R-for-ml

Esther Cuervo Fernández

30 Mayo, 2019

Introducción

En este infome se trata de exploar distintas técnicas de análisis sobre el dataset *student*, disponible en: https://archive.ics.uci.edu/ml/datasets/Student+Academics+Perfomance

Este dataset contiene datos sobre estudiantes de dos asignaturas: Matemáticas y Potugués. Los campos del dataset son los siguientes:

1. **school** - colegio del estudiante (binario: "GP" - Gabriel Pereira o "MS" - Mousinho da Silveira)
2. **sex** - género del estudiante (binario: "F" - mujer o "M" - hombre)
3. **age** - edad del estudiante (numérico: de 15 a 22)
4. **address** - tipo de dirección del estudiante (binario: "U" - urbano o "R" - rural)
5. **famsize** - tamaño de familia (binario: "LE3" - menor o igaul a 3 o "GT3" - más de 3)
6. **Pstatus** - estado de cohabitación de los padres (binario: "T" - viviendo juntos o "A" - separados)
7. **Medu** - educación de la madre (numérico: 0 - ninguna, 1 - educación primaria (4o), 2 – 5o a 9o, 3 – educación secundaria o 4 – educación superior)
8. **Fedu** - educación del padre (numérico: 0 - ninguna, 1 - educación primaria (4o), 2 – 5o a 9o, 3 – educación secundaria o 4 – educación superior)
9. **Mjob** - trabajo de la madre (nominal: "teacher" - profesor, "health" - salud, "services" - funcionario, "at_home" - am@ de casa o "other" - otro)
10. **Fjob** - trabajo del padre (nominal: "teacher" - profesor, "health" - salud, "services" - funcionario, "at_home" - am@ de casa o "other" - otro)
11. **reason** - razón por la que eligieron el colegio (nominal: "home" - cerca de casa, "reputation" - reputación del colegio, "course" - preferencia de curso o "other" - otro)
12. **guardian** - tutor legal del estudiante (nominal: "mother" - madre, "father" - padre o "other" - otro)
13. **traveltime** - tiempo de viaje entre casa y el colegio (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hour, o 4 - >1 hour)
14. **studytime** - tiempo de estudio semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, o 4 - >10 horas)
15. **failures** - número de clases suspensas anteriores (numérico: n si 1<=n<3, si no 4)
16. **schoolsup** - ayuda extra curricular (binario: yes o no)
17. **famsup** - ayuda familiar (binario: yes o no)
18. **paid** - clases extra pagadas en la asignatura del curso (matemáticas o portugués)) (binario: yes o no)
19. **activities** - actividades extra-curriculares (binario: yes o no)
20. **nursery** - fue a guardería (binario: yes o no)

21. **higher** - quiere educación superior (binario: yes o no)
22. **internet** - acceso a Internet en casa (binario: yes o no)
23. **romantic** - en una relación romántica (binario: yes o no)
24. **famrel** - calidad de relaciones familiares (numérico: de 1 - muy mala a 5 - excelente)
25. **freetime** - tiempo libre después de clase (numérico: de 1 - muy bajo a 5 - muy alto)
26. **goout** - sale con amigos (numérico: de 1 - muy bajo a 5 - muy alto)
27. **Dalc** - consumo de alcohol durante semana (numérico: de 1 - muy bajo a 5 - muy alto)
28. **Walc** - consumo de alcohol durante fines de semana (numérico: de 1 - muy bajo a 5 - muy alto)
29. **health** - estado de salud actual (numérico: de 1 - muy malo a 5 - muy bueno)
30. **absences** - número de ausencias (numérico: de 0 a 93)
31. **G1** - nota del primer periodo (numérico: de 0 a 20)
32. **G2** - nota del segundo periodo (numérico: de 0 a 20)
33. **G3** - nota final (numérico: de 0 a 20)

El fichero de descripción también destaca que hay ciertos alumnos (382) que están presentes en los datasets para ambas asignaturas.

Nuestro objetivo va a ser clasificar a los alumnos como aprobados o suspensos, considerando aprobado como G3>=10 y suspenso G3<10. Para ello vamos a utilizar el dataset que contiene información sobre la clase de matemáticas.

Lectura de datos

Los datos están formados por dos tablas, una para cada asignatura. Nos quedamos con la de matemáticas:

```
math=read.table("./data/student-mat.csv",sep=";",header=TRUE)
```

Creamos la variable a predecir, `pass`, binaria - "yes" para aprobados y "no" para suspensos:

```
math$pass <- as.factor(ifelse(math$G3>=10, "yes", "no"))
```

Estadísticas para la variable original G3:

```
summary(math$G3)
```

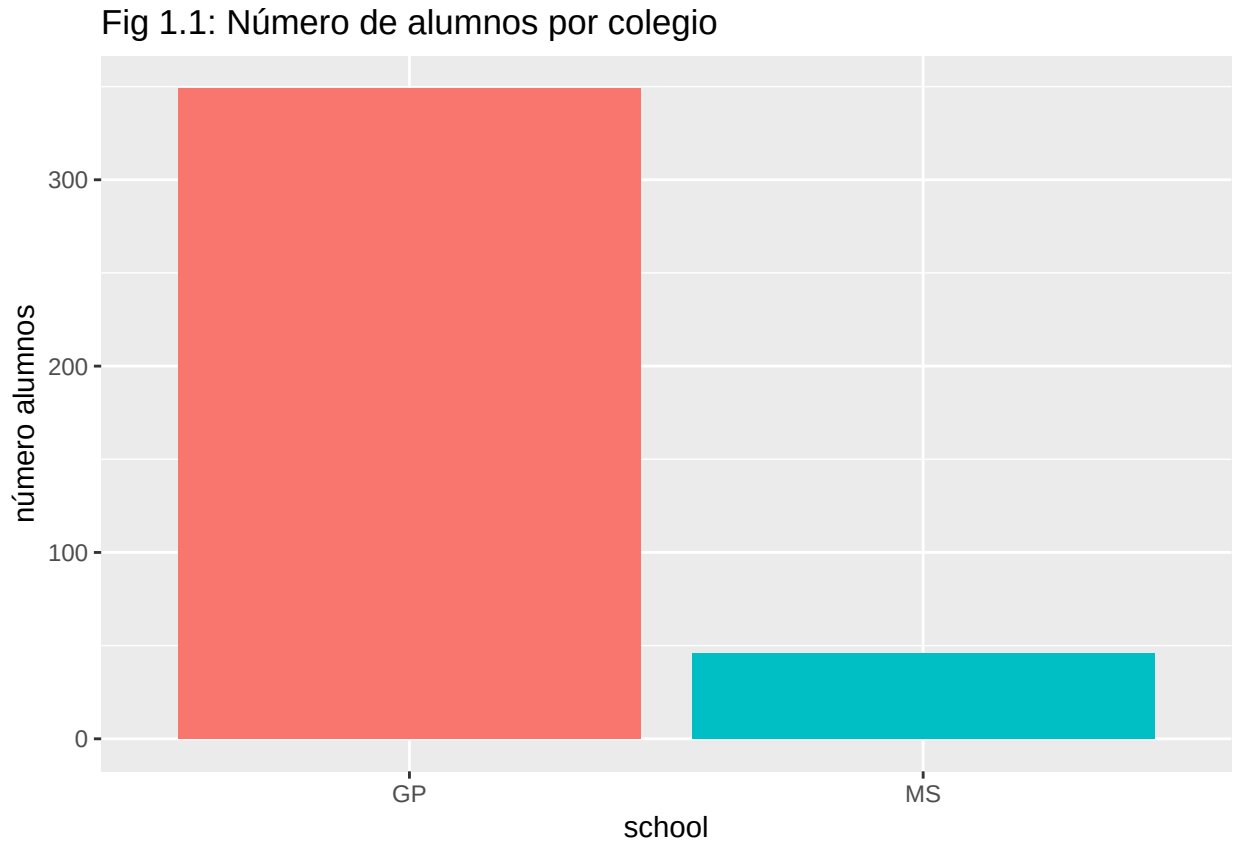
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.42   14.00   20.00
```

Podemos ver que existen notas tanto iguales a 0 como a 20. Los cuartiles también aportan información, por ejemplo que han aprobado algo más de un 50% de alumnos.

Exploración de variables

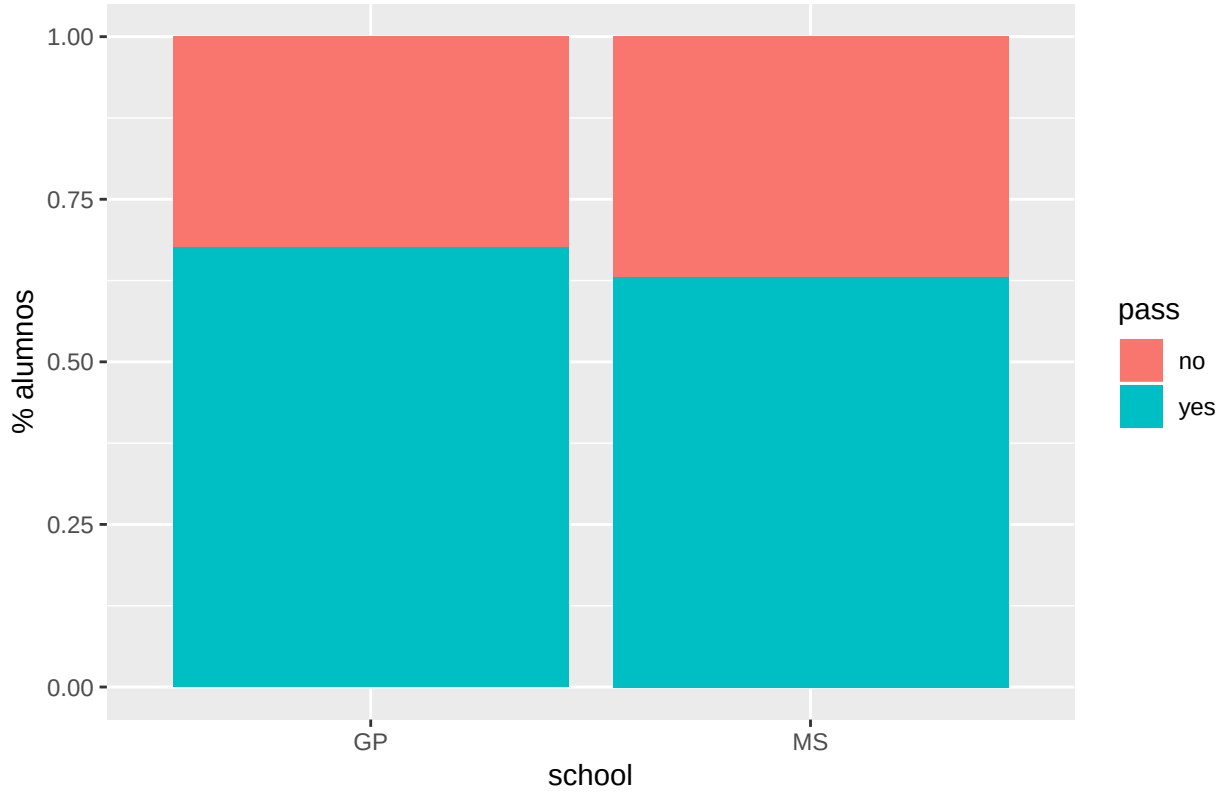
Exploramos algunas posibles relaciones entre variables así como con la variable a predecir, `pass`. Por ejemplo `school` y `reason`:

```r
if(! "ggplot2" %in% installed.packages()) install.packages("ggplot2", depend = TRUE)

library(ggplot2)

ggplot(math,mapping=aes(x=school,fill=school)) +
  geom_bar(show.legend = FALSE) +
  ggtitle("Fig 1.1: Número de alumnos por colegio") +
  ylab("número alumnos")
```



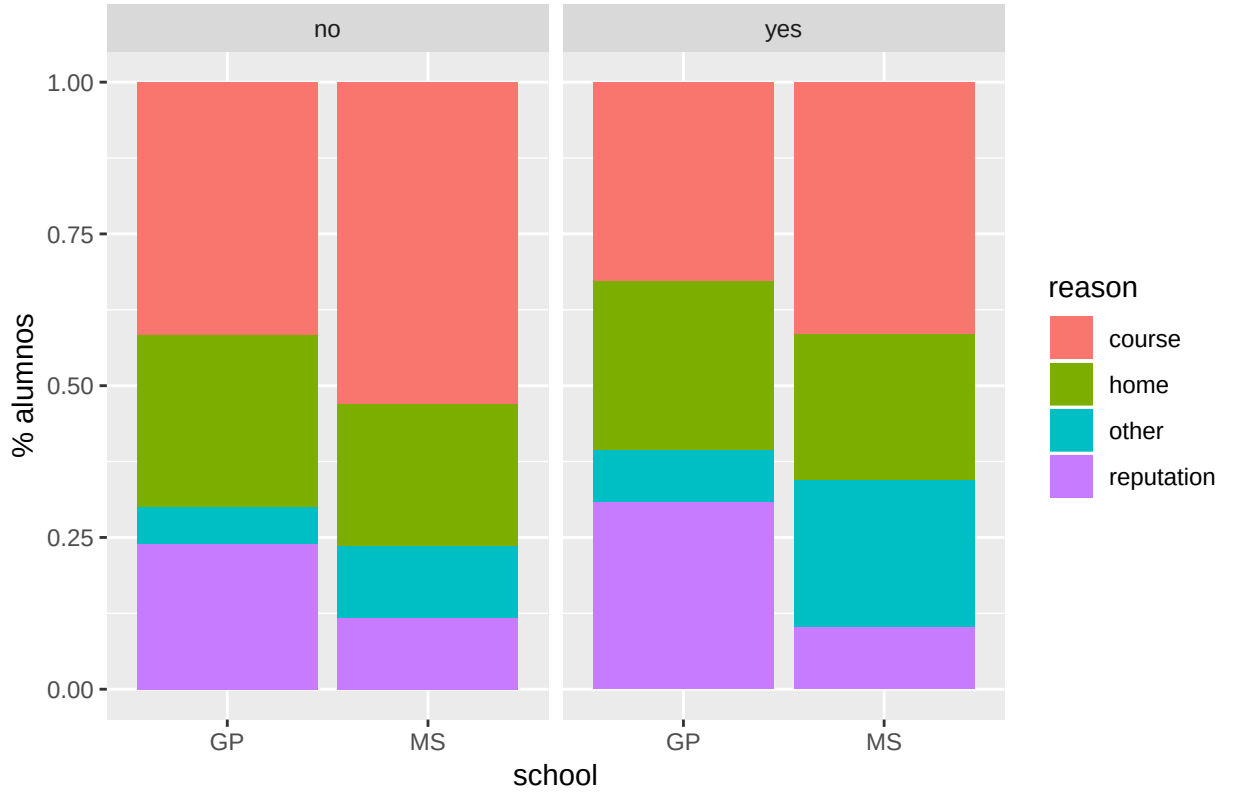

```r
ggplot(math,mapping=aes(x=school, fill=pass)) +
  geom_bar(position = "fill") +
  ggtitle("Fig 1.2: Distribución de aprobados por colegio") +
  ylab("% alumnos")
```

Fig 1.2: Distribución de aprobados por colegio




```
ggplot(math,mapping=aes(x=school, fill=reason)) +
  facet_wrap(~pass) +
  geom_bar(position = "fill") +
  ggtitle("Fig 1.3: Distribución de razón por colegio y aprobado") +
  ylab("% alumnos")
```

Fig 1.3: Distribución de razón por colegio y aprobado

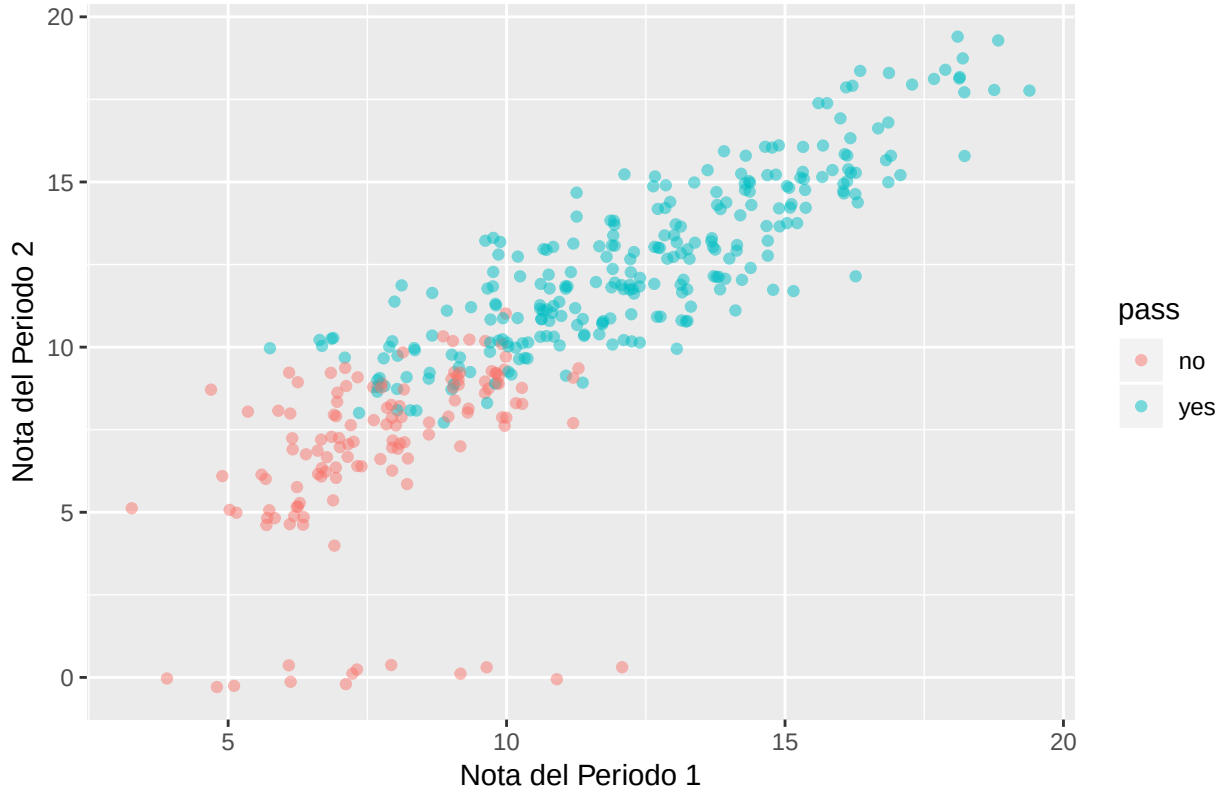


- La Fig 1.1. nos permite observar que el número de alumnos en el colegio `MS` es muy inferior al de alumnos en `GP`.
- La Fig 1.2. nos muestra que el porcentaje de aprobados es muy similar en ambos colegios, por lo que, aunque `MS` parece ser algo más difícil, no parece haber una diferencia significativa.
- La Fig 1.3. muestra las razones de elección del colegio por colegio y resultado. En el caso de los alumnos suspensos se ve un porcentaje significativamente alto de elección por causa `course`, mientras que a los alumnos aprobados parece importarles más la reputación en el caso del colegio `GP` y other en `MS`. En ambos casos parece que `GP` es más prestigioso.

Otra relación interesante puede ser la de `G1` y `G2`, las notas del primer y segundo periodo:

```
ggplot(math) +
  geom_jitter(mapping=aes(x=G1, y=G2, color=pass),alpha=0.5) +
  ggtitle("Fig 2: Nota del Periodo 2 contra Nota del Periodo 1 y aprobado") +
  xlab("Nota del Periodo 1") +
  ylab("Nota del Periodo 2")
```

Fig 2: Nota del Periodo 2 contra Nota del Periodo 1 y aprobado



Se pueden ver unos puntos que siguen una relación lineal, como era esperado, pero también algunos puntos que tienen valor de `G2` igual a 0, con valores de `G1` distribuidos. Podemos obtener a estos alumnos mediante un filtro:

```
if(! "dplyr" %in% installed.packages()) install.packages("dplyr", depend = TRUE)
if(! "knitr" %in% installed.packages()) install.packages("knitr", depend = TRUE)

library(dplyr)
library(knitr)
library(rmarkdown)

g2strange <- dplyr::filter(math,G2 == 0, G1 > 0)

kable(select(g2strange,"school","reason","age","higher","failures","traveltime","G1",
             "G2","G3","pass"))
```

school	reason	age	higher	failures	traveltime	G1	G2	G3	pass
GP	course	15	yes	2	2	12	0	0	no
GP	course	15	yes	0	3	8	0	0	no
GP	course	15	yes	0	4	9	0	0	no
GP	course	15	yes	0	1	11	0	0	no
GP	course	17	yes	0	3	10	0	0	no
GP	course	16	yes	2	2	4	0	0	no
GP	home	17	yes	3	1	5	0	0	no
GP	home	19	no	3	1	5	0	0	no

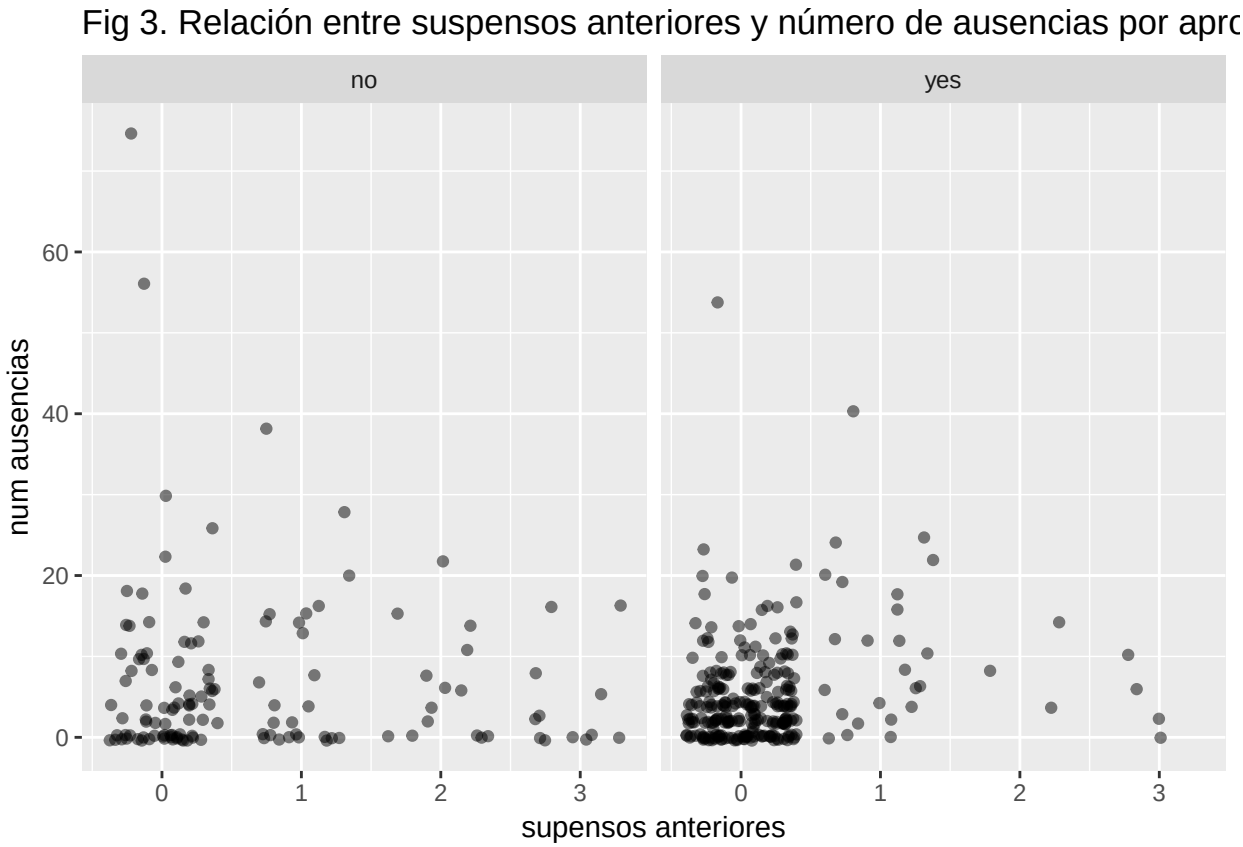
school	reason	age	higher	failures	traveltime	G1	G2	G3	pass
GP	course	16	yes	1	2	7	0	0	no
GP	course	16	yes	0	1	6	0	0	no
GP	course	18	yes	0	2	7	0	0	no
GP	reputation	18	no	0	2	6	0	0	no
GP	home	18	yes	0	1	7	0	0	no

Se puede observar que todos los alumnos son del mismo colegio, y que todos han obtenido una nota de 0 en `G3`. Especialmente curiosos son los alumnos que tenían notas de aprobado en `G1`. Una posible explicación son alumnos que han abandonado la clase antes del final del segundo periodo, y si el sistema educativo del país de origen de los datos es obligatorio sólo hasta la mayoría de edad, esto podría significar un abandono en el caso de los dos alumnos que no desean una educación superior, y ya tienen más de 18 años. También se puede observar que algunos de los alumnos tienen un número alto de suspensos anteriores.

En cualquier caso estos valores podrían ser outliers.

Otra variable que puede tener mucha relación con la nota final es el número de ausencias, `abscences` y el número de suspensos anteriores `failures`:

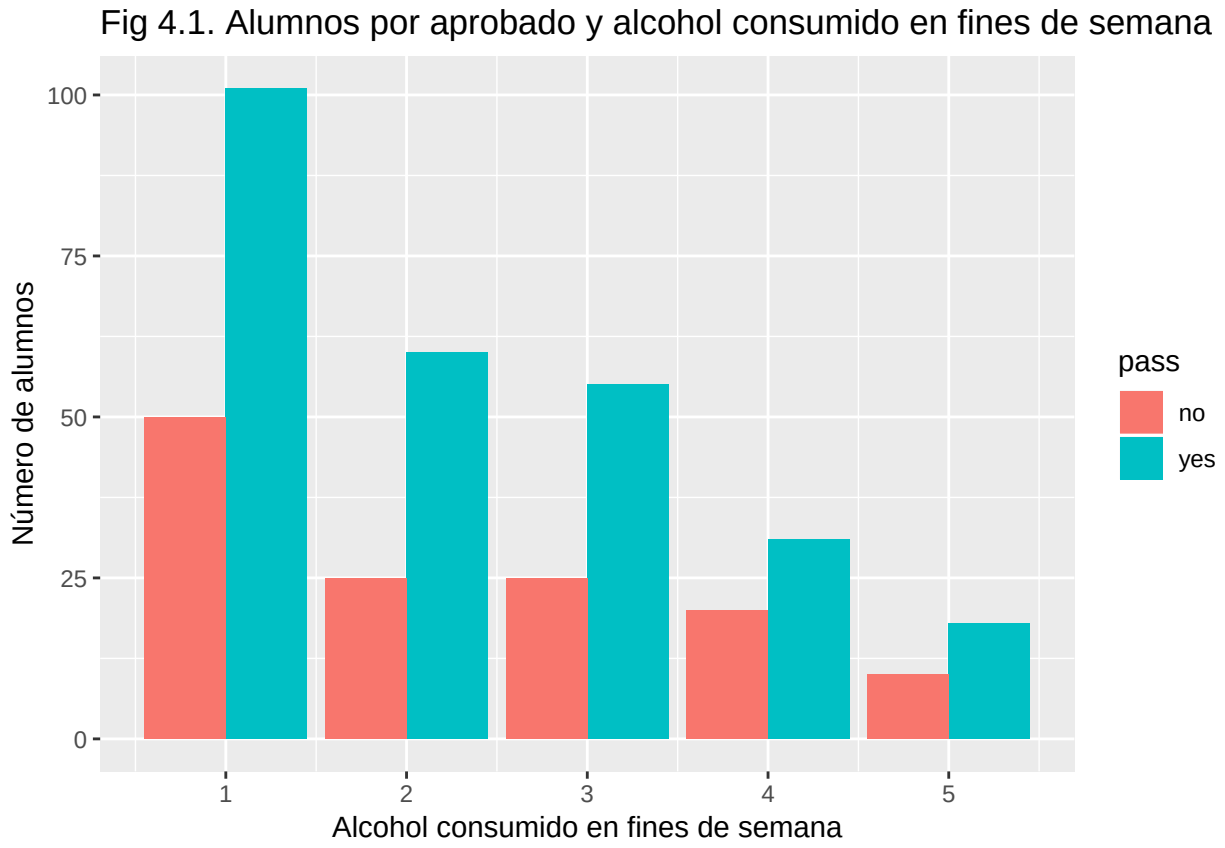
```
ggplot(math) +
  geom_jitter(mapping = aes(x=failures, y=absences),alpha=0.5) +
  facet_wrap(~factor(pass,labels=c("no","yes"))) +
  xlab("supensos anteriores") + ylab("num ausencias") +
  ggtitle("Fig 3. Relación entre suspensos anteriores y número de ausencias por aprobado")
```



Dentro de los no aprobados podemos ver un número alto de suspensos anteriores, aunque no se aprecia que la cantidad de ausencias sea especialmente mayor que en los aprobados. Existen algunos posibles outliers con más de 40 ausencias.

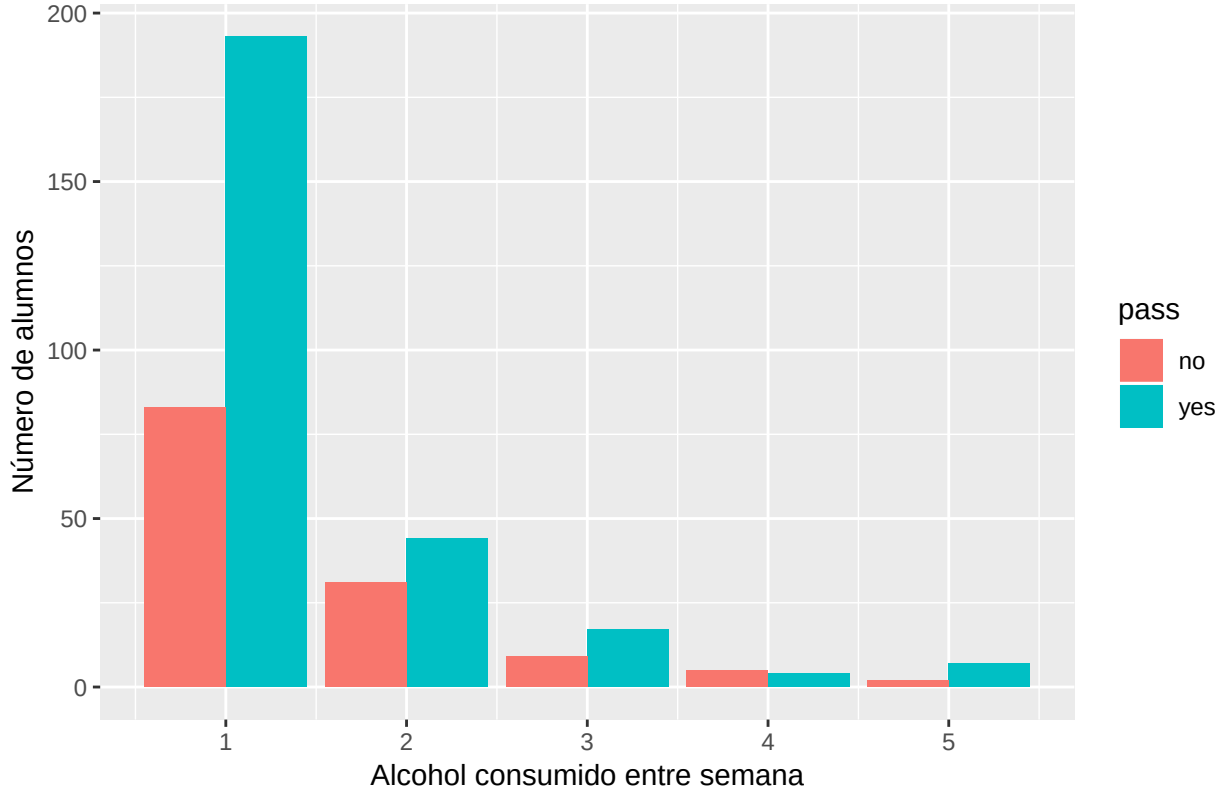
Otras variables como consumo de alcohol durante semana `Dalc` y fines de semana `Walc` puede tener relación con la nota.

```
ggplot(math) +
geom_bar(mapping = aes(x=Walc, fill=pass),position="dodge") +
ggtitle("Fig 4.1. Alumnos por aprobado y alcohol consumido en fines de semana") +
ylab("Número de alumnos") + xlab("Alcohol consumido en fines de semana")
```



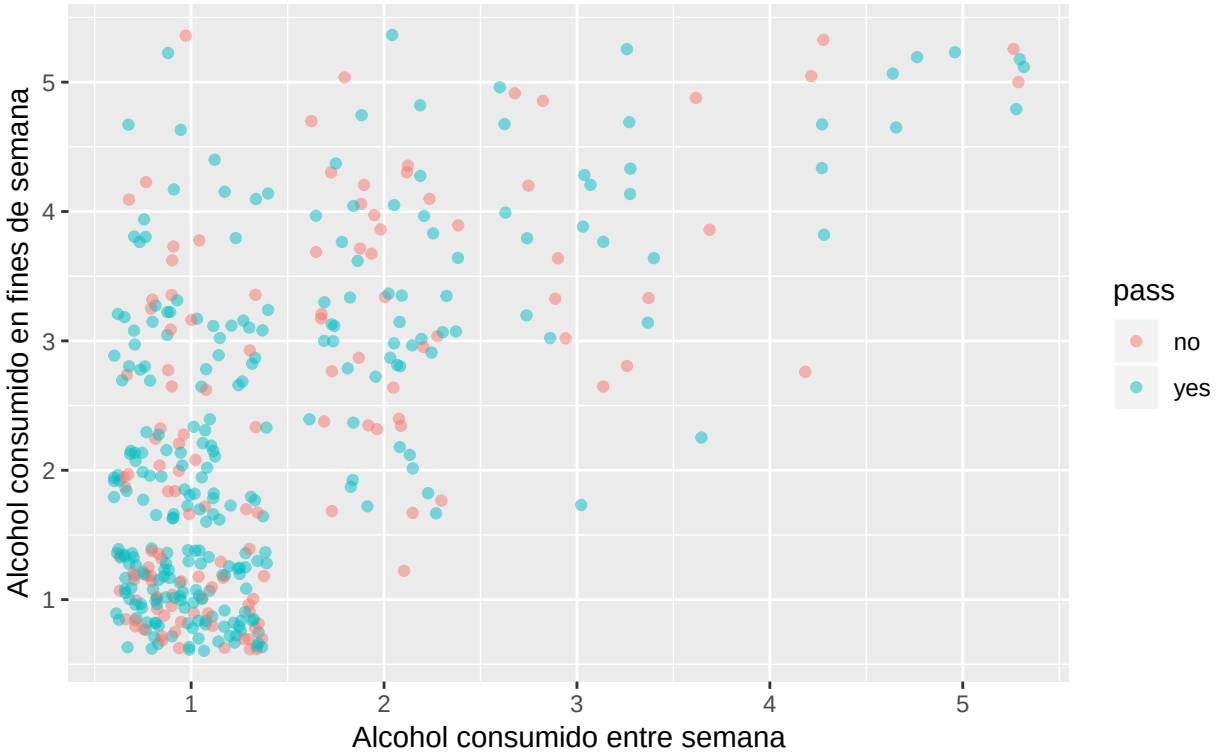

```
ggplot(math) +
geom_bar(mapping = aes(x=Dalc, fill=pass),position="dodge") +
ggtitle("Fig 4.2. Alumnos por aprobado y alcohol consumido entre semana") +
ylab("Número de alumnos") + xlab("Alcohol consumido entre semana")
```

Fig 4.2. Alumnos por aprobado y alcohol consumido entre semana




```
ggplot(math) +

geom_jitter(mapping = aes(x=Dalc, y=Walc,color=pass), alpha=0.5) +
ggtitle("Fig 4.3. Relación entre consumo de alcohol
        en fines de semana y entre semana y nota") +
  xlab("Alcohol consumido entre semana") +
  ylab("Alcohol consumido en fines de semana")
```

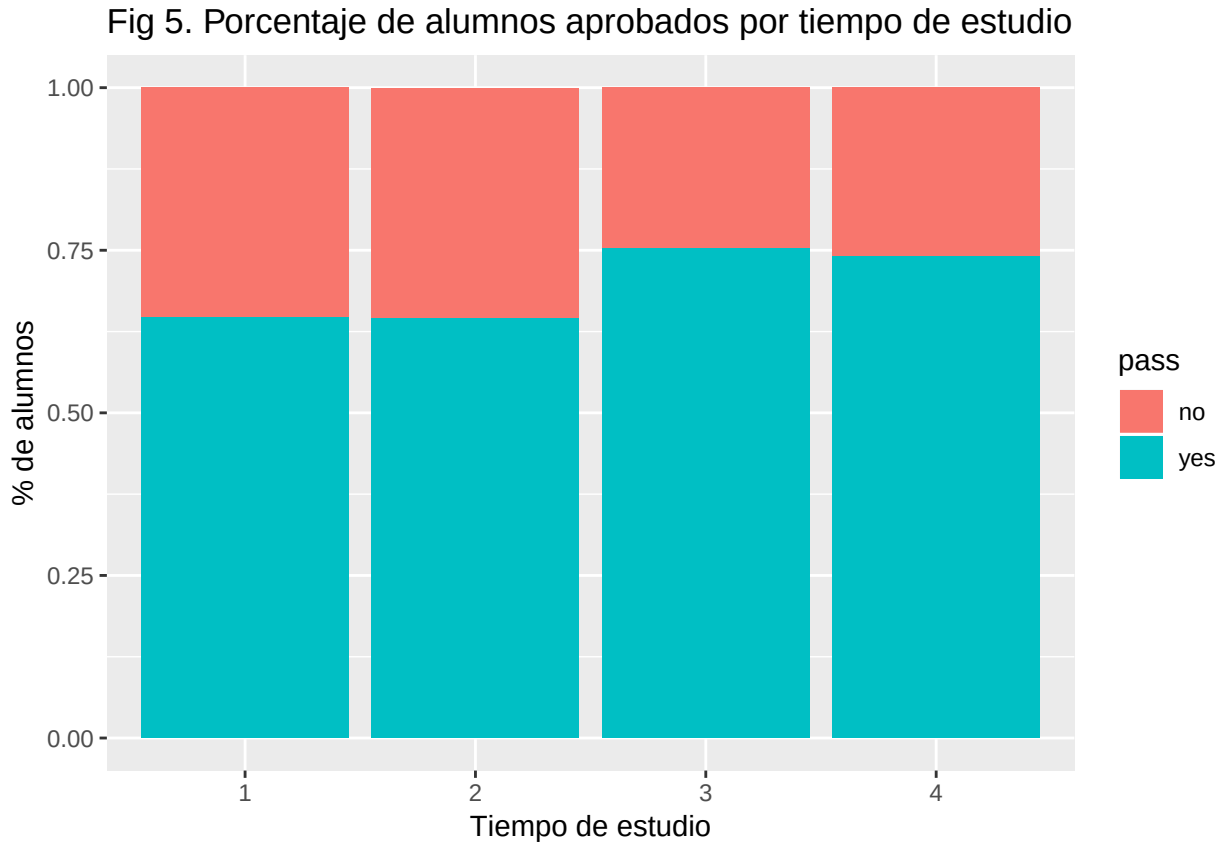
Fig 4.3. Relación entre consumo de alcohol
en fines de semana y entre semana y nota



- La Fig 4.1. nos indica que la mayoría de alumnos tanto suspensos como aprobados consumen muy poco alcohol durante fines de semana, aunque aún así existe una cantidad importante de ellos que consumen alcohol.
- La Fig 4.2. indica unos números mucho más bajos de consumo de alcohol entre semana, e incluso muestra que con consumos altos el número de suspensos supera a los aprobados, aunque con valores tan pequeños de muestra no se pueden hacer suposiciones.
- La Fig 4.3. muestra cierta relación entre consumir alcohol entre semana y en fines de semana, aunque es mayormente unilateral: los alumnos que consumen mucho alcohol entre semana tienden a consumir también mucho los fines de semana.

Otra posible relación es entre el tiempo de estudio y la nota:

```
ggplot(math) +
geom_bar( mapping = aes(x=factor(studytime),fill=pass), position = "fill") +
ggtitle("Fig 5. Porcentaje de alumnos aprobados por tiempo de estudio") +
xlab("Tiempo de estudio") + ylab("% de alumnos")
```

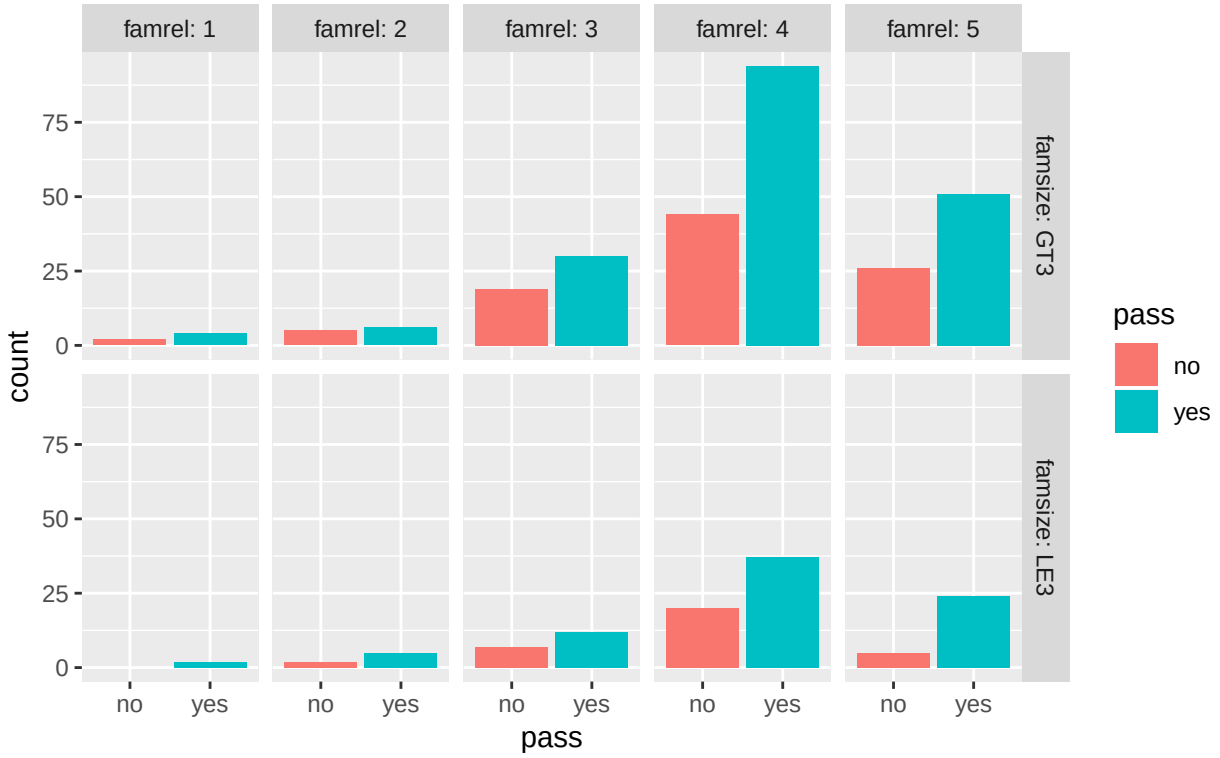


Si que es posible apreciar cierto aumento del porcentaje de alumnos aprobados en horas de estudio más altas, aunque este aumento sólo se aprecia por encima de las 5 horas de estudio.

También se pueden explorar algunas variables relacionadas con las familias de los alumnos, `famsize`, `Pstatus`, `Mjob`, `Fjob`, y `famrel`:

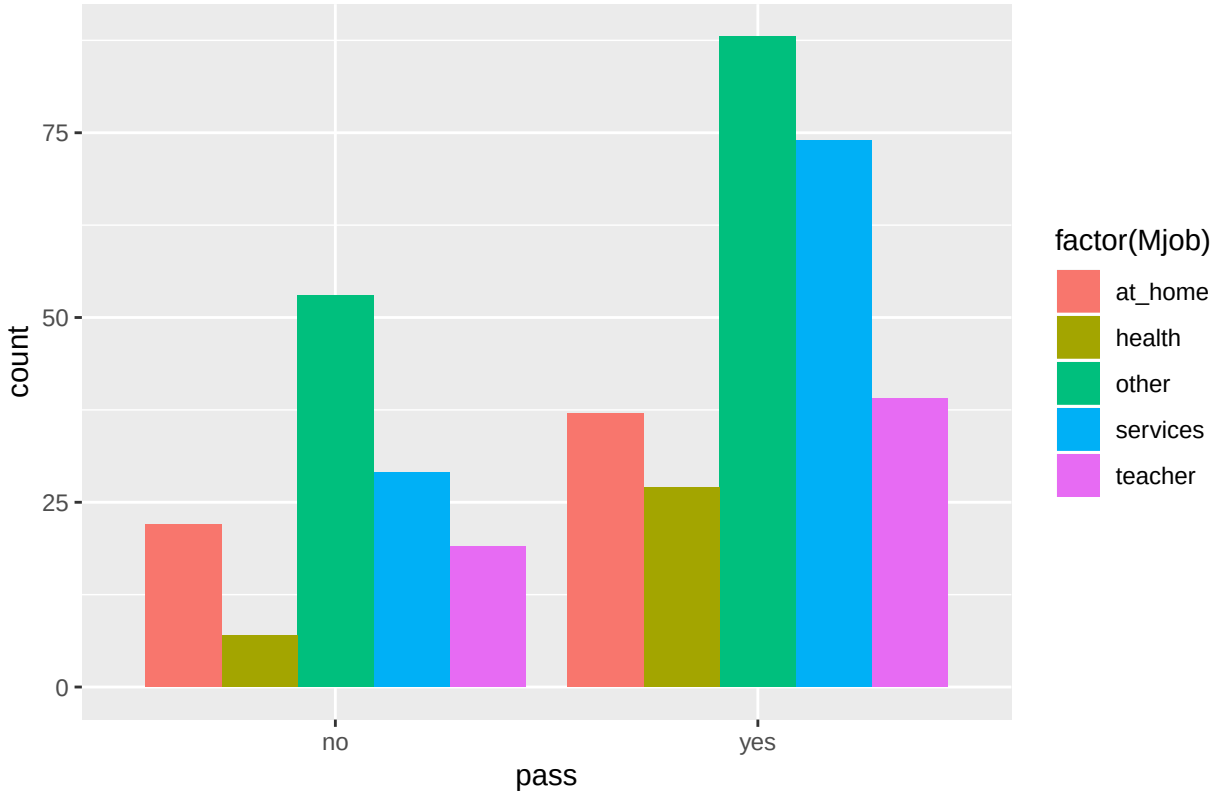
```
ggplot(math,mapping=aes(x=pass,fill=pass)) +
  geom_bar() +
  facet_grid(famsize ~ famrel, labeller = label_both) +
  ggtitle("Fig 6.1. Número de alumnos aprobados y suspensos por
          tamaño de familiia y relación con la familia")
```

Fig 6.1. Número de alumnos aprobados y suspensos por tamaño de familiia y relación con la familia



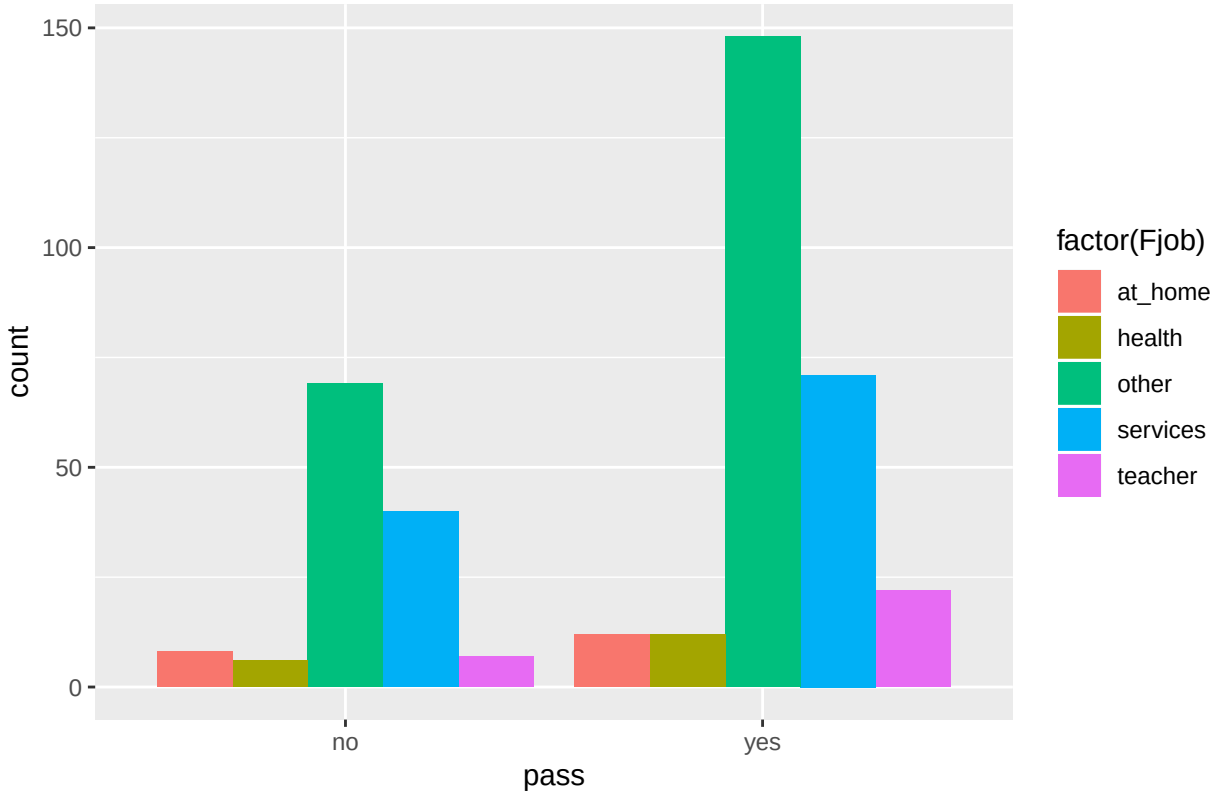

```
ggplot(math,mapping=aes(x=pass,fill=factor(Mjob))) +
  geom_bar(position="dodge") +
  ggtitle("Fig 6.2. Trabajo de la madre por aprobado y suspenso")
```

Fig 6.2. Trabajo de la madre por aprobado y suspenso

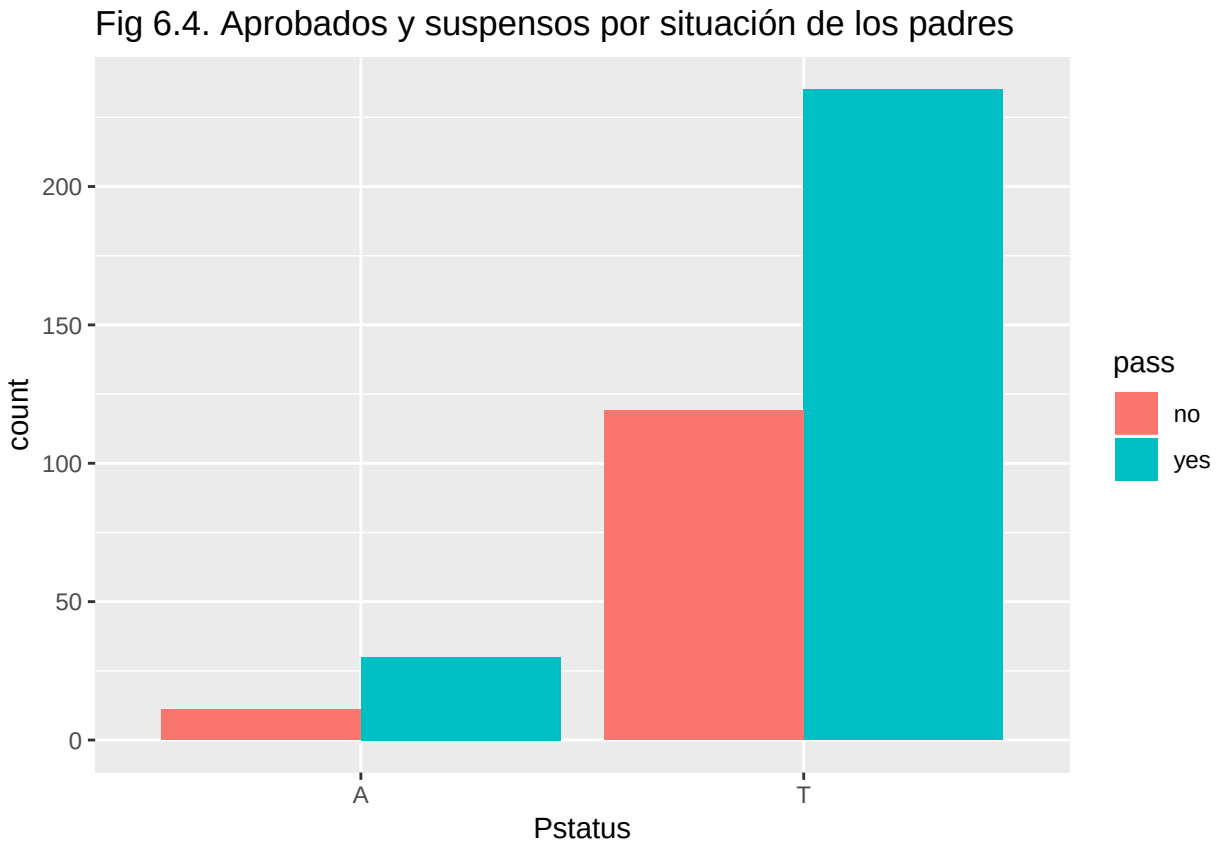



```
ggplot(math,mapping=aes(x=pass,fill=factor(Fjob))) +
  geom_bar(position="dodge") +
  ggtitle("Fig 6.3. Trabajo del padre por aprobado y suspenso")
```

Fig 6.3. Trabajo del padre por aprobado y suspenso




```
ggplot(math,mapping=aes(x=Pstatus,fill=pass)) +
  geom_bar(position="dodge") +
  ggtitle("Fig 6.4. Aprobados y suspensos por situación de los padres")
```

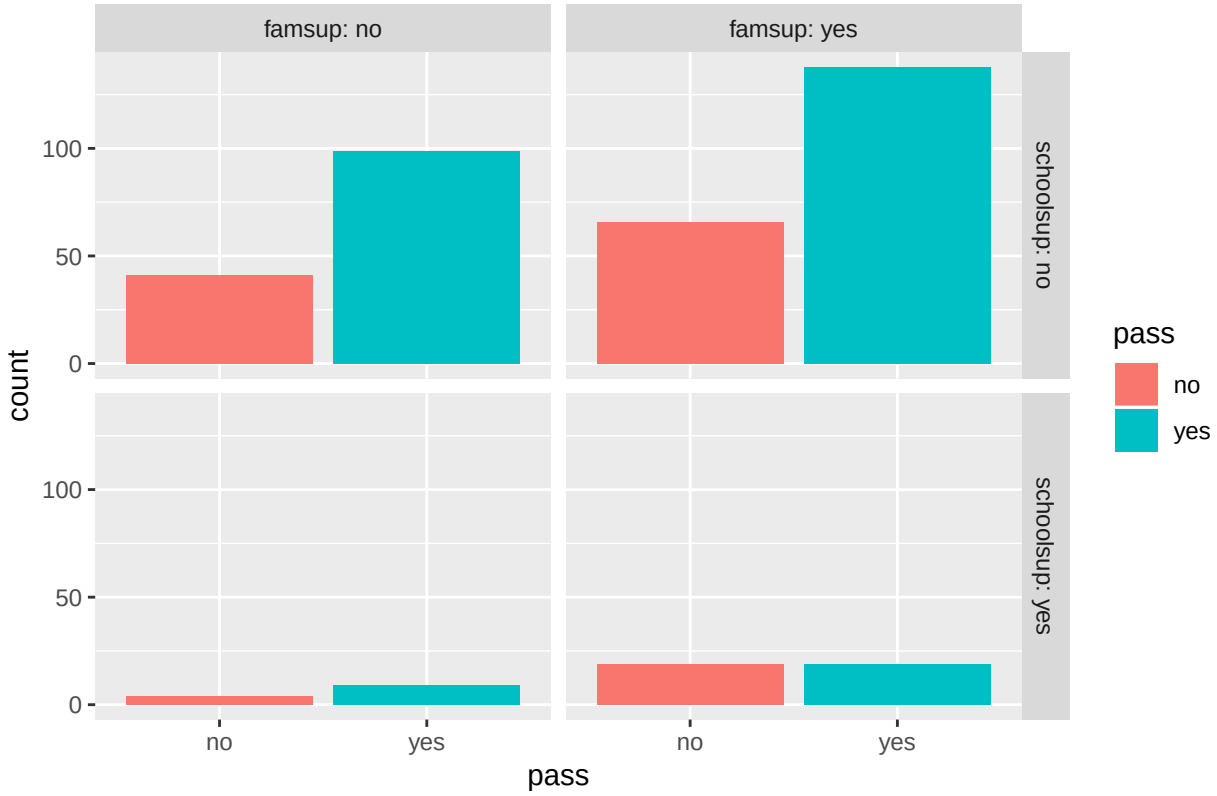


- Fig 6.1. muestra que las familias de más de tres miembros son mucho más comunes, y que las relaciones de calidad 4 (buenas) son más comunes en ambos tamaños.
- Fig 6.2. y 6.3. muestran que los trabajos más comunes son `other`, podría ser interesante ver el desglose de estos trabajos.
- En la Fig 6.4. se aprecia que lo más común son padres juntos, y no se observa una diferencia grande de distribución de aprobados por esta variable.

También se exploran variables relacionadas con la ayuda al alumno, tanto por parte de la escuela `schoolsup`, como su familia `famsup`, como clases extraescolares `paid`:

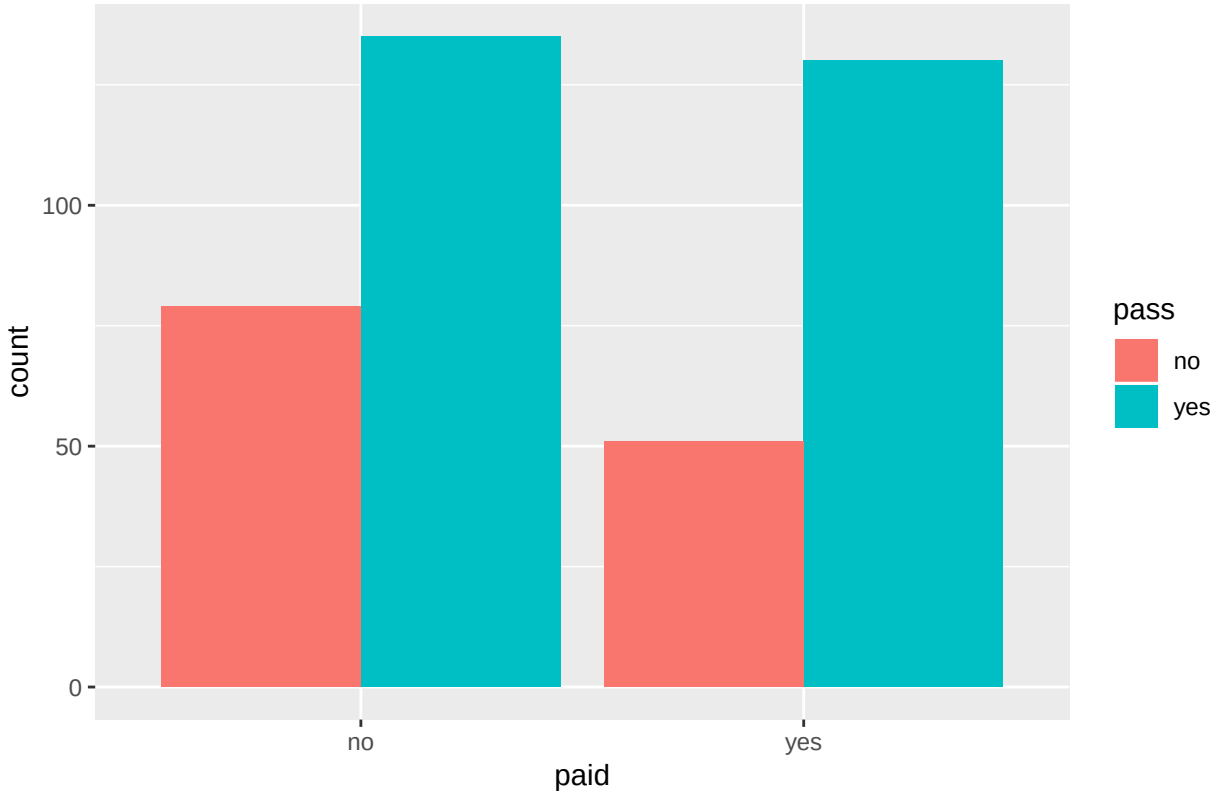
```
ggplot(math,mapping=aes(x=pass,fill=pass)) +
  geom_bar() +
  facet_grid(schoolsup ~ famsup, labeller = label_both) +
  ggtitle("Fig 7.1. Número de aprobados y suspensos por ayuda familiar y escolar")
```

Fig 7.1. Número de aprobados y suspensos por ayuda familiar y escolar




```
ggplot(math,mapping=aes(x=paid,fill=pass)) +
  geom_bar(position="dodge") +
  ggtitle("Fig 7.2. Aprobados y suspensos por asistencia a clases particulares o no")
```

Fig 7.2. Aprobados y suspensos por asistencia a clases particulares o no

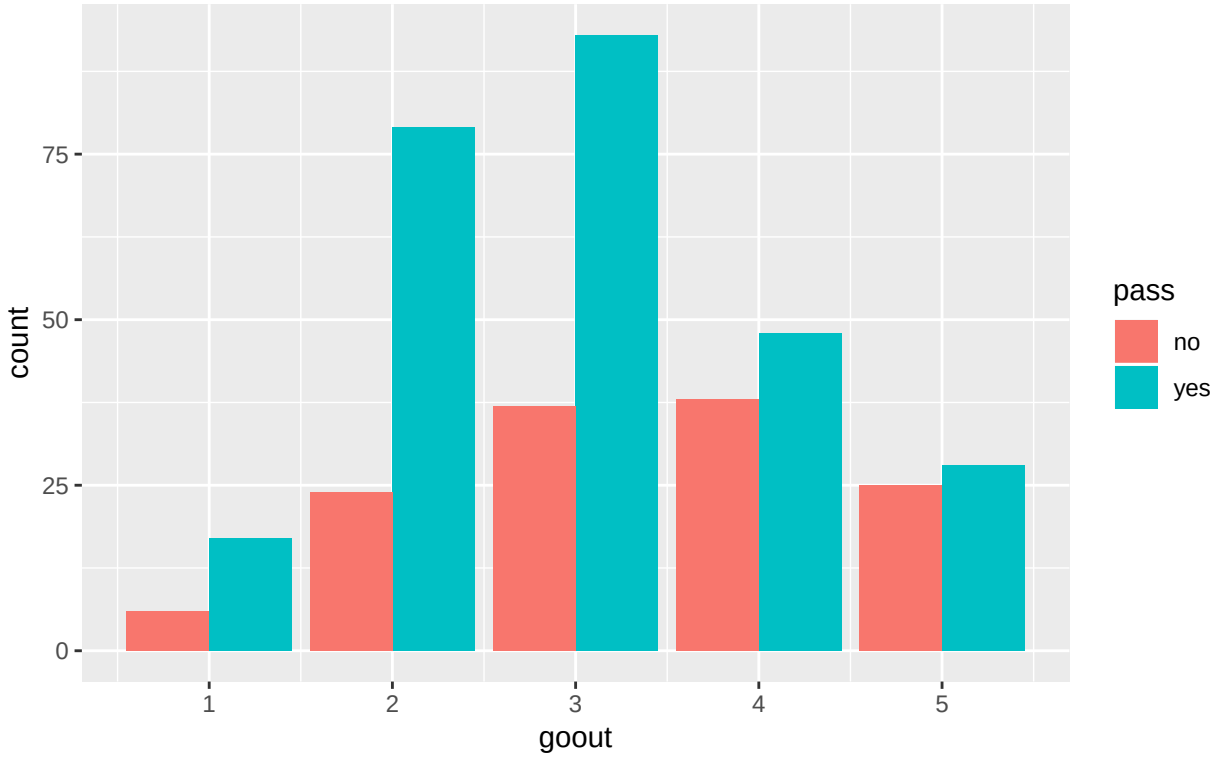


- Fig 7.1. muestra que muy pocos alumnos reciben ayuda del colegio, y entre aquellos que la reciben no parece mejorar el resultado, y de hecho parece haber un número de suspensos y aprobados similar. Respecto a la ayuda familiar, la mayoría de alumnos la tienen, pero tampoco parece resultar en una clara diferencia en aprobados.
- En Fig 7.2. se observa que las clases particulares si que parecen reducir el número de suspensos.

Respecto a las variables relacionadas con la vida social de los alumnos con sus notas, existen las variables `romantic`, `goout` y `activities`:

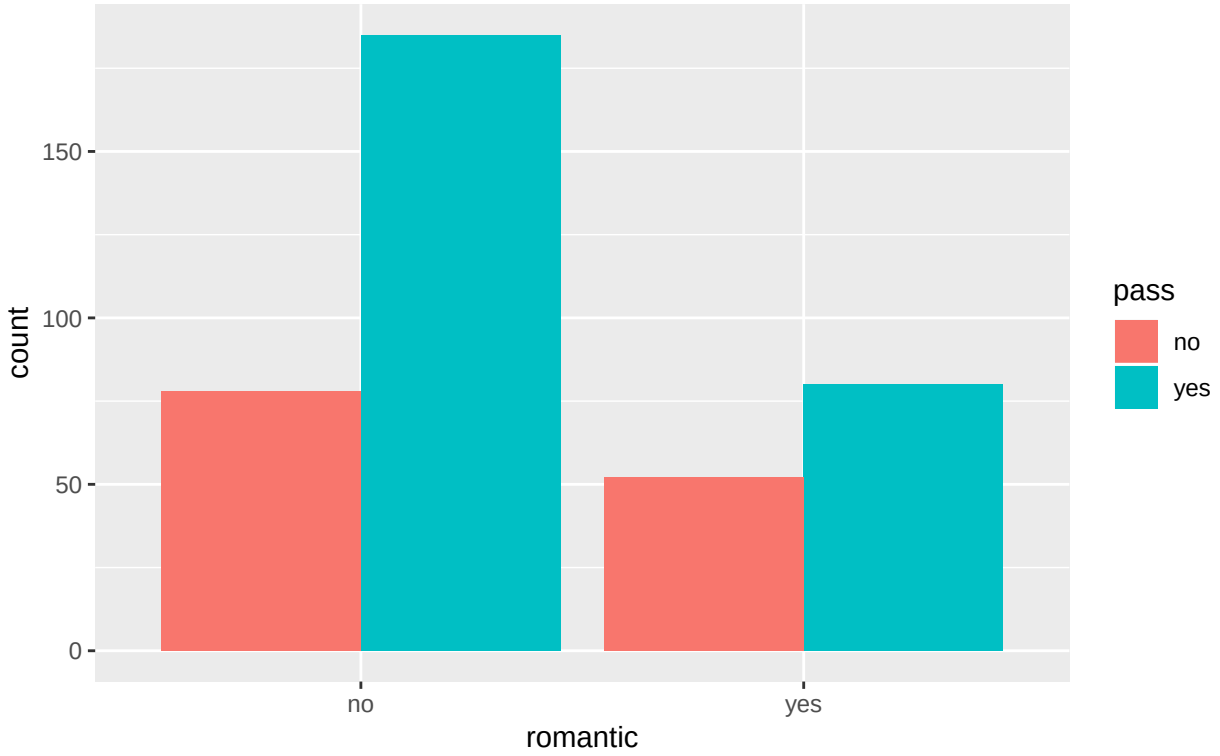
```
ggplot(math) +
  geom_bar( mapping=aes(x=goout, fill=pass ), position="dodge") +
  ggtitle("Fig 8.1. Número de alumnos aprobados y suspensos
          por la frecuencia con la que salen con amigos")
```

Fig 8.1. Número de alumnos aprobados y suspensos
por la frecuencia con la que salen con amigos

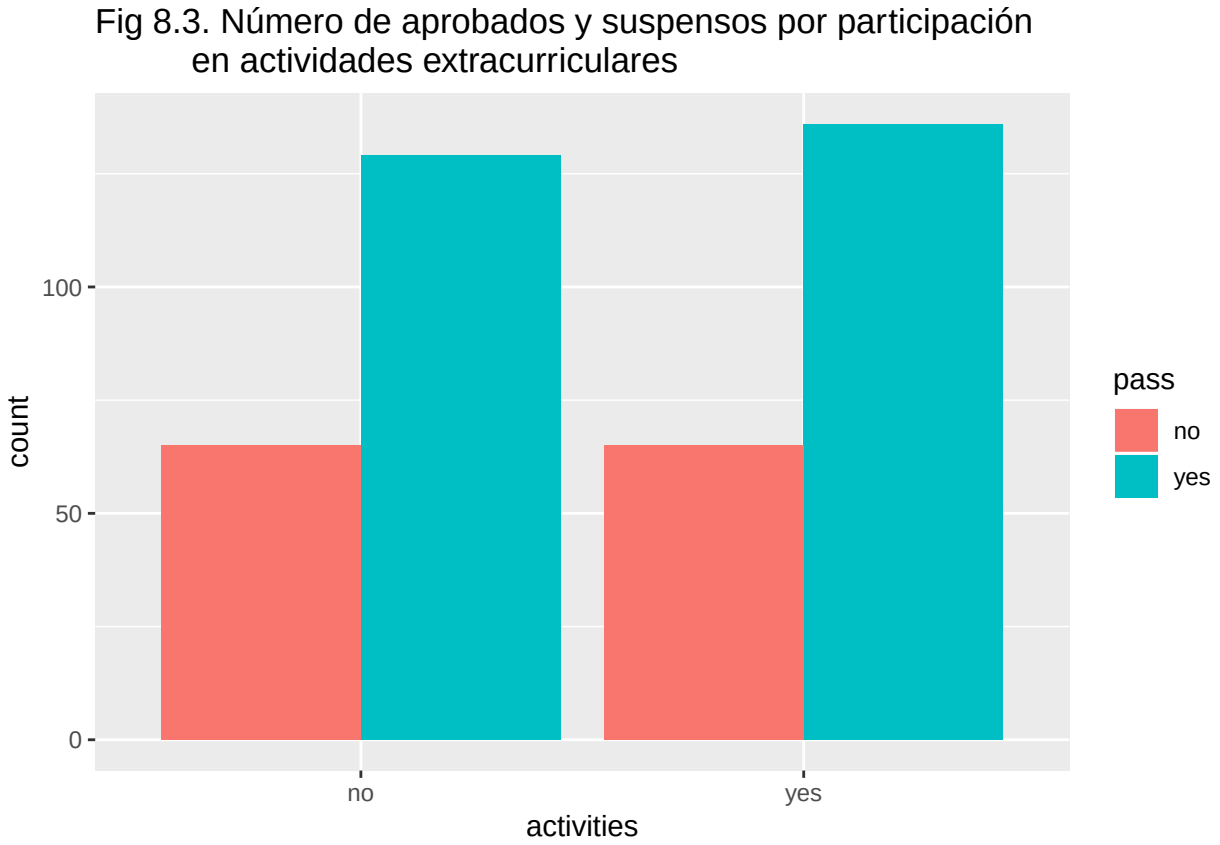



```
ggplot(math,mapping=aes(x=romantic,fill=pass)) +
  geom_bar(position="dodge") +
  ggtitle("Fig 8.2. Número de aprobados y suspensos según si el
          alumno tiene pareja")
```

Fig 8.2. Número de aprobados y suspensos según si el alumno tiene pareja




```
ggplot(math,mapping=aes(x=activities,fill=pass)) +
  geom_bar(position="dodge") +
  ggtitle("Fig 8.3. Número de aprobados y suspensos por participación
          en actividades extracurriculares")
```

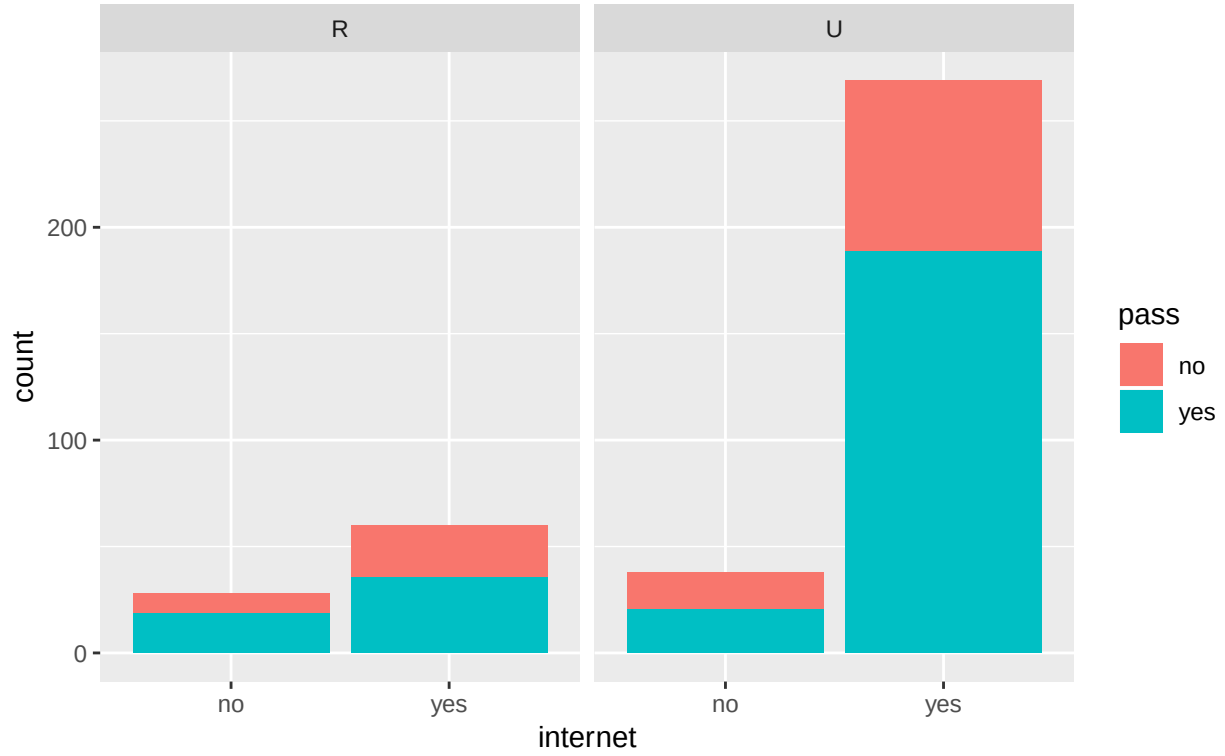


- Fig 8.1. Muestra que, mientras que la mayoría de alumnos salen con una frecuencia normal, frecuencias por encima parecen reducir en gran medida el número de aprobados.
- Fig 8.2. Muestra que la mayoría de alumnos no tienen pareja, pero que dentro de los que tienen pareja el número de aprobados es muy bajo, comparado con los alumnos sin pareja.
- Por el contrario Fig 8.3. mustra que las actividades extracurriculares no parecen incidir en los resultados del alumno.

Por último, se exploran otras variables, si la dirección en la que vive el alumno es rural o urbana `address` y si tiene acceso a `internet`:

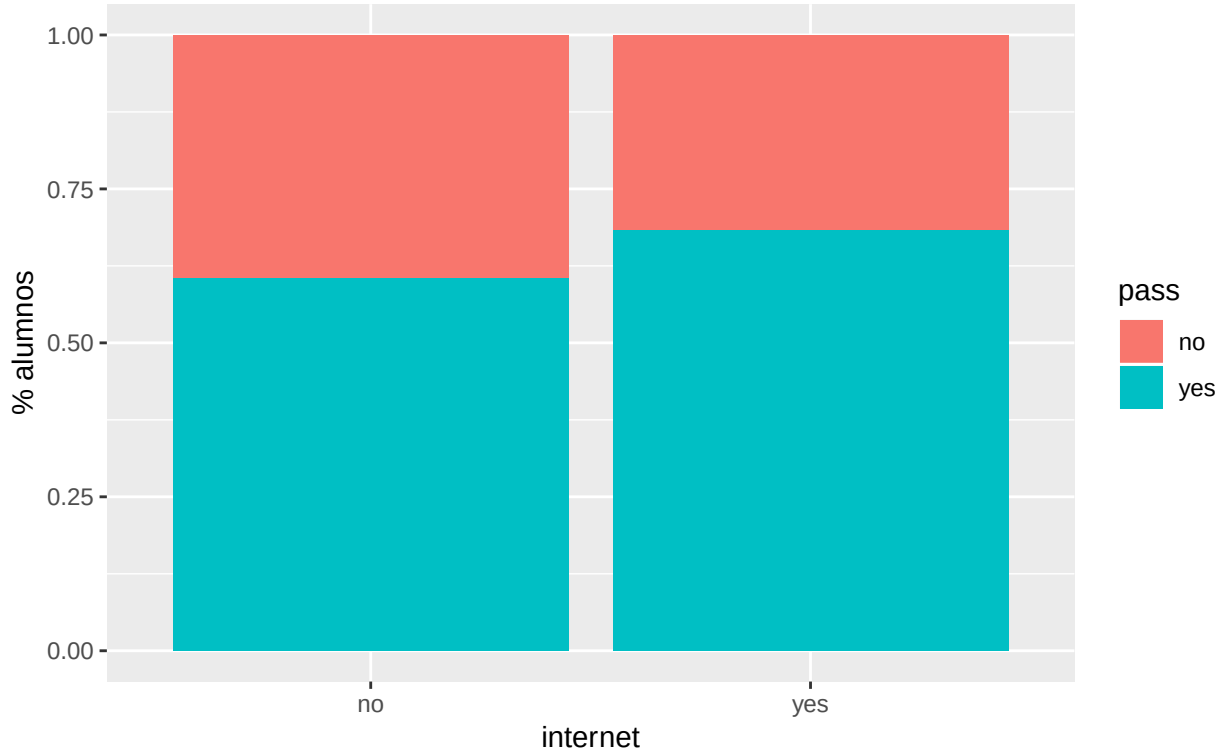
```
ggplot(math,mapping=aes(x=internet,fill=pass)) + facet_wrap(~address) +
  geom_bar() +
  ggtitle("Fig 9.1. Número de aprobados y suspensos según si el alumno tiene
          acceso a internet y zona")
```

Fig 9.1. Número de aprobados y suspensos según si el alumno tiene acceso a internet y zona




```
ggplot(math,mapping=aes(x=internet,fill=pass)) +
  geom_bar(position="fill") +
  ggtitle("Fig 9.2. Proporción de aprobados y suspensos según si tiene acceso a
          internet") +
  ylab("% alumnos")
```

Fig 9.2. Proporción de aprobados y suspensos según si tiene acceso a internet



- Fig 9.1. muestra que en zonas rurales la proporción de alumnos que no tienen acceso a internet es alta, mientras que en zonas urbanas es muy pequeña.
- En Fig 9.2. se puede apreciar cierto aumento en la proporción de alunmos suspensos cuando no tienen acceso a internet, aunque como se ha visto en la anterior figura, la mayoría de alumnos tienen acceso a internet, por lo que estos resultados pueden ser engañosos.

Modelos no supervisados

Eliminamos las variables `pass` y `G3`, ya que estos modelos serán no supervisados.

```
  x <- select(math, -pass, -G3)
```

La biblioteca `cluster` proporciona varios tipos de clustering. Probamos `clara`, que divide en k clusters y es útil para datasets grandes.

```
library(cluster)

clusters <- clara(x, 2, samples = 100)

clusters$clusinfo
```

```
##      size max_diss  av_diss isolation
## [1,]  284 14.10674 7.280026  1.298630
## [2,]  111 63.34824 9.627642  5.831679
```

```
math$cluster <- clusters$clustering
```

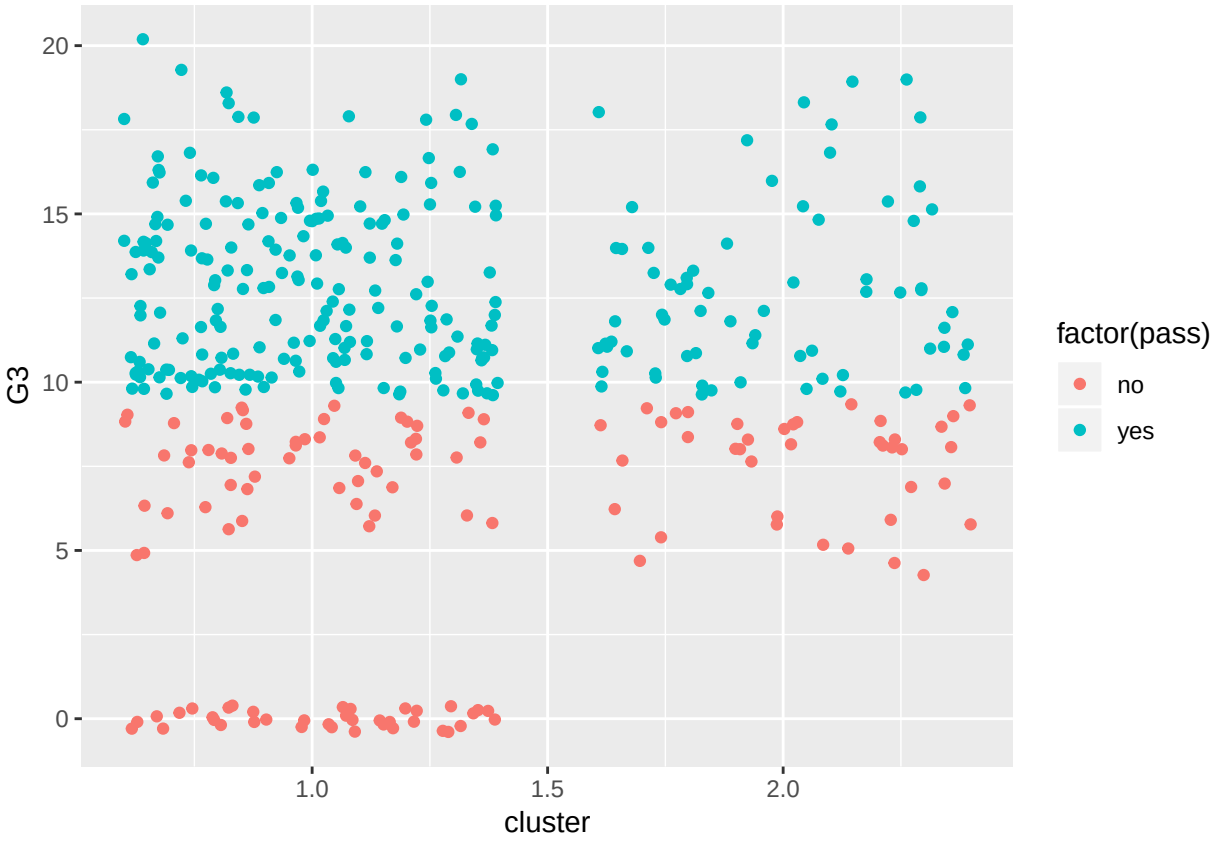
Podemos comprobar si la separación en clusters se parece a las clases asignadas:

```
library(caret)

truth1 <- factor(math$pass,levels=c("yes","no"))
pred1 <- factor(math$cluster,labels=c("yes","no"))

xtab1 <- table(pred1,truth1)

confusionMatrix(xtab1)
```

```
## Confusion Matrix and Statistics
##
##      truth1
## pred1 yes  no
##   yes 194  90
##   no   71  40
##
##                Accuracy : 0.5924
##                  95% CI : (0.5421, 0.6413)
##     No Information Rate : 0.6709
##     P-Value [Acc > NIR] : 0.9995
##
##                   Kappa : 0.0413
##
##  Mcnemar's Test P-Value : 0.1560
##
##             Sensitivity : 0.7321
##             Specificity : 0.3077
##          Pos Pred Value : 0.6831
##          Neg Pred Value : 0.3604
##              Prevalence : 0.6709
##          Detection Rate : 0.4911
##    Detection Prevalence : 0.7190
##       Balanced Accuracy : 0.5199
##
##        'Positive' Class : yes
##
```

Como podemos ver, casi un 60% de las instancias se clasifican correctamente simplemente utilizando clustering. Podemos representar los clusters contra la nota `G3`, para ver si existe alguna relación:

```
ggplot(math) +
  geom_jitter(mapping=aes(x=cluster,y=G3,color=factor(pass)))
```



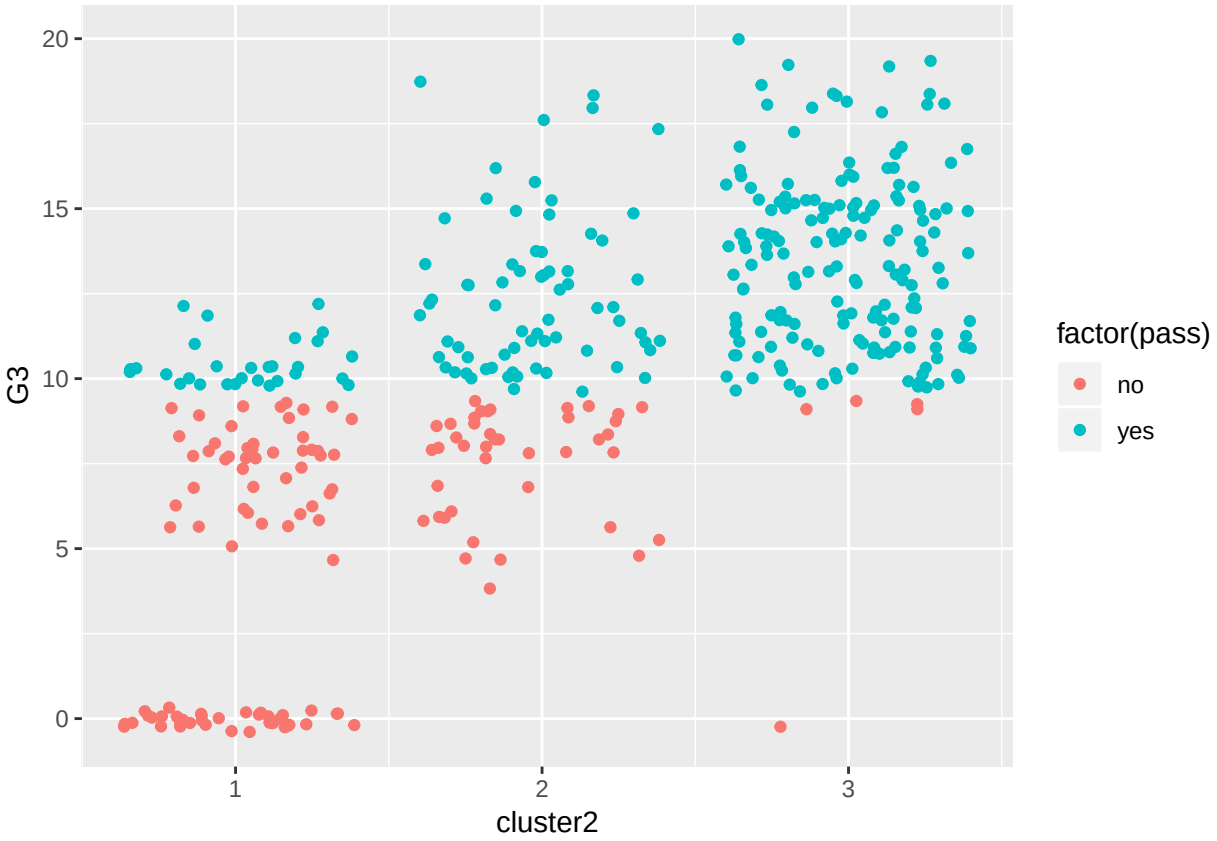
No se observa relación, aunque es interesante ver cómo todos los alumnos con nota final 0 están en el mismo cluster. Tratamos de crear 3 clústers, para ver si esos datos se separan completamente:

```
clusters <- clara(x, 3, samples = 100)

clusters$clusinfo
```

```
##      size  max_diss  av_diss isolation
## [1,]  113  9.848858 6.601493  1.152722
## [2,]  110 63.348244 9.610452  6.012749
## [3,]  172 10.677078 6.137257  1.249657
```

```
math$cluster2 <- clusters$clustering
```

```
ggplot(math) +
  geom_jitter(mapping=aes(x=cluster2,y=G3,color=factor(pass)))
```



No se ha conseguido separar los valores con G3=0 (de hecho, uno de ellos ha pasado al clúster 3), pero se puede apreciar cierta separación entre los tres clústers, con los alumnos en el clúster 1 teniendo notas bajas, casi todas suspensos, y el clúster 3 notas casi todas por encima de 10, excepto por el 0 anteriormente detectado. Por tanto, con 3 clústers, este modelo ya tiene cierta capacidad predictiva de la variable aprobado.

Sin embargo, lo interesante del clústering es su poder de dividir los alumnos en `k` grupos característicos. En este caso, juzgando por las notas, podrían ser alumnos con ciertas características que los dividen en buenos, mediocres, y malos estudiantes. Si esto fuera así, podría ser interesante explorar qué ha pasado con el alumno en el clúster 3 que ha obtenido una nota final de 0.

Modelos supervisados

Regresión lineal

Se puede utilizar regresión lineal para intentar predecir la variable `G3` a partir de los datos.

```
# eliminamos la variable pass (basada en G3) y las cluster obtenidas anteriormente
lm_math <- select(math,-pass,-cluster,-cluster2)

linear_reg_model <- lm(G3 ~ .,data=lm_math)
summary(linear_reg_model)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = lm_math)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9339 -0.5532  0.2680  0.9689  4.6461
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.115488   2.116958  -0.527 0.598573
## schoolMS          0.480742   0.366512   1.312 0.190485
## sexM              0.174396   0.233588   0.747 0.455805
## age              -0.173302   0.100780  -1.720 0.086380 .
## addressU          0.104455   0.270791   0.386 0.699922
## famsizeLE3        0.036512   0.226680   0.161 0.872128
## PstatusT         -0.127673   0.335626  -0.380 0.703875
## Medu              0.129685   0.149999   0.865 0.387859
## Fedu             -0.133940   0.128768  -1.040 0.298974
## Mjobhealth       -0.146426   0.518491  -0.282 0.777796
## Mjobother         0.074088   0.332044   0.223 0.823565
## Mjobservices      0.046956   0.369587   0.127 0.898973
## Mjobteacher      -0.026276   0.481632  -0.055 0.956522
## Fjobhealth        0.330948   0.666601   0.496 0.619871
## Fjobother        -0.083582   0.476796  -0.175 0.860945
## Fjobservices     -0.322142   0.493265  -0.653 0.514130
## Fjobteacher      -0.112364   0.601448  -0.187 0.851907
## reasonhome       -0.209183   0.256392  -0.816 0.415123
## reasonother       0.307554   0.380214   0.809 0.419120
## reasonreputation  0.129106   0.267254   0.483 0.629335
## guardianmother    0.195741   0.252672   0.775 0.439046
## guardianother     0.006565   0.463650   0.014 0.988710
## traveltime        0.096994   0.157800   0.615 0.539170
## studytime        -0.104754   0.134814  -0.777 0.437667
## failures         -0.160539   0.161006  -0.997 0.319399
## schoolsupyes      0.456448   0.319538   1.428 0.154043
## famsupyes         0.176870   0.224204   0.789 0.430710
## paidyes           0.075764   0.222100   0.341 0.733211
## activitiesyes    -0.346047   0.205938  -1.680 0.093774 .
## nurseryyes       -0.222716   0.254184  -0.876 0.381518
## higheryes         0.225921   0.500398   0.451 0.651919
## internetyes      -0.144462   0.287528  -0.502 0.615679
## romanticyes      -0.272008   0.219732  -1.238 0.216572
## famrel            0.356876   0.114124   3.127 0.001912 **
## freetime          0.047002   0.110209   0.426 0.670021
## goout             0.012007   0.105230   0.114 0.909224
## Dalc             -0.185019   0.153124  -1.208 0.227741
## Walc              0.176772   0.114943   1.538 0.124966
## health            0.062995   0.074800   0.842 0.400259
## absences          0.045879   0.013412   3.421 0.000698 ***
## G1                0.188847   0.062373   3.028 0.002645 **
## G2                0.957330   0.053460  17.907  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.901 on 353 degrees of freedom
## Multiple R-squared:  0.8458, Adjusted R-squared:  0.8279
```

```
## F-statistic: 47.21 on 41 and 353 DF,  p-value: < 2.2e-16
```

Como se puede ver a partir de los p-valores, las variables estadísticamente importantes a un 95% de confianza son `G2`, `absences`, `G1` y `famrel`, y al 90% `activities` y `age`. Un experimento más interesante puede ser intentar predecir la nota final a la mitad del curso, suponiendo que eso implica que se sabe la nota del primer periodo pero no del segundo:

```
lm_math2 <- select(lm_math,-G2)

linear_reg_model2 <- lm(G3 ~ .,data=lm_math2)
summary(linear_reg_model2)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = lm_math2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8485 -1.0234  0.2375  1.6580  5.7611
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.461954   2.913605   0.502  0.61614
## schoolMS          0.714499   0.505286   1.414  0.15823
## sexM              0.270532   0.322152   0.840  0.40161
## age              -0.297433   0.138697  -2.144  0.03268 *
## addressU          0.384201   0.372937   1.030  0.30362
## famsizeLE3        0.226821   0.312364   0.726  0.46823
## PstatusT         -0.491222   0.462151  -1.063  0.28855
## Medu              0.326059   0.206371   1.580  0.11501
## Fedu             -0.264031   0.177353  -1.489  0.13745
## Mjobhealth       -0.029068   0.715205  -0.041  0.96760
## Mjobother         0.508609   0.456833   1.113  0.26632
## Mjobservices      0.140905   0.509797   0.276  0.78241
## Mjobteacher      -0.218052   0.664252  -0.328  0.74290
## Fjobhealth        0.961408   0.918298   1.047  0.29584
## Fjobother         0.638958   0.655384   0.975  0.33026
## Fjobservices      0.636653   0.676442   0.941  0.34726
## Fjobteacher       0.009702   0.829649   0.012  0.99068
## reasonhome       -0.105154   0.353605  -0.297  0.76635
## reasonother       0.978041   0.521960   1.874  0.06178 .
## reasonreputation  0.120609   0.368679   0.327  0.74376
## guardianmother    0.014084   0.348283   0.040  0.96777
## guardianother    -0.210772   0.639391  -0.330  0.74186
## traveltime       -0.212412   0.216377  -0.982  0.32693
## studytime        -0.121165   0.185973  -0.652  0.51513
## failures         -0.266456   0.221959  -1.200  0.23076
## schoolsupyes      1.039899   0.438509   2.371  0.01825 *
## famsupyes         0.223607   0.309270   0.723  0.47015
## paidyes           0.453304   0.305005   1.486  0.13811
## activitiesyes    -0.271051   0.284034  -0.954  0.34059
## nurseryyes       -0.210118   0.350648  -0.599  0.54940
## higheryes         0.105435   0.690241   0.153  0.87868
```

```
## internetyes       0.214865   0.395680   0.543  0.58745
## romanticyes      -0.860234   0.299715  -2.870  0.00435 **
## famrel            0.203011   0.156988   1.293  0.19680
## freetime          0.019810   0.152020   0.130  0.89639
## goout            -0.134966   0.144723  -0.933  0.35167
## Dalc             -0.202195   0.211232  -0.957  0.33911
## Walc              0.291488   0.158318   1.841  0.06644 .
## health            0.009022   0.103104   0.088  0.93032
## absences          0.042675   0.018500   2.307  0.02164 *
## G1                1.109070   0.048766  22.743  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.622 on 354 degrees of freedom
## Multiple R-squared:  0.7057, Adjusted R-squared:  0.6724
## F-statistic: 21.22 on 40 and 354 DF,  p-value: < 2.2e-16
```

Como se puede ver, esto reduce a 0.67 el R-squared ajustado, con nuevas variables con p-valor menor que 0.05: `romantic` y `schoolsup`. Si no se consideran ninguna de las dos notas:

```
lm_math3 <- select(lm_math2,-G1)

linear_reg_model3 <- lm(G3 ~ .,data=lm_math3)
summary(linear_reg_model3)
```

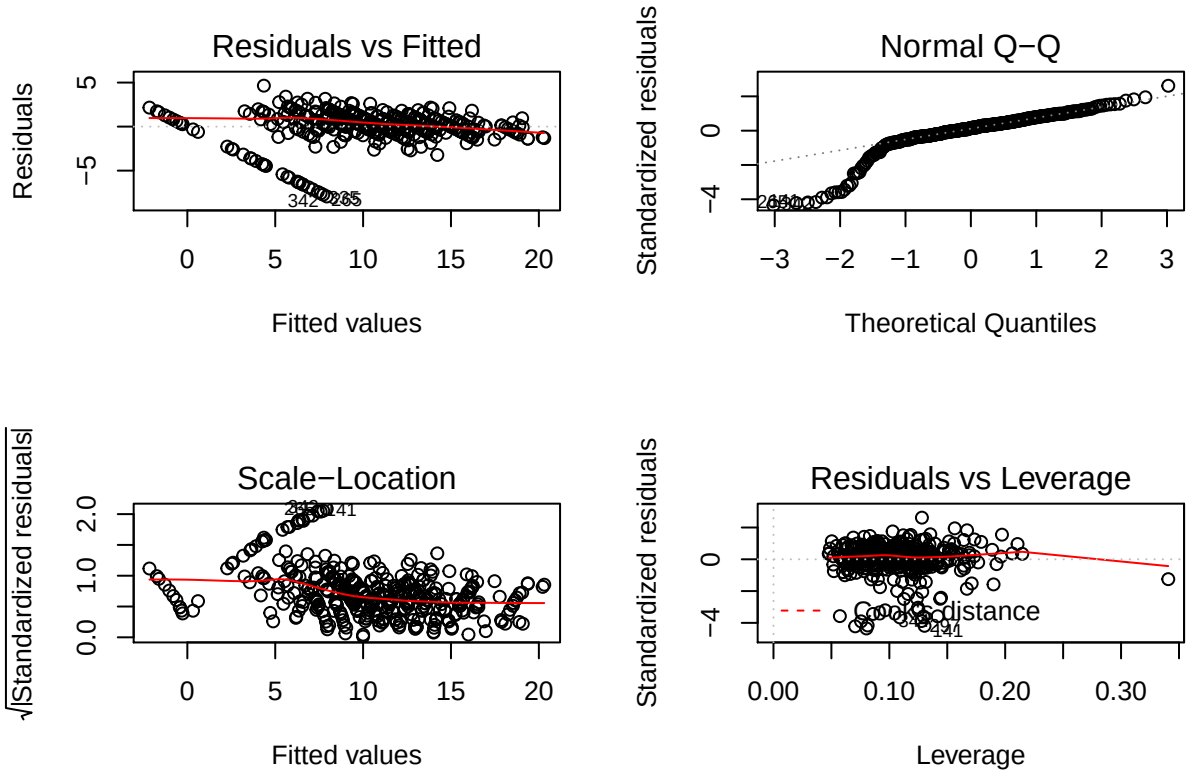
```
##
## Call:
## lm(formula = G3 ~ ., data = lm_math3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0442  -1.9028   0.4289   2.7570   8.8874
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.07769    4.48089   3.142  0.00182 **
## schoolMS          0.72555    0.79157   0.917  0.35997
## sexM              1.26236    0.50003   2.525  0.01202 *
## age              -0.37516    0.21721  -1.727  0.08501 .
## addressU          0.55135    0.58412   0.944  0.34586
## famsizeLE3        0.70281    0.48824   1.439  0.15090
## PstatusT         -0.32010    0.72390  -0.442  0.65862
## Medu              0.45687    0.32317   1.414  0.15833
## Fedu             -0.10458    0.27762  -0.377  0.70663
## Mjobhealth        0.99808    1.11819   0.893  0.37268
## Mjobother        -0.35900    0.71316  -0.503  0.61500
## Mjobservices      0.65832    0.79784   0.825  0.40985
## Mjobteacher      -1.24149    1.03821  -1.196  0.23257
## Fjobhealth        0.34767    1.43796   0.242  0.80909
## Fjobother        -0.61967    1.02304  -0.606  0.54509
## Fjobservices     -0.46577    1.05697  -0.441  0.65972
## Fjobteacher       1.32619    1.29654   1.023  0.30707
## reasonhome        0.07851    0.55380   0.142  0.88735
```

```
## reasonother       0.77707    0.81757   0.950  0.34252
## reasonreputation  0.61304    0.57657   1.063  0.28839
## guardianmother    0.06978    0.54560   0.128  0.89830
## guardianother     0.75010    0.99946   0.751  0.45345
## traveltime       -0.24027    0.33897  -0.709  0.47889
## studytime         0.54952    0.28765   1.910  0.05690 .
## failures         -1.72398    0.33291  -5.179 3.75e-07 ***
## schoolsupyes     -1.35058    0.66693  -2.025  0.04361 *
## famsupyes        -0.86182    0.47869  -1.800  0.07265 .
## paidyes           0.33975    0.47775   0.711  0.47746
## activitiesyes    -0.32953    0.44494  -0.741  0.45942
## nurseryyes       -0.17730    0.54931  -0.323  0.74706
## higheryes         1.37045    1.07780   1.272  0.20437
## internetyes       0.49813    0.61956   0.804  0.42192
## romanticyes      -1.09449    0.46925  -2.332  0.02024 *
## famrel            0.23155    0.24593   0.942  0.34706
## freetime          0.30242    0.23735   1.274  0.20345
## goout            -0.59367    0.22451  -2.644  0.00855 **
## Dalc             -0.27223    0.33087  -0.823  0.41120
## Walc              0.26339    0.24801   1.062  0.28896
## health           -0.17678    0.16101  -1.098  0.27297
## absences          0.05629    0.02897   1.943  0.05277 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.108 on 355 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.196
## F-statistic: 3.463 on 39 and 355 DF,  p-value: 3.317e-10
```

Como se puede apreciar, el fit del modelo a los datos ha caído en picado, con valores de R-squared ajustado de 0.2. Por tanto, sin acceso a las notas de los dos periodos, la regresión lineal no es apropiada para este problema.

Podemos observar las hipótesis del primer modelo:

```
par(mfrow=c(2,2))
plot(linear_reg_model)
```



En estos gráficos se observan ciertos valores extraños alrededor de las notas de 0, así como los residuos negativos. Se podría sospechar que dichos valores se corresponden a aquellos valores extraños que comentábamos en la exploración de variables. Los podemos eliminar y volver a analizar el modelo:

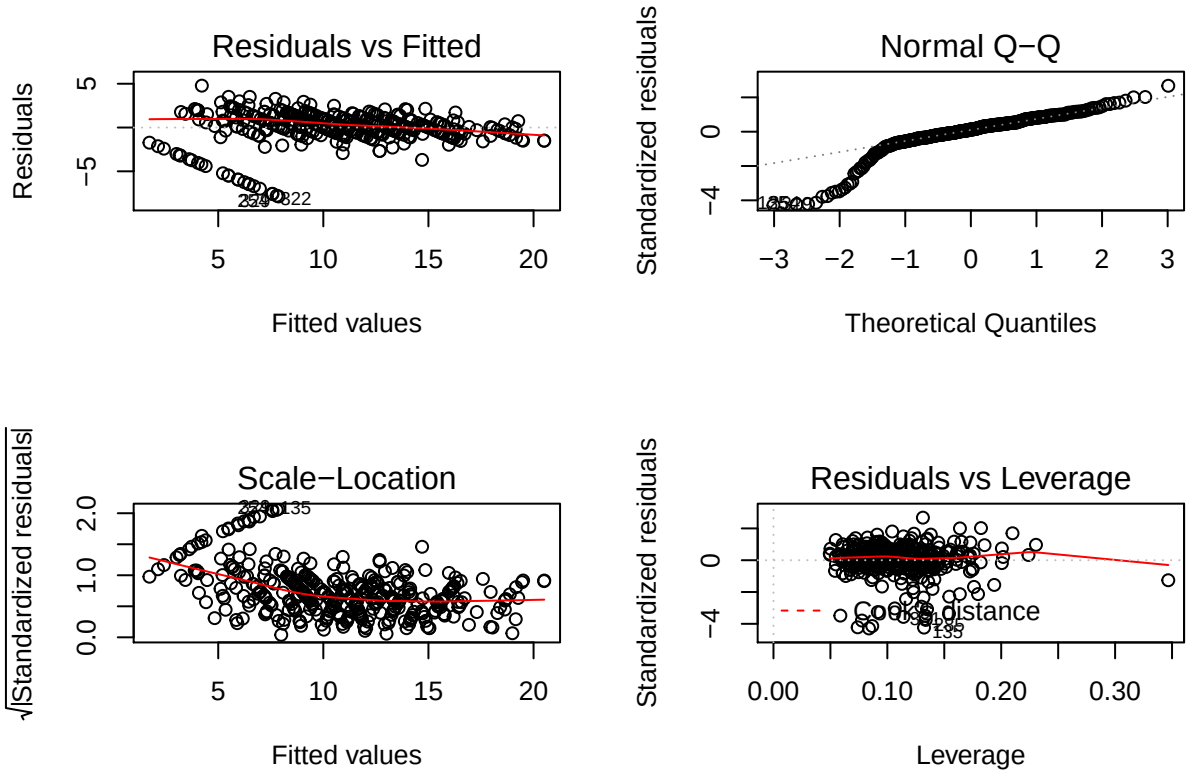
```
lm_clean <- filter(math, G2 > 0) %>% select(-pass,-cluster,-cluster2)

linear_reg_model_clean <- lm(G3 ~ .,data=lm_clean)
summary(linear_reg_model_clean)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = lm_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8715 -0.6064  0.2306  0.9580  4.7790
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.133312   2.231845  -0.956 0.339826
## schoolMS          0.564562   0.373509   1.512 0.131588
## sexM              0.204691   0.239819   0.854 0.393971
## age              -0.145764   0.106383  -1.370 0.171534
## addressU          0.108941   0.278220   0.392 0.695626
## famsizeLE3        0.038412   0.230670   0.167 0.867845
## PstatusT         -0.080350   0.343801  -0.234 0.815351
```

```
## Medu              0.085038   0.153433   0.554 0.579782
## Fedu             -0.125824   0.131716  -0.955 0.340120
## Mjobhealth       -0.163640   0.527384  -0.310 0.756533
## Mjobother        -0.025729   0.343047  -0.075 0.940259
## Mjobservices     -0.056506   0.381495  -0.148 0.882338
## Mjobteacher      -0.016168   0.495359  -0.033 0.973982
## Fjobhealth        0.492556   0.689390   0.714 0.475420
## Fjobother         0.066048   0.502746   0.131 0.895557
## Fjobservices     -0.171772   0.519741  -0.330 0.741229
## Fjobteacher       0.055639   0.636046   0.087 0.930344
## reasonhome       -0.236874   0.265282  -0.893 0.372536
## reasonother       0.281377   0.384432   0.732 0.464716
## reasonreputation  0.158881   0.274358   0.579 0.562906
## guardianmother    0.175665   0.257156   0.683 0.495006
## guardianother    -0.091019   0.481167  -0.189 0.850078
## traveltime        0.095241   0.163159   0.584 0.559790
## studytime        -0.080701   0.138139  -0.584 0.559472
## failures         -0.174057   0.171306  -1.016 0.310324
## schoolsupyes      0.530585   0.325042   1.632 0.103530
## famsupyes         0.144496   0.229689   0.629 0.529710
## paidyes           0.099035   0.225410   0.439 0.660685
## activitiesyes    -0.348224   0.212763  -1.637 0.102624
## nurseryyes       -0.187337   0.261886  -0.715 0.474892
## higheryes         0.417862   0.530146   0.788 0.431127
## internetyes      -0.239215   0.295979  -0.808 0.419531
## romanticyes      -0.288717   0.224931  -1.284 0.200163
## famrel            0.352995   0.115670   3.052 0.002454 **
## freetime          0.057603   0.112163   0.514 0.607893
## goout            -0.005975   0.108225  -0.055 0.956006
## Dalc             -0.179226   0.155024  -1.156 0.248445
## Walc              0.198657   0.117164   1.696 0.090888 .
## health            0.057348   0.075732   0.757 0.449427
## absences          0.053985   0.013905   3.882 0.000124 ***
## G1                0.099536   0.079592   1.251 0.211948
## G2                1.072006   0.080437  13.327  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.917 on 340 degrees of freedom
## Multiple R-squared:  0.8166, Adjusted R-squared:  0.7945
## F-statistic: 36.93 on 41 and 340 DF,  p-value: < 2.2e-16
```

Se puede observar una reducción en el R-squared, y como la variable `G1` ha dejado de ser relevante. Observamos las hipótesis:

```
par(mfrow=c(2,2))
plot(linear_reg_model_clean)
```



Como se puede ver nuestra suposición estaba equivocada, ya que siguen exisistiendo residuos con falta de normalidad.

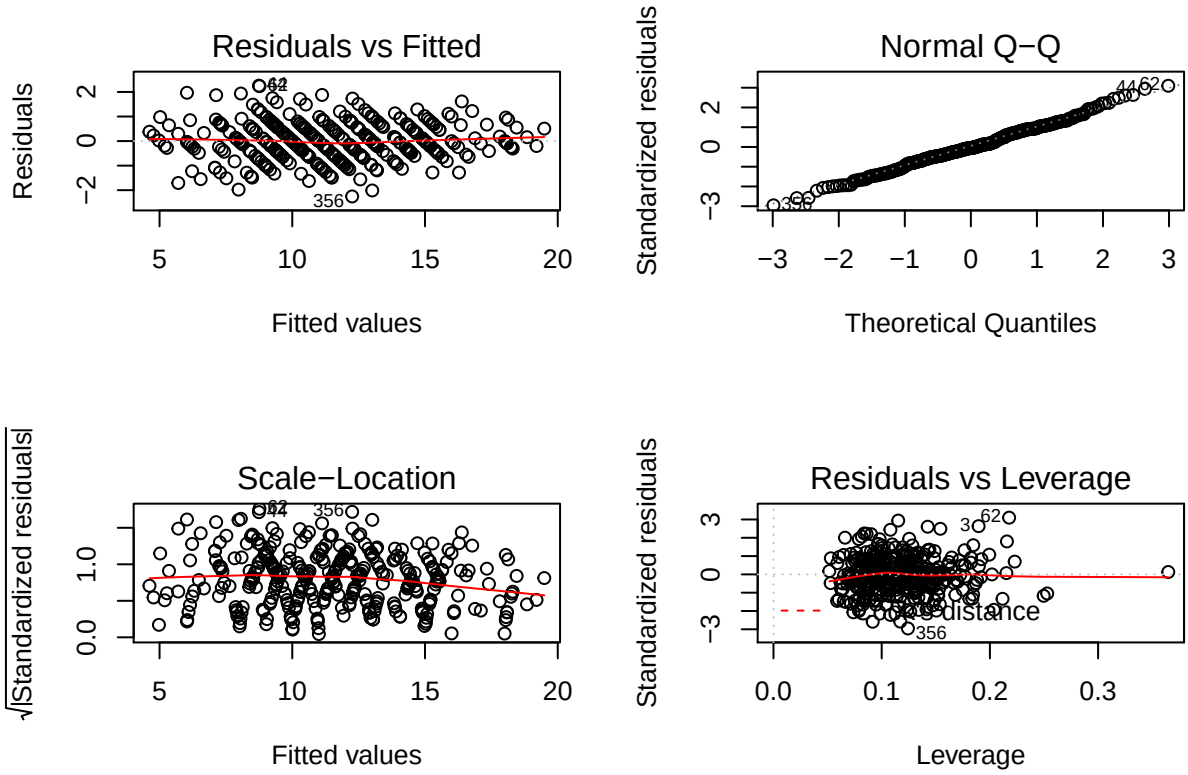
Podemos tratar de eliminar valores de `G3` igual a cero:

```
math_no0 <- filter(math,G3>0)

lm_trans <- select(math_no0, -pass,-cluster,-cluster2)

linear_reg_model_no0 <- lm(G3 ~ .,data=lm_trans)
summary(linear_reg_model_no0)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = lm_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25528 -0.46157 -0.04264  0.55154  2.25586
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.160153   0.994525  -0.161 0.872169
## schoolMS         -0.129408   0.166194  -0.779 0.436767
## sexM             -0.044812   0.104944  -0.427 0.669661
## age               0.040123   0.046967   0.854 0.393599
```

```
## addressU          0.156908   0.122269   1.283 0.200332
## famsizeLE3       -0.083691   0.101541  -0.824 0.410442
## PstatusT         -0.202391   0.149268  -1.356 0.176106
## Medu             -0.052943   0.066814  -0.792 0.428727
## Fedu             -0.027691   0.057855  -0.479 0.632539
## Mjobhealth        0.257715   0.235141   1.096 0.273915
## Mjobother        -0.179800   0.155994  -1.153 0.249946
## Mjobservices      0.056230   0.173034   0.325 0.745422
## Mjobteacher       0.236072   0.219856   1.074 0.283752
## Fjobhealth        0.225379   0.298382   0.755 0.450611
## Fjobother         0.263495   0.219969   1.198 0.231867
## Fjobservices      0.156067   0.227829   0.685 0.493837
## Fjobteacher       0.261549   0.278090   0.941 0.347672
## reasonhome        0.182901   0.118158   1.548 0.122641
## reasonother       0.017074   0.168701   0.101 0.919451
## reasonreputation  0.076043   0.121399   0.626 0.531515
## guardianmother   -0.015692   0.113625  -0.138 0.890244
## guardianother    -0.352319   0.219783  -1.603 0.109930
## traveltime        0.054524   0.071707   0.760 0.447600
## studytime         0.032501   0.061682   0.527 0.598630
## failures          0.036191   0.081554   0.444 0.657519
## schoolsupyes     -0.162106   0.142391  -1.138 0.255796
## famsupyes         0.095190   0.101215   0.940 0.347693
## paidyes          -0.206912   0.098395  -2.103 0.036271 *
## activitiesyes    -0.015363   0.094662  -0.162 0.871176
## nurseryyes       -0.165259   0.117074  -1.412 0.159064
## higheryes        -0.039675   0.255660  -0.155 0.876775
## internetyes       0.005063   0.131489   0.039 0.969311
## romanticyes       0.015926   0.100297   0.159 0.873941
## famrel            0.176589   0.052119   3.388 0.000793 ***
## freetime         -0.016981   0.049787  -0.341 0.733272
## goout            -0.093508   0.049473  -1.890 0.059664 .
## Dalc              0.030877   0.067913   0.455 0.649672
## Walc             -0.003257   0.052576  -0.062 0.950646
## health           -0.080123   0.033536  -2.389 0.017473 *
## absences         -0.012375   0.006263  -1.976 0.049053 *
## G1                0.100369   0.034602   2.901 0.003985 **
## G2                0.881188   0.035636  24.728  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8158 on 315 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9361
## F-statistic: 128.2 on 41 and 315 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(linear_reg_model_no0)
```



Como se puede observar, esto produce un muy buen R-squared además de corregir los errores en la falta de normalidad y heterocedasticidad.

Decision Trees

```
math_for_tree <- select(math,-G3)

#la clase debe ser de tipo factor para utilizar ctree
math_for_tree$pass <- as.factor(math_for_tree$pass)


set.seed(12345)
train_idx <- sample(1:395,size=(300),replace=F)
train_math <- math_for_tree[train_idx,]
test_math <- math_for_tree[-train_idx,]

summary(train_math)
```

```
##  school   sex          age        address famsize   Pstatus
##  GP:269   F:157   Min.   :15.00   R: 62   GT3:214   A: 32
##  MS: 31   M:143   1st Qu.:16.00   U:238   LE3: 86   T:268
##                   Median :17.00
##                   Mean   :16.63
##                   3rd Qu.:18.00
```

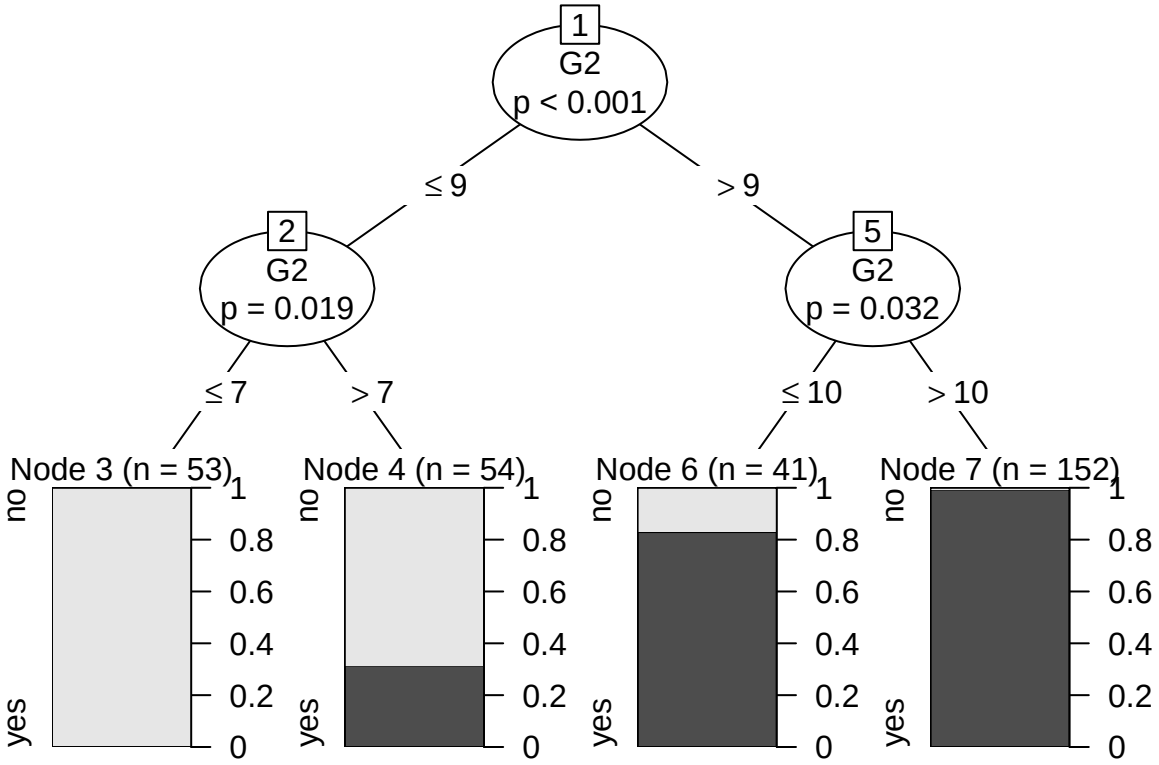
```
##                   Max.   :22.00
##       Medu            Fedu             Mjob           Fjob
##  Min.   :0.000   Min.   :0.000   at_home : 47   at_home : 16
##  1st Qu.:2.000   1st Qu.:2.000   health  : 25   health  : 15
##  Median :3.000   Median :3.000   other   :106   other   :167
##  Mean   :2.767   Mean   :2.527   services: 78   services: 84
##  3rd Qu.:4.000   3rd Qu.:3.000   teacher : 44   teacher : 18
##  Max.   :4.000   Max.   :4.000
##         reason      guardian     traveltime     studytime
##  course     :113   father: 66   Min.   :1.00   Min.   :1.000
##  home       : 84   mother:212   1st Qu.:1.00   1st Qu.:1.000
##  other      : 25   other : 22   Median :1.00   Median :2.000
##  reputation: 78                 Mean   :1.44   Mean   :2.033
##                                 3rd Qu.:2.00   3rd Qu.:2.000
##                                 Max.   :4.00   Max.   :4.000
##     failures      schoolsup famsup     paid     activities nursery
##  Min.   :0.0000   no :258   no :112   no :162   no :146    no : 59
##  1st Qu.:0.0000   yes: 42   yes:188   yes:138   yes:154    yes:241
##  Median :0.0000
##  Mean   :0.3333
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  higher    internet  romantic      famrel        freetime
##  no : 17   no : 46   no :198   Min.   :1.00   Min.   :1.000
##  yes:283   yes:254   yes:102   1st Qu.:4.00   1st Qu.:3.000
##                                Median :4.00   Median :3.000
##                                Mean   :3.94   Mean   :3.213
##                                3rd Qu.:5.00   3rd Qu.:4.000
##                                Max.   :5.00   Max.   :5.000
##      goout            Dalc            Walc           health
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :3.000   Median :1.000   Median :2.000   Median :4.000
##  Mean   :3.133   Mean   :1.477   Mean   :2.297   Mean   :3.493
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##     absences           G1              G2         pass
##  Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   no : 98
##  1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   yes:202
##  Median : 4.000   Median :10.00   Median :11.00
##  Mean   : 5.967   Mean   :10.81   Mean   :10.64
##  3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :75.000   Max.   :19.00   Max.   :19.00
##     cluster         cluster2
##  Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:1.00
##  Median :1.000   Median :2.00
##  Mean   :1.293   Mean   :2.16
##  3rd Qu.:2.000   3rd Qu.:3.00
##  Max.   :2.000   Max.   :3.00
```

```r
library(partykit)

tree_model <- ctree(pass ~ ., data=train_math)
```

```r
# Árbol:
print(tree_model)
```

```
##
## Model formula:
## pass ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + G1 + G2 + cluster + cluster2
##
## Fitted party:
## [1] root
## |   [2] G2 <= 9
## |   |   [3] G2 <= 7: no (n = 53, err = 0.0%)
## |   |   [4] G2 > 7: no (n = 54, err = 31.5%)
## |   [5] G2 > 9
## |   |   [6] G2 <= 10: yes (n = 41, err = 17.1%)
## |   |   [7] G2 > 10: yes (n = 152, err = 0.7%)
##
## Number of inner nodes:    3
## Number of terminal nodes: 4
```

```r
plot(tree_model)
```



Como se puede ver, sólo se ha utilizado la variable `G2` para elegir, partiendo simplemente en el valor de 9 y por tanto asignando "no" a los alumnos suspensos en el segundo periodo y "sí" a los alumnos aprobados. Respecto al conjunto de test, se obtiene la siguiente predicción:

```
pred <- predict(tree_model, newdata = test_math)
table(pred, test_math$pass)
```

```
##
## pred  no yes
##   no  32   7
##   yes  0  56
```

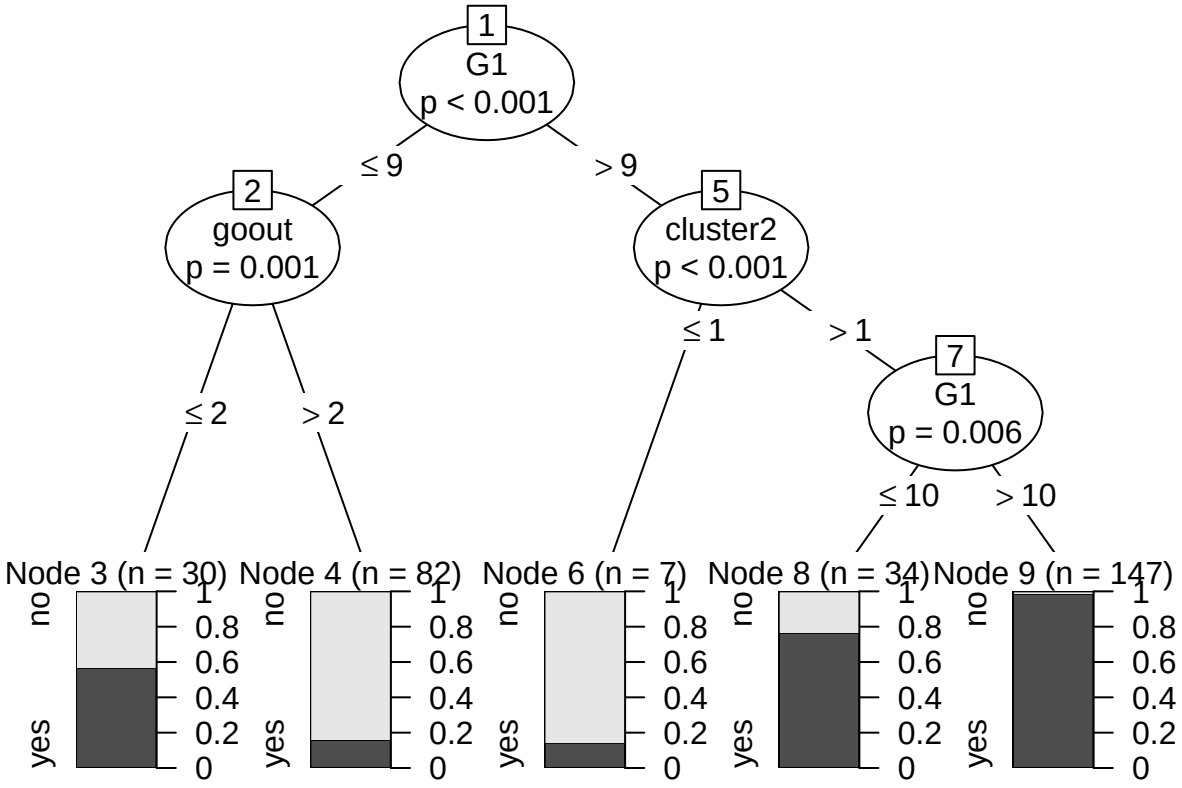
Como se puede observar el modelo sólo se equivoca en 7 instancias de 95 para el conjunto de test, y sólo en alumnos aprobados, predichos como suspensos.

Probamos un último modelo en el que no utilizamos `G2`:

```
tree_model2 <- ctree(pass ~ ., data=select(train_math,-G2))

# Árbol:
print(tree_model2)
```

```
##
## Model formula:
## pass ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + G1 + cluster + cluster2
##
## Fitted party:
## [1] root
## |   [2] G1 <= 9
## |   |   [3] goout <= 2: yes (n = 30, err = 43.3%)
## |   |   [4] goout > 2: no (n = 82, err = 15.9%)
## |   [5] G1 > 9
## |   |   [6] cluster2 <= 1: no (n = 7, err = 14.3%)
## |   |   [7] cluster2 > 1
## |   |   |   [8] G1 <= 10: yes (n = 34, err = 23.5%)
## |   |   |   [9] G1 > 10: yes (n = 147, err = 1.4%)
##
## Number of inner nodes:    4
## Number of terminal nodes: 5
```

```
plot(tree_model2)
```



Esta vez entran en juego las variables `G1`, `goout` y uno de los clusterizados realizados anteriormente, que separaba las instancias en tres conjuntos.

```
pred <- predict(tree_model2, newdata = test_math)
table(pred, test_math$pass)
```

```
##
## pred  no yes
##   no  18   7
##   yes 14  56
```

Como se puede observar, este modelo sigue clasificando 7 instancias con valor real aprobado como no aprobado, pero esta vez también asigna a 14 alumnos suspensos una predicción de aprobados.

Conclusiones

Como resumen del trabajo elaborado, se han explorado las variables del dataset `math`, encontrando ciertas relaciones interesantes entre variables. Respecto a la utilización de clústering, se han conseguido tres clústers con características potencialmente interesantes para separar a los alumnos con cierta relación con su nota final. Por último, los modelos supervisados han resultado en un modelo problemático en el caso de regresión lineal, y un modelo muy simple y relativamente efectivo en términos de la tasa de error en test.

Cabe la pena mencionar que en todos los casos se ha hecho obvia la importancia de las variables `G1`, y especialmente `G2` para predecir el aprobado o suspenso de un alumno. Sería de gran interés por tanto conseguir un modelo que clasifique de manera decente el dataset utilizando el resto de características.