14/06/2024

SAÉ 2.04 - EXPLOITATION
D'UNE BASE DE DONNÉES

Rapport Statistiques

FAIT PAR :

GUYOMARD Juline
DEROUESNE RIwan

Sommaire :

I - Les données Collèges.csv - Problématique :

1- Présentation des données

Le fichier Collèges.csv contient plusieurs séries statistiques sur l'ensemble de toutes les collèges répertoriés dans notre base de données :

- La population est l'ensemble des effectifs de 3èmes de chaque collège, représentés de manière unique par leur nom d'établissement et leurs codes (uai).
- ˆ La 1ère variable statistique correspond à la latitude exacte de chaque établissement.
- La 2ème correspond à la longitude exacte de chaque établissement.
- La 3ème correspond au nombre de lettres de chaque établissement.
- La 4ème correspond au taux de réussite générale de chaque établissement.
- La dernière correspond à l'effectif des 3ème de chaque établissement chaque année, c'est la variable endogène.

uai	nom_etablissement	nombre de lettre	_3eme_total	longitude	latitude	taux_de_reussite_g
9720495,00 F	Collège Trianon	15	73	-60.911570334238235	14.6205289328091	96.0
0691664J	Collège Jean Jaurès	19	188	4.883449409156436	45.75759880249377	75.0
0693093M	Collège du Tonkin	17	124	4.863710113971357	45.77637926436517	91.0
0694296V	Collège Simone Lagrange	23	89	4.903774500573558	45.77926108782248	84.0
251395,00 F	Collège Lou Blazer	18	156	6.791109347380717	47.49841073923035	93.0
0251397H	Collège les Hautes Vignes	25	86	6.85640985034442	47.47165765562169	83.0
0250056A	Collège Jouffroy d'Abbans	25	212	6.8414497402084065	47.512047196475024	84.0
0251599C	Collège les Bruyères	20	121	6.829900270160826	47.45420204696176	88.0
0390907Z	Collège Saint-Exupéry	21	157	5.565940517303166	46.67961909845252	84.0
0900017E	Collège de Châteaudun	21	80	6.852888130471386	47.645771115844525	90.0
0211357L	Collège Jean-Philippe Rameau	28	113	4.994749445961136	47.322807626205346	85.0

2- Problématique

En utilisant ces données, on va essayer de répondre à la problématique suivante : Comment le taux de réussite général, la position exacte de l'établissement (longitude et latitude) et le nombre de lettres dans le nom de l'établissement influencent-elles sur l'effectif des 3èmes?

En choisissant la 4eme série statistique comme variable endogène et les 4 autres séries comme variables explicatives, la régression linéaire multiple nous permettra d'obtenir une estimation des effectifs des 3ème en fonction des 4 autres séries statistiques.
Les paramètres de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus les effectifs des 3èmes.
Le coefficient de corrélation multiple nous permettra d'apporter une réponse à la problématique

II - Import des données, mise en forme

1- Importer les données en Python

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

Voici le code :

```
# Chargement des données avec le bon séparateur
CollegesDF = pd.read_csv("./colleges.csv", delimiter=',')
```

Voici le résultat :

```
           uai        nom_etablissement  ...   latitude  taux_de_reussite_g
0      9720495F          Collège Trianon  ...  14.620529                96.0
1      0691664J      Collège Jean Jaurès  ...  45.757599                75.0
2      0693093M        Collège du Tonkin  ...  45.776379                91.0
3      0694296V  Collège Simone Lagrange  ...  45.779261                84.0
4      0251395F       Collège Lou Blazer  ...  47.498411                93.0
...         ...                      ...  ...        ...                 ...
10989       NaN                      NaN  ...        NaN                84.0
10990       NaN                      NaN  ...        NaN                95.0
10991       NaN                      NaN  ...        NaN                71.0
10992       NaN                      NaN  ...        NaN                75.0
10993       NaN                      NaN  ...        NaN                84.0
```

2- Mise en forme

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

Voici le code :

```
17  # Suppression des lignes avec des valeurs manquantes
18  CollegesDF = CollegesDF.dropna()
19
20  # Conversion du DataFrame en array numpy
21  CollegesAr = CollegesDF.to_numpy()
```

Voici le résultat :

```
[10994 rows x 7 columns]
[['9720495F' 'Collège Trianon' 15.0 ... -60.91157033423824
  14.6205289328091 96.0]
 ['0691664J' 'Collège Jean Jaurès' 19.0 ... 4.883449409156436
  45.75759880249377 75.0]
 ['0693093M' 'Collège du Tonkin' 17.0 ... 4.863710113971357
  45.77637926436517 91.0]
 ...
 ['0710534V' 'Collège Robert Schuman' 22.0 ... 4.835703569402592
  46.32001804781104 78.0]
 ['0220054P' 'Collège Jean Racine' 19.0 ... -2.7510851598346853
  48.50911833490083 64.0]
 ['0350963G' 'Collège les Hautes Ourmes' 25.0 ... -1.649410128522291
  48.08946765302947 69.0]]
```

3- Centrer-réduire

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

Première étapes on supprimer les colonnes ne contenant pas des données numériques, ici on garde les colonnes 2 à 7 qui correspondent à l'effectif des 3ème, la longitude et latitude :

Voici le code :

```
#On ne garde que les valeurs numériques
CollegesAr0 = CollegesAr[:, 2:6]
print(CollegesAr0)
```

Voici le résultat :

```
[[15.0 73.0 -60.91157033423824 14.6205289328091]
 [19.0 188.0 4.883449409156436 45.75759880249377]
 [17.0 124.0 4.863710113971357 45.77637926436517]
 ...
 [22.0 129.0 4.835703569402592 46.32001804781104]
 [19.0 150.0 -2.7510851598346853 48.50911833490083]
 [25.0 115.0 -1.649410128522291 48.08946765302947]]
```

Maintenant nous pouvons centrer-réduire les données.

Voici le code :

```
# Fonction de centrage-réduction
def Centreduire(T):
    T = np.array(T,dtype=np.float64)
    lignes, colonnes = T.shape
    res = np.zeros((lignes, colonnes))
    moy = np.mean(T, axis=0)
    ecarttype = np.std(T, axis=0)
    for i in range(0,lignes):
        for j in range(0, colonnes):
            res[i][j] = (T[i][j]-moy[j])/ecarttype[j]
    return res


CollegesAr0_CR=Centreduire(CollegesAr0)
print(CollegesAr0_CR)
```

Voici le résultat :

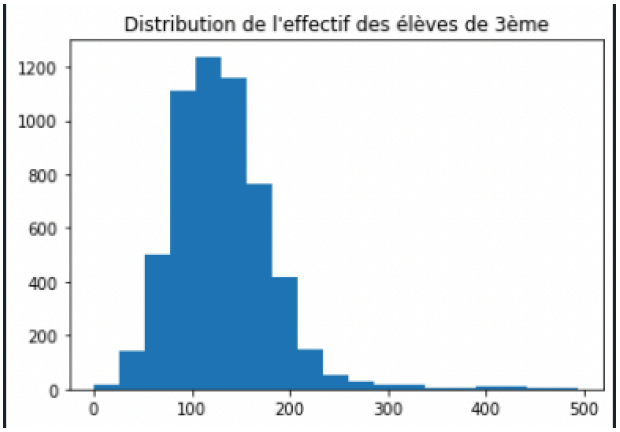
```
[[-1.42809433 -1.10091723 -3.5346792  -1.714571  ]
 [-0.54561458  1.09155011  0.14999992  0.20699118]
 [-0.98685445 -0.12860563  0.14889447  0.20815018]
 ...
 [ 0.11624523 -0.03328096  0.14732604  0.24169976]
 [-0.54561458  0.36708264 -0.27755235  0.37679572]
 [ 0.77810504 -0.30019003 -0.21585591  0.35089781]]
```

III - Exploration des données :

1- Représentation graphique

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

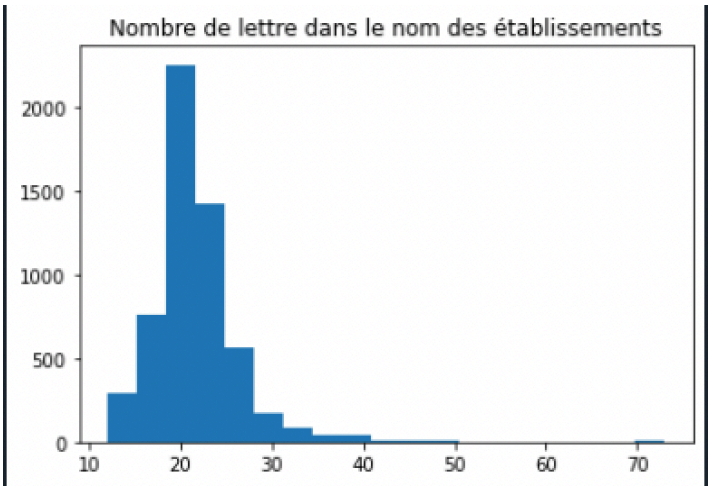
Diagramme bâton de notre variable endogène, l'effectif des 3ème :



On remarque que la plupart des effectifs des 3ème sont autour de 125 élèves. Certains collèges vont de presque 0 à quasiment 500 élèves.

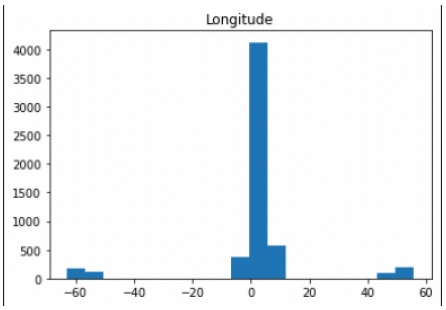
Diagramme bâton de nos variables explicatives :

- Le nombre de lettre dans le nom de l'établissement :



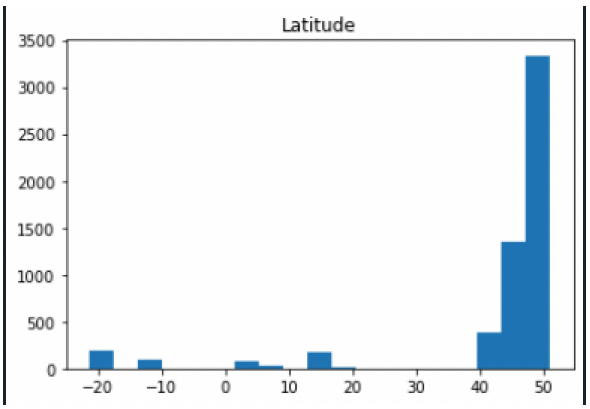
On remarque que la plupart des établissements sont autour de 20 lettres dans leurs noms, allant de 12 à 50 à peu près. Quelques exceptions s'ajoutent car certains établissements ont environ 70 lettres.

- La longitude de chaque établissement :



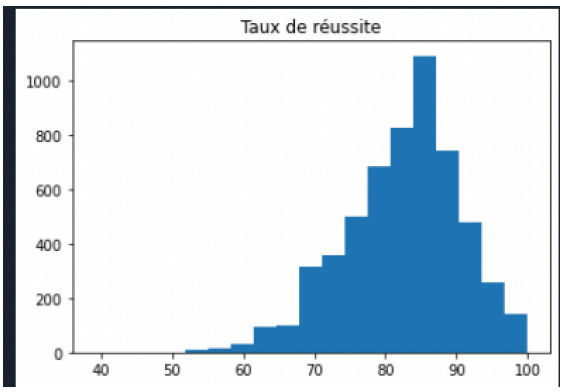
On peut remarquer que la longitude est principalement au alentours de 0-10 ou alors au extrème vers -60 ou 50.

- La latitude de chaque établissement :



On remarque que la latitude à une hausse très forte de 40 à 50, on passe de 400 environ à près de 3500. On a quelques établissements qui sont plus éparpillés : -20, -15/-10, de 2 environ à 9 environ et enfin de 13 environ à 20 a peu près.

- Le taux de réussite générale de chaque établissement :



On remarque que le taux de réussite est assez regroupé, et fortement au alentours de 85 environ. Quelques établissements touchent l'excellence à 100 tandis que d'autres sont à peine au-dessus de la moyenne, environ 52. On peut voir un gros écart entre les établissements.

2- Matrice de covariance

a. Démarche

Dans cette partie, on calcule la matrice de covariance pour nous permettre d'obtenir en un seul calcul les corrélations 2 à 2 de toutes nos variables. Si la valeur est positive alors les variables évoluent dans le même sens (augmentant toutes les deux ou diminuant toutes les deux) ou on un sens différent (l'une augmente et l'autre diminue et inversement). On pourra voir si les variables sont corrélées les unes aux autres.

Voici le code :

```
MatriceCov = np.cov(CollegesAr0_CR, rowvar=False)
print("Matrice de covariance :\n", MatriceCov)
```

b.Matrice de covariance

On obtient la matrice suivante :

Voici le résultat :

```
Matrice de covariance :
 [[ 1.00017702  0.02665865 -0.09080936 -0.15796307  0.02257205]
 [ 0.02665865  1.00017702  0.24801969 -0.36581415  0.02477904]
 [-0.09080936  0.24801969  1.00017702 -0.17201094  0.00861709]
 [-0.15796307 -0.36581415 -0.17201094  1.00017702 -0.00972582]
 [ 0.02257205  0.02477904  0.00861709 -0.00972582  1.00017702]]
```

	0	1	2	3	4
0	1.00018	0.0266586	−0.0908094	−0.157963	0.022572
1	0.0266586	1.00018	0.24802	−0.365814	0.024779
2	−0.0908094	0.24802	1.00018	−0.172011	0.00861709
3	−0.157963	−0.365814	−0.172011	1.00018	−0.00972582
4	0.022572	0.024779	0.00861709	−0.00972582	1.00018

Pour que les variables soient corrélées fortement il faut qu'elle soit supérieur à 0,88. Ici, on peut voir que les variables hors diagonales (corrélation de la mêmes variables) ne sont pas corrélées car elles sont inférieures à 0,88. La corrélation la plus forte est -0,365814 et la plus faible est -0,00972582.

IV - Régression linéaire multiple

1- Utilisation de la régression linéaire multiple : comment ?

La régression linéaire multiple nous permettra de savoir les variables explicatives influençant le plus notre variables endogène, l'effectif des 3ème de chaque établissement. Elle permet également d'obtenir une estimation de la moyenne des effectifs des 3eme dans les collèges en fonction d'autres variables sur ces collèges.

2- Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible l'effectif des 3 ème de chaque collège, qui se trouve dans la colonne 1 de CollegesAr0.

La colonne 1 de notre matrice de covariance (matriceCov) donne les coefficients de corrélation de l'effectif des 3eme avec chacune des autres variables/colonnes de CollegesAr0. Les coefficients de corrélation le plus grand en valeur absolue avec l'effectif des 3eme (colonne 1) sont dans l'ordre croissant : 0,365814, 0,24802, 0,24779, et 0,0266586 correspondant aux colonne de matriceCov 3,2,4, et 0.Ces colonnes correspondent respectivement à la latitude, la longitude, le taux de réussite et le nombre de lettre dans le nom de l'établissement.

3- Lien avec la problématique

Les paramètres de la régression linéaire multiple nous informeront des variables explicatives qui influencent le plus l'effectif des 3eme. 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 l'effectif des 3èmes.

4- Régression Linéaire Multiple en Python

On fait maintenant la régression linéaire multiple avec Python :

Voici le code :

```
# Régression linéaire
Y = np.array(CollegesAr0_CR[:, 1])
X = np.array(CollegesAr0_CR[:, 0:5])

RegressionLineaire = LinearRegression()
RegressionLineaire.fit(X, Y)
print(RegressionLineaire.coef_)
```

On a choisi la colonne 1 car c'est la colonne de notre variable endogène.

Voici le résultat :

```
[2.08194303e-16 1.00000000e+00 3.33066907e-16 3.33066907e-16
 1.38777878e-17]
```

5- Paramètres, interprétation

On obtient les paramètres a_0 = 2,082, a_1 = 1,000000, a_2 = 3,331, a_3 = 3,331, a3 = 1,388.

Le signe du paramètre a est positif, ils influencent positivement sur notre variable endogène. On peut voir que a1 est égale à 1 car c'est la variable endogène.

Comme les variables endogène et explicatives sont centrées-réduites, c'est - à - dire que plus a est grand plus l'influence est forte. Ici, les variables influençant le plus fort est la colonne a2 et a3, correspondant à la longitude et latitude, ce qui est intriguant c'est qu'ils influent pareil car leur est la même.

6- Coefficient de corrélation multiple, interprétation …

Voici le code :

```
print(RegressionLineaire.score(X, Y))
print(math.sqrt(RegressionLineaire.score(X, Y)))
```

Voici le résultat :

```
1.0
1.0
```

Selon les résultats, nous obtenons une corrélation parfaite car elle est égale à 1 mais si cela n'est pas possible. Cela voudrait dire que le nombre de lettres de l'établissement, le taux de réussite et la longitude, latitude est une forte influence sur le l'effectif des 3èmes de chaque établissement.

V - Conclusions

1- Réponse à la problématique ...

La problématique est : Comment le taux de réussite général, la position exacte de l'établissement (longitude et latitude) et le nombre de lettres dans le nom de l'établissement influencent-elles sur l'effectif des 3èmes?
D'après nos résultats, l'effectif des troisième est statistiquement fortement influencé car la longitude, la latitude, le taux de réussite ainsi que le nombre de lettres de chaque nom d'établissement.

2- Argumentation à partir des résultats de la régression linéaire

Les résultats de notre régression linéaire indiquent que les variables latitude, longitude, taux de réussite général et nombre de lettres dans le nom de l'établissement ont toutes une influence sur l'effectif des élèves de troisième. Plus précisément, les coefficients de régression montrent que la latitude et la longitude ont une influence significative et similaire sur l'effectif des troisièmes, suggérant que la position géographique de l'établissement joue un rôle important.

3- Interprétations personnelles

Selon nous, il semble peu probable que l'effectif des élèves de troisième puisse être significativement influencé par des variables telles que la latitude, la longitude et le nombre de lettres dans le nom de l'établissement. En effet, ces variables géographiques et nominales ne devraient pas logiquement avoir une relation directe avec le nombre d'élèves dans un niveau scolaire spécifique.
De plus, obtenir un résultat de régression linéaire avec un coefficient égal à 1 est extrêmement improbable en pratique. Un résultat de régression linéaire de 1 suggère une parfaite prédiction des valeurs de la variable dépendante par les variables indépendantes, ce qui est rarement, voire jamais, observé.