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 variabales 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 G3
4. Comparer des méthodes de machine learning pour prédire G3

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. Description 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))
# print(nrow(df))

## On rajoute la réussite scolaire comme variable qualitative
df$RS = factor(df$G3>=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    RS
## 1        6  5  6  6 FALSE
## 2        4  5  5  6 FALSE
## 3       10  7  8 10  TRUE
## 4        2 15 14 15  TRUE
## 5        4  6 10 10  TRUE
## 6       10 15 15 15  TRUE
```

```
data=df
data_quanti=data[c(3,7,8,13,14,15,25,26,27,28,29,30,31,32,33)]
data_quanti_mat=df.mat[c(3,7,8,13,14,15,25,26,27,28,29,30,31,32,33)]
data_quanti_por=df.por[c(3,7,8,13,14,15,25,26,27,28,29,30,31,32,33)]
head(data_quanti)
```

```
##   age Medu Fedu traveltime studytime failures freetime goout Dalc Walc health
```

```
## 1  18    4    4          2         2        0        3     4    1    1      3
## 2  17    1    1          1         2        0        3     3    1    1      3
## 3  15    1    1          1         2        3        3     2    2    3      3
## 4  15    4    2          1         3        0        2     2    1    1      5
## 5  16    3    3          1         2        0        3     2    1    2      5
## 6  16    4    3          1         2        0        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
```

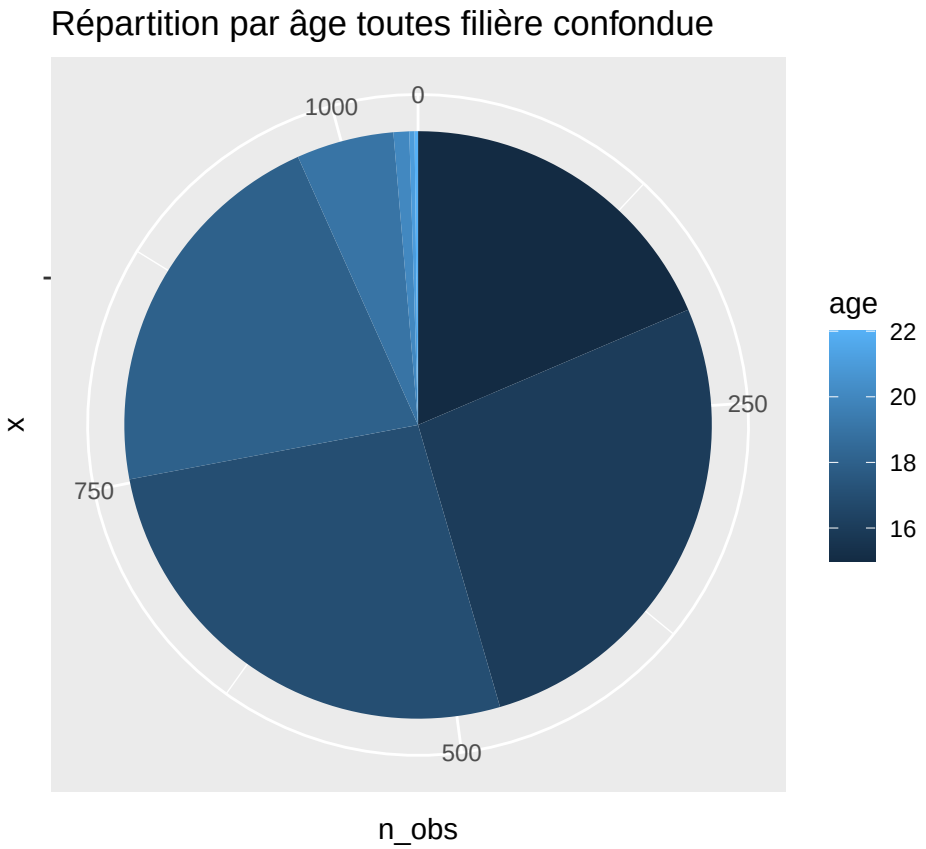
```
library(ggplot2)
library("dplyr")
```

```
##
## Attachement du package : 'dplyr'
```

```
## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag
```

```
## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
attach(data_quanti)
data_age=data_quanti

data_age=summarise(group_by(data_age,age),n_obs=n()) #on groupe par âge avec le nombre de personnes dan

#création du camembert
ggplot(data = data_age, aes(x = "", y = n_obs, fill = age)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Répartition par âge toutes filière confondue")
```

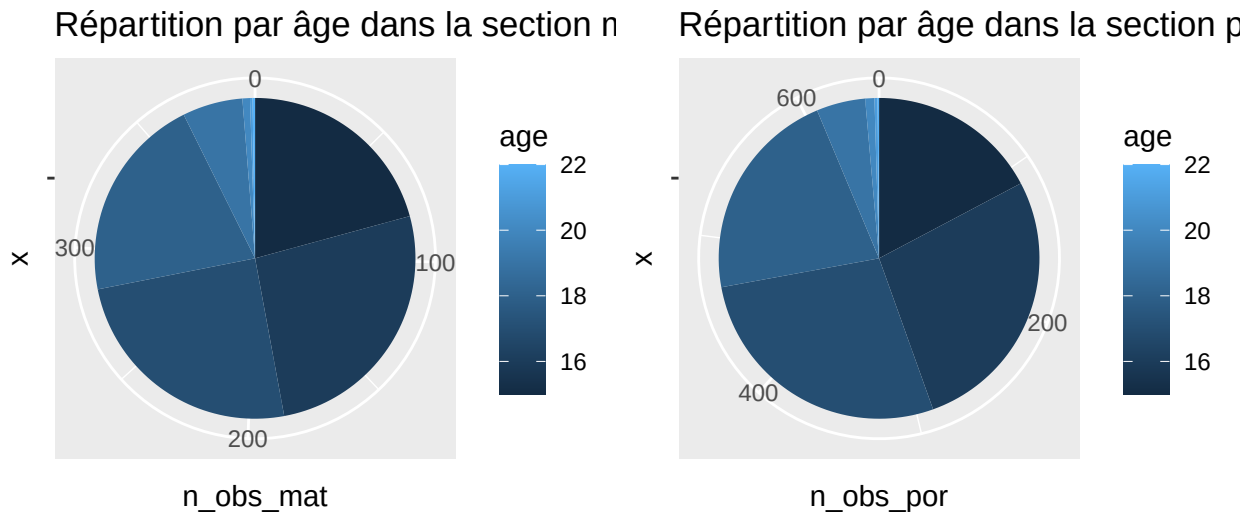


La couleur la plus claire correspond à l'âge le plus grand (22 ans), dès que l'on passe à une couleur plus foncée, on diminue l'âge de 1. On voit clairement ici que la majorité des étudiants ont entre 15 et 19 ans.

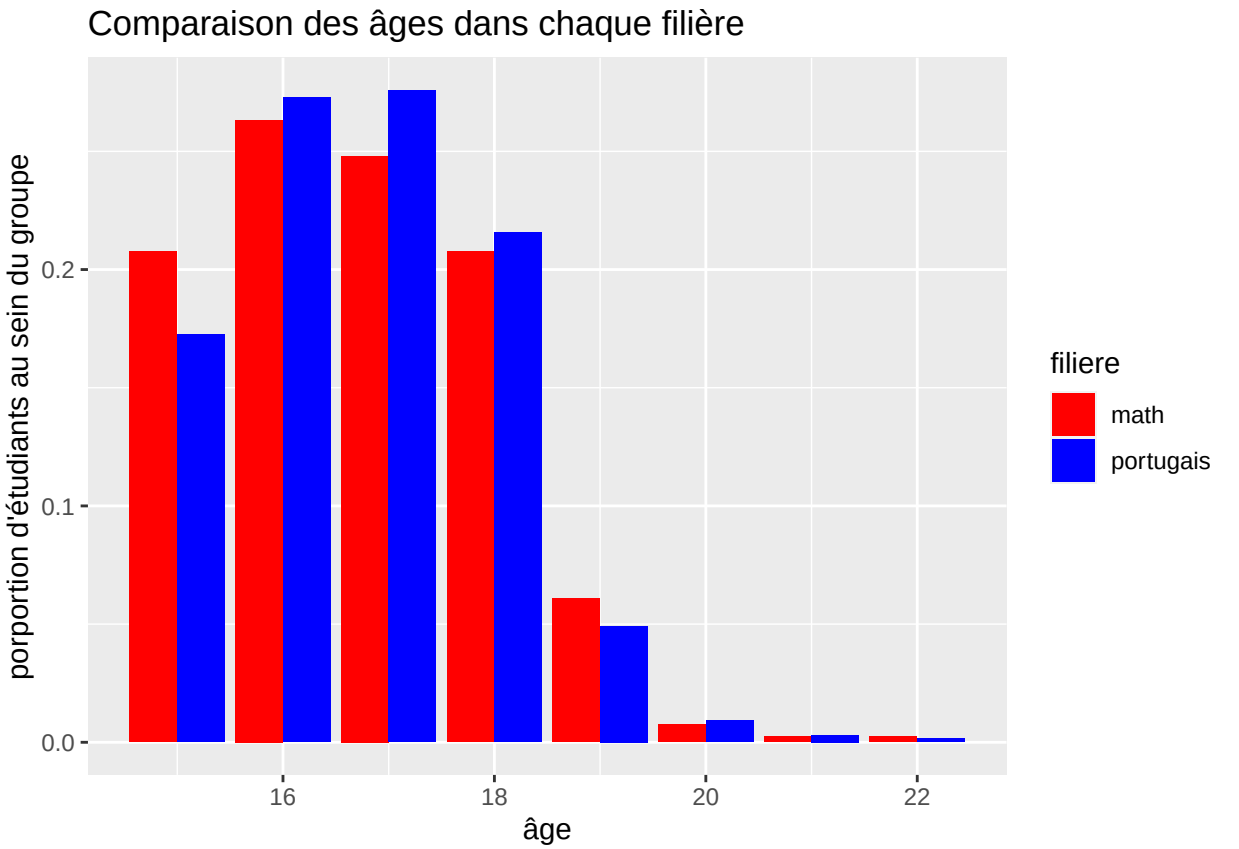
```
data_age_mat=data_quanti_mat
data_age_por=data_quanti_por

data_age_mat=summarise(group_by(data_age_mat,age),n_obs_mat=n()) #on groupe par âge avec le nombre de p
data_age_por=summarise(group_by(data_age_por,age),n_obs_por=n())

#création du camembert
p1=ggplot(data = data_age_mat, aes(x = "", y = n_obs_mat, fill = age)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Répartition par âge dans la section maths")

p2=ggplot(data = data_age_por, aes(x = "", y = n_obs_por, fill = age)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Répartition par âge dans la section portugais")


library(gridExtra)
```

```
##
## Attachement du package : 'gridExtra'

## L'objet suivant est masqué depuis 'package:dplyr':
##
```

```
##     combine
grid.arrange(p1, p2, ncol = 2)
```

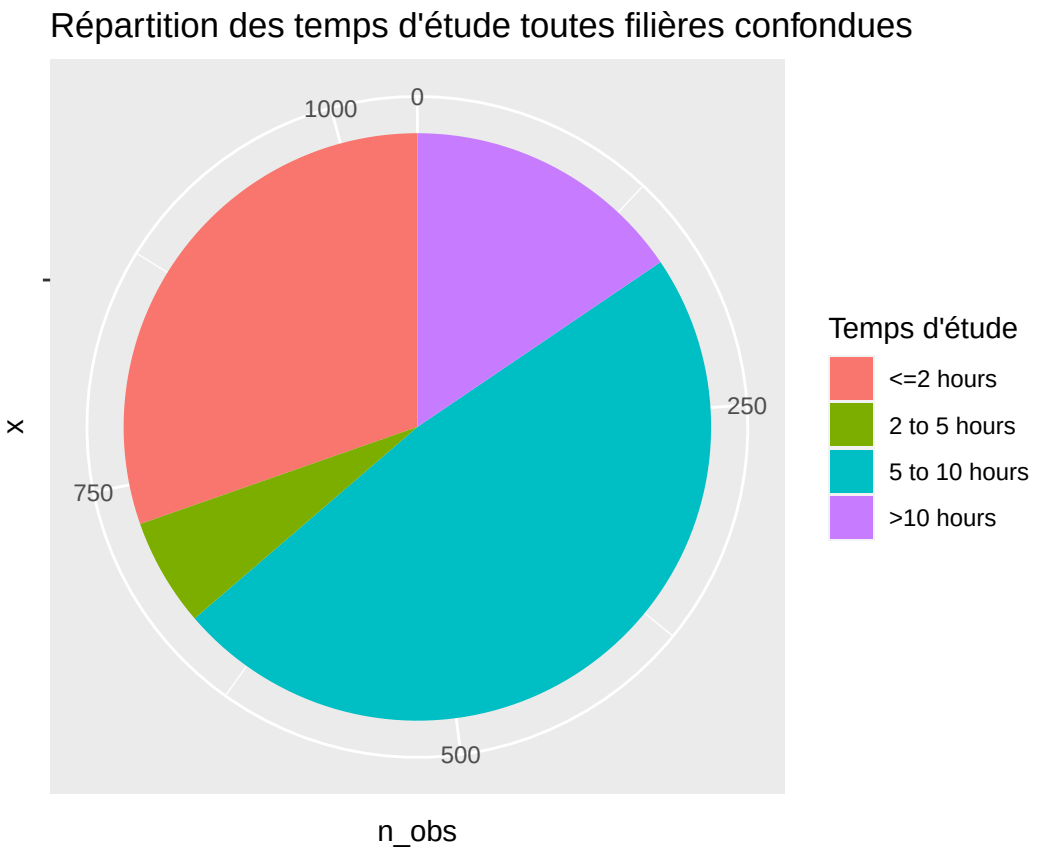



```
data_age_mat <- data_age_mat %>%
  mutate(proportion = n_obs_mat / sum(n_obs_mat))

data_age_por <- data_age_por %>%
  mutate(proportion = n_obs_por / sum(n_obs_por))

#on renome de la même manière les colonnes du nombre d'étudiants pour chaque obersvation
data_age_mat$filiere=c(rep("math",nrow(data_age_mat)))
data_age_por$filiere=c(rep("portugais",nrow(data_age_por)))
colnames(data_age_mat)[colnames(data_age_mat) == "n_obs_mat"] <- "n_obs"
colnames(data_age_por)[colnames(data_age_por) == "n_obs_por"] <- "n_obs"

#on concatène les deux datas frame
data_age=rbind(data_age_mat,data_age_por)

#Création du graphique

ggplot(data_age, aes(x = age, y = proportion, fill = filiere)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Comparaison des âges dans chaque filière", x ="âge", y = "porportion d'étudiants au sei
  scale_fill_manual(values = c("red", "blue"))
```



On voit que la répartion d'âge est la même dans chaque filière

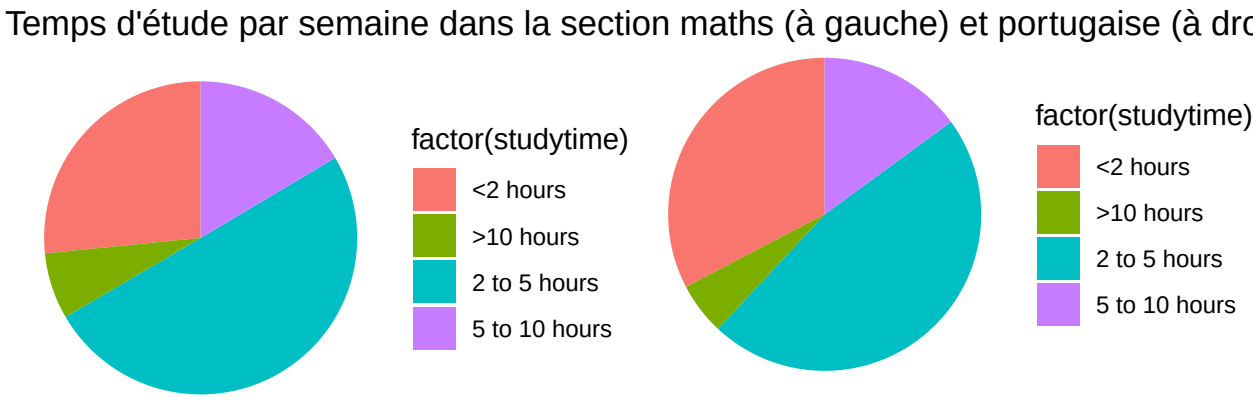
```
data_stud=data_quanti
data_stud=summarise(group_by(data_stud,studytime),n_obs=n()) #on groupe par temps d'étude par semaine

data_stud$studytime[data_stud$studytime == 1] <- "<2 hours"
data_stud$studytime[data_stud$studytime == 2] <- "2 to 5 hours"
data_stud$studytime[data_stud$studytime == 3] <- "5 to 10 hours"
data_stud$studytime[data_stud$studytime == 4] <- ">10 hours"


ggplot(data_stud, aes(x = "", y = n_obs, fill = factor(studytime))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Répartition des temps d'étude toutes filières confondues") +
  scale_fill_discrete(name = "Temps d'étude", labels = c("<=2 hours", "2 to 5 hours", "5 to 10 hours",
```



On voit clairement que les étudiants travaillent majoritairement moins de 2h00 ou entre 5h00 et 10h00 par semaines.

```
#creation data frame stud pour le groupe portugais
data_stud_por=data_quanti_por
data_stud_por=summarise(group_by(data_stud_por,studytime),n_obs_por=n()) #on groupe par temps d'étude p

data_stud_por$studytime[data_stud_por$studytime == 1] <- "<2 hours"
data_stud_por$studytime[data_stud_por$studytime == 2] <- "2 to 5 hours"
data_stud_por$studytime[data_stud_por$studytime == 3] <- "5 to 10 hours"
data_stud_por$studytime[data_stud_por$studytime == 4] <- ">10 hours"

#creation data frame stud pour le groupe mat b
data_stud_mat=data_quanti_mat
data_stud_mat=summarise(group_by(data_stud_mat,studytime),n_obs_mat=n()) #on groupe par temps d'étude p

data_stud_mat$studytime[data_stud_mat$studytime == 1] <- "<2 hours"
data_stud_mat$studytime[data_stud_mat$studytime == 2] <- "2 to 5 hours"
data_stud_mat$studytime[data_stud_mat$studytime == 3] <- "5 to 10 hours"
data_stud_mat$studytime[data_stud_mat$studytime == 4] <- ">10 hours"

library(gridExtra)

#création des camemberts pour les deux sections
p1=ggplot() +
  # Premier camembert
  geom_bar(data = data_stud_mat, aes(x = "", y = n_obs_mat, fill = factor(studytime)), stat = "identity
```

```
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "Temps d'étude par semaine dans la section maths (à gauche) et portugaise (à droite)")

  # Deuxième camembert
p2=ggplot() +
  geom_bar(data = data_stud_por, aes(x = "", y = n_obs_por, fill = factor(studytime)), stat = "identity
  coord_polar(theta = "y") +
  theme_void()

grid.arrange(p1, p2, ncol = 2)
```

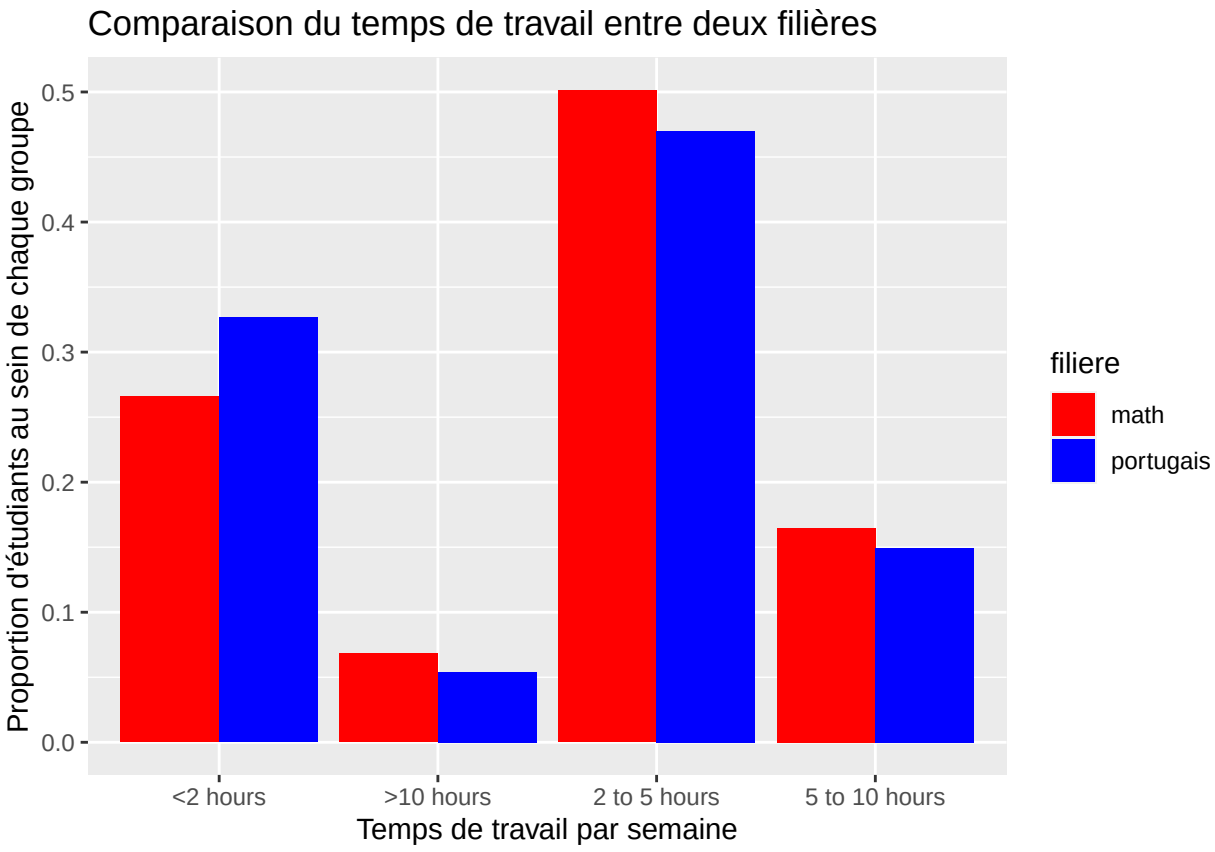



```
data_stud_mat
```

```
## # A tibble: 4 x 2
##   studytime     n_obs_mat
##   <chr>             <int>
## 1 <2 hours            105
## 2 2 to 5 hours        198
## 3 5 to 10 hours        65
## 4 >10 hours            27
```

On voit qu'il y a plus de personnes qui travaillent moins de deux heures par semaine dans la section portiguaise tandis qu'il y a moins de personnes qui travaillent plus de 10h00 dans cette même section. Le nombre d'étudiants travaillant entre 5 et 10 heures semble être a peu près le même. En effet:

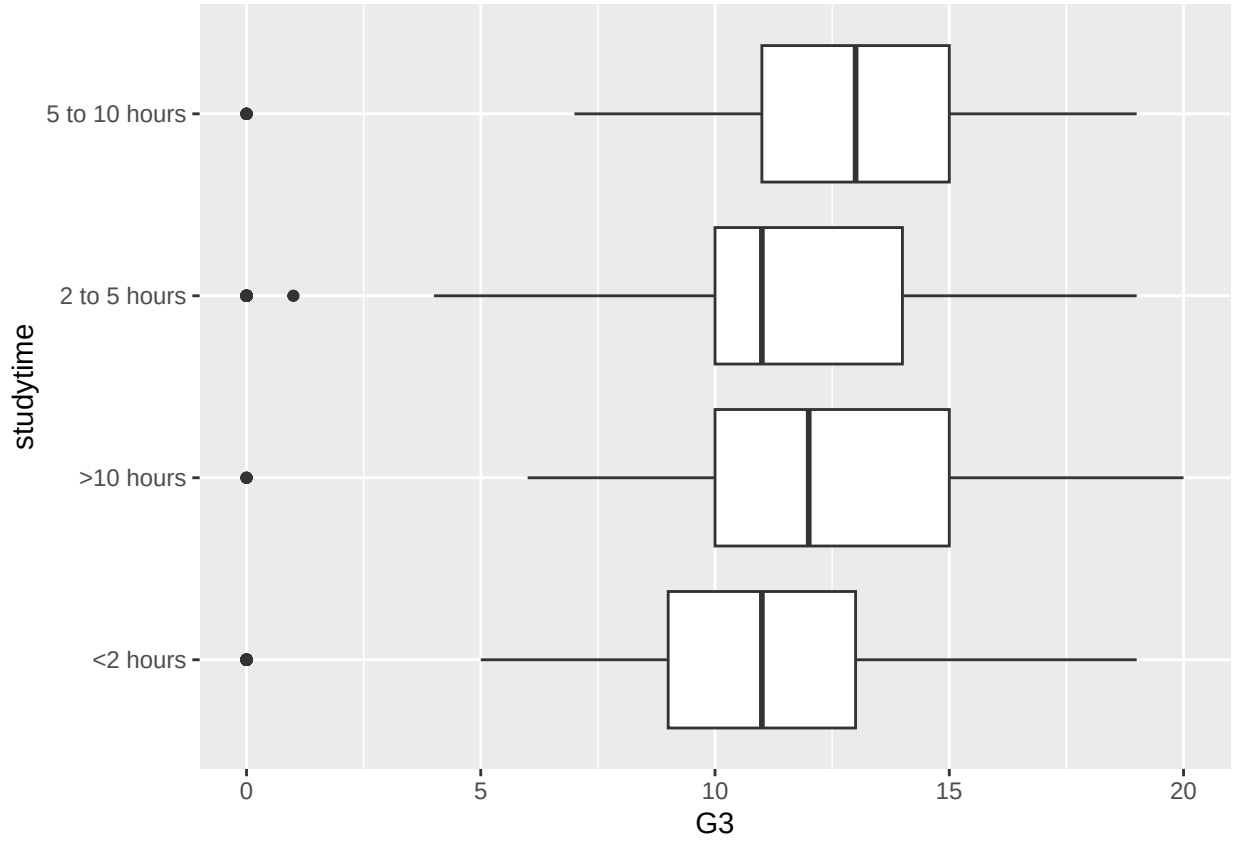
```r
#on calcul la porportion pour pouvoir comparer
data_stud_mat <- data_stud_mat %>%
  mutate(proportion = n_obs_mat / sum(n_obs_mat))

data_stud_por <- data_stud_por %>%
  mutate(proportion = n_obs_por / sum(n_obs_por))

#on renome de la même manière les colonnes du nombre d'étudiants pour chaque obersvation
data_stud_mat$filiere=c(rep("math",nrow(data_stud_mat)))
data_stud_por$filiere=c(rep("portugais",nrow(data_stud_por)))
colnames(data_stud_mat)[colnames(data_stud_mat) == "n_obs_mat"] <- "n_obs"
colnames(data_stud_por)[colnames(data_stud_por) == "n_obs_por"] <- "n_obs"

#on concatène les deux datas frame
data_stud=rbind(data_stud_mat,data_stud_por)

#Création du graphique

ggplot(data_stud, aes(x = studytime, y = proportion, fill = filiere)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Comparaison du temps de travail entre deux filières", x = "Temps de travail par semaine
  scale_fill_manual(values = c("red", "blue"))
```



On s'aperçoit donc que les élèves dans la filière mathématiques travaillent plus

```r
data_quanti$studytime[data_quanti$studytime == 1] <- "<2 hours"
data_quanti$studytime[data_quanti$studytime == 2] <- "2 to 5 hours"
```

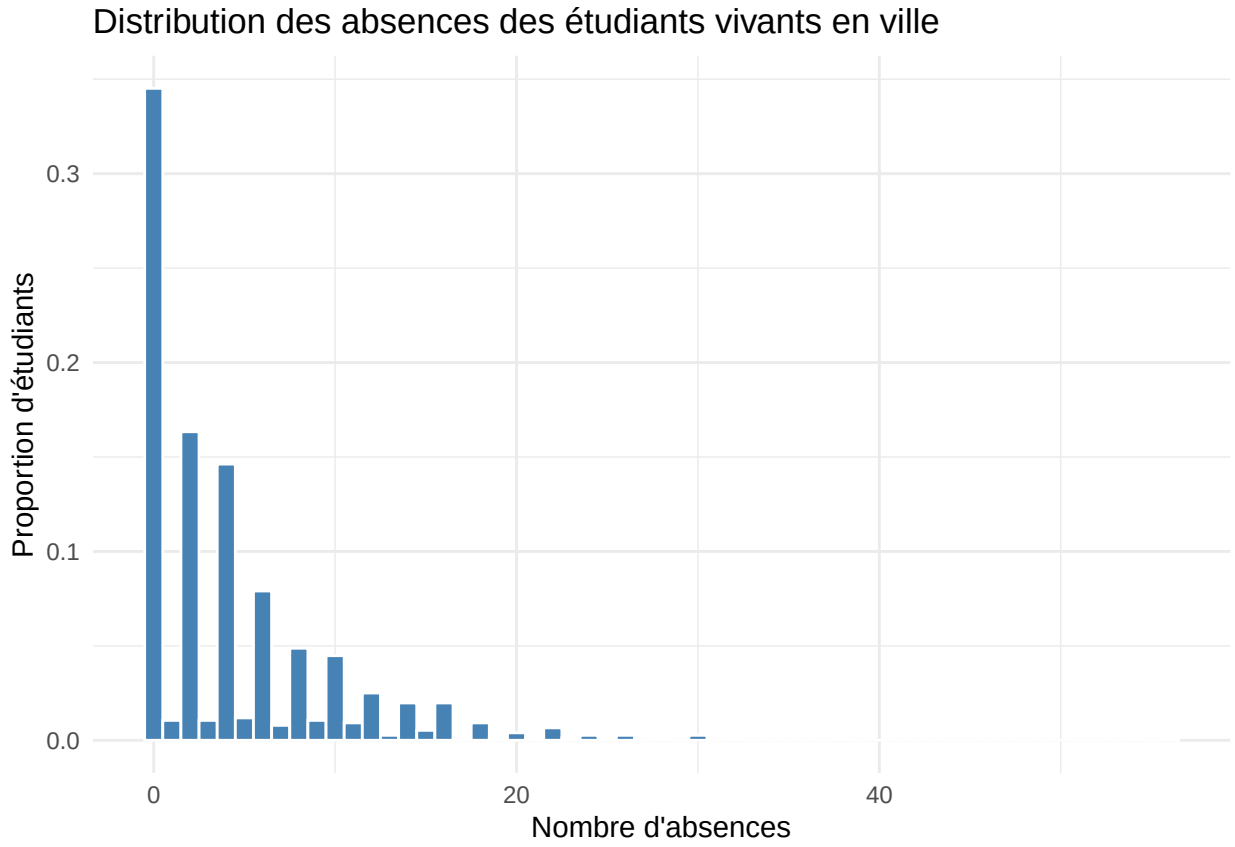
```
data_quanti$studytime[data_quanti$studytime == 3] <- "5 to 10 hours"
data_quanti$studytime[data_quanti$studytime == 4] <- ">10 hours"

ggplot(data_quanti, aes(x = G3, y = studytime)) +
  geom_boxplot()
```



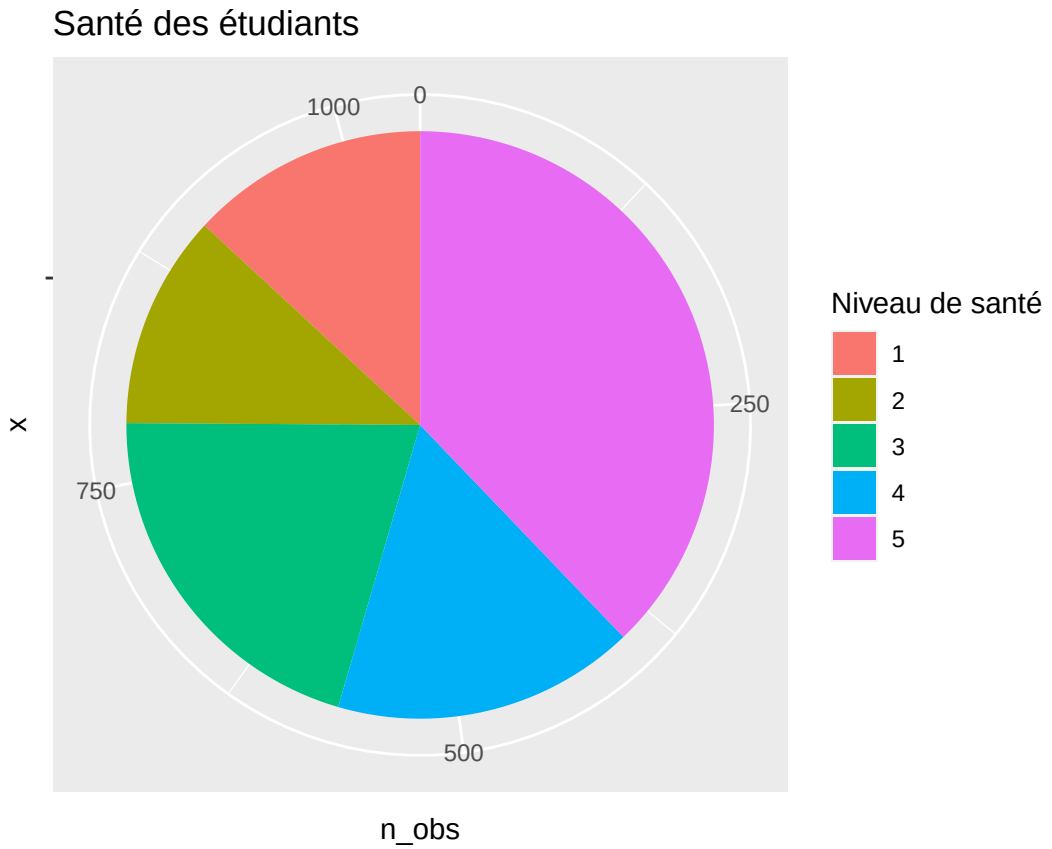
On voit que globalement, les élèves qui travaillent plus on de meilleures notes (comportement bizarre à vérifier)

```
#oui rudio j'ai fait un truc avec de qualis

ggplot(df[df$address == 'U',], aes(x=absences)) +
  geom_histogram(aes(y = ..count.. / sum(..count..)), binwidth=1, fill="steelblue", color="white") +
  labs(title="Distribution des absences des étudiants vivants en ville",
       x="Nombre d'absences", y="Proportion d'étudiants") +
  theme_minimal()
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```

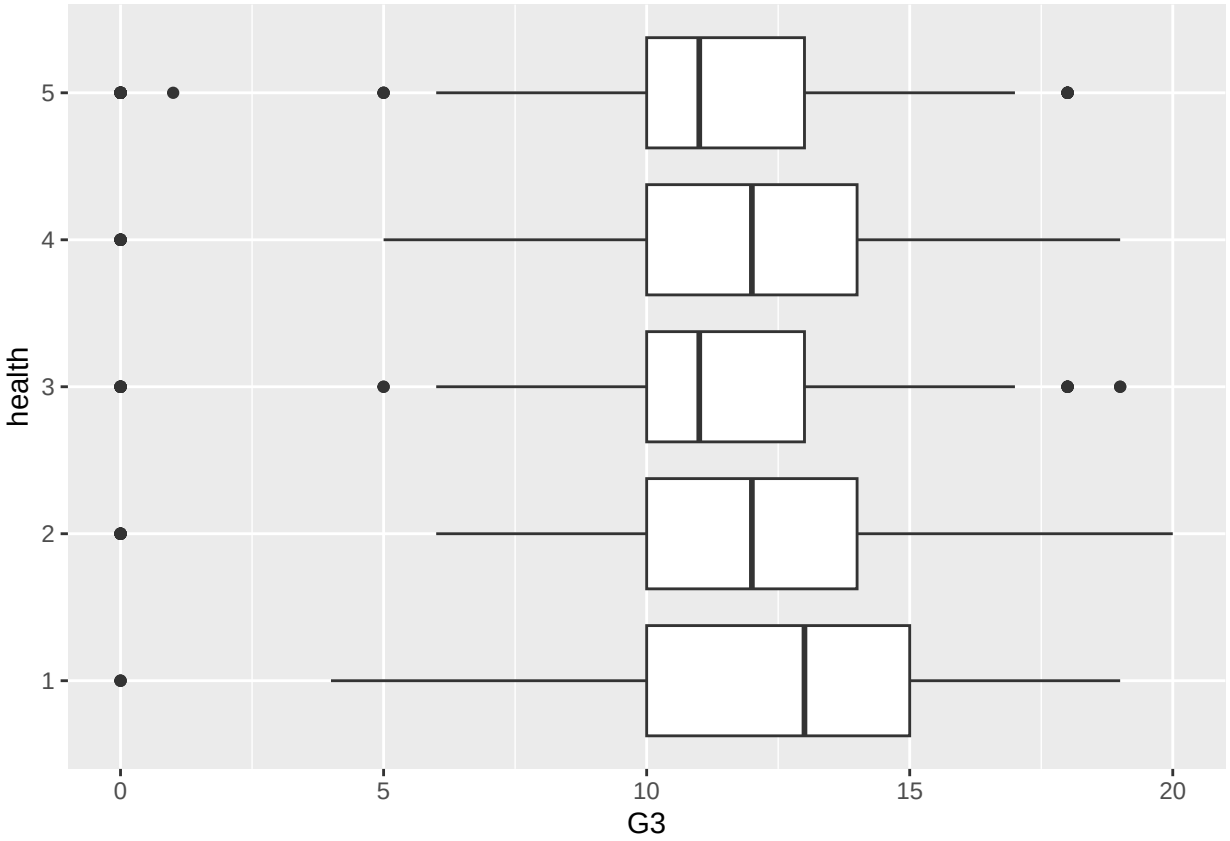



```
data_health=data_quanti
data_health=summarise(group_by(data_health,health),n_obs=n())


ggplot(data_health, aes(x = "", y = n_obs, fill = factor(health))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Santé des étudiants") +
  scale_fill_discrete(name = "Niveau de santé", labels = c(1,2,3,4,5))
```



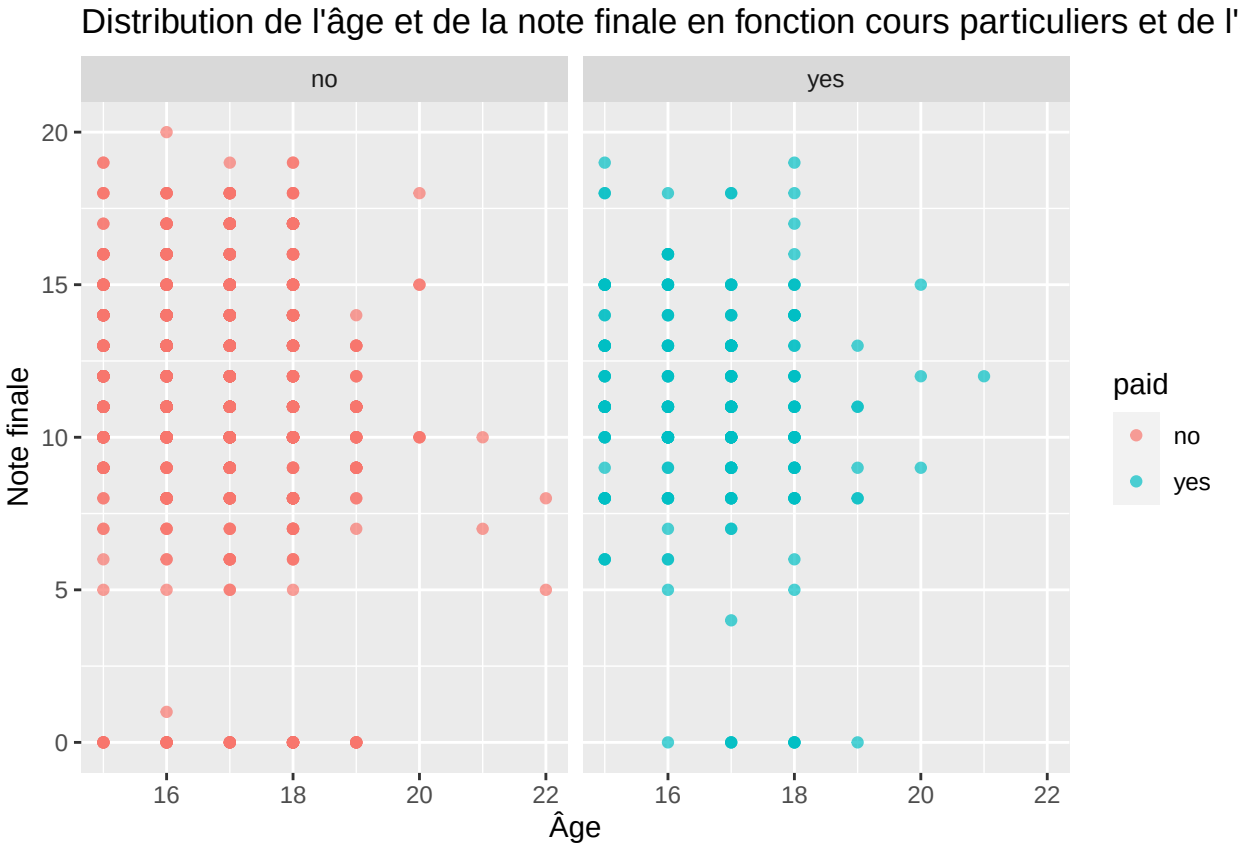
On voit que la plupart des étudiant sont en bonne santé

```
data_quanti$health=factor(data_quanti$health)
ggplot(data_quanti, aes(x = G3, y = health)) +
  geom_boxplot()
```



On voit que les étudiants en meilleure santé ont une meilleure réussite

```
ggplot(df, aes(x = age, y = G3, color = paid)) +
  geom_point(alpha = 0.7) +
  facet_wrap(~paid) +
  labs(title = "Distribution de l'âge et de la note finale en fonction cours particuliers et de l'âge",
       x = "Âge", y = "Note finale")
```



Curieusement, les résultats semblent meilleur pour ceux qui n'on pas pris de cours