Reproducible Research - Project 1

ThierryFauret

6 mai 2017

Loading and preprocessing the data

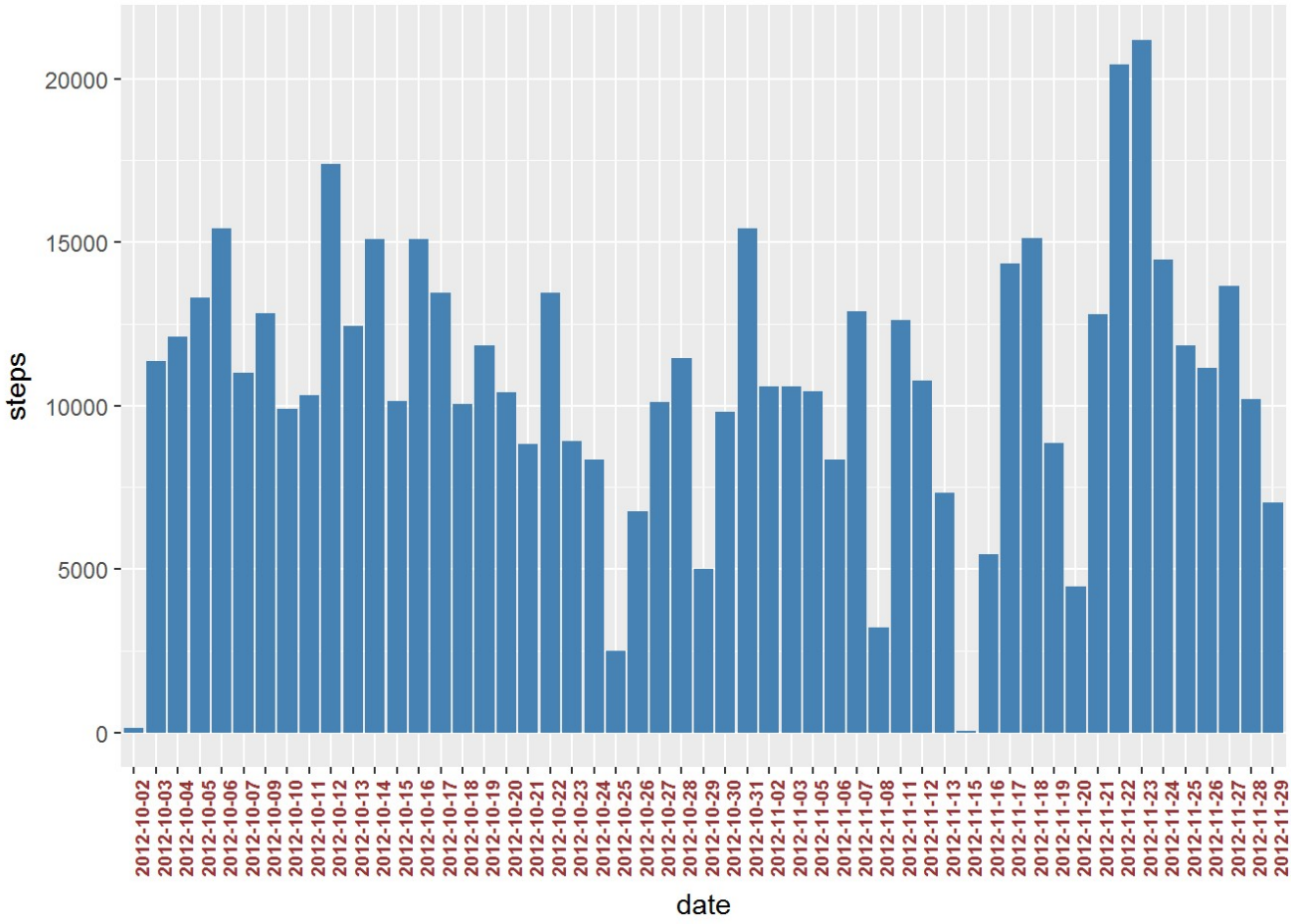
```
library(dplyr)
library(lubridate)
library(ggplot2)
activity<-read.csv("activity.csv",sep=",")
activity<-mutate(activity,steps=as.numeric(steps))

ok<-complete.cases(activity)
activity.complete<-activity[ok,]
```

We have created the data.frame activity.complete which does not have any NA.

What is mean total number of steps taken per day?

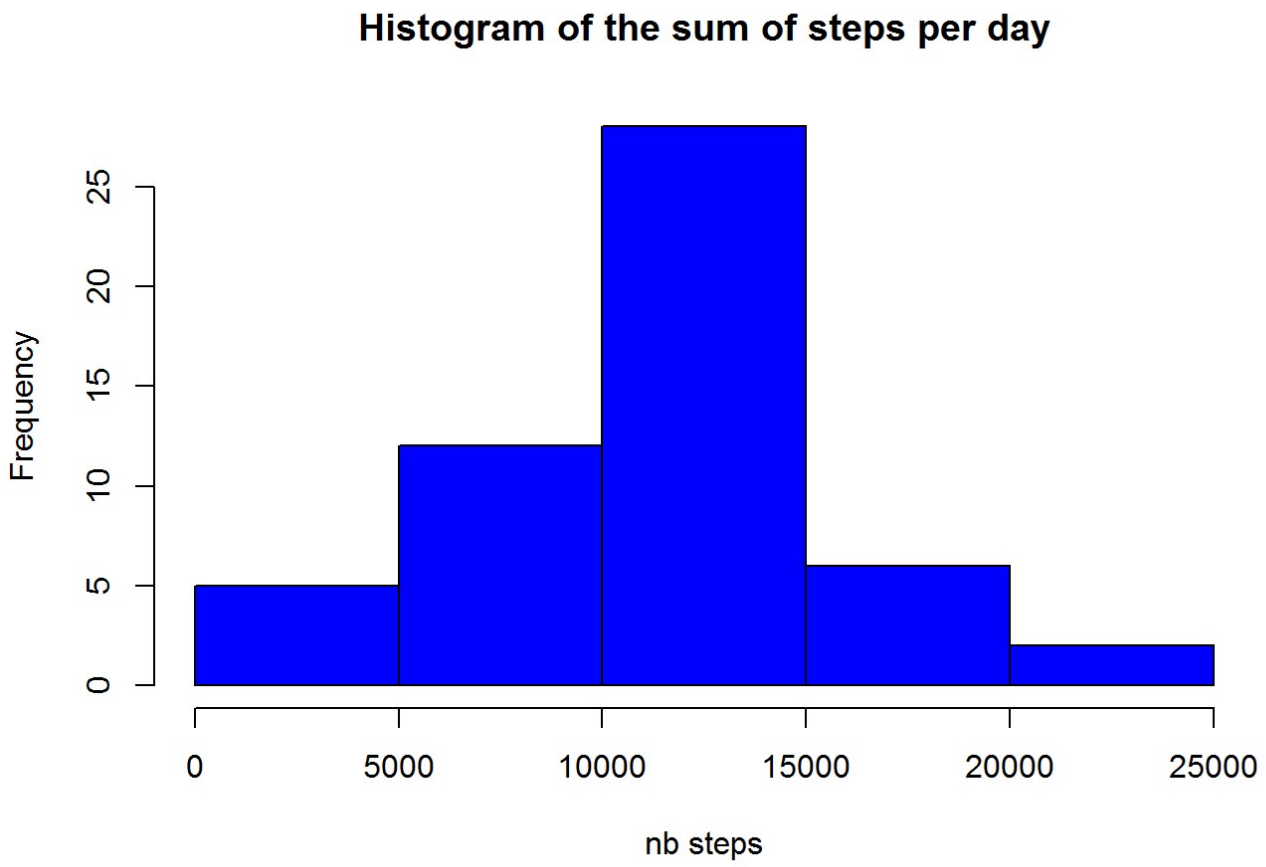
1. Calculate the total number of steps taken per day :

```
nb.ByDate.steps<-aggregate(steps~date,data=activity.complete,sum)
ggplot(data=nb.ByDate.steps, aes(x=date, y=steps)) +geom_bar(stat="identity", fill=
"steelblue")+
        theme(axis.text.x = element_text(face="bold", color="#993333",  size=7, ang
le=90))
```



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

```
hist(nb.ByDate.steps$steps,xlab="nb steps",main="Histogram of the sum of steps per day",col="blue")
```



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

```
print(mean(nb.ByDate.steps$steps))
```

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

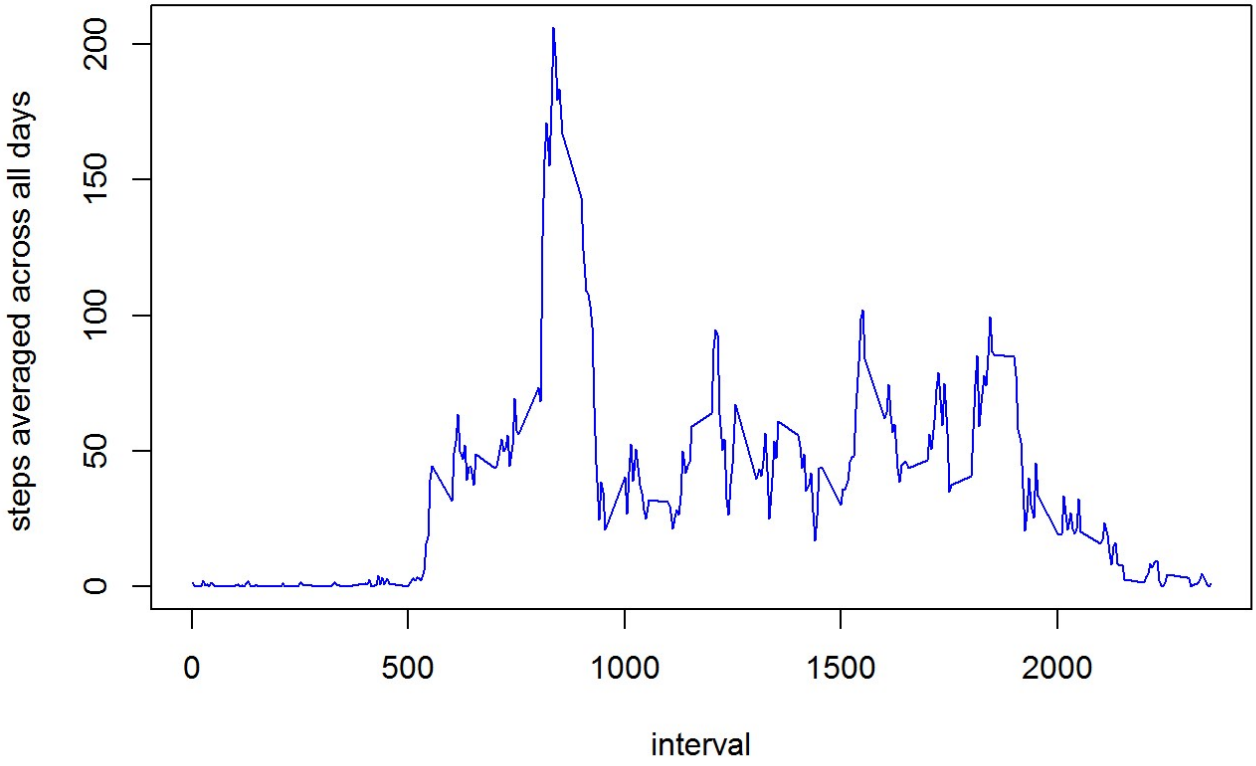
```
print(median(nb.ByDate.steps$steps))
```

```
## [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)

```
meanByInterval.steps<-aggregate(steps~interval,data=activity.complete,mean)
with(meanByInterval.steps,plot(interval,steps,type="l",ylab="steps averaged across
all days ",col="blue"))
```



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

```
interval.max.steps<-meanByInterval.steps$interval[which.max(meanByInterval.steps$steps)]
```

The 5-minute interval which contains the maximum number of steps is 835.

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)

```
nrow.missing.value<-nrow(activity)-nrow(activity.complete)
```

The number of Nas is : 2304.

2. 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.

In the following table, We can identify either days with only NA or days with complete data.

```
table(activity[ok,2])
```

```
##
## 2012-10-01 2012-10-02 2012-10-03 2012-10-04 2012-10-05 2012-10-06
##          0        288        288        288        288        288
## 2012-10-07 2012-10-08 2012-10-09 2012-10-10 2012-10-11 2012-10-12
##        288          0        288        288        288        288
## 2012-10-13 2012-10-14 2012-10-15 2012-10-16 2012-10-17 2012-10-18
##        288        288        288        288        288        288
## 2012-10-19 2012-10-20 2012-10-21 2012-10-22 2012-10-23 2012-10-24
##        288        288        288        288        288        288
## 2012-10-25 2012-10-26 2012-10-27 2012-10-28 2012-10-29 2012-10-30
##        288        288        288        288        288        288
## 2012-10-31 2012-11-01 2012-11-02 2012-11-03 2012-11-04 2012-11-05
##        288          0        288        288          0        288
## 2012-11-06 2012-11-07 2012-11-08 2012-11-09 2012-11-10 2012-11-11
##        288        288        288          0          0        288
## 2012-11-12 2012-11-13 2012-11-14 2012-11-15 2012-11-16 2012-11-17
##        288        288          0        288        288        288
## 2012-11-18 2012-11-19 2012-11-20 2012-11-21 2012-11-22 2012-11-23
##        288        288        288        288        288        288
## 2012-11-24 2012-11-25 2012-11-26 2012-11-27 2012-11-28 2012-11-29
##        288        288        288        288        288        288
## 2012-11-30
##          0
```

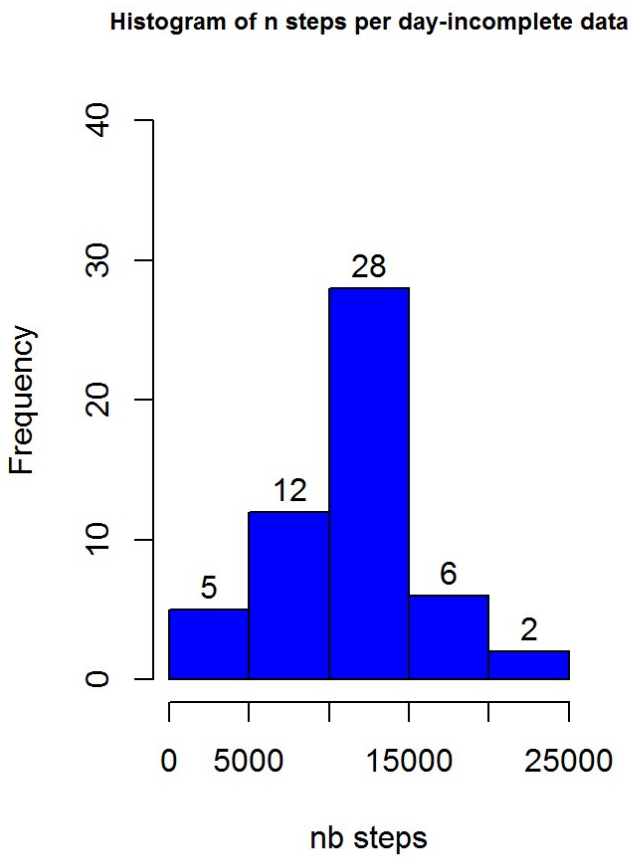
So we we calculate the averaged values of steps for each 5-minutes interval (averaged accross all days). It gives a vector with 288 values which we will report in each day without steps value. The new dataset with imput data will be activity.imput.

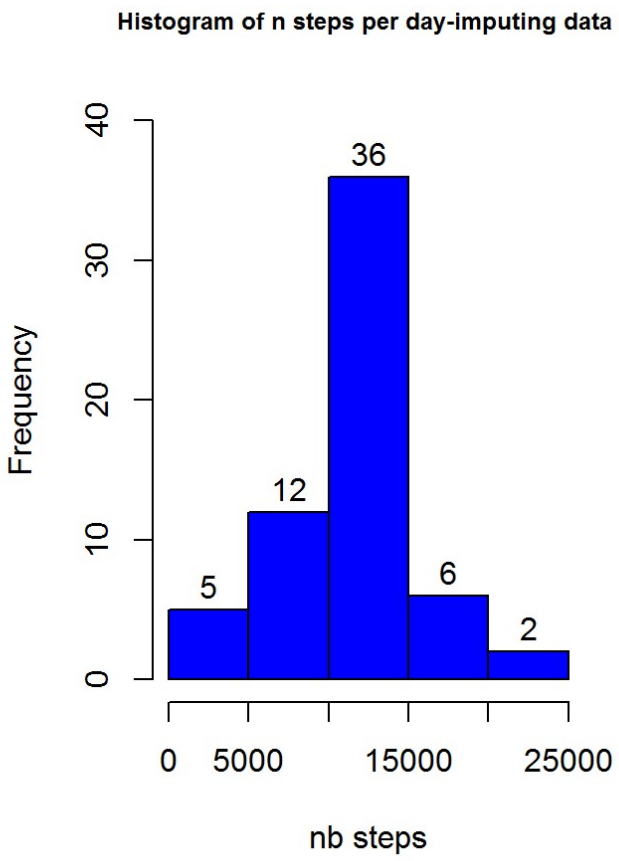
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
nb.NA.ByDate<-as.data.frame(table(activity$date[ok]))
colnames(nb.NA.ByDate)<-c("date","nb.NA")
nb.date<-nrow(nb.NA.ByDate)
activity.impute<-activity
for (i in 1:nb.date){
  if(nb.NA.ByDate[i,2] == 0){
    activity.impute[which(activity.impute$date==nb.NA.ByDate[i,1]),1]=meanByInterva
l.steps$steps
  }
}
```

4. 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.

```
nb.ByDate.steps.impute<-aggregate(steps~date,data=activity.impute,sum)
par(mfrow=c(1,2))
hist(nb.ByDate.steps$steps,xlab="nb steps",main="Histogram of n steps per day-incom
plete data",ylim=c(0,40),cex.main=0.75,col="blue",labels=T)
hist(nb.ByDate.steps.impute$steps,xlab="nb steps",main="Histogram of n steps per da
y-imputing data ",ylim=c(0,40),cex.main=0.75,col="blue",labels=T)
```






```
mean(nb.ByDate.steps$steps)
```

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

```
mean(nb.ByDate.steps.impute$steps)
```

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

```
median(nb.ByDate.steps$steps)
```

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

```
median(nb.ByDate.steps.impute$steps)
```

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

We can notice that means are the same and the medians are very closed.

Are there differences in activity patterns between weekdays and weekends?

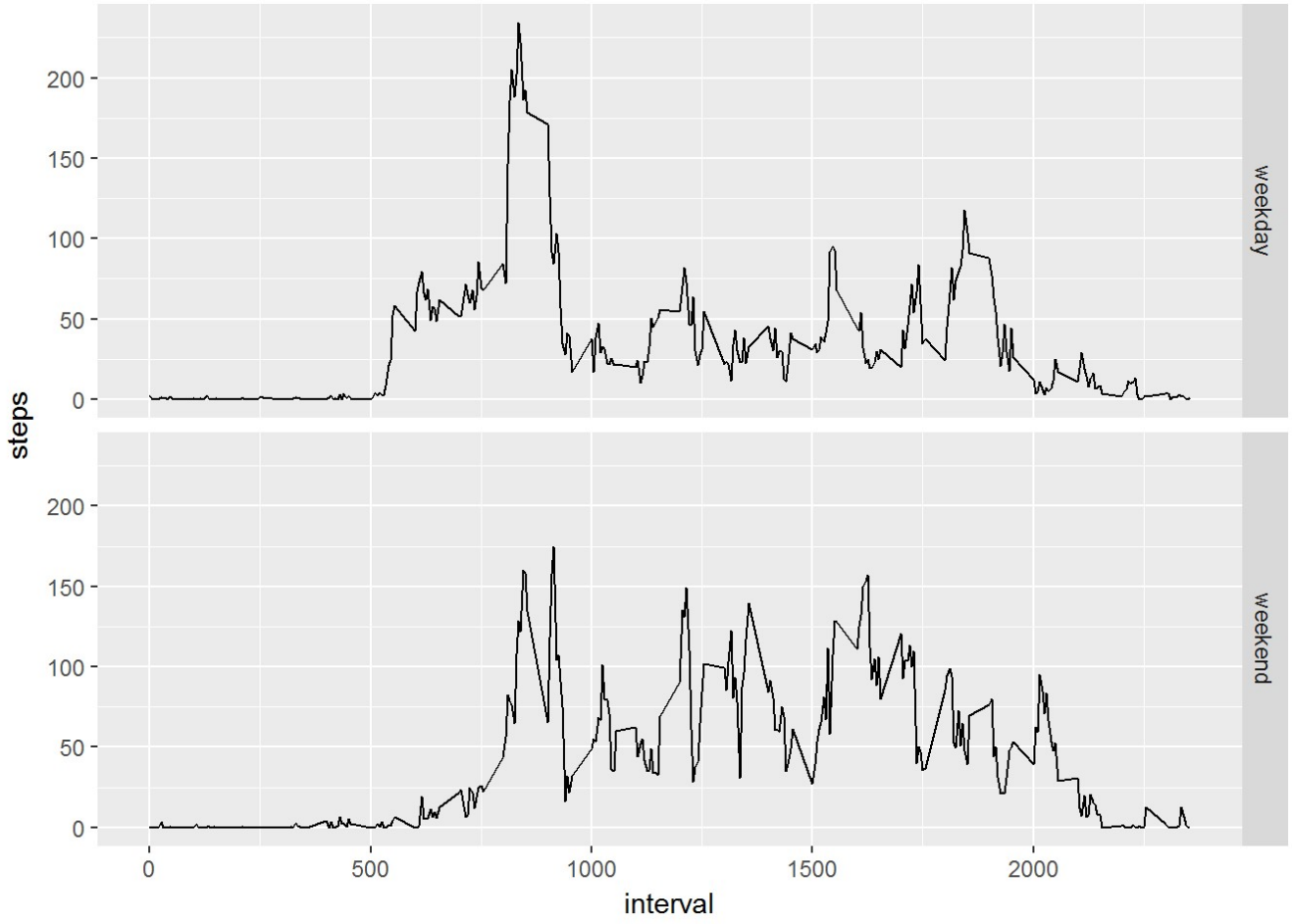
1. Create a new factor variable in the dataset with two levels - “weekday” and “weekend” indicating

whether a given date is a weekday or weekend day.

```
activity.complete<-mutate(activity.complete,date=ymd(date))
activity.complete<-mutate(activity.complete,weekday="weekday")

activity.complete[which(weekdays(activity.complete$date) %in% c("samedi","dimanche"
)),4]<-"weekend"
activity.complete<-mutate(activity.complete,weekday=as.factor(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)

```
nb.ByDate.steps<-aggregate(steps~interval+weekday,data=activity.complete,mean)
sp<-qplot(interval,steps,data=nb.ByDate.steps,geom="line")
sp + facet_grid(weekday ~ .)
```




```
nb.ByDate.steps.filter<-mutate(nb.ByDate.steps,index=interval %in% 900:2000)
nb.ByDate.steps.filter<-filter(nb.ByDate.steps.filter,index == TRUE)

mean.day<-aggregate(steps~weekday,data=nb.ByDate.steps.filter,mean)
```

For week days, we notice an high level at 08:35, after the average steps level (5 min interval) is 45.1696549 during the day (09:00 - 20:00 period). In the week-end, the sum of steps is more spread during the day. The average level between 09:00 and 20:00 is 74.5725027 the week end.