Visualization of data

Benoit Jallet, Cédric Lejay

En cherchant un jeu de données qui soulèverait un intérêt personnel particulier, nous nous sommes mis d'accord sur le thème du jeu vidéo. Nous avons alors trouvé une liste des ventes des différents jeux sur une période qui regroupe à peu près toutes l'histoire du médium toutes plateformes dédiées confondues.

Nous avons trouvé le jeu de données "vgsales.csv" (présent à la racine du repo) à l'adresse https://www.kaggle.com/gregorut/videogamesales

Nature du jeu de données :

Il contient la liste des jeux ayant fait plus de 100 000 ventes de 1980 à 2020. Chaque jeu est définie par : son nom, la plateforme sur lequel il est sortie, l'année de parution, le genre auquel on peut associer le jeu, l'éditeur du jeu, son nombre de vente dans dans le monde entier, en europe, en amérique du nord, au japon et dans le reste du monde.

VGChartz à l'origine de ces données a à la fin de 2018 arrêté de faire des estimations sur les ventes physique des jeux. La part marché numérique des ventes de jeu ayant pris une place trop conséquente, il était devenu trop compliqué d'estimer de manière précise le nombre de ventes réel des jeux. Ainsi la quantité de jeux sur les dernières années listés dans le jeu de données ne représente pas une réalité fiable.

Potentiel corrélation entre les variables :

Le nombre de ventes et l'année. Le marché du jeu vidéo s'est démocratisé au fil du temps, le nombre d'acheteur augmentant plus on se rapproche du présent.
Nous pouvons aussi nous attendre à ce que certaines plateformes aient plus de succès sur certains genres de jeu. Mais aussi que les plateformes aient de meilleures ventes en fonction des zones géographique et de l'origine de ces plateformes.
Les éditeurs devraient eux aussi être fortement liés à un nombre limité de consoles différentes.

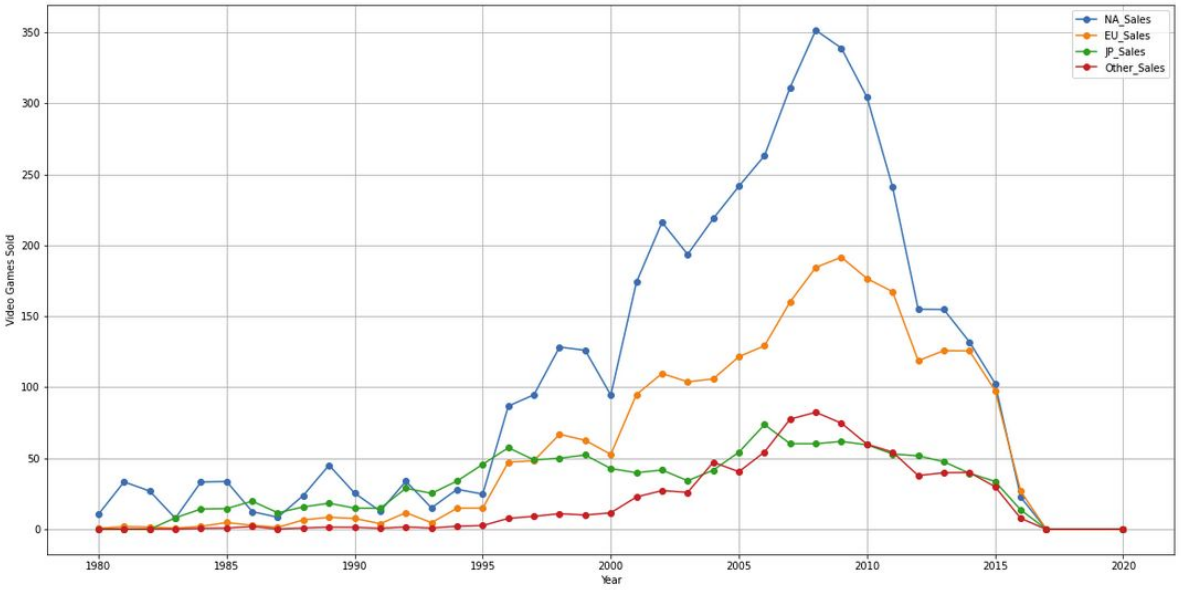
Explication de la visualisation et de l'analyse quantitative

Analyse du marché

Tout d'abord, nous avons voulu analyser la progression des ventes de jeux vidéo au cours du temps. Il est intéressant de voir les moments de notre histoire où les jeux vidéo sont devenus à la mode, mais aussi les moments où leur public s'en désintéressent.

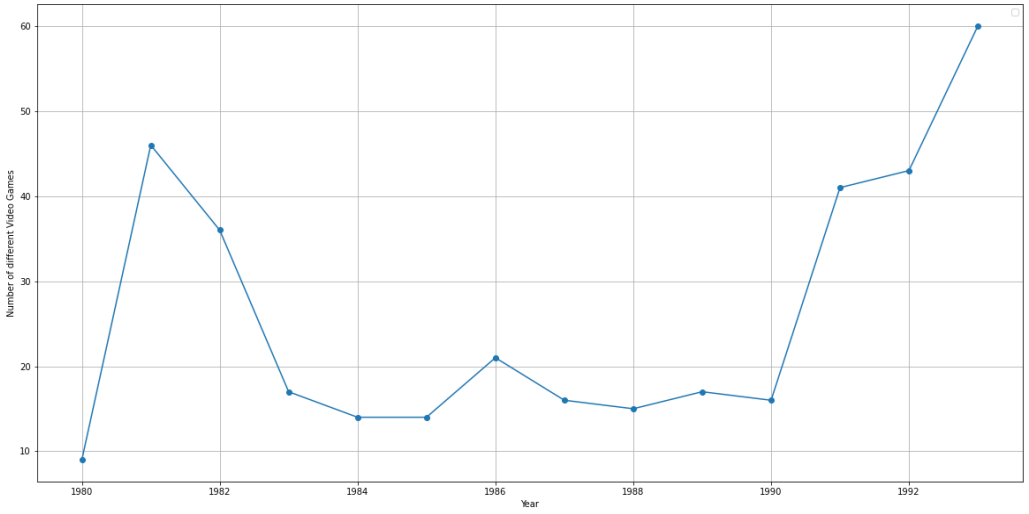
Nous avons donc lancé une analyse où l'on compare les ventes de jeux de tout genre dans les différentes régions du monde : Amérique du Nord, Europe, Japon, et le reste du monde.

Cela nous donne le graphe suivant :



A partir de ce graphe, nous pouvons constater plusieurs choses :

- Premièrement, nous pouvons remarquer que les débuts du jeu vidéo étaient plutôt instables. Nous pouvons voir que le crash du marché du jeu vidéo de 1983 n'as pas à long terme impacter le nombre total de ventes tout jeu confondu mais à diminuer drastiquement le nombre de jeu différent et cela jusqu'à la fin de la décennie (voir ci dessous).

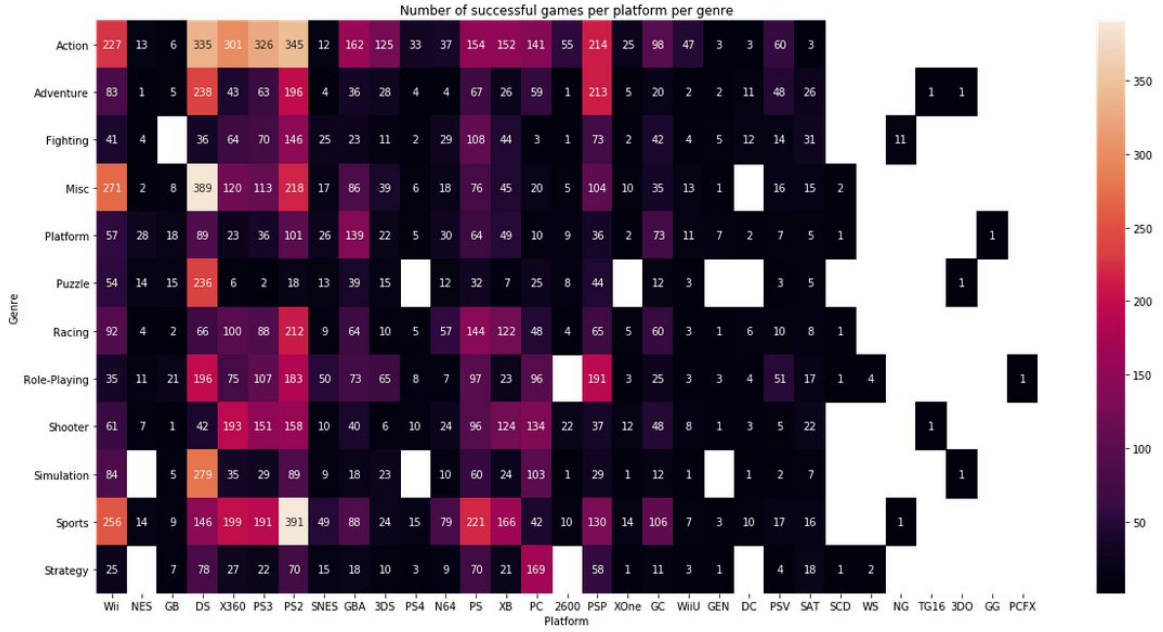


- Les ventes des jeux à succès varient beaucoup entre 1980 et 1993, surtout en Amérique du Nord. Cela peut s'expliquer par le fait que les jeux n'étaient pas encore démocratisés, n'ayant pas encore une bonne réputation. Mais également étant très cher et donc peu accessible à tous.
- C'est entre les années 1995 et 2000 que l'industrie des jeux vidéo semble avoir pris de l'ampleur.

- Nous pouvons également voir que l'Amérique du Nord est un grand consommateur de jeux vidéo. Les ventes en 2008 sont presque 2 fois plus importantes qu'en Europe.
- Enfin, nous constatons que depuis 2010 - 2011, des données semblent manquantes. Nous savons que l'industrie du jeu vidéo ne s'est pas effondrée, et pourtant le jeu de données indique le contraire. Comme indiqué par les créateurs de ce jeu de données, leurs recherches ont été difficiles à cause de la digitalisation des jeux. **Nous devrons donc prendre cette information en compte lors de nos futures analyses.**

Analyse sur les genres de jeux

Une seconde analyse que nous avons mené est sur les genres de jeux, afin de vérifier une certaine corrélation entre le genre de jeu, et la console sur laquelle il est sorti. Pour cela, nous avons décidé de générer une heatmap de la somme du nombre de jeux sortis par genre par console. Nous n'avons utilisé que les jeux datant d'au minimum 1980 et au maximum 2014, suite à la dernière analyse.
Cela nous donne la heatmap suivante :



Nous pouvons voir ici que pour certains genres de jeux, seulement quelques consoles ont sorti des exemplaires marquants. Ainsi, à l'époque de la Nintendo DS (entre 2004 et 2013), la plupart des jeux de puzzle ayant été des succès sont sortis sur cette console. Nous pouvons également constater cela avec les jeux de stratégie, dont la plateforme privilégiée semble être le PC.

Cependant d'autres genres de jeux, tels que les jeux de sport et les jeux d'action, ont fonctionné sur la plupart des plateformes disponibles sur le marché.

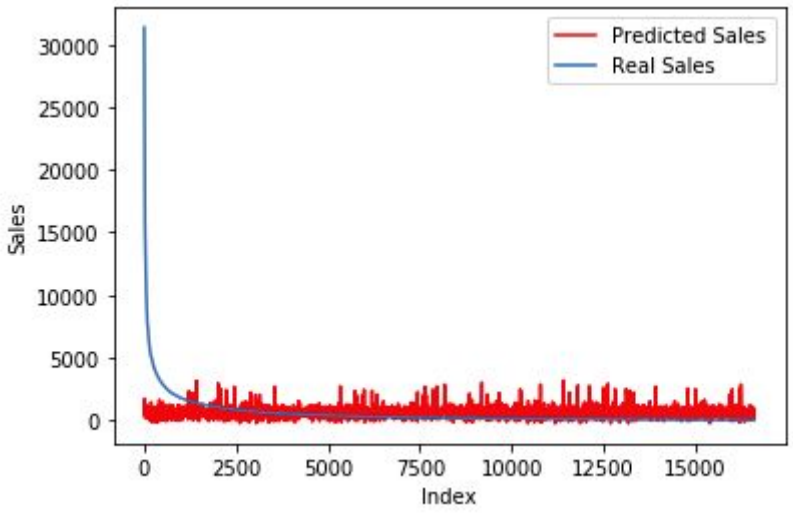
Comme nous n'avons ici que des jeux qui se sont vendus à au moins 100 000 exemplaires, nous pouvons donc en déduire que soit les consoles ayant peu de jeux listés pour un genre ont eu beaucoup de ratés, soit ces consoles n'étaient pas adaptés à ces genres de jeux.

Analyse quantitative - prediction des ventes globales

Maintenant que nous avons pu trouver quelques corrélations entre les données, nous pouvons tenter de produire un modèle prédictif.
Nous avons choisi de prédire les ventes globales de jeux vidéo, si nous devions en rajouter au jeu de données.
Pour ajouter un peu de complexité, nous avons décidé d'ajouter aux jeux un prix, générés avec une loi normale comprenant : un centre basé sur le prix moyen des jeux vendus sur ces consoles et une déviation standard de ce prix divisé par 10.

Tout d'abord, nous avons divisé notre jeu de données initial en 2 jeux différents : un jeu d'entraînement, qui nous servira à entraîner nos modèles, et un jeu de test afin de les valider.
Nous avons dû traiter les données de notre jeu, car certaines étaient inutiles (le nom, le rang), nous devions retirer les labels possibles, à savoir les ventes, mais également rendre exploitable les données non numériques.
Pour cela, nous avons utilisé un encodage one-hot, ce qui nous crée une matrice permettant de catégoriser les valeurs dans les colonnes non numériques.
Nous avons également échelonné les valeurs numériques avec un StandardScaler pour que chaque colonne soit traitée de la même manière par les modèles que nous utiliserons.
Nous avons réuni tout cela dans une pipeline, qui nous sert à appliquer ces traitements sur le jeu d'entraînement, puis sur les jeux de test.

Concernant les modèles utilisés, nous en avons testé plusieurs : un modèle de régression linéaire et un arbre de décision régresseur.
En analysant la RMSE et la déviance de ces deux modèles sur notre jeu de données, nous avons constaté que la régression linéaire s'adapte beaucoup mieux, et avait des prédictions plus précises.
Nous avons tout de même souhaité tester ces deux modèles prédictifs sur le jeu de test.
Mais les résultats furent les mêmes.
Cependant, malgré de meilleures précisions, les résultats ne nous convenaient pas, car trop aléatoires. Nous avons donc testé en enlevant certaines colonnes, et il nous a paru évident que la colonne des publieurs causait beaucoup d'aléatoire dans nos prédictions, à cause de leur diversité trop importante. Nous avons donc entraîné de nouveau notre modèle de régression linéaire, et voici le résultat sur le jeu de test :



Malgré tout, les résultats restent assez aléatoires. Ceci peut être dû à des outliers, notamment les fortes ventes de certains jeux par rapport à d'autres. Par exemple, le jeu WII Sports est bien au-dessus de la moyenne des ventes de notre jeu de données.
Ainsi, un traitement de la donnée encore plus poussé pourrait nous donner des prédictions plus précises, sans pour autant faire de sur-apprentissage.