Pràctica 2

Jordi Marsol López i Arnau Rafi Cuello

11/12/2021

Index

0. Treball previ de selecció 1
1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? 2
2. Integració i selecció de les dades d'interès a analitzar 2
Càrrega incial 2
3. Neteja de les dades. 3
3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos? 3
3.2. Identificació i tractament de valors extrems. 4
4. Anàlisi de dades 5
4.1. Planificació dels anàlisis a aplicar 5
4.2. Comprovació de la normalitat i homogeneïtat de la variància 7
4.3. Aplicació de proves estadístiques per comparar els grups de dades 9
4.3.1. Aplicació de proves paramètriques 9
4.3.2. Aplicació de proves no paramètriques 10
4.3.3. Creació d'un model de regressió logística 10
5. Representació dels resultats a partir de taules 11
6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema? 12
Participants 12

0. Treball previ de selecció

Inicialment volíem donar continuïtat a la pràctica 1 que vam treballar sobre *webscraping* però vam descartar les dades que vam extreure ja que no tenien prou varietat de camps per fer un anàlisi interessant. Seguidament vam trobar un *dataset* relacionat amb la nostra pràctica que podria ser més interessant sobre accidents en parcs d'atraccions (https://www.kaggle.com/stevenlasch/roller-coaster-accidents) i fins i tot el vam voler creuar amb la nostra extracció de dades però no hi havia un identificador clar per relacionar-los i vam descartar aquesta possibilitat.

Vam explorar més a fons les dades d'aquest nou conjunt de dades de manera independent del de la primera pràctica però també vam trobar diversos inconvenients que ens el van fer descartar finalment. Ens vam trobar que hi havia molt poques variables quantitatives, algunes d'elles podien ser interessants però vam trobar que tenien una gran quantitat de dades buides (mes del 90%) i altres com ara la edat dels accidentats tampoc li vam saber trobar interès amb creuar-lo amb altres variables.

Finalment vam trobar un *dataset* amb més interès i un bon equilibri de variables categòriques i numèriques què, sense ser molt gran, igualment li vam veure més potencial per desenvolupar l'estudi que farem a continuació.

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El conjunt de dades escollit és una enquesta realitzada a estudiants d'educació secundària dels cursos de matemàtiques i llengua portuguesa. El *dataset* recull diverses respostes sobre l'àmbit social i escolar dels estudiants. El podem trobar a: https://www.kaggle.com/uciml/student-alcohol-consumption

La pregunta principal que podem respondre és: quin son els factors que porten a obtenir les millors qualificacions? També podrem fer altres creuaments de dades que estimem oportuns segons anem veient a l'estudi.

Els atributs de què està format i les seves característiques son els següents com s'indica a la font original serien:

- 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 (Math or Portuguese) (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: from 0 to 93)
- 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

El *dataset* original separa en dos arxius amb els mateixos atributs els estudiants de matemàtiques i els de portuguès. A la font del *dataset* es comenta i dona la opció de fusionar els dos conjunts de dades en un de sol amb les dades dels estudiants que estiguin als dos cursos alhora. Tot i no haver un identificador únic d'estudiant si que se'ns proporcionen els camps que haurien de conformar un estudiant com a únic dins el conjunt per poder fer la fusió.

A priori inclourem tots els atributs de la taula ja que tots poden ser susceptibles d'interès per l'anàlisi. Tot i que només els compararem tots exhaustivament en alguns casos com ara l'anàlisi de correlacions, si que els revisarem al procés de neteja.

Càrrega incial

Carreguem l'arxiu amb les dades i en mostrem un resum de les seves variables

Per maximitzar l'estudi unirem els dos *datasets* que tenim un de la classe de matemàtiques i l'altre de portuguès i crearem la variable "tipus" per distingir-los, indicant "M" per matemàtiques o "P" per portuguès.

```
## 'data.frame':    1044 obs. of  34 variables:
## $ school    : chr  "GP" "GP" "GP" "GP" ...
## $ sex       : chr  "F" "F" "F" "F" ...
## $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
## $ address   : chr  "U" "U" "U" "U" ...
## $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
## $ Pstatus   : chr  "A" "T" "T" "T" ...
## $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
## $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
## $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
## $ Fjob      : chr  "teacher" "other" "other" "services" ...
## $ reason    : chr  "course" "course" "other" "home" ...
## $ guardian  : chr  "mother" "father" "mother" "mother" ...
## $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
## $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
## $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
## $ schoolsup : chr  "yes" "no" "yes" "no" ...
## $ famsup    : chr  "no" "yes" "no" "yes" ...
## $ paid      : chr  "no" "no" "yes" "yes" ...
## $ activities: chr  "no" "no" "no" "yes" ...
## $ nursery   : chr  "yes" "no" "yes" "yes" ...
## $ higher    : chr  "yes" "yes" "yes" "yes" ...
## $ internet  : chr  "no" "yes" "yes" "yes" ...
## $ romantic  : chr  "no" "no" "no" "yes" ...
## $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
## $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
## $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
## $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
## $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
## $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
## $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
## $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
## $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
## $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
## $ tipus     : chr  "M" "M" "M" "M" ...
```

Observem que hi ha variables tan de tipus categòric com numèric. En aquest *dataset* no hi ha variables amb valor de textos lliures que podrien tenir qualsevol contingut, per tant les que siguin tipus caràcter es poden tractar com a factors.

3. Neteja de les dades.

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

Seguidament farem un llistat de totes les variables i un recompte dels seus valors únics. No inclouríem aquelles que sabem que o bé no serien ni variables numèriques ni categòriques, com ara identificadors, dates o les que continguin textos lliures, observem en el resum de dades però que no seria el cas i totes poden ser útils.

En la següent funció en R es va fent un recompte de les variables, crearem un arxiu de sortida per a un primer anàlisi exploratori i mostraríem apart el percentatge de variables amb valors buits o "NA"s. La funció seria d'especial interès per altres estudis amb un gran nombre de variables.

```
## < table of extent 0 x 0 x 0 x 0 >
```

Descobrim que la taula no té cap valor buit, expressat com a una cadena buida en el cas de les dades tipus caràcter o com a valor NA en les variables numèriques.

Si veiem el resultat de l'arxiu de sortida obtindríem un resum de quants valors únics hi ha per cada variable. Ens interessa saber quina varietat de valors únics hi ha, així observem que en diverses variables numèriques només hi ha valors del 1 al 5, o del 0 al 20 per tant serien pràcticament categòriques ja que son variables amb un rang molt concret de valors numèrics fixes, igualment segueixen sent numèriques ja que es tracta de puntuacions i per tant es poden ordenar de menor a major o al revés.

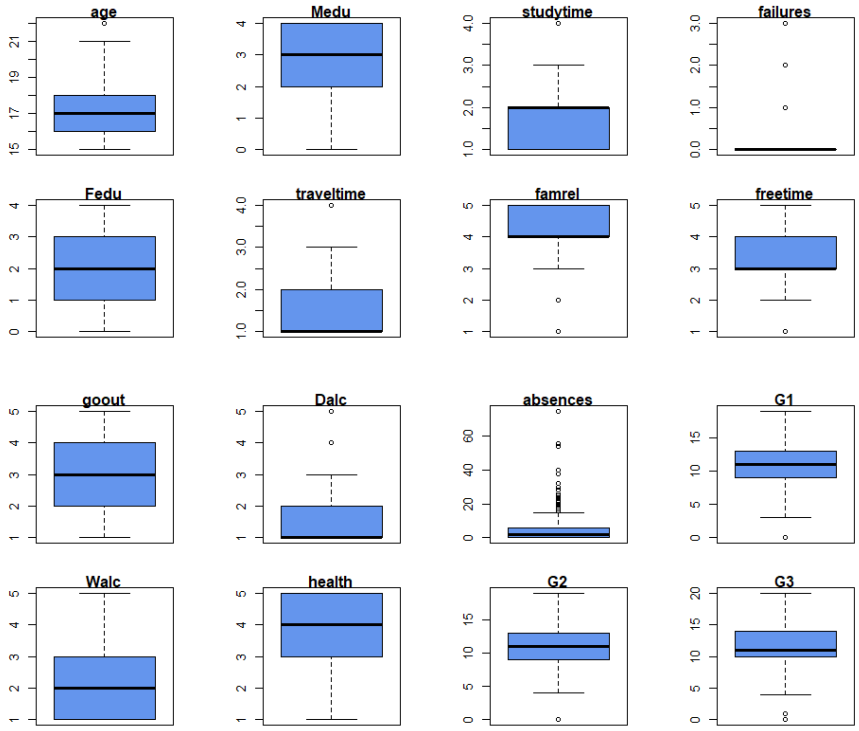
Com que hem obtingut un llista de valors únics, revisarem el resultat de la taula de recompte dels valors per veure si hi ha algun valor que pugui ser estrany o fora de la seva categoria, amb un format diferent, mal escrit, etc. Està documentat al *dataset* original quins valors poden prendre les variables però igualment cal verificar que sigui així. Observem que no hi ha valors erronis ni fora del rang que

s'espera que tinguin o valors expressats com a desconeguts. L'única variable que ens fa dubtar es *absences* ja que podria tenir *valors extrems*. Donat aquest resultat no serà necessari fer cap imputació de valors ja que les dades son completes.

Al següent pas convertirem les dades de tipus caràcter a factor.

3.2. Identificació i tractament de valors extrems.

Mostrarem gràfiques tipus *boxplot* per cadascuna de les variables numèriques per identificar *outliers* ràpidament i després ens centrarem en la variable *absences* que és la que veiem que té un valor màxim més elevat i podria contenir valors extrems.



Trobem que de les diverses variables numèriques algunes indiquen possibles valors extrems no obstant considerem que entren dins el rang previsible que puguin tenir i no les exclourem.

Veiem que les variables G1, G2 i G3 son similars i com que no caldria tractar-les individualment en crearem una de sola amb la mitjana de les tres.

Confirmem pel gràfic que la variable *absences* conté diversos valors extrems que sobresurten bastant. Els aïllem i els mostrem individualment, la resta de variables amb valors extrems son molt pocs i creiem que entren dins dels valors possibles que poden aparèixer.

```
##  [1] 16 16 25 54 18 26 20 18 16 16 56 24 18 28 22 16 18 20 16 21 75 22 30 19 20
## [26] 38 18 20 22 40 23 16 17 16 16 24 22 16 32 16 16 30 21 16 18 16 26 16 16 22
## [51] 18 18 16 21
```

Tenim que hi ha una cinquantena de valors extrems del total de 1044 registres de dades, així que decidim excloure aquests registres de la mostra per poder fer un anàlisi millor posteriorment. En un possible estudi també es podria analitzar com son de diferents els estudiants amb valors extrems de la resta que no ho son.

Per acabar el treball de neteja, eliminarem els *outliers* de la mostra i exportarem l'arxiu de sortida amb tot el *dataset* preparat.

4. Anàlisi de dades

4.1. Planificació dels anàlisis a aplicar

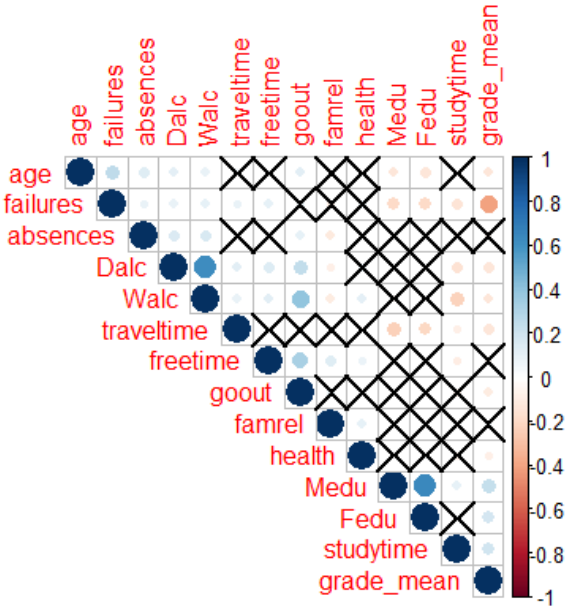
Selecciono les variables numèriques del *dataframe*:

```
## 'data.frame':    990 obs. of  14 variables:
## $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
## $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
## $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
## $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
## $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
## $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
## $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
## $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
## $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
## $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
## $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
## $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
## $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
## $ grade_mean: num  5.7 5.3 8.3 14.7 8.7 15 11.7 5.7 17.7 14.7 ...
```

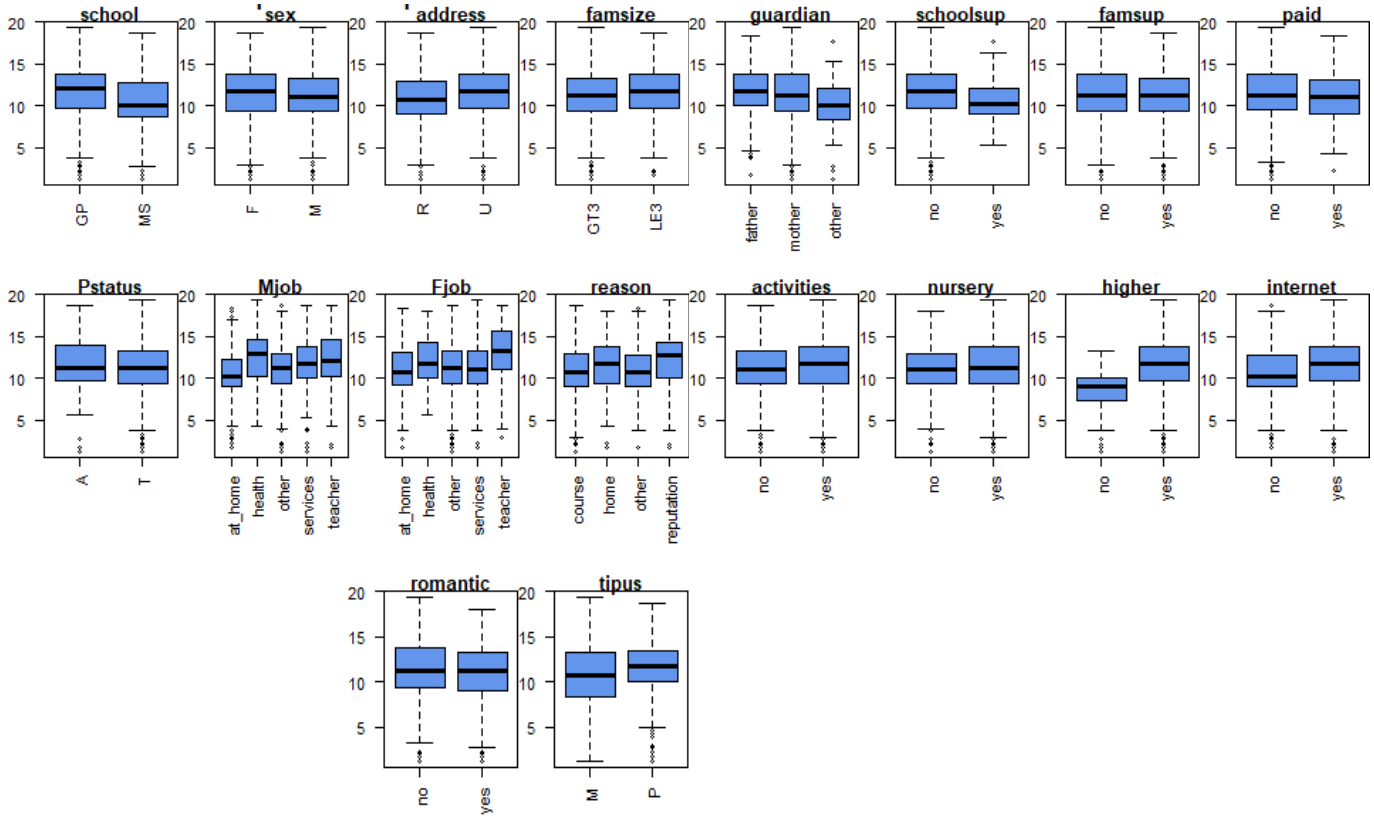
Comprovo possibles correlacions generals entre les variables numèriques utilitzant la llibreria `corrplot` i alguns correlogrames:

Font: <http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram>

```
##                 age        Medu        Fedu   traveltime    studytime
## age      0.000000e+00  2.958711e-05  1.360180e-05 8.685016e-02 9.766297e-01
## Medu     2.958711e-05  0.000000e+00 4.103940e-119 2.365988e-14 1.000901e-03
## Fedu     1.360180e-05 4.103940e-119  0.000000e+00 1.957421e-10 2.895666e-01
## traveltime 8.685016e-02  2.365988e-14  1.957421e-10 0.000000e+00 8.886286e-03
## studytime  9.766297e-01  1.000901e-03  2.895666e-01 8.886286e-03 0.000000e+00
## failures   1.734578e-16  3.973461e-10  2.436376e-10 2.949019e-03 2.404456e-06
```



Estudio ara possibles correlacions entre variables categòriques i numèriques. Faig diversos boxplots per veure de manera preliminar com varia el *grade_mean* en funció de diferents variables categòriques:



D'aquest dos conjunts de boxplots, trio els que semblen més interessants en relació amb el *grade* G:

- G ~ sex
- G ~ address
- G ~ tipus (assignatura)
- G ~ nursery
- G ~ internet

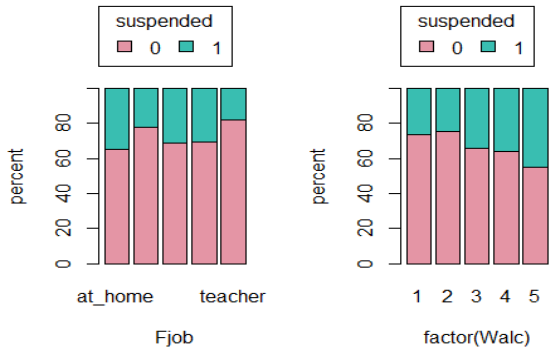
Per acabar el plantejament dels anàlisis, creo una variable dicotòmica anomenada *suspended* que, en funció de *grade_mean*, identifiqui si un estudiant ha suspès o no. Aquesta nova variable ens serà útil en el moment de plantejar models de regressió logística.

- Aprovat (suspended = 0): G => 10
- Suspens (suspended = 1): G < 10

```
##  [1] 1 1 0 1 1 0 0 0 0 0
## Levels: 0 1

##  [1]  0  0 15 11 10 10 16  9 10 11
```

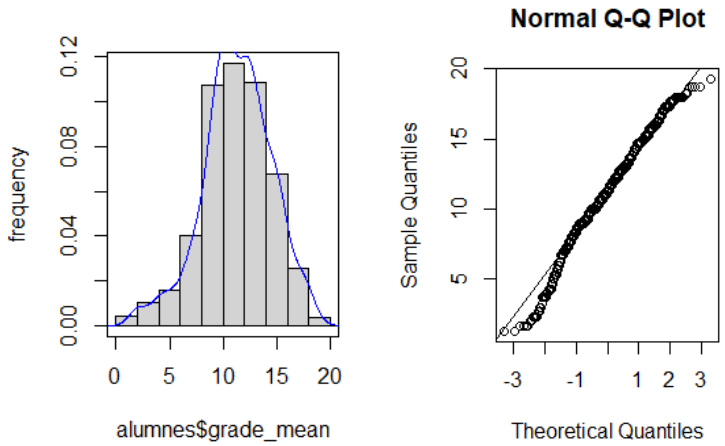
Així doncs, genero els següents gràfics de barres com a exemple:



Del segon diagrama es pot deduir clarament que proporció de suspensos augmenta quan més alt és el consum d'alcohol setmanal.

4.2. Comprovació de la normalitat i homogeneïtat de la variància

Comprovo la normalitat de la variable *Grade* que tractarem mitjançant un histograma amb corva de densitat i un diagrama de quantils qqplot:

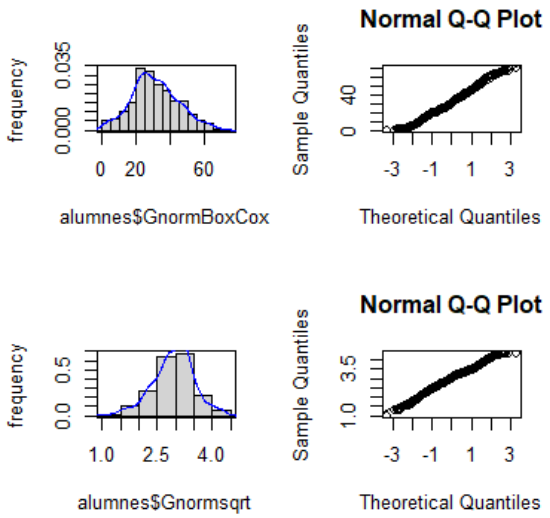



```
##  Shapiro-Wilk normality test
##
## data:  alumnes$grade_mean
## W = 0.98824, p-value = 3.958e-07
```

Es pot observar el següent:

- Forta tendència central al voltant de *Grade* = 10 .
- Alineació amb distribució normal sobretot entre els quantils -1 i 2, aproximadament.
- Malgrat que visualment la distribució no sembla distar d'una normal, el p value resultant de la aplicació del Shapiro-Wilk Test ens diu que hem de rebutjar la hipòtesi nul·la de normalitat de dades per la variable *Grade*.

Veient que la distribució de la mostra no podem considerar que segueixi una distribució normal, apliquem dues transformacions (*BoxCox* i *sqrt*) per intentar que s'ajusti a una distribució normal:



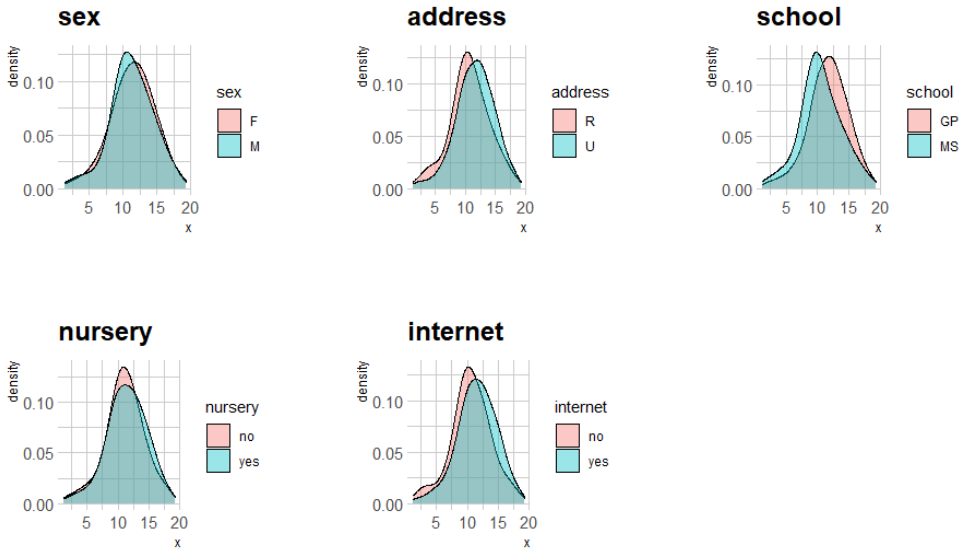

```
##
##  Shapiro-Wilk normality test
##
## data:  alumnes$GnormBoxCox
## W = 0.9911, p-value = 1.077e-05

##
##  Shapiro-Wilk normality test
##
## data:  alumnes$Gnormsqrt
## W = 0.98952, p-value = 1.639e-06
```

Malgrat haver fet dues transformacions, la distribució mostral de la variable segueix essent no normal pel test de Shapiro-Wilk. Així doncs, amb les observacions anteriors tenim dues opcions per seguir:

- A) Assumpció de normalitat seguint el TLC (n>30) i aplicació de proves paramètriques.
- B) Assumpció de no-normalitat i aplicació de proves no paramètriques.

Per acabar el plantejament dels anàlisis comprovo ara les gràfiques de densitat de les parelles de variables a analitzar del punt 4.1.:



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

4.3.1. Aplicació de proves paramètriques

Com hem explicat abans, assumim normalitat per fer proves paramètriques. Per assumir normalitat en la distribució poblacional, ens recolzem en l'aplicació del teorema del límit central (TLC) i en que la mida de cadascuna de les mostres és superior a 30 registres:

```
## [1] "Les mostres tenen una grandària de:"

## [1] "sex 1:2"        "sex c(559, 431)"
## [1] "address 1:2"        "address c(276, 714)"
## [1] "school 1:2"        "school c(719, 271)"
## [1] "nursery 1:2"        "nursery c(197, 793)"
## [1] "internet 1:2"        "internet c(214, 776)"
```

Comprovo ara l'homoscedasticitat de les diferents parelles de variables, és a dir, faré tests de variàncies que diran si les variàncies de les dues poblacions de cada variable són iguals.

```
## [1] "P-values resultants de var.test:"

## [1] "G ~ sex : 1 > 0.05? TRUE"
## [1] "G ~ address : 0.46 > 0.05? TRUE"
## [1] "G ~ school : 0.38 > 0.05? TRUE"
## [1] "G ~ nursery : 0.62 > 0.05? TRUE"
## [1] "G ~ internet : 0.91 > 0.05? TRUE"
```

Com queda demostrat, les parelles de variables establertes per l'estudi tenen un p-value superior al nivell de significància *alpha = 0.05*, fet que no ens permet refusar la hipòtesi nul·la de igualtat entre variàncies i, per tant, **es pot assumir que les variàncies poblacionals desconegudes són iguals (homoscedasticitat) en totes les parelles de variables.**

Formulem ara les preguntes d'estudi genèriques segons el que s'ha pogut observar anteriorment en els boxplots:

1. El sexe femení té *Grades* superiors al sexe masculí?
2. El *Grade* és superior en alumnes residents en l'àmbit urbà respecte alumnes residents en entorns rurals?
3. El *Grade* és superior en l'escola GP que en la escola MS?
4. Els alumnes que han passat per una guarderia/escola bressol tenen un valor de *Grade* més elevat que els que no hi han passat?
5. L'alumnat que té accés a internet té *Grade* major que el que no disposa d'internet?
6. El *Grade* és el mateix entre alumnes amb families nombroses i no?

Responem doncs a les següents preguntes:

```
## [1] "P-values resultants de t.test:"

## [1] "G ~ sex : 0.2958 > 0.05? TRUE"
## [1] "G ~ address : 2.043e-05 > 0.05? FALSE"
## [1] "G ~ school : 8.69e-08 > 0.05? FALSE"
## [1] "G ~ nursery : 0.08105 > 0.05? TRUE"
## [1] "G ~ internet : 4.677e-05 > 0.05? FALSE"
```

Com es pot comprovar, les respostes a les preguntes anteriors serien:

1. El valor p obtingut és major que el nivell de significança i, per tant, no es pot rebutjar la hipòtesi nul·la d'igualtat de mitjanes de notes (*Grades)* entre sexe femení i masculí. Amb un nivell de confiança del 95%, **es pot rebutjar estadísticament la idea que en general les alumnes de sexe femení tenen *Grades* més elevats que els alumnes de sexe masculí**.
2. El valor p obtingut és menor que el nivell de significança i es pot rebutjar la hipòtesi nul·la d'igualtat de mitjanes de notes (*Grades)* entre alumnes urbanites i rurals. Amb un nivell de confiança del 95%, **es pot afirmar estadísticament que, en general, l'alumnat que viu en àmbits urbans tenen *Grades* més elevats que l'alumnat que viu en l'àmbit rural**.

3. Com en l'anterior punt, el valor p obtingut és menor que el nivell de significança i, per tant, es pot rebutjar la hipòtesi nul·la d'igualtat de mitjanes de notes (*Grades)* entre les dues escoles. Amb un nivell de confiança del 95%, **es pot afirmar que els alumnes de l'escola GP tenen *Grades* més elevats que l'alumnat de l'escola MS**.
4. El valor p obtingut és major que el nivell de significança i, per tant, no es pot rebutjar la hipòtesi nul·la d'igualtat de mitjanes de notes (*Grades)* entre l'alumnat que ha anat a guarderia i el que no. Amb un nivell de confiança del 95%, **s'ha de rebutjar estadísticament la idea que els alumnes que han passat per una guarderia/escola bressol tenen un valor de *Grade* més elevat que els que no hi han passat**.
5. El valor p obtingut porta a rebutjar la hipòtesi nul·la d'igualtat de mitjanes de notes (*Grades)* entre alumnes amb accés a internet o no. Amb un nivell de confiança del 95%, **es pot afirmar estadísticament que l'alumnat que té accés a internet té un *Grade* major que el que no disposa d'accés a internet**.

4.3.2. Aplicació de proves no paramètriques

El test no paramètric equivalent al test t de dues mostres independents és l'anomenat test de suma de rangs de Wilcoxon. Aquest test es realitza per a dues mostres independents de mida n1 i n2. Tot i que no s'assumeix que les poblacions siguin normals, s'assumeix que la distribució de les poblacions és la mateixa (la qual cosa implica variància igual). Donada aquesta assumpció, el test es pot aplicar sobre les medianes i també sobre les mitjanes de la població.

Abans hem assumit que les variàncies poblacionals desconegudes són iguals (homoscedasticitat) en cadascuna de les parelles de variables. A més, com s'ha pogut veure en les gràfiques de densitat de les parelles de variables, s'observa que segueixen distribucions molt semblants.

Apliquem doncs, aquestes proves no paramètriques de manera anàloga a com hem aplicat abans les proves T en el apartat anterior:

```
## [1] "P-values resultants de Wilcoxon tests:"

## [1] "G ~ sex : 0.2095 > 0.05? TRUE"
## [1] "G ~ address : 7.435e-06 > 0.05? FALSE"
## [1] "G ~ school : 2.884e-09 > 0.05? FALSE"
## [1] "G ~ nursery : 0.07823 > 0.05? TRUE"
## [1] "G ~ internet : 1.805e-05 > 0.05? FALSE"
```

Podem comprovar que els p-value resultants de les proves no-paramètriques per aquestes parelles de variables, porta a les mateixes conclusions que els p-value resultants de les proves paramètriques. Per tant, es pot concloure que la assumpció (opció A detallada en el subapartat anterior) de normalitat poblacional -i la conseqüent desviació respecte a la normal- no genera cap impacte significatiu en els resultats dels anàlisis plantejats.

4.3.3. Creació d'un model de regressió logística

Plantegem diferents models logistics i en calculem els AIC:

Suspended ~ address + school + internet + Walc + failures + studytime + traveltime

AIC = 1025.123

Suspended ~ school + internet + Walc + failures + studytime + traveltime

AIC = 1023.307

Suspended ~ address + school + internet + Walc + failures + studytime

AIC = 1021.668

Dels tres models plantejats, el que ajusta millor les dades és el tercer ja que té el AIC més petit dels tres.

Comprovem els OR.

```
## (Intercept)    schoolMS internetyes       Walc    failures   studytime
          ##   0.4143914   1.7254593   0.6726934   1.1197712   4.5086460   0.7803065
```

Podem veure que estudiar a l'escola *MS* , no tenir internet, tenir un alt consum d'alcohol setmanal i, sobretot, haver suspès assignatures amb anterioritat, són factors de risc a l'hora de suspendre les assignatures de Matemàtiques i Portuguès.

5. Representació dels resultats a partir de taules

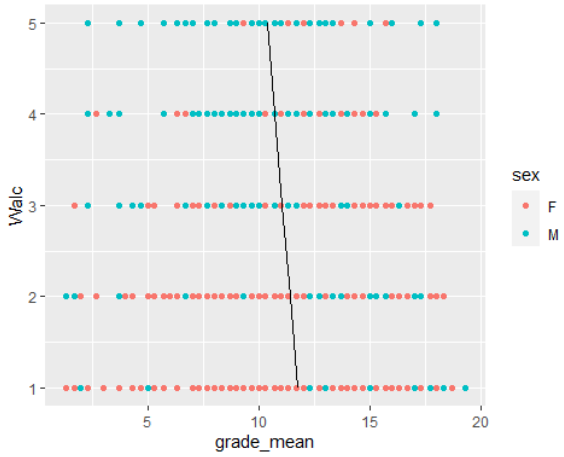
Durant el desenvolupament d'aquest document, s'ha pogut anar veient la visualització gràfica dels diferents resultats aportats. A continuació, resumim aquests resultats per arribar a les conclusions de l'apartat següent.

Taula 1. Taula de resultats

Pregunta d'estudi	**Test paramètric (T.test) p-valor resultant**	**Test no paramètric (Wilcoxon) p-valor resultant**	**Respostes amb un 95% de nivell de confiança *alpha (α) = 0.05***
El sexe femení té Grades superiors al sexe masculí?	*0.2958*	*0.2095*	**No es pot afirmar** que els Grades de sexe femení siguin superiors al sexe masculí
El Grade és superior en alumnes residents en l'àmbit urbà respecte alumnes residents en entorns rurals?	$2.043*10^{-5}$	$7.44*10^{-5}$	**Es pot afirmar** que el Grade **és superior** en alumnes residents en l'àmbit urbà respecte alumnes residents en entorns rurals.
El Grade és superior en l'escola GP que en la escola MS?	$8.69*10^{-8}$	$2.88*10^{-8}$	**Es pot afirmar** que el Grade **és superior** és superior en l'escola GP que en la escola MS.
Els alumnes que han passat per una guarderia/escola bressol tenen un valor de Grade més elevat que els que no hi han passat?	*0.08*	*0.08*	**No es pot afirmar** que els alumnes que han passat per una guarderia/escola bressol tenen un valor de Grade més elevat que els que no hi han passat.
L'alumnat que té accés a internet té Grade major que el que no disposa d'internet?	$4.68*10^{-5}$	$1.81*10^{-5}$	**Es pot afirmar** que l'alumnat que té accés a internet té Grade major que el que no disposa d'internet.

Conclusions de la comprovació dels ODD ratio del model de regressió logística que ajustava millor les dades en comparació amb altres 2 models amb major AIC.	
Quins són els factors de risc de suspendre?	- Estudiar a l'escola MS - No disposar d'accés a internet - Tenir un alt consum d'alcohol setmanal - Haver suspès assignatures amb anterioritat

Addicionalment, per més informació, la següent regressió lineal mostra la correlació negativa que indica que a menys consum d'alcohol el cap de setmana hi ha una lleugera millora en la qualificació en els estudis. També hem representat el sexe en colors on visualment identifiquem més punts masculins que femenins en els nivells més alts de consum d'alcohol en cap de setmana.



6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Hem començat l'estudi amb descripció i la la neteja de les dades. És necessari conèixer bé les variables amb què estem tractant, de quin tipus son i quina qualitat tenen. Per això hem iniciat el procés de neteja amb una integració de dos conjunts de dades i la selecció de totes les variables ja que a priori no s'ha vist necessari descartar-ne cap. També hem vist que es un conjunt net de valors buits o errors, no obstant cal fer la comprovació sempre per garantir-ne la qualitat. Hem acabat la neteja amb la cerca de valors extrems, on hem vist que no ha estat necessari tractar-los excepte per una variable que ens podria donar problemes per l'anàlisi posterior.

Seguidament hem passat a fer una tria de possibles variables d'interès mitjançant un gràfic de correlacions i una visualització de totes les variables categòriques i la seva relació amb la variable principal a estudiar que és la de les qualificacions anomenada *grade_mean* que hem calculat prèviament.

Posant el focus en aquesta variable n'hem fet proves de normalitat i homoscedasticitat per seguir amb les següents proves i optar principalment per proves paramètriques tot i no complir estrictament la normalitat. Hem estudiat algunes variables escollides respecte *grade_mean* i hem obtingut que el sexe femení no obté qualificacions estadísticament superiors al masculí però que el fet de viure en entorns urbans si que beneficia als alumnes en els seus estudis, entre d'altres comparacions.

També hem estudiat en un conjunt de variables concret quin model era millor per determinar el fet de suspendre els estudis.

Finalment a partir de l'obtingut a les correlacions de variables numèriques i els diagrames de caixa sobre la variable de *grade_mean* podrien concloure que els factors més determinants a aconseguir una bona nota son el nivell d'estudis dels pares així com la predisposició a voler cursar estudis superiors.

Participants

Contribucions	Signatura
Investigació prèvia	ARC,JML
Redacció de les respostes	ARC,JML
Desenvolupament del codi	ARC,JML