Mini-Projet – Analyse des contenus Netflix

Cours 8PRO408 – Outils de programmation pour la science des données

1. Introduction

Dans ce mini-projet, nous analysons un jeu de données réel provenant de Netflix, comprenant environ 8 800 titres (films et séries) décrits par une douzaine de variables (type, titre, pays d'origine, genres, année de sortie, date d'ajout sur la plateforme, etc.).

L'objectif est de comprendre la composition du catalogue, d'identifier les types de contenus les plus fréquents, d'examiner la répartition géographique et les genres dominants, ainsi que d'étudier l'évolution temporelle des ajouts de contenus sur Netflix.

L'analyse a été réalisée en Python à l'aide des bibliothèques Pandas, Matplotlib, Seaborn et Plotly, puis complétée par une mini application interactive développée avec Streamlit.

2. Structure et qualité du jeu de données

Le dataset contient des informations à la fois numériques (par exemple l'année de sortie) et textuelles (titre, pays, genres, casting, réalisateurs).

Une première inspection met en évidence la présence de valeurs manquantes dans certaines colonnes comme "director", "cast" ou "country", ce qui reflète le fait que tous les titres ne sont pas documentés de manière uniforme.

Nous avons effectué un nettoyage minimal : normalisation des noms de colonnes, conversion de la colonne "date_added" au format date, et création de nouvelles variables dérivées telles que "year_added" et "month_added".

Les doublons sont très rares ou inexistants, ce qui est cohérent avec un catalogue de titres uniques.

3. Principaux résultats

Types de contenus :

L'analyse de la colonne "type" montre que le catalogue est constitué d'une majorité de films par rapport aux séries télévisées. La série temporelle du nombre de contenus ajoutés par année met en évidence une forte croissance du catalogue à partir de certaines années récentes, avec une augmentation simultanée du nombre de films et de séries.

Genres :

En exploitant la colonne "listed_in", nous avons exploré les genres associés à chaque titre.

Les genres les plus fréquents appartiennent principalement aux catégories généralistes (par exemple drama, comedy, action), souvent combinées à d'autres sous-genres.

Cela suggère que Netflix vise un positionnement large, en couvrant des genres populaires et grand public.

Pays d'origine :

La répartition par pays (colonne "country") montre une forte présence de quelques pays dominants, notamment les États-Unis, mais aussi d'autres pays producteurs de contenus comme l'Inde ou le Royaume-Uni. Cette concentration géographique reflète la structure de l'industrie audiovisuelle mondiale, même si Netflix inclut également des titres provenant d'un grand nombre de pays différents.

Dimension temporelle :

L'analyse de "release_year" met en évidence que la majorité des titres sont relativement récents, ce qui est cohérent avec la stratégie de proposer des contenus modernes et attractifs.

En parallèle, la variable "date_added" montre une montée en puissance des ajouts de contenus à partir de certaines années, avec un rythme soutenu de nouveaux titres, aussi bien pour les films que pour les séries.

4. Discussion et conclusion

Cette analyse exploratoire met en lumière plusieurs caractéristiques du catalogue Netflix : une majorité de films, une forte représentation de certains pays dominants, des genres très populaires et une concentration sur des contenus récents.

Le catalogue apparaît à la fois diversifié (en termes de pays, de genres et de formats) et orienté vers des contenus à forte demande.

Le jeu de données présente toutefois certaines limites : la présence de valeurs manquantes dans des colonnes clés (réalisateurs, casting, pays), l'agrégation de plusieurs genres ou pays dans une même cellule, ainsi que l'absence d'informations sur la popularité ou l'audience des titres.

Dans des travaux futurs, il serait intéressant de croiser ce catalogue avec des données supplémentaires (notes des utilisateurs, temps de visionnage, classements par pays) afin d'étudier plus finement les préférences des abonnés et la performance réelle des différents types de contenus.

Dans l'ensemble, ce mini-projet a permis de mettre en pratique les outils de programmation pour la science des données, en appliquant une démarche complète

d'analyse exploratoire (chargement, nettoyage, visualisation, interprétation) sur un jeu de données réel.