Geoducation

Antoine Drouhin, Aurélien Garret, Cécile Hu, Lucas Morel

TODO, Note/CR de Réunion avec la prof :

- décrire la base de données : taille, variable quanti ou quali (à faire pour la présentation orale)
- Ségrégation scolaire
- Idée :
- Essayer de faire une classif sur les différents tx de réussite et les filières (pas étonné qu'en ts gros tx de réussite, qu'en L non etc,...) Essayer de voir pourquoi meilleur on est meilleur on sera et vice versa.
- Choisir un indcateur de réussite, créer un indictauer en addtionnant les tx de réussite par lycée, puis régréssion pouvoir si dépend de la filière et de la géographie.
- Secteur pivéé/public
- Rural/urbain
- Puis régréssion synthétique
- Cherche taille des communes pour joindre

Introduction

Base de données

Notre de base de données à été trouvée sur le site Data.gouv. Nous avons croisé deux jeux de données distincts. Le premier concerne des données sur la performance des lycées en France (taux de réussite etc). Le second présente des données géographiques pour l'ensemble des établissement scolaires francais (Coordonnées GPS, etc.).

https://www.data.gouv.fr/fr/

Nous avons choisi cette base car elle présente une approche intéressante sur la compréhension d'un sujet qui nous concerne tous, l'éducation. L'approche géographique des question de réussite scolaire nous a semblé être un champs suffisamment complet pour permettre l'utilisation des méthodes d'analyse vue en cours.

L'intérêt et le questionnement général porte sur la réussite scolaire de la France en fonction de la position géographique des établissements scolaires. Y'a t-il une corrélation entre la situation géographique des établissements et la réussite scolaire des étudiants ?

Plusieurs problématiques en découlent comme :

Y'a t-il des différences de réussites entre le top 10 des grandes villes en France et les villes de province ? Paris bénéficie-t-elle d'une réussite supérieure au reste de la France ? Quelles régions de France semble réussir mieux que les autres ?

Initialisation de la base de donnée

```
read.csv2("geoducation-data2.csv", sep=";", header=TRUE, na.strings = "")->bdd
#exists('bdd')
```

Khi-Deux

```r
bddKhiDeux = bdd[, c('Académie','Effectif.Présents.série.L','Effectif.Présents.série.ES','Effectif.Prés
# Petit clean des datas (Antoine)
bddKhiDeux[is.na(bddKhiDeux)] <- 0

# Cette portion de code supporse que bdd est ordonné par nom d'académie. (Antoine)

tableKhiDeux <- data.frame()
indiceCourant = 1
aca <- bddKhiDeux[1,"Académie"]

for(i in 1:nrow(bddKhiDeux)){
  if(aca != bddKhiDeux[i,"Académie"]){
    indiceCourant = indiceCourant + 1
    aca <- bddKhiDeux[i,"Académie"]
  }
  if (length(rownames(tableKhiDeux)) != 0 && !is.na(tableKhiDeux[indiceCourant,"Académie"]) && bddKhiDe
    tableKhiDeux[indiceCourant,"ES"] <- tableKhiDeux[aca,"ES"] + bddKhiDeux[i,"Effectif.Présents.série.
    tableKhiDeux[indiceCourant,"L"] <- tableKhiDeux[aca,"L"] + bddKhiDeux[i,"Effectif.Présents.série.L"]
    tableKhiDeux[indiceCourant,"S"] <- tableKhiDeux[aca,"S"] + bddKhiDeux[i,"Effectif.Présents.série.S"]
  } else {
    tableKhiDeux <- rbind(tableKhiDeux, data.frame(Académie = aca,ES = bddKhiDeux[i,"Effectif.Présents.
  }
}
```

Ici on a crée notre tableau pour effectuer notre test du Khi Deux (nous deux)

```r
print(tableKhiDeux)
```

```
##            Académie    ES     S    L
## 1     AIX-MARSEILLE  4509  7678 2311
## 2           AMIENS  2744  4651 1530
## 3         BESANCON  1721  3099  767
## 4         BORDEAUX  5003  8344 2632
## 5             CAEN  2360  3614 1353
## 6  CLERMONT-FERRAND  1797  2923 1118
## 7            CORSE   359   565  301
## 8          CRETEIL  7065 11182 3253
## 9            DIJON  2263  4053 1231
## 10         GRENOBLE  5856  9084 2491
## 11       GUADELOUPE   755  1248  570
## 12           GUYANE   328   438  274
## 13       LA REUNION  1300  2441  985
## 14            LILLE  6128 11026 2726
## 15          LIMOGES   830  1691  617
## 16            LYON  5374  8980 2149
## 17       MARTINIQUE   604  1021  418
## 18          MAYOTTE   577   408  395
## 19      MONTPELLIER  3633  6687 2268
## 20       NANCY-METZ  3349  6085 1578
## 21           NANTES  5841  9111 2921
## 22            NICE  3344  5504 1696
## 23    ORLEANS-TOURS  3756  6253 1904
```

```
## 24            PARIS  4556  7358 2535
## 25         POITIERS  2596  3913 1509
## 26            REIMS  1909  3448  955
## 27           RENNES  5707  8741 2426
## 28            ROUEN  2996  4805 1606
## 29       STRASBOURG  3102  5056 1124
## 30         TOULOUSE  4249  8152 2265
## 31       VERSAILLES 11720 17807 4720
```

On procede maintenant aux étapes du khi deux :

1 Calcul des effectifs théoriques

```
khideux <- chisq.test(tableKhiDeux[,c('S','ES','L')])
print(khideux)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tableKhiDeux[, c("S", "ES", "L")]
## X-squared = 1789.1, df = 60, p-value < 2.2e-16
```

La probabilité que la situation géographique soit indépendante de la répartition dans les différentes filiéres est donc inférieure à 2.2e-16.

Régréssion

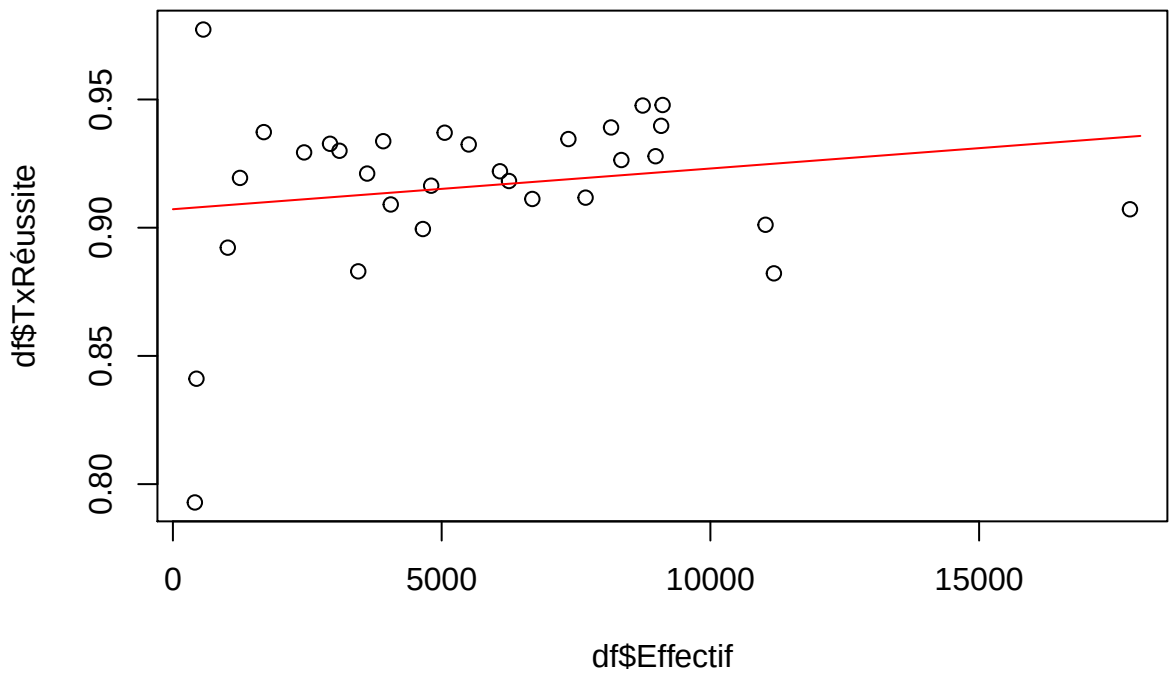
Est ce que le taux de réussite des èlèves en terminale S s'explique par la localisation académique ?

Commentaires Aurélien : A voir si la moyenne du tx de réussite c'est le bon truc, regarder et analyser les data et faire le bon choix de l'incateur.

```
bddReg = bdd[, c('Académie','Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S')]
bddReg[is.na(bddReg)] <- 0
bddReg[3] <- bddReg[3]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3])
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0),]

regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Académie), FUN=sum)
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Académie), FUN=mean)[2])


df = data.frame(regData[1], regData[2], regData[3])
col_headings <- c('Académie','Effectif', 'TxRéussite')
names(df) <- col_headings


model<-lm(df$TxRéussite~df$Effectif, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = df$TxRéussite ~ df$Effectif, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.114975 -0.011373  0.006419  0.018627  0.069191
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.072e-01  1.079e-02  84.058   <2e-16 ***
## df$Effectif 1.588e-06  1.581e-06   1.004    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03361 on 29 degrees of freedom
## Multiple R-squared:  0.0336, Adjusted R-squared:  0.0002809
## F-statistic: 1.008 on 1 and 29 DF,  p-value: 0.3236
```

```
plot(df$Effectif,df$TxRéussite)
x <- seq(0,18000)
lines(x,x*1.588e-06+9.072e-01,col="red")
```



Conclusion Aurélien: Tracer la droite

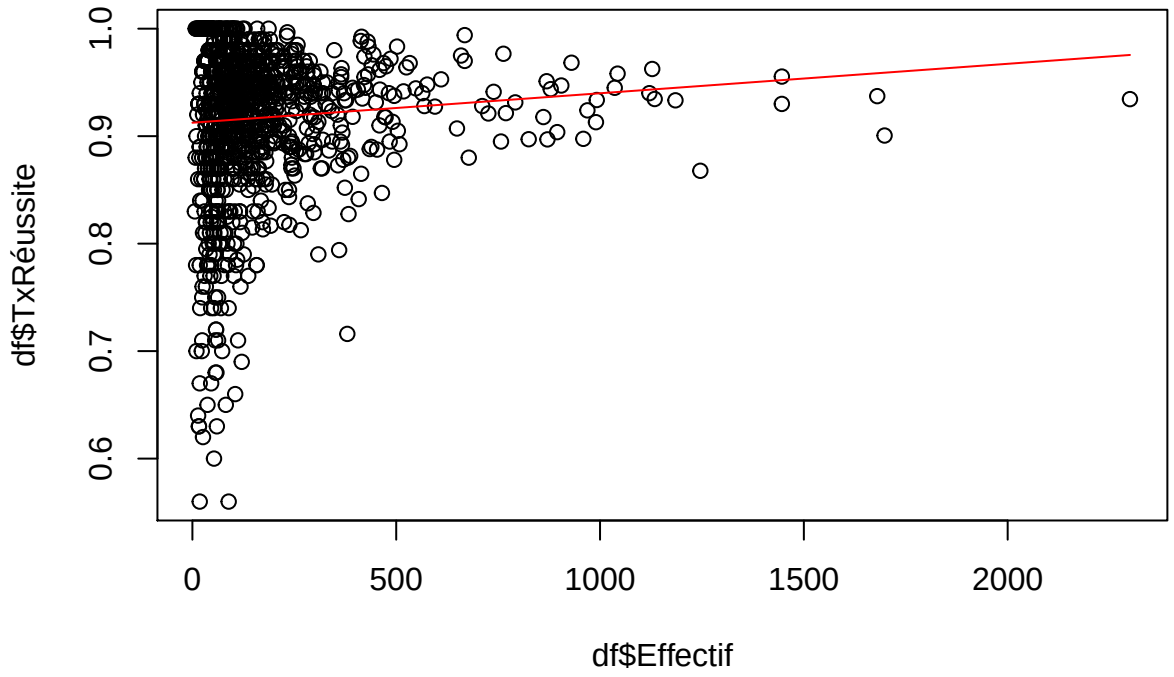
Est ce que le taux de réussite des èlèves en terminale S s'explique par la localisation des communes ?

```
bddReg = bdd[, c('Ville','Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S')]
bddReg[is.na(bddReg)] <- 0
```

```r
bddReg[3] <- bddReg[3]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3])
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0),]

regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Ville), FUN=sum)
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Ville), FUN=mean)[2])


df = data.frame(regData[1], regData[2], regData[3])
col_headings <- c('Ville','Effectif', 'TxRéussite')
names(df) <- col_headings
plot(df$Effectif,df$TxRéussite)
model<-lm(df$TxRéussite~df$Effectif, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = df$TxRéussite ~ df$Effectif, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35496 -0.02803  0.01452  0.04529  0.08726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.125e-01  2.640e-03 345.705  < 2e-16 ***
## df$Effectif 2.741e-05  1.025e-05   2.674  0.00761 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06718 on 1076 degrees of freedom
## Multiple R-squared:  0.0066, Adjusted R-squared:  0.005677
## F-statistic: 7.149 on 1 and 1076 DF,  p-value: 0.007613
```

```r
plot(df$Effectif,df$TxRéussite)
x <- seq(0,2300)
lines(x,x*2.741e-05+9.125e-01,col="red")
```



Tend vers zero donc non !

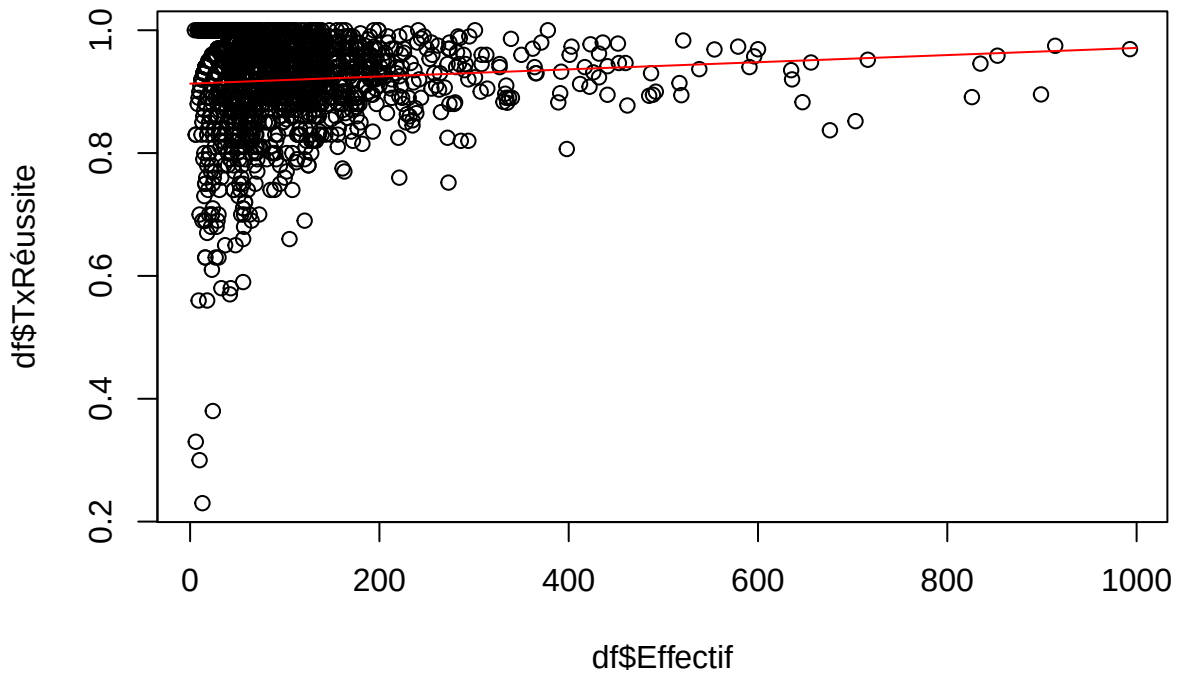
Est ce que le taux de réussite des èlèves en terminale S s'explique par les lycées dans lesquels les cours ont été suivis ?

```
bddReg = bdd[, c('Etablissement','Effectif.Présents.série.S','Taux.Brut.de.réussite.série.S')]
bddReg[is.na(bddReg)] <- 0
bddReg[3] <- bddReg[3]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3])
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0),]

regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Etablissement), FUN=sum)
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Etablissement), FUN=mean)[2])


df = data.frame(regData[1], regData[2], regData[3])
col_headings <- c('Ville','Effectif', 'TxRéussite')
names(df) <- col_headings
plot(df$Effectif,df$TxRéussite)
model<-lm(df$TxRéussite~df$Effectif, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = df$TxRéussite ~ df$Effectif, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.68382 -0.03177  0.01528  0.05293  0.08665
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.131e-01  2.783e-03 328.098  < 2e-16 ***
## df$Effectif 5.812e-05  1.862e-05   3.121  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07738 on 1620 degrees of freedom
## Multiple R-squared:  0.005977,   Adjusted R-squared:  0.005364
## F-statistic: 9.742 on 1 and 1620 DF,  p-value: 0.001833
```

```
plot(df$Effectif,df$TxRéussite)
x <- seq(0,1000)
lines(x,x*5.812e-05+9.131e-01,col="red")
```



Tend vers 0 donc non aussi !

p-value = risque d'érreur, probabilité d'érreur Quand très petite on a peu de chance de se tromper en infirmant qu'il y a lien entre les deux variables.

Il y a un risque de 7 pour 1000 de me tromper quand je dis que les effectifs par ville explique le tx de réussite sachant que le lien reste faible.

rajouter des variables explicatives

Est ce que les taux de réussite S, ES et L de chaque Acamédie explique le taux de réussite général de chaque académie ?

```
bddReg = bdd[, c('Académie',
                 'Taux.Brut.de.réussite.série.S',
```

```
                'Taux.Brut.de.réussite.série.ES',
                'Taux.Brut.de.réussite.série.L')]
bddReg[is.na(bddReg)] <- 0
bddReg[2] <- bddReg[2]/100
bddReg[3] <- bddReg[3]/100
bddReg[4] <- bddReg[4]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3],bddReg[4])
df<-df[ (df$Taux.Brut.de.réussite.série.S>0
         & df$Taux.Brut.de.réussite.série.ES>0
         & df$Taux.Brut.de.réussite.série.L>0),]

regData = c(aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Académie), FUN=mean))
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.ES, by=list(df$Académie), FUN=mean)[2])
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.L, by=list(df$Académie), FUN=mean)[2])

df = data.frame(regData[1], regData[2], regData[3], regData[4])
df[5] <- mean(sum(df[2], df[3], df[4]) / 100)
col_headings <- c('Ville','TxRéussite_S','TxRéussite_ES','TxRéussite_L', 'TxRéussiteGlobal' )
names(df) <- col_headings

df
```

```
##               Ville TxRéussite_S TxRéussite_ES TxRéussite_L
## 1     AIX-MARSEILLE    0.9135802     0.8970370    0.9165432
## 2            AMIENS    0.8985417     0.8987500    0.9318750
## 3          BESANCON    0.9327586     0.9248276    0.9251724
## 4          BORDEAUX    0.9265432     0.9243210    0.9300000
## 5              CAEN    0.9295652     0.9189130    0.9330435
## 6  CLERMONT-FERRAND    0.9307895     0.9284211    0.9207895
## 7             CORSE    0.9790000     0.9590000    0.9730000
## 8           CRETEIL    0.8797436     0.9037607    0.8864103
## 9             DIJON    0.9050000     0.9325000    0.9120455
## 10         GRENOBLE    0.9401064     0.9393617    0.9376596
## 11       GUADELOUPE    0.9187500     0.8975000    0.9037500
## 12           GUYANE    0.8411111     0.9022222    0.8988889
## 13       LA REUNION    0.9293548     0.9332258    0.9109677
## 14            LILLE    0.9116484     0.9082418    0.9282418
## 15          LIMOGES    0.9493333     0.9066667    0.9380000
## 16             LYON    0.9282955     0.9190909    0.9242045
## 17       MARTINIQUE    0.8914286     0.9407143    0.9150000
## 18          MAYOTTE    0.7850000     0.6850000    0.6266667
## 19      MONTPELLIER    0.9071186     0.9062712    0.9191525
## 20       NANCY-METZ    0.9218966     0.9117241    0.9243103
## 21           NANTES    0.9461682     0.9479439    0.9448598
## 22             NICE    0.9306383     0.9348936    0.9353191
## 23    ORLEANS-TOURS    0.9140000     0.9343333    0.9386667
## 24            PARIS    0.9208571     0.9462857    0.9432857
## 25         POITIERS    0.9358140     0.9281395    0.9197674
## 26            REIMS    0.8908824     0.9085294    0.9044118
## 27           RENNES    0.9501099     0.9364835    0.9349451
## 28            ROUEN    0.9221277     0.9153191    0.9240426
## 29       STRASBOURG    0.9440000     0.9462500    0.9632500
## 30         TOULOUSE    0.9403409     0.9234091    0.9413636
```

```
## 31       VERSAILLES    0.9147403     0.9288312    0.9237662
##    TxRéussiteGlobal
## 1         0.8524661
## 2         0.8524661
## 3         0.8524661
## 4         0.8524661
## 5         0.8524661
## 6         0.8524661
## 7         0.8524661
## 8         0.8524661
## 9         0.8524661
## 10        0.8524661
## 11        0.8524661
## 12        0.8524661
## 13        0.8524661
## 14        0.8524661
## 15        0.8524661
## 16        0.8524661
## 17        0.8524661
## 18        0.8524661
## 19        0.8524661
## 20        0.8524661
## 21        0.8524661
## 22        0.8524661
## 23        0.8524661
## 24        0.8524661
## 25        0.8524661
## 26        0.8524661
## 27        0.8524661
## 28        0.8524661
## 29        0.8524661
## 30        0.8524661
## 31        0.8524661
```

```
model<-lm(df$TxRéussiteGlobal ~ df$TxRéussite_S + df$TxRéussite_ES + df$TxRéussite_L, data = df)
summary(model)
```

```
## Warning in summary.lm(model): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = df$TxRéussiteGlobal ~ df$TxRéussite_S + df$TxRéussite_ES +
##     df$TxRéussite_L, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.781e-16 -6.882e-17 -3.574e-17  3.610e-18  1.113e-15
##
## Coefficients:
##                     Estimate Std. Error    t value Pr(>|t|)
## (Intercept)       8.525e-01  1.333e-15  6.394e+14   <2e-16 ***
## df$TxRéussite_S  -3.877e-16  2.136e-15 -1.810e-01    0.857
```

```
## df$TxRéussite_ES -4.499e-15  3.182e-15 -1.414e+00    0.169
## df$TxRéussite_L   3.690e-15  2.814e-15  1.312e+00    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.257e-16 on 27 degrees of freedom
## Multiple R-squared:  0.5118, Adjusted R-squared:  0.4575
## F-statistic: 9.434 on 3 and 27 DF,  p-value: 0.0001968
```