

Projet IF36

Koehler Theo, Delhomme Louis, Chivas Matthieu, Rubagotti Lucas

2024-05-05

Dans un premier temps, nous devons nous poser la question de comment allons-nous étudier nos données. En effet, on à notre disposition deux datasets : un sur les élèves suivant le cours de mathématiques (395) et l'autre de portugais (649). Or il est important de noter que la plupart des étudiants de mathématiques suivent également le cours de portugais (382 / 395).

De la sorte, si l'on essaie de joindre les deux datasets, nous auront des doublons, et même si on sait les identifier, il n'est pas pertinent de les fusionner pour les rendre unique. En effet, cela signifierait que l'on doive également fusionner des résultats scolaires de mathématiques et de portugais en une moyenne des deux, alors que les moyennes et répartitions des notes diffèrent entre dans ces deux domaines. Cela ne serait pas être pertinent.

Ainsi, en prenant en compte cette problématique, il nous est offert 3 chemins pour étudier notre dataset :

1. N'étudier que les élèves de mathématiques : des résultats scolaires issus d'une science exacte et évitant les doublons d'étudiants
2. N'étudier que les élèves de portugais : des résultats scolaires qui peuvent différer de ceux d'une science exacte et évitant les doublons
3. Etudier la jointure des étudiants de mathématiques et de portugais : des résultats scolaires mixtes et moins pertinents et considérant les doublons comme deux étudiants à part entière

Il peut être intéressant de répondre aux questions que l'on se pose sous différents angles et de comparer les résultats en fonction de ceux-ci. Nous allons donc considérer ces trois possibilités et les exploiter dans les réponses à nos interrogations.

Pour ceci, on a donc besoin du troisième dataset qui joint les deux autres, on le nommera `all_students`.

```
all_students <- full_join(tableMat,
                           tablePor,
                           by= c("school", "sex", "age", "address", "famsize", "Pstatus", "Medu", "Fedu", "Mjob",
                                 "Fjob", "reason", "guardian", "traveltime", "studytime", "failures", "schoolsups", "famsup", "paine",
                                 "activities", "nursery", "higher", "internet", "romantic", "famrel", "freetime", "goout", "Dalc",
                                 "Walc", "health", "absences", "G1", "G2", "G3"))
```

Question 1: Le fait de sortir souvent avec des amis entraîne-t-il nécessairement une plus grande consommation d'alcool ?

Pour continuer notre étude exploratoire des habitudes de vie des étudiants nous souhaitions analyser une possible influence des sorties entre amis sur la consommation d'alcool. Les jeunes adaptent-ils leur consommation d'alcool en fonction de leurs sorties? Quelle forme prend cette consommation d'alcool sur une semaine? Nous aurions tendance à penser que les personnes voyant souvent leurs amis en dehors des cours seraient plus fréquemment exposés à l'alcool lors de soirées par exemple. De ce fait, nous pourrions nous attendre à constater une forte consommation d'alcool chez les individus sortant beaucoup. Une consommation d'alcool plus importante le week-end qu'en semaine chez les étudiants serait également en accord avec la pensée commune.

Nous allons donc réaliser une analyse exploratoire de notre dataset pour comparer nos pensées avec la réalité des individus étudiés.

Les données qui m'interessent pour répondre à cette question sont :

- **goout** : fréquence de sorties entre amis (numérique : de 1 très faible à 5 très élevé)
- **Dalc** : consommation d'alcool en semaine (numérique : de 1 très faible à 5 très élevé)
- **Walc** : consommation d'alcool le week-end (numérique : de 1 très faible à 5 très élevé)
- **age** : age des individus (numérique)
- **sex** : sexe des individus (nominal : F ou M)

```
#Personnes sortant beaucoup
freq_dalc <- table(tableMat_filtered$Dalc)
total <- sum(freq_dalc)
freq_dalc <- (freq_dalc / total) * 100

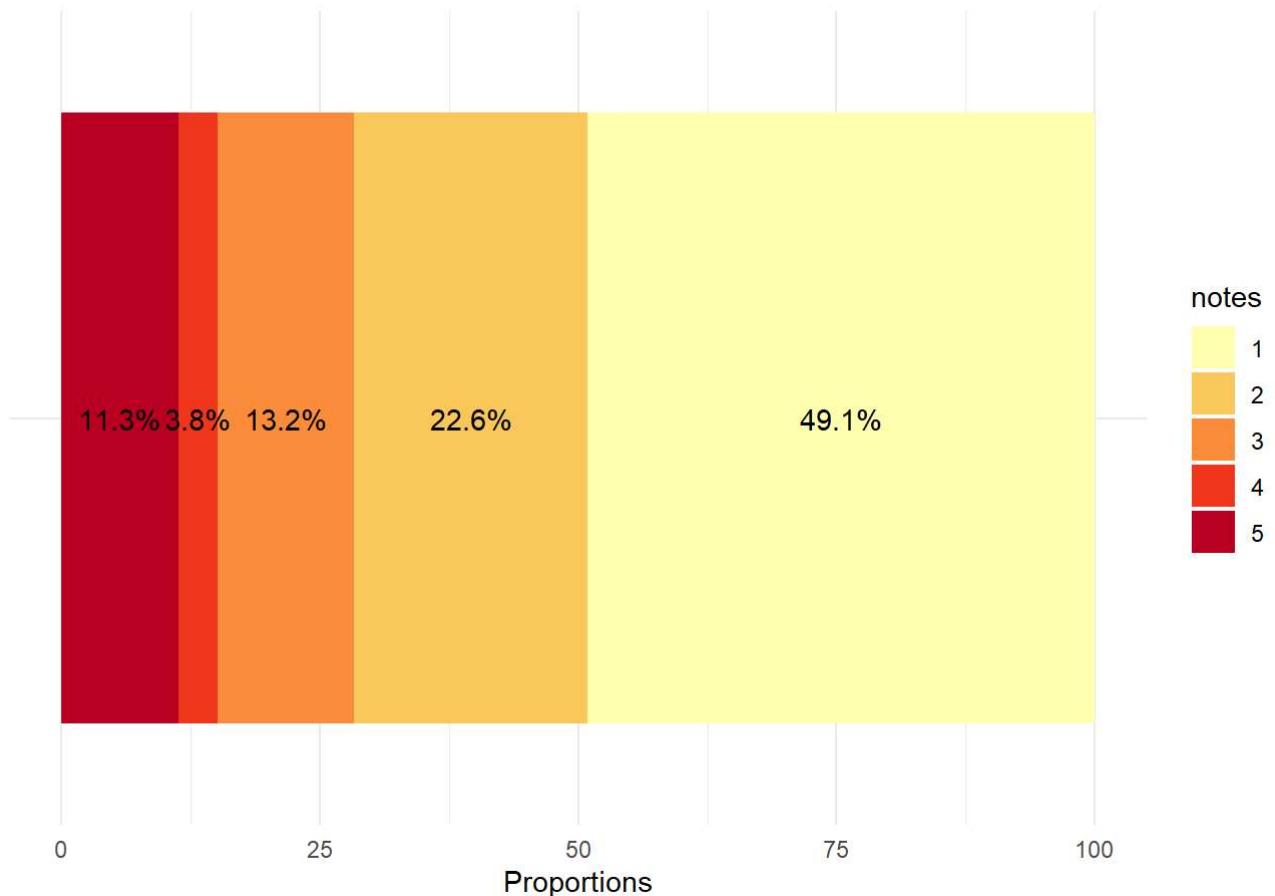
freq_walc <- table(tableMat_filtered$Walc)
total_walc <- sum(freq_walc)
freq_walc <- (freq_walc / total_walc) * 100

notes <- c(1, 2, 3, 4, 5)
df_dalc <- data.frame(notes = factor(notes), frequences = freq_dalc)
df_walc <- data.frame(notes = factor(notes), frequences = freq_walc)

ggplot(df_dalc, aes(x = "", y = freq_dalc, fill = notes, label = paste0(round(freq_dalc, 1),
"%")) +
  geom_bar(stat = "identity") +
  geom_text(position = position_stack(vjust = 0.5)) + # Ajouter les pourcentages centrés
  coord_flip() +
  labs(x = "", y = "Proportions", title = "Notes de consommation d'alcool en semaine chez les
personnes sortant souvent") +
  theme_minimal() +
  scale_fill_brewer(palette = "YlOrRd")

## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```

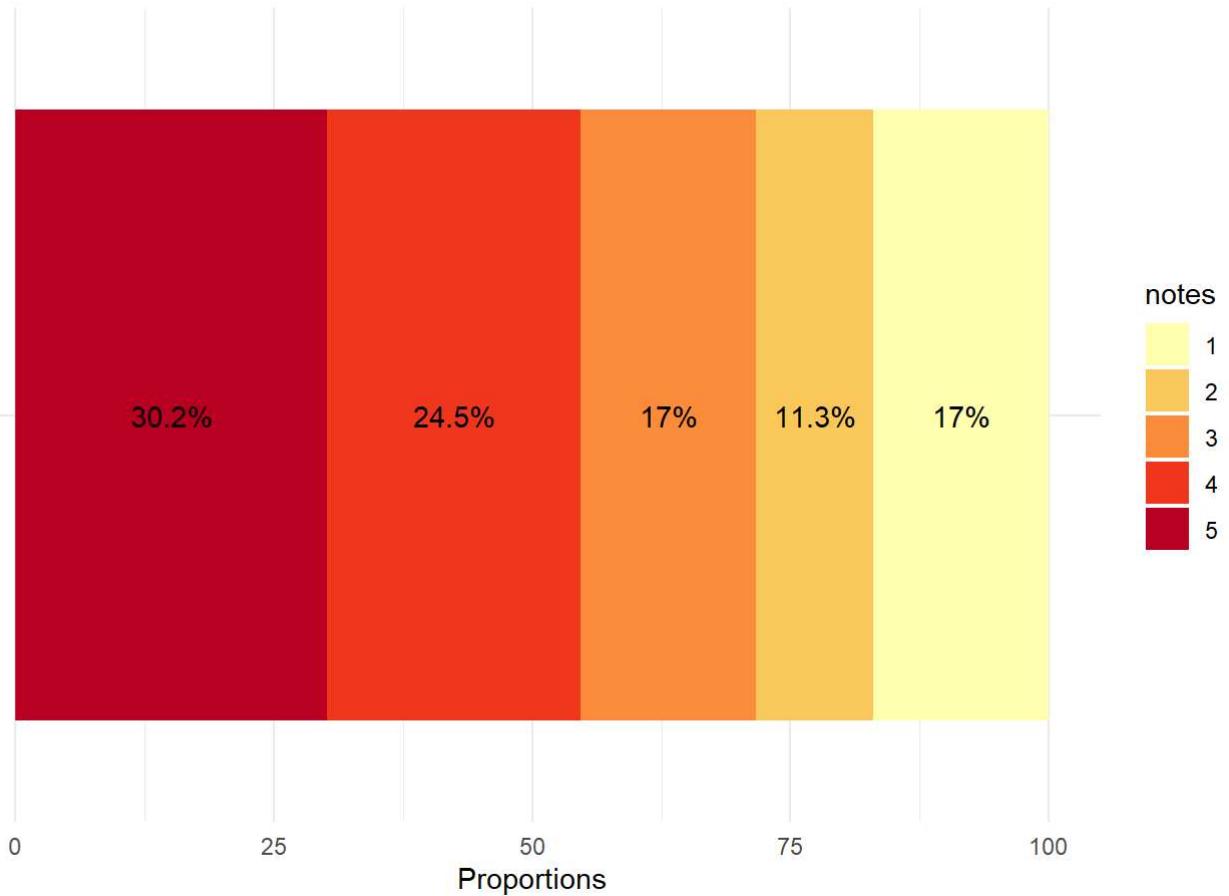
Notes de consommation d'alcool en semaine chez les personnes sortant souvent



```
ggplot(df_walc, aes(x = "", y = freq_walc, fill = notes, label = paste0(round(freq_walc, 1), "%"))) +  
  geom_bar(stat = "identity") +  
  geom_text(position = position_stack(vjust = 0.5)) + # Ajouter les pourcentages centrés  
  coord_flip() +  
  labs(x = "", y = "Proportions", title = "Notes de consommation d'alcool le week-end chez les personnes sortant souvent") +  
  theme_minimal() +  
  scale_fill_brewer(palette = "YlOrRd")
```

```
## Don't know how to automatically pick scale for object of type <table>.  
## Defaulting to continuous.
```

Notes de consommation d'alcool le week-end chez les personnes sortant souvent



```
#personnes ne sortant pas beaucoup

tableMat_filtered_inf <- tableMat %>% filter(goout < 4)

freq_dalc_inf <- table(tableMat_filtered_inf$Dalc)
total_inf <- sum(freq_dalc_inf)
freq_dalc_inf <- (freq_dalc_inf / total_inf) * 100

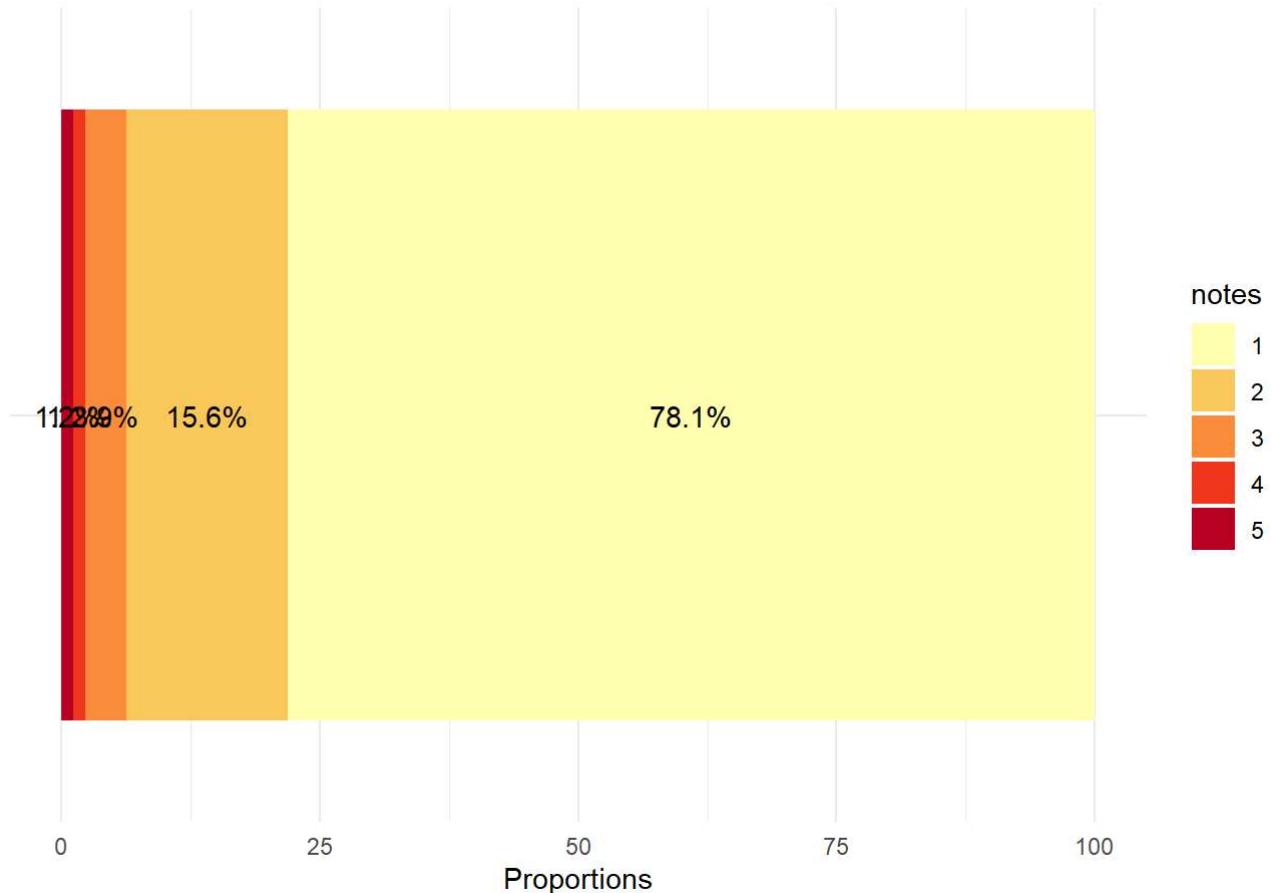
freq_walc_inf <- table(tableMat_filtered_inf$Walc)
total_walc_inf <- sum(freq_walc_inf)
freq_walc_inf <- (freq_walc_inf / total_walc_inf) * 100

df_dalc_inf <- data.frame(notes = factor(notes), frequences = freq_dalc_inf)
df_walc_inf <- data.frame(notes = factor(notes), frequences = freq_walc_inf)

ggplot(df_dalc_inf, aes(x = "", y = freq_dalc_inf, fill = notes, label = paste0(round(freq_da
lc_inf, 1), "%))) +
  geom_bar(stat = "identity") +
  geom_text(position = position_stack(vjust = 0.5)) + # Ajouter les pourcentages centrés
  coord_flip() +
  labs(x = "", y = "Proportions", title = "Notes de consommation d'alcool en semaine chez les
  personnes sortant peu") +
  theme_minimal() +
  scale_fill_brewer(palette = "YlOrRd")

## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```

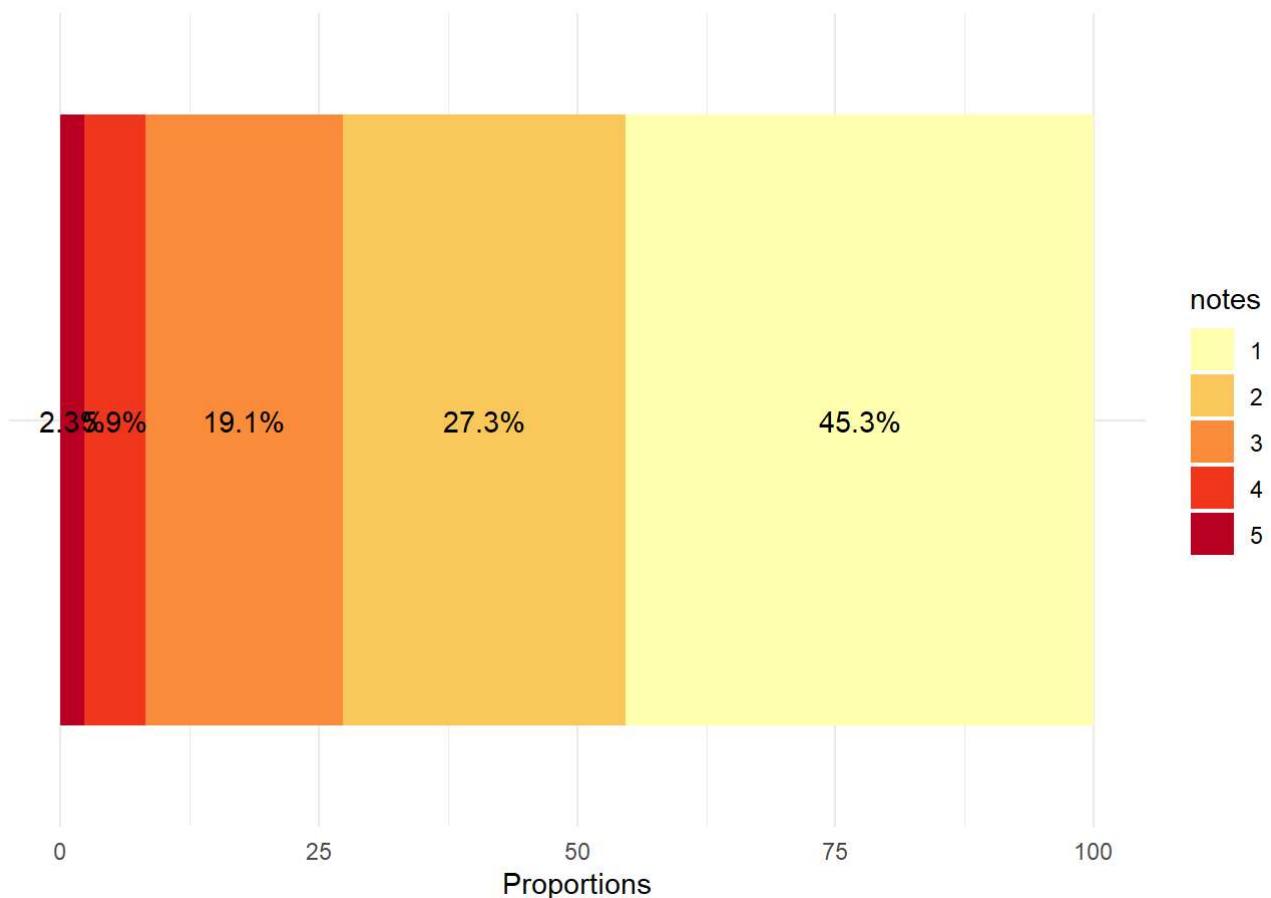
Notes de consommation d'alcool en semaine chez les personnes sortant peu



```
ggplot(df_walc_inf, aes(x = "", y = freq_walc_inf, fill = notes, label = paste0(round(freq_walc_inf, 1), "%"))) +  
  geom_bar(stat = "identity") +  
  geom_text(position = position_stack(vjust = 0.5)) + # Ajouter les pourcentages centrés  
  coord_flip() +  
  labs(x = "", y = "Proportions", title = "Notes de consommation d'alcool le week-end chez les personnes sortant peu") +  
  theme_minimal() +  
  scale_fill_brewer(palette = "YlOrRd")
```

```
## Don't know how to automatically pick scale for object of type <table>.  
## Defaulting to continuous.
```

Notes de consommation d'alcool le week-end chez les personnes sortant peu



Le premier objectif de notre étude est de visualiser la répartition des personnes sortant souvent entre amis en fonction de leur consommation d'alcool.

Nous avons donc, d'une part, uniquement gardé les individus ayant évalué leur fréquence de sortie à 5 sur 5. D'autre part, nous avons gardé ceux s'étant attribué une note de 1, 2 ou 3. Ces individus ont également évalué leur consommation d'alcool en semaine et le week-end.

Après avoir réalisé ces visualisations plusieurs choses sont constatables:

En semaine, la majorité (71.7%) des étudiants sortant souvent entre amis ne boit que très peu d'alcool tandis qu'un quart d'entre eux consomme fréquemment ou très fréquemment de l'alcool. Le week-end, la tendance s'inverse. En effet, 54.7% consomment fréquemment ou très fréquemment de l'alcool et seulement 28.3% que rarement ou très rarement. Nous constatons qu'un quart des étudiants sortant souvent consomment occasionnellement de l'alcool que ce soit en semaine ou le week-end.

Nous savons que les étudiants organisent souvent des sorties entre eux le week-end et moins en semaine en raison de leurs cours. De ce fait, le week-end devient le moment de la semaine où ils peuvent s'amuser sans avoir peur des conséquences et donc consommer plus d'alcool.

D'autre part, nous constatons clairement que les étudiants sortant rarement entre amis consomment beaucoup moins d'alcool que ceux sortant souvent et ce, en semaine mais aussi le week-end.

Il semble donc clair que le fait de sortir souvent avec des amis entraînerait une plus grande consommation d'alcool.

Pour approfondir encore notre étude nous pouvons tenter d'observer une divergence ou une similarité dans la consommation d'alcool chez les étudiants sortant souvent en fonction du sexe.

Analyse de la consommation d'alcool chez les étudiants sortant

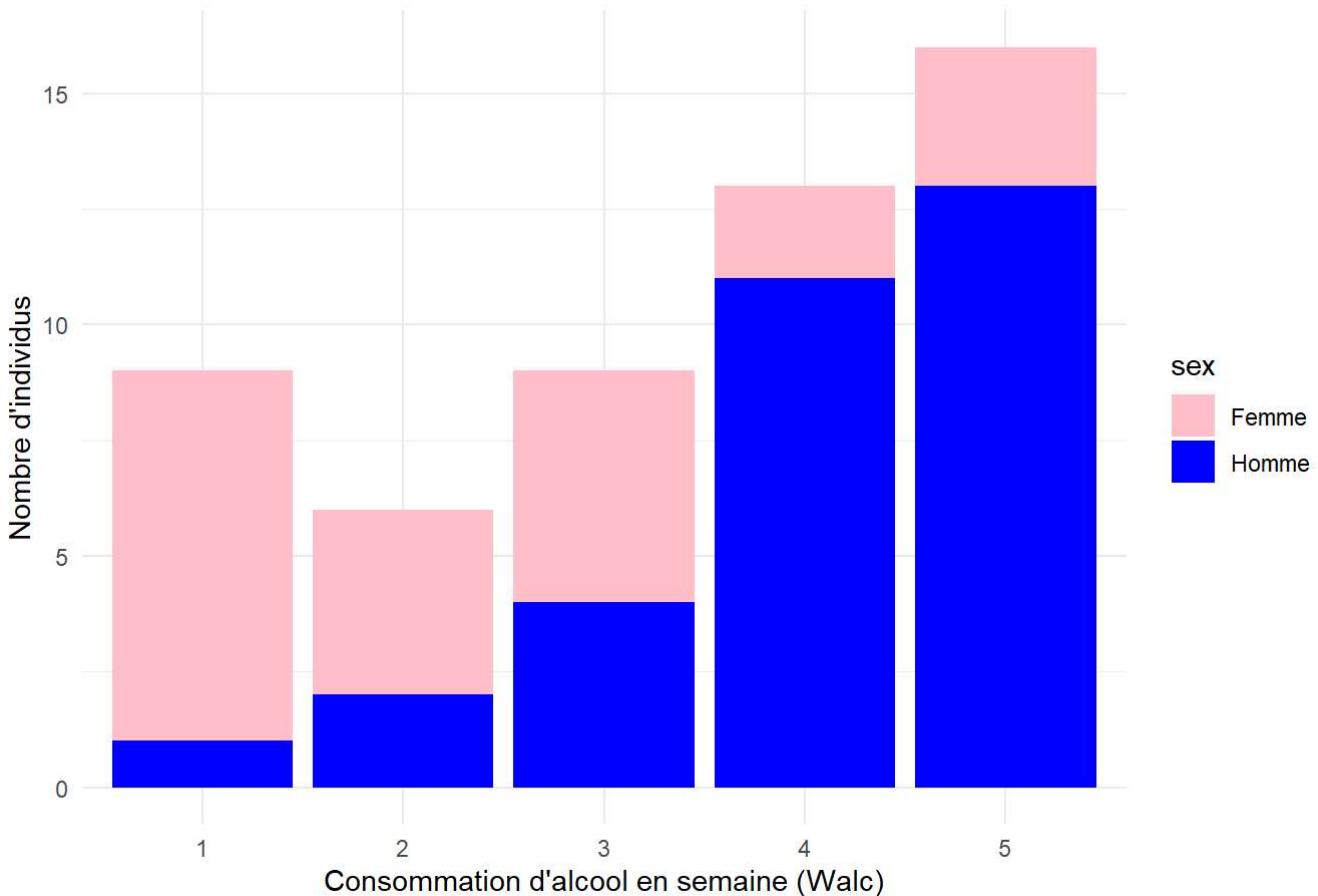
fréquemment en fonction de leur sexe

```
summary_data <- tableMat_filtered %>%
  group_by(Walc, sex) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Walc'. You can override using the
## `.groups` argument.
```

```
# Créer le graphique à barres empilées
ggplot(summary_data, aes(x = factor(Walc), y = count, fill = sex)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Consommation d'alcool en semaine (Walc)", y = "Nombre d'individus", title = "Consommation d'alcool le week-end chez les personnes sortant souvent par sexe") +
  scale_fill_manual(values = c("pink", "blue"), labels = c("Femme", "Homme")) + # Définir les couleurs et les labels pour le sexe avec l'ordre correct
  theme_minimal()
```

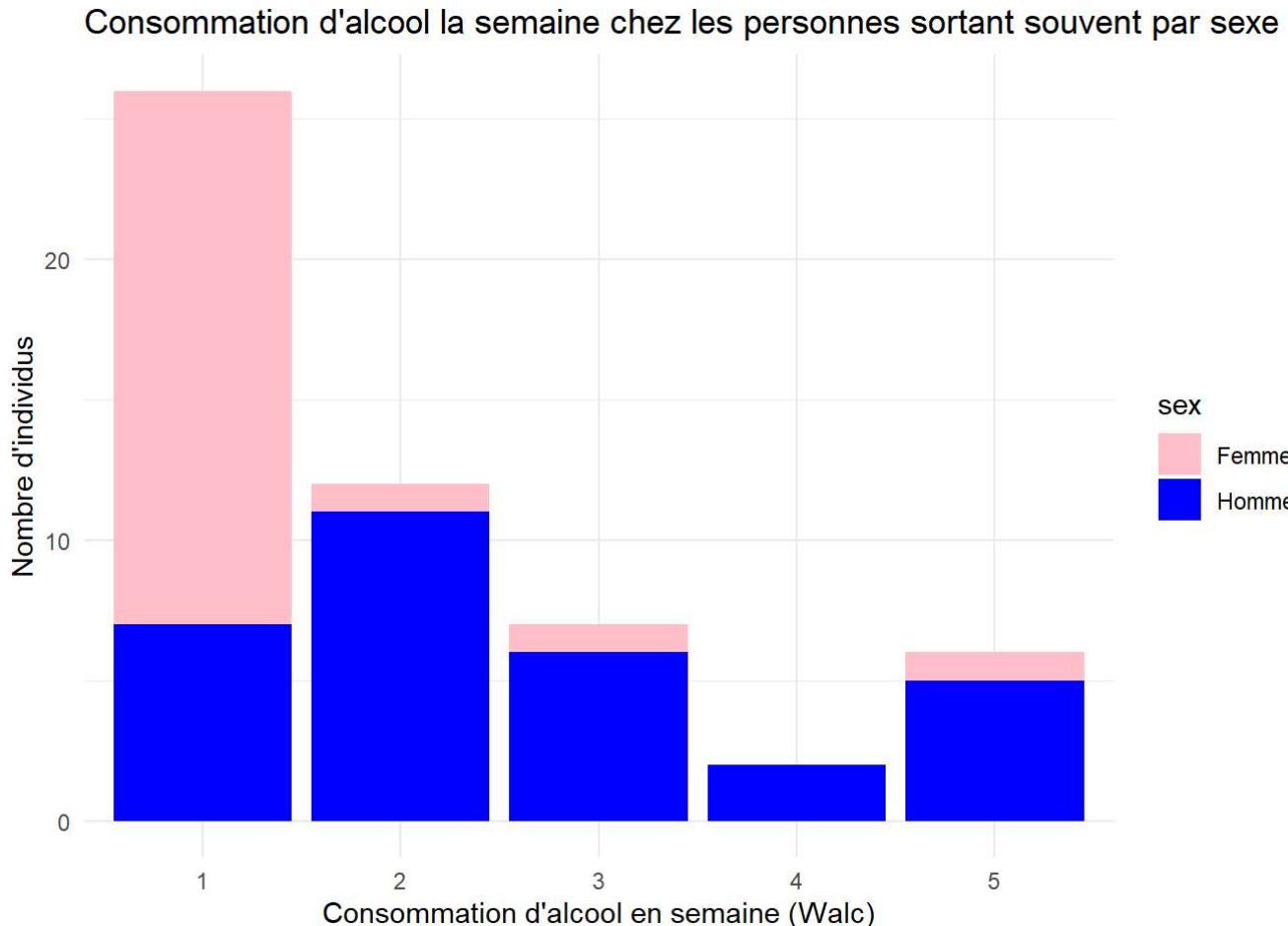
Consommation d'alcool le week-end chez les personnes sortant souvent par sexe



```
summary_data <- tableMat_filtered %>%
  group_by(Dalc, sex) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Dalc'. You can override using the
## `.groups` argument.
```

```
# Créer Le graphique à barres empilées
ggplot(summary_data, aes(x = factor(Dalc), y = count, fill = sex)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Consommation d'alcool en semaine (Walc)", y = "Nombre d'individus", title = "Consommation d'alcool la semaine chez les personnes sortant souvent par sexe") +
  scale_fill_manual(values = c("pink", "blue"), labels = c("Femme", "Homme")) + # Définir les couleurs et les labels pour le sexe avec l'ordre correct
  theme_minimal()
```



Analysons les visualisations obtenues.

En semaine, les étudiants sortant très fréquemment et consommant très rarement de l'alcool sont en majorité des femmes. La quasi totalité des femmes sortant fréquemment entre amis boivent très rarement. Les hommes ont quant à eux une consommation d'alcool plus répartie et uniforme. En effet, il y autant d'hommes qui en boivent très rarement que très fréquemment.

Le weekend, les hommes et les femmes consomment de l'alcool de manière beaucoup plus fréquente comme montré précédemment.

Les femmes qui sortent beaucoup entre amis ont donc tendances à boire très peu en semaine mais beaucoup plus fréquemment le week-end. Les hommes quant à eux ont en semaine une consomamtion d'alcool variant entre les individus mais s'accordent globalement le week-end pour boire très fréquemment.

Finalement, il existe une différence notable de comportement face à l'alcool entre les hommes et les femmes en semaine. Cependant, le week-end, les deux sexes ont tendance à consommer de manière identique.

Il est également possible de réaliser cette même étude en se penchant sur l'âge des individus plutôt que sur leur sexe.

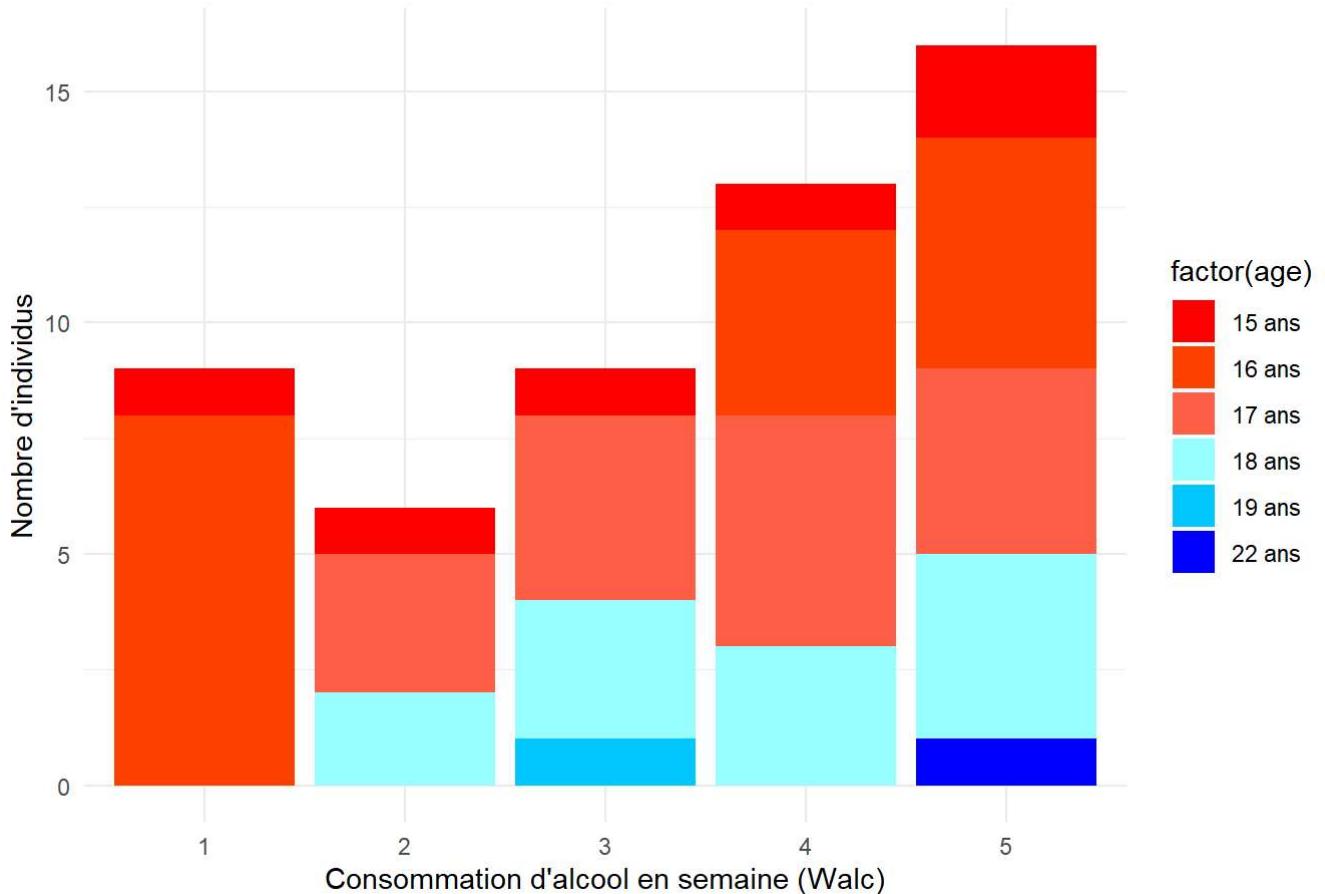
Analyse de la consommation d'alcool chez les étudiants sortant fréquemment en fonction de leur âge

```
summary_data <- tableMat_filtered %>%
  group_by(Walc, age) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Walc'. You can override using the
## `.groups` argument.
```

```
# Créer le graphique à barres empilées
ggplot(summary_data, aes(x = factor(Walc), y = count, fill = factor(age))) +
  geom_bar(stat = "identity", position = "stack") +
  scale_fill_manual(values = c("#FF0000", "#FF4500", "#FF6347", "#99FFFF", "#00ccFF", "#0000FF"),
                    labels = c("15 ans", "16 ans", "17 ans", "18 ans", "19 ans", "22 ans")) +
  labs(x = "Consommation d'alcool en semaine (Walc)", y = "Nombre d'individus", title = "Consommation d'alcool la semaine chez les personnes sortant souvent par sexe") +
  theme_minimal()
```

Consommation d'alcool la semaine chez les personnes sortant souvent par sexe

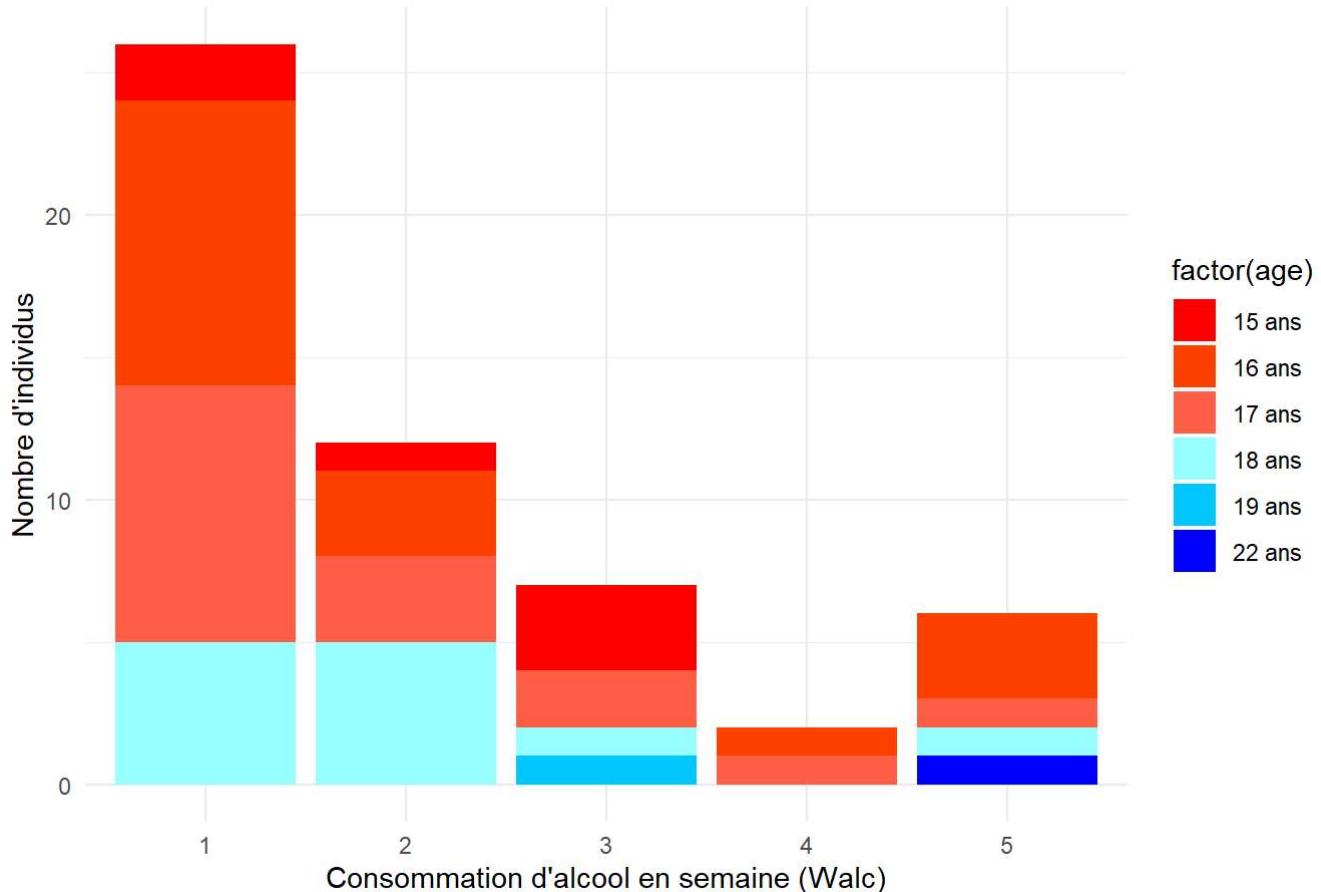


```
summary_data <- tableMat_filtered %>%
  group_by(Dalc, age) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Dalc'. You can override using the
## `.` argument.
```

```
# Créer le graphique à barres empilées
ggplot(summary_data, aes(x = factor(Dalc), y = count, fill = factor(age))) +
  geom_bar(stat = "identity", position = "stack") +
  scale_fill_manual(values = c("#FF0000", "#FF4500", "#FF6347", "#99FFFF", "#00ccFF", "#0000FF"),
                     labels = c("15 ans", "16 ans", "17 ans", "18 ans", "19 ans", "22 ans")) +
  labs(x = "Consommation d'alcool en semaine (Walc)", y = "Nombre d'individus", title = "Consommation d'alcool la semaine chez les personnes sortant souvent par sexe") +
  theme_minimal()
```

Consommation d'alcool la semaine chez les personnes sortant souvent par sexe



Grâce aux visualisations réalisées nous pouvons observer plusieurs phénomènes.

En semaine, la quasi totalité des étudiants majeurs sortant souvent entre amis boivent assez peu fréquemment de l'alcool. Cependant, ils ont tendance à tous boire plus souvent le week-end. Cette différence est uniquement dû aux étudiants ayant 18 ans car ceux ayant 19 ou 22 ans ne change pas leur fréquence de consommation en fonction du moment de la semaine.

Les étudiants mineurs sont intéressants à étudier. En effet, malgré qu'ils soient mineurs et n'ont donc pas le droit légal de boire de l'alcool, une partie non négligeable d'entre eux admet en consommer de manière plus ou moins fréquente. En semaine, la grande majorité des étudiants admettant consommer fréquemment de l'alcool est constituée de mineurs. Le week-end ils sont encore plus à en boire fréquemment.

Cependant, le weekend, les personnes ne buvant pas d'alcool sont uniquement des mineurs.

On constate donc une différence de comportement face à l'alcool en fonction de l'âge des individus sortant fréquemment entre amis. Cela pourrait être causé par le fait que les mineurs sont possiblement plus influençables et ont tendance à imiter les plus grands, entraînant ainsi une consommation d'alcool toujours plus précoce.

Question2 - les sportifs ont-ils de meilleurs résultats scolaires que les autres ?

Il nous est souvent raconté que faire du sport permet un meilleur apprentissage, une meilleure concentration et un meilleur développement. Suivant cette logique, il nous est expliqué que par extension, le sport permet d'obtenir de meilleurs résultats scolaires. Mais est-ce la vérité ?

Nous allons donc essayer de résoudre cette énigme en nous posant la question suivante : les sportifs ont-ils de meilleurs résultats scolaires que les autres ?

Pour répondre à cette question il faut donc tout d'abord identifier ce qu'est un sportif. Il nous a semblé concret d'identifier un sportif comme ayant un niveau de santé supérieur ou égal à 3/5 et participant à des activités extra-scolaires.

```
#Question 1 :  
#Faire du sport permet-il vraiment d'avoir de meilleurs résultats scolaires ?  
all_students2 <- all_students  
tablePor2 <- tablePor  
tableMat2 <- tableMat  
  
# Créer la variable sportif dans les 3 datasets  
all_students2$sportif <- ifelse(all_students2$health >= 3 & all_students2$activities == "yes", TRUE, FALSE)  
tablePor2$sportif <- ifelse(tablePor2$health >= 3 & tablePor2$activities == "yes", TRUE, FALSE)  
tableMat2$sportif <- ifelse(tableMat2$health >= 3 & tableMat2$activities == "yes", TRUE, FALSE)
```

Egalement, afin de comparer les résultats scolaires des étudiants, il nous faut une variable unique indiquant ces résultats. On a donc simplement créé la variable "mean_results", étant la moyenne des résultats scolaires des étudiants sur les 3 examens du semestre. Ne connaissant pas le poids de chacun des examens, nous avons posé une pondération linéaire : chacun des examens compte pour autant.

```
#Question 1 :  
#Faire du sport permet-il vraiment d'avoir de meilleurs résultats scolaires ?  
  
# Calculer la moyenne des variables G1, G2 et G3 dans les 3 datasets  
mean_results_global <- rowMeans(all_students[, c("G1", "G2", "G3")])  
mean_results_mat <- rowMeans(tableMat2[, c("G1", "G2", "G3")])  
mean_results_por <- rowMeans(tablePor2[, c("G1", "G2", "G3")])
```

Nous devrions maintenant avoir les clés en main pour essayer de répondre à notre interrogation.

Dans un premier temps, nous allons essayer d'étudier la question en regardant le dataset qui joint les étudiants de portugais et de mathématiques : all_students.

On doit donc essayer de mettre en évidence les résultats scolaires en fonction de deux facteurs : la santé et la sportivité.

Tout d'abord, il nous a semblé être le plus représentatif de poser les résultats scolaires en ordonnée et la santé en abscisse. On aura donc 5 colonnes indiquant les résultats scolaires en fonction de la santé, qui seront facilement comparables.

Ensuite, pour ce qui est de la sportivité, représentons simplement d'une autre couleur les étudiants considérés comme "sportifs".

Enfin, traçons la moyenne des résultats scolaires pour "sportifs" et "non-sportifs" afin de voir s'il y a en effet une différence de niveau.

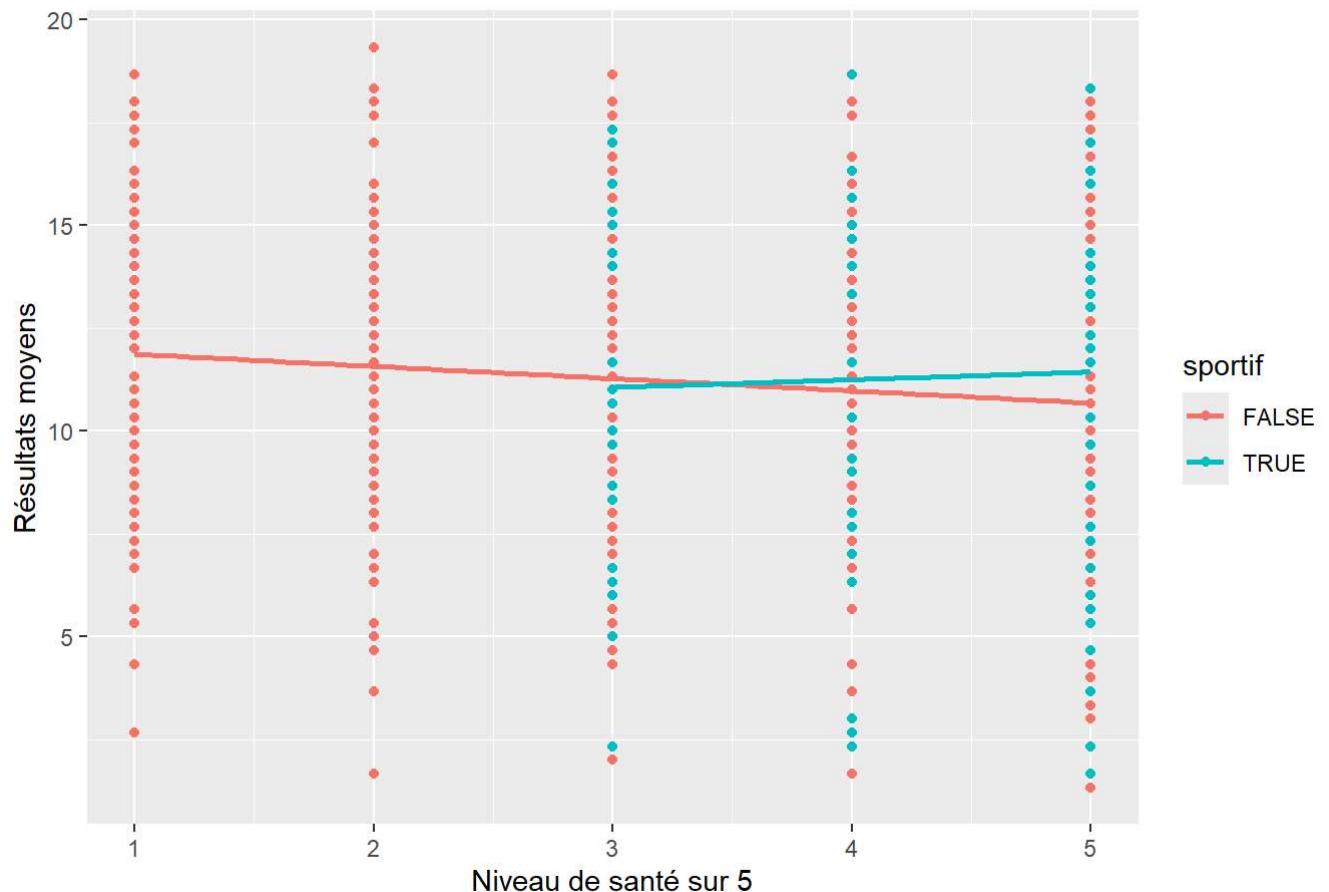
#Question 1 :

#Faire du sport permet-il vraiment d'avoir de meilleurs résultats scolaires ?

```
# Visualisation des résultats scolaires moyens en fonction de La santé et de La sportivité dans la jointure des deux classes de portugais et de mathématiques
ggplot(data = all_students2, aes(x = health, y = mean_results_global)) +
  geom_point(aes(color = sportif)) +
  geom_smooth(method = "lm", se = FALSE, aes(color = sportif)) +
  labs(x = "Niveau de santé sur 5", y = "Résultats moyens") +
  ggtitle("Résultats scolaires moyens en fonction de la santé et de la sportivité (GLOBAL)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Résultats scolaires moyens en fonction de la santé et de la sportivité (GLOBAL)



On remarque ainsi que si l'on regarde tous les étudiants, la santé seule ne semble pas jouer sur les résultats scolaires : les résultats moyens sont les mêmes quelque soit le niveau de santé des étudiants (si on combine sportifs + non-sportifs). La seule différence que l'on observe est celle entre les étudiants en bonne santé sportifs et les étudiants en bonne santé non-sportifs. Néanmoins même si cette différence est bien présente, elle n'est pas non plus très significative.

On n'observe donc pas de résultats flagrants. On peut remettre aussi en question nos résultats puisque nous comparons les résultats scolaires globaux en regardant du même oeil des résultats de mathématiques que ceux de portugais.

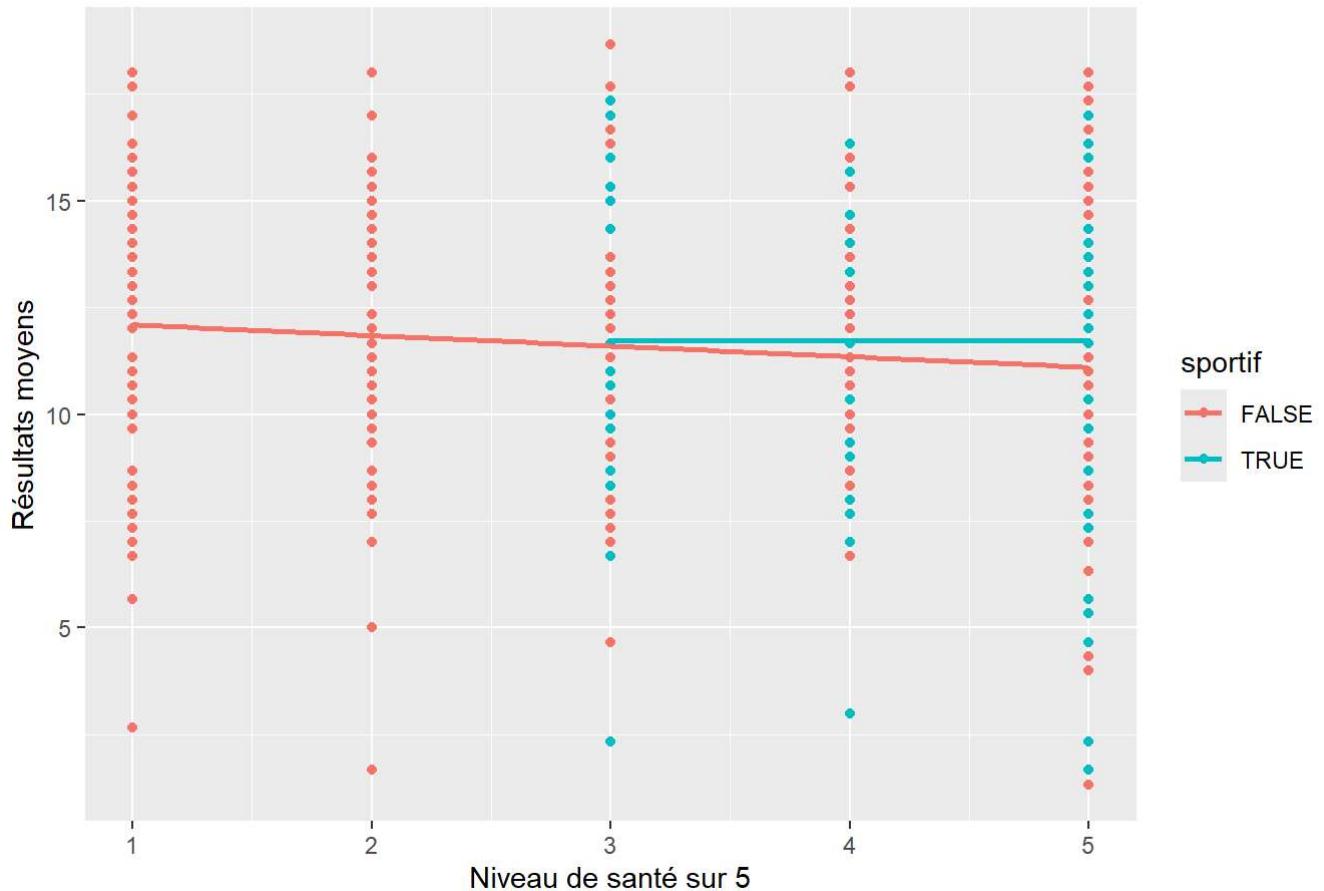
Il semble donc nécessaire de regarder ce que l'on pourrait obtenir en ne regardant que individuellement les classes de mathématiques et de portugais et en gardant le même modèle de visualisation.

```
#Question 1 :
#Faire du sport permet-il vraiment d'avoir de meilleurs résultats scolaires ?

# Visualisation des résultats scolaires moyens en fonction de la santé et de la sportivité dans la classe de portugais
ggplot(data = tablePor2, aes(x = health, y = mean_results_por)) +
  geom_point(aes(color = sportif)) +
  geom_smooth(method = "lm", se = FALSE, aes(color = sportif)) +
  labs(x = "Niveau de santé sur 5", y = "Résultats moyens") +
  ggtitle("Résultats scolaires moyens en fonction de la santé et de la sportivité (Portugais)")

## `geom_smooth()` using formula = 'y ~ x'
```

Résultats scolaires moyens en fonction de la santé et de la sportivité (Portugais)



#Question 1 :

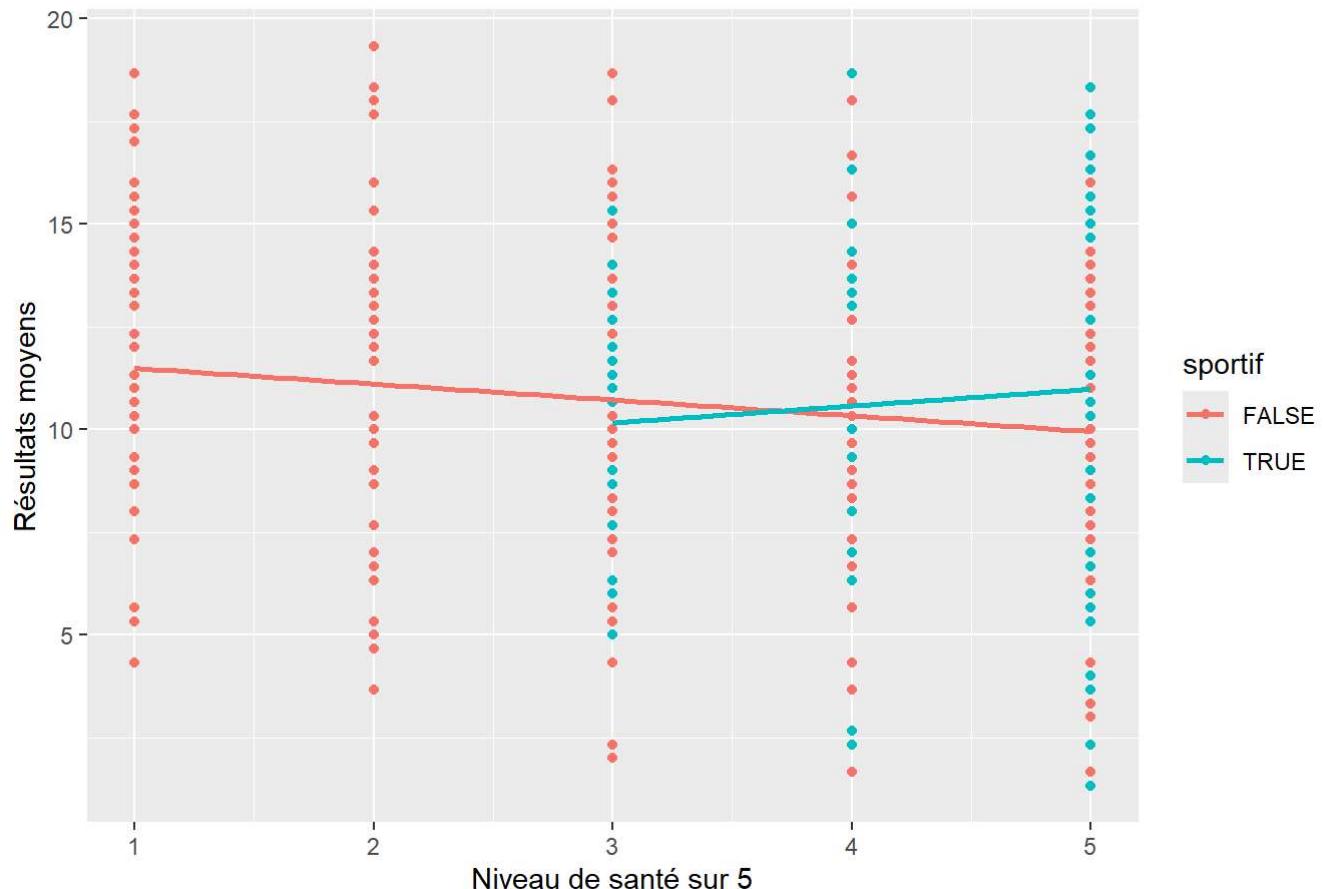
#Faire du sport permet-il vraiment d''avoir de meilleurs résultats scolaires ?

Visualisation des résultats scolaires moyens en fonction de la santé et de la sportivité dans la classe de mathématiques

```
ggplot(data = tableMat2, aes(x = health, y = mean_results_mat)) +  
  geom_point(aes(color = sportif)) +  
  geom_smooth(method = "lm", se = FALSE, aes(color = sportif)) +  
  labs(x = "Niveau de santé sur 5", y = "Résultats moyens") +  
  ggtitle("Résultats scolaires moyens en fonction de la santé et de la sportivité (Maths)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Résultats scolaires moyens en fonction de la santé et de la sportivité (Maths)



En analysant les résultats issus des classes de portugais de mathématiques nous retombons sur des conclusions plutôt similaires : il semblerait que plus les étudiants sont en bonne santé, plus le fait d'être sportif semble améliorer les résultats scolaires.

En revanche, cette amélioration peut-être remise en cause. Elle n'est que minime, et les résultats moyens des sportifs restent similaires à ceux obtenus par les étudiants en mauvaise santé. Cela montre-t-il que le sport n'apporte rien ?

Mais plus précisément, le problème ne vient-il pas simplement de notre définition de ce qu'est un sportif ? Notre approche de cette définition est finalement simpliste : sportif = santé + activités, mais un étudiant en bonne santé pratiquant des activités extra-scolaires peut très bien ne pas être sportif du tout. Il peut être passionné par l'art et de ce fait pratiquer des activités au delà de ses études.

Il faut donc essayer de trouver un angle d'approche à notre étude. Simplifions la problématique, évitons des erreurs d'interprétations et regardons directement si les étudiants pratiquent des activités extra-scolaires, et si l'on peut lier ceci au reste des données.

Nous allons à nouveau étudier la classe de portugais avec le même modèle puisque c'est celle qui nous donné les meilleurs résultats précédemment. Changeons simplement la variable "sportif" en la variable "goes_out" signifiant si oui ou non un étudiant pratique des activités extra-scolaires.

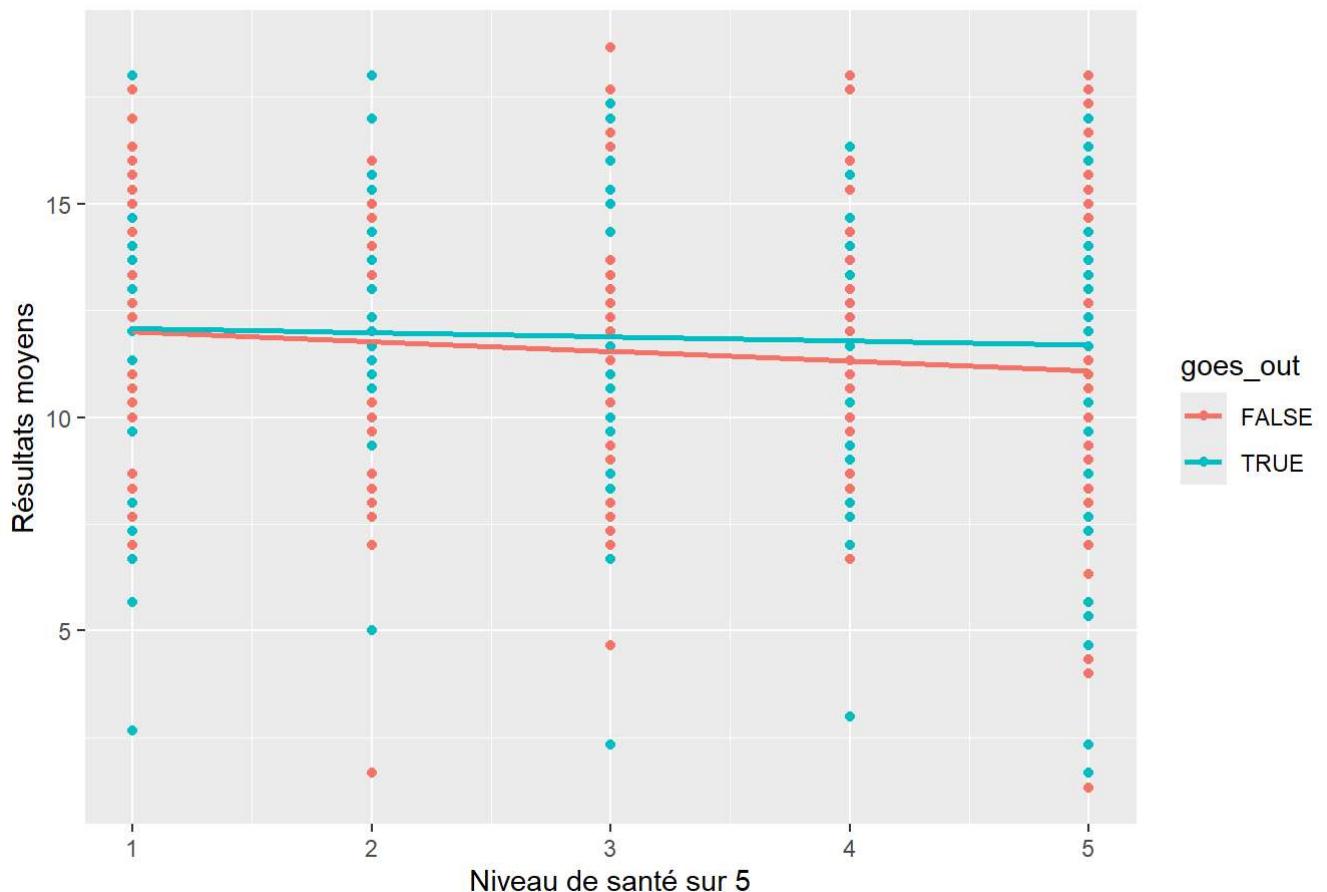
```
#Question 1 :
#Faire du sport permet-il vraiment d' avoir de meilleurs résultats scolaires ?

# Identifier les étudiants qui pratiquent des activités extra-scolaires dans une nouvelle variable pour plus de lisibilité
tablePor2$goes_out <- ifelse(tablePor2$activities == "yes", TRUE, FALSE)

# Visualisation des résultats scolaires moyens en fonction de La santé et des activités dans la classe de portugais
ggplot(data = tablePor2, aes(x = health, y = mean_results_por)) +
  geom_point(aes(color = goes_out)) +
  geom_smooth(method = "lm", se = FALSE, aes(color = goes_out)) +
  labs(x = "Niveau de santé sur 5", y = "Résultats moyens") +
  ggtitle("Résultats scolaires moyens en fonction de la santé et des activités (Portugais)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Résultats scolaires moyens en fonction de la santé et des activités (Portugais)



Ainsi, avec cette approche, nous observons tout de suite des résultats différents. On remarque aisément que plus un étudiant a un bon niveau de santé, plus il lui sera bénéfique par rapport aux autres de faire des activités extra-scolaires.

En revanche, les bénéfices ne sont pas vraiment existants. Le niveau n'évolue pas positivement, il se maintient juste. La pratique des activités extra-scolaires semble seulement permettre de maintenir un niveau scolaire équivalent quelque soit la santé, là où¹ la non-pratique semble le faire régresser. On peut tout de même dire qu'il faut donc pratiquer des activités extra-scolaires pour en moyenne avoir de meilleurs résultats que les autres.

[Hypothèses]

Pourquoi la différence activités / pas d'activités s'accentue avec la santé ?

Cela pourrait s'expliquer par le fait qu'utiliser son énergie dans des activités extra-scolaires peut permettre de majoritairement s'épanouir quand on est en bonne santé et peut-être de plus se fatiguer lorsque l'on est en mauvaise santé.

Pourquoi n'a-t-on pas des résultats si peu convaincants ?

Peut être que dans notre jeu de données, ceux pratiquant des activités extra-scolaires n'ont pas été tant meilleurs que les autres et qu'il nous faudrait plus de données pour avoir de meilleures résultats.

En conclusion, avec notre jeu de données, il n'est pas vraiment possible de répondre à la question "Faire du sport permet-il vraiment d'avoir de meilleurs résultats scolaires ?". En revanche, essayer de répondre à cette question a permis de mettre en évidence des différences scolaires en fonctions des niveaux de santé et de la pratique d'activités extra-scolaires. Cela a finalement permis de répondre à une question tout autre qui pourrait être "Est-il bon pour les études de pratiquer des activités extra-scolaires quelque soit sa santé ?". Question à laquelle la réponse est oui, mais il faut adapter ses activités à son niveau de santé afin de maximiser ses résultats scolaires.

Question 3: Existe-t-il une corrélation entre les étudiants considérés comme geek et leurs notes en mathématiques?

Partie exploratoire

Nous avons ce cliché en tête qui dit que les geeks sont des personnes qui ont de bonnes notes en mathématiques. Nous allons essayer de voir si cela est vrai en analysant un dataset d'étudiants.

Tout d'abord nous allons définir ce qu'est un geek. Pour cela nous allons considérer un étudiant comme geek si :

- **activities** : activités extra-scolaires sont égale à "no"
- **internet** : accès à Internet à domicile est égale à "yes"
- **goout** : sorties avec des amis est inférieur à 2 Ces critères de sélection sont parfaitement arbitraires et ne représentent en aucun cas la réalité.

Ensuite nous allons regarder les notes en mathématiques des élèves et nous verrons si corrélation il existe.

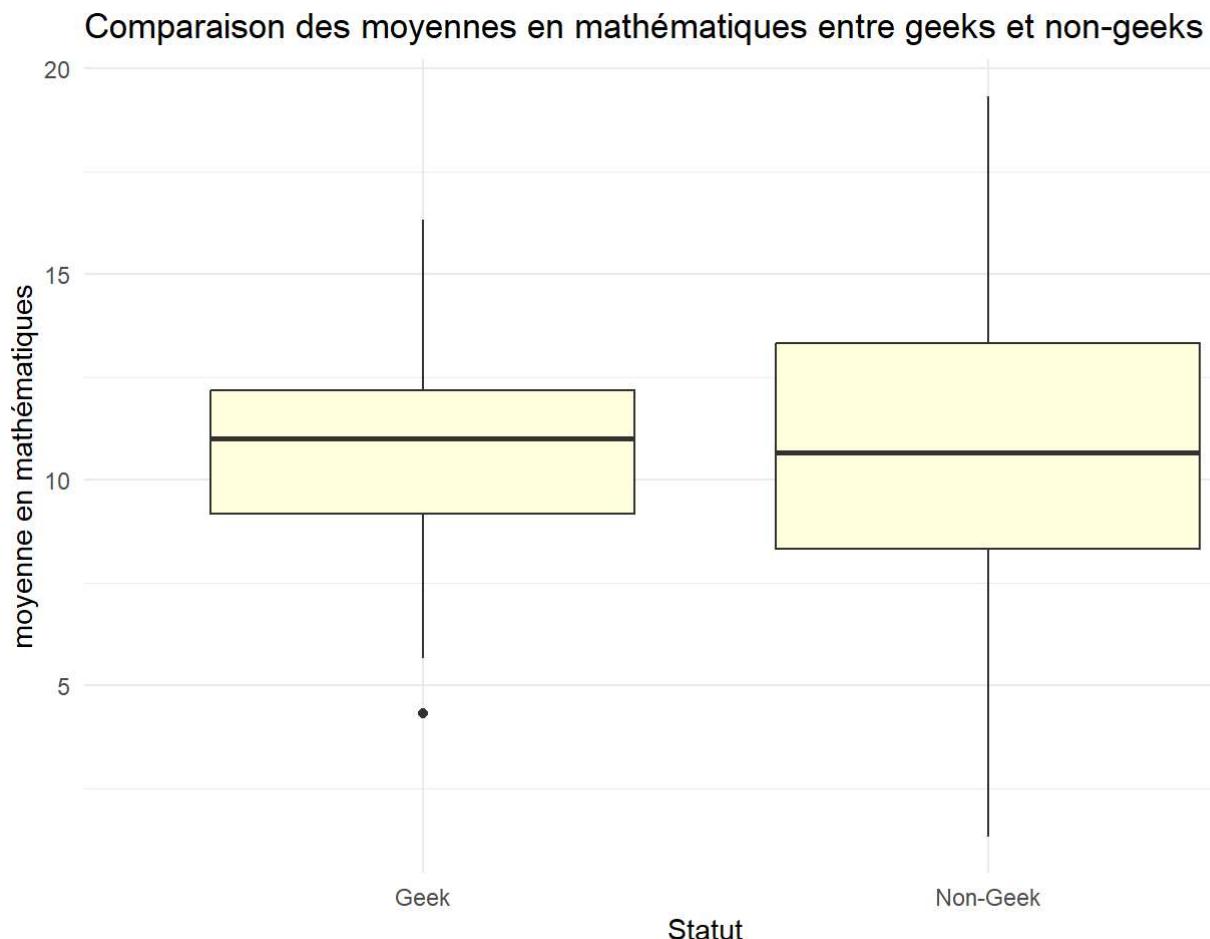
Les notes de maths sont les variables :

- **G1** : note du premier trimestre
- **G2** : note du deuxième trimestre
- **G3** : note du troisième trimestre

```
# Définir les critères pour considérer un étudiant comme geek
data <- tableMat %>%
  mutate(geek = ifelse(activities == "no" &
                      internet == "yes" &
                      goout < 2, "Geek", "Non-Geek"))

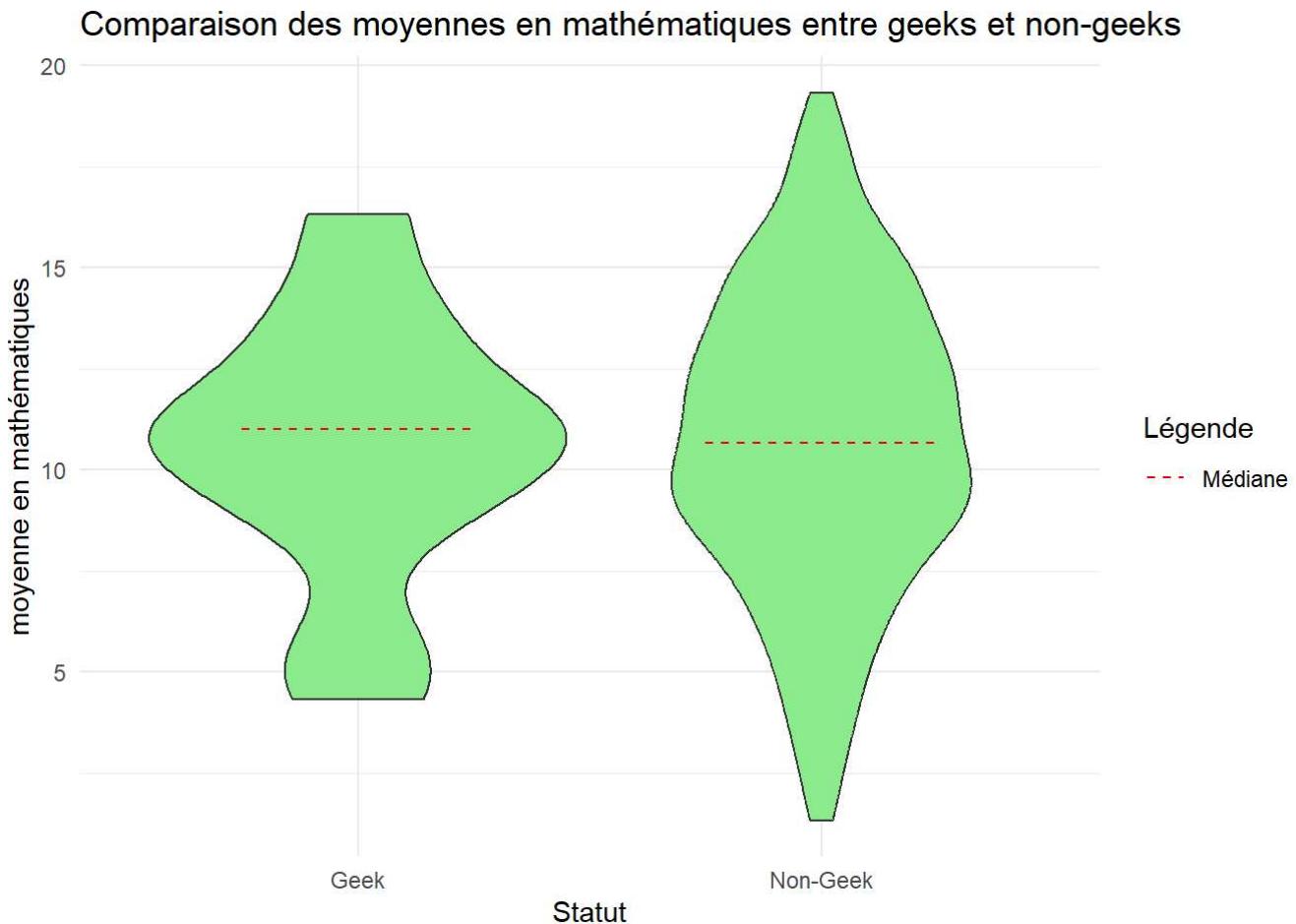
# Calcule de la moyenne des notes en mathématiques sur les trois semestres
moyenneNote <- rowMeans(data[,c("G1","G2","G3")])
```

```
# Créer un graphique pour comparer les notes
ggplot(data, aes(x = geek, y = moyenneNote)) +
  geom_boxplot(fill = "lightyellow") +
  labs(title = "Comparaison des moyennes en mathématiques entre geeks et non-geeks",
       x = "Statut",
       y = "moyenne en mathématiques") +
  theme_minimal()
```



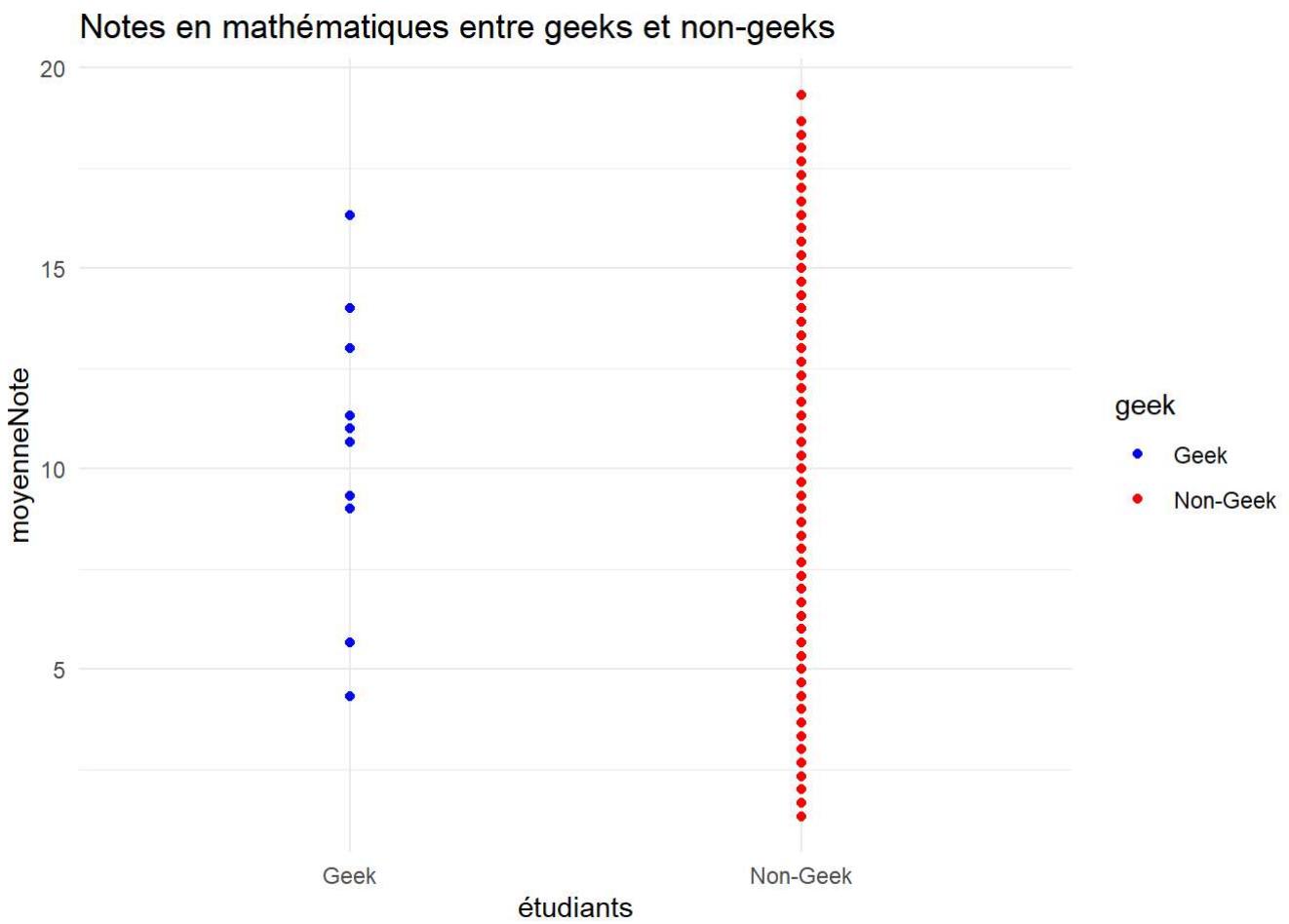
Dans ce premier graphique on utilise un geom boxplot car il est utile pour se rendre compte de la distribution des moyennes et de la médiane dans chaque catégorie. On remarque que les geeks ont une moyenne légèrement plus élevée mais cela n'est pas significatif et peut-être dû à d'autres facteurs. Il y a aussi une valeur très en dessous des autres, cela peut être une valeur aberrante.

```
# Créer un graphique pour comparer les notes
ggplot(data, aes(x = geek, y = moyenneNote)) +
  geom_violin(fill = "lightgreen") +
  stat_summary(fun = median, geom = "errorbar", aes(ymax = after_stat(y), ymin = after_stat(y), color = "Médiane"), width = 0.5, linetype = "dashed") +
  scale_color_manual(values = "red", name = "Légende", labels = "Médiane") +
  labs(title = "Comparaison des moyennes en mathématiques entre geeks et non-geeks",
       x = "Statut",
       y = "moyenne en mathématiques") +
  theme_minimal()
```



Ce deuxième graphique est identique au précédent à la différence qu'il est sous forme violon. Cela permet une meilleure estimation de la densité de moyenne pour une moyenne donnée. On remarque que pour les geeks la densité est la plus élevée autour de la médiane contrairement aux autres. On peut aussi remarquer anomalie pour les moyennes autour de 5. Cela montre une séparation en deux groupes parmi les geeks.

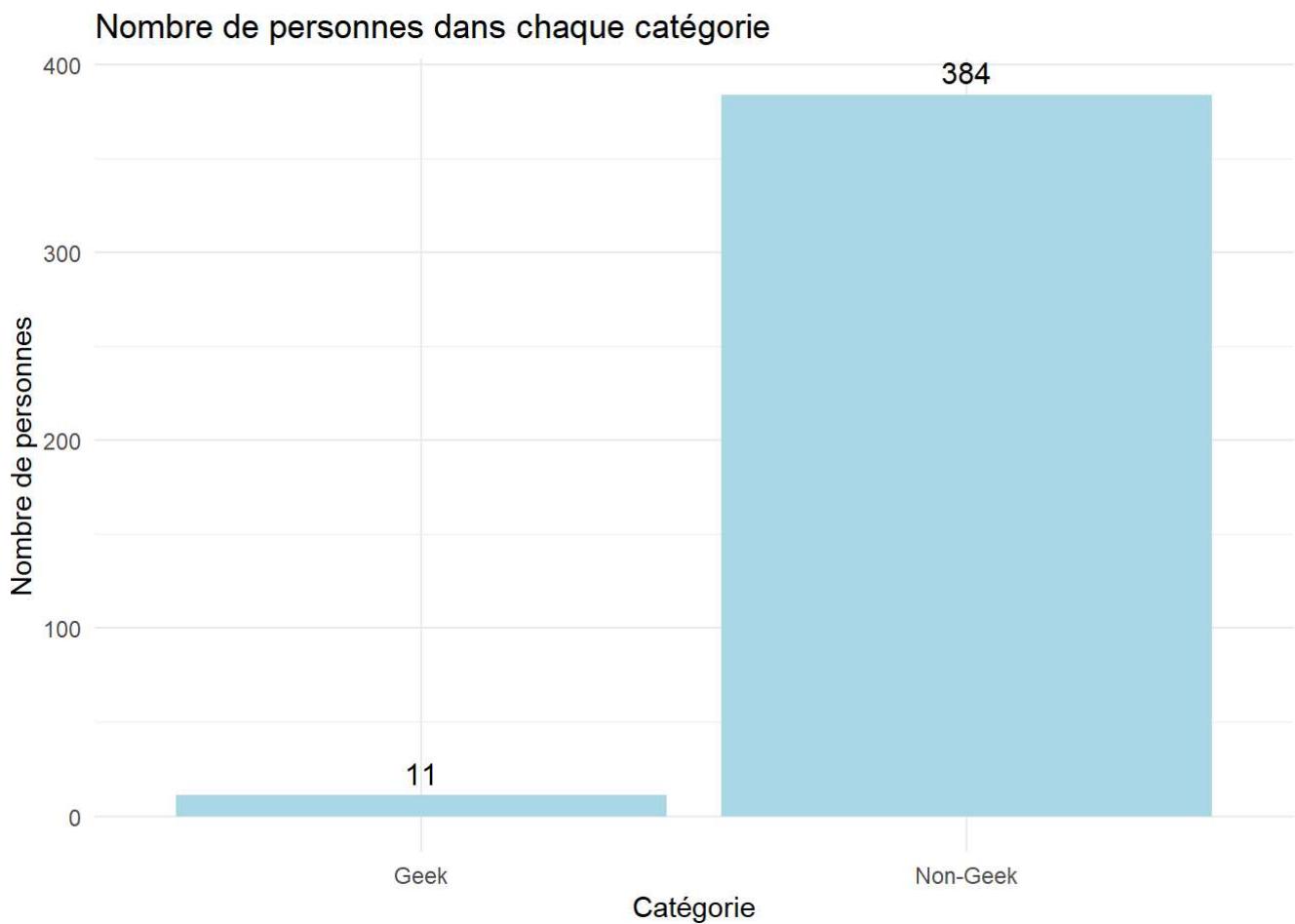
```
# Créer un nuage de points pour visualiser les notes en mathématiques entre les geeks et les non-geeks
ggplot(data, aes(x = geek, y = moyenneNote, color = geek)) +
  geom_point() +
  labs(title = "Notes en mathématiques entre geeks et non-geeks",
       x = "étudiants") +
  scale_color_manual(values = c("Geek" = "blue", "Non-Geek" = "red")) +
  theme_minimal()
```



Sur ce troisième graphique, on a décidé d'afficher chaque moyenne de manière individuelle pour pouvoir mieux se rendre compte de la répartition des notes. On remarque instantanément qu'il y a une grande différence d'échantillon parmi les deux catégories traitées. C'est est une première limite liée à cette analyse de données. Nous pouvons quand même dire que les données affichées sont cohérentes avec les deux premiers graphiques. On distingue bien deux groupes ainsi que la médiane.

```
count_data <- data %>%
  count(geek)

ggplot(count_data, aes(x = geek, y = n)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  geom_text(aes(label = n), vjust = -0.5, size = 4, color = "black") +
  labs(title = "Nombre de personnes dans chaque catégorie",
       x = "Catégorie",
       y = "Nombre de personnes") +
  theme_minimal()
```



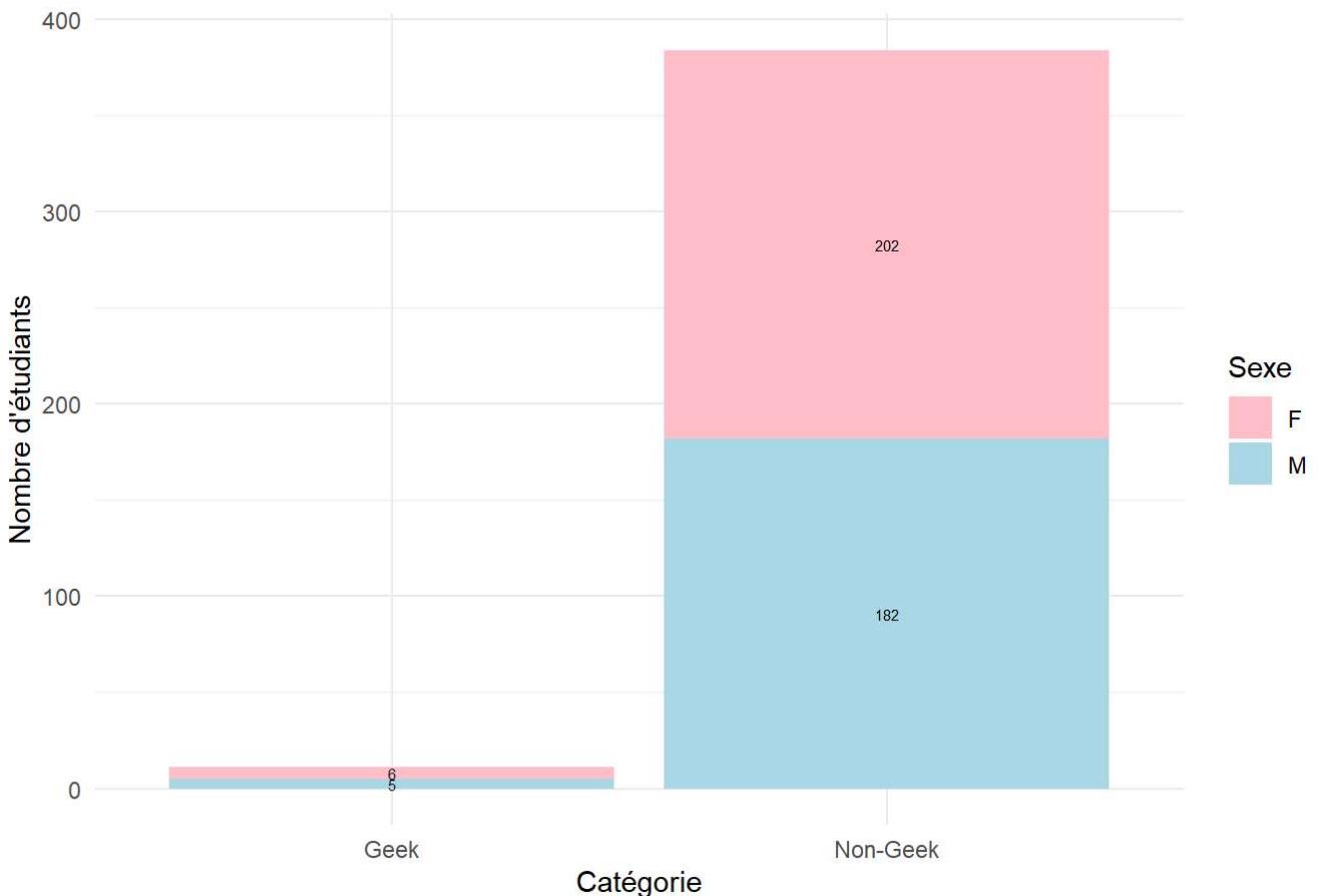
Sur ce dernier graphique, on a décidé de compter le nombre de personnes dans chaque catégorie. On remarque que le nombre de non-geeks est beaucoup plus élevé que le nombre de geek (facteur 35). Cela est une limite à notre analyse car les résultats ne sont pas exploitables sur un si petit échantillonnage.

```
# Créer un tableau croisé pour compter le nombre d'hommes et de femmes dans chaque catégorie
gender_counts <- table(data$geek, data$sex)

# Convertir le tableau croisé en un dataframe
gender_df <- as.data.frame(gender_counts)
names(gender_df) <- c("Catégorie", "Sexe", "Count")

# Créer un graphique à barres empilées pour visualiser la répartition des hommes et des femmes
ggplot(gender_df, aes(x = Catégorie, y = Count, fill = Sexe)) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = Count), position = position_stack(vjust = 0.5), color = "black", size = 2) +
  labs(title = "Répartition des hommes et des femmes parmi les catégories Geek et Non-Geek",
       x = "Catégorie",
       y = "Nombre d'étudiants",
       fill = "Sexe") +
  scale_fill_manual(values = c("pink", "lightblue")) + # Couleurs pour les hommes (bleu) et les femmes (rose)
  theme_minimal()
```

Répartition des hommes et des femmes parmi les catégories Geek et Non-Geek

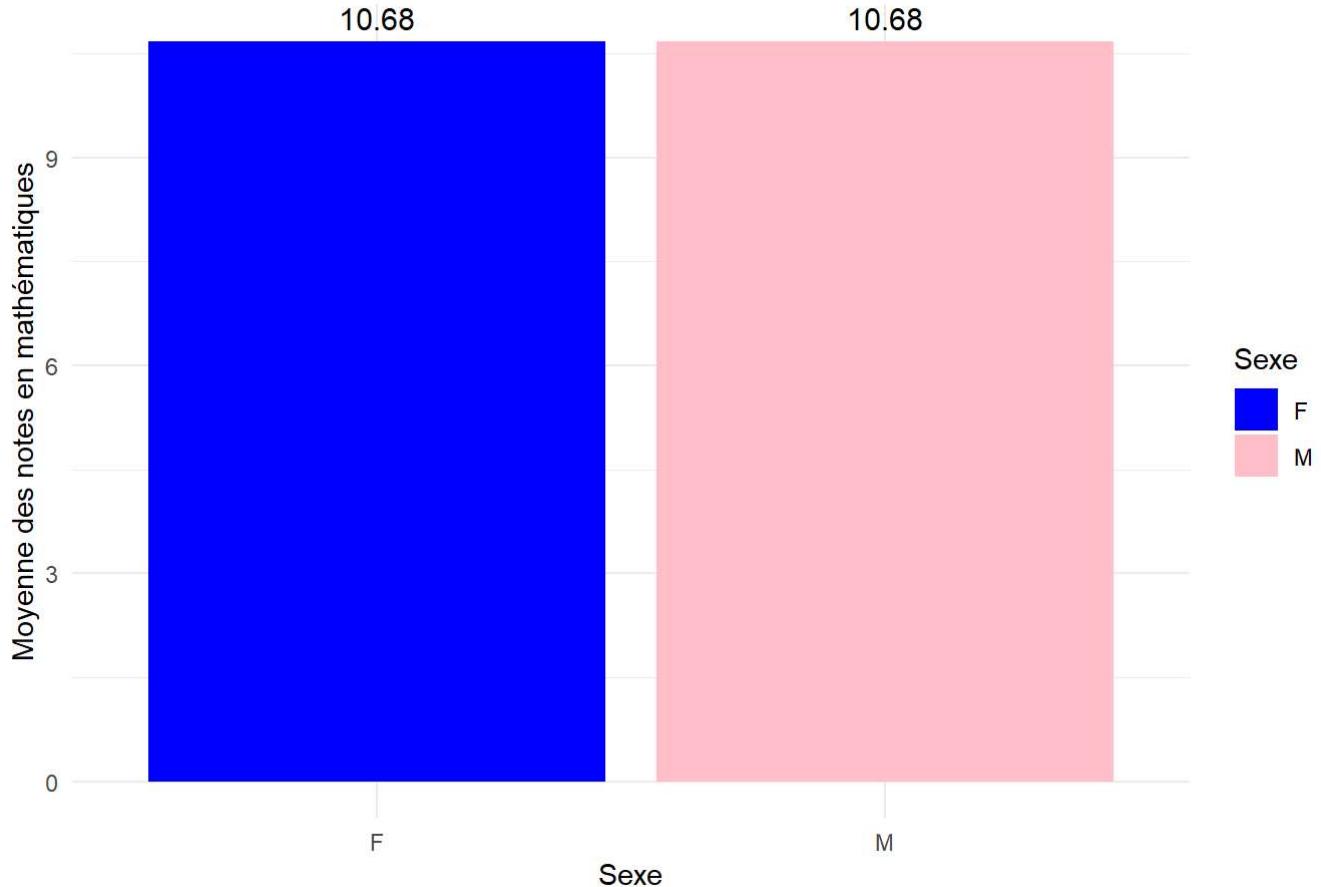


Sur ce graphique, on a décidé de compter le nombre d'hommes et de femmes dans chaque catégorie. On remarque que le nombre d'homme et de femme sont équivalents dans chaque catégorie. Donc cet aspect de l'analyse n'a pas d'importance.

```
means <- data %>%
  group_by(sex) %>%
  summarise(mean_math_score = mean(moyenneNote))

# Créer un graphique à barres montrant les moyennes des notes en mathématiques pour les hommes et les femmes
ggplot(means, aes(x = sex, y = mean_math_score, fill = sex)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(mean_math_score, 2)), vjust = -0.5, color = "black", size = 4) +
  labs(title = "Moyenne des notes en mathématiques entre hommes et femmes",
       x = "Sexe",
       y = "Moyenne des notes en mathématiques",
       fill = "Sexe") +
  scale_fill_manual(values = c("blue", "pink")) + # Couleurs pour les hommes (bleu) et les femmes (rose)
  theme_minimal()
```

Moyenne des notes en mathématiques entre hommes et femmes



Sur ce graphique, on a décidé de comparer les moyennes des notes en mathématiques entre les hommes et les femmes. On remarque que les femmes et les hommes ont exactement la même moyenne. Cela montre que le sexe n'a pas d'impact sur les notes en mathématiques.

Partie analyse

Dans cette partie nous tenterons d'analyser les différents résultats obtenus dans la partie exploratoire.

En effet, la précédente partie nous a permis de récolter beaucoup d'informations maintenant il s'agit de les comprendres.

On remarque que les geeks ont une moyenne en mathématiques légèrement plus élevée que les non-geeks. En effet, la médiane de ces derniers est légèrement plus haute. Cela s'est vu sur les deux premiers graphiques.

On a aussi pu remarquer que l'écart type des moyennes parmi les geeks était important à tel point que nous pouvions les séparer en deux groupes distincts. Cela est visible sur le deuxième graphique.

Cela nous montre qu'il y a peu de chances qu'une corrélation entre ces deux critères existe.

Mais il est important de noter que le nombre d'étudiants considérés comme geeks est très faible par rapport aux non-geeks. Cela fausse grandement l'analyse sachant que l'on parle de moyenne donc la population de chaque catégorie devrait être équivalente.

Donc, d'après ce dataset, la conclusion est que les geeks n'ont pas de meilleures notes en mathématiques que les non-geeks. Même si cette conclusion est erronée due au manque d'échantillonnage, elle est cohérente avec les données récoltées.

Question 4: Y-a-t-il un lien entre la consommation d'alcool et l'échec scolaire, l'absentéisme ?

Partie exploration

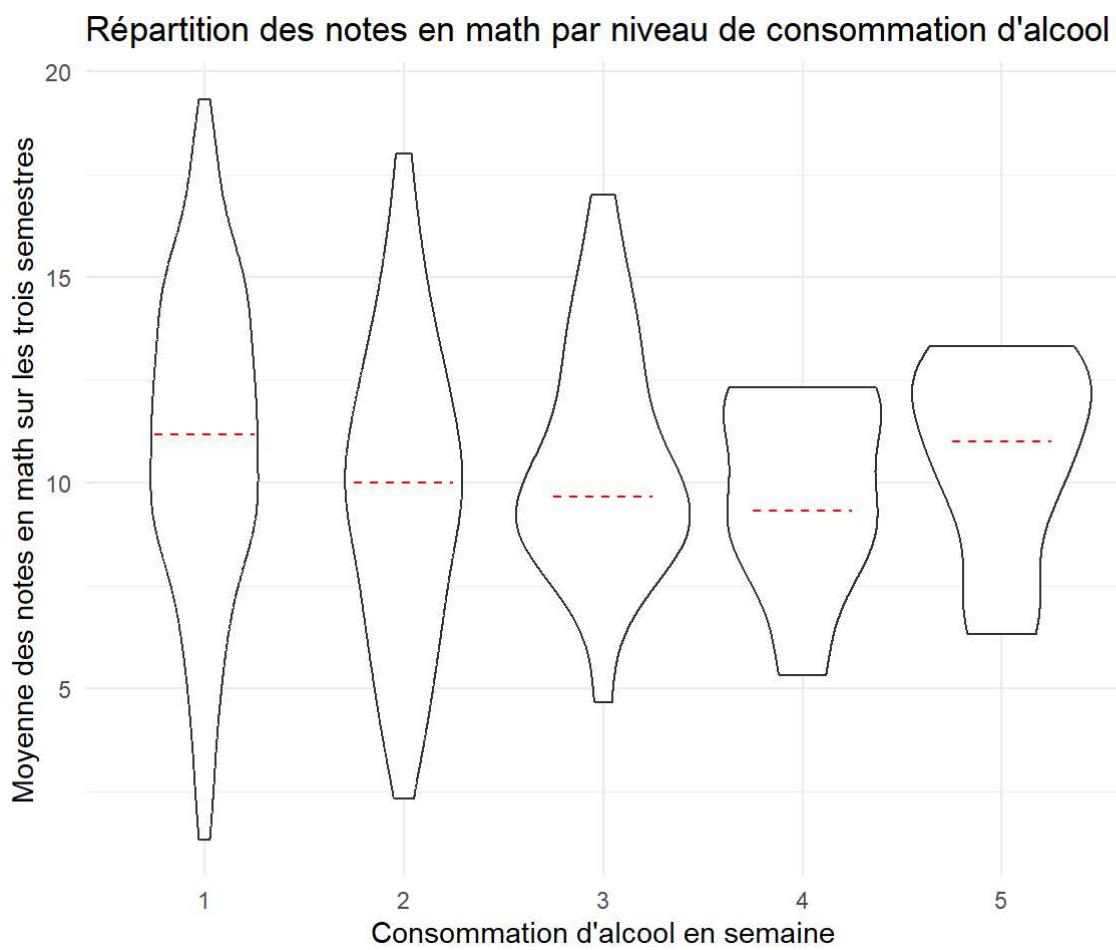
L'arrivée dans le monde des études est souvent le moment de la découverte de l'alcool. Or sachant le caractère destructeur que peut avoir l'alcool sur la santé. Nous nous sommes posé la question de l'impact de l'alcool sur l'échec scolaire. Nous voulons savoir si l'alcool a un impact à court terme sur ce groupe d'individu et pouvoir, peut être, le quantifier.

Nous voulons donc savoir si une forte consommation d'alcool a une influence négative sur les notes des individus. Nous pensons que l'alcool a une influence fortement négative sur les individus. Autrement formulés, nous pensons que les individus qui consomment le plus d'alcool sont ceux qui ont les moins bons résultats.

Les données qui m'interessent pour répondre à cette question sont :

- **Failures** : nombre d'échecs aux classes précédentes (numérique : n si 1 n 3, sinon 4)
- **Dalc** : consommation d'alcool en semaine (numérique : de 1 très faible à 5 très élevé)
- **Walc** : consommation d'alcool le week-end (numérique : de 1 très faible à 5 très élevé)
- **MoyenneNote** : qui est la moyenne que nous avons calculée en faisant la moyenne des notes des 3 trimestres
 - G1 : note du premier trimestre (numérique : de 0 à 20)
 - G2 : note du deuxième trimestre (numérique : de 0 à 20)
 - G3 : note finale (numérique : de 0 à 20, cible de sortie)

```
ggplot(data = tableMat, aes(x = factor(Dalc), y = MoyenneNote)) +  
  geom_violin() +  
  stat_summary(fun = median, geom = "errorbar", aes(ymin = after_stat(y), ymax = after_stat(y), color = "Médiane"), width = 0.5, linetype = "dashed") +  
  scale_color_manual(values = "red", name = "Légende", labels = "Médiane") +  
  labs(x = "Consommation d'alcool en semaine", y = "Moyenne des notes en math sur les trois semestres") +  
  theme_minimal() +  
  ggtitle("Répartition des notes en math par niveau de consommation d'alcool") +  
  theme(legend.position = "right")
```



Analyse statistique des données

On remarque que la répartition statistique des moyennes en math n'est pas vraiment impactée par la consommation d'alcool. En effet, les groupes d'individus qui se notent 5/5 pour la consommation d'alcool ont même une médiane légèrement supérieure à ceux qui se notent 1/5 (ceux qui consomment le moins).

Cependant, on peut noter que, plus la notation de la consommation d'alcool augmente, plus l'intervalle des notes à tendances à se resserrer et à se concentrer autour de la médiane.

On peut l'expliquer très simplement statistiquement par le nombre de personnes par catégorie. En effet, on sait que souvent les répartitions des notes suivent des lois normales. Or, les échantillons sont très grands pour une consommation d'alcool faible, mais très faibles pour une grosse consommation d'alcool. Donc les valeurs extrêmes seront alors moins probables. Cela peut alors affecter le nombre de valeur extrême.

Répartition par catégorie de consommation d'alcool (Math)

Voici la répartition du nombre de personnes par catégorie :

Consommation d'alcool en semaine	Nombre de personnes
1	276
2	75
3	26
4	9
5	9

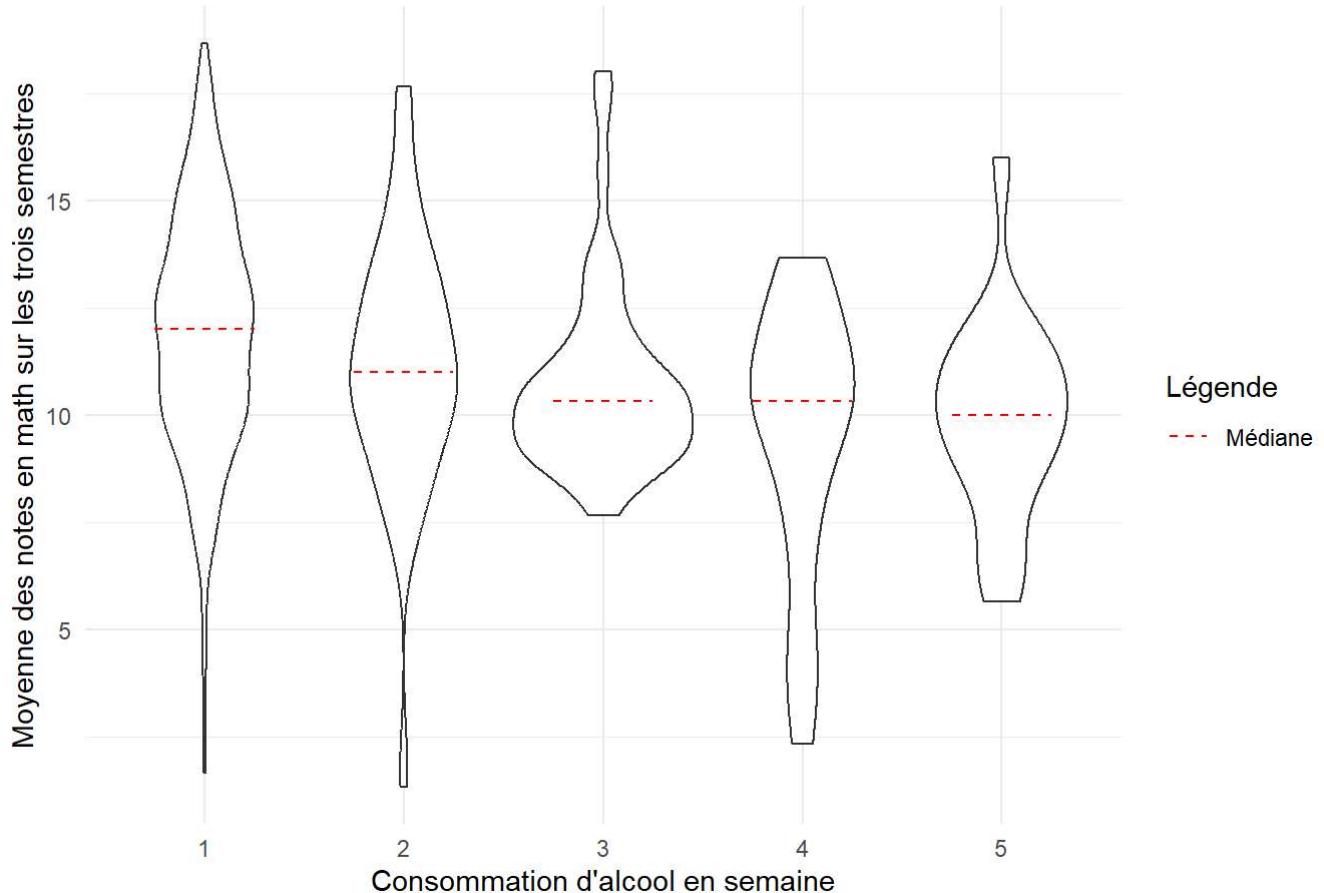
Répartition par catégorie de consommation d'alcool (Portugais)

Comparons avec les notes en portugais où¹ il y a deux fois plus d'individu par groupes:

Consommation d'alcool en semaine	Nombre de personnes
1	451
2	121
3	43
4	17
5	17

```
ggplot(data = tablePor, aes(x = factor(Dalc), y = MoyenneNote)) +  
  geom_violin() +  
  stat_summary(fun = median, geom = "errorbar", aes(ymin = after_stat(y), ymax = after_stat(y), color = "Médiane"), width = 0.5, linetype = "dashed") +  
  scale_color_manual(values = "red", name = "Légende", labels = "Médiane") +  
  labs(x = "Consommation d'alcool en semaine", y = "Moyenne des notes en math sur les trois semestres", title = "Répartition de la moyenne des notes en portugais par niveau de consommation d'alcool") +  
  theme_minimal() +  
  theme(plot.title = element_text(size=11, face="bold"))
```

Répartition de la moyenne des notes en portugais par niveau de consommation d'alcool



Analyse des données

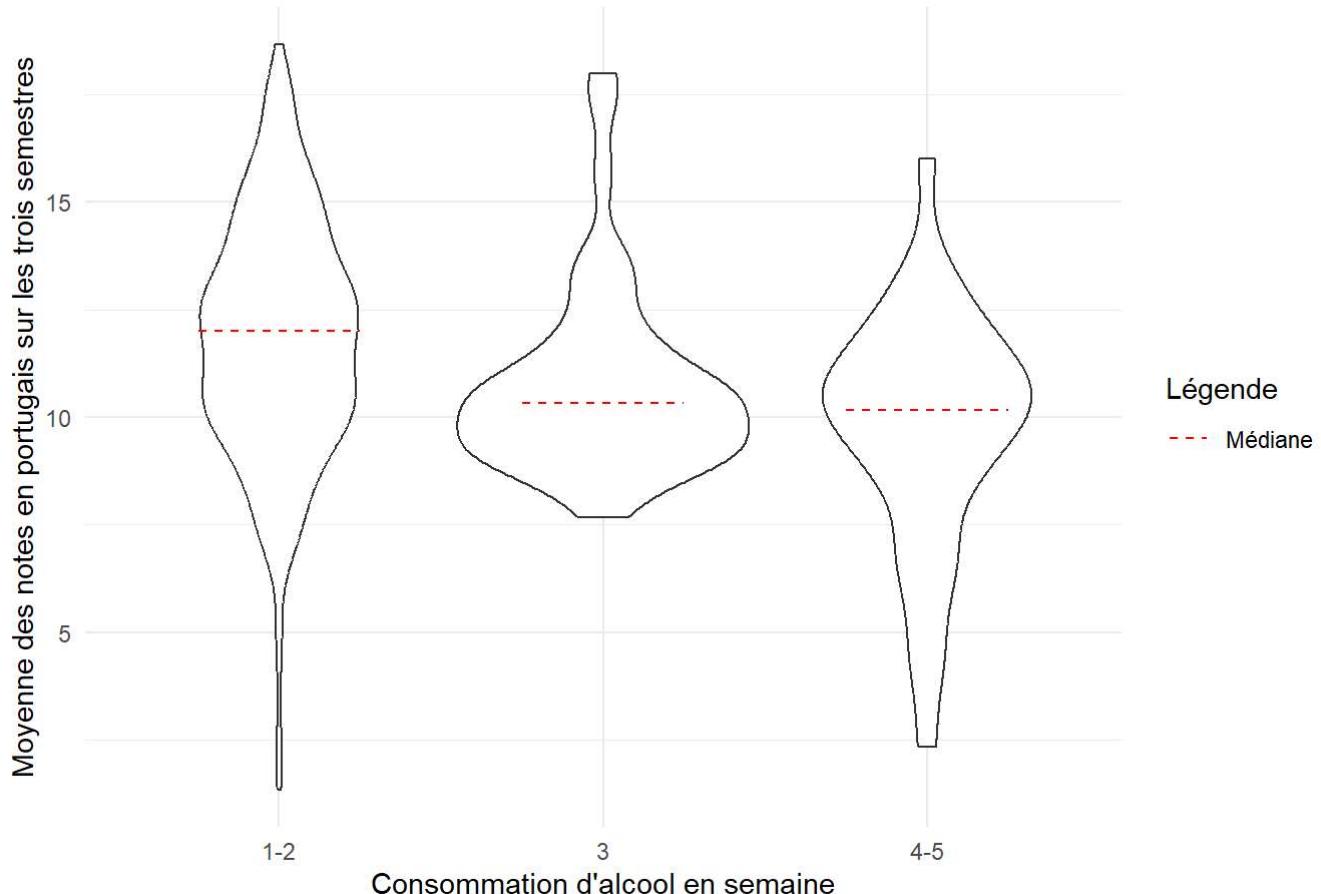
On remarque ici que la médiane des notes en Portugais chute légèrement quand la consommation d'alcool en semaine augmente. On retrouve ici, aussi le même phénomène de réduction de l'intervalle autour de la médiane quand la consommation d'alcool augmente.

Nous nous attendions à des résultats beaucoup plus significatifs. Il est fort probable aussi, au vu de l'auto-évaluation de la consommation d'alcool sur une échelle de 1 à 5, qu'un biais de la centralité est apparu chez certains étudiants, ce qui a peut-être faussé les données.

On va donc visualiser les catégories de consommation d'alcool 1 et 2 ensembles et les catégories 4 et 5 ensembles pour observer un résultat qui tiendrait compte de ce biais.

```
tablePor$GroupedDalc <- factor(ifelse(tablePor$Dalc %in% c(1, 2), "1-2", ifelse(tablePor$Dalc %in% c(4, 5), "4-5", tablePor$Dalc)))  
  
ggplot(data = tablePor, aes(x = GroupedDalc, y = MoyenneNote)) +  
  geom_violin() +  
  stat_summary(fun = median, geom = "errorbar", aes(ymin = after_stat(y), ymax = after_stat(y), color = "Médiane"), width = 0.5, linetype = "dashed") +  
  scale_color_manual(values = "red", name = "Légende", labels = "Médiane") +  
  labs(x = "Consommation d'alcool en semaine", y = "Moyenne des notes en portugais sur les trois semestres", title = "Répartition de la moyenne des notes en portugais par niveau de consommation d'alcool regroupé") +  
  theme_minimal() +  
  theme(plot.title = element_text(size=10, face="bold"))
```

Répartition de la moyenne des notes en portugais par niveau de consommation d'alcool regrou



Analyse des données

Avec ce graphique, on constate la même chute de la moyenne avec maintenant un échantillon plus important pour la catégorie 4-5. On remarque aussi, les gaussiennes grâce à cette visualisation en violin, ce qui nous conforte à croire que l'échantillon est un peu plus représentatif.

Répartition par catégorie de consommation d'alcool

Consommation d'alcool en semaine	Nombre de personnes
1/2	572
3	43
4/5	34

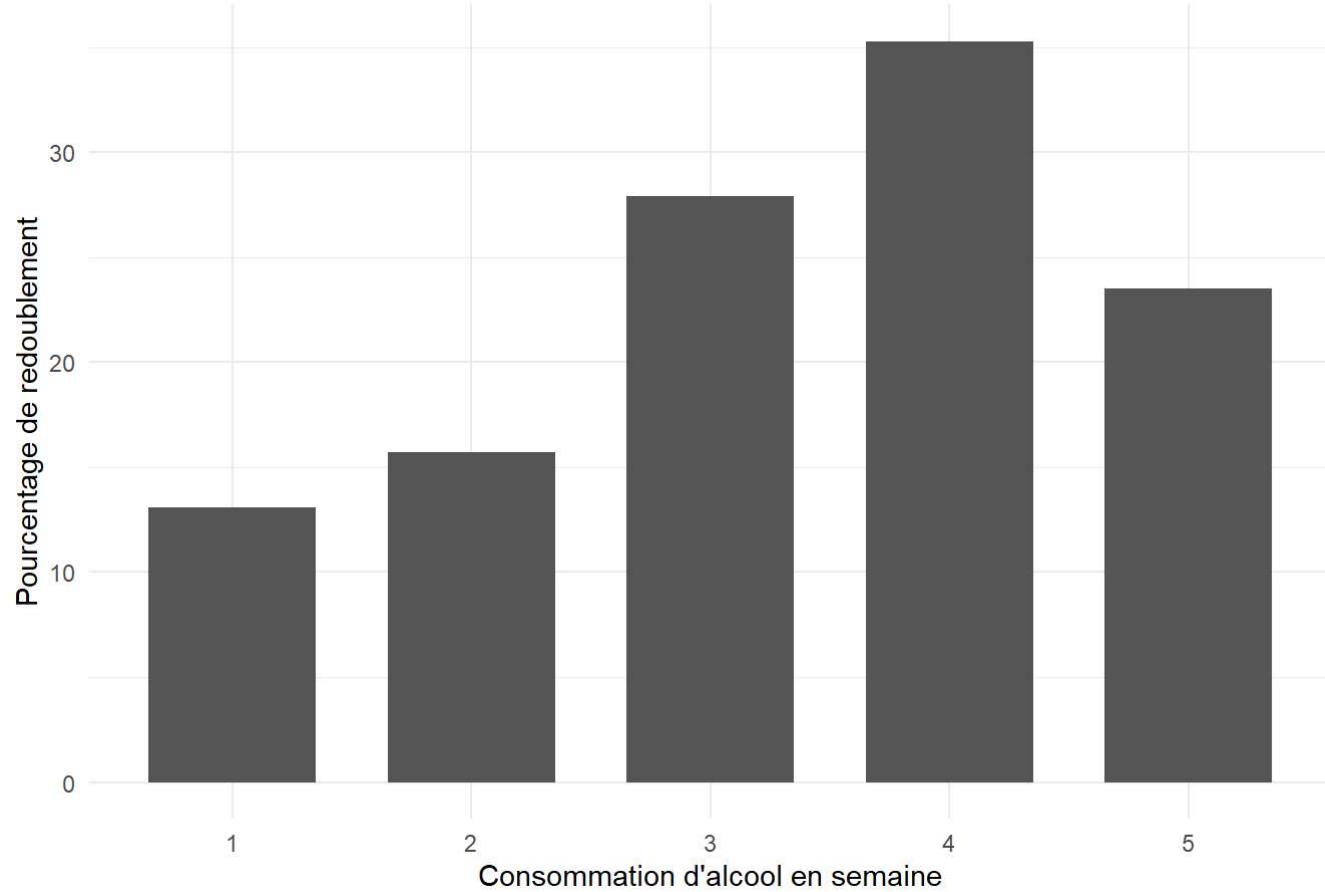
On peut maintenant tisser une relation entre ceux qui consomme le plus et ceux qui ont une moyenne au cours de l'année plus faible. Cela peut venir de plusieurs facteurs. Les causes possibles sont soit le manque de données, soit le fait que l'alcool rende moins performant, soit que les moins performant ont tendance à se tourner vers l'alcool/la sociabilité/l'extrascolaire, soit un mixte des deux derniers facteurs.

On peut également se poser la question du redoublement des personnes qui consomment le plus d'alcool. En effet, les personnes, qui consomment le plus, sont peut-être ceux qui ont déjà redoublé. Si c'est le cas, on pourrait s'interroger sur le rôle du redoublement dans la moyenne de leurs notes actuelles.

```
proportions_redoublement <- tablePor %>%
  group_by(Dalc) %>%
  summarise(Total = n(),
            Redoublements = sum(failures >= 1), # Assurez-vous que 'failures' est correctement défini pour redoublement
            PourcentageRedoublement = (Redoublements / Total) * 100) %>%
  ungroup()

# Créer Le graphique
ggplot(data = proportions_redoublement, aes(x = factor(Dalc), y = PourcentageRedoublement)) +
  geom_bar(stat = "identity", position = position_dodge(), width = 0.7) +
  scale_fill_brewer(palette = "Set1", name = "Consommation d'alcool en semaine") +
  labs(title = "Pourcentage de redoublement par niveau de consommation d'alcool en semaine",x ="Consommation d'alcool en semaine",y="Pourcentage de redoublement") +
  theme_minimal()
```

Pourcentage de redoublement par niveau de consommation d'alcool en semaine



On remarque que notre hypothèse a été validé. Il y a bien plus de redoublement proportionnellement pour une plus grande consommation d'alcool. Ce qui peut compenser, le fait que la médiane des moyennes des notes sur les groupes 4 et 5 ne soit pas énormément inférieur à la médiane des moyennes des notes des groupes 1 et 2.