CONSOMMATION D'ALCOOL CHEZ LES JEUNES ETUDIANTS PORTUGAIS

Rambaud & Machado

date rendu : 31 mars 2017

TABLE DES MATIERES

I) Introduction

Pour ce projet de stat nous avons choisi de travailler sur les données suivantes : La consommation d'alcool chez les jeunes étudiants portugais. Cette étude a été réalisée dans deux grandes écoles portugaises que sont Gabriel Pereira et Mousinho Da Silveira. L'échantillon sur lequel nous allons travaillé est constitué de personnes agées de 15 à 22 ans. Nous pensons que ces données peuvent étre intéressantes à analyser et à traiter.

Nous avons décidé de répondre à 3 questions :

- 1) La consommation d'alcool dépend-elle de la situation familiale ?(niveau d'éducation et d'étude, nombre de frères et soeurs, statut des parents, etc. . .)
- **2) Le caractère d'une personne peut-il influencer le niveau de consommation d'alcool chez un étudiant? (Sociabilité, ambition, sportif, nombre d'abscences scoalaires, etc. . .)**
- 3) L'environnement scolaire influence-il la consommation d'alcool?

Nous avons fait parvenir nos trois questions aux professeurs, ces derniers nous ayant conseillés de garder la seconde, celle-ci étant pour eux la plus complète.

II) Méthodologie

Pour répondre à cette problématique nous avons décidé de nous organiser de la façon suivante : Nous avons sélectionné tous les critères qui nous semblait cohérent pour définir le caractère d'une personne et plus précisémment d'un étudiant.

La liste des critères est donc la suivante : - Sexe - Age - En couple - Temps de travail pour les cours - Activité extra-scolaire - Volonté de poursuivre les études - Nombre de sortie entre amis - Nombre d'abscence en cours

Pour chaque critère nous avons fait une hypothèse et nous avons effectué une analyse factorielle. Le but étant de voir si l'analyse des résultats mettait en évidence une différence significative. Si différence significative il y a, il faut voir s'il n'y a pas de données qui faussent le résultat. Enfin on a pu valider ou non notre hypothèse de base.

La deuxième partie de notre travail a consisté à mettre en relation plusieurs critères que nous avons pu juger comme pertinent grâce à notre analyse factorielle. Le but étant de voir si on constate un lien qui permettrait d'affiner notre réponse à la problématique. Cette deuxième partie est donc une analyse multivariée.

Après ces deux analyses nous avons pu tiré une conclusion répondant à notre problématique.

III) Analyse Factorielle

Avant de se pencher sur notre problématique, nous avons trouver interessant de nous pencher sur certains critères. Tout d'abord nous allons voir si notre échantillon contient un nombre de filles et de garçons proche (il est nécessaire que le nombre s'equilibre afin de pouvoir faire des traitements sur ces données)

```
df <- read.csv("/Users/machado/student-mat.csv");
plot(df$sex, ylab="Nombre de personnes", xlab="sexe" ,main="Repartition Homme/Femme");
```


```
summary(df$sex);
```

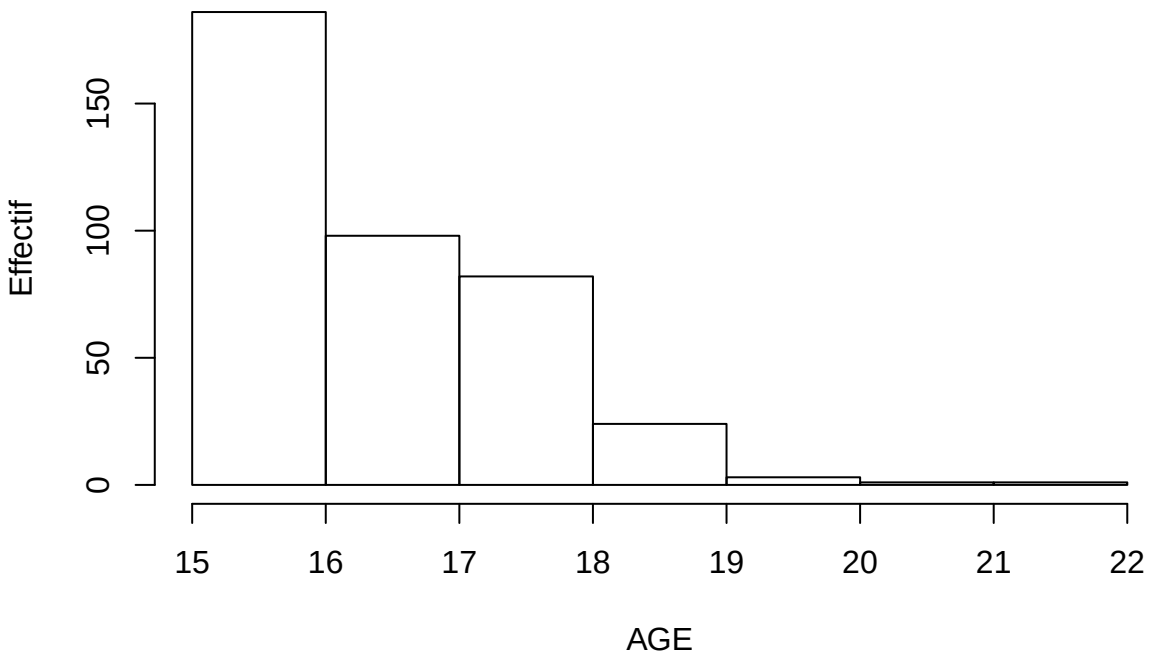
```
##   F   M
## 208 187
```

```
moy<-(df$Walc+df$Dalc)/2;
```

Nous pouvons constater à l'aide du graphe ci-dessus que le nombre de fille et de garçcon s'équilibre (208 F , 187 H), nous pourrons donc utiliser le sexe comme un critère pour diffèrentes analyses multivariées (ce critère est donc pertinent).

Nous trouvons important aussi d'analyser un critère qui pour nous est primordial, l'âge de notre population :

```
hist(df$age, breaks=7, xlab="AGE", ylab="Effectif", main="Répartition des âges")
```




```
summary(df$age);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    16.0    17.0    16.7    18.0    22.0
```

Nous travaillerons donc sur des personnes assez jeunes, la moyenne étant un peu plus que 16 ans et demi (16,7 pour être exact). L'age minimum étant 15 et l'age maximum 22 . Il pourrait être intéressant de regarder quelle est la tranche d'age ayant la consommation d'alcool la plus forte.

Notre seconde question est sans doute celle sur laquelle nous avons poussé le plus notre réflexion. Celle-ci portant sur la consommation d'alcool en fonction du caractère de la personne (Sociabilité, ambition, sportif, nombre d'abscences scoalaires, etc. . .)

Dans notre base nous possèdons deux variables concernant directement la consommation d'alcool, ces deux variables étant :

-Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

-Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

Après reflexion nous avons fait le choix de réunir ces deux critères afin de n'avoir qu'une seule variable concernant la consommation d'alcool. Nous aurions pu réaliser une somme mais nous avons trouvé plus pertinent de réaliser une moyenne de ces deux variables. Cela nous permettra de garder la même echelle de comparaison.

Celle-ci se présentant comme ci-dessous : -week alcohol consumption (from 1 - very low to 5 - very high)

```
moytot<-mean(df$Dalc+df$Walc)
moytot;
```

```
## [1] 3.772152
```

```
moy<-(df$Walc+df$Dalc)/2;
da<-data.frame(
```

```
            Semaine = df$Dalc,
            Weekend = df$Walc,
            Moyenne = moy);
moy
```

```
##   [1] 1.0 1.0 2.5 1.0 1.5 1.5 1.0 1.0 1.0 1.0 1.5 1.0 2.0 1.5 1.0 1.5 1.5
##  [18] 1.0 3.0 2.0 1.0 1.0 2.0 3.0 1.0 2.0 1.5 3.0 1.0 5.0 3.5 1.0 1.0 1.0
##  [35] 1.0 1.0 1.0 1.0 1.0 1.0 1.5 3.0 1.0 1.0 2.0 1.0 2.5 1.0 2.0 1.0 2.5
##  [52] 1.0 3.5 2.5 4.0 1.0 1.0 1.0 1.0 1.0 2.5 5.0 1.0 3.0 3.0 1.5 5.0 1.5
##  [69] 2.0 2.5 1.0 1.0 3.0 2.0 3.0 2.5 1.0 2.0 1.0 1.5 2.0 1.5 1.0 2.0 2.5
##  [86] 2.5 1.5 2.0 1.0 4.0 2.0 2.0 2.5 1.0 1.0 1.0 1.0 1.0 1.5 1.0 5.0 1.0
## [103] 1.0 1.0 1.0 1.0 1.0 1.0 4.0 1.0 1.0 1.0 1.0 1.0 1.0 1.5 1.0 1.0 2.5
## [120] 1.5 1.5 1.5 1.5 2.5 1.0 2.5 1.0 1.0 1.5 3.5 2.0 1.5 2.0 2.5 1.0 1.0
## [137] 3.0 1.0 2.0 1.0 1.0 2.0 1.0 3.0 1.5 1.5 1.0 1.0 1.5 3.5 3.5 4.0 2.5
## [154] 1.0 1.0 1.0 3.0 3.0 1.5 4.0 2.0 2.5 3.0 2.5 3.0 1.0 3.0 1.0 1.0 1.0
## [171] 3.0 1.0 2.0 1.0 1.0 4.0 2.5 2.5 3.5 1.0 2.5 1.5 2.5 2.5 1.5 2.5 1.5
## [188] 1.5 2.0 3.0 1.0 1.0 4.5 3.5 1.0 1.0 1.5 4.0 2.5 1.5 3.0 2.0 2.0 1.5
## [205] 1.0 3.5 2.0 1.0 2.5 1.0 1.5 4.5 1.0 3.0 1.5 2.0 3.0 3.0 2.5 1.0 1.5
## [222] 1.0 1.0 5.0 1.0 1.0 2.0 2.0 4.5 1.5 1.5 1.0 2.0 3.0 1.0 2.0 5.0 1.0
## [239] 1.0 4.0 2.5 2.5 1.0 1.5 1.0 1.0 1.0 5.0 2.0 3.0 3.0 2.0 3.5 2.0 3.0
## [256] 1.5 1.0 1.0 1.5 1.0 2.0 1.0 1.0 1.0 1.0 3.5 3.5 2.0 2.0 1.5 3.0 2.0
## [273] 1.0 2.0 1.0 2.5 1.0 2.5 1.0 1.5 3.0 3.5 1.0 1.0 1.5 1.5 1.5 1.0 2.0
## [290] 1.0 2.5 1.5 1.0 1.0 1.0 2.0 2.5 1.5 1.0 2.0 1.0 2.0 1.0 1.5 1.0 1.0
## [307] 1.0 1.0 1.5 2.0 2.0 1.0 2.0 1.5 1.0 1.0 1.5 1.0 3.5 3.0 1.5 1.0 2.0
## [324] 2.5 2.5 2.0 4.0 5.0 2.0 1.5 3.0 1.0 1.0 1.0 1.0 2.0 1.5 2.5 1.0 2.5
## [341] 2.0 2.0 2.0 1.5 1.5 2.5 1.5 2.5 2.0 5.0 3.0 2.5 2.5 3.0 2.0 1.0 1.5
## [358] 1.5 1.5 1.0 2.5 2.5 2.0 1.0 1.5 3.0 2.0 1.5 1.5 3.0 1.0 2.5 1.0 2.0
## [375] 1.0 1.5 1.0 3.5 1.5 2.5 2.5 2.0 1.0 2.0 3.5 2.0 2.0 1.5 1.0 1.0 4.5
## [392] 3.5 3.0 3.5 3.0
```

```
mean(moy)
```

```
## [1] 1.886076
```

Nous avons donc réalisé pour chaque étudiant la moyenne des deux critères concernant la consommation d'alcool. Nous tomberons donc sur des valeurs qui ne sont pas toutes entières, il y aura donc des valeurs comprises entre deux tranches. Nous pouvous voir que la consommation moyenne chez les étudiants (semaine+week-end) nous donne une moyenne d'environ 1,88 soit une consommation d'alcool proche de faible.

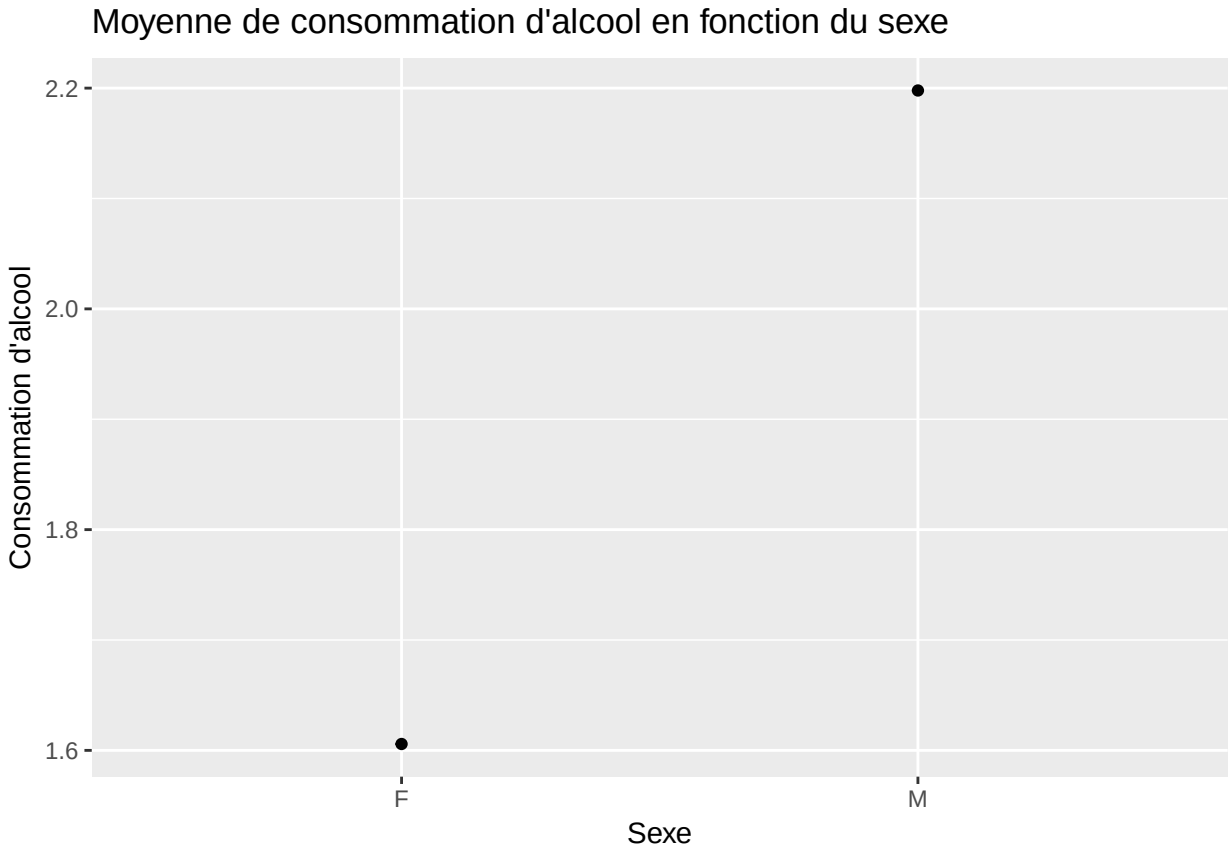
Nous allons procéder à **une analyse factorielle** portant sur le caractère d'une personne, nous avons donc séléctionné les critères que nous pensons en rapport avec le caractère :

- Si l'élève est serieux (temps de travail par semaine / nombres d'absences)
- S'il est ouvert à certaines activités (extra scolaire)
- S'il est ambitieux (volonté de poursuivre les études)
- S'il est "fêtard"" (sortie avec des amis)
- S'il est en couple

Nous allons donc voir que certains critères permettre de montrer leurs influences sur la consommation d'alcool Le premier que nous allons regarder est le sexe, celui-ci ne permet pas de décrire le caractère d'une personne cependant nous avions vu tout à l'heure que ce critère pouvait être interessant du fait que le nombre d'homme et de femme est equilibré.

```
t1<-df  %>% group_by(sex) %>% summarize(mean_sex = mean(Dalc+Walc)/2);
plot1<-ggplot(t1, aes(x = sex, y = mean_sex))+
               geom_point()+
               xlab("Sexe")+
```

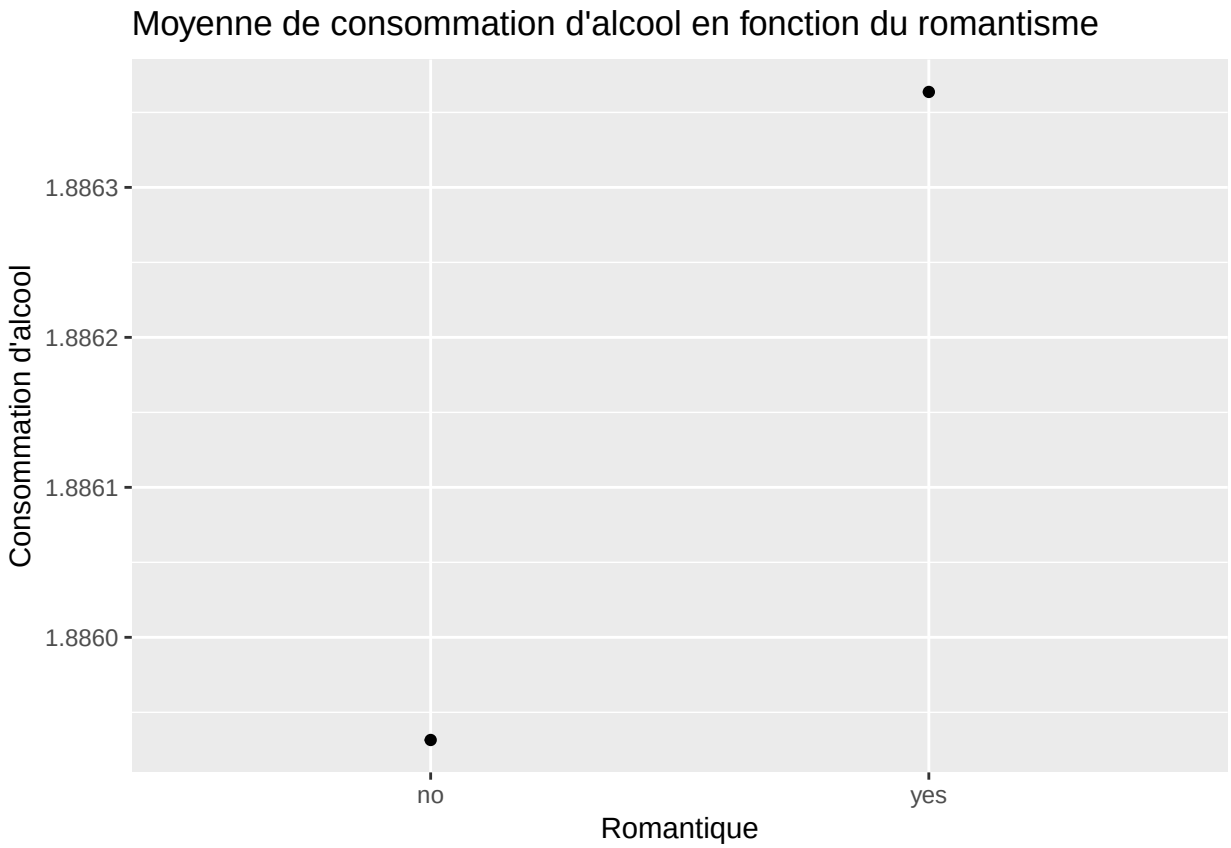
```
              ylab("Consommation d'alcool")+
              ggtitle("Moyenne de consommation d'alcool en fonction du sexe");
plot1
```



Le premier plot nous montre une certaine tendance, les hommes consomment en moyenne plus d'alcool que les femmes. En effet la consommation des femmes se situe entre faible et très faible tandis que les hommes, eux, ont une consommation moyenne "normale" (2.2, entre la forte et la faible consommation)

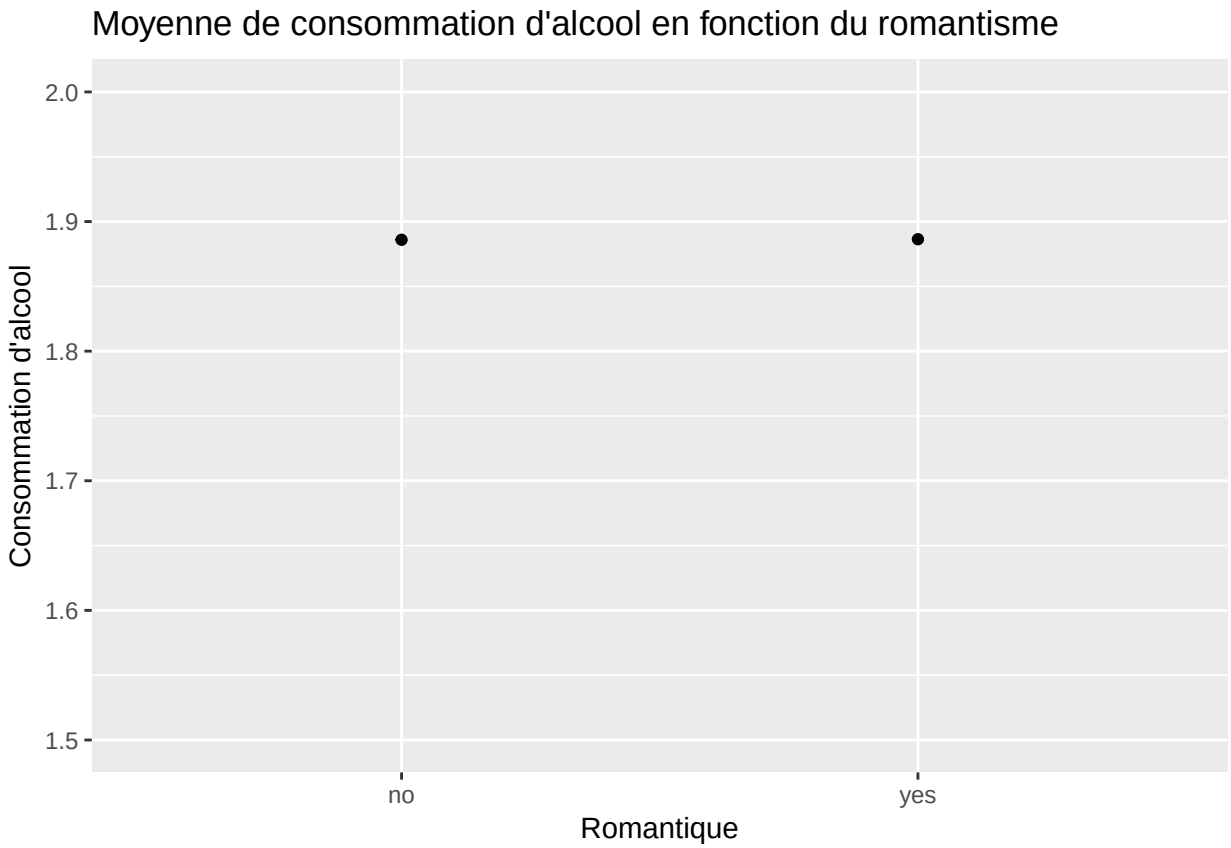
Nous allons désormais nous intéresser au critère de "romantisme". Ce critère sert tout simplement à dire si notre sujet est en couple ou pas. On pourrait penser qu'une personne qui est dans une relation amoureuse consomme moins d'alcool qu'une personne seule. Nous allons donc procéder de la même manière que pour le critère du sexe et analyser le résultat

```
t2<-df  %>% group_by(romantic) %>% summarize(mean_rom = mean(Dalc+Walc)/2);
plot2<-ggplot(t2, aes(x = romantic, y = mean_rom))+
              geom_point()+
              xlab("Romantique")+
              ylab("Consommation d'alcool")+
              ggtitle("Moyenne de consommation d'alcool en fonction du romantisme");
plot2
```



A première vue on pourrait interpréter ce graphique de la manière suivante : Les personnes en couples (“Yes”) consomment beaucoup plus d’alcool que les personnes célibataires (“No”). Cependant en regardant l’axe des ordonnées nous pouvons constater que l’échelle est vraiment réduite et que finalement les valeurs sont similaires. Nous allons donc garder ce même graphique mais changer l’échelle des ordonnées.

```
plot2 + scale_y_continuous(name="Consommation d'alcool", limits=c(1.5, 2));
```



En mettant l'échelle de 1.5 à 2 nous voyons tout de suite qu'il n'y a presque aucune différence entre la consommation d'alcool des étudiants en couple et celle des étudiants célibataires. De plus cette consommation d'alcool avoisine la consommation moyenne (environ 1.8). Notre hypothèse est donc fausse et grâce à ces deux graphiques nous savons que le critère de "Romantisme" n'est pas un critère très pertinent pour répondre à notre problématique.

Le prochain critère sur lequel nous allons nous pencher est le temps de travail. En effet nous avons selectionné ce critère pour montrer le serieux des etudiants. Avant l'analyse il semblerait logique que les personnes travaillant sont les personnes consommant le moins d'alcool.

```
t3<-df  %>% group_by(studytime) %>% summarize(mean_trav = mean(Dalc+Walc)/2);
plot3<-ggplot(t3, aes(x = studytime, y = mean_trav))+
               geom_point()+
               xlab("Temps de travail")+
               ylab("Consommation d'alcool")+
               ggtitle("Moyenne de consommation d'alcool en fonction du temps de travail");
plot3
```


```
t3count<-df  %>% group_by(studytime) %>% count();
t3count
```

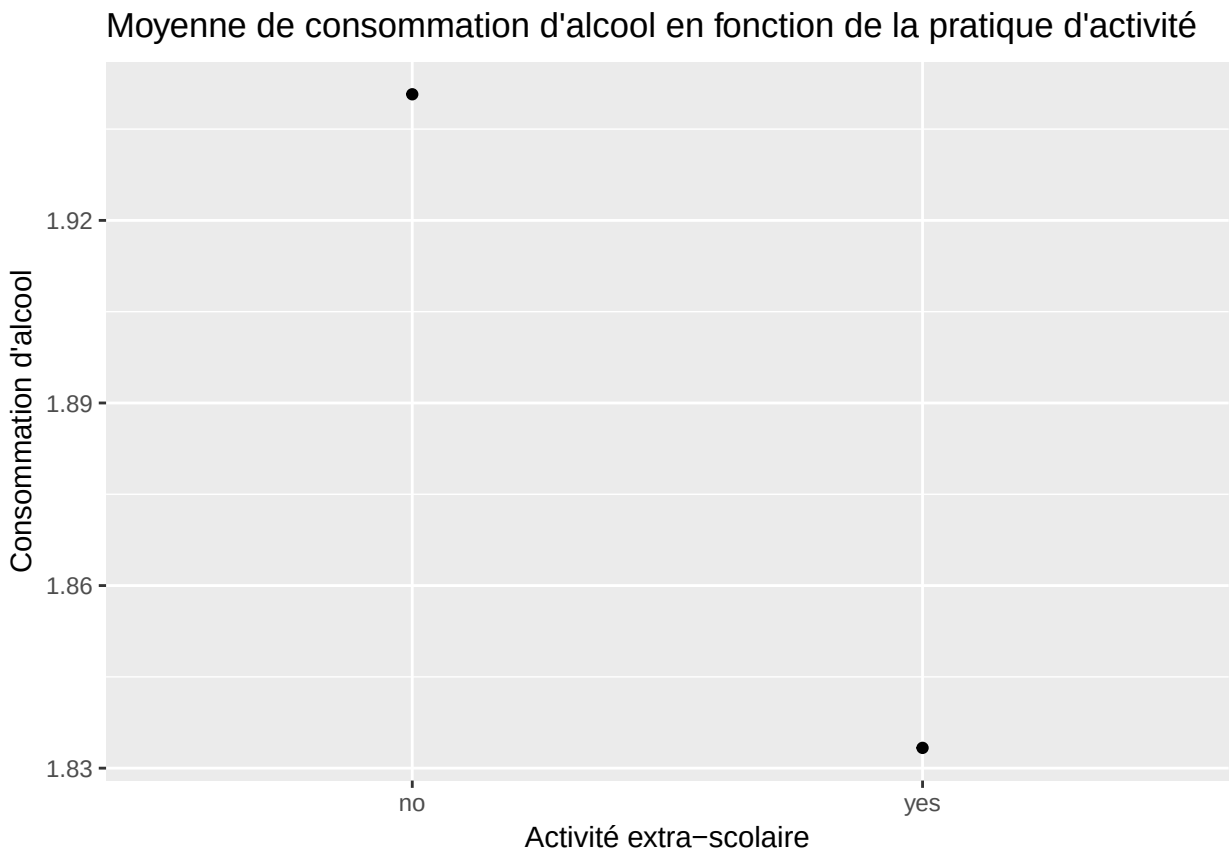
```
## # A tibble: 4 × 2
##   studytime     n
##       <int> <int>
## 1         1   105
## 2         2   198
## 3         3    65
## 4         4    27
```

L'hypothèse que nous avons effectué s'avère juste. On peut voir que les personnes travaillant moins de 2h par semaine ont une consommation d'alcool superieur à la moyenne (2.2 pour une moyenne à 1.88), que les personnes travaillant entre 2h et 5h sont eux très proches de la moyenne avec 1.85, ce qui est facilement compréhensible du fait que 198 personnes se trouvent dans cette catégorie, soit plus de la moitié de notre effectif total. Nous pouvons aussi remarquer que les personnes travaillant plus de 5h ont une consommation d'alcool entre faible et très faible.

Le critère suivant est celui de l'activité extra-scolaire. Les étudiants pratiquant une activité extra-scolaire ("Yes") consomment-ils plus que les autres ? Après analyse du graphique nous pouvons tirer la même conclusion que pour le critère de "Romantisme" à savoir que les valeurs sont très proches, c'est juste la mise en forme du graphique qui donne une impression de grosse différence. Les valeurs étant proches de la consommation moyenne, ce critère ne nous permet de tirer aucune conclusion pour répondre à notre problématique.

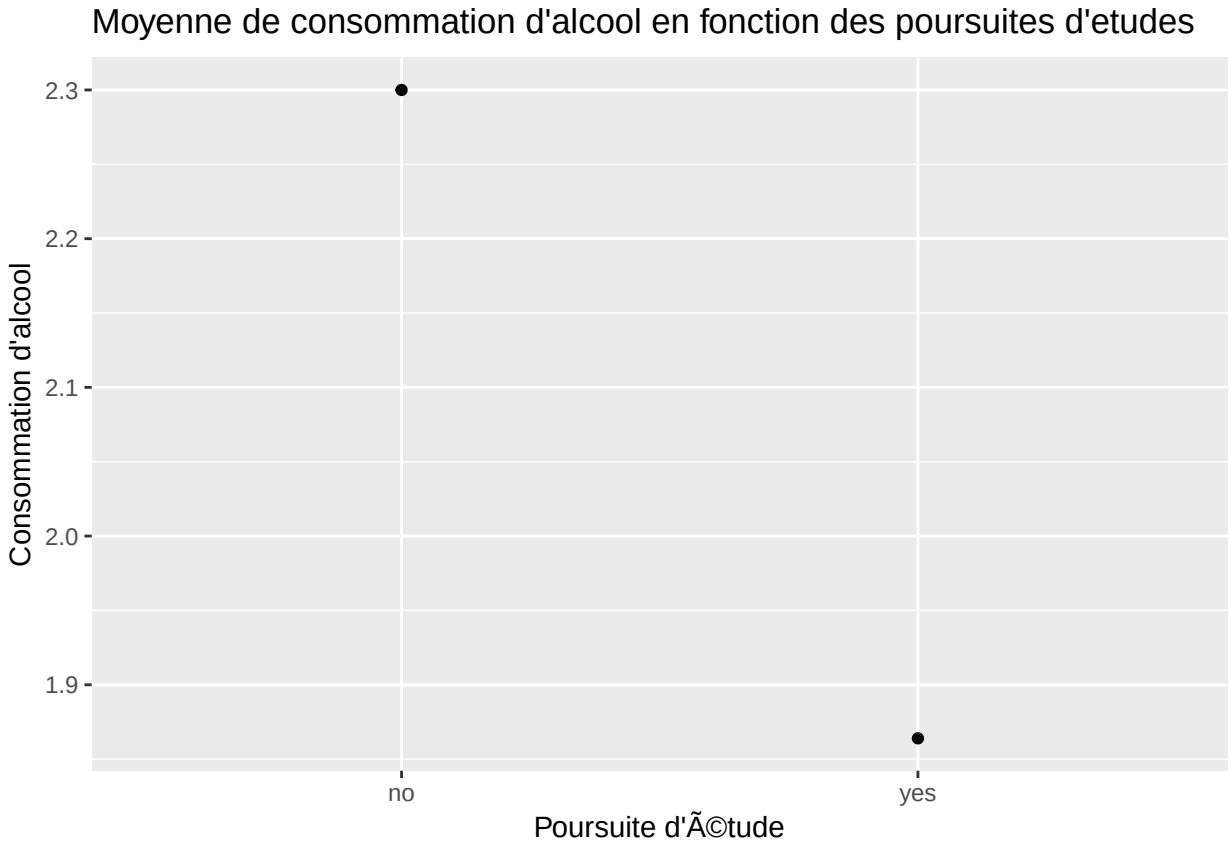
```
t4<-df  %>% group_by(activities) %>% summarize(mean_act = mean(Dalc+Walc)/2);
plot4<-ggplot(t4, aes(x = activities, y = mean_act))+
```

```
                geom_point()+
                xlab("Activité extra-scolaire")+
                ylab("Consommation d'alcool")+
                ggtitle("Moyenne de consommation d'alcool en fonction de la pratique d'activité");
plot4
```



Le critère suivant peut aussi être intéressant à analyser. Il s'agit de la volonté ou non de poursuivre ses études. On s'attend à voir que les personnes souhaitant s'investir dans leur études consomment moins d'alcool en général (semaine ou week-end) que les personnes souhaitant arrêter.

```
t5<-df  %>% group_by(higher) %>% summarize(mean_etu = mean(Dalc+Walc)/2);
plot5<-ggplot(t5, aes(x = higher, y = mean_etu))+
                geom_point()+
                xlab("Poursuite d'Ã©tude")+
                ylab("Consommation d'alcool")+
                ggtitle("Moyenne de consommation d'alcool en fonction des poursuites d'etudes");
plot5
```



On voit effectivement que l'étude de ce critère est pertinente, il y a une nette différence entre les deux différentes catégories. Cette différence est significative et elle nous permet de tirer une conclusion partielle : La volonté de l'étudiant à poursuivre ses études a un impact conséquent sur sa consommation d'alcool en générale.

‘‘‘

Nous allons désormais nous intéresser au lien qu'il peut y avoir entre la consommation d'alcool et la fréquence de sortie entre amis. Evidemment notre hypothèse est que l'analyse de ce critère va nous permettre de tirer quelques conclusions qui seront intéressantes pour notre problématique

```
t6<-df  %>% group_by(goout) %>% summarize(mean_ami = mean(Dalc+Walc)/2);
plot6<-ggplot(t6, aes(x = goout, y = mean_ami))+
               geom_point()+
               xlab("Sortie entre amis")+
               ylab("Consommation d'alcool")+
               ggtitle("Moyenne de consommation d'alcool en fonction des sorties entres amis");
plot6
```


```
t6count<-df %>% group_by(goout) %>% count();
t6count
```

```
## # A tibble: 5 × 2
##   goout     n
##   <int> <int>
## 1     1    23
## 2     2   103
## 3     3   130
## 4     4    86
## 5     5    53
```

L'analyse de ce graphique se fait très rapidement et facilement. Plus l'étudiant à tendance à sortir avec des amis, plus sa consommation d'alcool sera élevée. Notre hypothèse de départ est donc vraie. De plus en regardant les effectifs nous pouvons voir que la majorité des étudiants sortent moyennement (2-3). Il semblerait que leur consommation d'alcool se rapprochent de la consommation moyenne. Cependant les personnes qui sortent beaucoup (5) consomment beaucoup plus d'alcool que tous les autres groupes. Il peut être intéressant de regarder les chiffres de cette consommation.

```
t6<-df  %>% group_by(goout) %>% summarize(Conso_alcool_moyenne = mean(Dalc+Walc)/2);
t6
```

```
## # A tibble: 5 × 2
##   goout Conso_alcool_moyenne
##   <int>                <dbl>
## 1     1             1.391304
## 2     2             1.538835
## 3     3             1.726923
## 4     4             2.156977
```

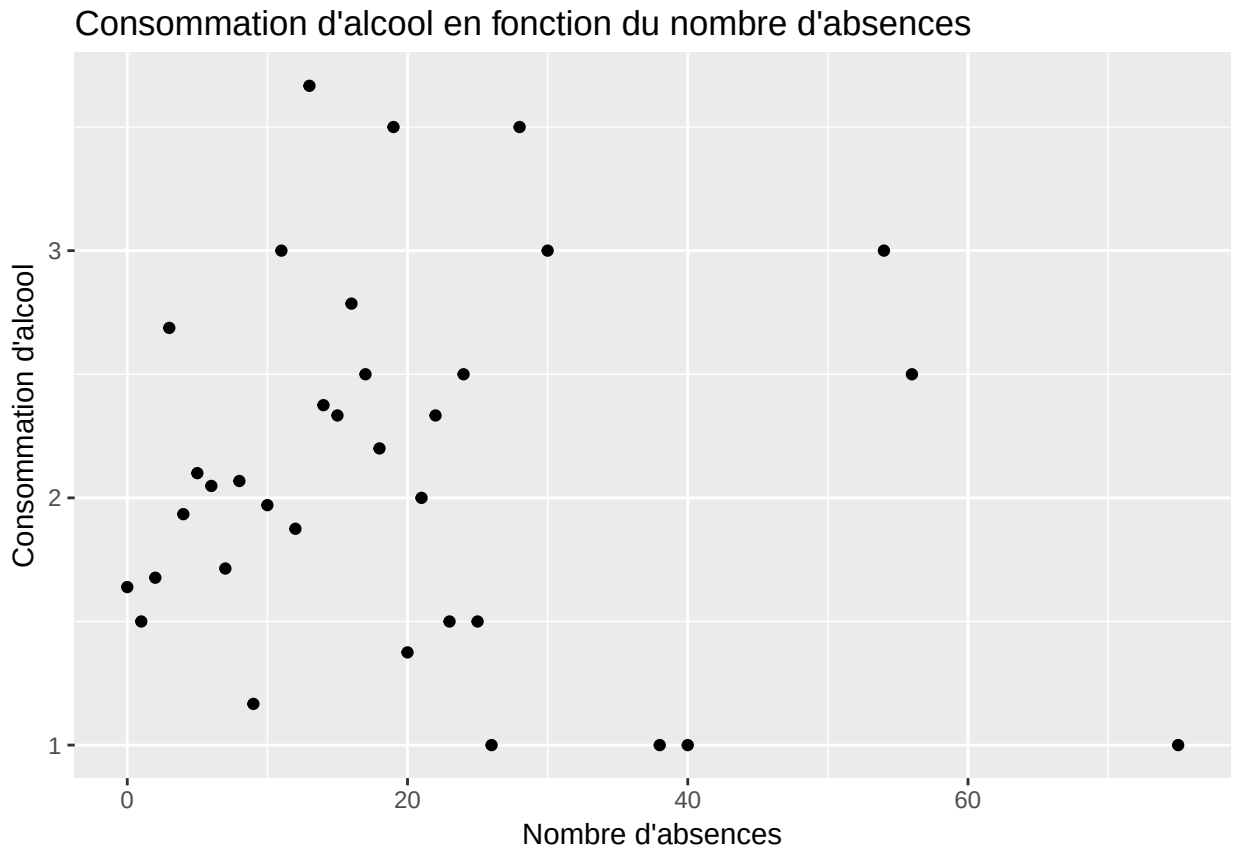
```
## 5     5              2.726415
```

Avec les valeurs nous voyons bien que la consommation d'alcool moyenne des étudiants qui sortent beaucoup de fois(4-5) est nettement supérieur à la consommation moyenne (1.8). On peut donc affirmer qu'il y a bien un lien entre le nombre de fois où un étudiant va sortir et sa consommation d'alcool.

Nous allons maintenant nous focaliser sur le nombre d'absences, ce critère complétant le nombre d'heures travaillées, le tout décrivant le sérieux d'une personne.

```
t7<-df  %>% group_by(absences) %>% summarize(mean_abs = mean(Dalc+Walc)/2);


plot7<-ggplot(t7, aes(x = absences, y = mean_abs))+
               geom_point()+
               xlab("Nombre d'absences")+
               ylab("Consommation d'alcool")+
               ggtitle("Consommation d'alcool en fonction du nombre d'absences");
plot7
```




```
t7count<-df  %>% group_by(absences) %>% filter(absences > 50)
t7count
```

```
## Source: local data frame [3 x 33]
## Groups: absences [3]
##
##   school    sex   age address famsize Pstatus  Medu  Fedu   Mjob     Fjob
##   <fctr> <fctr> <int>  <fctr>  <fctr>  <fctr> <int> <int> <fctr>   <fctr>
```

```
## 1     GP      F    16       U     GT3       T     3     3  other services
## 2     GP      F    17       U     LE3       T     3     3  other    other
## 3     GP      F    18       R     GT3       A     3     2  other services
## # ... with 23 more variables: reason <fctr>, guardian <fctr>,
## #   traveltime <int>, studytime <int>, failures <int>, schoolsup <fctr>,
## #   famsup <fctr>, paid <fctr>, activities <fctr>, nursery <fctr>,
## #   higher <fctr>, internet <fctr>, romantic <fctr>, famrel <int>,
## #   freetime <int>, goout <int>, Dalc <int>, Walc <int>, health <int>,
## #   absences <int>, G1 <int>, G2 <int>, G3 <int>
```

Le bilan que nous pouvoir faire de ce plot est le suivant. Nous aurions pu supposé que les étudiants les plus absents étaient ceux consommant le plus d'alcool. Notre idée de base se confirme, nous avons pu observer grace à la fonction filter() que 3 personnes ont un nombre d'absences supérieurs à 50 et que parmis ces personnes, 2 sont en très bonne santé et une en très mauvaise santé. Ce plot nous montre aussi une forte densité se trouvant entre 0 et 30 absences durant l'année avec des resultats très différents (la consomation d'alcool variant entre 1 et 5 dans cette zone). Dans cette zone on y trouve des personnes en bonne santé, d'autres non. Il est difficile de tirer des conclusions sur le sérieux d'une personne sans prendre en compte la maladie.

IV) Conclusion

En conclusion, si on résume l'aspect projet de ce travail, ce dernier nous a apporté de nombreuses choses. Ce projet nous a permi de travailler en groupe donc de pouvoir confronter nos points de vues et nos idées. De plus la perspective de pouvoir choisir son sujet et ne pas se le voir imposé permet d'être plus impliqué dans le projet du fait d'un sujet eveillant notre intéret. De plus il nous a permis d'utiliser le langage R et donc de revoir certaines notions.

Concernant le projet, nous pouvons tirer des conclusions de ce travail. En effet notre problématique était la suivante : le caractère d'une personne peut-il influencer le niveau de consommation d'alcool chez un étudiant ? A travers l'analyse de nos données (en grande partie à l'aide de l'analyse factorielle), en grande partie des données qualitatives, nous avons pu trouver des traits de caractères pouvant influer sur la consommation d'alcool. D'autre, comme le nombre d'absences pourrait avoir une influence mais seul celui-ci a très peu de sens (la santé influencant grandement ce critère).

V) Référence

https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION