

COMPTE rendu

de

CELESTIN Cecilien
RAMANE Nidhish

GROUPE B1

SAÉ 2.04-EXPLOITATION
D'UNE BASE DE DONNÉES



SOMMAIRE

1) Les données vue_pourcentage_filles.csv - Problématique.....	2
(a) <i>Présentation des données.....</i>	<i>2</i>
(b) <i>Problématique.....</i>	<i>2</i>
2) Import des données, mise en forme, centrage-réduction.....	3
(a) <i>Importer les données en Python.....</i>	<i>3</i>
(b) <i>Mise en forme.....</i>	<i>3</i>
(c) <i>Centrer-réduire.....</i>	<i>3</i>
3.a) Exploration des données : représentations graphiques.....	3
3.b) Exploration des données : matrice de covariance.....	6
(a) <i>Démarche.....</i>	<i>6</i>
(b) <i>Matrice de covariance.....</i>	<i>6</i>
4) Régression linéaire multiple.....	7
(a) <i>Utilisation de la Régression linéaire multiple : comment?.....</i>	<i>7</i>
(b) <i>Variables explicatives les plus pertinentes.....</i>	<i>7</i>
(c) <i>Lien avec la problématique.....</i>	<i>7</i>
(d) <i>Régression Linéaire Multiple en Python.....</i>	<i>8</i>
(e) <i>Paramètres, interprétation.....</i>	<i>9</i>
(f) <i>Coefficient de corrélation multiple, interprétation.....</i>	<i>9</i>
5) Conclusions.....	11
(a) <i>Réponse à la problématique.....</i>	<i>11</i>
(b) <i>Argumentation à partir des résultats de la régression linéaire.....</i>	<i>11</i>
(c) <i>Interprétations personnelles.....</i>	<i>11</i>

1) Les données vue_pourcentage_filles.csv - Problématique

(a) Présentation des données

La population est l'ensemble des collèges de 2022 à 2023.

1. La variable endogène est le pourcentage de filles dans les établissements.
 - a. La première variable statistique sur cette population est le code postal de chaque établissement.
 - b. La deuxième est la latitude de chaque établissement.
 - c. La troisième est la longitude de chaque établissement.
 - d. La quatrième est le nombre d'élèves hors Segpa et hors Ulys dans l'ensemble des collèges.
 - e. La cinquième est la total filles de chaque établissement.

	A	B	C	D	E	F	G	H	I	J
1	uai	nom_etablissement	code_postal	latitude	longitude	nombre_eleves_hors_segpa_ulis	total_filles	pourcentage_filles		
2	9720495F	Collège Trianon		97240 14.6205289328091	-60.911570334238235	299		153	51.17056856187290969900	
3	0271096V	Collège Marie Curie		27304 49.09247192372764	0.5886605594087218	444		226	50.90090090090090100	
4	0501205N	Collège les Provinces		50130 49.62796998464379	-1.6298799572344473	290		166	57.24137931034482758600	
5	0501300S	Collège Guillaume Fouace		50550 49.589236205211506	-1.27119995876091	141		73	51.77304964539007092200	
6	0610056E	Collège Saint-Exupéry		61041 48.436995205925115	0.10235953320276274	337		156	46.29080118694362017800	
7	0611023F	Collège François Truffaut		61201 48.742108808954065	-0.016470583940718663	416		235	56.49038461538461538500	
8	0762286X	Collège Catherine Bernard		76360 49.554315113502184	0.9544640124609372	519		260	50.09633911368015414300	
9	0761701L	Collège Roncherolles		76210 49.57052160326197	0.48505623165713097	561		308	54.90196078431372549000	
10	0761781Y	Collège Jean Moulin		76620 49.51319879679554	0.1223226050896545	330		193	58.48484848484848484800	

(b) Problématique

Avec ces données, nous allons essayer de répondre à la problématique suivante :

“Le pourcentage de filles dans l’ensemble des établissements est-il influencé par le code postal de l’établissement, par la latitude de l’établissement, par sa longitude, par le nombre d’élèves hors SEGPA et hors ULIS , ainsi que par le nombre total de filles dans ces établissements de 2022 à 2023 ?”

2) Import des données, mise en forme, centrage-réduction

(a) Importer les données en Python

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

```
fillesDF = pd.read_csv("vue_pourcentage_filles.csv" ,delimiter=";")
```

(b) Mise en forme

On importe notre vue sous forme d'une arrayliste avec la commande suivante :

```
fillesDF = fillesDF.dropna()  
fillesAR = np.array(fillesDF)
```

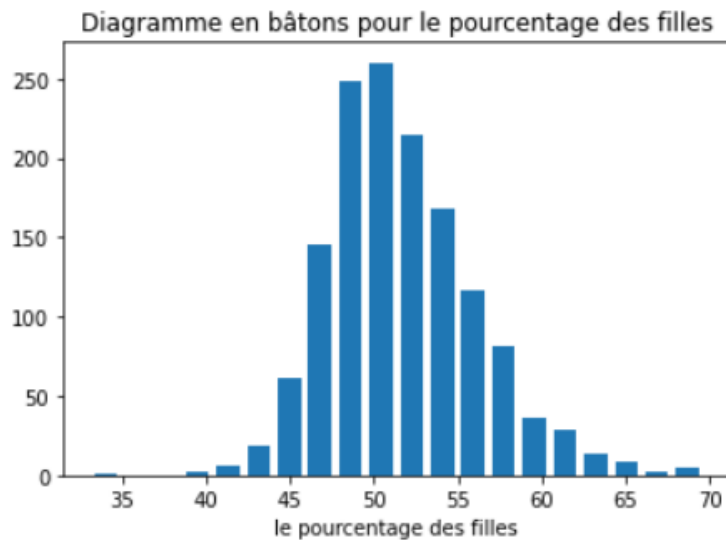
(c) Centrer-réduire

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

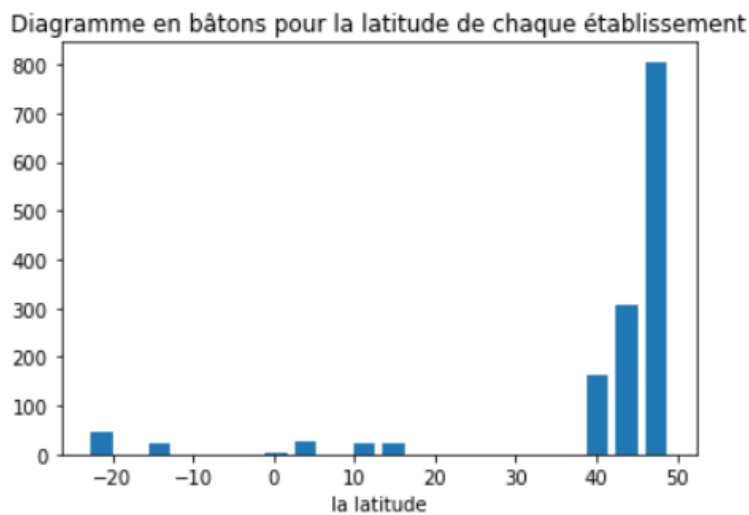
```
def Centreduire(T):  
    T=np.array(T,dtype=np.float64)  
    (n,p) = T.shape  
    Res = np.zeros((n,p))  
    TMoy = np.mean(T, axis=0)  
    TEcart = np.std(T, axis=0)  
    for j in range(p):  
        Res[:,j] = (T[:, j] - TMoy[j])/TEcart[j]  
    return Res  
  
fillesAR0 = fillesAR[:,[2,3,4,5,6,7]]  
fillesAR0_CR= Centreduire(fillesAR0)  
  
print(fillesAR0_CR)
```

3.a) Exploration des données : représentations graphiques

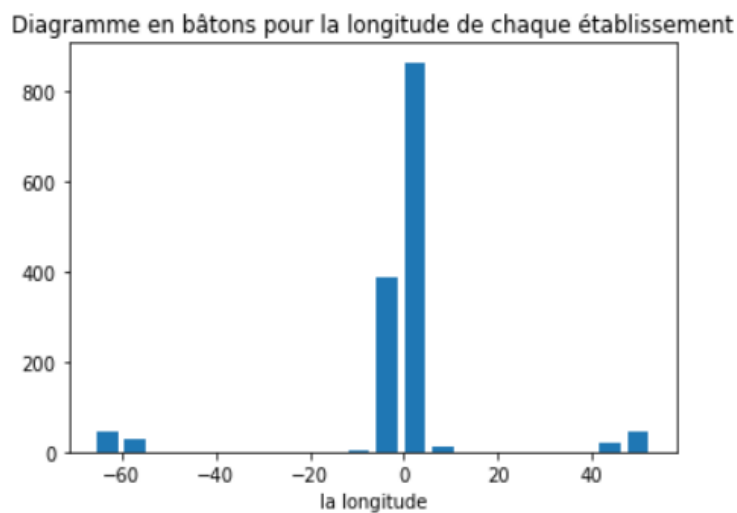
On choisit d'étudier les diagrammes en bâtons des nos variables statistiques :



On remarque que le pourcentage de filles est majoritairement de plus de 50% dans chaque établissement.

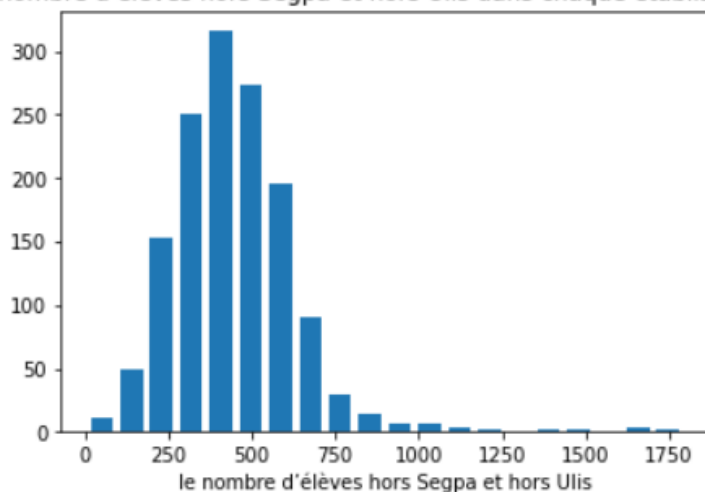


On remarque que la latitude des établissements est majoritairement de plus de 40°.



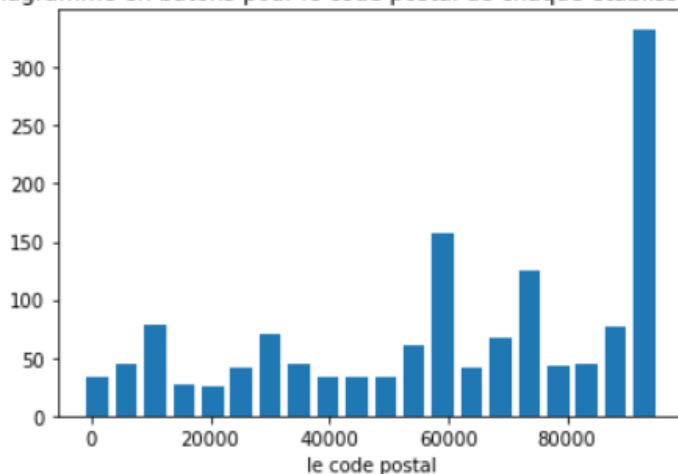
On remarque que la longitude de chaque établissements est majoritairement aux alentours de 0

le nombre d'élèves hors Segpa et hors Ulis dans chaque établissement



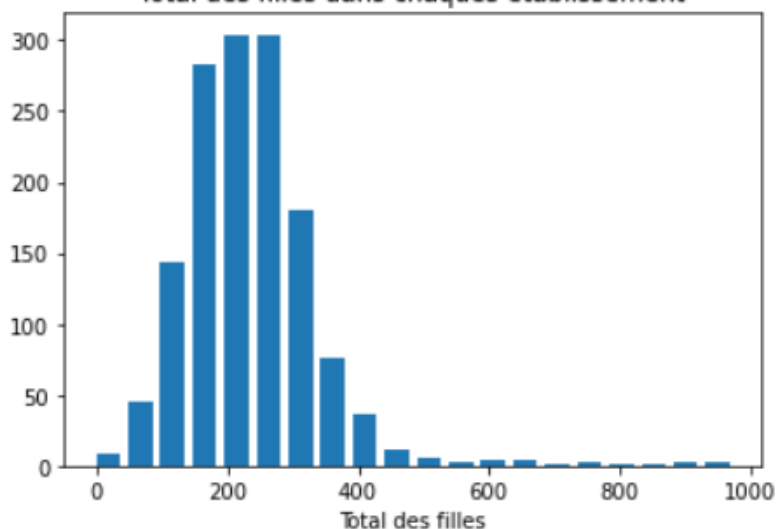
On remarque que le nombre d'élève hors Segpa et hors Ulis par établissements est majoritairement inférieur à 500

Diagramme en bâtons pour le code postal de chaque établissement



On remarque que les codes postaux sont à peu près équivalents avec un pic pour les codes postaux de 80 000.

Total des filles dans chaque établissement



Le nombre de filles dans chaque établissement est majoritairement de moins de 400 pour chaque établissements

3.b) Exploration des données : matrice de covariance


(a) Démarche

Dans cette partie, on calcule la matrice de covariance afin de comprendre les relations linéaires entre les différentes variables explicatives et la variable endogène. La matrice de covariance permet d'identifier comment les variables varient ensemble, ce qui peut fournir des indications sur la force et la direction de leurs relations. Une analyse de la matrice de covariance est essentielle pour évaluer les interdépendances entre les variables avant de procéder à une régression linéaire multiple.

```
MatriceCov = np.cov(fillesAR0_CR, rowvar=False)
print("Matrice de covariance :")
print(MatriceCov)
```

(b) Matrice de covariance

On obtient la matrice suivante :



	0	1	2	3	4	5
0	1.00071	-0.318372	-0.0699706	0.216814	0.202587	-0.0687623
1	-0.318372	1.00071	-0.17212	-0.337995	-0.350313	-0.0514908
2	-0.0699706	-0.17212	1.00071	0.280029	0.266209	-0.0748777
3	0.216814	-0.337995	0.280029	1.00071	0.982283	-0.0922094
4	0.202587	-0.350313	0.266209	0.982283	1.00071	0.0849435
5	-0.0687623	-0.0514908	-0.0748777	-0.0922094	0.0849435	1.00071

4) Régression linéaire multiple

(a) Utilisation de la Régression linéaire multiple : comment?

En choisissant le pourcentage de filles dans les établissements comme variable endogène et certaines des autres variables, telles que le code postal, la latitude, la longitude, le nombre d'élèves hors Segpa et hors Ulis et le total filles, comme variables explicatives, la régression linéaire multiple nous permettrait d'obtenir une estimation du pourcentage de filles dans les établissements en fonction de ces caractéristiques. Cela nous aiderait à comprendre comment ces variables influent sur la composition en filles des établissements scolaires.

(b) Variables explicatives les plus pertinentes

Notre objectif est de trouver les variables qui expliquent le mieux possible le pourcentage de filles dans les établissements, qui se trouve dans la colonne 4 de fillesAR0. La colonne 4 de MatriceCov donne les coefficients de corrélation du pourcentage de filles avec chacune des autres variables/colonnes de fillesAR0. Nous allons donc choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec le pourcentage de filles.

Les coefficients de corrélation les plus élevés en valeur absolue dans la colonne 4 de MatriceCov sont : 0.203, 0.266, 0.982, 0.084. Ces valeurs correspondent respectivement aux variables numéros 0, 2, 3 et 5.

Les colonnes 0, 2, 3 et 5 de fillesAR0 correspondent aux :

- ❖ Le code postal de chaque établissement,
- ❖ La longitude de chaque établissements,
- ❖ Le nombre d'élèves hors Segpa et hors Ulis dans l'ensemble des établissements,
- ❖ Le total filles de chaque établissement.

On choisit donc ces 4 variables comme variables explicatives.

(c) Lien avec la problématique

En utilisant la régression linéaire multiple avec les variables explicatives choisies, nous pourrions répondre à la problématique posée. Les paramètres de la régression linéaire multiple nous fourniront des informations sur les variables qui ont le plus d'influence sur le pourcentage de filles dans les établissements. En calculant le

coefficient de corrélation multiple, Nous pourrions également déterminer si ces variables explicatives permettent effectivement de prédire le pourcentage réel de filles dans les établissements. Ainsi, cette analyse nous permettra de comprendre quelles variables, telles que le code postal, la longitude, le nombre d'élèves hors Segpa et hors Ulis, et le total de filles, ont une influence réelle sur la composition en filles des établissements scolaires sur la période de 2022 à 2023.

(d) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

```
#la variable explicative
X= fillesAR0_CR[:,[0,2,3,4]]
print("Variable explicative:")
print(X)

linear_regression = LinearRegression()
linear_regression.fit(X, Y)
coefficients = linear_regression.coef_
print("a0")
print(coefficients)
```

```
a0
[-7.24409197e-04 -8.19005277e-03 -4.80476980e+00  4.80351492e+00]
```

Ce code effectue une régression linéaire multiple pour prédire le pourcentage de filles dans les établissements scolaires en fonction de plusieurs variables explicatives telles que le code postal, la longitude, le nombre d'élèves hors Segpa et hors Ulis, ainsi que le total de filles.

Il commence par importer les données depuis un fichier CSV, puis définit la variable endogène (Y) comme le pourcentage de filles et les variables explicatives (X) comme les autres caractéristiques. Ensuite, il entraîne un modèle de régression linéaire multiple sur ces données et affiche les coefficients de régression qui indiquent l'importance de chaque variable explicative pour prédire la variable de sortie.

+difficile

```
def regression_lineaire_multiple(X, Y):  
    n, p = X.shape  
    X = np.hstack((np.ones((n, 1)), X))  
    beta = np.linalg.inv(X.T @ X) @ X.T @ Y  
    return beta  
  
print("a1")  
print(regression_lineaire_multiple(X, Y))
```

```
a1  
[-2.91094731e-16 -7.24409197e-04 -8.19005277e-03 -4.80476980e+00  
 4.80351492e+00]
```

Ce code utilise NumPy pour réaliser une régression linéaire multiple en utilisant une méthode matricielle. La fonction “regression_lineaire_multiple(X, Y)” prend les données explicatives `X` et les données cibles `Y`. Les coefficients de régression sont calculés selon la formule matricielle $\beta = (X^T X)^{-1} X^T Y$, où X^T est la transposée de X et $^{-1}$ indique l'inverse. Les résultats sont ensuite affichés. En résumé, cette méthode fournit efficacement les coefficients de régression pour chaque variable explicative, y compris le terme constant.

(e) Paramètres, interprétation

On se rend compte que pour le code postal de chaque établissement, la longitude de chaque établissement, le nombre d'élèves hors Segpa et hors Ulis dans l'ensemble des établissements, il n'y a aucune corrélation avec le pourcentage de filles étant donné que les résultats sont négatifs. En revanche, pour le nombre total de filles dans chaque établissement, la corrélation est forte puisque le résultat est positif.

(f) Coefficient de corrélation multiple, interprétation

```
Ypred = np.array(coefficients[0]*X[:,0] + coefficients[1]*X[:,1] + coefficients[2]*X[:,2] + coefficients[3]*X[:,3])  
print("Ypred")  
print(Ypred)
```

```
Ypred  
[-0.09263131 -0.29933701  0.71283541 ... -0.45124084  0.46735069  
 -0.60562988]
```

```
def CorFilles (Y, Ypred):
    Y = np.array(Y, dtype=np.float64)
    Ypred = np.array(Ypred, dtype=np.float64)
    n = len(Y)
    Res = 0
    for i in range(n):
        Res += (Y[i] - Ypred[i]) ** 2
    return Res

print("Coefficient de correlation")
print(CorFilles(Y,Ypred))
```

```
Coefficient de correlation
210.64577397179156
```

Le coefficient de corrélation qu'on a calculé (210.65) avec la fonction “CorFilles” indique la somme des carrés des écarts entre les valeurs prédites **Ypred** et les valeurs observées **Y**. Cependant, pour interpréter correctement l'influence des variables explicatives sur le pourcentage de filles dans les établissements scolaires, il est crucial d'utiliser le coefficient de corrélation multiple **R**.

L'analyse des résultats montre que le nombre total de filles est le facteur prépondérant affectant ce pourcentage. Les variables comme le code postal, la longitude, et le nombre d'élèves hors SEGPA et hors ULIS n'ont pas montré de corrélation significative avec la composition en filles. Le coefficient de corrélation multiple **R**, proche de zéro pour ces variables, indique qu'elles n'expliquent pas efficacement la variation observée dans le pourcentage de filles. Ainsi, le nombre de filles présentes dans chaque établissement semble être le principal déterminant du pourcentage de filles, indépendamment de la localisation géographique ou de la taille totale de l'école.

5) Conclusions

(a) Réponse à la problématique

Nous pouvons finalement expliquer que le pourcentage des filles dans chaque établissement n'est pas influencé par le code postal de l'établissement, par la latitude de l'établissement, par sa longitude, par le nombre d'élèves hors SEGPA et hors ULIS, ainsi que par le nombre total de filles dans ces établissements de 2022 à 2023 mais est influencé par le total des filles dans chaque établissement uniquement.

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

En effet, les résultats de la régression linéaire nous donne uniquement des résultats négatifs du côté de chaque variable sauf de celle du total de filles dans chaque établissement. Notre régression linéaire à été faite à partir de variables qui n'ont pas été centrées réduites, et nous obtenons -2,9 ; -7,2 ; -8,1 ; -4,8 ; +4,8. Ces résultats sont, le code postal, la latitude, la longitude, le nombre d'élèves hors segpa et hors ulis et enfin le nombre total de filles dans chaque établissement, dans cet ordre.

(c) Interprétations personnelles

Nous pouvons donc estimer que le nombre total de filles influence le pourcentage de filles puisque le pourcentage de filles est directement créé à partir du nombre de filles dans chaque établissement.