SAÉ 2.04 -
Exploitation d'une base de données

Livrable n°3

Rapport

du travail effectué

Les données stats_mentions_g.csv - Problématique

a) Présentation des données

Le fichier stats_mentions_g.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.
- La variable statistique ou endogène sur cette population est le nombre de mentions des généraux au brevet en 2023 pour chaque collège.
- La 2e est le pourcentage de mention par candidat.
- La 3e est le taux de réussite du brevet dans le collège.
- La 4e est la note moyenne à l'écrit pour les généraux dans le collège.
- La 5e est le pourcentage de mention Bien pour les généraux dans le collège.
- La 6e est le secteur du collège (public ou privé/0 ou 1).
- La dernière est le code du département du collège.

b) Problématique

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

Parmi les données de notre fichier, certaines peuvent-elles permettre d'expliquer ce qui favorise la mention au brevet dans les différents collèges ?

Import des données, mise en forme

a) Importer les données en Python

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

```
CollegesDF=pd.read_csv("stats_mentions_g.csv")
```

b) Mise en forme

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

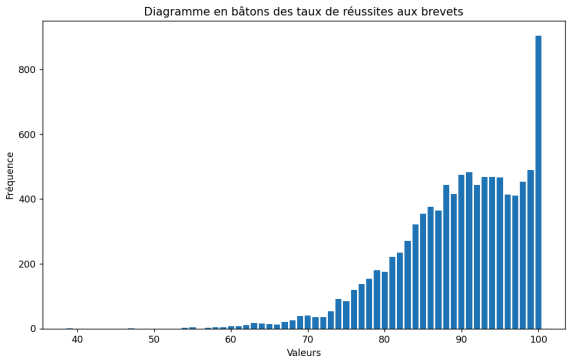
```
CollegesDF = CollegesDF.dropna()
CollegesAr = CollegesDF.to_numpy()
```

c) Centrer-réduire

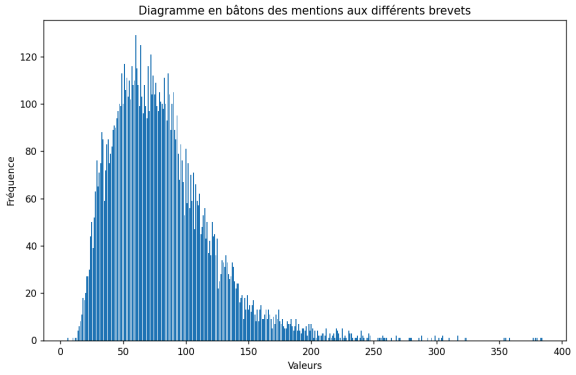
```
def CentreReduire(T):
    T = np.array(T,dtype = np.float64)
    (n,p) = T.shape
    res = np.zeros((n,p))
    TMoy = np.mean(T,axis = 0)
    TEcart = np.std(T,axis = 0)
    for j in range(p):
        res[:,j] = (T[:,j] - TMoy[j]) / TEcart[j]
    return res

CollegesAR0=CollegesAr
CollegesAR0_CR=CentreReduire(CollegesAR0)
```

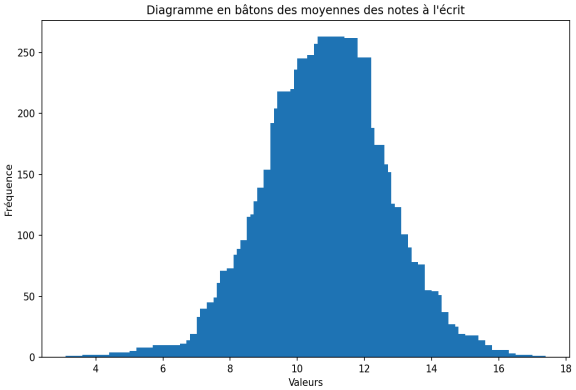
a. Exploration des données : représentations graphiques



On remarque que le taux de réussite est très souvent proche de 100% et n'est très majoritairement jamais en dessous de 60%.



On remarque que le graphique montre la répartition des mentions au brevet parmi les candidats. La moyenne remarqué est d' environ 100 mentions par établissement.



On remarque que les notes à l'écrit sont très majoritairement entre 6 et 16 sur 20 et que la moyenne des élèves se situe entre 10 et 12,5 sur 20.

b. Exploration des données : matrice de covariance

a) Démarche

Dans cette partie, on calcule la matrice de covariance afin de *à compléter*

```
print(np.cov(CollegesAR0_CR,rowvar=False)
```

b) Matrice de Covariance

```
On obtient la matrice suivante :
[[ 1.00010232  0.88618016  0.484438    0.36541023  0.47006612  0.87149836
  -0.09631524  0.22858563 -0.04320043]
 [ 0.88618016  1.00010232  0.07959637  0.01334321  0.10839815  0.77990631
  -0.09313951 -0.00330936  0.09923147]
 [ 0.484438    0.07959637  1.00010232  0.84615425  0.85204479  0.42803996
  -0.01465356  0.52466866 -0.22836792]
 [ 0.36541023  0.01334321  0.84615425  1.00010232  0.71386715  0.35029934
   0.05002517  0.49226044 -0.08924322]
 [ 0.47006612  0.10839815  0.85204479  0.71386715  1.00010232  0.38828334
  -0.04634459  0.47931905 -0.29962937]
 [ 0.87149836  0.77990631  0.42803996  0.35029934  0.38828334  1.00010232
   0.33206277  0.281395   -0.03931654]
 [-0.09631524 -0.09313951 -0.01465356  0.05002517 -0.04634459  0.33206277
   1.00010232  0.16745684 -0.00303239]
 [ 0.22858563 -0.00330936  0.52466866  0.49226044  0.47931905  0.281395
   0.16745684  1.00010232 -0.04277137]
 [-0.04320043  0.09923147 -0.22836792 -0.08924322 -0.29962937 -0.03931654
  -0.00303239 -0.04277137  1.00010232]]
```

Régression linéaire multiple

a) Utilisation de la Régression linéaire multiple : comment ?

Nous allons utiliser la régression linéaire multiple pour comprendre quelles variables expliquent le mieux le nombre de mentions obtenues. Pour cela, nous utiliserons les variables explicatives disponibles (pourcentage de mentions par candidat, taux de réussite, etc.) et la variable endogène (nombre de mentions).

b) Variables explicatives les plus pertinentes

Pour déterminer les variables explicatives les plus pertinentes, nous examinerons les coefficients de la régression linéaire. Les coefficients nous montrent l'impact de chaque variable explicative sur la variable endogène. Les variables avec des coefficients plus élevés en valeur absolue ont un impact plus significatif.

c) Lien avec la problématique

Les résultats de la régression linéaire multiple montrent quelles variables parmi celles disponibles dans le fichier influencent le plus le nombre de mentions au brevet. Par exemple, des variables comme le pourcentage de mentions par candidat et le taux de réussite peuvent avoir un impact significatif. Cela nous permet de cibler les facteurs spécifiques qui favorisent la mention au brevet dans les différents collèges, répondant ainsi à la problématique initiale.

d) Régression Linéaire Multiple en Python

On fait maintenant la r´egression lin´eaire multiple avec Python :

```
ar_endogene =  CollegesAr[:, [0]]
ar_explicative = CollegesAR0_CR[:, [2,3,4,6,7,8]]
linear_regression = lr()
linear_regression.fit(ar_explicative,ar_endogene)
a = linear_regression.coef_
```

e) Paramètres, interprétation

```
On obtient les paramétres :
a0 =  [19.07601856]
a1 =  [-7.16224389]
a2 =  [10.10181495]
a3 =  [-2.60708301]
a4 =  [-1.30955073]
a5 =  [4.90896391]
```

Le signe du paramètre a0 (positif) nous permet de voir qu'il y a une influence positive sur la variable endogène lorsqu'on ne prend en compte aucune des variables explicatives. Cela signifie que, de manière intrinsèque, le nombre de mentions a une tendance de base à être élevé dans les collèges.

Comme les variables endogène et explicatives sont centrées-réduites, nous pouvons de plus voir l'importance relative de chaque variable explicative sur la variable endogène. Voici ce que chaque paramètre nous indique :

- $a1 \approx -7$: Le pourcentage de mentions par candidat a un effet négatif significatif sur le nombre de mentions. Une augmentation d'un écart-type dans cette variable réduit le nombre de mentions d'environ 7 unités.
- $a2 \approx 10$: Le taux de réussite a un effet positif et assez fort. Une augmentation d'un écart-type dans le taux de réussite augmente le nombre de mentions d'environ 10 unités.
- $a3 \approx -22$: La note moyenne à l'écrit a un effet négatif modéré. Une augmentation d'un écart-type dans la note moyenne réduit le nombre de mentions d'environ 2 unités.

- $a4 \approx -1$: Le pourcentage de mentions Bien a un effet négatif léger. Une augmentation d'un écart-type dans cette variable réduit le nombre de mentions d'environ 1 unité.
- $a5 \approx 4$: Le secteur du collège (public ou privé) a un effet positif. Être dans un secteur privé (codé par 1) augmente le nombre de mentions d'environ 4 unités par écart-type.

Ces résultats montrent que, parmi les variables explicatives, le taux de réussite est le facteur le plus important pour expliquer le nombre de mentions obtenues au brevet. Cela suggère que les collèges avec un taux de réussite élevé ont tendance à obtenir plus de mentions. D'autres facteurs, comme le pourcentage de mentions par candidat et le secteur du collège, jouent également un rôle significatif.

En résumé, la régression linéaire multiple nous aide à identifier les facteurs clés qui favorisent l'obtention de mentions au brevet dans les différents collèges, en prenant en compte l'importance relative de chaque variable explicative après centrage-réduction.

f) Coefficient de corrélation multiple, interprétation

Le coefficient de corrélation multiple (R^2) mesure la proportion de la variance de la variable endogène expliquée par l'ensemble des variables explicatives. Il donne une idée de la qualité de la régression linéaire multiple.

```
r2 = linear_regression.score(ar_explicative,ar_endogene)
```

```
coefficient de corrélation multiple :
  0.2710981662442383
```

Un R^2 proche de 1 indique que le modèle explique bien la variance des données, tandis qu'un R^2 proche de 0 indique le contraire.

Conclusions

a) Réponse à la problématique

Les données fournies permettent effectivement d'identifier les facteurs qui favorisent l'obtention de mentions au brevet dans les différents collèges. Le modèle de régression linéaire multiple a montré que le taux de réussite, le pourcentage de mentions par candidat, et le secteur du collège (public ou privé) sont des variables significatives influençant le nombre de mentions obtenues.

b) Argumentation à partir des résultats de la régression linéaire

Les résultats de la régression linéaire multiple montrent que :

- Le taux de réussite ($a_2 \approx 10$) a un effet positif fort, indiquant que les collèges avec un taux de réussite élevé obtiennent plus de mentions.
- Le pourcentage de mentions par candidat ($a_1 \approx -7$) a un effet négatif significatif, ce qui pourrait suggérer que dans certains collèges, bien que beaucoup de candidats obtiennent des mentions, le nombre total de mentions reste influencé négativement par d'autres facteurs.
- Le secteur du collège ($a_5 \approx 4$) a également un effet positif, indiquant que les collèges privés ont tendance à avoir un plus grand nombre de mentions.

Ces coefficients nous montrent que les variables explicatives choisies sont pertinentes et permettent de modéliser le nombre de mentions au brevet avec un certain degré de précision, comme indiqué par le coefficient de corrélation multiple (R^2).

a) Interprétations personnelles

Cette analyse bien qu'intéressante nous laisse mitigés quant à la corrélation entre nos variables explicatives avec la variable endogène.

Donc, bien que cette analyse fournisse des indications utiles sur les facteurs influençant les mentions au brevet, elle souligne également la complexité et les limitations à l'utilisation des données quantitatives seules.