Formation Data Analyst

Objectifs : Acquérir des compétences appliquées relatives à l'analyse de données aussi bien quantitatives que qualitatives.
A l'issue de ce cycle de formation Data Analyst, vous saurez résumer l'information pertinente présente dans un fichier de données et en extraire celle utile à la prise de décision.

Durée : 14 jour(s) (98 heures)

Public : Toute personne désirant exploiter efficacement les données mises à sa disposition.

Pré-requis : Une formation scientifique de niveau BAC+2 ou supérieur est recommandée pour suivre cette formation dans de bonnes conditions.
La pratique d'un langage de programmation est un plus.

Méthode pédagogique : Pédagogie active mêlant exposés, exercices et applications pratiques dans le logiciel R.

Certification Data Value vous propose en option la certification « Valoriser ses données : collecter, prétraiter, analyser et interpréter ». Cette certification s'appuie sur un QCM d'évaluation de niveau ainsi que la réalisation d'une étude de cas afin de valider les connaissances et compétences acquises au cours du cycle de formation Data Analyst

Modalités d'évaluation : Un formulaire d'auto-évaluation proposé en amont de la formation nous permettra d'évaluer votre niveau et de recueillir vos attentes. Ce même formulaire soumis en aval de la formation fournira une appréciation de votre progression.
Des exercices pratiques seront proposés à la fin de chaque séquence pédagogique pour l'évaluation des acquis.
En fin de formation, vous serez amené(e) à renseigner un questionnaire d'évaluation à chaud.
Une attestation de formation vous sera adressée à l'issue de la session.
Trois mois après votre formation, vous recevrez par email un formulaire d'évaluation à froid sur l'utilisation des acquis de la formation.

Accessibilité : Vous souhaitez suivre notre formation Formation par ville et êtes en situation de handicap ? Merci de nous contacter afin que nous puissions envisager les adaptations nécessaires et vous garantir de bonnes conditions d'apprentissage

Tarifs :
- Présentiel : 6300 € HT
- Distanciel : 5600 € HT
(-10% pour 2 inscrits, -20% dès 3 inscrits)

Option(s) :
- Certification valoriser ses données : collecter, prétraiter, analyser et interpréter : 400 € HT
- Forfait déjeuners : 280 € HT

Nos prochaines sessions

Distance

- du 22 au 23 mai 2025
  du 26 au 28 mai 2025
  du 4 au 6 juin 2025
  du 18 au 20 juin 2025
  du 23 au 25 juin 2025
- du 16 au 17 octobre 2025
  du 3 au 5 novembre 2025
  du 17 au 19 novembre 2025
  du 24 au 26 novembre 2025
  du 1 au 3 décembre 2025

Lyon

- du 17 au 18 novembre 2025
  du 19 au 21 novembre 2025
  du 26 au 28 novembre 2025
  du 3 au 5 décembre 2025
  du 10 au 12 décembre 2025

Nantes

- du 29 au 30 septembre 2025
  du 1 au 3 octobre 2025
  du 8 au 10 octobre 2025
  du 12 au 14 novembre 2025
  du 17 au 19 novembre 2025

Paris

- du 12 au 13 mai 2025
  du 16 au 18 mai 2025
  du 26 au 28 mai 2025
  du 4 au 6 juin 2025
  du 16 au 18 juin 2025
- du 23 au 24 octobre 2025
  du 5 au 7 novembre 2025
  du 17 au 19 novembre 2025
  du 26 au 28 novembre 2025
  du 8 au 10 décembre 2025

Toulouse

- du 24 au 25 avril 2025
  du 28 au 30 avril 2025
  du 14 au 16 mai 2025
  du 26 au 28 mai 2025
  du 4 au 6 juin 2025
- du 8 au 9 septembre 2025
  du 10 au 12 septembre 2025
  du 17 au 19 septembre 2025
  du 1 au 3 octobre 2025
  du 8 au 10 octobre 2025

Programme :

Module 1 : Statistique descriptive (exploratoire) : savoir décrire des observations

Objectifs : Apprendre à décrire des jeux de données à l'aide de résumés numériques et de représentations graphiques

- Utilité des statistiques

- Recueillir les observations : observer et mesurer
- Précautions à prendre
- Organiser les informations : classements, tableaux
- Vocabulaire de base : variables, individus, échantillon, population

- La synthèse numérique

- Caractéristiques de la tendance centrale d'une distribution : moyenne, médiane, mode
- Caractéristiques de la dispersion d'une distribution : variance, écart type, fractiles

- La synthèse graphique

- Représenter les phénomènes pour les comprendre : l'histogramme, le diagramme en bâtons, le diagramme circulaire
- Identifier les phénomènes les plus importants : le diagramme de Pareto

- La statistique bivariée

- 2 variables quantitatives
    - Représentation graphique : le nuage de points
    - Interpréter le coefficient de corrélation
    - Ajuster une droite de régression
- 2 variables ordinales
    - Coefficient de corrélation sur les rangs de Spearmann
- 2 variables qualitatives
    - Représentation graphique : diagramme en barre des profils
    - Table de contingence
    - La statistique du Khi-2
- 1 variable quantitative et 1 variable qualitative
    - Représentation graphique : les boîtes à moustaches parallèles
    - Le rapport de corrélation

- Repérer et gérer les valeurs particulières (aberrantes, influentes)

Module 2 : Statistique décisionnelle (inférentielle) : savoir décider au vu des observations

Objectifs :Découvrir la statistique inférentielle permettant de généraliser à partir d'un échantillon (connaissance partielle d'un phénomène) afin de prendre une décision en sachant évaluer les deux types de risques associés. Maîtrise opérationnelle des notions d'estimation d'un paramètre, d'intervalle de confiance, de tests d'hypothèse, ...

- Rappels de statistique et probabilité

- Variables, individus, échantillons
- Tendance centrale et dispersion
- Calcul de probabilité et lois usuelles
- Tendance vers la loi normale ou « loi des grands nombres »

- Estimation à partir d'un échantillon

- Estimation d'une moyenne, d'une proportion
- Distribution des statistiques calculées sur échantillon
- Estimation par intervalle de confiance

- Évaluation des risques en fonction de la taille de l'échantillon

- Les tests d'hypothèses

- Mécanisme de la procédure de décision dite « test d'hypothèse »
- Mesure des risques d'erreurs associés à la décision. L'hypothèse « nulle » et les autres
- Tests classiques sur une moyenne, sur une proportion
- Tests de comparaisons de deux populations
- Échantillons indépendants et échantillons appariés
- Tests d'ajustement
- Test de liaison entre variables
- Tests non paramétriques, test du Khi-2

- Introduction au ré-échantillonnage Bootstrap

Module 3 : Régression linéaire, logistique et analyse de la variance

Objectifs : Acquérir la connaissance méthodologique et pratique des méthodes de modélisation que sont la régression linéaire, la régression logistique, l'analyse de la variance et de la covariance. Elles permettent d'obtenir une analyse explicative d'un phénomène, de confirmer des hypothèses, de prendre des décisions ou encore d'effectuer des prévisions

- Le modèle linéaire

- Introduction
- Le modèle linéaire, principe, écriture

- Régression simple et multiple

- Le modèle
    - Estimation des coefficients
- Validation du modèle
    - Tableau d'analyse de variance et coefficient de détermination (R^2)
    - Test global du modèle : le test de Fisher
    - Test de nullité de chacun des coefficients du modèle : le test de Student
    - Recherche de valeurs influentes
    - Etude graphique et statistique des résidus

- Liaisons entre variables explicatives : évaluer le degré de multicolinéarité, utilisation de l'analyse en composantes principales
- Critères de sélection de modèles concurrents
  - Critères de sélection de modèles : coefficient de détermination, coefficient de détermination ajusté, Cp de Mallow
  - Méthodes pas à pas de sélection de modèle : ascendante, descendante, mixte
- Utilisation du modèle en prévision
  - Intervalle de confiance et de prévision

- Régression logistique

- Spécificité et complémentarité avec la régression linéaire classique
- Spécification du modèle
  - Hypothèses
  - Fonction logit
  - Interprétation des paramètres du modèle
  - Intervalle de confiance
- Estimation des paramètres du modèle
- Tests d'hypothèses sur les paramètres du modèle
- Codage et interprétation des variables explicatives (binaire, qualitative)
- Comparaison de modèles et sélection de variables
- Validation des hypothèses du modèle et analyse des résidus

- Analyse de la variance et de la covariance

- Le modèle
  - Analyse de la variance à un ou plusieurs facteurs
  - Décomposition de la variance
  - Effets principaux et effets des interactions
  - Analyse de la covariance
- Vérification des hypothèses sur les données, validation du modèle
  - Tests de normalité des distributions, d'homogénéité des variances (homoscedasticité), transformation des données
  - Utilisation des boîtes à moustaches
  - Etude graphique et statistique des résidus
- Tests d'hypothèses, exploitation
  - Tests de comparaisons multiples de moyennes (Tukey, Bonferroni, ...)
  - Tests de type I, II, III
  - Analyse de contraste pour vérifier une hypothèse de départ
- Cas des plans déséquilibrés

- Les différents types de modèles
    - Modèles croisés
    - Modèles imbriqués
    - Mesures répétées

Module 4 : Analyse des données : méthodes exploratoires (ACP, AFC, classification)

Objectifs :Découvrir les principales méthodes exploratoires d'analyse des données (ACP, AFC, Classification) afin de mettre en évidence les liaisons entre paramètres, les similitudes et différences entre observations. Interpréter les résultats numériques et graphiques, éviter les pièges, savoir résumer l'information obtenue et communiquer les résultats importants

- Rappels de notions utiles

- Notions de statistique
- Panorama des méthodes
- Quelques rappels pour les visualisations
- Tronc commun des concepts

- Analyse en composantes principales (ACP)

- Représentations approchées optimales
- Interprétation des axes factoriels
- Variables actives et illustratives

- Analyse Factorielle des Correspondances (AFC)

- Tableaux de fréquences
- Comparaisons de profils et distances
- Représentation des lignes et des colonnes
- Points supplémentaires

- Analyse Factorielle des Correspondances Multiples (ACM)

- Type de données concernées
- Transformation des données : le tableau disjonctif complet
- Les variables actives et illustratives : le « thémascope »
- Les représentations graphiques et les indicateurs d’aide à l’analyse

- Analyses et interprétations

- Méthodes de classification

- Méthodes hiérarchiques
- Méthodes des centres mobiles
- Classifications mixtes
- Affectation des individus à des classes
- Typologies et description des classes obtenues

- Complémentarité des méthodes

- Panorama des logiciels

Module 5 : Analyse des données : méthodes décisionnelles

Objectifs :Découvrir les principales méthodes décisionnelles d'analyse des données (arbres de décision, règles d'association, régression multiple, analyse discriminante, ...), choisir celle appropriée au problème et aux données. Interpréter les résultats numériques et graphiques, éviter les pièges, savoir résumer l'information obtenue et communiquer les résultats importants

- Arbres de Décision

- Principe et algorithmes de construction
- Identification des variables discriminantes
- Arbre de régression et arbre de classement (discriminant)

- Règles d’association

- Recherche des règles d’association pertinentes dans une base de données
- Sélection des meilleures règles et leur utilisation
- Utilisation en Data Mining

- Modèle linéaire et régression multiple

- Modélisation de la relation entre la variable cible et les variables explicatives
- Interprétation des résultats et pièges à éviter

- Analyse Discriminante

- Analyse linéaire discriminante
- Qualité d'une discrimination
- Probabilité d'appartenance à un groupe

- Comparaisons, domaines d'application, conditions d'utilisation

- Comparaisons des propriétés, qualités et conditions d'application des familles de méthodes et des méthodes elles-mêmes.
- Complémentarité des méthodes
- Panorama des logiciels

- Réseaux de Neurones

- Principes des réseaux de neurones (perceptron)
- Techniques de calculs
- Applications à la résolution de nombreux problèmes dont la discrimination et la régression

Date de dernière modification : 5 novembre 2024