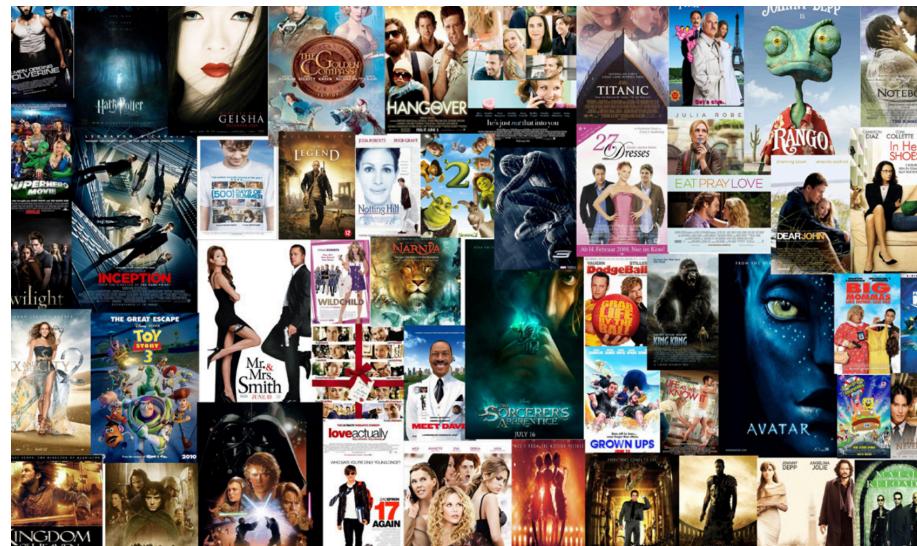


Parcours DataScientist Projet 3:

Recommendation de films



Jérôme d'Harveng

Mentor: Pierre Comalada

OpenClassrooms février 2019

Table des matières

PART I: Contexte de l'analyse de données

PART II: Nettoyage des données

PART III: Analyse exploratoire

PART IV : Pistes de modélisation

PART V: Modèle final et performances

PART VI: Conclusions



Part I:

Contexte de l'analyse



Problématique

- Grand Prince :

- * But final: via une API

à partir d'un titre de film recommander

=> 5 films similaires et intéressants pour le visiteur

- * Informations:

une **Base de données**

reprenant des films avec leurs caractéristiques

- * Points importants:

Notion de similarité entre films



Interprétation et déductions



- Analyse de la littérature

- * *Content Filtering*: « Si vous aimez cette item, vous aimerez peut-être ... »
- * *Item-Item Collaborative Filtering*: « Les clients qui aiment cette item, aiment aussi... »
- * *User-Item Collaborative Filtering*: « Les clients qui vous ressemblent aiment aussi... »

- Etude exploratoire nécessaire

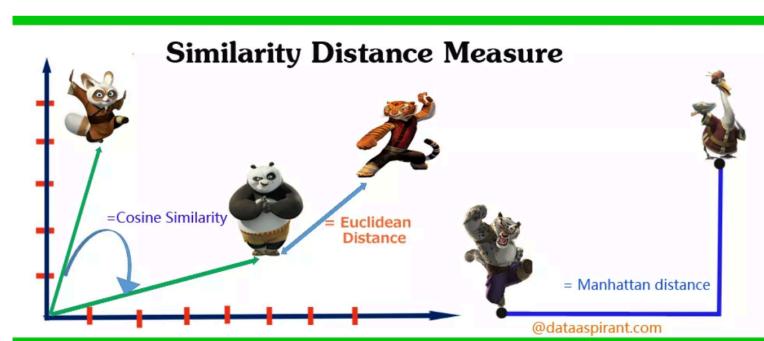
=> se focaliser les paramètres clé



- Pour permettre la **visualisation** => réduction dimensionnelle

- Notion de similarité:

- * clustering
- * notion de distance



Part II:

Nettoyage des données



Nettoyage de la Base (CLEANING : 1)

- **Valeurs manquantes** => différentes options

- * **Remplacer certaines valeurs:**

-> exemple actor_1_name, plot_keywords. ('')

- * **Supprimer certaines lignes:**

-> exemple (gross : 17%, budget 9%)

Idéalement étude : régression linéaire / polynomiale

- * **Supprimer certaines variables de la base:**

-> aspect_ratio (6.5 %)

- **Supprimer les doublons dans la base**

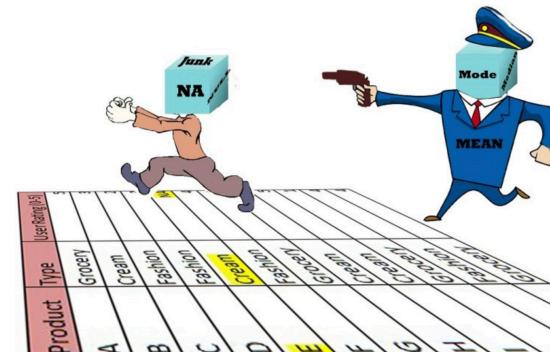
- **Conversion type de données:**

- * float vers int (ex. : likes , reviews)

- * string vers variable catégorielle (ex. content_rating)

- **Refactoring :**

- * plot_keywords : de « | » vers list []



Nettoyage de la Base (CLEANING : 2)

- *Feature Engineering:*

- * M =GP = PG => combiner en PG
- * X = NC-17 => on garde NC-17
- * « Not Rated », « Unrated », « Approved » et « Passed » vers R (catégorie la plus représentée)

df_movies_clean.content_rating.value_counts()

| | |
|-----------|------|
| R | 1699 |
| PG-13 | 1280 |
| PG | 560 |
| G | 91 |
| | 47 |
| Not Rated | 41 |
| Unrated | 24 |
| Approved | 17 |
| X | 9 |
| NC-17 | 6 |
| Passed | 3 |
| M | 2 |
| GP | 1 |



| | |
|-------|------|
| R | 1784 |
| PG-13 | 1280 |
| PG | 563 |
| G | 91 |
| NC-17 | 15 |

- *Outliers:*

Budget => si pas US (ou avant 2000 => pas euros) => Webscrapping mais
Problème de consistance sur les pages Web imdb

Details WINGED MIGRATION

Official Sites: BAC Films (French) | Official site [United States]
Country: France | Germany | Switzerland | Spain | Italy
Language: French
Release Date: 12 December 2001 (Belgium) See more »
Also Known As: Winged Migration See more »
Filming Locations: Monument Valley, Arizona, USA See more »

Box Office

Budget: FRF 160,000,000 (estimated)
Opening Weekend: \$2,402,605 (France), 14 December 2001
Opening Weekend USA: \$33,128, 20 April 2003, Limited Release
Gross USA: \$10,762,178, 14 December 2003
Cumulative Worldwide Gross: \$20,217,080, 6 November 2003
See more on IMDBPro »

Details MAD MAX

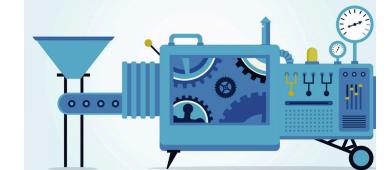
Official Sites: Official site | Official site [France]
Country: Australia | USA
Language: English | Russian
Release Date: 13 May 2015 (Belgium) See more »
Also Known As: Mad Max: Fury Road See more »
Filming Locations: Namib Desert, Namibia See more »

NO INFO OF BUDGET

Company Credits

Production Co: Warner Bros. Pictures, Village Roadshow Pictures, Kennedy Miller Productions See more »
Show more on IMDBPro »

Feature Engineering



En plus de *content_ratio* lors du CLEANING

- **Variable :** ***REVENUE***

= Gross - Budget



- **Variable :** ***SUCCESS***

= Revenue / Budget



Part III:

Analyse Exploratoire

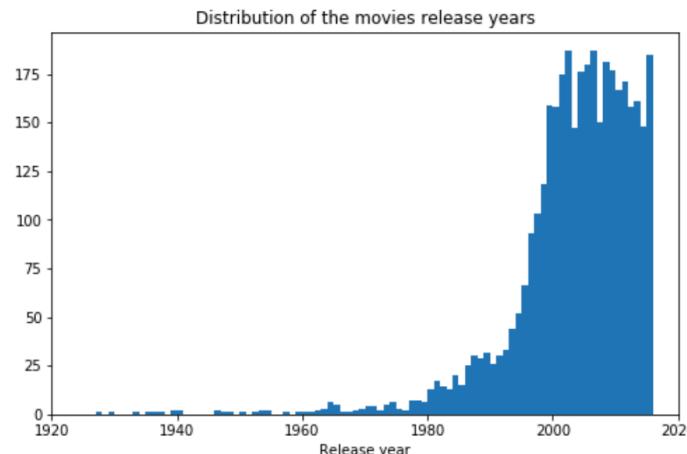


Analyse Univariée (1)

Observations:

- **Année de production:** films dans base produits entre 1927 - 2016

* 75 % des films réalisés après 1999



- **Données générales:**

* 96,7% en couleur

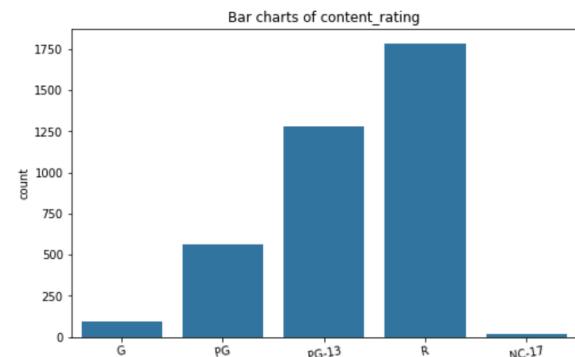
* 79,5 % produit aux US / 8,27% en UK

* 95,7% des films en anglais

- **Durée:**

* 75% des films moins de 2h

* **Film le plus long = 5h30** : « Blood In, Blood Out »



- **Content rating:**

* la plupart des films => catégorie « Restricted »

Analyse Univariée (2)

- Budget:

- * moyenne: 37 M\$
- * écart-type: 41 M\$ => grande variabilité
- * 75% des films budget < 50M\$
- * budget le plus élevé = 300 M\$ => « Pirates of the Caribbean: At Worlds end »
- * **2 films avec un budget < 2000\$**

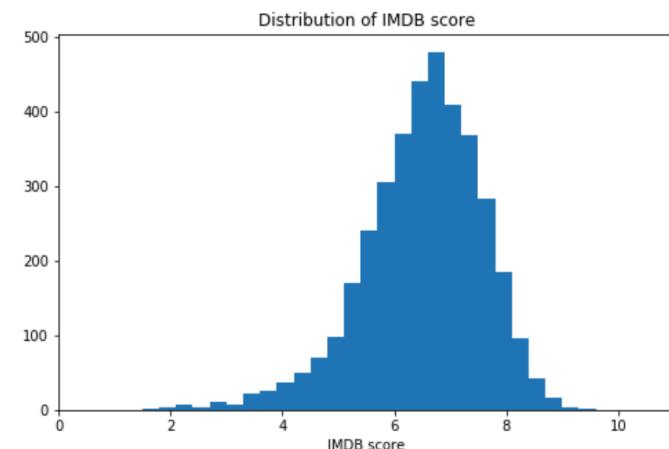
| | movie_title | budget | country |
|------|-------------------|--------|---------|
| 3652 | Tarnation | 218 | USA |
| 3732 | My Date with Drew | 1100 | USA |

- Gross:

- * moyenne: 50 M\$
- * écart-type: 68 M\$ => grande variation
- * 75% des films Gross < 65M\$
- * **Gross le plus élevé = 760M\$** => « Avatar » (James Cameron)

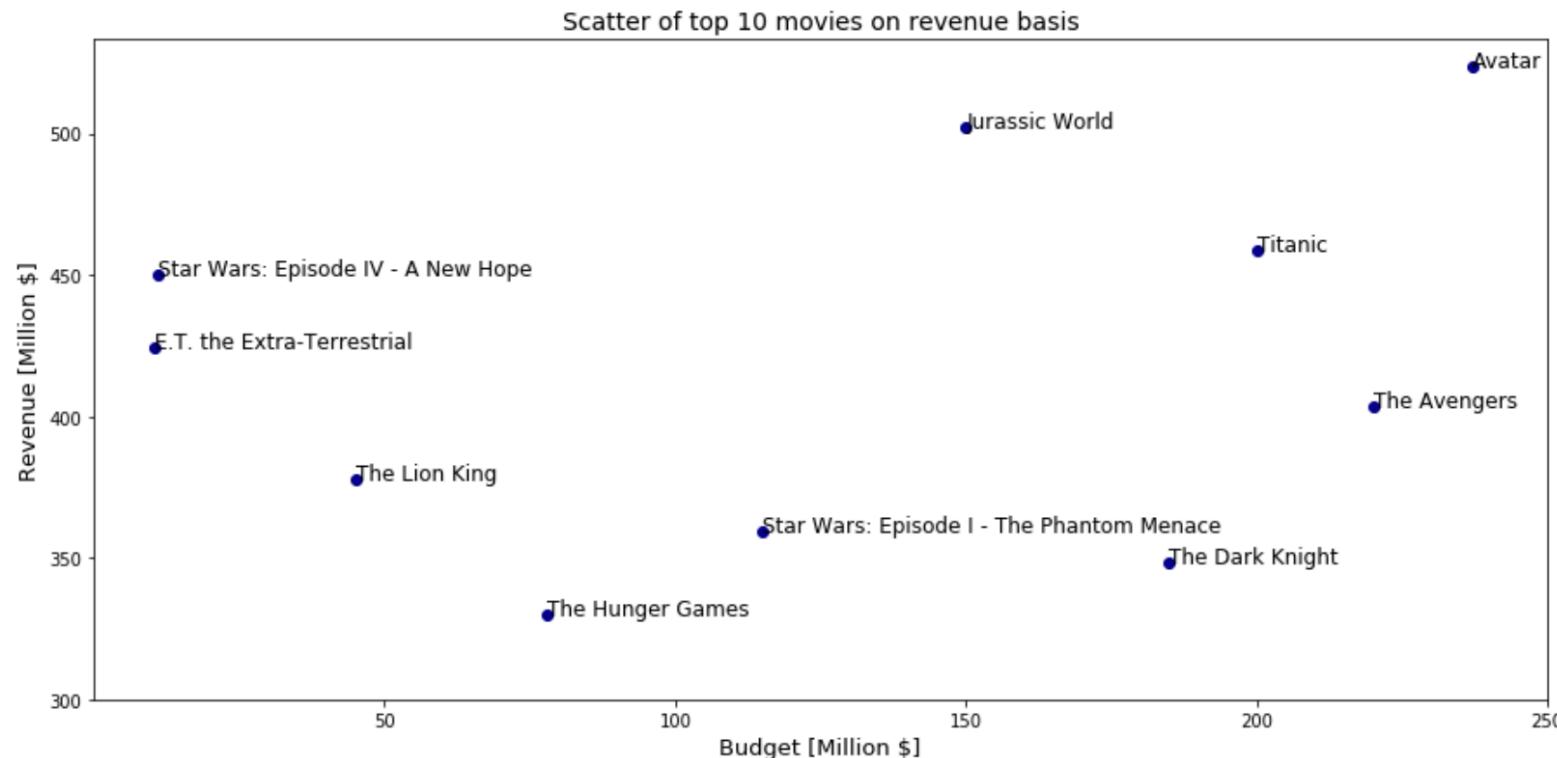
- Imdb Score:

- * moyenne: 6,45
- * écart-type: 1.05
- * **max = 9.3** : « The Shawshank Redemption »



Analyse Multivariée (1)

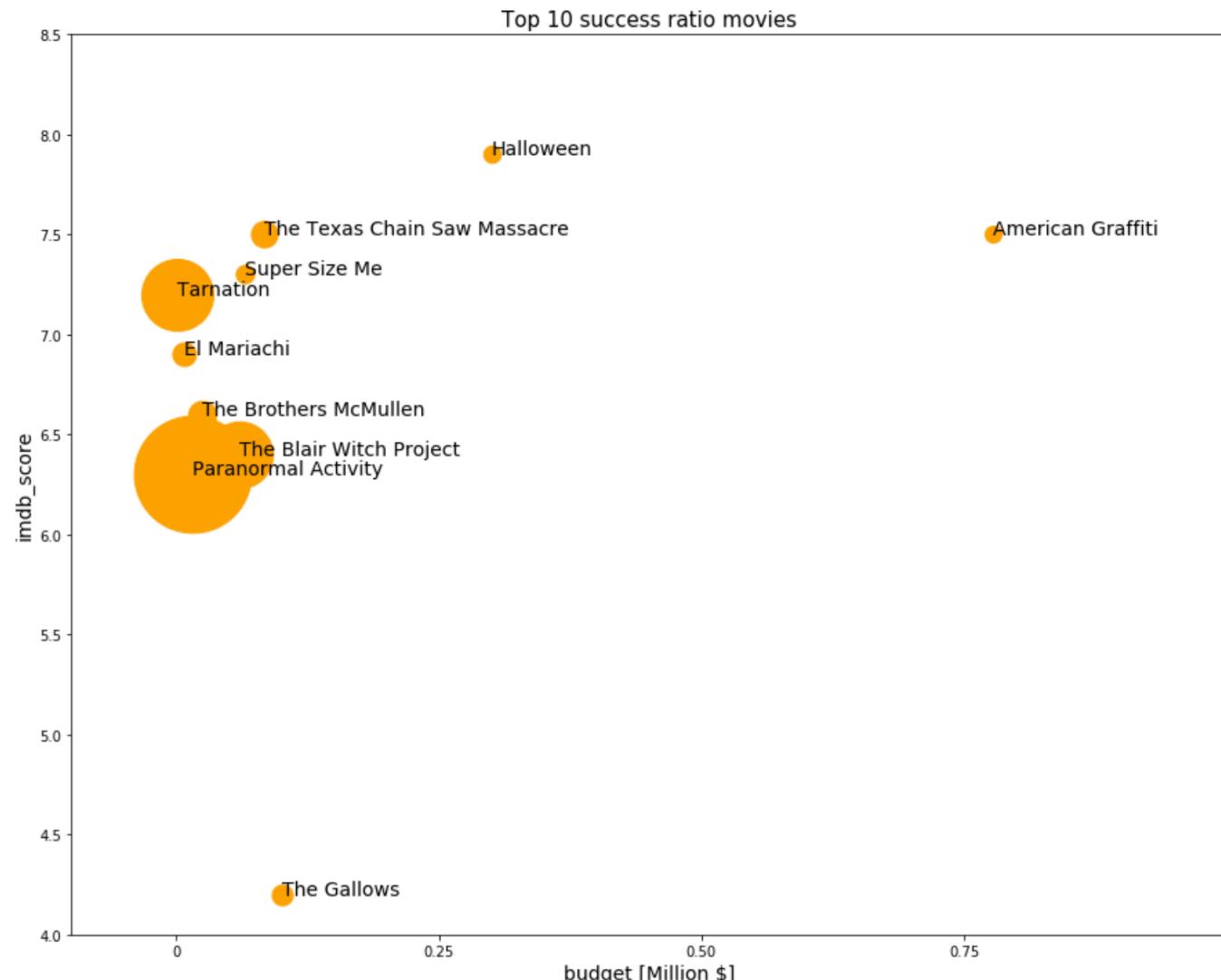
Lien Revenue vs Budget ? :



- d'où « **Feature Engineering** » => variable « Success »

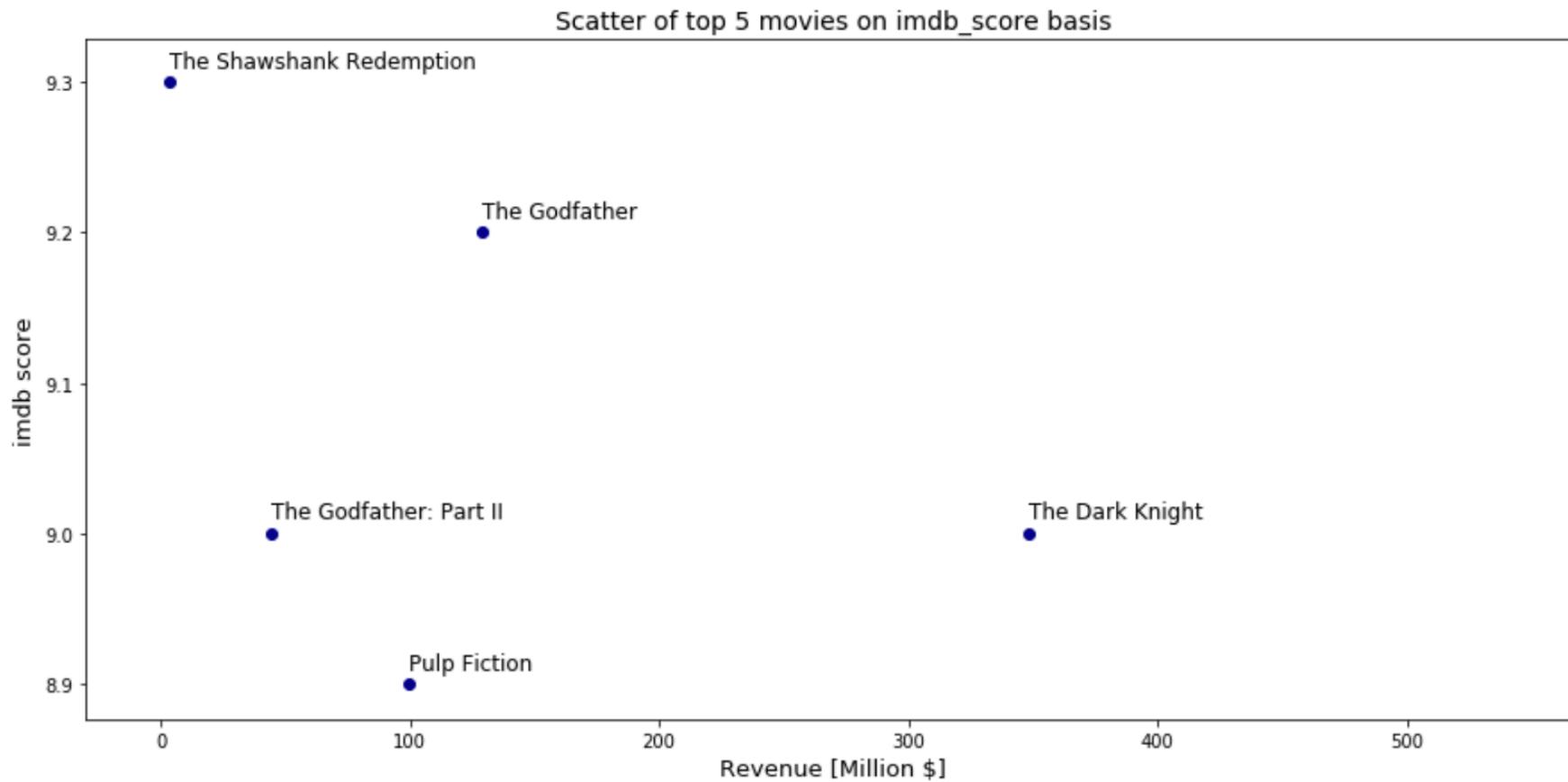
Analyse Multivariée (2)

Top 10 success movies



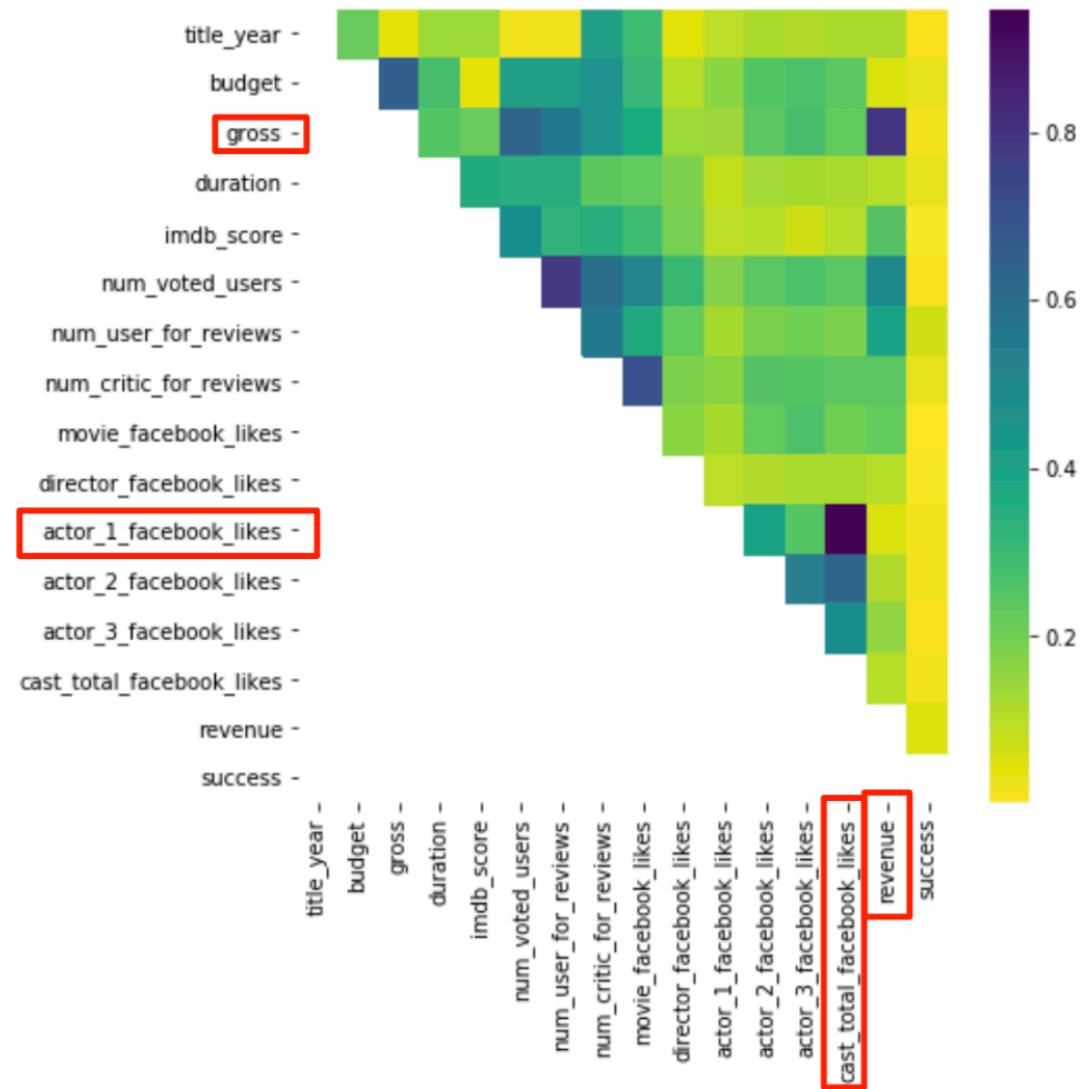
Analyse Multivariée (3)

Top 5 IMDB score vs Revenue



Feature Selection

Heatmap



Part IV:

Pistes modélisation

Encodage variables catégorielles

- One Hot Encoding vs Label Encoding:

* Label Encoding peut donner fausse impression d'ordre

- Langue et couleur => vue % pas déterminant pour l'étude

- Country:

* 'US', 'UK' et regroupement autres pays dans 'others'

* One Hot Encoding

- Content_rating => One Hot Encoding (avec Feature Engineering)

- Genres:

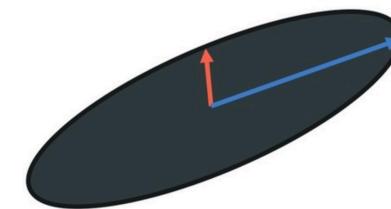
| | genre_action | genre_adventure | genre_animation | genre_biography | genre_comedy | genre_crime | genre_documentary | genre_drama | genre_family | genre_fantas |
|---|--------------|-----------------|-----------------|-----------------|--------------|-------------|-------------------|-------------|--------------|--------------|
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Idée de modélisation

- A partir des données nettoyées

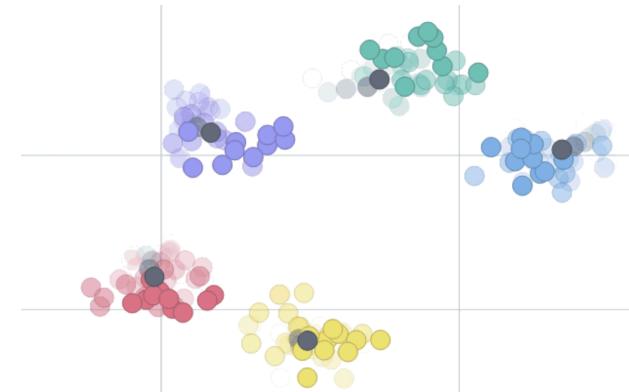
=> réduire la dimension des données:

- * permettre la visualisation
- * améliorer performances algorithmes



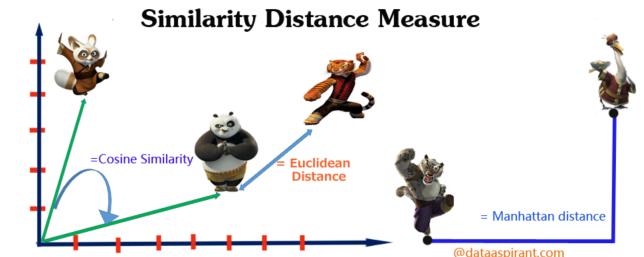
- A partir des données réduites

=> Clustering des données



- Dans le cluster correspondant au film

=> Similarités avec autres films intracluster



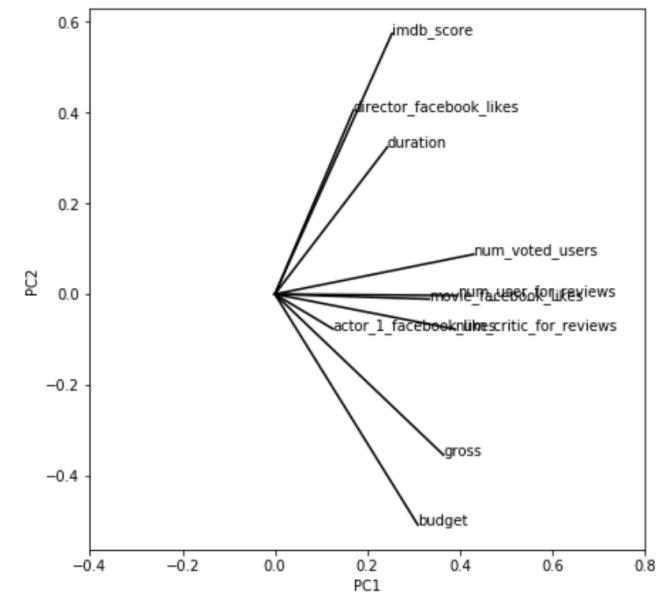
Réduction Dimensionnelle (1)

Analyse en composantes principales (ACP)

- Pas du tout adaptée aux *Variables Catégorielles*
- Seulement 53% variance expliquée par deux 1e composantes
- Si variance= 85% => garder 6 composantes principales
(sur 11 variables initiales)

=> *ACP pas idéal dans ce cas-ci*

=> essayer t-SNE



Réduction Dimensionnelle (2)

Stochastic Neighbour Embedding (t-SNE)

- **Notion de perplexité :**

entre 5 -50

* balance entre importance donnée à l'aspect local/global

- **Une des méthodes les plus utilisées mais interprétation difficile:**

* **Taille** des clusters ne signifie rien

* **Distances** entre clusters ne peuvent pas toujours être interprétées

* Pas toujours possible de tirer des conclusions de la **topologie**

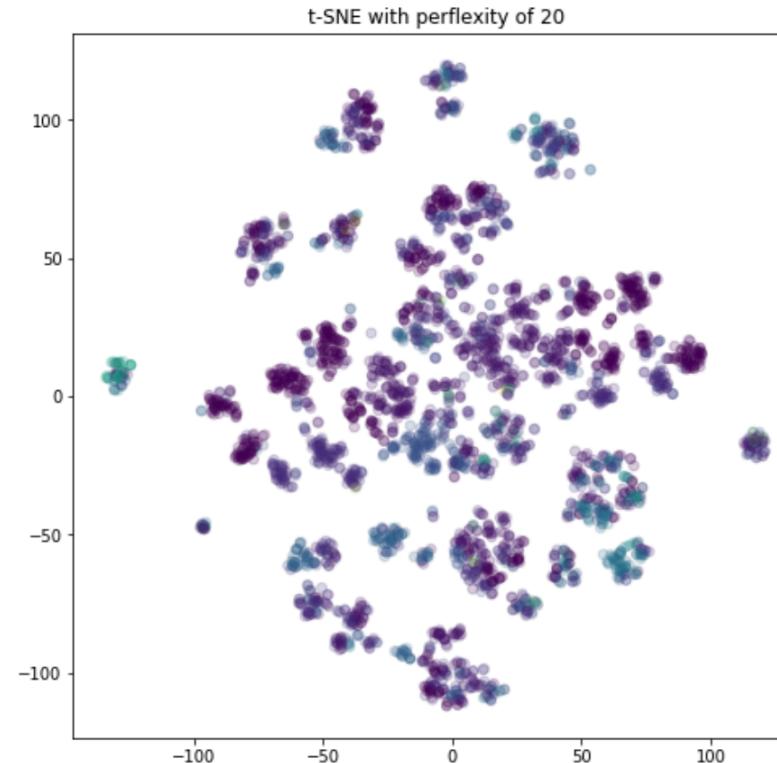
Réduction Dimensionnelle (3)

Stochastic Neighbour Embedding (t-SNE)

Comparaison coefficients de silhouette pour les perplexités choisies

```
Maximum silhouette for perplexity_5  
0.47517127  
Maximum silhouette for perplexity_10  
0.5334707  
Maximum silhouette for perplexity_20  
0.56310874  
Maximum silhouette for perplexity_30  
0.54580647  
Maximum silhouette for perplexity_40  
0.53075135  
Maximum silhouette for perplexity_50  
0.5251028
```

=> **choix Perplexité = 20**



Clustering (1)

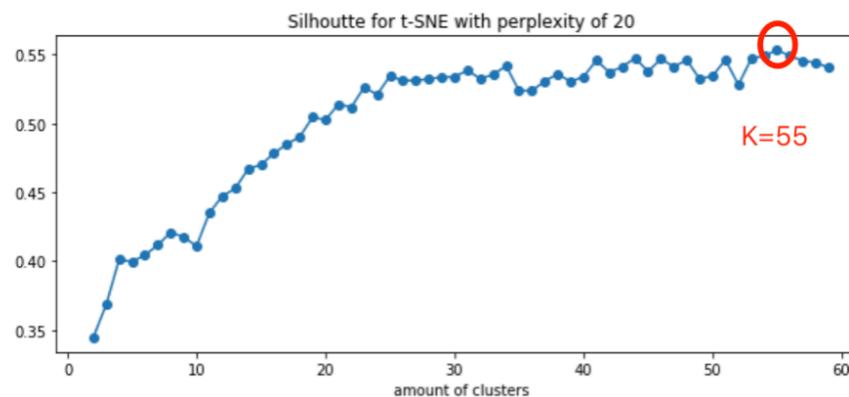
K-means++

- Inconvénient initialisation aléatoire => pas déterministe
=> centroïdes initiaux: les éparpiller un maximum dans les données

- Choix K cluster:

Analyse coefficient silhouette => **forme cluster**: denses + bien séparés

- ~1 : sample est éloigné des clusters voisins
- ~0 : sample est sur ou proche de la frontière 2 clusters

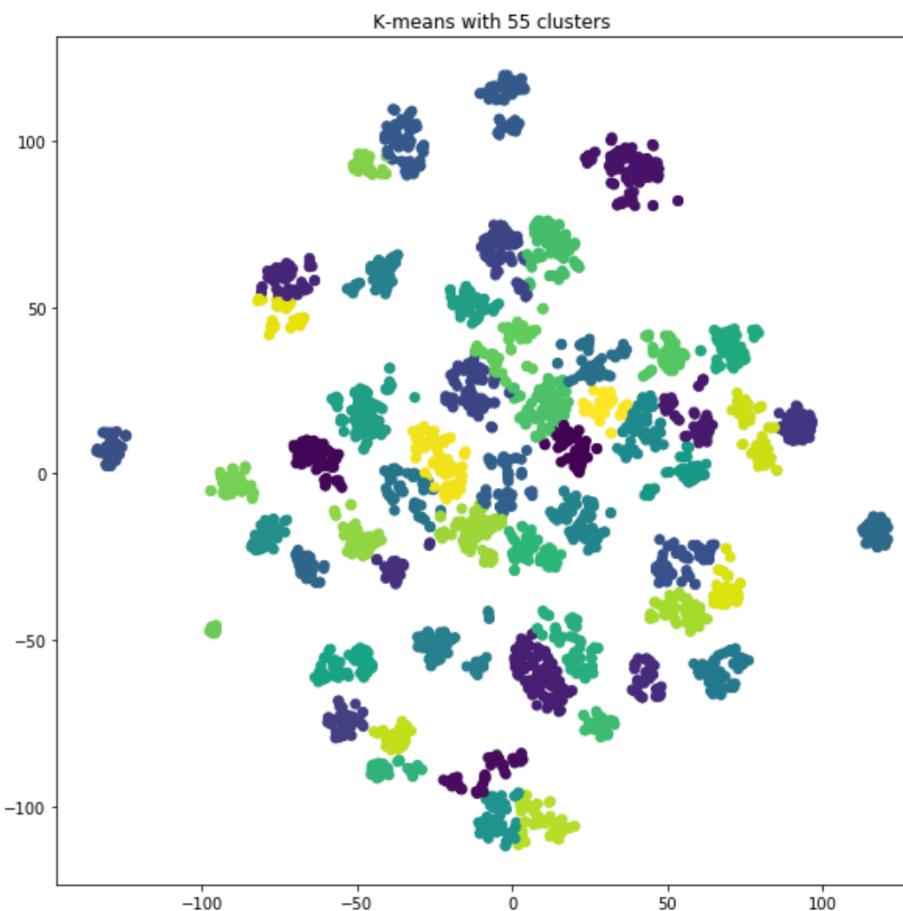
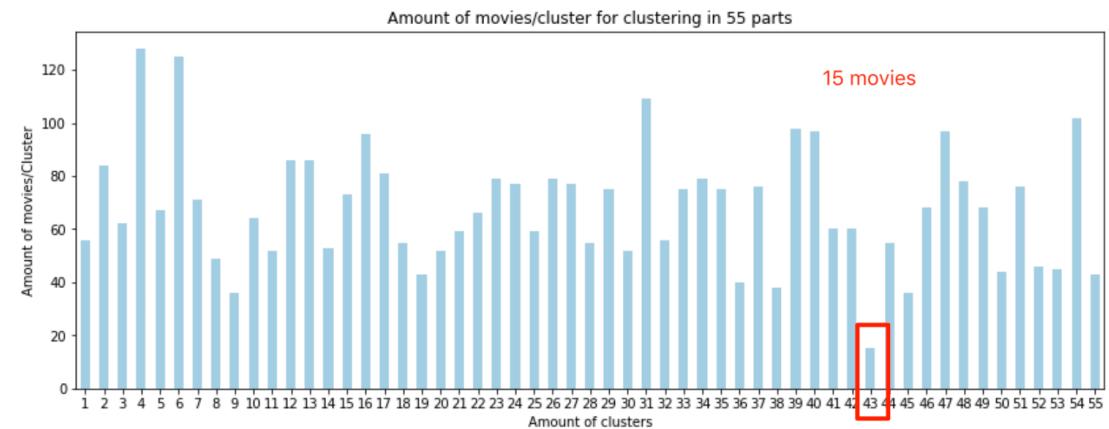


Clustering (2)

K-means++

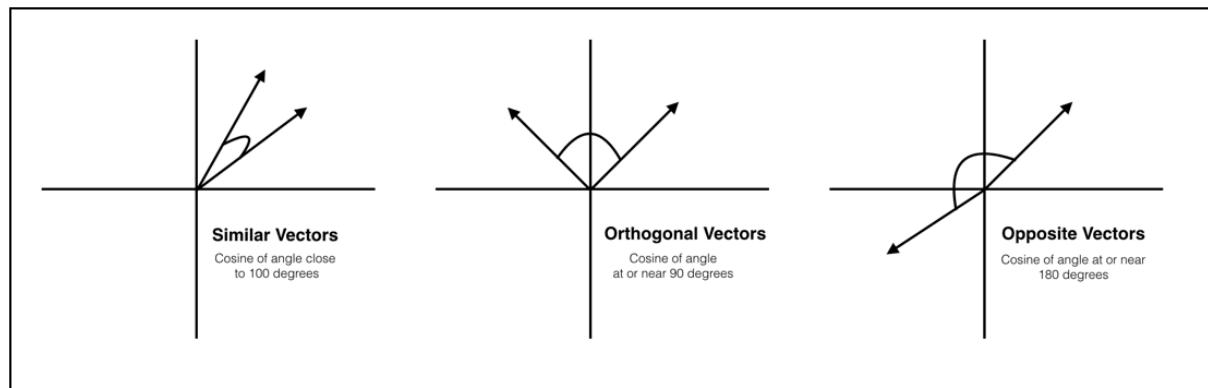
- choix $K = 55$

=> min= 15 films/cluster



Similarité dans cluster (1)

- besoin d'une mesure de similarité entre 2 films
- similarité en ML
 - ~ **distance** avec les dimensions représentant « features » des objets (films)
=> **distance faible** = degré élevé de similarité
- **Similarité cosinus:**
 - * *produit scalaire normalisé* entre 2 attributs
 - * popularité : très efficace même avec « **Vecteurs Creux** » (Sparse Vectors)



Similarité dans cluster (1)

Similarité Cosinus:

- appliquée sur: director_name, actor_1,2,3_name, plot_keywords

- **CountVectorizer**

| | 1980s | 1980s | monster | aaron | aaron | stanford | abrams | abrams | abrams | action | \ |
|---|-------|-------|---------|-------|-------|----------|--------|--------|--------|--------|---|
| 0 | 0 | | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | |
| 1 | 0 | | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | |
| 2 | 0 | | 0 | 0 | creux | | 0 | 0 | 0 | 0 | |
| 3 | 0 | | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | |

| | action | hero | adam | adam | baldwin | ... | arctic | arctic | captain | \ |
|---|--------|------|------|------|---------|-----|--------|--------|---------|---|
| 0 | 0 | 0 | | 0 | ... | | 0 | 0 | 0 | |
| 1 | 0 | 0 | | 0 | ... | | 0 | 0 | 0 | |
| 2 | 0 | 0 | | 0 | ... | | 0 | 0 | 0 | |
| 3 | 0 | 0 | | 0 | ... | | 0 | 0 | 0 | |
| 4 | 0 | 0 | | 0 | ... | | 0 | 0 | 0 | |

- **Similarité cosinus**

```
cosine_sim
[[1.          0.          0.          ... 0.          0.          0.        ]
 [0.          1.          0.          ... 0.          0.          0.        ]
 [0.          0.          1.          ... 0.          0.          0.        ]
 ...
 [0.          0.          0.          ... 1.          0.13483997  0.22226034]
 [0.          0.          0.          ... 0.13483997  1.          0.11773763]
 [0.          0.          0.          ... 0.22226034  0.11773763  1.        ]]
```

Part V:

Modèle final et performances

Recommendations (1)

Logique:

- 1. Récupérer label du cluster du film concerné
- 2. Récupérer tout le cluster correspondant à ce label
- 3. Dans ce cluster : calcule des similarités cosinus
- 4. Choisir les 5 films avec les similarités cosinus les plus élevées



Recommendations (2)

Action: SPECTRE

| | movie_title | title_year | budget | gross | imdb_score | movie_facebook_likes | country | content_rating | genres |
|---|---|------------|-----------|-----------|------------|----------------------|---------|----------------|----------------------------------|
| 0 | Spectre | 2015 | 245000000 | 200074175 | 6.8 | 85000 | UK | PG-13 | Action Adventure Thriller |
| 1 | Skyfall | 2012 | 200000000 | 304360277 | 7.8 | 80000 | UK | PG-13 | Action Adventure Thriller |
| 2 | Spider-Man 3 | 2007 | 258000000 | 336530303 | 6.2 | 0 | USA | PG-13 | Action Adventure Romance |
| 3 | Spider-Man 2 | 2004 | 200000000 | 373377893 | 7.3 | 0 | USA | PG-13 | Action Adventure Fantasy Romance |
| 4 | Spider-Man | 2002 | 139000000 | 403706375 | 7.3 | 5000 | USA | PG-13 | Action Adventure Fantasy Romance |
| 5 | Pirates of the Caribbean: On Stranger Tides | 2011 | 250000000 | 241063875 | 6.7 | 58000 | USA | PG-13 | Action Adventure Fantasy |

| | movie_title | director_name | actor_1_name | actor_2_name | actor_3_name | plot_keywords |
|---|---|---------------|-----------------|---------------|------------------|---|
| 0 | Spectre | sam mendes | christoph waltz | rory kinnear | stephanie sigman | [bomb, espionag, sequel, spi, terrorist] |
| 1 | Skyfall | sam mendes | albert finney | helen mccrory | rory kinnear | [childhood hom, intelligence ag] |
| 2 | Spider-Man 3 | sam raimi | j.k. simmons | james franco | kirsten dunst | [spider man, villain] |
| 3 | Spider-Man 2 | sam raimi | j.k. simmons | james franco | kirsten dunst | [death, doctor, scientist, super villain] |
| 4 | Spider-Man | sam raimi | j.k. simmons | james franco | kirsten dunst | [evil, spider, spider man, superhero] |
| 5 | Pirates of the Caribbean: On Stranger Tides | rob marshall | johnny depp | sam claflin | stephen graham | [captain, pirat, reveng, soldier] |

Recommendations (3)

Horreur: Paranormal Activity

| | movie_title | title_year | budget | gross | imdb_score | movie_facebook_likes | country | content_rating | genres |
|---|----------------------------|------------|----------|-----------|------------|----------------------|---------|----------------|---------------------------------|
| 0 | Paranormal Activity | 2007 | 15000 | 107917283 | 6.3 | 12000 | USA | R | Horror |
| 1 | Paranormal Activity 2 | 2010 | 3000000 | 84749884 | 5.7 | 14000 | USA | R | Horror |
| 2 | The Blair Witch Project | 1999 | 60000 | 140530114 | 6.4 | 0 | USA | R | Horror |
| 3 | The Devil Inside | 2012 | 1000000 | 53245055 | 4.2 | 12000 | USA | R | Horror |
| 4 | Wolf | 1994 | 70000000 | 65012000 | 6.2 | 0 | USA | R | Drama Horror Romance Thriller |
| 5 | The Ghost and the Darkness | 1996 | 55000000 | 38553833 | 6.8 | 0 | USA | R | Adventure Drama Horror Thriller |

| | movie_title | director_name | actor_1_name | actor_2_name | actor_3_name | plot_keywords |
|---|----------------------------|--------------------|--------------------|-------------------|---------------------|--|
| 0 | Paranormal Activity | oren peli | micah sloat | ashley palmer | amber armstrong | [found footag] |
| 1 | Paranormal Activity 2 | tod williams | sprague grayden | molly ephraim | micah sloat | [california, hous, nanni, security camera] |
| 2 | The Blair Witch Project | daniel myrick | heather donahue | joshua leonard | michael c. williams | [found footag, looking at the camera, maryland] |
| 3 | The Devil Inside | william brent bell | fernanda andrade | claudiu trandafir | brian johnson | [critically bash, demonic possess, exorc, exor...] |
| 4 | Wolf | mike nichols | christopher knight | peter gerety | ron rifkin | [blood, werewolf, wolf] |
| 5 | The Ghost and the Darkness | stephen hopkins | tom wilkinson | bernard hill | om puri | [bridg, engin, hunter, lion] |

Recommendations (4)

Comics: The Avengers

| | movie_title | title_year | budget | gross | imdb_score | movie_facebook_likes | country | content_rating | genres |
|---|----------------------------|------------|-----------|-----------|------------|----------------------|---------|----------------|-------------------------|
| 0 | The Avengers | 2012 | 220000000 | 623279547 | 8.1 | 123000 | USA | PG-13 | Action Adventure Sci-Fi |
| 1 | Avengers: Age of Ultron | 2015 | 250000000 | 458991599 | 7.5 | 118000 | USA | PG-13 | Action Adventure Sci-Fi |
| 2 | Captain America: Civil War | 2016 | 250000000 | 407197282 | 8.2 | 72000 | USA | PG-13 | Action Adventure Sci-Fi |
| 3 | Iron Man 2 | 2010 | 200000000 | 312057433 | 7.0 | 18000 | USA | PG-13 | Action Adventure Sci-Fi |
| 4 | Iron Man 3 | 2013 | 200000000 | 408992272 | 7.2 | 95000 | USA | PG-13 | Action Adventure Sci-Fi |
| 5 | Iron Man | 2008 | 140000000 | 318298180 | 7.9 | 10000 | USA | PG-13 | Action Adventure Sci-Fi |

| | movie_title | director_name | actor_1_name | actor_2_name | actor_3_name | plot_keywords |
|---|----------------------------|---------------|-------------------|--------------------|--------------------|--|
| 0 | The Avengers | joss whedon | chris hemsworth | robert downey jr. | scarlett johansson | [alien inva, assassin, battl, soldier] |
| 1 | Avengers: Age of Ultron | joss whedon | chris hemsworth | robert downey jr. | scarlett johansson | [artificial intellig, based on comic book, cap... |
| 2 | Captain America: Civil War | anthony russo | robert downey jr. | scarlett johansson | chris evans | [based on comic book, knife, marvel cinematic ...] |
| 3 | Iron Man 2 | jon favreau | robert downey jr. | scarlett johansson | jon favreau | [militari, reveng] |
| 4 | Iron Man 3 | shane black | robert downey jr. | jon favreau | don cheadle | [armor, explo, terrorist] |
| 5 | Iron Man | jon favreau | robert downey jr. | jeff bridges | jon favreau | [afghanistan, billionair, inventor, playboy, u...] |

Recommendations (5)

Améliorations:

| | movie_title | director_name | actor_1_name | actor_2_name | actor_3_name | plot_keywords |
|---|---|---------------|-----------------|---------------|------------------|---|
| 0 | Spectre | sam mendes | christoph waltz | rory kinnear | stephanie sigman | [bomb, espionag, sequel, spi, terrorist] |
| 1 | Skyfall | sam mendes | albert finney | helen mccrory | rory kinnear | [childhood hom, intelligence ag] |
| 2 | Spider-Man 3 | sam raimi | j.k. simmons | james franco | kirsten dunst | [spider man, villain] |
| 3 | Spider-Man 2 | sam raimi | j.k. simmons | james franco | kirsten dunst | [death, doctor, scientist, super villain] |
| 4 | Spider-Man | sam raimi | j.k. simmons | james franco | kirsten dunst | [evil, spider, spider man, superhero] |
| 5 | Pirates of the Caribbean: On Stranger Tides | rob marshall | johnny depp | sam claflin | stephen graham | [captain, pirat, reveng, soldier] |

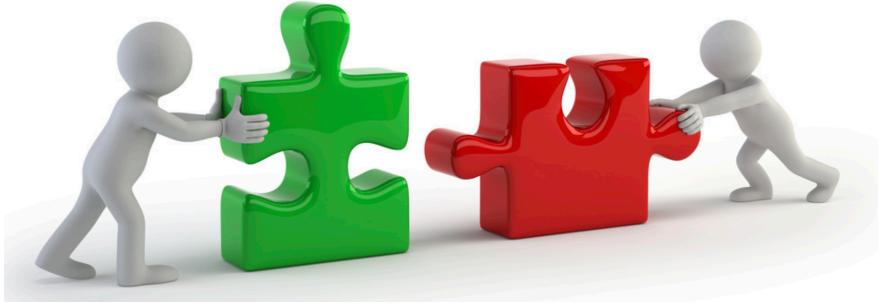
- Regrouper les **prénoms** et **noms** de familles des acteurs/directeur

| | movie_title | director_name | actor_1_name | actor_2_name | actor_3_name | plot_keywords |
|---|-----------------------|-------------------|-------------------|------------------|----------------------|--|
| 0 | Spectre | sam mendes | christoph waltz | rory kinnear | stephanie sigman | [bomb, espionag, sequel, spi, terrorist] |
| 1 | Skyfall | sam mendes | albert finney | helen mccrory | rory kinnear | [childhood hom, intelligence ag] |
| 2 | Iron Man 3 | shane black | robert downey jr. | jon favreau | don cheadle | [armor, explo, terrorist] |
| 3 | Armageddon | michael bay | bruce willis | steve buscemi | will Patton | [astronaut, bomb, outer spac] |
| 4 | The Dark Knight Rises | christopher nolan | tom hardy | christian bale | joseph gordon-levitt | [decept, imprison, police off, terrorist plot] |
| 5 | Star Trek Beyond | justin lin | sofia boutella | melissa roxburgh | lydia wilson | [hatr, sequel, space opera, star trek] |

Part VI:

Conclusions

Conclusions (1)



- Pour finaliser ce projet de recommandation de films:

Nous avons essayé d'opérer de façon structurée

1. Prise de connaissance de la Base:

Avec les étapes indispensables de :

- A. Nettoyage
- B. Exploration univariée et multivariée

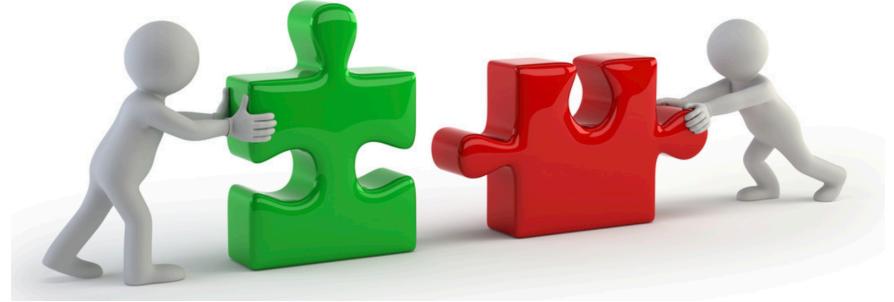
2. Analyse de la littérature (annexes proposées):

Avec différents types de systèmes de recommandation existants
(Content Based, Collaborative)

3. Réduction dimensionnelle des données:

D'abord **ACP** puis **t-SNE**

Conclusions (2)



4. Clustering sur données réduites:

K-means++

5. Similarités entre films d'un même cluster:

Similarité cosinus

6. Analyse des recommandations:

Sur base des paramètres utilisées pour la t-SNE et la similarité cosinus

Etude pour films de genres différents (action, aventure, horreur, ...) dans l'ensemble cohérent!

Une amélioration a été faite avec le regroupement des noms et prénoms

Pistes

Ici par rapport à la base de films disponible, on a fait du « **content Filtering** »
=> **tout le monde** obtient les *mêmes recommendations*.

Une piste d'amélioration serait d'investiguer la possibilité d'acquérir également des *informations sur les utilisateurs* afin de faire des recommandations *plus personnalisées*.

= « **Collaboratives Filtering** »



