Ejemplo_Analisis_Exploratorio

Isabel M. Izquierdo

6/10/2018

ANÁLISIS EXPLORATORIO EN R: Ejemplo con Student.zip dataset

En este informe, trataremos de ilustrar posibilidades de R para Análisis exploratorio de datos, sobre este dataset:

https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip,

descrito en:

https://archive.ics.uci.edu/ml/datasets/Student+Performance

No se trata de un proceso de Análisis exploratorio para una finalidad determinada, sino de contemplar un amplio abanico de ejemplos.

El target en una fase posterior de modelado, sería ver cómo contribuyen las otras variables a la nota final: G3

Este dataset de entrada contiene 2 archivos CSV, uno con datos de alumnos de portugués y otro con datos de alumnos de matemáticas, cuyos campos están descritos de este modo (en inglés):

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2 sex - student's sex (binary: "F" - female or "M" - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: "U" - urban or "R" - rural)

5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

12 guardian - student's guardian (nominal: "mother", "father" or "other")

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese: 31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets . These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

Creamos subdirectorio "datos"

El subdirectorio "datos" quedará creado en el directorio de trabajo actual, desde el que se esté ejecutando el script

```
if (!file.exists("./datos")) {
        dir.create("./datos")
}
```

Leemos los datos

Los cargamos en 2 objetos R (uno para los estudiantes de matemáticas y otro para los estudiantes de portugués)

```
fileURL <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip"
download.file(fileURL,destfile="./datos/student.zip",method="curl")
unzip("./datos/student.zip", exdir="./datos")
list.files("./datos")
```

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

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

con <- file("./datos/student-por.csv","r")

studentPor <- read.csv2(con)
close(con)
fechaDescarga <- date()
fechaDescarga
```

[1] "Sat Oct 6 17:51:10 2018"

Summarising

Algunos ejemplos de summarising

```
library(knitr)
# Para generar un doc bien formateado desde nuestro script R markdown

kable(head(studentMat[,1:5]))
```

school	sex	age	address	famsize
GP	F	18	U	GT3
GP	F	17	U	GT3
GP	F	15	U	LE3
GP	F	15	U	GT3
GP	F	16	U	GT3
GP	M	16	U	LE3

```
# kable imprime los resultados en forma tabulada en el informe
# head muestra las primeras filas

kable(tail(studentPor[,1:5]))
```

	school	sex	age	address	famsize
644	MS	F	18	R	GT3
645	MS	F	19	R	GT3
646	MS	F	18	U	LE3
647	MS	F	18	U	GT3
648	MS	M	17	U	LE3
649	MS	M	18	R	LE3

```
# tail muestra las últimas filas

# En el head y el tail mostramos solo las 5 primeras columnas por claridad
# en el informe

library(xtable)
TableInputData <- xtable(summary(studentPor))
print(TableInputData, type = "latex")
```

% latex table generated in R 3.4.2 by xtable 1.8-2 package % Sat Oct 6 17:51:10 2018

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	F
1	GP:423	F:383	Min. :15.00	R:197	GT3:457	A: 80	Min. :0.000	Min. :0.000	at_home :135	a
2	MS:226	M:266	1st Qu.:16.00	U:452	LE3:192	T:569	1st Qu.:2.000	1st Qu.:1.000	health : 48	h
3			Median :17.00				Median :2.000	Median :2.000	other :258	o
4			Mean :16.74				Mean :2.515	Mean :2.307	services:136	s
5			3rd Qu.:18.00				3rd Qu.:4.000	3rd Qu.:3.000	teacher : 72	t
6			Max. :22.00				Max. :4.000	Max. :4.000		

```
# kable(summary(studentPor))

names(studentMat)
```

[1] “school” “sex” “age” “address” “famsize”
[6] “Pstatus” “Medu” “Fedu” “Mjob” “Fjob”
[11] “reason” “guardian” “traveltime” “studytime” “failures”
[16] “schoolsup” “famsup” “paid” “activities” “nursery”
[21] “higher” “internet” “romantic” “famrel” “freetime”
[26] “goout” “Dalc” “Walc” “health” “absences”
[31] “G1” “G2” “G3”

```
# names muestra los nombres de las variables (columnas)

# Si queremos ver frecuencias, por ejemplo cuántos estudiantes de matemáticas tienen o no Internet en c

table(studentMat$internet)
```

no yes 66 329

```
any(studentMat$G3 == 20)
```

[1] TRUE

```
all(studentMat$G3 > 0)
```

[1] FALSE

```
# Comprobamos que tal como dice la descripción del dataset, hay 382 estudiantes que aparecen en ambos
# ficheros (matemáticas y portugués)

studentMatPor <- merge(studentMat,studentPor,
                       by=c("school","sex","age",
                            "address","famsize",
                            "Pstatus","Medu",
                            "Fedu","Mjob",
                            "Fjob","reason",
                            "nursery","internet"),
                       all=FALSE,
                       suffixes=c("mat","por"))
dim(studentMatPor)[1]
```

[1] 382

La máxima nota de la evaluación final es 20 para los estudiantes de matemáticas. Y no todos los estudiantes tienen una nota mayor que 0.

No observamos NA’s.

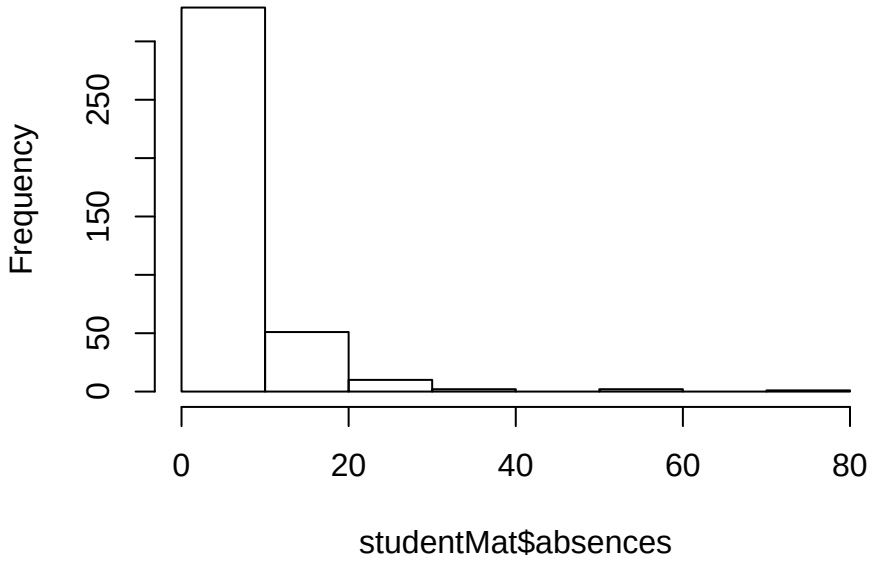
Comprobado que hay 382 estudiantes que aparecen en los dos data.frames

Analizamos la distribución de algunas variables

Primero utilizaremos Histogramas

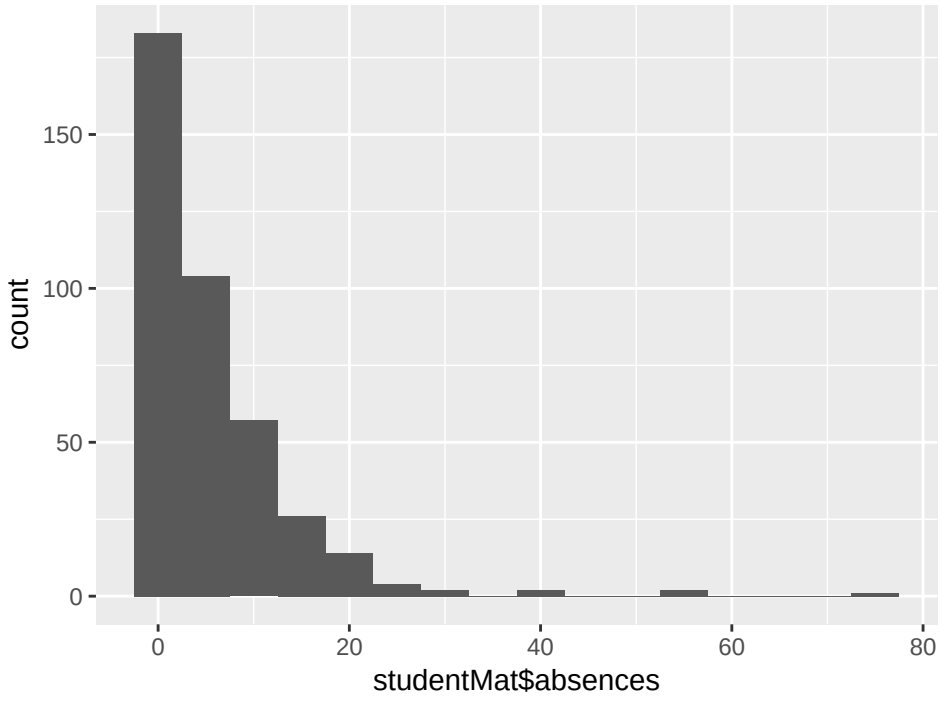
```
hist(studentMat$absences, main = "Fig 1.1: Histograma ausencias Matemáticas, plot 1", cex.main=0.7)

library(ggplot2)
```

Fig 1.1: Histograma ausencias Matemáticas, plot 1



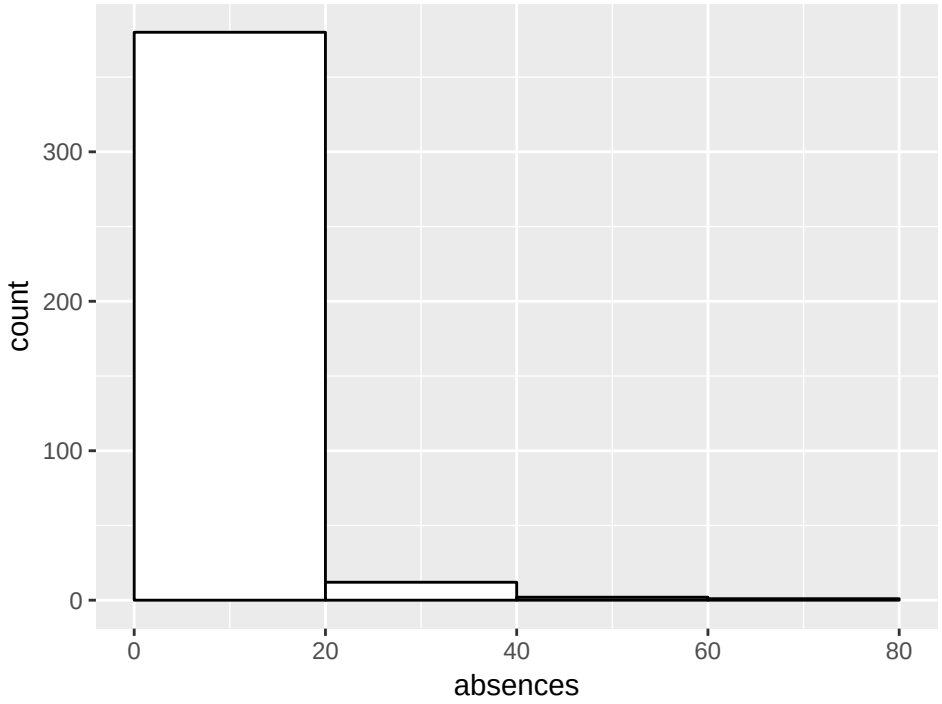

```
#plot.new()
qplot(studentMat$absences, binwidth = 5, main = "Fig 1.2: Histograma ausencias Matemáticas, plot 2")
```

Fig 1.2: Histograma ausencias Matemáticas, plot 2




```
#plot.new()
ggplot(studentMat, aes(x=absences)) +
        geom_histogram(binwidth = 20, fill = "white", colour = "black", origin = 0) +
        ggtitle ("Fig 1.1 Histograma ausencias matemáticas, plot 3") +
        theme(plot.title = element_text(vjust = +1.5, size = 12))
```

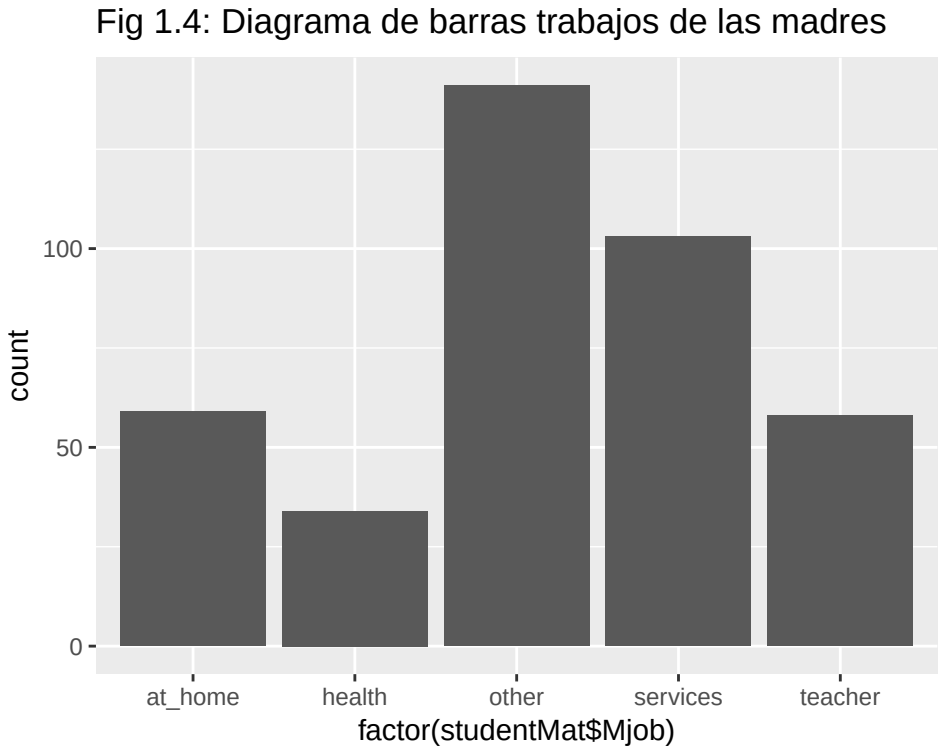
Fig 1.1 Histograma ausencias matemáticas, plot 3



Hemos utilizado 3 modos distintos de generar el mismo histograma para ver la distribución de las ausencias de los alumnos de matemáticos.

La mayoría tienen entre 1 y 20. Probablemente merecería la pena profundizar en esto, para ver si las ausencias influyen de algún modo en la nota final

```
#plot.new()
qplot(factor(studentMat$Mjob), main = "Fig 1.4: Diagrama de barras trabajos de las madres")
```



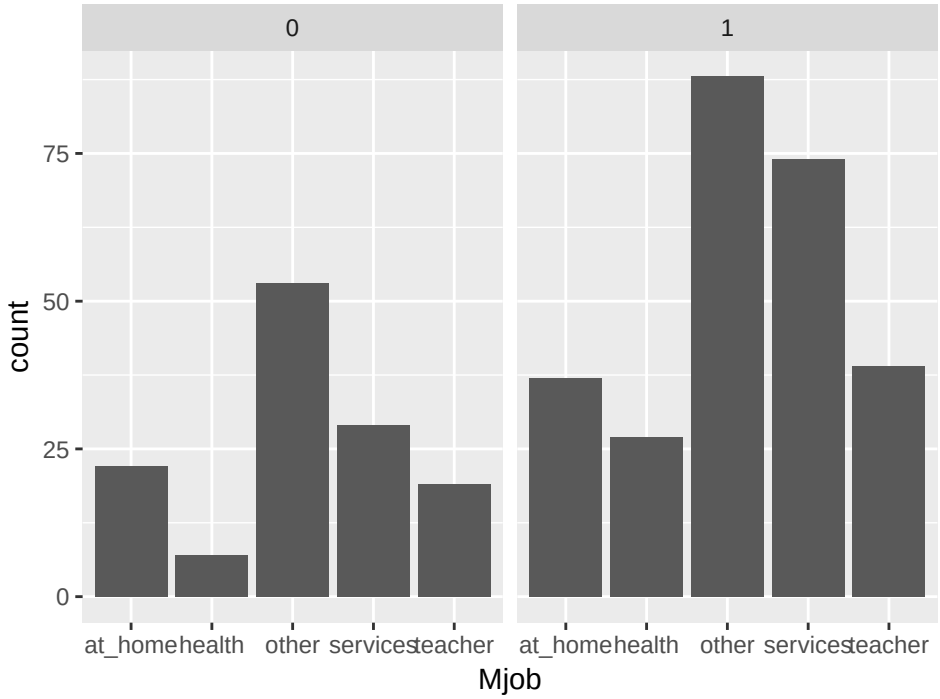
Hemos utilizado un diagrama de barras para ver la distribución de trabajos de las madres de los estudiantes de matemáticas. Vemos que la barra mayor es la correspondiente a "other", por lo que si hubiera posibilidad de obtener mejores datos, podríamos sugerir clasificar con más granularidad los trabajos.

Analizamos la relación de algunas variables con la variable objetivo

Consideramos que la variable objetivo es G3 = evaluación final.

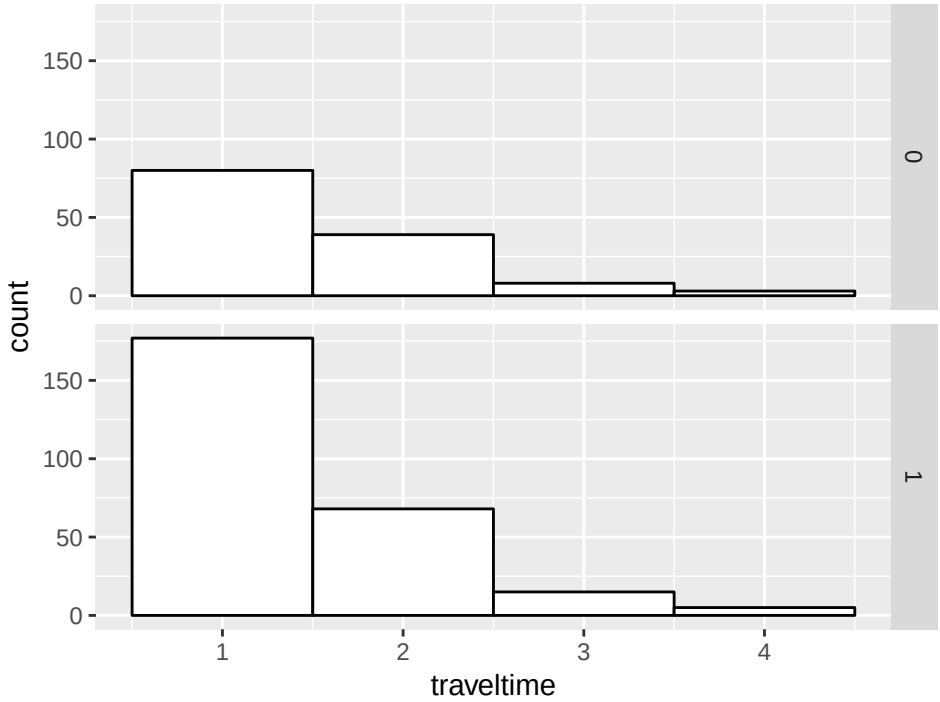
```
# Clasificamos en binario la variable target: Suponemos que "fail" es una nota
# final de 0 a 9. Y pass >9.

studentMat$pass <- ifelse(studentMat$G3>9, 1, 0)

# Para la variable tipo factor Mjob, podemos ver su distribución en relación al
# target, con diagramas de barras. Ejemplo:

 ggplot(studentMat, aes(Mjob)) + geom_bar() +
  facet_wrap(~ pass) +
        ggtitle ("Fig 2.1 Diagrama barras trabajo materno, por nota final") +
        theme(plot.title = element_text(vjust = +1.5, size = 12))
```

Fig 2.1 Diagrama barras trabajo materno, por nota final



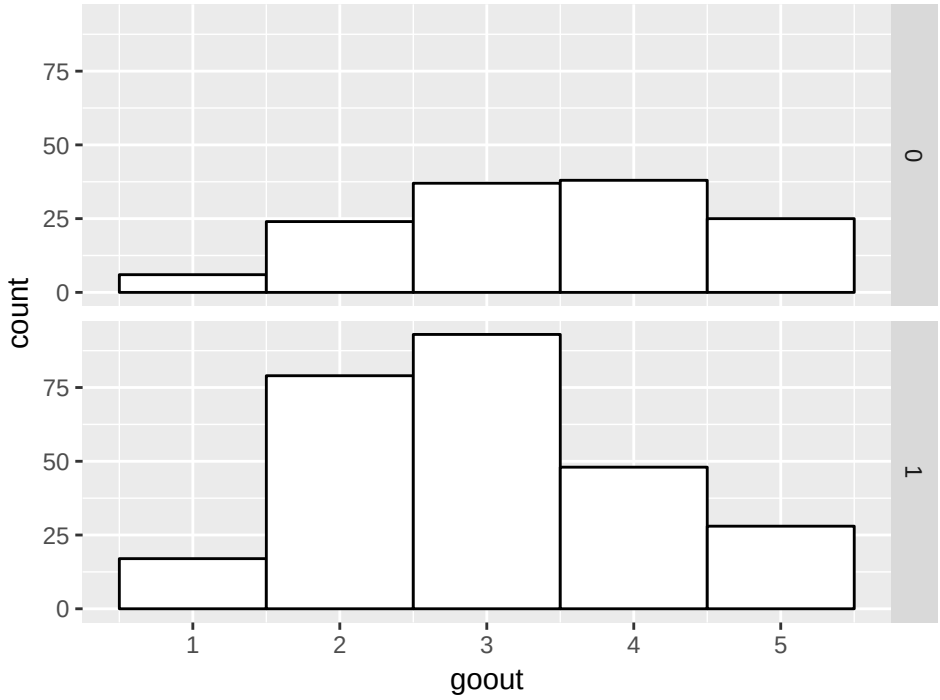

```
# Para las variables continuas, podemos ver su distribución en relación al target, con histogramas. Por

ggplot(studentMat, aes(x = traveltime)) + geom_histogram(binwidth = 1, fill = "white", colour = "black")
        facet_grid(pass ~ .) +
        ggtitle ("Fig 2.2 Histograma tiempo itinerario casa-escuela por nota final") +
        theme(plot.title=element_text(vjust = +1.5, size = 12))
```

Fig 2.2 Histograma tiempo itinerario casa−escuela por nota f



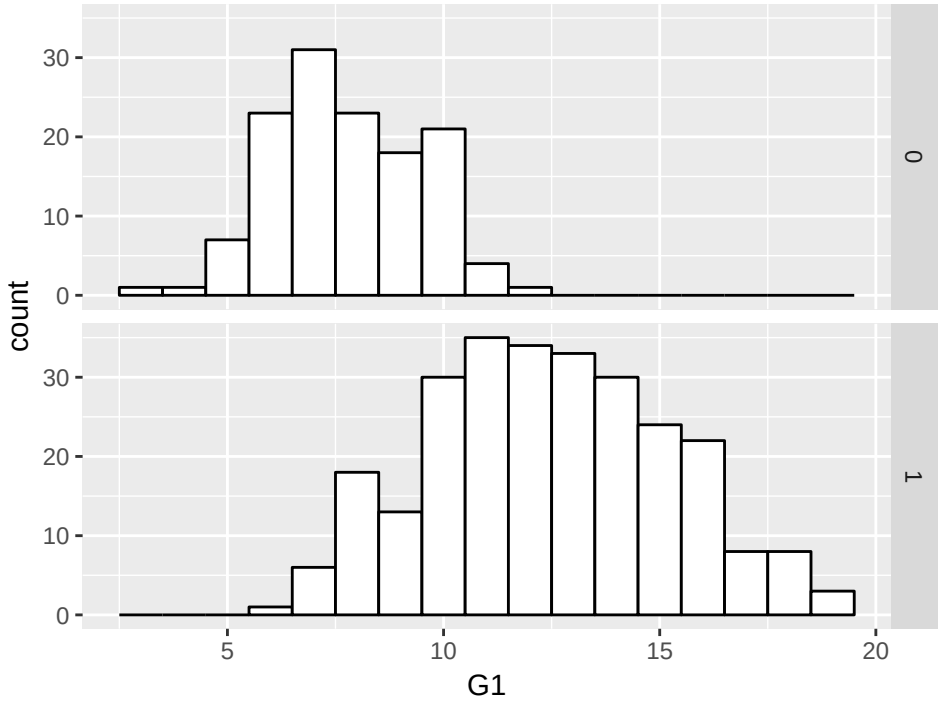
```
ggplot(studentMat, aes(x = goout)) + geom_histogram(binwidth = 1, fill = "white", colour = "black") +
        facet_grid(pass ~ .) +
        ggtitle ("Fig 2.3 Histograma salidas con amigos, por nota final") +
        theme(plot.title=element_text(vjust = +1.5, size = 12))
```

Fig 2.3 Histograma salidas con amigos, por nota final



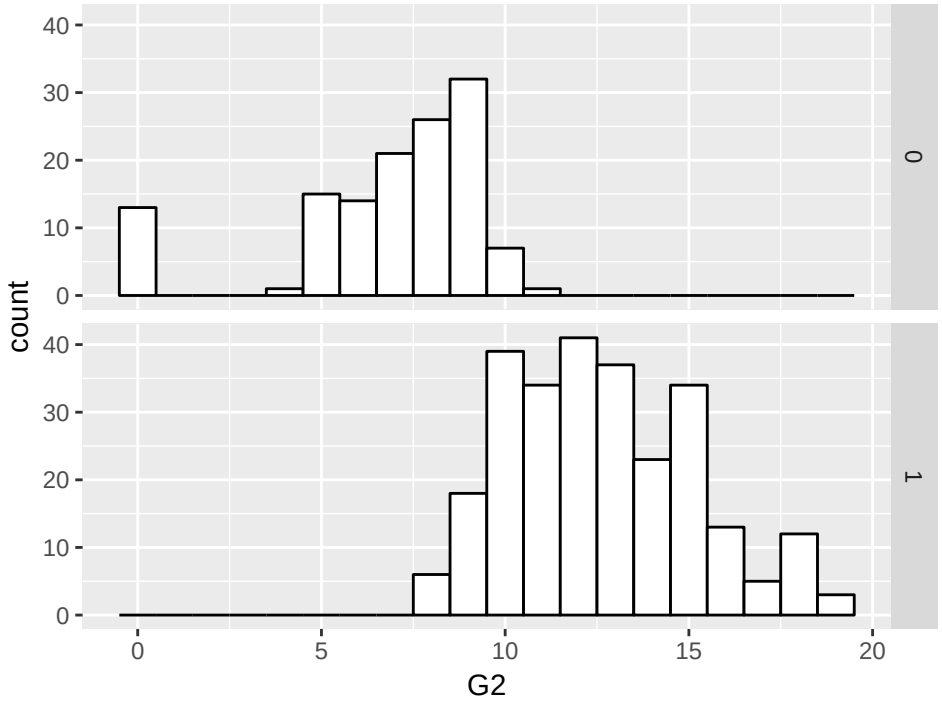

```
ggplot(studentMat, aes(x = G1)) + geom_histogram(binwidth = 1, fill = "white", colour = "black") +
        facet_grid(pass ~ .) +
        ggtitle ("Fig 2.4 Histograma G1 por nota final") +
        theme(plot.title=element_text(vjust = +1.5, size = 12))
```

Fig 2.4 Histograma G1 por nota final




```
ggplot(studentMat, aes(x = G2)) + geom_histogram(binwidth = 1, fill = "white", colour = "black") +
        facet_grid(pass ~ .) +
        ggtitle ("Fig 2.5 Histograma G2 por clasificación nota final") +
        theme(plot.title=element_text(vjust = +1.5, size = 12))
```

Fig 2.5 Histograma G2 por clasificación nota final



También podemos ver la relación de las variables continuas con el target, utilizando box plot

Vemos 2 de los ejemplos:

```
library(grid)
library(gridExtra)

plot1 = ggplot(studentMat, aes(factor(pass), traveltime, fill=factor(pass))) +
        geom_boxplot() +
   scale_colour_discrete(name = "Type") +
   scale_fill_discrete(name="Type", breaks=c("0", "1"),
            labels=c("fail", "pass")) +
   scale_x_discrete(breaks=c("0", "1"), labels=c("fail", "pass")) +
   xlab("") +
   ggtitle ("Fig 3.1. Tiempo casa-escuela, por nota final") +
        theme(plot.title=element_text(vjust = +2.5, size = 7),
              axis.text.x=element_blank(), axis.title.x=element_blank())

plot2 = ggplot(studentMat, aes(factor(pass), G1, fill=factor(pass))) +
        geom_boxplot() +
   scale_colour_discrete(name = "Type") +
   scale_fill_discrete(name="Type", breaks=c("0", "1"),
            labels=c("fail", "pass")) +
   scale_x_discrete(breaks=c("0", "1"), labels=c("fail", "pass")) +
   xlab("") +
   ggtitle ("Fig 3.2. G1 por nota final") +
        theme(plot.title=element_text(vjust = +2.5, size = 8),
              axis.text.x=element_blank(), axis.title.x=element_blank())

grid.arrange(plot1, plot2, nrow=1, ncol=2)
```

Fig 3.1. Tiempo casa−escuela, por nota final

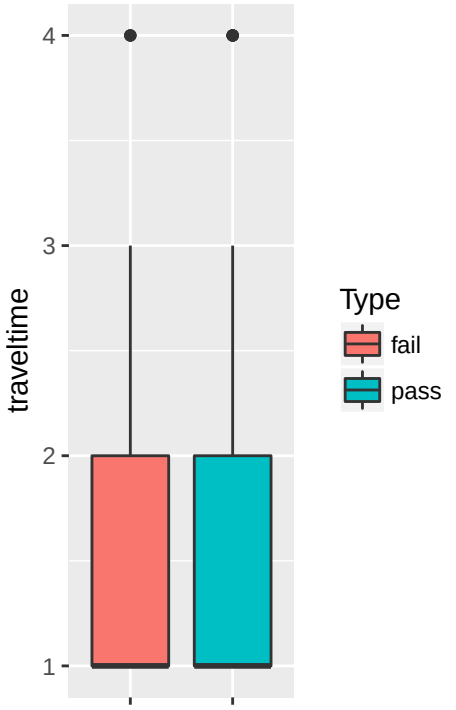
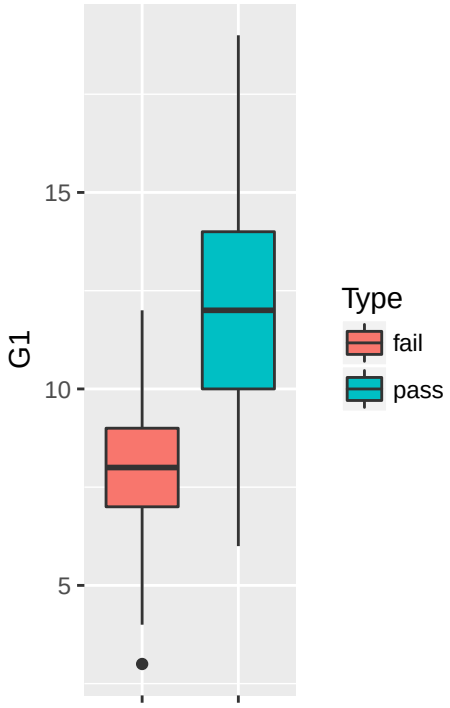


Fig 3.2. G1 por nota final



No se observan diferencias en la distribución de la variable Mjob (trabajo materno), para alumnos que aprueban (Fig. 2.1, gráfico de la derecha “1” y los que no “gráfico de la izquierda,”0“).

La distribución del tiempo que tarda el alumno en ir de casa a la escuela, parece similar para aquellos con “pass” y aquellos con “fail” en la nota final (G3).

Sin embargo, la distribución del tiempo que dedican a salir con amigos sí tiene distribuciones algo diferentes. Lo vemos tanto en los histogramas como en los box plots.

Y en el caso de las notas previas (G1 y G2) tienen distribuciones muy diferentes para los casos “pass” y “fail” en la nota final G3. En el caso de G1, lo hemos visto con box plot además de histograma. Con el box plot podemos decir que, en términos de mediana, G1 es mayor para los alumnos que aprueban en la nota final, que para aquellos alumnos que suspenden.

Este análisis de las distribuciones de distintas variables en relación al target, nos da pistas para nuestra fase posterior de Modelado.

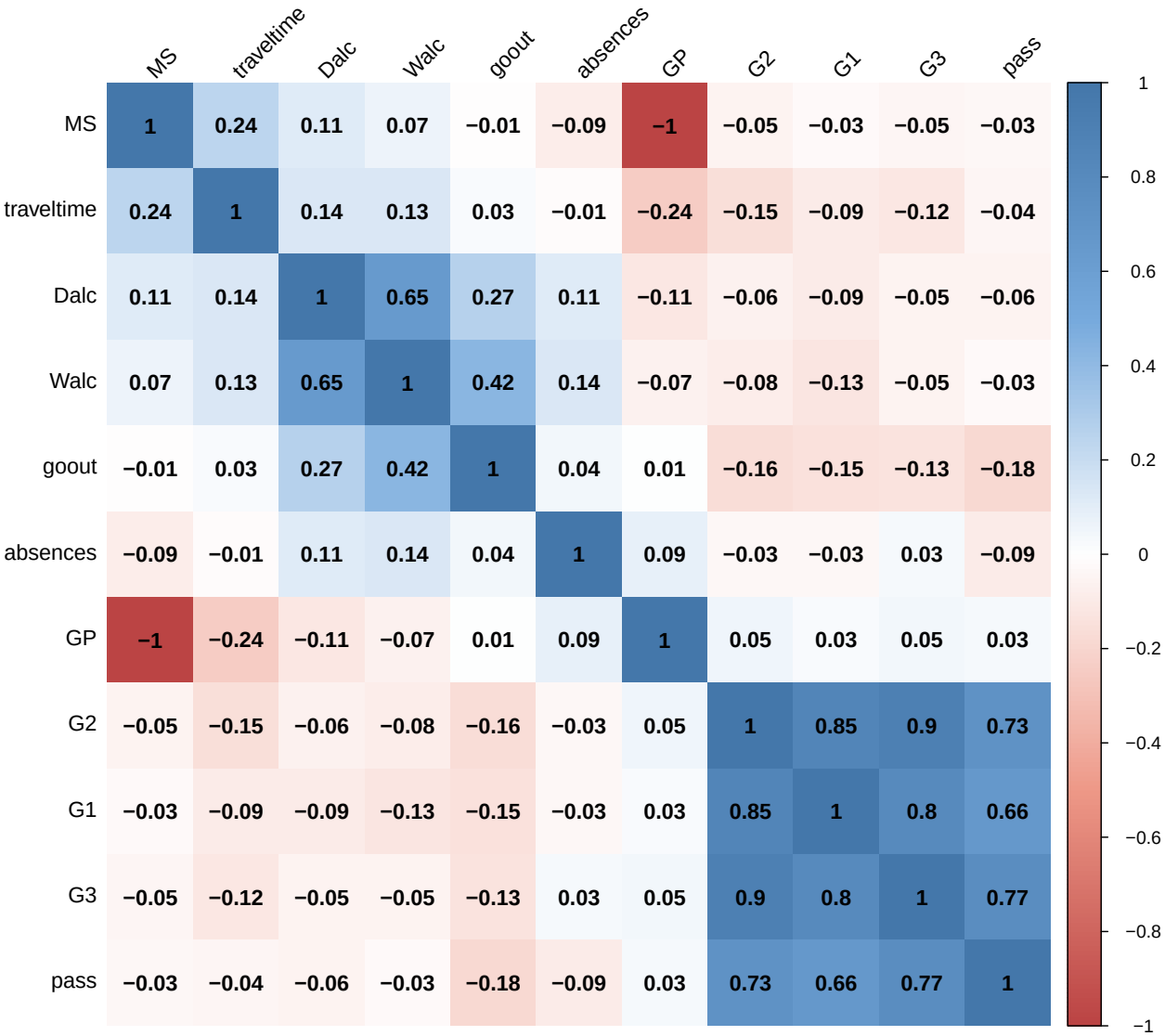
Matriz de correlación

Para ver lo correladas que están unas variables con otras, y también con el target, podemos calcular y dibujar una matriz de correlación.

```
# Primero creamos variables dummies para las variables de tipo factor que queramos
# incluir en la matriz de correlación. Por ejemplo:
studentMat$GP <- ifelse(studentMat$school == "GP", 1, 0)
studentMat$MS <- ifelse(studentMat$school == "MS", 1, 0)

# Por simplificación del ejemplo, sólo consideramos algunas variables para construir # la matriz de cor

matCor <- cor(studentMat[, c("GP","MS","absences","goout","Dalc","Walc","traveltime","G1","G2","G3","pas
```

```r
library(corrplot)
matCor[is.na(matCor)] <- 0

#plot.new()
# Generamos una paleta de colores más claros
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))

# Dibujamos la matriz de correlación con cuadrados de colores y etiquetas negras
  corrplot(matCor, method = "shade", shade.col = NA, tl.col = "black",
         tl.srt = 45, col = col(200), addCoef.col="black", order="AOE",
         mar = c(1,0,2,0), line=-2, is.corr=FALSE,
         main = "Fig 3.1. Predictors correlation matrix")
```

Fig 3.1. Predictors correlation matrix



	MS	traveltime	Dalc	Walc	goout	absences	GP	G2	G1	G3	pass
MS	1	0.24	0.11	0.07	−0.01	−0.09	−1	−0.05	−0.03	−0.05	−0.03
traveltime	0.24	1	0.14	0.13	0.03	−0.01	−0.24	−0.15	−0.09	−0.12	−0.04
Dalc	0.11	0.14	1	0.65	0.27	0.11	−0.11	−0.06	−0.09	−0.05	−0.06
Walc	0.07	0.13	0.65	1	0.42	0.14	−0.07	−0.08	−0.13	−0.05	−0.03
goout	−0.01	0.03	0.27	0.42	1	0.04	0.01	−0.16	−0.15	−0.13	−0.18
absences	−0.09	−0.01	0.11	0.14	0.04	1	0.09	−0.03	−0.03	0.03	−0.09
GP	−1	−0.24	−0.11	−0.07	0.01	0.09	1	0.05	0.03	0.05	0.03
G2	−0.05	−0.15	−0.06	−0.08	−0.16	−0.03	0.05	1	0.85	0.9	0.73
G1	−0.03	−0.09	−0.09	−0.13	−0.15	−0.03	0.03	0.85	1	0.8	0.66
G3	−0.05	−0.12	−0.05	−0.05	−0.13	0.03	0.05	0.9	0.8	1	0.77
pass	−0.03	−0.04	−0.06	−0.03	−0.18	−0.09	0.03	0.73	0.66	0.77	1

Los valores más cercanos a 1 (colores más oscuros), nos indican las variables más correladas.

Un resultado obvio, que no nos aporta nada, es la correlación entre "pass" y G3. En realidad, dependiendo de si después, en la fase de modelado, optamos por un "outcome" binario, elegiremos "pass", y si no, elegiremos

"G3".

Otros valores altos corroboran lo que habíamos visto en los histogramas, boxplots y diagramas de barras.

La matriz de correlación, puede usarse además, en la fase de modelado, para seleccionar las variables que queremos incluir en el modelo. A veces, si hay 2 muy correlacionadas, se decide incluir sólo una de ellas.

Conclusiones preliminares:

Sería conveniente una clasificación más detallada de las profesiones de las madres de los alumnos. Puede guardar alguna relación con la nota final, seguramente en combinación con otras variables, pero hay muchas instancias de "Other".

No se observan anomalías en los datos que puedan dar a entender que ha habido errores en la recolección de los mismos.

Se espera que las variables G1 y G2 sean determinantes para un modelo de predicción de G3 como target, es decir, que las calificaciones pasadas de los alumnos influirán en su nota final.

La dedicación de los alumnos a salir con amigos también puede resultar influyente a la hora de su calificación final.