**Rapport final**

**Projet Data Analysis - Jeux Vidéo**

*Victor Brass*

*Aline Dubucq*

*Issame Herkane*

*Florian Lys*

**Introduction au projet**

**Contexte**

En 2023, les ventes totales de jeux vidéo (toute plateforme et éditeur confondus) représentaient près de [188 milliards de $](https://www.lesechos.fr/tech-medias/hightech/lindustrie-du-jeu-video-repasse-en-marche-avant-en-2023-1973452#:~:text=Sur%20l'ensemble%20de%20l,les%20derni%C3%A8res%20projections%20de%20Newzoo.), soit un marché plus gros que celui de la musique et du cinéma réunis, notamment en France avec [5.5 milliards de €](https://www.leparisien.fr/economie/le-jeu-video-un-marche-plus-important-que-ceux-du-cinema-et-la-musique-reunis-22-10-2023-O544EA2EZ5AVTCHNL5IJOKY4FM.php) de revenus en 2023. L’industrie du jeu vidéo est donc une manne importante, riche en données à exploiter et où la concurrence est forte. Notre ambition à travers ce projet est de proposer une analyse des données du secteur, de reconnaître des corrélations et des disparités entre éditeurs, plateformes, distributeurs et de pouvoir élaborer un algorithme de machine-learning pouvant prédire le nombre de ventes d’un jeu vidéo.

**Objectifs**

Les principaux objectifs de ce projet sont:

1. Exploration, visualisation et pre-processing du jeu de données
2. Entraînement et évaluation des modèles de machine-learning pour la prédiction du nombre de ventes

Aucun des membres du groupe n’est un spécialiste de la data ni un expert pro du jeu vidéo ! Nous avons donc mis à profit toutes nos connaissances acquises au cours de la formation pour mener à bien ce projet. Nous sommes tous les 4 passionnés par le secteur que ça soit d’un point de vue économique ou par loisirs.

**Objectif 1 : Exploration, de data visualisation et de pre-processing des données**

**Compréhension et manipulation des données**

**Cadre**

L’équipe de DataScientest nous a redirigé vers un jeu de données disponible sur Kaggle, issu d’un scrap du site [vgchartz.com](http://vgchartz.com). Le site répertorie les ventes totales de jeux-vidéos pour toutes les principales plateformes, allant des années 70 à aujourd’hui. Le site fournit également des variables additionnelles pour chaque jeu, tel que l’éditeur, le développeur ou bien encore le genre pour chaque jeu listé.

Le scrap recommandé par DataScientest était toutefois quelque peu limité. Le document ayant été créé en 2016, il ne pouvait pas inclure les ventes de jeux sortis plus récemment. La variable développeur manquait également au jeu de données. Nous avons donc fait un travail de recherche supplémentaire pour trouver un scrap plus complet.

Au final, nous nous sommes tourné vers un autre jeu de données, toujours issu de vgchartz, allant jusqu’à 2019 et contenant les variables suivantes:

* Le ranking
* Le nom du jeu
* Le genre
* Sa notation ESRB
* La plateforme
* L’éditeur
* Le développeur
* Le score de la critique
* Le score des utilisateurs
* Les ventes totales
* Les ventes par marché (Amérique du Nord, PAL, Japon, Autre)
* L’année de sortie

## **Pertinence**

Pas toutes les variables listées ci-dessus sont pertinentes dans le cadre de l’objectif que nous nous sommes fixé. La notation ESRB notamment est utilisée uniquement sur le marché nord-américain et montre un volume important de valeurs manquantes. Nous avons décidé de travailler sans.

Le score de la critique et des utilisateurs a également un intérêt limité. En examinant ces variables de près nous nous sommes rendu compte que la plupart des scores affichés pour chaque jeu sont en réalité issus d’un à deux avis distincts. On peut donc difficilement partir du principe que ces données reflètent le sentiment global des joueurs et de la critique. De plus, de nombreux jeux vidéo n’ont pas récolté de score, résultant en un volume important de valeurs manquantes pour ces deux variables. Nous les avons donc également sorties du jeu de données.

Finalement,

Notre ambition étant de déterminer les ventes de jeux vidéo, l’estimated sales est devenue notre variable cible. Le publisher, le genre du jeu et la plateforme sont les 3 variables que nous percevons comme indispensable pour réaliser notre modèle.

## **Pre-processing et feature engineering**

Afin d’obtenir une prédiction pertinente et de mieux comprendre les variables importantes de notre dataset, nous avons ajouté une colonne aléatoire contenant des chiffres random. Nous avons également ajouté une colonne « Franchise » permettant de mieux appréhender le poids d’une franchise dans les prédictions de ventes. Ainsi, par exemple, nous avons plus de 130 occurrences du nom « Mario » parmi les noms de jeux contre 31 par exemple pour « Zelda ».

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Certains publishers ou développeurs n’étaient pas renseignés, nous avons fait du cas par cas pour les retrouver, idem pour l’année. Néanmoins, pour certains jeux « niche » où trop de données étaient manquantes, nous avons choisi de supprimer ces lignes.

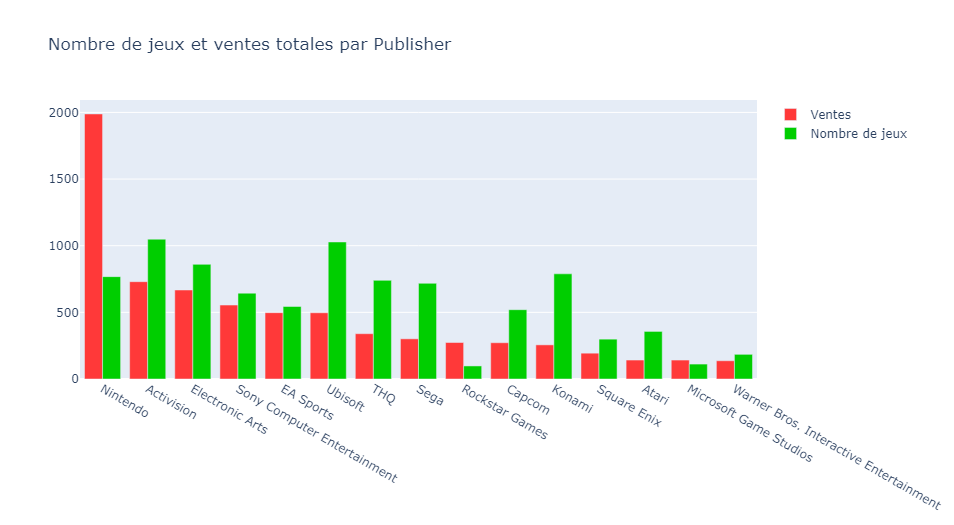
Nous avons transformé nos données avec notamment du OneHotEncoder pour les données catégorielles et de la standardisation pour les variables continues que nous avons gardées pour notre modèle.

Avez-vous dû procéder à des transformations de vos données de type normalisation/standardisation ? Si oui, pourquoi ?

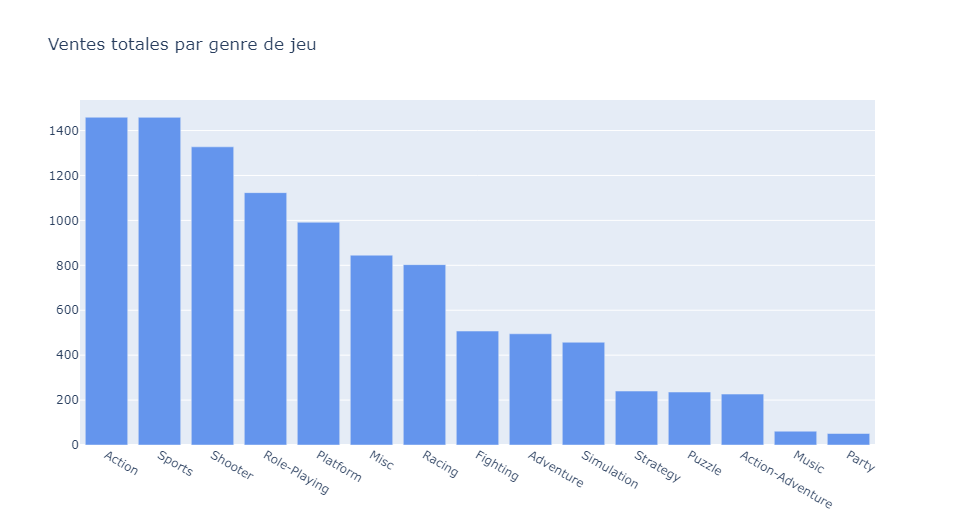
* Envisagez-vous des techniques de réduction de dimension dans la partie de modélisation ? Si oui, pourquoi ?

## **Visualisations et Statistiques**

Pour la partie visualisation et statistique, nous nous sommes penchés sur les relations entre les variables, notamment le genre du jeu, les publishers, les plateformes et les années. La visualisation était particulièrement centrée sur la compréhension entre le nombre de jeux, les ventes totales par publisher mais aussi quels genres de jeux étaient le plus vendus afin d’avoir un aperçu des forces en présence. Si par déduction et connaissance on peut dire que Nintendo est probablement le plus gros publisher au monde, il faut pouvoir le placer parmi les autres de par le nombre de ventes et le nombre de jeux sortis.

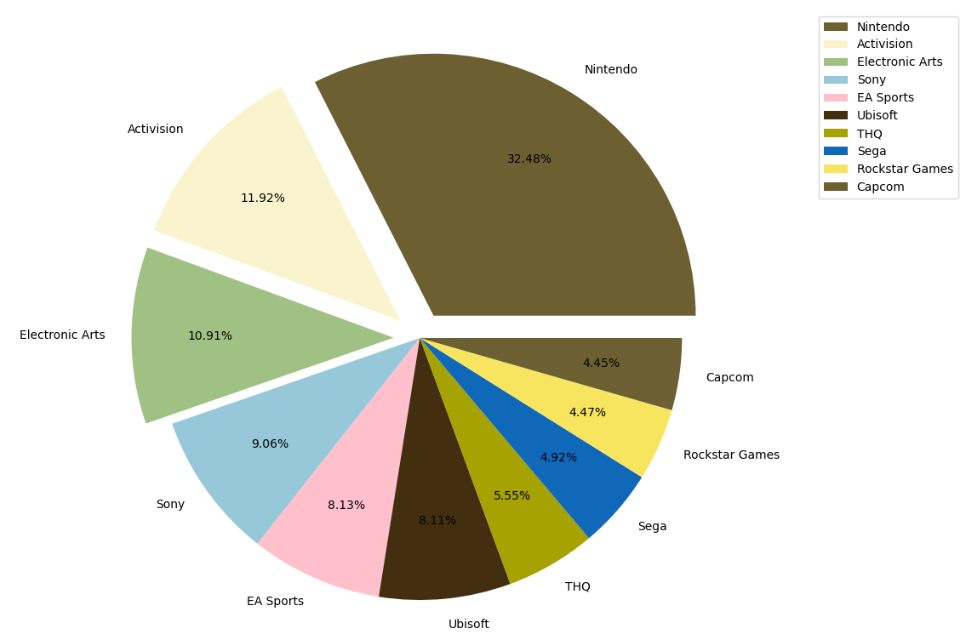


*Nintendo est effectivement le publisher qui a les ventes totales les plus importantes mais pas le nombre de jeux sortis. Pour cela, c’est Activision et Ubisoft qui ont produit le plus de jeux.*



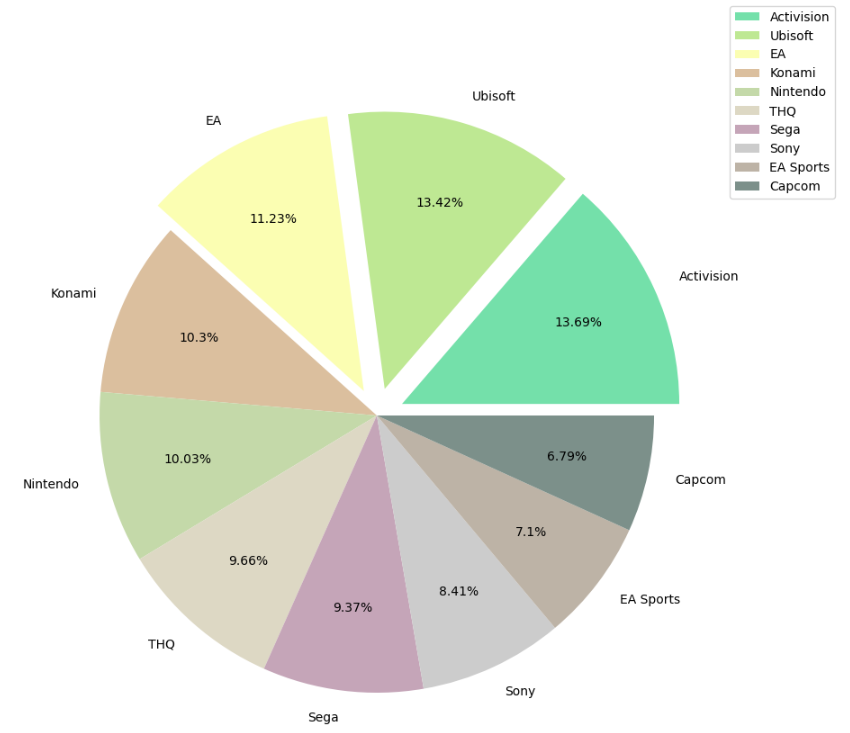
*Les jeux d’action et de sport sont les 2 genres qui ont le plus de ventes totales, suivis de près par les jeux de tirs. A l’opposé, les jeux de fêtes et de musique sont les genres qui se vendent le moins.*

Si l’on prend la répartition des ventes globales pour les 10 plus grands publishers mondiaux, on obtient ceci:



*Nintendo, EA et Activision représentent à eux trois près de 55% des ventes globales de jeux vidéo, sur les 10 premiers éditeurs de jeux en termes de ventes. Nintendo représentant à lui seul 32.5% de parts de marché.*

Si on représente le nombre de jeux sortis par publisher, on obtient:

**

*Ce qui est intéressant ici, c'est qu'en termes de nombre de jeux sortis, il y a moins de positions monopolistiques. Les 3 éditeurs ayant sorti le plus de titres représentent un peu moins de 40% de la totalité des 10 premiers éditeurs par nombre de jeux. Activision étant le plus gros avec environ 14% des jeux.*

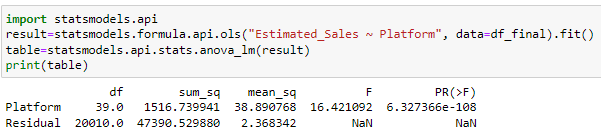
Ainsi, la comparaison avec le camembert précédent permet de montrer que ce n'est pas parce qu'un éditeur sort beaucoup de jeux qu'il a les plus grosses ventes. Nintendo a sorti 3% de titres en moins par rapport à Activision mais a réalisé 20% de ventes globales en plus.

Après ces observations, nous avons effectué des tests statistiques afin de comprendre si ces variables visualisées pouvaient avoir une influence sur notre variable cible, l’estimated sales.

Test Anova entre l'estimated\_sales et le type de plateforme

H0 : il n'y a pas d'effet significatif du type de plateforme sur les ventes

H1 : il y a un effet significatif du type de plateforme sur les ventes

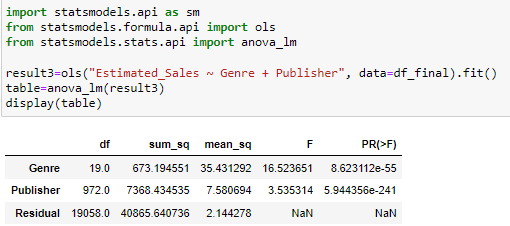


*La p-value étant inférieure à 5%, alors on rejette H0 et on conclut H1 : la plateforme a un effet significatif sur les ventes*

Test Anova entre l'estimated\_sales et le genre des jeux + le publisher

H0 : il n'y a pas d'effet significatif du genre et du publisher sur les ventes

H1 : il y a un effet significatif du genre et du publisher sur les ventes

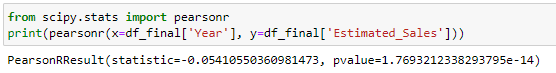


*La p-value étant inférieure à 5%, alors on rejette H0 et on conlut H1 : le genre du jeu et le publisher ont un effet significatif sur les ventes*

Voyons si l'année de sortie a un effet sur les ventes avec un test de Pearson

H0 : l'année n'a pas de corrélation sur les ventes

H1 : l'année a une corrélation sur les ventes

****

*# La p-value étant inférieure à 5%, alors on rejette H0 et on conclut H1 : l'année a un effet sur les ventes*

En conclusion, il est clair que notre modèle de machine-learning devra prendre en compte très fortement le genre, la plateforme et le publisher pour déterminer le nombre de ventes possibles.

**Objectif 2: Modélisation et évaluation des résultats**

# **Étapes de réalisation du projet**

# **Classification du problème**

Notre problème de machine-learning s’apparente à une régression : prédire un chiffre de ventes à partir de variables catégorielles.

* Quelle est la métrique de performance principale utilisée pour comparer vos modèles ? Pourquoi celle-ci ?
* Avez-vous utilisé d’autres métriques de performances qualitative ou quantitative ? Si oui, détaillez le.

## **Choix du modèle et optimisation**

Afin d’élaborer nos prédictions, nous avons essayé des algorithmes de régression linéaire, random forest et arbre de régression.

* Décrivez celui / ceux que vous avez retenu et pourquoi ?
* Avez-vous utilisé des techniques d’optimisation de paramètres de type Grid Search et Validation Croisée ?
* Avez-vous testé des modèles avancés ? Bagging, Boosting, Deep Learning… Pourquoi ?

## **Interprétation des résultats**

* Avez-vous analysé les erreurs de votre modèle ?
* Cela a-t-il contribué à son amélioration ? Si oui, décrivez.
* Avez-vous utilisé des techniques d’interprétabilité de type SHAP, LIME, Skater… (Grad-CAM pour le Deep Learning..)
* Qu’est ce qui a (ou non) engendré une amélioration significative de vos performances ?

**Conclusion et bilan**

# 

# **Difficultés rencontrées lors du projet**

* Quel a été le principal verrou scientifique rencontré lors de ce projet ?
* Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elles vous ont ralenti dans la mise en place de votre projet.
* Prévisionnel : tâches qui ont pris plus de temps que prévu, etc.
* Jeux de données : acquisition, volumétrie, traitement, agrégation, etc.
* Compétences techniques / théoriques : timing d’acquisition des compétences, compétence non proposée en formation, etc..
* Pertinence : de l’approche, du modèle, des données, etc..
* IT : puissance de stockage, puissance computationnelle, etc.
* Autres

# **Bilan**

* Détaillez quelle a été votre contribution principale dans l’atteinte des objectifs du projet.
* Avez-vous modifié le modèle depuis la dernière itération ? Si oui, détaillez.
* Présentez les résultats obtenus et comparez-les au benchmark
* Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.
* S’ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s’inscrire ? Détaillez.

**Suite du projet**

* Quelles pistes d’amélioration suggérez-vous pour augmenter les performances de votre modèle ?
* En quoi votre projet a-t-il contribué à un accroissement de la connaissance scientifique ?

## **Bibliographie**

* Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc) vous êtes vous appuyé pour réaliser votre projet ?

# **Annexes**

* Diagramme de Gantt.
* Description des fichiers de code.