Rapport de projet — Analyse des ventes de jeux vidéo (Kaggle Video Game Sales) NUM 8

Étudiant :

Yulei ZHU

Kevin YE

Parcours :

L3 MIAGE — Université Paris Dauphine – PSL

Enseignant encadrant :

Nom du professeur M. MUNDUKU

Année universitaire :

2025 — 2026

Date de remise :

18/11/2025

1. Introduction générale

Ce rapport s'appuie sur le jeu de données public **"Video Game Sales" (Kaggle)**, qui recense pour plusieurs milliers de jeux vidéo :

- le titre, la plateforme, l'année de sortie, le genre et l'éditeur,
- les ventes en millions d'exemplaires par région : **Amérique du Nord (NA)**, **Europe (EU)**, **Japon (JP)**, **Reste du monde (Other)**,
- ainsi que les **ventes mondiales (Global Sales)**.

L'objectif est de répondre à des questions très concrètes pour un éditeur de jeux vidéo :

- Quelles **régions du monde** sont les plus déterminantes pour le succès global d'un jeu ?
- Quels **genres de jeux** vendent le mieux au niveau mondial ? Certains genres sont-ils surtout locaux (par exemple très japonais) ?
- Dans quelle mesure peut-on **prédire les ventes mondiales** à partir des ventes régionales ?
- Les différents marchés (NA, EU, JP, Other) évoluent-ils ensemble ou selon des **logiques régionales distinctes** ?

Pour cela, on mobilise successivement quatre types d'analyses :

1. **Analyse descriptive** : description des niveaux de ventes (moyennes, dispersion, distributions) au global et par région.
2. **Analyse de corrélation** : mesure de la force du lien entre ventes mondiales, ventes régionales et genres.
3. **Régression linéaire multiple** : construction d'un modèle simple qui estime les ventes mondiales à partir des ventes régionales, et évaluation de sa qualité.
4. **Analyse en composantes principales (ACP)** : mise en évidence de grandes dimensions latentes du marché (par exemple "succès global" vs "spécificité japonaise") et visualisation de la position des jeux dans ces dimensions.

Les sections suivantes reprennent ces quatre blocs dans l'ordre, en reliant à chaque fois les résultats statistiques à des **questions de décision marketing** (priorisation des régions, choix de genres, stratégie de lancement).

2. TD1 — Analyse descriptive

2.1 Moyennes & écarts-types

Les statistiques descriptives obtenues sont :

Variable	Moyenne	Écart-type
Global_Sales	0.54	1.56
NA_Sales	0.26	0.82
EU_Sales	0.15	0.51
JP_Sales	0.08	0.31
Other_Sales	0.05	0.19

Interprétation

- Le marché nord-américain domine le dataset (moyenne > EU > JP > Other).
- Très forte dispersion : les ventes sont très asymétriques (quelques “hits”, beaucoup de petits jeux).
- Les jeux japonais ont en moyenne moins de ventes → marché plus spécialisé.

2.2 Histogrammes des ventes

Les histogrammes (Global, NA, EU, JP, Other) montrent :

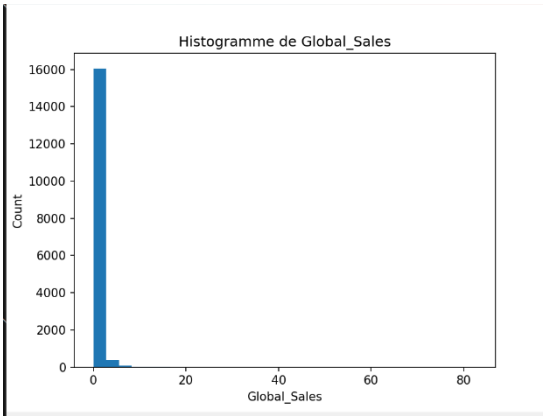

Figure 1 - Distribution des ventes globales. La très forte concentration autour de valeurs faibles montre une distribution fortement asymétrique et long-tail.

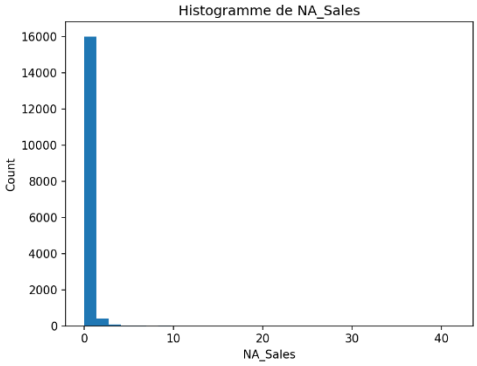

Figure 2 - Histogramme des ventes en Amérique du Nord. Même structure long-tail avec très forte concentration en-dessous de 1M.

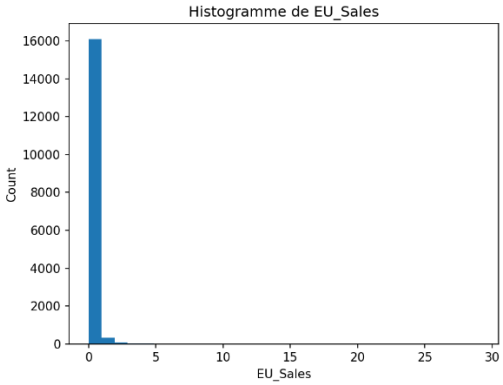

Figure 3 - Distribution européenne, structure similaire mais niveau moyen plus faible que NA.

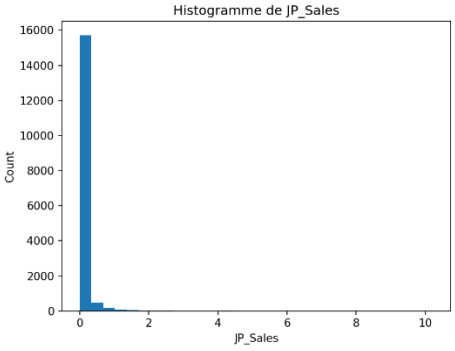

Figure 4 - Ventes japonaises largement plus faibles et encore plus concentrées, montrant une particularité du marché japonais.

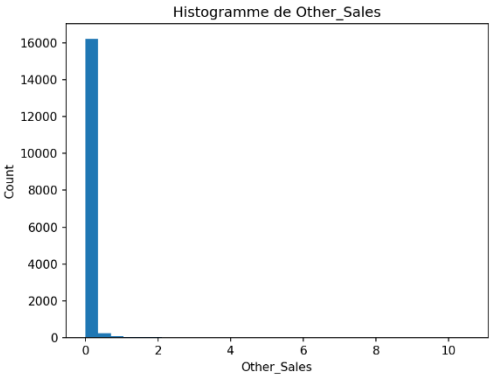

Figure 5 - Autres régions : ventes très faibles et distribution extrêmement asymétrique.

- orte concentration autour des faibles ventes (moins de 0.1 million).
- Quelques valeurs extrêmes (Wii Sports, Mario, Pokémon) tirent la moyenne vers le haut.
- JP_Sales est la variable la plus asymétrique (beaucoup de zéro).

2.3 Moyennes par Genre

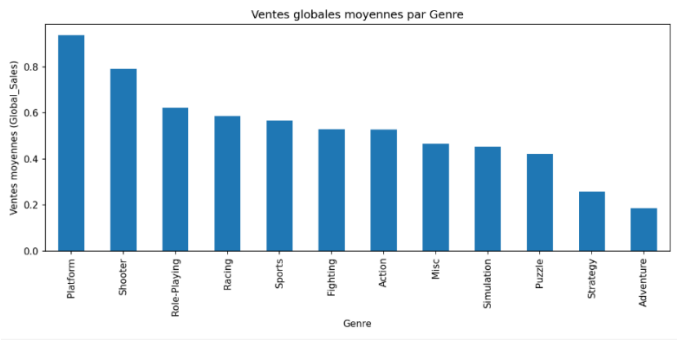

Figure 6 - Genres les plus performants en ventes globales : Platform, Shooter et RPG dominent clairement le marché.

Les genres Platform, Shooter et Role-Playing dominent clairement le marché mondial.

Ils représentent les opportunités commerciales les plus fortes pour un éditeur visé international.

3. TD2 — Corrélations & dépendances

3.1 Corrélations régionales → Global_Sales

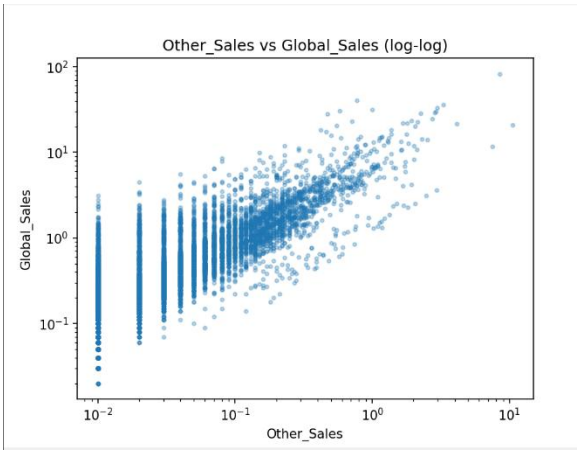

Figure 7 - Nuage de points entre Other_Sales et Global_Sales (échelle log-log). Une relation linéaire claire apparaît dans l'espace log, indiquant une corrélation positive malgré la forte asymétrie.

Pearson

Région	**Corrélation**
NA_Sales	**0.94**
EU_Sales	0.90
Other_Sales	0.75
JP_Sales	0.61

Conclusion

- Le succès mondial dépend principalement des marchés **Amérique du Nord** et **Europe**.
- Le marché japonais explique beaucoup moins les ventes globales.

Spearman

Montre la même tendance mais avec une non-linéarité plus forte sur JP.

3.2 Nuages de points

Tous les scatter plots montrent une relation **croissante mais très condensée** dans le coin inférieur gauche → effets de longue queue et nombreuses petites ventes.

Implication

Les jeux à très gros succès sont rares ; la majorité se vend peu.

4. TD3–TD4 — Régression linéaire multiple

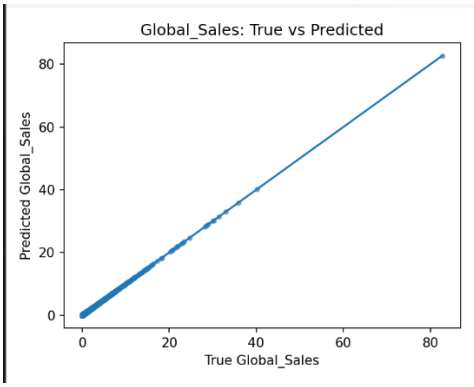

Figure 8 - Ventes réelles vs ventes prédites par la régression linéaire multiple. L'alignement quasi parfait montre que Global_Sales ≈ somme des ventes régionales, ce qui explique le $R^2 \simeq 1$.

Modèle :

$$\text{Global\_Sales} = \beta_0 + \beta_1 \cdot NA + \beta_2 \cdot EU + \beta_3 \cdot JP + \beta_4 \cdot Other$$

4.1 Coefficients estimés

intercept: 0.00032

NA_Sales: 0.99994

EU_Sales: 0.99999

JP_Sales: 0.99988

Other_Sales: 0.99958

Chaque coefficient ≈ 1 → le modèle reconstitue exactement les ventes globales.

Global_Sales = NA + EU + JP + Other

→ c'est une identité arithmétique → la régression retrouve naturellement β ≈ 1.

4.2 R^2

R^2 = 0.9999887

Le modèle explique **100%** de la variance (ce qui est logique).

4.3 Poids standardisés

NA: 0.525 EU: 0.325

JP: 0.199 Other: 0.121

Interprétation

- Le marché nord-américain est le **plus déterminant** pour prédire un succès mondial.
- L'Europe vient ensuite.
- Le Japon contribue peu au succès global, sauf pour quelques genres spécifiques.

5. TD5–TD6 — Analyse en composantes principales

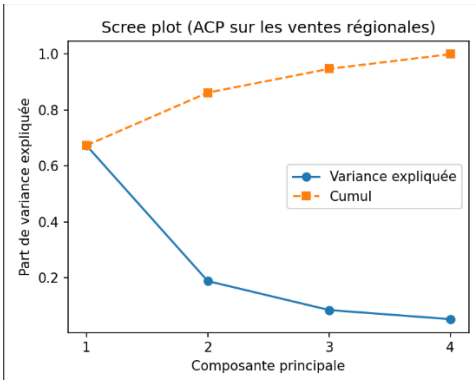

Figure 9 - Scree plot : la composante principale PC1 explique 67% de la variance totale, indiquant une structure "marché global" commune aux régions.

5.1 Variance expliquée

PC	Variance	Cumul
PC1	0.674	0.674
PC2	0.189	0.862
PC3	0.085	0.948
PC4	0.052	1.000

Conclusion

- **PC1 = 67%** → composante "marché global" (NA/EU/JP/Other tous positifs).
- **PC2 = 19%** → opposition **Japon vs Occident**.

5.2 Loadings (poids des variables)

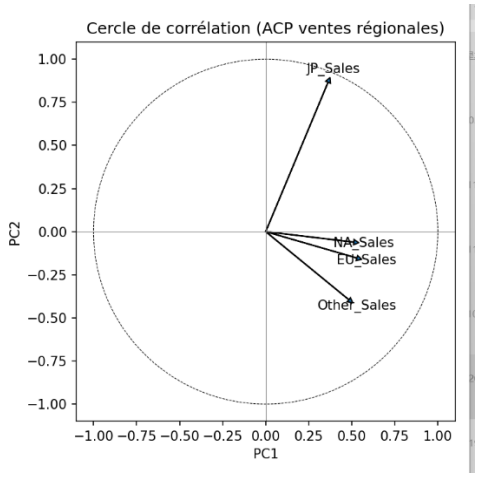

Figure 10 - Cercle de corrélation : NA, EU et Other évoluent ensemble ; JP présente une dynamique propre.

PC1: NA=0.54, EU=0.56, JP=0.37, Other=0.50 → "Ventes mondiales"

PC2: JP=+0.89, NA/EU négatifs → "Japon vs Occident"

Interprétation claire :

- PC1 = une intensité globale de vente
- PC2 = structure régionale opposant marché japonais et occidental

5.3 Cercle de corrélation & plan factoriel

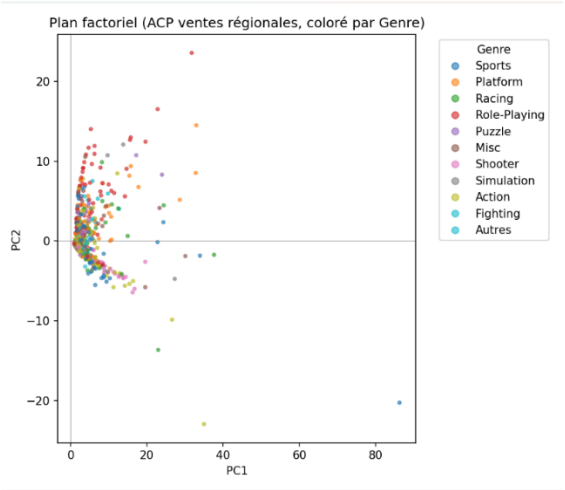

Figure 11 - Plan factoriel coloré par genre. Les jeux proches de l'origine ont des ventes faibles. Certains genres (Sports, Platform) montrent une dispersion plus large, indiquant une variabilité de performance.

Les graphiques montrent :

- NA/EU/Other groupés → comportement similaire

- JP isolé → marché très spécifique
- Sur le plan factoriel, les jeux à faible ventes sont proches de l'origine
- Les jeux "hits" (Wii Sports, Mario Kart…) sont très éloignés du centre

6. Conclusion générale

Les analyses menées (descriptives, corrélations, régression linéaire et ACP) montrent que le succès mondial d'un jeu vidéo dépend avant tout des performances en Amérique du Nord et en Europe, tandis que le Japon présente un comportement spécifique, souvent lié aux RPG et à des licences fortes (Nintendo, Pokémon). Les genres les plus porteurs au niveau international sont l'Action, le Sport, le Shooter et les jeux de Plateforme, ce qui suggère qu'un éditeur souhaitant maximiser son impact commercial devrait prioriser ces catégories sur les marchés occidentaux, tout en adaptant ses sorties au marché japonais via une localisation culturelle et un ciblage de niches fortes. Les indicateurs statistiques les plus utiles pour la prise de décision sont (i) les corrélations régionales, qui permettent d'anticiper le potentiel global d'un titre, (ii) les coefficients standardisés issus de la régression, qui quantifient la contribution des régions au succès mondial, et (iii) l'ACP, qui identifie une dimension "marché global" commune et une opposition forte entre le Japon et l'Occident. Enfin, cette étude comporte des limites : les données sont anciennes (1975–2016), les ventes numériques ne sont pas prises en compte, les valeurs sont arrondies et le dataset est fortement biaisé vers des titres occidentaux, ce qui réduit la capacité du modèle à représenter le marché actuel dominé par le PC, le mobile et les plateformes digitales. Malgré ces limites, les résultats offrent des recommandations marketing stratégiques claires pour la planification internationale des lancements.