Course project Reproducible research

Julian Chaves

30/5/2020

Reproducible Research Assignment 1

Prior the execution of R Script, is required to load the following libraries

```r
library(ggplot2)
library(dplyr)
library(lubridate)
```

Loading and preprocessing the data

In this process missing valuesare ignored.

1. Load the data

```r
unzip(zipfile="activity.zip")
data<-read.csv(file="activity.csv", header=TRUE, sep=",")
```

2. Visualize data (It must look like this)

```r
head(data)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

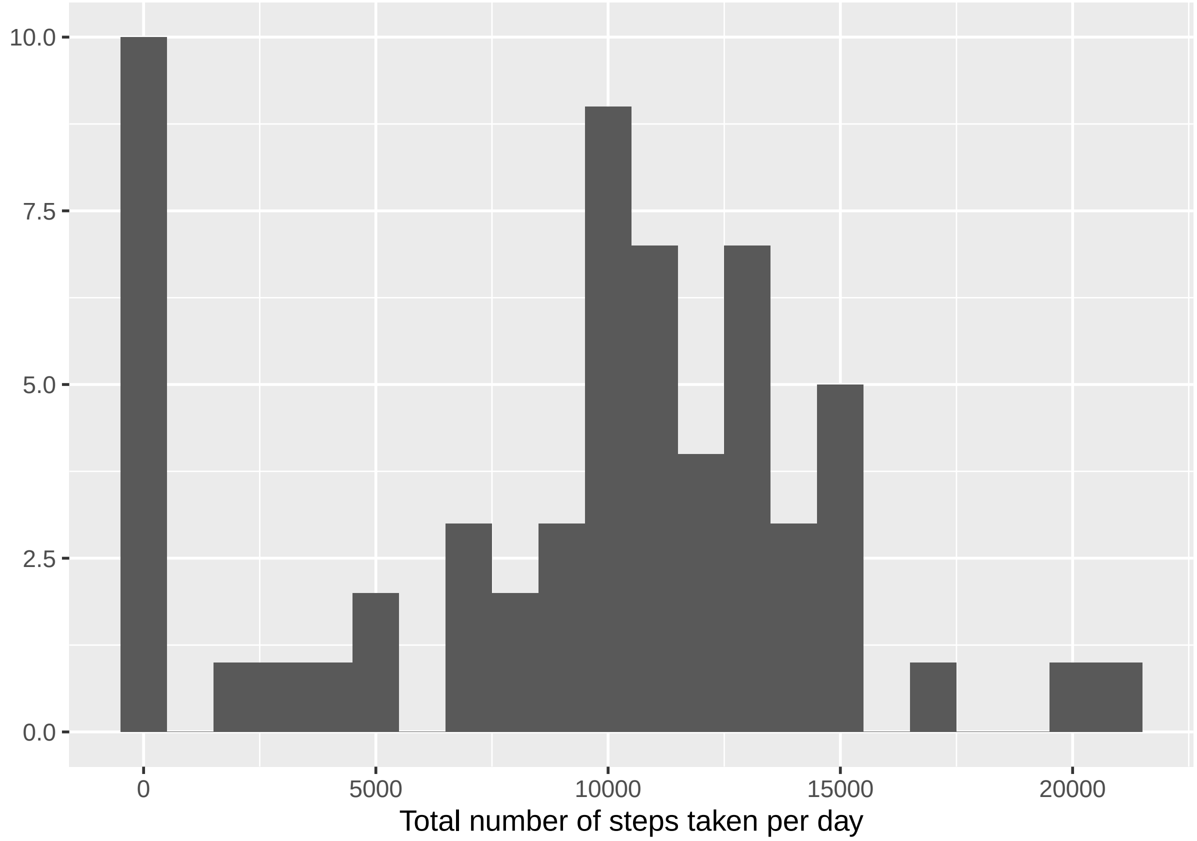
What is mean total number of steps taken per day?

To answer this question, let's calculate a histogram of the total number of steps taken each day using dplyr package:

```r
## calculating total steps
t_steps <- tapply(data$steps, data$date, FUN=sum, na.rm=TRUE)
```

now using Q Plot we make an histogram:

```
qplot(t_steps, binwidth=1000, xlab="Total number of steps taken per day")
```



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

```
averages <- aggregate(x=list(steps=data$steps), by=list(interval=data$interval),
                      FUN=mean, na.rm=TRUE)
## total steps mean:
mean(t_steps, na.rm=TRUE)
```

```
## [1] 9354.23
```

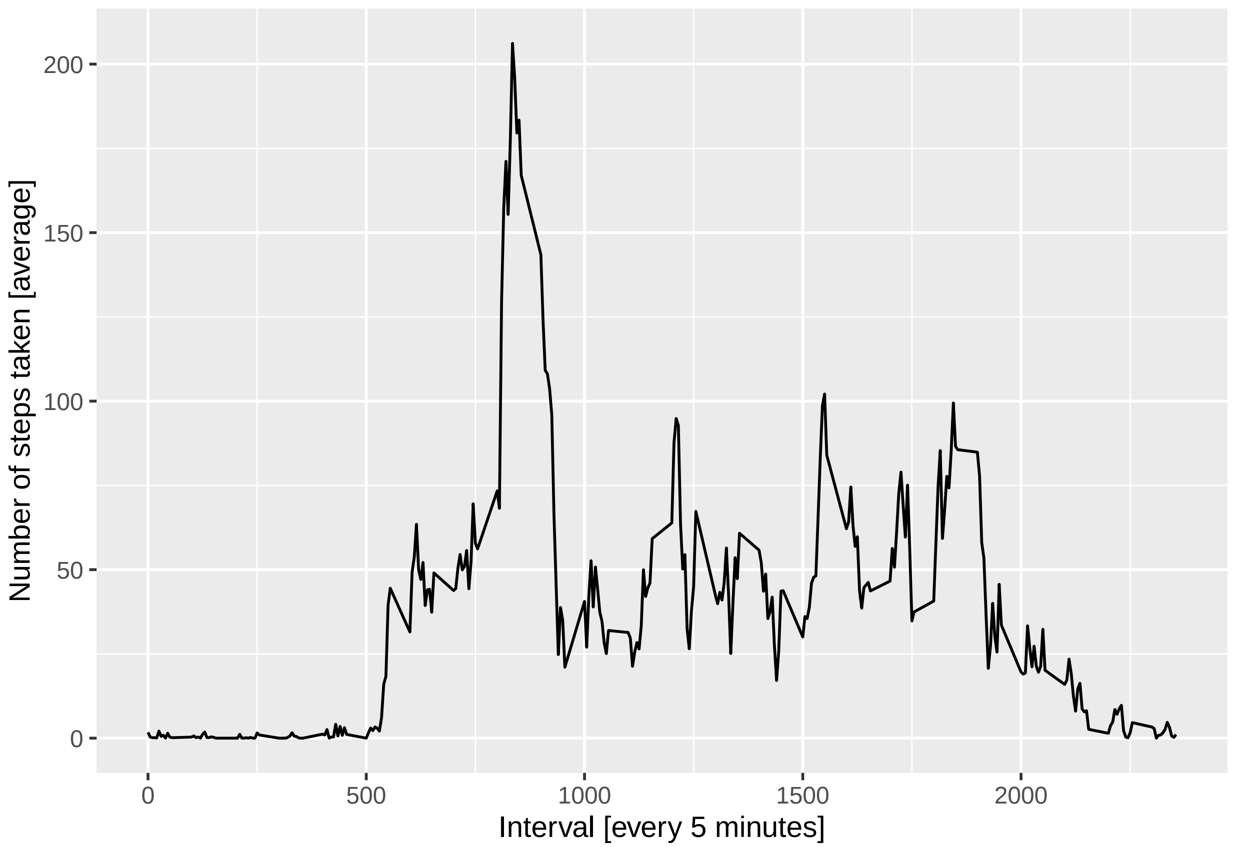
```
## total steps median:
median(t_steps, na.rm=TRUE)
```

```
## [1] 10395
```

Despite mean=**10,766** is close the median=**10,765** , the histogram is showing that the majority of the data points are in the interval [**10,000-15,000**]. Then distribution is almost uniform between 10,000 and 15,000.

What is the average daily activity pattern?

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



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

It seems that day interval=**835** is the larger interval of the day with **206** steps, if we compare it with the mean of **37** steps, it about **22.3** times larger

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as `NANA`). The presence of missing days may introduce bias into some calculations or summaries of the data.

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

```
missing<-sum(is.na(data$steps))
missing
```

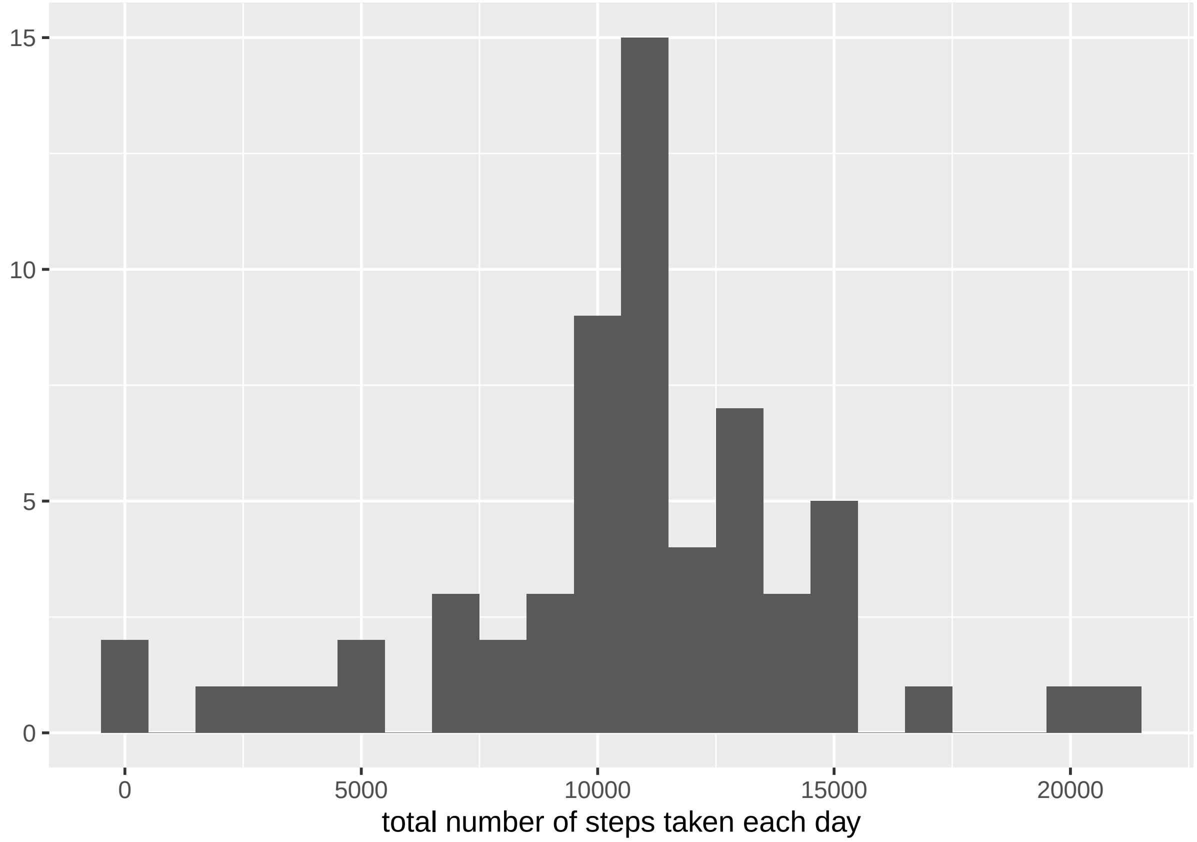
```
## [1] 2304
```

The dataframe contains **2,304** missing values,so let's fill in all of the missing values in the dataset and Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
# Replace each missing value with the mean calculated above
fill <- function(steps, interval) {
        filled <- NA
        if (!is.na(steps))
                filled <- c(steps)
        else
                filled <- (averages[averages$interval==interval, "steps"])
        return(filled)
}
completed.data <- data
completed.data$steps <- mapply(fill, completed.data$steps, completed.data$interval)
```

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?

```
t_steps <- tapply(completed.data$steps, completed.data$date, FUN=sum)
qplot(t_steps, binwidth=1000, xlab="total number of steps taken each day")
```



```
mean(t_steps)
```

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

```
median(t_steps)
```

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

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

```
# Filled data mean
mean(t_steps)
```

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

```
# Filed data median
median(t_steps)
```

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

The difference is with filled data, the higher frequency bar for the histogram is not the one at 10,000 steps but the following one.

Are there differences in activity patterns between weekdays and weekends?

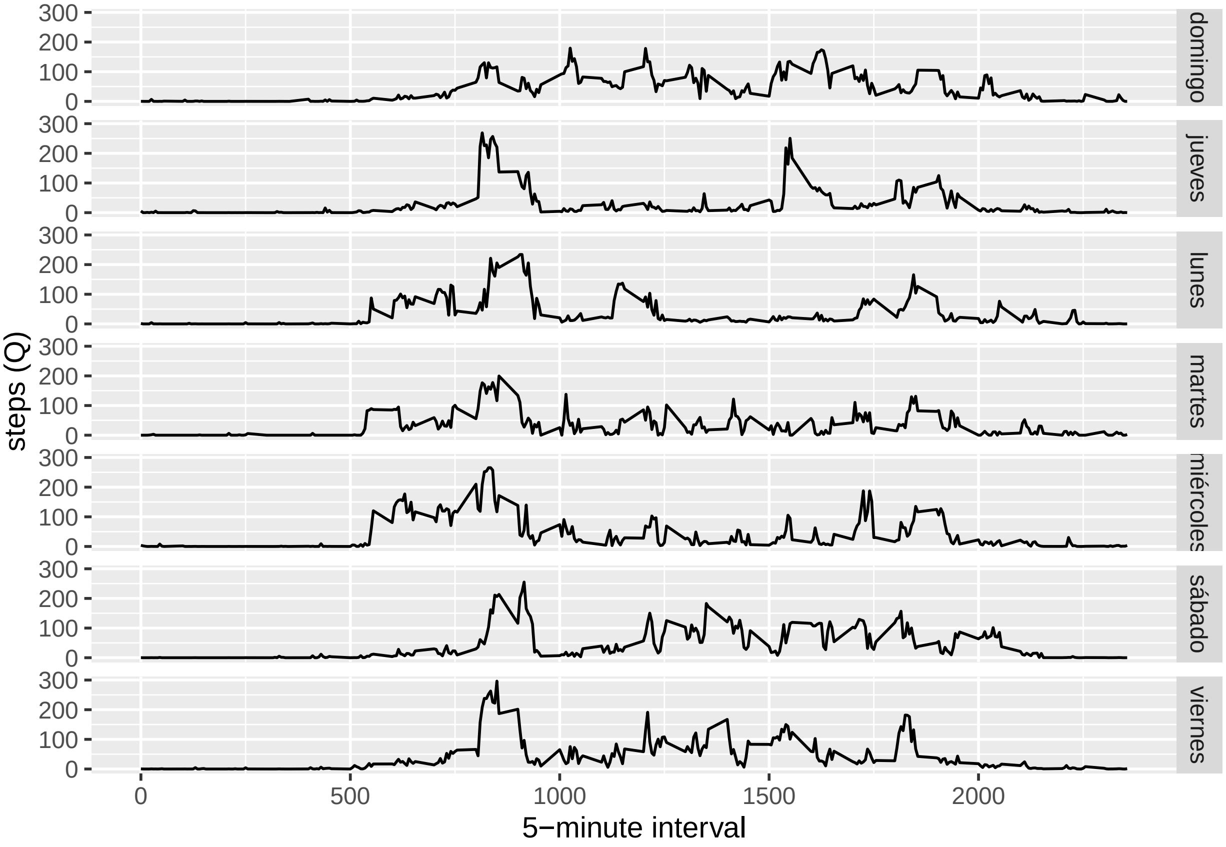
For this part the `weekdays()`weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

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.

```
completed.data$date<-as.Date(completed.data$date)
completed.data<-mutate(completed.data, weekdayType=0)
# as a good Latino, my system time is set in spanish, so be carefull to change weekend days when implem
completed.data$weekdayType <- ifelse(weekdays(completed.data$date) %in% c("sabado", "domingo"),
                                     "weekend", "weekday")
```

Make a panel plot containing a time series plot (i.e. `type = "l"`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).

```
day <- weekdays(completed.data$date)
averages <- aggregate(steps ~ interval + day, data=completed.data, mean)
ggplot(averages, aes(interval, steps)) + geom_line() + facet_grid(day ~ .) +
        xlab("5-minute interval") + ylab("steps (Q)")
```



From data we are able to see that differences between weekdays and weekends (included friday) are: 1. People tend to wake up later. 2. People use to go bed later. 3. Steps increase for lunchtime in weekends. 4. Sunday reveals lesser activity than others days. 5. Thursday sugests people focus more in work or study and limit their movements to go out home and come back.