

Quel est l'output de cette cellule ?	<code><class 'list'> <class 'str'></code>
Quel est l'output de cette cellule ?	<code>[('a', 'f'), ('b', 'g')]</code>
Quel est l'output de cette cellule ?	<code>Index(['age_of_subject', 'name_of_subject', 'time_since_last'], dtype='object')</code>
Quel est l'output de cette cellule ?	<code>{'hello': 2, 'how': 1, 'are': 1, 'you': 2, 'fine': 1, 'and': 1}</code>
A quoi est égale la liste l ?	<code>[0, 3, 6, 10, 8, 9]</code>
Si s est un objet pandas.Series, quel test d'égalité renvoie le booléen True ?	<code>list(set(s)) == s.unique().tolist()</code>
Plusieurs réponses sont justes à cette question. Quelles lignes de codes renvoient le nombre de lignes de df ?	<code>df.shape[0], len(df), len(df.index.unique())</code>
Qu'affiche la console ?	<code>kayak srola</code>
Quelles sont les colonnes de la DataFrame df ?	<code>"A" et "B"</code>
Quelles sont les les indexes de la DataFrame df ?	<code>"a" et "b"</code>
Qu'est-ce qu'un nombre magique en python ?	un nombre qui n'est pas contenu dans une variable et qui n'est pas clair sur son impact dans le code
SQL :	veut dire Structured Query Language, un langage de programmation
Sur un diagramme en boîte représentant une série statistique de N données,	Les extrémités des « moustaches » correspondent aux valeurs du 1er et du 9ème décile ou aux valeurs extrêmes de la série
La loi binomiale B(n, p) modélise :	Le nombre de succès obtenus lors de la répétition indépendante de n expériences aléatoires identiques, chacune ayant deux issues possibles : succès avec probabilité p, échec avec probabilité 1 - p.
Dans une série statistique de N expériences dont l'issue est une variable continue, l'écart-type mesure :	La dispersion de la série des valeurs autour de la moyenne

Les variables quantitatives peuvent être discrètes ou continues ?	Vrai
A propos des affirmations suivantes, indiquez celles qui sont exactes :	La profession est une variable catégorielle, La natrémie (concentration de sodium (Na) dans le plasma) est une variable continue, Le degré d'intensité de la douleur est une variable continue discrète
L'open data sont des données qui sont :	Libre de droit
A quoi sert la data viz ?	Communiquer efficacement, Raconter une histoire, Donner du sens à la donnée
Je veux comparer des données entre de nombreux éléments, quel type de graphique est conseillé ?	Le nuage de points (scatter plot)
Le Data Storytelling	Présente les résultats d'une réflexion déjà aboutie
Je veux fournir à mon client de quoi analyser de multiples données sous plusieurs formes avec Tableau, que vais-je faire ?	Un tableau de bord, Une histoire
On ne peut pas filtrer un graphique à partir d'un autre graphique dans tableau.	VRAI
Quelle est la bonne méthode ?	<code>pd.Series.value_counts()</code>
Laquelle des propositions suivantes n'est pas une loi de probabilité :	Loi Binormale
J'ai un dataset avec des données discrètes que je veux visualiser :	je fais un bar plot / scatter plot
Vrai ou faux ? Deux variables peuvent avoir un score de corrélation de zéro et pourtant avoir un lien.	Vrai
Laquelle des métriques suivantes n'est pas une métrique d'évaluation de classifieur :	p-value
L'accuracy est :	le taux de vraies prédictions sur toutes les prédictions, comprise entre 0 et 1
Que signifie ROC de la courbe ROC :	Receiver Operationnal Characteristic
Quel est le danger des arbres de décisions :	ils sont prones à l'overfitting si on ne les hypertune pas correctement

Qu'est-ce qui est vrai au sujet de la PCA ?	On perd en interprétabilité après la PCA, La PCA permet de réduire la dimensionalité des données
Imagine que tu développes une pipeline Machine Learning qui prédit le nombre de vues sur un article d'un journal. L'analyse se fonde sur des features comme l'auteur, le nombre d'articles écrit par cet auteur dans le passé, etc. Quelle métrique utiliserais tu ? 1) Mean Square Error, 2) Accuracy, 3) F1 Score	Seulement la 1
Qu'est ce qui est vrai à partir de ce tableau ?	L'accuracy vaut ~0.91, Le taux de vrai positif est de ~0.95
Laquelle des options suivantes est vraie pour une cross-validation avec $CV = K$?	Une augmentation de K entraînera un temps plus long nécessaire pour valider le résultat., Des valeurs plus élevées de K entraîneront une confiance accrue sur le résultat de la validation croisée par rapport à une valeur inférieure de K .
L'ajout d'une feature non importante à un modèle de régression linéaire entraîne	Une augmentation de R -carré
Disons que vous travaillez avec une ou plusieurs variables catégorielles et que vous n'avez pas examiné la distribution de la variable catégorielle dans les données de test. Vous souhaitez appliquer un one-hot encoding (OHE, NDJM: c-à-d une dummification) sur la ou les variables catégorielles. Quels sont les défis que vous pouvez rencontrer si vous avez appliqué une OHE sur le train ?	Les valeurs des variables catégorielles ne sont pas toutes présentes dans le jeu de données de train., La distribution de fréquence des variables est différente dans le train par rapport à l'ensemble de données de test.

<p>La cross-validation est une étape importante pour la recherche des hyperparamètres. Supposons que vous réglez un hyper-paramètre «max_depth» de XGBoost en le sélectionnant parmi 10 valeurs de profondeur différentes (les valeurs sont supérieures à 2) pour un modèle basé sur une arborescence utilisant une cross-val avec CV=5. Le temps pris par le training (sur un modèle avec max_depth = 2) sur un des plis "train" est de 10 secondes et 2 secondes pour la prédiction sur le pli "val" restant. Laquelle des options suivantes est vraie pour le temps d'exécution total pour une cross-val avec cv=5 avec 10 valeurs différentes de «max_depth»?</p>	<p>Plus de 600s</p>
<p>Vous souhaitez montrer la répartition des utilisateurs récurrents totaux de votre application entre les utilisateurs ayant un compte gratuit (70 %) et ceux ayant un compte payant (30 %). Quel type de graphique allez-vous utiliser ?</p>	<p>Un diagramme circulaire</p>