Chapitre 1 : Introduction au tidyverse

Malick SENE

2025-04-15

Chapitre 1 : Introduction au tidyverse

Le **tidyverse** est un écosystème de packages R spécialement conçu pour la **science des données moderne**. Créé et soutenu principalement par **Hadley Wickham** et l'équipe de RStudio, le tidyverse propose une grammaire unifiée pour la manipulation, la transformation, l'analyse et la visualisation des données.

Il repose sur une idée fondamentale : les données doivent être **bien organisées** (ou "tidy") pour permettre un traitement fluide et intuitif. Cette approche, bien que simple, transforme la manière dont les utilisateurs interagissent avec leurs jeux de données dans R.

Qu'entend-on par "données tidy" ?

Le concept de **données tidy**, formalisé par Hadley Wickham, repose sur trois principes :

1. Chaque **variable** forme une **colonne** ;
2. Chaque **observation** forme une **ligne** ;
3. Chaque **type d'unité d'observation** forme une **table**.

Ces règles permettent de structurer les données de manière standardisée, facilitant ainsi leur traitement et leur visualisation avec les outils du tidyverse.

Pourquoi adopter le tidyverse ?

Voici quelques raisons pour lesquelles le tidyverse est devenu une référence incontournable pour les data scientists et les analystes :

- Une suite **intégrée** de packages qui partagent une syntaxe, une logique et une documentation cohérentes ;
- Une **prise en main intuitive**, favorisée par une grammaire claire et descriptive ;
- Une performance optimisée pour le traitement de jeux de données volumineux ;
- Un support natif pour **l'enchaînement des opérations** (via les pipes `%>%` ou `|>`) ;
- Une synergie avec d'autres outils modernes comme `RMarkdown`, `Shiny`, ou encore `Quarto` ;
- Une vaste communauté et une abondance de ressources pédagogiques disponibles.

Les piliers du tidyverse

Le tidyverse comprend les packages suivants :

Package	Utilité principale
`ggplot2`	Visualisation des données
`dplyr`	Manipulation et transformation des données tabulaires
`tidyr`	Restructuration des données en format tidy
`readr`	Importation rapide de fichiers plats (CSV, TSV...)
`tibble`	Alternative moderne au `data.frame` classique
`stringr`	Traitement des chaînes de caractères
`purrr`	Programmation fonctionnelle sur des listes

Le tidyverse est **extensible** : des packages comme `lubridate` (gestion des dates), `forcats` (facteurs), `readxl` (Excel), `haven` (SPSS, Stata, SAS) en font également partie.

Syntaxe générale et philosophie

Le tidyverse encourage une écriture :

- **Déclarative** : on décrit ce que l'on veut faire, pas comment ;
- **Lisible** : le code devient un langage humain ;
- **Modulaire** : chaque étape est une fonction claire ;
- **Chaînée** : via l'opérateur pipe `%>%` ou `|>` qui passe le résultat d'une fonction à la suivante.

Ce que vous apprendrez dans ce guide

Ce document vous accompagne à travers toutes les grandes familles de tâches de la science des données avec le tidyverse :

- L'importation et l'exportation de données depuis différents formats ;
- La manipulation des données avec `dplyr` ;
- La visualisation avancée avec `ggplot2` ;
- Le nettoyage, la transformation et le recodage avec `tidyr` et `forcats` ;
- La manipulation de texte avec `stringr` ;
- L'automatisation des traitements avec `purrr`.

Chaque étape sera illustrée avec des exemples basés sur les données **EHCVM**.

Bienvenue dans l'univers du tidyverse : **simple, cohérent, élégant.**