

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 addionnant les tx de réussite par lycée, puis régréssion pouvoir si dépend de la filière et de la géographie.
- Secteur pivée/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 français (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

```

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

# Cette portion de code suppose 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"]) && bddKhiDeux[i,"Académie"] != aca){
    tableKhiDeux[indiceCourant,"ES"] <- tableKhiDeux[aca,"ES"] + bddKhiDeux[i,"Effectif.Présents.série.ES"]
    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.série.ES"],L = bddKhiDeux[i,"Effectif.Présents.série.L"],S = bddKhiDeux[i,"Effectif.Présents.série.S"]))
  }
}

```

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

```
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 filières est donc inférieure à 2.2e-16.

Régression

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

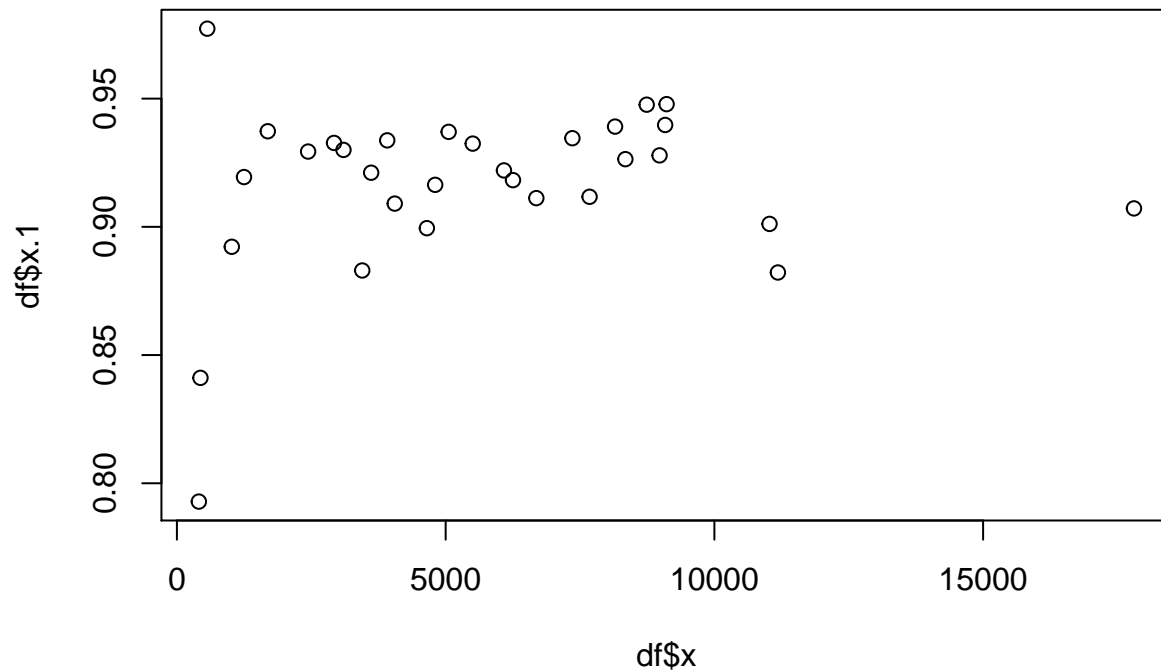
Commentaires Aurélien : A voir si la moyenne du tx de réussite c'est le bon truc, regarder et analyser les data et faire le bon choix de l'incateur.

```
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])
plot(df$x,df$x.1)
```



```
model<-lm(df$x.1~df$x, data = df)
model
```

```
##
## Call:
## lm(formula = df$x.1 ~ df$x, data = df)
##
## Coefficients:
## (Intercept)      df$x
##   9.072e-01    1.588e-06
```

Conclusion Aurélien: Coefficients: (Intercept) df\$x
9.072e-01 1.588e-06

Tend vers 0, donc la localisation n'explique pas le tx de réussite !!! Et ouais tonton !

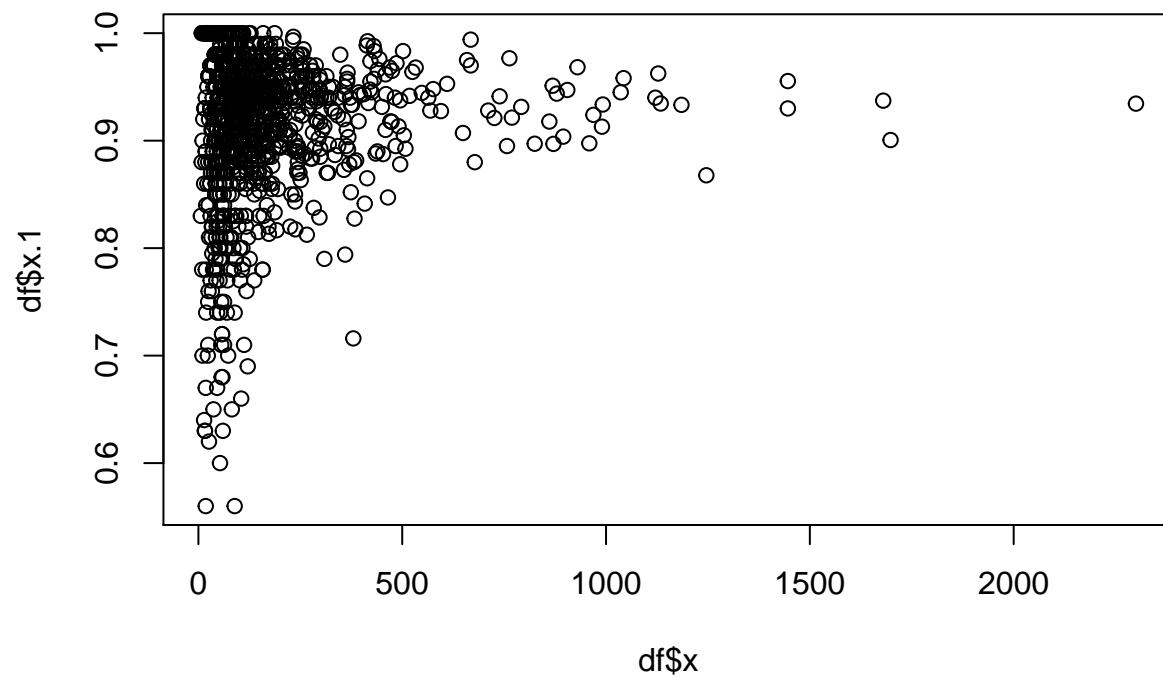
Est ce que le taux de réussite des élèves en terminale S s'explique par la localisation des communes ?

```
bddReg = bdd[, c('Ville', 'Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S')]
bddReg[is.na(bddReg)] <- 0
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])
plot(df$x, df$x.1)
```



```
model<-lm(df$x.1~df$x, data = df)
model
```

```
##
## Call:
## lm(formula = df$x.1 ~ df$x, data = df)
##
## Coefficients:
## (Intercept)      df$x
##  9.125e-01    2.741e-05
```

Tend vers zero donc non !