# **BUT Informatique 2023-2024**

SAé 2.04 - Modèle Livrable 3



# **1 Les données dataColleges.csv - Problématique**

## **(a) Présentation des données**

Le fichier dataColleges.csv contient plusieurs séries statistiques sur l’ensemble de toutes les collèges répertoriés dans notre base de données :

La population est l’ensemble des collèges, représentés de manière unique par leur code, et avec l’indication du nom du collège.

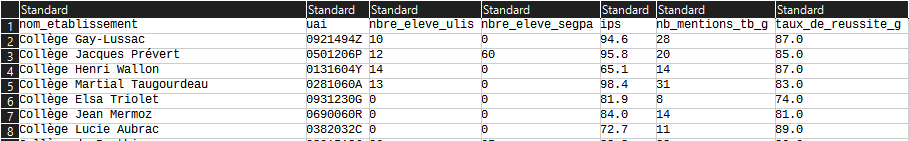
La 1er variable statistique sur cette population est le taux de réussite en filière général à l’écrit du brevet en 2020 pour chaque collège.

La 2e est le nombre d'élèves ulis, de chaque collège.

La 3e est le nombre d'élèves segpa, de chaque collège.

La 4e est l’indice de position sociale (IPS) moyen de chaque collège.

La 5e est le nombre de mention “très bien” obtenue lors de l’écrit du brevet en filière général et en 2020 pour chaque collège.



## **(b) Problématique**

En utilisant ces données, on va essayer de répondre à la problématique suivante :

*Parmi les données de notre fichier, certaines peuvent-elles permettre d’expliquer ce qui favorise le taux de réussite en filière général à l'écrit du brevet dans les différents collèges ?*

# **2 Import des données, mise en forme**

## **(a) Importer les données en Python**

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



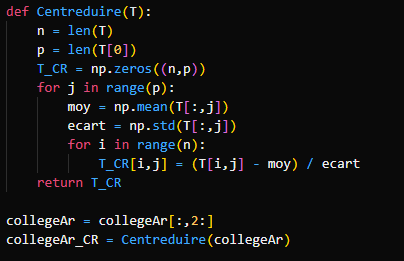
## **(b) Mise en forme**

On transforme notre DataFrame en Array :



## **(c) Centrer-réduire**

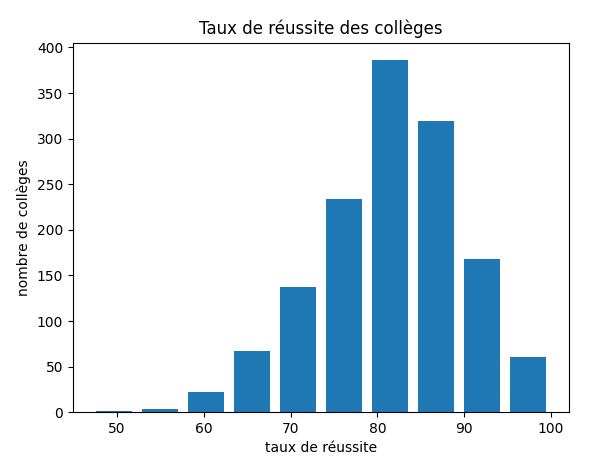
On ne garde que les colonnes de notre tableau qui contiennent des données numériques, on peut alors centrer-réduire ces données :



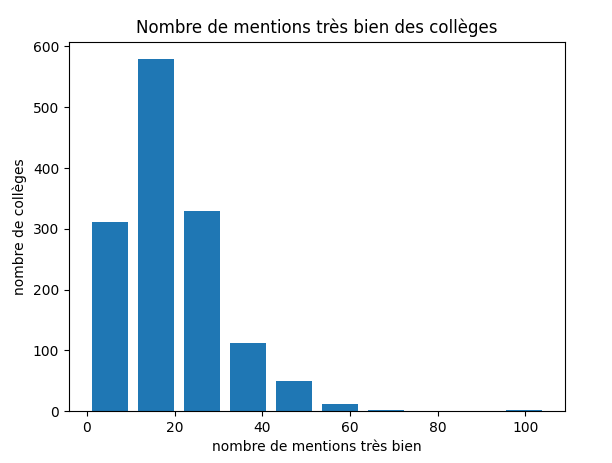
# **3 a. Exploration des données : représentations graphiques**

On choisit d’étudier les diagrammes en batons des nos variables statistiques :

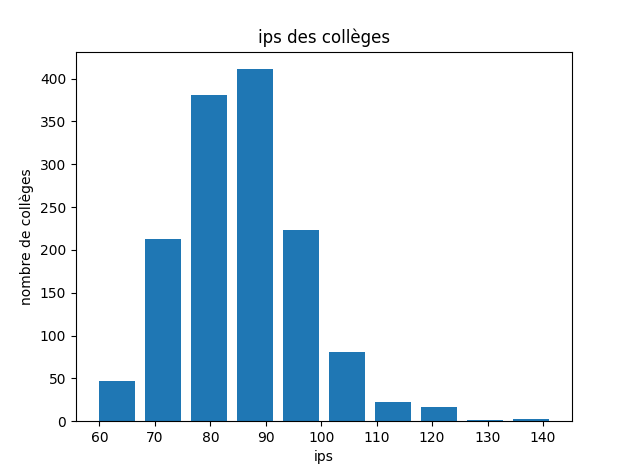
Diagramme en batons des notes au Brevet



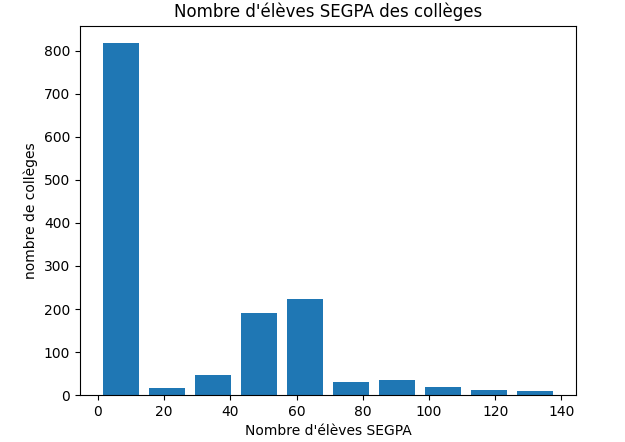
On remarque que la majorité des collège on un taux de réussite entre 70 et 90% de réussite mais qu’il descend jusqu'à 50%.



On remarque le la plupart du nombre de mention très bien sont en dessous de 40 par collèges même si quelques rares collèges monte à plus.

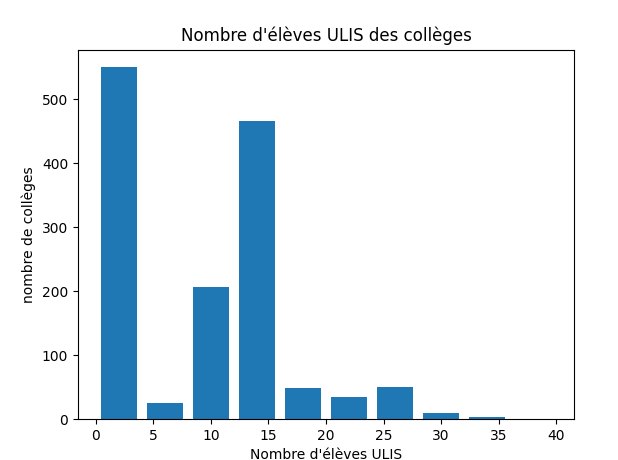


L’ips de la majorité des collège est entre 70 et 100 mais certains collèges monte jusqu’à 140.



On note que la plupart des collèges de notre base de données n'accueille pas d’élèves segpa.

Les collège en accueillant se situe majoriterement autour de 50 / 60 élèves segpa par collège même si une grande diversité se voit.



Quand au élève ulis, environ la moitié des collège de la base de donné n’en accueille pas mais ce qui en accueille en possède un nombre plus réduit entre 10 et 15 pour la majorité même si ici aussi une partie des collège en accueille jusqu'à 30.

# **3 b. Exploration des données : matrice de covariance**

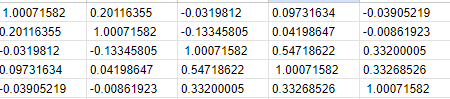
## **(a) Démarche**

Dans cette partie, on calcule la matrice de covariance afin de *à*



## **(b) Matrice de covariance**

On obtient la matrice suivante :



# **4 Régression linéaire multiple**

**(a) Utilisation de la Régression linéaire multiple : comment?**

En choisissant la 1ère variable statistique (taux de réussite en filière générale à l'écrit du brevet en 2020) 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.

## **(b) Variables explicatives les plus pertinentes**

Notre objectif est de trouver des variables qui expliquent le mieux possible le taux de réussite en filière générale à l'écrit au brevet des collèges, qui se trouve dans la colonne 4 de CollegesAr. La colonne 4 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 les 4 variables restantes .

Les coefficients de corrélation dans la colonne 4 de MatriceCov sont de : 0.039 ; 0.008 ; 0.332 ; 0.332. Ils correspondent aux variables numéro 2,3,4,5. Les colonnes 0,1,2,3 de CollegesAr correspondent aux :

l’ips de chaque collège,

le nombre de mention très bien de chaque collège,

le nombre d’élèves ulis de chaque collège,

et le nombre d’élèves segpa de chaque collège.

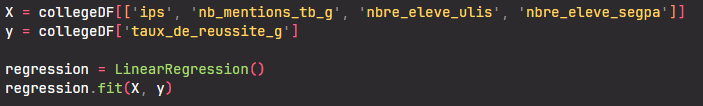
On choisit donc ces 4 variables comme variables explicatives.

## **(c) Lien avec la problématique**

Les **paramètres** de la régression linéaire multiple nous informeront des variables explicatives qui influencent le taux de réussite. 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 le taux de réussite.

## **(d) Régression Linéaire Multiple en Python**

On fait maintenant la régression linéaire multiple avec Python :

**

## **(e) Paramètres, interprétation**

Les résultats obtenus sont les suivants :

* **Coefficients** : [ 0.15523383, 0.15912003, -0.06254447, 0.00558167]

Interprétation des coefficients :

* **Coefficient pour l'IPS** : 0.15523382569393485
* **Coefficient pour le nombre de mentions TB** : 0.15912003316674028
* **Coefficient pour le nombre d'élèves ULIS** : -0.06254447129471503
* **Coefficient pour le nombre d'élèves SEGPA** : 0.005581668771783371

**(f) Coefficient de corrélation multiple, interprétation**

Le coefficient de corrélation multiple est de 0.14594325458493862, ce qui signifie qu'environ 14.59% de la variance du taux de réussite est expliquée par les variables explicatives.

# **5 Conclusions**

1. **Réponse à la problématique**

En utilisant la régression linéaire multiple, nous avons identifié les variables qui influencent le plus le taux de réussite au brevet dans les collèges. Les résultats montrent que l'IPS moyen, le nombre de mentions "très bien", le nombre d'élèves ULIS et le nombre d'élèves SEGPA sont des facteurs importants.

1. **Argumentation à partir des résultats de la régression linéaire**

Les coefficients de la régression linéaire multiple indiquent que :

* L'IPS moyen et le nombre de mentions "très bien" ont une influence positive significative sur le taux de réussite.
* Le nombre d'élèves ULIS a une influence négative modérée sur le taux de réussite.
* Le nombre d'élèves SEGPA a une influence très faible mais positive.

## **(c) Interprétations personnelles**

* 1. Ces résultats montrent que les collèges avec un IPS élevé et de bons résultats (mentions "très bien") tendent à avoir de meilleurs taux de réussite. Les politiques éducatives pourraient donc se concentrer sur le soutien aux élèves pour maximiser leur potentiel académique.