Ce document a été mis en ligne par l'organisme FormaV®.

Toute reproduction, représentation ou diffusion, même partielle, sans autorisation préalable, est strictement interdite.

Pour en savoir plus sur nos formations disponibles, veuillez visiter :
formav.co/explorer

BACHELOR UNIVERSITAIRE DE TECHNOLOGIE

SCIENCE DES DONNÉES

ÉVALUATION DE COMPÉTENCE C6 : METTRE EN ŒUVRE UNE ANALYSE EXPLORATOIRE

SESSION 2026

Durée : 2 h 30

SPÉCIALITÉ / UNITÉ	COEFF
SCIENCE DES DONNÉES	3,0 (INDICATIF)

Moyens de calculs et matériel autorisés :

L'usage de calculatrice avec mode examen actif est autorisé. L'usage de calculatrice sans mémoire, « type collège », est autorisé.

Aucun document ou matériel autorisé en dehors de ceux fournis avec le sujet. Une calculatrice autorisée selon la règle ci-dessus est admise.

Dès que le sujet vous est remis, assurez-vous qu'il est complet.

Ce sujet comporte 10 pages numérotées de 1/10 à 10/10.

Les documents réponses en fin de sujet sont à rendre avec la copie.

DOSSIER SUJET

MISE EN SITUATION PROFESSIONNELLE ET CAHIER DES CHARGES : OBSERVATOIRE DES FORMATIONS DE L'INSTITUT DATA LOIRE, TOURS

L'Observatoire des formations de l'Institut Data Loire souhaite étudier les résultats d'un groupe de dix étudiants à une évaluation de statistiques. Vous intervenez comme technicienne ou technicien en science des données. À partir du jeu de données fourni dans le dossier technique, vous devez contrôler les données, en décrire les distributions, explorer la relation entre le temps déclaré de révision et le score obtenu, puis communiquer des conclusions prudentes et utiles au responsable de formation.

Le score est exprimé sur 100. Le temps de révision est exprimé en heures. Les identifiants sont pseudonymisés. Les données sont fictives et construites pour cette étude. Aucune conclusion causale ne peut être tirée de ces seules observations.

Les réponses doivent distinguer les constats calculés, les éventuelles anomalies à vérifier et les limites de l'analyse. Les arrondis sont attendus à deux décimales lorsque cela est utile.

PARTIE 1 : Cadrage de l'analyse et contrôle des données (5 points)

Définition du besoin, qualité des données et cadre de traitement

Vous préparez l'analyse exploratoire destinée au responsable de formation. Les variables, les unités et les règles de contrôle figurent dans les documents DT 1 et DT 2. Le responsable demande une description fidèle du groupe, sans extrapolation à d'autres promotions.

Variables disponibles pour l'étude

Variable	Type	Unité ou modalités	Usage prévu
temps_revision	Quantitative	Heures	Étudier la distribution et la relation avec le score
score	Quantitative	Points sur 100	Décrire les résultats
absences	Quantitative discrète	Nombre de séances	Décrire les absences
modalite	Qualitative nominale	Présentiel ou distance	Décrire la composition du groupe

Synthèse des variables décrites dans le document DT 2.

1.1 (1,5 point) Formulez l'objectif statistique de l'étude et identifiez la population effectivement observée ainsi que l'unité statistique. Précisez pourquoi les résultats ne permettent pas, à eux seuls, de généraliser à l'ensemble des étudiants de l'établissement.

1.2 (2 points) À partir des règles du DT 2, vérifiez les contrôles de structure et de domaine à appliquer au fichier. Indiquez le résultat de ces contrôles sur les dix lignes fournies : nombre d'observations, valeurs manquantes, identifiants dupliqués et valeurs hors domaine pour le score et les absences.

1.3 (1,5 point) Indiquez deux précautions à prendre avant de transmettre les résultats du tableau de bord à un destinataire. Votre réponse doit prendre en compte la pseudonymisation et le risque d'interprétation abusive d'une association statistique.

PARTIE 2 : Description statistique et choix des représentations (6 points)

Calcul des indicateurs et représentation des distributions

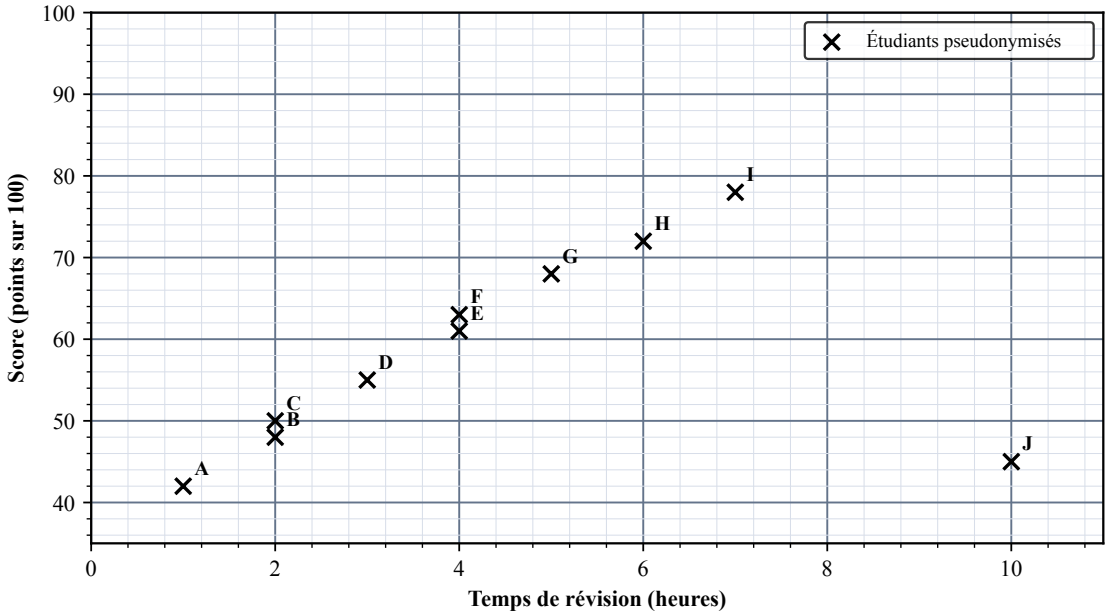
Utilisez les dix observations du DT 1. Pour les quartiles, appliquez la convention fixée au DT 2 : trier les valeurs, partager la série en deux moitiés de même effectif, puis prendre la médiane de chaque moitié. Pour l'écart interquartile, utilisez la relation $IQR = Q_3 - Q_1$. Pour les clôtures de Tukey, utilisez $[Q_1 - 1{,}5IQR ; Q_3 + 1{,}5IQR]$.

Résumé statistique à compléter

Indicateur	Valeur
Effectif	.....
Minimum	.....
Médiane	.....
Q1	.....
Q3	.....
Maximum	.....
Moyenne	.....
Écart interquartile	.....

À compléter pour la variable score à partir du DT 1.

Nuage de points : temps de révision et score



Les coordonnées sont issues du DT 1. Utiliser le nuage pour décrire la relation entre les deux variables.

2.1 (2,5 points) Complétez le tableau de résumé statistique pour la variable score. Présentez les étapes essentielles du calcul de la moyenne, de la médiane, de Q_1, de Q_3 et de l'écart interquartile.

2.2 (1,5 point) Déterminez les clôtures de Tukey pour les scores. Indiquez si une observation est signalée par cette règle et justifiez votre réponse. Une observation non signalée par cette règle ne doit pas être considérée automatiquement comme exempte de toute erreur.

2.3 (2 points) Décrivez la forme générale du nuage de points et la situation de l'observation J par rapport aux autres observations. Proposez une vérification à effectuer avant de qualifier cette observation d'erreur ou de la supprimer. Précisez si le nuage établit un lien de causalité.

PARTIE 3 : Analyse exploratoire, diagnostic et communication (5 points)

Choix des indicateurs, interprétation et préparation d'une restitution

Le responsable souhaite une note de synthèse courte à partir des résultats obtenus. Le document DT 3 rappelle les principes de choix des indicateurs et de communication. Les constats doivent porter uniquement sur les dix observations du jeu fourni.

Grille de diagnostic exploratoire

Axe de diagnostic	Constat appuyé par une donnée	Limite ou vérification nécessaire
Distribution des scores	.....	.....
Temps de révision et score	.....	.....
Qualité des données	.....	.....

Compléter la grille DR 1 à partir des données du DT 1 et des calculs réalisés.

3.1 (1,5 point) Choisissez une représentation adaptée pour décrire la distribution des scores et une représentation adaptée pour étudier la relation entre le temps de révision et le score. Justifiez brièvement chaque choix en fonction du type des variables.

3.2 (2 points) Complétez la grille de diagnostic DR 1 avec un constat chiffré ou directement vérifiable pour chaque axe et une limite ou vérification pertinente. Pour l'axe « temps de révision et score », tenez compte de l'observation J.

3.3 (1,5 point) Rédigez une conclusion professionnelle de trois à cinq phrases destinée au responsable. Elle doit comporter un résultat descriptif, une réserve sur l'observation J et une limite d'interprétation liée à l'effectif ou au caractère observationnel des données.

PARTIE 4 : Préparation du traitement et prévention des risques (4 points)

Protocole reproductible, protection des données et contrôle de la restitution

Avant de produire un tableau de bord, vous devez formaliser un protocole de traitement reproductible et prévenir les risques liés à l'usage des données. Les prescriptions de travail figurent dans le DT 3.

Protocole opératoire à compléter

Étape	Opération à effectuer	Contrôle ou résultat attendu
1	.....	.....
2	.....	.....
3	.....	.....
4	.....	.....

Compléter le document DR 2 dans l'ordre prévu.

4.1 (2 points) Complétez le protocole opératoire DR 2 en quatre étapes ordonnées, depuis la réception du fichier jusqu'à la validation de la restitution. Pour chaque étape, indiquez au moins un contrôle observable permettant de la valider.

4.2 (2 points) Identifiez deux risques professionnels ou organisationnels liés au traitement ou à la diffusion de ces données, puis proposez une mesure de prévention adaptée à chacun. Une mesure doit concerner la confidentialité ou la minimisation des données et l'autre la qualité ou l'interprétation des résultats.

DOSSIER TECHNIQUE ET RESSOURCES

Documents ressources nécessaires à la résolution du sujet

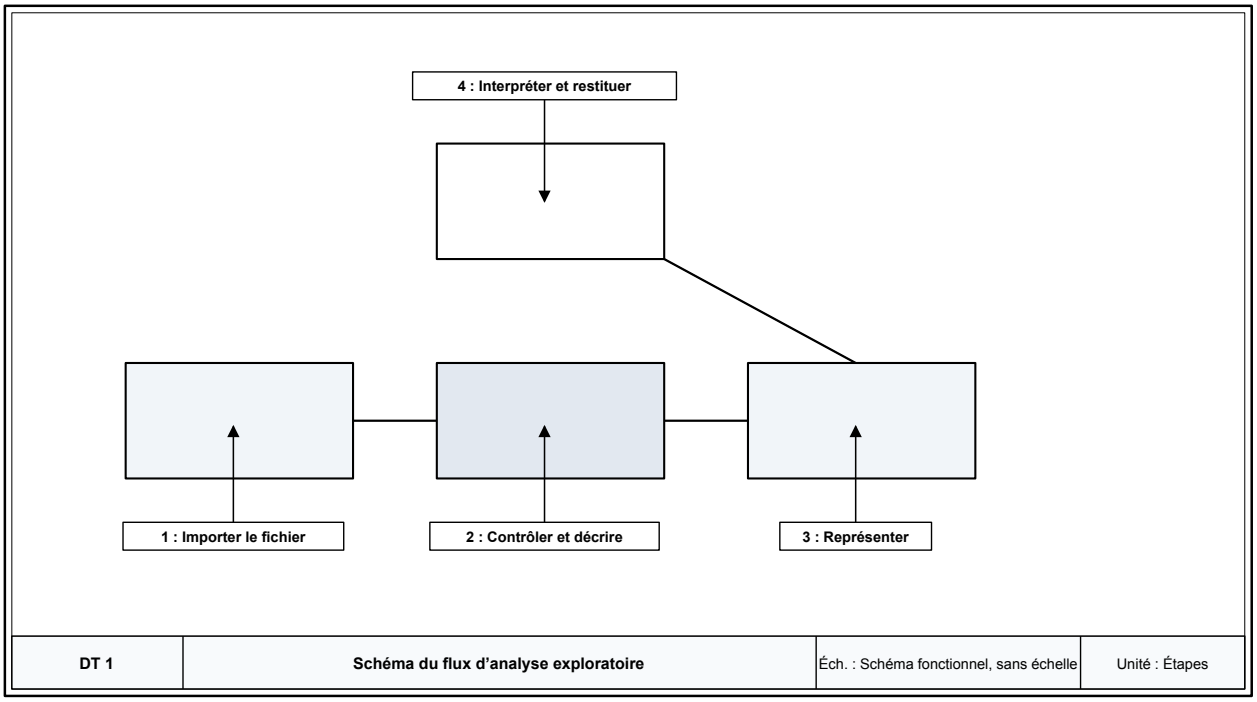
DT 1 : EXTRAIT DU JEU DE DONNÉES PSEUDONYMISÉ

Les dix lignes ci-dessous constituent l'intégralité du groupe observé pour cette étude. Les identifiants A à J sont des codes pseudonymisés. Le score est compris théoriquement entre 0 et 100 inclus. Le nombre d'absences est un entier positif ou nul. Le temps de révision est renseigné en heures. Les données présentées sont fictives.

Données observées

id	temps_revision (h)	score (/100)	absences (séances)	modalite
A	1	42	0	Présentiel
B	2	48	1	Présentiel
C	2	50	0	Distance
D	3	55	2	Présentiel
E	4	61	1	Distance
F	4	63	0	Distance
G	5	68	1	Présentiel
H	6	72	0	Distance
I	7	78	1	Présentiel
J	10	45	0	Distance

Source : extraction fictive de l'Observatoire des formations de l'Institut Data Loire.



DT 2 : DICTIONNAIRE DES VARIABLES ET RÈGLES DE CONTRÔLE

Le fichier attendu comporte une ligne par étudiant et une colonne par variable. L'identifiant doit être unique et non vide. Les valeurs de score doivent appartenir à l'intervalle $[0; 100]$. Le nombre d'absences doit être un entier positif ou nul. Le temps de révision est une durée positive ou nulle, exprimée en heures. Les modalités autorisées pour la variable modalite sont « Présentiel » et « Distance ».

Convention de calcul des quartiles pour cette épreuve : ordonner les valeurs ; pour un effectif pair, partager la série ordonnée en deux moitiés de même taille ; calculer la médiane de chaque moitié pour obtenir Q_1 et Q_3. Pour les dix valeurs de score, la médiane générale est la moyenne des deux valeurs centrales.

Dictionnaire des variables

Variable	Nature	Domaine attendu	Contrôle
id	Identifiant	Non vide et unique	Compter les identifiants distincts et repérer les doublons
temps_revision	Quantitative continue	Valeur numérique positive ou nulle, en heures	Contrôler le type, les valeurs manquantes et la cohérence d'unité
score	Quantitative	Valeur numérique de 0 à 100 inclus	Contrôler les valeurs manquantes et les bornes
absences	Quantitative discrète	Entier positif ou nul	Contrôler le type et les valeurs négatives
modalite	Qualitative nominale	Présentiel ou Distance	Contrôler les valeurs manquantes et les modalités inattendues

Typage, domaine attendu et contrôle à effectuer.

DT 3 : REPÈRES MÉTHODOLOGIQUES, RESTITUTION ET PROTECTION DES DONNÉES

Pour une variable quantitative, des indicateurs de position et de dispersion peuvent être mobilisés. La moyenne d'une série $x_1, \ldots, x_n$ est $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. La médiane partage la série ordonnée en deux parties de même effectif. L'écart interquartile est $IQR = Q_3 - Q_1$. Le repérage de Tukey utilise les bornes $Q_1 - 1{,}5IQR$ et $Q_3 + 1{,}5IQR$. Cette règle signale des valeurs à examiner ; elle ne prouve pas qu'elles sont erronées.

Un histogramme ou une boîte à moustaches convient à la description d'une variable quantitative. Un nuage de points convient à l'étude descriptive de deux variables quantitatives. Une association observée ne démontre pas un effet causal. Une observation atypique doit être vérifiée avec la source avant toute correction ou exclusion, et les résultats doivent être présentés avec les limites liées à l'effectif et au mode de collecte.

Le traitement doit respecter les principes de minimisation des données, de limitation des accès, de sécurisation des fichiers et de diffusion de résultats agrégés lorsque cela suffit à répondre au besoin. Les identifiants pseudonymisés restent des données à protéger. Avant diffusion, vérifier les destinataires, les unités, les intitulés, les arrondis, les effectifs et la cohérence des graphiques avec les calculs.

Règles de choix et de contrôle avant restitution

Besoin	Outil possible	Point de vigilance
Décrire une variable quantitative	Moyenne, médiane, quantiles, dispersion, histogramme ou boîte à moustaches	Vérifier l'unité, l'effectif et les valeurs extrêmes
Étudier deux variables quantitatives	Nuage de points	Ne pas conclure à une causalité à partir d'une association
Préparer une diffusion	Tableau ou graphique avec titre, unités et effectif	Limiter les données diffusées et sécuriser les accès

Repères méthodologiques applicables au travail demandé.

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 1 : Grille de diagnostic à rendre avec la copie, questions 3.2 et 3.3 **À RENDRE AVEC LA COPIE**

Compléter directement les cases avec des constats issus du DT 1 et des calculs de la partie 2. Indiquer une limite ou une vérification concrète pour chaque axe. Ne pas remplacer une vérification par une supposition.

Grille de diagnostic exploratoire à compléter

Axe de diagnostic	Constat appuyé par une donnée	Limite ou vérification nécessaire
Distribution des scores	.....	.....
Temps de révision et score	.....	.....
Qualité des données	.....	.....

Un constat et une limite ou vérification sont attendus dans chaque ligne.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

.....

.....

.....

.....

.....

.....

.....

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 2 : Protocole opératoire à rendre avec la copie, question 4.1 **À RENDRE AVEC LA COPIE**

Compléter les quatre étapes dans l'ordre chronologique. Pour chaque opération, indiquer un contrôle ou un résultat observable. Le protocole doit permettre à une autre personne de reproduire l'analyse.

Protocole de traitement et de restitution

Étape	Opération à effectuer	Contrôle ou résultat attendu
1	.....	.....
2	.....	.....
3	.....	.....
4	.....	.....

Quatre étapes attendues, de la réception du fichier à la validation finale.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

Copyright © 2027 FormaV®. Tous droits réservés.

FormaV® est une marque déposée auprès de l'INPI sous le numéro 5227897.

Ce document a été élaboré par FormaV® avec le plus grand soin afin d'accompagner chaque apprenant vers la réussite de ses examens. Son contenu (textes, graphiques, méthodologies, tableaux, exercices, concepts, mises en forme) constitue une œuvre protégée par le droit d'auteur.

Toute copie, partage, reproduction, diffusion ou mise à disposition, même partielle, gratuite ou payante, est strictement interdite sans accord préalable et écrit de FormaV®, conformément aux articles L.111-1 et suivants du Code de la propriété intellectuelle. Dans une logique anti-plagiat, FormaV® se réserve le droit de vérifier toute utilisation illicite, y compris sur les plateformes en ligne ou sites tiers.

En utilisant ce document, vous vous engagez à respecter ces règles et à préserver l'intégrité du travail fourni. La consultation de ce document est strictement personnelle.

Merci de respecter le travail accompli afin de permettre la création continue de ressources pédagogiques fiables et accessibles.