

# Reproducible Research: Peer Assessment 1

author: "Pedro Rebelo"

date: "7 Feb 2016"

github repo with RMarkdown source code: [https://github.com/trashmanp1/RepData\\_PeerAssessment1](https://github.com/trashmanp1/RepData_PeerAssessment1)  
([https://github.com/trashmanp1/RepData\\_PeerAssessment1](https://github.com/trashmanp1/RepData_PeerAssessment1))

output: html\_document —

This is project 1 of Reproducible Research course by Coursera (<https://www.coursera.org/learn/reproducible-research> (<https://www.coursera.org/learn/reproducible-research>)) The objective of this work is to make a Literate Statistical Program that loads a data set, make some statistical analysis in a human and computer readable form. The data set can be obtained here: <https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip> (<https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip>)

## Loading and preprocessing the data

Loading the data is easy, just set the work directory to the directory where data is, unzip the data and load it

```
setwd("/Users/trashman/Documents/RepData_PeerAssessment1")#path to directory where data is
unzip("activity.zip")
dados <- read.csv("activity.csv")
```

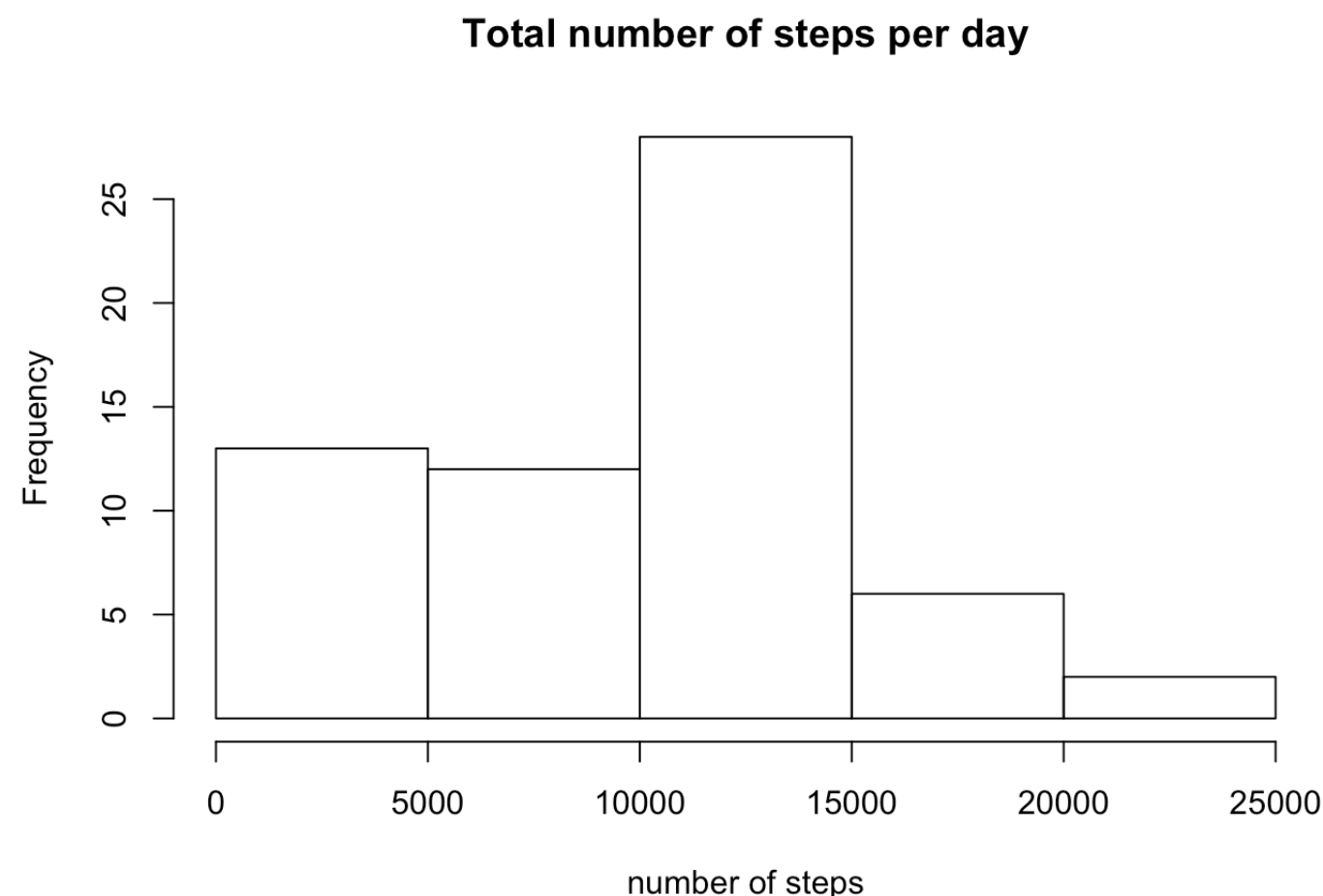
## What is mean total number of steps taken per day?

Because in a day there are 288 periods of 5 minutes, we must aggregate the result by day, before take the mean and median:

```
passos_por_dia <- aggregate(dados[, "steps"], by=as.list(dados[ "date" ]), FUN=sum, na.rm=TRUE)
names(passos_por_dia) <- c("date", "steps")
media<-round(mean(passos_por_dia$steps)) #integer number of steps
mediana<-median(passos_por_dia$steps)
```

The general distribution is given by the hystogram:

```
hist(passos_por_dia$steps, breaks=round(log2(dim(passos_por_dia)[1])+1), #Sturges rule
      xlab="number of steps",
      main="Total number of steps per day")
```



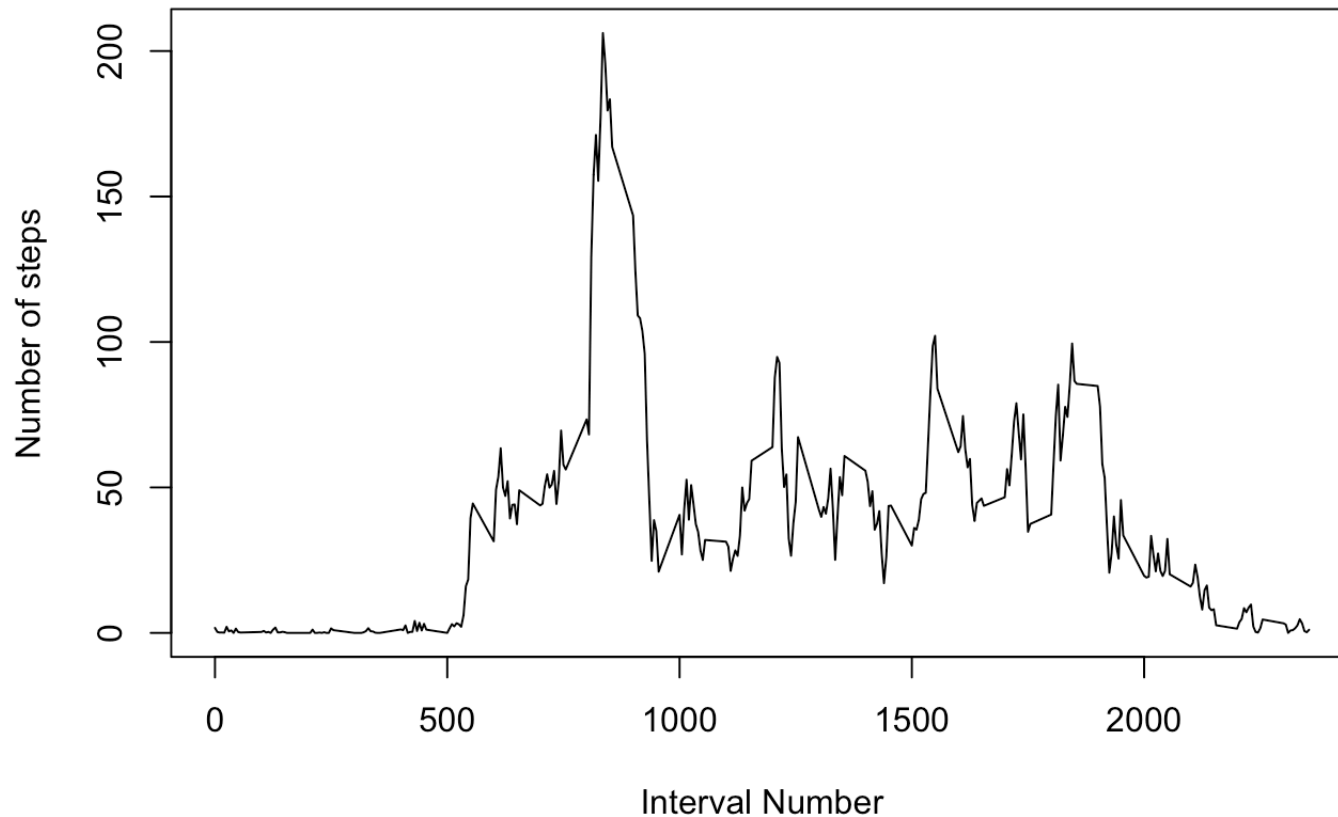
so, the mean is 9354 steps per day, and the median is 10395 steps per day.

## What is the average daily activity pattern?

We can get the number of steps for the 5 minute intervals averaged over all days

```
passos_por_5_min <- aggregate(dados[, "steps"], by = as.list(dados["interval"]),
                             FUN = mean, na.rm=TRUE)
plot(passos_por_5_min, type = "l",
     xlab="Interval Number",
     ylab="Number of steps",
     main="Daily 5 min average steps")
```

### Daily 5 min average steps



The 5-minute interval, on average across all the days in the dataset that contains the

```
maximo<-passos_por_5_min$interval[which.max(passos_por_5_min$x)]
```

maximum number of steps occurs at 835 interval.

## Imputing missing values

```
n_falhas<-sum(is.na(dados$steps))
tamanho<-length(dados$steps)
percentagem<-n_falhas/tamanho*100
```

The total sum of missing values is 2304 that corresponds 13.1147541 % of the total data points.

If we choose the mean to fill the points without registration, we do not change the overall mean, but we will decrease the variance of the overall data. On the other hand if we use the median it will not only decrease the variance but also change the overall mean. If we use zero for the missing values we increase the variance and decrease the overall mean. Because the median of all 5 min interval is zero, I chose zero. So I will increase the dispersion of data and decrease the mean.

```
dados_martelados<-dados
dados_martelados$steps[is.na(dados$steps)]<-0
summary(dados$steps)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.00   0.00   0.00   37.38  12.00   806.00   2304
```

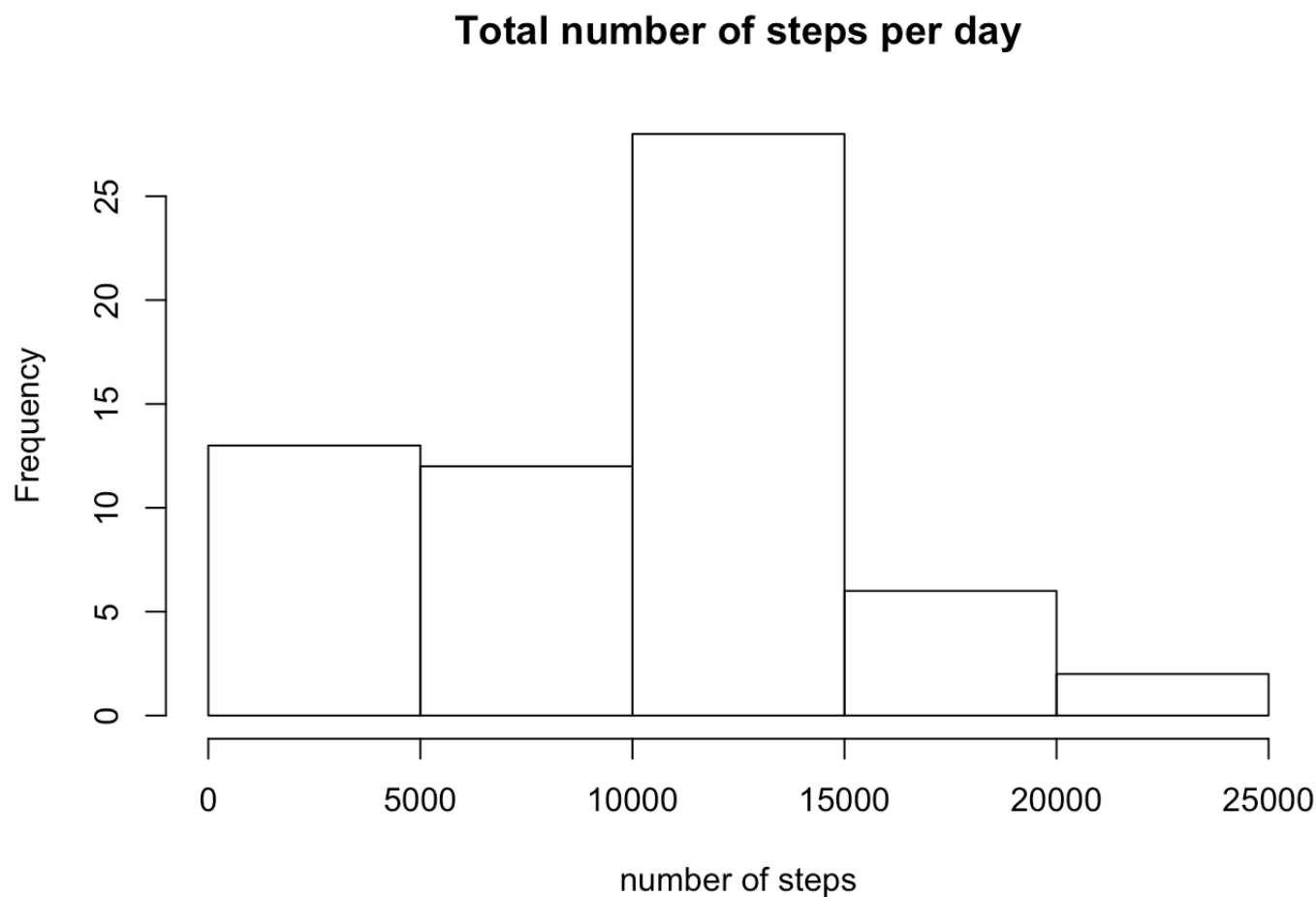
```
summary(dados_martelados$steps)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   0.00   0.00   32.48   0.00   806.00
```

The histogram of the new data set is:

```
n_passos_por_dia <- aggregate(dados_martelados[, "steps"], by=as.list(dados_martelados[ "date" ]), FUN=sum, na.rm=TRUE)
names(n_passos_por_dia) <- c("date", "steps")

hist(passos_por_dia$steps, xlab="number of steps",
      main="Total number of steps per day")
```



```
n_media<-round(mean(n_passos_por_dia$steps))#integer number of steps
n_mediana<-round(median(n_passos_por_dia$steps))#integer number of steps
```

and the new mean is 9354 steps per day, and the new median is  $1.0395 \times 10^4$  steps per day. So the results are equal to the previous, because we do not change the sum of the total number of steps by day.

## Are there differences in activity patterns between weekdays and weekends?

Based on the next figure, we see that the beginning of the day on a weekend have less activity then a weekday but have more activity along the day.

```
#create a new factor from date
wd <- function(date) {
  if (weekdays(as.Date(date)) %in% c("Saturday", "Sunday")) {
    "weekend"
  } else {
    "weekday"
  }
}
dados$wd <- as.factor(sapply(dados$date, wd))

#calculate the 5 min mean by wd
steps_wd<-aggregate(dados[, "steps"], by=as.list(c(dados[ "wd" ], dados[ "interval" ])), FUN=mean, na.rm=TRUE)

#create a ggplot
library(ggplot2)
graph <- ggplot(steps_wd, aes(interval, x)) + theme_bw()
graph <- graph + geom_line() + facet_wrap(~ wd, nrow = 2)
graph <- graph + xlab("Interval number")
graph <- graph + ylab("Number of steps")
graph
```

