

4- Introduction au langage R pour Data Scientist

DATASCIENCE3

Formation disponible en intra-entreprise, nous contacter pour plus de détails

Présentiel - Synchrones

Objectifs

Prendre en main le langage R

Public Visé

Personnes ayant suivi un enseignement de statistique théorique de base ou de machine learning dans un autre environnement (par exemple SAS), et qui sont désireuses d'évaluer le langage R ou de pratiquer la statistique et le machine learning dans un environ

Pré Requis

Pré-requis : Connaissances de bases en Statistiques et Machine Learning
Pré-requis conseillé : Mise à niveau théorique avec les Fondements de la Data Science

Les + métier

Cette formation convient notamment pour aider des Statisticiens à faire la transition vers le Big Data et la Data Science et de véritablement prendre en mains le langage R (ou ou R sous KNIME, connaissances KNIME nécessaires dans ce cas).
La formation es

Objectifs pédagogiques et d'évaluation

Prendre en main le langage R

Méthodes pédagogiques

- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions

Méthodes d'évaluation des Acquis

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation de fin de stage adressé avec la facture ou remise par le formateur aux stagiaires

Parcours pédagogique

Présentation de R

- But de la formation
- Présentation du langage et de son environnement, de sa philosophie
- R et R Studio

Types de données et opérateurs de base

- Types: int, float, complexe... etc.
- Opérateurs: assignation, égalité, multiplication... etc.

Objets R

- Vecteurs
- Matrices et aires
- Tables et Data Frames
- Listes et Facteurs

Manipulations de base du langage

- Fonctions d'aperçu
- Accès aux éléments et colonnes des objets R
- Fonctions et attributs de base: names, diff, sort, length... etc.

Entrées/Sorties de base

- Partie de travail: créer, charger, modifier
- Lire/écrire dans un fichier plat
- Installer et charger un package R
- Bonus: lire un fichier SAS avec le package sas7bdat

Manipulation de données avancée

- Manipulations de formes
- Filtrage et compréhension de listes
- Méthodes table () et aggregate ()
- Jointures avec le package dplyr

Structures de contrôle

- Construction if – else
- Constructions for et while
- Opérateurs de boucles: lapply, sapply... etc.

Fonctions

- Déclaration et arguments d'une fonction
- Exemples
- Exercices

Statistiques de base

- Fonctions statistiques basiques: mean, var, quantile... etc.
- PDF et histogrammes
- Quelques tests statistiques: test de normalité, Student, Fischer
- Statistiques et graphiques basiques

Méthodes statistiques avancées: première partie

- Analyse en Composantes Principales: rappels sur l'ACP
- ACP avec les packages Stats et FactoMineR
- Détection d'outliers à l'aide d'une ACP
- Introduction à l'analyse de correspondances multiples: ACM
- Clustering: k-means

Méthodes statistiques avancées: deuxième partie

- Modèles prédictifs rappels méthodologiques: variables explicatives, variables cibles, échantillons d'apprentissage, échantillons test et validation.
- Régressions: régression linéaire, régression logistique, estimations d'erreurs (rmse).
- Classification par arbres de décision: matrices de confusion, importance des variables, indicateurs de performance.
- Bonus: algorithme RandomForest

Graphiques avancées

- Introduction aux graphiques avancées avec le package ggplot2
- Bonus: visualisation d'ACP avec le package ggbiplot

Use Case complet

- Données réelles: fichier issue d'une étude universitaire
- Statistiques descriptives
- ACP
- Détection d'outliers
- Clustering
- Régression et Classification
- Comparaison de performance entre les modèles
- Visualisation des résultats

Méthodes et modalités d'évaluation

Test des connaissances acquises et délivrance d'une attestation de fin de formation

Durée

35.00 Heures **5** Jours