

Formation R pour la data analyse

(3 jours)

Claire Della Vedova (claire@delladata.fr) Data Value (formation@datavalue.fr)

Objectif

S'approprier les outils, et les méthodes permettant de rendre son processus d'analyse de données sous R, aisé, reproductible, et performant.

Public

Toute personne amenée, dans le cadre de son activité professionnelle, à analyser des données avec le logiciel R, et souhaitant rendre son processus de travail plus performant.

Prérequis

Le participant a des connaissances de base sur le logiciel R : les types de données, les structures de données, la création et la manipulation d'objets, l'importation de données. Dans l'idéal, le participant a suivi la formation "Bien débiter avec R"

Méthode

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

Programme (3 jours)

Introduction

- Origine et évolution de R
- Place de R dans la data analyse

Rappels

- Les outils de R Studio
- Organiser son travail sous R : projet R et architecture
- Importations de données, de scripts et d'environnements

Manipulation de données avec le package dplyr

- Introduction au package tidyverse et à la notion de pipe
- Filtrer des lignes avec la fonction filter()
- Sélectionner des colonnes (variable) avec la fonction select()
- Création de nouvelles variables avec la fonction mutate()
- Renommer ses variable avec la fonction rename()
- Calcul de paramètres par sous groupes : fonctions group_by() et summarise()

Manipulation des variables catégorielles avec le package forecats

- Inspection des variables catégorielles avec les fonctions levels(), fct_count et fct_unique()
- Modifier l'ordre des modalités
- Modifier le nom des modalités
- Réaliser des regroupements de modalités

Manipuler les chaînes de caractères avec le package stringr

- Détection de patterns
- Découpage
- Gestion des longueurs
- Remplacement

Manipuler des données de date : utilisation du package lubridate

- Convertir les données au format YYYY-MM-DD et HH:MM:SS
- Décomposer les éléments d'année, de mois et de jour
- Calculer des différences de dates et les exprimer en jours, ou heure

Assemblage et pivot de tables

- Jointure de tables: left join, right join, inner join et full join
- Combinaison de tables par les lignes
- Les formats wide et long
- Passage d'un format wide à long et inversement

Réaliser des représentations graphiques performantes avec le package ggplot2

- Le principe des couches successives de ggplot2
- Réalisation des graphiques de base : scatterplot, barplots, line plot, boxplots
- Représentation des séries temporelles
- Utilisation du format long et facetting
- Gestion des couleurs, titres, axes et légendes

Générer dynamiquement son rapport d'analyse avec rmarkdown

- Principe, formats de sorties (html, docx, pdf)
- Les différents éléments d'un fichier Rmd : en-tête, chunk, etc. . .
- Gestion des éléments de texte : gras, italique, titre

- Gestion des tables
- Gestion des images
- Gestion des graphiques
- Gestion des options des éléments de code
- Gestion de la table des matières et numérotation

Introduction à la programmation fonctionnelle avec le package purrr (optionnel)

- Les lists
- Les fonctions `apply()`, `lapply()`, `tapply()`
- Les fonctions `map()`
- Les Nested data