**Problématique :**

Prévision des entrées de cinéma en fonction de :

Données du film : date de sortie, casting, budget, réalisateur, durée du film …

Données externes : météo , critères géographiques…

Données graphiques : affiches

**Sources de données :**

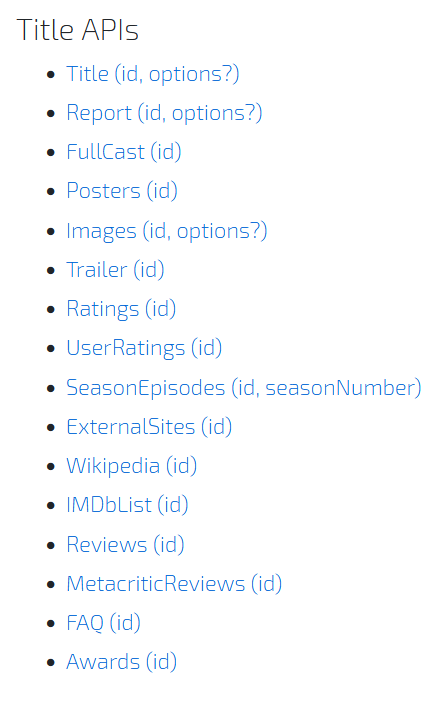
UNE MINE D’OR A PREMIERE VUE : <http://www.jpbox-office.com/v9_filmlst.php>

* Listes de films par ordre alphabétique (beaucoup beaucoup de films)
* Budget
* Date de sortie
* Nombre d’entrées
* Genre
* Durée
* Nationalité

API  IMDB🡪 OK c’est dingo tu récupères l’id d’un film et après tu as autant de fonction que tu veux :

- <https://imdb-api.com/api/#FullCast-header> 🡪 tout le casting d’un film

Et plein d’autres :



* LIMITER A 100 PAR JOUR EN GRATOS. SINON FAUT PATER 24 DOLLARS POUR ATTEINDRE 5000 JOURS DONC LARGE… A VOIR.

Webscrapping wikipedia / allociné

Cnc nombre de cinéma par ville

Data.culture.gouv (fréquentation annuelle)

Liste des établissements cinématographiques actifs

Company.boxoffice.com

Lien à tester : https://www.cbo-boxoffice.com/v4/page000.php3

# IMPORT DES DONNEES

Je propose qu’on mette à jour un script R au fur et à mesure avec l’import des données.   
On met un maximum et après on avise sur les variables que l’on garde et croise pour modéliser.

J’ai créé le fichier IMPORT\_DATA.R avec un premier jeu de donnée Cf ci-dessous. Ce serait pas mal de l’incrémenter au fur et à mesure.

## Variable à prévoir Y  :

Source données publiques du CNC MIS 12/06/2022 VARIABLE Y : [https:\\www.data.gouv.fr\fr\datasets\films-ayant-realise-plus-dun-million-dentrees\](https://www.data.gouv.fr/fr/datasets/films-ayant-realise-plus-dun-million-dentrees/)

Données importées : TITRE, ENTREES, NATIONALITES, SORTIE.

Limite : Film avec + de 1 millions d’entrées de 2003 à 2021

857 films.

TRAITEMENTS : Données totalisées par TITRE, renommage colonne

Entrée semble ok , nationalités c’est le bordel (plein de données différentes)

* Resultat CSV sur le Git : ./DATA\_CLEAN/FILMS\_1MILLIONS.csv

**PLANNING :**

14/12 : DATA SET

* LISTE SUR EXCEL FICHIER + CARAC DONNEES (HISTORIQUE + GEOGRAPHIQUE)
* RECUPERER DONNEES BRUTES API

18/01 MODELE

20/02 SHINY