UQAC

Rapport – Mini-Projet : Analyse des contenus Netflix

8PRO408 – Outils de programmation pour la science des données

Étudiant : MOUCTAR DAFFE CODE PERMANENT : DAFM01119900

Sujet : Analyse exploratoire du catalogue Netflix (Movies & TV Shows)

1. Introduction

L'objectif de ce mini-projet est de réaliser une analyse exploratoire du catalogue Netflix à partir d'un jeu de données public provenant de Kaggle. Le dataset contient plusieurs milliers de titres (films et séries), accompagnés d'informations telles que le type de contenu, le pays d'origine, l'année de sortie, la date d'ajout sur la plateforme, le casting, le réalisateur, les genres et une courte description.

Cette analyse vise à mieux comprendre la composition du catalogue Netflix, la répartition entre films et séries, les pays dominants, les genres les plus représentés ainsi que certaines tendances temporelles. Le travail a été réalisé en Python à l'aide de Pandas, Matplotlib, Seaborn et Plotly, et complété par une mini-application Streamlit permettant d'explorer les données de manière interactive.

2. Méthodologie

Le travail s'est déroulé en plusieurs étapes :

1. **Chargement et inspection des données**
   Le fichier netflix_titles.csv a été chargé dans un DataFrame Pandas. Une première inspection (head(), info(), describe()) a permis d'identifier la structure du dataset (colonnes, types de données, taille) et de repérer la présence de valeurs manquantes dans certaines colonnes (notamment director, cast, country et date_added).
2. **Nettoyage et préparation**
   Un nettoyage léger a été effectué :
   - conversion de la colonne date_added en format de date ;
   - création de variables dérivées (year_added, month_added, main_country) ;
   - gestion des valeurs manquantes par remplissage avec la valeur "Inconnu" pour certaines colonnes textuelles ;
   - extraction du pays principal lorsque plusieurs pays étaient listés.
3. **Analyses descriptives**
   Plusieurs analyses ont été menées pour :
   - comparer la proportion de films et de séries ;
   - étudier la distribution des années de sortie (release_year) et des années d'ajout sur Netflix (year_added) ;
   - analyser les pays les plus représentés dans le catalogue ;
   - identifier les genres et les ratings les plus fréquents ;
   - comparer la durée typique des films et le nombre de saisons des séries.
4. **Visualisations**
   Des visualisations ont été produites à l'aide de :
   - **Matplotlib / Seaborn** : histogrammes, countplots, diagrammes en barres ;
   - **Plotly** : graphiques interactifs permettant de zoomer et filtrer les données (répartition par année, type, genre, rating).

5. **Application Streamlit**
   Une mini-application Streamlit a été développée, offrant :
   - un filtrage dynamique par type (film / série), pays, et année de sortie ;
   - des visualisations interactives des contenus par année, pays, genre et rating.

3. Résultats principaux

Les principales observations sont les suivantes :

- **Répartition films vs séries**
  Le catalogue contient davantage de films que de séries, même si le nombre de séries a significativement augmenté au cours des dernières années. Après 2015, la croissance des séries est très marquée, ce qui reflète la stratégie de Netflix d'investir dans les contenus sériels.
- **Tendances temporelles**
  La majorité des contenus ont été produits après les années 2000. La distribution des années de sortie montre une forte concentration de titres récents, avec un pic pour les années 2010. L'analyse de la date d'ajout (date_added) indique également une accélération du nombre de titres ajoutés à partir de 2015, ce qui correspond à la phase d'expansion rapide de la plateforme.
- **Répartition géographique**
  Les États-Unis constituent le pays le plus représenté dans le catalogue, ce qui n'est pas surprenant compte tenu du poids de l'industrie audiovisuelle américaine. Cependant, on observe également une présence notable de pays comme l'Inde, le Royaume-Uni, le Canada et divers pays européens ou asiatiques, ce qui témoigne d'une certaine diversification de l'offre.
- **Genres**
  L'analyse de la colonne listed_in montre que les genres les plus fréquents incluent des catégories telles que *International Movies*, *Dramas*, *Comedies*, *TV Dramas* ou *Children & Family Movies*. Les genres sont souvent combinés, ce qui enrichit la description mais complique légèrement l'analyse automatique.
- **Ratings (classifications d'âge)**
  Les ratings les plus courants sont TV-MA, TV-14 et TV-PG. Cela suggère que le catalogue vise principalement un public adolescent et adulte, avec une proportion non négligeable de contenus potentiellement matures.
- **Durée des contenus**
  Les films ont majoritairement une durée comprise entre 80 et 120 minutes, ce qui correspond au format classique des longs-métrages. Les séries comptent le plus souvent entre 1 et 3 saisons, indiquant une forte présence de mini-séries ou de séries relativement courtes.

4. Limites du dataset et perspectives

Quelques limites importantes ont été identifiées :

- Le dataset ne contient aucune information sur la **popularité** des contenus (notes des utilisateurs, nombre de vues, etc.). Il est donc impossible de relier la structure du catalogue aux préférences réelles des abonnés.
- Certaines colonnes sont **incomplètes** ou peu structurées (par exemple cast, director, country, listed_in), ce qui nécessite des prétraitements plus avancés pour des analyses poussées (NLP, regroupement de catégories, etc.).
- Les dates d'ajout peuvent être manquantes ou imprécises, ce qui limite la précision de certaines analyses temporelles.

Malgré ces limites, le dataset offre une base intéressante pour :

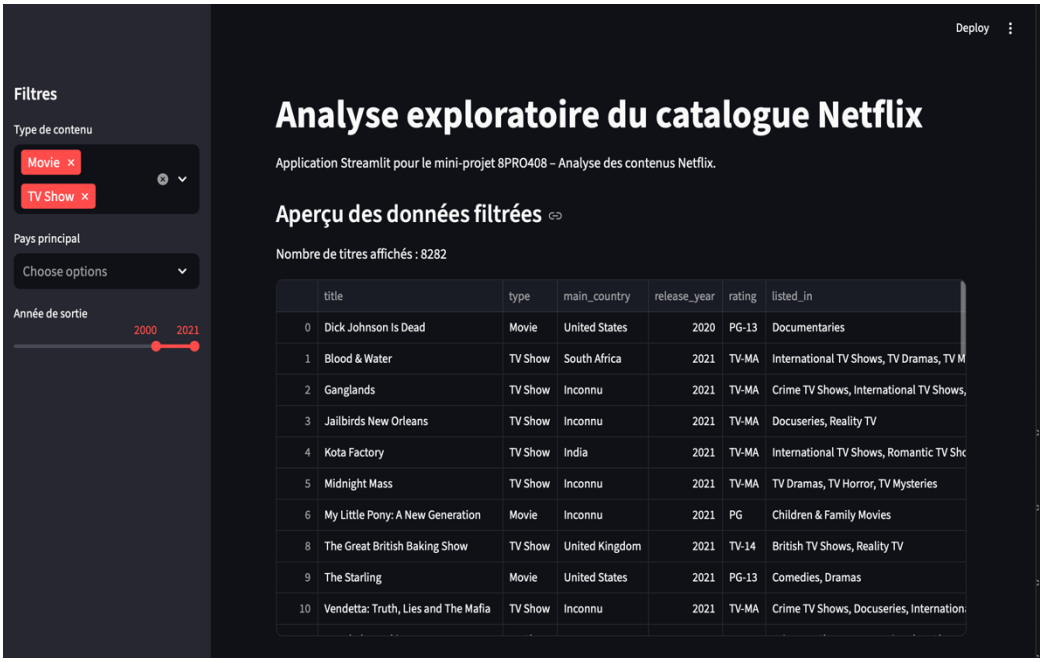
- construire des systèmes de **recommandation** simples,
- étudier la **diversité géographique** ou générique de l'offre,
- analyser l'évolution du catalogue dans le temps.

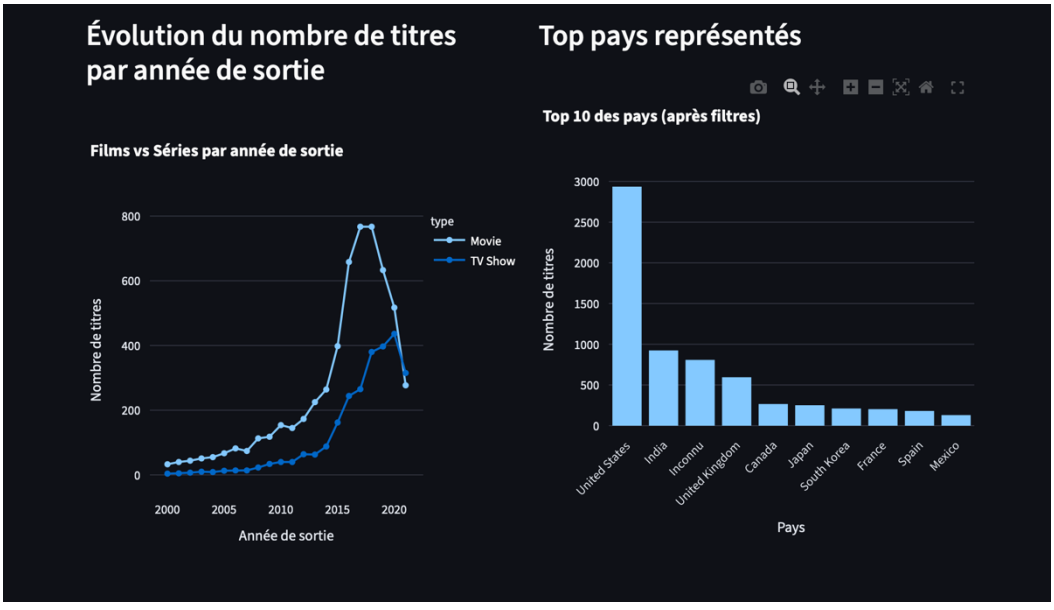
5. Conclusion

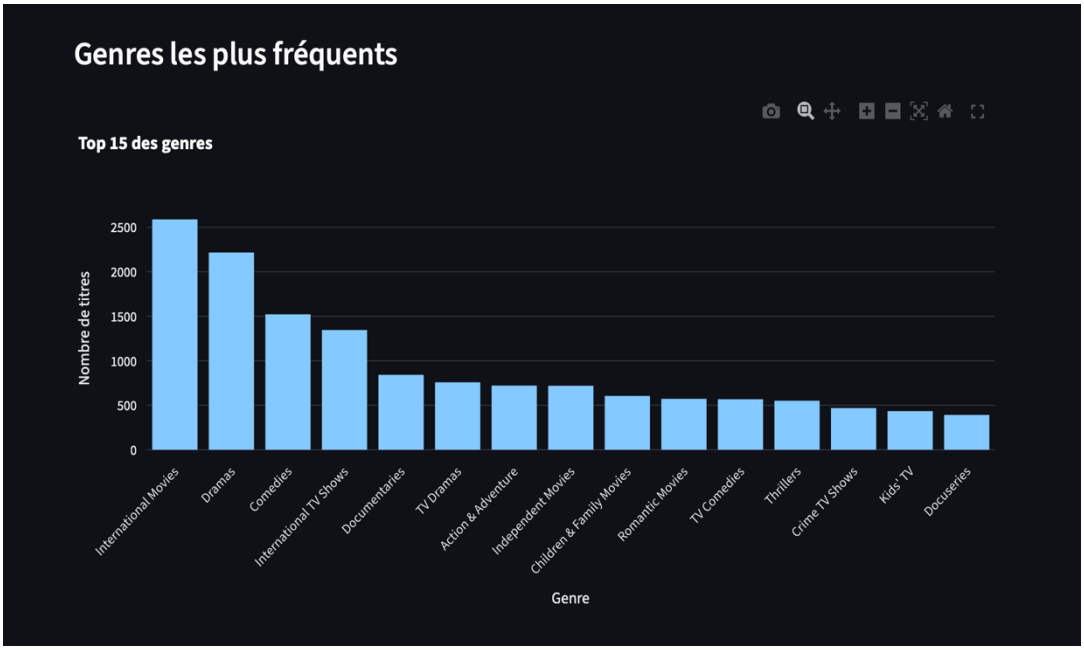
Ce mini-projet a permis de mettre en pratique les outils de programmation pour la science des données (Pandas, Matplotlib, Seaborn, Plotly, Streamlit) sur un cas concret et proche du monde réel. L'analyse exploratoire a fourni une vue d'ensemble claire de la composition du catalogue Netflix : répartition entre films et séries, pays dominants, genres populaires, tendances temporelles et caractéristiques générales des contenus.

La mini-application Streamlit développée permet de prolonger ce travail en offrant une exploration interactive du dataset, ce qui constitue un complément utile au notebook et au rapport écrit.

Voici quelques images du Streamlit :






Genres les plus fréquents
Top 15 des genres
Nombre de titres
2500
2000
1500
1000
500
0
International Movies
Dramas
Comedies
International TV Shows
Documentaries
TV Dramas
Action & Adventure
Independent Movies
Children & Family Movies
Romantic Movies
TV Comedies
Thrillers
Crime TV Shows
Kids' TV
Docuseries
Genre