Méthodologie de Nettoyage des Données et Modèle en Étoile Modulaire

Nettoyage Préliminaire des Données (Power Query)

Dans une phase initiale, les données brutes ont été nettoyées et préparées en utilisant **Power Query**. Cette étape a assuré que les ensembles de données soient cohérents et de qualité avant l'analyse en Python. Les opérations effectuées incluent:

Suppression des colonnes inutiles ou vides	Les colonnes sans information significative ou complètement vides ont été éliminées de l'ensemble de données original, réduisant le bruit dans les données.
Gestion des erreurs et des types de données	Les valeurs reconnues comme des erreurs ont été remplacées par null (valeurs manquantes) pour éviter les incohérences. De plus, les types de données (formats de dates, nombres, texte) ont été corrigés pour l'uniformité.
Séparation des champs composés	Les colonnes contenant des listes de valeurs (par exemple, genres multiples, plateformes supportées, noms de développeurs ou d'éditeurs concaténés) ont été séparées en champs distincts pour chaque élément. Cela permet d'analyser chaque attribut séparément sans conflits.
Dédoublonnage préliminaire et normalisation	Des tables distinctes ont été créées pour les entités clés comme **les Genres, les Plateformes, les Développeurs et les Éditeurs,** éliminant les doublons et assurant la cohérence des noms. Par exemple, pour les plateformes, l'utilisation des majuscules/minuscules a été uniformisée et les noms équivalents ont été standardisés (comme "Switch" standardisé en "Nintendo Switch"). De même, pour les genres, une nomenclature cohérente a été choisie (par exemple, "Role-Playing" converti en "RPG"). Ces transformations améliorent la **cohérence** des données, facilitant leur intégration et leur comparaison.
Standardisation des formats numériques	Tous les champs numériques (comme les évaluations, les prix, etc.) ont été formatés de manière uniforme (par exemple, séparateurs décimaux, unités de mesure) pour prévenir les ambiguïtés lors de l'analyse.

Ce pré-traitement a posé des bases solides pour l'analyse, produisant des données "propres" c'est-à-dire sans erreurs grossières et **plus fiables pour les phases suivantes**.

Traitements des Données en Python

Après le nettoyage préliminaire, les ensembles de données ont été importés en Python pour des traitements avancés supplémentaires. À cette étape, diverses techniques de nettoyage et de transformation des données ont été appliquées:

Gestion des données manquantes	Les lignes avec des portions majoritaires de valeurs absentes ont été automatiquement supprimées. Supprimer les enregistrements trop incomplets est une stratégie courante pour assurer la qualité globale de l'ensemble de données (privilégiant l'analyse de données suffisamment informatives par rapport aux remplissages incertains).

Élimination des doublons	Un contrôle a été effectué sur les identifiants uniques (comme slug, nom du jeu et ID) pour supprimer les éventuels doublons présents. La déduplication garantit l'unicité de chaque jeu dans la base de données et préserve l'intégrité des informations.
Normalisation du texte	Les champs textuels ont été nettoyés des caractères spéciaux ou des formatages incohérents, rendant les majuscules/minuscules uniformes si nécessaire. Par exemple, les espaces ou symboles redondants dans les noms des jeux ont été supprimés, pour permettre des comparaisons précises et des jointures fiables entre les tables.
Standardisation des plateformes et des genres	Comme anticipé, en Python, la standardisation initiée dans Power Query a été encore affinée. Par exemple, toutes les occurrences de plateformes ont été mappées à une dénomination standard ("switch" → "Nintendo Switch"), et les genres ont été agrégés sous des abréviations cohérentes ("Action RPG" → "GDR Action", etc.). Cette uniformité sémantique facilite les agrégations et les filtrages corrects dans les analyses futures.
Conversion des variables booléennes	Les champs avec des réponses oui/non (par exemple, indiquant si un jeu supporte des fonctionnalités spécifiques) ont été convertis en valeurs booléennes standard True/False. Cela simplifie les requêtes ultérieures et l'utilisation de ces champs dans des conditions logiques ou des filtres.
Alignement des dates	Les formats de date anormaux ou les années inconsistantes (par exemple, les années à 2 chiffres interprétées de manière erronée) ont été identifiés et corrigés. Toutes les dates de sortie des jeux sont maintenant enregistrées dans un format cohérent (YYYY-MM-DD), assurant un bon ordonnancement temporel et des agrégations par année/mois sans erreurs.

Ces traitements en Python ont encore renforcé la cohérence et la qualité de l'ensemble de données consolidé. Nous disposons maintenant de données standardisées, sans doublons et sans valeurs anormales, prêtes à être intégrées avec des sources externes et analysées en profondeur.

Intégration de Données Additionnelles (sources externes Kaggle)

Pour enrichir l'analyse stratégique, des données provenant de sources externes (principalement des ensembles de données **Kaggle**) qui offrent des informations complémentaires ont été intégrées. Ces ajouts fournissent des insights sur les comportements des utilisateurs, les tendances du secteur et les préférences des joueurs, augmentant la profondeur de l'analyse. En particulier, les éléments suivants ont été ajoutés:

Steam Games Dataset 2025 (Steam et Steam Spy)	Un large ensemble de données public (plus de 90 000 jeux publiés sur Steam) obtenu à partir des API de Steam et des données de Steam Spy. Il inclut des métriques détaillées pour les jeux PC (par exemple, nombre de joueurs, évaluations Steam, etc.), utiles pour analyser le comportement des utilisateurs sur la plateforme **Steam** et le succès des titres sur cette vitrine.
Metacritic Videogames Data (1995–2025) et Metacritic's Best Games and Reviews 2025	Deux ensembles de données contenant une liste historique de jeux vidéo (de 1995 jusqu'aux années récentes) avec les dates de sortie correspondantes et les scores des critiques ainsi que des utilisateurs. Intégrer ces données permet d'étudier l'évolution des évaluations critiques au fil du temps et de corréler le **metascore** d'un jeu avec ses performances commerciales ou de satisfaction.
Gaming Preferences and Habits Player Survey 2024	Résultats d'une enquête menée en 2024 sur **les habitudes et préférences de jeu** des joueurs. Cet ensemble de données fournit des informations démographiques et comportementales (par exemple, genre préféré,

	plateformes possédées, habitudes de dépense) directement de la voix des consommateurs. Ces insights aident à segmenter le public et à ajuster les stratégies de marketing sur les différents profils de joueurs.
Steam Hardware & Software Survey	Statistiques provenant de l'enquête matérielle/logicielle mensuelle de Steam (**avril 2022**). Valve (Steam) mène régulièrement cette enquête sur une base volontaire pour recueillir des données sur les configurations matérielles (GPU, CPU, RAM, etc.) et logicielles (systèmes d'exploitation) utilisées par les utilisateurs de Steam. En analysant ces données, nous pouvons identifier les tendances technologiques prédominantes dans la base de joueurs PC (par exemple, la diffusion d'un certain modèle de carte vidéo) et prévoir les besoins futurs (par exemple, le support à certaines technologies).

Toutes ces sources complémentaires ont été acquises et **chargées dans le modèle de données**. Actuellement, chaque ensemble de données externe est maintenu dans une structure dédiée en attendant d'être corrélé avec l'ensemble de données principal. L'intégration des données a été simplifiée grâce à une importante opération de normalisation des titres (suppression des caractères spéciaux, élimination des espaces et conversion en minuscules). Cependant, n'ayant pas réussi à unifier parfaitement tous les jeux entre les différentes sources, je ne peux pas garantir une correspondance complète entre RAWG, Steam et Metacritic.

Choix du Modèle en Étoile Modulaire

Considérant l'hétérogénéité et la provenance multiple des données, nous avons opté pour une **modélisation en étoile** (star schema) modulaire au lieu d'agréger tout dans une seule table plate. Dans un modèle en étoile, les données sont organisées avec une table centrale des faits liée à plusieurs tables dimensionnelles environnantes. Cette configuration offre plusieurs avantages dans notre contexte:
*Exemple simplifié de modèle en étoile: une table de **Faits** centrale (par exemple, ventes) liée à plusieurs tables de **Dimensions** (Produit, Client, Temps, etc.), chacune contenant des attributs descriptifs pertinents.*

Moindres pertes de données dans les unions	En unissant différents ensembles de données dans une seule super-table, on risquerait de perdre des enregistrements qui ne correspondent pas parfaitement (par exemple, jeux présents dans un ensemble de données mais pas dans un autre). Avec un modèle en étoile modulaire, chaque source alimente un fait spécifique en maintenant **toute la granularité** disponible, évitant les exclusions de données importantes. Les relations entre différents faits sont gérées via des dimensions communes ou des tables pont, plutôt qu'en écartant des informations non alignées.
Réduction des duplications et des incohérences	Insérer toutes les informations dans une seule table entraînerait des redondances (par exemple, répéter les noms de genres ou de plateformes pour chaque ligne de jeu) et des incohérences possibles dans les données mises à jour. En séparant les données en dimensions normalisées (jeux, genres, plateformes, etc.), on évite de dupliquer des informations partagées et on garantit que chaque entité soit enregistrée en un seul endroit avec une seule définition véridique. Cela améliore l'intégrité référentielle et simplifie les mises à jour futures (un genre renommé doit être modifié en un seul point, pas dans des milliers d'enregistrements).
Préservation de la granularité pour des analyses spécifiques	Grâce à différentes tables de faits, nous pouvons mener des analyses approfondies sur chaque aspect (ventes, utilisation Steam, critiques, enquêtes utilisateurs, etc.) au niveau de détail natif de cette source. En même temps, via les dimensions communes (comme le jeu ou le genre), il est possible de combiner ces perspectives en analyses croisées sans perdre de détails. En

	résumé, le modèle en étoile offre une **flexibilité analytique**, permettant à la fois des vues agrégées transversales et des explorations drill-down dans les faits individuels.
Amélioration des performances et de la clarté	Un modèle en étoile bien conçu rend les requêtes **plus simples et plus rapides** car il réduit la complexité des jointures nécessaires par rapport à un schéma hautement normalisé. Chaque requête peut se concentrer sur le fait d'intérêt en unissant quelques tables de dimensions, au lieu de scanner une énorme table plate avec de nombreux champs inutiles. Cela optimise les performances, surtout dans les outils de BI comme Tableau, et rend le modèle intuitif à comprendre même pour les utilisateurs non techniques (tables de faits centrales avec des dimensions définies autour).

En définitive, la décision d'utiliser un schéma en étoile modulaire vise à **optimiser l'analyse multi-sources:** nous maintenons les avantages de l'intégration (données liées par des clés substitutives) sans compromettre l'exhaustivité des données originales, dont la trace reste préservée dans des tables séparées.

Création de Dimensions et de Tables Pont (Bridge)

Pour implémenter le modèle en étoile, diverses **tables de dimensions** ont été définies, servant de références partagées entre les différentes sources, ainsi que des tables pont pour gérer les relations plusieurs-à-plusieurs lorsque nécessaire. Voici les principales **dimensions créées** jusqu'à présent dans le modèle:

dim_jeu	Fiche unifiée des jeux vidéo, contenant les identifiants principaux (ID interne, nom du jeu) et les attributs descriptifs de base (par exemple, date de sortie, etc.). Cette dimension est cruciale car elle relie de nombreux faits entre eux via la clé du jeu.
dim_jeu_2025	Liste complete des titres présentes dans le catalogue Steam en 2025 (issues notamment du *Steam Games Dataset 2025*). On y retrouve: ID interne, ID Steam, nom officiel, date de sortie prévue, principaux tags. Une table bridge relie **Dim_Jeu_2025** à **Dim_Jeu** pour consolider l'historique complet du catalogue.
dim_genre	Liste normalisée des genres de jeux vidéo (par exemple, Action, Aventure, RPG, Stratégie...). Chaque genre est représenté une seule fois avec un ID unique, lié aux jeux correspondants. Cela permet de filtrer ou d'agréger les métriques par genre de manière cohérente sur différents faits.
dim_plateforme	Liste des plateformes/consoles de jeu (par exemple, PC, PlayStation 5, Xbox Series, Nintendo Switch, etc.). Elle est liée aux jeux disponibles sur chaque plateforme. (Note: dans certains cas, on distingue entre les plateformes matérielles et les boutiques numériques – ici, nous avons consolidé les deux lorsque cela était pertinent, par exemple, Steam est considéré comme une plateforme PC).
dim_editeur et dim_developpeur	Contiennent respectivement la liste unique des éditeurs (publishers) et des développeurs des jeux présents au catalogue, avec des attributs comme la nation, l'année de fondation, etc. Ces dimensions permettent des analyses par maison de production (par exemple, comparer les performances de différents éditeurs).

dim_metacritic	Table dimensionnelle dérivée de l'ensemble de données Metacritic, qui liste les jeux présents sur Metacritic avec leur Metascore moyen et d'autres informations agrégées. Comme tous les titres ne sont pas parfaitement alignés avec notre liste principale (il existe des variations de noms, des éditions différentes, etc.), cette dimension est tenue séparée et liée via une table pont aux jeux correspondants. Le **Metascore** reste plus populaire en Amérique du Nord, plutôt qu'en Europe.
dim_user_survey	Dimension dérivée de l'enquête utilisateurs, qui peut contenir des catégories ou des segments de joueurs identifiés (par exemple, "casual", "hardcore", tranches d'âge, plateforme préférée, etc., selon ce qui est ressorti de l'enquête). Elle sert à croiser les résultats de l'enquête avec d'autres dimensions (par exemple, genre préféré vs genre du jeu acheté).
dim_metric et dim_category	Dimensions qui listent les configurations possibles ou catégories relevées par l'enquête hardware de Steam (par exemple, catégories de GPU, gamme de RAM installée, OS utilisé). Cela permet de lier les données d'utilisation matérielle avec d'éventuelles segmentations (pour comprendre, par exemple, le rapport entre les performances de jeu et la diffusion de certaines GPU).
dim_console	Inventaire des consoles physiques avec leur constructeur (Sony, Microsoft, Nintendo, etc.), génération, année de lancement. Utile pour analyser les ventes hardware.

En plus de ces dimensions, des tables pont ont été prévues pour représenter les **relations plusieurs-à-plusieurs** entre les entités listées ci-dessus. En particulier:

Pont Jeu-Genre	Lie les jeux aux genres (un jeu peut appartenir à plusieurs genres).
Pont Jeu-Plateforme	Lie les jeux aux plateformes sur lesquelles ils sont disponibles (un titre sort souvent sur plusieurs plateformes).
Pont Jeu-Éditeur et Jeu-Développeur	Gèrent les cas de co-développement ou de co-édition (un jeu peut avoir plus d'un développeur ou éditeur associé).
Pont Jeu-Metacritic	Relie chaque jeu de notre bibliothèque à l'entrée correspondante dans la table Metacritic (au cas où les noms ne coïncideraient pas exactement).
Pont Jeu-Eco	Associer chaque jeu PC aux thématiques « eco-friendly » pour analyser la thematiqe environnementale et filtrer les titres « verts » dans les tableaux de bord.

Étant donné la complexité d'aligner les titres entre différentes sources sans clés naturelles communes, nous avons procédé avec un mappage semi-automatique (basé sur le nom du jeu, la plateforme, etc.) pour remplir ces tables de liaison.
Les dimensions et les ponts définis assurent que le **modèle en étoile** maintienne la cohérence référentielle: chaque fait pourra se référer aux entités communes via des clés substitutives uniques. Cela évite les ambiguïtés (par exemple, le genre "Sport" a un seul ID, même s'il provient de différents

ensembles de données) et permet de naviguer à travers les relations complexes (plusieurs-à-plusieurs) de manière simplifiée en utilisant les tables pont.

Tables de Faits Finales

Dans le modèle, plusieurs **tables de faits** ont été définies, chacune spécialisée dans un certain type de données/analyse, toutes liées aux dimensions communes décrites ci-dessus. En particulier, les tables de faits finales prévues (et en cours de remplissage) sont:

Fact_Ventes	Contient les données de ventes historiques des jeux vidéo (par exemple, copies vendues, revenus) sur diverses plateformes et régions, avec une granularité temporelle (par exemple, année ou trimestre). Cette table permet d'analyser la tendance commerciale des titres au fil du temps et de la comparer avec d'autres indicateurs (par exemple, évaluations, investissement marketing, etc.).
Fact_Jeux_RAWG	Inclut les indicateurs provenant de la plateforme RAWG (notre source primaire initiale de données sur les jeux). Peut contenir des métriques comme le nombre d'utilisateurs qui ont marqué le jeu comme "joué" ou "à jouer", les évaluations de la communauté RAWG, et d'autres données collectées à partir de l'API RAWG. Étant basée sur le catalogue RAWG, elle sert de fait principal pour les informations générales sur chaque jeu.
Fact_SteamSpy	Contient les métriques d'utilisation et de popularité tirées de **Steam Spy** (et des ensembles de données brutes de Steam), par exemple le nombre de propriétaires sur Steam, le pic de joueurs simultanés, le temps moyen de jeu, etc. Ces données fournissent des insights sur le succès des titres sur Steam et l'engagement de la communauté PC.
Fact_Steam_2025	Comprend les données d'utilisation de Steam mises à jour en 2025 (incluses dans le Steam Games Dataset 2025 de Kaggle). Cette table permet des analyses détaillées et mises à jour sur l'écosystème Steam pour l'année en cours.
Fact_Metacritic	Rassemble les évaluations **Metacritic** par jeu, incluant à la fois le metascore (critique) et le score utilisateur, ainsi que les comptes de critiques. Chaque enregistrement est typiquement identifié par jeu et plateforme (étant donné que Metacritic distingue par plateforme) et année. En reliant cette table aux ventes ou aux métriques d'utilisation, on peut étudier les corrélations entre les évaluations et les performances commerciales ou entre la critique et l'appréciation du public.
Fact_Enquete_Utilisateurs	Contient les résultats élaborés de **l'enquête sur les joueurs de 2024**. Par exemple, elle pourrait avoir pour chaque répondant (anonyme) ou groupe de répondants des informations telles que la plateforme préférée, les genres préférés, la dépense moyenne, les heures de jeu hebdomadaires, etc., collectées à partir de l'enquête Kaggle. Cette table permet de croiser les profils utilisateurs avec les données de vente ou d'utilisation (par exemple, voir si les genres les plus joués coïncident avec les genres les plus vendus).
Fact_Steam_HW_Survey	Inclut les principaux résultats de **l'enquête hardware de Steam** (par exemple, la distribution en pourcentage des utilisateurs par type de

GPU, par quantité de RAM, par OS, etc. en **avril 2022**). Chaque enregistrement représente une catégorie technique dans une certaine période avec sa part d'utilisation. Cette table est isolée des autres (non liée à Dim_Jeu, mais éventuellement à la dimension technologie) et sert pour l'analyse des tendances technologiques.

Toutes les tables de faits ci-dessus sont liées aux dimensions communes par des clés étrangères (foreign key). Par exemple, Fact_Ventes et Fact_Jeux_RAWG partagent la clé du jeu avec Dim_Jeu, tout comme Fact_Metacritic est liée via la table pont Jeu-Metacritic, etc. Cela signifie que nous pouvons naviguer du fait des ventes aux détails du jeu, et de là à ses évaluations Metacritic ou à sa popularité sur Steam, le tout grâce au modèle relationnel en étoile.

État actuel

La définition du schéma (dimensions, faits, clés) a été complétée et les données nettoyées sont prêtes. À la fin du processus, le **modèle en étoile** résultant offre une vue unifiée et cohérente de toutes les informations. Il est donc possible d'importer facilement le modèle de données dans des outils de BI comme *Tableau* pour réaliser des tableaux de bord interactifs. Nous nous attendons à ce que ce modèle multidimensionnel garantisse une **analyse rapide et flexible** (grâce à des requêtes simplifiées sur le schéma en étoile: *medium.com*) et **une traçabilité complète** des données (chaque métrique est décomposable dans ses dimensions d'origine). En conclusion, l'approche adoptée permet de tirer le meilleur parti des données collectées à partir de différentes sources, en maintenant le projet évolutif et prêt pour les futures mises à jour et approfondissements.

Sondage personnel

En complément des données publiques, un **sondage personnel** a été réalisé auprès des joueurs via *Google Form*. Les réponses ont été exportées depuis *Google Sheets* et importées directement dans *Tableau*, permettant une analyse ad hoc des préférences et comportements du panel de joueurs pour affiner la segmentation client et valider certaines hypothèses de terrain.

Rapport fait le 5 juillet 2025,
par **Giulia Governatori**

Business Intelligent Analyst
(giuliagovernatori@hotmail.com)