# Problématique

On essaie d’expliquer les notes à l’écrit en fonction du genre, de la déficience mentale, et de l’emplacement des établissements scolaires

# Présentation des données

* La population est l’ensemble des collèges
* La variable endogène est la moyenne des notes à l’écrit (note\_a\_l\_ecrit\_g).
* La première variable explicative est le nombre d’élèves de Segpa (nbre\_eleves\_segpa).
* La deuxième variable explicative est le nombre d’élèves d'Ulis (nbre\_eleves\_ulis).
* La troisième variable explicative est la longitude de la position des établissements (longitude).
* La quatrième variable explicative est l’effectif de filles dans les établissements (effectifs\_filles).
* La cinquième variable explicative est l’effectif de garçons dans les établissements (effectifs\_garçons).

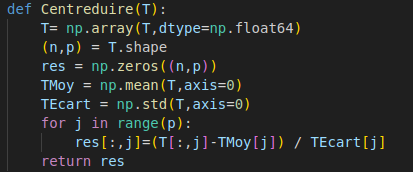
# Import des données et mise en forme

## On importe notre vue sous forme de DataFrame avec la commande suivante :



## On a besoin de supprimer les cases vides (qui contiennent nan en Pythons), puis on transforme notre DataFrame en Array :

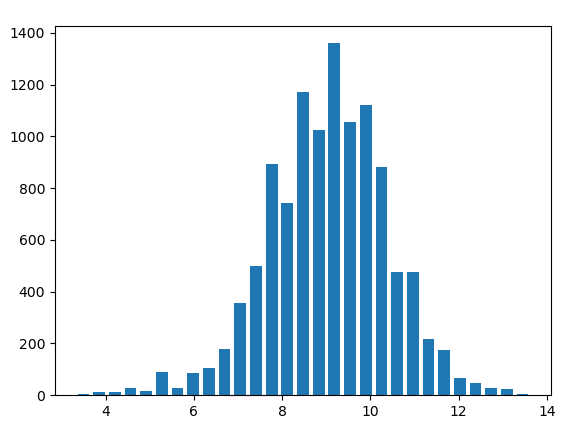
## Centrer-réduire, on ne garde que les colonnes de notre tableau qui contiennent des données numériques, on peut alors centrer-réduire ces données :

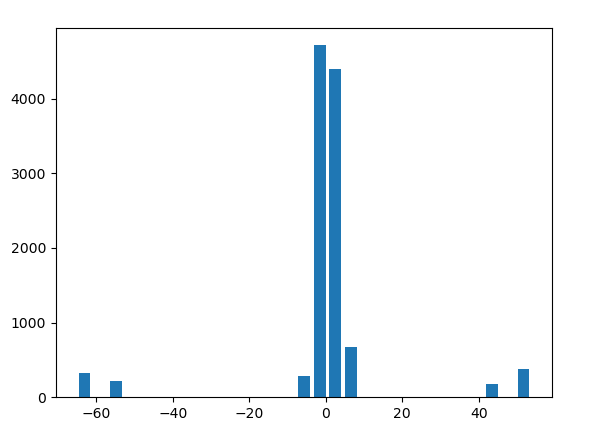


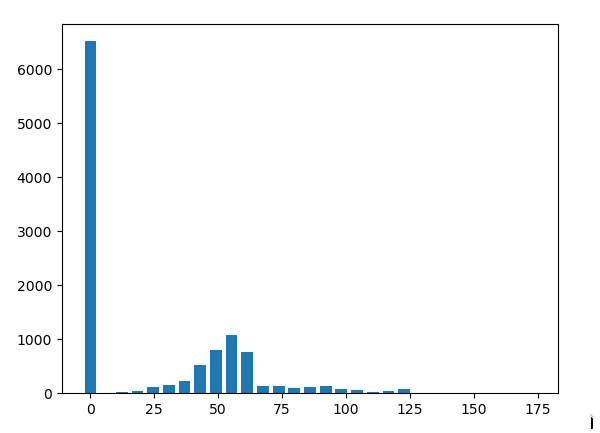
# Exploitation des données

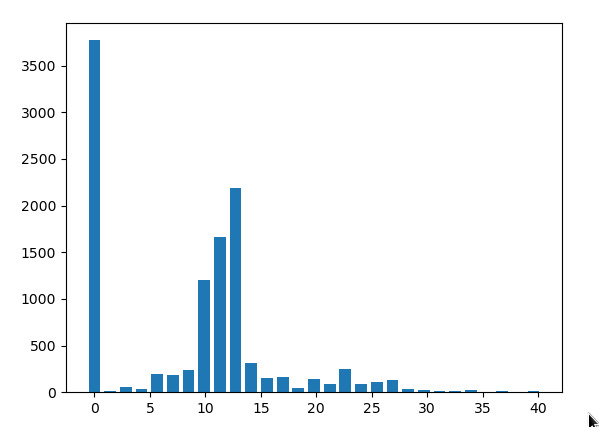
## Représentation graphique

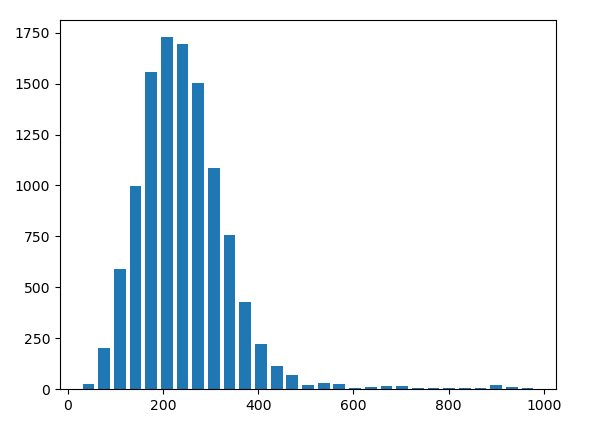
On choisit d’étudier les diagrammes en bâtons de nos variables statistiques :

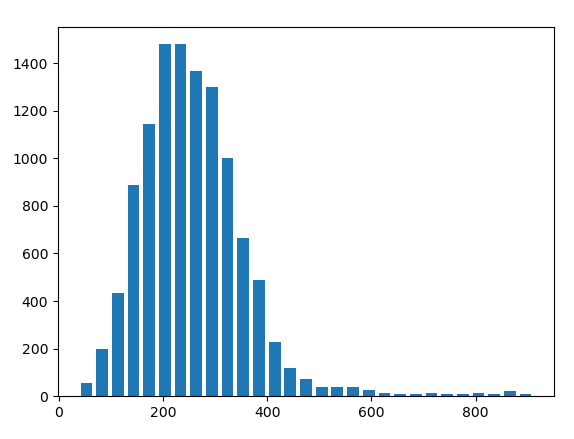
Note à l’écrit

Longitude

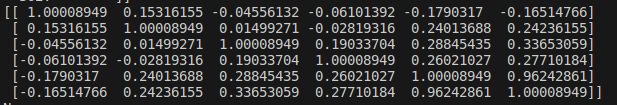
Nombre eleves segpa

Nombre eleves ulis

Effectifs filles

Effectifs garcons

## Matrice de covariance



# Régression linéaire multiple

## Utilisation de la Régression linéaire multiple : comment ?

En choisissant la 1e variable statistique comme variable endogène et certaines des autres variables comme variables explicatives, la régression linéaire multiple nous permettrait d'obtenir une estimation de la moyenne au brevet dans les collèges en fonction d’autres informations sur ces collèges.

## Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible la note moyenne au brevet des collèges, qui se trouve dans la colonne 0 de CollegesAr. La colonne 0 de MatriceCov donne les coefficients de corrélation de la note au brevet avec chacune des autres variables/colonnes de CollegesAr. On va choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec la note au brevet.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de MatriceCov sont : 0.153, -0.179, -0.165 et –0.061. Ils correspondent aux variables numéro 1, 4, 5 et 3. Les colonnes 1, 4, 5 et 3 de CollegesAr correspondent aux :

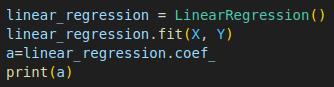
* Longitude de chaque collège
* Effectif filles de chaque collège
* Effectif garçons de chaque collège
* Nombre d’élève Ulis de chaque collège

## Lien avec la problématique

Les paramètres de la régression linéaire multiple nous informeront des variables explicatives qui influencent le plus la note au brevet. En calculant le coefficient de corrélation multiple, on saura de plus si cette influence permet de prédire la réalité, on saura ainsi ce qui influence réellement la note moyenne au brevet.

## Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :



## Paramètres, interprétation

On obtient les paramètres a0 = 0.016, a1 = 0.000, a2 = 0.000, a3 = -0.003, a4 = 0.000

Le signe des paramètres a0, a1, a2, a4 nous permet de voir qu’ils influencent positivement les note de l’écrit

Alors que le signe du paramètre a3 nous permet de voir qu’il influence négativement les note de l’écrit

Comme les variables endogène et explicatives sont centrées-réduite, on peut de plus voir que a0 a une plus grande influence parce qu’il est le plus éloigner de 0. Donc les variables qui ont le plus d’influence dans l’ordre décroissant :

a0, a3, a4, a1, a2

## Coefficient de corrélation multiple, interprétation





Le coefficient de corrélation multiple < 0.866, il est donc faiblement corrélé.

# Conclusion

## Réponse à la problématique

Les variables explicatives choisies exercent une influence limitée sur la note moyenne à l'écrit. Bien que certaines variables montrent des signes d'influence (comme les effectifs de garçons exerçant une influence négative et la longitude une influence positive), la corrélation globale reste faible, avec un coefficient de corrélation multiple inférieur à 0.866. On peut en conclure que le genre, la localisation de l’établissement ou la déficience mentale n’exercent pas une grande influence sur les notes à l’écrit du brevet des collèges.

## Argumentation à partir des résultats de la régression linéaire

Les coefficients nous permettent de conclure que :

* La longitude et les effectifs de filles exercent une influence positive sur la note à l'écrit.
* Le nombre d'élèves Ulis exerce une influence négative.
* Les effectifs de garçons semblent exercer une influence relativement faible mais positive quand même.

## Interprétations personnelles

Ces analyses montrent que les performances scolaires dépendent de nombreux autres facteurs non inclus dans notre sujet comme la qualité de l'enseignement, le soutien familial...Les données démographiques et les effectifs ne suffisent pas à démontrer l’influence sur les notes à l’écrit.