Campus Ynov Aix-en-Provence - B3 IA/Data

Module « Exploration et analyse de données »

Février 2026

Nom : _____ Prénom : _____

Cochez les affirmations vraies. Il peut y avoir plusieurs réponses correcte pour une question.

Cotation : Le nombre de point que rapporte chaque question est indiqué. Attention lorsque vous cochez plusieurs réponses, une réponse (une case cochée) incorrecte annule une réponse (case cochée) correcte. Des réponses incomplètes (pas toutes les cases possibles cochées) rapportent des points au pro-rata. Interdit de cocher toutes les cases d'une question (malus supplémentaire de -1).

Question 1 - 1 point
Quelle méthode Pandas permet d'obtenir rapidement les statistiques descriptives (moyenne, écart-type, quartiles) d'un DataFrame ?

☐ A. `df.stats()` ☐B. `df.describe()` ☐C. `df.summary()` ☐D. `df.info()`

Réponse B

Question 2 - 1 point
Dans Matplotlib, quelle méthode permet d'afficher un nuage de points ?

☐ A. `plt.plot()` ☐ B. `plt.scatter()` ☐ C. `plt.points()` ☐ D. `plt.cloud()`

Réponse B

Question 3 - 1 point
Dans le processus d'EDA, quelles étapes sont essentielles avant de créer un modèle de Machine Learning ?

☐ A. Gérer les valeurs manquantes
☐ B. Identifier et traiter les outliers
☐ C. Visualiser les distributions et corrélations
☐ D. Déployer le modèle en production

Réponses A, B et C. Le déploiement n'intervient qu'une fois le modèle créé et validé.

Question 4 - 1 point
Dans Seaborn, quel paramètre de la fonction `sns.histplot()` permet d'afficher la courbe de densité estimée ?

☐ A. `density=True`
☐ B. `kde=True`
☐ C. `curve=True`
☐ D. `distribution=True`

Réponse B

Question 5 - 1 point
Quelle est la principale différence entre un barplot et un histogramme ?

☐ A. Le barplot représente des catégories discrètes, l'histogramme des variables continues
☐ B. Le barplot est en couleur, l'histogramme en noir et blanc
☐ C. Le barplot utilise des lignes, l'histogramme des surfaces

☐ D. Il n'y a pas de différence, ce sont des synonymes

Réponse A

Question 6 - 2 point
Parmi les types de graphiques suivants, lesquels sont adaptés pour visualiser la distribution d'une variable quantitative ?

☐ A. Histogramme ☐ B Boxplot (boîte à moustaches)
☐ C. Barplot ☐ D. KDE (Kernel Density Estimation)

Réponses A, B et D. Le barplot est utilisé dans les variables catégorielles.

Question 7 - 2 points
Quelle est la méthode standard pour repérer les outliers en utilisant l'IQR (InterQuartile Range) ?

☐ A. Valeurs en dehors de [Q1 - 1.5×IQR, Q3 + 1.5×IQR]
☐ B. Valeurs en dehors de [Q1 - 3×σ, Q3 + 3×σ]
☐ C. Valeurs en dehors de [moyenne - 2×IQR, moyenne + 2×IQR]
☐ D. Valeurs en dehors de [médiane - IQR, médiane + IQR]

Réponse A (méthode standard, 1,5×IQR visualisé via les boxplots).

Question 8 - 2 points
Concernant les valeurs manquantes (NaN), quelles affirmations sont correctes ?

☐ A. `df.dropna()` supprime les lignes contenant des valeurs manquantes
☐ B. Une chaîne vide `''` est automatiquement considérée comme NaN par Pandas
☐ C. On peut remplacer les NaN par la moyenne ou la médiane de la variable
☐ D. Les valeurs `np.inf` sont automatiquement détectées comme NaN

Réponses : A et C. Chaînes vides et np.inf (et tout autre encodage) ne sont pas reconnues automatiquement, il faut les traiter spécifiquement.

Question 9 - 2 points
Dans Seaborn, quelle fonction permet de visualiser les relations entre toutes les paires de variables d'un DataFrame ?

☐ A. `sns.relplot()` ☐ B. `sns.pairplot()` ☐ C. `sns.jointplot()` ☐ D. `sns.corrplot()`

`Réponse B`

Question 10 - 2 points
Que représente le coefficient de corrélation de Pearson ?

☐ A. La différence entre deux moyennes
☐ B. L'intensité et la direction de la relation linéaire entre deux variables
☐ C. La probabilité qu'une relation existe entre deux variables
☐ D. La variance d'une variable par rapport à une autre

Réponse B (voir déf.)

Question 11 - 2 points
Pour créer des subplots avec Matplotlib, quelles approches sont valides ?

☐ A. Utiliser `plt.subplot(nrows, ncols, index)`
☐ B. Utiliser `fig, axs = plt.subplots(nrows, ncols)`

☐ C. Utiliser `plt.GridSpec()`

☐ D. Utiliser `plt.multiplot()`

Réponses A, B et C (cf. doc). Multiplot() n'existe pas.

Question 12 - 2 points
Selon le Théorème Central Limite (TCL), que se passe-t-il lorsque la taille de l'échantillon augmente ?

☐ A. La moyenne d'échantillon s'éloigne de la moyenne de la population
☐ B. La distribution des moyennes d'échantillons tend vers une loi normale
☐ C. L'écart-type de l'échantillon augmente proportionnellement
☐ D. Les outliers deviennent plus fréquents

Réponse B

Question 13 - 2 points
Quelle est la relation entre l'erreur standard (SE) et la taille de l'échantillon (n) ?

☐ A. $SE = \sigma \times \sqrt{n}$ ☐ B. $SE = \sigma / n$ ☐ C. $SE = \sigma / \sqrt{n}$ ☐ D. $SE = \sigma \times n$

Réponse C

Question 14 - 2 points
Dans un test d'hypothèse, quelles affirmations sont correctes ?

☐ A. L'hypothèse nulle (H0) représente l'absence d'effet ou de différence
☐ B. Si p-value < α, on rejette H0
☐ C. Une p-value de 0.03 avec α=0.05 est statistiquement significative
☐ D. Rejeter H0 prouve que l'hypothèse alternative est vraie

Réponses A, B et C (C est juste une spécification de B). Rejeter H0 signifie juste que nos données ne nous permettent pas de trancher avec une probabilité coorespondant à notre critère de décision (α).

Question 15 - 2 points
Pourquoi utilise-t-on la distribution de Student plutôt que la loi normale dans certains cas ?

☐ A. Quand la taille de l'échantillon est grande (n > 100)
☐ B. Quand la taille de l'échantillon est petite (n < 30) et σ inconnu
☐ C. Quand les données ne sont pas corrélées
☐ D. Quand on veut visualiser des distributions

Réponse B

Question 16 - 3 points
Vous observez une forte corrélation (r = 0.85) entre deux variables X et Y. Quelles conclusions pouvez-vous tirer ?

☐ A. Il existe une relation linéaire forte entre X et Y
☐ B. X cause nécessairement Y
☐ C. Des outliers pourraient influencer cette corrélation
☐ D. Une troisième variable cachée pourrait expliquer cette corrélation

Réponse A, C et D. Corrélation n'est pas causalité. Une corrélation est sensible aux outliers et n'exclut pas l'intervention d'une troisième variable.

Question 17 - 3 points
Dans le contexte de la "winsorisation" des outliers, quelle affirmation est correcte ?

☐ A. On supprime complètement les outliers du dataset
☐ B. On remplace les outliers par des valeurs seuils (plafonds) en maintenant la taille de l'échantillon
☐ C. On déséquilibre un peu la symétrie de la distribution
☐ D. On multiplie les outliers par un facteur de correction

Réponse B. Il faut que la transformation respecte la symétrie.

Question 18 - 3 points
Un intervalle de confiance à 95% de [178.74 cm ; 181.26 cm] pour la taille moyenne d'une population signifie : (Plusieurs réponses possibles)

☐ A. Il y a 95% de chances que la vraie moyenne de la population soit dans cet intervalle
☐ B. Si on répétait l'échantillonnage 100 fois, environ 95 intervalles contiendraient la vraie moyenne
☐ C. On a 95% de confiance dans la méthode de calcul de l'intervalle
☐ D. La taille moyenne de la population est forcément de 180 cm

Réponse B et C. On ne peut pas parler de « chance » que la moyenne soit égale à ceci ou cela (elle a une vraie valeur fixée, mais qu'on ne connaît pas, et qui peut être n'importe quelle valeur en réalité).

Question 19 - 3 points
Vous testez si un nouveau moteur réduit les émissions de CO2. Avec une moyenne de 2900g/h (vs 3000g/h pour l'ancien) et une p-value de 0.045 (α=0.05), quelle est la meilleure interprétation ?

☐ A. Le nouveau moteur réduit définitivement les émissions de 100g/h
☐ B. Il y a 4.5% de chances que l'hypothèse nulle soit vraie
☐ C. On rejette H0 : les données suggèrent une différence significative, mais ne prouvent pas la causalité
☐ D. On ne rejette pas H0 car la différence est trop faible

Réponse C

Question 20 - 3 points
En cartographie géospatiale avec Geopandas, quelles considérations sont importantes ? (Plusieurs réponses possibles)

☐ A. Le système de projection (CRS) doit être cohérent entre les différentes sources de données
☐ B. La projection de Mercator conserve les aires, idéale pour comparer les superficies
☐ C. Pour la France métropolitaine, Lambert 93 (EPSG:2154) est le système officiel
☐ D. Les coordonnées en WGS 84 (EPSG:4326) sont en longitude/latitude

Réponse A, C (conique pour la Métropole) et D (GPS). Mercator conserve les angles (utile pour la navigation).