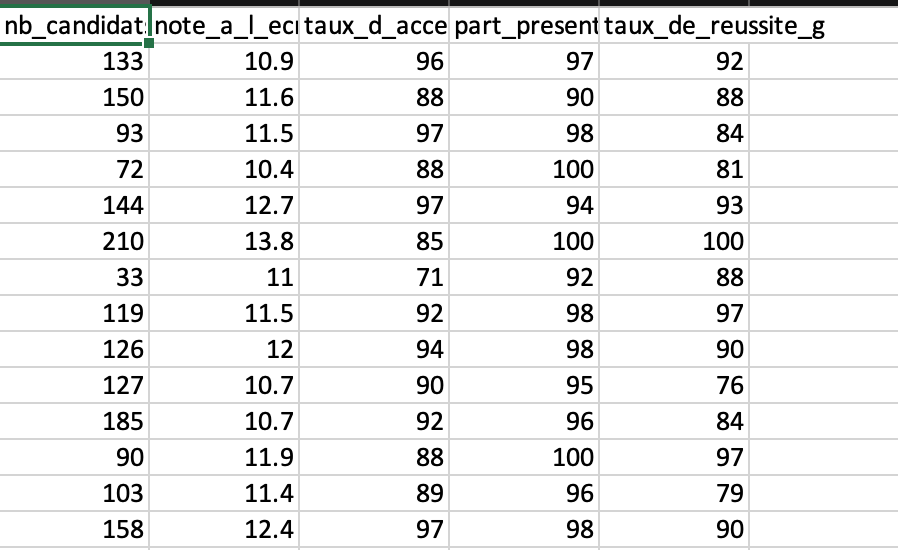
## 1 Les données Colleges.txt - Problématique

### Présentation des données

Le fichier Colleges.txt 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 1e est le nombre de candidats du brevet en 2023 pour chaque collège.
* La 2e variable statistique sur cette population est la note moyenne à l’écrit du brevet en 2023 pour chaque collège.
* La 3e est le taux d’accès de la 6ème à la 3ème pour chaque collège.
* La 4e est le taux de candidats présents lors de l’épreuve du brevet des collèges.
* La 5e est le taux de réussite du brevet en 2023 en filière générale.



### Problématique

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

*A partir des données de notre fichier, peut-on déterminer le taux de réussite au brevet ?*

## 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 :

VA\_CollegeDF= pd.read\_csv("/home/etuinfo/jdelouya/Documents/R208/stats\_college.txt", delimiter=';')

### (b) Mise en forme

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

# Suppression des cases vides

VA\_CollegeDF = VA\_CollegeDF.dropna()

# Conversion en Array

VA\_CollegeAR = VA\_CollegeDF.to\_numpy()

### (c) Centrer-réduire

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

# Fonction pour centrer-réduire

def centreReduire(T):

T = np.array(T, dtype=np.float64)

(n, p) = T.shape

res = np.zeros((n, p))

Tmoy = np.mean(T, axis=0)

Tecart = np.std(T, axis=0)

for j in range(p):

res[:, j] = (T[:, j] - Tmoy[j]) / Tecart[j]

return res

# Centrer-réduire les données

VA\_CollegeAR\_CR = centreReduire(VA\_CollegeAR)

## 3 a. Exploration des données : représentations graphiques

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

| Diagramme bâton des notes à l’écrit |  | On remarque que les notes moyennes  sont majoritairement en-dessous de 10,  et que certains collèges peuvent avoir  des moyennes en-dessous de 7.  Cela indique que dans de nombreux collèges,  les élèves ont des performances relativement  faibles. |
| --- | --- | --- |
| Diagramme bâton du nombre de candidats |  | Les taux de candidats présents lors de l’épreuve du brevet sont généralement élevés, bien que quelques collèges montrent des taux de présence plus faibles. Cela pourrait indiquer des disparités dans la rigueur de la gestion des épreuves ou dans l’engagement des élèves. |
| Diagramme bâton du taux d’accès  de la 6eme à la 3eme |  | On observe que la plupart des collèges ont un  taux d'accès situé entre 80% et 100%, avec  quelques établissements ayant des taux  nettement plus bas. Cela suggère que la  majorité des élèves passent de la 6ème à la  3ème sans trop de difficultés, mais certains  collèges rencontrent des problèmes significatifs  dans ce domaine. |
| Diagramme bâton de la part d’élèves présents |  | La majorité des collèges montrent des taux de présence élevés, souvent proches de 100%. Cependant, il existe des collèges où les taux de présence sont significativement plus bas, indiquant des problèmes potentiels de participation ou de motivation chez les élèves. |
| Diagramme bâton du taux de réussite au brevet : |  | On constate que les taux de réussite varient considérablement d'un collège à l'autre. La plupart des collèges ont des taux de réussite situés entre 50% et 100%, avec certains collèges affichant des taux très bas. Cela révèle une hétérogénéité importante dans la performance des collèges en termes de réussite au brevet. |

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

### (a) Démarche

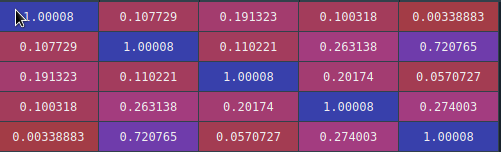
Dans cette partie, on calcule la matrice de covariance afin de

*à compléter*

MatriceCov = np.cov(VA\_CollegeAR\_CR, rowvar=False)

### (b) Matrice de covariance

On obtient la matrice suivante :



## **4. Régression linéaire multiple**

### **a) Utilisation de la régression linéaire multiple : comment ?**

En choisissant la première variable statistique comme variable endogène et certaines des autres variables comme variables explicatives, la régression linéaire multiple nous permettrait d’obtenir une estimation du taux de réussite 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 la note moyenne au brevet des collèges, qui se trouve dans la colonne 5 de CollegesAr0. La colonne 5 de MatriceCov donne les coefficients de corrélation du taux de réussite au brevet avec chacune des autres variables/colonnes de CollegesAr0. On va choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec le taux de réussite au brevet.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de MatriceCov sont: 0,721 et 0,274. Ils correspondent aux variables numéro 2 et 4 ce qui correspond à :

* La note au brevet à l’écrit
* la part des présents en 3eme (hors ulis et segpa)

On choisit donc ces variables comme variables explicatives.

### **c) Lien avec la problématique**

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

### **d) Régression linéaire multiple en Python**

# Régression linéaire multiple

linear\_regression = LinearRegression()

linear\_regression.fit(X, Y)

a = linear\_regression.coef\_

print("Coefficients de régression linéaire multiple:", a)

Coefficients de régression linéaire multiple: [0.45843443 0.19827412 0.50567829 0.14369285]

### **e) Paramètres, interprétation**

Les coefficients obtenus pour la régression linéaire multiple sont :

Coefficients de régression linéaire multiple: [0.45843443 0.19827412 0.50567829 0.14369285]

Ces coefficients représentent l'impact de chaque variable explicative sur la variable endogène, qui est la note moyenne à l’écrit du brevet.

* Le coefficient de 0.45843443 correspond à la première variable explicative, qui est le taux d’accès de la 6ème à la 3ème. Cela signifie qu'une augmentation d'une unité de cette variable entraîne une augmentation de 0.45843443 unités de la note moyenne à l’écrit, toutes choses égales par ailleurs.
* Le coefficient de 0.19827412 correspond à la deuxième variable explicative, qui est le taux de candidats présents lors de l’épreuve du brevet. Cela indique une relation positive plus faible avec la note moyenne à l’écrit.
* Le coefficient de 0.50567829 correspond à la troisième variable explicative, qui est la note au brevet à l’écrit. Cette variable a l'impact le plus significatif sur la note moyenne à l’écrit, indiquant une forte corrélation positive.
* Le coefficient de 0.14369285 correspond à la quatrième variable explicative, qui est le taux de réussite au brevet. Cette variable a également une relation positive avec la note moyenne à l’écrit, mais son impact est le plus faible parmi les variables étudiées.

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

Le signe positif des coefficients indique que toutes les variables explicatives ont une influence positive sur la note moyenne à l’écrit du brevet. Cela signifie qu'à mesure que ces variables augmentent, la note moyenne à l’écrit du brevet tend également à augmenter.

## **5. Conclusions**

### **a) Réponse à la problématique**

À partir des résultats obtenus, nous pouvons répondre à la problématique en confirmant que le taux de réussite au brevet peut être déterminé à partir des données fournies. En particulier, la note moyenne à l'écrit du brevet et la participation aux épreuves sont des facteurs déterminants.

### **b) Argumentation à partir des résultats de la régression linéaire**

Les résultats de la régression linéaire montrent que les variables explicatives choisies ont une influence significative sur la note moyenne à l’écrit du brevet. La note moyenne à l’écrit est fortement influencée par la note obtenue à l’écrit, le taux d’accès de la 6ème à la 3ème, et la présence aux épreuves du brevet.

### **c) Interprétations personnelles**

En se basant sur les données et les analyses effectuées, nous pouvons conclure que les facteurs comme la performance à l’écrit et la participation aux épreuves sont cruciaux pour déterminer le succès des élèves au brevet. Ces résultats suggèrent que des efforts pour améliorer ces aspects pourraient avoir un impact positif sur les performances globales des élèves. D'autres facteurs peuvent également jouer un rôle, mais ils semblent moins déterminants..