# **Présentation et exploration des données**

## Description détaillée du dataset TMDB (films, synopsis, genres, données numériques)

Le dataset TMDB contient des informations sur environ 5000 films issus de The Movie Database, une base collaborative et riche en métadonnées cinématographiques. Ce jeu de données contient 4803 lignes et 20 colonnes et est structuré autour de plusieurs types de variables :

### Films et Identifiants

id : identifiant unique attribué à chaque film.

title : titre générique du film.

original\_title : titre original dans sa langue de sortie.

homepage : URL officielle du film (peut être vide).

### Synopsis et descriptions textuelles

overview : résumé ou synopsis du film, qui présente l’intrigue ou le sujet principal en quelques phrases. Cette variable est importante pour le traitement automatique du langage et l’analyse thématique.

tagline : phrase slogan souvent utilisée pour le marketing ou les affiches.

### Genres et mots-clés

genres : liste des genres associés au film (ex. Action, Drame, Comédie). Stockés sous forme JSON, il faut parser cette chaîne pour obtenir les catégories.

keywords : mots-clés extraits décrivant des thèmes ou objets importants.

Ces informations catégorielles aident à comprendre la classification thématique des films.

### Données numériques et financières

budget : budget de production en dollars américains.

revenue : recettes totales générées par le film.

popularity : indice de popularité calculé sur la base de recherches et interactions des utilisateurs.

runtime : durée du film en minutes.

vote\_average : note moyenne attribuée par les utilisateurs (sur une échelle de 0 à 10).

vote\_count : nombre total de votes recueillis.

### Production et langues

production\_companies : sociétés de production (liste JSON).

production\_countries : pays producteurs (liste JSON).

spoken\_languages : langues parlées dans le film (liste JSON).

original\_language : code langue originale du film (ex. en pour anglais, fr pour français).

### Dates

release\_date : date de sortie officielle (format texte,à convertir pour analyses temporelles).

* Les données textuelles (overview, tagline, keywords) seront nettoyées et vectorisées (TF-IDF ou embeddings) pour représenter chaque film numériquement.
* Les genres et productions serviront à classifier et analyser les films par catégories.
* Les données numériques telles que budget, popularity et vote\_average seront normalisées pour le clustering.
* La date permettra d’étudier l’évolution sur le temps ou de filtrer par période.

## Analyse exploratoire des données

L’analyse exploratoire vise à comprendre la distribution, les caractéristiques et les relations principales des variables dans le dataset TMDB. Elle permet de révéler des tendances, des anomalies éventuelles, ainsi que la structure globale des données.

### Statistiques descriptives des variables numériques

Les principales statistiques sur des variables clés montrent :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Moyenne** | **Écart-type** | **Min** | **Q1** | **Médiane** | **Q3** | **Max** |
| budget | 29 M$ | 40.7 M$ | 0 | 0.79 M$ | 15 M$ | 40 M$ | 380 M$ |
| popularity | 21.5 | 31.8 | 0 | 4.67 | 12.9 | 28.3 | 875.6 |
| revenue | 82.3 M$ | 163 M$ | 0 | 0 | 19.2 M$ | 92.9 M$ | 2.79 G$ |
| runtime | 107 minutes | 22.6 | 0 | 94 | 103 | 118 | 338 |
| vote\_average | 6.1 / 10 | 1.19 | 0 | 5.6 | 6.2 | 6.8 | 10 |
| vote\_count | 690 | 1235 | 0 | 54 | 235 | 737 | 13,752 |

Les budgets et revenus présentent des valeurs minimales à 0 et sont très dispersés, la médiane étant bien inférieure à la moyenne (effet d’outliers).  
La popularité varie fortement, comme le nombre de votes.  
Les durées sont cohérentes pour un long-métrage, mais il y a des films à 0 minute (erreur ou données manquantes).  
La note moyenne des films tourne autour de 6, mais il existe quelques films extrêmes (0 ou 10).

### Visualisations principales

#### Distribution des budgets

* Histogramme montrant une majorité de films à petit ou moyen budget, quelques très gros budgets rares.

#### Distribution des notes moyennes

* Distribution centrée sur des notes entre 5 et 7, avec peu de films très mal ou très bien notés.

#### Repartition des genres

* Les genres les plus populaires sont Action, Drame, Thriller, Comédie.

#### Evolution du nombre de films par année

* La production de films varie selon les années, avec une tendance à l’augmentation sur les dernières décennies.

L’analyse exploratoire a permis de :

* Mettre en lumière la diversité forte des budgets, popularités et recettes.
* Confirmer la nécessité d’un traitement adapté des valeurs extrêmes (outliers).
* Identifier les genres dominants sur lesquels concentrer certaines analyses.
* Préparer le dataset pour les étapes suivantes : nettoyage approfondi, prétraitement textuel, vectorisation et clustering.

## Nettoyage des données (valeurs manquantes, doublons)

Le nettoyage des données est une étape cruciale afin de garantir la qualité et la fiabilité des analyses ultérieures sur le dataset TMDB, qui contient des informations sur environ 5000 films. Cette étape vise à éliminer les doublons, à gérer les valeurs manquantes, et à s’assurer que les données soient cohérentes et exploitables.

### Suppression des doublons

Un contrôle a été effectué pour détecter la présence d’entrées dupliquées.

* **Résultat** : aucun doublon n’a été trouvé, ce qui témoigne d’une bonne qualité initiale du dataset.
* **Conclusion** : aucune suppression n’a été nécessaire à ce stade.

### Gestion des valeurs manquantes

* Synthèse des valeurs manquantes

Après inspection, voici les colonnes contenant des données manquantes et leur quantité respective :

|  |  |  |
| --- | --- | --- |
| **Colonne** | **Valeurs manquantes** | **% par rapport au total** |
| homepage | 3,088 | ≈ 64% |
| tagline | 841 | ≈ 17.5% |
| release\_date | 1 | ≈ 0.02% |
| Autres | 0 | 0% |

* Analyse
* **homepage** (URL officielle) et **tagline** (phrase d’accroche) sont des champs non critiques pour le clustering ; leur absence ne nuit pas à l’analyse principale.
* La colonne **release\_date** possède une seule valeur manquante, facilement écartable si nécessaire.
* Toutes les autres colonnes essentielles (budget, genres, synopsis, scores, etc.) ne contiennent aucune valeur manquante, assurant une base solide.
* Traitement effectué
* Les films sans synopsis (`overview`) ont été retirés (3 films), car le synopsis est fondamental pour le traitement automatique du langage et la vectorisation textuelle.
* Les valeurs à 0 dans le budget, les recettes et la durée ont été imputées par la médiane des valeurs non-nulles pour préserver l’intégrité des données numériques.

Le nettoyage a permis d’obtenir un dataset de haute qualité, prêt à être exploité pour les analyses exploratoires et les traitements avancés. L’absence de doublons et la quasi-totalité de valeurs renseignées renforcent la pertinence des résultats futurs.

Les champs non essentiels avec valeurs manquantes peuvent être soit ignorés, soit traités, sans risque de biais majeur.

## Prétraitement textuel des synopsis

Dans le cadre de l’analyse de données issues du dataset TMDB, le prétraitement textuel des synopsis est une étape cruciale permettant d’améliorer la qualité des analyses sémantiques ultérieures, telles que la vectorisation et le clustering des films. Ce prétraitement comprend essentiellement le nettoyage du texte, la tokenization et la suppression des mots vides (stop words).

Outils utilisés

* **NLTK (Natural Language Toolkit)** : bibliothèque Python reconnue pour le traitement du langage naturel, utilisée pour la tokenization des textes.
* **StopwordsISO** : bibliothèque fournissant des listes de stop words multilingues basées sur la norme ISO 639-1, offrant une couverture plus large que les stop words natifs de NLTK.
* **Setuptools** : utilisé pour assurer la bonne intégration de StopwordsISO dans l’environnement Python.
* **Langdetect** : bibliothèque permettant de détecter automatiquement la langue d’un texte, utilisée pour confirmer ou corriger la langue indiquée dans le dataset.

Installation des bibliothèques

Les commandes suivantes ont été exécutées pour installer les outils requis :

pip install nltk  
pip install stopwordsiso  
pip install setuptools

pip install langdetect

Lien vers le dépôt officiel de StopwordsISO :

<https://github.com/stopwords-iso/stopwords-iso>

Description du prétraitement réalisé

1. Nettoyage du texte

* Conversion des textes en minuscules pour uniformiser la casse.
* Suppression de la ponctuation, des chiffres et des caractères spéciaux, en prenant en compte les caractères accentués via une expression régulière adaptée.

2. Détection de la langue

* Utilisation de la bibliothèque langdetect pour identifier automatiquement la langue du synopsis.
* Comparaison avec la colonne original\_language fournie dans le dataset pour corriger les incohérences.
* Création d’une colonne langue\_effective intégrant la langue détectée lorsqu’elle diffère de la valeur initiale.

3. Tokenization

* Découpage du texte en mots élémentaires (tokens) à l’aide de NLTK.
* Cette étape prépare le texte pour la suppression des mots non significatifs.

4. Suppression des stop words

* Utilisation de **StopwordsISO** pour la suppression des mots vides, adaptée à la langue effective de chaque synopsis.
* Cette approche remplace l’utilisation des stop words de NLTK, car StopwordsISO fournit une liste plus étendue et adaptée à la diversité linguistique importante du dataset TMDB.
* La suppression ciblée des mots fréquents inutiles permet de réduire le bruit et de concentrer l’analyse sur les termes à forte charge sémantique.

Avantages de l’utilisation de StopwordsISO vs NLTK

* Couverture multilingue bien plus large (plus de 80 langues supportées), correspondant mieux à la diversité du dataset TMDB (37 langues identifiées).
* La détection automatique via **langdetect** permet de corriger les erreurs ou manques de métadonnées (original\_language)
* Alignement sur la norme ISO 639-1 garantissant la cohérence des codes langue pour le filtrage.
* Mise à jour fréquente des listes de stop words grâce à la communauté open-source, assurant une meilleure qualité des données.

Conclusion

Le prétraitement textuel des synopsis est une étape essentielle pour la réussite de l’analyse sémantique et du clustering dans le projet FilmClusterTMDB.   
L’utilisation conjointe de NLTK pour la tokenization et de StopwordsISO pour la suppression des stop words multilingues garantit une préparation propre et adaptée à la diversité linguistique du dataset.   
Cette méthodologie assure un nettoyage efficace des données textuelles, favorisant des résultats d’analyse plus précis et robustes.