Rapport CYS

26 mars, 2020

Contents

Etude des données

```
CYS = read.csv("DonneeS3_filtre.csv")
dif_snap=CYS$snapshot.2-CYS$snapshot.1
ratio_snap=CYS$snapshot.2/CYS$snapshot.1
summary(CYS)
```

```
##        prénom          Semestre   Filière     snapshot.1       snapshot.2
##  ALEXANDRE:  3   S3 2017-18: 38   EEA:181   Min.   : 1.000   Min.   : 1.75
##  ALEXIS   :  3   S3 2018-19:143             1st Qu.: 4.500   1st Qu.: 7.00
##  HUGO     :  3                              Median : 6.750   Median : 8.75
##  LUCAS    :  3                              Mean   : 6.442   Mean   : 8.82
##  NICOLAS  :  3                              3rd Qu.: 8.000   3rd Qu.:10.50
##  VINCENT  :  3                              Max.   :13.000   Max.   :15.75
##  (Other)  :163
##  Snapshot.2...4m CYS.S3    CYS.S4    TP.S3    TP.S4     CMI        Groupe.S3
##         :127     non:120      :137   FR:132     :143   non:150   Siuban :49
##  -      : 15     oui: 61   non: 21   GB: 49   FR: 15   oui: 31   Akane  :34
##  11     :  3               oui: 23            GB: 23             Alba   :29
##  11.75  :  3                                                     Nadia  :24
##  14.5   :  3                                                     Steven :23
##  6.75   :  3                                                     Yolanda:15
##  (Other): 27                                                     (Other): 7
##     Groupe.S4     Prof.TP
##          :154         :168
##  Nadia   : 15   Didier:  6
##  Virginia: 12   Pierre:  7
##
##
##
##
```

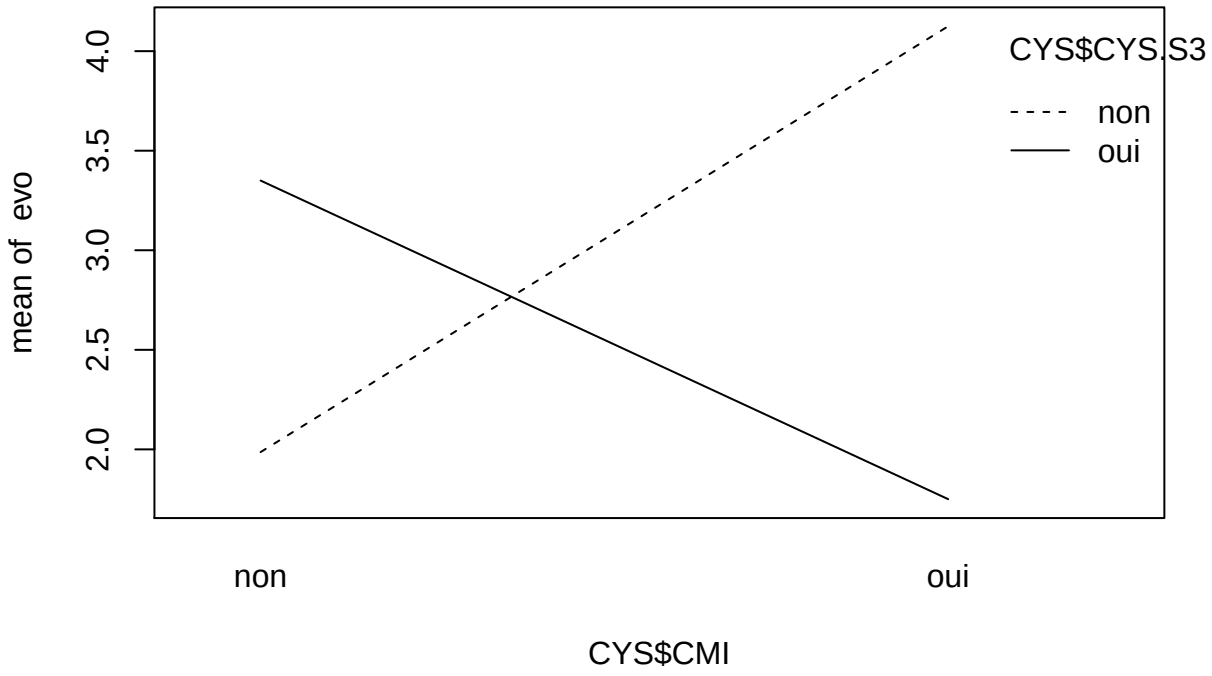
En fait, en semestre 3, on effectue un bilan compris de 181 étudiants dont:

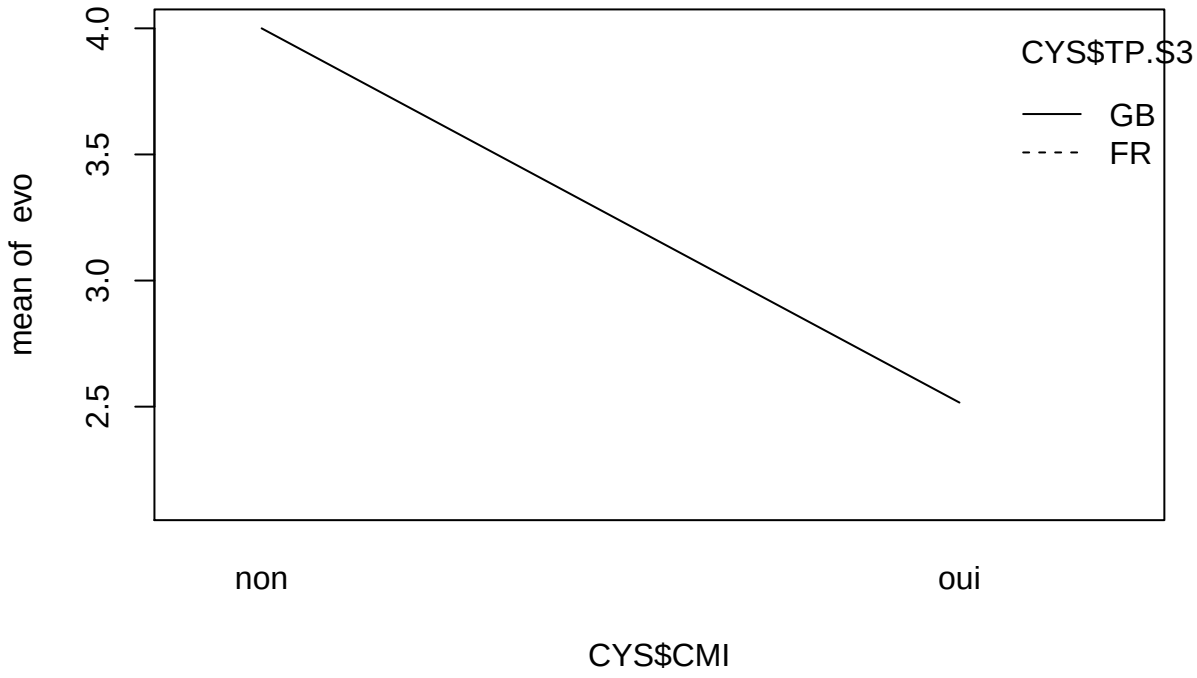
- 38 en 2017-2018 et 143 en 2018-2019

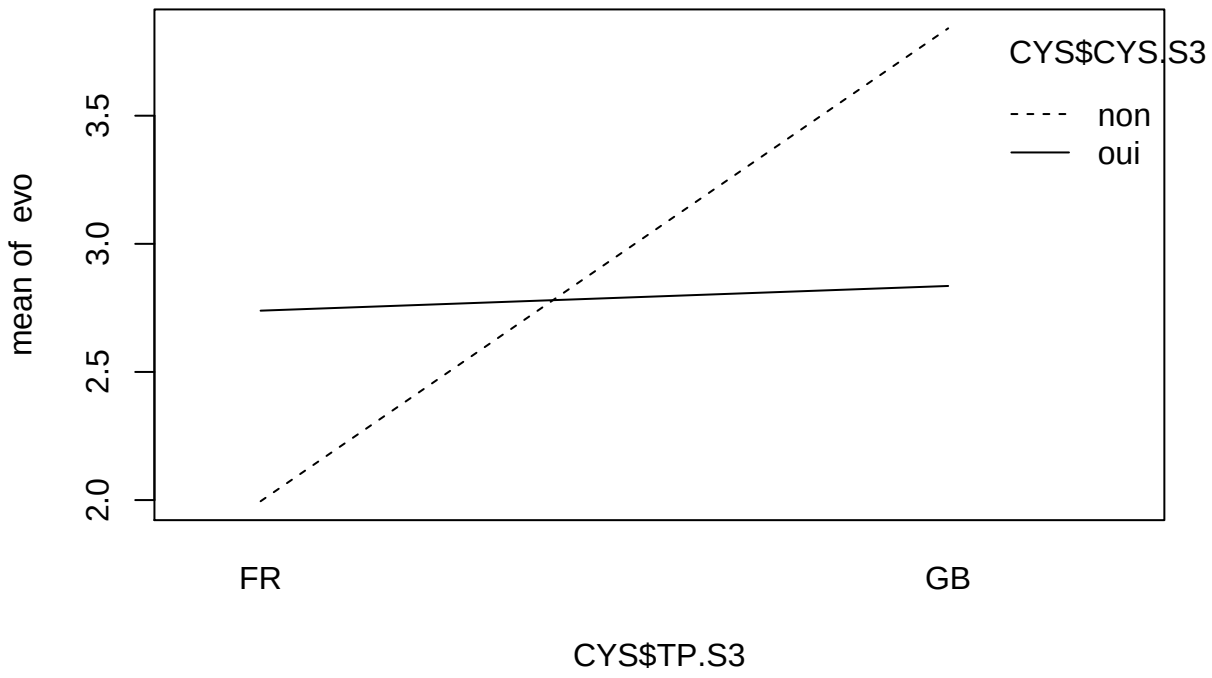
- Tout 181 en filière EEA
- 120 utilisent l'outil CYS et 41 ne l'utilisent pas
- 132 font des TPs en français et 49 les font en anglais
- 160 sont en CMI alors que 31 n'y sont pas

Tableau croisé de 2 variables CMI et TP.S3

```
##
##      non oui
##   FR 132   0
##   GB  18  31
```

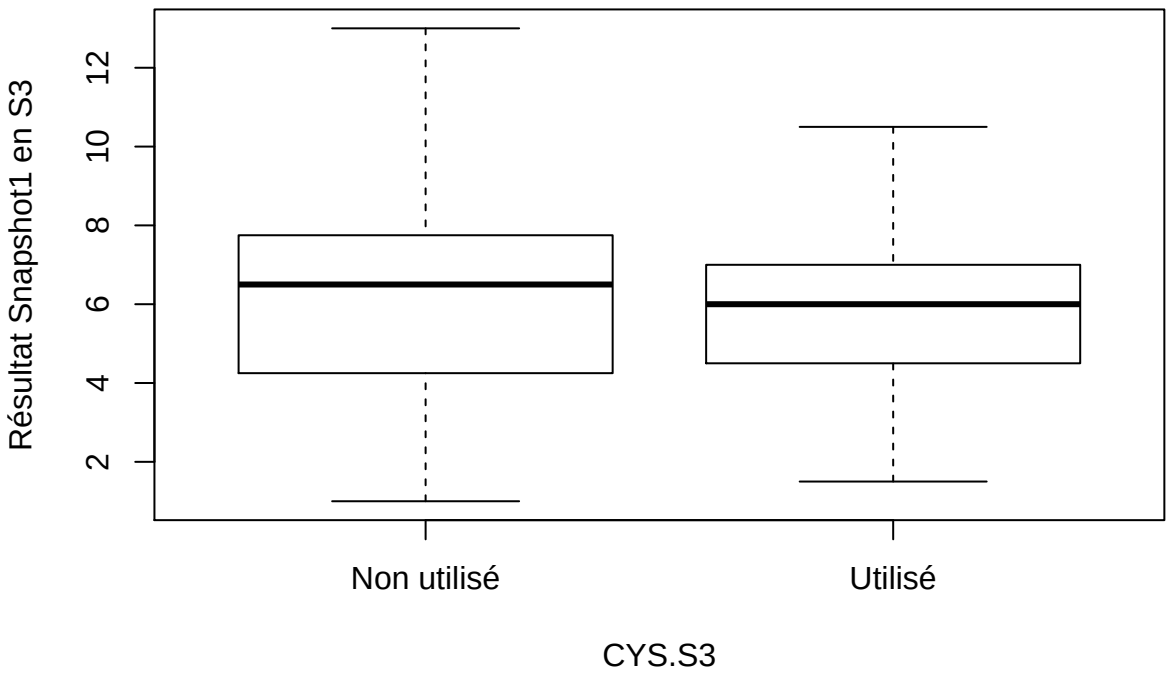


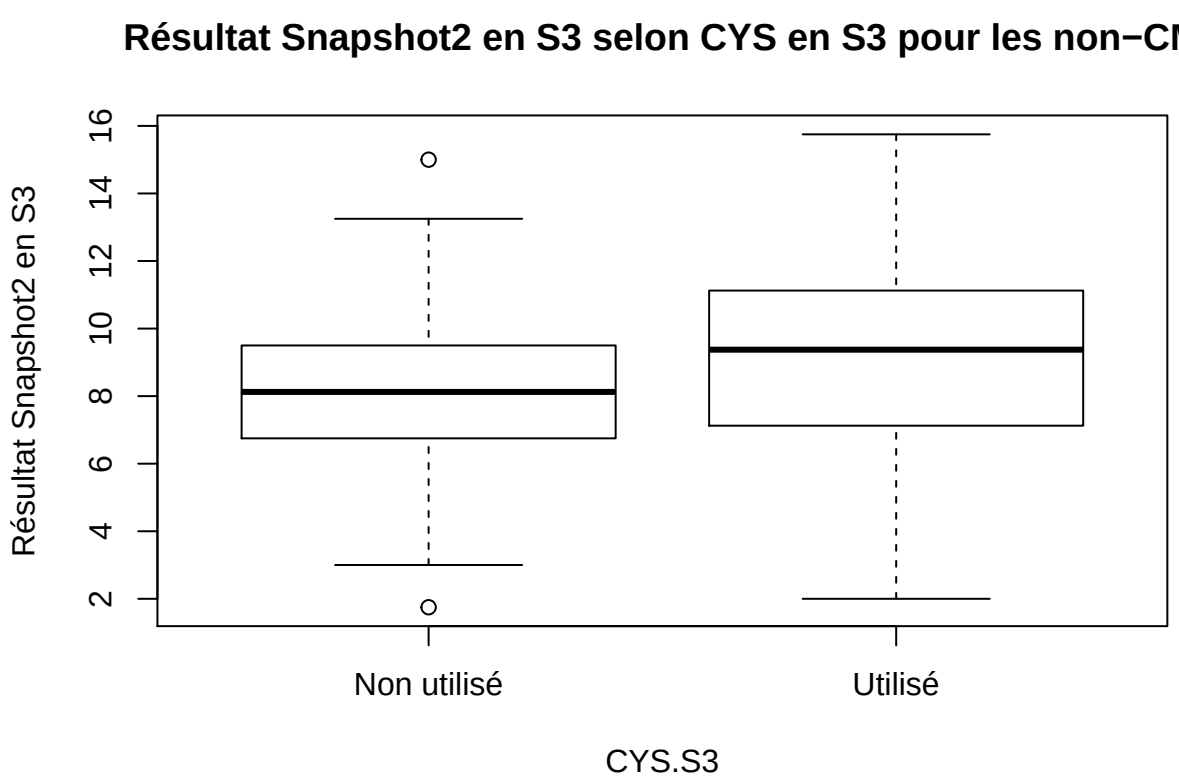

mean of evo
4.0
3.5
3.0
2.5
CYS$TP.$3
GB
FR
non
oui
CYS$CMI



On a trouvé que:

- L'interaction entre CYS et CMI est importante: Un étudiant utilisant CYS progresse moins s' il est en CMI. En revanche, un étudiant n'utilisant pas CYS progresse plus s'il est en CMI.
- L'interaction entre TP et CMI n'existe pas car il n'y a pas d'étudiant en CMI utilisant français pour les TPs.
- L'interaction entre CYS et CMI existe: Un étudiant utilisant CYS progresse un peu plus s' il pratique les TPs en anglais. Or, un étudiant n'utilisant pas CYS progresse bien plus s'il pratique les TPs en anglais.

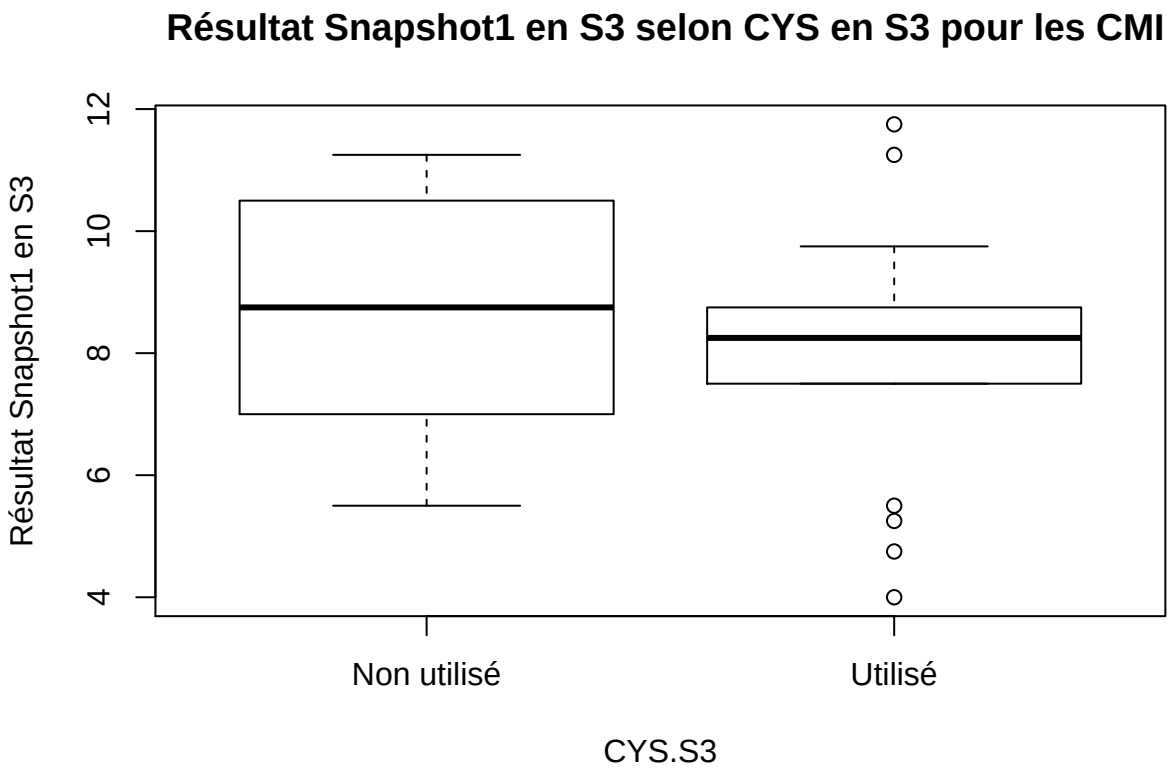

Résultat Snapshot1 en S3 selon CYS en S3 pour les non−CMI
Résultat Snapshot1 en S3
2
4
6
8
10
12
Non utilisé
Utilisé
CYS.S3

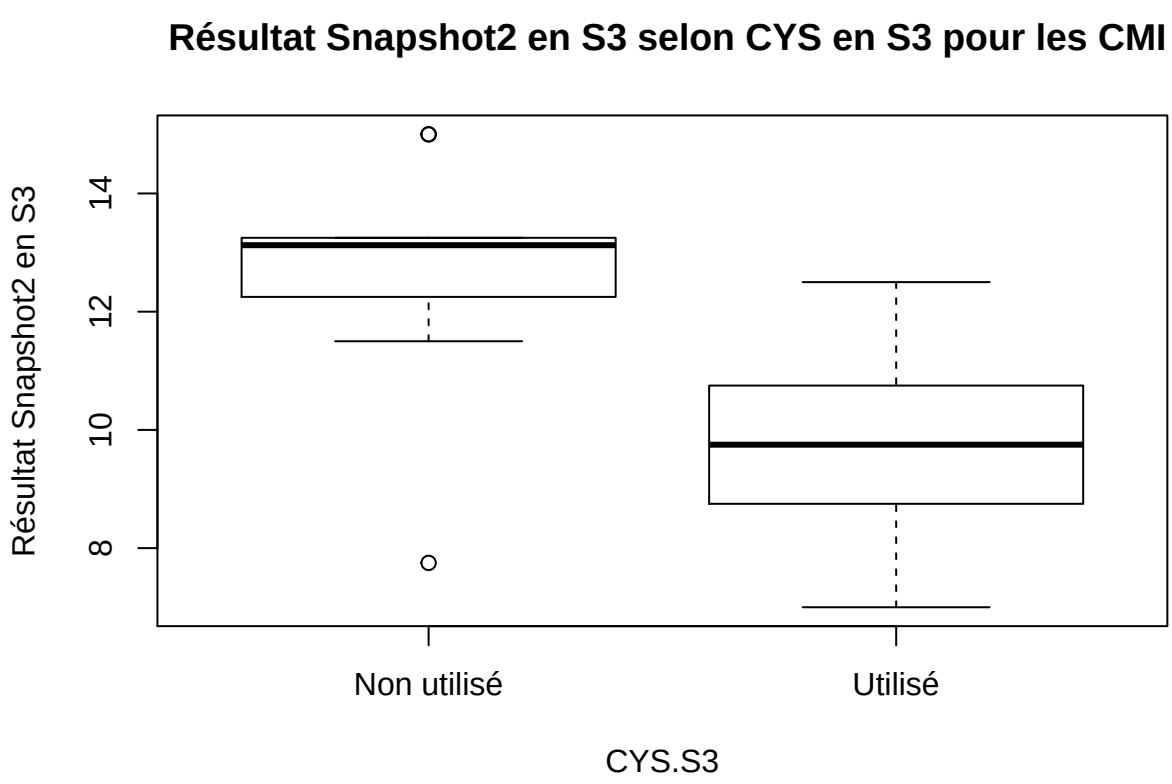


Selon les boxplots, on a trouvé le fait que:

- Généralement, un étudiant non en CMI a un note de Snapshot 1 plus élevé quand il n'utilise pas l'outil CYS.
- Pourtant,parmi les étudiant non en CMI, ceux utilisant CYS prend des notes de Snapshot 2 plus élevés que ceux ne l'utilisant pas.

On a observé ici une éfficaité de l'outil CYS appliqués sur les étudiant non en CMI

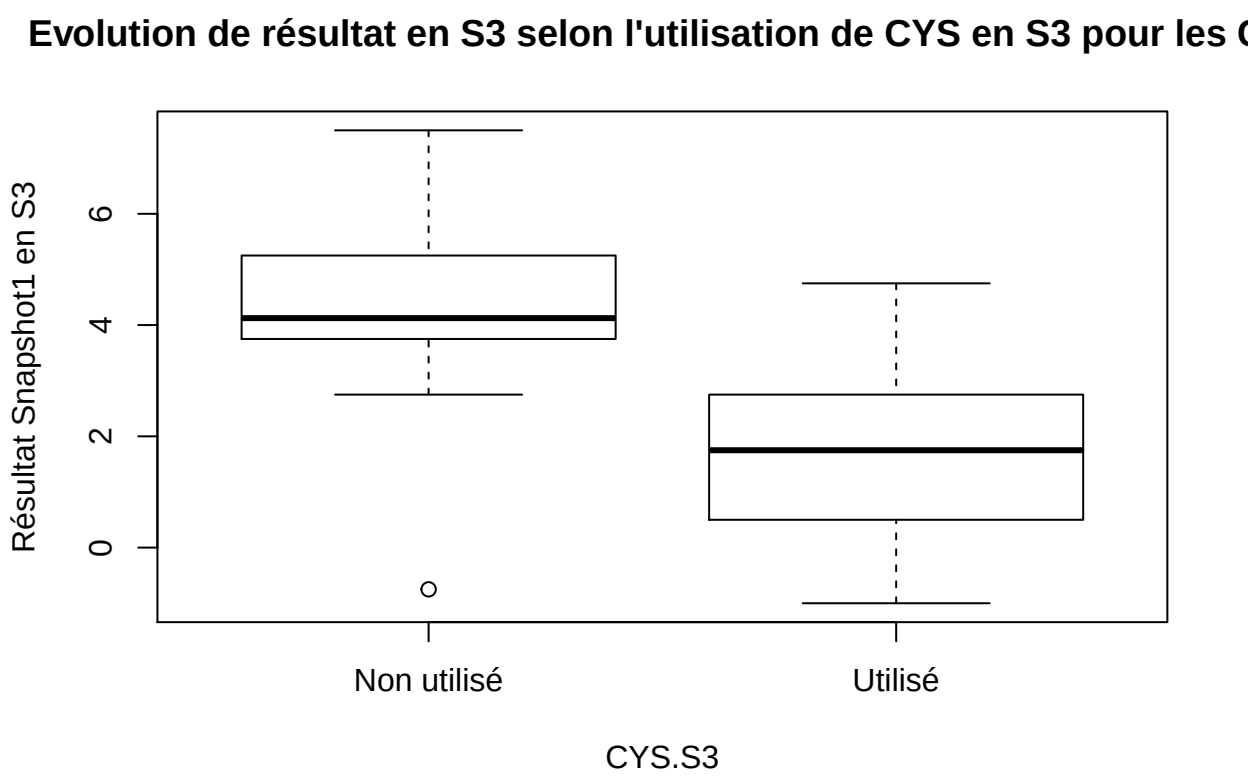

Résultat Snapshot1 en S3 selon CYS en S3 pour les CMI
Résultat Snapshot1 en S3
4
6
8
10
12
Non utilisé
Utilisé
CYS.S3



Selon les boxplots, on a trouvé le fait que:

- Généralement, un étudiant en CMI a un note de Snapshot 1 moins élevé quand il n'utilise pas l'outil CYS.
- Parmi les étudiant non en CMI, ceux utilisant CYS prend des notes de Snapshot 2 moins élevés que ceux ne l'utilisant pas.

On va étudier pour ce cas l'évolution de résultat



Ainsi, les étudiant en CMI progressent mieux quand ils n'utilisent pas CYS.

Modèle linéaire

On commence à mener des modèle ANOVA pour étudier l'impact des facteurs CMI, CYS et la langue de TP sur l'évolution de résultat entre Snapshot 1 et Snapshot 2 en semestre 3

Modèle ANOVA à 3 facteurs

```
mod1=lm(dif_snap~(CYS$CYS.S3+CYS$TP.S3+ CYS$CMI)^2,data=CYS)
#step.backward = step(mod1)
modBIC1=lm(dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,data=CYS)
summary(modBIC1)
```

```
##
## Call:
## lm(formula = dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,
##     data = CYS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.875 -1.600  0.000  1.264  5.400
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.9864     0.2068   9.606  < 2e-16 ***
```

```
## CYS$CYS.S3oui              1.3636     0.4005   3.405 0.000818 ***
## CYS$CMIoui                 2.1386     0.7164   2.985 0.003233 **
## CYS$CYS.S3oui:CYS$CMIoui  -3.7386     0.9245  -4.044 7.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.169 on 177 degrees of freedom
## Multiple R-squared:  0.1009, Adjusted R-squared:  0.08569
## F-statistic: 6.623 on 3 and 177 DF,  p-value: 0.000289
```

```
# A completer
mod2=lm(ratio_snap~(CYS$CYS.S3+CYS$TP.S3+ CYS$CMI)^2,data=CYS)
#step.backward = step(mod2,direction="backward",k=log(nrow(CYS)))
modBIC2=lm(ratio_snap ~ CYS$CMI,data=CYS)
summary(modBIC2)
```

```
##
## Call:
## lm(formula = ratio_snap ~ CYS$CMI, data = CYS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8666 -0.4505 -0.1435  0.1981  3.4334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.56657    0.05769  27.153   <2e-16 ***
## CYS$CMIoui  -0.20431    0.13941  -1.466    0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7066 on 179 degrees of freedom
## Multiple R-squared:  0.01186,    Adjusted R-squared:  0.006336
## F-statistic: 2.148 on 1 and 179 DF,  p-value: 0.1445
```

En applquant le critère BIC sur le modèle complet, on trouve le modèle “modBIC2”, où on ne trouve pas l’impacte de la langue de TP:

- Quand on considère la différence entre 2 snapshots on obitent le modèle modBIC1 sous lequel cette différence dépend de l’utilisation de CYS, le fait d’étre en CMI et un terme d’interaction entre ces deux derniers.
- Quand on considère la ratio entre 2 snapshots on obitent le modèle modBIC2 sous lequel cette ratio ne dépend que du fait d’étre en CMI .

Pourtant on a les valeurs R-ajustées trop petites (inférieure à 0.1) donc on a décidé d’aller plus loin vers le modèle ANCOVA

Modèle ANCOVA de 3 facteurs qualitatives (CYS S3, CMI, TP S3) et 1 facteur quantitative(Snapshot1)

```
##
## Call:
## lm(formula = CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 +
##     CYS$CMI + CYS$CYS.S3:CYS$CMI, data = CYS)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.8965 -1.2725 -0.0205  1.1728  5.8232
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.62384    0.44064  10.493  < 2e-16 ***
## CYS$snapshot.1             0.56912    0.06511   8.740 1.85e-15 ***
## CYS$TP.S3GB                1.90361    0.59802   3.183  0.00172 **
## CYS$CYS.S3oui              0.43506    0.43744   0.995  0.32132
## CYS$CMIoui                 1.28154    0.86629   1.479  0.14084
## CYS$CYS.S3oui:CYS$CMIoui -3.02652    0.85983  -3.520  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.932 on 175 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4615
## F-statistic: 31.85 on 5 and 175 DF,  p-value: < 2.2e-16
```

Selon le test d'BIC, on trouve le meilleur modèle modBIC2:

$CYS\$snapshot.2 \sim CYS\$snapshot.1 + CYS\$TP.S3 + CYS\$CYS.S3 + CYS\$CMI + CYS\$CYS.S3 : CYS\$CMI$

En fait, on veut estimer le résultat de Snapshot2 par le modèle:

$$(modbest) : Snapshot2_{ijkl} = \mu + \alpha Snapshot1_{ijkl} + \beta_i + \gamma_j + \theta_k + \delta_{jk} + \varepsilon_{ijkl}, \forall i = 1,\ 2,\ \forall j = 1, 2, \forall k = 1, 2$$

où:

i,j,k sont les indices de modalité pour les variables qualitatives TP.S3, CYS.S3 et CMI, respectivement.(1 pour la réponse Non et 2 pour la réponse Oui, dans le cas de TP 1 pour FR et 2 pour GB)

L'indice ijkl est pour indiquer l'individu l-ième ayant des modalités i,j,k pour TP.S3, CYS.S3 et CMI, respectivement. ε_{ijkl} est des erreurs de l'estimation de l'individu ayant l'indice ijkl.

D'où:

$$\mu = 4.62384 \alpha = 0.56912 \beta_1 = \gamma_1 = \theta_1 = \delta_{11} = \delta_{12} == \delta_{21} = 0 \beta_2 = 1.90361 \gamma_2 = 0.43506 \delta_{22} = -3.02652$$

On a décidé de modéliser la note de Snapshot2 en fonction de Snapshot1, l'utilisation de l'outil CheckYourSmile, la langue de TP et le fait que l'étudiant est en CMI ou pas.

Alors, sous le modèle ANCOVA on a trouvé que les trois facteurs qualitatives ont des impactes sur le résultat Snapshot2. Or, le Snaphot1 a un gros effet sur le résultat de Snapshot2. On y trouve aussi un terme d'interaction entre la variable CYS.S3 et la variable CMI et celui dernier a un effet important négatif sur le Snapshot2.

ie, un étudiant en CMI utilisant l'outil Check Your Smile a tendance de dégrader environ 2,6 points (-3,02652+0,43506) et un étudiant non CMI utilisant l'outil Check Your Smile a tendance de progresser environ 0,4 points (0,43506)

On voit que le modèle ANCOVA nous donne un R-ajusté bien meilleur que le modèle ANOVA. (0,4615»0,006 et 0,4615»0,09)

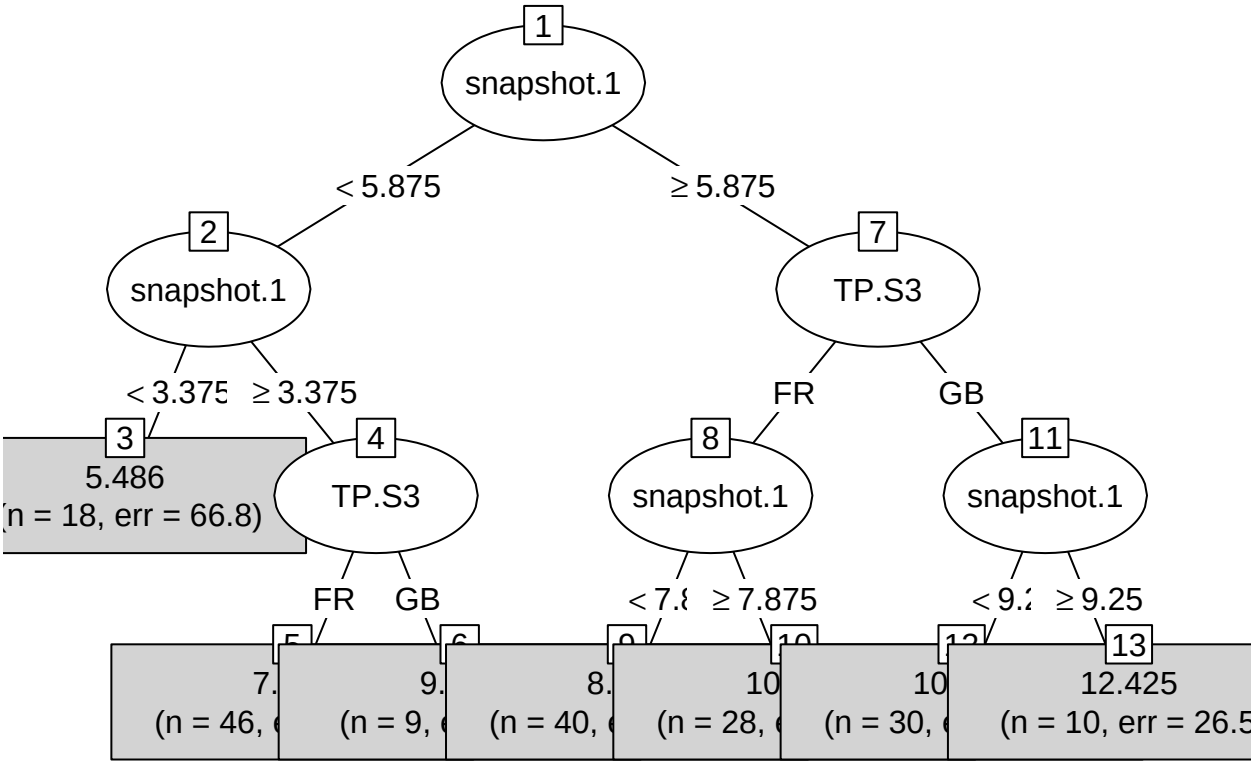
Pour le but d'obtenir une valeur de R plus élevé on va passer sous des modèles non linéaires

Erreur de validation croisée:

```
## [1] 69.78652
```

Arbre binaire de décision

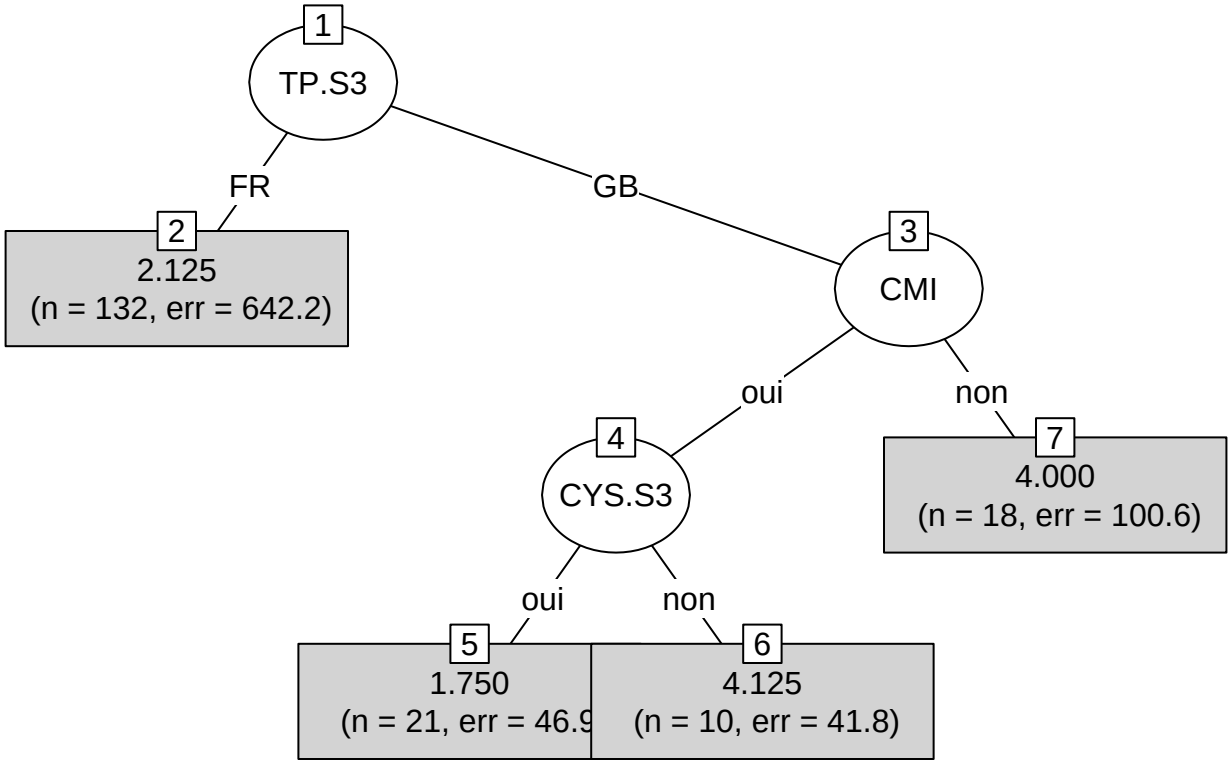
En cas on veut construire un arbre de régression de la note Snapshot 2 en fonction d'autres variables



Erreur de validation croisée:

```
## [1] 80.01503
```

En cas on veut construire un arbre de régression de l'évolution de la note en fonction d'autres variables



Erreur de validation croisée:

```
## [1] 86.3406
```

On a essayé plusieurs fois pour étudier le comportement des erreurs de validation croisée de ces deux types de l'arbre et a gardé le premier donc on ne voit pas l'impacte de CYS.

Lorsau'on prend le second:

- Parmi les étudiants faisant les TP en anglais, un étudiant utilisant l'outil CYS progresse moins qu'un autre n'utilisant pas CYS. (cf des feuilles 5,6,7). De plus, parmi les étudiants faisant les TP en anglais et utilisant l'outil CYS, un étudiant en CMI progresse moins qu'un autre non CMI.
- L'arbre binaire de régression nous permet d'observer l'effet de l'outil CYS dans la progression des étudiants n'est pas remarquable.
- L'erreur de validation croisé du modèle ANCOVA est plus petite que celle du l'arbre de décision. Pourtant, on trouve le même phénomène pour les CMI sur l'effet de CYS sur la progression des étudiants.