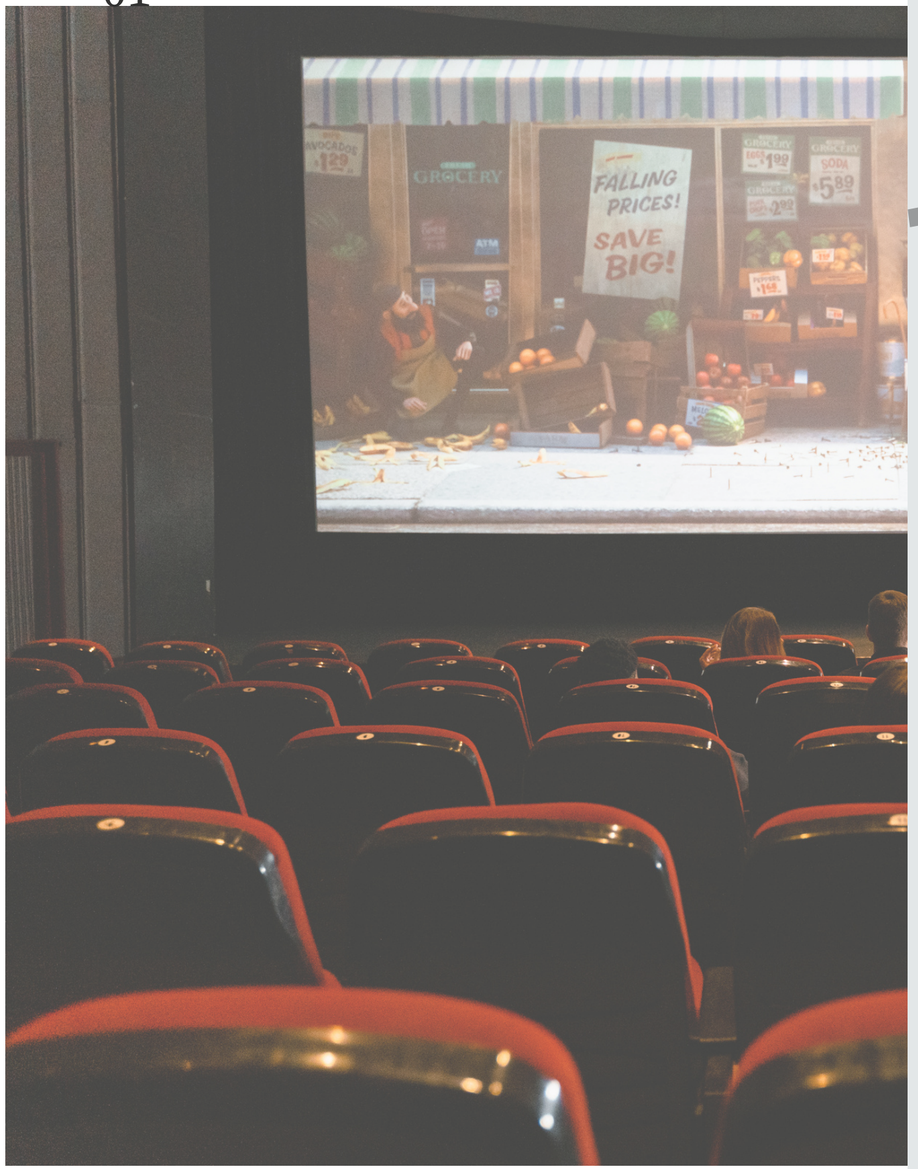
01

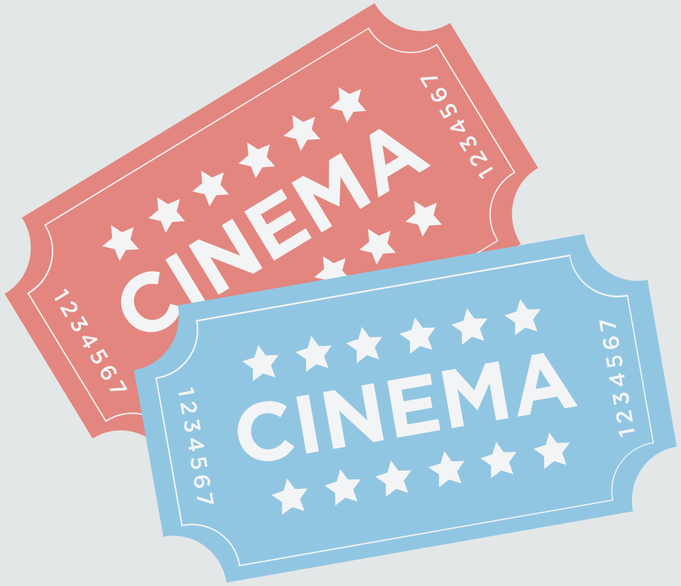


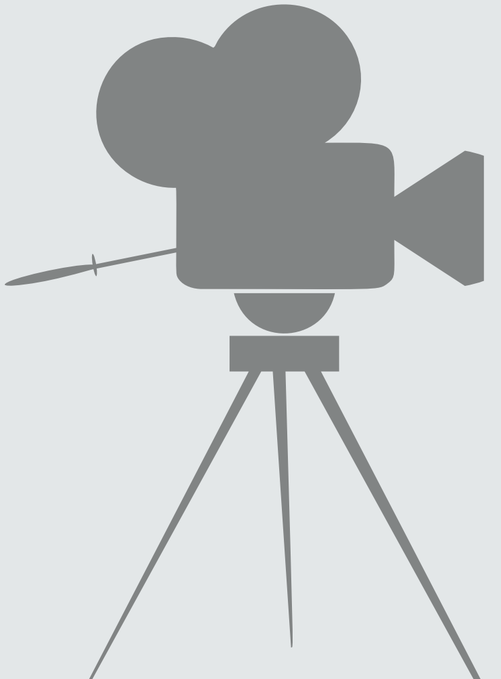
Data Project : June oral Presentation

Bouchez Hugo, Guilliams Laurie, Lemoine Estelle

Motivations

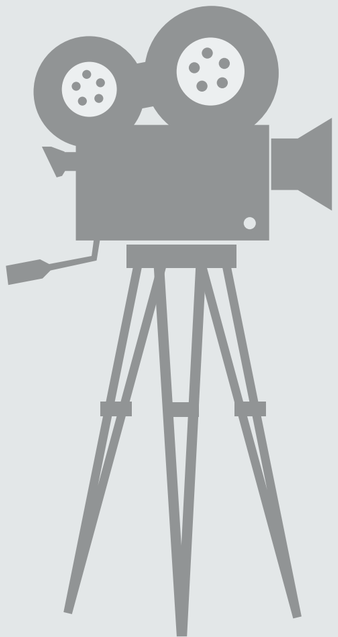
- We are interested in the cinema.
- Nowadays, films of all genres, all nationalities and with all types of actors are released every month.
- There are many websites that classify and rate these films, but on what basis?





Question we are insterested in

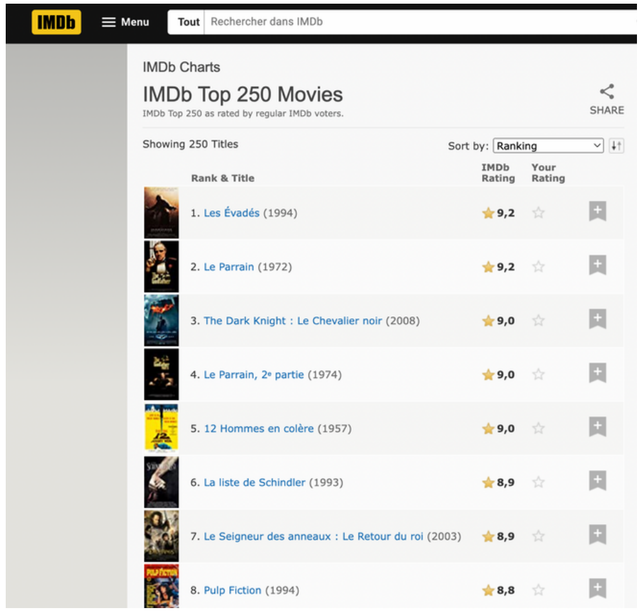
Can you predict the rating of upcoming movies ?

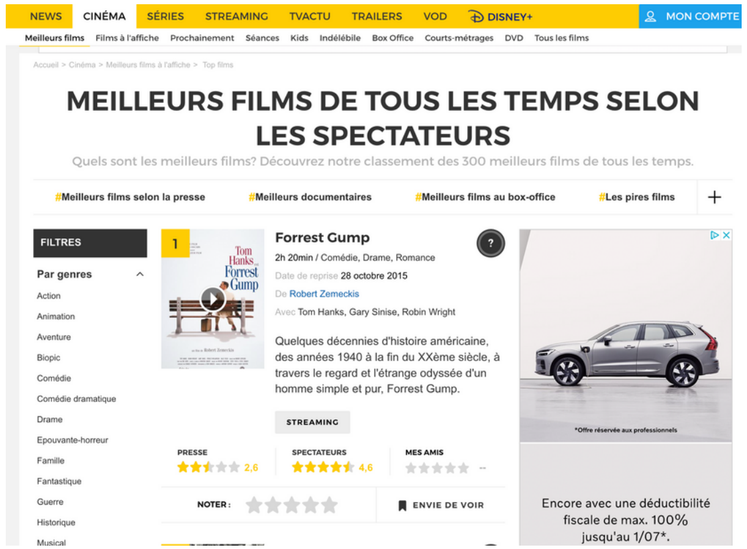


Data sources and data work

- **IMBD : We scraped the top 250 movies from this site**
- **Allociné : We scraped the list of the most famous actors on the Allociné site**
- **IMBD : We scraped the data of the upcoming movie releases on the IMBD website in order to predict their rating based on success factors such as genre, director or actors**

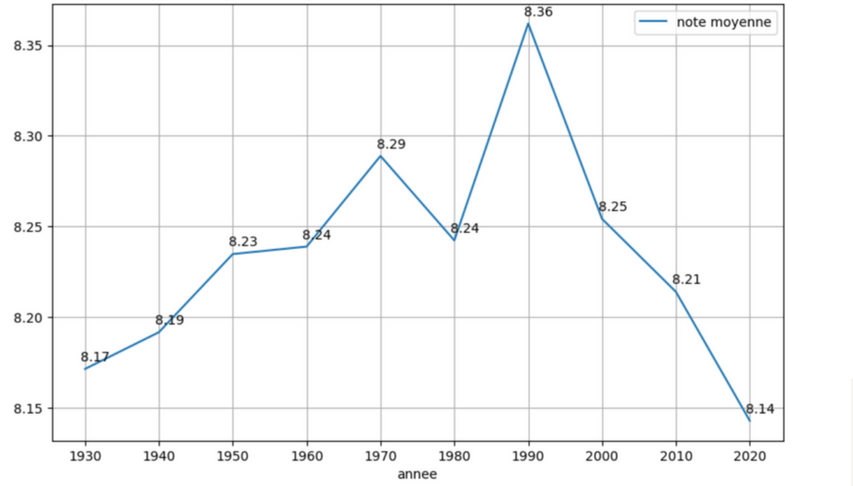
IMBD site

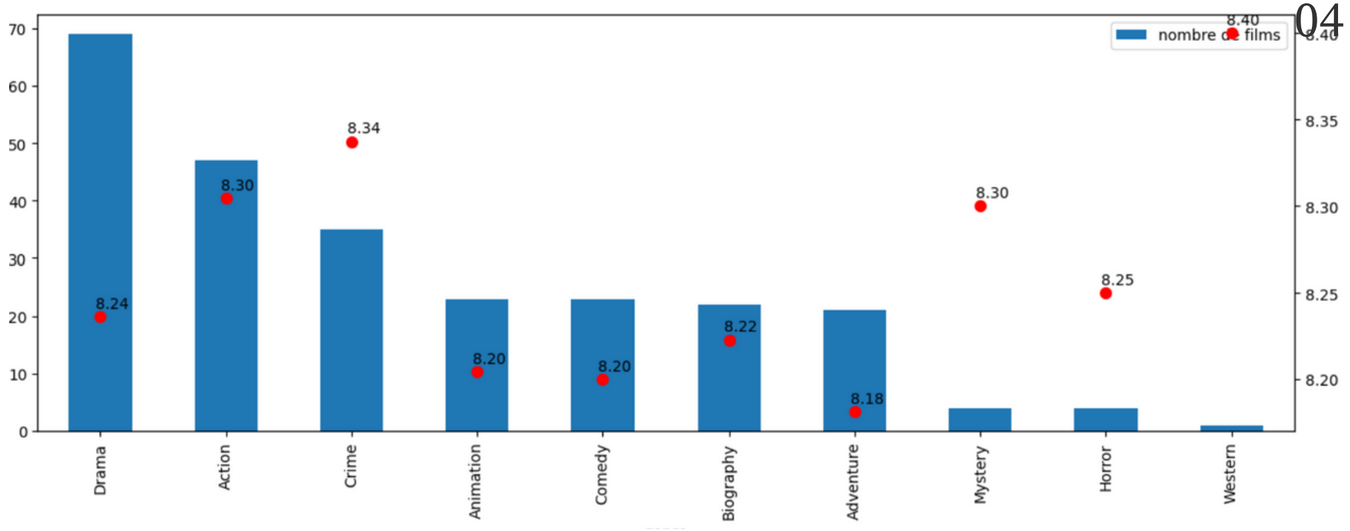


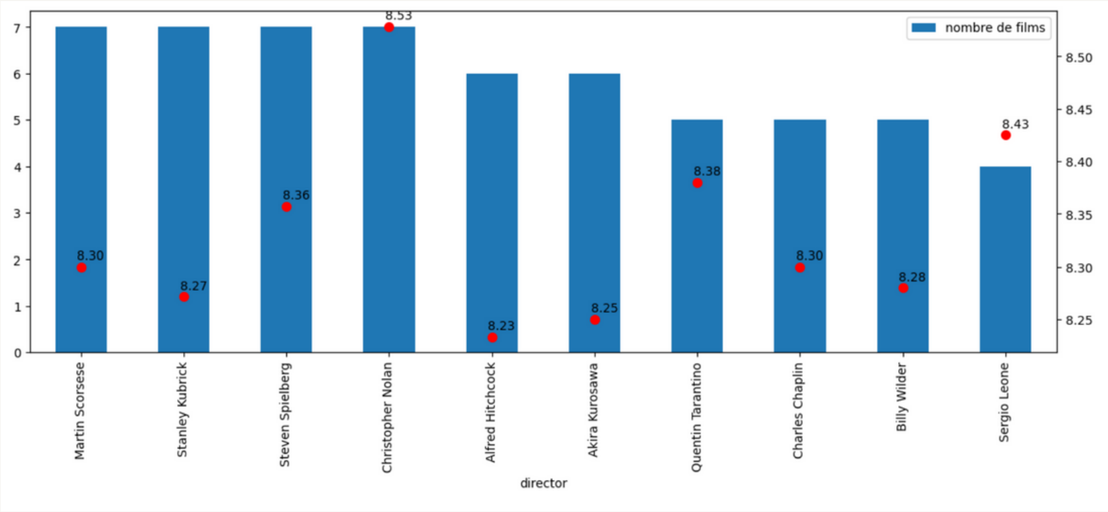


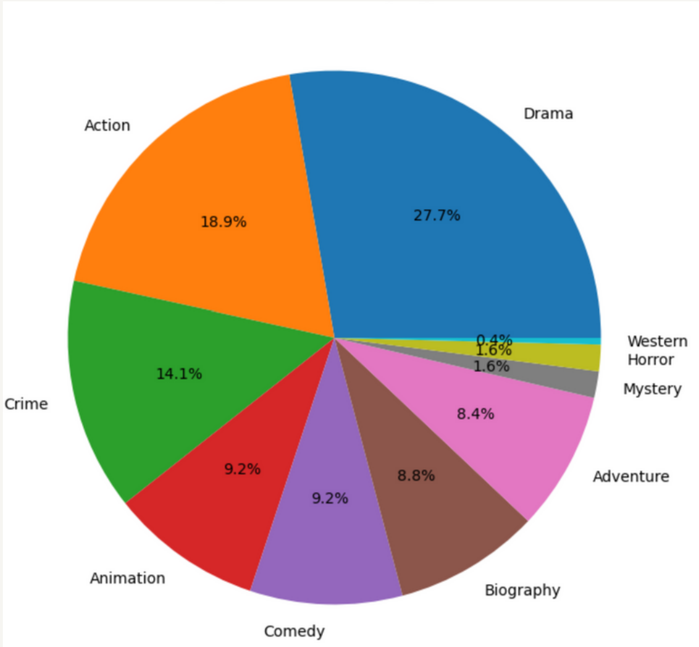
Allociné site

visualisation









Modeling

- **Training : Training a linear regression model to predict the movie ratings of a complete database (top 250)**
- **Evaluating : Evaluating the model with a test dataset**
- **Predicting : Using the linear regression model found during the training to find the movie rating of an upcoming movie dataset in which there is no rating yet**

```
# Encoder les variables catégorielles (genre) en variables numériques
label_encoder = LabelEncoder()
df['genre_encoded'] = label_encoder.fit_transform(df['genre'])
```

```
# Préparer les variables d'entrée (features) et de sortie (target)
X = df[['director get a golden globes ?', 'how many famous actors are in it?', 'genre_encoded']]
y = df['rating']
```

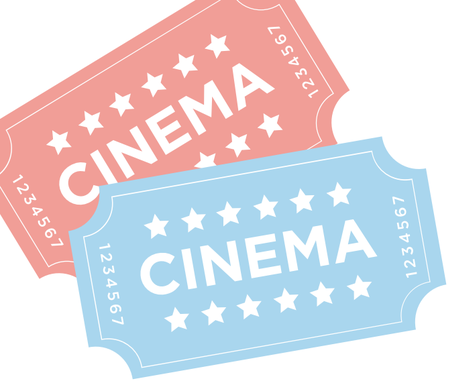
```
# Encoder la variable binaire 'director get a golden globes ?' en variable numérique
X['director_get_gg_encoded'] = X['director get a golden globes ?'].map({'yes': 1, 'no': 0})
```

```
# Supprimer la colonne originale 'director get a golden globes ?'
X.drop('director get a golden globes ?', axis=1, inplace=True)
```

```
# Effectuer un encodage one-hot pour la variable catégorielle 'genre_encoded', Cela permet de traiter les différentes catég
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), ['genre_encoded'])], remainder='passthrough')
X_encoded = ct.fit_transform(X)
```

```
# Créer un modèle de régression linéaire
model = LinearRegression()

# Entraîner le modèle
model.fit(X_encoded, y)
```



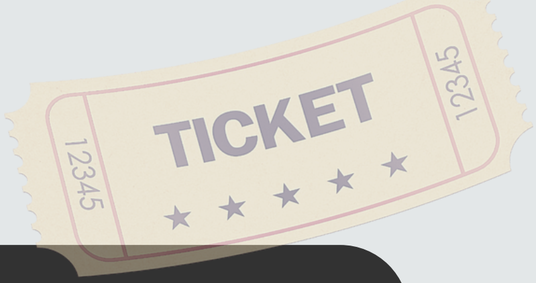
The result of our question

Can you predict the rating of upcoming movies ?


```python
# Afficher les résultats
print(df_sorties[['title', 'predicted_rating']])
```

```
                                                title  predicted_rating
0                Spider-Man: Across the Spider-Verse (2023)          8.314575
1                                  The Boogeyman (2023)          8.340605
2                              To Catch a Killer (2023)          8.353687
3                 Transformers: Rise of the Beasts (2023)          8.353687
4                                       Strays (2023)          8.314575
5                                  Spinning Gold (2023)          8.270464
6                                      The Flash (2023)          8.353687
7                               About My Father (2023)          8.296912
8                                     Elemental (2023)          8.314575
9                                   Asteroid City (2023)          8.296912
10                               No Hard Feelings (2023)          8.296912
11          Indiana Jones and the Dial of Destiny (2023)          8.353687
12                   Ruby Gillman, Teenage Kraken (2023)          8.314575
13                   Harold and the Purple Crayon (2023)          8.314575
14                        Insidious: The Red Door (2023)          8.340605
15  Mission: Impossible - Dead Reckoning Part One ...              8.353687
16                               The Last Warrior (2021)          8.314575
17                                   Oppenheimer (2023)          8.270464
18                                        Barbie (2023)          8.247797
19                                 Coyote v. Acme (2023)          8.314575
20                                Haunted Mansion (2023)          8.296912
21                      How to Blow Up a Pipeline (2022)          8.411425
22                          The Meg 2: The Trench (2023)          8.353687
23  Teenage Mutant Ninja Turtles: Mutant Mayhem (2...              8.314575
24                                   Gran Turismo (2023)          8.353687
25                               The Inseparables (2023)          8.314575
26                                    Blue Beetle (2023)          8.353687
27                                The Equalizer 3 (2023)          8.353687
```

Conclusion



In conclusion, we scraped information on the top actors, the best films, and upcoming movie releases. Using this information, we were able to make progress in the visualization and modeling aspects, ultimately arriving at an answer to our research question: "Can you predict the rating of upcoming movies?".

Moreover, all three of us were novices in the skills involved. It hasn't always been easy, but we conducted research to get up to date.