1 Les données Colleges.csv - Problématique

(a) Présentation des données

Le fichier Colleges.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 1e variable statistique sur cette population est la note moyenne à l'écrit du brevet en 2023 pour chaque collège.
- La 2e est la valeur ajoutée estimée sur la note au brevet moyenne pour ce collège.
- La 3e est le pourcentage de collégiens hors segpa et ulis dans chaque collège
- La 4e est le pourcentage de collégiens en LV1 allemand an 3e.
- La 5e est le pourcentage de collégiennes en 3e.
- La 6e est l'indice de position sociale (IPS) moyen du collège.
- La 7e est l'écart-type de l'IPS dans le collège.
- La 8e donne le classement REP/REP+ ou hors Education Prioritaire (EP) du collège.
- La dernière est l'académie dans laquelle se trouve le collège.

	A	B	C	D	E	F	G	H	I	J	K
1	numero_college	nom_etablissement	note_a_l_ecrit_g	va_de_la_note_g	pourcentage_hors_segpa_hors_ulis	pourcenetage_allemand_lv1	pourcentage_filles	ips	ecart_type_de_l_ips	ep_2022_2023	region_academique
2	0080046G	Collège Elisabeth de Nassau	10.5	-0.8	96.875000000000000000	42.187500000000000000	54.687500000000000000	95.7	34.2	HORS EP	GRAND-EST
3	0100807Y	Collège Pierre Brossolette	9.2	0.3	89.320388349514563107000	0.000000000000000000000	48.543689320388349515000	82.4	30.2	REP+	GRAND-EST
4	0520049W	Collège Anne Frank	8.1	-0.5	77.227722772277227723000	0.000000000000000000000	51.485148514851485149000	71.6	25.1	REP+	GRAND-EST
5	0511474A	Collège Pierre Gilles de Genne	9.8	NULL	93.406593406593406593000	0.000000000000000000000	47.252747252747252747000	85.4	28.5	HORS EP	GRAND-EST
6	0510044W	Collège Colbert	9.2	-0.6	97.560975609756097561000	13.008130081300813008000	46.341463414634146341000	76.8	23.6	REP+	GRAND-EST

(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 de bonnes notes au 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 :

```
CollegesDF=pd.read_csv("Colleges.csv")
```

(b) Mise en forme

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

```
CollegesDF = CollegesDF.dropna()
CollegesAr=CollegesDF.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 :

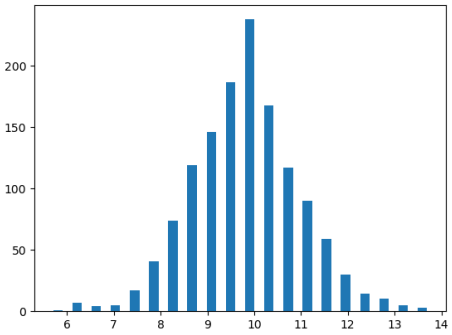
```
def Centreduire(T):
    T=np.array(T,dtype=np.float64)
    (n,p)=T.shape
    res=np.zeros((n,p))
    TMoy=np.mean(T,axis=0)
    TEcart=np.std(T,axis=0)
    for j in range(p):
         res[:,j]=(T[:,j]-TMoy[j])/TEcart[j]
    return res

CollegesAr0=CollegesAr[:,2:9]
CollegesAr0_CR=Centreduire(CollegesAr0)
```

3 a. Exploration des données : représentations graphiques

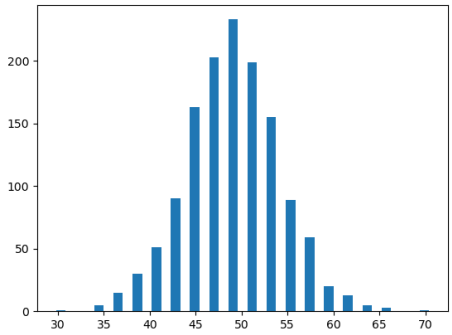
On choisit d'étudier les diagrammes en batons des nos variables statistiques :

Diagramme en batons des notes au Brevet



On remarque que les notes moyennes sont majoritairement en-dessous de 10, et que certains collèges peuvent avoir des moyennes en-dessous de 7 (!)

Diagramme en batons des pourcentages de filles



On remarque que la plupart des collèges sont autour de la parité (entre 45 et 55% de filles), mais certains collèges ont un fort déséquilibre (jusqu'à 2/3 de filles, ou 2/3 de garçons).

etc...

(à compléter)

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(CollegesAr0_CR,rowvar=False)
```

(b) Matrice de covariance

On obtient la matrice suivante :

	0	1	2	3	4	5	6
0	1.00075	0.486237	0.0180225	0.00903441	0.0305552	0.617076	0.605737
1	0.486237	1.00075	0.00741649	0.0557707	-0.019603	-0.168779	-0.0987854
2	0.0180225	0.00741649	1.00075	-0.0181191	0.108842	0.169584	0.117024
3	0.00903441	0.0557707	-0.0181191	1.00075	0.0228675	-0.0906089	-0.0267153
4	0.0305552	-0.019603	0.108842	0.0228675	1.00075	-0.0236116	-0.0355643
5	0.617076	-0.168779	0.169584	-0.0906089	-0.0236116	1.00075	0.825416
6	0.605737	-0.0987854	0.117024	-0.0267153	-0.0355643	0.825416	1.00075

4 Régression linéaire multiple

(a) 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.

(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 0 de `CollegesAr0`. La colonne 0 de `MatriceCov` donne les coefficients de corrélation de la note 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 la note au brevet.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de MatriceCov sont : 0.617, 0.606, 0.486 et 0.031. Ils correspondent aux variables numéro 5, 6, 1 et 4. Les colonnes 5, 6, 1 et 4 de `CollegesAr0` correspondent aux :

- IPS de chaque collège,
- écart-type de l'IPS dans le collège,
- valeur ajoutée du collège,
- pourcentage de filles.

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 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.

(d) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

...

(e) Paramètres, interprétation

On obtient les paramètres $a_0 = \ldots$, $a_1 = \ldots$, $\ldots\ldots$
Le signe du paramètre a_0 nous permet de voir...

...

Comme les variables endogène et explicatives sont centrées-réduite, on peut de plus voir ...

...

(f) Coefficient de corrélation multiple, interprétation

...

5 Conclusions

(a) Réponse à la problématique

...

(b) Argumentation à partir des résultats de la régression linéaire

...

(c) Interprétations personnelles

...