SAÉ 2.04 - Livrable 3

PFRANGER Mathéo - FAUCHET Malo

1 Les données Colleges.csv - Problématique

1.1 Présentation des données

Le fichier Colleges.csv contient plusieurs séries statistiques sur l'ensemble de tout les collèges répertoriés dans notre base de données :

- La population est l'ensemble des collèges.
- La 1ère variable statistique sur cette population est le nombre de candidats provenant de filière générale.
- La 2ème variable statistique est le taux de réussite au brevet.
- La 3ème variable statistique est la moyenne des notes du brevet à l'écrit.
- La 4ème variable statistique est la part des élèves de 3ème présent au brevet

	A	B	C	D	E
1	nb_mentions_tb_g	nb_candidats_g	taux_de_reussite_g	note_a_l_ecrit_g	part_presents_3eme_ordinaire_g
2	17	113	86.0	9.7	94.0
3	12	61	90.0	10.6	87.0
4	7	158	66.0	7.5	98.0
5	12	128	89.0	8.9	81.0
6	55	169	95.0	11.0	88.0
7	31	126	90.0	10.6	98.0

1.2 Problématique

En utilisant ces données, on va essayer de répondre à la problématique suivante :
Parmi les données de notre fichier, certaines influent-elles sur le nombre de mentions "Très bien" obtenu au brevet ?

2 Import des données, mise en forme

2.1 Importer les données en Python

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

```
CollegesDF = pd.read_csv("export.csv", sep=";")
```

2.2 Mise en forme

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

```
CollegesDF = CollegesDF.dropna()
CollegesAr = CollegesDF.to_numpy(dtype=np.float64)
```

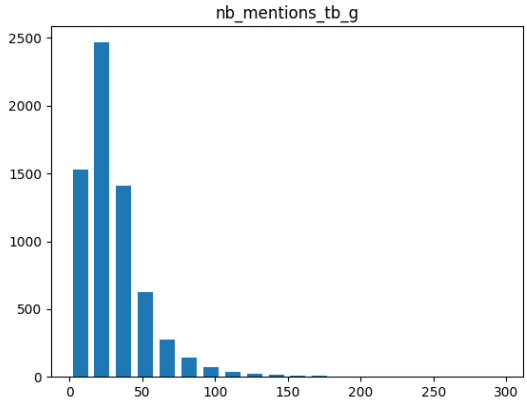
2.3 Centrer-réduire

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

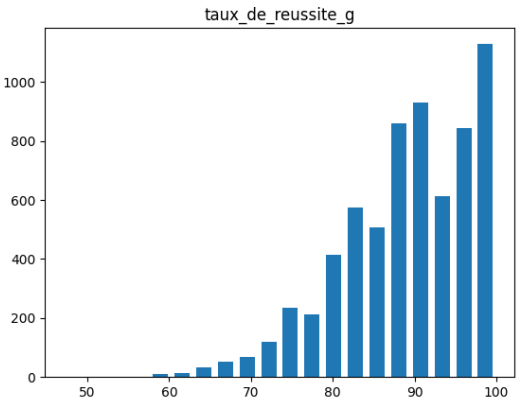
```
def Centreduire(T):
    Res = (T - np.mean(T, axis=0)) / np.std(T, axis=0)
    return Res

CollegesAr_CR = Centreduire(CollegesAr)
```

3 Exploration des données

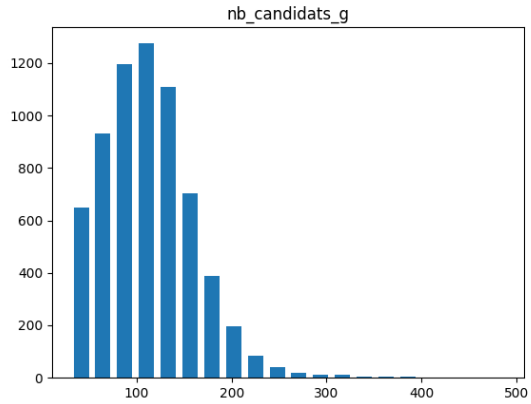
3.1 Représentations Graphiques



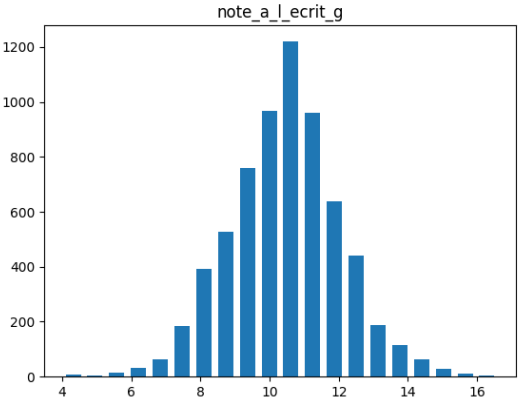
On remarque que le nombre de mentions "Très bien" obtenue par collèges est majoritaire ment d'environ 25. Très peu de collège obtiennent plus de 50 mentions.



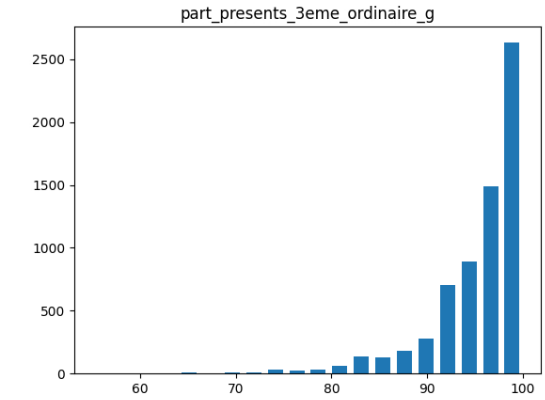
On remarque que beaucoup de collège ont un taux de réussite proche de 100%. De plus très peu de collège ont un taux de réussite inférieur à 70%



On remarque que la moyenne du nombre de candidat de filière générale au brevet par collège se situe entre 90 et 130 élèves.



On remarque que les notes moyennes sont majoritairement entre 10 et 11, que certains collèges peuvent avoir des moyennes en-dessous de 7, et que d'autre réussissent à monter jusqu'à 15 de moyenne.



On remarque que la part des élèves de 3ème de filière générale présent est majoritairement très proche de 100%. Certains collèges tombent sous 80%.

3.2 Matrice de Covariance

3.2.1 Démarche

Dans cette partie, on calcule la matrice de covariance afin de mesurer la relation linéaire entre différentes variables. Si deux variables ont une covariance positive, cela signifie qu'elles tendent à augmenter ou diminuer ensemble. Une covariance négative indique qu'elles évoluent en sens inverse. De plus, plus deux variables ont une covariance forte, plus elles évoluent rapidement.

```
coVarMat = np.cov(CollegesAr_CR, rowvar=False)
```

3.2.2 Matrice de Covariance

On obtient la matrice suivante :

	0	1	2	3	4
0	1.00015	0.701974	0.389442	0.605841	0.207924
1	0.701974	1.00015	0.0183992	0.125582	0.110071
2	0.389442	0.0183992	1.00015	0.715814	0.215875
3	0.605841	0.125582	0.715814	1.00015	0.292995
4	0.207924	0.110071	0.215875	0.292995	1.00015

4 Régression Linéaire Multiple

4.1 Utilisation de la Régression Linéaire Multiple : comment ?

En choisissant la 1ère 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.

4.2 Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui influent le plus possible sur le nombre de mentions "Très bien" obtenue au brevet des collèges, qui se trouve dans la colonne 0 de CollegesAr.

La colonne 0 de MatriceCov donne les coefficients de corrélation du nombre de mentions obtenues 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 : 0.702, 0.389, 0.606 et 0.208. Ils correspondent aux variables numéro 1, 2, 3, et 4. Les colonnes 1, 2, 3 et 4 de CollegesAr correspondent à :

— le nombre de candidats provenant de filière générale.

— le taux de réussite au brevet.

— la moyenne des notes du brevet à l'écrit.

— la part des élèves de 3ème présent au brevet

On choisit toutes nos variables comme variables explicatives, car le coefficient de corrélation de chaque variable statistique avec notre variable endogène est élevé.

4.3 Lien avec la problématique

Les paramètres de la régression linéaire multiple nous informeront des variables explicatives qui influencent le plus le nombre de mentions "Très bien" obtenu. 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 nombre de mentions obtenues.

4.4 Régression Linéaire Multiple avec Python

On fait maintenant la régression linéaire multiple avec Python :

```
Y = CollegesAr[:, 0]  # variable endogène
X = CollegesAr[:, 1:]  # variable statistiques
Y = Centreduire(Y)
X = Centreduire(X)

linearRegression = LinearRegression()
linearRegression.fit(X, Y)
coefs = linearRegression.coef_
CorS = math.sqrt(linearRegression.score(X, Y))
```

4.5 Paramètres, Interprétation

On obtient les paramètres suivant (variable *coefs* dans le script Python précédent) :

$$a_0 = 0.637, a_1 = 0.003, a_2 = 0.529, a_3 = -0.018$$

Le signe du paramètre a0 (positif) nous permet de voir qu'une augmentation du nombre de candidats provenant de filière générale est associée à une augmentation du nombre de mentions "Très bien". Plus ce nombre est élevé, plus il y a de mentions "Très bien".

Le signe du paramètre a1 (positif) nous permet de voir qu'une augmentation du taux de réussite au brevet est associée à une augmentation (bien que très faible) du nombre de mentions "Très bien". Un taux de réussite plus élevé tend donc légèrement à augmenter le nombre de mentions "Très bien".

Le signe du paramètre a2 (positif) nous permet de voir qu'une augmentation de la moyenne des notes du brevet à l'écrit est associée à une augmentation du nombre de mentions "Très bien". Une meilleure moyenne à l'écrit est donc un facteur important pour obtenir des mentions "Très bien".

Le signe du paramètre a3 (négatif) nous permet de voir qu'une augmentation de la part des élèves de 3ème présents au brevet est associée à une légère diminution du nombre de mentions "Très bien". Plus la proportion d'élèves de 3ème passant le brevet est élevée, moins il y a de mentions "Très bien".

De plus, comme les variables endogènes et statistiques sont centrées-réduites, on peut voir que plus la valeur absolue d'un coefficient est élevée, plus l'influence de celui-ci sur la variable endogène est grande. L'ordre des variables statistiques de la plus influente à la moins influente est :

1. le nombre de candidats provenant de filière générale.
2. la moyenne des notes du brevet à l'écrit.
3. la part des élèves de 3ème présent au brevet
4. le taux de réussite au brevet.

4.6 Coefficient de corrélation multiple, interprétation

On obtient le coefficient de corrélation multiple suivant (variable *CorS* dans le script Python précédent) :

$$CorS = 0.875$$

Ce coefficient de corrélation multiple est élevé (supérieur à 0.866), ce qui indique que le modèle explique une grande partie de la variance du nombre de mentions "Très bien". En d'autres termes, les variables explicatives utilisées dans le modèle (nombre de candidats provenant de filière générale, taux de réussite au brevet, moyenne des notes du brevet à l'écrit, part des élèves de 3ème présents au brevet) combinées expliquent bien la variation du nombre de mentions "Très bien".

5 Conclusions

5.1 Réponse à la problématique

Les résultats de l'analyse de régression multiple montrent que certaines données influent significativement sur le nombre de mentions "Très bien" obtenues au brevet :

1. Le nombre de candidats provenant de filière générale à la plus forte influence positive sur le nombre de mentions "Très bien".
2. La moyenne des notes du brevet à l'écrit a également une influence positive significative.
3. La part des élèves de 3ème présents au brevet a une influence négative, mais modeste.
4. Le taux de réussite au brevet a une influence positive très faible, presque négligeable.

En résumé, parmi les données de votre fichier, le nombre de candidats provenant de filière générale et la moyenne des notes du brevet à l'écrit sont les variables qui influencent le plus fortement le nombre de mentions "Très bien" obtenues au brevet.

5.2 Argumentation à partir des résultats de la Régression Linéaire

1. Nombre de candidats provenant de la filière générale : Les élèves de la filière générale sont souvent mieux préparés académiquement, ce qui augmente leurs chances d'obtenir une mention "Très bien".
2. Moyenne des notes du brevet à l'écrit : Une meilleure performance à l'écrit est un indicateur clé de l'obtention de la mention, reflétant une compréhension et une préparation solides.
3. Part des élèves de 3ème présents au brevet : Une plus grande proportion d'élèves passant le brevet peut diluer le nombre de mentions "Très bien", mais cet effet reste faible.
4. Taux de réussite au brevet : Le taux de réussite global a une influence négligeable, indiquant que la simple réussite ne suffit pas pour obtenir la mention.

5.3 Interprétation Personnelles

Les résultats montrent que le nombre de candidats de la filière générale et la moyenne des notes à l'écrit sont les facteurs déterminants pour l'obtention de mentions "Très bien" au brevet. Cette analyse met en évidence l'importance de la filière suivie et des performances académiques spécifiques plutôt que des indicateurs plus généraux comme la simple présence ou le taux de réussite global. Ces conclusions permettent de mieux comprendre les dynamiques qui sous-tendent les performances scolaires élevées et d'identifier les leviers les plus efficaces pour promouvoir l'excellence académique.