Ewen JAIN
Enzo CAZ

1A2
11/06/2024

Peut-on prédire le nombre de filles qui passent le brevet par collège ?

Présentation des données et de la problématique

1) Présentation des données

Le fichier *colleges_stats.csv* contient plusieurs séries statistiques sur l'ensemble des collèges répertoriés dans la base de données, certains y sont plusieurs fois car ils ont des données de plusieurs années.
La population y est l'ensemble des données sur les collèges répertoriés chaque année (de 2020 à 2022).

Les variables présentes dans ce fichier sont respectivement les suivantes :

- La 1ère variable désigne le nombres de filles en classe de 3e dans le collège étudié.
- La 2de variable désigne la présence d'un quartier prioritaire à proximité de l'établissement (Hors QP / Dans QP / Dans QP+200m / Dans QP+200-300m).
- La 3e donne le nombre de candidats inscrits au DNB général dans ce collège.
- La 4e variable nous donne la taux de réussite au diplôme du DNB.
- La 5e variable nous donne le nombre de mentions obtenues au DNB.
- La 6e variable est le taux d'élèves de 6e qui sont restés jusqu'en 3e.

Ces données ont été choisies car nous pensons qu'elles ont une corrélation plus ou moins forte avec la donnée du nombre de filles ayant passé le DNB en cette même année dans le collège.

Voici un extrait des données contenues dans la base. La variable qp_a_proximité sera traitée avec les valeur Hors QP, Dans QP +200-300m, dans QP +200m et Dans QP considérées respectivement comme les valeurs 0, 1, 2 et 3, afin de pouvoir faire des calculs.

_3eme_filles	qp_a_proximite_o_n	nb_candidats_g	taux_de_reussite_g	nb_mentions_tb_g	taux_d_acces_6eme_3eme
33	Hors QP	59	93.0	7	88.0
33	Hors QP	70	81.0	6	86.0
100	Hors QP	168	71.0	14	85.0
100	Hors QP	187	77.0	16	79.0
59	Dans QP+200m	88	73.0	10	91.0
59	Dans QP+200m	86	95.0	15	92.0
41	Dans QP	111	72.0	17	87.0
41	Dans QP	86	76.0	7	92.0
87	Dans QP+200m	140	70.0	13	87.0
87	Dans QP+200m	135	76.0	17	91.0
38	Dans QP+200m	81	86.0	17	90.0
38	Dans QP+200m	71	92.0	11	89.0
109	Dans QP+200m	160	77.0	34	94.0
109	Dans QP+200m	159	78.0	27	94.0
64	Hors QP	113	80.0	26	96.0
64	Hors QP	110	82.0	20	94.0
72	Dans QP+200-300m	97	93.0	33	91.0
72	Dans QP+200-300m	116	87.0	30	92.0
44	Dans QP+200-300m	68	62.0	13	90.0
44	Dans QP+200-300m	83	77.0	15	91.0
59	Dans QP+200m	118	87.0	23	90.0
59	Dans QP+200m	108	91.0	24	91.0

2) Présentation de la problématique

À partir de ces données, nous allons tenter de répondre à la problématique suivante:

Est-il possible de prédire le nombre de filles qui passent le Diplôme National du Brevet (DNB) par collège en fonction d'autres statistiques concernant le collège ?

Pour étudier cette question, nous allons étudier la corrélation entre le nombre de filles passant le Brevet et les autres variables choisies, ce qui nous permettra potentiellement de deviner plus ou moins précisément ce nombre en fonction de toutes les autres variables.

Import des données, mises en forme, centrage-réduction

On importe les bibliothèques utiles

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
```

A) Importation des données

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

```
Colleges2DF = pd.read_csv("colleges2_stats.csv", delimiter=';')
```

B) Mise en forme

On a besoin de supprimer les cases vides (qui contiennent nan en Python), puis on transforme notre DataFrame en Array.
De plus, on remplace les variables textuelles par des chiffres pour la colonne quartier prioritaire :

```
Colleges2DF = Colleges2DF.dropna()
Colleges2AR = Colleges2DF.to_numpy()

for ligne in Colleges2AR:

    if ligne[1] == "Dans QP          ":
        ligne[1] = 3

    elif ligne[1] == "Dans QP+200m     ":
        ligne[1] = 2

    elif ligne[1] == "Dans QP+200-300m":
        ligne[1] = 1

    #La valeur des QP non renseignés sera de 0 (comme pour Hors QP)
    else:
        ligne[1] = 0
```

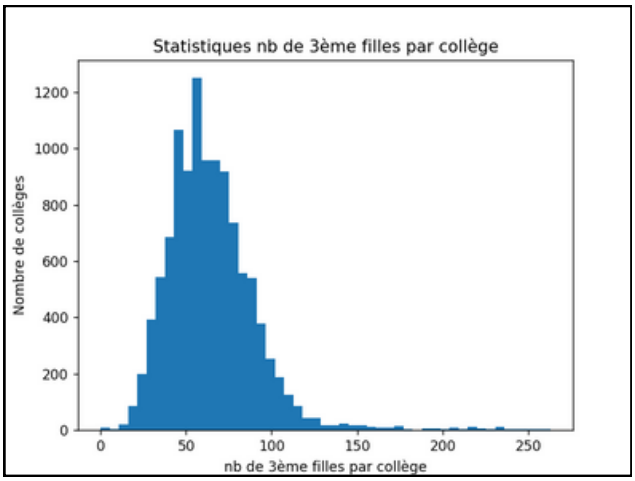
C) Centrer-réduire

On centre-réduit

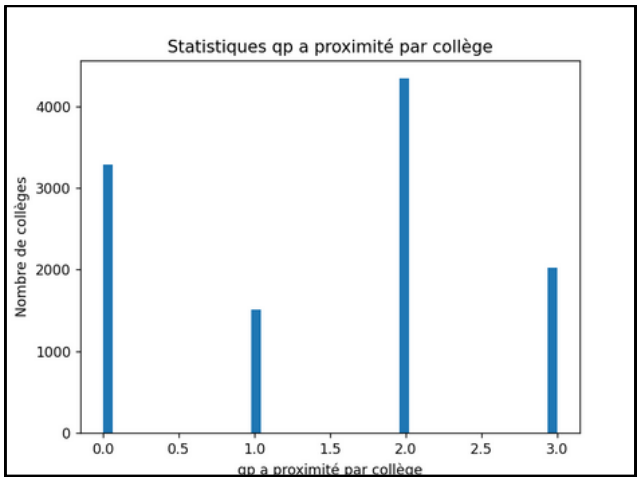
```
def centrerReduire(arr):
    arr=np.array(arr,dtype=np.float64)

    nv_arr = np.zeros((len(arr),len(arr[1])))
    for i in range(len(arr)):
        for j in range(len(arr[0])):
            nv_arr[i,j] = (arr[i,j] - np.average(arr[:,j])) / np.std(arr[:,j])
    return nv_arr

Colleges2AR_CR = centrerReduire(Colleges2AR)
```

Exploration des données : représentations graphiques

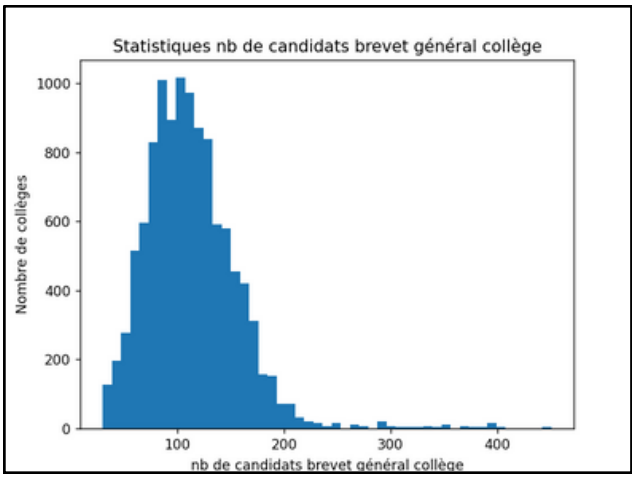
On choisit d'étudier les diagrammes en batons des nos variables statistiques :



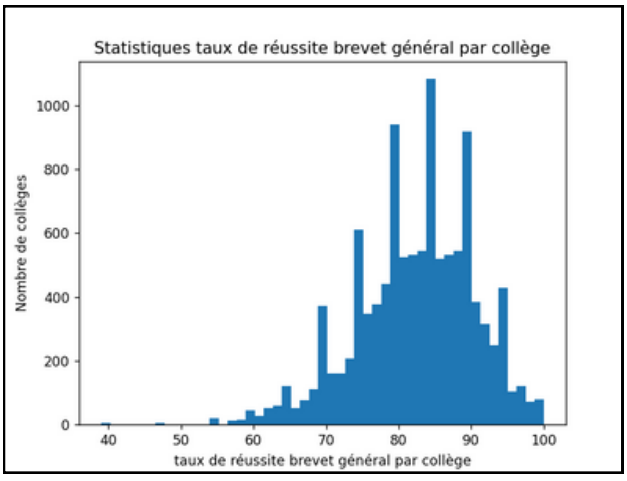
La majorité des collèges ont entre 30 et 90 filles en classe de 3e. Il y a cependant des exceptions qui en ont de très grand nombres



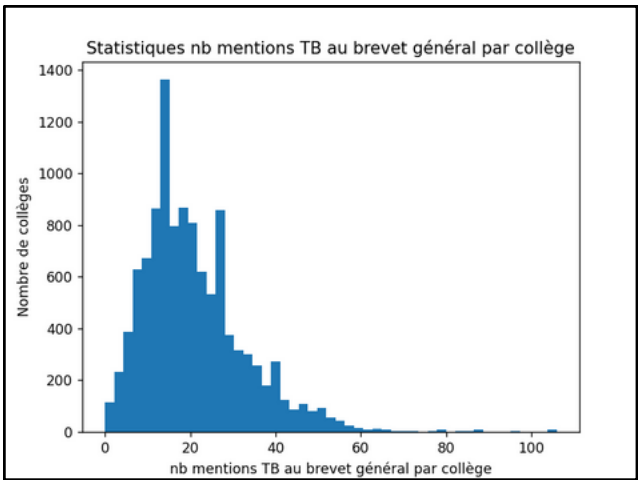
Les collèges ont plus souvent comme valeur 'QP +200m' ou 'Hors QP', mais a répartition n'est pas totalement hétérogène.



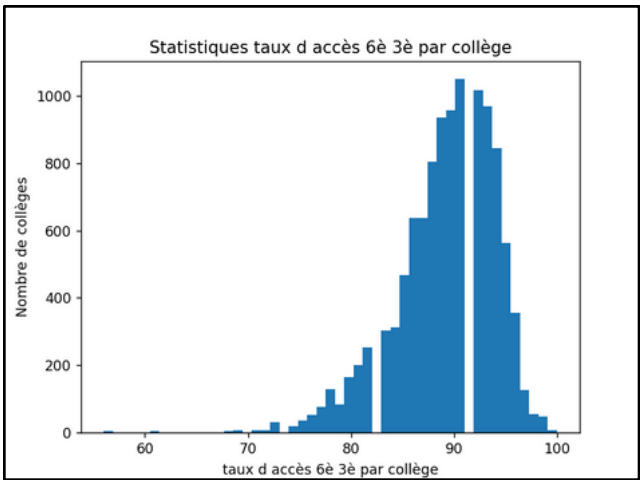
Ce graphique est logiquement bien corrélé au graphique au dessus, avec environ les mêmes statistiques (avec des nombres doublés car c'est filles + garçons)



Ce diagramme est plutôt étonnant quand on voit les pics de fréquences pour certaines valeurs, mais si on les oublie, c'est cohérent avec des valeurs entre 70 et 95% de taux de réussite pour la plupart des collèges



Le nombre de mentions très bien par collège est assez bien réparti, avec un pic aux alentours des 15. il y a des exceptions très éloignées (vers les 110).



Le taux d'accès de la 6e a la 3e est bien réparti avec une presque courbe. Un trou est présent au milieu, peut-être du a un bug car il est très étonnant placé ici.

Exploration des données : matrice de covariance

A) Démarche

Dans cette partie, on calcule la matrice de covariance afin de repérer les variables les plus corrélées entres elles. On s'intéresse surtout a notre variable du nombre de filles, donc la 1ere ligne / colonne. Les données les plus corrélées auront une valeur se rapprochant de 1

```
MatriceCov = np.cov(Colleges2AR_CR,rowvar=False)
print(MatriceCov)
```

B) Matrice de covariance

On obtient la matrice suivante :

1.00008957	-0.00930244	0.88220813	-0.09351943	0.38423570	0.17620029
-0.00930244	1.00008957	-0.01173451	-0.03002337	-0.00047595	-0.13994855
0.88220813	-0.01173451	1.00008957	-0.10680438	0.49019837	0.21587764
-0.09351943	-0.03002337	-0.10680438	1.00008957	0.35204184	0.13979818
0.38423570	-0.00047595	0.49019837	0.35204184	1.00008957	0.19539796
0.17620029	-0.13994855	0.21587764	0.13979818	0.19539796	1.00008957

Régression linéaire multiple

A) Utilisation de la Régression linéaire multiple

En choisissant la 1e variable statistique comme variable endogène et les autres variables comme variables explicatives, la régression linéaire multiple nous permettrait d'obtenir une estimation de la du nombre de filles en 3e dans chaque collèges en fonction des autres informations sur ces collèges.

B) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible le nombre de filles qui passent le Diplôme National du Brevet par collège, qui se trouve dans la 1ère colonne de nos listes.
La 1ère colonne de la matrice de covariance donne les coefficients de corrélation entre le nombre de filles et chacune des autres variables.
Les variables ayant une valeur plus grande dans ce tableau sont les plus fortement correlées, c'est donc celles que nous allons choisir.
Les plus grands coefficients de corrélation (valeur absolue) dans la colonne 0 de la matrice de covariance sont :

- Le nombre de candidats(3e colonne), coef 0,8822
- Le nombre de mentions très bien(5e colonne), coef 0,3842
- Le taux d'accès 6e 3e(6e colonne), coef 0.1762
- Le taux de réussite au brevet général(4e colonne), coef 0.0935

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 le nombre de filles qui passent le Diplôme National du Brevet par collège.

D) Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

```
Y = Colleges2AR[:,0]
X = Colleges2AR[:,1:]

Colleges2LR = LinearRegression()
Colleges2LR.fit(X, Y)

coefficients = Colleges2LR.coef_
resultat = Colleges2LR.intercept_

print("Coefficients:", coefficients)
print("Intercept:", resultat)
```

E) Paramètres, interprétation

On obtient les paramètres:

a0 = 0.01853329
a1 = 0.55131191
a2 = 0.11697106
a3 = -0.17458668
a4 = -0.07298654

Le signe de chaque paramètre nous permet de voir s'il influence de manière positive ou négative le résultat. a3 et a4 ont une influence négative, elle est positive pour les autres.

F) coefficient de corrélation multiple

Le coefficient de corrélation multiple obtenu avec les données précédentes est de 0.782272.
On se rend donc compte que l'association des différentes variable semble donner une estimation vague, ce n'est pas très précis (précis a 78.23% donc) mais c'est quand même assez bien corrélé.

Conclusions

A) Réponse à la problématique

La problématique étant de savoir si il était possible de prédire le nombre de filles qui passent le Brevet par collège en fonction d'autres statistiques concernant le collège.
Je pense qu'on peut répondre que oui, c'est possible. Mais je pense aussi qu'il existe d'autres variables corrélées à cette statistiques, et qu'il serait possible d'avoir, avec plus de données disponibles, un c coefficient de corrélation multiple fort.

B) Argumentation à partir des résultats de la régression linéaire

On peut en déduire d'après les résultats de la régression linéaire que certaines variables sont (plus ou moins étonnamment) corrélées au nombre de filles.
Le nombre de candidats y est évidemment fortement corrélé, car en général, plus il y a de candidats, plus y il a de filles.
Le nombre de mentions très bien a un coefficient de corrélation de 0,3842, montrant donc que le nombre de filles influence le nombre de mentions (est-ce juste car il y en a plus ou tout simplement car elles sont meilleures ? on ne le saura pas…).
le taux d'accès de la 6e a la 3e est également lié, même si faible, signifiant peut-être que les filles décrochent un peu moins souvent au cours du collège.
Enfin, le taux de réussite au brevet est très très très faiblement lié, mais il y a peut-être un très faible signe que les filles ont un tau de réussite différent que les garçons.

C) Interprétations personnelles

On peut peut-être, avec cette étude, émettre l'hypothèse que les filles et les garçons ont des différences au brevet, que ce soit dans le taux de réussite ou dans les mentions très bien.
L'étude n'étant pas focalisé dessus, c'est à prendre avec des pincettes.
Ce dont on est sûr, c'est que plus il y a d'élèves dans un collège, alors plus il y a de filles.