Rapport de Projet : Visualisation de Données Multivariées

Analyse et Visualisation Interactives avec D3.js

Cours : Information de visualisation

Superviseur : Marco Winckler

Étudiants : Mathis Hartmann, Alexis Dubarry, Noel Shanti, Cherigui Allah-Eddine

Date : 2 novembre 2025

UNIVERSITÉ CÔTE D'AZUR

Table des matières

1 Introduction

Ce projet a pour objectif d'analyser et de visualiser un jeu de données multivarié issu de Kaggle à l'aide de la librairie **D3.js**. Le rapport décrit les différentes étapes du pipeline de visualisation, les utilisateurs cibles, les objectifs de visualisation, ainsi que la conception et l'implémentation d'une interface interactive permettant l'exploration de données.

2 Choix du Jeu de Données

2.1 Source du Dataset

- **Lien Kaggle :** `https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-m`
- **Nom du dataset :** Netflix TV Shows and Movies
- **Type de données :** numériques, catégorielles et temporelles
- **Nombre de variables :** 15
- **Nombre d'observations :** 5850

2.2 Description du Jeu de Données

Le dataset contient des informations sur les films et séries disponibles sur Netflix, incluant leurs caractéristiques, notes, pays de production, durée et popularité. Les principales variables incluent : identifiant, titre, type (film ou série), description, année de sortie, âge recommandé, durée ou nombre de saisons, genres, pays de production, scores et votes IMDB, popularité et note TMDB.

Variable	Type	Description
id	Chaîne	Identifiant unique du contenu
title	Chaîne	Nom du film ou de la série
type	Chaîne	Type : "MOVIE" ou "SHOW"
description	Chaîne	Résumé du contenu
release_year	Numérique	Année de sortie
age_certification	Chaîne	Classification par âge (ex : TV-MA)
runtime	Numérique	Durée en minutes (films)
seasons	Numérique	Nombre de saisons (séries)
genres	Chaîne	Liste de genres du contenu
production_countries	Chaîne	Pays de production
imdb_id	Chaîne	Identifiant IMDB
imdb_score	Numérique	Note IMDB
imdb_votes	Numérique	Nombre de votes IMDB
tmdb_score	Numérique	Note TMDB
tmdb_popularity	Numérique	Score de popularité TMDB

TABLE 2.1 – Aperçu des variables principales du dataset Netflix TV Shows and Movies

3 Pipeline de Visualisation

3.1 Pipeline de Transformation des Données

Le projet utilise un notebook Jupyter (`preprocessing.ipynb`) pour préparer le dataset Netflix avant visualisation. Le processus se déroule en 5 étapes principales :

3.1.1 Processus de Transformation

1. Chargement des données brutes : Import de `titles.csv` contenant 5800+ films/séries avec métadonnées (genres, pays, scores IMDB/TMDB).

2. Parsing des listes JSON : La fonction `ensure_list_of_strings()` convertit les chaînes de caractères `"['drama', 'crime']"` en vraies listes Python `['drama', 'crime']` avec `ast.literal_eval()`.

3. Conversion codes pays : La fonction `code_to_name()` utilise `pycountry` pour transformer les codes ISO (`US, FR, JP`) en noms complets (`United States, France, Japan`).

4. Mapping géographique : La fonction `get_regions_from_names()` agrège 100+ pays en 8 régions continentales (Europe, Asia, North America, South America, Africa, Oceania, Middle East, Central America) via un dictionnaire prédéfini. Les co-productions génèrent plusieurs régions (ex : `US + UK → ['North America', 'Europe']`).

5. Nettoyage final : Suppression des colonnes inutiles (`imdb_id`, `tmdb_popularity`, `tmdb_score`), conversion des listes vides en `NaN`, et export vers `preprocessed.csv`.

Résultat : Dataset optimisé prêt pour D3.js avec 13 colonnes (contre 15 initialement), taille réduite de 8%, et hiérarchie géographique claire pour les visualisations. Le fichier final contient des données structurées directement exploitables par `d3.csv()`.

Transformation	**Avant**	**Après**
Genres	`"['drama', 'crime']"`	`['drama', 'crime']`
Pays	`['US', 'GB']`	`['United States', 'United Kingdom']`
Régions	Absent	`supprimé`
Colonnes	15	13

TABLE 3.1 – Transformations clés du preprocessing

4 Analyse des Utilisateurs et Objectifs

4.1 Utilisateurs Cibles

4.1.1 Journalistes

- **Overview :** Vue d'ensemble de la production et répartition des contenus.
- **Identify :** Identifier les films/séries marquants.
- **Details on demand :** Obtenir des informations détaillées sur un contenu.
- **Relate :** Mettre en relation plusieurs aspects (succès critique vs audience).
- **Export / History :** Extraire des données ou sauvegarder un état de la visualisation.

4.2 Objectifs et Tâches Utilisateurs – Journalistes

Objectif de Visualisation	**Tâches Utilisateurs**
Obtenir une vue d'ensemble sur la production et la répartition des films/séries (Overview)	Explorer la répartition des productions par pays, genre et période ; observer les tendances globales.
Identifier les contenus marquants (Identify)	Rechercher les films/séries les plus populaires, les mieux notés ou les plus regardés.
Explorer les détails à la demande (Details-on-Demand)	Cliquer ou survoler un film/série pour obtenir des informations détaillées (fiche complète, résumé, acteurs, note).
Mettre en relation plusieurs indicateurs (Relate)	Croiser des variables (ex. succès critique, popularité, durée) pour comprendre leurs corrélations.
Exporter et sauvegarder les visualisations (Export / History)	Extraire ou sauvegarder des graphiques et des jeux de données pour usage rédactionnel ou publication.

TABLE 4.1 – Objectifs de visualisation et tâches utilisateurs pour le profil journaliste (dataset Netflix).

5 Technique de Visualisation avec D3.js

5.1 Type de Visualisation

— Alexis : Treemap Hiérarchique Interactif avec Comparaison Temporelle (Carte à cases avec drill-down et mode dual-view).
— Shanti : Scatterplot Dynamique Animé (Nuage de points interactif avec contrôle temporel).
— Mathis : Chord diagram.
— Allah-Eddine : Sunburst (explorer la répartition des classifications d'âge selon genre et type de contenu, 3 niveaux : classification d'âge, type de contenu, genre).

5.2 Interactions

Le dashboard propose une architecture multi-niveaux d'interactions pour permettre une exploration fluide et interactive du dataset.

5.2.1 Filtres Globaux

— Sidebar avec navigation rapide et accès aux sections.
— Checkboxes multi-sélection (genres et régions) avec logique "All".
— Slider temporel pour sélectionner une plage d'années.
— Propagation instantanée des filtres à toutes les visualisations.

5.2.2 Interactions par Visualisation

— **Scatterplot** : Zoom/Pan, hover tooltip, click pour modal détails, Play/Pause pour animation temporelle, configuration axes/taille/couleur.
— **Treemap** : Drill-Down, hover tooltip, click pour panneau latéral, comparaison temporelle.
— **Sunburst** : Navigation hiérarchique, zoom arcs, hover tooltip, click pour détails.
— **Chord Diagram** : Hover et click sur rubans, coloration adaptative selon filtres.

5.2.3 Chargement de Dataset / Onglets

— 4 onglets correspondant aux visualisations.
— Lazy initialization pour performance.
— Les filtres restent actifs lors du changement d'onglet.

5.2.4 Résumé des Patterns d'Interaction

— **Drill-Down :** Exploration hiérarchique (Treemap et Sunburst).
— **Brushing & Linking :** Hover → mise en évidence synchronisée (Scatterplot, Chord Diagram).
— **Overview + Detail :** Tooltip + Modal / panneau latéral.
— **Focus + Context :** Zoom sémantique et rescaling (Scatterplot, Sunburst).
— **Animate Transition :** Transitions fluides lors de l'animation temporelle ou du drill-down.

5.3 Niveaux de Visualisation

Pour permettre une exploration efficace du dataset, la visualisation propose deux niveaux principaux :

5.3.1 Vue d'ensemble (Overview)

— Permet d'obtenir une vision globale du dataset.
— Visualise les tendances générales : répartition des contenus par pays, genre, type ou classification d'âge.

— Idéale pour identifier rapidement des patterns, des clusters ou des anomalies.
— Exemple : treemap par pays et genre, sunburst montrant la hiérarchie classification d'âge → type → genre.

5.3.2 Vue détaillée (Detail)

— Permet d'accéder à des informations spécifiques sur un film ou une série.
— Affichage des détails via : tooltip (au survol), panneaux latéraux ou modals (au clic).
— Inclut les métriques individuelles : note IMDb, popularité, durée, année de sortie, genres, pays de production.
— Exemple : cliquer sur un genre dans le sunburst ou sur une bulle du scatterplot pour afficher les informations complètes.

5.3.3 Interaction entre Overview et Detail

— Les interactions permettent de passer facilement de la vue globale au détail (drill-down).
— Les filtres globaux et la sélection dans la vue d'ensemble mettent à jour la vue détaillée en temps réel.
— Les deux niveaux sont synchronisés pour offrir un contexte tout en permettant un focus précis.

6 Implémentation Technique

6.1 Organisation du Projet

Le projet suit une architecture modulaire garantissant la séparation claire entre les différentes composantes : le tableau de bord principal, les filtres, les visualisations, les données et les styles.

Structure Générale

Le répertoire racine contient :
— **dashboard.html** : point d'entrée du tableau de bord ;
— **dashboard.js** : coordination entre les filtres et les visualisations ;
— **filters.js** : gestion des filtres interactifs ;
— **style.css** : feuille de style globale ;
— **data/** : contient les jeux de données brut et prétraité ;
— **preprocessing.ipynb** : script de nettoyage et de structuration des données ;

— **scatterplot/, treemap/, sunburst/, chord/** : répertoires dédiés à chaque visualisation.

Chaque dossier de visualisation contient le code JavaScript spécifique à la figure, ainsi que ses styles locaux lorsque nécessaire. Le fichier `dashboard.js` assure la synchronisation des filtres globaux et gère le chargement des données prétraitées (`preprocessed.csv`).

6.2 Technologies et Bibliothèques Utilisées

L'ensemble du tableau de bord repose sur la bibliothèque **D3.js (Data-Driven Documents)**, utilisée pour manipuler dynamiquement les données et générer des représentations graphiques interactives. Les visualisations sont intégrées dans un environnement web (HTML/CSS/JS) et orchestrées par un script principal pour assurer la cohérence entre les vues.

6.3 Principes Clés d'Implémentation

6.3.1 Chargement et Prétraitement des Données

Les données sont chargées de manière asynchrone à partir de fichiers CSV et converties en structures exploitables. Des opérations de nettoyage et de normalisation sont effectuées via le notebook `preprocessing.ipynb` : suppression des valeurs manquantes, homogénéisation des genres, et filtrage des types de contenus.

6.3.2 Gestion des Échelles et Mappings

Les échelles D3.js permettent de transformer les données en valeurs visuelles :

— Échelles linéaires ou logarithmiques pour les axes numériques ;
— Échelles ordinales pour les catégories (genres, régions, classifications) ;
— Échelles racine carrée pour les tailles de bulles, afin de préserver la perception visuelle des surfaces.

6.3.3 Axes et Grilles

Les axes sont générés automatiquement à partir des échelles à l'aide des fonctions de D3. Leur formatage est adapté à chaque visualisation (notation scientifique abrégée, étiquettes lisibles, unités cohérentes).

6.3.4 Hiérarchies et Layouts Spécifiques

Plusieurs visualisations s'appuient sur des structures hiérarchiques :

— **Treemap** et **Sunburst** : construits à partir d'objets hiérarchiques (genre, type, classification d'âge) via les fonctions `d3.hierarchy()`, `d3.treemap()` et `d3.partition()`.
— **Chord Diagram** : repose sur une matrice de relations (genres × régions) traitée par `d3.chord()` et `d3.ribbon()`.

6.3.5 Transitions et Animations

Les transitions permettent de fluidifier les mises à jour et d'accompagner les interactions. Chaque changement de filtre déclenche une transition contrôlée par `.transition()`, avec des fonctions d'interpolation (`d3.easeCubicInOut`, `d3.interpolate()`) pour adoucir les mouvements.

6.3.6 Interactions et Comportements Dynamiques

Les visualisations intègrent différents types d'interactions :
— Zoom et déplacement sur le Scatterplot (`d3.zoom()`) ;
— Sélection ou survol d'éléments avec affichage de tooltips ;
— Mises à jour synchronisées entre filtres et graphiques ;
— Exploration hiérarchique (drill-down) dans le Treemap et le Sunburst.

6.3.7 Pattern Update/Enter/Exit

Chaque visualisation suit le paradigme fondamental de D3.js :
— **Enter** : ajout d'éléments pour les nouvelles données ;
— **Update** : mise à jour des éléments existants ;
— **Exit** : suppression des éléments obsolètes.
Ce modèle assure la cohérence visuelle lors des changements de filtres ou d'année.

6.4 Synthèse par Type de Visualisation

Visualisation	Fonctionnalités et Approches Principales
Scatterplot	Échelles logarithmiques pour les votes IMDB, racine carrée pour la popularité, zoom et transitions animées.
Treemap	Structure hiérarchique basée sur la répartition des genres et types de contenus, avec navigation drill-down et mise à jour interactive.
Sunburst	Représentation circulaire hiérarchique (classification d'âge, type, genre) permettant l'exploration fluide entre niveaux.
Chord Diagram	Visualisation des relations entre genres et régions à partir d'une matrice de cooccurrences.

TABLE 6.1 – Résumé des approches techniques par visualisation

6.5 Pipeline Commun

Tous les modules partagent un pipeline d'exécution cohérent :

1. **Chargement des données** : lecture du CSV prétraité ;
2. **Filtrage dynamique** selon les choix de l'utilisateur (année, type, genre, région) ;
3. **Transformation et agrégation** via des fonctions de regroupement (`rollup`, `group`) ;
4. **Mise à jour des visualisations** : recalcul des échelles et redessin via le pattern Update/Enter/Exit ;
5. **Synchronisation globale** : chaque changement de filtre déclenche une mise à jour coordonnée de toutes les visualisations.

7 Démonstration et Résultats

7.1 Lien du GitHub

URL : alai06/Information_visualisation

7.2 Instructions d'Exécution

1. Télécharger ou cloner le projet depuis le dépôt GitHub.

2. Ouvrir le fichier `dashboard.html` dans un navigateur moderne (Chrome, Firefox, Edge).
3. Attendre le chargement du dataset `preprocessed.csv`.
4. Naviguer entre les onglets pour accéder aux différentes visualisations (Treemap, Scatterplot, Sunburst, Chord Diagram).
5. Utiliser les filtres (années, genres, pays, type de contenu) pour explorer les sous-ensembles.

7.3 Résultats et Observations

7.3.1 Vue d'ensemble du Dataset

L'analyse du dataset Netflix a permis de dégager plusieurs tendances générales :

— La production de contenus sur Netflix a fortement augmenté après 2015, avec un pic autour de 2019.

— Les films représentent une majorité du catalogue, mais les séries connaissent une croissance plus rapide ces dernières années.

— Les genres dominants sont le *drama*, la *comedy* et le *documentary*, tandis que les films d'action ou d'horreur sont moins représentés.

— Les classifications d'âge les plus fréquentes sont `TV-MA` et `TV-14`, traduisant une orientation vers un public adolescent et adulte.

7.3.2 Analyse par Visualisation

Treemap Hiérarchique : Cette visualisation a permis d'identifier la distribution des contenus par pays et genre. Par exemple, les États-Unis et l'Inde représentent les plus gros volumes de production, avec une forte diversité de genres. Le mode de comparaison permet de repérer rapidement les différences entres deux années.

Scatterplot Dynamique : Le nuage de points animé montre la relation entre la *popularité TMDB*, le *nombre de votes IMDB* et la *note moyenne*. On observe que :

— Les contenus populaires ne sont pas forcément les mieux notés.

— Certains genres, comme les documentaires, ont de bonnes notes mais peu de popularité.

— La dimension temporelle permet de suivre l'évolution des productions au fil des années.

Sunburst : Le diagramme en rayons de soleil illustre la hiérarchie entre la classification d'âge, le type de contenu et le genre. Il montre notamment que :

— Les séries `TV-MA` sont majoritairement des drames et comédies.

— Les films `PG` concernent surtout les genres familiaux ou d'animation.

Chord Diagram : Cette visualisation révèle les connexions entre les genres et les régions de production. Par exemple :

— Les genres *drama* et *comedy* sont présents dans presque toutes les zones géographiques.
— Certains genres comme *anime* sont fortement associés à la région Asie.

7.3.3 Synthèse

L'ensemble des visualisations permet une compréhension globale et approfondie du catalogue Netflix :

— Les **vues d'ensemble** (Treemap, Sunburst) aident à percevoir la structure et les volumes.
— Les **vues analytiques** (Scatterplot, Chord) révèlent les relations et corrélations entre variables.
— Les **interactions** (filtres, drill-down, hover) offrent une exploration fluide et personnalisée.

8 Conclusion et Perspectives

8.1 Bilan du Projet

Ce projet a permis de concevoir et de mettre en œuvre une application complète de visualisation de données multivariées à l'aide de la librairie **D3.js**. À travers l'étude du catalogue Netflix, nous avons pu :

— Structurer un pipeline de visualisation complet, de l'importation à l'interaction.
— Expérimenter différentes techniques de représentation (hiérarchique, temporelle, relationnelle).
— Développer des interactions avancées (drill-down, filtres globaux, animation temporelle).
— Offrir une plateforme intuitive permettant à différents profils d'utilisateurs (data scientists, journalistes, grand public) d'explorer librement les données.

Ce travail démontre la puissance et la flexibilité de D3.js pour créer des visualisations dynamiques et expressives, tout en soulignant l'importance d'une conception centrée sur l'utilisateur.

8.2 Améliorations Futures

Plusieurs pistes d'amélioration ont été identifiées pour enrichir le projet :

— **Intégration de nouvelles sources de données** : relier les données Netflix avec celles d'autres plateformes (Amazon Prime, Disney+) pour des comparaisons croisées.
— **Ajout d'un moteur de recherche intelligent** : filtrage par acteur, réalisateur ou mots-clés dans la description.
— **Optimisation des performances** : chargement progressif (*lazy loading*) et agrégation des données pour les grands volumes.
— **Visualisation narrative** : création d'un mode "storytelling" guidé pour mettre en évidence certaines tendances clés.
— **Analyse de sentiment** : intégration de l'analyse textuelle des descriptions ou critiques IMDB pour étudier la perception des contenus.

8.3 Conclusion Générale

En conclusion, ce projet a permis de combiner rigueur analytique et créativité visuelle pour produire un outil interactif et informatif. La visualisation de données n'est pas seulement un moyen d'afficher des chiffres, mais un véritable vecteur de compréhension et de découverte. Les outils développés ici ouvrent la voie à des explorations plus riches et à une prise de décision éclairée, illustrant la valeur ajoutée de la visualisation interactive dans le traitement des données complexes.

Références

— Kaggle Dataset : `https://www.kaggle.com/...`
— Documentation D3.js : `https://d3js.org/`