Parcours Datascientist : projet 3

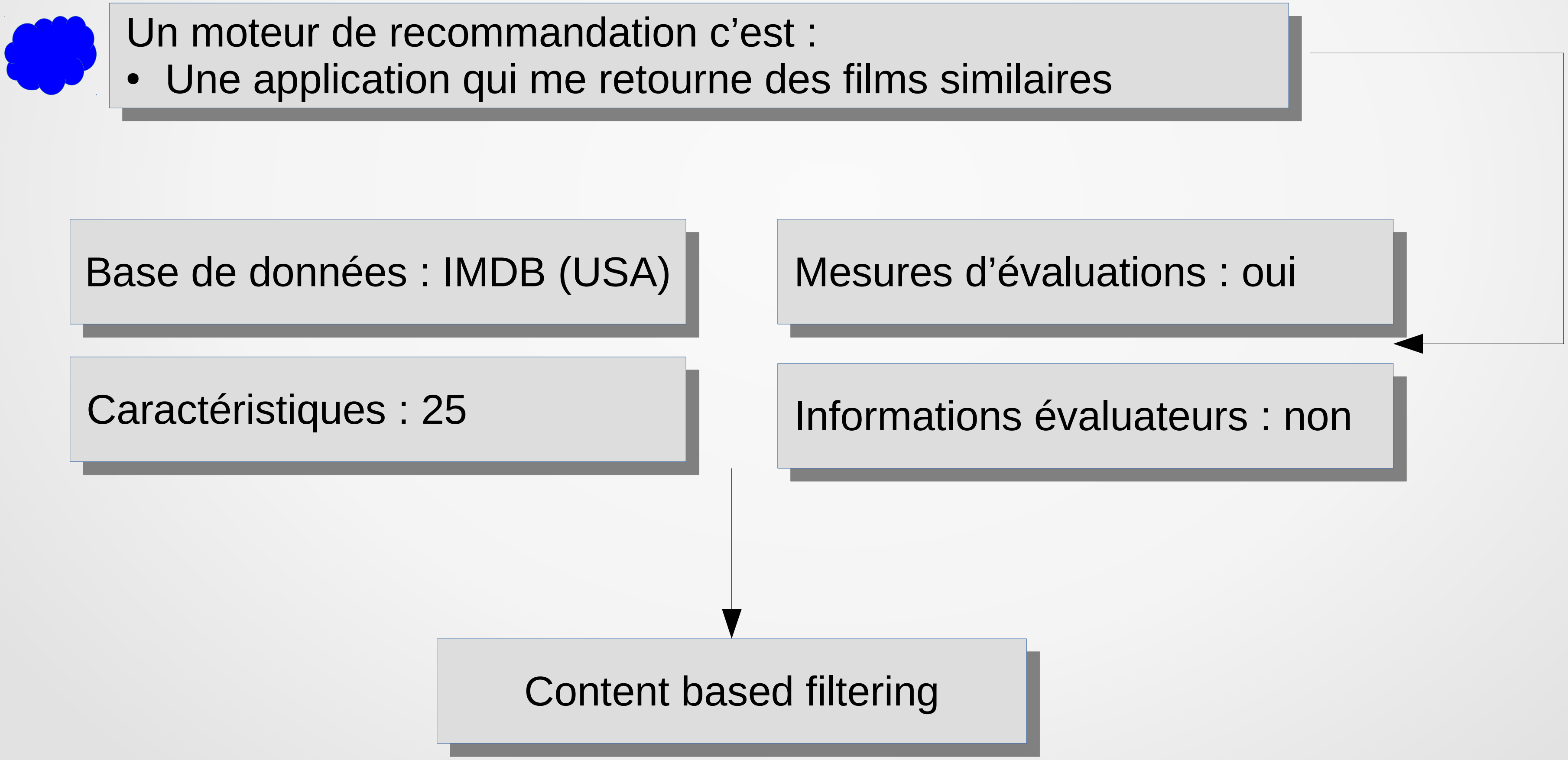
Moteur de recommandations

Francois BANGUI

Formulation du problème

♥ Un moteur de recommandation c'est :

- Une application qui me retourne les films que j'apprécie



☁ Un moteur de recommandation c'est :

- Une application qui me retourne des films similaires

Base de données : IMDB (USA)

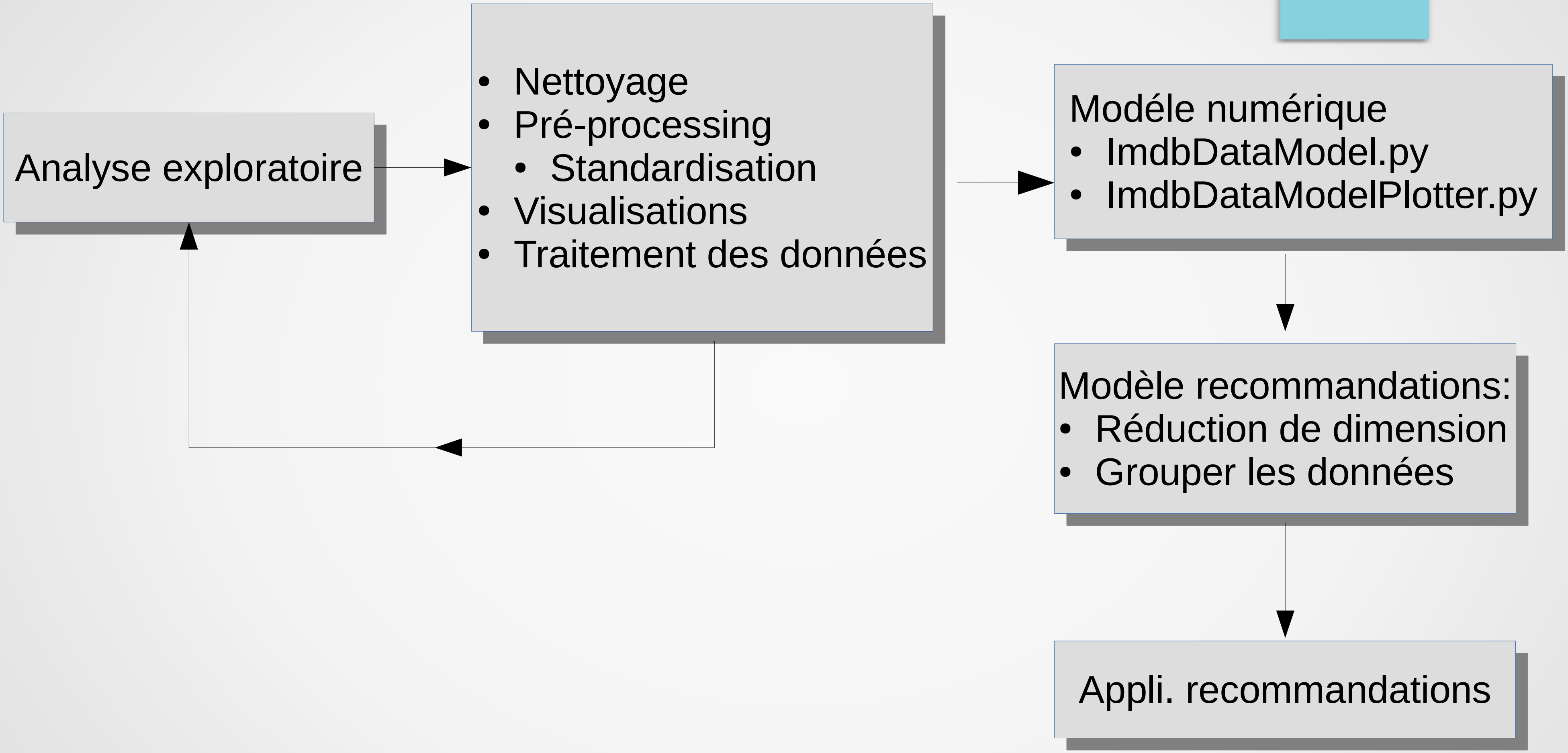
Mesures d'évaluations : oui

Caractéristiques : 25

Informations évaluateurs : non

Content based filtering

Méthodologie



Analyse exploratoire

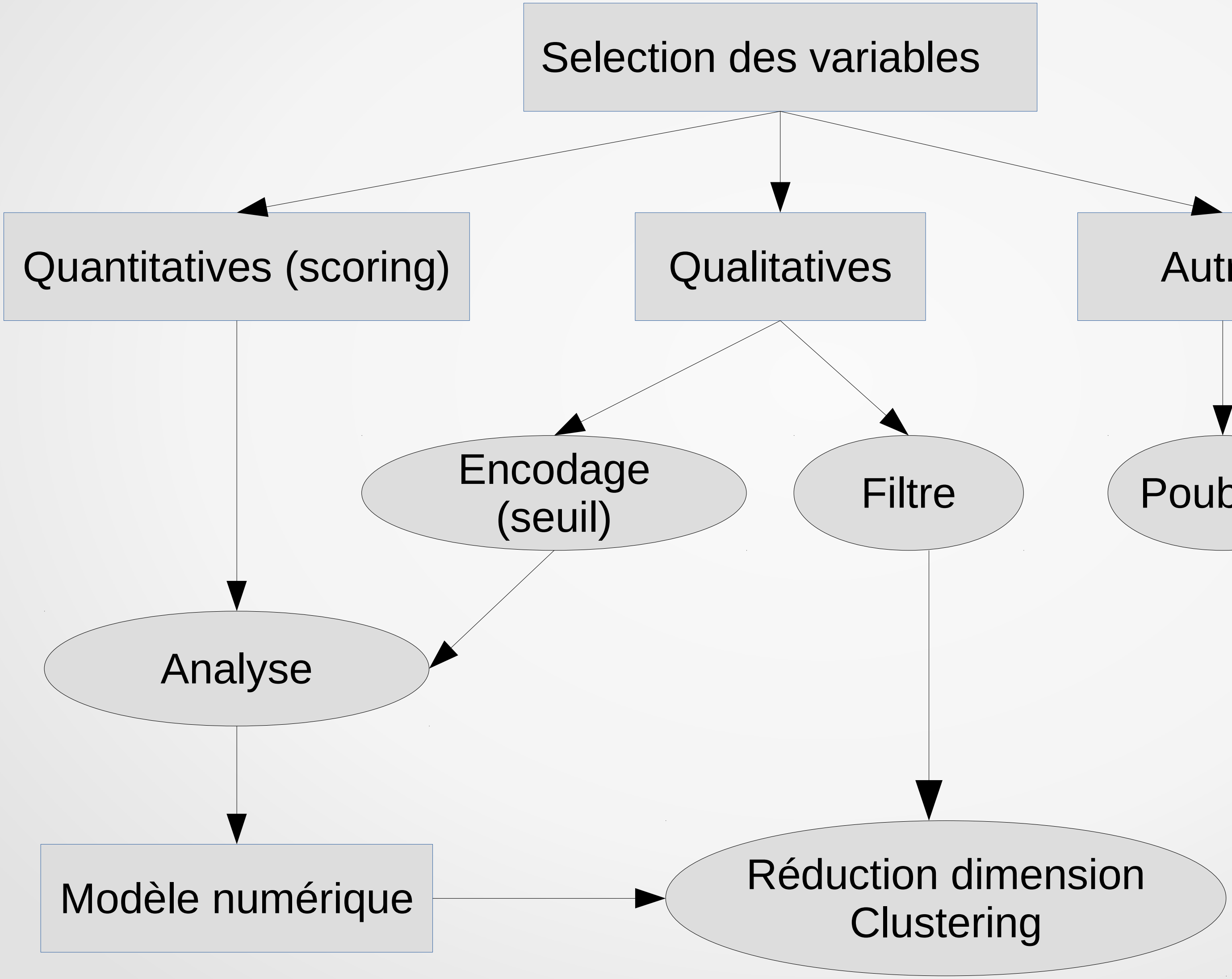
- Nettoyage
- Pré-processing
  - Standardisation
- Visualisations
- Traitement des données

Modéle numérique
- ImdbDataModel.py
- ImdbDataModelPlotter.py

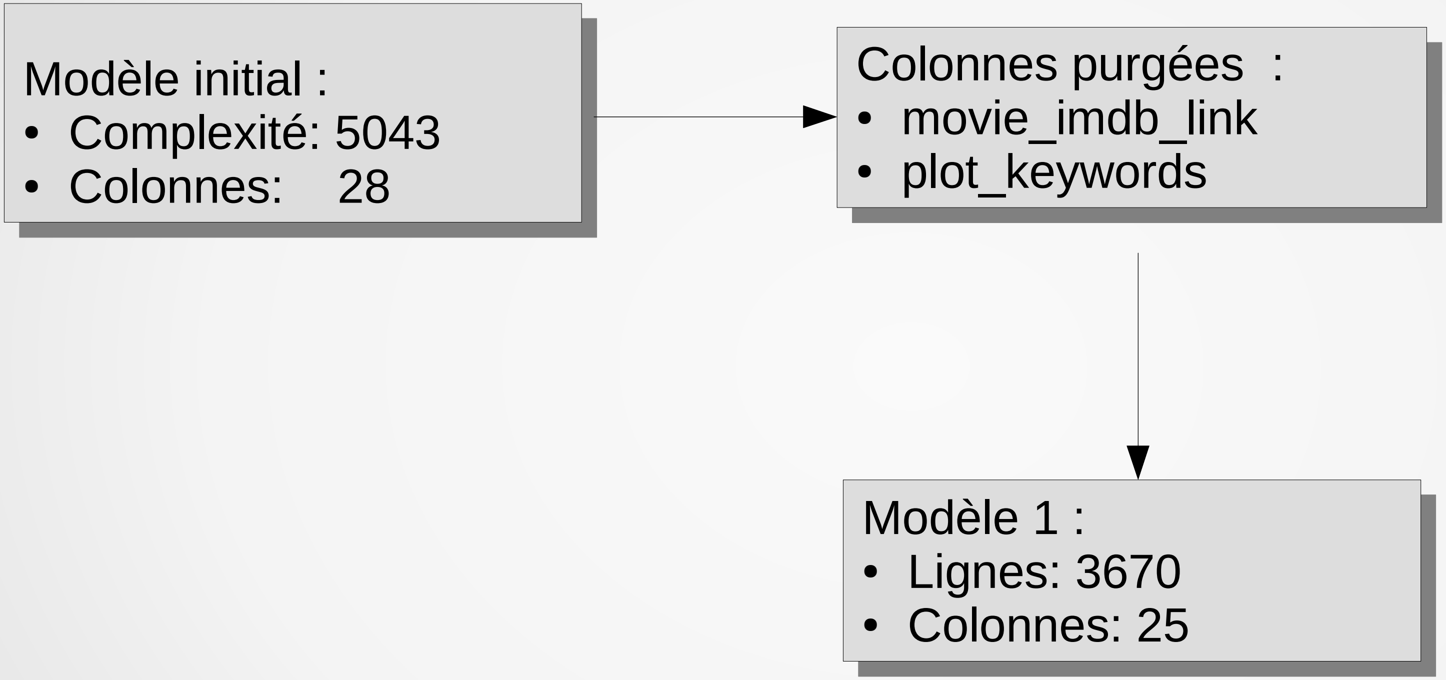
Modèle recommandations:
- Réduction de dimension
- Grouper les données

Appli. recommandations

Traitement des variables



Nettoyage 1 : lignes nan purgées



Modèle initial :

- Complexité: 5043
- Colonnes: 28

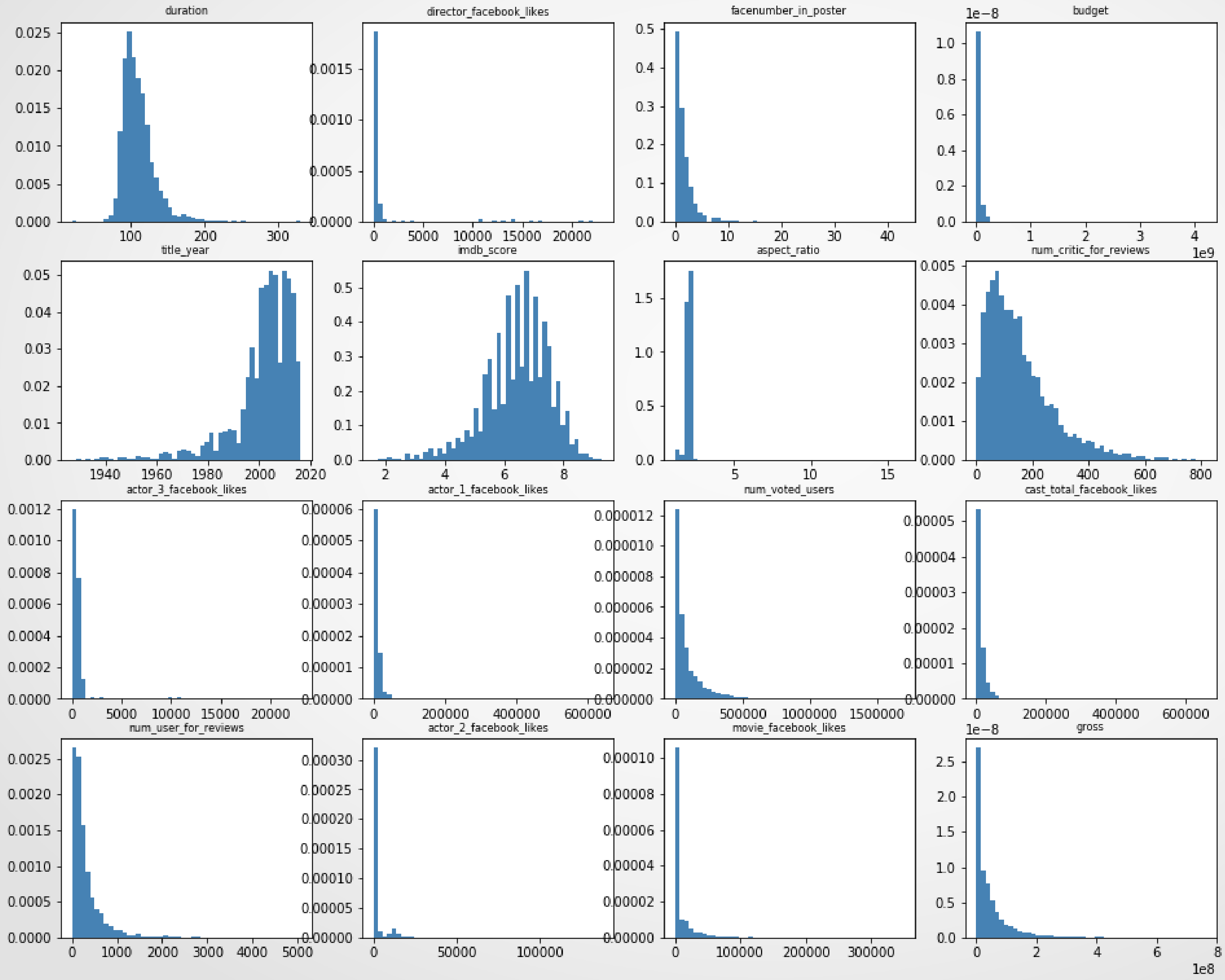
Colonnes purgées :

- movie_imdb_link
- plot_keywords

Modèle 1 :

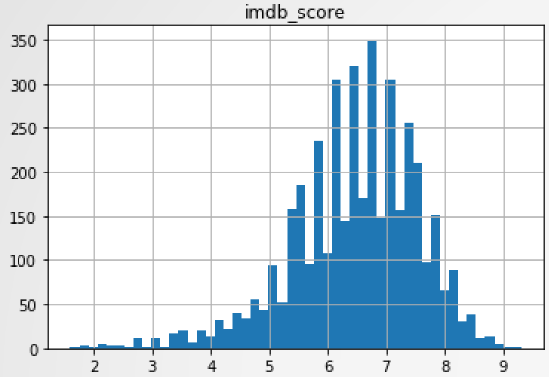
- Lignes: 3670
- Colonnes: 25

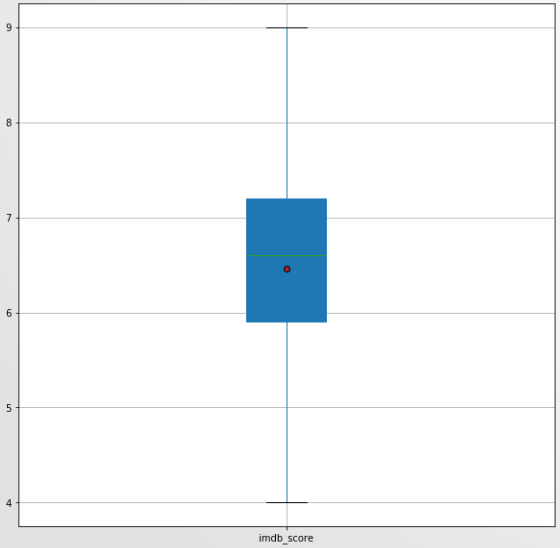
Distribution des scores

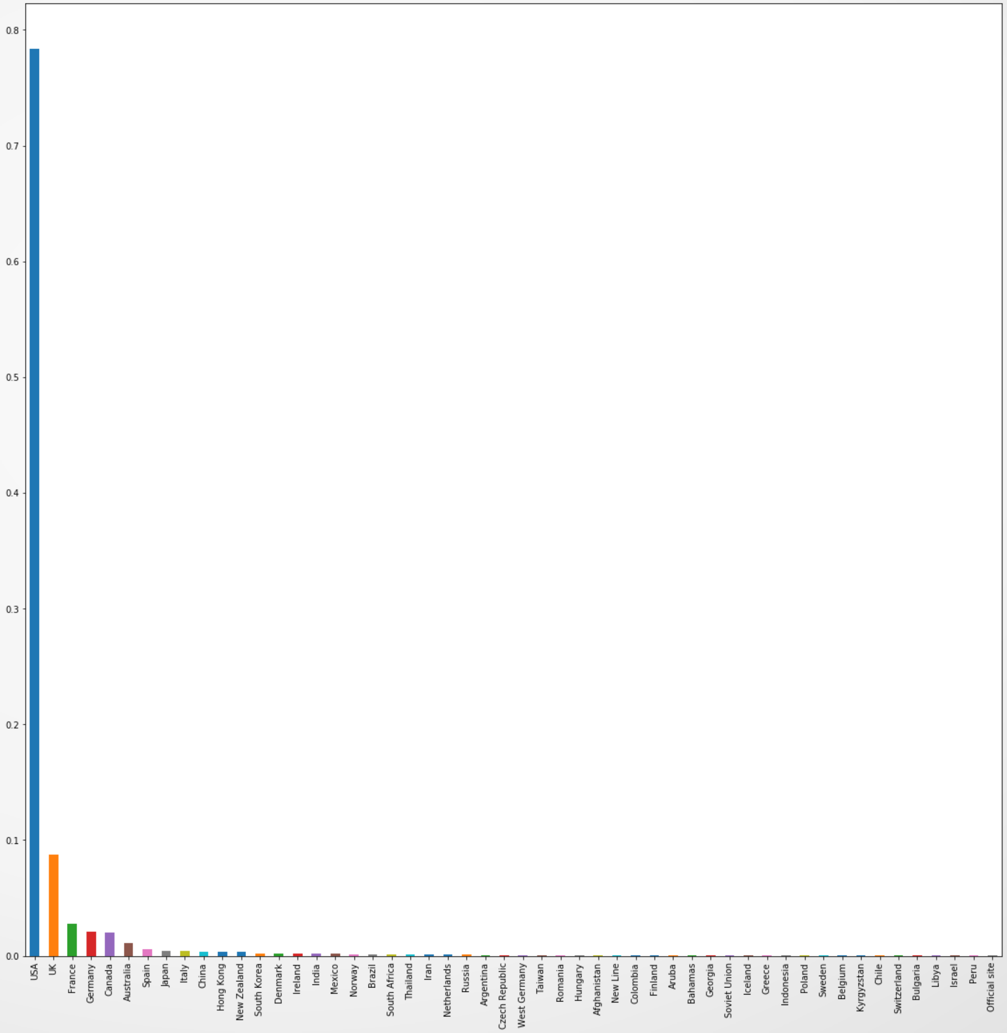


Analyse exploiratoire : IMDB

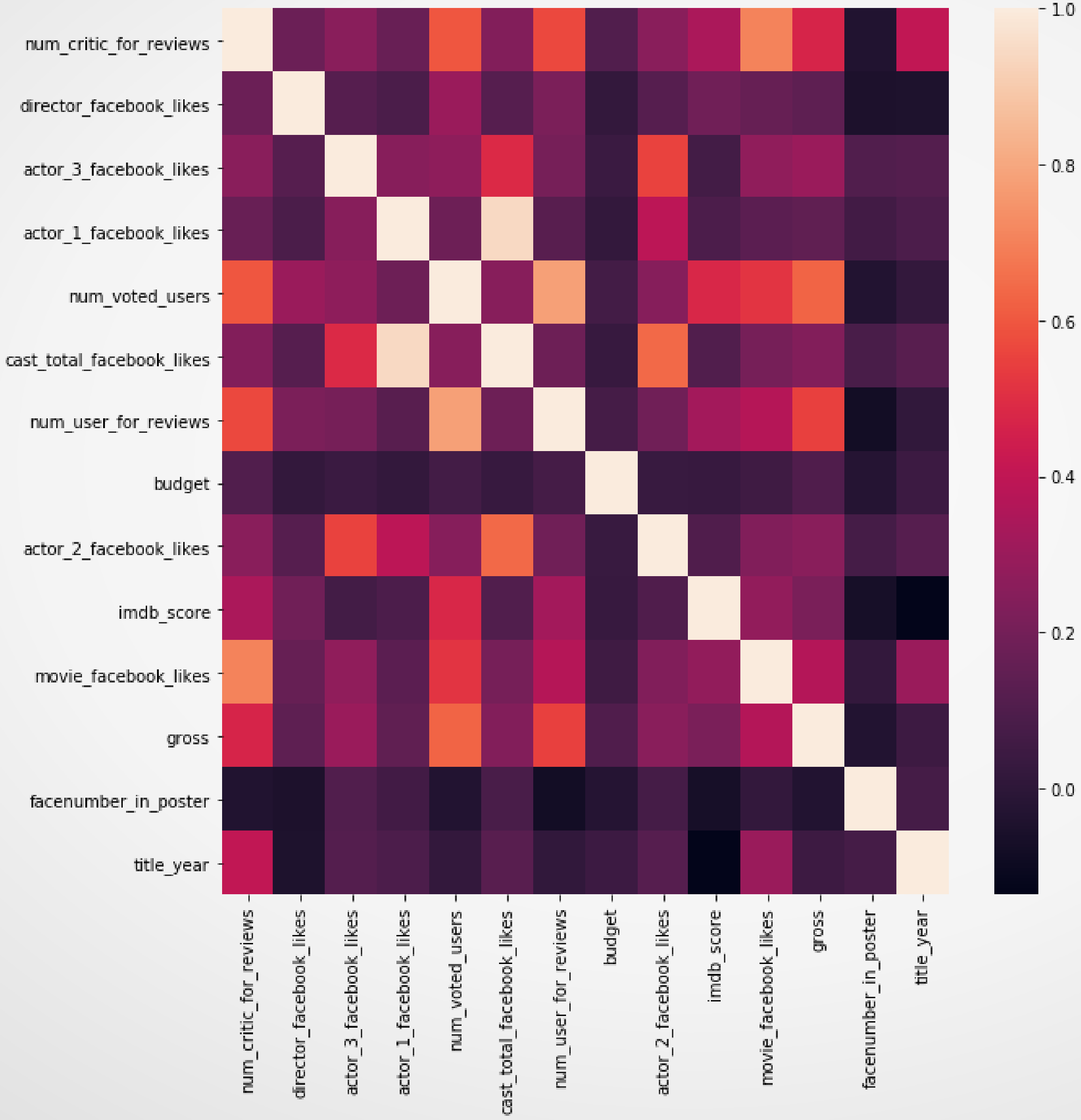
Score IMDB sur ~4000 films :



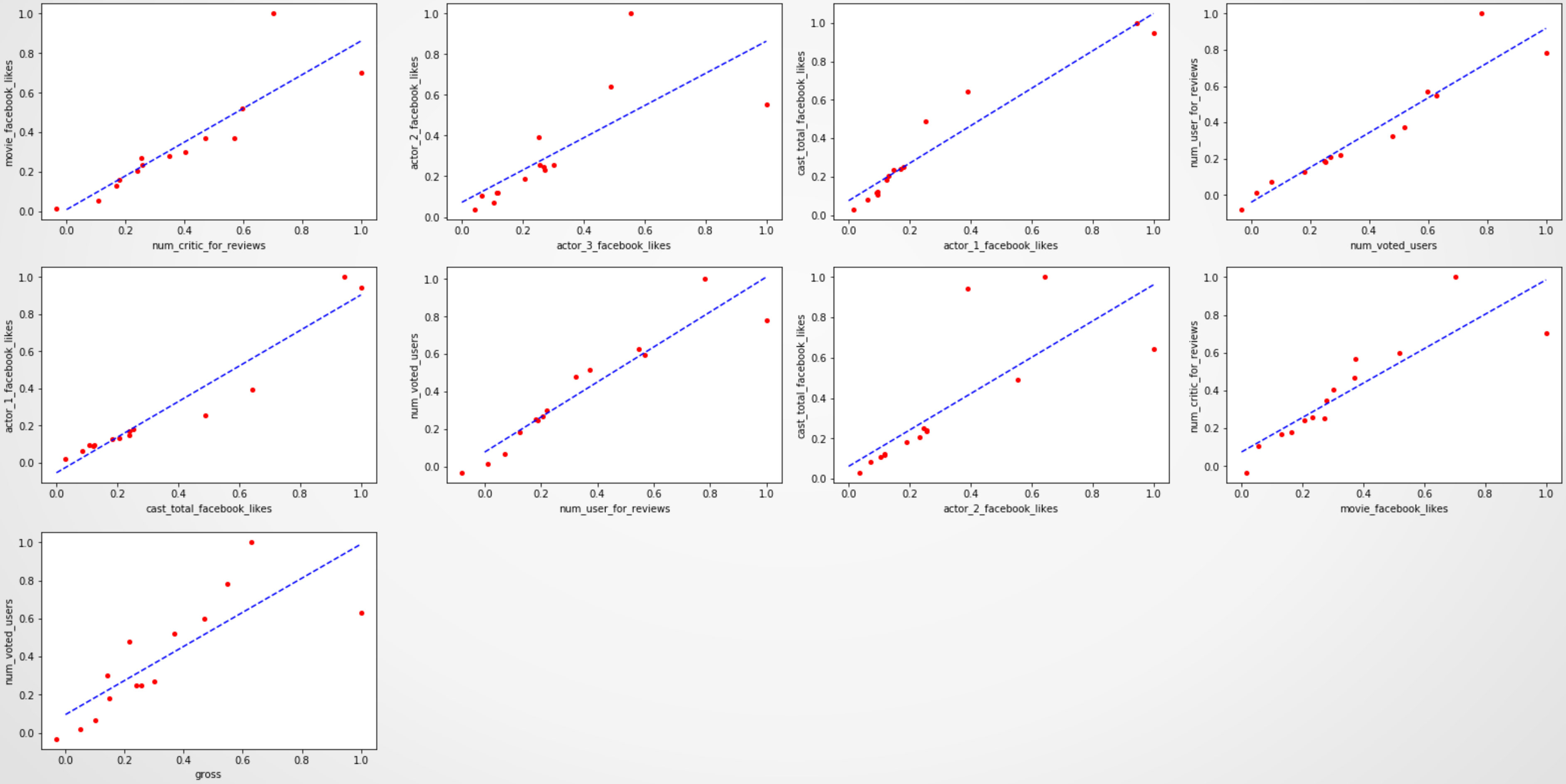




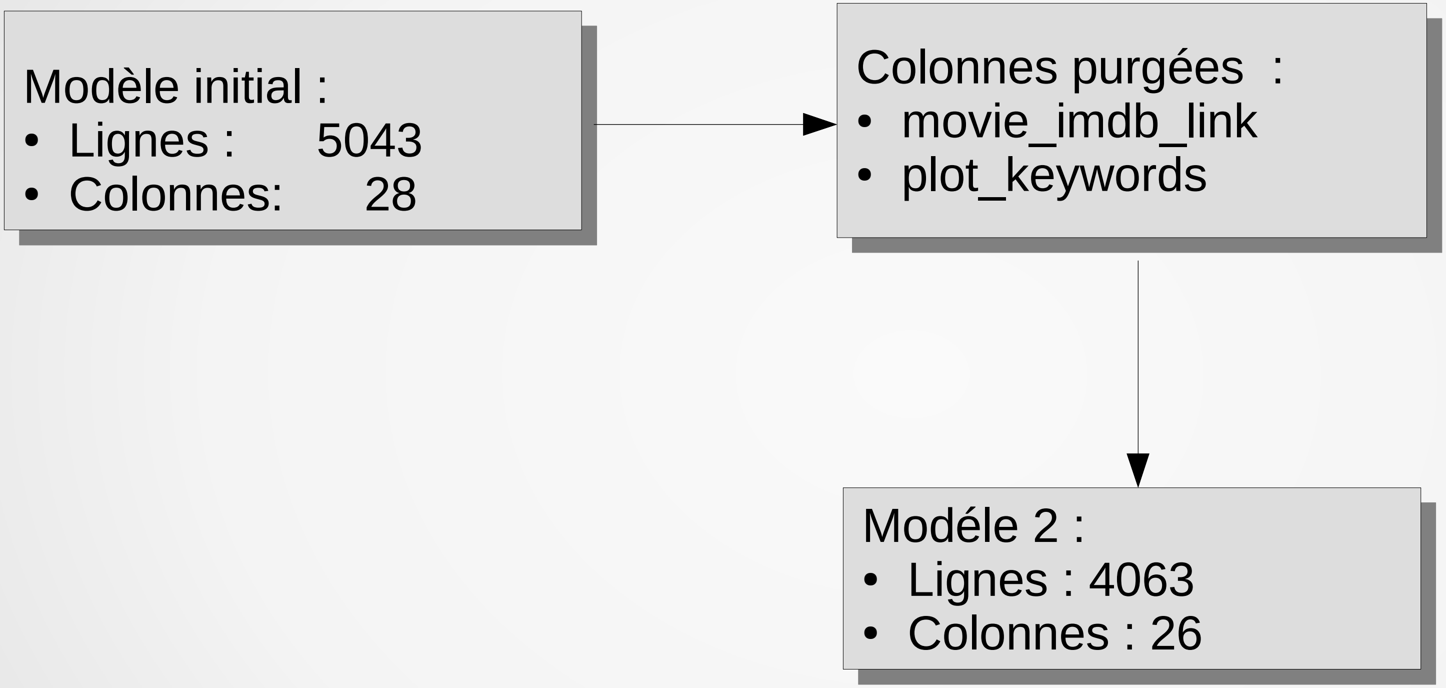
Analyse exploratoire : corrélations



Analyse exploratoire : Imputations



Nettoyage 2 : lignes nan purgées + imputations



Modèle initial :

- Lignes : 5043
- Colonnes: 28

Colonnes purgées :

- movie_imdb_link
- plot_keywords

Modéle 2 :

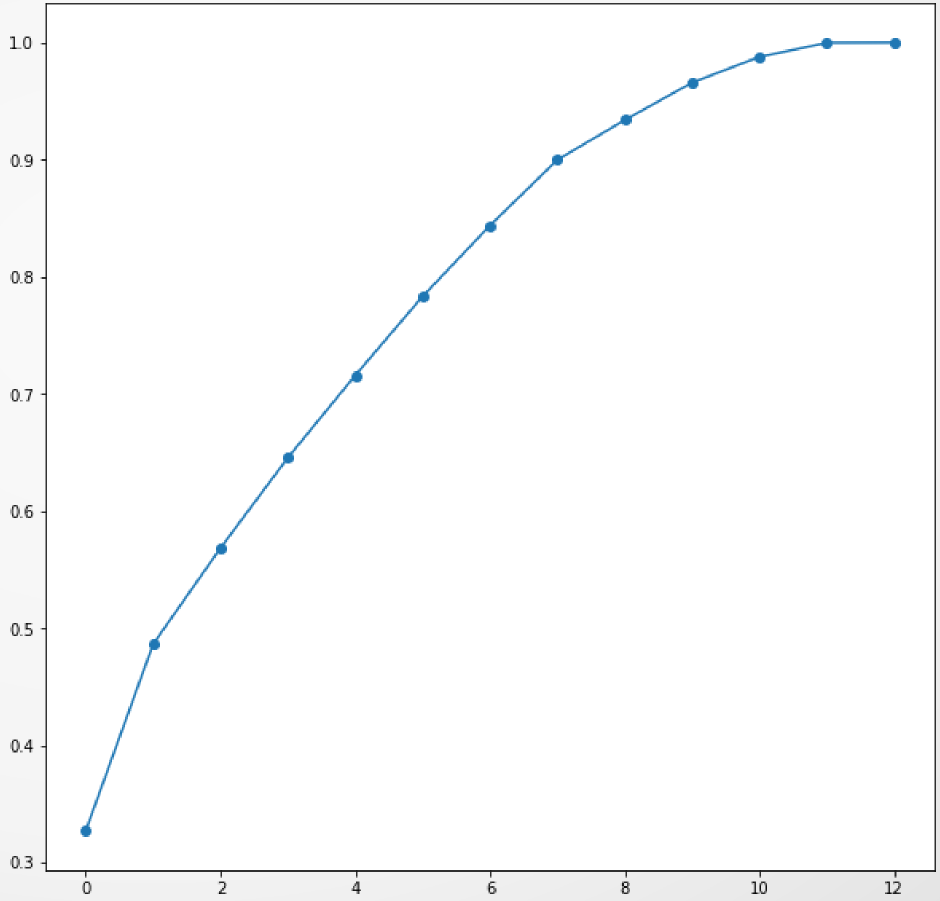
- Lignes : 4063
- Colonnes : 26

Régression linéaire à seuil : Pearson >=50 %   ⇒ Perte d’informations : de 28 % à 20 %

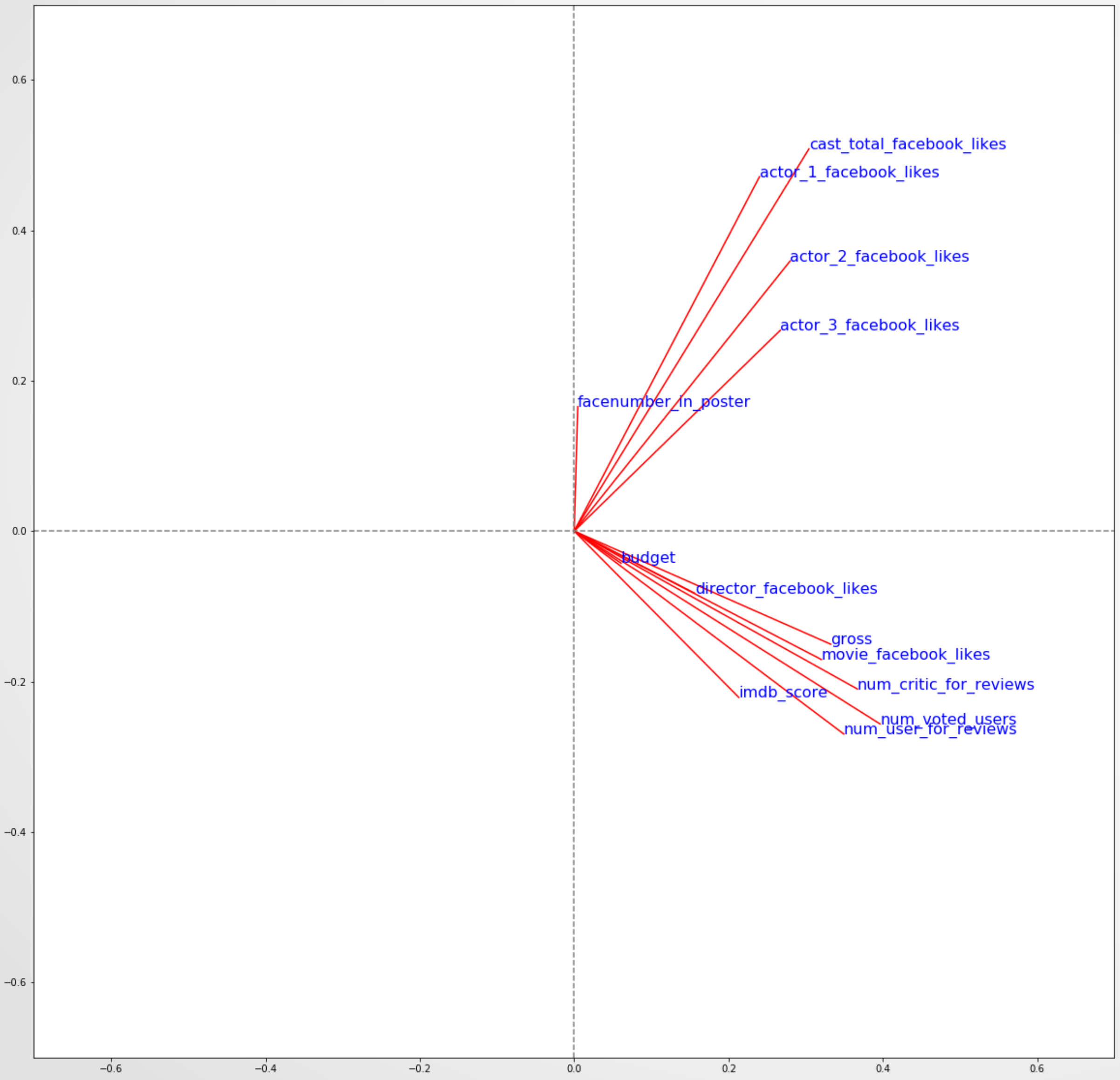
Analyse exploratoire : ACP (scoring)

Modéle numérique des scores standardisés:

- Lignes : 4068
- Colonnes : 13



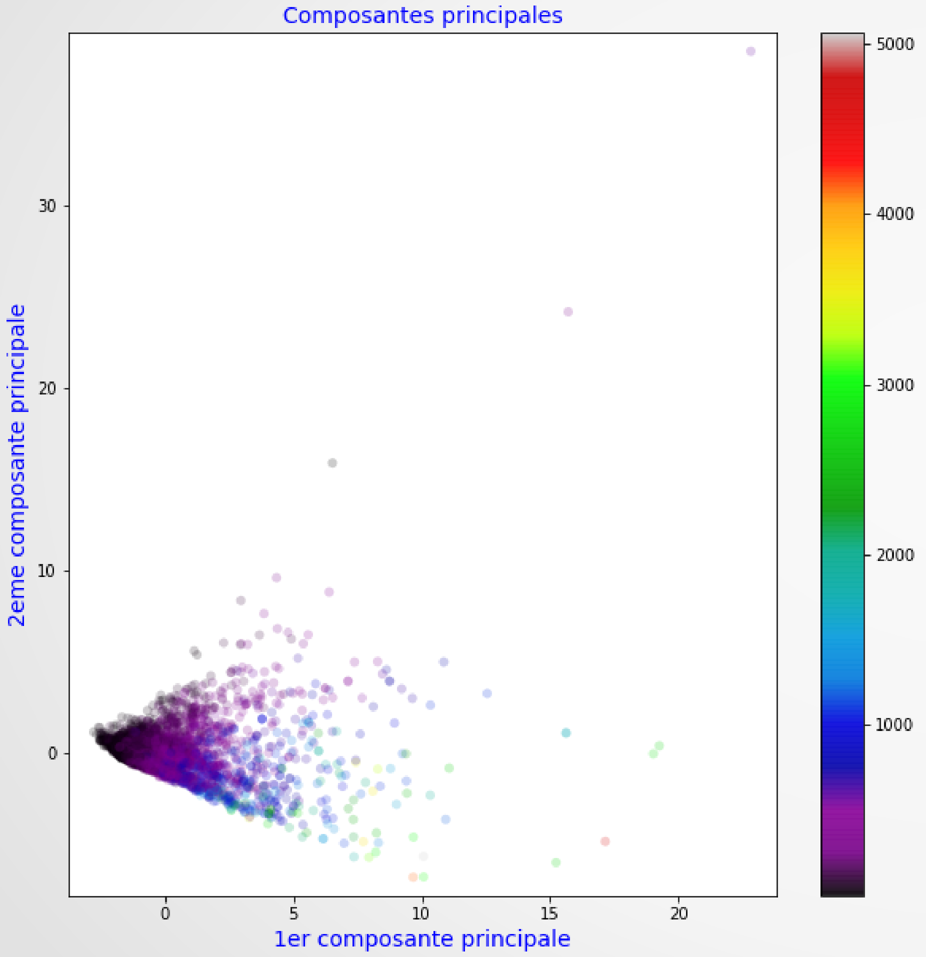
Analyse exploratoire : ACP (2)

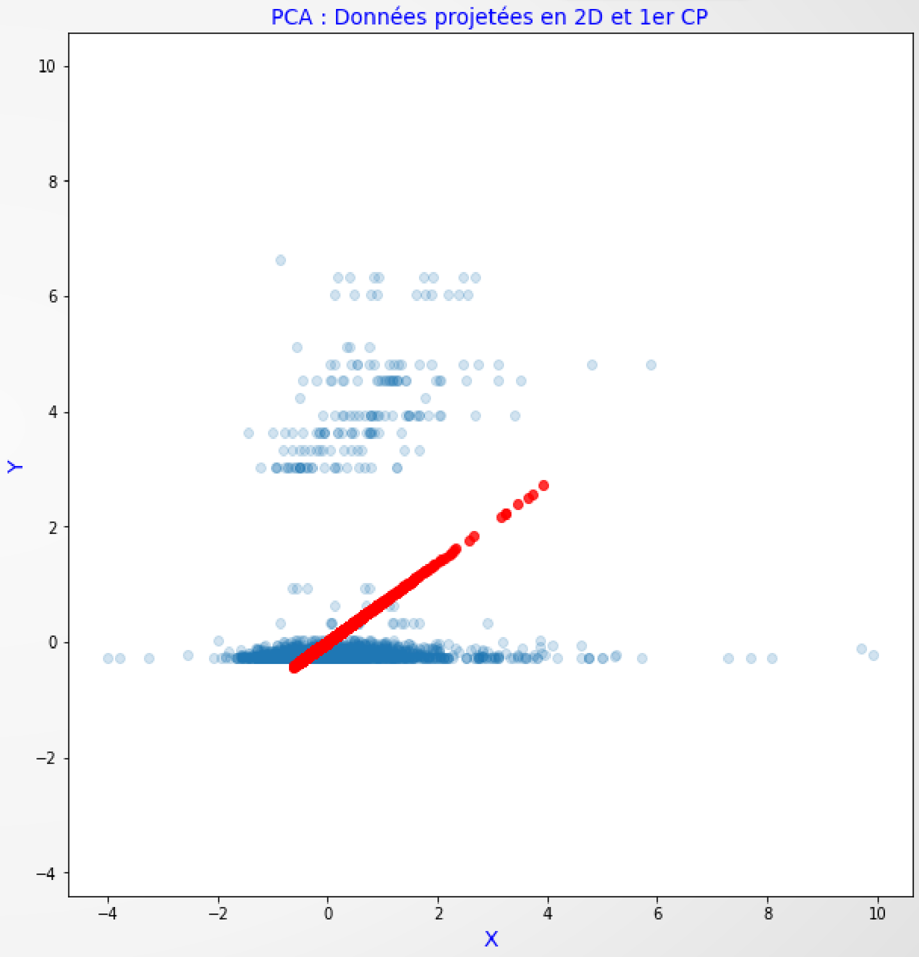


Corrélations :

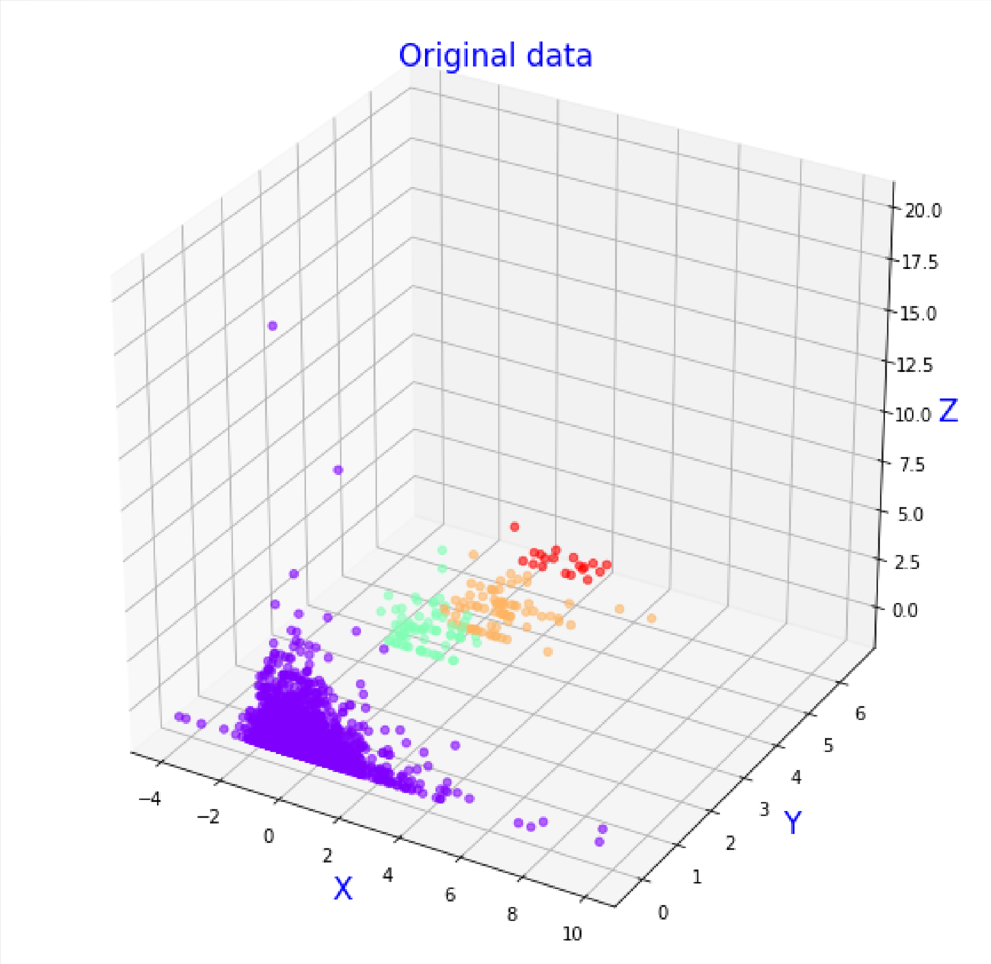
- movie_facebook_likes
- num_critic_for_reviews

Analyse exploratoire : ACP (3)

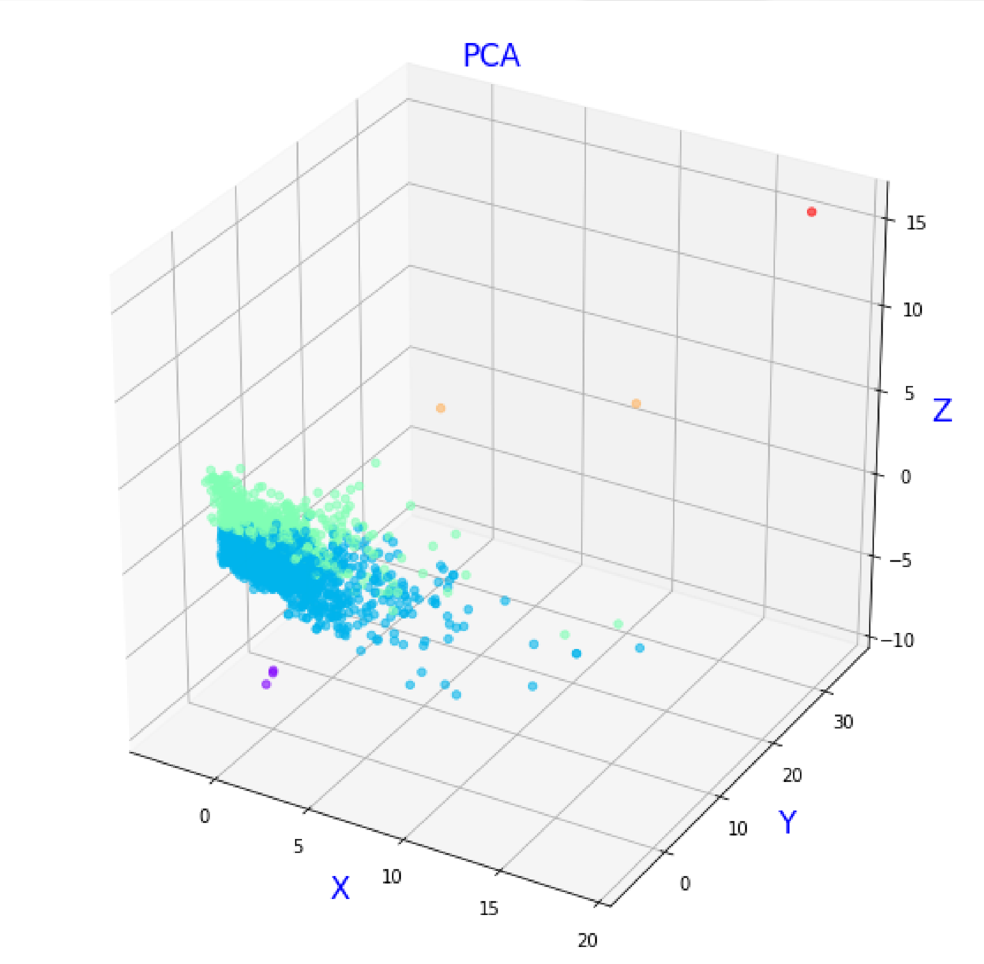




Analyse exploratoire : ACP (4)



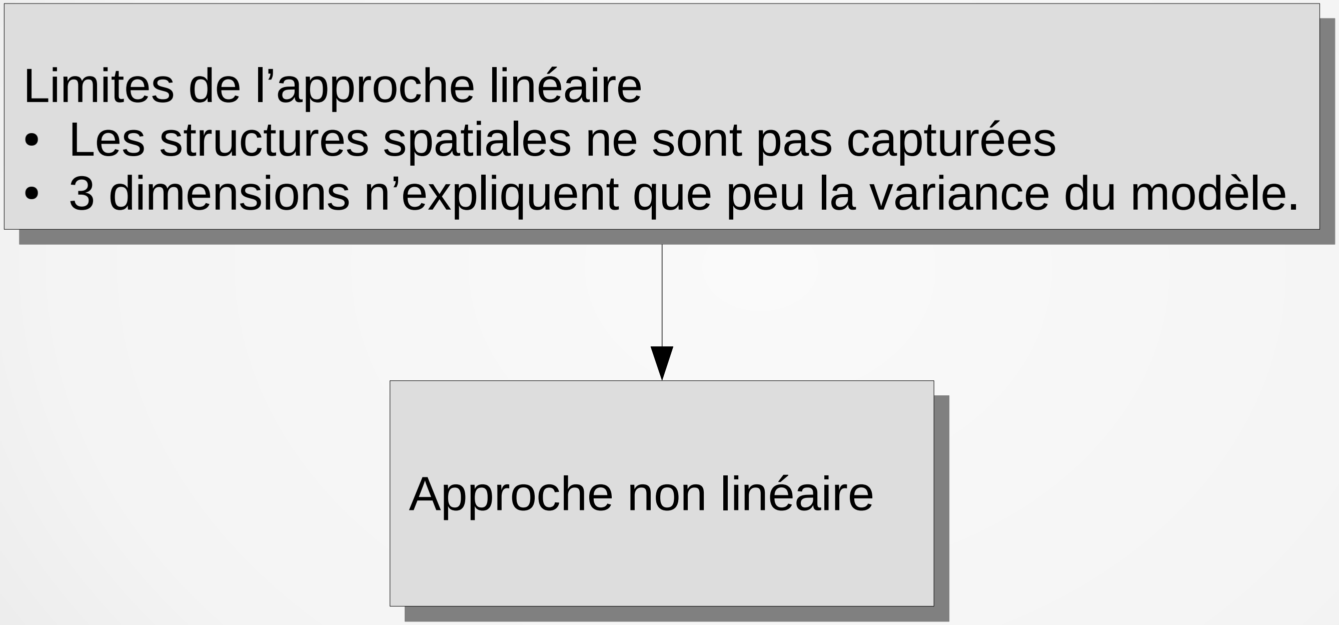
Original data



PCA

La représentation des données dans l’ espace originel 3D présente une structure géométrique non linéaire

Analyse exploratoire : bilan 1



Limites de l’approche linéaire

- Les structures spatiales ne sont pas capturées
- 3 dimensions n’expliquent que peu la variance du modèle.

Approche non linéaire

Traitement : one-hot encoding

	Avant	Augmentation	Après
color	26	+1	27
genres	27	+22	49
langues	49	+34	83
country	83	+53	136
content_rating	136	+14	150
Dimension du modèle	**150**		

Exclues
- director_name
- actor_1_name
- actor_2_name
- actor_3_name

Encodées
- movie_title
- Duration
- Genres
- Langage
- Country
- content_rating
- title_year
- color

Variables exclues
- movie_title
- plot_keywords

Acteurs et directeurs : distribution

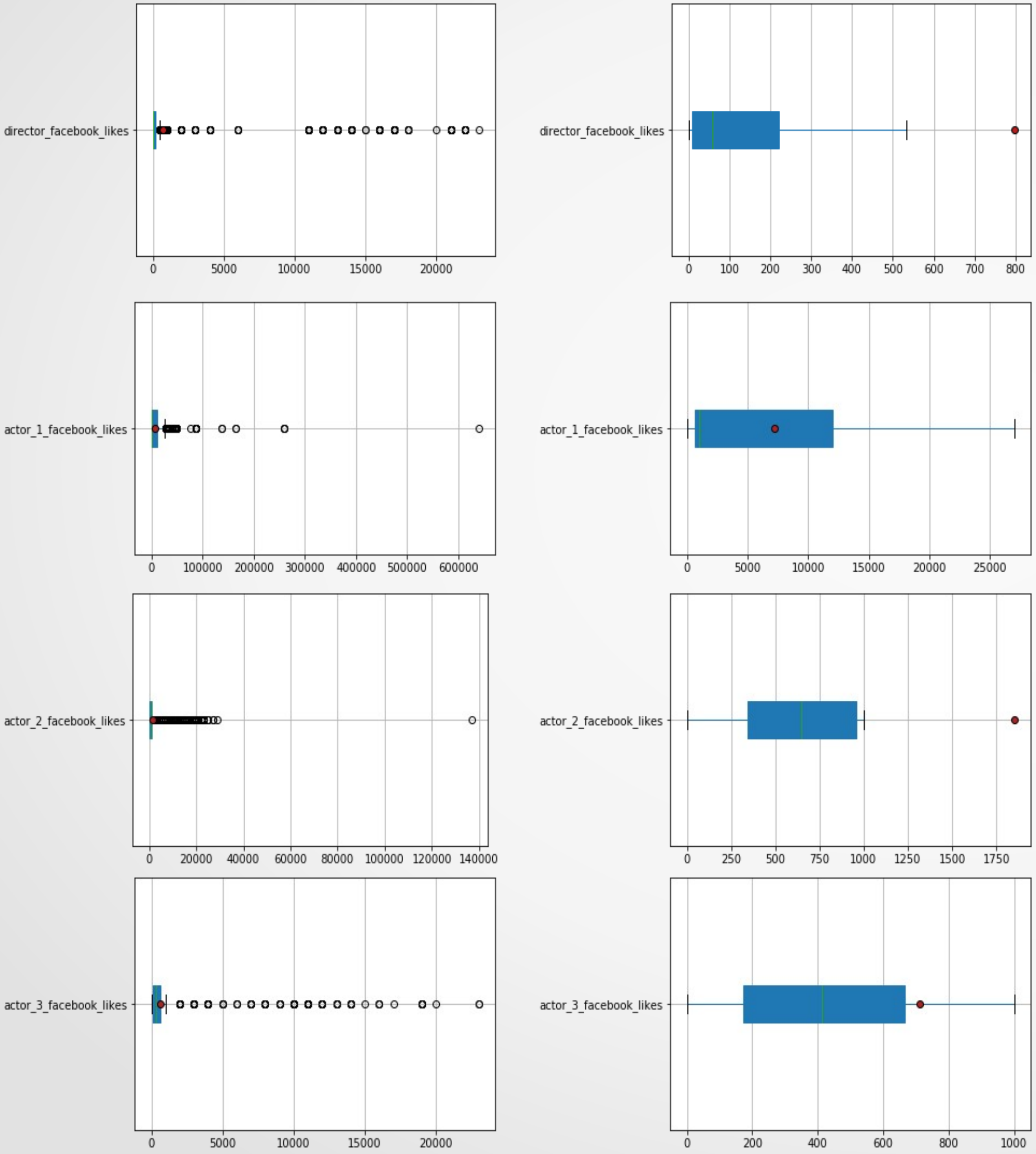
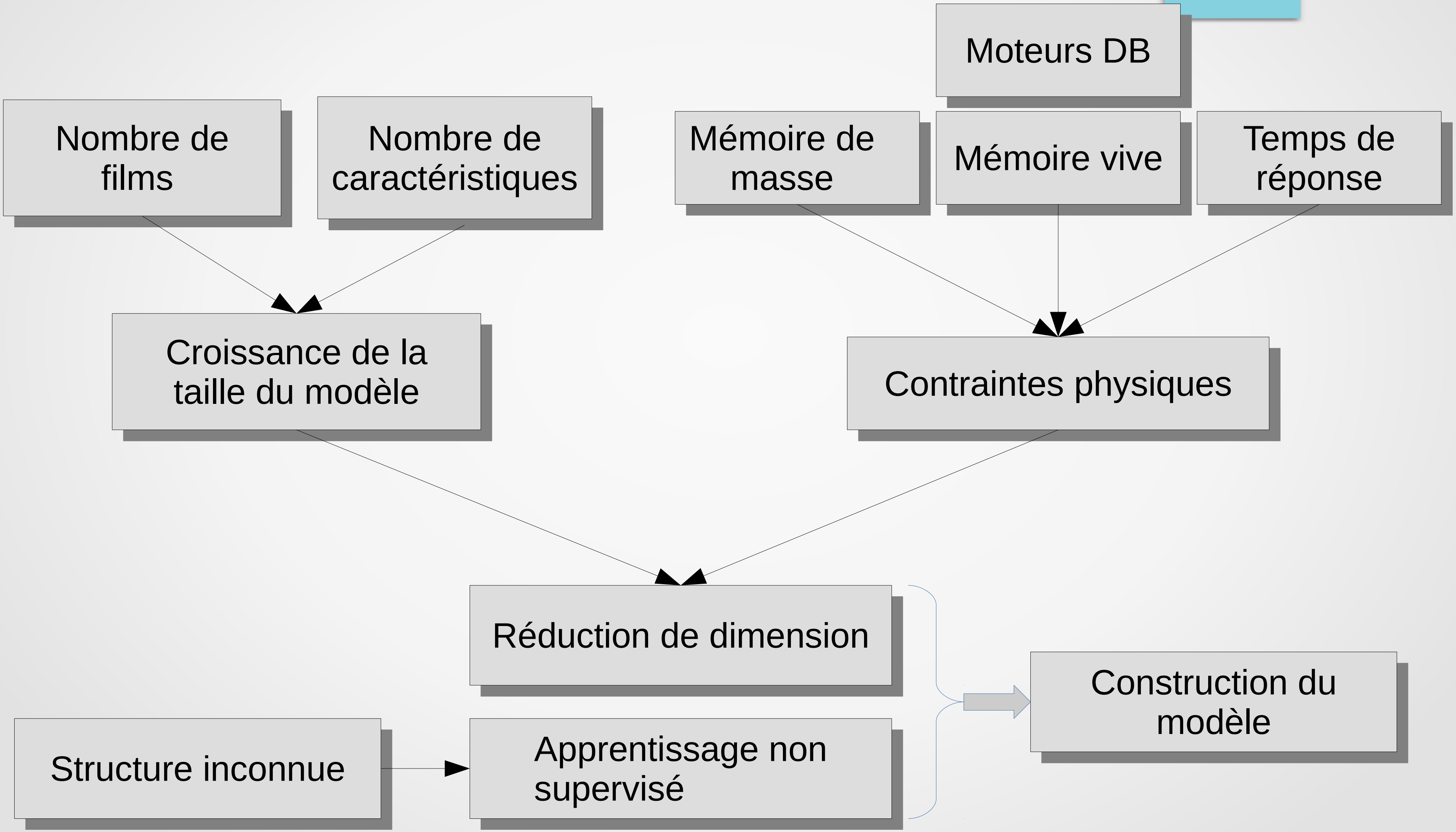


Schéma d’encodage à seuil :

Outliers : Facebook likes

	Avant	Augmentation	Après
director_name actor1_name actor2_name actor3_name	150	+283	433
Dimension du modèle	**433**		

Réduction de dimension et app. Non supervisé



Moteurs DB

Nombre de films

Nombre de caractéristiques

Mémoire de masse

Mémoire vive

Temps de réponse

Croissance de la taille du modèle

Contraintes physiques

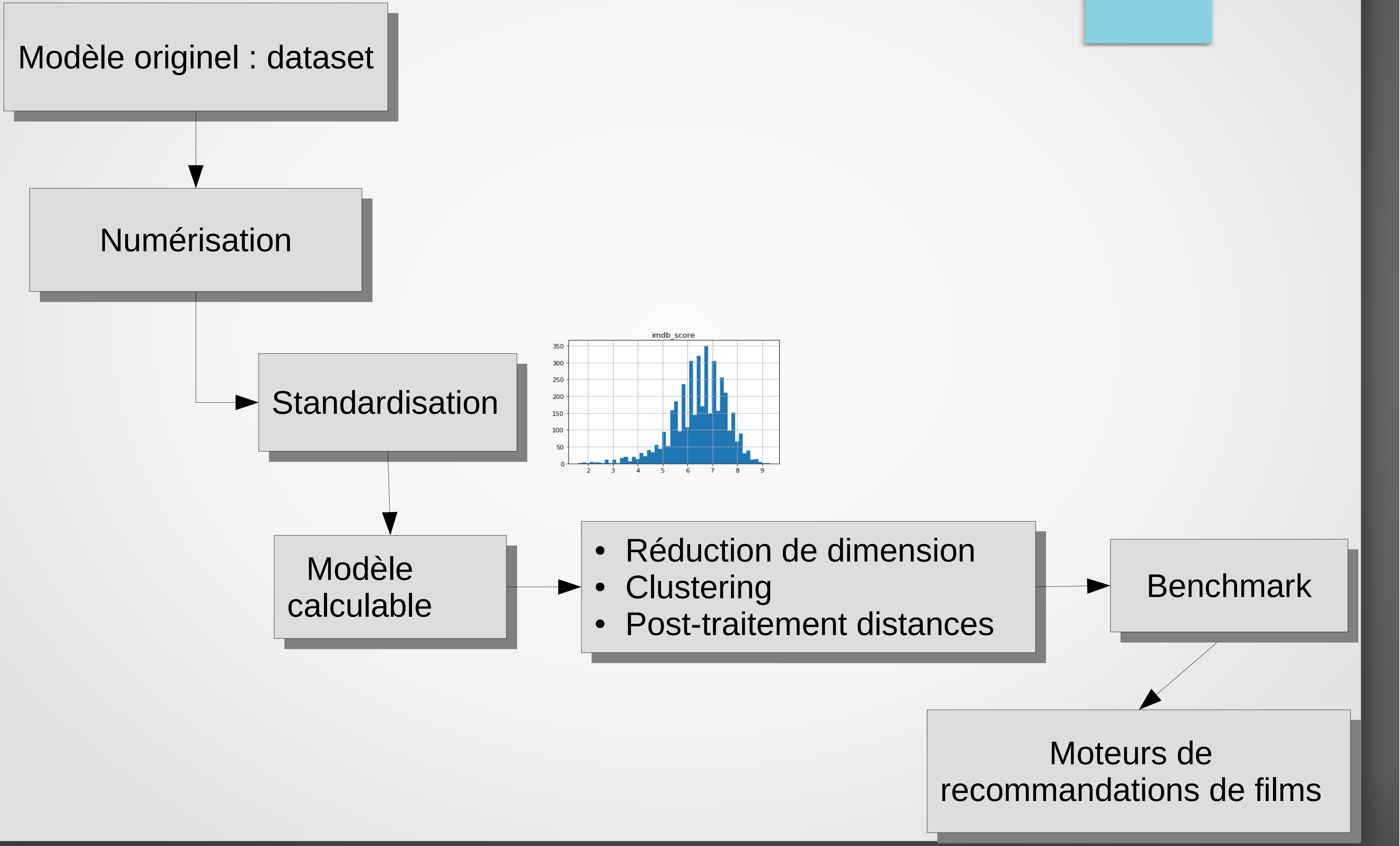
Réduction de dimension

Structure inconnue

Apprentissage non supervisé

Construction du modèle

Construction du modèle



Modèle originel : dataset

Numérisation

Standardisation

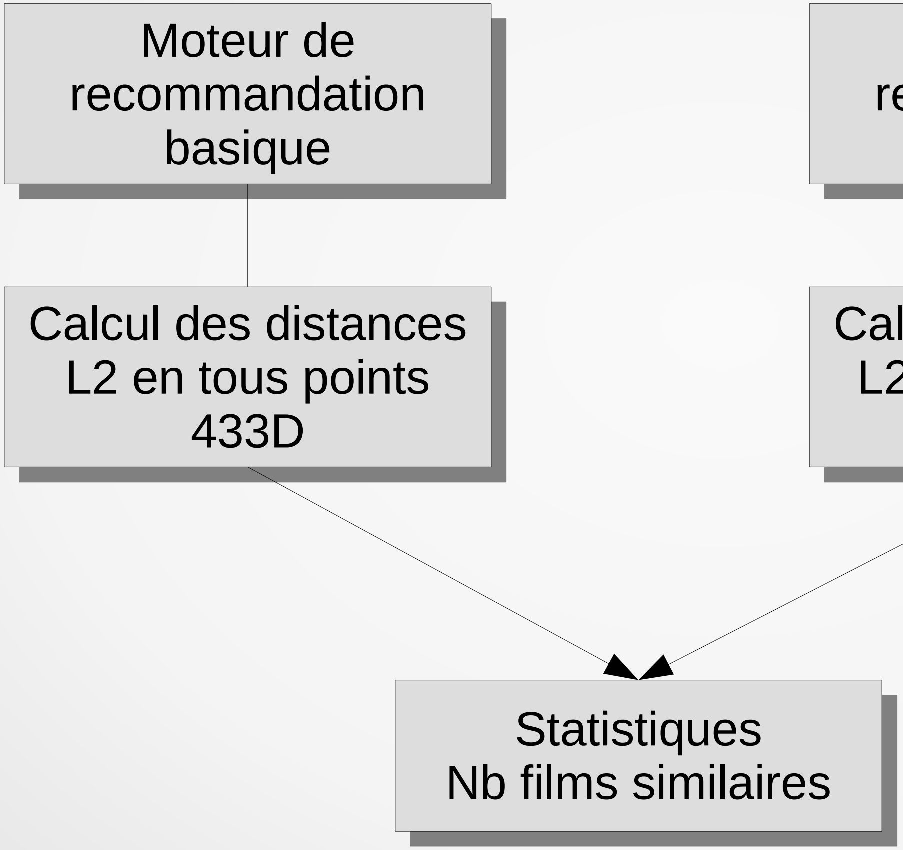
Modèle calculable

- Réduction de dimension
- Clustering
- Post-traitement distances

Benchmark

Moteurs de recommandations de films

Benchmark des modèles



Moteur de recommandation basique

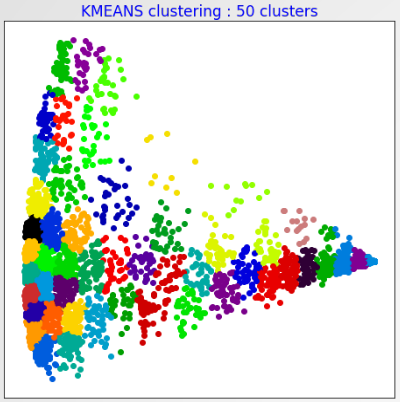
Calcul des distances L2 en tous points 433D

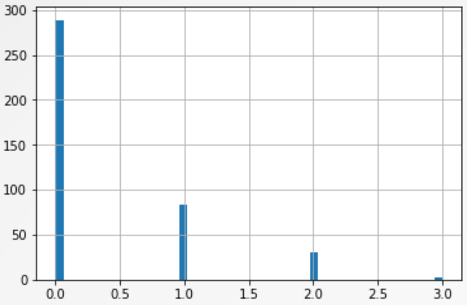
Moteur de recommandation non-supervisé

Calcul des distances L2 dans un cluster 2D

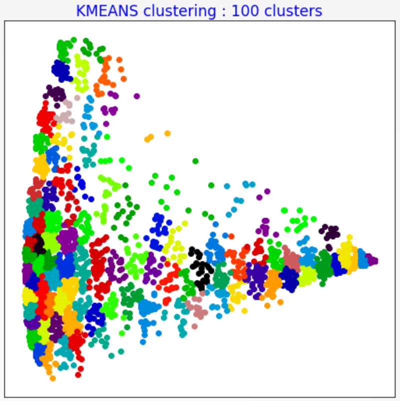
Statistiques
Nb films similaires

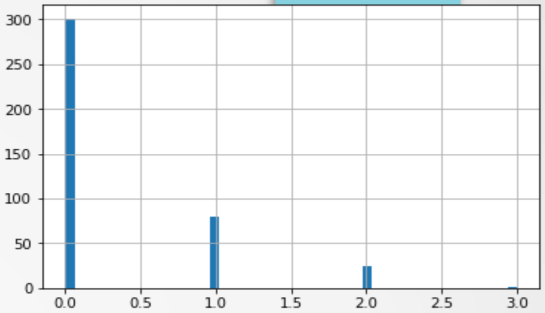
Benchmark : KPCA / K-means : 406 films



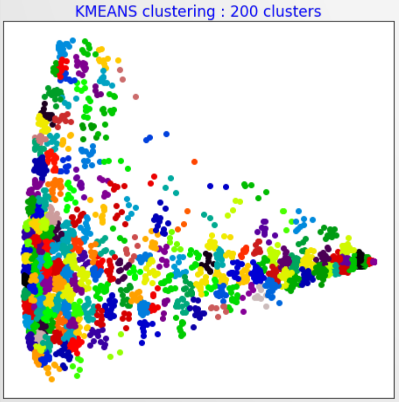


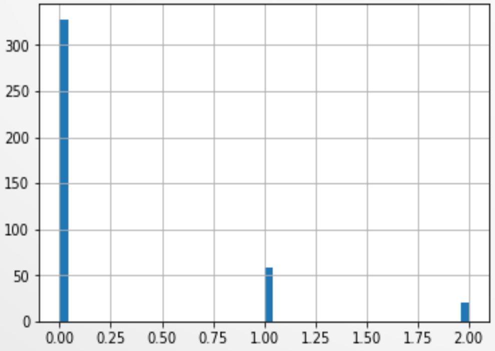
0 : 71 %
1 : 21 %
2 :   7 %
3 :   1 %
4:    0 %
5:    0 %





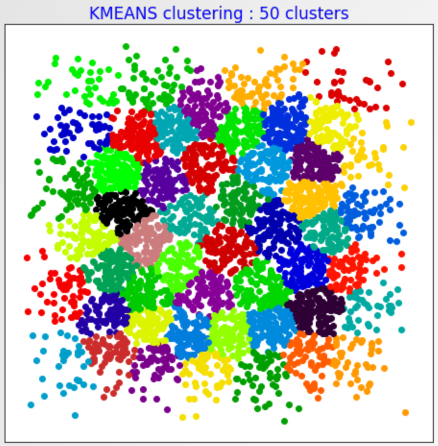
0 : 74 %
1 : 20 %
2 :   6 %
3 :   1 %
4:    0 %
5:    0 %

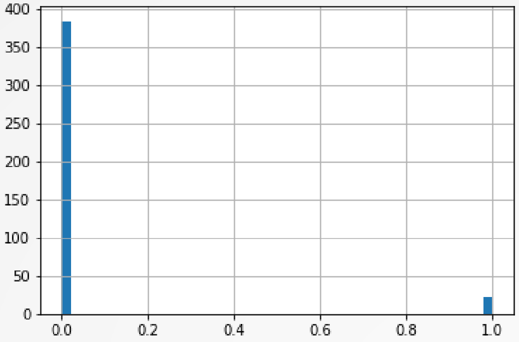




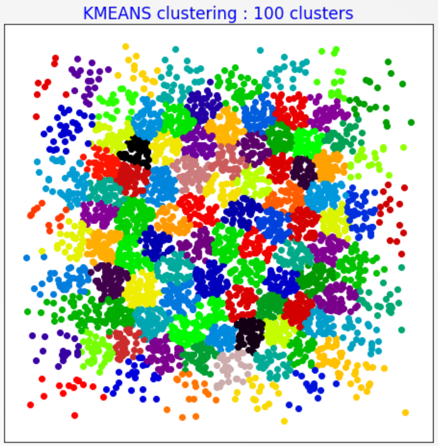
```
0    80.788177
1    14.285714
2     4.926108
```

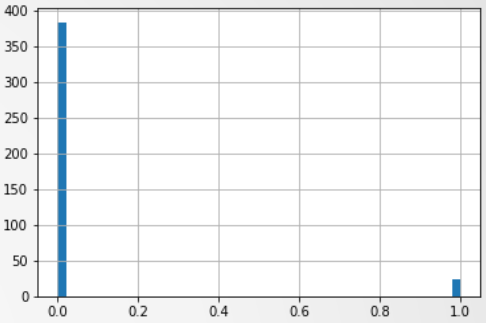
Benchmark: MDS+ kmeans : 406 films



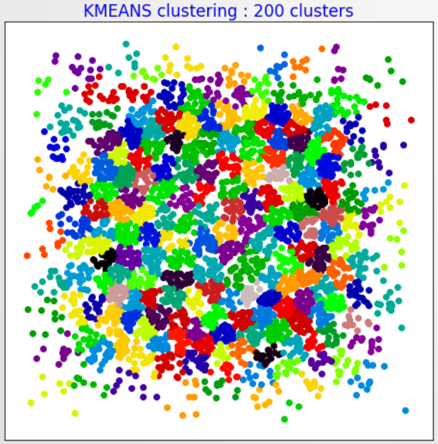


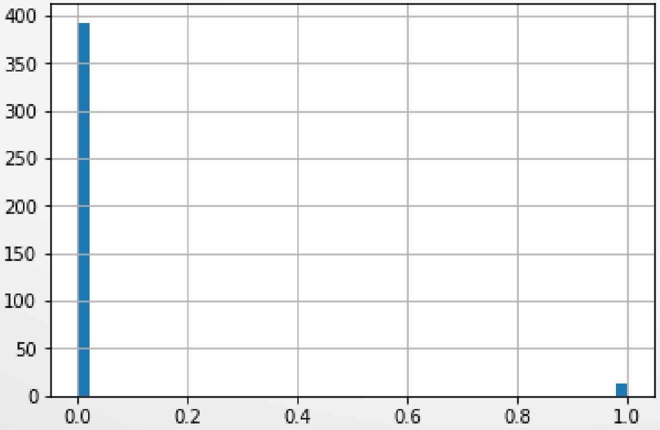
```
0     94.581281
1      5.418719
```





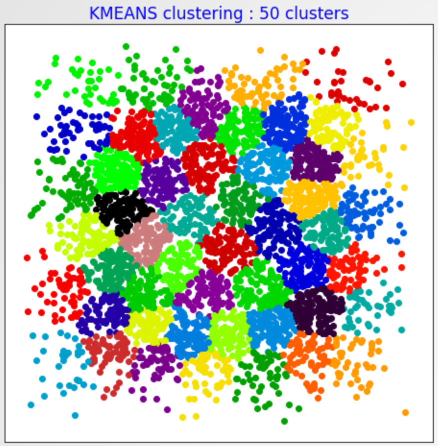
```
0     94.334975
1      5.665025
```

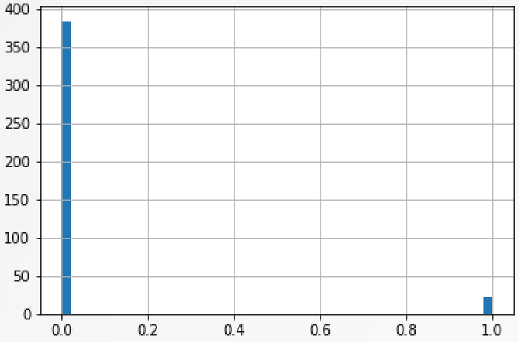




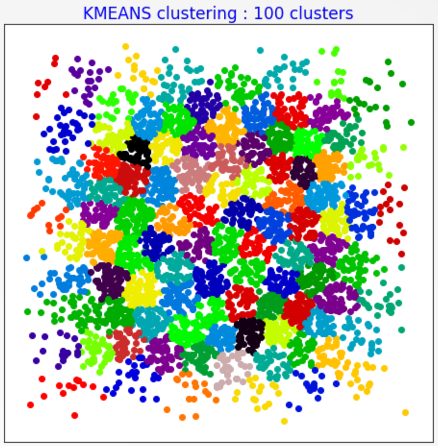
```
0     96.79803
1      3.20197
```

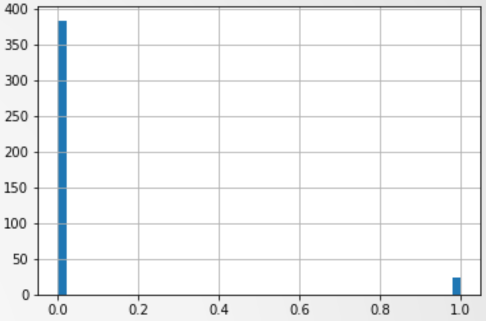
Benchmark: MDS+ kmeans : 406 films



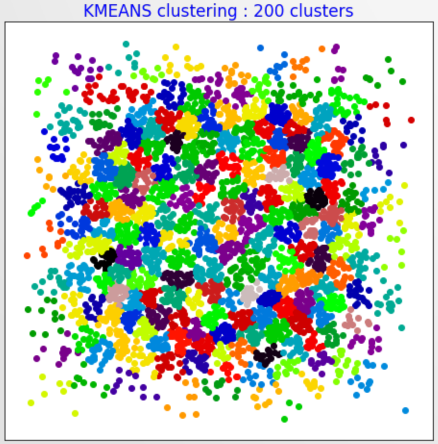


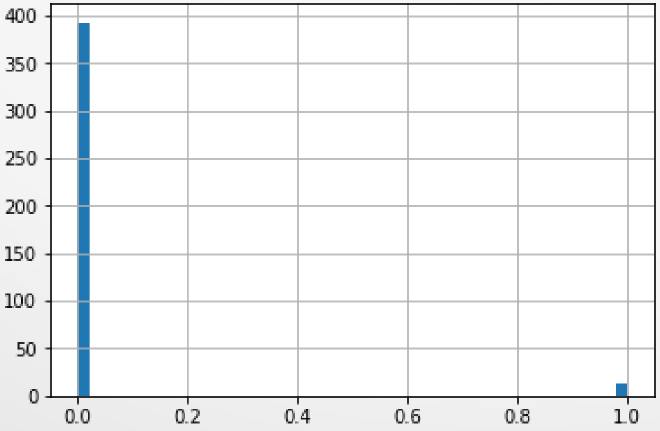
```
0     94.581281
1      5.418719
```





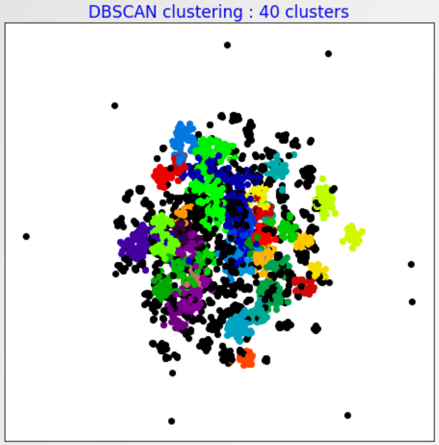
```
0     94.334975
1      5.665025
```

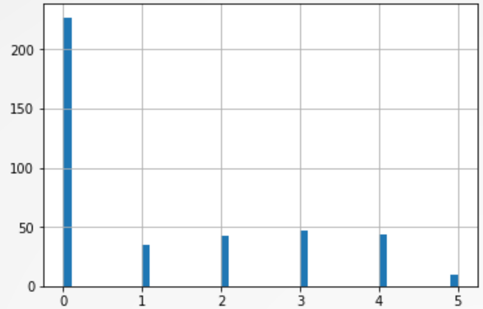




```
0     96.79803
1      3.20197
```

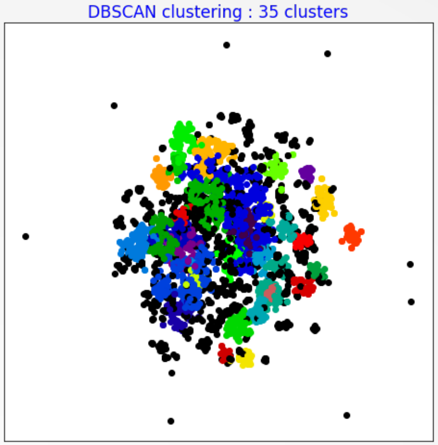
Benchmark : t-SNE + DBSCAN : 406 films

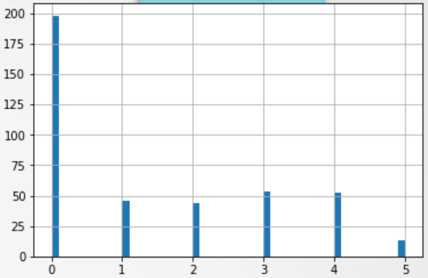




```
DBSCAN eps :           5
DBSCAN min samples :   25
DBSCAN clusters :      40
```

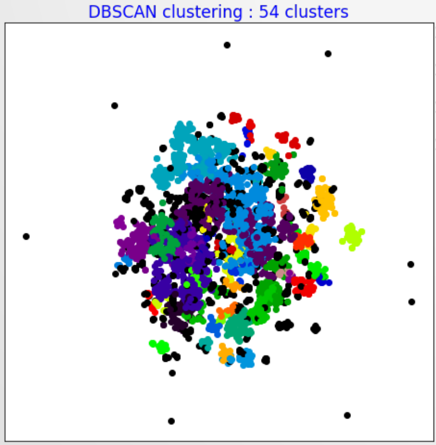
```
0     55.911330
3     11.576355
4     10.837438
2     10.591133
1      8.620690
5      2.463054
```

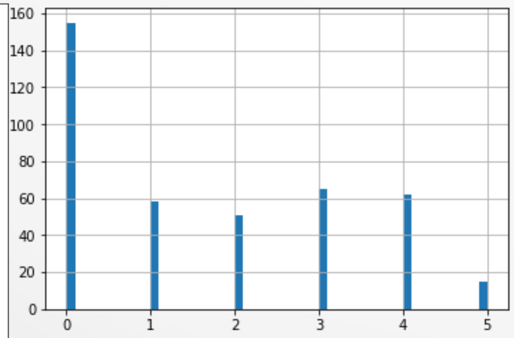




```
DBSCAN eps :           5
DBSCAN min samples :   20
DBSCAN clusters :      35
```

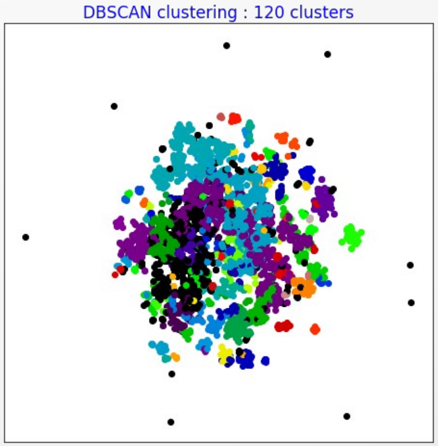
```
0     48.768473
3     13.054187
4     12.807882
1     11.330049
2     10.837438
5      3.201970
```

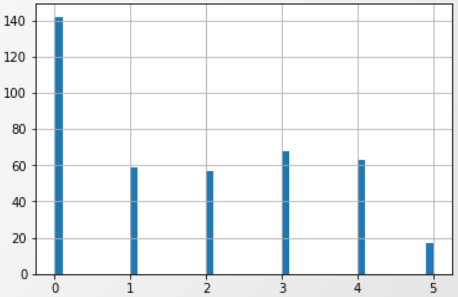




```
DBSCAN eps :           5
DBSCAN min samples :   10
DBSCAN clusters :      54
```

```
0     38.177340
3     16.009852
4     15.270936
1     14.285714
2     12.561576
5      3.694581
```

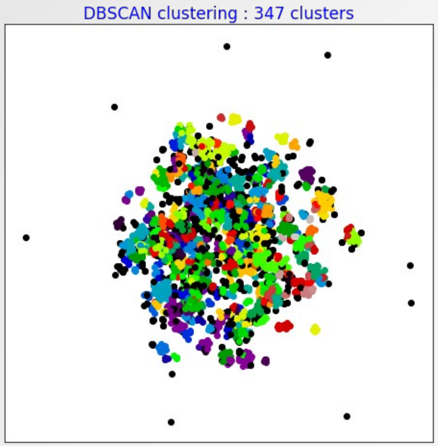


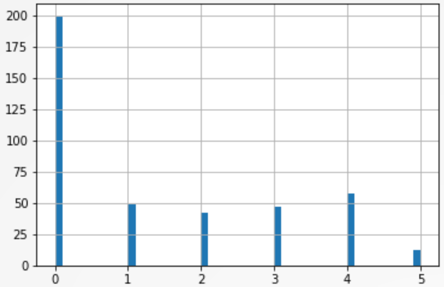


```
DBSCAN eps :           5
DBSCAN min samples :   4
DBSCAN clusters :      120
```

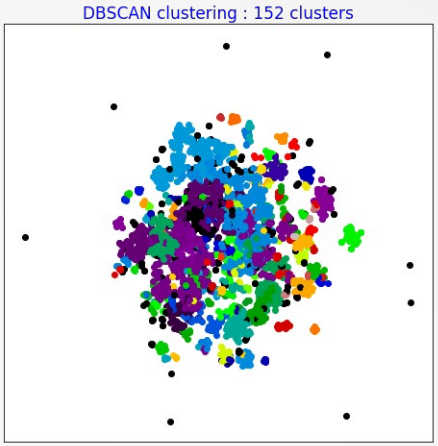
```
0     34.975369
3     16.748768
4     15.517241
1     14.532020
2     14.039409
5      4.187192
```

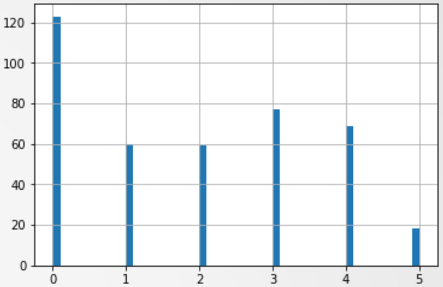
Benchmark : t-SNE + DBSCAN : 406 films





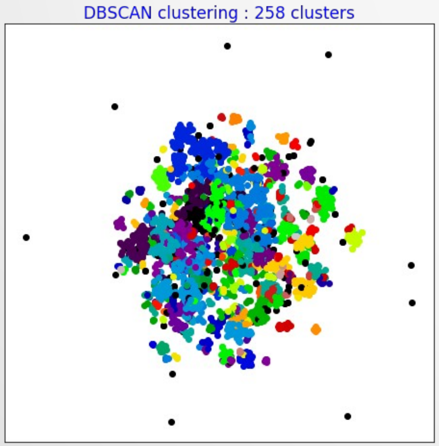
```
0    49.014778
4    14.039409
1    12.068966
3    11.576355
2    10.344828
5     2.955665
```

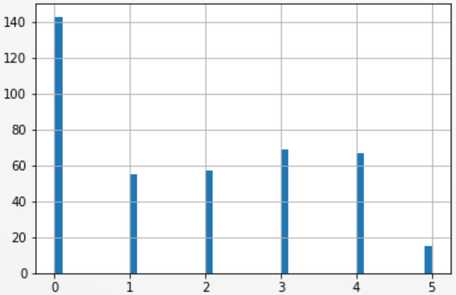




```
DBSCAN eps :            4
DBSCAN min samples :    4
DBSCAN clusters :       152
```

```
0    30.295567
3    18.965517
4    16.995074
1    14.778325
2    14.532020
5     4.433498
```





```
0    35.221675
3    16.995074
4    16.502463
2    14.039409
1    13.546798
5     3.694581
```

```
DBSCAN eps :            3
DBSCAN min samples :    3
DBSCAN clusters :       258
```

Bruit : 4,5 %

```
DBSCAN eps :            3
DBSCAN min samples :    3
DBSCAN clusters :       318
```

Cas tests : Titanic , Avatar

```
Reference = Titanic
{
         "_model":{tsne_dbscan}
         "_results":[
                { "id": "26","name": "Titanic"   },
                { "id": "2535","name": "Sense and Sensibility"   },
                { "id": "1011","name": "The Life of David Gale"   },
                { "id": "1114","name": "Revolutionary Road"   },
                { "id": "144","name": "Flushed Away"   }
        ]
}
{
         "_model":{scaled_none}
         "_results":[
                { "id": "26","name": "Titanic"   },
                { "id": "2535","name": "Sense and Sensibility"   },
                { "id": "1011","name": "The Life of David Gale"   },
                { "id": "144","name": "Flushed Away"   },
                { "id": "990","name": "The Beach"   },
                { "id": "641","name": "Body of Lies"   }
        ]
}
```

```
Reference = Avatar
{
         "_model":{tsne_dbscan}
         "_results":[
                { "id": "0","name": "Avatar"   },
                { "id": "339","name": "The Lord of the Rings: The Return of the King"   },
                { "id": "112","name": "Transformers"   },
                { "id": "3024","name": "Star Wars: Episode IV - A New Hope"   },
                { "id": "36","name": "Transformers: Revenge of the Fallen"   },
                { "id": "1536","name": "Star Wars: Episode VI - Return of the Jedi"   }
        ]
}
{
         "_model":{scaled_none}
         "_results":[
                { "id": "0","name": "Avatar"   },
                { "id": "339","name": "The Lord of the Rings: The Return of the King"   },
                { "id": "3024","name": "Star Wars: Episode IV - A New Hope"   },
                { "id": "112","name": "Transformers"   },
                { "id": "36","name": "Transformers: Revenge of the Fallen"   },
                { "id": "53","name": "Transformers: Dark of the Moon"   }
        ]
}
```

Cas tests : Indiana Jones , Star Trek

Reference = Indiana Jones and the Kingdom of the Crystal Skull

```
{
        "_model":{tsne_dbscan}
        "_results":[
                { "id": "54","name": "Indiana Jones and the Kingdom of the Crystal Skull"   },
                { "id": "1749","name": "Indiana Jones and the Temple of Doom"   },
                { "id": "1039","name": "Indiana Jones and the Last Crusade"   },
                { "id": "2154","name": "Close Encounters of the Third Kind"   },
                { "id": "189","name": "War of the Worlds"   },
                { "id": "697","name": "Jurassic Park"   }
        ]
}
{
        "_model":{scaled_none}
        "_results":[
                { "id": "54","name": "Indiana Jones and the Kingdom of the Crystal Skull"   },
                { "id": "1749","name": "Indiana Jones and the Temple of Doom"   },
                { "id": "1039","name": "Indiana Jones and the Last Crusade"   },
                { "id": "114","name": "Harry Potter and the Order of the Phoenix"   },
                { "id": "626","name": "Sky Captain and the World of Tomorrow"   },
                { "id": "3570","name": "A Home at the End of the World"   }
        ]
}
```

Reference = Star Trek II: The Wrath of Khan

```
{
        "_model":{tsne_dbscan}
        "_results":[
                { "id": "2923","name": "Star Trek II: The Wrath of Khan"   },
                { "id": "1803","name": "Star Trek VI: The Undiscovered Country"   },
                { "id": "1631","name": "Star Trek V: The Final Frontier"   },
                { "id": "3868","name": "R.L. Stine's Monsterville: The Cabinet of Souls"   },
                { "id": "1409","name": "Star Trek: The Motion Picture"   },
                { "id": "1343","name": "Star Trek: Generations"   }
        ]
}
{
        "_model":{scaled_none}
        "_results":[
                { "id": "2923","name": "Star Trek II: The Wrath of Khan"   },
                { "id": "1803","name": "Star Trek VI: The Undiscovered Country"   },
                { "id": "1631","name": "Star Trek V: The Final Frontier"   },
                { "id": "2396","name": "Star Trek III: The Search for Spock"   },
                { "id": "2018","name": "Star Trek IV: The Voyage Home"   },
                { "id": "3868","name": "R.L. Stine's Monsterville: The Cabinet of Souls"   }
        ]
}
```

Cas tests : GoldenEye, Spider-Man

```
Reference = GoldenEye
{
        "_model":{tsne_dbscan}
        "_results":[
                { "id": "717","name": "GoldenEye"   },
                { "id": "252","name": "Tomorrow Never Dies"   },
                { "id": "172","name": "The World Is Not Enough"   },
                { "id": "1166","name": "Licence to Kill"   },
                { "id": "1230","name": "The Count of Monte Cristo"   },
                { "id": "169","name": "Sahara"   }
        ]
}
{
        "_model":{scaled_none}
        "_results":[
                { "id": "717","name": "GoldenEye"   },
                { "id": "252","name": "Tomorrow Never Dies"   },
                { "id": "172","name": "The World Is Not Enough"   },
                { "id": "1166","name": "Licence to Kill"   },
                { "id": "169","name": "Sahara"   },
                { "id": "1230","name": "The Count of Monte Cristo"   }
        ]
}
```

```
Reference = Spider-Man 3
{
        "_model":{tsne_dbscan}
        "_results":[
                { "id": "6","name": "Spider-Man 3"   },
                { "id": "31","name": "Spider-Man 2"   },
                { "id": "161","name": "Spider-Man"   }
        ]
}
{
        "_model":{scaled_none}
        "_results":[
                { "id": "6","name": "Spider-Man 3"   },
                { "id": "31","name": "Spider-Man 2"   },
                { "id": "161","name": "Spider-Man"   },
                { "id": "2109","name": "Homefront"   },
                { "id": "669","name": "Mona Lisa Smile"   },
                { "id": "3180","name": "Deuces Wild"   }
        ]
}
```

Cas tests : films sud coréen et japonais

```
Bruit : 4,5 %
DBSCAN eps :            3
DBSCAN min samples :    3
DBSCAN clusters :       318
```

```
'The Last Godfather'
{
      "_model":{tsne_dbscan}
      "_results":[
           { "id": "2833","name": "The Last Godfather"   }, ◄ Comedy
           { "id": "1564","name": "Dragon Wars: D-War"   }, ◄ Action|Drama|Fantasy|Horror|Thriller
           { "id": "1072","name": "Inchon"   },       ◄ Drama|History|War
           { "id": "1325","name": "Snowpiercer"   }   ◄ Action|Drama|Sci-Fi|Thriller
      ]
}
Reference = 'One Missed Call' (Japan)
{
      "_model":{tsne_dbscan}
      "_results":[
           { "id": "2220","name": "One Missed Call"   },  ◄ Horror|Mystery
           { "id": "1447","name": "Street Fighter"   }, ◄ Canada /   Action|Crime|Drama|Mystery|Thriller
           { "id": "1413","name": "Trainwreck"   }, ◄ Comedy|Romance
           { "id": "519","name": "The Secret Life of Pets"   }, ◄ Animation|Comedy|Family
           { "id": "1933","name": "Tora! Tora! Tora!"   },  ◄ Action|Drama|History|War
           { "id": "1560","name": "The Quick and the Dead"   } ◄ Action|Thriller|Western
      ]
}
```

Cas tests : films français & indiens

```
Bruit : 4,5 %
DBSCAN eps :            3
DBSCAN min samples :    3
DBSCAN clusters :       318
```

```
'Hitman'
{
        "_model":{tsne_dbscan}
        "_results":[
              { "id": "2481","name": "Hitman"   },            ← France
              { "id": "2477","name": "Wolves"   },            ← France
              { "id": "1776","name": "Pride & Prejudice"   }, ← France
              { "id": "2723","name": "Mulholland Drive"   },  ← France
              { "id": "2204","name": "Babel"   },             ← France
              { "id": "3179","name": "The Straight Story"   } ← France
       ]
}
```

```
'Monsoon Wedding'
{
        "_model":{tsne_dbscan}
        "_results":[
              { "id": "4490","name": "Monsoon Wedding"   },
              { "id": "4385","name": "The Lunchbox"   },
              { "id": "4160","name": "Lage Raho Munna Bhai"   },
              { "id": "3075","name": "Kabhi Alvida Naa Kehna"   },
              { "id": "3208","name": "Krrish"   },
              { "id": "3344","name": "My Name Is Khan"   }
       ]
}
```

Conclusions

Problème fondamentalement non-linéaire

t-SNE + DBSCAN : meilleurs résultats Mais : + lent

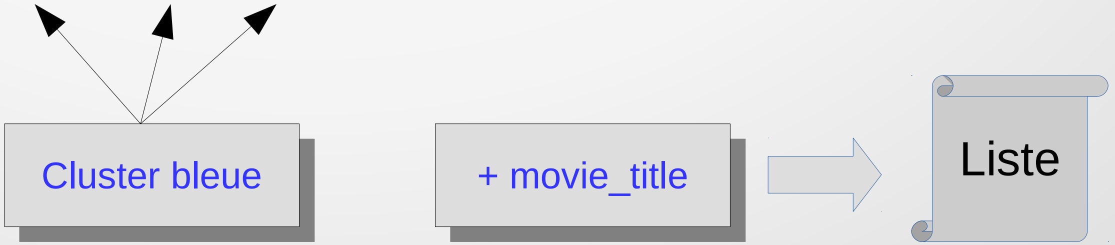
Axes d’améliorations :

- Performance t-SNE + DBSCAN
  - Dimension >2
  - Diminution du bruit (~10 %)
- Purge des pays avec < 5 films
- Traitement des « outliers » atypiques
- Variable plot_keywords
- Dimension culturelle du moteur

Ingénierie logicielle: calcul de similarité

Reconstitution de la matrice des distances L2 issue de X_xxx

	id_1	id_2	id_p	id_k	id_n	movie_id
id_1						
id_k	x_k1	x_k2	x_kp	0.0	x_kn	movie_k



Cluster bleue

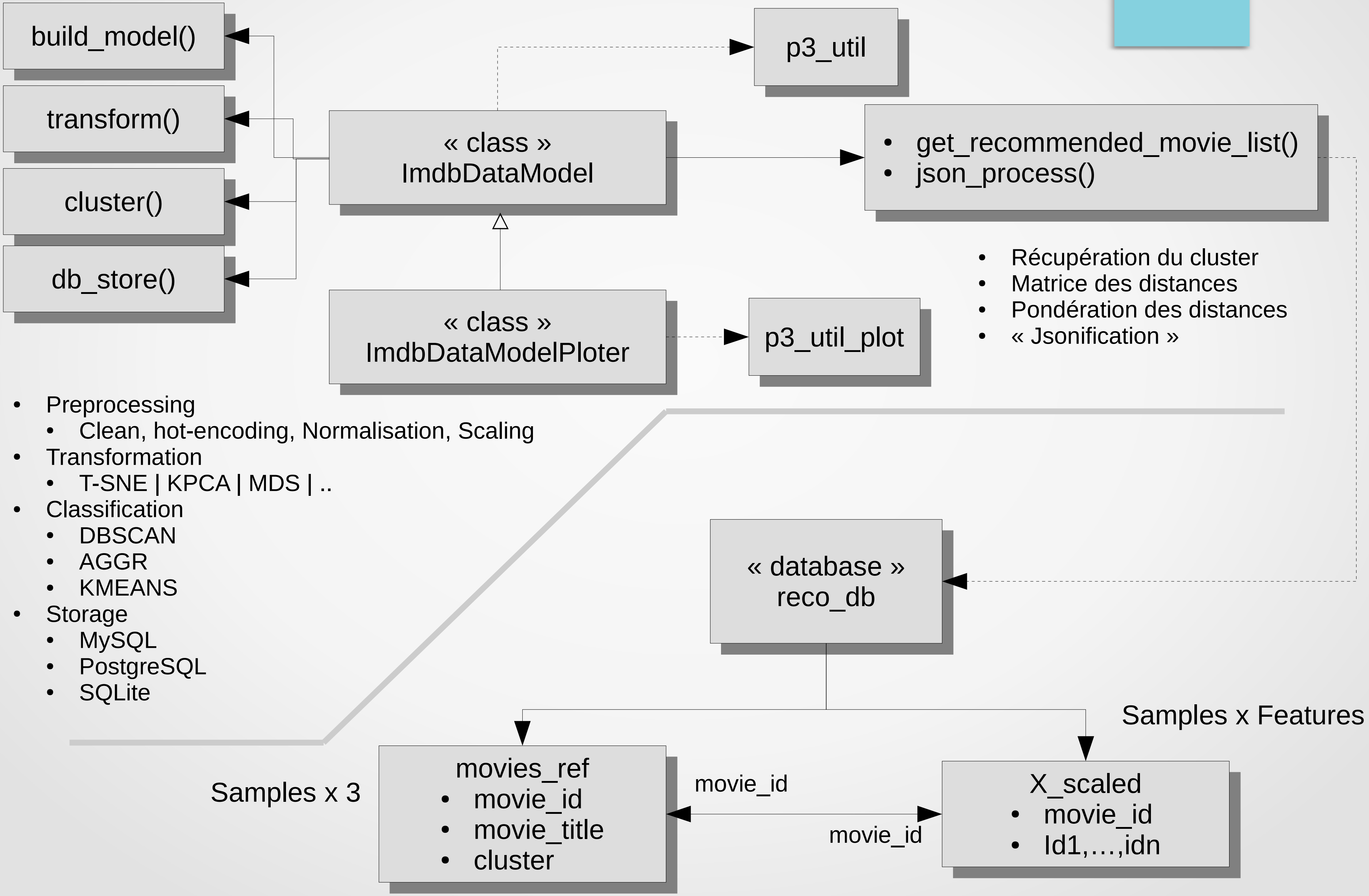
+ movie_title

Liste

Annexe

- **Fichiers source python :**
  - heroku/recomovies/recomovies/ImdbDataModel.py
  - heroku/recomovies/recomovies/p3_util.py
  - ImdbDataModelPlotter.py
  - p3_util_plot.py
- **Notebook de l’alnalyse exploratoire :**
  - P3.ipynb : Nettoyage, Exploration, Modélisation
- **Notebook des approches de modélisation :**
  - ImdbDataModel.ipynb : Évaluation, pré-production
- **Rapport sous forme de présentation pdf:**
  - Openclassrooms_ParcoursDatascientist_P3.pdf
- **Point d’entrée de l’API :**
  - Pour récupérer toutes les références avec leur nom :
    - https://recomovies.herokuapp.com/recommend?’*’
  - Pour récupérer une liste de films recommandés à partir d’un identifiant :
    - https://recomovies.herokuapp.com/recommend?movie_id=0
  - Pour récupérer une liste de films recommandés à partir d’un titre :
    - https://recomovies.herokuapp.com/recommend?movie_title= »Avatar »

Annexe : Ingénierie logicielle



build_model()

transform()

cluster()

db_store()

p3_util

« class »
ImdbDataModel

- get_recommended_movie_list()
- json_process()

« class »
ImdbDataModelPloter

p3_util_plot

- Récupération du cluster
- Matrice des distances
- Pondération des distances
- « Jsonification »

- Preprocessing
  - Clean, hot-encoding, Normalisation, Scaling
- Transformation
  - T-SNE | KPCA | MDS | ..
- Classification
  - DBSCAN
  - AGGR
  - KMEANS
- Storage
  - MySQL
  - PostgreSQL
  - SQLite

« database »
reco_db

Samples x Features

Samples x 3

movies_ref
- movie_id
- movie_title
- cluster

movie_id

movie_id

X_scaled
- movie_id
- Id1,…,idn

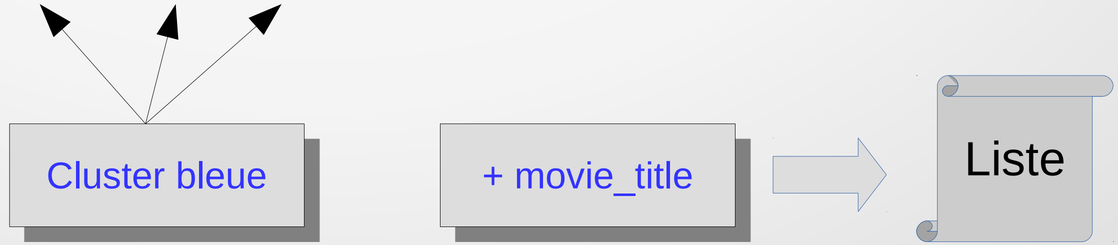
Annexe : Structure des tables de reco_db

```
recomovies::DATABASE=> \d

 Schema |    Name      | Type  |     Owner
--------+--------------+-------+--------------
 public | movies_ref | table  | akmtryukierams
 public | x_tsne       | table  | akmtryukierams


recomovies::DATABASE=> select * from x_tsne limit 2;
 index |    0        |     1          | movie_id
--------+------------+--------------+-------------
     0  | -28.0314 |   7.58088  |       0
     1  | -14.6499 | -0.163551 |       1


 index |               movie_title                                   | movie_id | cluster
--------+----------------------------------------------------+------------+---------
     0  | Avatar                                                    |       0    |       0
     1  | Pirates of the Caribbean: At World's End |       1    |       1
```

Annexe: calcul de similarité

Reconstitution de la matrice des distances L2 issue de X_xxx

	id_1	id_2	id_p	id_k	id_n	movie_id
id_1						
id_k	x_k1	x_k2	x_kp	0.0	x_kn	movie_k



Cluster bleue

+ movie_title

Liste