

New

Analyse et automatisation de données avec R

Manipuler, visualiser avec ggplot2 et automatiser les analyses statistiques avec R pour exploiter et valoriser les données

 Présentiel ou en classe à distance



2 jours (14 h)

Prix inter : 1.750,00 € HT
Forfait intra : 5.690,00 € HT

Réf.: BI118

La formation **Analyse et automatisation de données avec R** permet de maîtriser les principaux outils de l'**analyse de données** et de la **data visualisation** avec **R**. Elle couvre la manipulation de données avec **tidyverse**, la création de graphiques avec **ggplot2** et l'automatisation de rapports avec **R Markdown** et **Quarto**. Une formation idéale pour produire des analyses fiables, reproductibles et industrialiser les traitements de données dans un environnement professionnel.

A qui s'adresse cette formation ?



Pour qui

- Data analysts, statisticiens



Prérequis

- Notions statistiques et bases R
- **Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !**

Programme

1 - Environnement R moderne et écosystème tidyverse

- Revoir les fondamentaux de RStudio et de l'écosystème R : Posit, packages CRAN et projets RStudio
- Découvrir la philosophie du tidyverse et ses principaux packages : dplyr, tidyr, ggplot2, readr, purrr, lubridate et stringr
- Utiliser les opérateurs pipe %>% de magrittr et |> natif de R pour construire des pipelines lisibles
- Importer différents formats de fichiers avec readr (read_csv), readxl (read_excel) et les fonctions associées
- Appliquer les conventions et bonnes pratiques : tidy data, nommage des variables et organisation d'un projet d'analyse

2 - Sélectionner, filtrer et agréger les données

- Utiliser select() pour sélectionner des colonnes par nom, index ou helpers : starts_with, contains et where
 - Filtrer les lignes selon des conditions logiques combinées avec filter()
 - Créer et modifier des colonnes et réaliser des calculs vectorisés avec mutate()
 - Effectuer des agrégations par groupe (sum, mean, n, n_distinct)
 - Trier les données sur une ou plusieurs colonnes avec arrange()
 - Enchaîner les opérations avec le pipe afin de créer des pipelines lisibles et maintenables
- Atelier

Charger un fichier CSV et explorer les transactions avec readr, glimpse() et summary()

Filtrer les données selon plusieurs critères : date, catégorie et montant

Créer des colonnes calculées, telles que la taxe et le segment client, avec mutate()

Agréger le chiffre d'affaires par catégorie et par mois avec group_by() et summarise()

3 - Préparer et consolider des données multisources

- Restructurer les données avec `pivot_longer()` et `pivot_wider()` de `tidyr` pour passer du format large au format long et inversement
 - Réaliser des jointures avec `dplyr` : `left_join`, `inner_join`, `full_join` et `anti_join`, sur des clés simples ou multiples
 - Concaténer des DataFrames avec `bind_rows` et `bind_cols`
 - Convertir et extraire les dates avec `lubridate` : `ymd`, `dmy`, `year`, `month` et `wday`
 - Manipuler les chaînes de caractères avec `stringr` : `str_detect`, `str_replace` et `str_extract`
 - Traiter les valeurs manquantes avec `na_if`, `replace_na`, `drop_na` et `complete`
- Atelier

Charger deux fichiers de transactions et de clients, puis les fusionner avec `left_join`

Transformer une table du format large au format long avec `pivot_longer`

Extraire le mois et le jour de la semaine d'une colonne de dates avec `lubridate`

Traiter les valeurs manquantes avec `drop_na` et `replace_na`, puis exporter le dataset nettoyé

4 - Concevoir des visualisations de données professionnelles

- Comprendre la grammaire des graphiques : données, propriétés esthétiques, géométries, échelles, facettes et thèmes
 - Utiliser les géométries fondamentales : `geom_point`, `geom_line`, `geom_bar`, `geom_col`, `geom_histogram` et `geom_boxplot`
 - Configurer les propriétés esthétiques avec `aes()` : couleur, remplissage, forme, taille et transparence
 - Comparer des sous-groupes avec `facet_wrap` et `facet_grid`
 - Personnaliser les échelles et les thèmes avec `scale_x_continuous`, `scale_color_brewer` et `theme_minimal`
 - Combiner plusieurs graphiques dans une figure unique avec `patchwork`
- Atelier

Représenter l'évolution mensuelle du chiffre d'affaires avec `geom_line` et personnaliser les axes, le titre et le thème

Comparer la distribution des montants par catégorie avec `geom_boxplot` et `geom_violin`

Créer une matrice de facettes avec `facet_wrap` afin de visualiser un indicateur par segment client

Combiner trois graphiques avec `patchwork` et exporter la figure finale au format PNG haute résolution

5 - Automatiser des analyses statistiques reproductibles

- Réaliser des statistiques descriptives dans R : somme, moyenne, médiane, écart-type, quantiles et tables de fréquences
 - Appliquer les tests d'hypothèses courants : test t, test du khi-deux et corrélations
 - Construire une régression linéaire simple avec `lm()`, interpréter les coefficients et produire un résumé du modèle
 - Découvrir R Markdown et Quarto pour créer des analyses reproductibles aux formats HTML, PDF et Word
 - Structurer un document avec un en-tête YAML, des blocs de code et les options `echo`, `eval` et `fig.cap`
 - Paramétrer des rapports avec les paramètres YAML afin de générer automatiquement plusieurs versions
- Atelier

Créer un document Quarto avec un en-tête YAML, des blocs de code et des graphiques `ggplot2`

Intégrer des statistiques descriptives, une corrélation et un test t sur le dataset préparé

Ajouter un paramètre « segment client » dans le fichier YAML afin de générer un rapport propre à chaque segment

Compiler le rapport aux formats HTML et PDF, puis générer trois versions paramétrées en une seule commande



Les objectifs de la formation

- Manipuler et préparer des données avec le langage R
- Concevoir des visualisations de données avec `ggplot2`
- Automatiser des traitements et des analyses statistiques à l'aide de scripts R



Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.



Les points forts de la formation

- Une approche pédagogique fondée sur un cas d'étude fil rouge, permettant de manipuler un jeu de données réaliste, de son chargement jusqu'à la génération de rapports automatisés
- Une formation couvrant l'ensemble du cycle d'analyse de données avec R, de la préparation des données avec le tidyverse à leur visualisation avec ggplot2 et à la production de rapports avec Quarto
- Des ateliers pratiques mettant en oeuvre les principaux packages du tidyverse (dplyr, tidyr, lubridate, stringr et ggplot2) selon les bonnes pratiques actuelles
- Un apprentissage des méthodes de reproductibilité et d'automatisation grâce à Quarto, le standard moderne pour la génération de rapports et l'analyse reproductible avec R
- Les contenus pédagogiques numériques remis aux participants après la formation prolongent les acquis et facilitent la mise en pratique des compétences en environnement professionnel



Dates et villes 2026 - Référence BI118



Dernières places disponibles



Session garantie

A distance

du 12 oct. au 13 oct.

du 9 nov. au 10 nov.

Paris

du 12 oct. au 13 oct.

du 9 nov. au 10 nov.