0. Les données sae - Problématique

a) Présentation des données

Le Fichier sae.csv contient plusieurs séries statistiques sur l'ensemble de toutes les formations répertoriées dans Parcoursup :

- La population est l'ensemble des formation, représentées par leur code_aff_form
- La 1ère série correspond à l'effectif total d'admis
- La 2e série correspond au nombre de places disponibles par formation
- La 3e série correspond au nombre total de candidats
- La 4e série correspond au nombre total de candidates
- La 5e série correspond au rang du dernier appelé de la formation
- La 6e série correspond à l'effectif d'amis avec la mention très bien, très bien avec félicitations du jury et mention bien

b) Problématique

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

Comment influe la capacité d'accueil, l'effectif total des candidats, l'effectif total des candidates, le rang du dernier appelé, le nombre d'admis qui ont obtenu la mention bien, mention très bien avec et sans félicitations du jury sur l'effectif total d'admis dans une formation spécifique ?

c) Utilisation de la régression linéaire multiple : comment ?

En choisissant la 1ère série statistique comme **variable endogène** et les autres séries comme **variables explicatives**, la **régression linéaire multiple** nous permettrait d'obtenir une estimation de l'effectif total d'admis dans une formation spécifique en fonction d'autres informations.

d) Utilisation de la régression linéaire multiple : pour quoi ?

Les **paramètres** de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus l'effectif total d'amis dans une formation spécifique. En observant si cette **estimation** est proche de la réalité, on aura une réponse à la problématique.

1. Import des données, mise en forme

a) Importer les données

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

```
vue_DF=pd.read_csv("sae.csv")
```

b) Mise en forme

Nous n'avons pas besoin de supprimer les cases vides (qui contiennent nan en Python) car nous n'en avons pas.

Nous avons transformé notre DataFrame en Array :

```
vue_A=vue_DF.to_numpy()
```

c) Centrer-réduire

On enlève la première colonne de notre tableau, qui contient les codes des formations et qui n'est donc pas une donnée statistique :

```
def CentreReduire(a):
    a=np.array(a,dtype=np.float64)
    res=np.empty(a.shape[0])
    moy=a.mean(axis=0)
    e_t=a.std(axis=0)
    res=a-moy
    res=res/e_t
    return res

vue_CR = CentreReduire(vue_A[:,1:])
```

2. Choix des variables explicatives

a) Démarche

Dans cette partie, on réduit le nombre de variables explicatives pour ne garder que les plus pertinentes. On commence par calculer la matrice de covariance :

```
matriceCov=np.cov(vue_CR,rowvar=False)
```

b) Matrice de covariance

On obtient la matrice suivante :

	0	1	2	3	4	5
0	1.00002	0.678265	0.593849	0.543481	0.467772	0.662891
1	0.678265	1.00002	0.342278	0.316096	0.27055	0.378113
2	0.593849	0.342278	1.00002	0.924397	0.693167	0.541495
3	0.543481	0.316096	0.924397	1.00002	0.675163	0.462653
4	0.467772	0.27055	0.693167	0.675163	1.00002	0.370638
5	0.662891	0.378113	0.541495	0.462653	0.370638	1.00002

c) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible le nombre de candidats admis dans une formation qui se trouve dans la colonne 0 de vue_A.
La colonne 0 de matriceCov donne les coefficients de corrélation du nombre de candidats admis dans une formation avec chacune des autres variables/colonnes de vue_A. On prend toutes les autres variables comme variables explicatives car elles ont un coefficient de corrélation grand (en valeur absolue) avec le nombre de candidats admis dans une formation.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de matriceCov sont : 0.678, 0.662, 0.593, 0.543 et 0.467. Ils correspondent aux variables 1,5,2,3 et 4. Les colonnes 1,5,2,3 et 4 de vue_A correspondent aux :

- la capacité de la formation
- l'effectif admis selon leurs mentions
- l'effectif total de candidats
- l'effectif total de candidates
- le rang du dernier appelé

On choisit donc toutes ces variables comme variables explicatives.

3. Régression linéaire multiple pour sae.csv

a) Régression linéaire multiple

On fait maintenant la régression linéaire multiple avec la série du nombre de candidats admis comme variable endogène et les 5 variables ci-dessus comme variables explicatives.

```
Y = vue_CR[:, 0]
X = vue_CR[:, 1:]

linear_regression = LinearRegression()

linear_regression.fit(X, Y)

a = linear_regression.coef_

Cor = linear_regression.score(X, Y)
```

b) Paramètres, interprétation

Le coefficient 0.456 représente la capacité d'une formation, on peut voir que cette variable explicative influe positivement sur l'effectif total de candidats admis car le coefficient est supérieur à 0.
Le coefficient 0.151 représente l'effectif total des candidats, on peut voir qu'elle influe positivement mais faiblement notre variable endogène.
Le coefficient 0.038 représente l'effectif total de candidates, on peut voir qu'elle influe à peine positivement la variable endogène car le coefficient est très proche de 0.
Le coefficient 0.079 représente le rang du dernier appelé, on peut voir que cette variable explicative influe à peine sur l'effectif total de candidats admis.
Le coefficient 0.361 représente l'effectif d'admis selon les mentions qui influe positivement sur l'effectif total de candidats admis.

```
Paramètres a :
 [0.45619455 0.15132404 0.03863563 0.07953314 0.3610957 ]
```

Aucun de nos paramètres n'a une très grande influence car ils ne sont jamais très proche de 1 sur l'effectif d'admis cependant ils influent quand même.

c) Coefficient de corrélation multiple, interprétation

Le coefficient de corrélation multiple est plutôt élevé ce qui signifie qu'il est possible de prédire l'effectif admis assez précisément.

```
Coefficient de corrélation :
 0.6968373273860246
```

4. Conclusions

a) Réponse à la problématique
b) Argumentation à partir des résultats de la régression linéaire
c) Interprétation personnelles