Projet Analyse de données

Rudio et Léo-Paul

2023-05-09

Présentation du projet et du jeu de données

Le jeu de données est constitués d'informations sur la vie d'étudiants dans une université du Portugal. Ces informations vont de leur résultats universitaires, leur vie familiale à leur consommation d'alcool. Le jeu a été construit à partir d'une enquête menée auprès d'étudiant en mathématiques et en portugais.

L'objectif serait alors d'analyser le jeu de données afin de comprendre les facteurs qui impactent la réussite scolaire de ces étudiants. L'intérêt du jeu est la grande variété de facteurs proposée qui permet de courvrir un maximum d'hypothèsesn, notamment celle sur la consommation d'alcool proposée directement par le nom du jeu de données.

Voici les variables présentent dans ce jeu de données ;

- **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)

These grades are related with the course subject, Math or Portuguese: - **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)

Au cours de ce projet, nous nous concentrons sur la variable G3 qui est la variable de sortie représentant la note finale des élèves. Il s'agirait donc d'un problème de régression sur la variables G3 ou même plus généralement un problème de classification.

Voici les étapes que nous allons suivre :

1. Identifier les variables significatives
2. Appliquer des méthodes de classification sur la réussite scolaire
3. Effectuer une regression linéaires pour prédire les notes/la réussite
4. Comparer des méthodes de machine learning pour prédire les notes/la réussite

1.Chargement des données

```
# Chargement de la base de données
df.mat=read.table("student-mat.csv",sep=",",header=TRUE,as.is = FALSE)
df.por=read.table("student-por.csv",sep=",",header=TRUE,as.is = FALSE)

# Etudiants qui appartiennent aux deux cours
both= merge(df.mat,df.por,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob",

# Concaténation des deux dataframes
df = rbind(df.mat,df.por)
head(df)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
```

```
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

2. Nettoyage et vérification des données

Le jeu est composé de 33 variables dont 17 qualitatives et 16 quantitatives. On rajoute une variable en plus pour la réussite scolaire.

```
print(str(df))
```

```
## 'data.frame':    1044 obs. of  33 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ 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      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ 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 : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ 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 ...
## NULL
```

```
print(nrow(df))
```

```
## [1] 1044
```

```r
## On calcule la moyenne des étudiants
df$Moy = (df$G1+df$G2+df$G3)/3

## On rajoute la réussite scolaire comme variable qualitative
df$RS = factor(df$Moy>=10)
head(df)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3       Moy    RS
## 1        6  5  6  6  5.666667 FALSE
## 2        4  5  5  6  5.333333 FALSE
## 3       10  7  8 10  8.333333 FALSE
## 4        2 15 14 15 14.666667  TRUE
## 5        4  6 10 10  8.666667 FALSE
## 6       10 15 15 15 15.000000  TRUE
```

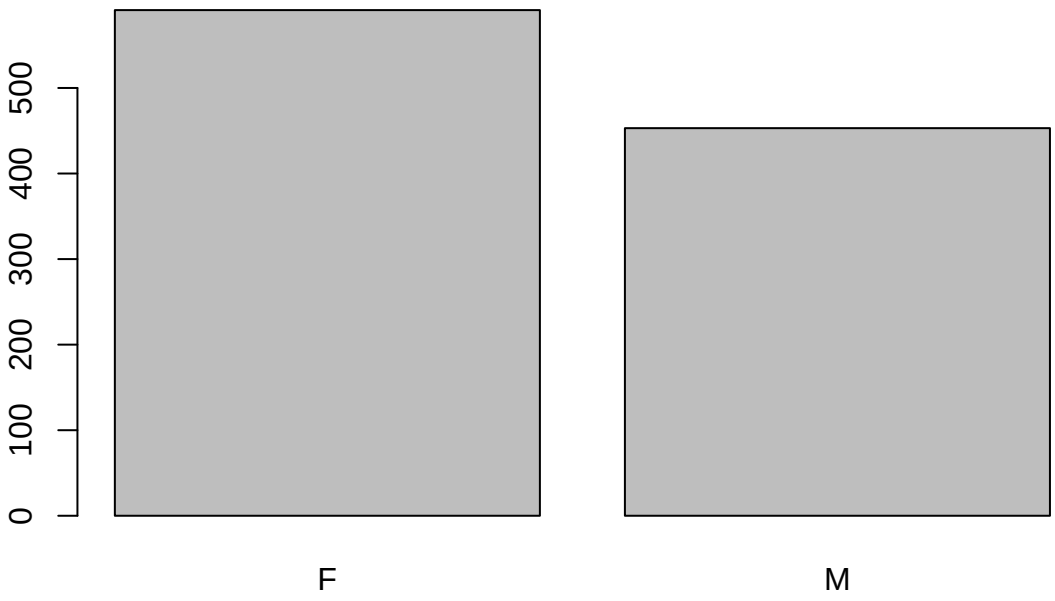
3. Exploration des données : études des variables

Cette partie consiste à appliquer des méthodes de statistiques descriptives afin de mieux comprendre le jeu de données. On se concentre sur l'analyse de la distribution des variables et leur corrélation avec les résultats scolaires.

Le sexe des étudiants

D'après le diagramme, le dataset est plutôt équilibré en terme d'hommes et de femmes. On étudie ensuite le lien entre le sexe et les notes en effectant une ANOVA1. D'après le test de Fisher, p-value $> 5\%$ donc il n'y a pas d'effet du sexe sur les notes.

```r
S = table(df$sex)
barplot(S,main="Répartion des sexes")
```



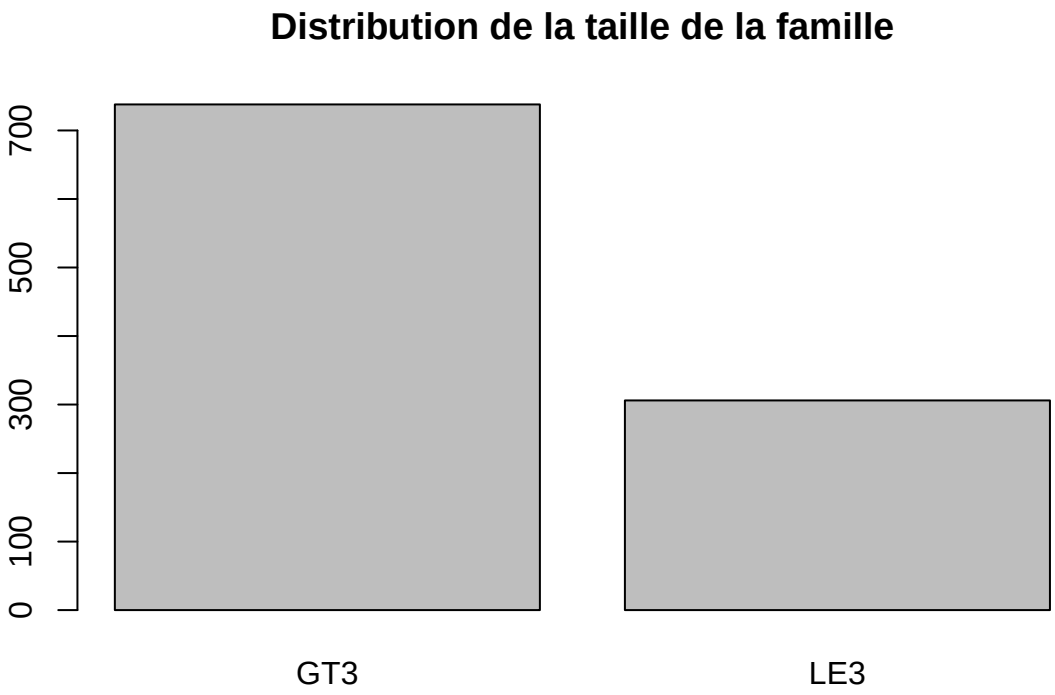

```
res = lm(Moy ~ sex,data=df)
summary(res)
```

```
##
## Call:
## lm(formula = Moy ~ sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0152  -2.0152  -0.0152   2.1722   8.1722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3486     0.1324  85.706   <2e-16 ***
## sexM         -0.1874     0.2010  -0.932    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.219 on 1042 degrees of freedom
## Multiple R-squared:  0.0008335,  Adjusted R-squared:  -0.0001254
## F-statistic: 0.8693 on 1 and 1042 DF,  p-value: 0.3514
```

La taille de la famille

On a deux fois plus de grandes familles que de petites familles. D'après le test de Fisher, il y a bien un impact de taille de la famille sur les notes.

```
Fam = table(df$famsize)
barplot(Fam,main="Distribution de la taille de la famille")
```



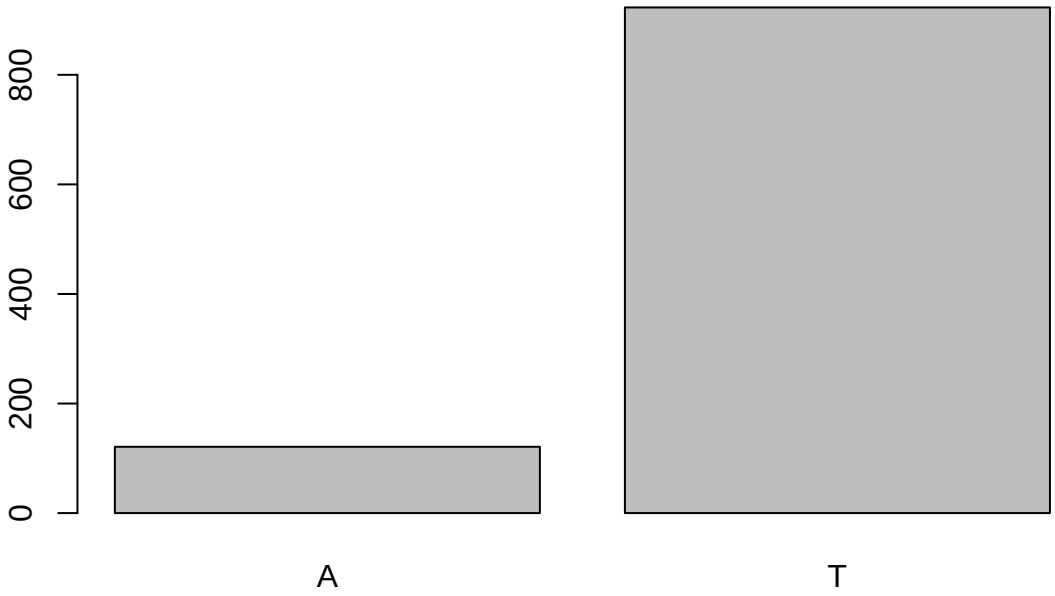

```
summary(lm(Moy ~ famsize,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ famsize, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9096 -1.9096 -0.1391  2.1942  8.1942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1391     0.1183   94.15   <2e-16 ***
## famsizeLE3    0.4371     0.2185    2.00   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.214 on 1042 degrees of freedom
## Multiple R-squared:  0.003825,   Adjusted R-squared:  0.002869
## F-statistic: 4.001 on 1 and 1042 DF,  p-value: 0.04573
```

Situation familliale : séparation des parents

Le jeu est très déséquilibré au sujet de la situation famille : il y a 4 fois plus d'étudiants qui ont leurs parents qui vivent ensemble. De plus, le test de Fisher indique que la situation familliale n'a pas d'impact sur les notes.

```
barplot(table(df$Pstatus),main="Distribution de la situation familiale")
```



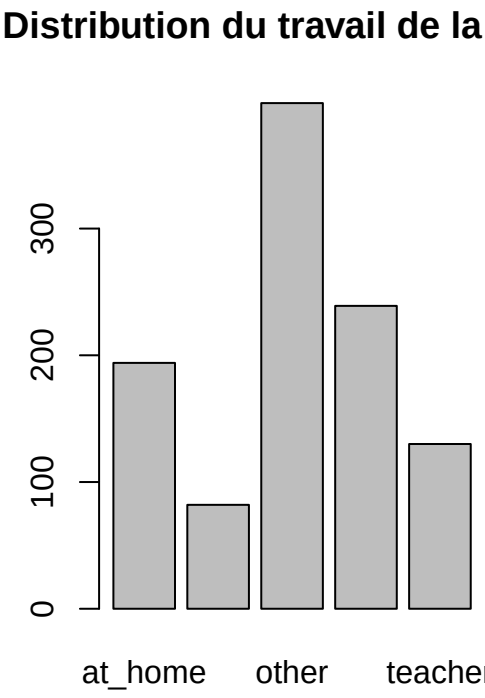

```
summary(lm(Moy ~ Pstatus,data=df))
```

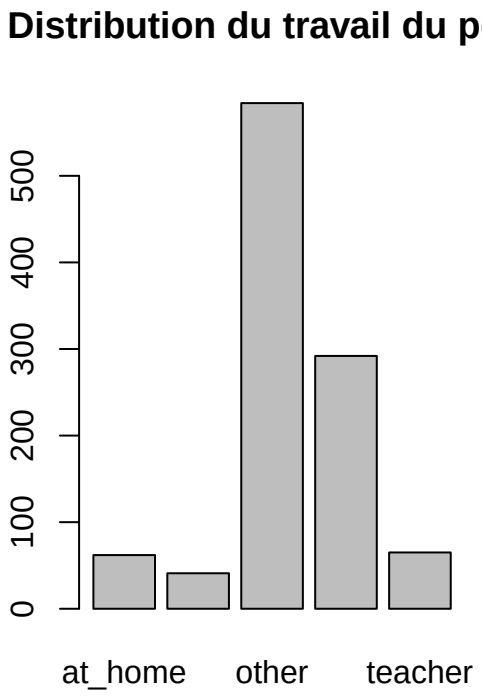
```
##
## Call:
## lm(formula = Moy ~ Pstatus, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0744  -1.9155   0.0845   2.0845   8.0845
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4077     0.2927   38.97   <2e-16 ***
## PstatusT     -0.1589     0.3113   -0.51     0.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.22 on 1042 degrees of freedom
## Multiple R-squared:  0.0002499,  Adjusted R-squared:  -0.0007095
## F-statistic: 0.2605 on 1 and 1042 DF,  p-value: 0.6099
```

Travail des parents

Dans les deux cas, others et services sont les catégories qui dominent. Une différence notable est la que la proportion de femme au-foyer est bien plus élevée que celle des hommes. D'après le test de Fisher, le travail de la mère a un impact sur les notes, contrairement à celui du père.

```
par(mfrow=c(1,2))
barplot(table(df$Mjob),main="Distribution du travail de la mère")
barplot(table(df$Fjob),main="Distribution du travail du père")
```





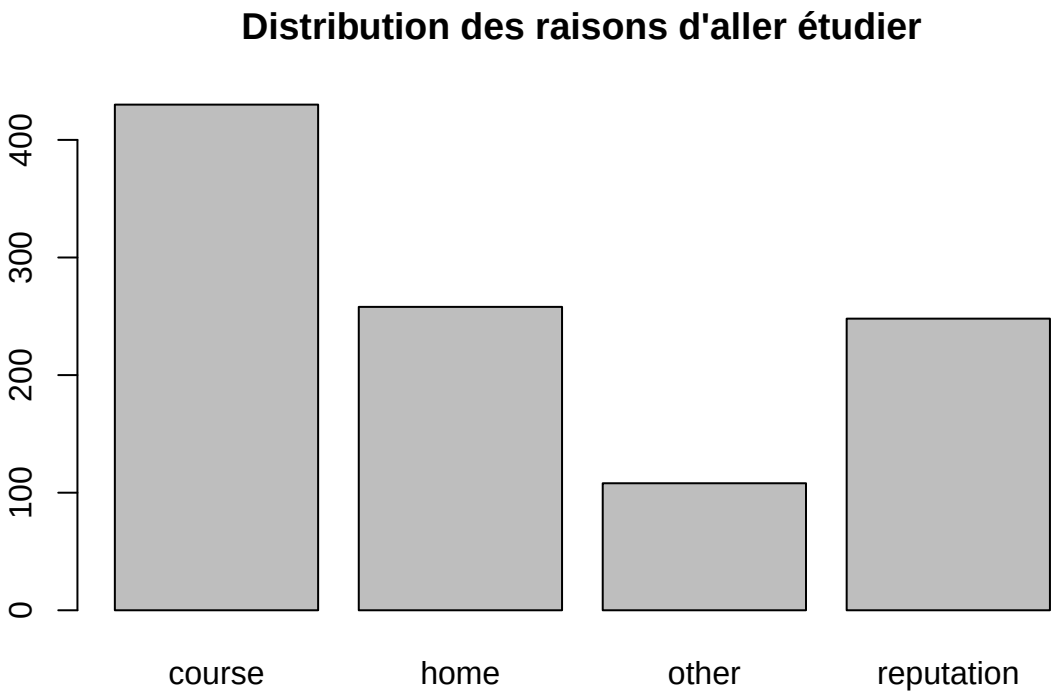

```
summary(lm(Moy ~ Medu+Fedu,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ Medu + Fedu, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.265  -1.732   0.068   2.126   7.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4233     0.2595  36.316  < 2e-16 ***
## Medu          0.5214     0.1125   4.635 4.02e-06 ***
## Fedu          0.2037     0.1150   1.771   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.133 on 1041 degrees of freedom
## Multiple R-squared:  0.05434,    Adjusted R-squared:  0.05252
## F-statistic: 29.91 on 2 and 1041 DF,  p-value: 2.344e-13
```

Les raisons d'aller en cours

D'après le digramme circulaire, seule "other" possède un petit effectif alors que "course" domine. Ainsi, les élèves vont majoritairement en cours car ils les apprécient. D'après l'ANOVA1, il est clair que la raison d'aller en cours impacte les notes des étudiants (p-value $< 5\%$). Cela paraît cohérent étant donné que cela détermine leur motivation à avoir de bonnes notes.

```r
barplot(table(df$reason),main="Distribution des raisons d'aller étudier")
```




```r
summary(lm(Moy~ reason,data=df))
```

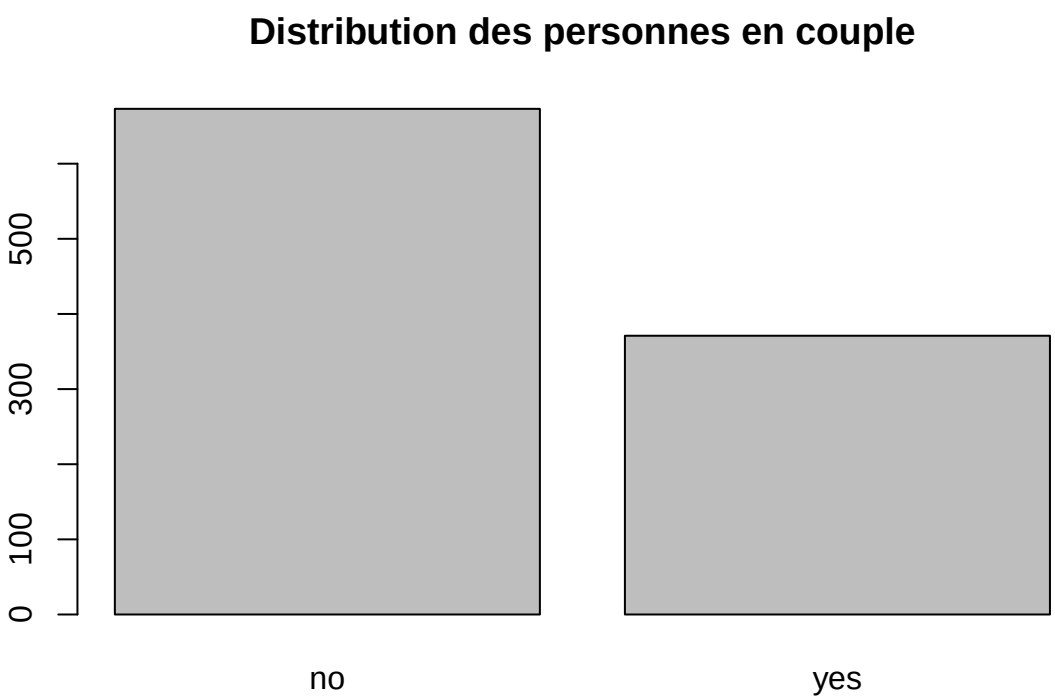
```
##
## Call:
## lm(formula = Moy ~ reason, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3858  -1.8791  -0.0052   2.1209   7.7876
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.87907    0.15372  70.771  < 2e-16 ***
## reasonhome        0.45943    0.25103   1.830   0.0675 .
## reasonother      -0.03956    0.34309  -0.115   0.9082
## reasonreputation  1.17335    0.25417   4.616 4.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.188 on 1040 degrees of freedom
## Multiple R-squared:  0.02209,    Adjusted R-squared:  0.01927
## F-statistic: 7.832 on 3 and 1040 DF,  p-value: 3.587e-05
```

Les relations

Il y a environ deux fois plus de jeunes célibataires que de jeunes en couple. On peut penser qu'être en couple réduit le temps passé à étudier et rajoute des distractions, donc il devrait avoir un impact négatif sur les notes. D'après le test de Fisher, la p-value est fortement inférieure à 5%, donc on rejette H0: il y a bien un lien entre situation romantique et notes, ce qui rejoint bien l'idée de départ. Il serait donc intéréssant d'étudier la distribution des notes selon la situation romantique. D'après les boxplots, les différences sont

assez minimes, même si on peut aperçevoir que les notes des célibataires sont légèrement meilleures.

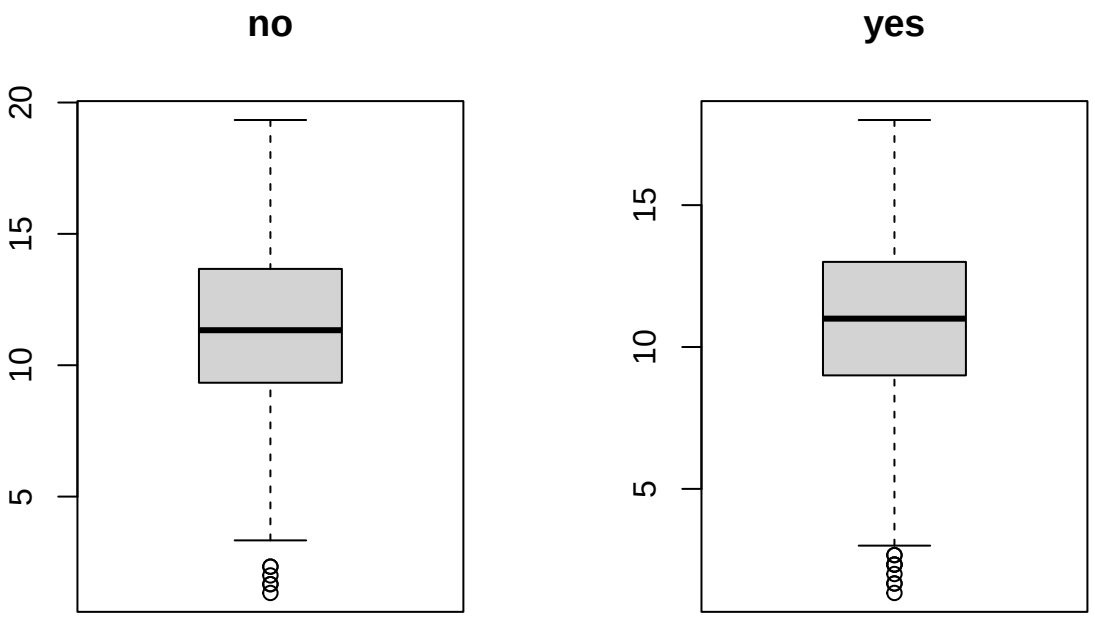
```
barplot(table(df$romantic),main="Distribution des personnes en couple")
```




```
summary(lm(Moy~ romantic,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ romantic, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1486  -1.9455   0.1222   2.1847   7.8514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4819     0.1236  92.871  < 2e-16 ***
## romanticyes  -0.6041     0.2074  -2.913  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.207 on 1042 degrees of freedom
## Multiple R-squared:  0.008077,   Adjusted R-squared:  0.007125
## F-statistic: 8.485 on 1 and 1042 DF,  p-value: 0.003658
```

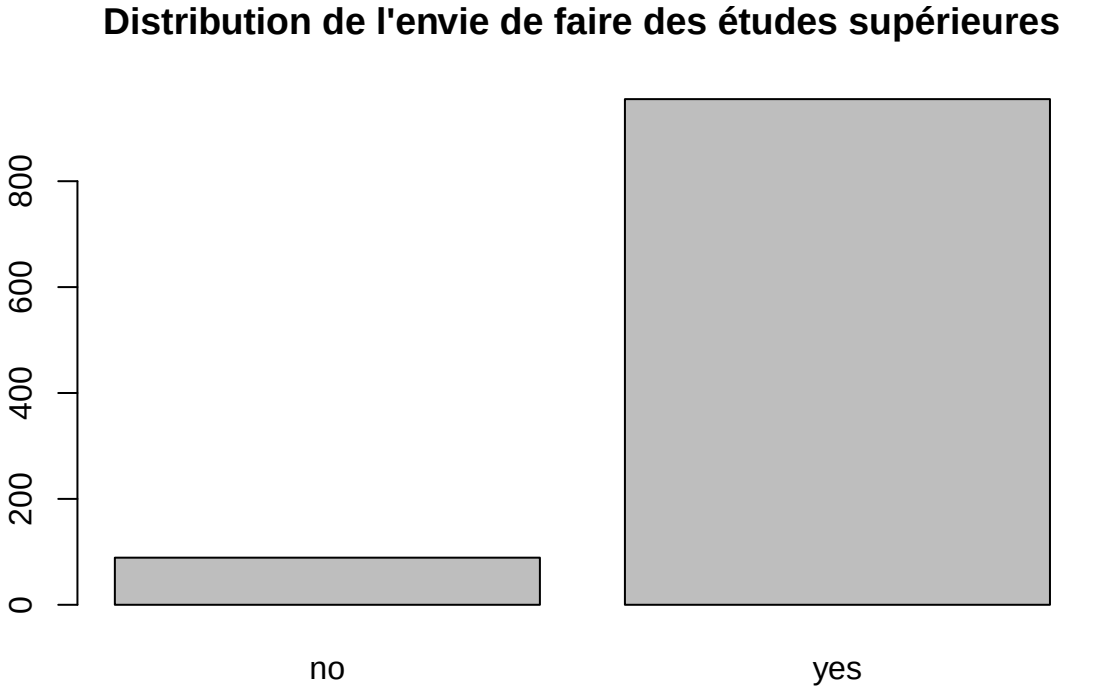
```
yes = df$Moy[df$romantic=='yes']
no = df$Moy[df$romantic=='no']

# Boxplot des notes
par(mfrow=c(1,2))
boxplot(no,main="no")
boxplot(yes,main="yes")
```



Volonté de faire des études supérieures

On observe qu'au moins 80% des élèves veulent continuer leur études après le lycée, ce qui est plutôt rassurant. De plus, d'après le test de Fisher, les deux variables sont corrélées. On peut également annoncer que ceux qui veulent faire des études supérieures tendent à avoir de meilleures notes grâce au test unilatéral.

```
barplot(table(df$higher),main="Distribution de l'envie de faire des études supérieures")
```




```
summary(lm(Moy ~ higher,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ higher, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1930  -1.8597   0.1403   2.1403   7.8070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.4869     0.3293  25.775   <2e-16 ***
## higheryes     3.0395     0.3443   8.829   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.106 on 1042 degrees of freedom
## Multiple R-squared:  0.0696, Adjusted R-squared:  0.06871
## F-statistic: 77.95 on 1 and 1042 DF,  p-value: < 2.2e-16
```

```r
yes = df$Moy[df$higher=='yes']
no = df$Moy[df$higher=='no']

# Boxplot des notes
par(mfrow=c(1,2))
hist(no,main="no")
hist(yes,main="yes")
```