Analyse de données (parcours débutant)
Rendu (Valentina Proust M1 SCT)

Séance 2

1. Les statistiques sont utilisées comme un outil d'analyse, on prélève des données que l'on analyse, interprète, calcul pour décrire, comparer ou encore expliquer des phénomènes spatiaux.
2. Oui le hasard existe dans tout ce qui touche le vivant mais en terme statistique on parlera plus de variabilité et d'aléa.
3. Il existe plusieurs types d'information géographique : données quantitatives, données qualitatives, données spatiales (localisées), données temporelles.
4. On utilise les données en géographie pour décrire, comparer, expliquer, modéliser et représenter spatialement les phénomènes.
5. La statistique descriptive décrit les données, donc crée par exemple des graphiques ou des moyennes alors que la statistique explicative cherche des causes et des relations.
6. Selon le type de donnée ainsi que de l'échelle et de l'objectif, on peut choisir différents types de visualisation. Cela peut être une carte, notamment pour les choroplèthes, points, et flux. Ou des graphiques pour représenter des barres, courbes ou par secteur.
7. Les méthodes d'analyse de données possibles sont : les statistiques descriptives, les analyses multivariées, la corrélation et la régression ainsi que l'analyse spatiale (SIG).
8. La population statistique est l'ensemble des individus étudiés, tandis que l'individu statistique correspond à l'unité d'observation. Le caractère statistique est la variable observée sur chaque individu et la modalité est la valeur que peut prendre ce caractère. Les caractères peuvent être qualitatifs (nominaux ou ordinaux) ou quantitatifs (discrets ou continus), et il existe une hiérarchie allant des caractères qualitatifs aux caractères quantitatifs.
9. Pour mesurer l'amplitude on soustrait la valeur minimale de la valeur maximale. Pour la densité on divise l'effectif par la surface.
10. Les formules de sturges et de Yule permettent de déterminer le nombre de classes dans une distribution statistique.
11. L'effectif correspond au nombre d'individus observés. La fréquence se calcule en divisant l'effectif par le total de la population, tandis que la fréquence cumulée est la somme progressive des fréquences. Une distribution statistique est la répartition des individus selon les différentes modalités d'un caractère statistique.

Pour cette première séance : résultats du premier tour de l'élection présidentielle de 2022.. J'ai ensuite d'abord utilisé la fonction `len()` pour calculer le nombre de lignes du tableau. Au début, je pensais que cette fonction permettait également d'obtenir le nombre de colonnes. J'ai compris par la suite que `len()` ne renvoie que le nombre de lignes.

Une étape importante a consisté à identifier la nature statistique des variables. À l'aide d'une boucle, j'ai affiché le type de chaque colonne du DataFrame. Cette étape a été difficile , car

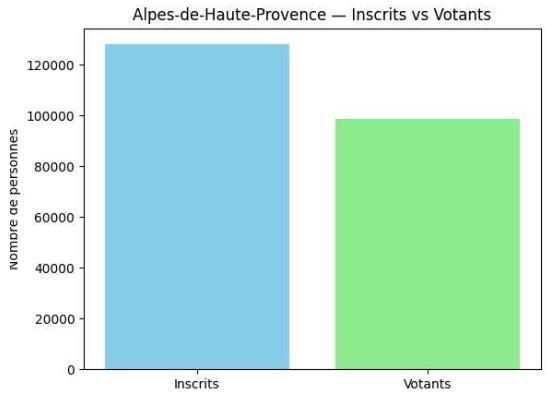
les types retournés par Pandas (object et float64) n'étaient pas immédiatement compréhensibles. Après avoir demandé à plusieurs amis, ils m'ont expliqués que les variables de type float64 correspondaient à des variables quantitatives continues (nombre d'inscrits, de votants, de voix, etc.), tandis que les variables de type object correspondaient à des variables qualitatives, comme les noms de départements ou de candidats.

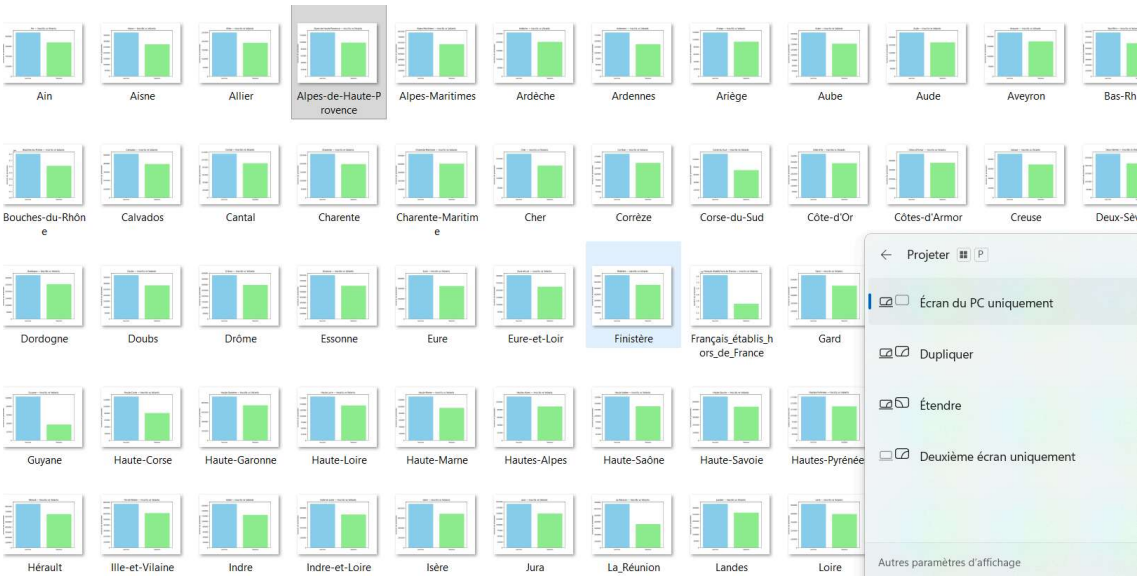
L'utilisation de la méthode `head()` m'a ensuite permis d'afficher les premières lignes du tableau et donc d'identifier clairement les noms des colonnes. Cela m'a facilité la sélection de variables spécifiques, comme la colonne Inscrits, afin de les analyser séparément.

Pour le calcul des effectifs, j'ai utilisé une boucle permettant de sommer les valeurs de chaque colonne. En premier, j'ai tenté de calculer la somme de toutes les colonnes sans distinction, ce qui produisait des résultats incohérents pour les variables qualitatives. J'ai dû ajouter une condition sur le type des colonnes afin de ne sommer que les variables numériques pour que ça marche.

La partie visualisation a été réalisée à l'aide de Matplotlib. J'ai d'abord regroupé les données par département à l'aide de la méthode `groupby()`, puis j'ai créé des diagrammes en barres comparant le nombre d'inscrits et de votants pour chaque département. La compréhension de `groupby()` a été une difficulté, notamment parce qu'il fallait utiliser `sum()` et `reset_index()` pour obtenir un DataFrame exploitable dans la boucle de création des graphiques. Les diagrammes produits permettent de visualiser clairement l'écart entre le nombre d'inscrits et le nombre de votants, et de mettre en évidence l'abstention.

Voici des exemples des images que j'ai obtenues.





Sceance 3

1. Le caractère qualitatif est le plus général, car il permet de décrire des catégories ou des qualités (comme un nom de département ou une profession), alors que le caractère quantitatif est un cas particulier qui mesure des quantités numériques.
2. Les caractères quantitatifs discrets prennent des valeurs isolées et dénombrables (par exemple le nombre d'élèves ou le nombre de gâteaux dans une boulangerie). Les caractères quantitatifs continus peuvent prendre une infinité de valeurs dans un intervalle (la taille ou la température). On les distingue car ils ne se traitent pas de la même manière, notamment pour les calculs et les graphiques (intégrales pour les continus, sommes pour les discrets).
3. a)Il existe plusieurs types de moyenne car toutes les situations ne se résument pas de la même façon. Certaines moyennes sont plus adaptées selon le type de données (arithmétique, harmonique, géométrique) ou selon ce que l'on veut mesurer.

   b)La médiane permet de représenter la valeur centrale d'une série sans être influencée par les valeurs extrêmes. Elle est donc plus représentative que la moyenne dans les distributions dissymétriques.

   c)Le mode peut être calculé lorsqu'une valeur apparaît plus souvent que les autres. Il peut exister pour des variables qualitatives ou quantitatives.

4. La médiale permet de mesurer la concentration d'une variable en tenant compte de la masse totale (par exemple les salaires).
   L'indice de C. Gini permet de mesurer les inégalités de répartition : plus il est élevé, plus la concentration est forte.
5. a)La variance est calculée à partir des carrés des écarts à la moyenne, ce qui évite les compensations entre valeurs positives et négatives.
   L'écart type est utilisé à la place de la variance car il est exprimé dans la même unité que la moyenne, donc plus facile à interpréter.

b)L'étendue mesure l'amplitude des valeurs en calculant la différence entre la valeur maximale et la valeur minimale.

c)Un quantile sert à diviser une série ordonnée en parts égales.
Les quantiles les plus utilisées sont les quartiles (le premier quartile, la médiane (Q2) et le troisième quartile).

d)La boîte de dispersion permet de visualiser la répartition des données, la médiane, la dispersion et les valeurs extrêmes.
Elle s'interprète en observant la taille de la boîte (dispersion) et la position de la médiane (symétrie ou dissymétrie).

6. a)Les moments absolus sont calculés à partir des valeurs brutes et les moments centrés sont calculés par rapport à la moyenne.
Ils sont utilisés pour décrire la forme d'une distribution, comme la dispersion, l'asymétrie et l'aplatissement.

   b)Vérifier la symétrie permet de savoir si la moyenne est représentative de la distribution. On peut la vérifier en comparant la moyenne, la médiane et le mode, ou à l'aide des coefficients d'asymétrie.

Pratique :
J'ai commencé par lire le fichier CSV des résultats des élections, puis j'ai sélectionné uniquement les colonnes quantitatives à l'aide de
:`df_elec.select_dtypes(include=np.number)`
Cela m'a permis de calculer : la moyenne, la médiane, le mode, l'écart-type, l'écart absolu à la moyenne, l'étendue, la distance interquartile, la distance interdécile.

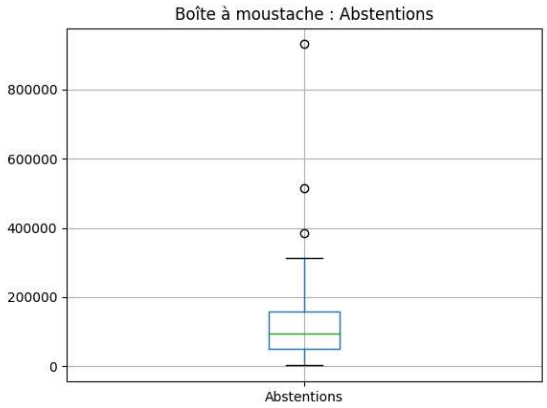
À l'aide de Matplotlib, j'ai créé une boîte à moustaches pour chaque colonne quantitative en utilisant une boucle `for`.
Chaque image est sauvegardée dans le dossier `img/`.
Les boxplots permettent de visualiser rapidement la dispersion des données, la médiane et d'éventuelles valeurs atypiques.

Certaines colonnes montrent une forte dispersion, ce qui indique des différences importantes selon les territoires.

J'ai ensuite travaillé sur le fichier `island-index.csv`.
J'ai sélectionné la colonne Surface (km²) et j'ai utilisé `pd.cut()` pour classer les îles selon les intervalles demandés (0–10, 10–25, …, ≥ 10000 km²).
Après j'ai dû compter le nombre d'îles dans chaque catégorie avec `value_counts()`.

Au début, je me suis trompée dans le chemin du fichier CSV, ce qui provoquait une erreur de type FileNotFoundError.
J'avais aussi fait une erreur sur le calcul de l'écart absolu à la moyenne en oubliant d'utiliser `abs()`, ce qui donnait des résultats bizarres. J'ai dû demander plusieurs fois à chat pour comprendre.
Pour le mode, je n'avais pas compris tout de suite que Pandas pouvait renvoyer plusieurs

valeurs, ce qui causait une erreur d'affectation.
J'ai également eu un problème avec le nom exact de la colonne Surface (km²) car moi j'avais mis km2. J'ai du le copier coller.



exemple d'images :

La boîte à moustache montre une dispersion notable des valeurs, avec une médiane située à l'intérieur de l'intervalle interquartile.La présence de valeurs atypiques indique que certaines observations se distinguent fortement du reste de l'échantillon.

séance 4

1. Le premier point à considérer, c'est le type de variable qu'on étudie. Quand on compte des événements, des occurrences ou des succès, on travaille avec une variable discrète qui ne peut prendre que certaines valeurs bien définies. À l'inverse, si on mesure une durée, une distance ou une surface, on a affaire à une variable continue qui peut prendre n'importe quelle valeur dans un intervalle donné. Ensuite, il faut observer comment se répartissent concrètement nos données. Est-ce que la distribution est plutôt symétrique ? Y a-t-il une asymétrie marquée ? Les valeurs sont-elles concentrées ou dispersées ? Est-ce qu'il y a des valeurs extrêmes ? Ces constatations vont nous orienter vers telle ou telle loi, même si on pourra le vérifier statistiquement par la suite.

   Les caractéristiques statistiques qu'on calcule jouent aussi leur rôle : moyenne, variance, écart-type, asymétrie, aplatissement... Parfois, certaines lois s'imposent naturellement parce qu'elles collent bien à ces paramètres. Par exemple, la loi normale devient pertinente quand la moyenne et la médiane coïncident.
   Il y a aussi la question du nombre de paramètres de la loi. Des lois comme la gamma ou la Weibull, avec leurs multiples paramètres, offrent plus de souplesse pour s'ajuster à des distributions compliquées.
   Tout ça rejoint d'ailleurs ce que disaient Gnedenko et Kolmogorov en 1954 : « l'action collective régularise tout ». Autrement dit, certains phénomènes finissent naturellement par suivre des lois de probabilité bien identifiables.

2. En géographie, on privilégie les lois qui nous aident à comprendre les dynamiques spatiales, démographiques ou territoriales. Ces phénomènes sont rarement homogènes, d'où le besoin de modèles adaptés à cette diversité.La loi normale reste

un outil de référence, particulièrement quand on analyse des phénomènes influencés par de multiples facteurs indépendants, avec des valeurs qui se concentrent autour d'une moyenne. C'est l'héritage de Laplace et Gauss.
Mais ce sont surtout la loi log-normale et la loi de Pareto qui occupent une place importante en géographie. Elles servent notament à analyser la taille des villes, la répartition des populations ou des richesses. La loi de Pareto, formulée initialement par Vilfredo Pareto en 1897 puis dévelopée par Mandelbrot dans les années 1960-1990, capture particulièrement bien les fortes inégalités et la présence de valeurs extrêmes.
Au final, la géographie mobilise surtout des lois capables de rendre compte de distributions déséquilibrées, où quelques valeurs dominent largement. On est loin de la belle symétrie de la courbe normale.

je me suis occupée des distributions discrètes demandées : la loi de Dirac, la loi uniforme discrète, la loi binomiale, la loi de Poisson et la loi de Zipf (approchée pour la Zipf-Mandelbrot).

j'ai eu un problème sur la log-normal, j'avais fait :

```
x = np.linspace(-5, 20, 500)

y_logn = lognorm(s=0.5).pdf(x)
```

j'ai demandé à Chat et il ma répondu :

```
x_logn = np.linspace(0.001, 20, 500)

y_logn = lognorm(s=0.5).pdf(x_logn)

plot_distribution(x_logn, y_logn, "Loi log-normale", "lognormale")
```

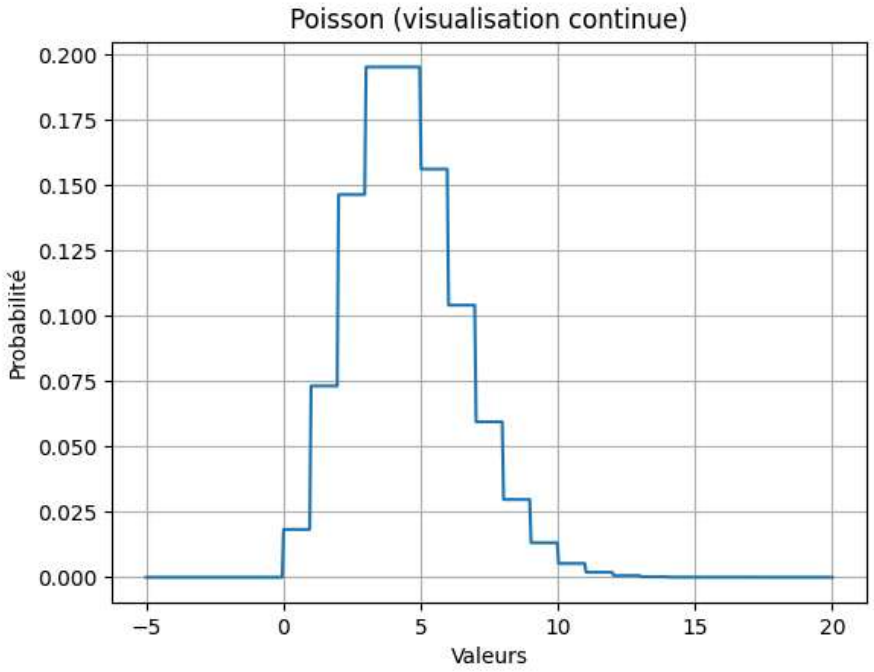
J'ai également rencontré des difficultés sur la loi de Dirac. J'avais mis :

```
distributions = {"Dirac": x_dirac}
```

Mais `x_dirac = [0..10]` ne correspond pas à un échantillon de Dirac (dirac = toutes les valeurs égales à x0).
Il fallait faire :

```
distributions["Dirac"] = np.full(10_000, dirac_x0)
```

J'ai aussi eu un problème au niveau de la loi poisson qui était en réalité discrète et pas continue.
Pour la 2, j'ai écrit une fonction qui calcule la moyenne et l'écart-type à partir d'un ensemble de valeurs. Pour les lois, j'ai généré des échantillons aléatoires à l'aide de la méthode `rvs()` afin d'obtenir des estimations numériques de ces paramètres. Les résultats sont affichés directement dans le terminal pour chaque distribution.



Les graphiques obtenus permettent de visualiser clairement la forme des distributions. Par exemple, la loi de Poisson on voit une forme en escaliers parcque la loi discrète s'appliqe ici sur un axe continu (`floor(x)`). La distribution est asymétrique à droite. Après à peu près 10, la probabilité devient faible : les événements "beaucoup plus grands que λ" deviennent rares.

Sceance 5
ATTENTION : JE N'AI PAS RÉUSSIS À METTRE LES IMAGES SUR GITHUB (a chaque fois que j'uploadais, je recevais un message d'erreur pour une image, j'ai du donc supprimer le dossier image pour déposer ma séance. Il n'y a donc pas d'images.

1. L'échantillonnage consiste à sélectionner une partie de la population afin de tirer des conclusions sur l'ensemble. On n'utilise pas toujours la population entière car cela peut être long voire même impossible (population trop grande, données inexistantes ou indisponibles).
   Les principales méthodes sont : l'échantillonnage aléatoire simple, l'échantillonnage stratifié, l'échantillonnage par quotas, et l'échantillonnage par grappes.
2. Un estimateur est une fonction/formule qui permet d'estimer un paramètre inconnu de la population à partir d'un échantillon.
   Une estimation est la valeur numérique obtenue lorsque l'on applique l'estimateur à un échantillon donné.
3. L'intervalle de fluctuation sert à vérifier si un échantillon est compatible avec un modèle un certain modèle théorique.
   L'intervalle de confiance sert à encadrer un paramètre inconnu de la population avec un certain niveau de confiance (par exemple 95 %).

L’un sert donc pour tester la cohérence d’un échantillon, l’autre pour estimer un paramètre.

4. Un biais correspond à une erreur systématique dans l’estimation d’un paramètre. Un estimateur est biaisé lorsque sa valeur moyenne ne correspond pas à la vraie valeur du paramètre dans la population.
5. Une statistique calculée sur la population entière est appelée un paramètre.
   Avec les big data, on se rapproche des fois à une situation où l’on observe presque toute la population, ce qui réduit le recours à l’inférence, mais pose d’autres problèmes (qualité des données, biais, interprétation).
6. La précision des résultats, la présence ou non d’un biais, la stabilité des estimations est influencée par le choix d’un estimateur. C’est pour cela que son choix est important. Un mauvais estimateur peut conduire à des conclusions incorrectes, même avec beaucoup de données.
7. Les méthodes principales sont : la méthode des moments, la méthode du maximum de vraisemblance, l’estimation bayésienne. Il faut la sélectionner en prenant en compte les hypothèses sur la distribution des données, la taille de l’échantillon et la complexité du modèle.
8. Il existe de nombreux tests statistiques (tests de moyenne, de proportion, tests du χ^2, tests non paramétriques, etc.). Ils servent à prendre une décision à partir de données. Pour créer un test il faut passer par plusieurs étapes tel que : formuler une hypothèse nulle, choisir une statistique de test, fixer un seuil de risque, comparer la statistique observée à une loi théorique.
9. Les critiques sont souvent légitimes, surtout quand on vérifie mal les hypothèses de départ ou qu'on tire des conclusions un peu trop hâtives. Mais la statistique inférentielle reste quand même indispensable pour analyser des phénomènes incertains, du moment qu'on l'utilise avec prudence et un minimum de recul critique.

Vu qu’il est difficile d’obtenir des données “réelles” car elles sont pas trop accessibles. Les données utilisées ici sont créées de simulations aléatoires, ce qui permet de comprendre l’intérêt des méthodes d’échantillonnage, d’estimation et de décision statistique. La population mère étudiée est composée de 2 185 individus, répartis en trois modalités d’opinion :

- Pour: 852
- Contre: 911
- Sans opinion: 422

L’objectif est de montrer comment, sans connaître directement cette population, on peut approcher ses caractéristiques grâce à des échantillons, puis vérifier certaines hypothèses avec des tests statistiques.

J’ai utilisée : `Echantillonnage-100-Echantillons.csv`. Il contient 100 échantillons tirés au hasard sans remise. Chaque ligne correspond à un échantillon, avec le nombre de Pour, contre et sans opinion. j’ai calculé la moyenne de chaque colonne sur les 100 échantillons. Comme ce sont des individus, j’ai arrondi à l’entier.

Après, j'ai calculé les fréquences à partir des moyennes. J'ai fait la somme des trois moyennes, puis j'ai divisé chaque moyenne par cette somme. J'ai fait la même chose pour la population mère à partir des valeurs données dans l'énoncé. Les fréquences ne sont pas exactement les mêmes que celles de la population mère parce que les échantillons sont aléatoires.
Après, j'ai calculé les intervalles de fluctuation à 95 % avec z = 1,96z = 1{,} 96z = 1,96. Ces intervalles permettent de voir si les fréquences réelles de la population peuvent être expliquées par la variabilité due à l'échantillonnage. En général, les valeurs de la population mère tombent dans ces intervalles, ce qui signifie que les échantillons sont cohérents.

Pour l'estimation on essaie d'être dans un cas plus réaliste où on n'a qu'un seul échantillon. J'ai pris le premier échantillon du fichier avec `iloc(0)` et je l'ai transformé en liste.J'ai calculé la taille de l'échantillon (somme de la ligne), puis les fréquences pour chaque opinion. À partir de ces résultats, j'ai calculé les intervalles de confiance à 95 %.
La formule est la même que pour l'intervalle de fluctuation, mais ce n'est pas la même interprétation.
J'ai demandé à chat de me comparer les résultats de ce premier échantillon avec les moyennes sur les 100 échantillons et avec la population mère. Selon l'échantillon choisi, les résultats changent, ce qui montre bien l'incertitude.

Les intervalles ne suffisent pas toujours, donc on utilise des tests statistiques. j'ai utilisé le test de Shapiro-Wilk pour vérifier si une distribution suit une loi normale.
J'ai chargé les deux fichiers `Loi-normale-Test-1.csv` et `Loi-normale-Test-2.csv`. et fais `scipy.stats.shapiro()`.

Si la p-value est supérieure à 0,05, la distribution est normale. Sinon, on rejette la normalité. Une seule des deux distributions vérifie cette condition.

RETOUR SUR LE COURS

Honnêtement, c'était très dur et il y a énormément d'informations à prendre en compte. Lorsque je bloquais au départ, j'essayais de demander à mes amies, mais en fait c'est impossible, car parfois on bloque sur chaque question. ChatGPT a vraiment été d'une grande aide. Il m'expliquait mes erreurs et me réécrivait mes lignes de code, sinon ça aurait été impossible. Je trouve cela dommage. J'aurais préféré avoir un réel accompagnement et faire des exercices beaucoup plus légers, plutôt que d'acquérir une masse d'informations en même temps et de demander à plusieurs intelligences artificielles de m'expliquer.
J'ai essayé de regarder des tutos YouTube, mais c'était beaucoup trop long comparé à l'efficacité de l'IA et au fait qu'elle pouvait me répondre sur des problèmes précis.
Également, cela a énormément ralenti mon ordinateur. Je trouve que c'est un réel problème à gérer avec la fac, l'accès aux ordinateurs. Je pense que cette matière serait mieux en option (moins de personnes dans les cours, donc les professeurs peuvent mieux s'occuper des élèves, et également des élèves qui ont vraiment envie de faire cette matière). Ou sinon, comme je l'ai dit, faire un programme plus léger.

Merci,
Bien cordialement,
Valentina Proust