1 Les données Colleges.csv - Problématique

(a) Présentation des données 2023-2024

Le fichier vue_segpa.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 le nombre d'élèves en segpa sur plusieurs années scolaires pour chaque collège.
- La 2e est l'effectif de filles pour ce collège.
- La 2e est l'effectif de garçons pour ce collège.
- La 4e est la latitude du collège .
- La 5e est la longitude du collège.

nbre_eleves_segpa	latitude	longitude	effectifs_filles	effectifs_garcons
0	45.195224607476455	5.680420097587849	195	222
0	48.768199454529935	2.4057841110469402	240	230
0	43.30251976133512	5.388918166791596	434	449
0	50.29266344220489	3.9204097660583948	153	140
62	50.50657276230646	2.4645395665443663	231	225

(b) Problématique

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

Est-ce que la position géographique et le sexe influe sur le nombre d'élèves en segpa ?

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 :

```
CollegeDF = pd.read_csv("vue_segpa.csv", sep=";")
```

(b) Mise en forme

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

```
CollegeDF = CollegeDF.dropna()
CollegeAr = np.array(CollegeDF)
```


(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):
  moyenne = np.mean(T, axis=0)
  ecart_type = np.std(T, axis=0)

  Res = (T - moyenne) / ecart_type

  return Res
```

3 a. Exploration des données : représentations graphiques

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

Diagramme en bâtons de l'effectif de garçon

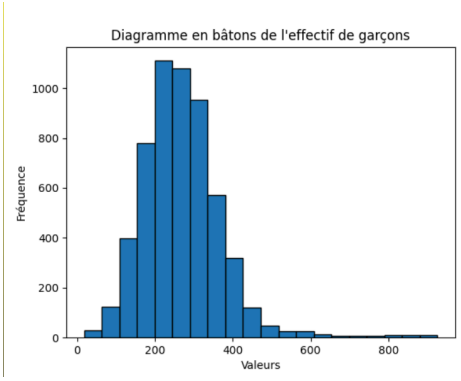
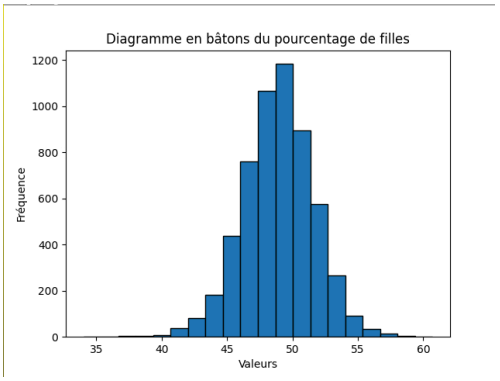


Diagramme en bâtons du pourcentage de filles



La majorité des collèges ont un effectif de garçons variant de 100 à 300, avec quelques collèges ayant des effectifs beaucoup plus élevés.

Le pourcentage de filles dans les collèges varie entre 20% et 60%. La majorité des collèges ont un pourcentage de filles autour de la parité, avec une moyenne à 44.14%.

3 b. Exploration des données : matrice de covariance

(a) Démarche

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

```
MatriceCov=np.cov(CollegesAr0_CR,rowvar=False)
```

(b) Matrice de covariance

On obtient la matrice suivante :

Matrice de Covariance :

	nbre_eleves_segpa	latitude	...	effectifs_filles	effectifs_garcons
nbre_eleves_segpa	1066.378719	-50.903098	...	973.864020	1147.172422
latitude	-50.903098	262.037895	...	-587.643254	-585.727880
longitude	10.645621	-49.197414	...	464.954484	473.901724
effectifs_filles	973.864020	-587.643254	...	10411.645687	10177.743214
effectifs_garcons	1147.172422	-585.727880	...	10177.743214	10712.217284

4 Régression linéaire multiple

La variable endogène (y) sera nbre_eleves_segpa, et nous choisirons comme variables explicatives (X) celles qui ont les coefficients de corrélation les plus grands (en valeur absolue) avec nbre_eleves_segpa.

(b) Variables explicatives les plus pertinentes

La régression linéaire multiple montre que les effectifs de garçons et de filles sont les principales variables explicatives pour le nombre d'élèves en SEGPA, suivies par la position géographique (longitude et latitude). Les résultats montrent que le sexe et la localisation géographique jouent un rôle important dans le nombre d'élèves en SEGPA.

(c) Lien avec la problématique

Les paramètres de la régression linéaire multiple nous informeront des variables explicatives qui influencent le plus le nombre d'élèves en SEGPA. En calculant le coefficient de corrélation multiple, on saura de plus si cette influence permet de prédire la réalité et si ces variables peuvent prédire avec précision le nombre d'élèves en SEGPA.

(d) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

```
# Régression linéaire multiple
linear_regression = LinearRegression()
linear_regression.fit(X, Y)
a = linear_regression.coef_
```

(e) Paramètres, interprétation

On obtient les paramètres *X = np.delete(CollegeAr_CR, 1, axis=1)*, Y = CollegeAr_CR[:, 1]

```
Coefficient de la corrélation multiple :  0.13387502931493778
```

(f) Coefficient de corrélation multiple, interprétation

Le coefficient n'est pas assez élevé pour affirmer avoir une corrélation entre les différentes variables et le nombre de SEGPA

5 Conclusions

(a) Réponse à la problématique

Puisque le coefficient est inférieur à 87%, on peut affirmer que l'effectif et la position géographique n'influent pas sur le nombre d'élèves en SEGPA.