Projet outils et logiciels d'analyse des données

Tekgoz Sumeyye

2022-12-11

Chemin importation des bases de données

```
path <- file.path("C:", "Users", "tekgo", "Documents", "GitHub", "dataviz", fsep="\\")
setwd(path)
```

Version de R : R version 4.1.3 (2022-03-10)

Téléchargement des librairies

```
library(readr)
library(coefplot)
```

```
## Le chargement a nécessité le package : ggplot2
```

```
library(ggplot2)
library(labelled)
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
```

```
library(RColorBrewer)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(scales)
```

```
##
## Attachement du package : 'scales'

## L'objet suivant est masqué depuis 'package:readr':
##
##     col_factor
```

Importation de la base de données

```
maths <- read_csv("Maths.csv")
Portuguese <- read_csv("Portuguese.csv")
```

I- Introduction

Ces données portent sur les résultats des élèves dans l'enseignement secondaire de deux écoles portugaises. Les attributs des données comprennent les notes des élèves, les caractéristiques démographiques, sociales et scolaires, et ont été collectés a l'aide de rapports et de questionnaires scolaires. Deux ensembles de données sont fournis concernant les performances dans deux matières distinctes : Les mathématiques (mat) et la langue portugaise (por).

A partir de l'analyse du lien qui pourrait exister entre la consommation d'alcool et les résultats scolaires, l'idée est de déterminer plus généralement quels pourraient être les facteurs affectant la réussite scolaire dans le contexte des jeux de données dont nous disposons.

```
N1 <- nrow(maths)
N2 <- nrow(Portuguese)
N1
```

```
## [1] 395
```

```
N2
```

```
## [1] 649
```

Pour la base de donnée maths, il y a 395 observations et 33 variables. Pour la base de donnée portuguese, il y a 649 observations et 33 variables identiques avec la première base de donnée.

Nous avons décidé de rassembler les deux bases de données Avec str, on visualise les premières données des variables, et on regarde si ce sont des variables qualitative ou quantitatives

```
fulldt<-bind_rows(maths, Portuguese)
nrow(fulldt)
```

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

```
str(fulldt)
```

```
## spc_tbl_ [1,044 x 33] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ school    : chr [1:1044] "GP" "GP" "GP" "GP" ...
##  $ sex       : chr [1:1044] "F" "F" "F" "F" ...
```

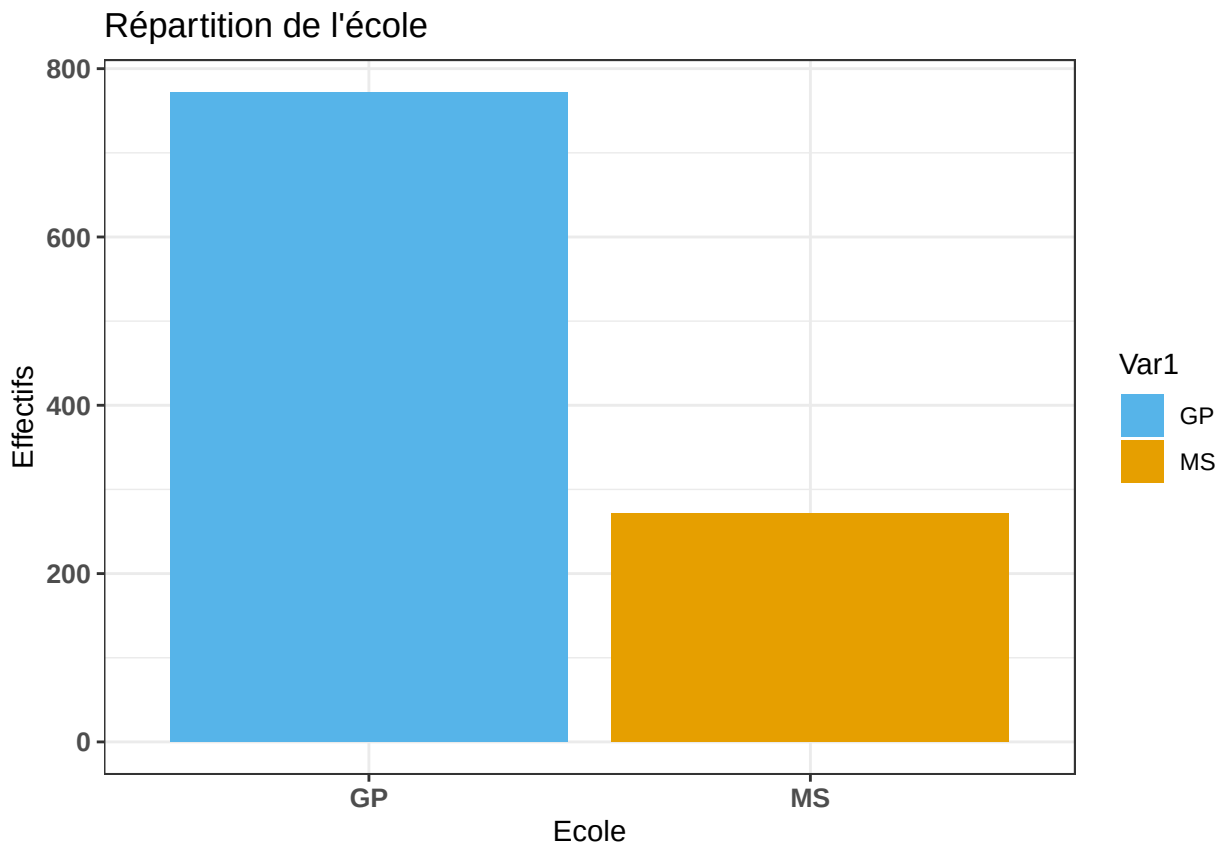
```
##  $ age       : num [1:1044] 18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr [1:1044] "U" "U" "U" "U" ...
##  $ famsize   : chr [1:1044] "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr [1:1044] "A" "T" "T" "T" ...
##  $ Medu      : num [1:1044] 4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : num [1:1044] 4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr [1:1044] "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr [1:1044] "teacher" "other" "other" "services" ...
##  $ reason    : chr [1:1044] "course" "course" "other" "home" ...
##  $ guardian  : chr [1:1044] "mother" "father" "mother" "mother" ...
##  $ traveltime: num [1:1044] 2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : num [1:1044] 2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : num [1:1044] 0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr [1:1044] "yes" "no" "yes" "no" ...
##  $ famsup    : chr [1:1044] "no" "yes" "no" "yes" ...
##  $ paid      : chr [1:1044] "no" "no" "yes" "yes" ...
##  $ activities: chr [1:1044] "no" "no" "no" "yes" ...
##  $ nursery   : chr [1:1044] "yes" "no" "yes" "yes" ...
##  $ higher    : chr [1:1044] "yes" "yes" "yes" "yes" ...
##  $ internet  : chr [1:1044] "no" "yes" "yes" "yes" ...
##  $ romantic  : chr [1:1044] "no" "no" "no" "yes" ...
##  $ famrel    : num [1:1044] 4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : num [1:1044] 3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : num [1:1044] 4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : num [1:1044] 1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : num [1:1044] 1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : num [1:1044] 3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : num [1:1044] 6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : num [1:1044] 5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : num [1:1044] 6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : num [1:1044] 6 6 10 15 10 15 11 6 19 15 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   school = col_character(),
##   ..   sex = col_character(),
##   ..   age = col_double(),
##   ..   address = col_character(),
##   ..   famsize = col_character(),
##   ..   Pstatus = col_character(),
##   ..   Medu = col_double(),
##   ..   Fedu = col_double(),
##   ..   Mjob = col_character(),
##   ..   Fjob = col_character(),
##   ..   reason = col_character(),
##   ..   guardian = col_character(),
##   ..   traveltime = col_double(),
##   ..   studytime = col_double(),
##   ..   failures = col_double(),
##   ..   schoolsup = col_character(),
##   ..   famsup = col_character(),
##   ..   paid = col_character(),
##   ..   activities = col_character(),
##   ..   nursery = col_character(),
##   ..   higher = col_character(),
```

```
##   ..   internet = col_character(),
##   ..   romantic = col_character(),
##   ..   famrel = col_double(),
##   ..   freetime = col_double(),
##   ..   goout = col_double(),
##   ..   Dalc = col_double(),
##   ..   Walc = col_double(),
##   ..   health = col_double(),
##   ..   absences = col_double(),
##   ..   G1 = col_double(),
##   ..   G2 = col_double(),
##   ..   G3 = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

II- Visualisation des données

a) Ecole

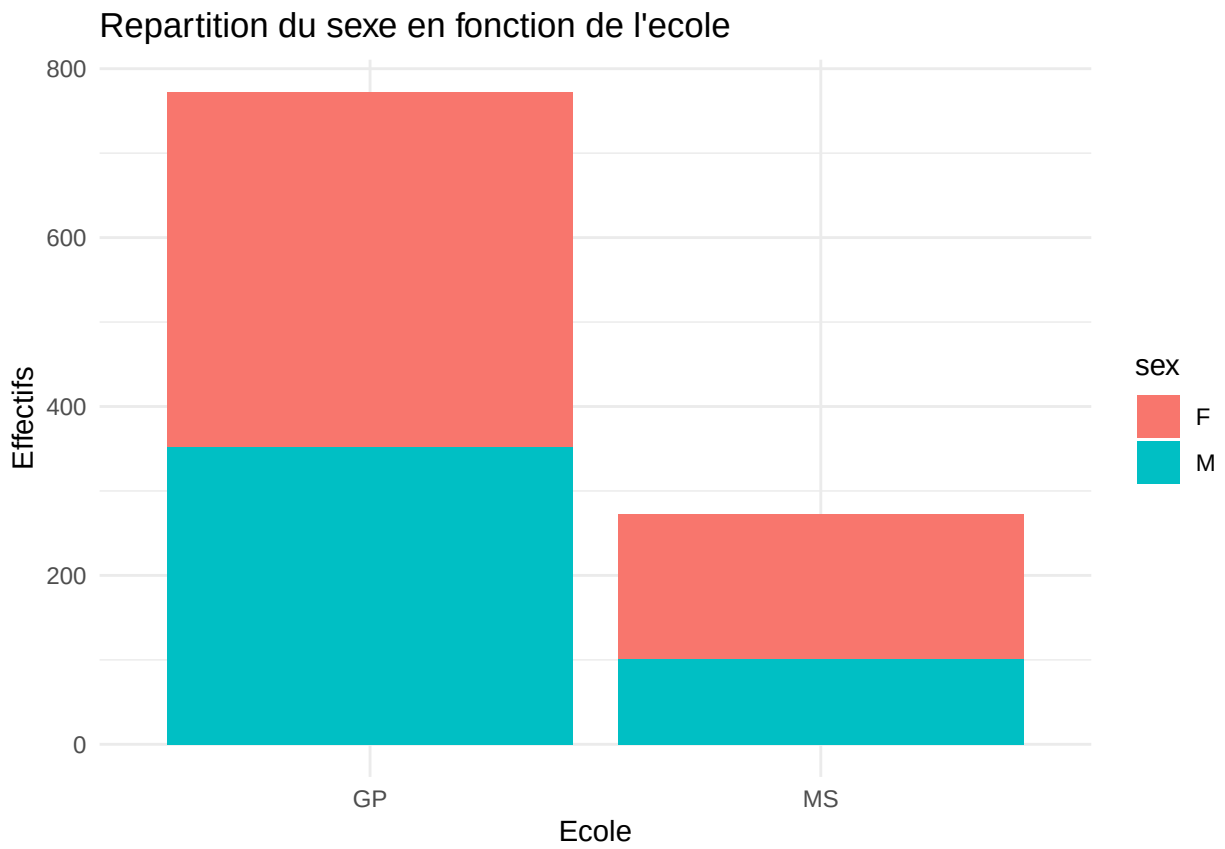
```
ggplot(as.data.frame(table(fulldt$school))) +
     geom_bar(aes(x = Var1, y = Freq, fill = Var1),
                             stat = 'identity') +
     scale_fill_manual(values=c("#56B4E9", "#E69F00")) +
     ggtitle("Répartition de l'école") +
     xlab("Ecole") +
     ylab("Effectifs") +
     theme_bw() +
     theme(axis.text.x = element_text(face = 'bold', size = 10),
                      axis.text.y = element_text(face = 'bold', size = 10))
```



Pour la base fulldt, il y a 772 élèves scolarises à Gabriel Pereira et 272 à Mousinho da Silveira.

b) Le sexe en fonction de l'école

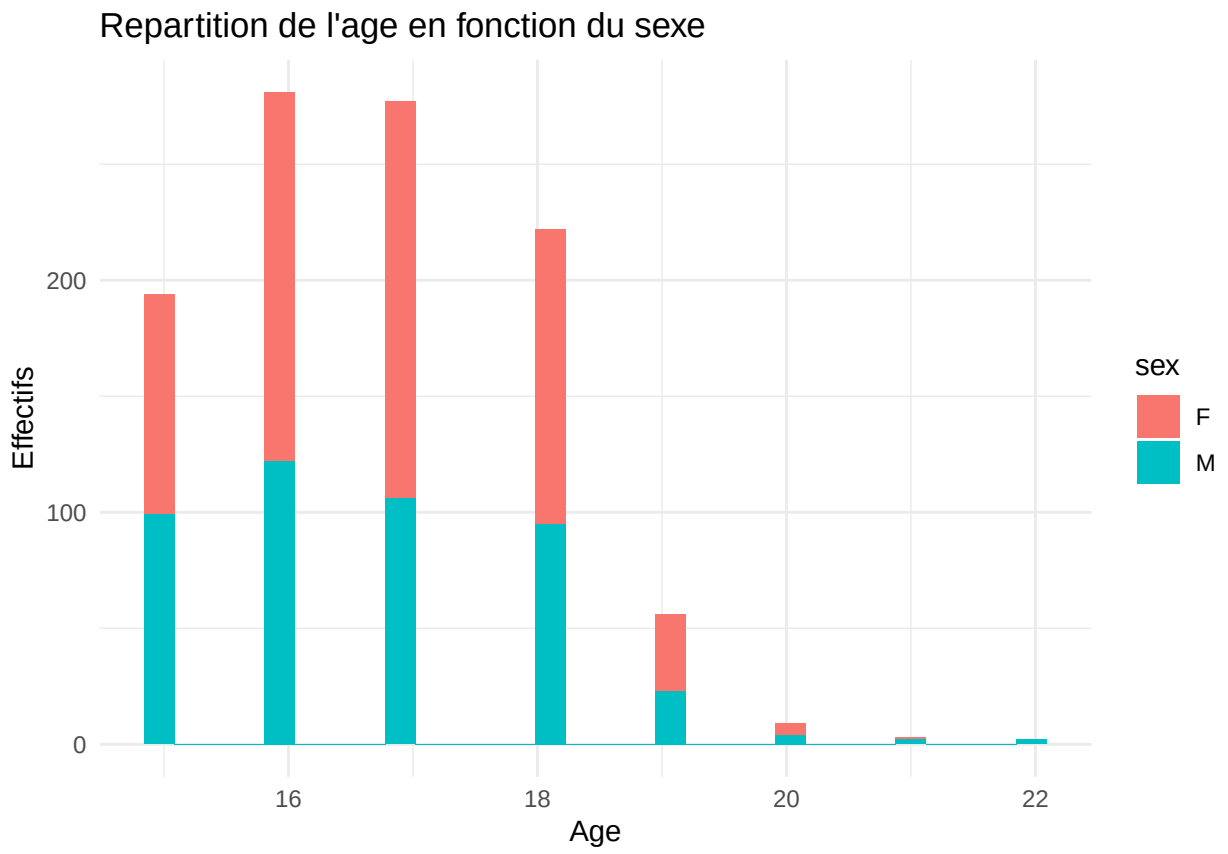
```
ggplot(fulldt) +
  aes(x = school, fill = sex) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Ecole",
    y = "Effectifs",
    title = "Repartition du sexe en fonction de l'ecole"
  ) +
  theme_minimal()
```



Il y a 591 filles et 453 garçons au sein des 2 écoles. La proportion de fille est plus importante au sein de l'école Mousinho da Silveira.

c) Age en fonction du sexe

```
ggplot(fulldt) +
  aes(x = age, fill = sex) +
  geom_histogram(bins = 30L) +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Age",
    y = "Effectifs",
    title = "Repartition de l'age en fonction du sexe"
  ) +
  theme_minimal()
```




```
table(fulldt$age)
```

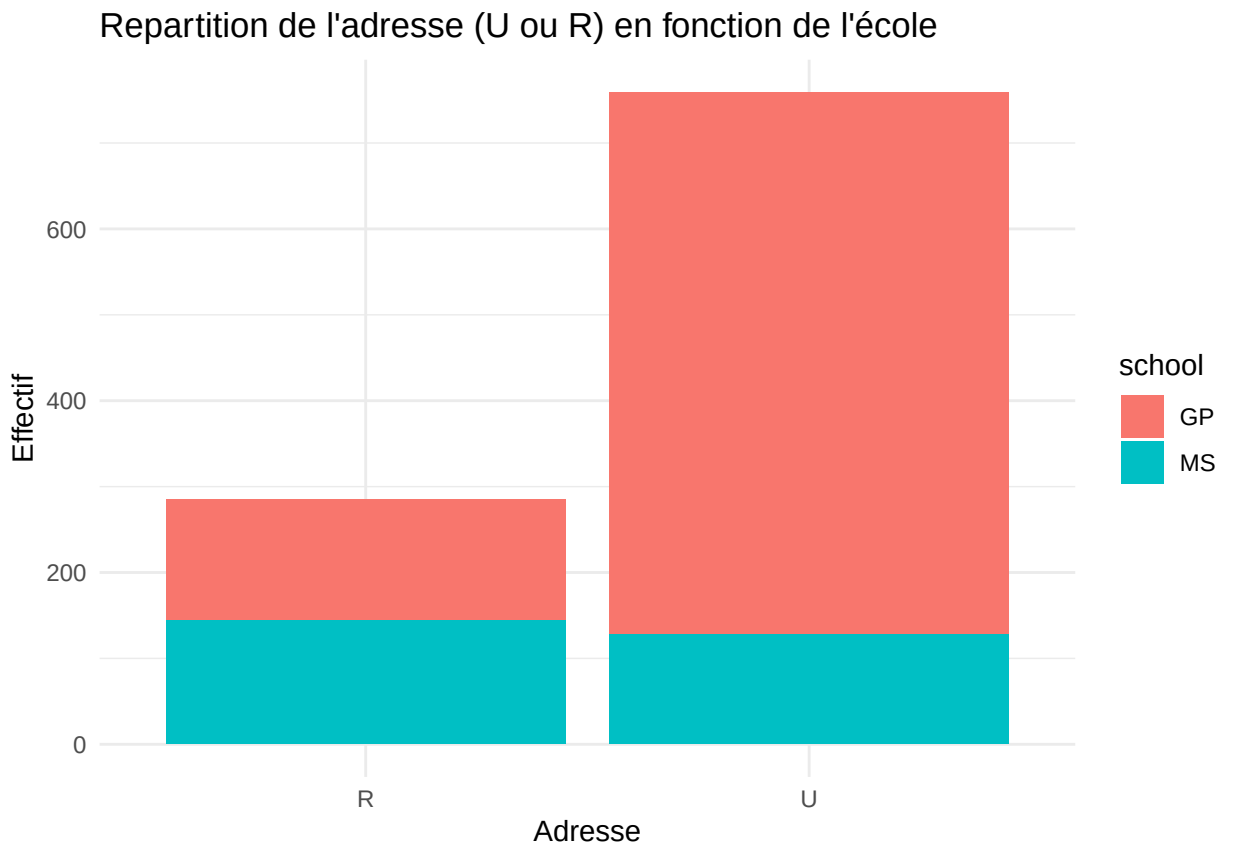
```
##
##  15  16  17  18  19  20  21  22
## 194 281 277 222  56   9   3   2
```

Il y a très peu d'élèves au delà de 19 ans, les élèves au delà de cette âge sont sûrement dû à des redoublements au cours de leurs scolarités.

d) Adresse

Adresse en fonction de l'école

```
ggplot(fulldt) +
  aes(x = address, fill = school) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Adresse",
    y = "Effectif",
    title = "Repartition de l'adresse (U ou R) en fonction de l'école"
  ) +
  theme_minimal()
```



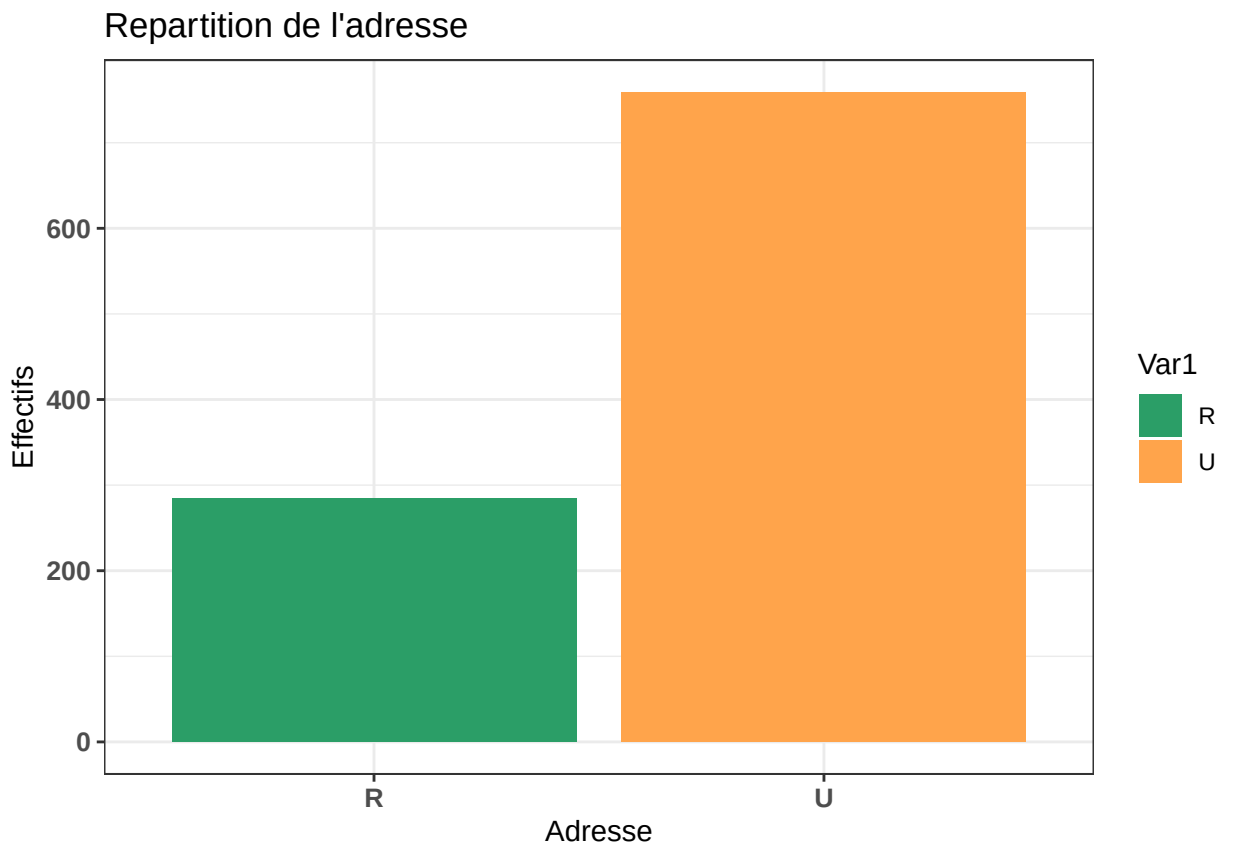
Pour l'école MS, il y a une répartition approximative des élèves vivant dans une zone urbaine ou rurale alors que au contraire pour l'école GP, une grande partie vivent dans une zone urbaine.

Seulement adresse

```
as.data.frame(table(fulldt$address))
```

```
##   Var1 Freq
## 1    R  285
## 2    U  759
```

```
ggplot(as.data.frame(table(fulldt$address))) +
  geom_bar(aes(x = Var1, y = Freq,  fill = Var1),
           stat = 'identity') +
  scale_fill_manual(values=c( '#2B9E67', '#FFA44B')) +
  ggtitle("Repartition de l'adresse") +
  xlab("Adresse") +
  ylab("Effectifs") +
  theme_bw() +
  theme(axis.text.x = element_text(face = 'bold', size = 10),
        axis.text.y = element_text(face = 'bold', size = 10))
```




```
table(fulldt$address)
```

```
##
##   R   U
## 285 759
```

Il y a 285 élèves vivant dans une zone rurale et 759 dans une zone urbaine.

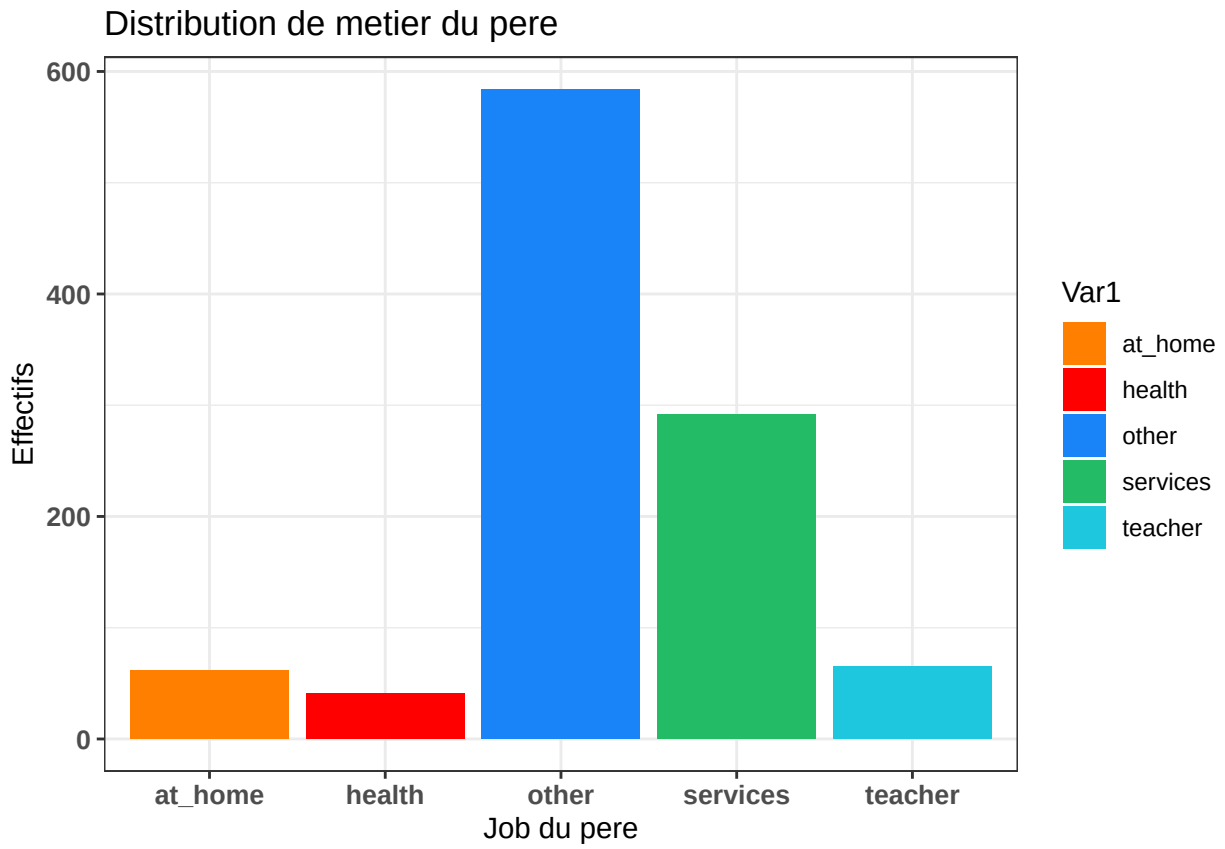
e) Emploi du père

```
as.data.frame(table(fulldt$Fjob))
```

```
##       Var1 Freq
## 1  at_home   62
## 2   health   41
## 3    other  584
## 4 services  292
## 5  teacher   65
```

```
ggplot(as.data.frame(table(fulldt$Fjob))) +
  geom_bar(aes(x = Var1, y = Freq, fill = Var1),
           stat = 'identity') +
```

```
  scale_fill_manual(values=c("#ff8000", "#FF0000", "#1884F7", "#23BB66", "#1EC7DE")) +
  ggtitle("Distribution de metier du pere") +
  xlab("Job du pere") +
  ylab("Effectifs") +
  theme_bw() +
  theme(axis.text.x = element_text(face = 'bold', size = 10),
        axis.text.y = element_text(face = 'bold', size = 10))
```



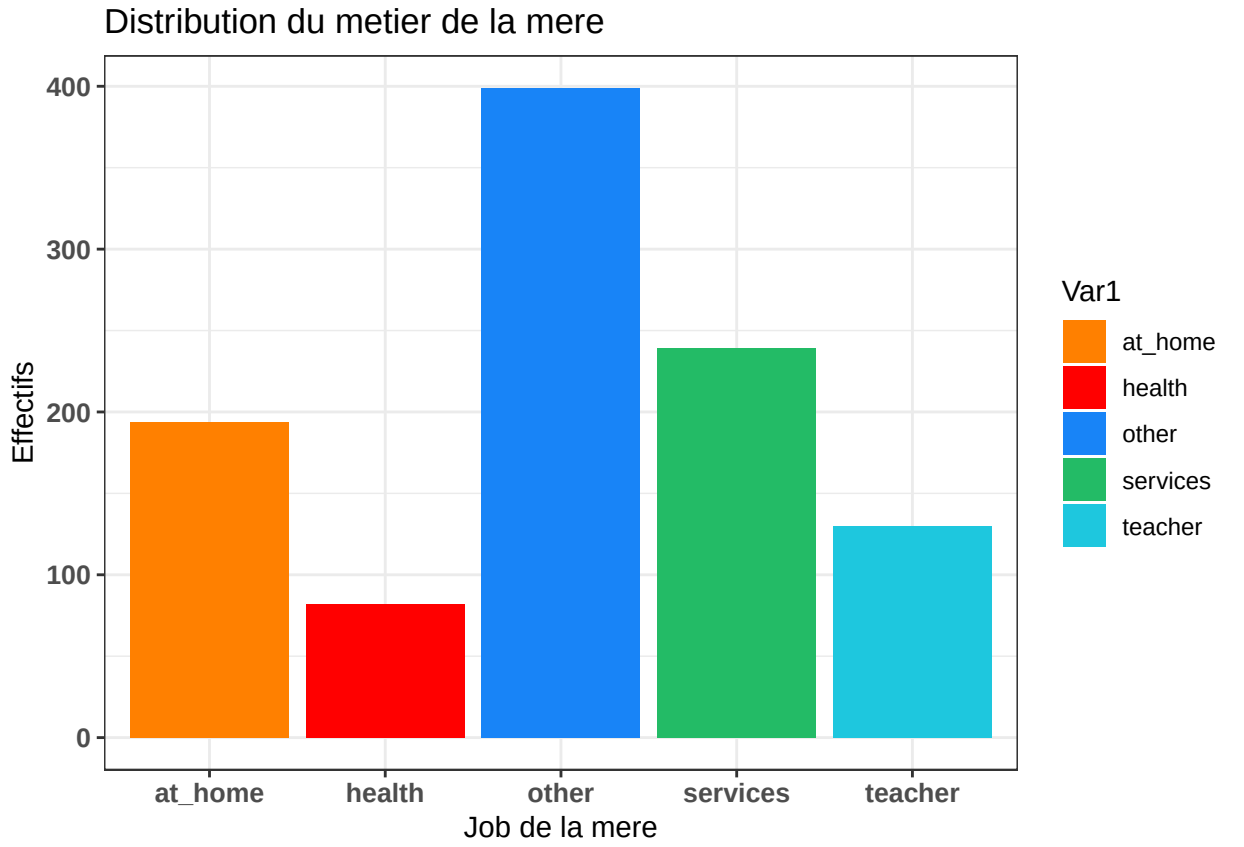
Nous observons que la catégorie “les autres” se démarque des 3 autres catégories (à la maison, santé et professeurs)

f) Emploi de la mère

```
as.data.frame(table(fulldt$Mjob))
```

```
##       Var1 Freq
## 1  at_home  194
## 2   health   82
## 3    other  399
## 4 services  239
## 5  teacher  130
```

```
ggplot(as.data.frame(table(fulldt$Mjob))) +
  geom_bar(aes(x = Var1, y = Freq, fill = Var1),
           stat = 'identity') +
  scale_fill_manual(values=c("#ff8000", "#FF0000", "#1884F7", "#23BB66", "#1EC7DE")) +
  ggtitle("Distribution du metier de la mere") +
  xlab("Job de la mere") +
  ylab("Effectifs") +
  theme_bw() +
  theme(axis.text.x = element_text(face = 'bold', size = 10),
        axis.text.y = element_text(face = 'bold', size = 10))
```

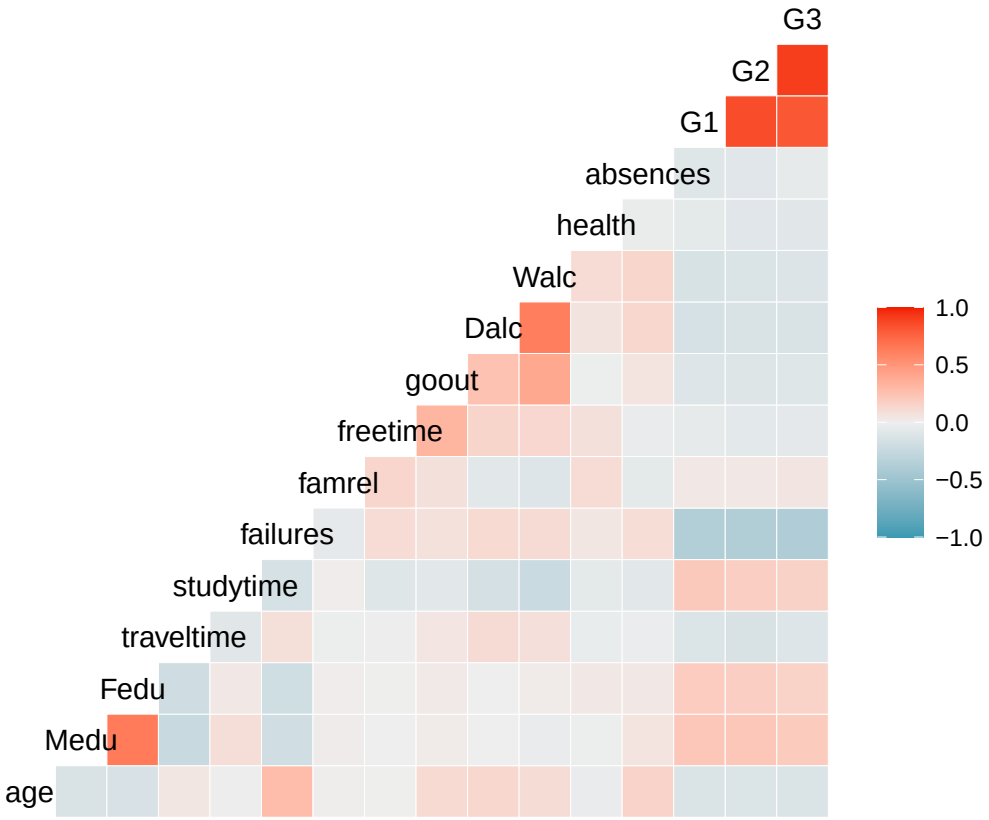


Ici aussi, la catégorie “autres” se démarque des autres, suivi de services. Contrairement aux mères, les pères travaillent beaucoup plus dans des métiers de catégories “autres” ou services, alors qu’il y a une part importante de mère travaillant dans éducation, santé ou des mères aux foyers.

h) Maitrice de corrélation entre variables quantitatives

```
ggcorr(fulldt)
```

```
## Warning in ggcorr(fulldt): data in column(s) 'school', 'sex', 'address',
## 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', 'guardian', 'schoolsup',
## 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic' are
## not numeric and were ignored
```



Grâce à cette matrice de corrélation, nous pouvons apercevoir les variables quantitatives corrélées, par exemple, il y a une forte corrélation entre les notes G1, G2 et G3. La consommation d'alcool en semaine et le weekend a une corrélation positive d'environ 0.5. Au contraire, il n'y a pas de corrélation entre les variables age et le temps libres par exemples. Il y a aussi une corrélation négative entre la variables échecs scolaires et les notes G1, G2 et G3, qui montre que l'échec scolaires impactes négativement les notes.

III. Visualision des données afin d'établir une eventuelle corrélation entre la consommation d'alcool et les résultats scolaires

Profil général des consommateurs d'alcool, afin d'établir une premiè re typologie general :

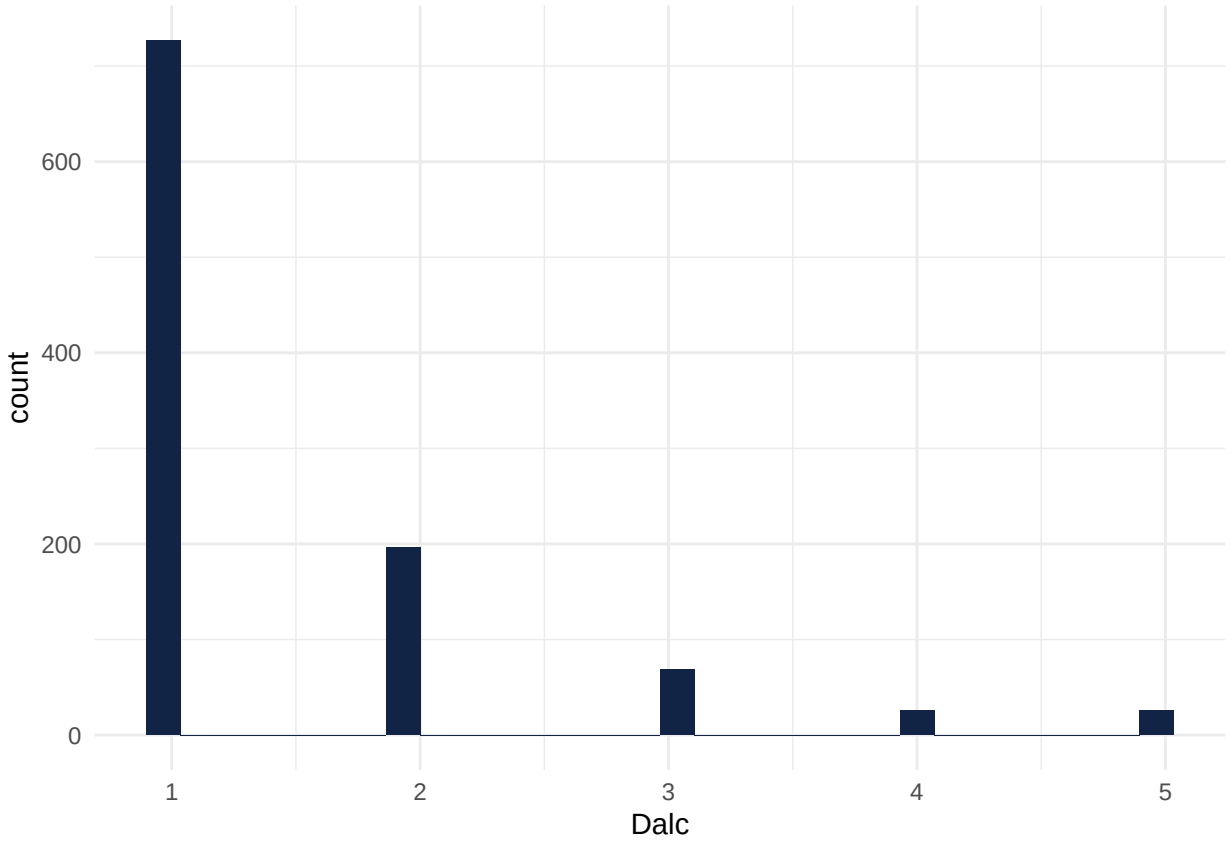
1. Visualisation des consommations d'alcool weekend et semaine

```
table(fulldt$Dalc)
```

Consommation d'alcool en semaine

```
##
##   1   2   3   4   5
## 727 196  69  26  26
```

```
ggplot(fulldt) +
  aes(x = Dalc) +
  geom_histogram(bins = 30L, fill = "#112446") +
  theme_minimal()
```



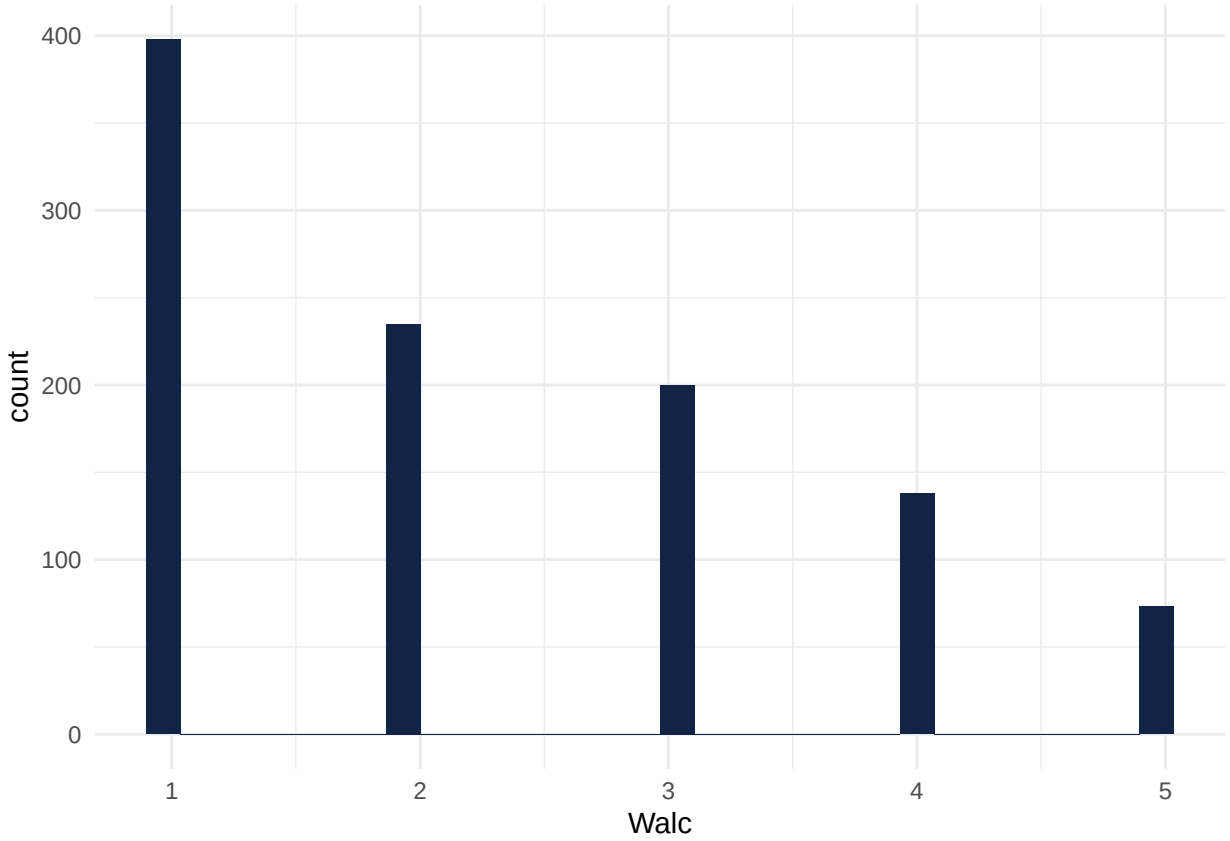
En semaine, la consommation d'alcool est plutôt modérer (très faible consommation).

```
table(fulldt$Walc)
```

Consommation d'alcool le weekend

```
##
##   1   2   3   4   5
## 398 235 200 138  73
```

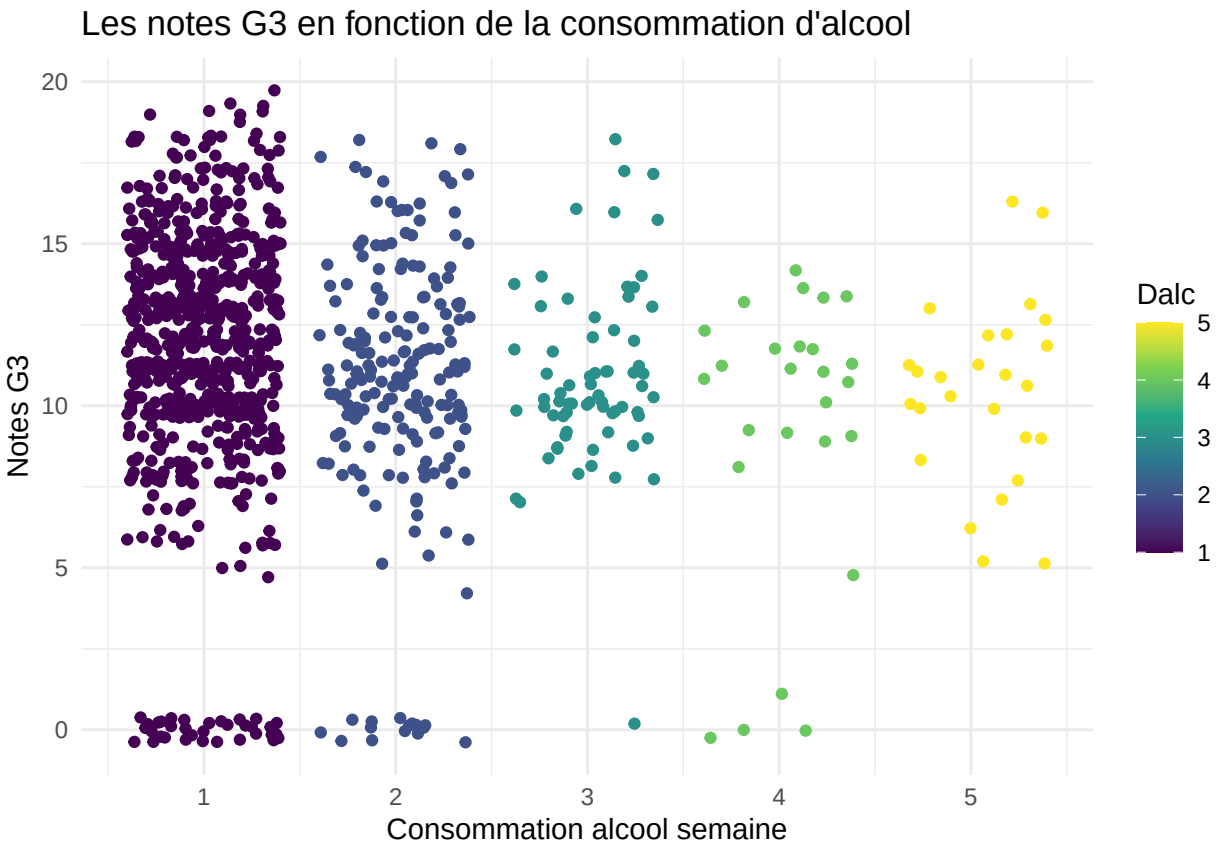
```
ggplot(fulldt) +
  aes(x = Walc) +
  geom_histogram(bins = 30L, fill = "#112446") +
  theme_minimal()
```



Il y a beaucoup plus de consommation le weekend, étant donné qu'il n'y a pas cours, plus de personnes consomment excessivement de l'alcool.

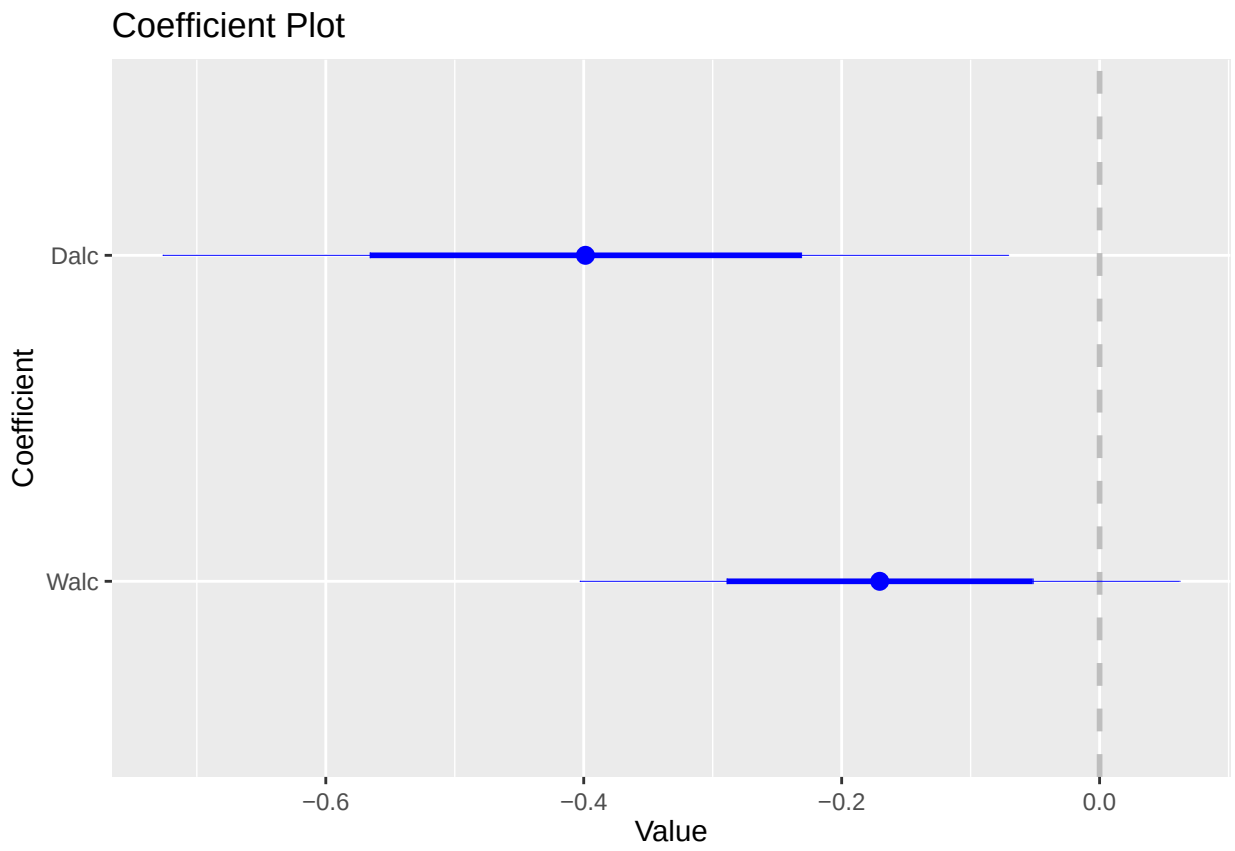
```
ggplot(fulldt) +
  aes(x = Dalc, y = G3, colour = Dalc) +
  geom_jitter(size = 1.5) +
  scale_color_viridis_c(option = "viridis", direction = 1) +
  labs(
    x = "Consommation alcool semaine",
    y = "Notes G3",
    title = "Les notes G3 en fonction de la consommation d'alcool"
  ) +
  theme_minimal()
```

Impact de l'alcool sur les notes G3



Grace au nuage de point, on remarque que les élèves qui ont obtenue de mauvaises notes (0) sont généralement des élèves consommant très peu d'alcool, on ne retrouve pas d'élèves consommant beaucoup d'alcool dans les mauvaises notes. De plus, très peu d'élèves consomment excessivement de l'alcool en semaine, et ces élèves ont des notes obtenue entre 5 et 15. De plus, les consommateurs modérés, ont généralement des notes compris entre 7.5 et 17.5, quelques personnes se démarques en ayant obtenue de très bonne note, ce sont souvent des élèves consommant très peu d'alcool.

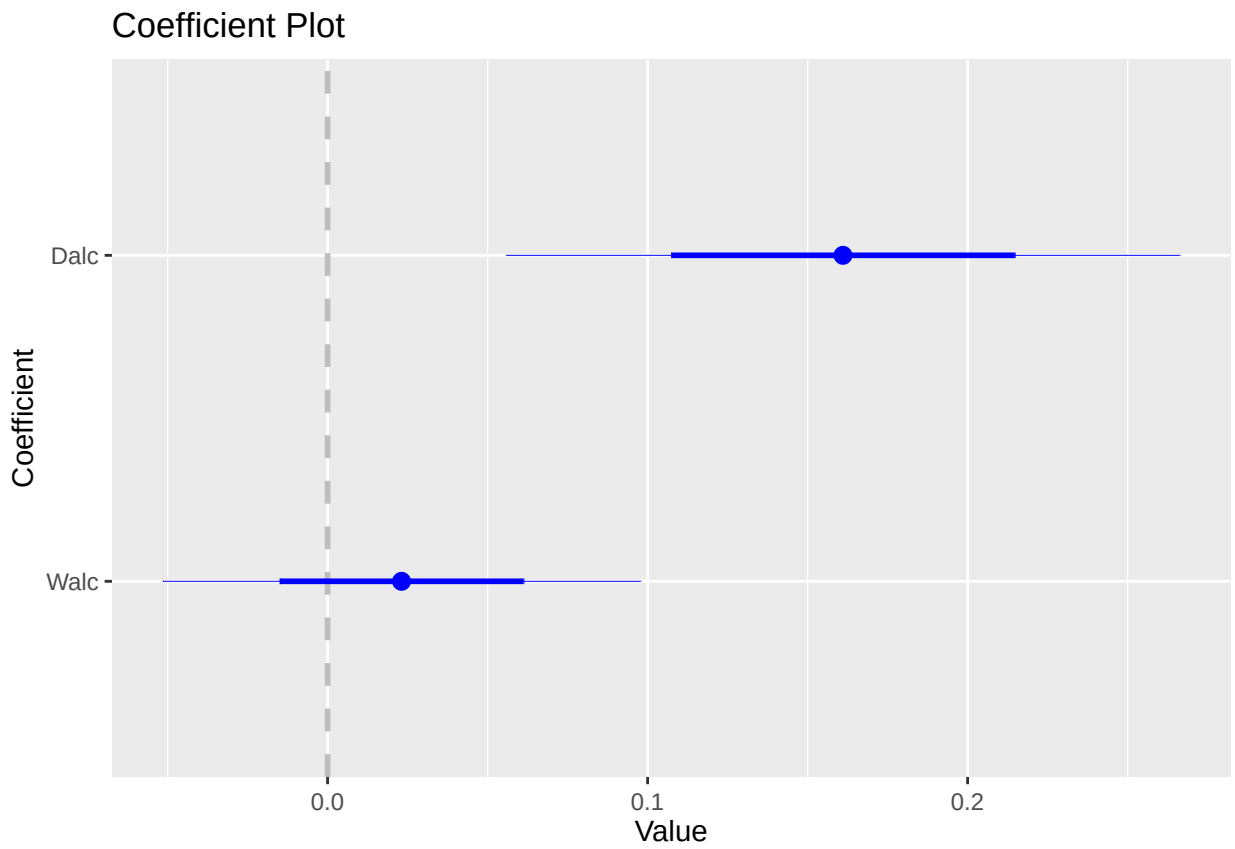
```
cor1 <- G3 ~ Walc + Dalc
lm1<-lm(cor1 , data = fulldt)
coefplot(lm1 , outerCI = 1.96 , intercept = FALSE)
```



La consommation d'alcool en semaine a plus d'impact négatif sur les résultats scolaires G3 que la consommation le weekend. Les deux coefficients est négative, mais l'intervalle de confiance de Walc couvre 0, donc le coefficient de Walc n'est pas significatif. Cela peut être dû au fait qu'en semaine, les personnes consommant de l'alcool ne se concentre pas aux révisions donc leurs notes sont impactés.

```
cor2 <- age ~ Walc + Dalc
lm2<-lm(cor2 , data = fulldt)
coefplot(lm2 , outerCI = 1.96 , intercept = FALSE)
```

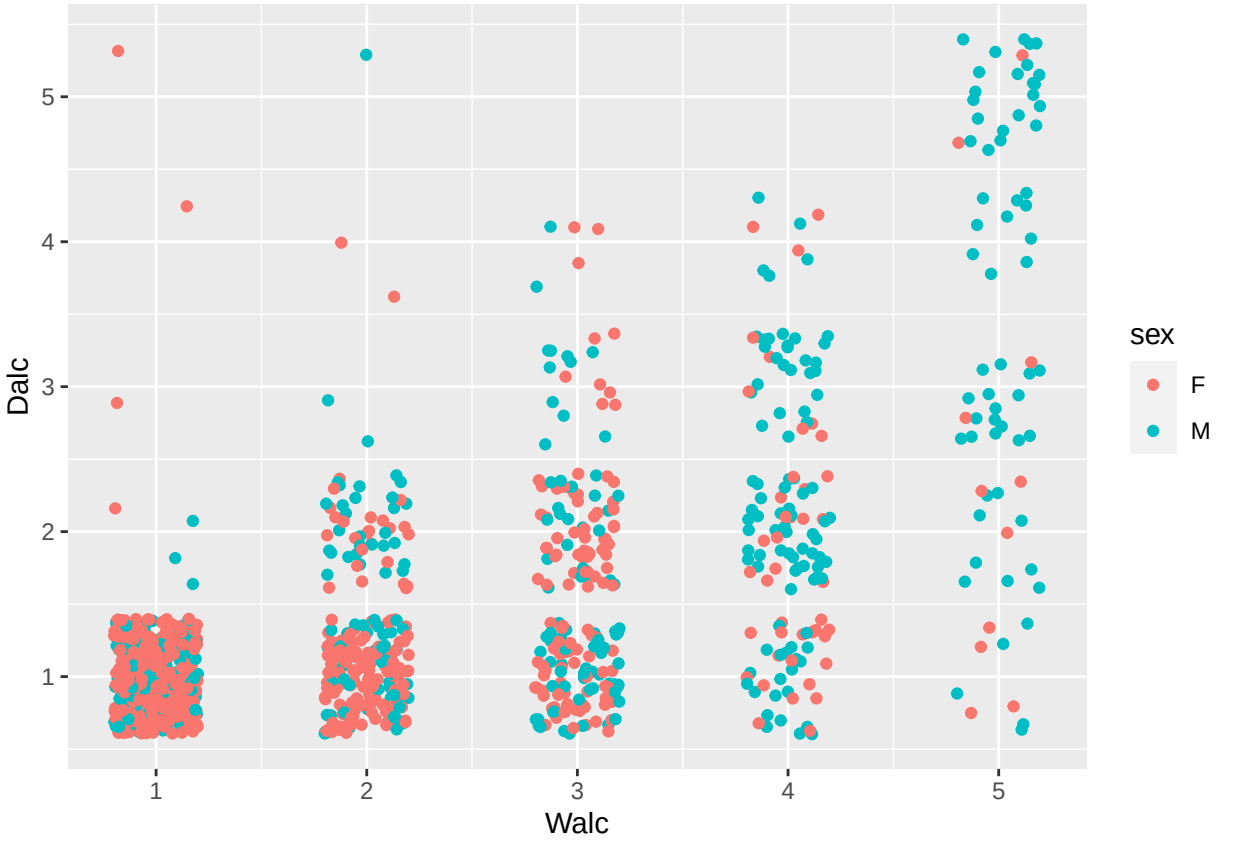
Impact de l'âge sur la consommation de l'alcool



Ici, nous cherchons à comprendre si la consommations d'alcool en semaine ou le weekend dépend de l'âge. Les deux coefficient sont positifs, mais étant donné que l'intervalle de confiance de Walc couvre 0, le coefficent n'est pas significative contrairement a Dalc. La consommation d'alcool en semaine dépend plus de l'âge que la consommation le weekend, c'est à dire que par exemple, en fonction de l'age, les personnes font plus attention à leurs consommations d'alcool en semaine que le weekend.

```
ggplot(fulldt, aes(x = Walc, y = Dalc , color = sex))+   geom_jitter(position=position_jitter(0.2))
```

Impact du sexe sur la consommation de l'alcool



Analyse de la consommation d'alcool durant le travail en semaine (Dalc) et la consommation d'alcool durant le week-end (Walc) selon le genre : Valeurs numérique: 1 est équivaut a très faible et 5 est équivaut à très élevé. A travers le nuage de points ce que l'on peut voir selon la consommation d'alcool pendant le travail en semaine (Dalc) et en week-end (Walc) c'est qu'il y a peu de consommation excessive d'alcool que ce soit pour les hommes ou pour les femmes, les points étant principalement concentrer entre le nombre 1 dalc et 3 walc. Les femmes à part quelques occurrences (certain point compris au-dela de 5), boivent moins que les hommes qui sont plus nombreux à boire en semaine (une importante répartition de points entre le 4 et le 5).

Pour la consommation d'alcool durant le week-end (Walc), il y a également peu de consommation (majorité étant compris dans le 1) que ce soit pour les hommes ou pour les femmes . Il y a plus de consommation de la part des hommes durant le week end les points bleu étant compris entre le 4 et le 5. Il y a une augmentation de la consommation d'alcool pour les hommes le week-end (points bleu compris entre le 4 et le 5).

Selon notre nuages de points les hommes consomme plus d'alcool le week-end (la majorité étant comprises dans le 4) que la semaine (la majorité étant comprises entre le 3 et 4).

Les femmes consommerai plus d'alcool durant le week-end (entre le 2 et le 3) qu'en semaine (majorité en 1).

```
table(fulldt$address, fulldt$Dalc)
```

Impact de l'adresse sur la consommation de l'alcool

```
##
##        1   2   3   4   5
```

```
##   R 182  59  30   7   7
##   U 545 137  39  19  19
```

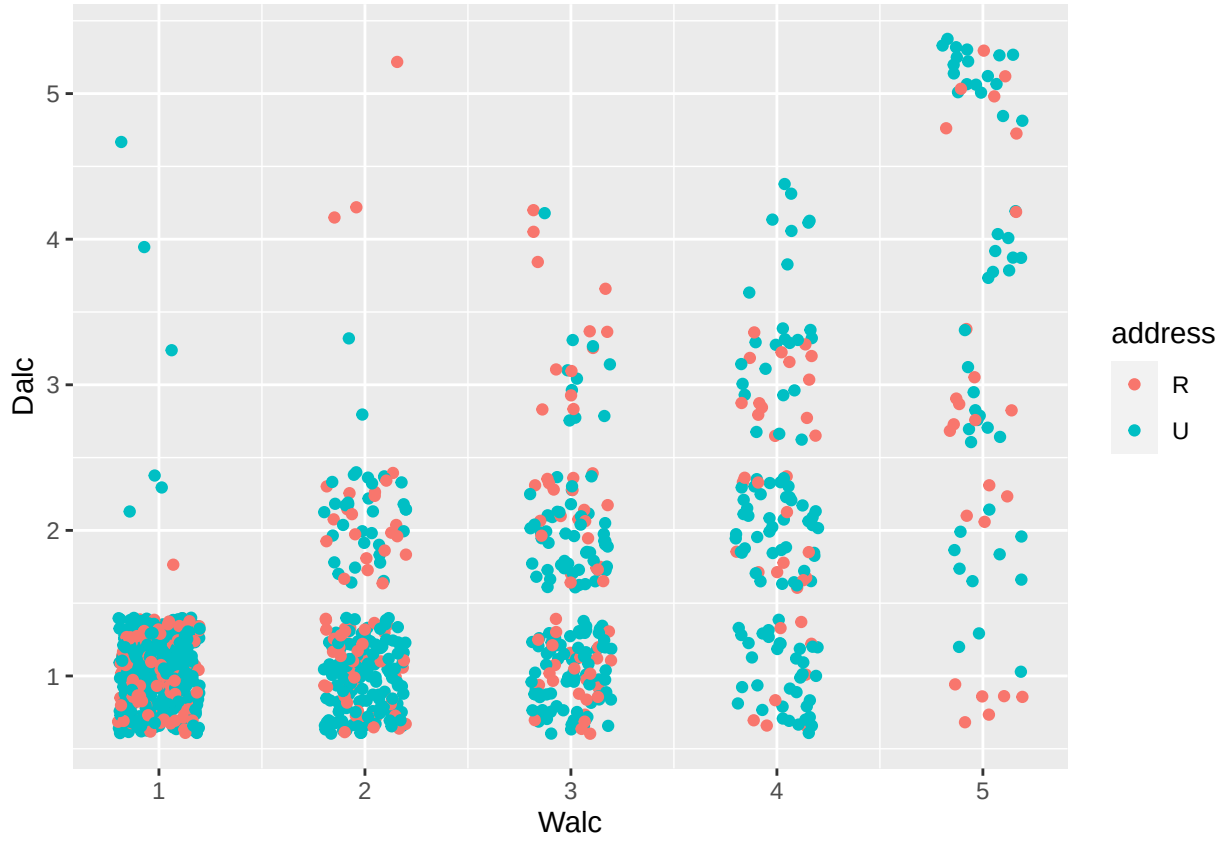
```
table(fulldt$address, fulldt$Walc)
```

```
##
##       1   2   3   4   5
##   R  97  69  59  35  25
##   U 301 166 141 103  48
```

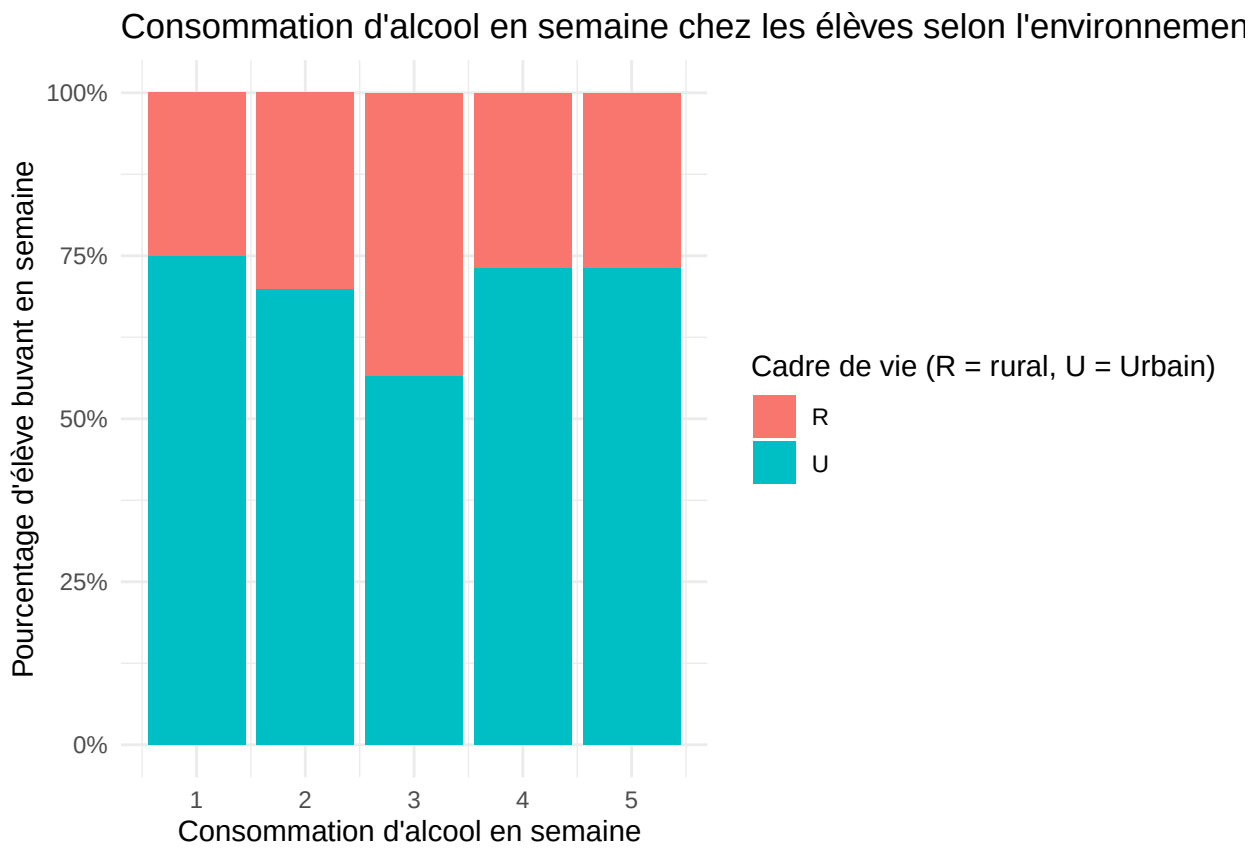
```
Dalc<-c(table(fulldt$Dalc))
Walc<-c(table(fulldt$Walc))
data.frame(Dalc, Walc)
```

```
##   Dalc Walc
## 1  727  398
## 2  196  235
## 3   69  200
## 4   26  138
## 5   26   73
```

```
ggplot(fulldt, aes(x = Walc, y = Dalc , color = address))+ geom_jitter(position=position_jitter(0.2))
```

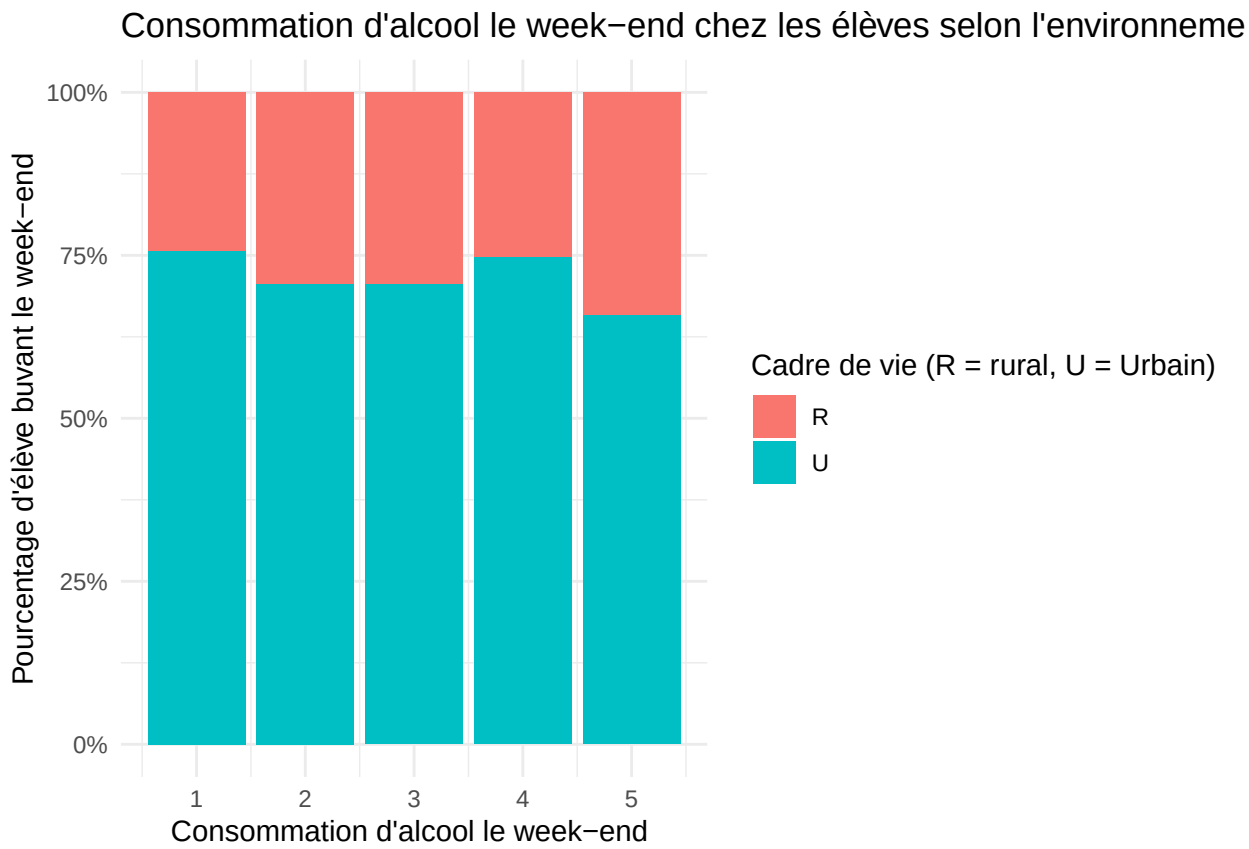


```
ggplot(fulldt) +
 aes(x = Dalc, fill = address) +
 geom_bar(position = "fill") +
 scale_fill_hue(direction = 1) +
 labs(x = "Consommation d'alcool en semaine ", y = "Pourcentage d'élève buvant en semaine",
 title = "Consommation d'alcool en semaine chez les élèves selon l'environnement ", fill = "Cadre de vi
    scale_y_continuous(labels = percent) +
 theme_minimal()
```



Le pourcentage d'élèves consommant de l'alcool en semaine est plus élevé pour les élèves habitant dans une zone urbaine (environ 75% pour les catégories 1, 2, 4 et 5).

```
ggplot(fulldt) +
 aes(x = Walc, fill = address) +
 geom_bar(position = "fill") +
 scale_fill_hue(direction = 1) +
 labs(x = "Consommation d'alcool le week-end ", y = "Pourcentage d'élève buvant le week-end",
 title = "Consommation d'alcool le week-end chez les élèves selon l'environnement ", fill = "Cadre de v
    scale_y_continuous(labels = percent) +
 theme_minimal()
```

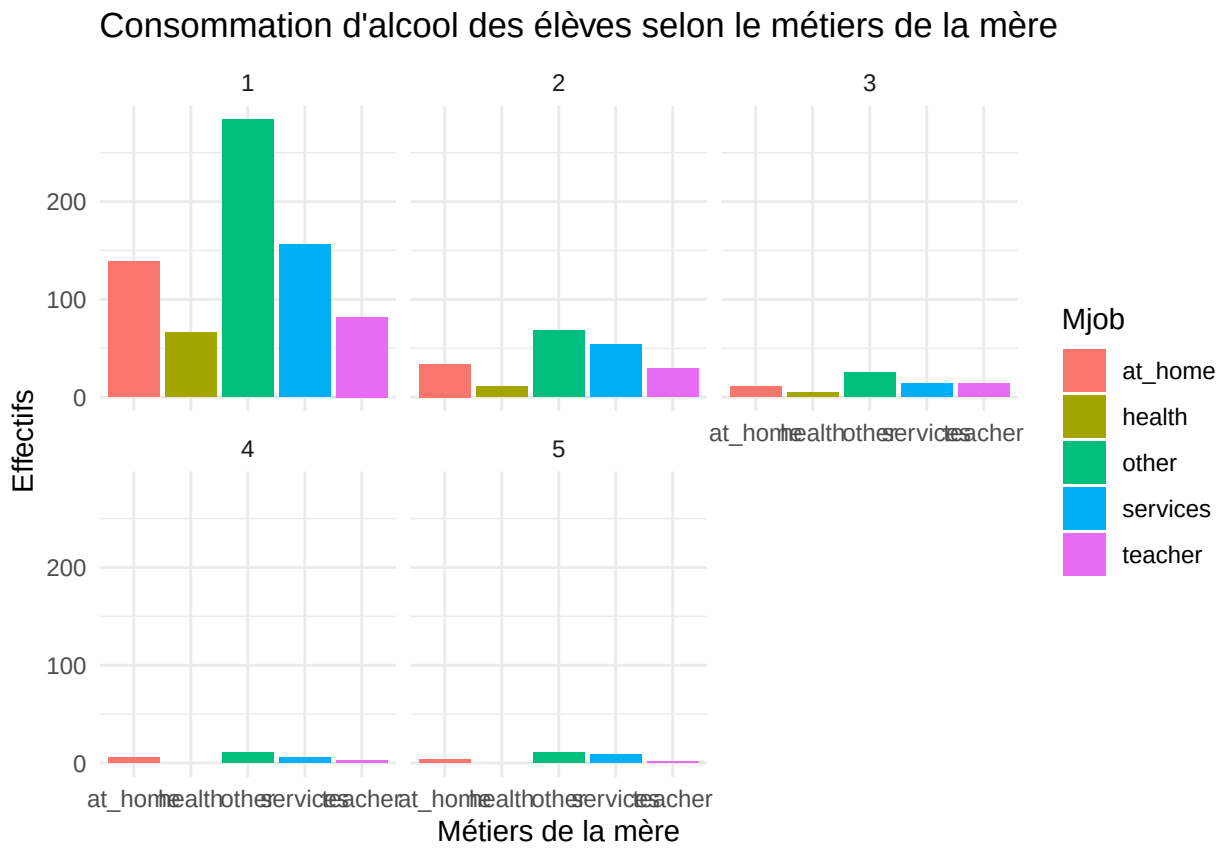


Le pourcentage d'élèves consommant de l'alcool le weekend est presque identique que le pourcentage d'élèves consomment de l'alcool en semaine, ici aussi, les élèves sont plus nombreux à déclarer habiter dans une zone urbaine.

La consommation d'alcool le week-end selon l'origine sociale Origine sociale (Medu / Fedu / Mjob / Fjob)

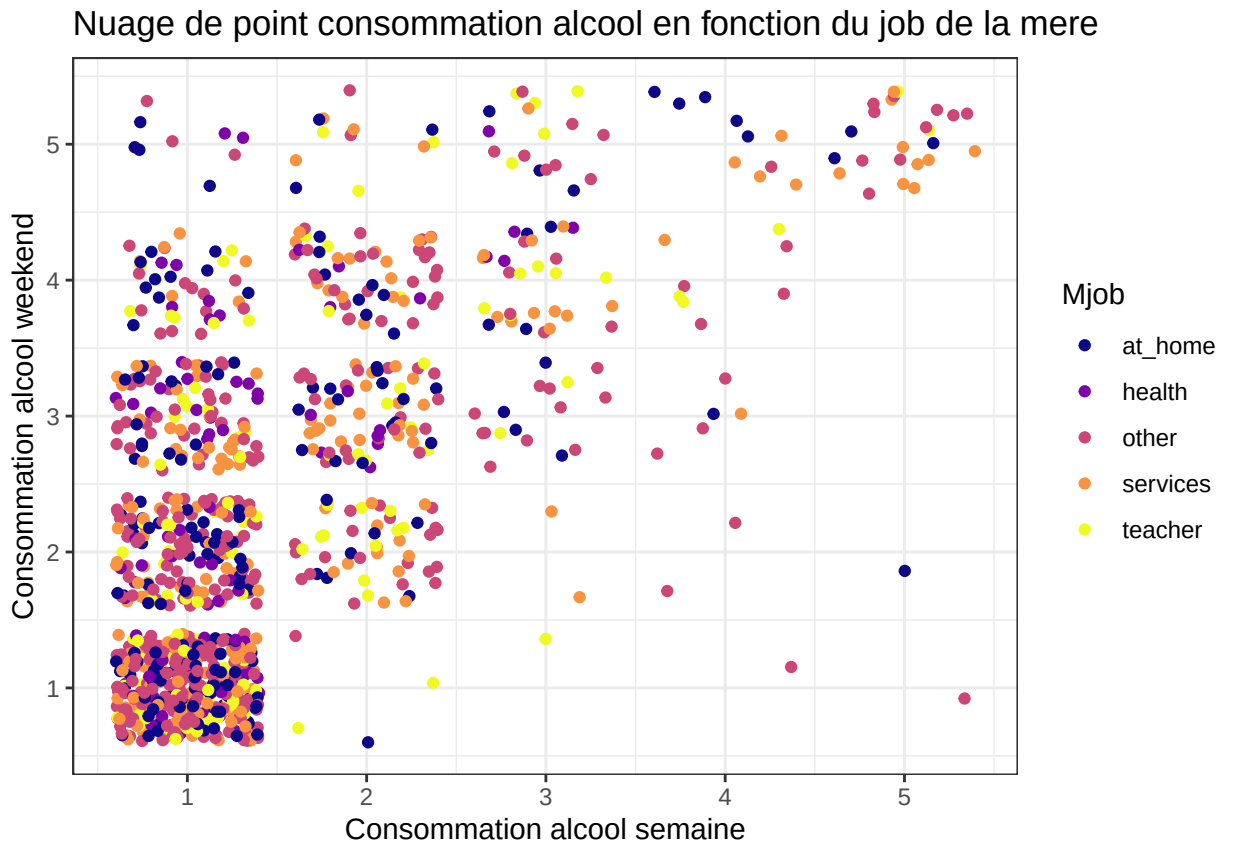
Mjob (geom_bar, geom_jitter)

```
ggplot(fulldt) +
  aes(x = Mjob, fill = Mjob) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Métiers de la mère",
    y = "Effectifs",
    title = "Consommation d'alcool des élèves selon le métiers de la mère "
  ) +
  theme_minimal() +
  facet_wrap(vars(Dalc))
```



Etant donné que les élèves déclarant boire de l'alcool est plus élèves que les autres, nous avons meilleur graphique. Nous remarquons que les mères travaillant dans "les autres", services ou encore des mères à la maison ont des enfants qui consomment beaucoup plus que les enfants ayant une mère travaillant dans la santé ou professeur.

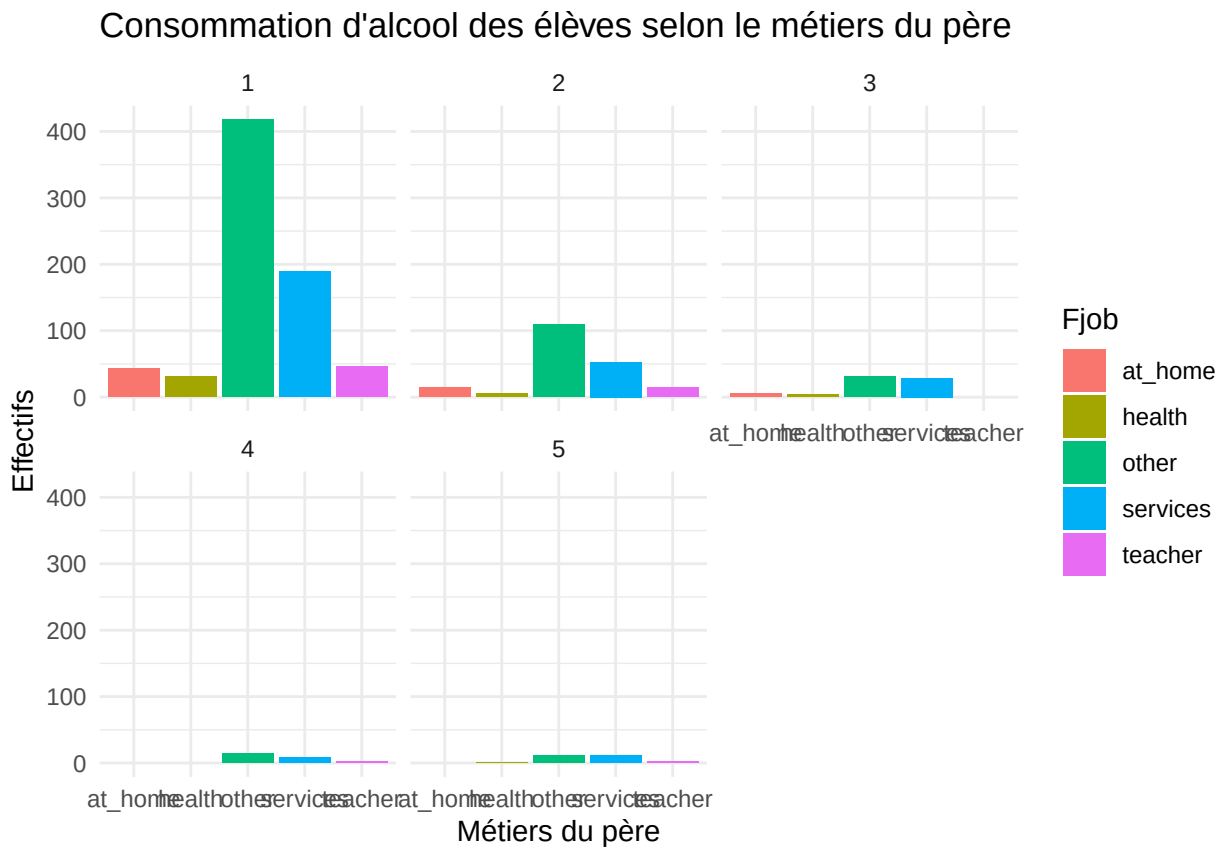
```
#Consommation d'alcool selon job mere
ggplot(fulldt) +
  aes(x = Dalc, y = Walc, colour = Mjob) +
  geom_jitter() +
  scale_color_viridis_d(option = "plasma", direction = 1) +
  labs(
    x = "Consommation alcool semaine",
    y = "Consommation alcool weekend",
    title = "Nuage de point consommation alcool en fonction du job de la mere"
  ) +
  theme_bw()
```



Comme analyser précédemment, si on observe dans ce nuage de point la partie consommant beaucoup d'alcool, on retrouve surtout des mères travaillant dans les services, autres ou des mères au foyer. De plus, nous observons qu'il y a un regroupement de points pour les échelles de consommation de le weekend et principalement en 1 pour la consommation en semaine.

Fjob (geom_bar)

```
ggplot(fulldt) +
  aes(x = Fjob, fill = Fjob) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Métiers du père",
    y = "Effectifs",
    title = "Consommation d'alcool des élèves selon le métiers du père "
  ) +
  theme_minimal() +
  facet_wrap(vars(Dalc))
```



Nous observons que comme pour le graphique du métier de la mère vu précédemment, on a un graphique distincte pour les élèves consommant très peu d'alcool. Pour les pères travaillant dans la catégories "autres", nous observons que leur enfants sont plus nombreux à boire de l'alcool que les autres métiers, suivi de services.

la structure familiale (Pstatus / famsize / famrel / guardian /)

```
table(fulldt$Pstatus)
```

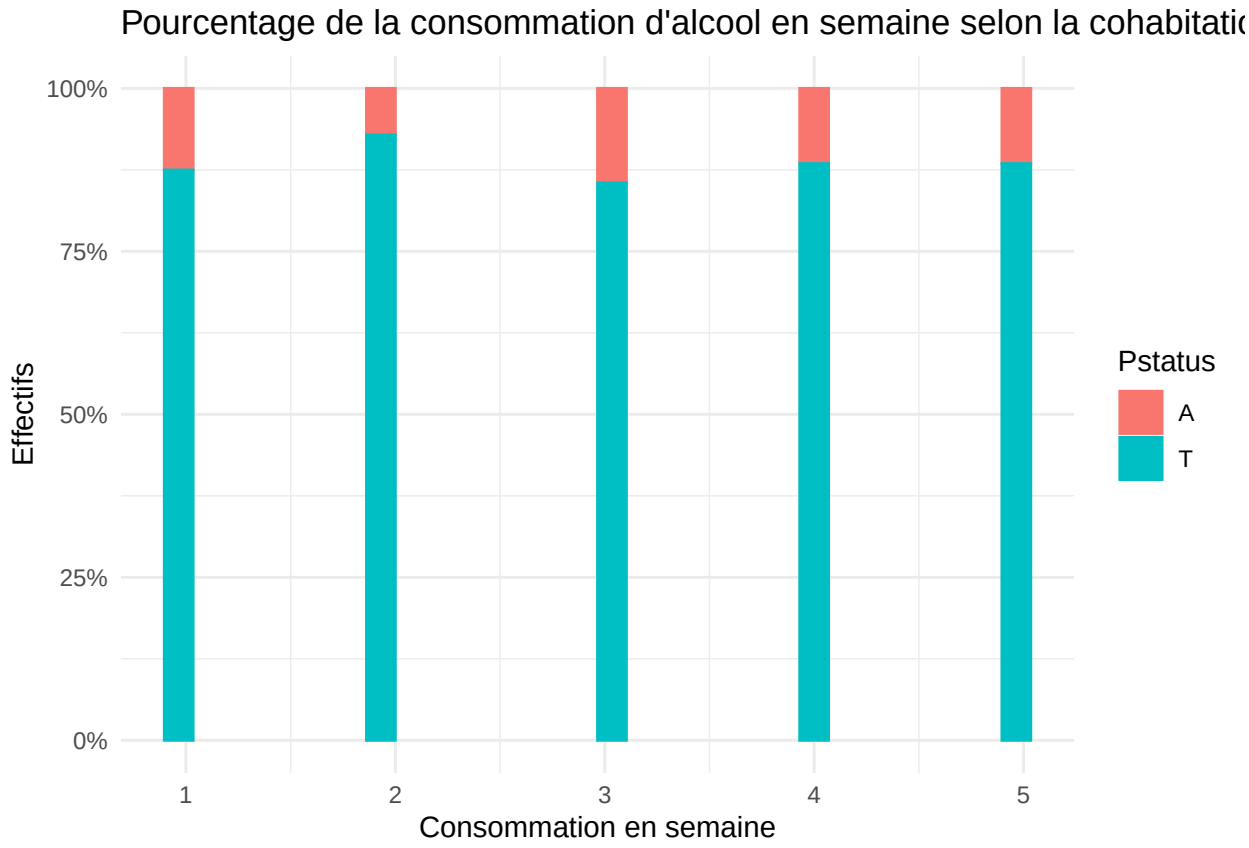
```
##
##   A   T
## 121 923
```

923 élèves ont des parents vivant toujours ensemble et 121 ont des parents séparés.

```
ggplot(fulldt) +
  aes(x = Dalc, fill = Pstatus, colour = Pstatus) +
  geom_histogram(position = "fill") +
  scale_fill_hue(direction = 1) +
  scale_color_hue(direction = 1) +
  labs(
    x = "Consommation en semaine",
    y = "Effectifs",
    title = "Pourcentage de la consommation d'alcool en semaine selon la cohabitation parentale"
  ) +
    scale_y_continuous(labels = percent) +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 50 rows containing missing values (geom_bar).
```

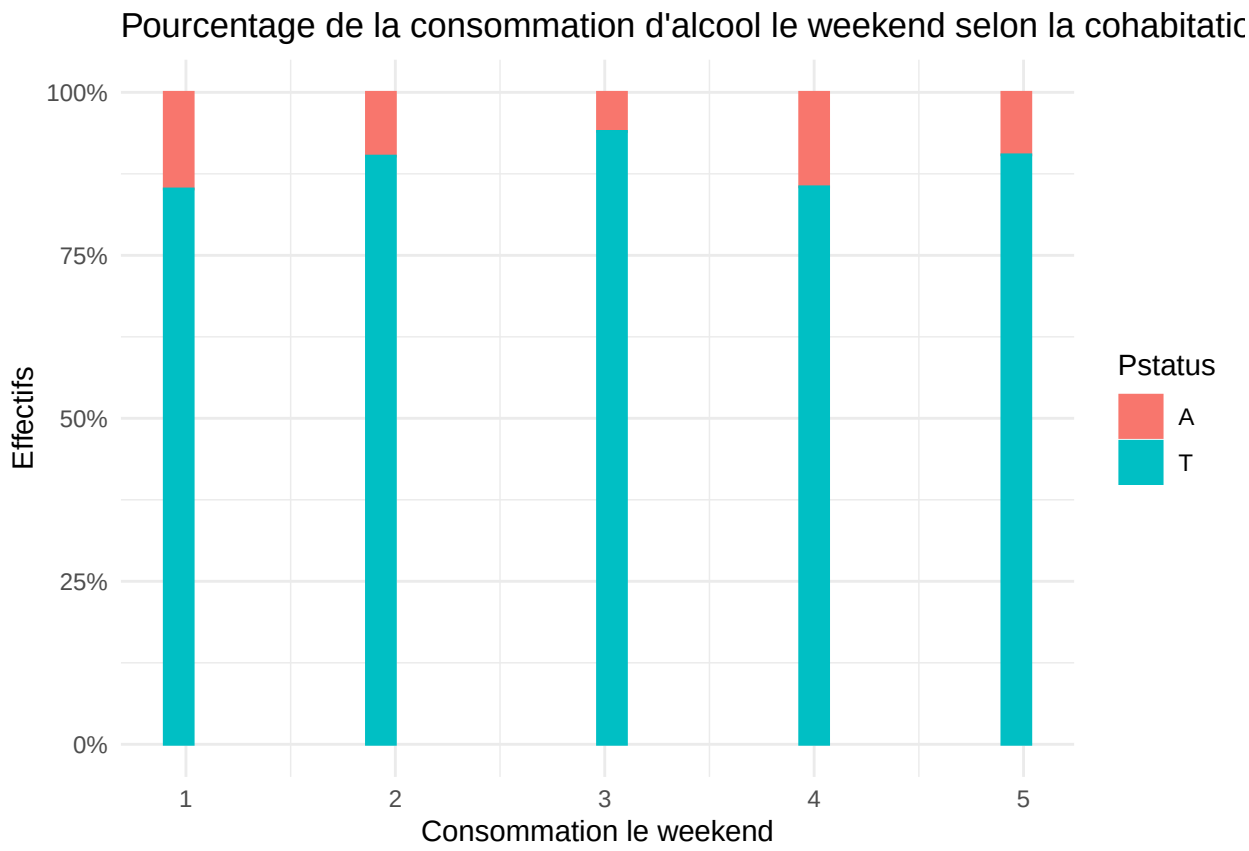


On ne peut pas conclure qu'avoir des parents séparés influent la consommation d'alcool en semaine, étant donné qu'environ 85% ont des parents vivant ensemble.

```
ggplot(fulldt) +
  aes(x = Walc, fill = Pstatus, colour = Pstatus) +
  geom_histogram(position = "fill") +
  scale_fill_hue(direction = 1) +
  scale_color_hue(direction = 1) +
  labs(
    x = "Consommation le weekend",
    y = "Effectifs",
    title = "Pourcentage de la consommation d'alcool le weekend selon la cohabitation parentale"
  ) +
    scale_y_continuous(labels = percent) +
  theme_minimal()
```

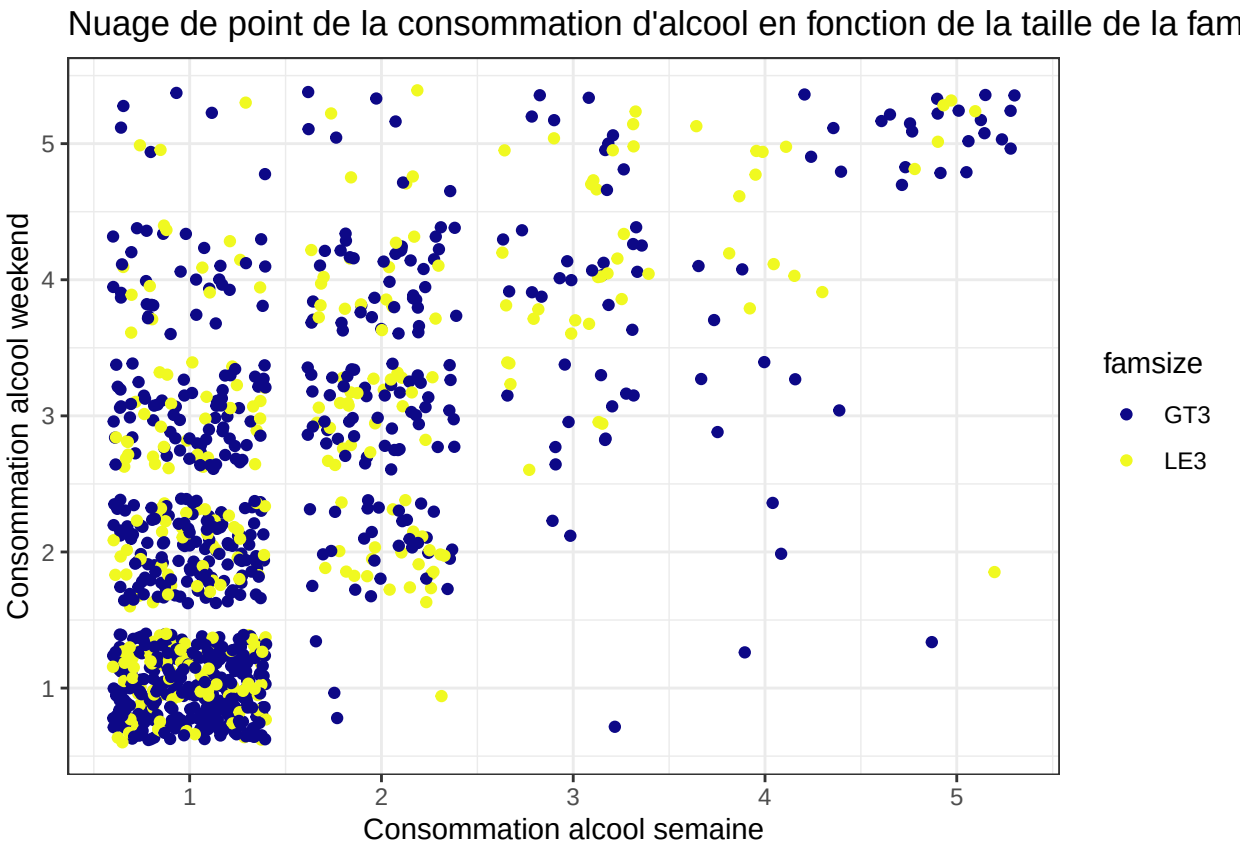
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 50 rows containing missing values (geom_bar).
```



La conclusion est identique pour la consommation d'alcool le weekend, nous pouvons affirmer que le fait d'avoir des parents séparés n'affecte pas la consommation d'alcool en générale.

```
ggplot(fulldt) +
  aes(x = Dalc, y = Walc, colour = famsize) +
  geom_jitter() +
  scale_color_viridis_d(option = "plasma", direction = 1) +
  labs(
    x = "Consommation alcool semaine",
    y = "Consommation alcool weekend",
    title = "Nuage de point de la consommation d'alcool en fonction de la taille de la famille "
  ) +
  theme_bw()
```

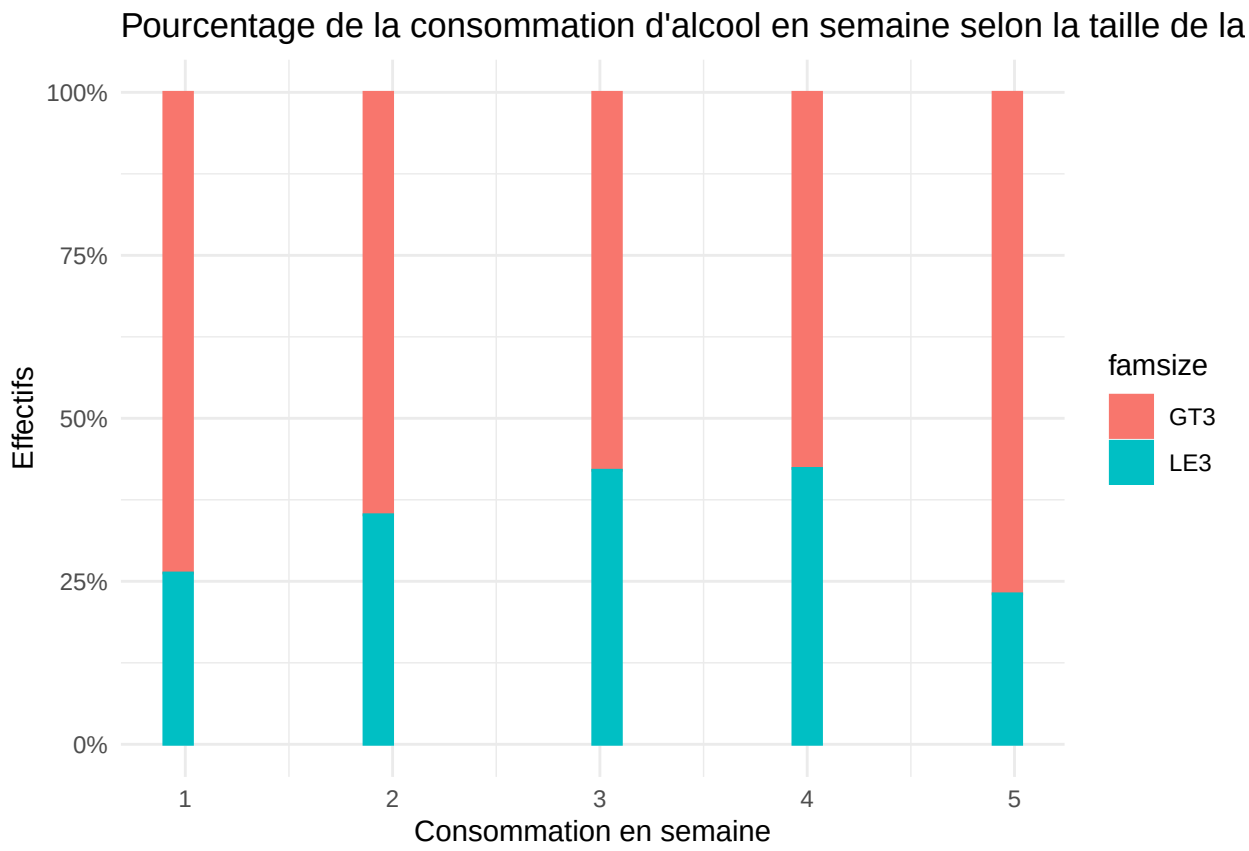


Nous observons que pour la consommation d'alcool excessive, nous retrouvons majoritairement des points bleus, c'est à dire des élèves ayant une famille de taille supérieur à 3.

```
ggplot(fulldt) +
  aes(x = Dalc, fill = famsize, colour = famsize) +
  geom_histogram(position = "fill") +
  scale_fill_hue(direction = 1) +
  scale_color_hue(direction = 1) +
  labs(
    x = "Consommation en semaine",
    y = "Effectifs",
    title = "Pourcentage de la consommation d'alcool en semaine selon la taille de la famille"
  ) +
    scale_y_continuous(labels = percent) +
  theme_minimal()
```

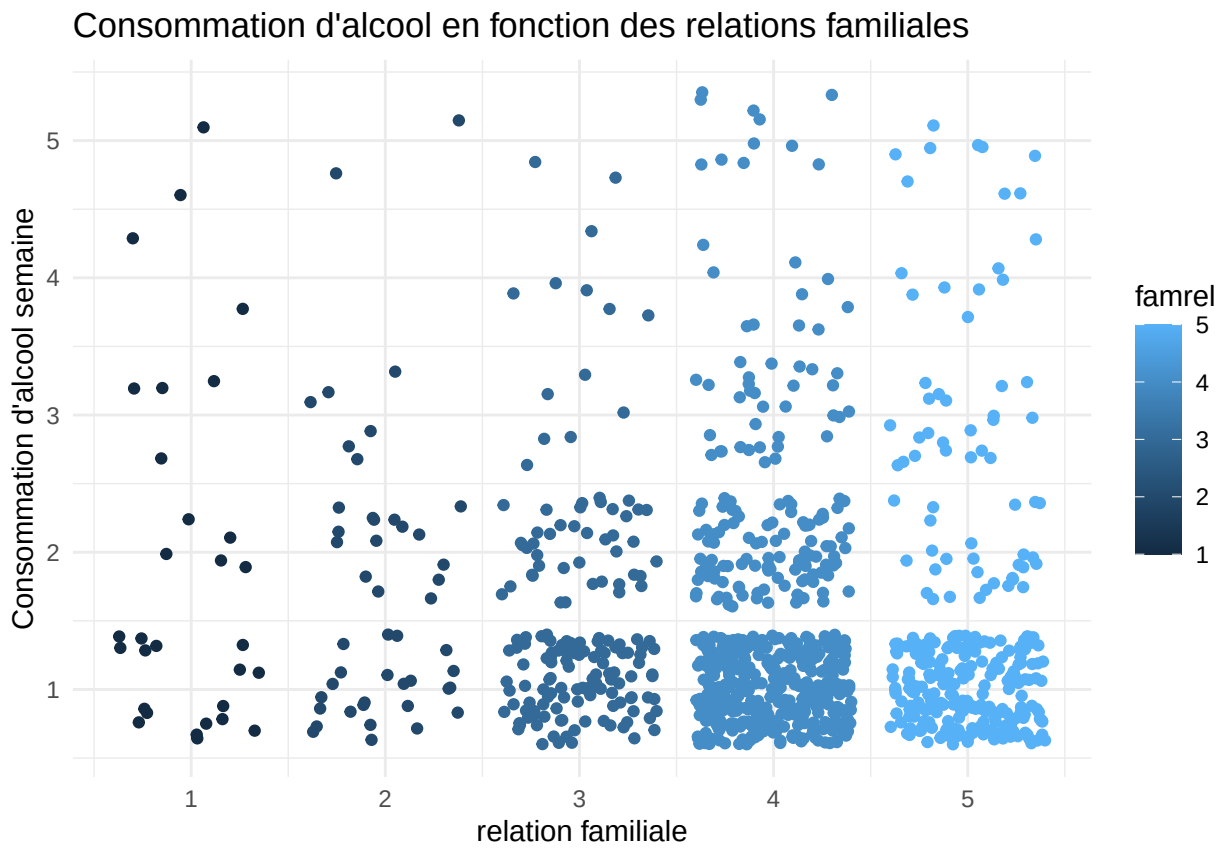
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 50 rows containing missing values (geom_bar).
```



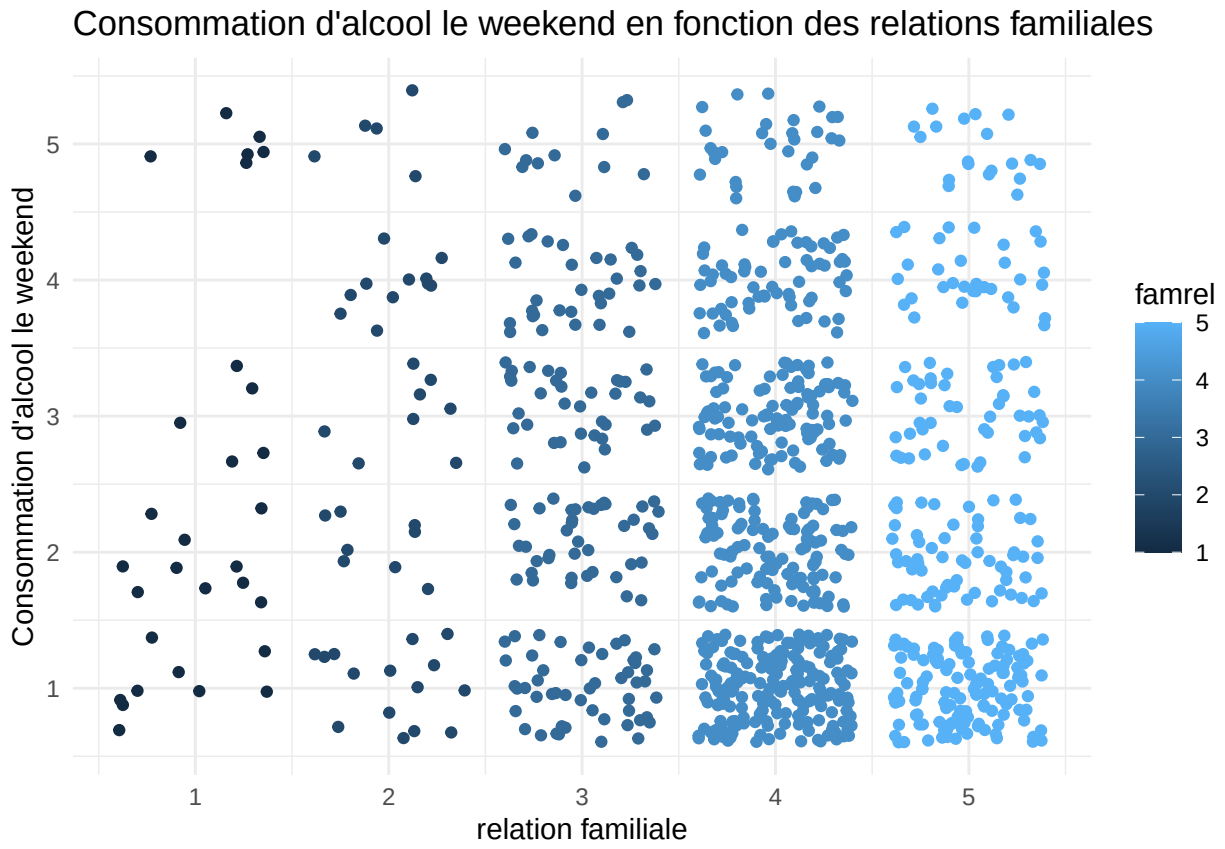
Grace à cet histogramme, on observe que généralement plus de la moitié des consommateurs sont issues de familles nombreuses. 75% des personnes déclarant boire très peu ou beaucoup d'alcool sont issues de famille nombreuses.

```
ggplot(fulldt) +
  aes(x = famrel, y = Dalc, colour = famrel) +
  geom_jitter(size = 1.5) +
  scale_color_gradient() +
  labs(
    x = "relation familiale",
    y = "Consommation d'alcool semaine",
    title = "Consommation d'alcool en fonction des relations familiales"
  ) +
  theme_minimal()
```



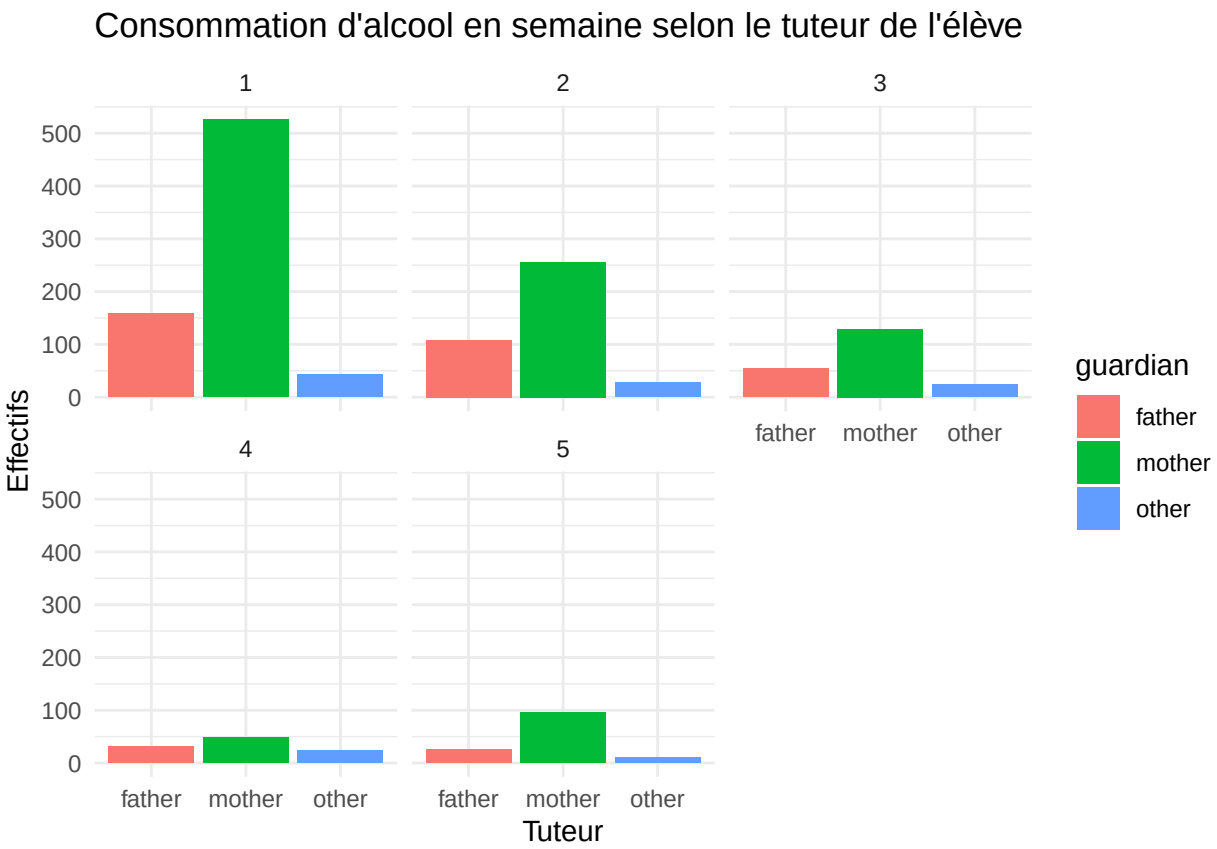
Nous observons que les personnes consomment très peu d'alcool en semaine ont une très bonne relation familiale majoritairement. Généralement, les nuages de points pour les relations familiales mauvaises se regroupent pour la consommation d'alcool très faible. Nous pouvons conclure que les relations familiale n'est peu être pas forcement un facteur de la consommation d'alcool chez les élèves, étant donné qu'il n'y a que 30 élèves déclarant avoir des relations mauvaises, nous pouvons pas faire une conclusion.

```
ggplot(fulldt) +
  aes(x = famrel, y = Walc, colour = famrel) +
  geom_jitter(size = 1.5) +
  scale_color_gradient() +
  labs(
    x = "relation familiale",
    y = "Consommation d'alcool le weekend",
    title = "Consommation d'alcool le weekend en fonction des relations familiales"
  ) +
  theme_minimal()
```



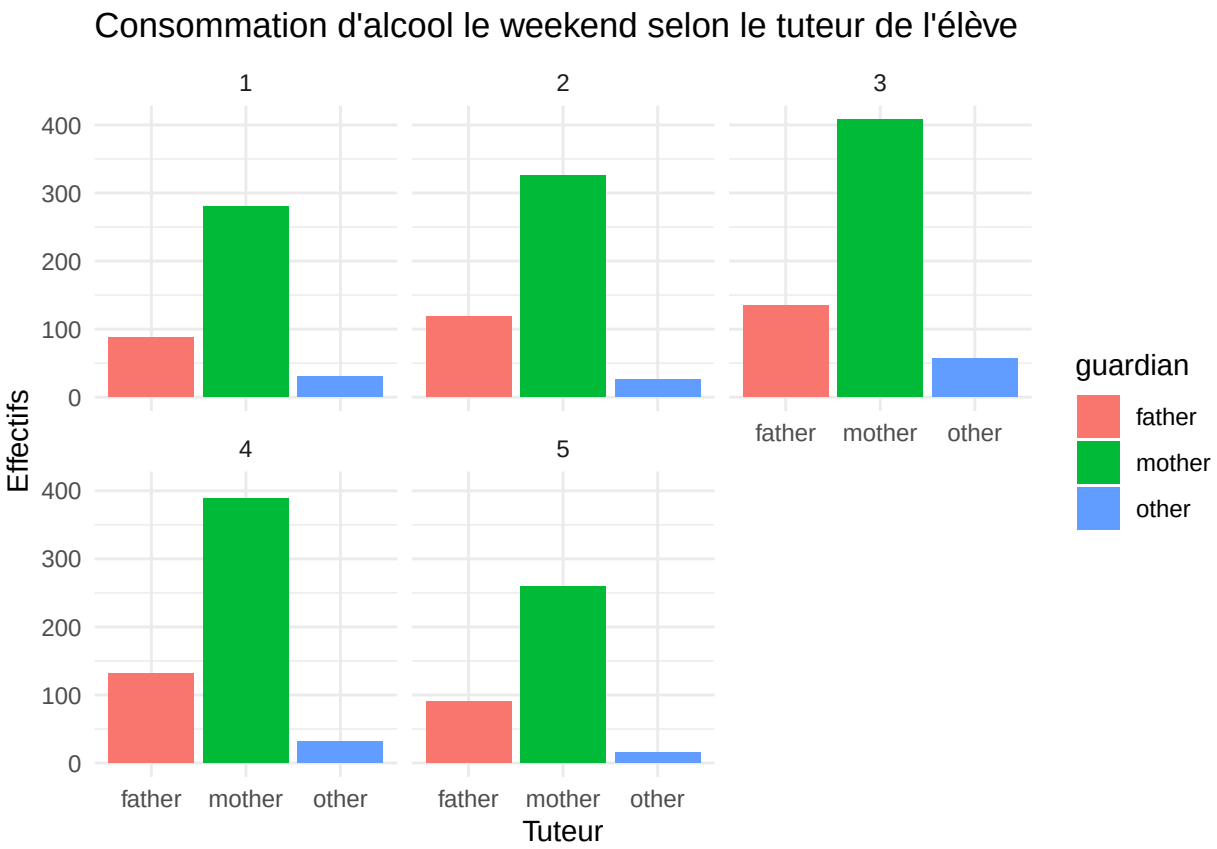
Identique à la précédente interprétation, étant donné qu'il y a très peu d'élèves déclarant avoir une mauvaise relation familiale, le nuage de point n'est pas forcement approprié pour savoir si les relations familiale est un facteur de la consommation d'alcool.

```
ggplot(fulldt) +
  aes(x = guardian, fill = guardian, weight = Dalc) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Tuteur",
    y = "Effectifs",
    title = "Consommation d'alcool en semaine selon le tuteur de l'élève"
  ) +
  theme_minimal() +
  facet_wrap(vars(Dalc))
```



Généralement, les élèves ayant comme tuteur leur mère consomment plus d'alcool, suivi du tuteur père. Dans la catégorie autres très peu consomment d'alcool.

```
ggplot(fulldt) +
  aes(x = guardian, fill = guardian, weight = Walc) +
  geom_bar() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Tuteur",
    y = "Effectifs",
    title = "Consommation d'alcool le weekend selon le tuteur de l'élève"
  ) +
  theme_minimal() +
  facet_wrap(vars(Walc))
```

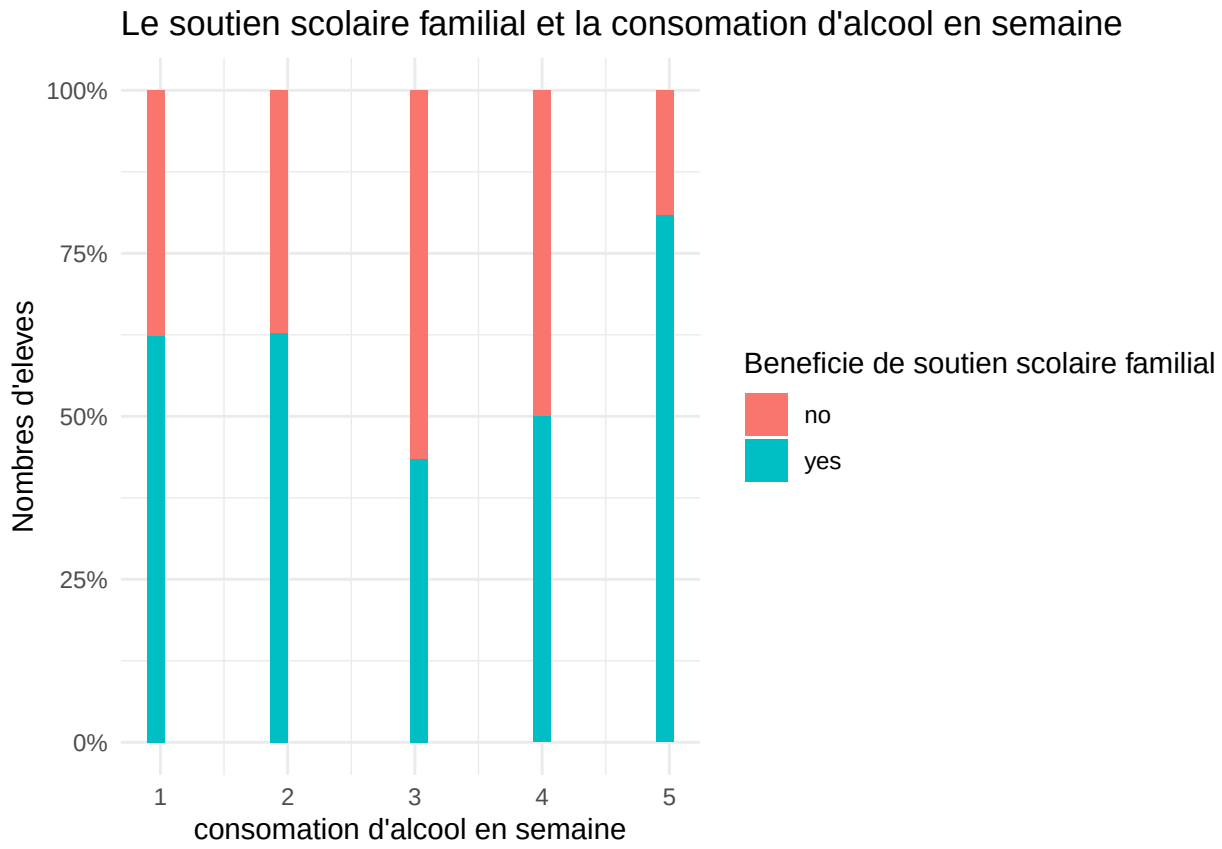


Ce graphique est un peu près identique que la consommation en semaine, sauf que les effectifs sont beaucoup plus important.

L'accompagnement (variables schoolsup / famsup / paid)

```
ggplot(fulldt) +
  aes(x = Dalc, fill = famsup) +
  geom_histogram(position = "fill", bins = 30) +
  scale_fill_hue(direction = 1) +
  labs (
    x = "consomation d'alcool en semaine" ,
    y = "Nombres d'eleves" ,
    title = "Le soutien scolaire familial et la consomation d'alcool en semaine",
    fill = "Beneficie de soutien scolaire familial"
  ) +
      scale_y_continuous(labels = percent) +
  theme_minimal()
```

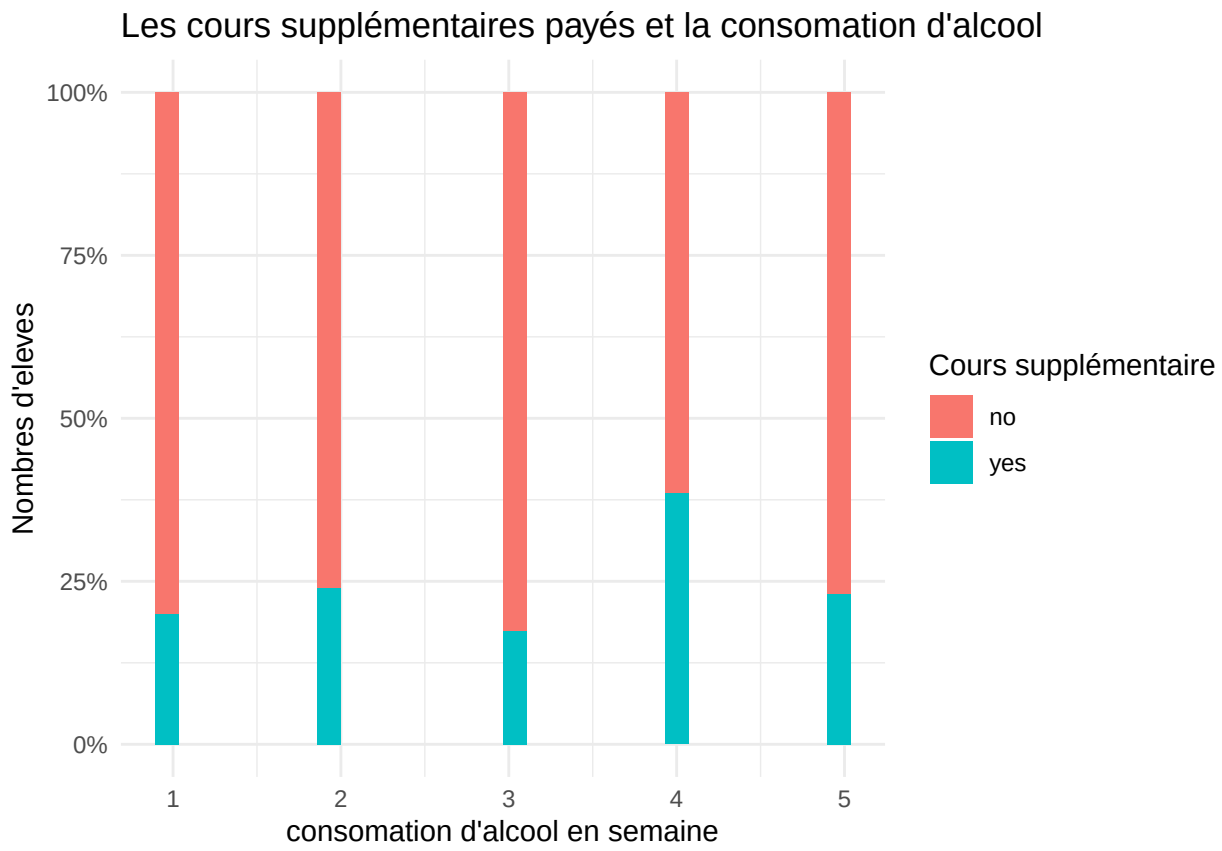
```
## Warning: Removed 50 rows containing missing values (geom_bar).
```



Les élèves consommant beaucoup d'alcool en semaine ont environ 80% de soutiens familiale, contre 65% pour des élèves consommant très peu d'alcool. Pour les personnes ayant une consommation ayant une consommation normale (3), environ 55% ne bénéficie pas de soutient familiale.

```
ggplot(fulldt) +
  aes(x = Dalc, fill = paid) +
  geom_histogram(position = "fill", bins = 30L) +
  scale_fill_hue(direction = 1) +
  labs (
    x = "consomation d'alcool en semaine",
    y = "Nombres d'eleves" ,
    title = "Les cours supplémentaires payés et la consomation d'alcool",
    fill = "Cours supplémentaire"
  ) +
  scale_y_continuous(labels = percent) +
  theme_minimal()
```

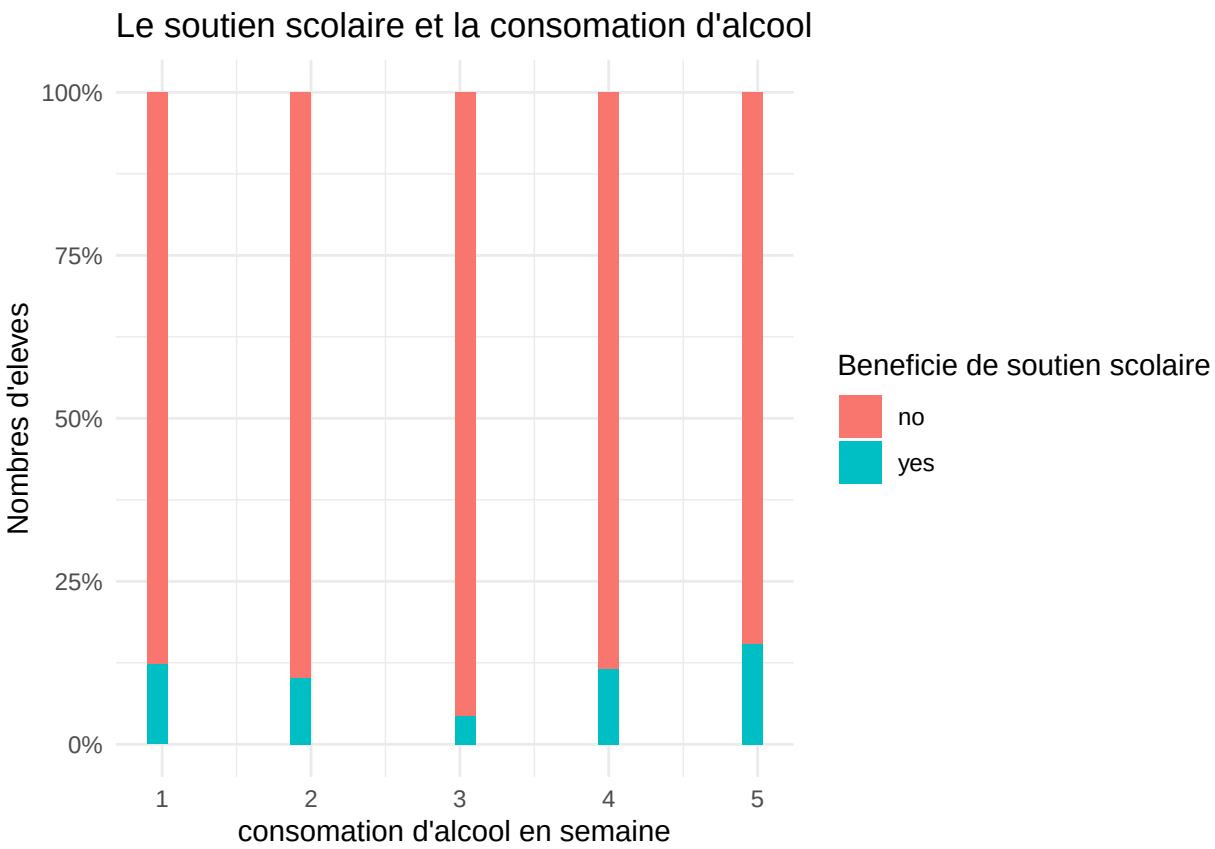
```
## Warning: Removed 50 rows containing missing values (geom_bar).
```



Les élèves bénéficiant du soutient scolaire consomme moins d'alcool que les élèves ne bénéficiant pas de soutient scolaire, en effet, les élèves déclarant consommés très peu d'alcool est de 20%, et de d'environ 35% pour les élèves consommant beaucoup d'alcool.

```
ggplot(fulldt) +
  aes(x = Dalc, fill = schoolsup ) +
  geom_histogram(position = "fill", bins = 30L) +
  scale_fill_hue(direction = 1) +
  labs (
    x = "consomation d'alcool en semaine" ,
    y = "Nombres d'eleves" ,
    title = "Le soutien scolaire et la consomation d'alcool",
    fill = "Beneficie de soutien scolaire"
  ) +
  scale_y_continuous(labels = percent) +
  theme_minimal()
```

```
## Warning: Removed 50 rows containing missing values (geom_bar).
```

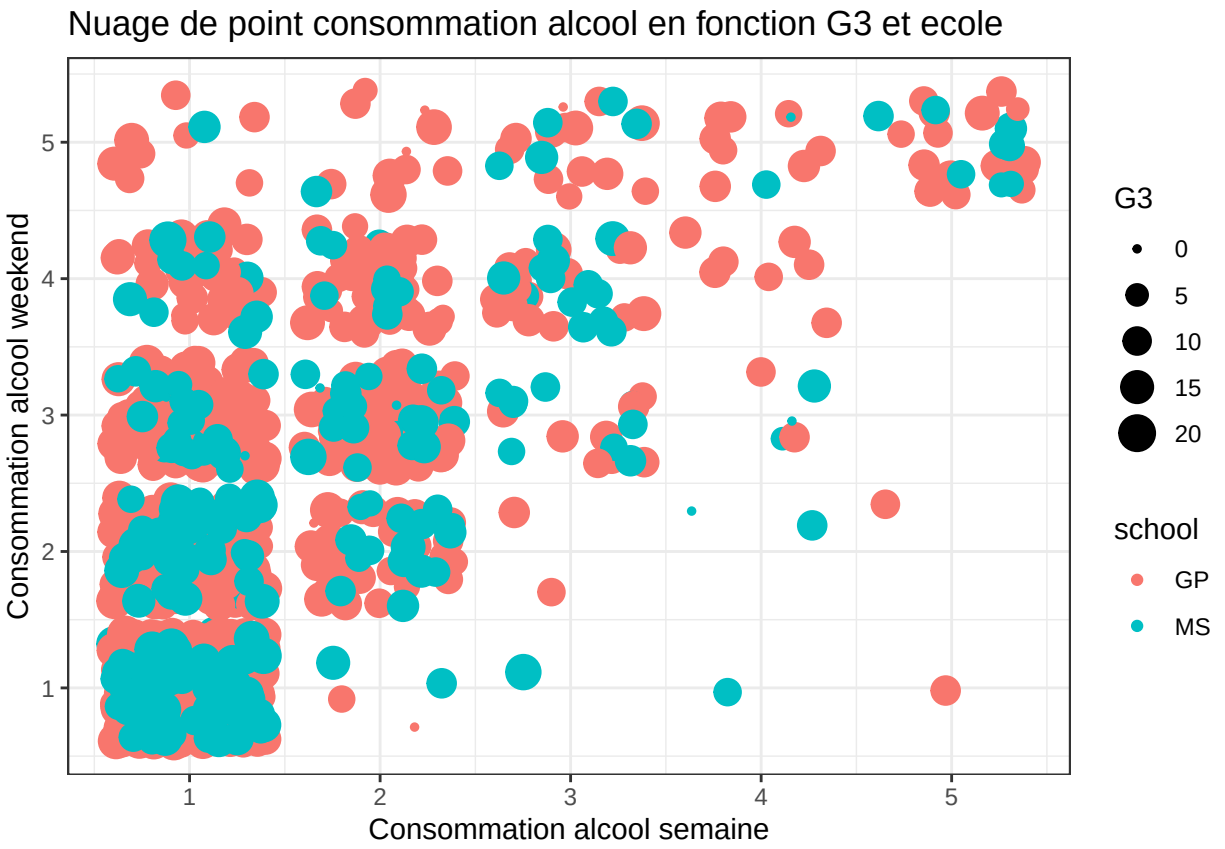


Les personnes qui bénéficie d'un soutient scolaire ne boivent presque pas d'alcool en semaine, seulement 12,5% consomment très peu et 15% consomment beaucoup d'alcool en semaine.

2. Visualisation des facteurs impactant les notes G3

```
ggplot(fulldt) +
  aes(x = Dalc, y = Walc, colour = school, size = G3) +
  geom_jitter() +
  scale_color_hue(direction = 1) +
  labs(
    x = "Consommation alcool semaine",
    y = "Consommation alcool weekend",
    title = "Nuage de point consommation alcool en fonction G3 et ecole"
  ) +
  theme_bw()
```

Impact de l'école sur les notes G3



Ici on utilise le nuage de point pour montrer la différence entre école et l'influence de la consommation d'alcool sur les notes aux examens finaux (G3) car la fonction nuage de point nous permet ici de facilement montrer (ou non) la disparité entre les deux écoles et surtout de faire ressortir les tendances statistiques sur la consommation d'alcool des élèves et peut être d'établir une influence sur les résultats finaux aux examens des élèves. Le graphique se lit comme suit : les points en haut a droite sont ceux qui consomment le plus d'alcool a la fois en weekend et en semaine (ceux se rapprochant le plus du 5 sur l'abscisse et l'ordonnée). A l'inverse, ceux étant le plus en bas a gauche (les plus proches de 1) sont ceux consommant le moins, voir pas d'alcool. En haut a gauche, se situent les personnes consommant exclusivement le weekend (1 a l'abscisse et 5 a l'ordonnée), et ceux étant le plus en bas a droite sont ceux qui consomment de l'alcool exclusivement en semaine (5 a l'abscisse et 1 a l'ordonne). Le premier résultat que l'on remarque est que la tendance a la consommation d'alcool est beaucoup plus importante durant le weekend que durant la semaine malgré quelques exceptions. On peut expliquer cette tendance par le fait que la consommation d'alcool a l'adolescence et pour les jeunes adultes soient surtout liée a des moments sociabilités entre groupes de pair. Les étudiants ayant cours en semaine, la plupart des activités sociales sont donc organisées en fin de semaine, en weekend et ces activités sont le moment propices a la consommation d'alcool en vue de sociabiliser (sorties en bars, boites, soirée chez quelqu'un). Le deuxième résultat que l'on peut noter est que les élèves de Gabirel Pereira (en rouge) sont plus nombreux a consommer de l'alcool que ceux de Mousinho da Silveira (en bleu). Les élèves de GP sont aussi surtout beaucoup plus représenter dans des consommations intensives et notamment celles faites en semaines. Difficile de dire si la différence est assez grande pour être significative. Un début d'analyse pour expliquer cette différence serait que Gabriel Pereira est une école publique et donc que les élèves aient une plus grande liberté et moins d'attente au niveau du corps enseignant. Le troisième résultat est que l'on remarque que les résultats les plus bas (0, 5, 10) ont tendance a augmenter avec la consommation d'alcool et surtout avec la consommation d'alcool en semaine. On ne peut pas affirmer que ce résultat soit significatif au vu de la présence de très nombreux non-buveurs ou buveurs occasionnelle. Cependant on peut imaginer une hypothèse qui est que ceux qui consomment en semaine sont probablement victime d'addiction, d'alcoolisme étant donne que l'addiction a l'alcool se caractérise par une consommation

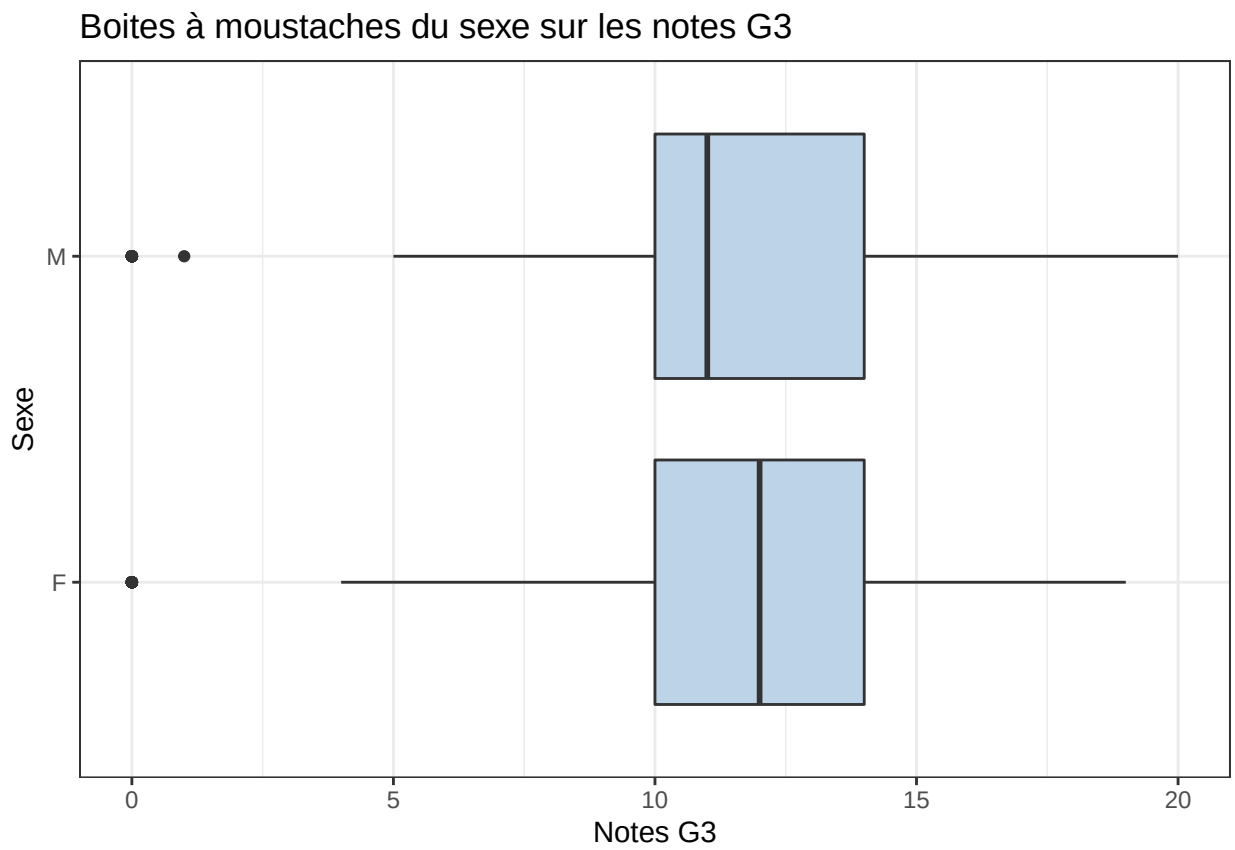
presque journalière. On manque d'élément pour approuver la validation d'une tel hypothèse, mais tout comment l'addiction a la marijuana a des conséquences sur les notes, l'addiction a l'alcool pourrait entraîner des conséquences sur les notes finales comme le montre le graphique. Cependant, tout comme l'addiction a la marijuana, il faut voir si c'est l'addiction a l'alcool en elle même qui cause la baisse des notes ou un environnement social particulièrement précaire qui causerait cette baisse des notes et cette addiction a la fois. C'est pour cela que nous analyseront la consommation en fonction de l'origine sociale, de la structure familiale et de l'accompagnement des élèves.

```
table(fulldt$G3)
```

Impact du sexe sur les notes G3

```
##
##  0  1  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
## 53  1  1  8 18 19 67 63 153 151 103 113 90 82 52 35 27 7 1
```

```
ggplot(fulldt) +
  aes(x = G3, y = sex) +
  geom_boxplot(fill = "#BDD3E8") +
  labs(x = "Notes G3", y = "Sexe", title = "Boites à moustaches du sexe sur les notes G3") +
  theme_bw()
```

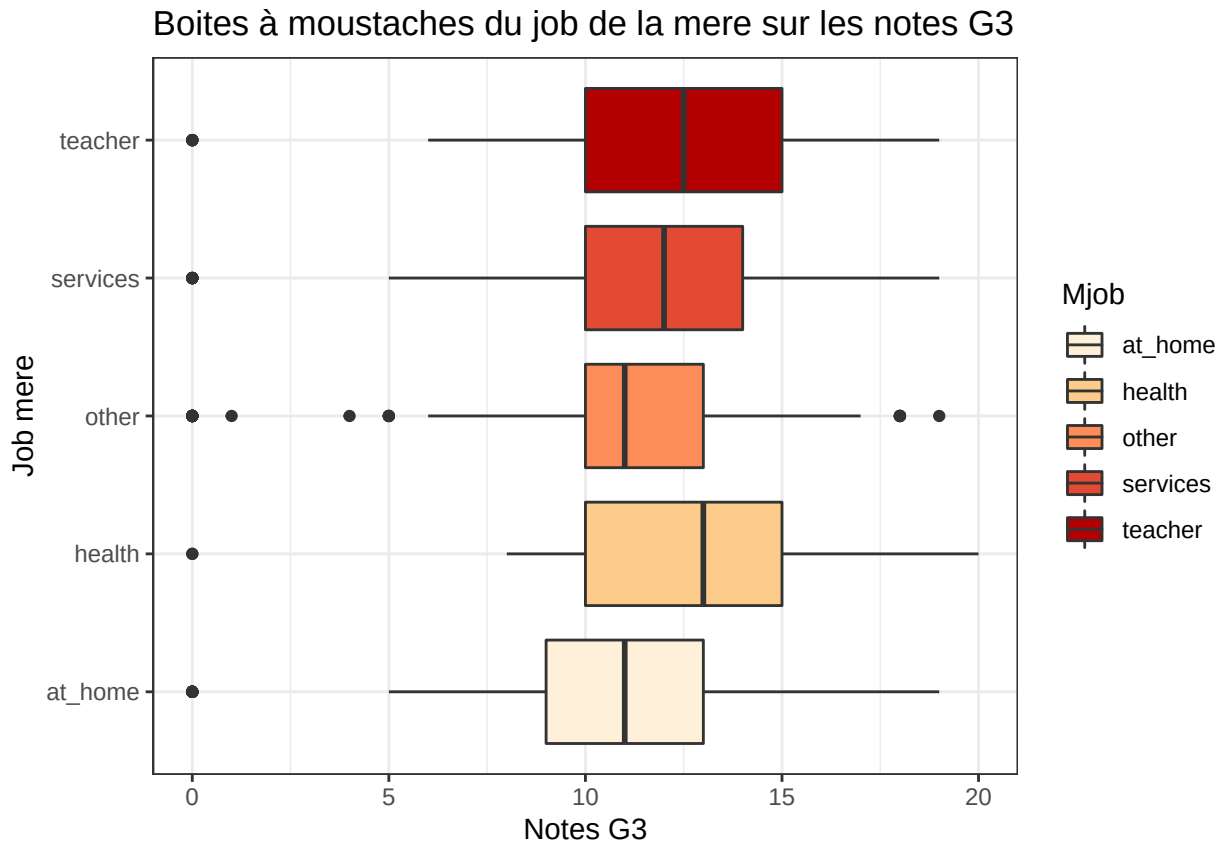


Pour le sexe masculin, il y a 2 valeurs aberrantes et pour le sexe féminin seulement 1. Les notes médians sont plus élèves pour les filles avec environ 12 et 11 pour les garçons. Par contre, le minimum est plus élevé

chez les garçon, le maximum aussi est plus élevé pour les garçons. Cela s'explique que les notes des filles sont similaires et que globalement les filles ont des notes élevées alors que chez les garçons, malgré quelques bonnes notes, ils ont plutôt des notes moyennes.

```
ggplot(fulldt) +
  aes(x = G3, y = Mjob, fill = Mjob) +
  geom_boxplot() +
  scale_fill_brewer(palette = "OrRd", direction = 1) +
  labs(x = "Notes G3", y = "Job mere", title = "Boites à moustaches du job de la mere sur les notes G3 
  theme_bw()
```

Impact du métier de la mère sur les notes G3

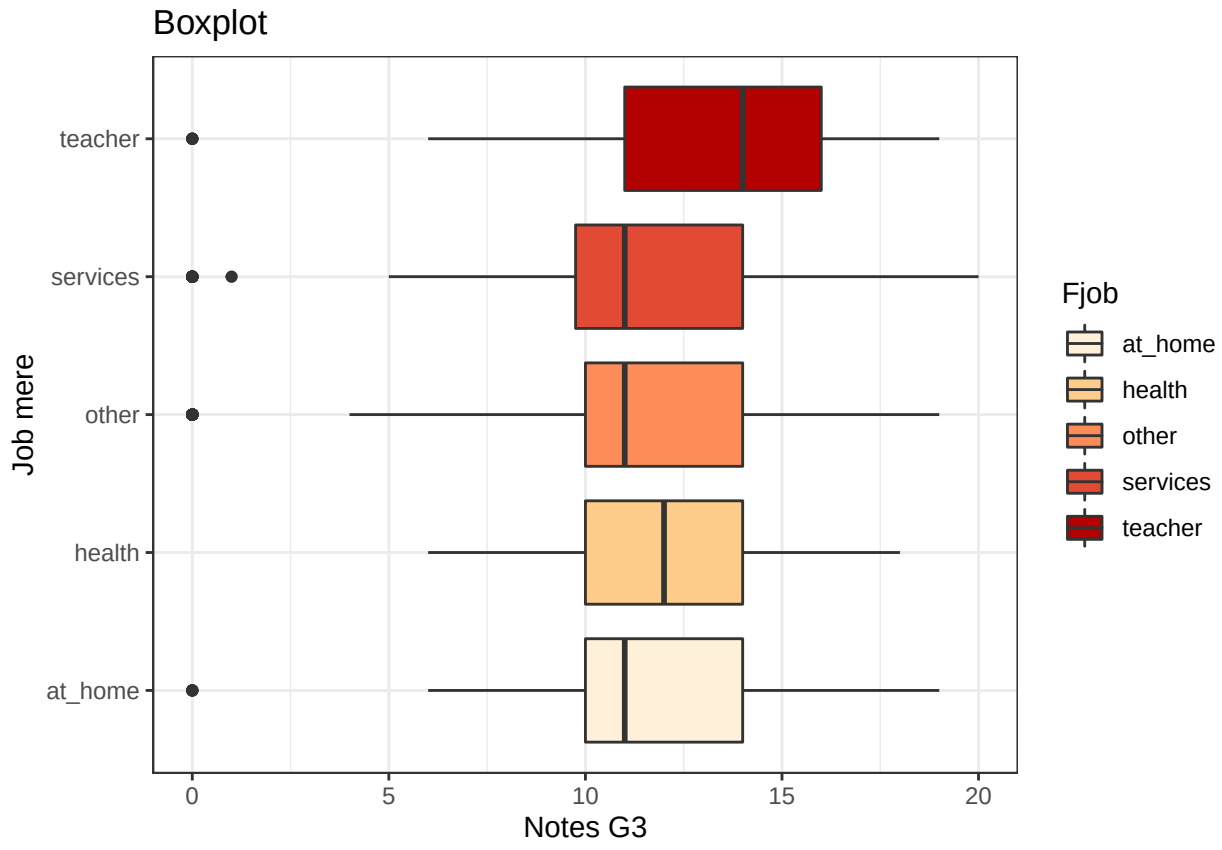


Pour les boites à moustache du métier de la mère, il y a 6 valeurs aberrantes pour la catégorie "autres", ce sont des valeurs qui sont supérieures ou inférieures aux limites définies par les moustaches. On observe selon la boite à moustache que les élèves ayant une mère qui est soit enseignante, soit dans la sante ou à la maison ont leur notes minimums comprise entre 5 et 7. Ceux qui ont les notes les plus élèves sont ceux dont leur mères est enseignante (médiane supérieure aux autres ainsi que le troisième quartile). Les élèves qui ont leur mères qui est "autre" (other) ont leur notes minimum qui est plus large (4,5 environ), surement liées au fait que cette catégorie est très large (beaucoup de qualificatif large à lintérieur). Les élèves ayant une mère enseignante (teacher) ont de meilleures notes (ayant leur médiane se rapprochant plus de 15 et leur 3ème quartile dépassant les 15). Ceux qui ont une mère à la maison (at_home) en majorité ont des notes

comprise entre 10 et 14 (la médiane étant environ à 11). On peut conclure alors que la profession de la mère à une influence sur les résultats scolaires de l'élève. Cela peut s'expliquer notamment par le temps disponible liées à la profession mais aussi les aides qui peuvent être apporter (par exemple le fait d'autre enseignant est plus susceptible d'aider). Le minimum de mère-sante est de 7 et le maximum est de 20 contrairement aux mères-maison ou mères-service qui est de 5 et 18,5 respectivement. Par contre, l'écart interquartile est vraiment important pour les mères-sante, ce qui signifie que les notes sont variables.

```
ggplot(fulldt) +
  aes(x = G3, y = Fjob, fill = Fjob) +
  geom_boxplot() +
  scale_fill_brewer(palette = "OrRd", direction = 1) +
  labs(x = "Notes G3", y = "Job mere", title = "Boxplot") +
  theme_bw()
```

Impact du métier du père sur les notes G3

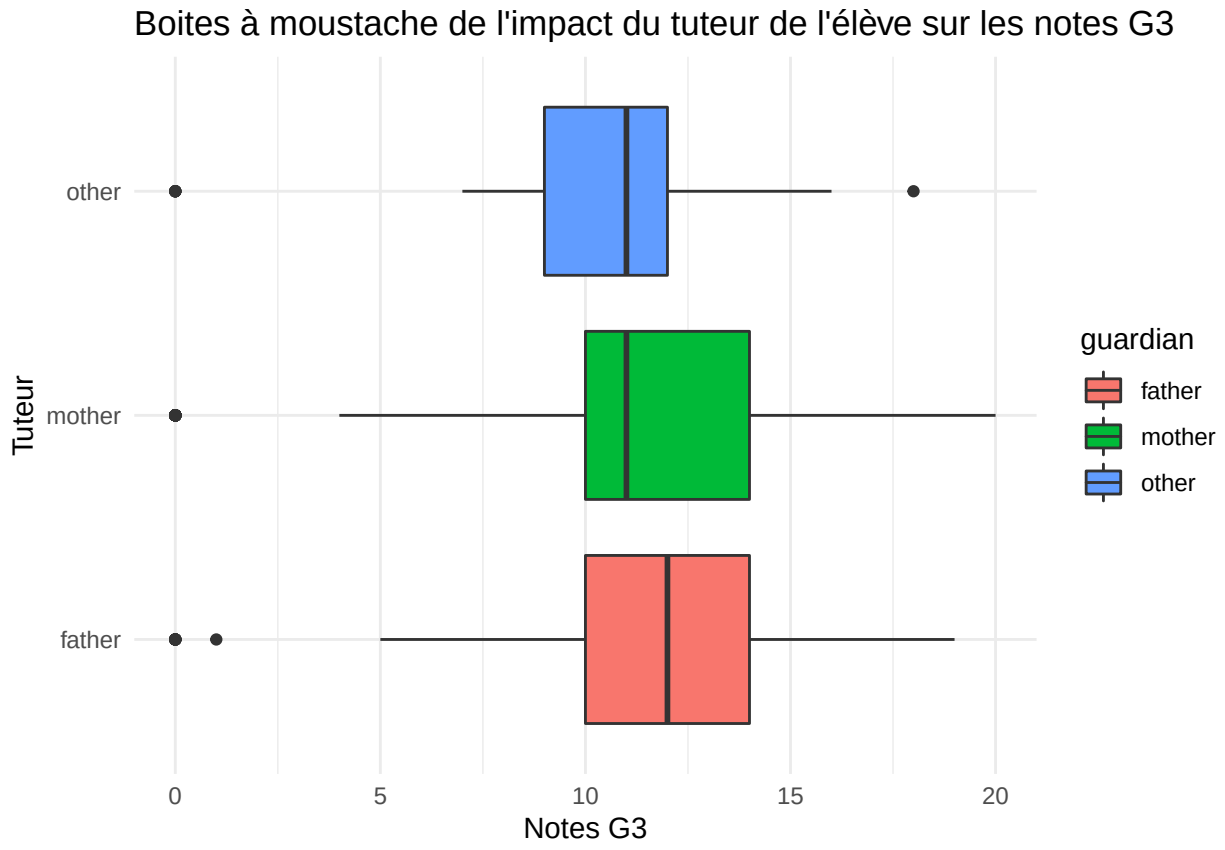


Pour les boites à moustache du métier du père, il y a 2 valeurs aberrantes pour la catégorie "services", ce sont des valeurs qui sont supérieures ou inférieures aux limites définies par les moustaches. On observe selon la boite à moustache que les élèves ayant un père travaillant dans les catégories services, à la maison ou autres ont des médians similaire égales à environ 11. Le maximum le plus élevé appartient à un élève qui a un père travaillant dans les services.
La boites à moustache de père-professeur est particulièrement différentes des autres, car le premier et

troisième quartile sont supérieur aux autres, les notes sont généralement compris entre 11 et 16. Egalement les pères-santé ont un médian légèrement supérieur aux 3 autres catégories, le fait de faire des études peut donc impactés les notes de leurs enfants. Les notes minimum de autres, à la maison et professeurs sont identiques.

```
ggplot(fulldt) +
  aes(x = G3, y = guardian, fill = guardian) +
  geom_boxplot() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Notes G3",
    y = "Tuteur",
    title = "Boites à moustache de l'impact du tuteur de l'élève sur les notes G3"
  ) +
  theme_minimal()
```

Impact du tuteur sur les notes G3



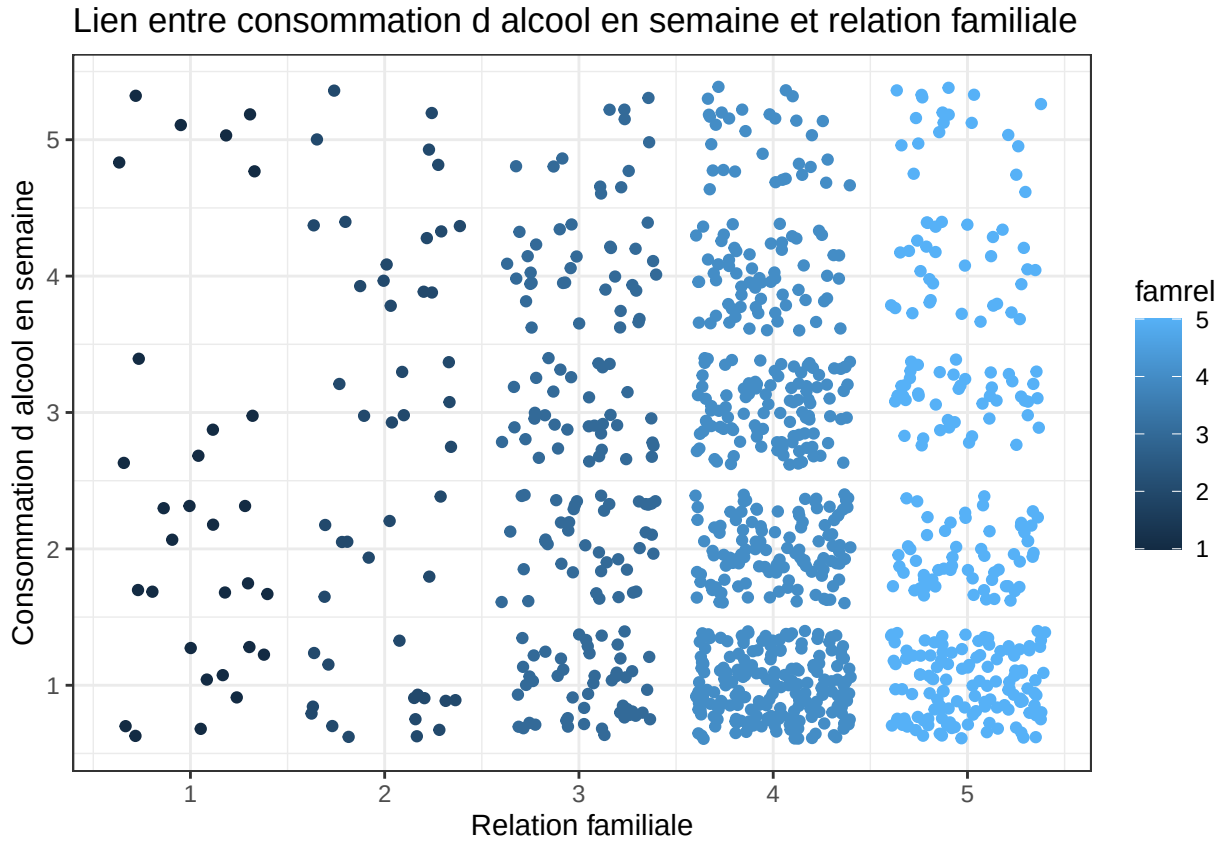
Les élèves ayant comme père et mère tuteur sont identiques au niveau des répartitions de notes, la note médian des élèves ayant un père tuteur est de 12 et celui des mères tuteur et autres sont de 1. Le minimum de tuteur autre est de 7 alors que pour mère et père tuteur celui ci est égale ou inférieur à 5. Par contre, le maximum est de 16 alors que celui du tuteur mère est de 20 et tuteur père est de 18. La boite à moustache

de autre tuteur est plus petit, ce qui signifie que les notes se concentre généralement autour de 9 et 12 alors que les 2 autres sont autour de 10 et 14.

```
table(fulldt$famrel, fulldt$Dalc)
```

```
##
##        1   2   3   4   5
##    1  17   5   4   2   2
##    2  24  15   6   0   2
##    3 113  42   6   6   2
##    4 353 106  32  10  11
##    5 220  28  21   8   9
```

```
ggplot(fulldt)+
  aes(x=famrel, y=Walc, colour=famrel)+
  scale_fill_continuous()+
  geom_jitter()+
  labs(x="Relation familiale", y = "Consommation d alcool en semaine", title="Lien entre consommation d
  theme_bw()
```



Concernant les deux tableaux sur le lien entre relation familiale et consommation d'alcool (un pour la semaine et un pour le week-end). On remarque une tendance générale sur les deux tableaux, étant que les réponses sont distribuées de façon assez uniforme a travers le nuage de points. Si les relations familiales impactaient la consommation d'alcool, on aurait du voir une concentration de point en haut à gauche et en bas a droite des tableaux. Au lieu de ca, les points sont distribues de façon similaire entre ceux ayant les pires relations et ceux ayant les meilleures relations familiales. La seule différence est que l'on remarque que les gens ont tendance a plus consommer en week-end qu'en semaine comme le confirme un de nos tableaux précédents.