Reproducible research assignment 1

Azat Gabdolla

29 декабря 2018 г

```
suppressMessages(suppressWarnings(library(data.table)))
suppressMessages(suppressWarnings(library(lattice)))
suppressMessages(suppressWarnings(library(ggplot2)))
```

Loading and preprocessing the data

1. Code for reading in the dataset and/or processing the data

```
setwd("~/Edu/Reproducible research")

activity <- setDT(read.csv("activity.csv",sep = ",", header = TRUE, na.strings = c("NA", " ",
 '#DIV/0!')))
```

What is mean total number of steps taken per day?

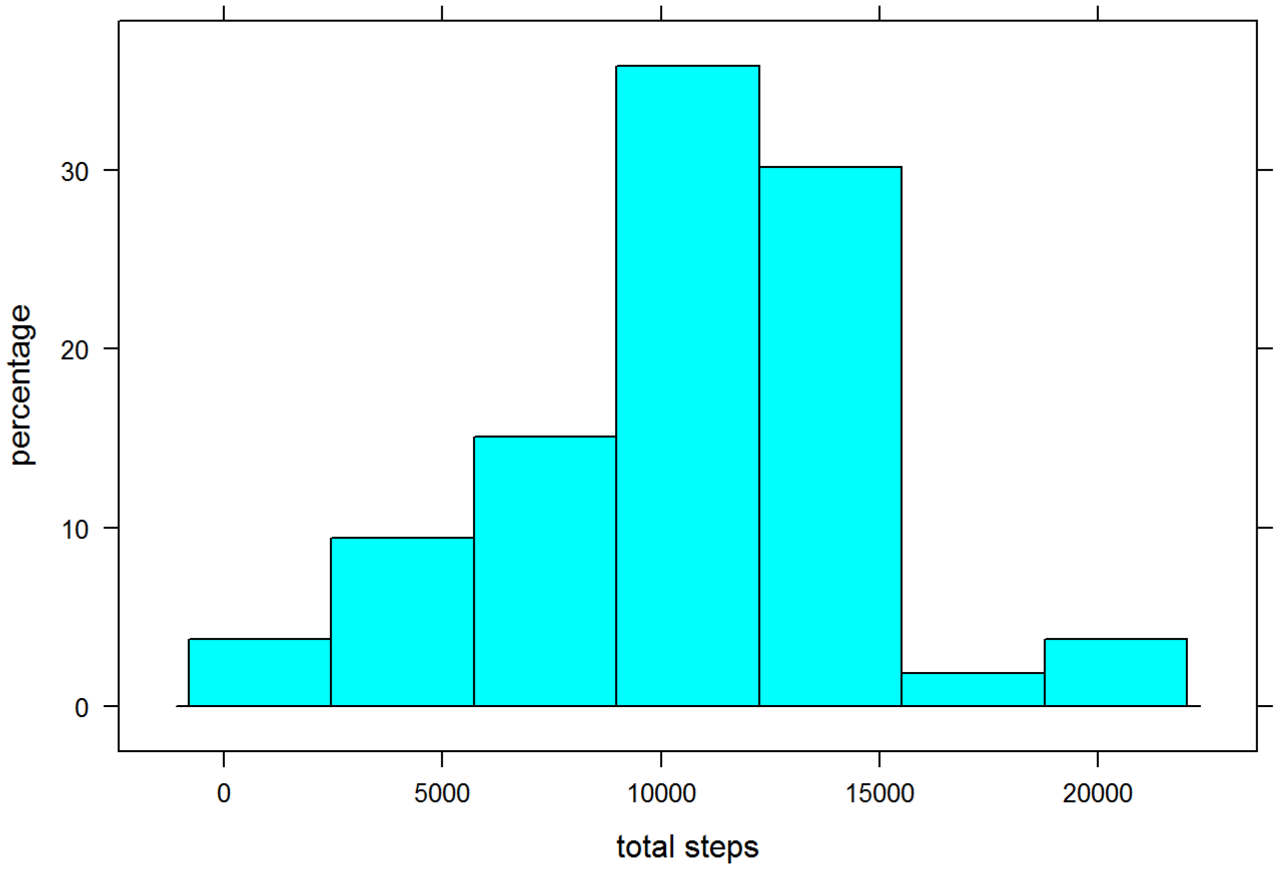
Table of total steps taken per day

```
activity[!is.na(steps) , sum(steps), date]
```

```
##           date    V1
##  1: 2012-10-02   126
##  2: 2012-10-03 11352
##  3: 2012-10-04 12116
##  4: 2012-10-05 13294
##  5: 2012-10-06 15420
##  6: 2012-10-07 11015
##  7: 2012-10-09 12811
##  8: 2012-10-10  9900
##  9: 2012-10-11 10304
## 10: 2012-10-12 17382
## 11: 2012-10-13 12426
## 12: 2012-10-14 15098
## 13: 2012-10-15 10139
## 14: 2012-10-16 15084
## 15: 2012-10-17 13452
## 16: 2012-10-18 10056
## 17: 2012-10-19 11829
## 18: 2012-10-20 10395
## 19: 2012-10-21  8821
## 20: 2012-10-22 13460
## 21: 2012-10-23  8918
## 22: 2012-10-24  8355
## 23: 2012-10-25  2492
## 24: 2012-10-26  6778
## 25: 2012-10-27 10119
## 26: 2012-10-28 11458
## 27: 2012-10-29  5018
## 28: 2012-10-30  9819
## 29: 2012-10-31 15414
## 30: 2012-11-02 10600
## 31: 2012-11-03 10571
## 32: 2012-11-05 10439
## 33: 2012-11-06  8334
## 34: 2012-11-07 12883
## 35: 2012-11-08  3219
## 36: 2012-11-11 12608
## 37: 2012-11-12 10765
## 38: 2012-11-13  7336
## 39: 2012-11-15    41
## 40: 2012-11-16  5441
## 41: 2012-11-17 14339
## 42: 2012-11-18 15110
## 43: 2012-11-19  8841
## 44: 2012-11-20  4472
## 45: 2012-11-21 12787
## 46: 2012-11-22 20427
## 47: 2012-11-23 21194
## 48: 2012-11-24 14478
## 49: 2012-11-25 11834
## 50: 2012-11-26 11162
## 51: 2012-11-27 13646
## 52: 2012-11-28 10183
## 53: 2012-11-29  7047
##           date    V1
```

2. Histogram of the total number of steps taken each day

```
histogram(activity[!is.na(steps) , sum(steps), date][,V1], xlab = c("total steps"), ylab = c(
"percentage") )
```



3. Mean and median number of steps taken each day

```
paste("Mean - ", activity[!is.na(steps) , sum(steps), date][,mean(V1)])
```

```
## [1] "Mean -  10766.1886792453"
```

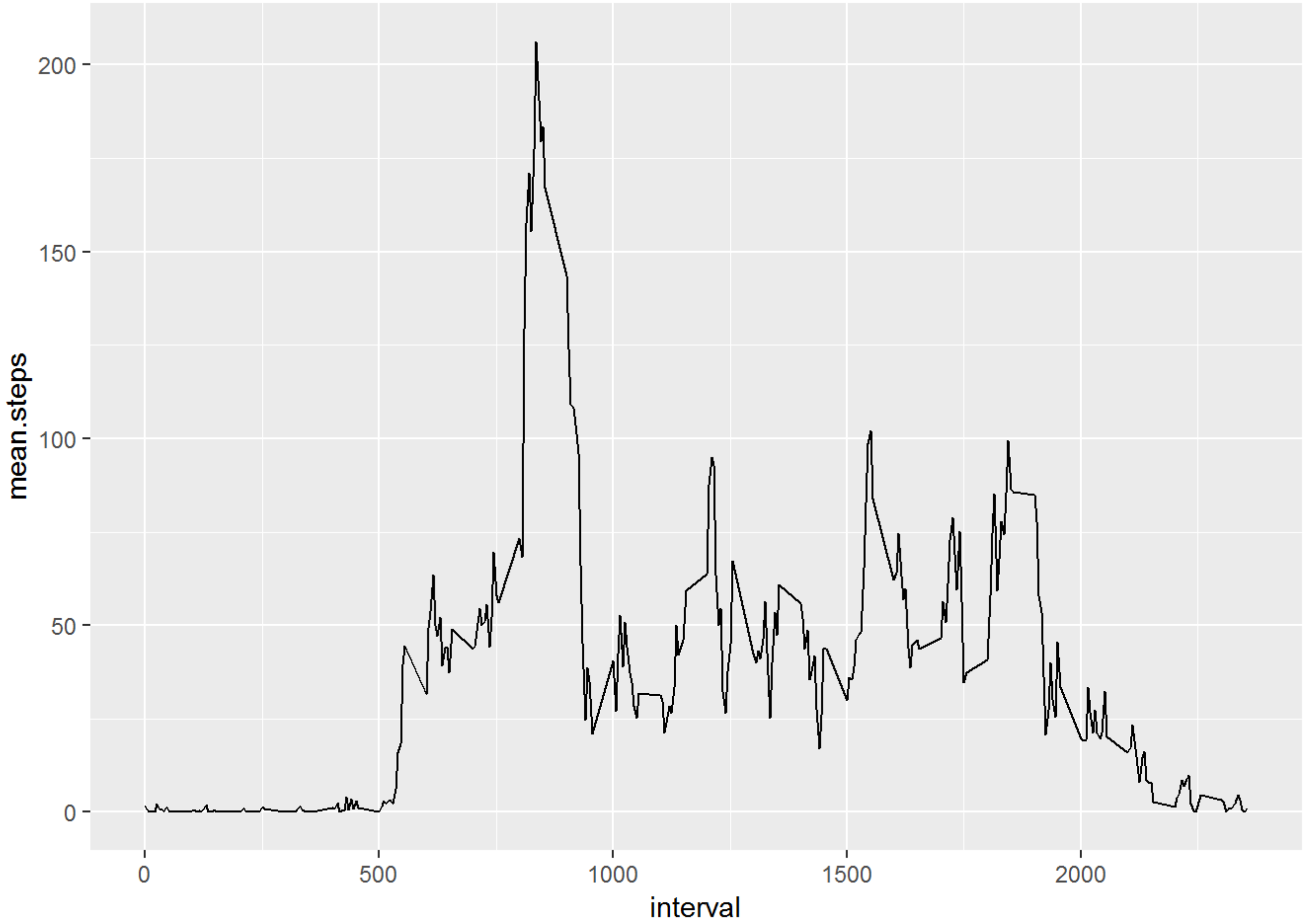
```
paste("Median - ", activity[!is.na(steps) , sum(steps), date][,median(V1)])
```

```
## [1] "Median -  10765"
```

What is the average daily activity pattern?

4. Time series plot of the average number of steps taken

```
tsactivity <-  aggregate(activity$steps, by = list(activity$interval), mean, na.rm=TRUE)
tsactivity <- setDT(tsactivity)
setnames(tsactivity, c("interval" , "mean.steps"))
ggplot(tsactivity,  aes(x=interval, y = mean.steps))+geom_line()
```



5. The 5-minute interval that, on average, contains the maximum number of steps

```
paste("Interval on which maximum steps are done is ", tsactivity[mean.steps == max(mean.step
s), .(interval)][,interval])
```

```
## [1] "Interval on which maximum steps are done is  835"
```

Imputing Missing values

6. Code to describe and show a strategy for imputing missing data Code that is used is attached in markdown, so in this part I explain the way

7. We calculate number of missing value

```
paste( activity[is.na(steps),.N], "of cells are missing in steps")
```

```
## [1] "2304 of cells are missing in steps"
```

2. I calculate average number of steps per one unit of interval done on measures that are not 0 and empty
3. I impute in missing values of steps the average step per one unit of measurement multiplied by the interval, so I get new dataset without missing values

```
Oneunit <- activity[!is.na(steps) & ! interval == 0, mean(steps/interval)]
# 0.0350705
activity[is.na(steps), steps := Oneunit*interval]
```

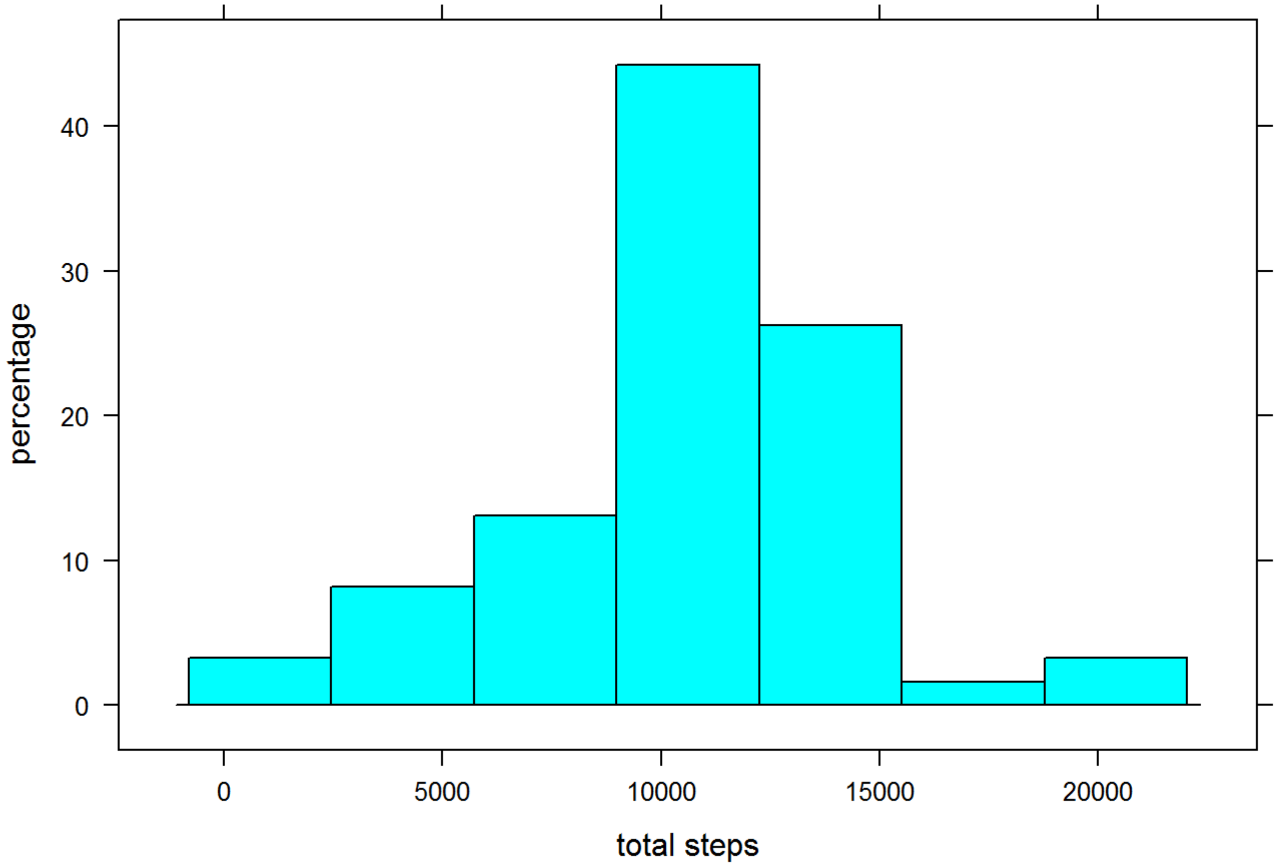
```
## Warning in `[.data.table`(activity, is.na(steps), `:=`(steps, Oneunit
## * : Coerced double RHS to integer to match the type of the target column
## (column 1 named 'steps'). One or more RHS values contain fractions which
## have been lost; e.g. item 2 with value 0.175353 has been truncated to 0.
```

```
paste("Number of missing values in new dataset -", activity[is.na(steps), .N])
```

```
## [1] "Number of missing values in new dataset - 0"
```

7. Histogram of the total number of steps taken each day after missing values are imputed

```
histogram(activity[!is.na(steps) , sum(steps), date][,V1], xlab = c("total steps"), ylab = c(
"percentage") )
```



```
paste("New Mean - ", activity[!is.na(steps) , sum(steps), date][,mean(V1)])
```

```
## [1] "New Mean -  10895.3442622951"
```

```
paste("New Median - ", activity[!is.na(steps) , sum(steps), date][,median(V1)])
```

```
## [1] "New Median -  11458"
```

There is no big difference in patterns how data behave. However new mean and new median are larger than in dataset with missing values

Are there differences in activity patterns between weekdays and weekends?

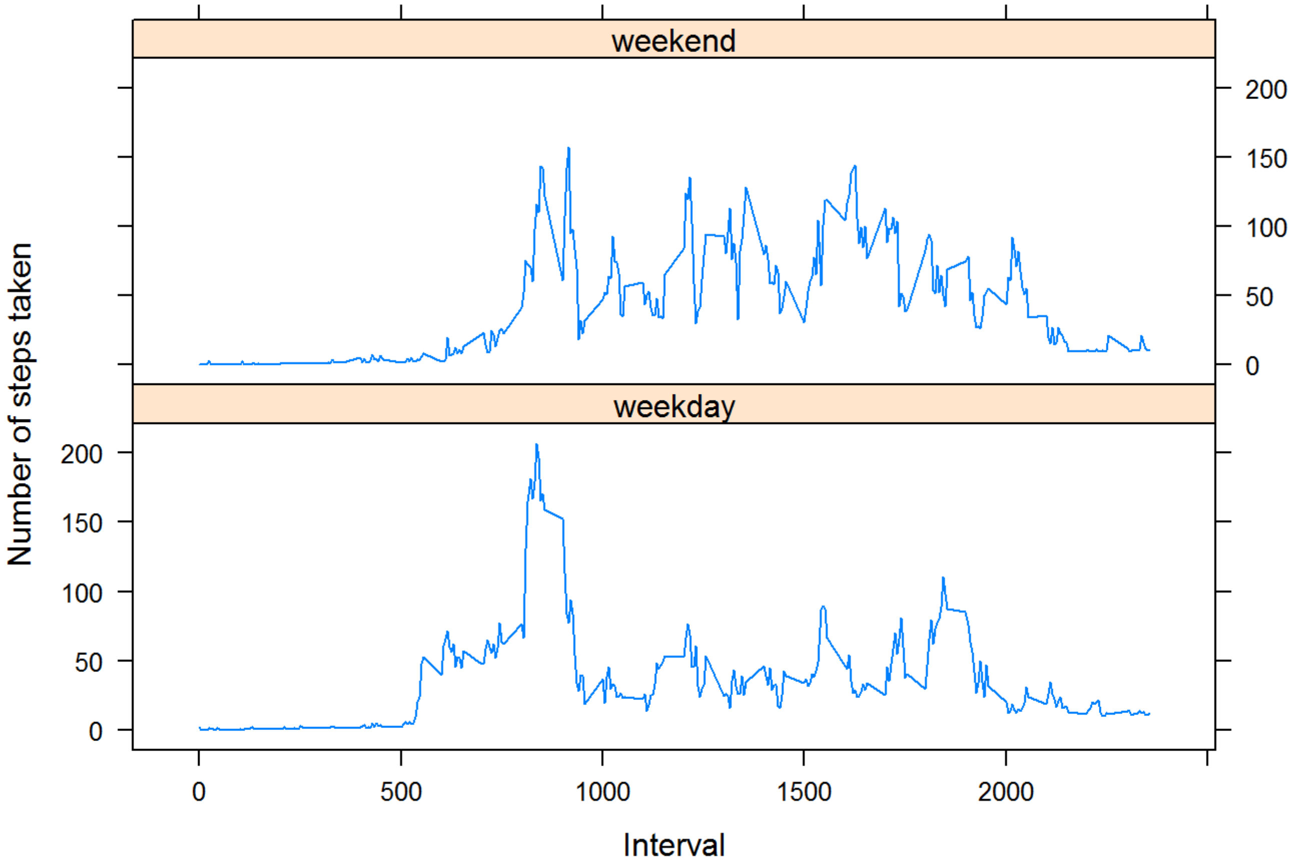
```
  activity[,date:= as.Date(date)]

activity$weekdays <- as.factor(ifelse(weekdays(activity$date) %in% c("суббота","воскресенье"
), "weekend", "weekday"))
```

8. Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

```
steps_interval_mean <- aggregate(steps ~ interval + weekdays, FUN=mean,
                                    data=activity)

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



In order to make code reproducible, please set russian language on your computer, and locate code, data on folder “Reproducible Research”" in Folder “Edu” on your computer. In case of absence of necessary package, download them with install.packages() function. In these cases Markdown will be reproducible