THIS IS THE FIRST ASSIGMENT.

- title: "assigment1" author: "christian uribe" date: "Friday, September 25, 2015" output: html_document --- ## About *This was the first project for the Reproducible Research course in Coursera's Data Science specialization track. The purpose of the project was to answer a series of questions using data collected from a FitBit.

Synopsis

The purpose of this project was to practice:

-loading and preprocessing data -imputing missing values -interpreting data to answer research questions

Data

*The data for this assignment was downloaded from the course web site:

Dataset: Activity monitoring data [52K] The variables included in this dataset are:

steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)

date: The date on which the measurement was taken in YYYY-MM-DD format

interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

Loading and preprocessing the data

Download, unzip and load data into data frame <- datos. Is more simple if you download and unzip without code, then when you have the excel, you only need to put this:

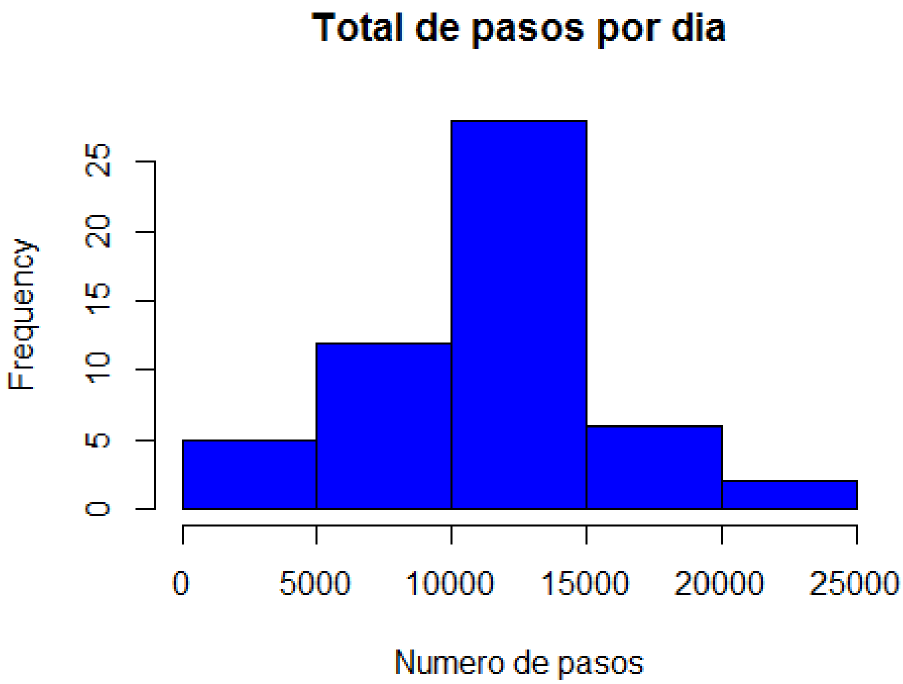
```
datos <- read.csv("activity.csv")
summary(datos)
```

```
##      steps                date          interval
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8
##  Median :  0.00   2012-10-03:  288   Median :1177.5
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5
##  3rd Qu.: 12.00   2012-10-05:  288   3rd Qu.:1766.2
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0
##  NA's   :2304     (Other)   :15840
```

What is mean total number of steps taken per day?

Sum steps by day, create Histogram, and calculate mean and median.

```
pasos_diarios <- aggregate(steps~date, datos, sum)
hist(pasos_diarios$steps, main=paste("Total de pasos por dia"),
col="blue", xlab="Numero de pasos")
```



```
media <- mean(pasos_diarios$steps)
mediana <- median(pasos_diarios$steps)
```

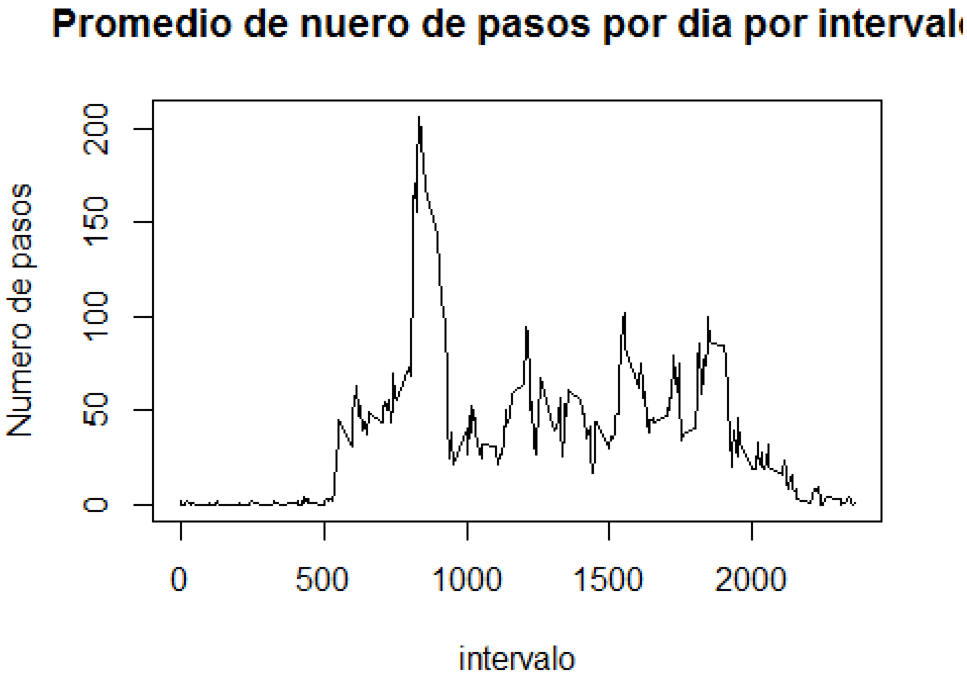
The mean is r media and the median is r mediana.

What is the average daily activity pattern?

-Calculate average steps for each interval for all days. -Plot the Average Number Steps per Day by Interval. -Find interval with most average steps.

```
pasos_intervalo <- aggregate(steps~interval, datos, mean)

plot(pasos_intervalo$interval,pasos_intervalo$steps, type="l",
xlab="intervalo", ylab="Numero de pasos", main="Promedio de nuero de
pasos por dia por intervalo")
```



```
maximo_intervalo <- pasos_intervalo[which.max(pasos_intervalo$steps),1]
```

The 5-minute interval, on average across all the days in the data set, containing the maximum number of steps is r maximo_intervalo.

Impute missing values. Compare imputed to non-imputed data.

Missing data needed to be imputed. Only a simple imputation approach was required for this assignment. Missing values were imputed by inserting the average for each interval. Thus, if interval 10 was missing on 10-02-2012, the average for that interval for all days (0.1320755), replaced the NA.

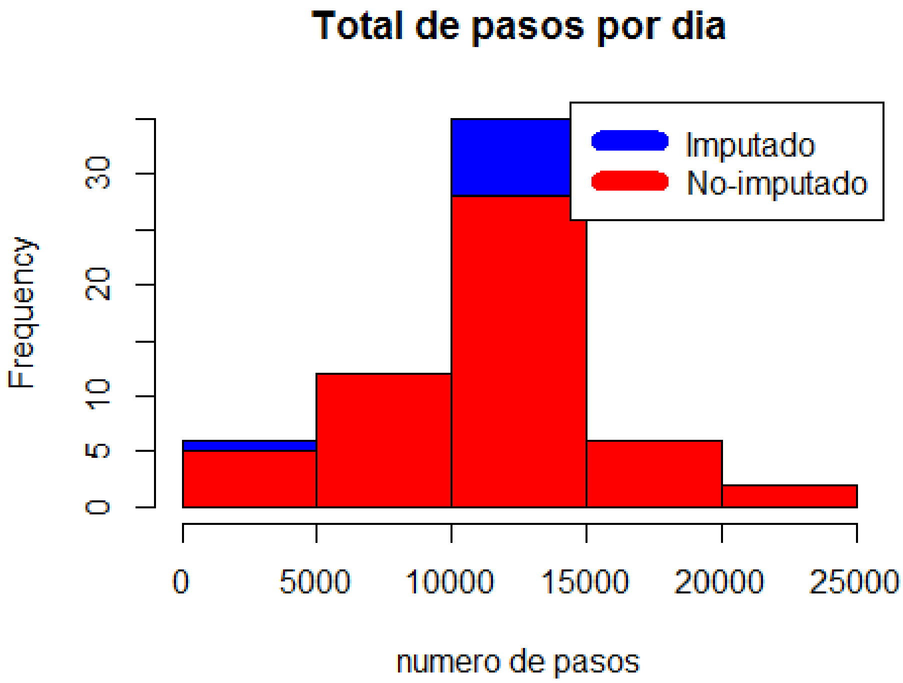
```
incompletos <- sum(!complete.cases(datos))
datos_imputados <- transform(datos, steps = ifelse(is.na(datos$steps),
pasos_intervalo$steps[match(datos$interval, pasos_intervalo$interval)],
datos$steps))
```

Zeroes were imputed for 10-01-2012 because it was the first day and would have been over 9,000 steps higher than the following day, which had only 126 steps. NAs then were assumed to be zeros to fit the rising trend of the data.

```
datos_imputados[as.character(datos_imputados$date) == "2012-10-01", 1] <-
0
```

Recount total steps by day and create Histogram.

```
pasos_diarios_i <- aggregate(steps~date, datos_imputados, sum)
hist(pasos_diarios_i$steps, main=paste("Total de pasos por dia"),
col="blue", xlab="numero de pasos")
#Crear un histograma para mostrar la diferencia
hist(pasos_diarios$steps, main=paste("Total de pasos cada dia"),
col="red", xlab="Numero de pasos", add=T)
legend("topright", c("Imputado", "No-imputado"), col=c("blue", "red"),
lwd=10)
```



Calculate new mean and median for imputed data.

```
media_i <- mean(pasos_diarios_i$steps)
mediana_i <- median(pasos_diarios_i$steps)
```

Calculate difference between imputed and non-imputed data.

```
diferencia_media <- media_i - media
diferencia_mediana <- mediana_i - mediana
```

Calculate total difference.

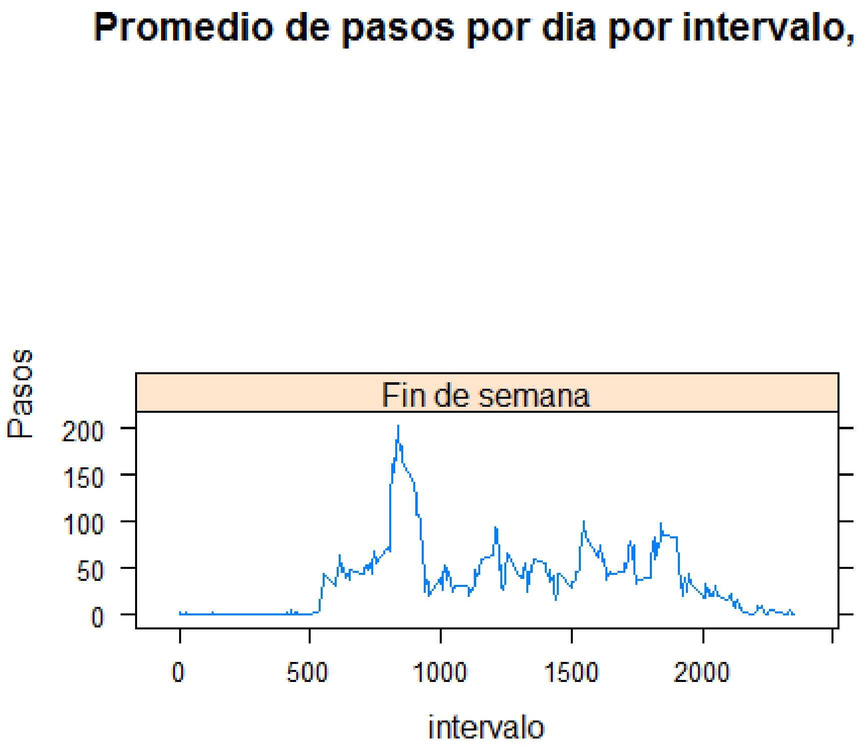
```
diferencia_total <- sum(pasos_diarios_i$steps) - sum(pasos_diarios$steps)
```

The imputed data mean is media_i The imputed data median is mediana_i The difference between the non-imputed mean and imputed mean is r diferencia_media The difference between the non-imputed mean and imputed mean is r diferencia_medias The difference between total number of steps between imputed and non-imputed data is r diferencia total, there were r diferencia_total more steps in the imputed data.

Are there differences in activity patterns between weekdays and weekends?

Created a plot to compare and contrast number of steps between the week and weekend. There is a higher peak earlier on weekdays, and more overall activity on weekends.

```r
weekdays <- c("Monday", "Tuesday", "Wednesday", "Thursday","Friday")
datos_imputados$dow =
as.factor(ifelse(is.element(weekdays(as.Date(datos_imputados$date)),
weekdays),"Dias entre semana", "Fin de semana"))
pasos_intervalo_i <- aggregate(steps ~ interval + dow, datos_imputados,
mean)
library(lattice)
xyplot(pasos_intervalo_i$steps ~ pasos_intervalo_i$interval |
pasos_intervalo_i$dow, main="Promedio de pasos por dia por intervalo,",
xlab="intervalo", ylab="Pasos", layout= c(1,2), type="l")
```


Promedio de pasos por dia por intervalo,
Pasos
Fin de semana
intervalo