RAPPORT D'ACTIVITÉ

ANALYSE DE DONNÉES

Python : parcours débutant

Rapport rédigé par **Zara HUSTON**.
M1 GAED Parcours Géosuds. 2025-2026

SOMMAIRE

SÉANCE 1

I - Mise en route de Python.

1. Installation de Docker.

J'ai tout de suite trouvé ça étrange quand vous avez dit qu'il fallait utiliser Docker pour faire du Python, alors qu'il s'agit d'un logiciel très complexe à utiliser, essentiellement fait pour le codage. De plus, mes amis qui ont suivi une formation en programmation et dont certains sont actuellement déjà en entreprise, ils ont été très surpris quand je leur ai dit que notre professeur voulait qu'on installe Docker pour pouvoir faire du Python, surtout pour, à priori, des raisons de « sécurité », en sachant que les manipulations données à faire ne peuvent pas impacter ou faire planter les ordinateurs, à moins qu'il date de 1990.

Néanmoins, j'ai tout de même cherché à suivre votre méthode, et je m'y suis mise très tôt, très sérieusement. Or, j'ai très vite été confrontée à des erreurs, des ralentissements sur mon appareil (alors qu'il est assez puissant). Je vous avais demandé de l'aide sur Discord dès septembre, et malgré vos suggestions, rien n'avait changé. Il faut savoir que je n'ai pas de difficulté quant à la manipulation des ordinateurs (sauf MacOS), donc j'étais frustrée de ne pas réussir. Au bout d'au moins 6h, j'ai décidé d'arrêter cet acharnement, surtout que je ne voyais pas l'intérêt (*très honnêtement)* de faire ça, j'ai décidé d'utiliser Python traditionnellement, avec VS Code et l'application Python en 3.13.9.

2. Critique de Docker.

Ayant échoué, et ne voulant pas, initialement, me prendre la tête avec Docker, j'ai utilisé VS Code, une application intuitive et dont l'installation **est** (et je maintiens mon propos), plus rapide que Docker.

En effet, absolument toute ma classe, à l'exception de Louka je crois, n'ont pas réussi à installer Docker en suivant votre méthode, du moins, certainement pas en 5 minutes. Très vite, la classe était résignée, dépitée et ça a été le sujet de nombreuses discussions entre nous.

Docker n'a rien de facile et ce n'est certainement pas l'application à faire installer à des étudiants qui n'ont jamais fait de programmation Python.
Vos intentions étaient très bonnes, et vos indications se voulaient être claires, mais très peu ont réussi à faire tourner Docker sur leur ordinateur.
Ce n'est pas une application adaptée pour faire du Python, surtout quand ce sont des étudiants en M1 géographie, venus pour la plupart d'une formation classique en Géographie, ou de CPGE A/L, ou globalement en sciences humaines. (*je reviendrai sur ça dans la partie commentaire du cours).*

Je vous suggère, si ce cours se poursuit l'année prochaine (et en soi, j'y vois un bon intérêt, notamment pour la production graphique des données géographiques récoltées), de ne pas utiliser Docker, mais Python + VS CODE.

L'installation est simple :

- Installer Python en cochant toutes les cases proposées.
- Installer VS CODE et une fois installée, ajouter l'extension « Python » (procédure qui prend *réellement* 30 secondes).
- Vérifier l'installation de python avec la commande : python -- version
- Installer les bibliothèques (*et uniquement celles nécessaires...)* dans l'invite de commande :

    - pip install pandas
    - pip install matplotib
    - pip install numpy
    - et autres **SI** c'est nécessaire

Le plus important selon moi : abandonner l'idée de pédagogie inversée, du moins, pas avant la troisième séance. Prenez le temps, de faire les étapes d'installation sur le PC, en projetant les étapes, ainsi les étudiants feront exactement la même chose que vous, simultanément. Et ça prend 5 minutes, pas plus. Vous le faites lors de la deuxième heure de cours, après la présentation générale.

Il ne faut pas abandonner les étudiants, à la fin du premier cours. Nous avons beaucoup de travail, et Python a causé énormément de stress à la plupart d'entre-nous. Certains, encore fin novembre, n'avaient pas réussi l'installation. Certes, s'y prendre tard, vous n'êtes certainement pas responsable. Néanmoins, il aurait fallu s'assurer que tout le monde ait bien compris et faire une installation commune (et par conséquent équitable) sur tous les PC des étudiants. C'est votre rôle en tant qu'enseignant. J'espère que vous prendrez vraiment en considération mes remarques. Elles ont uniquement pour but d'améliorer le cours pour les prochains étudiants.

3. Critique de GitHub

Dans le cadre de ce cours, l'utilisation de GitHub m'a semblé largement disproportionnée par rapport aux objectifs pédagogiques visés. Les travaux demandés auraient pu être déposés et organisés de manière bien plus simple via Moodle, plateforme déjà utilisée dans les autres enseignements et parfaitement adaptée à un contexte universitaire. À l'inverse, GitHub impose une série d'étapes techniques (création de dépôts, installation de logiciels, gestion des chemins de fichiers, manipulation de dossiers) qui prennent un temps non négligeable et n'apportent pas de bénéfice évident à l'analyse des données. Autrement dit, ça ne sert pas à grand-chose pour nous, étudiants, surtout pour ceux qui ne disposent pas d'une formation informatique approfondie, ces manipulations s'avèrent souvent plus complexes que formatrices (puis que vous nous avez donné aucune indication, et bien souvent, vous haussiez le ton quand certains posaient des questions sur GitHub), au point de devenir un obstacle plutôt qu'un outil. J'ai ainsi eu le sentiment que l'apprentissage des méthodes statistiques passait au second plan derrière la gestion technique de l'environnement de travail, ce qui interroge la pertinence réelle de ce choix méthodologique dans un cours pour des étudiants en M1 Géographie.

4. Mon utilisation de Python.

Alors, vous avez pu remarquer que j'étais assez à l'aise avec Python. Je ne suis absolument pas programmatrice, je précise. J'ai des compétences assez limitées, mais dans le cadre de ce cours, je pouvais gérer, surtout les premières séances. Je n'ai pas voulu prendre

les autres parcours, car je trouvais que le parcours débutant suffisait pour un usage de Python, en tant que géographe. Il n'y a pas besoin d'en faire autant avec la Séance 10 et....à moins d'être en Master Data Analyst, ce qui n'est pas le cas ici, en GAED.

Puis, avec la quantité de cours, et de devoirs en Géosuds (plus que les autres parcours), je ne voulais pas accorder trop de temps là-dessus, je pense que c'est assez évident. Cependant, je n'ai pas négligé les exercices, j'ai raté aucune séance, j'ai réalisé toutes mes manipulations, et comme vous pouvez le voir, je rédige ce rapport avec beaucoup de rigueur et de sincérité.

Comme je disais, je ne suis pas une programmatrice, encore moins une experte en Python. Je suis une « littéraire » comme vous dites souvent. Et pour le coup, j'en suis vraiment une, car j'arrive de prépa A/L et cela fait qu'un an que je suis à Paris. Je suis arrivée dans ce Master après avoir validée, l'année dernière, une double licence Histoire-Géographie, ici-même. J'ai acquis ces compétences car au lycée j'ai fait des spécialités scientifiques et littéraires, et parce que je suis une « geek ». Et comme toute geek qui se respecte, je sais quelques bricoles en informatique.

5. Commentaires sur le premier cours.

Si vous pouvez essayer d'être plus compréhensif. On sait que vous êtes un professionnel. On sait que c'est pas si compliqué. Mais en réalité, si, ça l'est. Votre ressenti n'est pas le nôtre. On a tous fait de notre mieux. Et toutes les personnes que j'ai aidées, surtout pour l'installation, les premières séances, et les codes très difficiles, je tiens à vous assurer qu'elles ont pris sur elles, et ont toutes énormément progressé. J'espère que vous allez considérer cet effort collectif pour les étudiants du Master parcours Géosuds.

SÉANCE 2

I - Questions de cours.

1. Remarque sur le cours.

Il y avait beaucoup de questions pour cette séance, qui semblaient compliquées au début, au vue de la formulation de ces dernières alors qu'il s'agit, en réalité, de questions de statistiques niveau collège. Votre cours est assez hermétique et complique un peu les choses, surtout pour des étudiants en Géographie (et non en Statistiques mathématiques).

C'est donc un cours parfait pour des étudiants en CPGE MPSI/PCSI, mais il n'est adapté pour des M1 Géographie dont 3/4 viennent de prépa littéraires. De plus, vous avez fait le choix d'utiliser des termes anciens, très scientifiques (notamment les formules mathématiques), alors qu'on pouvait faire beaucoup plus simple pour illustrer les formules. Vos cours ont effrayé plus d'un, moi-même, je l'étais.

- Par exemple, au lieu d'utiliser « ensemble des nombres Réels », « X », « t », mettez des chiffres.
- La question 9 sur la densité, mettez simplement : densité pour 1000hab dans une ville de $5km^2$ => $1000/5km^2 = 200km^2$ = Simple, et clair.

2. Réponses aux questions de cours.

1. La géographie souhaiterait mettre de la distance par rapport aux statistiques malgré le fait qu'elle soit une discipline au croisement des sciences dures et humaines. Elle réalise des données qui peuvent faire l'objet de statistiques.

2. Le hasard existe en géographie. La plupart des géographes sont unanimes sur la question. Le hasard fait partie des statistiques, par conséquent, il existe également en géographie, et il est même essentiel. Il est impossible de prévoir, notamment dans le cas de la géographie humaine. Il faut donc, même en géographie, considérer les lois du hasard (c'est-à-dire, la contingence des faits et non la nécessité en opposition au déterminisme philosophique).

3. Il y a les entrées territoriales claires et précises (géographie humaine et physique) et la morphologie même des espaces étudiés (données géométriques).

4. Pour l'analyse de données, les géographes ont besoin de données (ils ne les collectent pas eux-même, à l'exception des géomorphologues) récupérées auprès d'organismes. Ensuite, ils

doivent réaliser une nomenclature pour traiter ces données ainsi que décrire ces données (méta-données) avant de pouvoir passer à l'analyse.

5. La statistique descriptive correspond à l'étude des données. Elle permet de montrer et décrire ces données de manière simplifiée. Elle est utile à la statistique mathématique (c'est-à-dire les prédictions mathématiques, niveau avancé). De plus, elle permet l'établissement de lois de probabilité, aboutit à des calculs et variables, et établit des relations entre plusieurs informations. Elle facilite la lecture/interprétation.
La statistique explicative permet de comprendre les relations entre les variables (le lien inférentiel/causal).

6. Histogramme (variables continues), représentation sectorielle, des intervalles...(tableaux de synthèse).
Pour la lisibilité.

7. Les méthodes descriptives, explicatives et de prévisions.

8. Les types de caractères : qualitatifs (nominal ou ordinal) ou quantitatifs (relatif donc continu (=évolutif) ou absolu donc discret(=fini)), soit le type de variable.
-population statistique : c'est un ensemble (sens mathématique = E). Ex : nombre de personne sur un territoire (spatial), nombre entreprise (non spatial).
-individu statistique : un élément de cet ensemble (de la pop statistique). Ex : une ville parmi l'ensemble des villes.
-caractères statistiques : les caractéristiques de l'individu. Ex : salaire, âge, nombre d'enfants...
-modalités statistiques : valeurs prises par un caractère c'est-à-dire chaque individu est associé à une modalité.
Hiérarchie entre eux : oui.

9. Amplitude : étendue de valeurs d'une donnée (classe). Ex : 30-15 = 15 étendue.
Densité : le nombre d'unité par surface (le rapport entre l'effectif= nombre d'élément) et l'étendue (30-15=15). ex : 1000hab dans une ville de 5km² => **1000/5km² = 200km²**

$$d = \frac{n_i}{b - a}$$

10. Sturges sert à déterminer le nombre d'effectif (classe) pour un histogramme.

Yule sert à calculer l'étendue d'un effectif (classe) pour un histogramme.

11. L'effectif : le nombre de fois qu'une valeur apparaît dans une donnée

La fréquence : le pourcentage (soit la proportion totale). Pour un caractère quantitatif (donc les chiffres), les écrire par ordre croissant (et donc on peut calculer l'effectif cumulé pour arriver à 1).
L'effectif cumulé : la somme des valeurs (la proportion des éléments/à une valeur).
Distribution statistique : c'est faire un tableau qui reprend toutes ces valeurs.

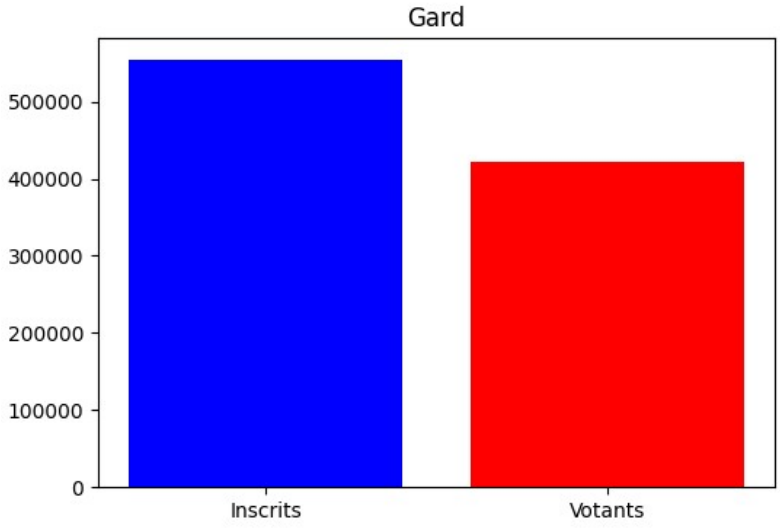
II - Manipulations python.

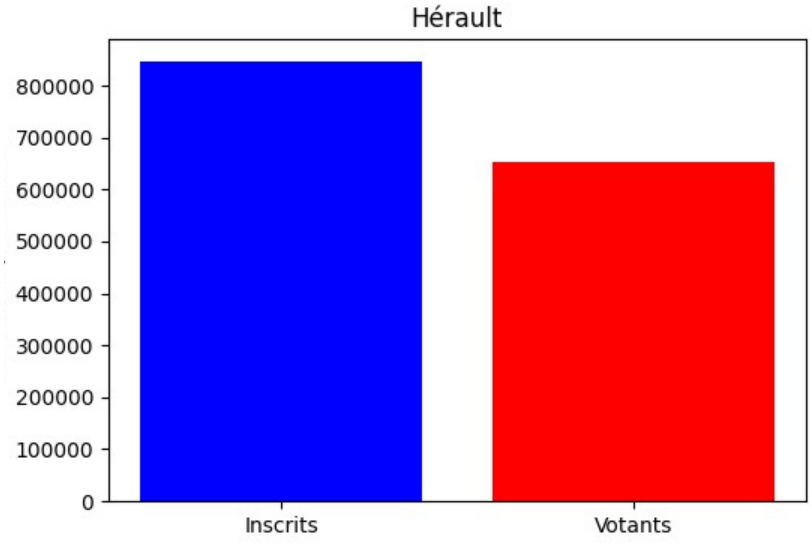
Les manipulations n'étaient pas extrêmement compliquées pour la première séance. Au début, j'avais du mal à faire apparaître le tableau car le ./ avant data n'était pas bon, mais j'ai très vite réalisé le problème et donc j'ai pu finir la séance assez rapidement en faisant directement data/.

```
105      24414
106    1435746
Name: Inscrits, Length: 107, dtype: int64
Nombre total d'inscrits 48747876
[np.int64(48747876), np.float64(12824169.0), np.float64(35923707.0), np.float64(543609.0), np.float64(247151.0), np.float64(35132947.0), np.float64(197094
), np.float64(1101387.0), np.float64(8133828.0), np.float64(2485226.0), np.float64(7712520.0), np.float64(616478.0), np.float64(1627853.0), np.float64(167
6.0)]
Inscrits : 48747876
Abstentions : 12824169.0
Votants : 35923707.0
Blancs : 543609.0
Nuls : 247151.0
```

Bon, globalement je n'ai rien de spécial à dire sur ce code, car il n'y avait pas vraiment de calcul à faire, ni de gros changements, il fallait surtout afficher les données du csv sur python et faire les graphiques à partir de ces dernières, donc les résultats graphiques sont beaucoup plus intéressants que les lignes de codes.III - Résultats graphiques.

Question 11 : diagramme en barre des inscrits et votants par département.

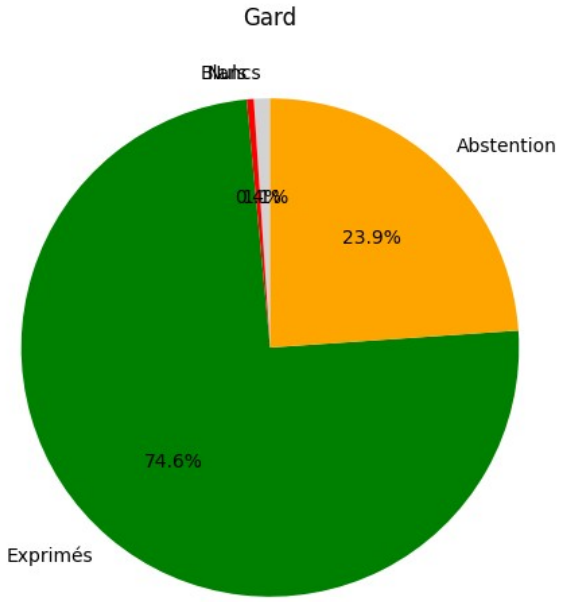




J'ai choisi ces deux diagrammes car je suis Montpelliéraine (de naissance, et j'y ai fait ma CPGE au lycée Joffre) et Nîmoise (là où j'ai grandi et où je vis actuellement).

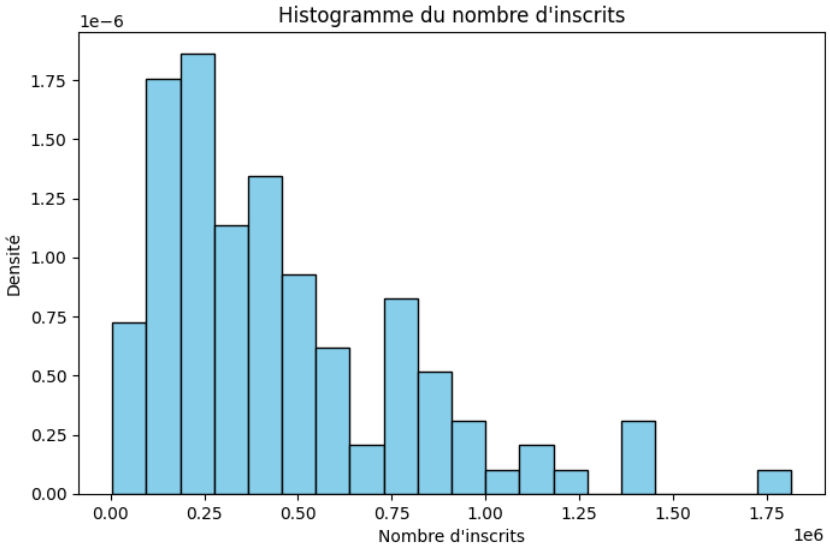
Concernant les graphiques, j'ai mal calibré mon *plt.figsize (6,4)*, de fait la colonne avec indiqué « nombre d'inscrits » a été à moitié supprimée. Je reconnais n'avoir pas pris le temps de le refaire, car ce qui importe ce sont les données, puis j'ai corrigé cette erreur pour les prochains graphiques. Pour une brève analyse, on remarque qu'il y a plus d'inscrits dans l'Hérault que dans le Gard (ce qui est logique étant donné qu'il y a plus d'habitants dans l'Hérault, une différence de 300 000 personnes), mais d'un point de vue des votants, la différence est moins importante (620 000 dans l'Hérault et 450 000 dans le Gard).

Question 12 : diagramme circulaire avec les votes blancs, nuls, exprimés et abstentions par département.

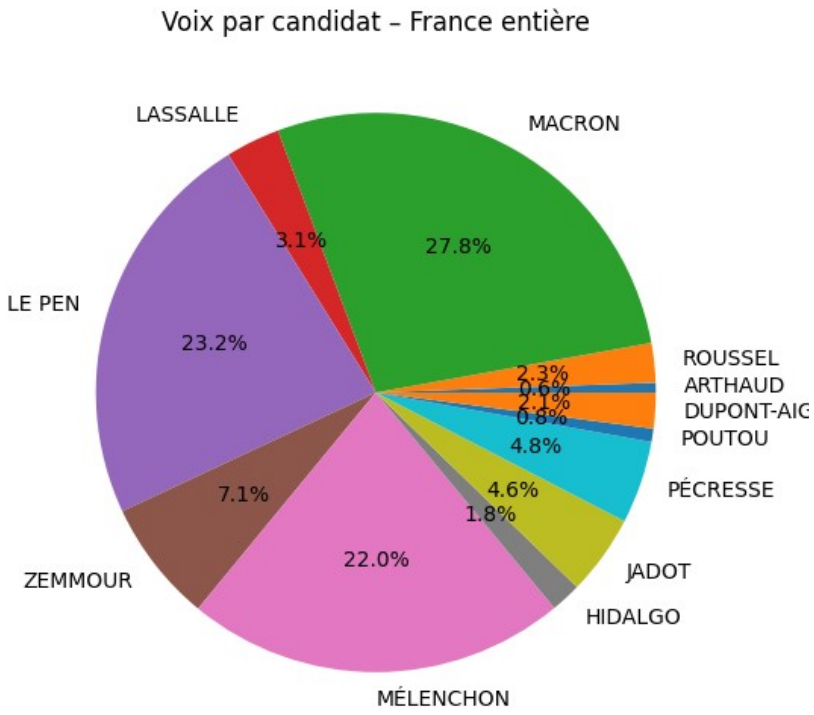


Ce diagramme n'est pas très beau. Il permet de bien voir les exprimés et les abstentions car ce sont les données les plus importantes. Mais dans le cas des votes nuls et blancs, c'est quasiment illisible car ce sont des petites données très similaires en terme de pourcentage.

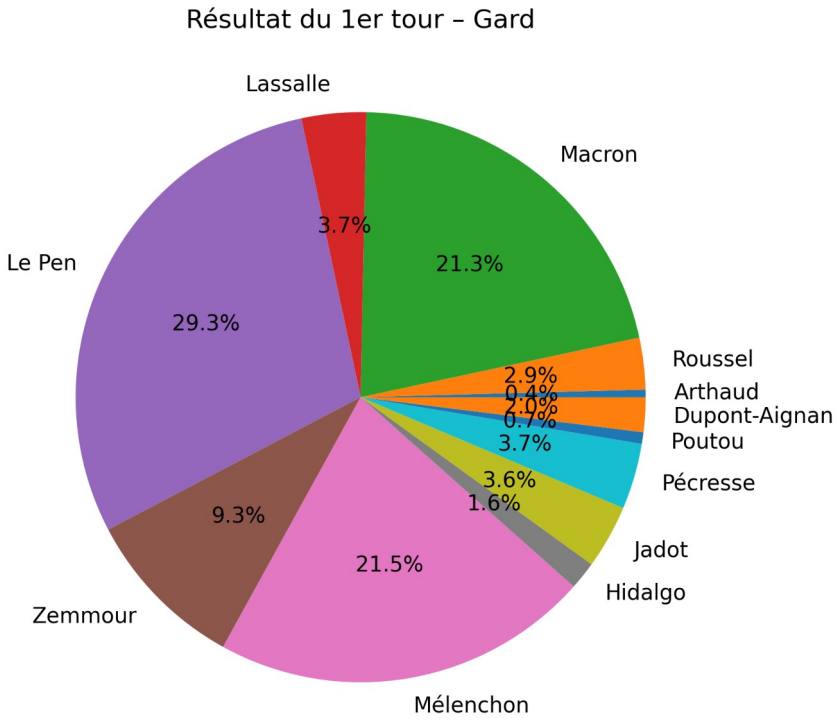
Question 13 : histogramme en barre de la distribution des inscrits.



BONUS : diagramme circulaire représentant le nombre de voix par candidat pour toute la France et pour chaque département.



Bon, M. Dupont-Aignan est aussi à moitié coupé, mais c'était un peu énervant de faire ce diagramme circulaire parce qu'avant il fallait s'assurer que les noms dans le fichier csv soit tous bien indiqué, respecter l'orthographe, voir s'il y a des espaces, puis puisque dans le fichier csv, il y a une colonne différente pour les Voix (Voix 1, 2, 3) et noms des candidats, c'était un peu pénible (à mon niveau), mais finalement j'ai pu réaliser ces graphiques.



J'ai pu rattrapé le coup sur Dupont-Aignan. Le graphique est assez clair, on voit bien les noms et là, ils sont en minuscule. Je ne sais pas trop pourquoi ça a changé, j'ai pas tellement cherché. Je ne suis pas contente des résultats au premier tour dans mon département, mais pas si étonnée...(*c'est vraiment la honte).*

IV - Commentaire sur la séance.

Séance intéressante, pour ma part j'ai bien aimé, et je l'ai trouvé très utile car on traite des données, on fait des graphiques utiles en soi, donc vraiment je suis satisfaite de cette séance et de mon code.
Néanmoins, et là je parle au nom de ma classe, elle était « difficile » pour des personnes qui n'ont jamais fait de Python, même si, à la fin, ils l'ont, pour la plupart, trouvée pertinente, comme moi.
Par contre, quel plaisir d'exécuter le code pour faire des vérifications et avoir 150 images de départements qui se retélécharge car on avait oublié que ça faisait ça (et que j'ai pas fait la fonction def au début).C'est vraiment pas nécessaire...

SÉANCE 3

I - Questions de cours.

1. Remarque sur le cours.

Le cours est trop compliqué à lire pour des questions niveau collège et lycée. Il faut faire moins long et plus simple, intuitif. Il faut penser qu'on a tous des emplois du temps chargé, et même si on voit l'effort donné à la réalisation des cours, il n'est pas adapté pour nous.

2. Réponses aux questions de cours.

1. Le caractère quantitatif car qualitatif est par définition singulier/particulier à un caractère.

2. Les caractères quantitatifs discrets sont des quantités en nombre absolu alors que les caractères quantitatifs continus sont relatifs. La différence est que les quantitatifs absolus expriment un chiffre fixe qui n'est pas amener à changer et les quantitatifs relatifs expriment un chiffre amener à changer (comme un pourcentage sur telle année).

3. Paramètres de position
1) Car la moyenne seule ne permet pas de voir une évolution/régularité (données extrêmes) donc il faut utiliser d'autres paramètres comme la médiane.
2) Elle permet d'observer la « moyenne du milieu », la séparation au milieu d'une série de données. Elle permet de représenter une valeur centrale non influencée par les valeurs extrêmes (de la moyenne).
3) Il correspond à la valeur minimale souscrite à la valeur maximale. Cela permet d'établir la fréquence quand des valeurs se répètent qualitativement.

4. Le coef GINI et la médiale mesurent les inégalités (revenus, etc.) d'un lieu, population et la répartition. Donc ce sont des indices très utiles .

5. Paramètres de dispersion
1) La variance mesure la dispersion globale
2) L'écart-type la rend plus lisible car exprimée dans les mêmes unités que les données(pas de problèmes de conversion).
3) On calcule l'étendue pour connaître l'amplitude totale entre la plus petite et la plus grande valeur.

4) Les quantiles servent à découper les résultats des données en des parties égales (25 %, 50 % et 75%), pratique pour la boîte à moustaches. Les plus utilisés sont donc les quartiles (Q1, Q2, Q3).
5) a boîte à moustaches sert à visualiser la répartition, la symétrie et les valeurs (minimum-maximum) d'une variable (à l'aide des quantiles).

6. Paramètres de forme
1) Les moments centrés mesurent la forme (symétrie, aplatissement) alors que les moments absolus mesurent la dispersion sans signe.
2) Il faut vérifier la symétrique d'une distribution pour savoir si les valeurs de la distribution sont équilibrées : moyenne ≈ médiane = symétrie ; moyenne > médiane = asymétrie à droite.

II - Manipulations python.

Je n'ai rien de spécial à dire. Dans cette séance, il fallait faire des boîtes à moustache pour visualiser les abstentions, les exprimés, les nuls, etc...C'est un exercice en continuité avec la Séance 2 et ça permet de voir les différentes représentations graphiques pour la même source de données.

```
for col in colonnes_quantitatives:
    data = contenu[col].dropna()
    plt.figure(figsize=(6,6))
    plt.boxplot(data)
    plt.title(f"Boîte à moustache - {col}")
    plt.ylabel("Valeurs")
    plt.savefig(f"img/boxplot_{col}.png") #choisir l'emplacement
    plt.close()
```

Je noterai simplement la petite subtilité, où certains ont peut-être été bloqués : bien vérifier que notre ligne de code soit écrit exactement de la même manière que le fichier csv (surface (km^2)).

```
contenu["Surface (km²)"] = pd.to_numeric(contenu["Surface (km²)"], errors='coerce')
```

Ici, on retrouve les manipulations indiquées pour faire apparaître des colonnes dans le terminal python.

	Colonne	Moyenne	Médiane	Mode	Ecart-type	Ecart absolu moy	Etendue	IQR	IDR
0	Inscrits	455587.63	366859.0	5045.0	351003.78	272240.72	1808861.0	401050.0	793988.8
1	Abstentions	119852.05	95369.0	2272.0	117017.80	74959.07	929183.0	106489.0	193676.2
2	Votants	335735.58	274372.0	2773.0	258393.81	201517.17	1297100.0	301770.5	602687.2
3	Blancs	5080.46	4001.0	4577.0	3492.52	2817.95	17389.0	4852.5	8845.8
4	Nuls	2309.82	2039.0	17.0	1501.38	1131.99	8236.0	1917.0	3240.6
5	Exprimés	328345.30	268568.0	2701.0	253758.58	197762.20	1272080.0	296870.5	590169.2
6	Voix	1842.00	1627.0	1203.0	1268.37	977.36	7651.0	1517.5	3015.6
7	Voix.1	7499.27	5968.0	19.0	6501.29	4474.96	45883.0	6264.5	13104.2
8	Voix.2	91430.45	67831.0	534.0	77226.14	59929.14	372286.0	101317.0	177340.2

```
nombre d'île / catégorie de surface :
Surface (km²)
0-10          78423
10-25          2327
25-50          1164
50-100          788
100-2500       1346
2500-5000        60
5000-10000       40
10000+           71
Name: count, dtype: int64
```

Question 10 : organigramme pour visualiser le nombre d'île par rapport aux surfaces.

C'était le code le plus « difficile » et «long» à faire.

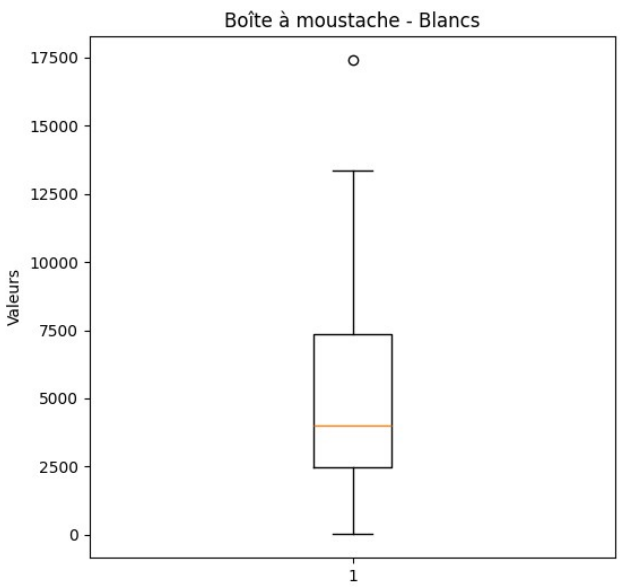
```
contenu["Surface (km²)"] = pd.to_numeric(contenu["Surface (km²)"], errors='coerce')# bien co
surface = contenu["Surface (km²)"].dropna() #sélection de la colonne
bins = [0, 10, 25, 50, 100, 2500, 5000, 10000, float('inf')]
labels = ("0-10", "10-25","25-50","50-100", "100-2500", "2500-5000","5000-10000", "10000+")
categories = pd.cut(surface, bins=bins, labels=labels, right=True, include_lowest=True)
compte = categories.value_counts().sort_index()
print("nombre d'île / catégorie de surface :")
print(compte)
```

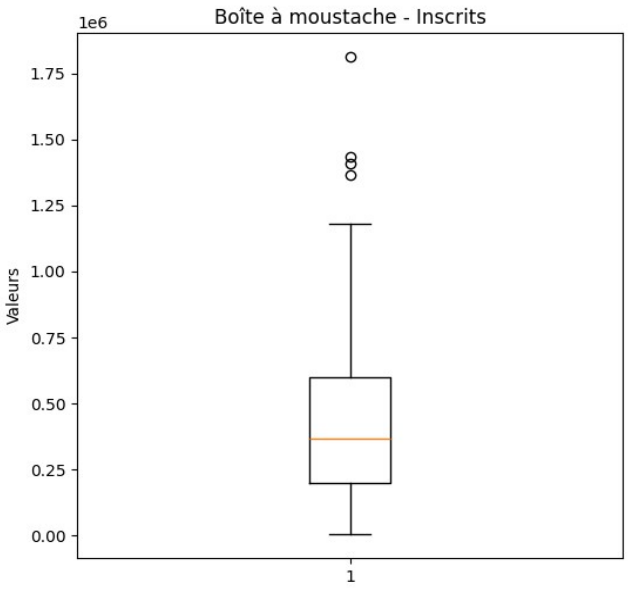
J'ai sélectionné uniquement la colonne de surface en faisant attention de bien noter la « Surface (km²) » sinon ça aurait envoyé un message d'erreur. Puis j'ai précisé les différentes catégories de surface dans des intervalles (bins et labels). Ensuite, python se charge du calcul avec la formule *categories.value_counts().sort_index()* et donc il répartit toutes les données dans la colonne surface par rapport au bins et labels donnés. Ensuite j'ai fait apparaître l'organigramme en faisant *print(compte).*

III - Résultats graphiques.

1. Boîtes à moustaches.

Les boîtes à moustaches sont des représentations graphiques intéressantes, et sympa à produire. J'ai bien aimé les faire sur python, et je suis vraiment satisfaite du résultat donné. Il y a toujours mieux à faire, et on peut toujours s'améliorer, mais ce sont les premiers résultats que j'ai obtenus et ils sont suffisamment clairs et facile à interpréter selon moi.
Ils ont été crées à partir d'une boucle (fonction col) à partir du document csv donné.





IV - Commentaires sur la séance.

La séance s'inscrit parfaitement dans la continuité de la première. On reprend les données de la Séance 2, et on utilise des nouvelles fonctions pour globalement montrer la même chose.
Elle a un intérêt pour les productions graphiques en géographie à partir d'une sélection de données. C'est bien.

SÉANCE 4

I - Questions de cours.

1. Remarque sur le cours.

Le cours était beaucoup trop long et compliqué. Il m'a perdue à de nombreuses reprises. Difficile d'essayer de comprendre, de prendre des notes et de répondre aux questions quand un cours fait plus de 55 pages, qu'il est hermétique et présente des statistiques mathématiques et ses formules. J'ai envoyé le cours à mon cousin en PCSI, en deuxième année, il a eu du mal à le déchiffrer et à comprendre pourquoi je devais lire ça en M1 de Géographie.

3.1.8 Loi hypergéométrique ou loi du tirage exhaustif

La loi hypergéométrique concerne un **tirage sans remise**, c'est-à-dire pour lequel les proportions changent à chaque tirage. Soit une urne contenant N boules de deux couleurs différentes. On note m le nombre de boules blanches et $N-m$ le nombre de boules noires. On tire n boules. Avec remise, la variable aléatoire X représentant les boules blanches suit une loi binomiale de type $X \sim \beta\left(n, \frac{m}{N}\right)$. Sans remise, la loi de X une loi hypergéométrique de paramètre N, m et n, notée $H(N, m, n)$. La probabilité $\Pr(X=k)$ vaut :

$$\Pr(X=k) = \frac{\mathrm{C}_m^k \mathrm{C}_{N-m}^{m-k}}{\mathrm{C}_N^n} \tag{3.41}$$

avec $m = Np$. Il s'agit d'un tirage équiprobable sans remise (ou tirage exhaustif). Les valeurs extrêmes sont : $\min k = \max\{0, n-N(1-p)\}$ et $\max k = \min\{N, Np\}$. Le quotient $\frac{n}{N}$ est appelé **taux de sondage**.

L'espérance vaut :

$$\mathbb{E}(X) = \frac{nm}{N} \tag{3.42}$$

Soit X_1 la variable aléatoire du premier individu, on a alors :

$$\left.\begin{array}{c} \Pr(X_1=1)=p \\ \Pr(X_1=0)=1-p \end{array}\right\} \Rightarrow \left\{\begin{array}{c} \mathbb{E}(X_1)=p \\ \mathbb{V}(X_1)=p(1-p) \end{array}\right. \tag{3.43}$$

Soit X_2 la variable aléatoire du premier individu, on a alors :

$$\left.\begin{array}{c} \Pr(X_2=1)=p \\ \Pr(X_2=0)=1-p \end{array}\right\} \Rightarrow \left\{\begin{array}{c} \mathbb{E}(X_2)=p \\ \mathbb{V}(X_2)=p(1-p) \end{array}\right. \tag{3.44}$$

Pour X_i, $\mathbb{E}(X) = \sum_{i=1}^{n} X_i = np$

La variance vaut :

$$\mathbb{V}(X) = \frac{nm}{N}\left(1-\frac{n}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{3.45}$$

ou

$$\mathbb{V}(X) = \frac{N-n}{N-1} np(1-p) \tag{3.46}$$

Tout ça n'était absolument pas nécessaire, même pas pour comprendre les graphiques python (étant donné que les graphiques aident à comprendre et à visualiser les différents phénomènes statistiques.

2. Réponses aux questions de cours.

1. Je pense que le choix dépend principalement de la nature de la donnée qu'on obtient.
S'il s'agit de valeurs que l'on peut compter, alors c'est une variable discrète. Si la variable correspond à des mesures pouvant prendre toutes les valeurs possibles dans un intervalle, comme la température, il s'agit d'une variable continue. Après, certaines variables normalement continues peuvent être regroupées en classes et traitées comme discrètes pour simplifier l'analyse et la reproduction graphique (car c'est plus simple avec des données absolues).

2. Les lois statistiques les plus utilisées en géographie varient beaucoup du type de phénomène observé.

- La loi normale est la plus fréquente et utilisées pour des variables mesurables (ex : les altitudes, les températures ou les tailles de population à partir d'une moyenne).
- La loi de Pareto est souvent utilisée pour les phénomènes avec des valeurs beaucoup plus petites (ex : la distribution des villes par taille).
- Les lois binomiale et de Poisson servent à modéliser des événements plus rares, ponctuelles (ex : le nombre de catastrophes naturelles dans une région donnée à un moment donné=valeur absolue).

II - Manipulations python.

La séance 4 visait à introduire les **distributions statistiques théoriques** à travers leur **simulation informatique** en Python. L'objectif principal était de comprendre la différence entre **variables discrètes et continues** et d'observer concrètement la forme des distributions à l'aide d'échantillons simulés.
À l'aide de la bibliothèque scipy.stats, différentes lois discrètes (Dirac, uniforme discrète, binomiale, Poisson, Zipf-Mandelbrot) et continues (normale, log-normale, uniforme, χ^2, Pareto) ont été générées sous forme d'échantillons aléatoires. Ces simulations permettent de visualiser la structure des distributions et d'identifier leurs caractéristiques principales (concentration, dispersion, asymétrie).
Des fonctions Python ont ensuite été écrites afin de calculer **la moyenne et l'écart-type** de chaque distribution. Cette étape permet de relier les propriétés théoriques des lois statistiques aux résultats numériques obtenus à partir des échantillons simulés.

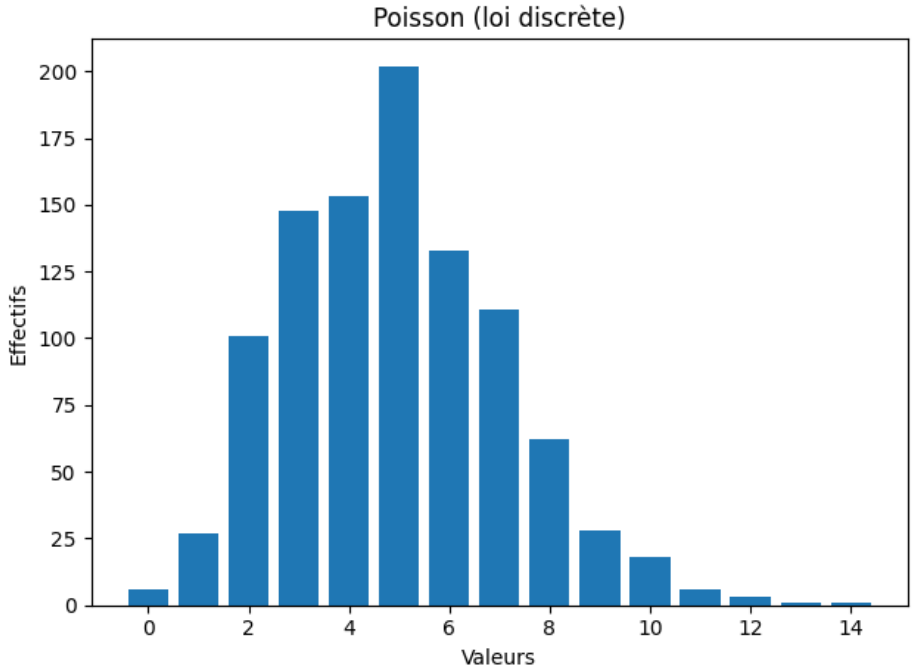
L'utilisation conjointe de scipy.stats, numpy et de fonctions personnalisées permet ainsi de comprendre comment les distributions statistiques peuvent être **modélisées, simulées et analysées** à l'aide d'outils informatiques. Cette approche constitue une base méthodologique pour les séances suivantes, notamment pour la comparaison entre distributions théoriques et données observées.

III - Résultats graphiques.

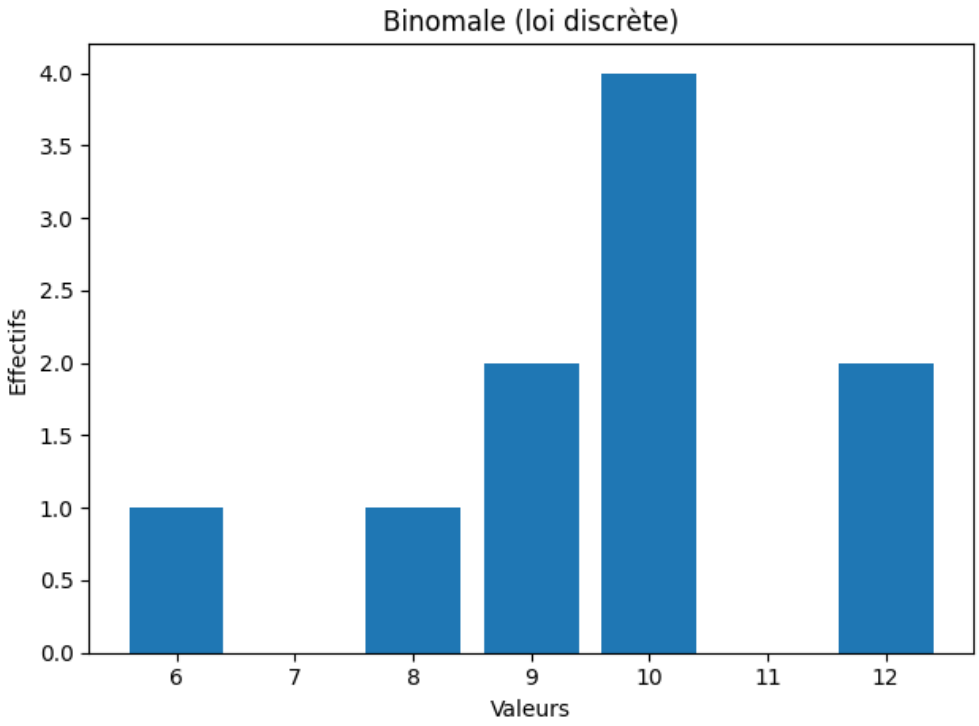
J'ai choisi, à titre d'exemples, les graphiques représentants les lois statistiques les plus utilisées en géographie

1. Distributions statistiques pour les variables discrètes.

1.1. Loi du Poisson.

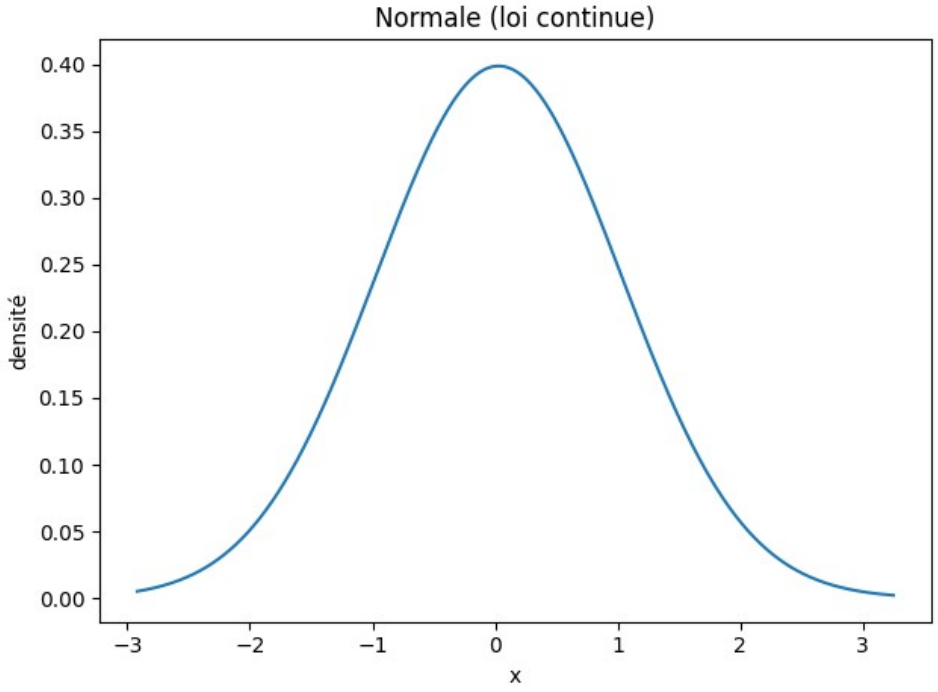


1.2. Loi Binomale.

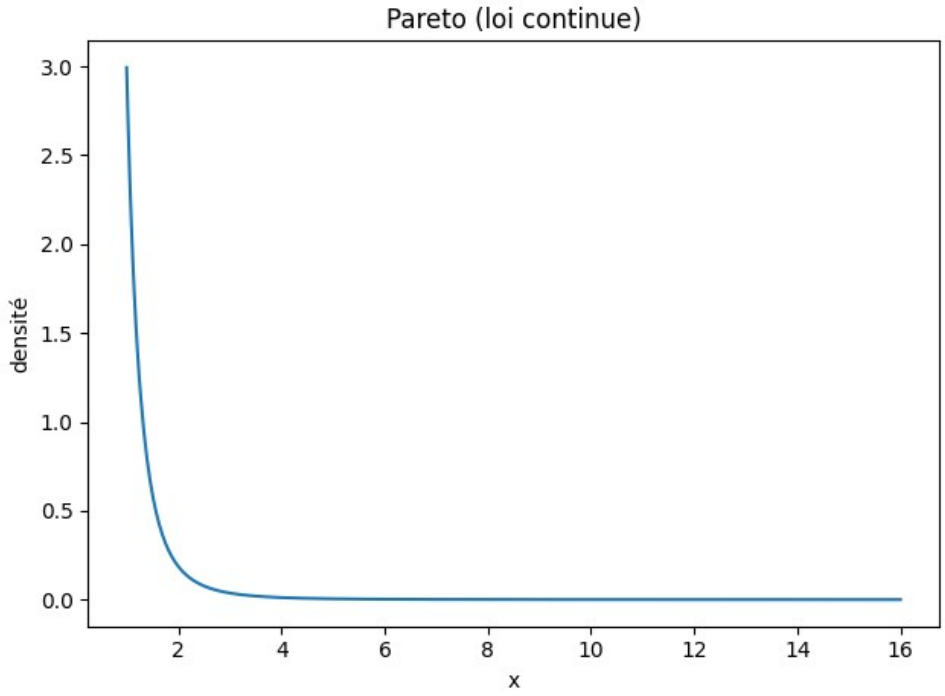


2. Distributions statistiques pour les variables continues.

2.1. Loi Normale.



2.2. Loi de Pareto.



IV - Commentaires sur la séance.

Je n'ai pas vraiment compris l'intérêt de la séance au début, puis j'ai compris son intérêt après la production graphique, mais cette séance est trop mathématique. Venant d'un parcours essentiellement littéraire, ce n'est pas assez intuitif. Il aurait fallu nous expliquer en cours son intérêt, les différentes formules et ce qu'elles impliquent car le cours est toujours autant hermétique.

Globalement, on peut la trouver intéressante lorsqu'on l'applique à la géographie (et que surtout on comprend comment cela s'applique), et encore, on ne fait pas ça dans nos Master respectif... Je pense que cet intérêt est uniquement personnel, et pas à l'échelle de toute la promo GAED (on doit être très peu à avoir aimé cette séance).

SÉANCE 5

I - Questions de cours.

1. Remarque sur le cours.

La séance 5 avait pour objectif d'introduire les fondements de la **statistique inférentielle** à travers trois étapes complémentaires : l'échantillonnage, l'estimation et la prise de décision statistique. On a pu tester ces notions à travers les exercices réalisés sur Python.
Sinon, le cours n'est pas intuitif, je reconnais n'avoir pas cherché à le lire, ainsi que celui de la séance 4 et 6. Trop long, trop mathématiques...On n'est pas en études de statistiques. Pour ma part, ça fait des années que je n'ai pas fait de mathématiques, je sors d'une prépa littéraire, comme la plupart d'entre-nous, je suis en Master de Géographie sociale et culturelle...Comprenez que vos cours ne sont pas du tout adaptés, et clairement ils ne sont pas utiles pour notre compréhension de python, la preuve, j'ai réussi les manipulations python sans le cours. Je pense que la meilleure chose à faire, c'est un tout petit cours, écrit simplement, qui explique les différentes théories d'échantillonnage et les formules Python à utiliser, rien de moins, rien de plus.

2. Réponses aux questions de cours.

1. L'échantillonnage consiste à sélectionner une partie représentative d'une population pour réaliser des analyses statistiques. On ne peut pas toujours utiliser la population entière, car cela peut être long à sonder et même impossible sauf si c'est un village de 50 habitants. Les principales méthodes d'échantillonnage sont :

- Aléatoire
- Stratifié
- Systématique
- Par grappes

Le choix dépend de la structure de la population, du niveau de précision souhaité et des ressources disponibles.

2. Un estimateur est une statistique calculée à partir d'un échantillon qui permet d'estimer un paramètre inconnu de la population (ex : la moyenne).
Une estimation est la valeur numérique obtenue en appliquant cet estimateur à un échantillon précis.

3. L'intervalle de fluctuation sert à prédire la variabilité d'un échantillon autour d'une valeur connue.
L'intervalle de confiance sert à estimer avec un certain niveau de certitude un paramètre inconnu de la population à partir d'un échantillon.

4. Un biais est la différence entre la valeur moyenne de l'estimateur et le paramètre réel de la population. Un estimateur est non biaisé si sa valeur moyenne correspond exactement au paramètre qu'il estime.

5. Une statistique calculée sur la population entière s'appelle un paramètre. Avec les données massives (big data), il devient parfois possible de travailler sur la population totale plutôt que sur un échantillon, réduisant ainsi le besoin d'estimation et les erreurs.

6. Le choix d'un estimateur influence la précision et la fiabilité des résultats. Un mauvais estimateur peut conduire à des résultats trompeurs.

7. Les différentes méthodes :

- Méthode des moments : on fait correspondre les moments de l'échantillon à ceux de la population.
- Méthode de vraisemblance : on choisit le paramètre qui rend les données observées les plus probables.
- Méthode bayésienne : on combine des données avec des informations a priori sur le paramètre.

Le choix dépend de la nature des données, des hypothèses sur la population, des méthodes calculatoires.

8. Les test statistiques :

- Test t de Student
- Test du chi^2
- ANOVA
- Tests non paramétriques

Ils servent à vérifier une hypothèse statistique à partir de données.

Pour créer un test, on définit une hypothèse, puis on choisit une statistique de test, on calcule sa valeur à partir de l'échantillon, puis on compare les valeurs.

9. Je pense qu'on peut se méfier de ces statistiques car elles reposent des hypothèses, et que les résultats sont probablement faux (or, on cherche des résultats justes). Donc, si on fait que des test mais qu'en plus, on a pas la garantie qu'ils soient pas erronés, c'est problématiques. Mais, ces statistiques sont essentielles pour tirer des conclusions sur une population à partir d'un échantillon, surtout quand l'analyse de la population entière est impossible.

II - Manipulations python.

1. Théorie de l'échantillonnage.

Dans un premier temps, un jeu de données simulant **100 échantillons** issus d'une population mère a été analysé. À l'aide de Pandas, les moyennes par opinion ont été calculées et comparées aux valeurs réelles de la population mère telle qu'indiquée. Après, il fallait calculer les fréquences et des intervalles de fluctuation à 95 %. Cela a permis de montrer que les résultats issus de l'échantillonnage varient autour d'une valeur théorique, tout en restant globalement cohérents avec celle-ci. Cette étape met en évidence que les échantillons sont toujours très incertain (et c'était l'intérêt de cette séance).

```
Intervalles de fluctuation à 95 %
Pour : [0.369, 0.41]
Contre : [0.396, 0.438]
Sans opinion : [0.177, 0.21]
```

2. Théorie de l'estimation.

La seconde partie repose sur l'analyse d'un échantillon unique, représentant une situation plus réaliste en géographie. À partir de cet échantillon, les fréquences ont été calculées puis on a réalisé des intervalles de confiance. Cette méthode permet d'estimer les paramètres de la population mère sans en connaître les valeurs exactes, et d'évaluer la fiabilité des résultats obtenus.

```
Intervalle de confiance à 95 % (du premier échantillon)
Pour : [0.37, 0.43]
Contre : [0.37, 0.43]
Sans opinion : [0.185, 0.235]
```

3. Théorie de la décision.

Enfin, la séance a montré que les tests statistiques sont des outils d'aide à la décision. Le test de Shapiro–Wilk (méthode de scipy.stats) a été utilisé afin de déterminer si une distribution suit une loi normale.

Mon interprétation a reposé sur la valeur de la p-value :
Si p-value<0,05, l'hypothèse de normalité est rejetée (ce qui était le cas dans les fichiers). J'ai donc eu besoin de réaliser un graphique pour visualiser et ainsi déterminer la normalité d'un des fichiers.

```
Théorie de la décision
Résultats du test de shapiro
Fichier 1 → p = 0.0000 → Distribution non normale (p < 0.05)
Fichier 2 → p = 0.0000 → Distribution non normale (p < 0.05)
```

4. Explications test de Shapiro.

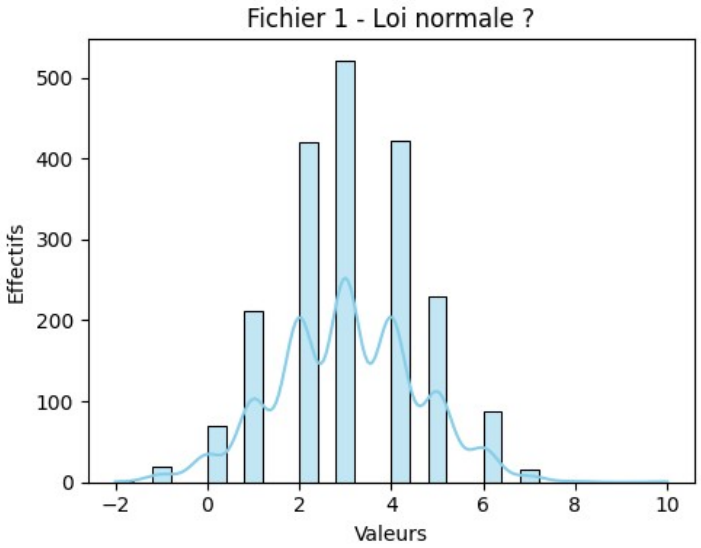
J'ai été perturbée au début car le test de Shapiro–Wilk indique des p-values très faibles pour les deux fichiers ($p < 0.05$). J'étais donc partie du principe qu'aucune des deux lois était normale. Cependant, après avoir effectué des recherches et ayant relu les manipulations, j'ai compris que ce test était sensible à la taille des échantillons, donc qu'il était possible que les distributions soient visuellement proches d'une loi normale (en réalisant des graphiques avec *matplotlib*) malgré ce résultat statistique.

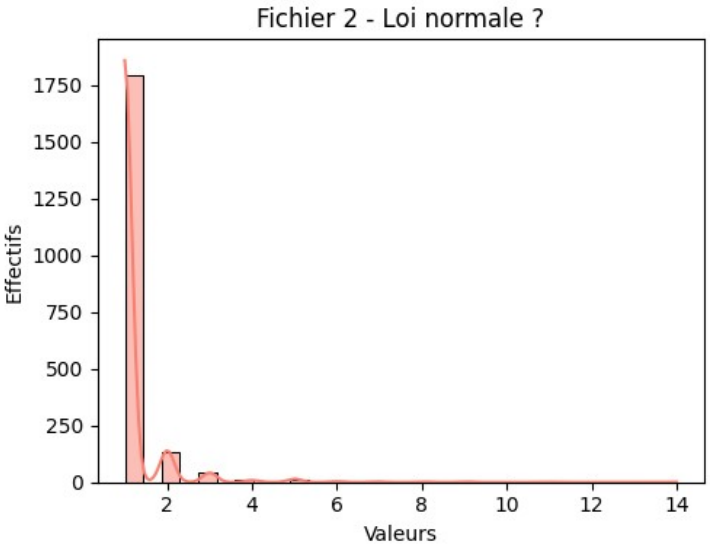
J'ai donc vérifié avec un histogramme dessinant une courbe. Les deux ne semblent pas normales dans le sens que ce n'est pas une courbe en cloche parfaite mais tout de même, visuellement, le **fichier 1** présente une répartition proche d'une courbe en cloche, donc d'une loi normale.

III - Résultats graphiques.

1. Test de Shapiro : Loi normale ?

J'ai effectué ce test afin de visualiser laquelle des deux lois était normale étant donné que le test mathématique n'a pas été concluant.

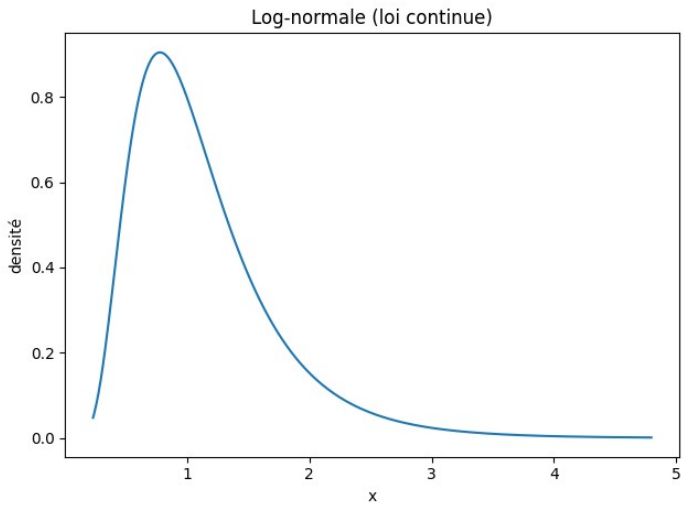


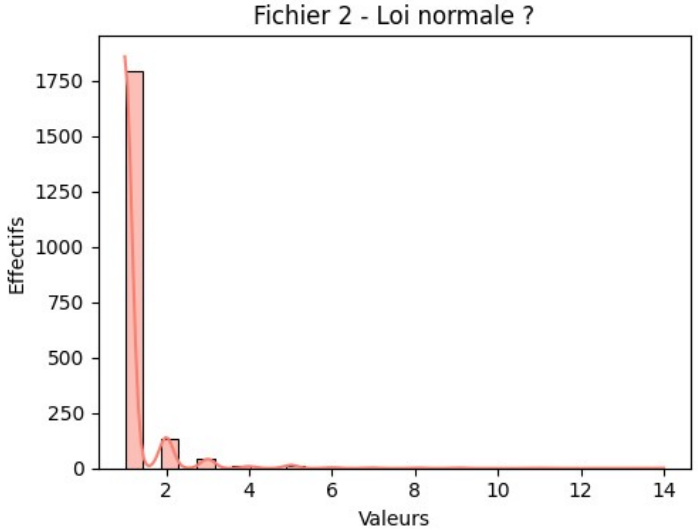


Voici les rendus graphiques pour le test de Shapiro. On voit que le graphique de gauche (fichier 1) présente une courbe qui peut prendre l'allure d'une cloche, donc de normalité (même s'il y a des imprécisions). Dans tous les cas, le fichier 2, à droite, n'est clairement pas une distribution normale, elle s'en éloigne complètement. Ainsi, je suis partie du principe que le fichier 1 représentait une loi normale, visuellement.

2. BONUS. Quelle distribution pour le deuxième graphique ?

Le test de Shapiro–Wilk a montré à travers la représentation graphique que seule la première série suit une loi normale (fichier 1). La seconde série, on voit que les valeurs sont concentrées à gauche et s'étendent en hauteur. Ce modèle ne suit pas celui d'une loi normale. En observant sa distribution, on constate une asymétrie typique d'une loi log-normale, qui décrit des données positives et très étalées vers les grandes valeurs. J'ai vérifié cet hypothèse en observant le graphique que j'avais obtenu lors de la séance 4, sur la loi log-normale (graphique de gauche).





On aperçoit assez clairement une ressemble (certes pas parfaite), mais bon, je pense qu'il s'agit d'une loi log-normale pour le fichier 2.

IV - Commentaires sur la séance.

J'ai globalement apprécié la séance 5, notamment parce qu'elle mobilise des outils statistiques relativement rigoureux (fréquences, intervalles de fluctuation, intervalles de confiance, tests de normalité) qui relèvent davantage des sciences dites « dures ». Ce cadre méthodologique m'a intéressée intellectuellement, car il permet de comprendre comment on peut quantifier l'incertitude et évaluer la fiabilité d'un échantillon par rapport à une population mère. En revanche, cette séance m'a aussi paru très fortement mathématique et relativement déconnectée des problématiques géographiques. Les manipulations réalisées en Python (calculs de moyennes, de fréquences, tests de Shapiro-Wilk) relèvent davantage d'un exercice de statistique appliquée que d'une véritable analyse spatiale ou territoriale. Si j'ai pris du plaisir à effectuer ces calculs et à comprendre la logique probabiliste sous-jacente, j'ai eu plus de difficulté à percevoir l'intérêt concret de ce type d'approche dans le cadre de la géographie, en particulier lorsque les résultats ne sont pas véritablement mis en relation avec des dynamiques spatiales ou sociales.

SÉANCE 6

I - Questions de cours.

1. Remarque sur le cours.

Non, je risquerai de me répéter.

2. Réponses aux questions de cours.

1. Une statistique ordinale est une statistique qui classe les individus ou unités selon un ordre ou un rang, mais sans mesurer l'écart exact entre eux.
Elle s'oppose à une statistique nominale, qui ne fait que regrouper les individus en catégories sans hiérarchie. Elle utilise des variables ordinales, comme la densité de population classée par quartiles. Elle permet de mettre en évidence une hiérarchie spatiale, (ex : ordonner la population, comme on l'a fait dans les manipulations).

2. Dans les classifications, il est préférable d'adopter un ordre croissant ou décroissant

3. La corrélation des rangs mesure la force et le sens de la relation entre deux variables classées par ordre (coefficient de Spearman). La concordance de classements évalue juste le degré «d'accord» *(je ne sais pas comment dire)* entre deux classements, c'est-à-dire combien d'éléments occupent des positions similaires dans les deux listes, sans quantifier l'intensité de la relation (contrairement au coefficient de Spearman).

4. La différence entre les deux :

- Le test de Spearman se base sur la différence des rangs et calcule un coefficient de corrélation, souvent plus sensible aux extrêmes.
- Le test de Kendall se base sur le nombre de paires concordantes et discordantes, et est plus robuste pour de petits échantillons ou les égalités.

5. Le but :

- Le coefficient de Goodman-Kruskal mesure la force de l'association entre deux variables ordinales, afin de réduire l'incertitude d'une des variables.
- Le coefficient de Yule sert à quantifier l'association entre deux variables différentes, en évaluant la probabilité qu'un événement se produise conjointement avec un autre par rapport à leur occurrence individuelle.

II - Manipulations sur python.

Question 6 : Fonction pour convertir les données en données logarithmiques ?

La fonction ordrePopulation(pop, etat) associe chaque valeur numérique à un pays, supprime les valeurs manquantes (np.isnan()), puis ordonne les pays par rang décroissant. Le rang est ensuite explicitement reconstruit (element + 1), ce qui permet de travailler non plus sur les valeurs mais sur leur position relative. Cette étape est essentielle car les corrélations de rangs ne comparent pas des quantités mais des ordres.

Question 7 : Tests sur des rangs ?

On peut pas faire des tests sur des rangs...car ce sont des rangs ? donc pas des données statistiques.
Un rang = position dans un classement à partir des valeurs que j'ai triées c'est tout.

Question 11 – Fonction *def* OrdrePopulation.

```python
#Question 11 : Fonction pour obtenir le classement des listes spécifiques aux populations
def ordrePopulation(pop, etat):
    ordrepop = []
    for element in range(0, len(pop)):
        if np.isnan(pop[element]) == False:
            ordrepop.append([float(pop[element]), etat[element]])
    ordrepop = ordreDecroissant(ordrepop)
    for element in range(0, len(ordrepop)):
        ordrepop[element] = [element + 1, ordrepop[element][1]]
    return ordrepop
etats = monde["État"].tolist()
pop2007 = monde["Pop 2007"].astype(float).tolist()
pop2025 = monde["Pop 2025"].astype(float).tolist()
dens2007 = monde["Densité 2007"].astype(float).tolist()
dens2025 = monde["Densité 2025"].astype(float).tolist()

classement_pop2007 = ordrePopulation(pop2007, etats)
classement_pop2025 = ordrePopulation(pop2025, etats)
classement_dens2007 = ordrePopulation(dens2007, etats)
classement_dens2025= ordrePopulation(dens2025, etats)
```

La structure du code proposée repose sur des boucles et des conversions manuelles qui, si elles permettent de comprendre la logique de base du tri, alourdissent considérablement la manipulation. L'usage des méthodes natives de Pandas permettrait d'obtenir les mêmes résultats en quelques lignes, tout en rendant le code plus lisible. Ce que vous aviez proposé était assez...chaotique.

Question 12 – Comparaison et sort.

```
           État    Pop 2007  Pop 2025  Densité 2007  Densité 2025
0     Afghanistan  31100000.0  42600000          48.0         65.24
1  Afrique du Sud  47300000.0  64000000          39.0         52.37
2         Albanie   3200000.0   2700000         110.0         93.10
3         Algérie  33500000.0  46800000          14.0         19.63
4       Allemagne  82400000.0  84500000         231.0        236.69
[[1, 'Chine'], [2, 'Inde'], [3, 'États-Unis'], [4, 'Indonésie'], [5, 'Brésil'], [6, 'Pakistan'], [7, 'Bangladesh'], [8, 'Russie'], [9, 'Nigeria'], [10, 'Japon']]
```

La fonction classementPays() permet de mettre en correspondance deux classements (par exemple population et densité) en s'appuyant sur le nom des États. Le résultat est une liste de paires de rangs, nécessaire pour appliquer les tests statistiques ultérieurs.
Les listes de rangs sont ensuite isolées à l'aide d'une boucle, car les fonctions de corrélation exigent deux listes numériques distinctes.

Question 14: Coef de corrélation et concordance des rangs.

```python
from scipy.stats import spearmanr, kendalltau

print("Kendall et Spearman")

coef_spear_2007, p_spear_2007 = spearmanr(rangs_pop_2007, rangs_dens_2007)
coef_kendall_2007, p_kendall_2007 = kendalltau(rangs_pop_2007, rangs_dens_2007)
print("Spearman :", coef_spear_2007, "p-value :", p_spear_2007)
print("Kendall :", coef_kendall_2007, "p-value :", p_kendall_2007)

coef_spear_2025, p_spear_2025 = spearmanr(rangs_pop_2025, rangs_dens_2025)
coef_kendall_2025, p_kendall_2025 = kendalltau(rangs_pop_2025, rangs_dens_2025)
print("Spearman :", coef_spear_2025, "p-value :", p_spear_2025)
print("Kendall :", coef_kendall_2025, "p-value :", p_kendall_2025)
```

```
Kendall et Spearman
Spearman : 0.09727602179182566 p-value : 0.20561225645362893
Kendall : 0.06928104575163399 p-value : 0.17856904513455385
Spearman : -0.027540929971689544 p-value : 0.7037988164679578
Kendall : -0.008419689119170985 p-value : 0.8619617935380242
```

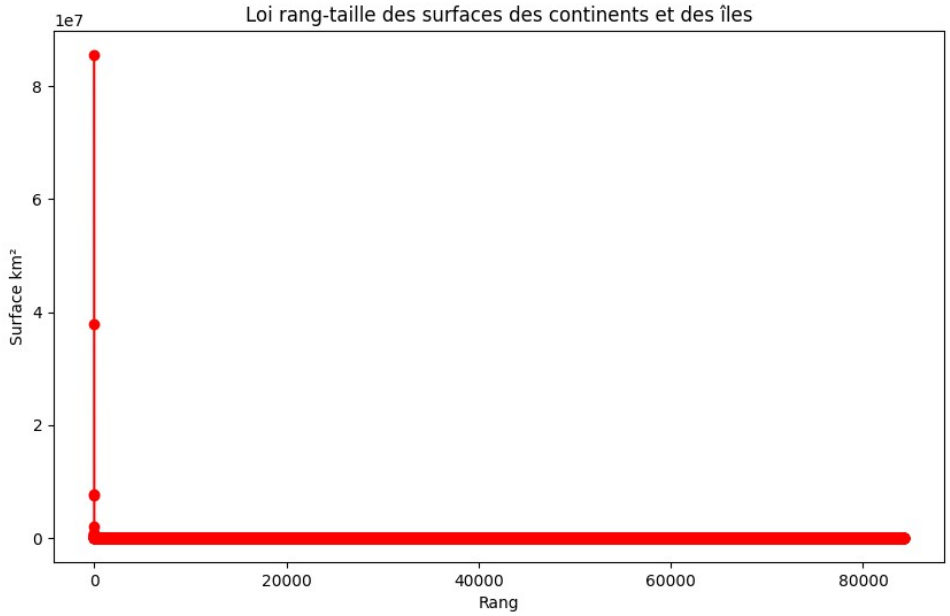
Enfin, les fonctions spearmanr() et kendalltau() de scipy.stats ont été utilisées.

- **Spearman** mesure la corrélation monotone entre deux classements.
- **Kendall** mesure le degré de concordance entre les rangs.

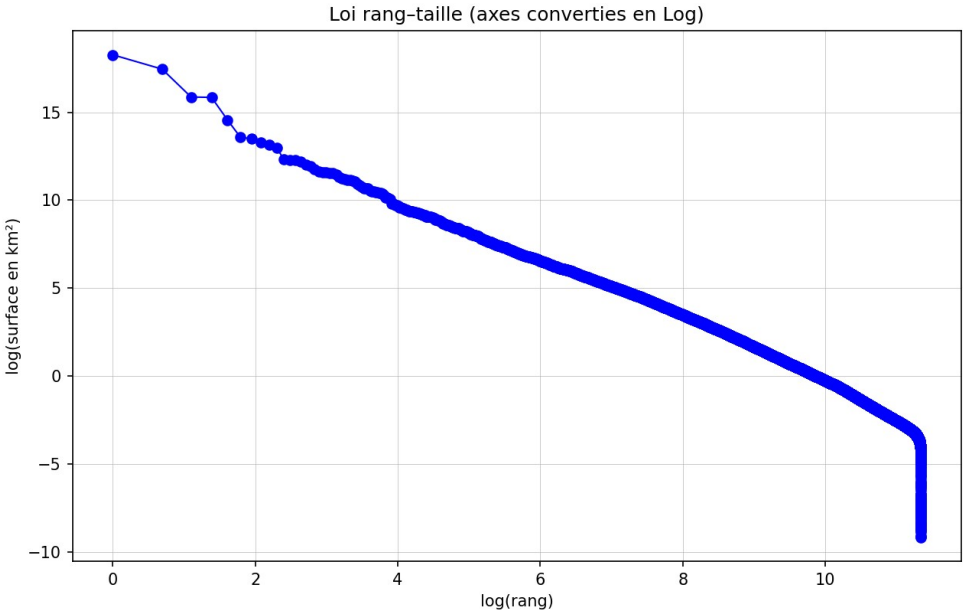
Ces tests sont particulièrement adaptés ici car les données analysées sont des **classements** et non des variables quantitatives continues.

III - Résultats graphiques.

1. La loi rang-taille.



2. La loi rang-taille avec la formule (log).



IV - Commentaires sur la séance.

La séance 6 est celle que j'ai le moins appréciée. Les objectifs généraux (comparaison de classements, analyse des rangs, utilisation de coefficients de corrélation) sont intéressants sur le plan théorique, mais leur mise en œuvre m'a semblé inutilement complexe. Les fonctions proposées dans le fichier *main.py* sont longues, peu lisibles et reposent sur une écriture du code que j'ai trouvée datée, voire contre-productive pour l'apprentissage. Certaines fonctions auraient pu être remplacées par des méthodes beaucoup plus simples et plus claires, notamment via les bibliothèques Pandas ou SciPy déjà utilisées ailleurs dans le cours. Cette complexité excessive a rendu la compréhension du raisonnement statistique plus difficile, l'attention se portant davantage sur le décryptage du code que sur l'analyse des résultats. J'ai utilisé cette fois l'IA car je n'arrive pas à trouver de solution pour la deuxième partie de l'exercice. Quant au bonus, il m'a semblé dépourvu d'intérêt réel : il demande un investissement important en temps et en programmation sans rien réellement apporter. J'ai ainsi eu le sentiment que cette séance privilégiait la technicité informatique au détriment de la compréhension géographique et statistique, ce qui a rendu le travail à la fois frustrant et peu formateur.

CONCLUSION

Message pour les camarades que j'ai aidés.

Vous le savez, et tout le monde le sait : j'ai été en quelque sorte la batterie qui a alimenté toute ma classe durant ce semestre en Analyse de Données.
Au départ, j'ai voulu aider les camarades avec qui je m'entendais le mieux et de bouche à oreilles, tout le monde s'est tourné vers moi...même le 18 décembre au soir (et oui, ce n'est pas une blague).

Il n'y a pas **un jour** où je n'ai pas été sollicitée par quelqu'un par rapport à l'Analyse de Données. Essentiellement pour :
- L'installation (et oui...Docker, toujours et l'unique Docker).
- Les bibliothèques
-Explications des lignes de codes et du but des séances (parfois).
-Lignes de codes
-Résultats graphiques

Bref, sur tout finalement.
Je pense, très honnêtement, avoir installé Python sur plus de 25 ordis différents au sein de ma classe, et 4-5 ordis en plus (personnes des autres Master).
J'ai fait la séance 2 sur plus d'une vingtaine d'ordis, et je compte pas le nombre de problèmes résolus (ouverture du fichier avec la commande data, affichage des tableaux dans le terminale, absence de bibliothèques, vérification de lignes de code).

J'ai réservé, à la demande de mes camarades, de nombreuses salles de travail à Clignancourt, où j'organisais des séances « python » qui consistait à projeter mon ordi et montrer la progression/lignes de code, question par question. La dernière date du 13 décembre.
Je n'ai jamais arrêté, et à l'heure actuelle, je suis dans le train, en direction du sud de la France (mon chez-moi), je rédige ce rapport en retard car j'étais encore en train d'aider mes camarades.

Comprenez que si je vous dis tout ça, ce n'est pas pour « me la raconter », au contraire, je suis une personne nonchalante, qui jouait aux Sims devant vous (mais au moins, j'étais présente).

Seulement, c'est pour vous dire l'ampleur du désespoir des gens de ma classe (et du mieux accessoirement) vis-à-vis de votre cours. Rien n'allait, personne n'avait python, personne ne comprenait rien, honnêtement, ça m'a fatigué d'aider tout le monde, mais bon, l'essentiel, comme je le dis toujours : « c'est que ça finisse par marcher ». Et ça marche !

Ainsi, je voulais mettre un petit mot pour mes camarades qui ont vraiment cherché à faire un rapport propre malgré des pleurs, cris et panique (oui, oui). Je sais que beaucoup m'ont citée, mais j'aimerais le faire aussi, et j'aimerais beaucoup **que vous serez indulgents envers eux, et plus globalement, envers tout le monde.** Mais notez réellement l'effort de certains qui, une fois lancés, ont pu continuer le travail avec mon soutien.

Je pense à Myriam Menard, à Anaé Delliere, à Philippine Nourrit, à Inès Gbale, à Roxane Célestine qui étaient toujours présentes aux cours, assises au fond avec moi. Je les ai beaucoup aidées, et elles ont fini par réussir, étape par étape, avec mon appui, certes, mais je suis vraiment contente pour elles. Elles étaient angoissées, elles ne pensaient pas pouvoir vous envoyer de rapport, comme quoi...

Il y aussi évidemment Matthieu Elizéon, à Loris Kauffmann, Ababakar Fall (qui a perdu son ordinateur, malheureusement, soyez vraiment indulgent s'il vous plaît), Alida Tuedom, Karla Alfaro, qui n'étaient pas dans le même groupe que moi, mais ce sont aussi mes amis, et ils étaient les premiers à vouloir aussi réserver des salles pour travailler.

Puis à ceux que j'ai aidé brièvement, globalement entièreté de ma classe, à un moment où un autre.

Bilan et critique générale sur le cours d'Analyse de données.

Je pense qu'il y a **beaucoup** de choses à revoir pour améliorer ce cours.
Déjà, abandonner l'idée de Docker, totalement inutile et inadapté pour des étudiants en Géographie.
Puis, abandonner Github, à la rigueur, on peut l'utiliser pour récupérer les fichiers des séances, le reste, on le met sur Moodle ou par mail (dossier ZIP), mais là, c'était très pénible.
Aider à l'installation de Python, jusqu'au bout.
Faire des cours simples, limite niveau collège, pas des cours quasiment incompréhensible même pour des étudiants en maths sup maths spé.

Et s'il vous plaît, arrêtez de nous rabâcher qu'on est des « littéraires », même si c'est le cas de beaucoup d'entre-nous, ce n'est pas une généralité. Pire, ça nous rabaisse, et nous décourage à faire du python. La pédagogie, c'est important.

Globalement, de ce que je retiens des cours en Analyse de Données...J'ai aimé faire les graphiques, et me remettre un peu à l'informatique, mais parce que **moi** je connaissais (du moins un peu), et je pense bien être une des seules à avoir un minimum compris ce qu'il fallait faire (et qui ait le moins utilisé l'IA, même si je sais que mon langage peut sembler robotique, c'est la prépa qui m'a formatée ainsi).

Je n'ai rien d'autre à ajouter, j'espère que tout est clair. Je pense que tout le monde est déçu, frustré de ce cours, et ça s'est vu lorsque l'effectif diminuait chaque séance.

J'ai fait mon maximum, j'espère que ça se verra et que vous apprécierez mon rapport même si, je l'admets, moi aussi je ne suis absolument pas satisfaite de ce cours et c'est dommage, car il y avait un bon potentiel, et je suis frustrée que ce dernier n'ait pu être totalement exploité.

Je vous souhaite, sincèrement, de passer de belles fêtes de fin d'année.

Zara **HUSTON.**