What is the average daily activity pattern?

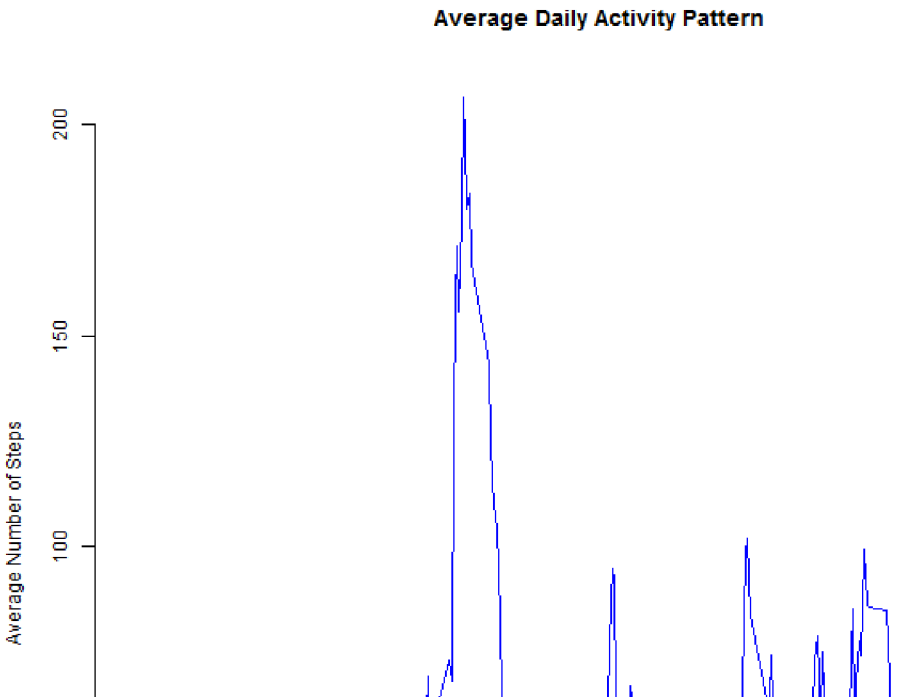
1. Make a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

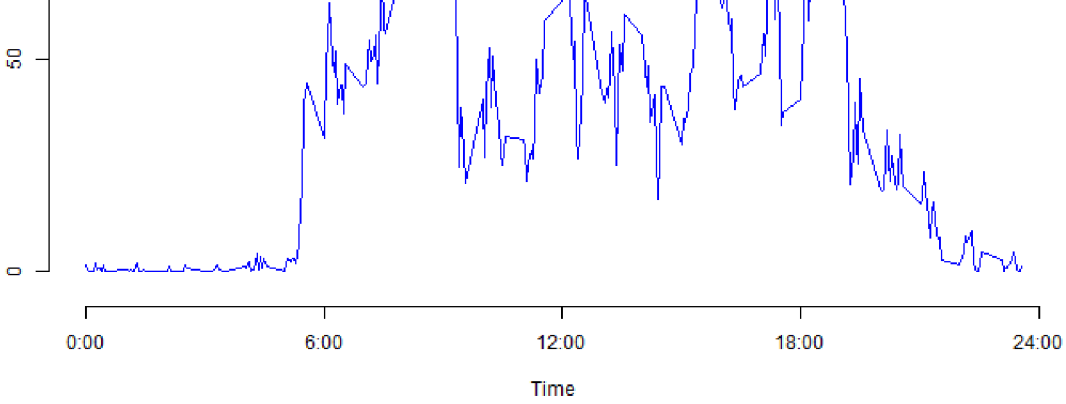
Calcluating the average number of steps taken for each 5-min intervals

```
averagePerInterval <- ddply(cleandata, .(interval), summarise, steps = mean(steps))
```

Creating the corresponding plot

```
plot(averagePerInterval$interval, averagePerInterval$steps, axes = F, type = "l",
    col = "blue", xlab = "Time", ylab = "Average Number of Steps", main = "Average Daily
Activity Pattern")
axis(1, at = c(0, 600, 1200, 1800, 2400), label = c("0:00", "6:00", "12:00",
    "18:00", "24:00"))
axis(2)
```





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

```
averagePerInterval[which.max(averagePerInterval$steps), ]
```

```
##     interval steps
## 104      835 206.2
```

It should be the interval from 8:35 to 8:40

Imputing missing values

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

```
sum(is.na(data$steps))
```

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

2.Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated.

I will fill the NA with average value for that 5-min interval

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

```
imputed <- data

for (i in 1:nrow(imputed)) {
    if (is.na(imputed$steps[i])) {
        imputed$steps[i] <- averagePerInterval$steps[which(imputed$interval[i] ==
            averagePerInterval$interval)]
    }
}

imputed <- arrange(imputed, interval)
```

3. 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?

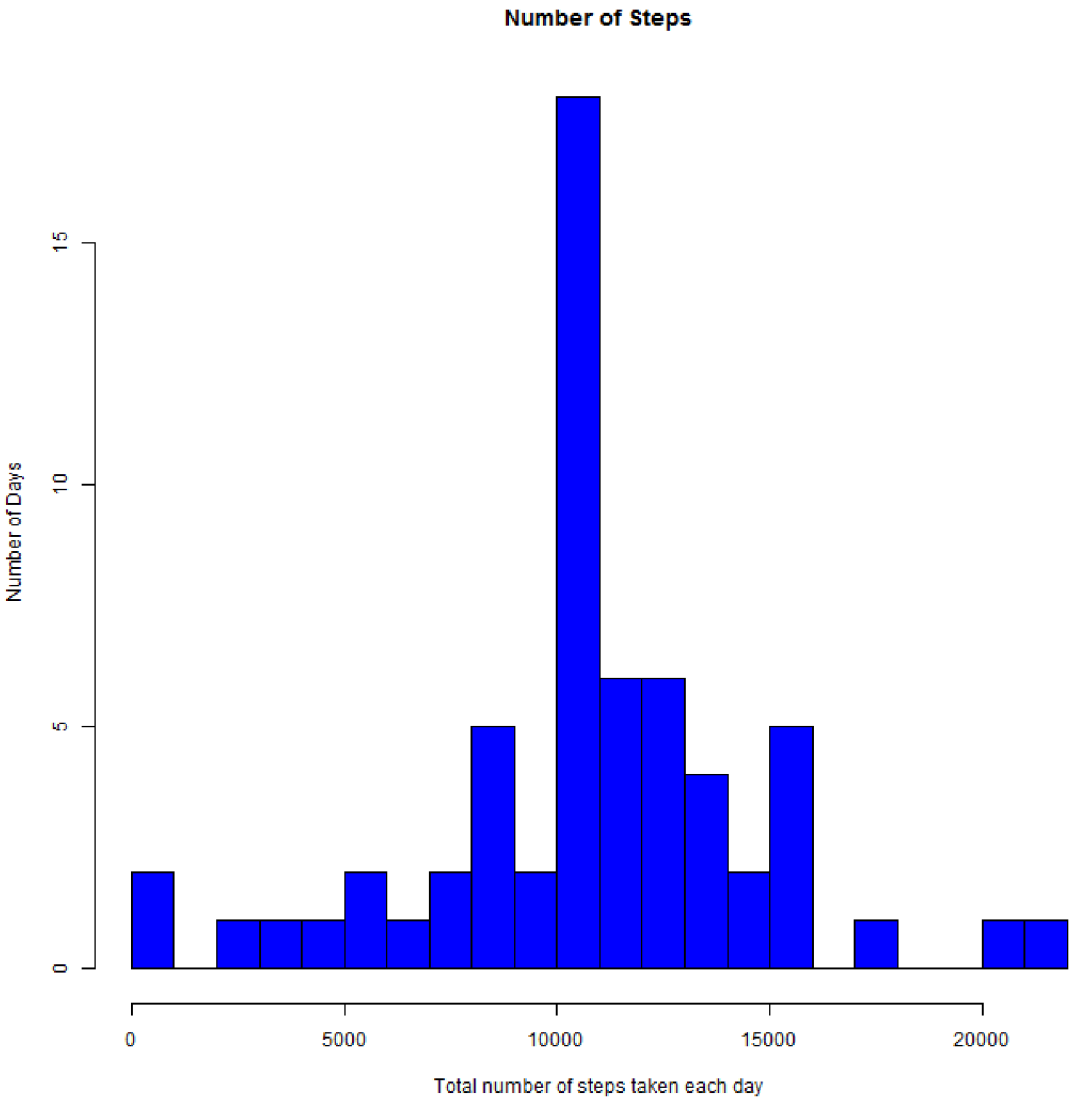
Calculating the total number of steps taken each day

```
totalPerDayImputed <- ddply(imputed, .(date), summarise, steps = sum(steps))
```

Creating the corresponding plot

```
hist(totalPerDayImputed$steps, breaks = 20, main = "Number of Steps", xlab = "Total number
```

```
of steps taken each day",
    ylab = "Number of Days", col = "blue")
```



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

```
mean(totalPerDayImputed$steps)
```

```
## [1] 10766
```

```
median(totalPerDayImputed$steps)
```

```
## [1] 10766
```

Test: Do these values differ from those in the first part

```
abs(mean(tpd$steps) - mean(totalPerDayImputed$steps))
```

```
## [1] 0
```

```
abs(median(tpd$steps) - median(totalPerDayImputed$steps))/median(totalPerDay$steps)
```

```
## Error: Objekt 'totalPerDay' nicht gefunden
```

The mean didn't change. tThe median is slightly changed about 0.1% of the original value.

Test: How do total steps taken per day differ

```
totalDifference <- sum(imputed$steps) - sum(cleandata$steps)
```

```
totalDifference
```

```
## [1] 86130
```

Impute the dataset cause the estimation on total steps per day to increase

Are there differences in activity patterns between weekdays and weekends?

1. 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.

```
Sys.setlocale("LC_TIME", "English")
```

```
## [1] "English_United States.1252"
```

```
imputed$weekdays <- weekdays(as.Date(imputed$date))
imputed$weekdays <- ifelse(imputed$weekdays %in% c("Saturday", "Sunday"), "weekend",
    "weekday")
```

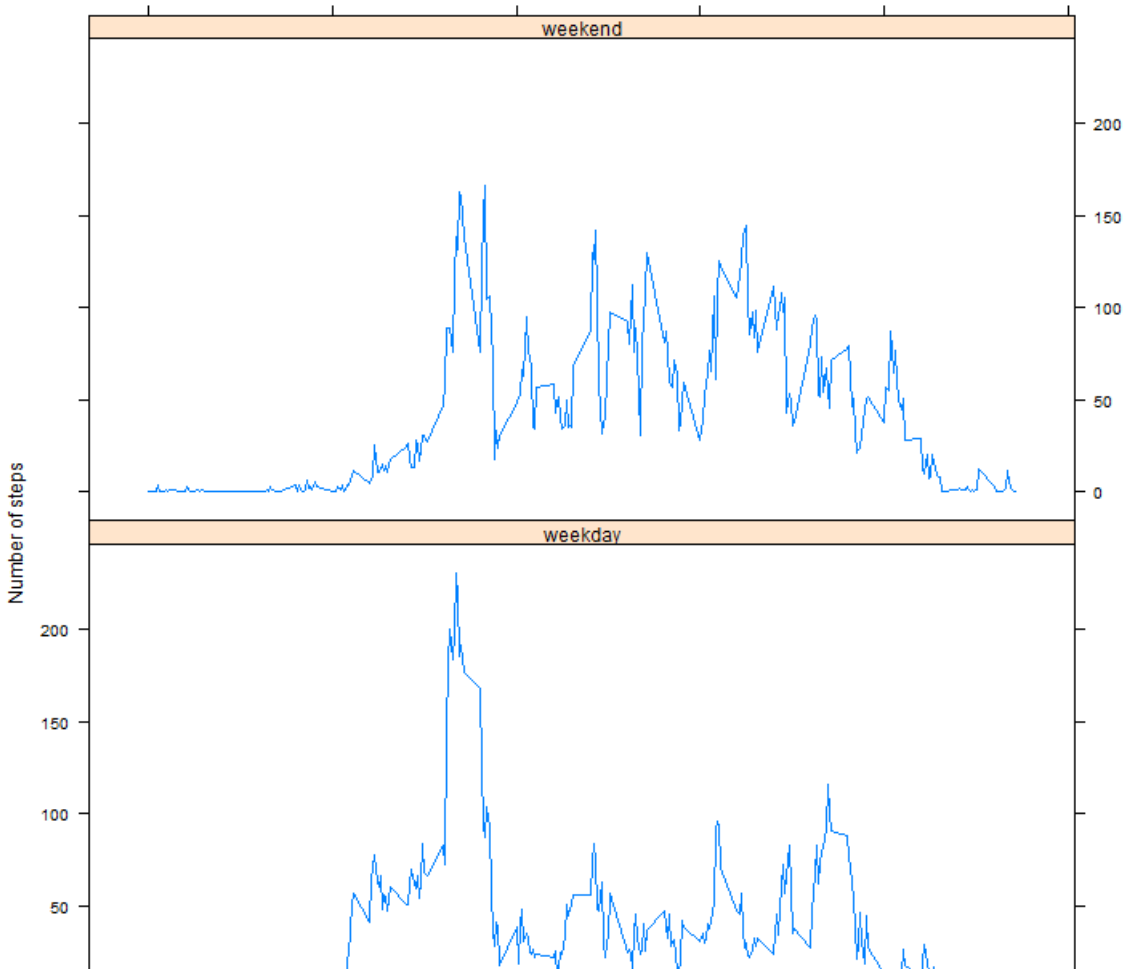
2.Make a panel plot containing a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

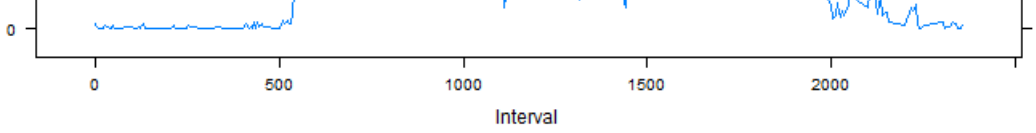
Calculating the average for each interval

```
average <- ddply(imputed, .(interval, weekdays), summarise, steps = mean(steps))
```

Creating the corresponding plot

```
xyplot(steps ~ interval | weekdays, data = average, layout = c(1, 2), type = "l",
    xlab = "Interval", ylab = "Number of steps")
```




0
0
500
1000
1500
2000
Interval