GamoScopy

Rapport d’exploration, de data visualisation et de pre-processing des données

Introduction au Projet

Le projet consiste à analyser la vente de jeux vidéos à travers le monde à partir de données existantes obtenues via le web scraping ou encore le data mining. Les orientations porteront sur l'aspect économique avec un volet prédictif, un aspect socio-culturel et un aspect sociétal.

01

Contexte

Insertion projet d'un point de vue technique, économique et scientifique.

02

Objectifs

Description des objectifs du projet et niveau d'expertise de la problématique.

Introduction au Projet

Contexte Projet

Mise en situation : Mandatés par un éditeur de jeux vidéos qui souhaite se positionner sur le marché, nous devons réaliser une étude pour déterminer les facteurs de succès en terme de ventes. Par le biais d'une analyse approfondie, l'objectif est d'orienter la stratégie de développement vers des choix pertinents.

ETAPES

01 **Collecte des données pertinentes**

02 **Analyse des données recueillies**

03 **Prédictions et estimations**

04 **Résultat de l'étude / Conclusion**

Introduction au Projet

Principaux Objectifs

01

Analyse des ventes

Et détermination des facteurs-clés qui influent sur le volume des ventes

02

Profilage des joueurs

Pour mieux connaître la cible et les stratégies marketing à mettre en place

03

Analyses corollaires

Scores PISA (performance scolaire), temps passé à jouer

Introduction au Projet

Expertise de la Problématique

Guillaume

Joueur PC

Parcours en Sciences Sociales et Analyse de Discours

Céline

Entourée de gamers

02

Educatrice accompagnant les atypiques et membre d'une association ALERTE ECRANS qui lutte contre les écrans récréatifs

Adeline

Passionnée de jeux vidéo

Consultante en webmarketing et accompagnement en transformation digitale

Compréhension et Manipulation des Données

Analyse et rapport d'exploration des jeux de données recueillis pour atteindre les objectifs préalablement décrits.

01

Cadre

Descriptif des jeux de données utilisés au sein du projet

02

Pertinence

Variables, particularités et limitation des données recueillies

03

Pre-processing / feature engineering

Processus de traitement des données, standardisation

04

Visualisation / Statistiques

Relations entre les données, distribution et rapports statistiques

Données : Cadre

VG Chartz

Metacritic

Wikipédia

JeuxVideo.com

Twitter (à venir)

Twitch / IGDB

Web scraping

Le scraping en lui-même n'est pas illégal tant que les données sont accessibles publiquement. En revanche la réutilisation des données, présente des risques et doit faire l'objet d'une étude juridique fine axée sur le niveau de transformation des données collectées. Attention, de nombreux sites web se protègent des robots et leur bloquent ainsi l'accès (utilisation de proxies pour contourner le blocage),

- **Vgchartz** : Dataset initial étant la seule source trouvée à faire état du volume des ventes
- **Metacritic** pour obtenir les notes et les résumés des jeux
- **Wikipédia** pour la liste des développeurs et leurs pays respectifs, les scores PISA
- **JeuxVideo.com** pour la classification (PEGI)

API (Application Programming Interface)

Accès à des modèles sans impacter les systèmes opérationnels. Cela facilite grandement la maintenabilité, la mise en place d'A/B test et la sécurisation des models.

- **Twitch** – documentation – Profilage des joueurs
- **Twitter** – documentation – Analyse de sentiments

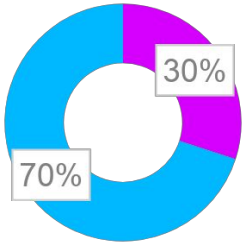
Autres Sources

- **Excel** : liste des consoles

Compréhension et Manipulation des données

Pertinence des données

Description
Sharing des sources



VG Chartz
Dataset de base

Autres sources
Compléments

Variables pertinentes

- **Volume des ventes (cible)**
- Plateforme / console
- Genre
- Développeur
- Année de sortie
- Notes utilisateurs

Particularités / Limites

- **Beaucoup de données manquantes** notamment sur les ventes
- **Données relativement fiables** (sources obscures – voir article Wikipédia) mais aucun jeu de données concurrent connu
- **Problème de typage des données** : jeu (année qui se rajoute, caractères spéciaux ...), mélange de franchises, jeux multi-plateformes (non identifiables) et jeux uni-plateforme, éditeurs et développeurs pas catégorisés à retravailler
- **Impossibilité d'établir une répartition des ventes par région** suite à la dématérialisation des ventes.

62253
Lignes
Dataset de base

21885
Lignes après cleaning
=> 35 % des données

Compréhension et Manipulation des données

Pre-processing et feature engineering

Types

Re-typage des données selon le type de variable. Float ou integer pour les variables numériques dont la variable cible, string ou object pour les variables descriptives, datetime pour les années. Plus d'infos dans les notebooks

Valeurs Manquantes

Suppression ou remplacement en fonction de leur pertinence. Lignes sans ventes supprimées d'office la variable cible étant le volume des ventes. Récupération des années manquantes sur Metacritic.

Doublons

Suppression des doublons après analyse. Le site VGChartz a quelques lignes en doublon d'origine. Exemple : Jeu comprenant 2 lignes car 2 genres attribués. Nécessité de séparer le dataset en 3 jeux de données : Franchises, Multi-plateformes, et Jeux par ligne du fait de la redondance et des merges pertinents

Transformation des données

Pour faciliter les merges entre datasets, replace avec dictionnaires, rename, regex, formatage des données (majuscules etc.). Création de datasets intermédiaires avec méthode Groupby pour collecter un maximum de données

Chart Data Slide

Evolution des Ventes (2000-2022)

Le nombre moyen de jeux vendus par titre semble augmenter à compter de 2017.

On passe d'une Moyenne de 2 Millions d'exemplaires en 2016 VS plus de 4 Millions d'exemplaires en 2022

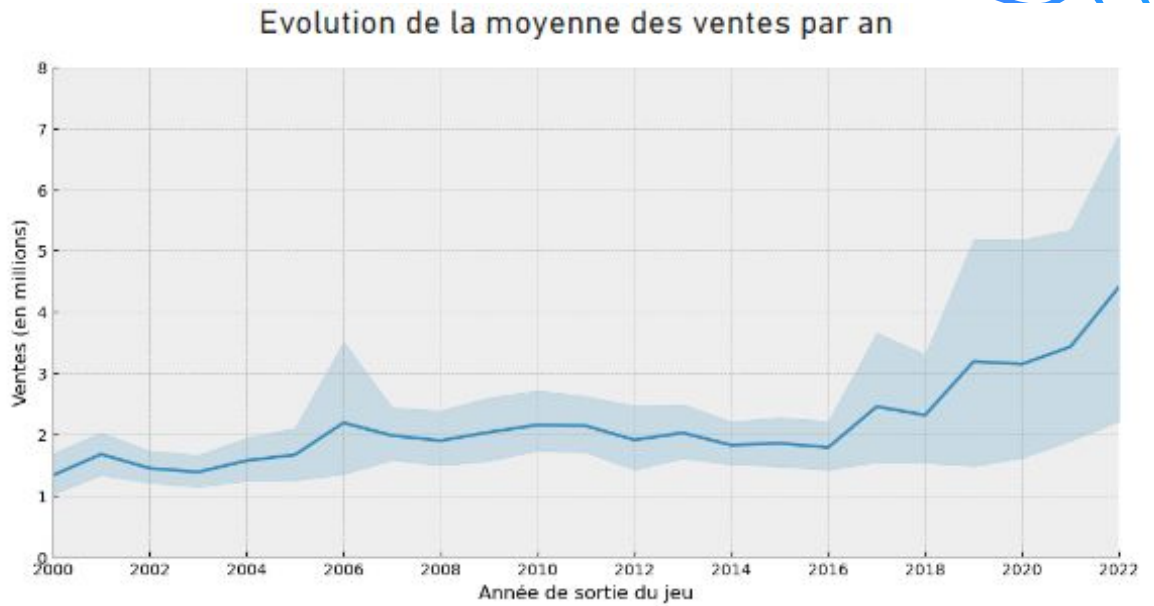
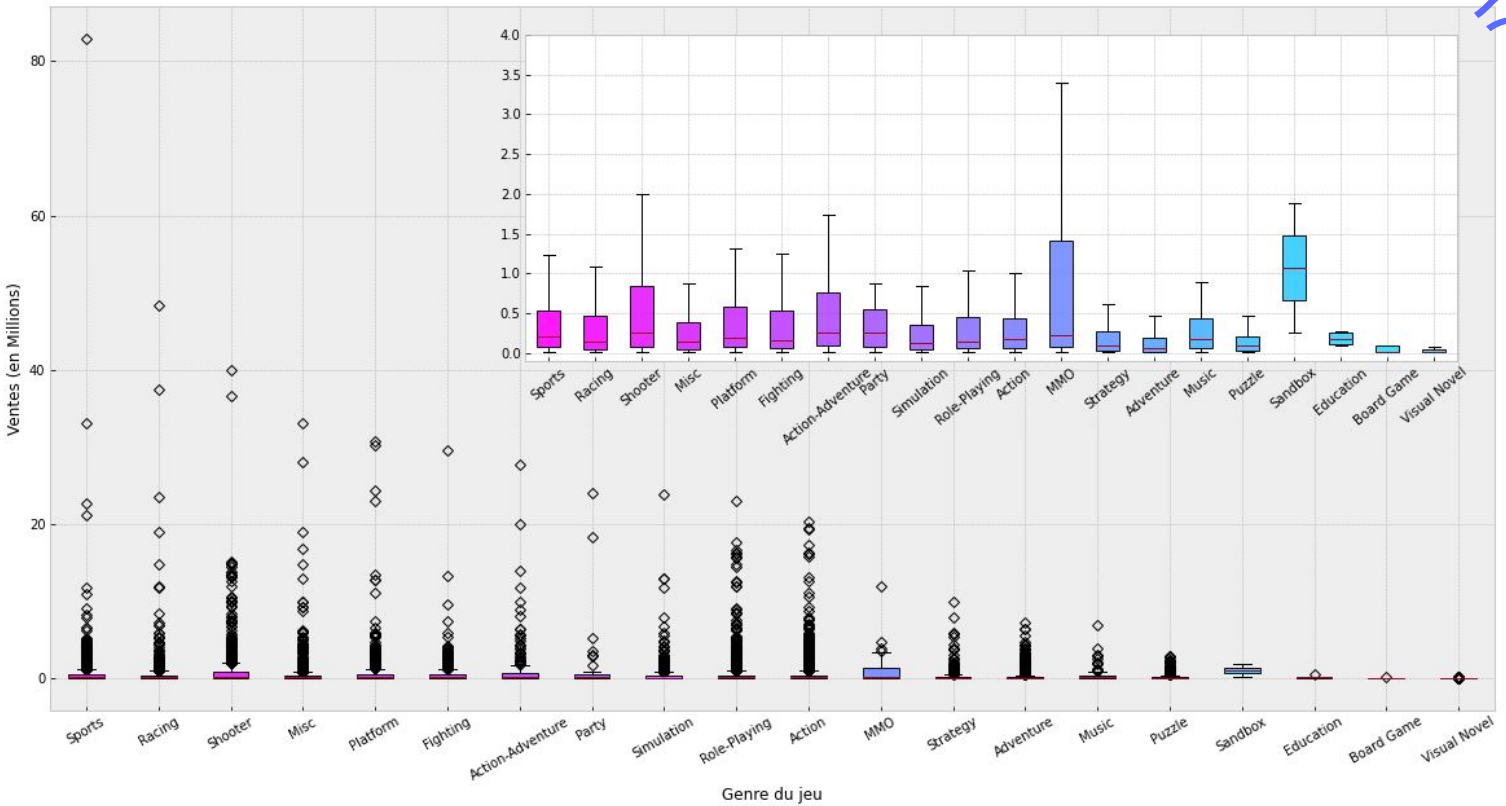


Chart Data Slide

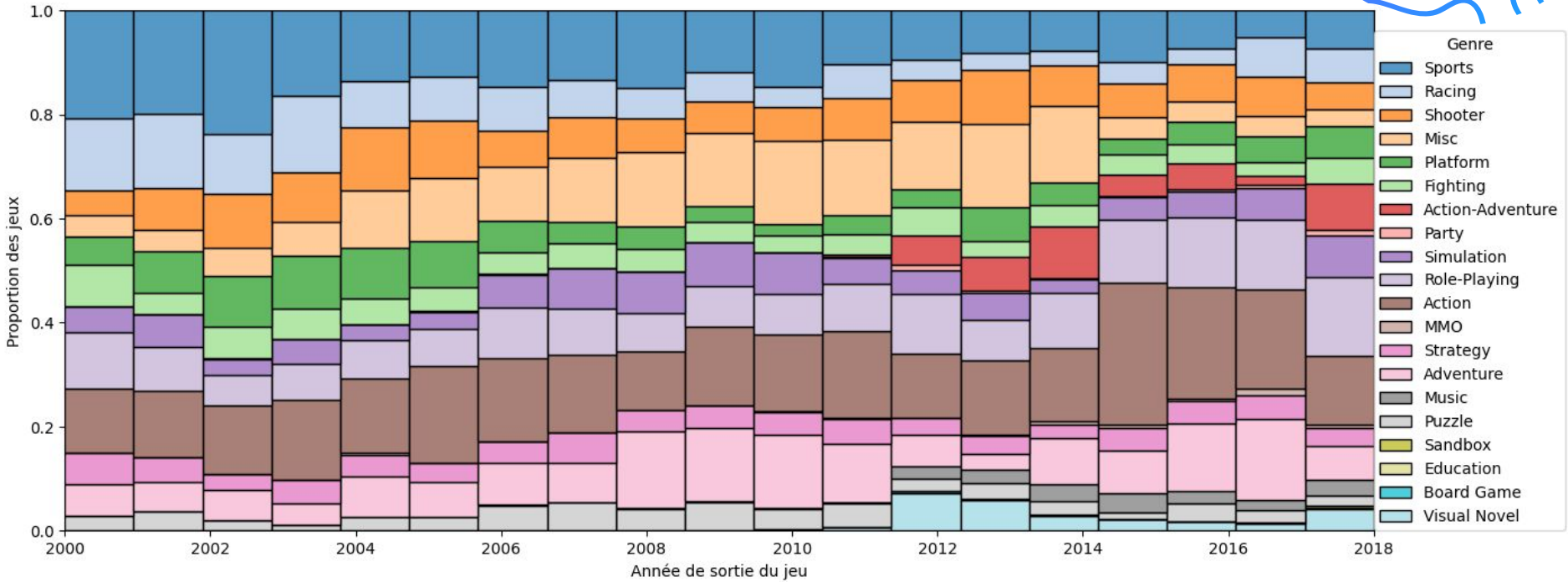
Répartition des Ventes par Genre



Le genre du jeu a un effet statistique significatif sur les ventes du jeu (P-value = 2.406063e-42)

Chart Data Slide

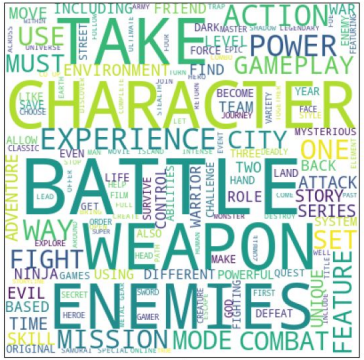
Répartition des Genres par années



On observe par exemple l'apparition des genres "Action-Aventure" et "Visual Novel" en 2012, ou le déclin des genres "Sports", "Misc" et "Racing" sur la période.

Chart Data Slide

Analyse Sémantique par Genre

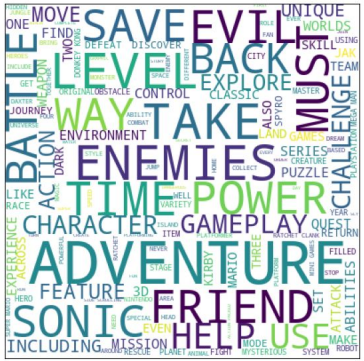


Action

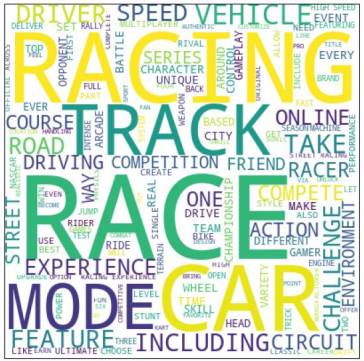
Shooter

Sport

Role-play



Platform



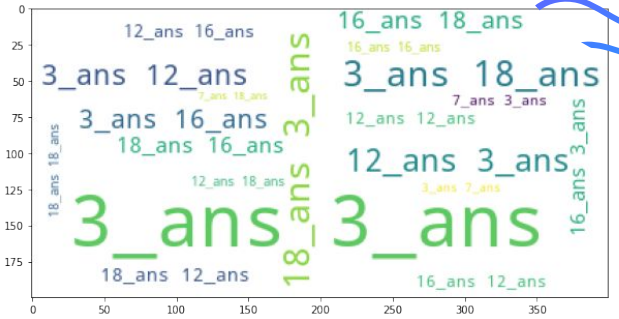
Racing

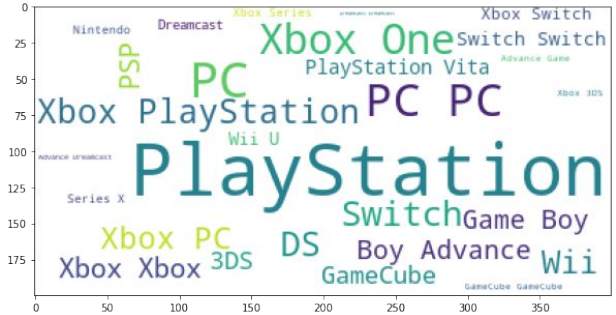
Mots les plus utilisés dans les descriptions des jeux des genres les plus populaires.

Chart Data Slide

Exploration Sémantique par Jeu, Plateforme et Classification PEGI

Les mots les plus utilisés dans la nomination des jeux vidéos, des plateformes et des catégories d'âges.

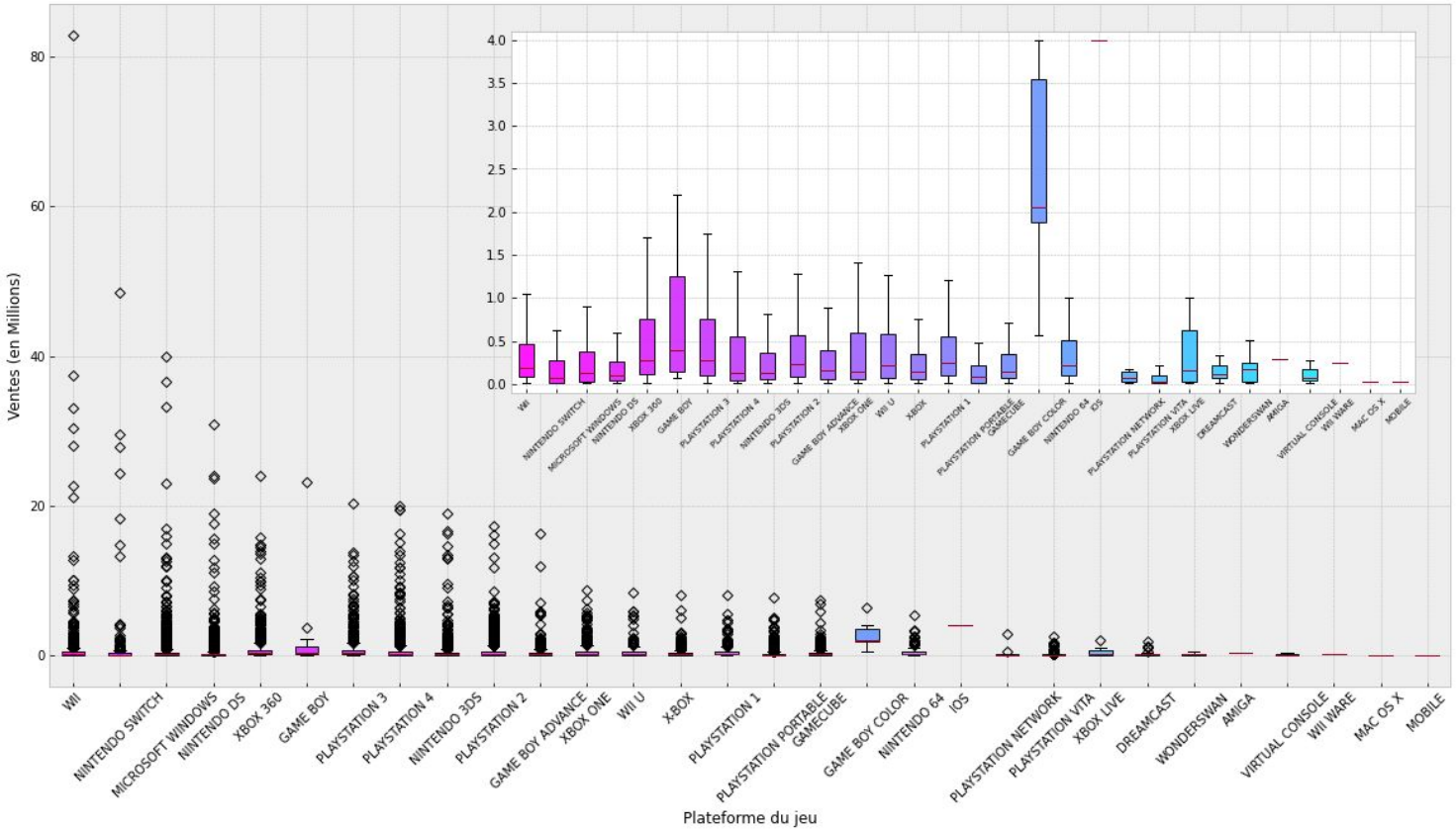




Source : dataset All_gameswithpegi

Chart Data Slide

Répartition des Ventes par Plateforme

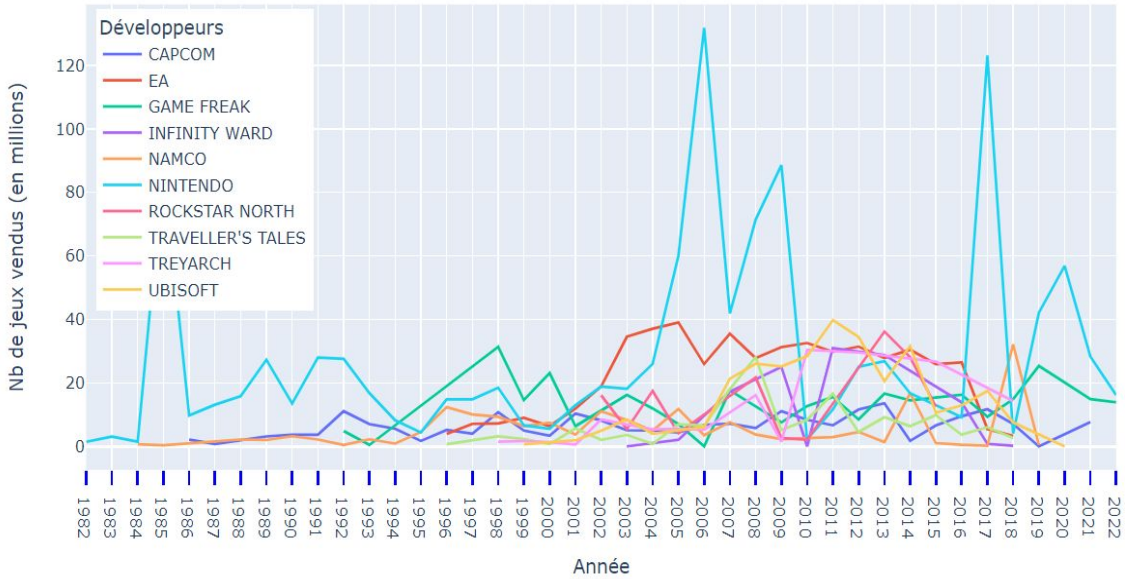


La console de sortie du jeu a un effet statistique significatif sur les ventes du jeu (P-value = 2.556184e-37)

Chart Data Slide

Ventes par Développeur - Top 10

Évolution des ventes de jeux vidéo par développeur de 1982 à nos jours avec échantillonnage top 10 des développeurs.



Infos-clés

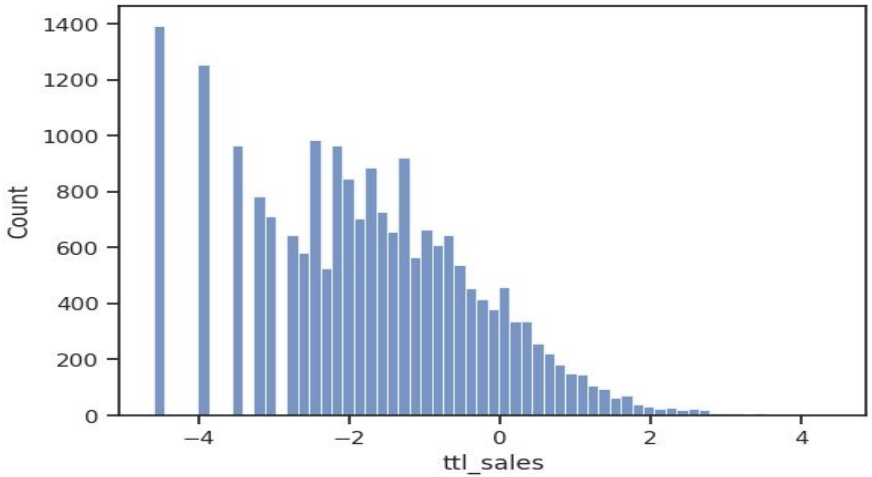
Nintendo a su maintenir une grosse part du marché face à l'arrivée de nombreux studios de développement au fil des ans.

Malgré une entrée plus récente sur le marché (1992), **Game Freak** dispose d'une belle part du marché, certainement liée à des franchises telles que Mario Bros ou Pokémon.

EA après une belle entrée dans l'industrie, semble en régression depuis 2017 tandis que **Capcom** se maintient sur le marché depuis plus de 30 ans

Chart Data Slide

Distribution du journal de vente
Dataset modélisation



L'histogramme représente la **distribution de ventes de jeux vidéos à l'échelle logarithmique**.

- Axe horizontal : valeurs de ventes à l'échelle logarithmique
- Axe vertical : fréquence de chaque valeur de ventes.

La majorité de jeux vidéos ont des ventes relativement faibles, tandis qu'un petit nombre de jeux vidéos ont des ventes relativement élevées : un petit nombre de titres très populaires représente une grande part des ventes totales

Etapes de Réalisation Du Projet

Analyse et rapport d'exploration des jeux de données recueillis pour atteindre les objectifs préalablement décrits.

01

Classification du problème

Types de Machine Learning et métriques de performance

02

Choix du modèle / Interprétation

Algorithmes utilisés et optimisation (analyse erreurs et amelioration performances)

Etapes de Réalisation Du Projet

Classification du problème

Types de Machine Learning

Classification
Régression

Métriques de performance pour comparaison

Accuracy, (F1-score), Recall, RMSE, MAE, R2

Prédiction

Savoir estimer la vente de jeux vidéos en fonction des features

Etapes de Réalisation Du Projet

Choix du modèle

Premier choix de modèle :

Résultat du “.score” de différents modèles de machine learning sur un premier dataset test.

	Linear	**DecisionTree**	**RandomForest**
Regressor	-13.58	0.07	0.32
Classifier (5 classes)	0.45	0.43	0.47

Le **RandomForestClassifier** semble être le plus prometteur !

Etapes de Réalisation Du Projet

Amélioration du modèle

Évolution de l'accuracy du modèle.

Modification ajoutée **(comprend toujours les modifications précédentes)**	**Accuracy**
Modèle de base	0.475
Réduction du nombre de catégories à 3	0.581
Re-sample	0.586
RandomGridSearch	**0.603**

Modèles Avancés en test

Top 10K , éditeur de 3 ou plus Jeux vidéos
Variable cible : ttl_sales (total des ventes en millions)

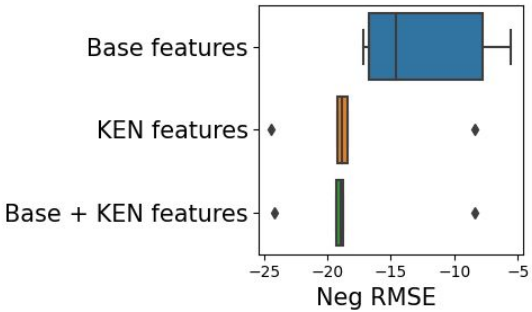
Essais avec plusieurs modèles pour trouver ceux qui donnent de meilleures performances

Modèles	Performances
SVR/ GridSearchCV	RMSE – 2,96
Select KBest et régression linéaire	R2 - 1
Ridge	MAE – 1,5272
XGBoost	RMSE – 1,7153
Neural Network Model (tensorflow/keras)	2,86
Neural Network Model Sequential/keras)	8,7754

Confirmation : Le RandomForestClassifier EST le plus performant - Score : 0.603

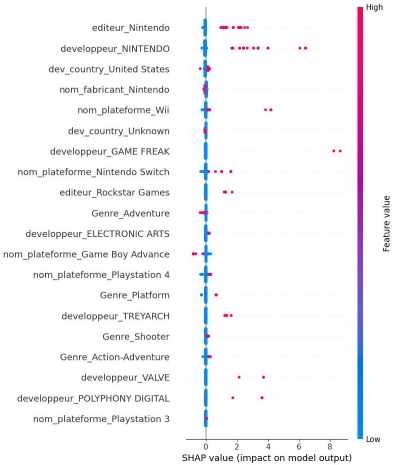
Modèles avancés
Visualisation des résultats

Dirty Cat / Get Ken Embedding



XGBoost et SHAP

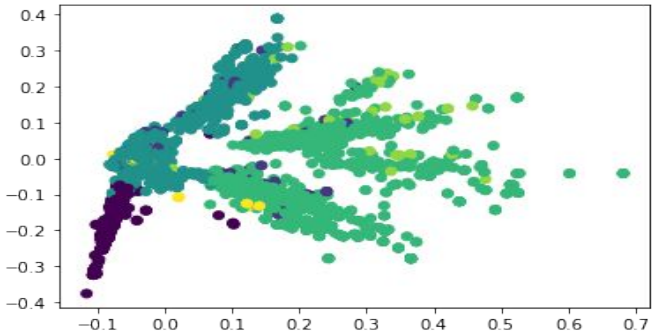
ici graphique de répartitions des forces pour visualiser les valeurs SHAP

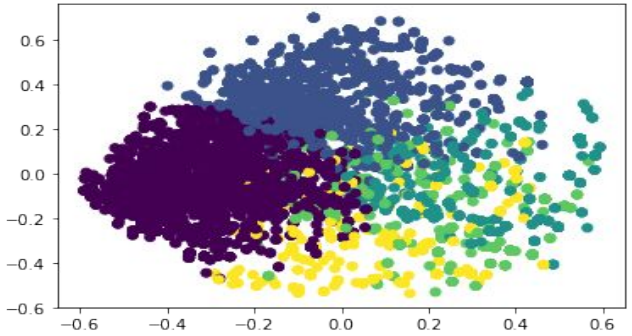


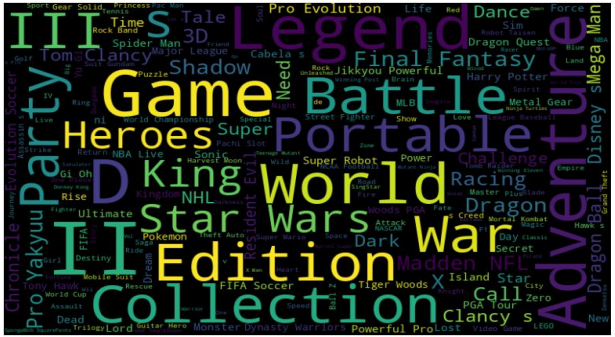
Base features : performance du modèle avec uniquement les caractéristiques de base (éditeurs, plateformes, genres..)
Ken features : features catégorielles encodées avec ken embedding

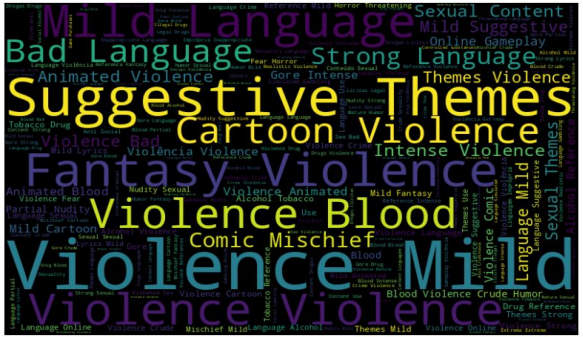
Jeux et description du contenu en fonction de l'âge

Tokenisation, vectorisation de la colonne (jeu / contenu description), **réduction de dimension** (PCA)/nuages de mots tokenisé pour en faire ressortir les mots qui ressortent le plus









Conclusion Tirées

Analyse et rapport d'exploration des jeux de données recueillis pour atteindre les objectifs préalablement décrits.

01 **Difficultés rencontrées**
Descriptif des jeux de données utilisés au sein du projet

02 **Bilan**
Contributions au projet et atteinte des objetcifs

03 **Suite du projet**
Pistes d'amelioration et apports du projet

04 **Bibliographie**
Supports bibliographiques utilisés lors du projet

Conclusions Tirées

Difficultés rencontrées

Tâche plus longue que prévu pour la construction du jeux de données multi-sources, l'objectif étant de collecter un maximum de features potentiellement determinants.

Retraitement (fastidieux) des données necessaire pour l'obtention de données exploitables et merge des différentes sources de données (jeuxvideo.com, metacritic, igdb, twitch, etc). Exploration des données et decision de les aggréger par type (franchises, multi plateformes et uniques)

Temps très court pour assimiler les méthodes et compétences pour scraping qui sont à la base du jeu de données qui ont été construits, et compléments de formation / méthodes obtenus au fil des entretiens avec nos mentors (Romain et Gaspard)

Difficultés d'optimisation de récupération des données (scraping) et **randomgrid** (45 minutes de test 100 modèles), éviter le ban en scraping / API, Prise en main colab / Anaconda, Ressources CPU Power BI

Chacun d'entre nous a rencontré des difficultés techniques selon ses aptitudes mais au travers de nos échanges de partager nos difficultés pour se challenger et donc performer pour trouver des solutions et avancer au mieux dans notre projet.

Conclusions Tirées

Bilan

Pour atteindre les objectifs fixés, nous avons cherché à compléter et updater le jeu de données initial avec des données qui nous semblaient pertinentes

Nous avons prévu d'améliorer nos résultats (si possible) au travers de nouvelles itérations avant la soutenance.

Les résultats obetnus par d'autres data analysts sur le même type de jeux de données semblent meilleurs que les nôtres, toutefois nous avons cherché à améliorer nos résultats.

Atteinte des objectifs

01	**Analyse des ventes**	Nous avons concentré nos efforts sur cette partie pour determiner les critères principaux influençant le volume des votes
02	**Profilage des joueurs**	Nous n'avons pas eu le temps de developer ce sujet, et le type de données collectées à ce jour ne permet pas une analyse pertinente
03	**Analyses corollaires**	Scores PISA (performance scolaire), temps passé à jouer. Nous n'avons pas pu récupérer à ce jour des données exploitables pour analyse.

Conclusions Tirées

Suite du Projet

1. Améliorer le jeu de données grâce à :

- L'obtention de données de professionnels du jeu vidéo, données non publiées et non publiques,
- La récupération données de type commentaires joueurs pour analyse de sentiments,
- Du temps ☺️

2. Développer nos compétences théoriques / Pratiques

Ce projet du fait de sa complexité, nous a permis de mettre en pratique les modules théoriques et d'appréhender des difficultés « sur le terrain », mais aussi d'aller chercher d'autres méthodes / connaissances en dehors du cadre de la formation pour une meilleure pertinence / cohérences des données et modèles.

3. Compléter l'analyse des ventes avec des objectifs complémentaires

Pour une étude plus complète et approfondie, ce qui est logiquement attendu dans un cadre professionnel (et donc plus de temps ☺️)

Conclusions Tirées

Bibliographie

Benchmark :

https://github.com/CraigKelly/steam-data (particularité est le merge un peu pénible de datasets)
https://github.com/leomaurodesenv/game-datasets (particularité : le personne a gardé tous les datasets de jeux qu'il a trouvé)
https://steamdb.info/sales/ (toutes le ventes de jeux, particularité le site ne veut pas être scrapé mais fournisse dans le cadre éducatif)
https://github.com/SteamRE/SteamKit (api de steam)
https://www.kaggle.com/code/alanhsu/notebook66669898b5 (notebook jeux steam)
https://steamspy.com/ (data et stats sur les jeux vidéos)
https://data.world/yansian/popular-video-games (data jeux vidéos)
https://dirty-cat.github.io/stable/auto_examples/06_ken_embeddings_example.html#sphx-glr-auto-examples-06-ken-embeddings-example-py (dataset jeux vidéos et modelisation)

Modules :

131-Text Mining
133-WebScraping avec BeautifulSoup
134-Text Summerization
137-WebScraping avec Sélénium
120-Machine learning pour les Data Analysts
121-Algorithmes et méthodes de classification avec scikit learn
123-Méthode de clustering
124-Méthode de régression
125-Méthode de réduction de dimension
127-Series temporelles avec statsmodels
Masterclass Text mining -Avis d'internautes

Conclusions Tirées

Bibliographie

Ressources

ScikitLearn2.pdf
ScikitLearn1.pdf
Feature_Scaling.pdf
file:///C:/Users/celine/Downloads/Machine_Learning.pdf

Ref BPI:
https://www.youtube.com/watch?v=rcFFEEHPeE8
Guy in a Cube - YouTube
SQLBI - YouTube
https://learn.microsoft.com/fr-fr/power-bi/connect-data/desktop-python-scripts

Overfitting :
Random forest overfitting - Crunching the Data

Models ML
API Reference — scikit-learn 1.2.2 documentation
1.13. Feature selection — scikit-learn 1.2.2 documentation

Interprétabilité :
Welcome to the SHAP documentation — SHAP latest documentation
SHAP notebook example

Articles

Score pisa

- https://www.education.gouv.fr/pisa-programme-international-pour-le-suivi-des-acquis-des-eleves-41558
- https://www.linternaute.com/actualite/education/1310839-pisa-2022-quand-est-prevue-la-prochaine-publication-analyse-des-derniers-resultats/
- https://fr.wikipedia.org/wiki/Programme_international_pour_le_suivi_des_acquis_des_%C3%A9l%C3%A8ves

Ventes de jeux vidéos (france)

- https://ciliabule.fr/analyse-marketing-jeux-video/#:~:text=Dans%20l'Hexagone%2C%20ce%20march%C3%A9,dans%20la%20vie%20des%20fran%C3%A7ais.
- https://www.sell.fr/sites/default/files/essentiel-jeu-video/essentiel_du_jeu_video_mars_2022_sell.pdf
- https://www.afjv.com/chiffres_jeux_video.php?c=ventes