Marius CHARTIER—LE GOFF
Paolo TOE

SAE 2.04

Exploitation d'une base de données

Analyse de données

6/16/2024

IUT DE LANNION

Sommaire

1 - Problématique

Population : Etablissements

Variables statistiques :

- 1e : Nombre d'occurrences de la lettre "a" dans le nom de la commune
- 2e : Latitude de l'établissement
- 3e : Longitude de l'établissement
- 4e : Taux de réussite g de l'établissement
- 5e : Nb candidats g de l'établissement,

	A	B	C	D	E	F	G
1	uai	nb_a_in_mot	latitude	longitude	taux_de_reussite_g	nb_candidats_g	nb_qp_academie
2	0010035H	1	46.25485364646008	5.638289823414171	90.0	125	193
3	0010039M	2	45.9415466100717	5.431009651453726	79.0	62	193
4	0010802S	1	46.253597238029414	5.637920168031979	75.0	153	193
5	0010974D	0	46.20705911951749	5.24540999711648	64.0	118	193
6	0010975E	1	46.247648621001346	6.031359448747479	83.0	160	193

PROBLEMATIQUE :

Est-ce que ces variables peuvent-elles permettre d'expliquer le nombre total de quartiers prioritaires (qp) d'une académie selon ses établissements ?

Exemple :

Lycée "LeDantec" a 3 quartiers prioritaires à proximité et collège "Diwan Penn ar Bed" en a 2 donc, avec les variables explicatives du Lycée "LeDantec" on devrait pouvoir trouver que l'Académie de rennes à 5 quartiers prioritaires à proximité de ses établissements.

2 – Import des données, mise en forme

A - Importation des données :

```
data_debase= pd.read_csv("data_s204_p3.csv")
```

B - Mise en forme des données :

```
data_debase_np = data_debase.to_numpy()
```

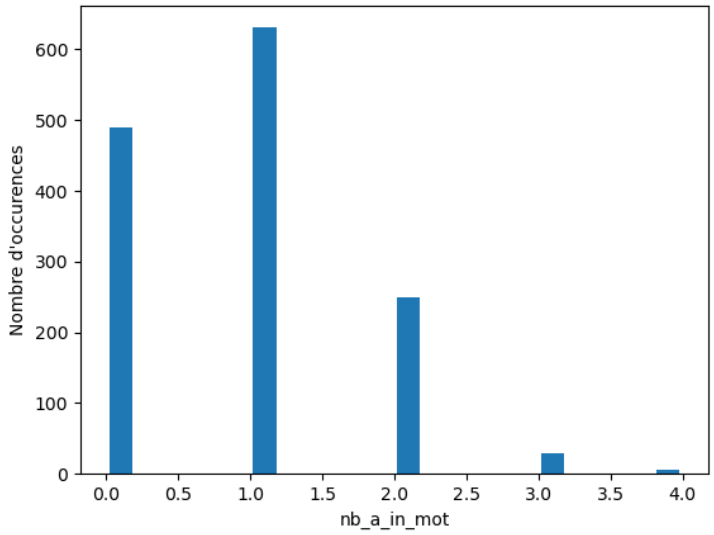
C - Centrer-réduire :

```
def Centreduire(T):
T = np.array(T,dtype=np.float64)
(n,p) = T.shape
Moyennes = np.mean(T, 0)
EquartTypes = np.std(T, 0)
Res = np.eye(n,p)
for j in range(p):
Res[:,j]= (T[:,j]-Moyennes[j])/EquartTypes[j]
```

3 - Exploration des données

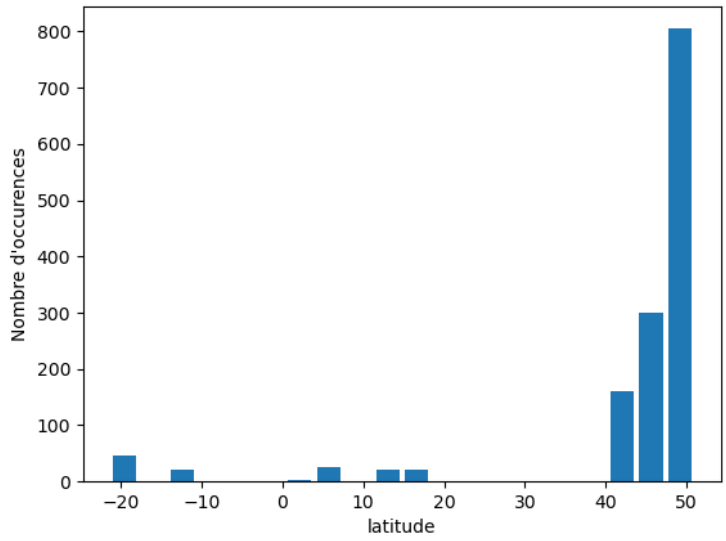
A - Exploration des données : représentations graphiques

Diagramme en bâtons du nombre de A dans le nom de commune où se situe l'établissement :



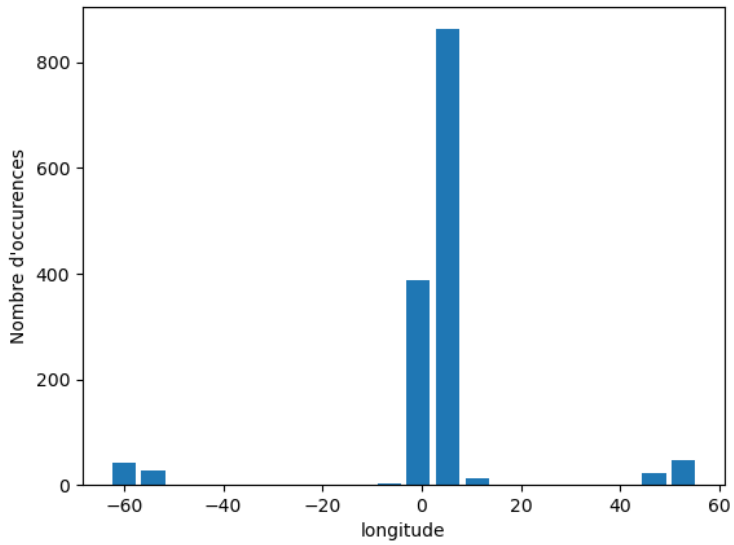
On remarque que la majorité des établissements sont dans des communes avec un seul A.

Diagramme en bâtons de la latitude des établissements :



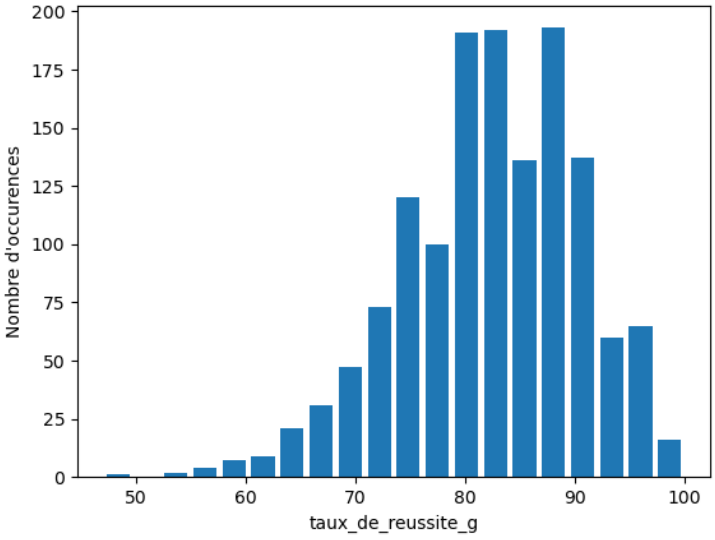
La majorité des établissements se situent à la latitude 50, vers l'Ile-de-France, et ce qui est surprenant est que certains semblent être dans l'hémisphère sud.

Diagramme en bâtons de la longitude des établissements :



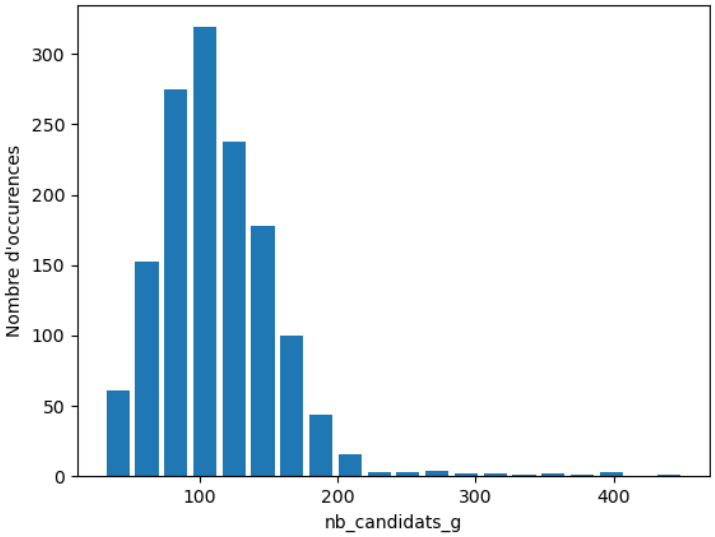
Les établissements se situent près du premier méridien, en France, avec quelques-uns en outre-mer.

Diagramme en bâtons du taux de réussite des établissements :



Ici on peut voir que la grande majorité des établissements ont un taux de réussite au-dessus de 80% pourcent, avec certains approchant 100%.

Diagramme en bâtons du nombre de candidats des établissements :



Le nombre de candidats approchent les 100% pour la plupart des établissements.

B - Exploration des données : matrice de covariance

(a) Démarche

Dans cette partie, on calcule la matrice de covariance avec la fonction numpy cov.

```
MatCovar = np.cov(X,rowvar=False)
```

(b) Matrice de covariance

	A	B	C	D	E	F	G
1		nb_qp_academie	nb_a_in_mot	latitude	longitude	taux_de_reussite_g	nb_candidats_g
2	nb_qp_academie	1.00071225071225	-0.0689276598680313	0.292059877003812	0.0765471497418443	0.0317376195670296	0.134146716156051
3	nb_a_in_mot	-0.0689276598680313	1.00071225071225 0562e+00	-1.34293671206494 5470e-01	-2.50618479986413 6920e-02	4.67223919898347 1620e-02	-1.05582916323133 3107e-02
4	latitude	0.292059877003812	-1.34293671206494 5470e-01	1.00071225071225 1006e+00	-1.95216576196013 7358e-01	7.95451749486546 6498e-02	-2.88788978898121 7731e-01
5	longitude	0.0765471497418443	-2.50618479986413 6920e-02	-1.95216576196013 7358e-01	1.00071225071225 1006e+00	-1.10954649666692 3981e-01	2.33754709568669 7551e-01
6	taux_de_reussite_	0.0317376195670296	4.67223919898347 1620e-02	7.95451749486546 6498e-02	-1.10954649666692 3981e-01	1.00071225071225 1006e+00	-1.16086065533685 1954e-01
7	nb_candidats_g	0.134146716156051	-1.05582916323133 3107e-02	-2.88788978898121 7731e-01	2.33754709568669 7551e-01	-1.16086065533685 1954e-01	1.00071225071225 1006e+00

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 du nombre quartier prioritaire dans l'académie de l'établissement en fonction d'autres informations sur ces établissements.

(b) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible le nombre quartier prioritaire dans l'académie de l'établissement, qui se trouve dans la colonne 6 de **data_debase_np**. La colonne 0 de matriceCov donne les coefficients de corrélation du nombre quartier prioritaire dans l'académie de l'établissement avec chacune des autres variables/colonnes de **data_debase_np**

Notre objectif est de trouver des variables qui expliquent le mieux possible le nombre quartier prioritaire dans l'académie de l'établissement qui se trouve dans la colonne de **data_debase_np.**

Les coefficients de corrélation correspondant sont :

-0.0689276598680313	0.292059877003812	0.0765471497418443	0.0317376195670296	0.134146716156051

Les colonnes de **data_debase_np** correspondent aux :

- Nombre de ”a” dans le nom de la commune
- Latitude
- Longitude
- Taux de réussite général
- Nombre de candidats général

(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 le nombre quartier prioritaire dans l’académie de l’établissement.

(d) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

data_debase_np.

Le résultat de la fonction :

linear_regression = LinearRegression()

linear_regression.fit(X, Y)

a = linear_regression.coef_

(e) Paramètres, interprétation

On obtient les paramètres a0 = -0.016198, a1 = 0.37019038, a2 = 0.10093067, a3 = 0.03996232, a4 = 0.22177069.

Le signe du paramètre a0 nous permet de voir que le nombre de "a" dans le nom de la commune influence négativement la variable Y.

Le signe du paramètre a1 nous permet de voir que la latitude influence positivement la variable Y.

Le signe du paramètre a2 nous permet de voir que la longitude influence positivement la variable Y.

Le signe du paramètre a3 nous permet de voir que le taux de réussite influence positivement la variable Y.

Le signe du paramètre a4 nous permet de voir que le nombre de candidat influence positivement la variable Y.

Comme les variables endogène et explicatives sont centrées réduites, on peut voir que l'influence des diverses informations ne sont pas très grandes.

(f) Coefficient de correlation multiple, interpretation

```
def Ypred(coeffs, x0, x1, x2, x3, x4):
     return coeffs[0]*x0 + coeffs[1]*x1 + coeffs[2]*x2 +coeffs[2]*x2
+coeffs[3]*x3 +coeffs[4]*x4

def Var(X):
    N=X.shape[0]
    Xm=sum(X)/N
    return sum((X-Xm)*(X-Xm))/N

linear_regression = LinearRegression()
linear_regression.fit(X, Y)
coeffs = linear_regression.coef_

def CoeffMult(coeffs, X, Y):
    res = 0
    N = len(Y)
    for i in range(N):
        res += (Ypred(coeffs,X[i,0],X[i,1],X[i,2],X[i,3],X[i,4]) - Y[i])**2
    res = res /  (Var(Y)*N)
    res = 1 - res
    return sqrt(res)
```

Le résultat de ces fonctions est le coefficient de corrélation multiple :
0.37106063905236236

Puisque le coefficient est en dessous de $\sqrt{(3)} / 2 = 0.87$, c'est une corrélation faible.

5 – Conclusions

(a) Réponse à la problématique

Les variables étudiées n'ont aucune corrélation avec le nombre total de quartiers prioritaires d'une académie selon ses établissements.

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

De toutes les séries statistiques étudiées, celle des latitudes semble avoir le plus d'influence, ce qui est surprenant.

(c) Interprétations personnelles

On comprend donc qu'il n'est pas fiable d'estimer le nombre total de quartiers prioritaires d'une académie selon ses établissements à partir du nombre de "a" dans le nom de la commune, la Longitude, la Latitude, le Taux de réussite de l'établissement et le Nombre de candidats.

Malgré cela les résultats de cette étude ont été intéressants, comme le fait que de toutes les différentes variables statistiques, la latitude influençait le plus.