Université Claude Bernard Lyon I
Master II Informatique - Data Science

Steam Games

Analyse de données sur les recommandations de jeux vidéos Steam

Gernido HANAMPATRA
p2021498

17 Novembre 2024

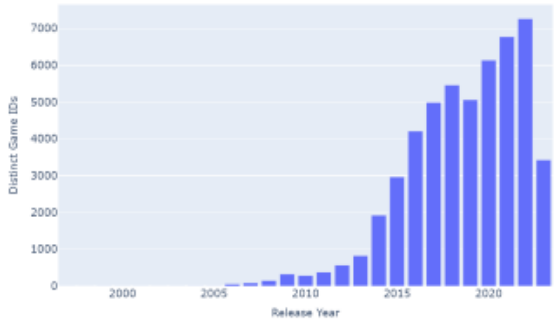
Contents

1 Introduction

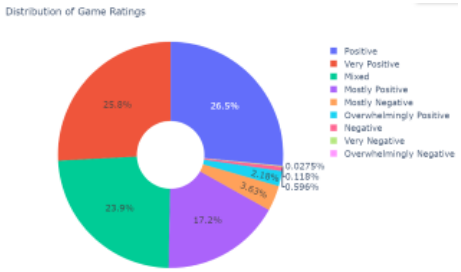
Ce rapport analyse un jeu de données Steam, une des plus grandes plateformes de distribution de jeux vidéos au monde. L'objectif est d'explorer les données des utilisateurs, des jeux et des recommandations afin d'extraire des informations significatives sur le comportement des utilisateurs et les relations entre les différents jeux.

Le jeu de données Steam [1] utilisé comprend quatre fichiers principaux :

- **users.csv** : 14 millions d'utilisateur. Il renseigne leur identifiant unique anonymisé (`user_id`), le nombre de produits qu'ils possèdent (`products`) et leur nombre de critiques publiées (`reviews`).
- **games.csv** : 50k jeux Steam avec leur (`app_id`), (`title`), (`date_release`), leur compatibilité avec différents OS, leur évaluation globale (`rating`), le ratio d'évaluations positives (`positive_ratio`), le nombre total de critiques d'utilisateurs (`user_reviews`), et leur prix avant et après réduction.

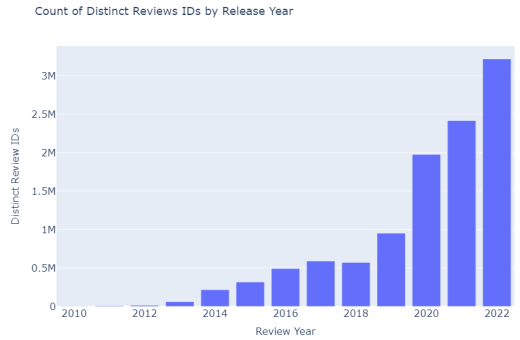


(a) Histogramme sur le nombre de jeux publiés sur Steam par année de sortie, elle montre l'évolution de la prépondérance de Steam sur le marché. Tendance d'abord exponentielle, puis linéaire à partir de 2015. Elle est maximale pendant la période du covid. Elle s'arrête le 10/23, ce qui explique en partie la légère baisse.

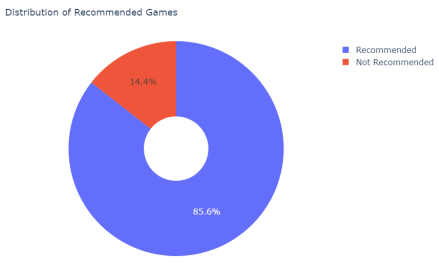


(b) Pie chart sur la distribution des évaluations des jeux. Les ratings sont en majorité positives. Ce label dépend à la fois du ratio d'évaluations positives et du nombre total de reviews.

- **recommendations.csv** : 10 millions de recommendations. Pour chaque jeu on a son identifiant (`app_id`), celui de l'utilisateur associé à la review (`user_id`), le nombre d'heures jouées par l'utilisateur au moment de la review (`hours`), la date de la critique (`date`) et la recommandation du joueur (`is_recommended`).



(a) Nombre de recommendations sur Steam par année. On remarque encore une fois que la popularité de la plateforme suit une loi de puissance : plus il y a de jeux, plus il y a de joueurs, et vice-versa.



(b) Distribution des recommendations positives/négatives. Les recommendations sont grandement déséquilibrées en faveur des avis positifs.

- **games_metadata.json** : Détail des métadonnées supplémentaires sur les jeux avec leur (`description`) et une liste de tags associés au jeu (`tags`).

Pour analyser ce jeu de données, différentes méthodes seront utilisées, incluant les statistiques descriptives, la visualisation des données (histogrammes, diagrammes de dispersion), l'analyse de corrélation, la modélisation de sujets, la réduction de dimensionnalité (PCA et t-SNE), le clustering (DBSCAN, Gaussian Mixture), l'extraction de règles d'association (Apriori) et la visualisation de graphes.

2 Analyse Descriptive des Données

2.1 Prétraitement des Données

Le prétraitement des données est une étape cruciale pour garantir la qualité et la cohérence des données avant de les analyser. Différentes opérations ont été effectuées sur les fichiers de données :

users.csv

- **Suppression des lignes incohérentes ou inutiles :** Les lignes correspondant à des utilisateurs n'ayant aucun jeu, mais 0 ou plus d'une critique ont été supprimées. `10% des lignes supprimées.`
- **Normalisation des données :** Les colonnes `products` et `reviews` ont été normalisées à l'aide de `RobustScaler`. Cette normalisation permet de minimiser l'impact des valeurs aberrantes et de mieux comparer les utilisateurs.

games.csv

- **Conversion des évaluations :** Les évaluations textuelles des jeux `("Overwhelmingly Positive", "Very Positive", etc.)` ont été converties en valeurs numériques allant de 0 à 8. Cela facilite l'analyse quantitative des évaluations.

recommendations.csv

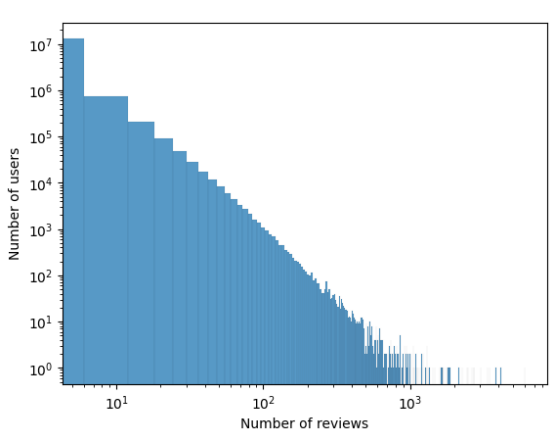
- **Traitement des dates :** Conversion en format `datetime` et extraction de l'année.
- **Suppression des doublons :** Les critiques dupliquées ont été ignorées en conservant la critique la plus récente pour chaque utilisateur et chaque jeu. Cela garantit que chaque avis, même après modification, soit pris en compte une seule fois.
- **Conversion des recommendations:** La recommendation des utilisateurs (`False, True`) est convertie en 0 ou 1.
- **Traitement des valeurs manquantes :** Les valeurs manquantes ont été remplacé par des 0. La suite portant essentiellement sur la recommendation, cette approche suppose qu'une absence d'évaluation équivaut à une absence ou d'interaction, ce qui simplifie l'analyse mais peut introduire un biais en assimilant toutes les absences à des retours négatifs.
- **Jointure entre `games.csv` et `recommendations.csv` :** Ajout d'informations sur les jeux (`"title", "date", "tags"`).

Problèmes rencontrés lors du prétraitement des données

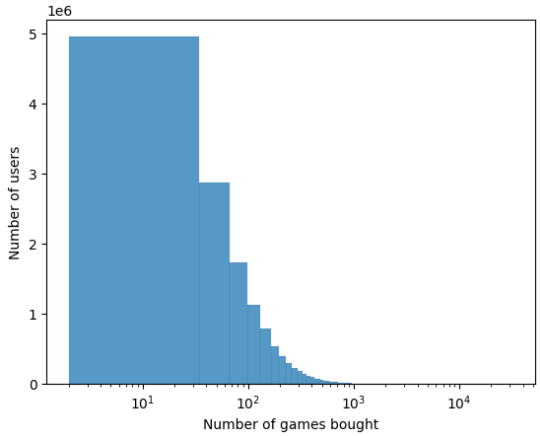
- Des valeurs aberrantes ont été identifiées dans les prix des jeux (prix original à 0) 5b. On pourrait penser que ce sont les jeux gratuits, mais ce n'est pas le cas, d'autant plus que leur prix final est supérieur à 0. (`ex: Lies of P, Forgive Me Father 2, One Piece Odyssey etc.`). Cela est probablement dû à des erreurs dans la récupération de données.
- Des incohérences entre `games.csv` et `recommendations.csv` ont été observées, notamment en ce qui concerne le nombre d'heures jouées pour certains jeux. Par exemple, le jeu avec l'ID `495050` a 991 critiques, mais il a été joué pendant seulement 6 minutes au total selon `recommendations.csv`. Ainsi il n'y a pas de correspondances exactes entre ces deux datasets.
- Les dimensions étant importantes, il était nécessaire de réduire la taille de l'échantillon lors de la création de matrices pivot, afin d'éviter une surcharge de la mémoire. La solution adoptée a été de se restreindre à un échantillon aléatoire stratifié.

2.2 Analyse des Utilisateurs

Statistiques descriptives Le fichier `users.csv` contient plus de 14 millions d'utilisateurs. En moyenne, chaque utilisateur possède 116 jeux, tandis qu'il ne publie que 3 critiques.



(a) Histogramme du nombre de critiques par utilisateur (échelle logarithmique)



(b) Histogramme du nombre de jeux possédés par utilisateur (abscisse logarithmique)

L'histogramme (3a) révèle une distribution en loi de puissance, indiquant que la grande majorité des utilisateurs publient moins de 10 critiques, tandis qu'une très faible proportion d'utilisateurs (0,01%) publie plus de 100 reviews. La matrice pivotée des recommendations sera donc très clairsemée.

L'histogramme (3b) montre une distribution similaire, avec une densité décroissant plus lentement. Il n'est pas étonnant de constater que 32% des utilisateurs ont plus de 100 jeux dans leur ludothèque pour les raisons suivantes :

- Il y a souvent des jeux vendus en bundle à des prix compétitifs.
- La baisse du coût du stockage, la démocratisation de la fibre, la confiance en la plateforme (pas de perte de données), la communauté gargantuesque en expansion

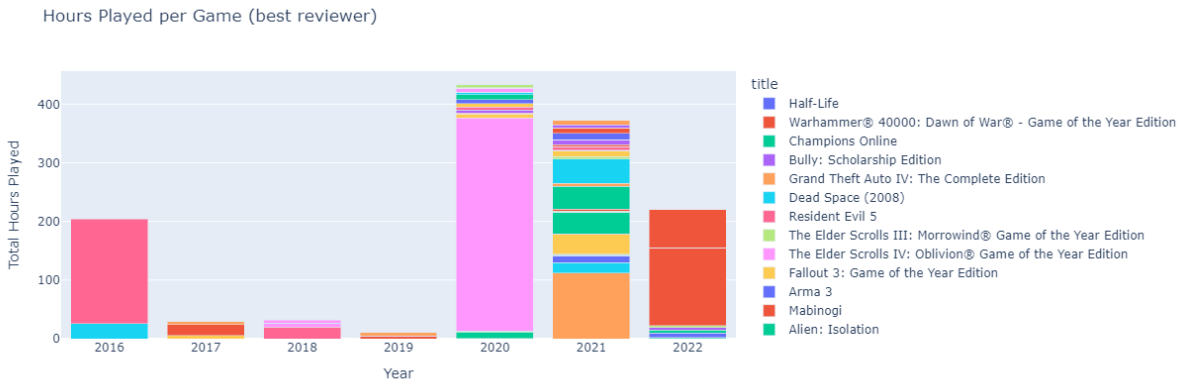
(1a, 2a) qui pousse à s'investir sur cette platforme et pas une autre..

- En donnant des badges en fonction du nombre de jeux possédés, des accomplissemnts faits en jeu etc. Steam encourage la collection de jeux vidéos. Steam peut dès lors devenir une vitrine, public ou personnel ; tendance souvent discutée dans les forums en ligne.

Cas particulier

- L'utilisateur le plus actif en termes de critiques a publié 6045 critiques, ce qui représente 76% des jeux qu'il possède
- L'utilisateur possédant le plus de jeux (plus de 32 000) n'a évalué que 0,03% d'entre eux.

On pourrait alors se demander si ces utilisateurs sont en fait des bots. Cependant, comme il n'y a pas de correspondances exactes entre **users.csv** et **recommendations.csv** il est difficile de généraliser. Par exemple, pour l'utilisateur avec le plus de critiques: dès 6045 reviews, on ne peut en récupérer que 83. Cet utilisateur semble toutefois être une vraie personne si l'on regarde ses temps de jeux au moment des reviews par année (4a).



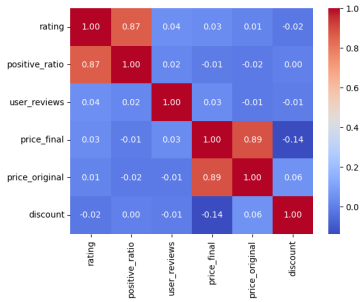
(a) Nombre d'heures jouées au moment d'une review par l'utilisateur 11764552 dans le dataset `users.csv` (la légende est coupée).

Conclusion La plupart des utilisateurs de Steam consomment des jeux vidéo sans laisser de review, ce qui est typique des plateformes de ce type (3a). Ce qui est étonnant en revanche, c'est que lorsque des avis sont émis, ils sont généralement positifs (1b, 2b). Toutefois, ce biais positif pourrait également venir de la façon dont les jeux ont été explorés lors de la construction du dataset. Steam mettant plutôt en avant des jeux avec beaucoup de recommendations positives.

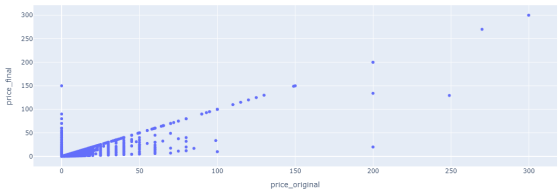
2.3 Analyse des Jeux

L'analyse descriptive des jeux a été réalisée en utilisant certaines variables de **games.csv**, comme la date de sortie, le prix et les évaluations.

- Les ratings globaux ("Extrêmement positifs"...) semblent corrélés au ratio des évaluations positives. C'est le cas : plus le rating est élevé, plus le ratio l'est également. Mais on remarque aussi que les ratings peuvent se chevaucher (ratings



(a) Matrice de corrélation de **games.csv** (avec Pearson)



(b) Prix final en fonction du prix original de chaque jeu

différents mais même ratio). Cela est dû au fait que le nombre de reviews entre également en compte dans le calcul du rating par Steam.

- L'évolution des prix originaux et finaux est linéaire. Ce qui suggère que le prix final est seulement après réduction, il ne prend donc pas en compte les éventuels augmentations de prix.

3 Analyse Approfondie

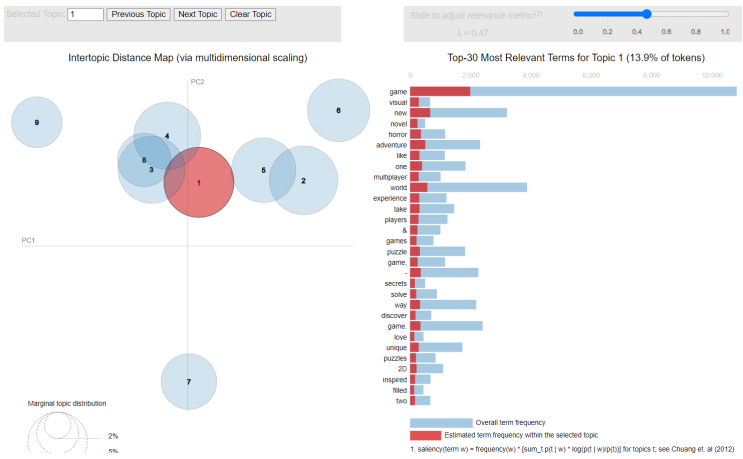
3.1 Topic Modeling

L'algorithme LDA (Latent Dirichlet Allocation) a été appliqué aux descriptions des jeux pour découvrir leurs thèmes dominants.

3.1.1 Prétraitement des descriptions

- **Suppression des stop words** : Les descriptions étant en anglais, on supprime les "stop words" pour cette langue. Cela permet de ne pas prendre en compte les mots vides ("it, this, a, is, are..") lorsque l'on calcule la fréquence de chaque mot.
- **Tokenisation des mots** : Utilisation de la librairie `gensim` pour générer des vecteurs de chaque mot.

3.1.2 Topics



(a) Top 30 des termes les plus représentatifs du Topic 1.

Analyse Topic 1 - Jeux d'aventure narratifs et immersifs

- Les mots **Adventure**, **discover**, **experience**, **puzzle**, et **secrets** suggèrent des jeux qui se concentrent sur l'exploration, la découverte et la résolution d'énigmes.
- **Multiplayer** et **players** impliquent une dimension sociale dans certains de ces jeux.
- **Novel** et **unique** évoquent une expérience originale ou créative.

Analyse Topic 2 - Jeux VR et multijoueur axés sur l'action et la stratégie

- **VR, HTC, Oculus, Vive** : Indiquent clairement une association avec des jeux en réalité virtuelle, compatibles avec des casques VR.
- **Multiplayer, player, online** : Jeux multijoueurs.
- **Racing, FPS** : Genres populaires en VR, comme les jeux de tir à la première personne (FPS) et les jeux de course.
- **Puzzle, strategy, fast-paced, action** : Styles de gameplay variés, allant de jeux de réflexion et stratégie à des expériences rapides et orientées action.

Analyse Topic 4 - Jeux de science-fiction axés sur l'exploration, la construction et l'aventure spatiale

- **Alien, planet, space, Earth, world, sci-fi, life** : Évoquent des environnements extraterrestres et des thématiques de science-fiction.
- **Exploration, adventure, journey, discover, mysterious** : Accent sur l'exploration et la découverte de mondes inconnus..
- **Build, strategy, city** : Indiquent une dimension de construction, de gestion ou de planification stratégique.

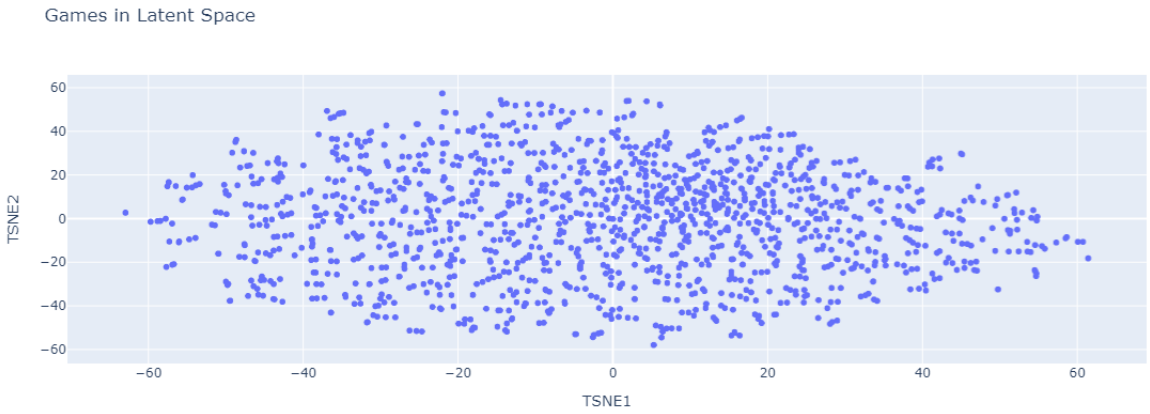
Les topics 1, 2 et 4 mettent en lumière la richesse et la diversité des expériences vidéoludiques proposées sur Steam. Ces topics montrent les efforts de l'industrie du jeu vidéo à s'adapter aux attentes d'un public varié et exigeant sur fond d'avancées technologiques.

3.2 Recommendation (User-Item)

L'objectif est de capturer les informations latentes de nos jeux vidéos. Pour cela, on ne conserve que les colonnes `user_id, title, is_recommended` de notre User-Item dataset.

3.2.1 Extraction des variables latentes et affichage

On entraîne un SVD avec 15 variables pour trouver nos variables latentes. Puis on réduit notre matrice avec une méthode de réduction non-linéaire, t-SNE.



(a) Variables latentes après t-SNE. Chaque point est un jeu vidéo.

Observation En regardant à la main (7a), le lien entre les jeux n'est pas clair. La plupart ont l'air d'être mélangé que ce soit en genre, en année, en type de gameplay ou autres. Cette méthode n'a donc pas été concluante sur ce jeu de données à priori.

3.3 Carte des jeux vidéos (Item-User)

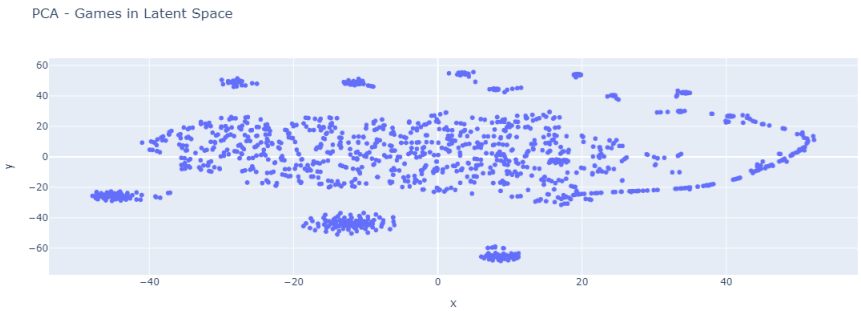
Cette fois-ci on pivote le dataset obtenu précédemment de sorte à ce que chaque ligne soit un jeu, et une colonne un utilisateur. On ne conserve que la recommendation de l'utilisateur.

La matrice étant à la fois clairsemée et de très grandes dimensions, on choisit de faire les optimisations suivantes :

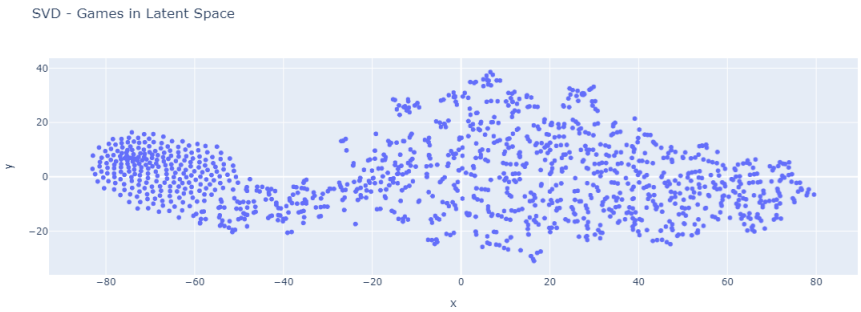
- Stocker la matrice clairsemée avec `scipy.sparse` pour la compresser (seules les valeurs non nulles restent en mémoire).
- Prendre 1 millions d'échantillons par tirage aléatoire stratifié avant de pivoter la matrice. L'idée étant de conserver la proportion de jeux différents de notre dataset initial. Par ailleurs, chaque jeu apparaît au moins une fois pour un résultat plus complet. Grâce à cela, après pivot, 1299 jeux différents sont conservés.

3.3.1 Réduction de dimensions et affichage

On compare deux approches différentes **TruncatedSVD** et **PCA**. On utilise les deux avec 50 dimensions, puis un t-SNE pour afficher le résultat en 2D.



(a) PCA en 50 dimensions sur la matrice pivotée, puis t-SNE



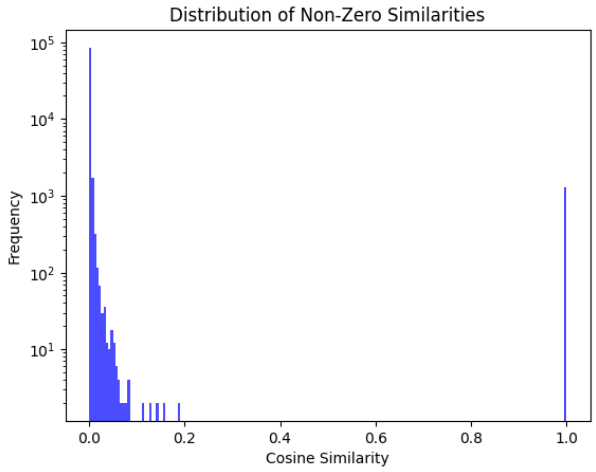
(b) TruncatedSVD en 50 dimensions sur la matrice pivotée, puis t-SNE

Observation Les jeux vidéos sont visiuellement mieux regroupés. Si on vérifie à la main, les jeux appartenant à une même série, ou relativement similaires, sont quelques fois ensembles, contrairement à (7a).

3.3.2 Transformation en Network

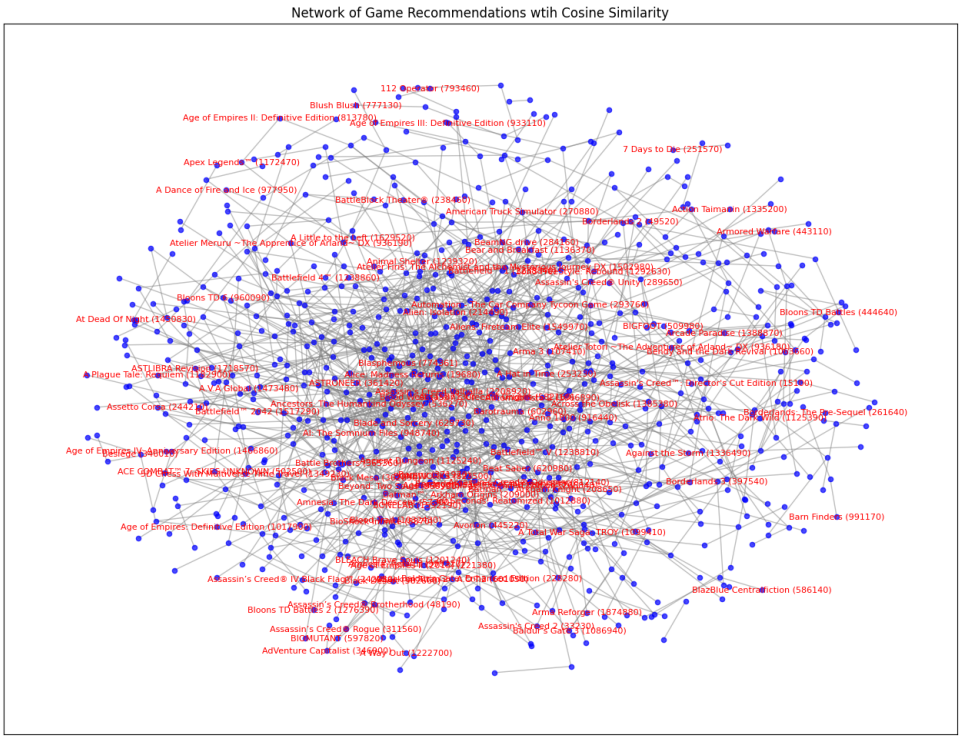
On calcule la matrice des similarités cosinus entre deux jeux pour savoir s'ils doivent être reliés ou non dans le graphe.

Le problème est que le threshold doit être fixé relativement bas si on ne veut pas se retrouver avec la matrice identité (9a). La connexion peut en fait être du bruit.



(a) Distribution des valeurs de similarités cosinus.

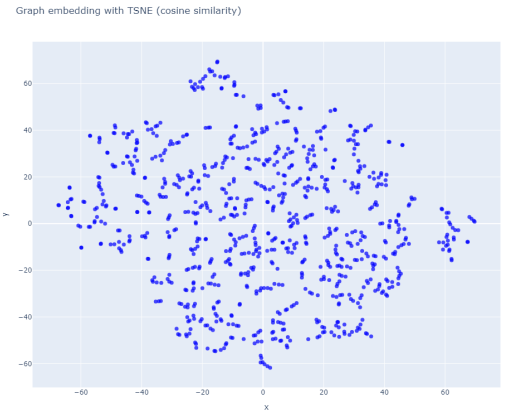
Finalement on obtient le graphe suivant.



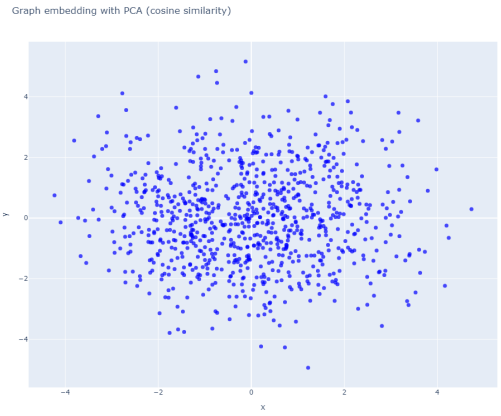
(a) Graphe des jeux vidéos. Il y a lien s'ils sont suffisamment similaires selon leur distance cosinus.

3.3.3 Encodage du graphe

La proximité des nœuds dans l'espace d'encodage du graphe reflète non seulement la similarité entre deux éléments, mais aussi leur degré de connexion au sein de l'ensemble du graphe. L'encodage a été fait avec `Node2Vec` puis réduit avec t-SNE et PCA pour comparer.



(a) Graphe Embedding après réduction (t-SNE)



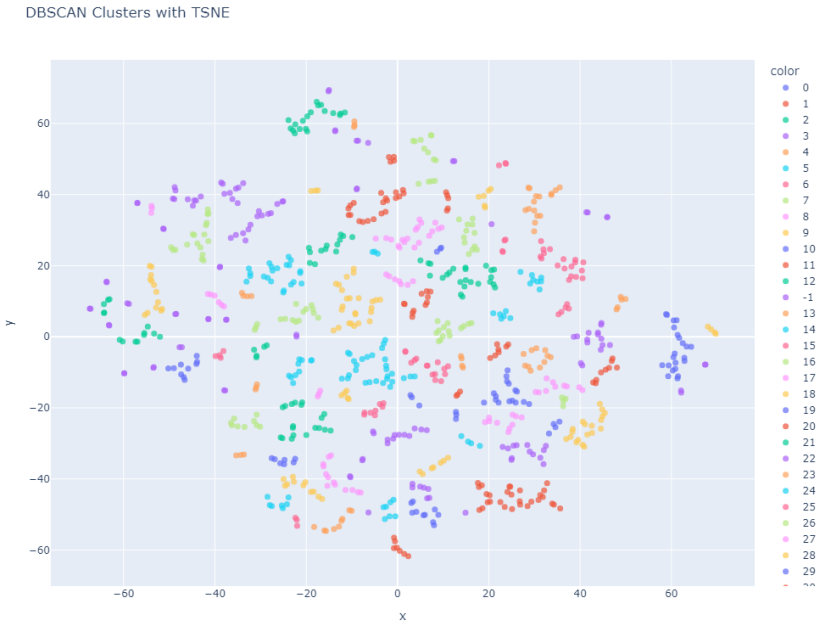
(b) Graphe Embedding après réduction (PCA)

Observation t-SNE semble bien fonctionner cette-fois : la plupart des jeux dans une même série ou du même genre sont regroupés ensembles.

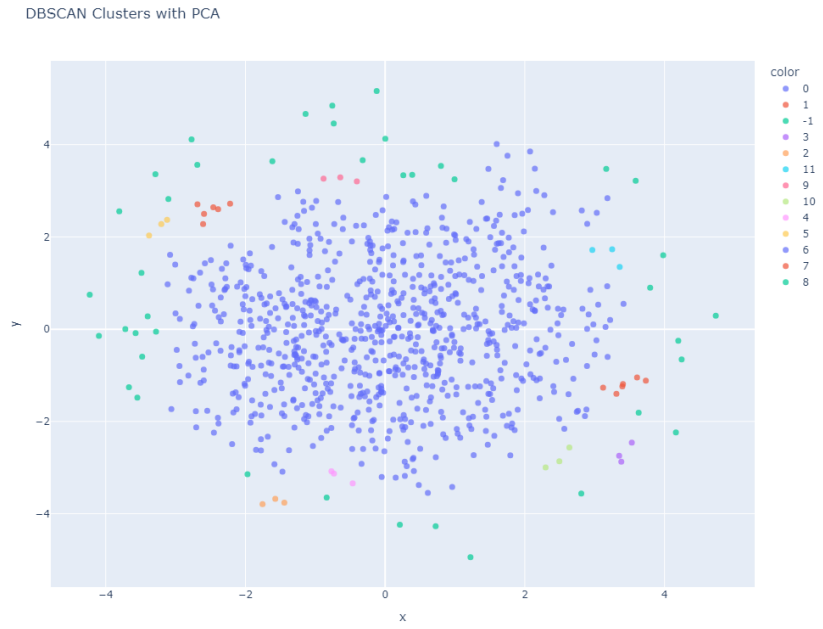
3.3.4 Clustering

On souhaite maintenant trouver nos clusters dans l'espace d'encodage obtenu.

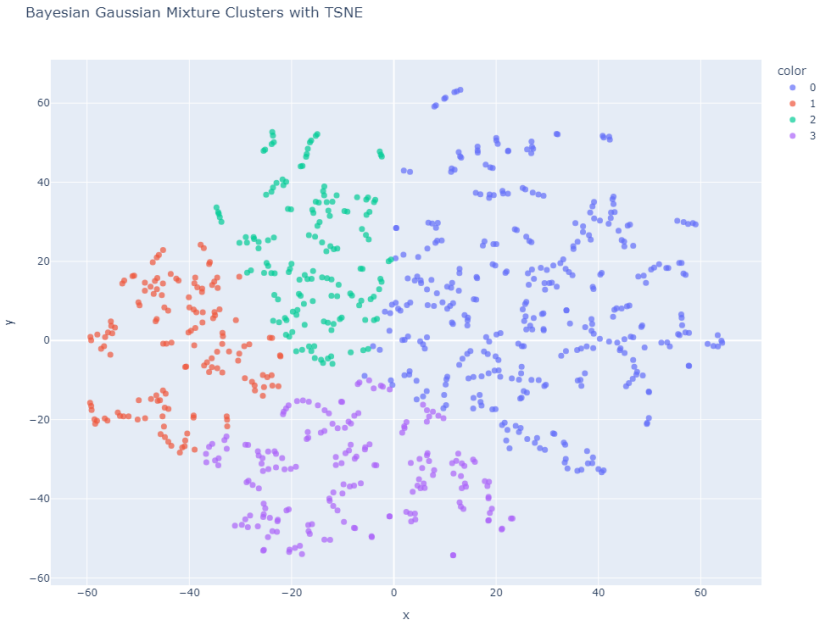
Les algorithmes DBScan et Bayesian Gaussian Mixture ont été appliqués pour découvrir des clusters de jeux sans renseigner de nombres exactes. La silhouette score a été utilisé pour calculer la performance de chaque cluster. Les scores ne sont pas très élevés, surtout le DBScan-PCA qui assigne la mauvaise classe aux points (12b).



(a) DBSCAN (t-SNE). La classe -1 est du bruit. Score Silhouette: 0.36



(b) DBSCAN (PCA). La classe -1 est du bruit. Score Silhouette: -0.19



(a) BGM (t-SNE). La classe -1 est du bruit. Score Silhouette: 0.33

(b) BGM (PCA). La classe -1 est du bruit. Score Silhouette: 0.28

3.4 Frequent Pattern

3.4.1 Algorithme Apriori

On cherche à trouver les k-itemsets les plus souvent recommandés ensembles. Ainsi, on pourra découvrir d'autres relations propres à ces jeux.

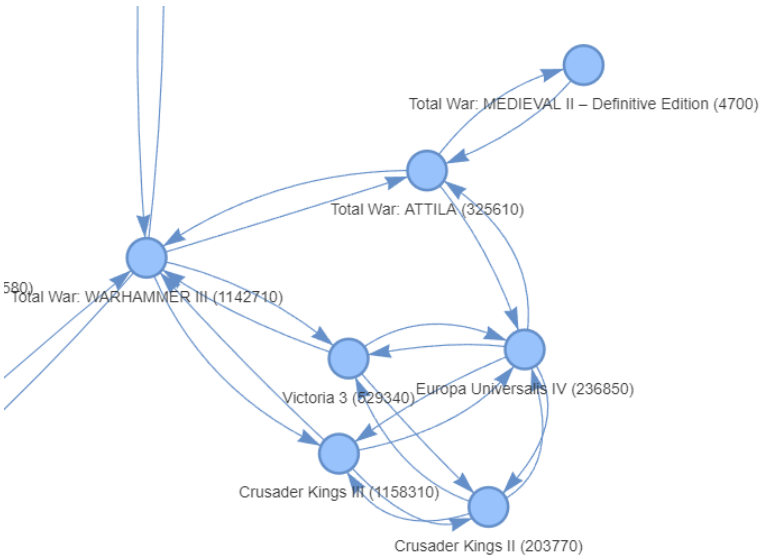
Pour cela, on applique la méthode `apriori` sur la matrice des transactions avec les users en ligne, et les titres des jeux en colonne.

Le problème qui se pose est la faible fréquence des jeux dans le dataset. Le jeu le plus fréquent est Rust, et pourtant il n'a qu'une fréquence d'apparition de 2% sur 899055 transactions. Cela implique que le support minimal pour que l'item soit pris en compte doit être très bas, ici 0.0005%. Non seulement cela a un impact sur la mémoire, mais aussi, avec un support si faible, on pourrait potentiellement ne capturer que du bruit.

3.4.2 Graphe d'association

Toutefois, l'algorithme `apriori` a permis d'identifier des ensembles d'éléments fréquents et des règles d'association entre les jeux.

D'autant plus que parmis toutes les associations trouvées, plus de 12.5% ont un lift supérieur à 2 et 22,5% supérieur à 1. Ce qui démontre que malgré tout, une partie de ces associations sont très significatives.



(a) Jeux Paradox

Observation Les jeux liés ont beaucoup en commun, c'est le cas de Europa Universalis IV, Victoria 3, Crusader Kings III & II :

- Ils ont été publié par le même studio : Paradox Interactive, réputé pour ses jeux complexes de grande stratégie.
- Ce sont tous les trois des "jeux de carte" historique, se déroulant avant le XIX^{e} siècle.
- Dans chacun d'entre eux, on doit gérér l'économie de son état, sa diplomatie, etc.

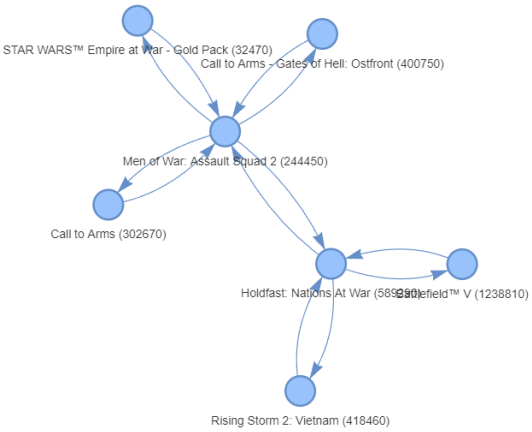


Figure 15. Jeux de simulation de guerre

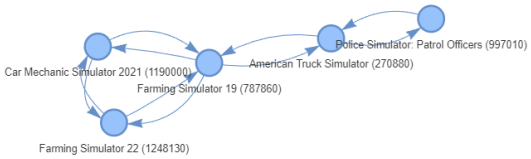


Figure 16. Jeux de simulation métier

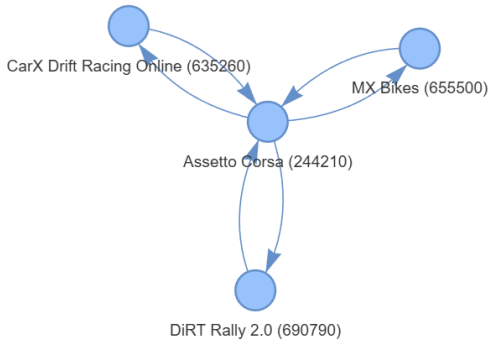
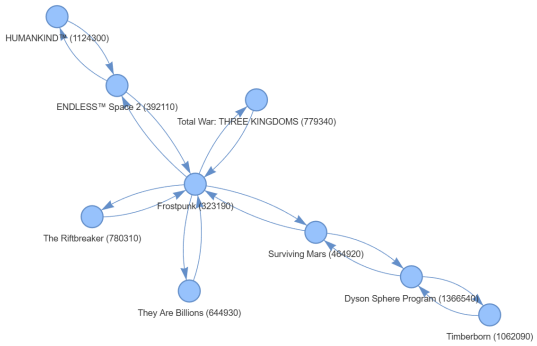
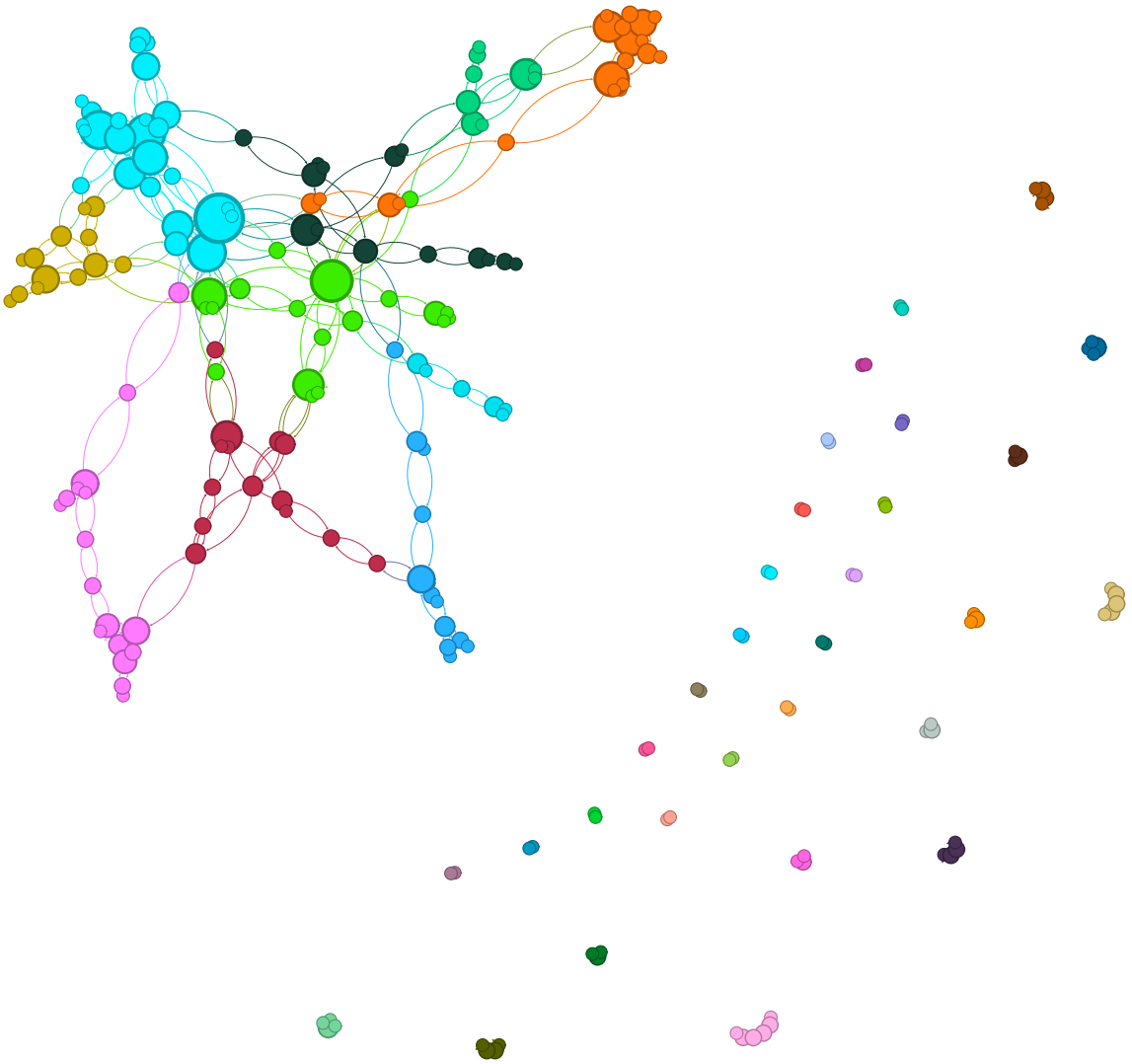


Figure 17. Jeux de simulation de courses



Jeux de constructution de ville/civilisation



(a) Graphe dirigé des associations résultantes sous Gephi. La taille d'un nœud est proportionnelle à son degré, reflétant sa popularité ou son caractère généraliste. La couleur de chaque nœud est déterminée par la communauté à laquelle il appartient, identifiée grâce à la modularité.

4 Conclusion

4.1 Résumé des Résultats

L'analyse des données Steam a permis de mettre en lumière plusieurs aspects importants du comportement des utilisateurs et des relations entre les jeux :

- La plupart des utilisateurs de Steam consomment des jeux sans laisser de commentaires.
- La modélisation de sujets a permis d'identifier les thèmes dominants dans les descriptions des jeux.
- La réduction de dimensionnalité et le clustering ont révélé des regroupements potentiels de jeux similaires, mais les résultats ne sont pas très concluants.
- L'extraction de règles d'association a permis de découvrir des relations intéressantes entre les jeux recommandés.

4.2 Limites

L'analyse a été limitée par plusieurs facteurs :

- Biais potentiels dans les données.
- Incohérences entre les fichiers de données, notamment en ce qui concerne le nombre d'heures jouées.
- Pour le topic modeling, les caractères spéciaux auraient dû être retirés.

4.3 Perspectives

Des analyses futures pourraient approfondir certains aspects de l'analyse :

- Analyse des genres de jeux et de leur évolution au fil du temps.
- Étude de l'impact des facteurs externes (événements, promotions) sur les recommandations.
- Comparaison des résultats avec d'autres plateformes de jeux vidéos.

References

[1] Anton Kozyriev. “Game Recommendations on Steam”. 2024. URL: https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam/data (page 1).