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

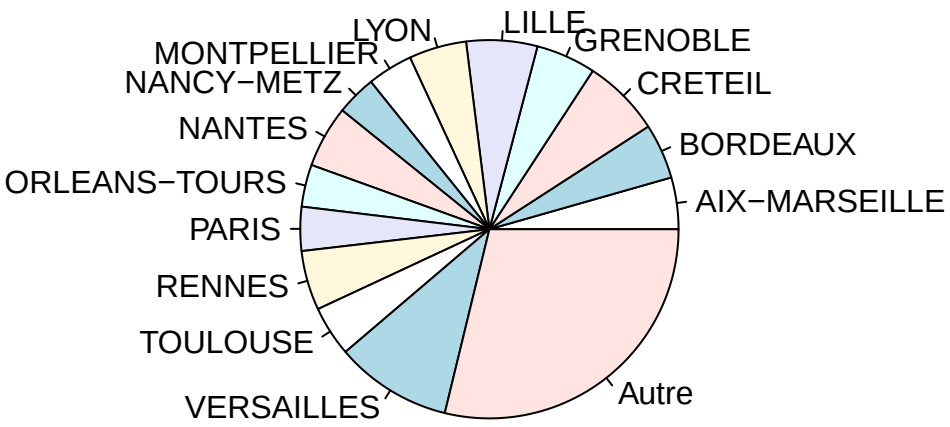
Notre de base de données à été trouvée sur le site Data.gouv. Nous avons croisé deux jeux de données distincts. Le premier concerne des données sur la performance des lycées en France (taux de réussite etc). Le second présente des données géographiques pour l'ensemble des établissement scolaires francais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

La base de données comporte de nombreuses données qui sont réparties sur toutes la France. Nous avons des répartitions de données par établissements, villes, départements et académies. Par ailleurs l'ensemble des bac généraux et technologique ainsi que certains bac professionnels sont représentés.

```
print(pie(slices, labels = lbls, main="Répartition des effectifs par académie"))
```




```
## NULL
```

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 = "", encoding = "UTF-8")->bdd
#exists('bdd')
```

Khi-Deux

```
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"]) && bddKhiDeu
    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.s
  }
}

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 dispose ici de la table de départ pour calculer le KhiDeux. Cette table nous donne l'agrégation des effectifs par Filiaire et par Académie.

```
# Calcul de la table des Abstrait
abstraitKhiDeux <- tableKhiDeux

for(i in 1:nrow(abstraitKhiDeux)){
  abstraitKhiDeux$ES[i] = sum(tableKhiDeux$ES)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeu
```

```
  abstraitKhiDeux$S[i] = sum(tableKhiDeux$S)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeux[
  abstraitKhiDeux$L[i] = sum(tableKhiDeux$L)*sum(tableKhiDeux[i,'ES'],tableKhiDeux[i,'S'],tableKhiDeux[
}

#Calcul de la table des ecarts
ecartsKhiDeux <- tableKhiDeux

for(i in 1:nrow(abstraitKhiDeux)){
  ecartsKhiDeux$ES[i] = tableKhiDeux$ES[i] - abstraitKhiDeux$ES[i]
  ecartsKhiDeux$S[i] = tableKhiDeux$S[i] - abstraitKhiDeux$S[i]
  ecartsKhiDeux$L[i] = tableKhiDeux$L[i] - abstraitKhiDeux$L[i]
}

#Caclul de la table des contributions
contribKhiDeux <- tableKhiDeux
i=1
for(i in 1:nrow(abstraitKhiDeux)){
  contribKhiDeux$ES[i] = ecartsKhiDeux$ES[i]*ecartsKhiDeux$ES[i] / abstraitKhiDeux$ES[i]
  contribKhiDeux$S[i] = ecartsKhiDeux$S[i]*ecartsKhiDeux$S[i] / abstraitKhiDeux$S[i]
  contribKhiDeux$L[i] = ecartsKhiDeux$L[i]*ecartsKhiDeux$L[i] / abstraitKhiDeux$L[i]
}

print(contribKhiDeux)
```

```
##            Académie           ES            S           L
## 1     AIX-MARSEILLE  2.258210485 7.056569e-01   0.3631176
## 2            AMIENS  3.150555771 1.986713e-01  11.1330769
## 3          BESANCON  1.760279322 9.677770e+00  14.3858369
## 4          BORDEAUX  1.230219800 1.685375e-01   5.4101212
## 5              CAEN  0.377787557 1.367897e+01  34.5472304
## 6  CLERMONT-FERRAND  1.923209177 6.332123e+00  43.0948651
## 7             CORSE  2.404496402 9.361501e+00  60.6730302
## 8           CRETEIL  7.535248899 8.096613e-01   5.1044715
## 9             DIJON  7.853642509 2.247062e+00   1.5552036
## 10         GRENOBLE 17.573030493 3.834758e-01  23.3115699
## 11       GUADELOUPE  4.901726708 7.653678e+00  67.1922477
## 12           GUYANE  0.023175128 2.119144e+01  74.2972403
## 13       LA REUNION 27.440309990 5.814052e-01  78.1061081
## 14            LILLE  5.999835840 3.431562e+01  52.0066428
## 15          LIMOGES 28.290086932 1.230328e+00  30.6420672
## 16             LYON  2.989952960 1.209383e+01  77.5452107
## 17       MARTINIQUE  3.224003272 2.392065e+00  28.8958467
## 18          MAYOTTE 43.449851924 1.378283e+02 145.4695221
## 19      MONTPELLIER 34.300492976 1.071730e+00  41.4092052
## 20       NANCY-METZ  6.712826255 1.650792e+01  13.9422166
## 21           NANTES  4.311357601 7.437503e+00   4.1081607
## 22             NICE  0.026830616 1.291225e-01   0.7899129
## 23    ORLEANS-TOURS  0.285534967 2.802474e-03   0.4394440
## 24            PARIS  0.338804687 6.446688e+00  29.8354127
## 25         POITIERS  0.826191992 2.037636e+01  48.2748095
## 26            REIMS  4.834099205 5.678761e+00   1.5001534
## 27           RENNES 21.575404587 1.368208e+00  19.9556804
## 28            ROUEN  0.005711176 3.389475e+00  10.5838682
```

```
## 29       STRASBOURG  7.610441089 7.200998e+00  77.7877942
## 30         TOULOUSE 37.007187202 2.740241e+01   0.8256151
## 31       VERSAILLES 62.920279491 1.369277e+00  83.5206781
```

Ici nous avons appliqué les étapes successives permettant de calculer le KhiDeux. Soit la corrélation entre les deux variables qualitatives : Académies et Filliaires.

Sur la table des contributions(ci-dessus) on peut observer que certaines régions et séries ont une contribution fortes a rendre dépendante ces deux variables.

On peut constater que certaines académies correspondantes a des zones géographiques périphériques ont une influance forte sur le khiDeux. Mayotte, Corse et Guadeloupe notamment. Dans ces régions la répartition entre les filières est modifiées et on trouve notamment une plus grande proportions de personnes en filiére Litéraire.

Certaines académies de métropoles ont également des comportement particulier, par exemple l'académie de Versailles a une proportion particulièrement forte de ES et faible de L. Les académies de Limoges, montpellier et Strasbourg ont également des comportement qui s'écartent des standards.

On constate finallement que la proportion de filliaire L a une forte tendance a varier alors que les filliaires ES et S ont souvent une proportions stable l'une par rapport à l'autre (environs un peux moins de deux fois plus de S que de ES). Ainsi de nombreuses académies ont une proportion de L élevée (Domtom etc..) ou faible (Lyon, Lille, Strasbourg etc..)

Finallement on calcule le score global de khideux

```
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
```

Cet indicateur nous permet de dire que la situation géograpique est certainement fortement dépendante de la répartition entre les filliaires. En effet la probabilité que la situation géographique soit indépendante de la répartition dans les différentes filiéres est inférieure à 2.2e-16.

Régréssion

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

Problématique

Une interrogation récurrente vis à vis de la réussite scolaire est de se demander si la situation géographique d'un étudiant tend à lui offrir des chances supplémentaire d'obtenir son baccalauréat.

Pour essayer de déterminer si l'acadamie a un rôle prédominant dans la réussite de l'élève nous allons chercher à connaitre l'impact de l'académie sur le taux de réussite au bac S, mais nous nous interrogerons aussi sur l'influence émise par les taux de réussite au baccalauréat L. Ainsi, un environnement, crée par la jointure entre une situation géographique donnée et un taux de réussite dans une autre fillière donné, a-t-il un fort impact sur la réussite d'un éléve passant son baccalauréat scientifique ?

Ainsi, nous allons crée une matrice comportant l'académie, l'effectif présent en série scientifique, le taux brut de réussite dans cette même série et le taux dans la série L.

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

Nous procédons ensuite au nettoyage de notre matrice en retirant les valeurs nulles et en transformant les taux à une forme 0<x<1.De plus nous retirons les valeurs aberrantes, soit celles où il n'y a pas d'éléve inscrit dans les fillières étudiées.

```
#valeur non définies mise à 0
bddReg[is.na(bddReg)] <- 0
#transformation des taux
bddReg[3] <- bddReg[3]/100
bddReg[4] <- bddReg[4]/100

df=data.frame(bddReg[1],bddReg[2],bddReg[3], bddReg[4])
#suppression des données aberrantes
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0 & df$Taux.Brut.de.réussite.s
```

Pour mener une étude par académie nous devons agréger l'ensemble des établissements scolaire appartenant à la même académie.

```
#regroupement des effectifs par académie
regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Académie), FUN=sum)
#moyenne de l'ensemble des taux de réussite des lycées par académie
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Académie), FUN=mean)[2])
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.L, by=list(df$Académie), FUN=mean)[2])
```

Nous créeons maitenant notre model linéaire et nous allons procéder à la regression.

```
#création du modèle
df = data.frame(regData[1], regData[2], regData[3], regData[4])
col_headings <- c('Académie','Effectif', 'TxRéussiteS', 'TxRéussiteL')
names(df) <- col_headings
model<-lm(df$TxRéussiteS~df$Effectif+df$TxRéussiteL, data = df)
#affichage des résultats
summary(model)
```

```
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$Effectif + df$TxRéussiteL,
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.066971 -0.009052  0.002738  0.011954  0.032609
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.434e-01  5.949e-02   7.453 4.06e-08 ***
## df$Effectif    -5.343e-08  1.049e-06  -0.051     0.96
## df$TxRéussiteL  5.170e-01  6.588e-02   7.847 1.51e-08 ***
```

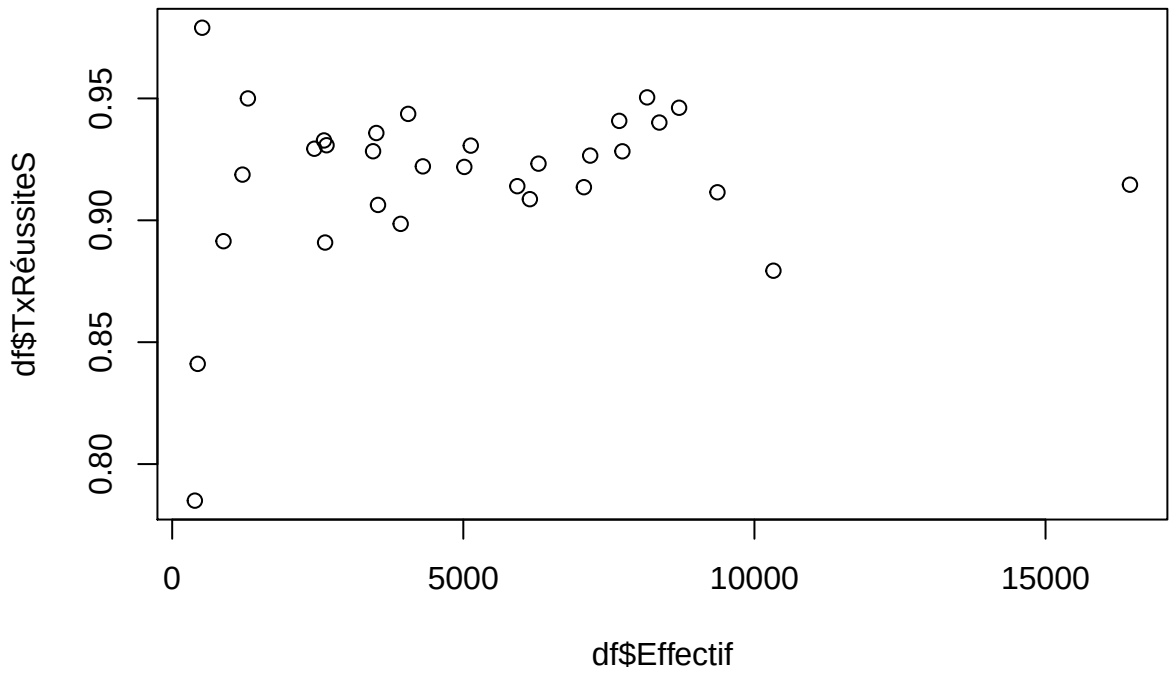
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01989 on 28 degrees of freedom
## Multiple R-squared:  0.699,  Adjusted R-squared:  0.6775
## F-statistic: 32.51 on 2 and 28 DF,  p-value: 5.02e-08
```

Analyse

On constate que la valeur de R carré est élevé impliquant que le modéle a une importance sur le taux de réussite, selon une p-value extrémement faible (5.02e-08) soit une précision de 1 sur 1 milliard. Cependant nous pouvons aussi observer que les deux variables utilisées n'ont pas le même impact sur notre résultat. En effet, l'effectif semble avoir un faible impact (-5.343e-08), tandis que le Taux de réussite en série L a un impact fort (5.170e-01).

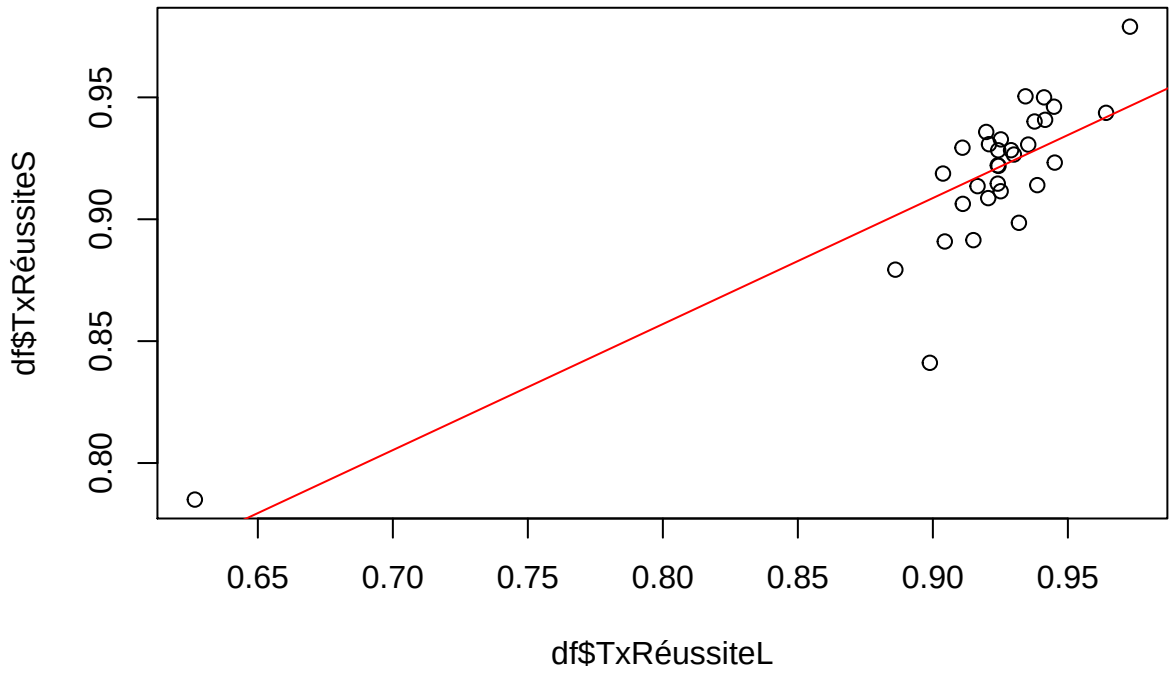
Nous pouvons représenter l'impact de l'effectif par le nuage de point suivant, et nous constatons que la droite de la fonction de regression

```
plot(df$Effectif,df$TxRéussiteS)
x <- seq(0,18000)
lines(x,x*-5.343e-08+4.434e-01,col="red")
```



Nous pouvons représenter l'impact du taux de réussite de la série L par le nuage de points suivant, et nous constatons que la droite de la fonction de regression montre une augmentation forte.

```
plot(df$TxRéussiteL,df$TxRéussiteS)
x <- seq(0,18000)
lines(x,x*5.170e-01+4.434e-01,col="red")
```



Conclusion

Ainsi, selon notre étude de donnée nous pouvons affirmer que l'environnement a un impact sur la réussite d'un élève, mais ce n'est pas la localisation qui crée cette empreinte mais la réussite des paires dans une série différente est-elle un facteur déterminant. On peut ainsi estimer qu'un environnement où une fillière a un taux de réussite élevé impactera bénéfiquement les chances de réussite d'un élève.

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

Nous cherchons à savoir ici su le fait qu'un étudiant inscrit au baccalauréat d'une commune a plus de chance de réussir que dans une autre commune. Nous allons regarder spécialement la série scientifique.

La première étape de ce cas de test réunie dans une nouvelle matrice, les colonnes "Ville", "Effectif.Présents.série.S" et "Taux.Brut.de.réussite.série.S". Nous allons tenter d'expliquer par la suite le taux de réussite de chaque ville par les effectifs inscrit dans ces mêmes localisations.

```
bddReg = bdd[, c('Ville','Effectif.Présents.série.S', 'Taux.Brut.de.réussite.série.S', 'Taux.Brut.de.ré
```

Maintenant que nous disposons des données propre à l'étude de ce cas, nous avons besoin de nettoyer les données. Il faut notemment mettre des valeurs nulles dans les champs non remplis et ramèner le taux à des valeurs comprises entre 0 et 1.

```
# Permet de mettre 0 dans les cases non remplies
bddReg[is.na(bddReg)] <- 0
# Ramène le pourcentage du taux de réussite à une valeur entre 0 et 1
bddReg[3] <- bddReg[3]/100
bddReg[4] <- bddReg[4]/100
```

Pour palier à des villes où aucun candidat ne serait inscrit dans la série S, nous supprimons volontairement ces enregistrements qui sont considérés comme des individus abérants pour notre études. C'est ce que fait la portion de code suivante.

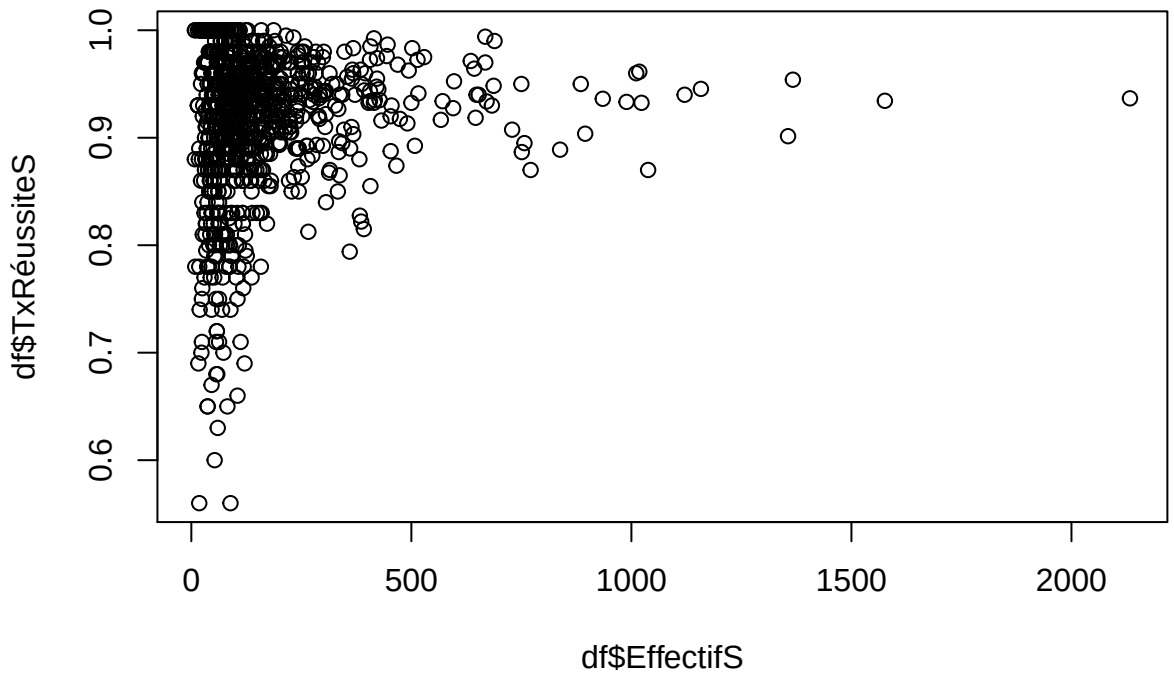
```
df=data.frame(bddReg[1],bddReg[2],bddReg[3], bddReg[4])
df<-df[(df$Effectif.Présents.série.S>0 & df$Taux.Brut.de.réussite.série.S>0 & df$Taux.Brut.de.réussite.
```

Ce que nous allons faire maintenant, c'est faire un groupement par ville en faisant la somme des effectifs et la moyenne des taux de réussite de chaque établissement pour avoir un seul enregistrement par ville.

```
# Addition des efefctifs groupé par Ville
regData = aggregate(df$Effectif.Présents.série.S, by=list(df$Ville), FUN=sum)
# Moyenne des taux de réussite des séries S assimilée
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.S, by=list(df$Ville), FUN=mean)[2])
# Moyenne des taux de réussite des séries L assimilée
regData = c(regData,aggregate(df$Taux.Brut.de.réussite.série.L, by=list(df$Ville), FUN=mean)[2])
```

On peut donc maintenant s'apercevoir de la repartition du taux de réussite S en fonction des effectifs par ville.

```
df = data.frame(regData[1], regData[2], regData[3], regData[4])
col_headings <- c('Ville','EffectifS', 'TxRéussiteS', 'TxRéussiteL')
names(df) <- col_headings
plot(df$EffectifS,df$TxRéussiteS)
```



Nous allons donc effectuer une régréssion linéaire sur ces données pour tenter d'expliquer le taux de réussites par le lieu d'inscription du candidat au baccalauréat.

```
model<-lm(df$TxRéussiteS~df$EffectifS + df$TxRéussiteL, data = df)
summary(model)
```

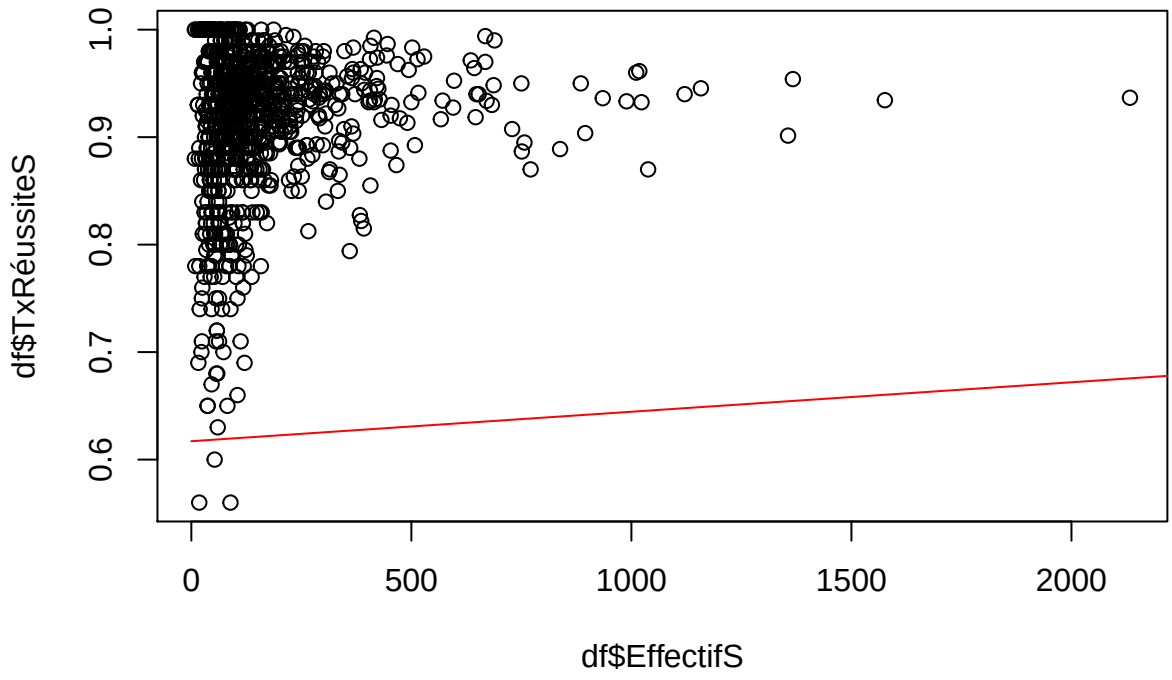
```
##
## Call:
## lm(formula = df$TxRéussiteS ~ df$EffectifS + df$TxRéussiteL,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33674 -0.02687  0.00833  0.04207  0.14147
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.171e-01  2.256e-02  27.356   <2e-16 ***
## df$EffectifS   2.369e-05  1.069e-05   2.216   0.0269 *
## df$TxRéussiteL 3.228e-01  2.460e-02  13.118   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06015 on 1014 degrees of freedom
## Multiple R-squared:  0.155,  Adjusted R-squared:  0.1533
## F-statistic: 92.98 on 2 and 1014 DF,  p-value: < 2.2e-16
```

On s'aperçoit que l'effectif explique peu le taux de réussite. En effet, pour une unité du taux de réussite, l'effectif change de 2.741e-05 ce qui est très petit.

Le R carré ajusté en tendant vers 0 (adjusted R sqare = 0.005677) nous indique aussi que l'effectif explique faiblement le taux de réussite avec environ 7 chance sur 1000 de se tromper donc cette prédiction est plutot forte (p-value = 0.007613).

En traçant la droite ax + b correspondant au modèle (2.741e-05*x + 6.171e-01), on remarque sa faible pente et sa répresentation plutot horizontale ce qui indique aussi par le visuel un faible lien.

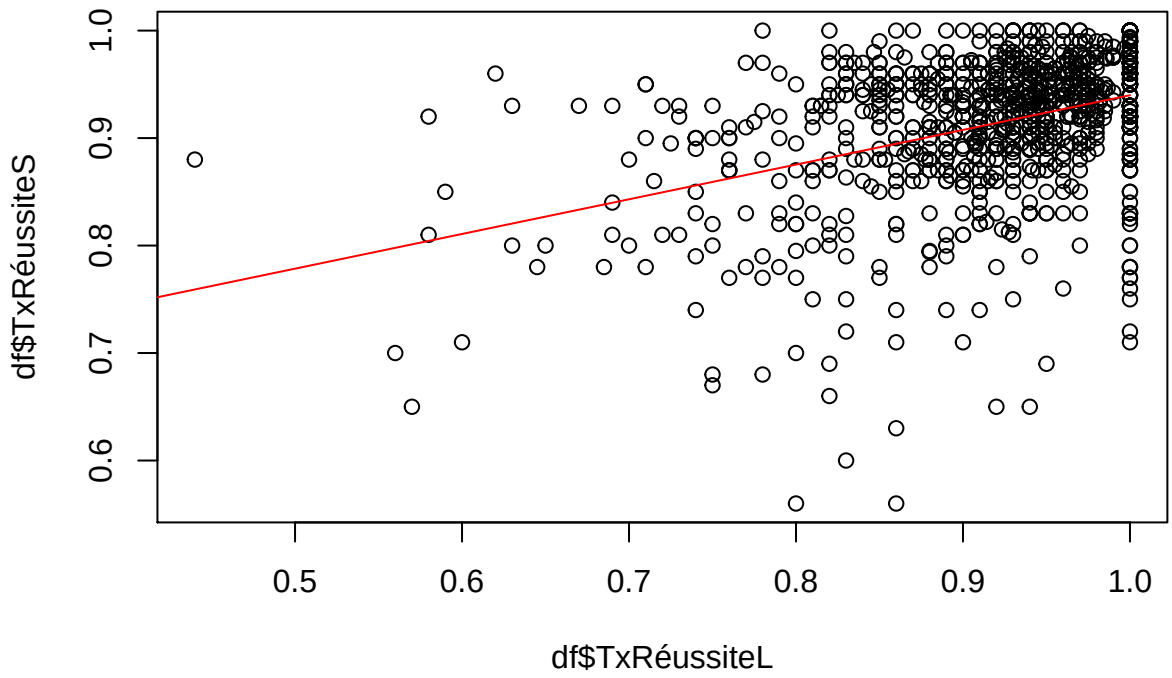
```
plot(df$EffectifS,df$TxRéussiteS)
x <- seq(0,2300)
lines(x,x*2.741e-05+6.171e-01,col="red")
```



En revanche pour le taux de réussite des séries L explique encore une fois le taux de réussite des séries S. On peut imaginer que le niveau général d'une ville irradie de L en S. Ce lien exuste avec un R carré adjusté de 0.1533 avec une p-value infinitesimalement petite. Il y a donc quasiement aucune chance de se tromper sur ce point.

A titre informatif voici la droite du modèle qui représente ce lien :

```
plot(df$TxRéussiteL,df$TxRéussiteS)
x <- seq(0,1)
lines(x,x*3.228e-01+6.171e-01,col="red")
```



Conclusion Générale