Analyse des données de Steam

J. CORITON 03/2022

Table des matières

1. Introduction

L'objectif du projet est de réaliser des recommandations sur les caractéristiques et le prix de vente de jeu qui sont ou seront développés par Able Bear Studio, en se basant sur le contenu de la base de données mise à disposition par Steam, qui est une plateforme de distribution de contenus en ligne. Cette base de données comprend plus de 50 000 applications ainsi que de nombreuses informations sur celle-ci.

L'analyse se concentrera sur différents axes :

- L'analyse des Tags, pour déterminer les Tags les plus populaires et à mettre en avant lors de l'annonce de la date de sortie du jeu, ainsi que les caractéristiques du jeu à développer qui pourrait augmenter l'attrait du jeu. Il faudra porter une attention particulière sur ces Tags, car ceux-ci sont choisis uniquement par les joueurs.
- L'analyse des Genres, pour connaitre les types de jeux qui fonctionnent le mieux sur Steam, et proposer des types de jeux attrayants pour le développement de futurs jeux.
- L'analyse des prix, pour choisir un prix attractif pour un maximum de joueurs ainsi qu'avoir un équilibre commercial.
- L'analyse du succès des jeux en fonction de leur dates de sorties pour obtenir une date optimale.

Ainsi je vais orienter mes recherches principalement en fonction du nombre de personnes possédant chacun des jeux et qu'elles sont les variables qui influencent cette donnée. Le jeu actuellement développé par le studio, sera un Jeu de rôle (RPG) médiéval narratif, et en fonction du succès de celui-ci, le prochain pourrait être la suite ou un jeu du même genre. Mes analyses iront donc dans ce sens.

2. Récupération des données

La Base de données utilisées pour ce projet est fournie par l'API Steam, qui contient des informations sur plus de 50 000 applications. Pour récupérer ces différentes données, j'ai mis en place le fichier : **steamExctractor.py**, qui permet de faire des requêtes à deux API différentes, celle de Steam et celle de SteamSpy.

2.1 Récupération des données Steam

Une partie du programme récupère les données via l'API Steam. Elle récupère la liste de toutes les applications disponibles sur la plateforme, ainsi que de nombreuses variables qui lui sont associées pour les stocker dans le fichier : **steam_app_data.csv**. La base de données contient des informations sur 51 000 applications qui sont décrites avec 39 variables :

'type', 'name', 'steam_appid', 'required_age', 'is_free', 'controller_support', 'dlc', 'detailed_description', 'about_the_game', 'short_description', 'fullgame', 'supported_languages', 'header_image', 'website', 'pc_requirements', 'mac_requirements', 'linux_requirements',

'legal_notice', 'drm_notice', 'ext_user_account_notice', 'developers', 'publishers', 'demos', 'price_overview', 'packages', 'package_groups', 'platforms', 'metacritic', 'reviews', 'categories', 'genres', 'screenshots', 'movies', 'recommendations', 'achievements', 'release_date', 'support_info', 'background', 'content_descriptors'.

Les données sont plutôt complètes, avec moins de 3% des données manquantes sur les variables qui nous intéressent. Parmi celles-ci, la moins remplie est **'categories'**, correspondant aux tags des jeux, avec 2.15 % d'informations manquantes.

2.2 Récupération des données de Steamspy

La seconde partie du programme récupère les données à partir de l'API de Steamspy. Celle-ci permet d'éviter de réaliser des requêtes sur l'ensemble des utilisateurs publiques de la plateforme pour connaitre leurs possessions, leurs différents avis, ainsi que leurs habitudes de jeu. Cette base de données synthétise de nombreuses informations pertinentes pour notre analyse pour chaque application. Comme pour la première partie, les données sont stockées dans le fichier : **steamspy_data.csv**,. La base de données contient des informations sur 50 418 applications, qui sont décrites avec 20 variables (les données manquantes concernent des applications non disponibles actuellement):

'appid', 'name', 'developer', 'publisher', 'score_rank', 'positive', 'negative', 'userscore', 'owners', 'average_forever', 'average_2weeks', 'median_forever', 'median_2weeks', 'price', 'initialprice', 'discount', 'languages', 'genre', 'ccu', 'tags'.

Les données sont quasiment complètes avec 0.54 % de données manquantes, si l'on exclue le **'score_rank'**, qui correspond au score donné au jeu sur le site de metacritique.com. A noter que le nombre de possesseurs de jeux est approximatif en raison de l'existence de profil d'utilisateur privé.

3. Préparation des données

Afin d'analyser plus facilement les données, celles-ci doivent être transformées pour être exploitables. Certaines données (comme les dates de sorties) n'ont pas toutes de formats standardisés, il est donc important de récupérer chaque variable de manière pertinente.

3.1 Créer une colonne par Genre / Tag / Langage

J'ai décidé de créer une colonne par Genre, Tag et langage différents avec une entrée booléenne pour pouvoir observer les occurrences. Les genres ont des variables commençant par **« gis_ »,** les tags commencent par **« tis_ »** et les langues par **« lis_ »**

```
def Bool_tag(tag):
# """function wich search each game with the tag asked

#     Args:
#         param1: tag, use string value

#     Returns:
#         transform tag_steamspy, add columns named "tis_tagname",
#         with True for success about research tag, False otherwise.
# """
    value = tag_steamspy["tags"].str.contains(tag)
    tag_steamspy["tis_" + str(tag)] = value
for j in range(len(count_tag[0][:100])):
    Bool_tag(count_tag[0][j])
```

Ci-contre, la partie du code concernant les tags. Il a été décidé de se limiter au 100 Tags les plus utilisés, qui représentent 76% des votes totaux effectués par les joueurs.

3.2 Créer un indice de popularité

L'indice de popularité sera basé sur le pourcentage de recommandations positives pour chaque jeu ainsi que sur le pourcentage du nombre de possesseurs de jeux maximum pour les jeux indépendants.

3.3 Récupérer les dates de sortis

J'ai récupéré un maximum de formats de dates **(jj/mm/yyyy, jj/mm/yy, jj mon. year, jj month year, etc…),** jusqu'à avoir seulement 5,2% de perte.

4. Nettoyage des données

Etape	Description	Nb d'entrées	Nb de variables	Nb données manquantes
0	Fusion des deux datasets	56530	254	118669
1	Nettoyage des entrées dupliquées	48477	254	49615
2	Nettoyage par mots clés	48208	168	7529
3	Nettoyage des erreurs	48101	168	2625
4	Nettoyage par date de sortie	45613	168	137

Figure 1 – Evolution des données du dataset après les différentes étapes de nettoyage.

4.1 Suppression des données en doubles

Les données sont nettoyées pour se concentrer sur celles qui nous intéressent. Les variables qui ne sont pas pertinentes sont éliminées. Les deux Dataframes provenant des deux API sont fusionnées. Puis je considère la présence d'applications qui apparaissent plus d'une fois dans mes données. Je supprime les doublons en ne conservant que la première entrée (qui correspond aux données de Steamspy, qui sont plus pertinentes).

4.2 Suppression des données par mots clés

Avec la mise en forme du Dataframe des genres, il est facile de supprimer les applications qui ne sont pas des jeux, en supprimant les données liées à d'autres genres.

4.3 Suppression des applications avec des données erronées

Certaines applications ont été récupérées en plus des jeux. J'ai supprimé les 107 échantillons avec le prix le plus élevé. Cela a permis d'épurer la plupart des applications tiers utilisées pour faire de la modélisation 3D ou de la création de contenu (musique, jeu vidéos, vidéos …), qui était classée dans le genre « Action ».

4.4 Suppression des échantillons sans date de sortie

Les jeux qui n'avaient pas de date de sortie ont été eux aussi supprimés, car ils rassemblent deux grandes catégories de jeux : ceux qui ont juste été annoncés et qui n'ont pas de date de sortie officielle, et ceux qui sont anciens (antérieur à 2010) et dont le format ne respectait pas de forme standard.

Pour finir nous obtenons donc un Dataframe avec 45 613 échantillons de jeux, sur les 51 000 initiaux. Le Dataframe est presque complet avec seulement 0.3% de données manquantes.

5. Création de catégorie de jeu

Après l'étape de nettoyage, il s'agira de définir le degré de similarité existant entre deux entrées et pour cela, j'opte pour un "calcul de distance" entre les tags sélectionnés par les utilisateurs pour les différents jeux de la base de données. Je sélectionne donc dans le Dataframe les variables que je juge être pertinentes pour ce calcul. Les variables que j'ai retenues sont donc les 100 tags les plus utilisés par les joueurs, ce qui correspond à 76% des Tags utilisés.

Pour mieux visualiser les différents groupes, j'ai réalisé un dendrogramme qui utilise la même distance pour séparer les groupes.

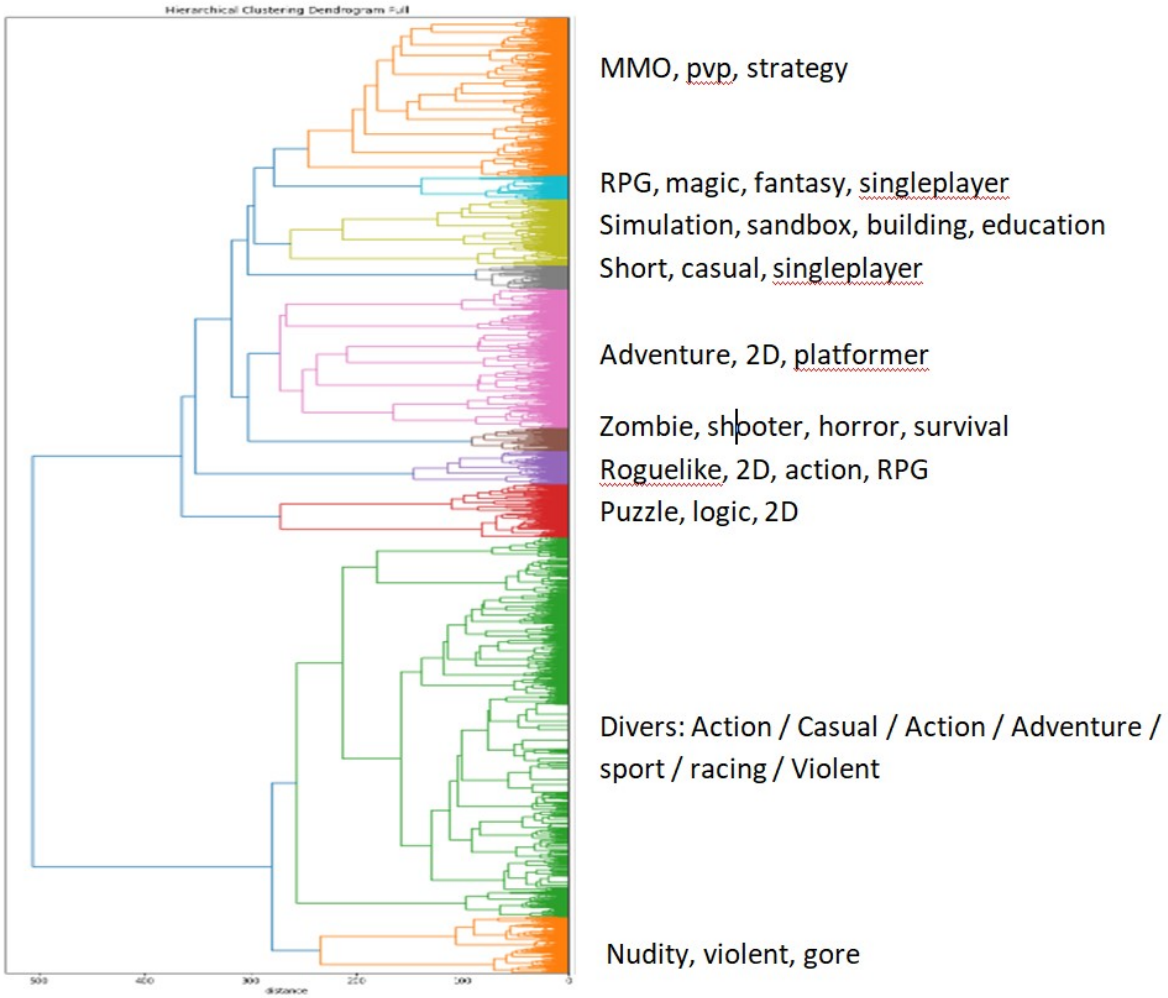


Figure 2 – Création de 10 groupes de jeux à partir des tags

6. Exploration des données

Enfin, après avoir bien préparé le Dataset, nous arrivons avec des données presque complètes. Dans les valeurs manquantes, il y a juste 29 jours de sorties qui restent indisponibles et 108 avis des possesseurs du jeu (valeur manquante car il n'y a aucun avis libre d'accès pour le calculer). Il reste 45 613 échantillons à analyser, ce qui représente 89.4 % du total d'applications disponibles sur Steam.

Après l'étape de nettoyage, j'ai visualisé le contenu des variables les plus pertinentes pour la suite de l'analyse. A titre illustratif, j'ai juste reporté 5 figures qui indiquent le nombre d'occurrences des différents mots-clés, ainsi que l'évolution du nombre de sorties des jeux par an afin d'avoir un aperçu des contenus les plus populaires.

6.1 Traduction et définition des différentes variables

Certains termes des variables pouvant être assez obscures, je vais donc définir ici ceux étant les moins clairs ou utilisant un langage spécifique ou des sigles pour faciliter la compréhension des analyses suivantes.

Owners_mini et owners_maxi : qui correspondent respectivement à l'estimation minimale et maximale du nombre des possesseurs pour chaque jeu. Il n'y a pas de chiffre précis en raison, ici aussi, de l'existence de profil privé.

Owners_positif : représente le pourcentage d'avis positifs laissés par les joueurs.

Price / initialprice / discount : représente le prix actuel, celui à la sortie et le plus bas qui a été disponible.

Day / month / year : correspond au jour, mois et année de sorties du jeu.

Indie : permet de savoir si le jeu a été réalisé par des studios indépendants (c'est-à-dire ne faisant pas parti des grosses entreprises de production vidéo ludique).

Casual : représente les jeux destinés au grand public de joueurs occasionnels. Ce type de jeu est caractérisé par l'accessibilité et la facilité de prise en main pour attirer le plus de joueurs possible.

RPG : Abréviation de « role playing game », représente les jeux vidéo de rôle. Ce genre de jeu a les caractéristiques de jouer un ou plusieurs personnages capables d'évoluer au fil du temps tout en avançant dans une histoire plus ou moins développée.

Sandbox : soit jeu « bac à sable » en français, correspond au jeu généralement non linéaire, sans objectif prédéterminé et faisant appel à la curiosité et la créativité des joueurs.

Rogue like : c'est un genre de jeu vidéo caractérisé par l'exploration de donjons créés de manière procédurale, avec des morts du personnage joueur presque permanentes, et une carte constituée d'un ensemble de tuile.

Co-op : Correspond à l'abréviation de coopération, ce qui implique de participer avec d'autres joueurs à une œuvre commune.

Pvp : Sigle signifiant « player versus player » soit « joueur contre joueur ». C'est un mode de jeu permettant aux joueurs de s'affronter les uns les autres.

Massively multiplayer : traduit en français par massivement multiplayers en ligne, et sera appelé MMO dans la suite de ce document. C'est un genre faisant participer un très grand nombre de joueurs simultanément à travers internet.

Hand-drawn : C'est une caractéristique de jeu mettant en avant les dessins fait à la main pour une partie ou la totalité du jeu. Ce type de jeu est en opposition avec ceux réalisés en 3D et cherchant à être toujours plus proche du réalisme.

Singleplayers : Jeu pouvant être joué seul.

6.2 Analyse exploratoire des Genres

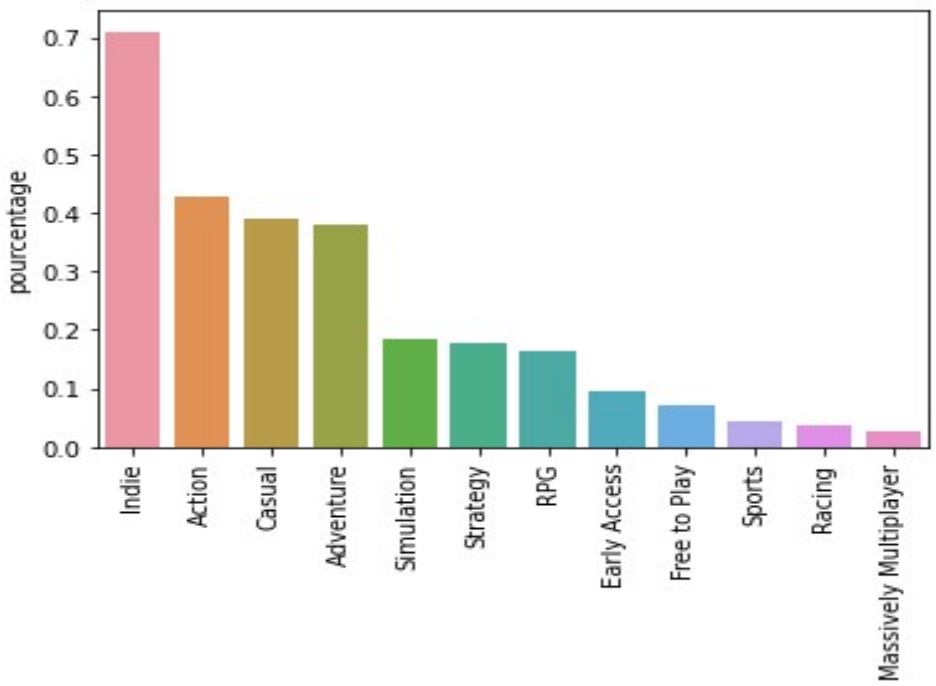


Figure 3 - Répartition du nombre de Genre par jeux

On peut voir que plus de 70% des jeux Steam proviennent de studios indépendants et que les parts de marché des jeux RPG représente moins de 20% du total des jeux.

6.3 Analyse exploratoire des Tags

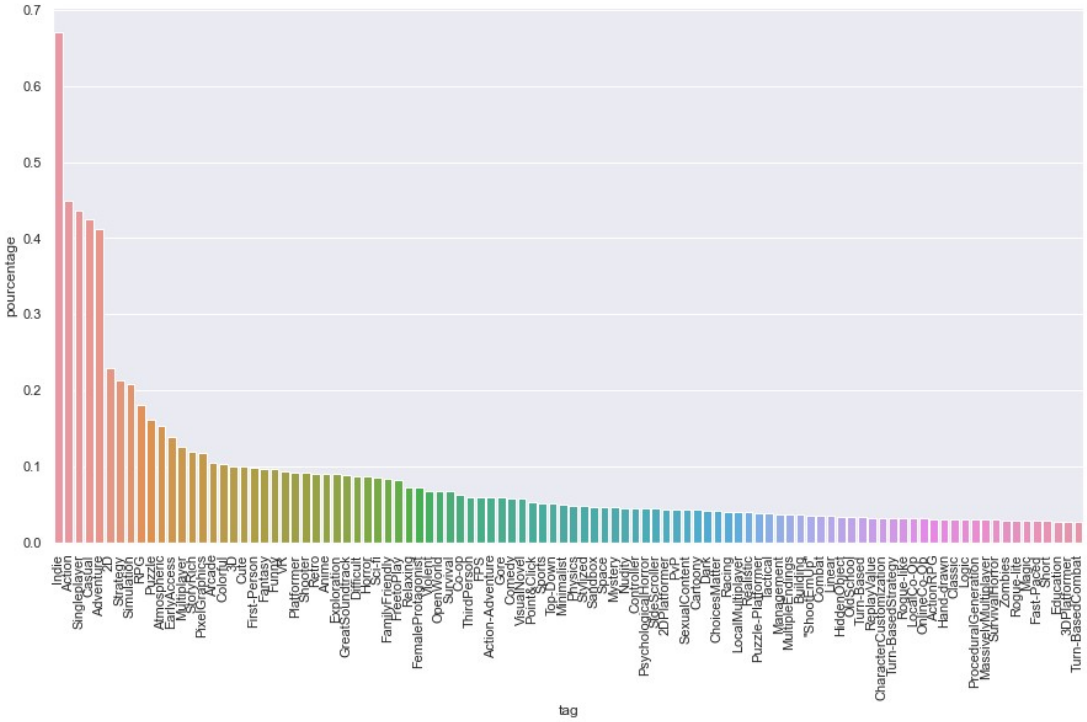


Figure 4 - Répartition des 100 Tag les plus populaires par jeux

En comparant cette figure avec la précédente, on peut voir que la répartition des tags est très corrélée aux genres. C'est pour cela que pour la suite de l'analyse, nous nous concentrerons plus sur les Tags car ils ont des données beaucoup plus variées.

6.4 Analyse exploratoire des langues

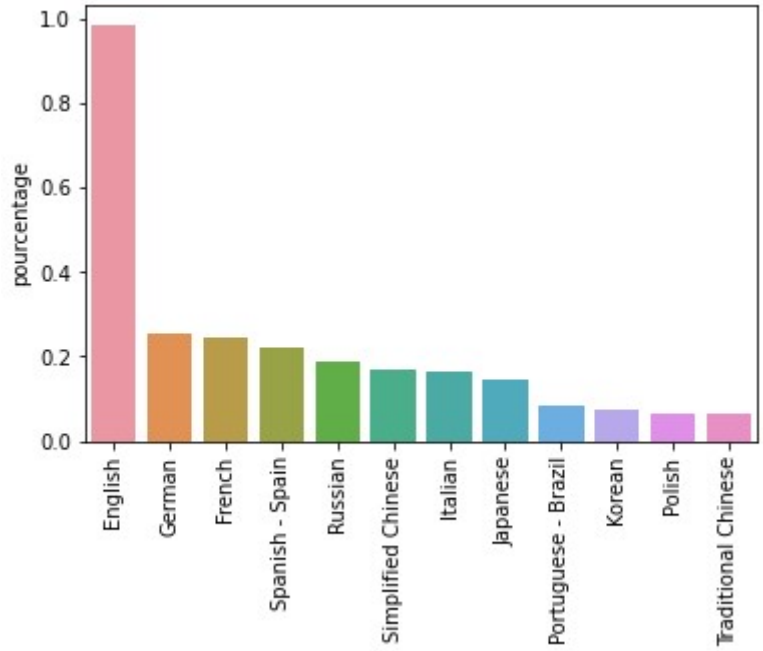


Figure 5 - Répartition des langues disponible par jeux

On peut observer que presque la totalité des jeux sont traduits en anglais, ce qui peut facilement s'expliquer par la plus grande part de marché que représente cette langue.

6.5 Analyse exploratoire des sorties et des prix des jeux RPG indépendants, traduit en français

Pour la suite de l'analyse, nous nous focaliserons sur les RPG français réalisés par des studios indépendants, car c'est ce type de jeu que le studio Able Bear a pour objectif de réaliser.

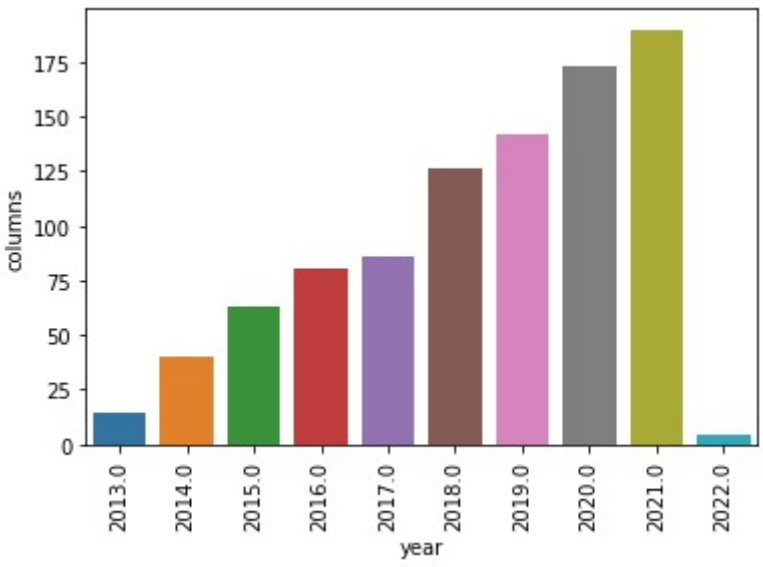


Figure 6 - Evolution du nombre de sorties de jeux en langue français, RPG et indépendants

La politique de Steam a changé en 2013, en remplaçant l'analyse des respects des critères Steam réalisée par du personnel par un algorithme. Ceci permet d'accepter beaucoup plus de jeux chaque année. On voit que depuis 2013 le nombre de jeux RPG possède une forte croissance.

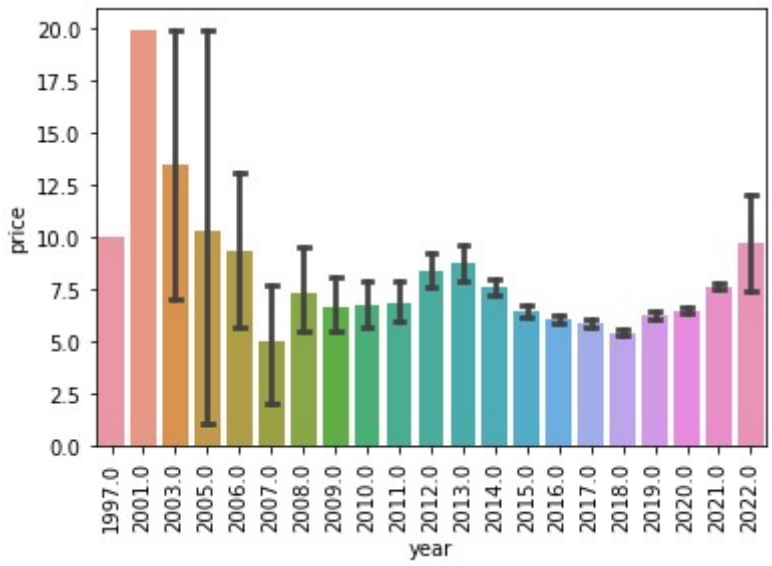


Figure 7 - Evolution du prix moyen en € de sorties de jeux RPG et indépendants

Avec cette figure, on peut voir que l'ouverture du marché des jeux vidéos avec le distributeur Steam en 2013 a fait chuter le prix moyen avec l'augmentation de la concurrence jusqu'en 2018. Mais il y a une tendance à l'augmentation des prix après cette période.

7. Analyses des Corrélations pour les RPG français indépendants

Pour réaliser les analyses de corrélations, je vais uniquement sélectionner l'ensemble des jeux possédant les tags/genres RPG et indépendants, ainsi que ceux disponibles en français, pour être le plus représentatif possible des styles de jeux que pourrait produire Able Bear Studio.

7.1 Corrélation entre les genres et le nombre de possesseurs

	owners_mini	price	gis_Action	gis_Casual	gis_Adventure	gis_Strategy	gis_Early Access	gis_Free to Play
owners_mini	1.000000	0.136257	0.033043	-0.072947	-0.007155	0.036483	-0.035799	0.101917
price	0.136257	1.000000	0.052604	-0.199638	-0.035213	0.068080	0.110722	-0.287285
gis_Action	0.033043	0.052604	1.000000	-0.068951	0.144392	-0.120035	0.131228	-0.007661
gis_Casual	-0.072947	-0.199638	-0.068951	1.000000	0.092181	0.103665	-0.046102	0.054187
gis_Adventure	-0.007155	-0.035213	0.144392	0.092181	1.000000	-0.081710	-0.003218	-0.044367
gis_Strategy	0.036483	0.068080	-0.120035	0.103665	-0.081710	1.000000	0.049467	0.025786
gis_Early Access	-0.035799	0.110722	0.131228	-0.046102	-0.003218	0.049467	1.000000	-0.007822
gis_Free to Play	0.101917	-0.287285	-0.007661	0.054187	-0.044367	0.025786	-0.007822	1.000000

Figure 8 - Corrélation de Pearson entre les possesseurs, le prix et les genres

Le seul genre favorisant un grand nombre de possesseurs à un terme significatif est Free to Play. Ce qui est logique, car les jeux gratuits sont beaucoup accessibles. A noter que l'on peut voir des corrélations significatives entre les genres. Avec cette information, il a été décidé d'essayer de sélectionner des jeux plus représentatifs des jeux d'Able Bear, car à ce stade, de nombreux jeux possèdent les Tag RPG et indépendants, mais ne correspondent pas au style de jeu souhaité. Il y a

trop de MMO, qui sont assez éloignés en terme de moyen financier ou de graphisme, et très peu de RPG possèdent le tag **singleplayer,** alors qu'ils ne sont pas multi-joueurs.

7.2 Examen des corrélations entre les tags et les genres homonymes

	gis_Action	tis_Action	gis_Casual	tis_Casual	gis_Adventure	tis_Adventure	gis_Strategy	tis_Strategy	gis_RPG	tis_RPG
gis_Action	1.000000	0.891833	-0.152185	-0.172299	0.056974	0.043977	-0.137860	-0.149257	-0.018720	-0.018918
tis_Action	0.891833	1.000000	-0.171856	-0.172815	0.057046	0.079745	-0.142488	-0.139258	-0.004169	0.006274
gis_Casual	-0.152185	-0.171856	1.000000	0.876652	-0.038332	-0.057962	0.011330	0.007916	-0.115919	-0.129461
tis_Casual	-0.172299	-0.172815	0.876652	1.000000	-0.052066	-0.052430	0.006146	0.016490	-0.128464	-0.130459
gis_Adventure	0.056974	0.057046	-0.038332	-0.052066	1.000000	0.856393	-0.108501	-0.119354	0.185937	0.179188
tis_Adventure	0.043977	0.079745	-0.057962	-0.052430	0.856393	1.000000	-0.124522	-0.114374	0.194195	0.204894
gis_Strategy	-0.137860	-0.142488	0.011330	0.006146	-0.108501	-0.124522	1.000000	0.902270	0.093003	0.099550
tis_Strategy	-0.149257	-0.139258	0.007916	0.016490	-0.119354	-0.114374	0.902270	1.000000	0.099774	0.116941
gis_RPG	-0.018720	-0.004169	-0.115919	-0.128464	0.185937	0.194195	0.093003	0.099774	1.000000	0.892766
tis_RPG	-0.018918	0.006274	-0.129461	-0.130459	0.179188	0.204894	0.099550	0.116941	0.892766	1.000000

Figure 9 - Coefficients de corrélations de Pearson entre les Genres et les Tags homonymes

Afin de déterminer s'il existe des variables redondantes et donc inutiles dans le Dataframe, j'examine les corrélations entre les différentes variables. Notons que sur cette figure, les valeurs des coefficients de corrélations de Pearson sont indiquées pour chaque couple de variables. L'examen de ces coefficients montre que les mots-clés des Genres et des Tags sont fortement corrélés. Il a été décidé, pour éviter la redondance d'informations, de ne pas placer les mots-clés des genres dans la suite des analyses.

7.3 Corrélation entre les Tags et le nombre de possesseurs

	owners_mini
owners_mini	1.000000
owners_maxi	0.995835
tis_Multiplayer	0.193894
tis_Co-op	0.176630
tis_Singleplayer	0.140121
initialprice	0.139706
price	0.136257
tis_Sandbox	0.132136
tis_PvP	0.131792
gis_Massively Multiplayer	0.128486
tis_FPS	0.121374
tis_Survival	0.107176
gis_Free to Play	0.101917
tis_First-Person	0.075820
tis_Difficult	0.071674
tis_Atmospheric	0.070014
tis_Horror	0.066403
tis_Fantasy	0.063965
tis_Sci-fi	0.063837
tis_Exploration	0.063578
tis_Building	0.061920
tis_Action	0.059917
tis_Tactical	0.057792
tis_Turn-Based	0.055837
tis_2D	0.055430
owners_positif	0.052158
tis_Rogue-like	0.050915
tis_Shooter	0.049879
tis_Funny	0.049490

On peut observer que les tags **singleplayer**, **multiplayer** et **co-op** sont ceux qui ont le plus d'influence sur le nombre d'acheteurs. Ils ont encore plus d'influence que le prix du jeu. Cela peut s'expliquer par la présence presque obligatoire d'au moins un de ces trois Tags dans chaque jeux.

Sans surprise, on peut voir que le prix et les jeux gratuits favorisent grandement le nombre de possesseurs de jeux en favorisant l'accessibilité. Les types de jeux les mieux placés sont :

- **Survival** (jeu accès sur la survie)
- **Horror** (jeu à suspense ou d'horreur)
- **Fantasy** (jeu avec un part de fantastique)
- **Sci-fi** (jeu futuriste ou de Science-fiction)
- **Exploration**
- **Building** (jeu de construction ou de gestion)
- **Action**

Figure 10 - Corrélation de Pearson entre les possesseurs
et les Tags correspondant au jeu développé par Able Bear Studio

8. Analyses des prix

8.1 Comparaison des différents modèles de prédiction

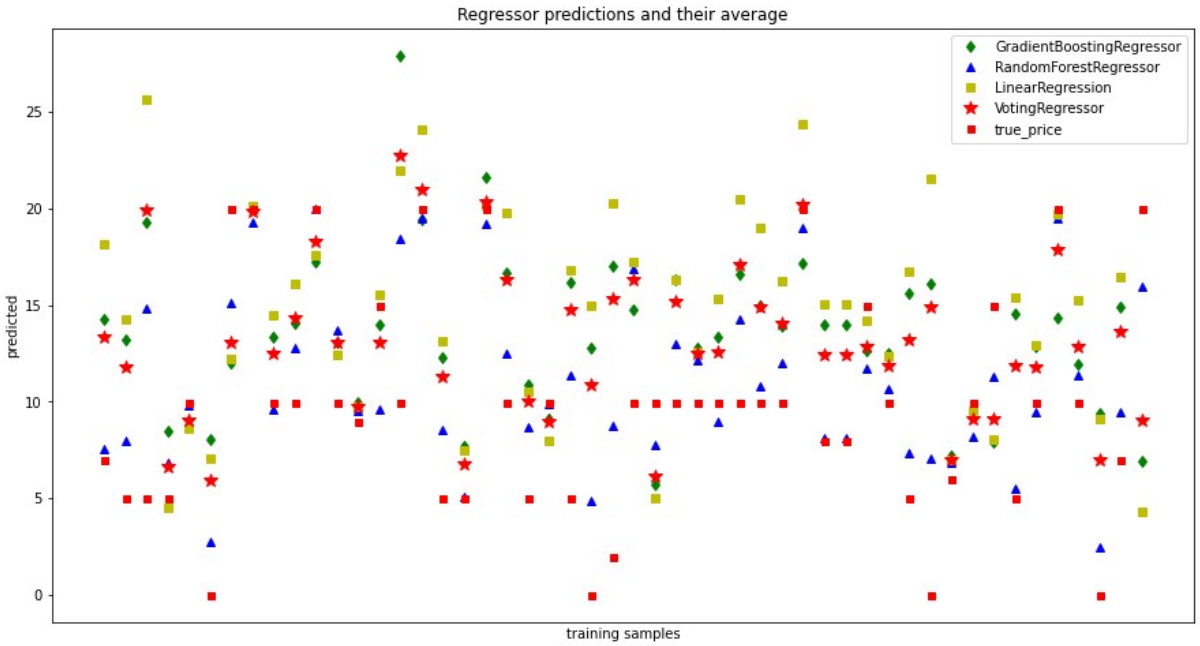


Figure 11 - Comparaison de l'efficacité de la prédiction de modèle pour déterminer le prix de vente en fonction des tags pour 50 jeux sélectionnés aléatoirement

Afin de déterminer le prix permettant de cibler un nombre d'acheteurs maximal en fonction des tags, nous avons testé plusieurs algorithmes. Dans la figure ci-dessus, j'ai visualisé ceux ayant les meilleurs résultats, c'est-à-dire ceux qui sont le plus proche du prix réel des jeux. Le vrai prix est représenté par des carrés rouge ■, les différents algorithmes de prédiction de booster de Gradient, de forêt aléatoire et de régression linéaire sont respectivement représentés par les symboles : ◆, ▲ et ■. Le dernier symbole ★, correspond à la moyenne des résultats des 3 algorithmes testés.

```
print("GradientBooster : " + str(abs((pred1 - y[:].values.reshape(1,-1)).round(0)).sum()))
print("RandomForestRegressor : " + str(abs((pred2 - y[:].values.reshape(1,-1)).round(0)).sum()))
print("LinearRegression : " + str(abs((pred3 - y[:].values.reshape(1,-1)).round(0)).sum()))
print("VotingRegressor : " + str(abs((pred4 - y[:].values.reshape(1,-1)).round(0)).sum()))
```

```
GradientBooster : 25369.0
RandomForestRegressor : 10675.0
LinearRegression : 27668.0
VotingRegressor : 20803.0
```

Nous avons en suite calculé le résultat de la somme des écarts par rapport au prix réel pour chaque méthode de prédiction.

```
print("GradientBooster : " + str(abs((pred1 - y[:].values.reshape(1,-1)).round(0)).mean()))
print("RandomForestRegressor : " + str(abs((pred2 - y[:].values.reshape(1,-1)).round(0)).mean()))
print("LinearRegression : " + str(abs((pred3 - y[:].values.reshape(1,-1)).round(0)).mean()))
print("VotingRegressor : " + str(abs((pred4 - y[:].values.reshape(1,-1)).round(0)).mean()))
```

```
GradientBooster : 6.271693448702101
RandomForestRegressor : 2.639060568603214
LinearRegression : 6.840049443757725
VotingRegressor : 5.142892459826947
```

Le résultat de la moyenne des écarts par rapport aux prix réel a aussi été calculé. Il apparaît clairement que la méthode de forêt aléatoire est la plus performante. Nous avons donc décidé d'utiliser cette méthode de prédiction et avons calculé la précision des prédictions.

8.2 Amélioration de l'algorithme de prédiction de forêt aléatoire

```
précision de la prédiction à 10%
0.26378244746600743

précision de la prédiction à 2€
0.511495673671199

précision de la prédiction à 5€
0.8707045735475896
```

Initialement, la précision du modèle de foret aléatoire pour les prix était relativement assez précise à 5€ prêt.

```
précision de la prédiction à 10%
0.37726586102719034

précision de la prédiction à 2€
0.5845921450151057

précision de la prédiction à 5€
0.9082326283987915
```

Nous avons continué d'améliorer le modèle en supprimant les Outliers (jeux gratuits, ou jeux avec prix élevé) et nous avons entraîné le modèle avec des prix compris entre 1 centimes et 30 €, car c'est la tranche de prix visés par le studio pour la mise en vente de leur jeu. Nous avons obtenu un résultat précis à moins de 2€ pour plus de 58% des jeux.

8.3 Tag conseiller pour maximiser le prix

Avec les Tag prévus initialement (Indie, Singleplayer, Adventure, RPG, Fantasy), mon modèle propose un prix de 5.57€. Mais si on ajoute les Tag liés aux combats (Tactical, Combat, Turn-Based), le prix passe à 12.52€. Il remonte en mettant des combats tactiques à 10.77 €. Les jeux avec des combats tactiques au tour par tour sont, en moyenne, les plus chers vendus (12.52 €). Les deux autres tags qui ont le plus d'impact sur le prix sont : **Story_rich**, **Multiple_ending** et **Fantasy**. Avec l'ensemble des tags cités précédemment, nous obtenons un prix de vente idéal à 13.51€.

A noter que d'autres Tags ont une influence, mais ne sont pas ici car ils ne font pas partie des Tags visés par le Studio. Par exemple, le jeu a le Tag **Stratégie** (correspond à la stratégie en temps réel de plusieurs unités), ce qui baisse considérablement le prix (-3€). Ou encore le Tag **Casual**, avec qui le prix passe à 5€ même avec les combats tactiques.

Avec ses prédictions, le Studio a décidé d'incorporer des combats tactiques en tour par tour dans leur RPG narratif, ainsi que de mettre en avant le côté fantaisie, la richesse de leur univers et de leurs histoires. A noter que cette prédiction de prix permet aussi d'évaluer le coût moyen de développement d'une caractéristique (ici les combats au tour par tour).

9. Analyse des dates de sorties

9.1 Les prix et les dates de sorties

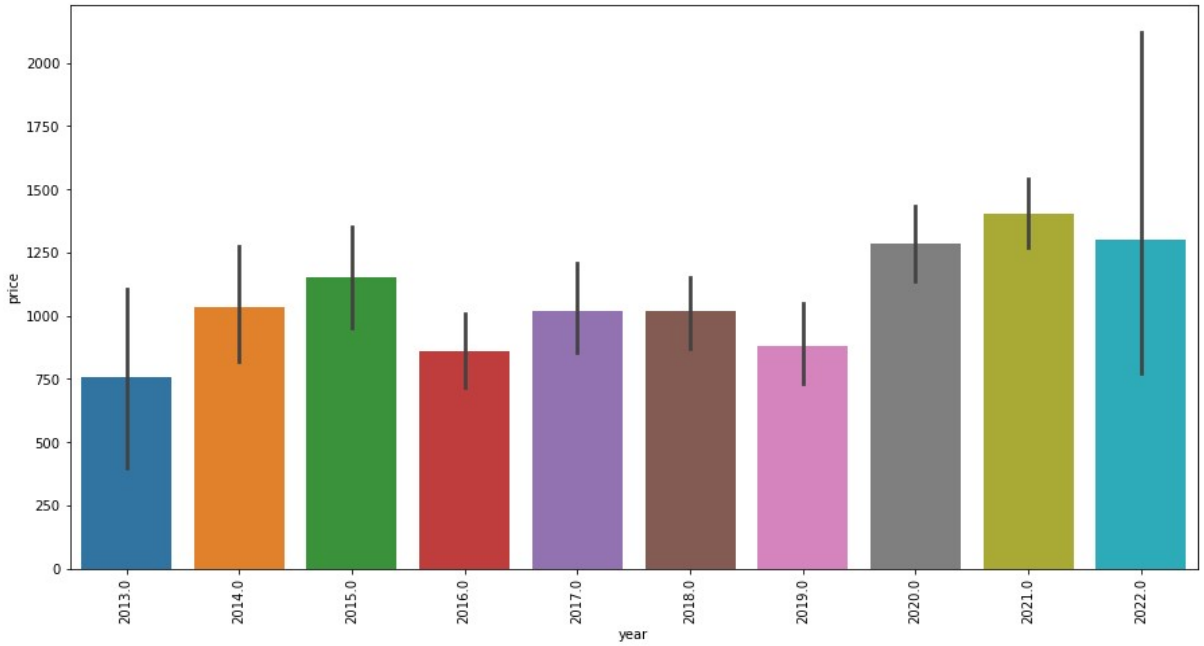


Figure 12 - Evolution du prix de sortie moyen des jeux au cours des années

On peut voir que depuis 2019, nous constatons une croissance significative des prix des jeux RPG **singleplayers** indépendants. En continuant dans ce sens, on peut estimer que le prix moyen en 2022 sera approximativement de 15€. Actuellement, les prix de 2022 sont plus faibles et ont un écart type beaucoup plus grand, car mon analyse utilise seulement les données des jeux présentement prévus pour cette année.

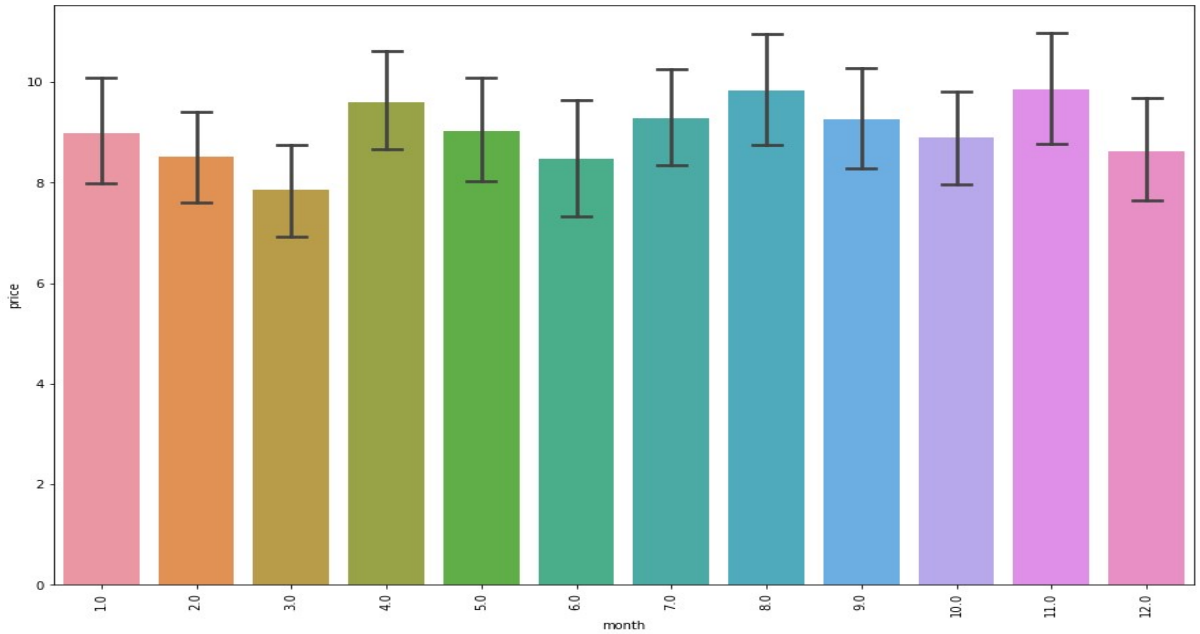
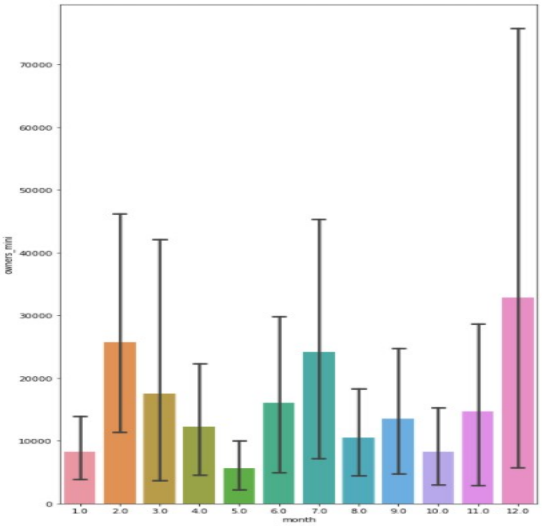


Figure 13 - Répartition du prix de sortie moyen des jeux au cours des années

Après avoir testé les différences entre chaque mois, il apparaît que les jeux qui sortent au mois de Mars son significativement moins chers que les autres mois. Aucune autre différence n'a été trouvée entre les autres mois.

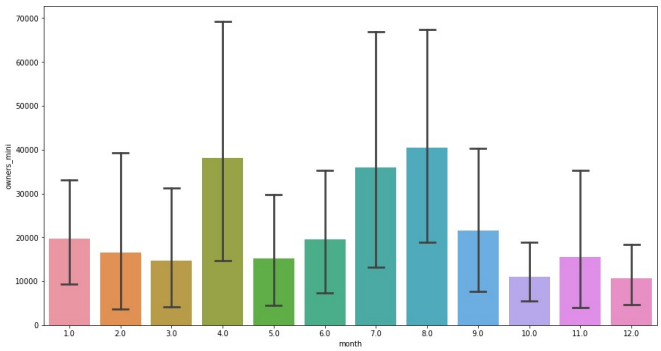
9.2 Le nombre d'acheteurs et les date de sorties



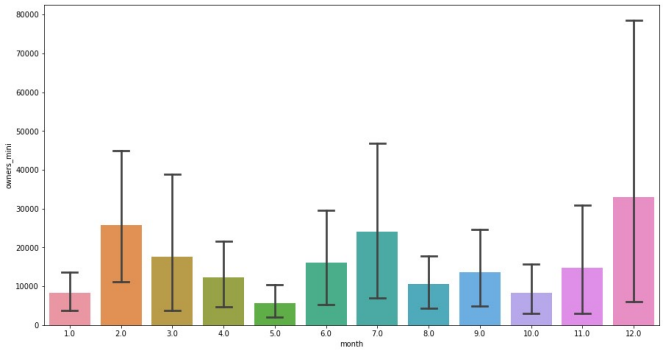
De manière générale, on peut observer que le mois de Décembre semble être celui qui attire le plus les acheteurs, sûrement en corrélation avec les fêtes de fin d'années. Sinon, les mois qui semblent être les plus intéressants pour publier des jeux RPG indépendants sont ceux de Février et d'Août, mais sans aucune différence significative.

*Figure 14 - Répartition du nombre d'acheteurs de jeux RPG **singleplayers** sortis depuis 2008 en fonction de leur mois de sorties*

N'ayant pas pu trouver de mois de prédiction pour la sortie du jeu de Able Bear Studio avec cette analyse généraliste, nous avons décidé d'essayer de trouver des corrélations sur les deux dernières années.



2020



2021

*Figure 15 - Répartition du nombre d'acheteurs de jeux RPG **singleplayers** sortis au cours de l'année 2020 et 2021 en fonction de leur mois de sorties*

Nous constatons l'effet du début du 1er confinement au mois d'Avril 2020 dû au Covid-19, ainsi que la fin du 3ème confinement, le 3 mai 2021. La différence de possesseurs entre les deux années peut s'expliquer par le temps supplémentaire qu'ont eu les jeux de 2020 par rapport à 2021 pour acquérir des joueurs. Le succès des jeux ne semble pas corrélé à leur date de sortie mais plutôt par des influences extérieures (économique, sociale, politique …)

Malheureusement, en raison d'une trop grande variance des résultats, il n'y a aucune différence significative pour nous permettre de déterminer une date de sortie optimale. Ne pouvant facilement prévoir ces différents critères, nous avons préconisé une analyse des sorties des jeux concurrents peu de temps avant l'annonce de la date de sortie, ainsi que d'éviter de sortir le jeu pendant la fin de l'année, qui est une période ou la concurrence est particulièrement forte.

10. Prédiction du nombre de ventes

Tout comme pour l'analyse de prix, j'ai essayé plusieurs algorithmes avec différents hyper-paramètres. Et une fois encore c'est la méthode de forêt aléatoire qui est la plus performante. Pour prédire quel serait le prix initial optimal, j'ai entraîné mon modèle avec des jeux possédant les tags choisis précédemment et utiliser le nombre de possesseurs maximum. L'utilisation du prix optimal permettra par la suite au studio de faire des promotions lors de la sortie du jeu afin d'attirer un maximum d'acheteurs.

	price	owners
0	0.0	95138.0
1	500.0	28675.0
2	1000.0	54150.0
3	1500.0	56155.0
4	2000.0	97948.0
5	2500.0	81841.0
6	3000.0	174537.0
7	3500.0	149856.0
8	4000.0	148111.0
9	4500.0	151325.0
10	5000.0	160694.0

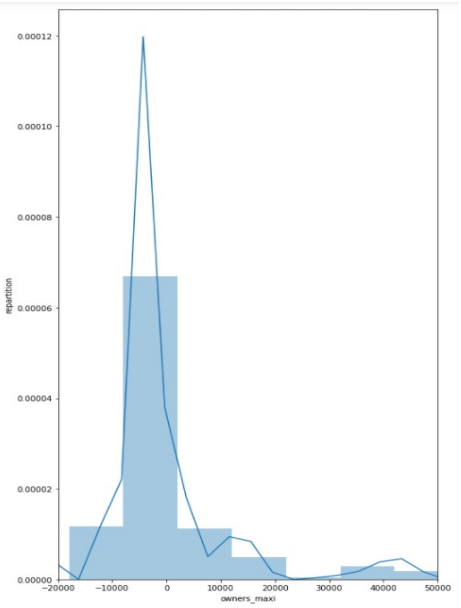


Figure 16 - Prédiction du modèle RF sur les acheteurs maximaux en fonction du prix et répartition des résidus du modèle

On peut facilement voir que l'accessibilité favorise le nombre de possesseurs. En effet, un jeu gratuit a le potentiel d'attirer plus qu'un jeu à 5€. Après analyse, les jeux qui sont à 30€ ou plus attirent beaucoup de joueurs, mais ce sont généralement des studios qui n'en sont pas à leur premier jeu ou qui constitue une suite avec déjà un certain nombre de joueurs de base, déjà acquis lors des jeux précédents. Comme on peut le voir en observant les résidus, notre modèle est précis à moins de 10 000 possesseurs pour 83% des jeux. Mais il a tendance à minimiser le nombre d'acheteurs pour les 76% des jeux n'étant pas dans l'intervalle entre -10 000 et 10 000 acheteurs.

La cible du studio étant un prix entre 10 et 30 €, le prix initial que nous avons préconisé est de 20€ pour viser un maximum de joueurs, qui est plus en adéquation avec le prix initial des jeux similaires qui a été déterminé par les Tag qui est de 13.51€.

11. Conclusion

Je tire un bilan très positif de ce stage, qui fût une expérience très intéressante autant sur le plan professionnel que personnel. Sur le plan professionnel d'abord, j'ai pu appréhender toutes les facettes du métier de data analyst, notamment le nettoyage et la préparation des données, les divers analyses en utilisant de nombreuses méthodes différentes comme la création de groupes homogènes, la recherche de corrélation ou la comparaison d'algorithmes de prédiction.

J’ai donc rempli la majorité des objectifs fixés, à savoir :
- Déterminer les caractéristiques qui intéressent particulièrement les joueurs à travers les différents tags/genres des jeux. Cela a permis de mettre en avant l’apport que pourrait potentiellement apporter la gestion de combat tactique au tour par tour, qui après discussion, sera présent dans le jeu.
- Obtenir un prix par rapport aux caractéristiques (tags/genres), pour permettre au studio de savoir où son jeu va se placer par rapport aux autres. Nous avons donc obtenu un prix concurrentiel de 13,51€ après l’incorporation des combats tactiques.
- Prédire le nombre d’acheteurs maximal en fonction du prix de vente initial du jeu. Pour maximiser le nombre d’acheteurs il a été décidé de vendre initialement le jeu à 20 €, qui permettraient de toucher un maximum de 97 000 joueurs et de réaliser des offres promotionnelles à 15 € pour rester cohérent par rapport à la concurrence.

Malheureusement je n’ai pas pu proposer de date de sortie optimale en raison de la trop grande variabilité de nos résultats. Il a malgré tout été donné quelques préconisations, comme éviter le mois de Décembre et essayer d’analyser les variables économiques et sociales, ainsi que les dates de sorties des jeux concurrents pour trouver une date adéquate.

De plus j’ai réalisé de nombreuses analyses complémentaires dont les résultats ne sont pas présentés ici, comme l’analyse de l’influence des avis positifs des joueurs sur le succès d’un jeu, les liens entre les limites d’âges (Pegi) ainsi que la moyenne du nombre de joueurs depuis toujours ou des deux dernières semaines avant la récupération des données. Ces analyses n’ayant pas apporté de résultats intéressants, elles n’ont pas été présentées ici.

Sur le plan personnel ensuite, j’ai grandement apprécié mon implication dans le développement du jeu, notamment en me demandant régulièrement mon avis sur des aspects graphiques ou de Gameplay. Au cours de cette période, comme dans toute phase d’apprentissage, il m’est par ailleurs arrivé de faire quelques erreurs comme avoir seulement utilisé les Tags et les Genres pour créer un groupe de jeux sur lesquels travaillait mon modèle, qui n’était pas représentatif du jeu actuellement développé. J’ai pu rapidement les corriger en utilisant un algorithme de Clustering pour obtenir un groupe plus ressemblant aux jeux souhaités et ainsi améliorer la précision de mes prédictions.

J’ai grandement apprécié mon stage grâce à la méthodologie de travail forte que l’entreprise d’Able Bear Studio m’a transmise, notamment la méthode Agile, et j’ai vraiment pu observer ce que j’apportais à l’entreprise. Aujourd’hui, après deux mois passés auprès de Frédéric BAST dans l’entreprise Able Bear Studio, je sais que je souhaite plutôt intégrer une structure à taille humaine, afin de travailler sur des projets moins "cloisonnés" et véritablement transversaux, pour ainsi démultiplier mon champ de compétences.

Toute mon analyse, bien que répondant au besoin du studio, pourrait être approfondie pour améliorer les prédictions en se limitant aux premiers jeux de chaque studio sur Steam pour entraîner mon modèle. De plus, à la fin de mon stage, le studio m’a annoncé vouloir des analyses complémentaires sur les performances des différents éditeurs de jeux vidéo et l’impact de passer par ceux-ci.

Après avoir analysé le jeu médian, nous nous sommes penchés sur d'autres centiles, pour lesquels les résultats étaient plus mitigés. Commençons par le point positif : le nouveau jeu qui correspond au 75e centile (c'est-à-dire le jeu qui a généré plus de revenus que 75 % des nouvelles sorties de l'année, mais moins de revenus que les 25 % restants) a rapporté 56 % de plus lors des deux premières semaines suivant sa sortie en 2019 que le jeu équivalent en 2018. Toutefois, la nouveauté qui correspond au 25e centile a généré 17 % de revenus en moins.

De manière générale, nous avons constaté que les nouveautés au-dessus du 35e centile ont généré plus de revenus en 2019 qu'en 2018, et que les nouveautés en dessous en ont généré moins.