# CaeaFlix project

*Algorithme de recommandation de films*

# Sommaire

# Qui sommes-nous ?

# Equipe

**Scrum master :**

Arnaud Le Naourès

**Product Owner :**

César Ozeer

**Team Members :**

Antoine Rafflegeau

Esrin Erdem

# CaeaFlix … C’est quoi ?

# L’objectif du projet

**Projet Caeaflix :**

- Relancer les ventes d’un cinéma dans la Creuse

**Demande :**

- Mettre en place un système de recommandations de films

# Quels films ?

**Approche business :**

- S'appuyer sur une étude consommateurs afin de prendre les bonnes décisions

**Sélection films:**

- Films(movies)
- diffusé en français ou en anglais *(mais seulement pour ceux aux USA)* …
- … dans plus de 7 pays
- note moyenne > Q1
- nombre de votes > Q1

La base de données finale contient près de **15 000 films**.

# Constitution de la base de données

**Données de départ :**

- *title_akas, title_basics, title_principals, title_ratings, name_basics*

  ⇒ *Fusion des tables*

**Manque d'informations :**

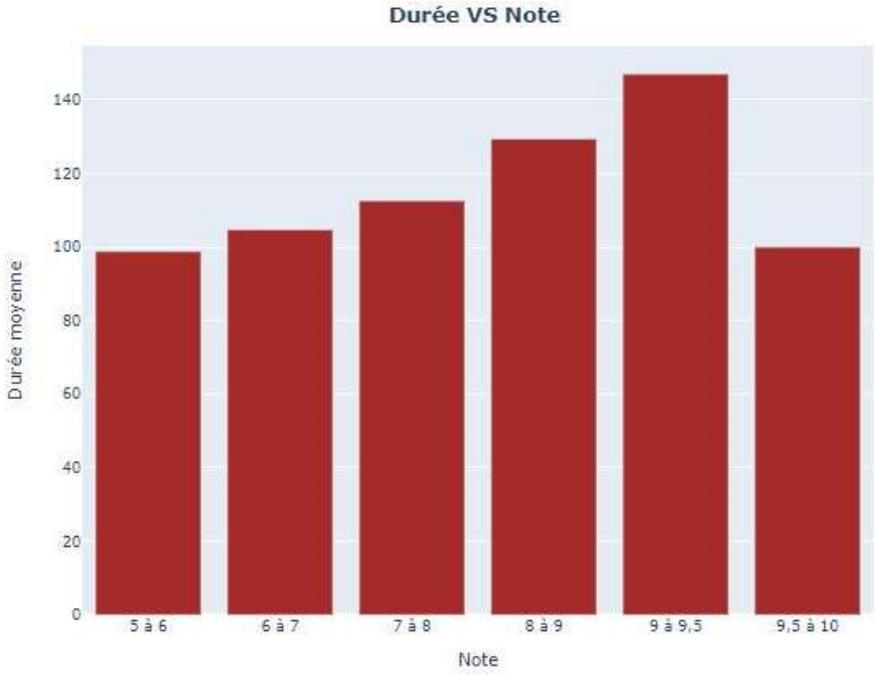
- recettes, budget, affiche du film, mots-clés, lien vers le site du film

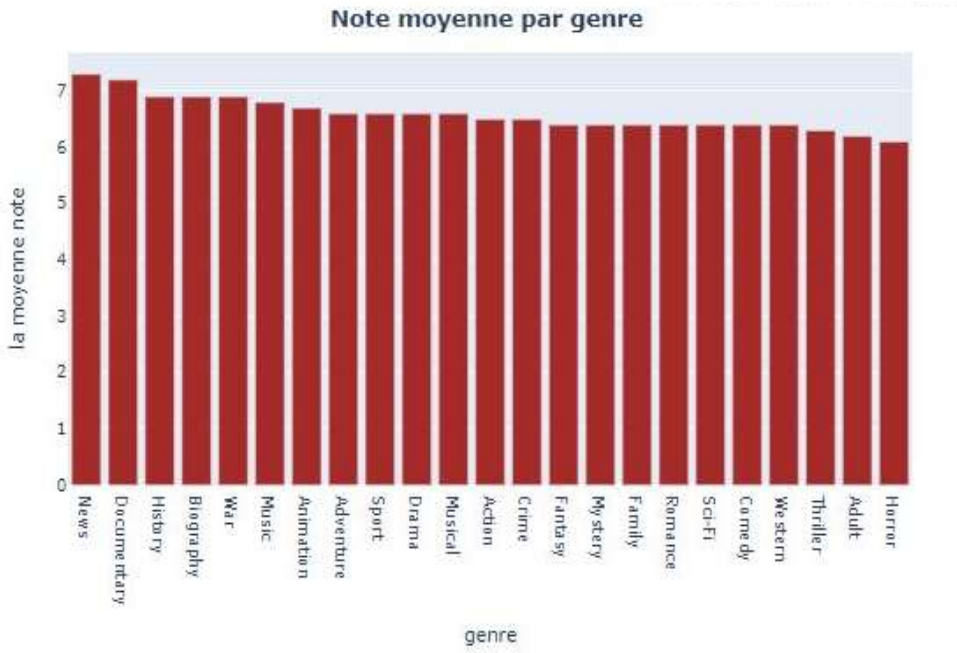
  ⇒ recours à des API pour récolter ces informations

**Data cleaning :**

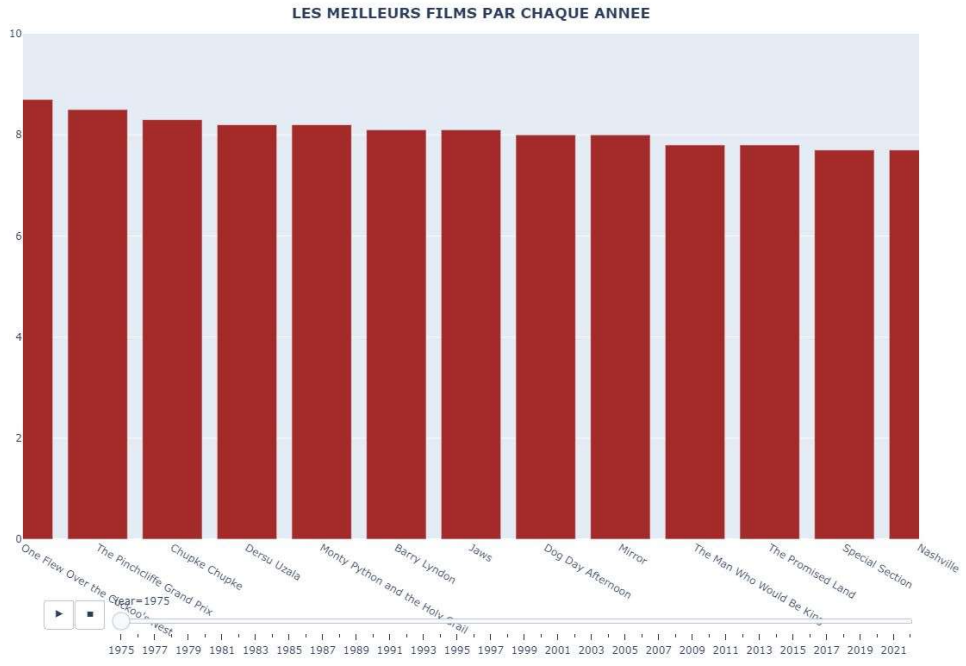
- Regroupement avec une ligne par film et une colonne par catégorie (acteur, actrice, réalisateur)
- Suppression des "NaN"

# Data viz (1/2)


Durée VS Note
Durée moyenne
0
20
40
60
80
100
120
140
5 à 6
6 à 7
7 à 8
8 à 9
9 à 9,5
9,5 à 10
Note


Note moyenne par genre
la moyenne note
0
1
2
3
4
5
6
7
News
Documentary
History
Biography
War
Music
Animation
Adventure
Sport
Drama
Musical
Action
Crime
Fantasy
Mystery
Family
Romance
Sci-Fi
Comedy
Western
Thriller
Adult
Horror
genre

# Data viz (2/2)


LES MEILLEURS FILMS PAR CHAQUE ANNEE
10
8
6
4
2
0
One Flew Over the Cuckoo's Nest
The Pinchcliffe Grand Prix
Chupke Chupke
Dersu Uzala
Monty Python and the Holy Grail
Barry Lyndon
Jaws
Dog Day Afternoon
Mirror
The Man Who Would Be King
The Promised Land
Special Section
Nashville
Year=1975
1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015 2017 2019 2021


Répartition des Genres
200
150
100
50
0
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Sport
Thriller
War
Western
Year=1975
La Moyenne Durée
1975 1978 1981 1984 1987 1990 1993 1996 1999 2002 2005 2008 2011 2014 2017 2020

# Modèle

**Modèle : K-Nearest Neighbors (K-NN)**

- Variables ⇒ positionnent les films les uns par rapport aux autres

- Recommandation ⇒ film le plus proche en tenant compte de toutes les variables

# Modèle final (1/2)

**Mise en place de 4 dataframes :**

- Films “anciens” (avant 1999 compris)
- Films “récents” (après 1999)
- Anime (seulement les anime)
- Indiens (seulement les films de Bollywood)

**Création de variables :**

| **Binaires** | **Scaled** |
|---|---|
| Genre | Runtime minutes |
| Key words | Average rating |
| Pays / ISO / nom de la (des) société(s) de production | Number of votes |
| Nom réalisateur / acteurs / actrices | Start year |
| | Diffusion |

# Modèle final (2/2)

- **3 modèles :**
  - Modèle général (films les plus similaires) :
    - X = Keywords + Nom des producteurs + Nom des pays des producteur + Nom des ISO des pays des producteurs + Nom réalisateur + Nb diffusion + Genres + Langue
  - Modèle par réalisateur :
    - X = Nom réalisateur + note moyenne
  - Modèle par studio de production :
    - X = Nom producteurs

# Démo

# Limites

# Limites

**Base de données trop lourde** (30K+ colonnes) **:**

⇒ Lenteurs dans l'exécution du code

**Volonté de donner un poids différents aux variables :**

⇒ Le modèle KNN tel qu'on l'utilise ne le permet pas

# Recommandations

# Recommandations

- **Trouver d'autres modèles** afin de ne pas avoir autant de colonnes et pouvoir donner plus d'importance à certaines variables

  ⇒ un modèle possible serait le Scipy scalar (hot encoding)

- **Utiliser la méthode GridSearch / Randomized Search**

# Next Steps

# Next Steps

**Pour apporter plus de services:**

- Création d'un système de recommandation fondé sur la consommation des autres clients ⇒ User based
- Création d'un système qui récupère les recommandations faites aux utilisateurs et donne des indications sur les films à prochainement mettre à l'affiche
- Création d'un système de recommandation fondé sur les deux systèmes précédents ⇒ Hybrid based
- SMS des recommandations de films après avoir acheté une place pour une séance au cinéma.

# CaeaFlix project

*Algorithme de recommandation de films*