Análisis Exploratorio

## 1. Introducción con el objetivo del análisis:

El primer objetivo es clasificar a los alumnos como aprobado o suspenso dependiendo de su nota final A partir de ello voy a ver si influyen en esta nueva variable y en la nota meda final, la edad y si tienen una relación de pareja

Para ello voy a usar los datos de los alumnos de matemáticas, para obtener estos datos de los alumnos de portugués seguiría el mismo proceso ya que la estructura de los ficheros es la misma (También se podría hacer con la combinación de ambos ficheros)

## 2. Carga de los datos

Creo un directorio y descargo en el los datos que voy a utilizar

#Instalo y cargo las librerias que voy a utilizar  
libs <- c("ggplot2","knitr", "dplyr")  
   
for (i in libs){  
 print(i)  
 if(!require(i, character.only = TRUE)) { install.packages(i, character.only = TRUE); }  
 library(i, character.only = TRUE)  
}

## [1] "ggplot2"  
## [1] "knitr"  
## [1] "dplyr"

currentDir <- getwd()  
  
if(!file.exists("./datos"))  
 dir.create("./datos")  
  
#Descargo el fichero con los datos   
fileUrl <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip"  
download.file(fileUrl,destfile="./datos/student.zip")  
unzip("./datos/student.zip", exdir="./datos")  
list.files("./datos")

## [1] "student-mat.csv" "student-merge.R" "student-por.csv" "student.txt"   
## [5] "student.zip"

rm(fileUrl)  
  
fechaDescarga <- date()

Creo el dataset

student.mat <- read.csv2("./datos/student-mat.csv",row.names=NULL, sep=";", header=TRUE)

## 3. Análisis descriptivo

Analizamos el nombre de cada una de las variables y las pasamos a minúsculas para que sean más fáciles de manejar

names(student.mat) <- tolower(names(student.mat))

Muestro las primeras lineas del dataset para ver que datos tengo

head(student.mat[,1:5])

## school sex age address famsize  
## 1 GP F 18 U GT3  
## 2 GP F 17 U GT3  
## 3 GP F 15 U LE3  
## 4 GP F 15 U GT3  
## 5 GP F 16 U GT3  
## 6 GP M 16 U LE3

Muestro las dimensiones del dataset y un resumen de los datos que hay en el

dim(student.mat)

## [1] 395 33

summary(student.mat)

str(student.mat)

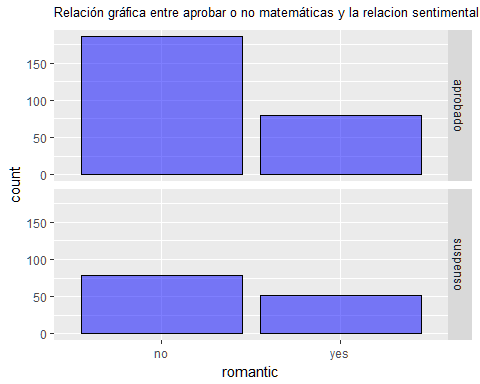
## 4. Análisis exploratorio

Para hacer análisis exploratorio solo voy a coger las variables age, romantic y g3 que son las que necesitare para este análisis

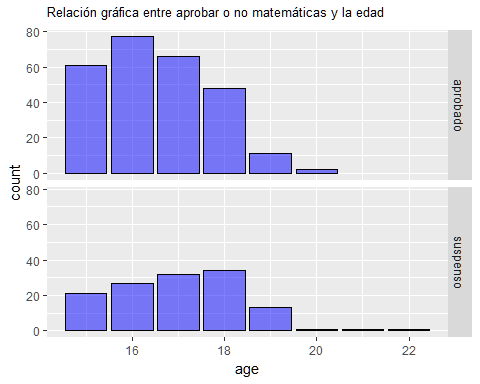
student.mat.exploratorio <- student.mat %>% select(age, romantic, g3)  
  
#añado una variable de aprobado o suspenso dependendo del valor de g3  
aprobados <- ifelse(student.mat.exploratorio$g3>=10,'aprobado','suspenso')  
student.mat.exploratorio <- data.frame(student.mat.exploratorio,aprobados)  
summary(student.mat.exploratorio)

## age romantic g3 aprobados   
## Min. :15.0 no :263 Min. : 0.00 aprobado:265   
## 1st Qu.:16.0 yes:132 1st Qu.: 8.00 suspenso:130   
## Median :17.0 Median :11.00   
## Mean :16.7 Mean :10.42   
## 3rd Qu.:18.0 3rd Qu.:14.00   
## Max. :22.0 Max. :20.00

#Hago un gráfico que nos saca la relación que hay entre los aprobados y los suspensos dependiendo de si tienen o no una relación sentimental  
ggplot(student.mat.exploratorio, aes(x = romantic)) + geom\_bar(colour="black", fill="blue", alpha=.5, stat="count") +  
facet\_grid(aprobados ~ .) +  
ggtitle ("Relación gráfica entre aprobar o no matemáticas y la relacion sentimental") +  
theme(plot.title=element\_text(vjust = +1.5, size = 10))



#Hago lo mismo con la edad  
ggplot(student.mat.exploratorio, aes(x = age)) + geom\_bar(colour="black", fill="blue", alpha=.5, stat="count") +  
facet\_grid(aprobados ~ .) +  
ggtitle ("Relación gráfica entre aprobar o no matemáticas y la edad") +  
theme(plot.title=element\_text(vjust = +1.5, size = 10))



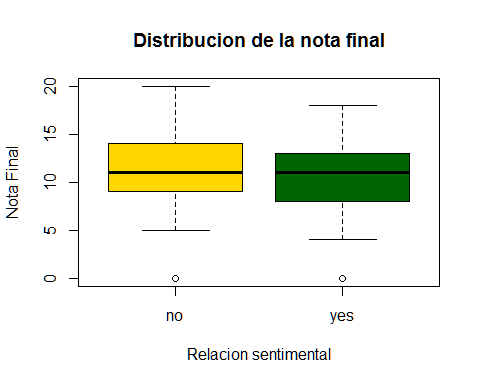
#Comparo la media de las notas finales  
aggregate(student.mat.exploratorio["g3"], student.mat.exploratorio["romantic"], mean)

## romantic g3  
## 1 no 10.836502  
## 2 yes 9.575758

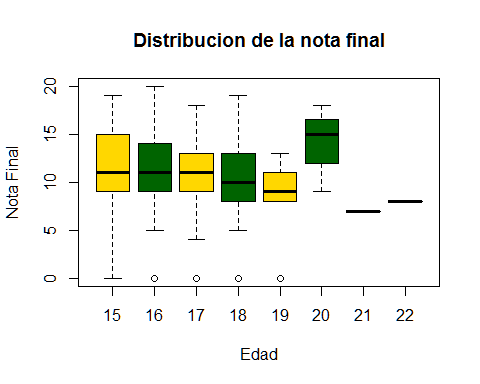
aggregate(student.mat.exploratorio["g3"], student.mat.exploratorio["age"], mean)

## age g3  
## 1 15 11.256098  
## 2 16 11.028846  
## 3 17 10.275510  
## 4 18 9.548780  
## 5 19 8.208333  
## 6 20 14.000000  
## 7 21 7.000000  
## 8 22 8.000000

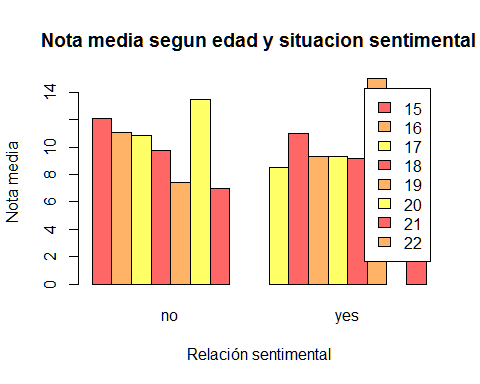
#Hago un gráfico que nos saca la relación que hay entre las notas medias dependiendo de si tienen o no una relación sentimental  
boxplot(g3~romantic,data=student.mat.exploratorio, main="Distribucion de la nota final", col=(c("gold","darkgreen")),  
 xlab="Relacion sentimental", ylab="Nota Final")



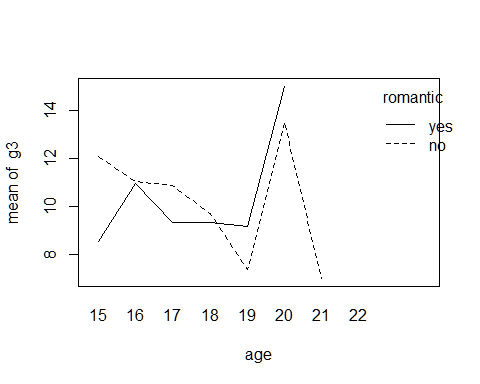
#Hago lo mismo con la edad  
boxplot(g3~age,data=student.mat.exploratorio, main="Distribucion de la nota final", col=(c("gold","darkgreen")),  
 xlab="Edad", ylab="Nota Final")



#Convierto la variable aprobados a numérica  
student.mat.exploratorio$aprobados <- as.numeric(student.mat.exploratorio$aprobados)  
  
#Hasta aquí he visto gráficos que me relacionan la nota final con una de las otras variables, ahora voy a ver como lo hace con las dos a la vez  
#Calculo la nota media por edad y relacion sentimental  
x <- tapply(student.mat.exploratorio$g3, list(student.mat.exploratorio$age, student.mat.exploratorio$romantic), mean, na.rm=TRUE)  
  
barplot(x,  
 legend=TRUE,   
 beside=TRUE,  
 main="Nota media segun edad y situacion sentimental",  
 ylab="Nota media",  
 xlab="Relación sentimental",  
 col=heat.colors(3,alpha=.6),  
 xpd=F)



#Este grafico de líneas me muestra gráfico las interacciones entre los dos factores y la nota final  
with(student.mat.exploratorio, interaction.plot(age, romantic, g3))



## 5. Conclusión

Observando las tablas y los gráficos que he realizado llego a siguiente conclusión.

Con las variables que yo he seleccionado no hay una relación clara con los aprobados ni con la nota media.

Al comparar la variable romantic con los aprobados y los suspensos veo que el número de aprobados sin relación sentimental si es mayor que los que no tienen relación, pero en el caso de los suspensos es muy similar. La nota media final es similar entre ambos casos, aunque un poco mayor en el caso de los que no tienen relación sentimental

Al usar la variable age veo que las notas medias van bajando según aumenta la edad, exceptuando el caso de la edad igual a 20 que aumenta significativamente. En el caso de aprobados y suspensos no hay una relación clara.

Al combinar ambas variables vemos que la media de la nota final sube y baja sin tener una tendencia clara.

Habría que añadir alguna variable más al análisis o seleccionar otras variables ya que con estas no se puede sacar ninguna conclusión clara.