**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

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.

Par ailleurs pour fiabiliser faudrait croiser ensuite les différentes sources.

J’ai créé le fichier IMPORT\_DATA.R avec un premier jeu de donnée Cf ci-dessous.

## 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 : 1\_ENTREES

**PLANNING :**

14/12 : DATA SET

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

18/01 MODELE

20/02 SHINY