Etude d’un dataset :
Video Game Sales with Ratings

28/03/2023

Théo PATRAC, RIADO Bastien, Damien TORNAMBÉ, Levi CORMIER

Table des matières

Introduction

I. Contexte

Metacritic est un site web américain lancé en 2001. Il s'occupe d'agréger les avis et les notes, provenant de la presse américaine, sur les médias en général et plus particulièrement les jeux vidéos, et cela dans le but d'aider les utilisateurs à choisir les meilleurs contenus pour dépenser leur temps et leur argent.

Avant 2020, il était possible pour les utilisateurs de voter pour un jeu avant qu'il ne soit réellement sorti, si la plateforme avait inclus sa page sur le site. Les metascore est la note de la presse donnée à un jeu, cette note pouvant aller de 0 à 100. La note provenant des utilisateurs s'échelonne quant à elle sur une échelle de 0 à 10.

Le processus d'ajout de reviews est permanent. La note est donc sujette à évoluer. Toutefois, seule la première note publiée et récupérée d'une presse est ajoutée. Elle ne sera pas mise à jour si cette dernière la change après.

II. But du dataset

Rush Kirubi, l'auteur de ce dataset, a travaillé en tant que Data Scientist dans trois services de l'entreprise Verisk, située dans le Massachusetts aux USA.
Il a écrit qu'il aimerait voir appliquées à son jeu de données des techniques de machine learning ou des méthodes de visualisation continue. Également, il explique qu'il a été motivé par un autre dataset scrappé avec Python par Gregory Smith, intitulé "Video Game Sales", basé sur le site VGChartz qui est un site de suivi des ventes de jeux vidéo. Cet ensemble de données contient une liste de jeux vidéo dont les ventes sont supérieures à 100 000 exemplaires.
Rush Kirubi a fusionné le travail de Gregory Smith avec une extension, générée elle aussi par scrapping Python, en ajoutant le nombre et la moyenne des notations des utilisateurs et celles de Metacritic. Cependant, il y a des observations manquantes car Metacritic ne couvre pas toutes les plateformes effectivement couvertes par VGChartz.

Le dataset contient des informations sur les ventes de jeux vidéo et leurs notes entre 1985 et 2016. Il met à disposition plus de 16500 données dont environ 6900 sont des données complètes incluant les ventes et les notes.

Le dataset contient donc 3 catégories de données :

- Les informations sur le jeu : nom, année de sortie, éditeur, genre et classification ESRB ;
- Le nombre d'unités vendues par région : Japon, Amérique du Nord, Europe, mondial et autre ;
- Les critiques : celles des utilisateurs et celles de la presse, déterminée en faisant la moyenne des critiques de différents professionnels, ainsi que leur décompte.

III. Description de la question

L'arrivée d'internet et notamment des réseaux sociaux amena avec elle l'ère de l'information en temps réel accessible à tous. Cela imposa une énorme pression qui entraîna une crise structurelle de la presse. En effet, Il est maintenant très facile pour les utilisateurs de partager leurs avis et de les consulter, court-circuitant ainsi ceux de personnes dont c'est le métier.
Il est généralement admis dans la communauté des joueurs, qu'afin de survivre, les acteurs de la presse vidéoludique sur-noteraient les jeux des gros éditeurs en échange de pot-de-vin.

Y a-t-il eu réellement un changement des différences entre les notes de la presse et des utilisateurs au cours du temps, sur les gros éditeurs en particulier ?

Méthodologie

I. Préparation des données

Il y a plusieurs points problématiques dans ce jeu de données brutes qui nous obligent à effectuer un travail de préparation en amont avant de pouvoir les utiliser.

Tout d'abord, nous avons changé le type de 4 des 16 variables qui composent ce dataset. Notamment les scores des utilisateurs qui étaient enregistrés en tant que caractère et non réel. Nous avons aussi mis les scores sur la même échelle, donc de 0 à 100.

Ensuite, nous voulions que nos données soient au format tidy 'long'. On a donc ajouté une colonne Origin_of_the_Score qui discrimine les colonnes Score et Count par leur provenance ('User' ou 'Critic'). De plus, nous avons éliminé toutes les lignes ne comportant pas à la fois une note utilisateur et critique ainsi qu'une année de sortie ce qui nous a conduit à écarter les jeux sortis avant 1996.

À la fin de notre processus de mise en forme, nous sommes passés d'un jeu de données contenant 16719 jeux distincts à seulement 6890 utilisables et au format tidy 'long'.

II. Étude du dataset

Le point de départ de notre étude vise à observer nos données d'un point de vue global.
Nous allons alors nous intéresser aux différences entre les notes moyennes par an données par la presse par rapport à celles des utilisateurs entre 1996 et 2016.

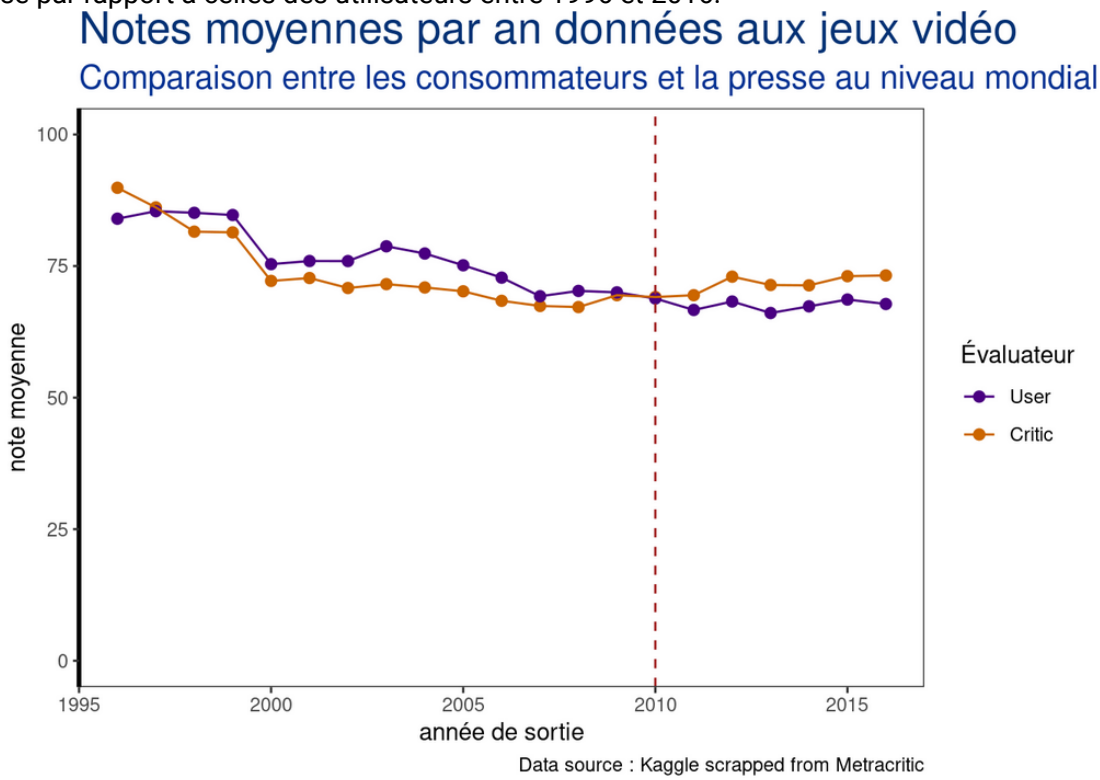

Figure 1 : Graphique des notes moyennes par an données au jeux vidéos au niveau mondial

Cette visualisation des données montre les notes moyennes sur tous les jeux sans distinction, données par les utilisateurs entre 1996 et 2016.

Ce qui nous apparaît clair en premier lieu, c'est qu'il semble y avoir une corrélation entre les moyennes des utilisateurs et de la presse. Il y a une cassure de la continuité entre 1999 et 2000.

Il y a quelques fluctuations mais elles ne sont pas très significatives : avant 1999 les notes moyennes des jeux se trouvent autour de 84/100, après cela elles sont plutôt autour de 71/100. Ce n'est cependant pas à exclure, car cela pourrait traduire une baisse dans la qualité des jeux sortis après 1999, ou bien des votes par effet de nostalgie, par exemple.

Nous pouvons constater par 2 fois, en 1997 et en 2010, le croisement des valeurs avant l'inversement de la tendance.

Nous avons alors décidé d'observer plus précisément cette tendance en comparant les différences de notations pour les 6 plus gros éditeurs, choisis en fonction de leurs ventes et nombre de jeux.

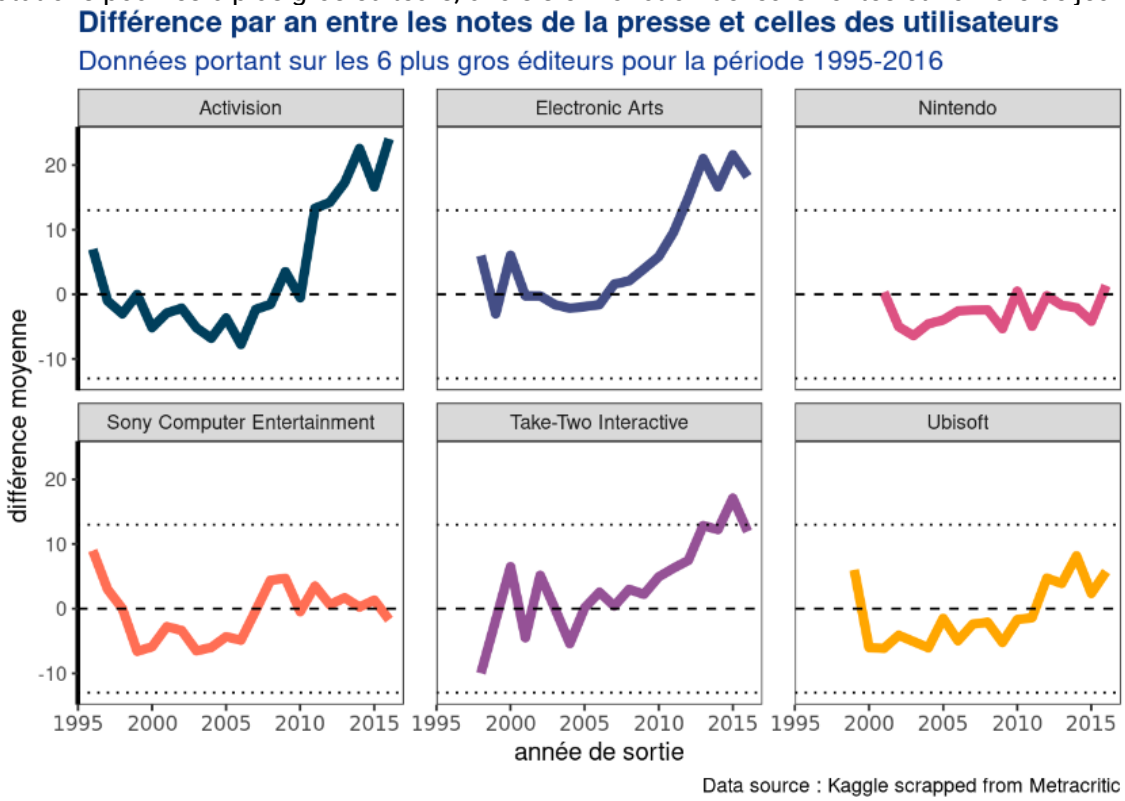

Figure 2 : Graphique des différences de notations sur les 6 plus gros éditeurs de jeux vidéo

Pour mieux appréhender le graphique, il faut savoir que :

- La courbe colorée en fonction de l'éditeur représente la différence moyenne de notation entre la presse et les utilisateurs sur tous ses jeux, au fil des années ;
- Quand la courbe est située en dessous de 0 cela signifie que les utilisateurs ont mieux noté le jeu que la presse ;
- Quand la courbe est située au dessus de 0, cela signifie que la presse a mieux noté le jeu ;
- Les lignes en pointillés espacés reflètent l'écart-type entre les notes de tous les jeux.

Avant 1997, la presse notait en moyenne légèrement mieux, en revanche, entre 1998 et 2008, soit une période de 10 ans, c'était plutôt les utilisateurs. Après 2010 la presse est de nouveau repassée devant.

Un indicateur intéressant qui ne saute pas forcément aux yeux est que l'écart-type entre les différences des notes de nos 2 sources est de 13. Ce qui, au regard d'un score, reste une différence non négligeable. En effet, sur une échelle de 100, cela représente une variation de plus de 10%, qui fait facilement passer un jeu d'une catégorie à une autre.

Les courbes d'Activision et d'Electronic Arts suivent un pattern plus particulier que les autres courbes étudiées précédemment, car même si elles fluctuent, elles sortent clairement de l'écart-type moyen des différences globales après 2010, la différence de notation entre les utilisateurs et la presse est plus grande que la normale, avoisinant une différence moyenne de 20. La presse noterait alors nettement mieux les jeux que les utilisateurs. On a vu que les éditeurs Activision et Electronic Arts avaient un différentiel entre les notes nettement supérieur aux autres après 2012.

Pour conclure nos analyses, nous allons observer maintenant l'évolution des notes données aux jeux de la licence Call of Duty édité par Activision. Sur la période de 2004 à 2016 ils ont sorti un épisode par an, sauf en 2015. On veut voir l'évolution de leurs notations afin de voir si aujourd'hui, la licence phare de ce studio a eu un impact sur les divergences observées post 2010.

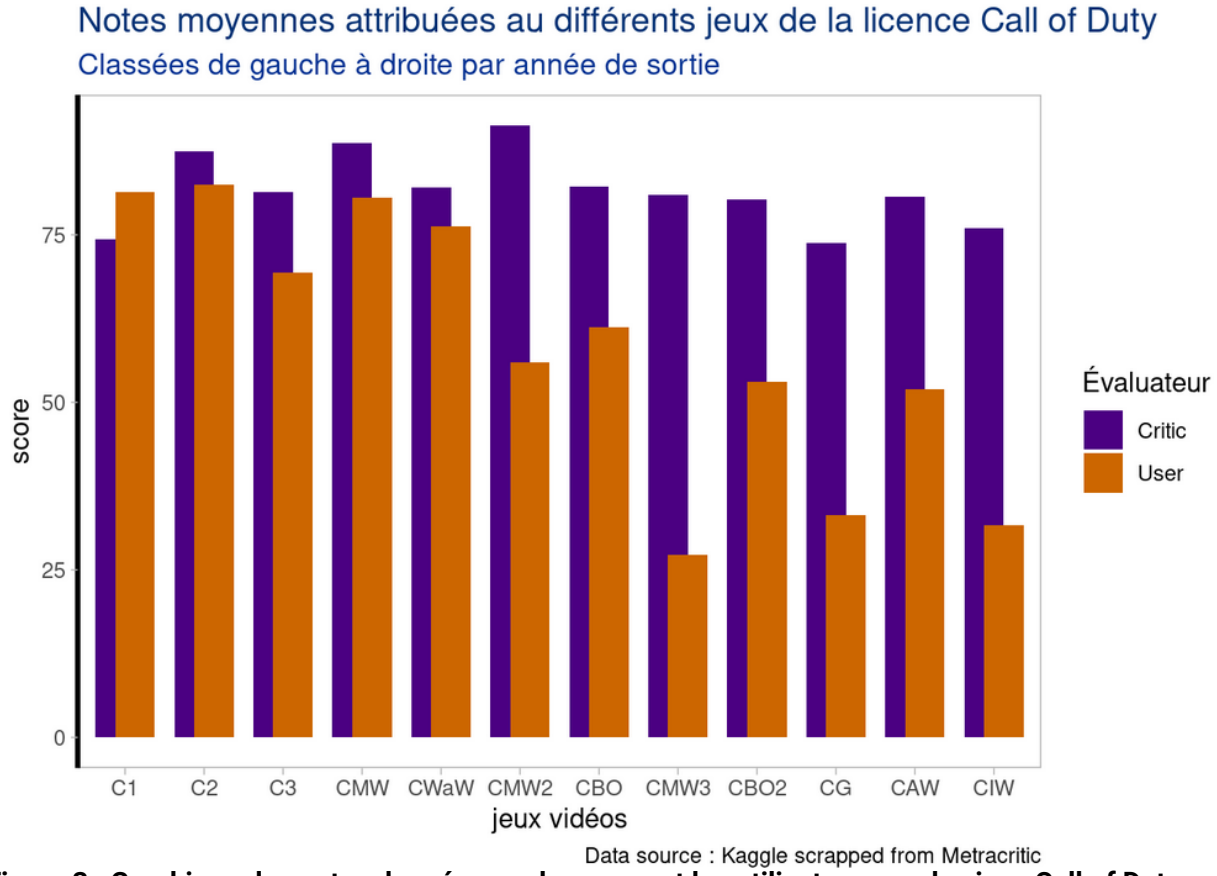

Figure 3 : Graphique des notes données par la presse et les utilisateurs sur les jeux Call of Duty

On voit bien que le public a ici tranché, c'est une hécatombe. On passe d'une note moyenne de 59 pour les utilisateurs à 82 pour la presse. On observe un pic au niveau du jeu Call of Duty: Modern Warfare 3 (CMW3) avec une différence entre les notes de 54 points.

Ce qui est très intéressant ici, c'est l'énorme constance au cours du temps des notes attribuées par la presse sur les jeux de cette licence. En effet l'écart-type est ici d'environ 5,5 contre plus de 20 pour

les utilisateurs. On ne peut que s'étonner devant une telle constance sur 12 jeux s'échelonnant sur un intervalle de 13 ans, quand on les compare aux autres.

Conclusion

Finalement, il est très difficile de se prononcer. L'observation des différences sur les 6 plus gros éditeurs ne permet pas de répondre précisément à notre question. En effet, on observe de bien trop grosses fluctuations, sur un intervalle bien trop grand pour que cela soit pertinent.

A cela s'ajoute nombre de biais que nous ne pouvons quantifier à savoir:

- L'impact de la crise des subprimes en 2008 qui a fortement impacté l'économie ;
- La montée en flèche des réseaux sociaux et des moyens de communications qui a et qui continue à obliger ce milieu à trouver des moyens de financement ;
- Les utilisateurs sanctionnant des jeux en leur donnant des notes très faibles (souvent 0) ; sans y avoir joué pour protester contre des pratiques de monétisation à l'intérieur des jeux ou des affaires juridiques entachant l'image du studio ;
- L'effet de nostalgie qui pousse des gens à sur-noter les très vieux jeux.

Pour un point de vue personnel, il nous paraît évident que la presse tire vers le haut les notes des gros éditeurs. Aujourd'hui, les sites doivent même faire appel à la générosité de leurs utilisateurs pour pouvoir survivre. Il serait donc illusoire de penser que face aux difficultés et à cette concurrence farouche, les médias jouent franc-jeu.

Toutefois, quand on voit que la presse à gratifié un jeu comme Call of Duty: Modern Warfare 2 d'un 91/100, simple copier-coller de l'épisode précédent, et que le chef d'œuvre qu'est Call of Duty: Black Ops n'a récolté que d'un 61/100, on est en droit de se demander si finalement, les notes attribuées ne relève pas d'un lancé de dés.

Références

Notre étude se base sur un dataset scrappé par Rush Kirubi :

https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

Qui lui même a repris le dataset de Gregory Smith en le mettant à jour:

https://www.kaggle.com/datasets/gregorut/videogamesales