Pràctica 2: Neteja i anàlisi de dades

Miquel Fraire i Carla Manzanares

Gener 2022

1. Descripció del dataset.

En el següent treball analitzarem les dades reals de dues escoles de secundària portugueses per tal d'aproximar-nos al rendiment dels estudiants amb l'objectiu d'identificar les variables clau que afecten en l'èxit i/o el fracàs educatiu. Concretament en la matèria de matemàtiques.

Les dades contenen la informació sobre tendències i patrons que es poden utilitzar per millorar la presa de decisions i optimitzar l'èxit. Modelar el rendiment dels estudiants és una eina important ja que pot ajudar a comprendre millor aquest fenomen i, en última instància, a millorar-lo.

L'objectiu final serà predir el rendiment dels estudiants i esbrinar els factors que els afecten.

El fitxer CSV correspon al dataset de Student Performance de UCI Machine Learning. Per llegir-lo fem servir la funció read.csv.

```
students <- read.csv("/Volumes/GoogleDrive/La meva unitat/2020-2021/SEMESTRE 2/MINERIA/PRA1/student-mat.csv", sep=";", stringsAsFactors = TRUE)
```

Diccionari de variables

A continuació adjuntem el diccionari de variables amb l'explicació de què indica cadascuna i el rang de valors que té en cas de ser categòrica.

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

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

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

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

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

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

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

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

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

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

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

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

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)

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

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

schoolsup extra educational support (binary: yes or no)

famsup family educational support (binary: yes or no)

paid extra paid classes within the course subject (binary: yes or no)

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

nursery attended nursery school (binary: yes or no)

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

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

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

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

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

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

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

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

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

absences number of school absences (numeric)

These grades are related with the course subject, Math: **G1** first period grade (numeric: from 0 to 20) **G2** second period grade (numeric: from 0 to 20) **G3** final grade (numeric: from 0 to 20, output target).

2. Integració i selecció de les dades d'interès a analitzar.

De vegades ens trobem que necessitem manipular dades recollides en diferents fitxers i fusionar-les per tal de crear una estructura de dades coherent i única que contingui tota l'informació. Però no és el nostre cas.

Volem analitzar la influència que té l'entorn familiar, el sexe de l'estudiant, la dedicació a l'estudi i l'interés per seguir estudiant en les qualificacions finals dels estudiants. Per tant, treballarem amb un subcojunt de la base de dades original amb les variables del nostre interés que són: `sex`, `Medu`, `Fedu`, `Mjob`, `Fjob`, `studytime`, `paid`, `higher`, `absences` com a variables independents i `G3` com a variable dependent.

```
students2 <- subset(students, select = c(sex, Medu, Fedu, Mjob, Fjob, stu
dytime, paid, higher, absences, G3))
```

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Un dels problemes que poden presentar els conjunts de dades és que hi hagi valors absents, atès que les dades no disponibles poden causar errors i alterar el resultat de les anàlisis. Les diferents estratègies per solucionar aquest problema són:

- Eliminar les variables amb un percentatge alt de valors buits
- Eliminar les files amb variables no disponibles
- Imputar les dades o omplir-les amb un valor per defecte

La funció colMeans ens mostra quina proporció de dades no disponibles tenim per columna.

```
sort(colMeans(is.na(students2)), decreasing = TRUE)
```

```
##       sex      Medu      Fedu      Mjob      Fjob studytime      paid    higher
##         0         0         0         0         0         0         0         0
##  absences        G3
##         0         0
```

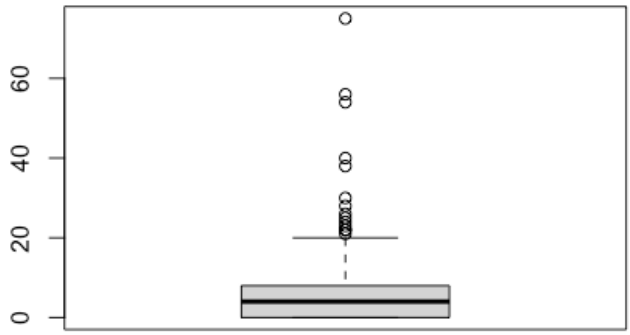
Com que no falta cap dada, no caldrà modificar res.

3.2. Identificació i tractament de valors extrems.

Els valors extrems són aquelles dades que es troben molt allunyades de la distribució normal d'una variable o població. Són observacions que es desvien tant de la resta que aixequen sospites. Aquests valors poden afectar de manera adversa els resultats de les anàlisis posteriors i poden aparèixer per diferents raons, per la qual cosa s'apliquen diferents solucions en funció del context.

Amb la gràfica resultant de la funció `boxplot()` s'identifiquen els outliers, representats en forma de cercles i el valor dels quals es poden recuperar del resultat `out`. Aquesta funció només té sentit aplicar-la sobre variables numèriques contínues i sense acotació. L'única variable del nostre subconjunt que reuneix aquestes característiques és la variable `absences`.

```
valors.extrems <-boxplot(students2$absences)
```




```
valors.extrems$out
```

```
##  [1] 25 54 26 56 24 28 22 21 75 22 30 38 22 40 23
```

En els casos en què els outliers fossin errors en les dades, complicats de corregir, generalment es tractarien com a valors perduts, de manera que s'optaria per eliminar o corregir el registre mitjançant els mètodes d'imputació de dades esmentades en l'exercici anterior.

En aquest cas però, es tracta de valors atípics però legítims, perquè el valor més alt de tots és de 75 i, malgrat que, és una xifra molt alta d'absentisme escolar, no deixa de ser possible. Altrament seria si aquest valor superès el màxim de dies lectius en tot un curs.

4. Anàlisi de les dades.

Comencem l'anàlisi amb una visió general de les variables i la seva distribució. La funció `summary` ens permet fer una descriptiva ràpida de totes les variables. Pel que fa a les variables numèriques, ens mostra la mitjana, la desviació estàndard, el mínim, el màxim i els quartils. Pel que fa a les variables categòriques (i en aquest cas llegides en tipus factor) ens fa un recompte de cada valor.

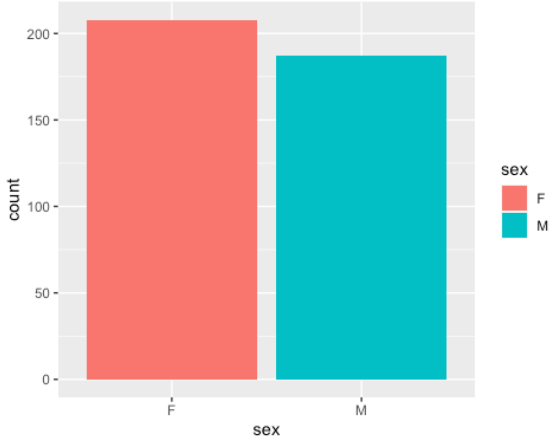
Prèviament hem hagut de convertir el format d'algunes variables a categòriques, com és el cas de `Medu`, `Fedu` i `studytime`, perquè el format en què s'han carregat és numèric però, segons el diccionari de variables que hem adjuntat més amunt, han de ser categòriques.

```
students2$Medu <- factor(students2$Medu, labels = c('none', 'primary educati
on', '5th to 9th grade', 'secondary education', 'higher education'))
students2$Fedu <- factor(students2$Fedu, labels = c('none', 'primary educati
on', '5th to 9th grade', 'secondary education', 'higher education'))
students2$studytime <- factor(students2$studytime, labels = c('<2 hours', '2
to 5 hours', '5 to 10 hours', '>10 hours'))

summary(students2)
```

```
##  sex                      Medu                       Fedu           Mjob
##  F:208   none               :  3   none               :  2   at_home : 59
##  M:187   primary education  : 59   primary education  : 82   health  : 34
##          5th to 9th grade   :103   5th to 9th grade   :115   other   :141
##          secondary education: 99   secondary education:100   services:103
##          higher education   :131   higher education   : 96   teacher : 58
##
##        Fjob             studytime    paid     higher       absences
##  at_home : 20   <2 hours     :105   no :214   no : 20   Min.   : 0.000
##  health  : 18   2 to 5 hours :198   yes:181   yes:375   1st Qu.: 0.000
##  other   :217   5 to 10 hours: 65                       Median : 4.000
##  services:111   >10 hours    : 27                       Mean   : 5.709
##  teacher : 29                                           3rd Qu.: 8.000
##                                                         Max.   :75.000
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

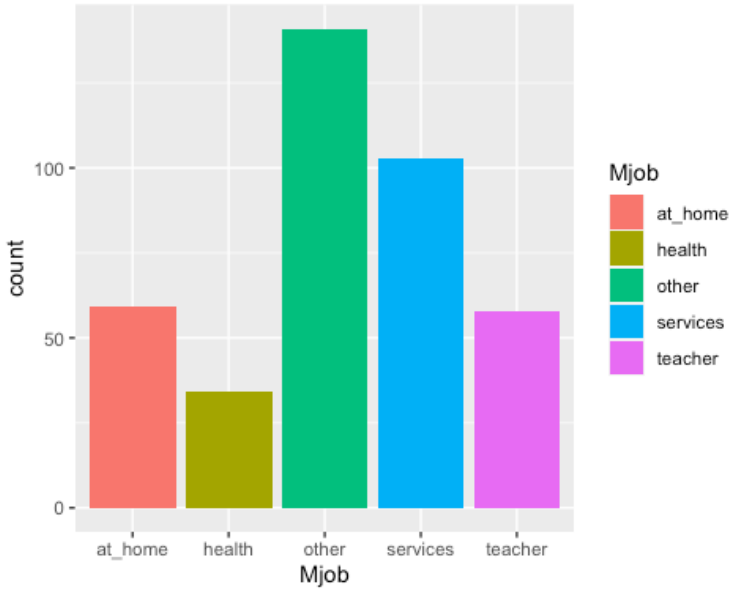
Tot seguit examinarem la distribució de valors per a cada variable. En primer lloc, les categòriques.

```
library(ggplot2)
ggplot(students2, aes(x = sex, fill = sex)) + geom_bar()
```



Pel que fa a la variable sex, la distribució és força paritària entre tots dos valors.

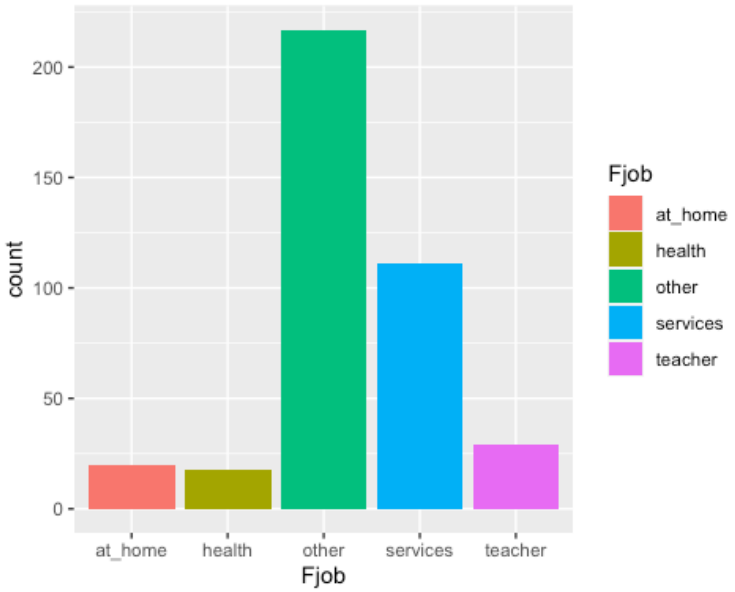
```
ggplot(students2, aes(x = Mjob, fill = Mjob)) + geom_bar()
```



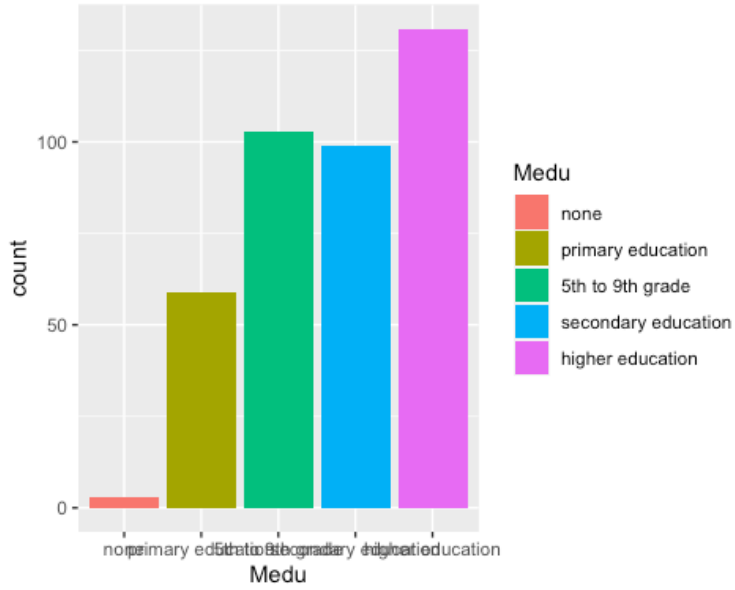
Pel que fa al treball del pare, la majoria d'observacions es concentren en la categoria 'other' seguida de forma gradualment descendent per les categories 'services', 'teacher' i 'at_home' al mateix nivell, i finalment 'health'.

```
ggplot(students2, aes(x = Fjob, fill = Fjob)) + geom_bar()
```

Pel que fa al treball de la mare, la majoria d'observacions també es concentren en la categoria 'other' seguida de forma abruptament descendent per les categories 'services', 'teacher', i finalment 'at_home' i 'health' al mateix nivell.



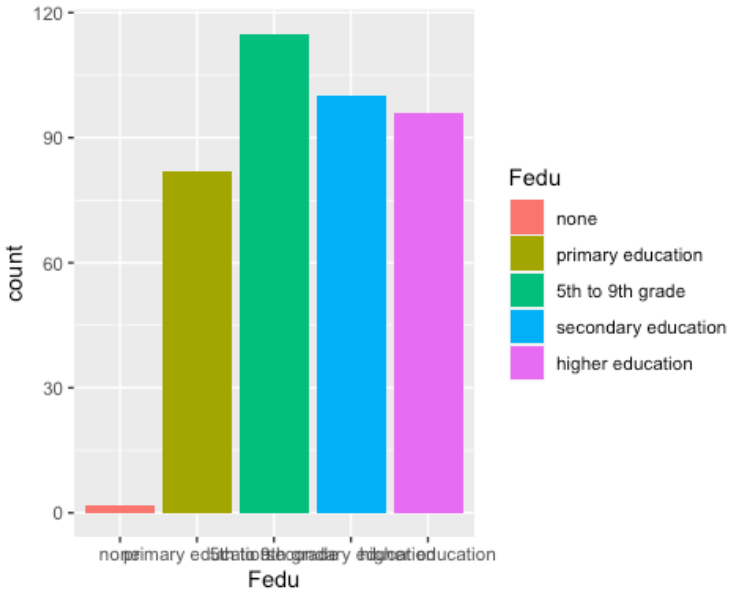
```
ggplot(students2, aes(x = Medu, fill = Medu)) + geom_bar()
```



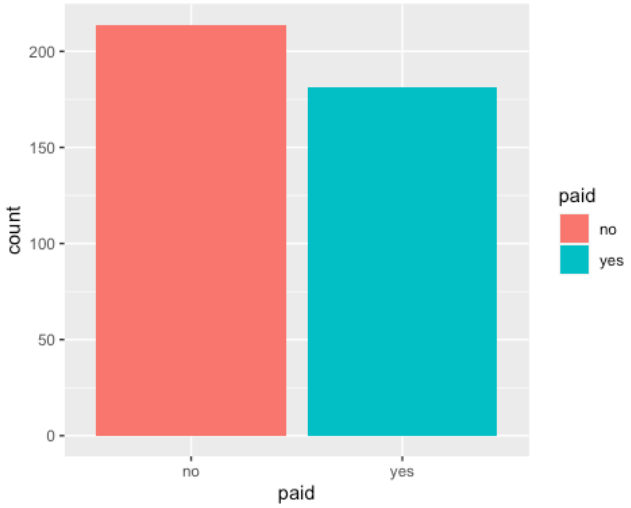
Pel que fa a l'educació del pare, la majoria d'observacions es concentren en la categoria 'higher education' seguida per les categories '5th to 9th grade' i 'secondary education' gairebé al mateix nivell, a continuació, a força distància 'primary education' i finalment 'none' amb una concentració gairebé testimonial.

```
ggplot(students2, aes(x = Fedu, fill = Fedu)) + geom_bar()
```

Pel que fa a l'educació de la mare, la majoria d'observacions es concentren en la categoria '5th to 9th grade' seguida en un descens gradual per les categories 'secondary education', 'higher education', una mica més distant 'primary education' i finalment 'none' amb una concentració gairebé testimonial.



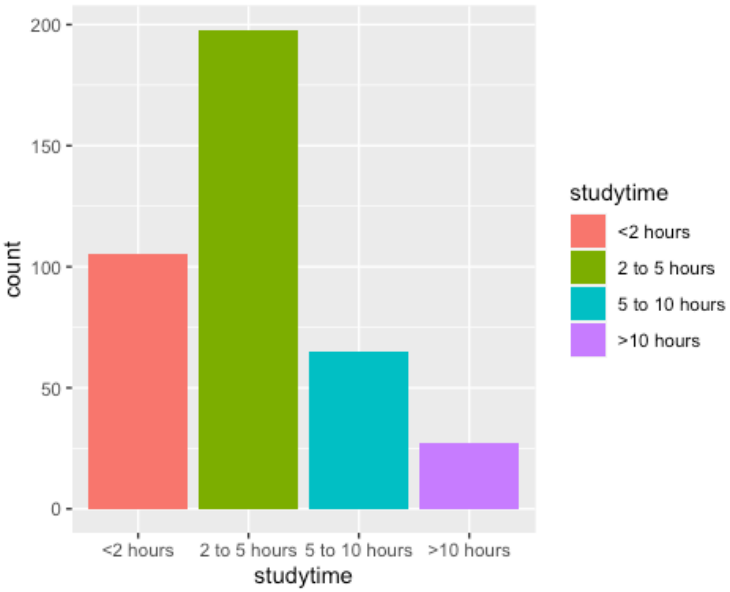
```
ggplot(students2, aes(x = paid, fill = paid)) + geom_bar()
```



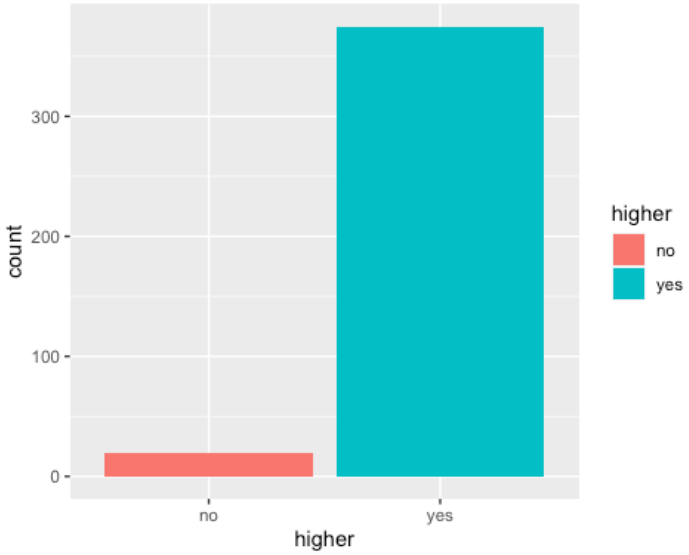
Pel que fa a les classes extraescolars dels alumnes, s'observa força paritat entre aquells alumnes que en fan i els que no, tot i que la columna del no supera la del si.

```
ggplot(students2, aes(x = studytime, fill = studytime)) + geom_bar()
```

Pel que fa a la dedicació de l'alumne als estudis, la majoria d'observacions es concentren entre les dues i les cinc hores d'estudi. La següent categoria més poblada concentra aproximadament la meitat d'observacions que la primera i aquests hi dediquen menys de dues hores a l'estudi. Les següents categories es troben en l'ordre de 5 a 10 hores i més de 10 hores en descens gradual.



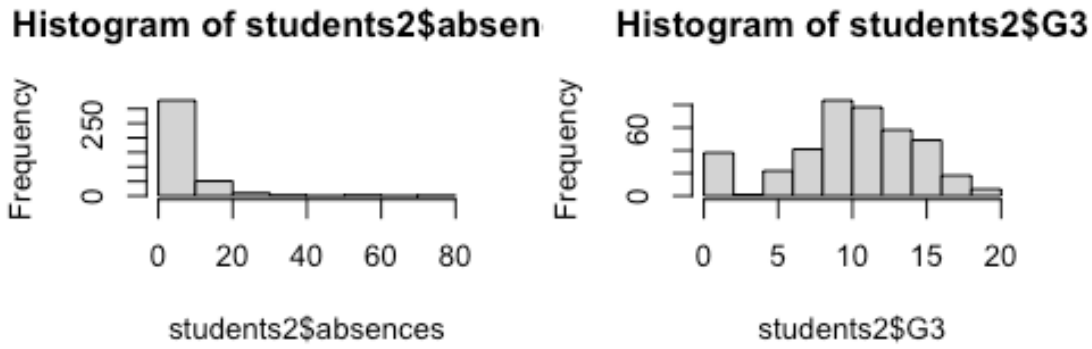

```
Ggplot(students2, aes(x = higher, fill = higher)) + geom_bar()
```



Pel que fa a les aspiracions acadèmiques, la immensa majoria d'estudiants tenen la intenció de seguir estudiant, mentre que una ínfima minoria no la té. Per tant, la distribució és completament asimètrica.

A continuació grafiquem la distribució de valors de les variables numèriques.

```
Par(mfrow=c(2,2))
hist(students2$absences)
hist(students2$G3)
```



Pel que fa a l'absentisme escolar, la gran majoria d'alumnes falta molt poc a classe no sobrepassant els 10 dies. Molts pocs alumnes falten entre 10 i 20 dies, i absències més grans de fins a 80 dies són molt esporàdiques.

Pel que fa al rendiment escolar, la distribució s'apropa visualment a la normalitat, ja que la majoria d'observacions es concentren al voltant de la mitjana i a partir d'aqui els valors de rangs més allunyats de la mitjana van disminuint gradualment. Com a excepció a aquest comportament hi ha el cas d'alguns alumnes que no superen el curs.

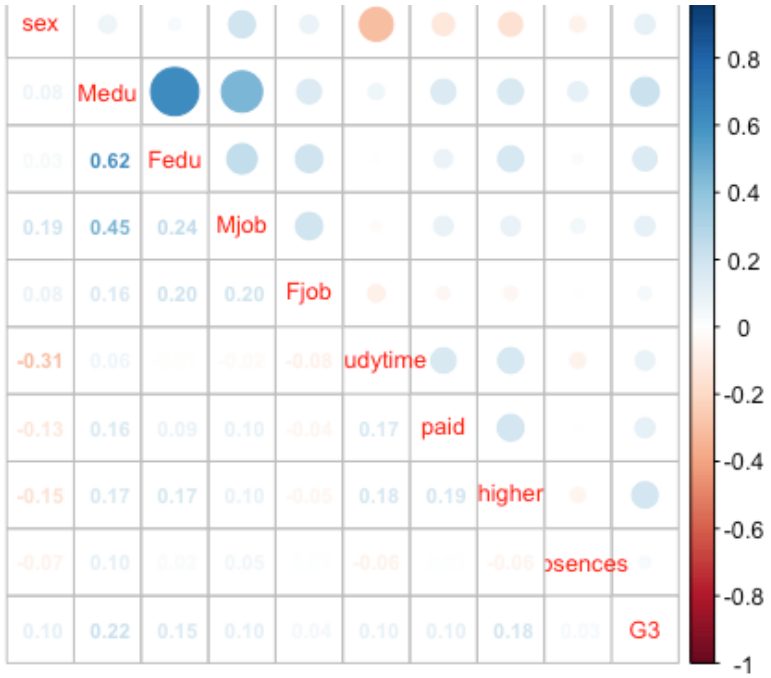
4.1. Selecció dels grups de dades que es volen analitzar/comparar

Fem una matriu simètrica per mostrar el valor de correlació per a cada parella de variables i un gràfic per visualitzar tota aquesta informació. Per poder calcular les correlacions entre variables necessitem transformar les variables categòriques a numèriques.

```
# Transformem les variables
students.num <- students2
students.num$sex <- as.integer(students2$sex)
students.num$Medu <- as.integer(students2$Medu)
students.num$Fedu <- as.integer(students2$Fedu)
students.num$Mjob <- as.integer(students2$Mjob)
students.num$Fjob <- as.integer(students2$Fjob)
students.num$studytime <- as.integer(students2$studytime)
students.num$paid <- as.integer(students2$paid)
students.num$higher <- as.integer(students2$higher)

# Realitzem la matriu de correlacions
matriu.cor <- cor(students.num)
matriu.cor
```

```
##                    sex       Medu         Fedu        Mjob         Fjob
## sex         1.00000000 0.07822789  0.034878306  0.19142075  0.084957444
## Medu        0.07822789 1.00000000  0.623455112  0.45480549  0.157920391
## Fedu        0.03487831 0.62345511  1.000000000  0.24332713  0.200168813
## Mjob        0.19142075 0.45480549  0.243327126  1.00000000  0.196758381
## Fjob        0.08495744 0.15792039  0.200168813  0.19675838  1.000000000
## studytime -0.30626762 0.06494414 -0.009174639 -0.02309735 -0.076816652
## paid       -0.12912562 0.15970038  0.086981416  0.09656612 -0.040434613
## higher     -0.15105552 0.16884453  0.174565520  0.09788679 -0.045250208
## absences   -0.06696176 0.10028482  0.024472887  0.05025366  0.008195729
## G3          0.10345565 0.21714750  0.152456939  0.10208180  0.042285873
##                studytime         paid      higher     absences         G3
## sex       -0.306267618 -0.129125619 -0.15105552 -0.066961757 0.10345565
## Medu       0.064944137  0.159700381  0.16884453  0.100284818 0.21714750
## Fedu      -0.009174639  0.086981416  0.17456552  0.024472887 0.15245694
## Mjob      -0.023097354  0.096566122  0.09788679  0.050253661 0.10208180
## Fjob      -0.076816652 -0.040434613 -0.04525021  0.008195729 0.04228587
## studytime  1.000000000  0.167219880  0.17508142 -0.062700175 0.09781969
## paid       0.167219880  1.000000000  0.18921385  0.007435174 0.10199624
## higher     0.175081425  0.189213846  1.00000000 -0.056085227 0.18246462
## absences  -0.062700175  0.007435174 -0.05608523  1.000000000 0.03424732
## G3         0.097819690  0.101996241  0.18246462  0.034247316 1.00000000
```

```
# Representació gràfica de les correlacions
library(corrplot)

## corrplot 0.88 loaded

corrplot.mixed(matriu.cor,upper="circle",number.cex=.7,tl.cex=.8)
```



S'observa que hi ha dos valors de correlació més propers a 1 que la resta, entre Medu i Fedu; les variables presenten una dependència lineal en sentit directe (correlació positiva). La següent parella de variables amb més correlació ja se situa a un coeficient per sota del 0,5.

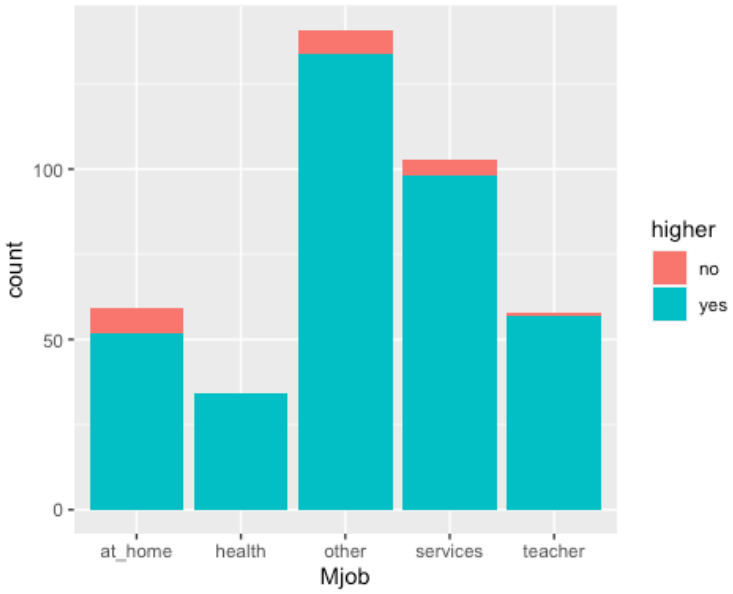
Per analitzar un conjunt de dades necessitem tenir en compte més d'una variable alhora. L'anàlisi bivariant permet identificar les relacions entre dues

variables, i fins i tot veure de quina manera una pot predir l'altra.

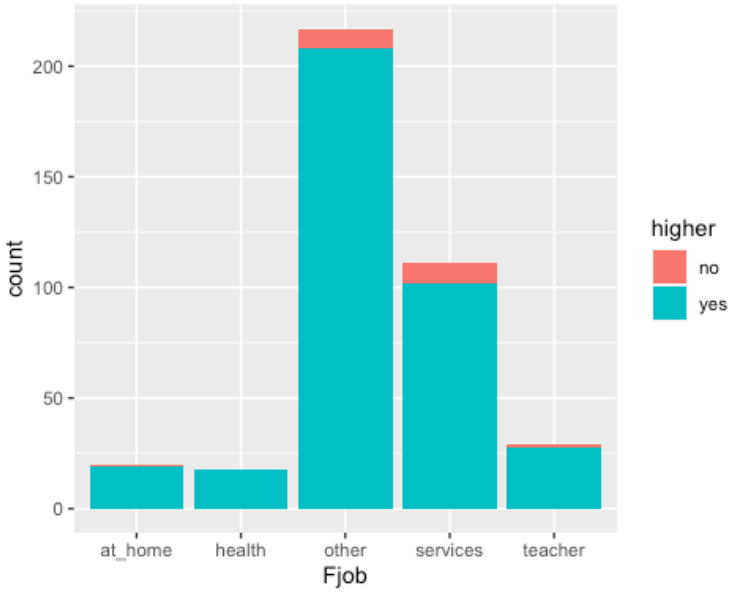
El grup de variables de les quals estudiarem la relació són aquelles relacionades amb el context familiar i les aspiracions dels alumnes.

```
Ggplot(data=students2, aes(x=Mjob, fill=higher)) + geom_bar()
```

Pel que fa a la feina del pare en relació a les ganes de seguir estudiant de l'alumne, la majoria d'observacions es troba en la categoria 'other' amb una diferència abismal entre els que volen seguir estudiant i els que no a favor dels primers. La següent categoria amb més observacions és 'services' i presenta una distribució interna similar a l'anterior. A continuació les categories 'at_home' i 'teacher' concentren un nombre similar d'observacions però, mentre la primera té una petita franja de partidaris a no seguir estudiant, la segona encara mostra aquesta tendència més acusada. Per últim, tots els alumnes amb pares dedicats a la salut tenen la intenció de seguir estudiant.



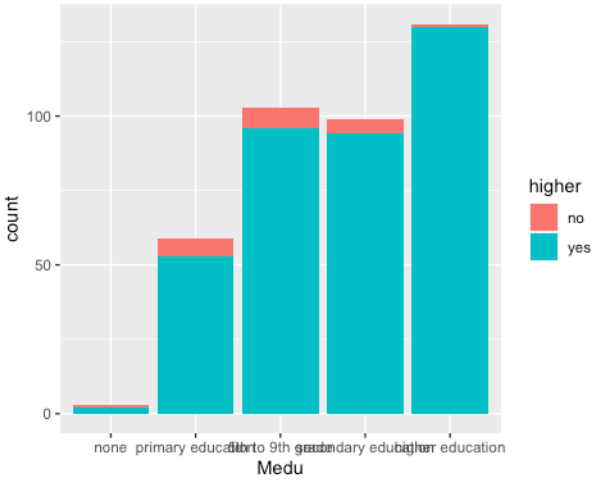

```
Ggplot(data=students2, aes(x=Fjob, fill=higher)) + geom_bar()
```



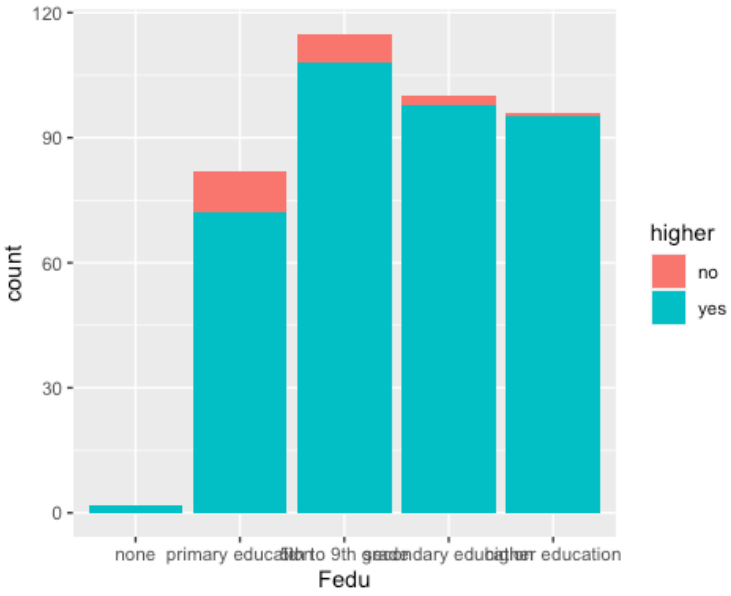
Pel que fa a la feina de la mare en relació a les ganes de seguir estudiant de l'alumne, la majoria d'observacions també es troba en la categoria 'other' amb una diferència abismal entre els que volen seguir estudiant i els que no a favor dels primers. La següent categoria amb més observacions també és 'services' i presenta una distribució interna similar a l'anterior. A continuació les categories 'teacher', 'at_home' i 'health' concentren un nombre similar d'observacions en descens, però mentre les dues primeres tenen una minúscula franja de partidaris a no seguir estudiant, l'última conté només estudiants que pretenen continuar els estudis.

```
Ggplot(data=students2, aes(x=Medu, fill=higher)) + geom_bar()
```

Pel que fa a l'educació del pare en relació a les ganes de seguir estudiant de l'alumne, la majoria d'observacions es concentren en la categoria 'higher education' amb una diferència abismal entre els que volen seguir estudiant i els que no a favor dels primers. Seguida per les categories '5th to 9th grade' i 'secondary education' gairebé al mateix nivell, amb una diferència encara molt gran entre els que volen seguir estudiant i els que no, a favor dels primers. A continuació i a força distància se situa 'primary education' amb una distribució interna similar a les dues anteriors; i finalment 'none' amb una concentració gairebé testimonial i amb el doble de partidaris a seguir estudiant que de deixar-ho.




```
Ggplot(data=students2, aes(x=Fedu, fill=higher)) + geom_bar()
```



Pel que fa a l'educació de la mare en relació a les ganes de seguir estudiant de l'alumne, la majoria d'observacions es concentren en la categoria '5th to 9th grade' seguint amb una distribució interna igual que les anteriors. Seguida en un descens gradual per les categories 'secondary education' i 'higher education' que presenten una mínima quantitat d'alumnes que no volen seguir estudiant. Una mica més distant, 'primary education', amb una distribució una mica diferent a les anteriors però que segueix primant els alumnes que volen seguir estudiant. Finalment 'none' amb una concentració gairebé testimonial amb l'excepció que no hi ha cap alumne que no vulgui seguir estudiant, fet que contrasta amb l'anterior gràfica.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

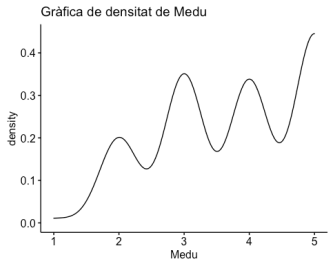
Amb l'objectiu de verificar la suposició de la normalitat, el test de Shapiro-Wilk es considera un dels mètodes més potents per contrastar la normalitat. "Assumint com a hipòtesi nul·la que la població està distribuïda normalment, si el p-valor és més petit que el nivell de significació, generalment α=0,05, llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no compten amb una distribució normal. Si, per contra, el p-valor és major a α, es conclou que no es pot rebutjar aquesta hipòtesi i s'assumeix que les dades segueixen una distribució normal." (M.Calvo, D.Pérez, L.Subirats, 2019: 30)

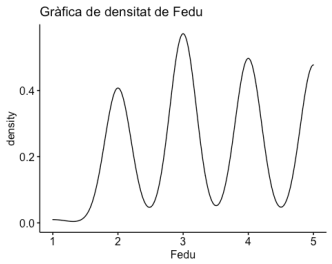
```
library("ggpubr")

shapiro.test(students.num$Medu)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$Medu
## W = 0.86103, p-value < 2.2e-16

shapiro.test(students.num$Fedu)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$Fedu
## W = 0.87555, p-value < 2.2e-16

shapiro.test(students.num$Mjob)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$Mjob
## W = 0.89223, p-value = 4.676e-16

shapiro.test(students.num$Fjob)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$Fjob
## W = 0.83053, p-value < 2.2e-16

shapiro.test(students.num$studytime)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$studytime
## W = 0.8342, p-value < 2.2e-16
```

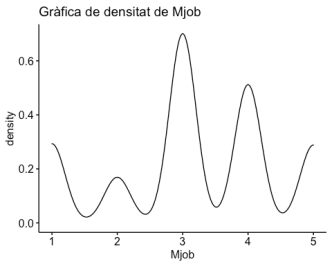
```
shapiro.test(students.num$absences)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$absences
## W = 0.66683, p-value < 2.2e-16

shapiro.test(students.num$G3)

##
##  Shapiro-Wilk normality test
##
## data:  students.num$G3
## W = 0.92873, p-value = 8.836e-13
```

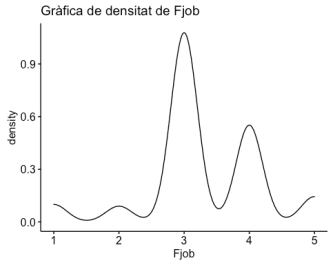
Segons el test de Saphiro, es rebutgen totes les hipòtesis nul·les i es considera que les dades NO segueixen una distribució normal. I amb les gràfiques següents queda demostrat visualment.

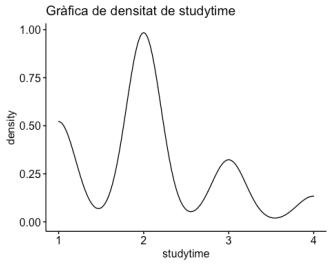
```
Ggdensity(students.num$Medu,
          main = "Gràfica de densitat de Medu",
          xlab = "Medu")

ggdensity(students.num$Fedu,
          main = "Gràfica de densitat de Fedu",
          xlab = "Fedu")

ggdensity(students.num$Mjob,
          main = "Gràfica de densitat de Mjob",
          xlab = "Mjob")

ggdensity(students.num$Fjob,
          main = "Gràfica de densitat de Fjob",
          xlab = "Fjob")

ggdensity(students.num$studytime,
          main = "Gràfica de densitat de studytime",
          xlab = "studytime")

ggdensity(students.num$absences,
          main = "Gràfica de densitat de absences",
          xlab = "absences")

ggdensity(students2$G3,
          main = "Gràfica de densitat de G3",
          xlab = "G3")
```

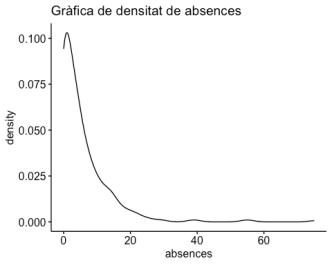


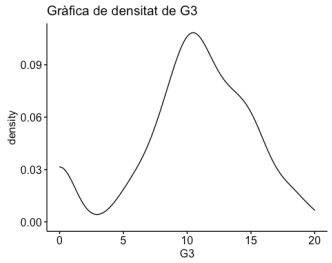












Per comprovar l'homogeneïtat en la variància de les dades, el test de Filgner-Killen és el més utilitzat quan les dades no compleixen amb la condició de normalitat, que és el nostre cas. La hipòtesi nul·la assumeix igualtat de variàncies en els diferents grups de dades, de manera que p-valors inferiors al nivell de significació indicaran heteroscedasticitat.

```
library(car)

## Loading required package: carData

library(stats)
fligner.test(G3 ~ interaction(sex, Mjob, Fjob, Medu, Fedu, paid, studytim
e, absences, higher), data = students2)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  G3 by interaction(sex, Mjob, Fjob, Medu, Fedu, paid, studytime,
absences, higher)
## Fligner-Killeen:med chi-squared = 339.9, df = 379, p-value = 0.9261
```

És confirma homocedasticitat.

4.3. Aplicació de proves estadístiques per comparar els grups de dades.

Ens disposem a posar en pràctica tres mètodes d'analisi estadístic: contrast d'hipòtesis, regressió lineal i test de Kruskal-Wallis.

Constrast d'hipòtesis

El contrast d'hipòtesis el farem sobre dues mostres per determinar si el sexe dels estudiants determina el seu rendiment.

Es planteja el següent contrast d'hipòtesis de les dues mostres:

- Hipòtesi nul·la: La nota mitja de les noies és igual que la dels nois.
- Hhipòtesi alternativa: La nota mitja de les noies és superior a la dels nois.

```
students.girls.G3 <- students2[students2$sex == 'F',]$G3
students.boys.G3 <- students2[students2$sex == 'M',]$G3

t.test(students.girls.G3, students.boys.G3, alternative = 'less')
```

```
##
##  Welch Two Sample t-test
##
## data:  students.girls.G3 and students.boys.G3
## t = -2.0651, df = 390.57, p-value = 0.01979
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.191125
## sample estimates:
## mean of x mean of y
##  9.966346 10.914439
```

Atès que p-valor és menor que el nivell de significació, rebutgem la hipotesi nula i, per tant, afirmem la nostra propia hipòtesi: les noies tenen una mitja més alta que la dels nois.

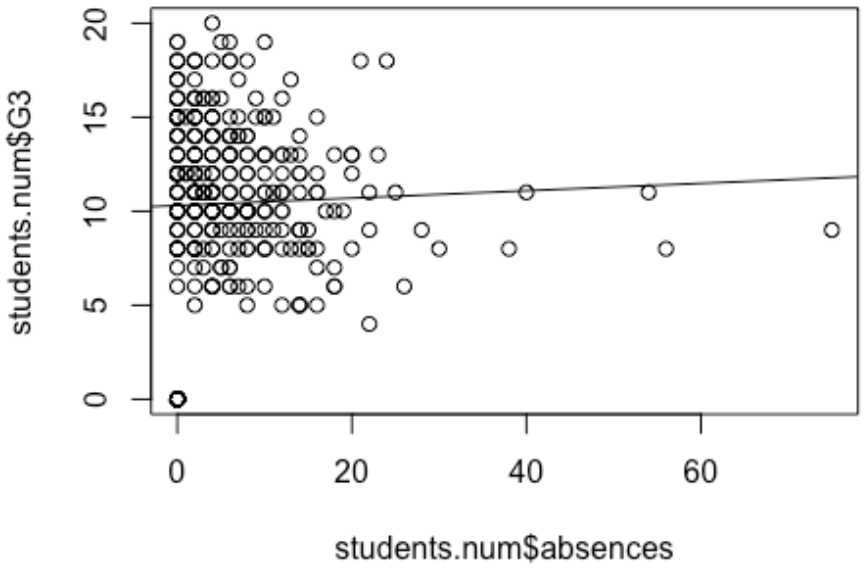
Regressió lineal

Anem a comprovar la regressió lineal entre les variables numèriques, ja que obtindrem la relació de dependència lineal entre la variable dependent i la independent. Si volem visualitzar alhora les relacions creuades entre les variables, podem fer un pairplot, i així tenir una guia visual més fàcil per començar.

```
regressio <- lm(G3~absences,data=students.num)
summary(regressio)
```

```
##
## Call:
## lm(formula = G3 ~ absences, data = students.num)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3033  -2.3033   0.5007   3.4811   9.6183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30327    0.28347  36.347   <2e-16 ***
## absences     0.01961    0.02886   0.679    0.497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.585 on 393 degrees of freedom
## Multiple R-squared:  0.001173,   Adjusted R-squared:  -0.001369
## F-statistic: 0.4615 on 1 and 393 DF,  p-value: 0.4973

plot(students.num$absences,students.num$G3)
abline(lm(G3~absences,data=students.num))
```



Com que el coeficient de determinació és molt proper a 0 no presenten correlació entre les variables. A la gràfica es pot veure com la linea és bastant plana ja que les observacions que es destaquen de la resta per les seves absències, estan situats en els valors mitjans de la nota. Aquesta regressió explica com el fet d'acumular moltes absències no condiciona el resultat final.

L'anàlisi de variància unidireccional

Amb l'anàlisi de variància unidireccional comparem les mitjanes entre més de dos grups de dades. Es tracta de saber si hi ha cap relació entre el rendiment de l'alumne i les diferents feines i formacions dels pares.

```
shapiro.test(students2$G3)

##
##  Shapiro-Wilk normality test
##
## data:  students2$G3
## W = 0.92873, p-value = 8.836e-13

fligner.test(G3 ~ Fjob, data = students2)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  G3 by Fjob
## Fligner-Killeen:med chi-squared = 4.7741, df = 4, p-value = 0.3113

kruskal.test(G3 ~ Fjob, data = students2)

##
##  Kruskal-Wallis rank sum test
##
## data:  G3 by Fjob
## Kruskal-Wallis chi-squared = 6.2764, df = 4, p-value = 0.1794
```

Atès que el p-valor obtingut és major que el nivell de significació, es pot concloure que el nivell de rendiment que reflecteixen les notes NO mostra diferències significatives segons la feina de la mare.

```
shapiro.test(students2$G3)

##
##  Shapiro-Wilk normality test
##
## data:  students2$G3
## W = 0.92873, p-value = 8.836e-13

fligner.test(G3 ~ Mjob, data = students2)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  G3 by Mjob
## Fligner-Killeen:med chi-squared = 0.95368, df = 4, p-value = 0.9167

kruskal.test(G3 ~ Mjob, data = students2)

##
##  Kruskal-Wallis rank sum test
##
## data:  G3 by Mjob
## Kruskal-Wallis chi-squared = 16.127, df = 4, p-value = 0.002853
```

Atès que el p-valor obtingut és menor que el nivell de significació, es pot concloure que el nivell de rendiment que reflecteixen les notes mostra diferències significatives segons la feina del pare.

```
shapiro.test(students2$G3)

##
##  Shapiro-Wilk normality test
##
## data:  students2$G3
## W = 0.92873, p-value = 8.836e-13

fligner.test(G3 ~ Fedu, data = students2)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  G3 by Fedu
## Fligner-Killeen:med chi-squared = 2.3796, df = 4, p-value = 0.6663

kruskal.test(G3 ~ Fedu, data = students2)

##
##  Kruskal-Wallis rank sum test
##
## data:  G3 by Fedu
## Kruskal-Wallis chi-squared = 14.677, df = 4, p-value = 0.005419
```

Atès que el p-valor obtingut és menor que el nivell de significació, es pot concloure que el nivell de rendiment que reflecteixen les notes mostra diferències significatives segons la formació de la mare.

```
shapiro.test(students2$G3)

##
##  Shapiro-Wilk normality test
##
## data:  students2$G3
## W = 0.92873, p-value = 8.836e-13

fligner.test(G3 ~ Medu, data = students2)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  G3 by Medu
## Fligner-Killeen:med chi-squared = 2.7752, df = 4, p-value = 0.5961

kruskal.test(G3 ~ Medu, data = students2)

##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  G3 by Medu
## Kruskal-Wallis chi-squared = 24.104, df = 4, p-value = 7.613e-05
```

Atès que el p-valor obtingut és menor que el nivell de significació, es pot concloure que el nivell de rendiment que reflecteixen les notes mostra diferències significatives segons la formació del pare.

Contribuciones	**Firma**
Investigació prèvia	MFF, CMC
Redacció de les respostes	MFF, CMC
Desenvolupament codi	MFF, CMC