



RAPPORT PROJET VIDEO GAMES

DEC23 BDA

Dans ce rapport, vous trouverez l'ensemble de la méthodologie appliquée, ainsi que les résultats obtenus afin de répondre à la demande initiale de ce projet

Nous estimerons la tendance des ventes totales de la franchise « Grand Theft Auto » à l'aide d'informations descriptives, de données issues d'un dataset fournis et de webscraping.

De plus, une étude d'analyse de sentiments sera effectuée afin de quantifier l'engouement généré par cette franchise sur l'axe de Storytelling et celui de GTA VI avant sa sortie en 2025.

Par Sonita Arjoon , Aymeric Chalmot de la Mesliere, Sabri Guery et Ezzat Saoud, Février 2024

Mentor : Yaniv Benichou



SOMMAIRE

1.	Présentation de l'équipe et du projet	Page 2
1.1	La Team Projet	
1.2	Le Projet Video Games	
2.	Exploration des données	Page 4
2.1	Exploration préliminaire des variables du jeu de données	
2.2	Focus sur Take two Interactive	
2.3	Axe de recherche complémentaire	
2.4	Grand Theft Auto : en quelque mots	
3.	Webscrapping	Page 6
3.1	Méthodologie	
3.2	Données retenus	
3.3	Autres données retenues sur internet	
4.	Dataset description	Page 9
4.1	Dataset toutes variable hors "commentaires"	
4.2	Dataset variable "commentaires"	
5.	Analyse du sentiment	Page 11
5.1	Text Mining	
5.2	Wordcloud	
5.3	Sentiment Analysis avec "vader"	
6.	Data visualisation	Page 14
6.1	Évolution des ventes des GTA à aujourd'hui	
6.2	Évolution des avis au fil du temps	
6.3	Corrélation entre les variables	
7.	Modélisation	Page 19
7.1	Modèle de prédiction	
7.2	Autre supposition	
7.3	Méthodologie du Scrapping	
	Conclusion	Page 20
	Projet/Résultat - Perspectives	
	Les difficultés rencontrées – Et si c'était à refaire ?	



1. PRÉSENTATION DE L'ÉQUIPE ET DU PROJET

1.1 La Team Projet (avant/après)

Ce projet "fil rouge" a été réalisé dans le cadre de notre formation de Data Analyste, afin de mettre en pratique l'ensemble des concepts et méthodologies apprises tout au long de cette formation de 3 mois.

Notre équipe était composée de 4 personnes, avec des profils différents. Aucun d'entre nous, n'avaient pas d'expérience métier en Data.

Certains sont ingénieurs avec quelques connaissances théoriques acquis lors de leurs études supérieures et d'autres en reconversion professionnelle sans apprentissage spécifique sur le domaine.

Le projet Video Games, nous a permis d'acquérir de l'expérience significative sur :

- La gestion de projet Data
- Le codage en langage Python
- Les librairies Python
- Les méthodologies suivantes :
 - ✓ Data cleaning
 - ✓ Pré-processing
 - ✓ Data visualisation
 - ✓ Webcsrapping
 - ✓ Sentiments Analysis via du text Mining
 - ✓ Modélisation

De plus, nous avons été amenés à utiliser plusieurs types d'outils, tels que :

- Les outils de codage :
 - ✓ Anaconda
 - ✓ VS Code
 - ✓ Jupyter Notebook
- Les outils de stockage :
 - ✓ GitHub
 - ✓ Google Drive
- Les outils de communication :
 - ✓ Zoom
 - ✓ Google meet
 - ✓ Slack

1.2 Le Projet Video Games

Le Projet sur l'analyse des ventes des jeux vidéo avait le plus haut niveau de difficulté proposés par DataScientest.

En effet, la grande difficulté de ce projet résidait dans les données fournies au départ qui ne contenaient qu'un nombre restreint d'informations pour la réalisation de ce projet.

Nous avions des données non à jour, car l'arrêté des chiffres de ventes totales de jeux, selon des régions du monde étaient en 2020. Même si nous avions les informations concernant les studios, le pays et les éditeurs, ceci n'allait pas être suffisant.

Il a fallu apprendre rapidement, peu de temps après le début de la formation, la méthodologie d'extraction de données sur les sites internet existants pour répondre aux questions suivantes :

Les avis, commentaires laissés par des joueurs sur internet, ont-ils un impact significatif sur l'évolution des ventes de la franchise Grand Theft Auto ?

Qu'est-ce qui fait de ce jeu une valeur sûre pour son éditeur ? Quels sont leurs caractéristiques ?
Comment les gamers se projettent-ils sur le prochain GTA VI prévu en 2025 ?

Comment a été accueilli son nouveau trailer sorti le 5 décembre dernier ?

Pourquoi cette franchise est devenue mythique ?

Autant de questions auxquelles nous répondrons tout au long de ce rapport dans l'axe du Storytelling plus que de celui de la prédiction. Prédire des ventes aurait été difficile, car il existe pléthore de circonstances externes qui participent à la réussite de la vente d'un produit, comme le montant alloué à la communication, le marketing, qui sont des données confidentielles à l'entreprise difficile à obtenir.

Avec l'ensemble des cours dispensés par DataScientest et de recherche personnelle,
Veuillez trouver ci-après, le détail et le résultat de nos travaux.

Nous restons à votre disposition pour des informations complémentaires

@Team.Project

#GTA#Missions#Parédelatêteaupied#MerciYaniv 😊



2. EXPLORATION DES DONNÉES

2.1 Exploration préliminaire des variables du jeu de données

Pour atteindre les objectifs de notre projet, nous avons focalisé notre analyse sur les jeux de la série Grand Theft Auto (GTA). Dans cette optique, nous avons utilisé trois jeux de données distincts :

- **Le jeu de donnée « vgsales »** disponible sur Kaggle. Ce jeu de données recueille des informations sur les ventes mondiales de jeu vidéo y compris les chiffres de vente par région (Amérique du Nord, Japon et Europe), les genres de jeux, les plateformes, les années de sortie et les éditeurs.
- **Les données descriptives** des jeux Grand Theft Auto (GTA) collectées à partir de différentes sources. Ces variables comprennent le nom du jeu, la durée de vie moyenne du jeu, son niveau de difficulté, le nombre de missions, la taille de la carte de jeu.
- **Les commentaires et avis** : nous avons extrait des commentaires et des avis des joueurs des jeux Grand Theft Auto/ Grand Theft Auto II / Grand Theft Auto III/ Grand Theft Auto IV/ Grand Theft Auto V à partir du site « Sens Critique », afin d'enrichir notre analyse avec des données qualitatives et des opinions d'utilisateurs.

Concernant la disponibilité des données, oui, elles sont librement accessibles. Le jeu de données « vgsales » est disponible publiquement sur Kaggle, où les utilisateurs peuvent le télécharger et l'utiliser à des fins d'analyse et de recherche. De plus, les commentaires et avis des joueurs sur GTA sont accessibles sur le site « Sens Critique », sans besoin d'adhésion spécifique. Cette accessibilité permet à des chercheurs ou à des passionnés de jeu vidéo de consulter et d'utiliser ces données pour leurs propres travaux.

En ce qui concerne la volumétrie de notre jeu de données, le jeu de données « vgsales » sur Kaggle contient plusieurs milliers de lignes de données, avec des informations détaillées sur des centaines de jeux vidéo, couvrant différentes régions et années.

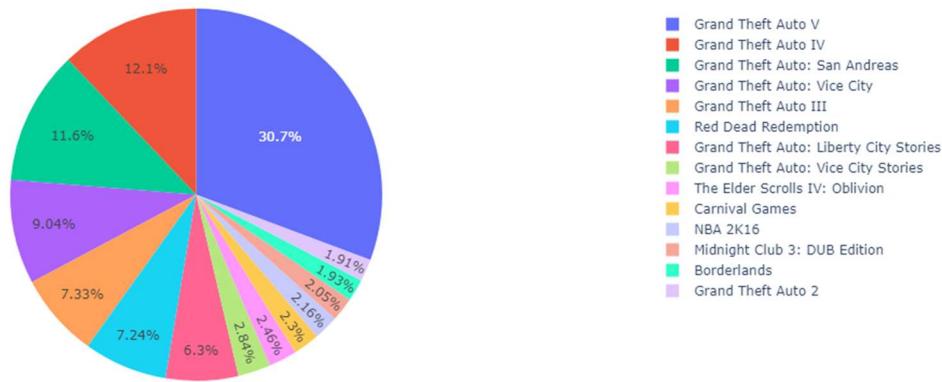
Les données extraites des commentaires et avis sur GTA peuvent varier en taille en fonction du nombre de commentaires collectés, ceci rajoute une dimension qualitative à notre analyse. Entre autres, notre jeu de données offre une volumétrie conséquente avec un vaste ensemble d'informations pour notre exploration et nos analyses.

2.2 Focus sur Take two Interactive

Dans notre analyse sur les différents GTA, plusieurs variables se distinguent comme étant pertinentes pour atteindre nos objectifs. Premièrement, les ventes mondiales de jeu vidéo forment une variable clé, avec des chiffres par région (Amérique du Nord, Japon et Europe). Ces données permettront de mesurer le succès financier de l'entreprise dans différentes parties du globe. Mais également d'identifier les préférences géographiques et de comprendre le marché pour Take-Two Interactive.

En ce qui concerne les jeux GTA, les variables telles que la durée de vie moyenne, le niveau de difficulté, le nombre de missions et la taille de la carte sont indispensables. Ces caractéristiques des GTA permettront d'analyser en profondeur les éléments du jeu, la complexité des missions, la variété des environnements, fournissant ainsi des connaissances précieuses sur les éléments qui ont contribué au succès de la franchise.

Part des ventes des 20 meilleurs jeux de "Take-Two Interactive"



La particularité de notre jeu de données réside dans le regroupement des informations à partir de différentes sources. Notre jeu de données « vgsales » offre une vue d'ensemble des ventes mondiales de jeux vidéo. Tandis que les données descriptives des GTA, extraites de diverses sources, fournissent des détails spécifiques sur les jeux emblématiques. Les commentaires et avis des joueurs extraits du site « Sens Critique » ajoutant une dimension qualitative à notre analyse. Cette combinaison de données quantitatives et qualitatives nous permet d'avoir une vision complète et nuancée des performances de Take-Two Interactive en mettant l'accent sur la franchise de GTA.

2.3 Axe de recherche complémentaire

Bien que nos jeux de données soient riches et variés, nous pourrions rencontrer certaines limites dans notre analyse.

En ce qui concerne les commentaires et avis des joueurs du site « Sens Critique », elles peuvent être sujets à des opinions individuelles qui pourraient ne pas représenter l'opinion générale des joueurs. La quantité de données recueillies peut varier d'un jeu à l'autre. Cela pourrait influencer la dureté de notre analyse qualitative.

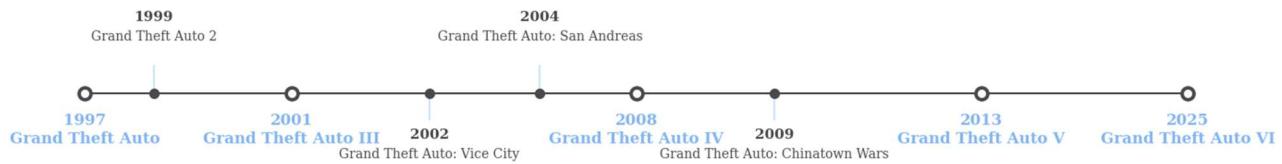
Malgré ces limites, nous nous attacherons à utiliser au mieux nos données et de combler les lacunes par des analyses complémentaires. De ce fait, nous veillerons à interpréter nos résultats de manière judicieuse et tirer des conclusions qui reflètent aux mieux la réalité du marché des jeux GTA



2.4 Grand Theft Auto : en quelque mots

Grand Theft Auto au fil du temps

De la 2D à la 3D - du jeu à une anthologie



- **Grand Theft Auto (1997)** : Le jeu avec lequel tout a commencé, avec un environnement en monde ouvert et des activités criminelles.
- **Grand Theft Auto 2 (1999)** : Une suite se déroulant dans une ville futuriste avec des graphismes et un gameplay améliorés
- **Grand Theft Auto III (2001)** : Un jeu révolutionnaire en 3D qui a introduit Liberty City et a bouleversé le jeu en monde ouvert
- **Grand Theft Auto Vice City (2002)** : Un voyage nostalgique dans les années 80, avec un cadre inspiré de Miami et une Intrigue Captivante
- **Grand Theft Auto San Andreas (2004)** : Situé dans l'État fictive de San Andreas, il offrait un vaste monde ouvert et un gameplay varié.
- **Grand Theft Auto IV (2008)** : Situé à Liberty City, ce jeu suivait l'Histoire de Niko Bellic, un immigrant en quête d'une vie meilleure
- **Grand Theft Auto Chinatown Wars (2009)** : Un jeu portable (Nintendo DS exclusivement à la sortie), se déroulant à Liberty City, axé sur le trafic de drogue et la criminalité
- **Grand Theft Auto V (2013)** : L'entrée la plus réussie de la série avec cette fois, 3 protagonistes et la ville fictive de Los Santos

Ces jeux ont laissé une empreinte durable dans l'industrie du jeu vidéo, alliant action, narration et exploration. Que l'on soit fan des classiques ou des titres modernes, la Série Grand Theft Auto continue de captiver les joueurs et joueuses du monde entier.





3. WEBCSRAPPING

Dans cette partie, il a fallu scrapper les données sur le Web des différents jeux GTA. Pour rappel, notre étude se concentre sur les franchises principales de GTA afin de suivre une logique directrice et afin de réaliser un meilleur comparatif. L'objectif dans cette partie a été d'ajouter des informations en plus à notre DataFrame de départ. En effet, il manquait sur notre DataFrame de départ un nombre considérable d'informations, ce qui aurait abouti à une conclusion peu parlante. Par exemple scrapper les commentaires et les notes des jeux sur des sites connues tel que 'Sens Critique' nous a permis d'effectuer de meilleures analyses des franchises principales et donc d'en ressortir de précieuses informations

3.1. Méthodologie

Dans un premier temps, nous avons scrappé les notes de jeux sur le site 'Sens Critique'. Nous avons utilisé en tant que première méthode BeautifulSoup. Malheureusement, le code source HTML qui a été importé est différent du code source du site. Il nous a été impossible de récupérer les liens href menant aux jeux GTA. L'hypothèse est qu'une protection anti Scrapping a été mise en place. De plus, récupérer les notes des jeux a été la deuxième difficulté. En effet, les notes étaient présentes dans un diagramme interactif et donc impossible à récupérer en utilisant BeautifulSoup. Cependant, BeautifulSoup s'est avéré très efficace pour récupérer les titres, note test, note avis, description et date de sortie des jeux du site [jeuxvidéos.com](https://www.jeuxvideo.com/). Concernant le site [Senscritique](https://www.senscritique.com/), une autre approche a été nécessaire. La méthode Sélénum nous a permis de récupérer le reste des informations.

L'objectif de l'utilisation de Sélénum sur le site Sens Critique est de scrapper les commentaires des différents jeux GTA afin de pouvoir effectuer un sentiment analyses par la suite.

- Pour récapituler la démarche effectuée :

1. Premier Scrapping :

Scrapper les informations (Titre du jeu, Note test des testeurs du jeu, Note des avis des joueurs, Description du jeu, Sortie du jeu) de tous les jeux sur le site [jeuvideo.com](https://www.jeuxvideo.com) à l'aide de BeautifulSoup

Deuxième Scrapping :

Scrapper les informations (les notes des jeux GTA, les commentaires, les dates des commentaires, le nombre de fois que ces commentaires ont été lus) des jeux GTA sur le site Sens critique à l'aide de Sélénum

L'essentiel de la structure du code

1.Premier Scrapping :

- Faire appel au driver menant à l'url
- Récupérer les textes se trouvant dans les balises
- Coder la pagination pour aller aux pages suivantes
- Refaire la même démarche quand le 2ème tiret (une boucle for a été nécessaire)

2.Deuxième Scrapping :

- Faire appel au driver menant à l'url où se trouve tous les jeux GTA
- Récupérer tous les liens des jeux de GTA
- Aller sur un lien qui mène vers un jeu GTA
- Récupérer les notes du jeu
- Simuler le Click() sur l'interface qui mène ensuite vers tous les commentaires du jeu
- Récupérer dans les balises les textes voulus
- Aller à la page suivante et reproduire la même démarche que précédemment
- Reproduire la même démarche jusqu'à la page finale pour ainsi avoir tous les commentaires et autres informations du jeu
- Refaire la même méthode pour chacun des liens récupérer

3.2 Données retenus

Sur le site jeuvideo.com les données retenues ont été seulement les descriptions des jeux, car il y avait beaucoup de valeurs manquantes sur les autres informations voulues.

Sur le site Sens Critique les données retenues ont été les notes des jeux GTA ainsi que les commentaires et la date des commentaires.

3.3Autres données retenues sur internet

Des variables descriptives ont été trouvées sur internet pour l'ensemble des GTA principaux. Ces variables comprennent le nom du jeu, la durée de vie moyenne du jeu, son niveau de difficulté, le nombre de missions, la taille de la carte de jeu. De plus, des données sur l'évolution des ventes ont été aussi récupérées.

4.DATASET FINAL DESCRIPTION

4.1 Dataset toutes variable hors "commentaires"

4.1.1 Le Dataset "vgsales"

Il comporte plusieurs variables nécessaires à notre étude.

Après une étude rapide des données, nous avons utilisé les variables suivantes pour trouver notre sujet principal :

- **Name** : Le nom des jeux (*11493 valeurs unique*)
- **Publisher** : Le nom de l'éditeur du jeu (*578 valeurs unique*)
- **Global Sales** : Le nombre de vente Global en million (*Somme de [NA_Sales, EU_Sales, JP_Sales, Other_Sales]*)

Ce jeu de données avait très peu de valeur manquantes (329/16598), une suppression simple a été réalisée.

4.1.2 Le Dataset " GTA_officiel_concatened"

Il est composé de données du Dataset "vgsales" et des notations de joueurs, scrappées sur le site "sens critique.com"

- **Nom** : Le nom du jeu
- **Annee** : L'année de sortie du jeu
- **Genre_x** : La typologie du jeu
- **Developpeur** : Le nom du développeur
- **Date** : La date complète de la sortie du jeu
- **Console** : Les consoles de jeu sur lesquels le jeu est disponible
- **Note_Site** : La note moyenne calculée donné par les joueurs
- **Nb_note_1 à Nb_note_10** : Le nombre de votes par notation de 1 à 10 (*10 étant la plus élevée*)
- **Link** : Le lien du site internet qui a permis le Scrapping
- **Moyenne des ventes estimée (en million)** : Le nombre moyen des ventes en million en 2023 (*Scrappé depuis Wikipédia*)
- **Platform** : La console sur lequel le jeu est disponible
- **Publisher** : Le nom de l'éditeur
- **NA_Sales** : Total des ventes en Amérique du Nord en million à 2020
- **EU_Sales** : Total des ventes en Europe en million à 2020
- **JP_Sales** : Total des ventes au Japon en million à 2020
- **Other_Sales** : Total des ventes dans d'autres régions en million en 2020
- **Global_Sales** : Total des ventes en Amérique du Nord, Europe, Japon et autres régions en million à 2020

4.1.3 Le Dataset " GTA_Description"

Il est composé de données issue d'informations provenant d'internet. Il n'existe pas un site spécifique où trouver ce type de donnée.

Les informations sont cohérentes avec ce que nous savons de GTA, nous utilisons donc ces données dans la poursuite de nos travaux de visualisation.

- Nom du jeu : Le nom du jeu
- Nb Heure : Le nombre d'heure estimée de temps de jeu
- Mode de jeu : Le mode de jeu, seul ou à plusieurs
- Style de jeu : La vue caméra du jeu
- Gameplay : Les actions possibles dans le jeu
- Nombre de missions : Le nombre de missions à faire pour finir le jeu
- Nombre de personnages jouables : Le nombre de personnage à incarner
- Nom des personnages : Le nom des personnages à incarner
- Environnement : Le lieu où se déroule le jeu
- Musique : le nombre de musique disponible
- Dimensions : Les dimensions du jeu
- Pixel : La qualité de graphisme
- Taille de Map : La taille de la carte dans laquelle le personnage peut se déplacer

4.1.4 Le Dataset " GTA_Cumul des ventes"

Il est composé de données issue d'informations provenant d'internet. Il n'existe pas un site spécifique où trouver ce type de donnée. Les sources ont été diverses. (Breakflip, jeux.video, capital, les echos, Rockstarmag, sens critique)

Les informations sont cohérentes avec ce que nous savons de GTA, nous utilisons donc ces données dans la poursuite de nos travaux de visualisation.

- Nom : Le nom du jeu
- Année de sortie : L'année de sortie du jeu
- Vente 1997 à Vente 2023 : Cumul d'exemplaires vendus pour chaque jeu principale de la franchise de 1997 à 2023 (en millions)

4.2 Dataset variable "commentaires"

Le Dataset est réparti sur plusieurs fichiers csv. Chaque fichier csv correspond à un GTA. Dans le fichier csv est compris le DataFrame avec le nom du jeu, les commentaires du jeu, les dates des commentaires qui ont été au préalable convertis au format Date et ainsi que le nombre de fois que les commentaires ont été lues et dont la colonne a été convertie en int.

La difficulté dans le DataFrame est de faire comprendre au code que le format date scrappé n'était pas dans le format adéquat.

Sa conversion a été rude. Concernant le nombre de fois que les commentaires ont été lus, le format n'était pas adapté pour convertir la colonne en int.

Il fallait récupérer que la partie chiffrée du texte et ensuite multiplié par 1000 ceux qui possédaient la lettre K.



5. ANALYSE DU SENTIMENT

Après le Scrapping, nous avons analysé les commentaires, pour cela nous avons créé en premier un Word cloud affiche de façon design les mots les plus employés dans les commentaires

5.1 Text Mining

Un prétraitement a été nécessaire afin de nettoyer et de rendre les données utilisables pour l'analyse de sentiment

Nous avons utilisé la tokenisation, en réalisant un découpage des phrases en mots afin de faciliter l'automatisation du traitement de la donnée.

Pour rendre l'analyse plus efficace, l'utilisation d'un Stop Word a été nécessaire. Les pronoms personnels, par exemple ont été supprimés.

Après avoir terminé ce « pre-processing » de texte, vient enfin le moment de l'analyse de données. On utilise alors différents algorithmes de Text Mining pour dégager des informations dans un Worcloud.

5.2 Word cloud

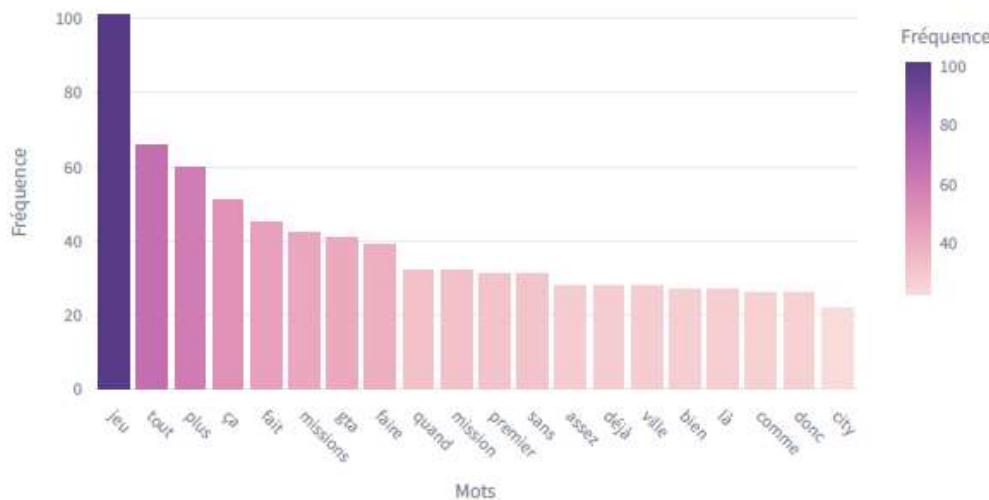


Ce Word cloud permet de visualiser rapidement les mots les plus couramment utilisés dans les commentaires.

Plus le mot apparaît en grand et en premier plan, cela signifie qui fait partie des mots le plus souvent cités. Ici nous voyons que les mots les plus importants sont : « Mission », « Ville », « GTA », « Jeu », « voiture », « Chapitre. »

Ensuite, nous affichons un histogramme représentant les 20 mots les plus couramment employés. Cela nous permet de classer les mots et de connaître leurs nombres d'apparition plus précisément.

20 mots les plus fréquemment employés par les internautes pour 30 commentaires Grand Theft Au



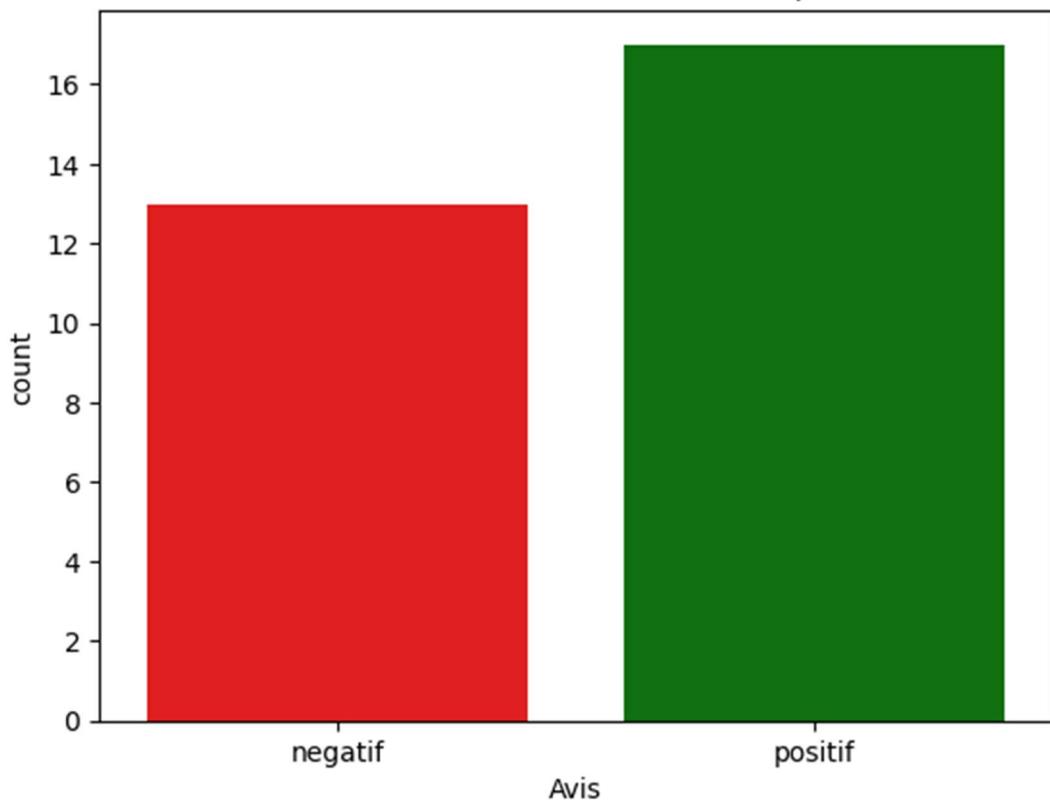
5.3 Sentiment Analysis avec "vader"

Pour pousser l'analyse un peu plus loin, nous avons utilisé Vader. L'objectif est de comprendre qu'elles sont les avis des joueurs, s'ils ont globalement aimé, détesté ou non le jeu en question.

Vader est un analyseur lexical utilisé pour attribuer des scores de sentiment à une phrase. Il analyse chaque mot et attribut une note si c'est positif, négatif ou neutre et une moyenne.

Après avoir analysé tous les commentaires avec Vader, nous avons trié les commentaires en fonction de moyenne en trois cas pour chaque jeu :

Nombre de commentaire en fonction du sentiment pour Grand Theft Auto II



Les limites rencontrées avec cette librairie :

Vader possède un dictionnaire par mot qui ne permet pas la compréhension des expressions issues des commentaires que nous avons récoltés.

Il va, par exemple, considérer l'expression « super nul » comme étant positif en faisant une moyenne entre « super » et « nul » en considérant le mot « super » comme plus fort que « nul ».

De même, qu'il va considérer des mots comme « vol », « guerre » ou encore des superlatifs en langage très familier comme négatif, alors qu'il s'agit de commentaire positif dans sa globalité.

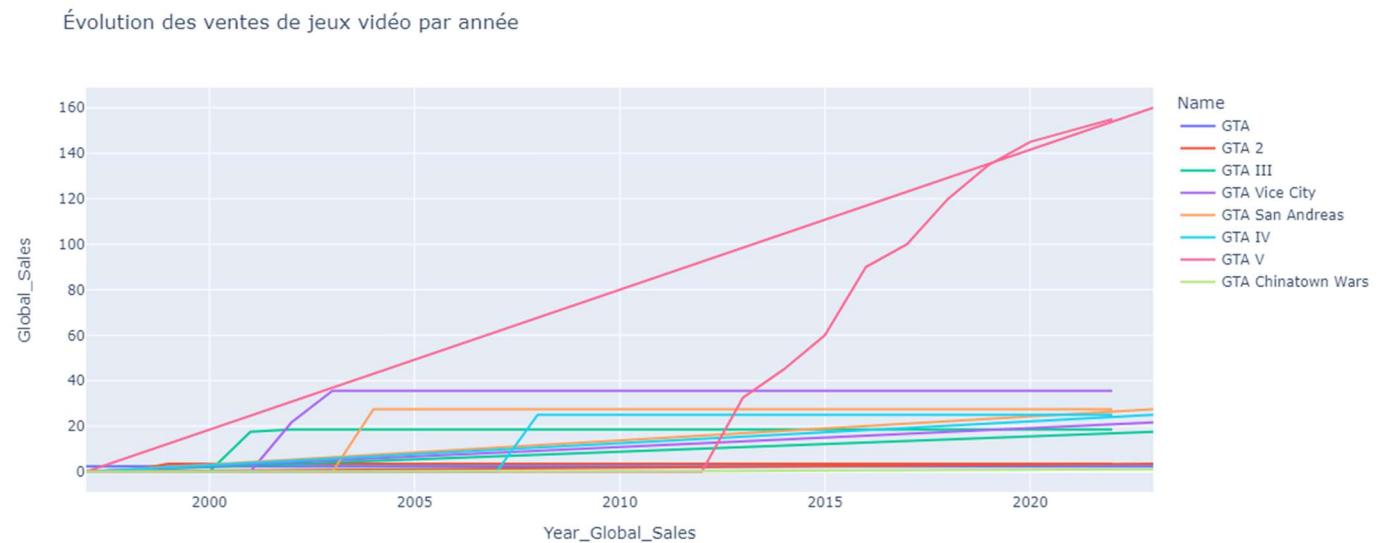
Même en affinant le code, il serait difficile d'arriver à une précision totale, car ce jeu est en lui-même apprécié pour sa violence et son langage très familier.

6. DATA VISUALISATION

Dans cette partie de Data visualisation, nous verrons l'évolution des ventes de chaque jeu de la série de 1997 à 2023. Nous verrons également la description des jeux et les notes qu'ont mis les joueurs.

Pourrons-nous établir des corrélations entre elles ?

6.1 Évolution des ventes des GTA à aujourd'hui



Dans le graphique en ligne ci-dessous, nous visualisons que pour chaque jeu sauf GTA V, le cumul des ventes de 1997 à 2023 reste linéaire. Le pic pour chacun de ces jeux représente le nombre d'unité vendues en millions, lors de sa sortie pour ensuite ne pas continuer à se vendre. Ce qui paraît logique pour un jeu ordinaire, pour lequel ont fini le jeu et on ne le rachète pas. (Ne sont pas compris dans ces chiffres le marché de l'occasion)

Cependant, GTA V, ne poursuit pas cette logique, car même après sa sortie en 2014, celui-ci continue de se vendre encore 10 ans plus tard. Le commerce de ce jeu représente 32 millions d'unités vendues en 2014 à 160 environs, selon les sources en 2023 et continue de s'accroître.

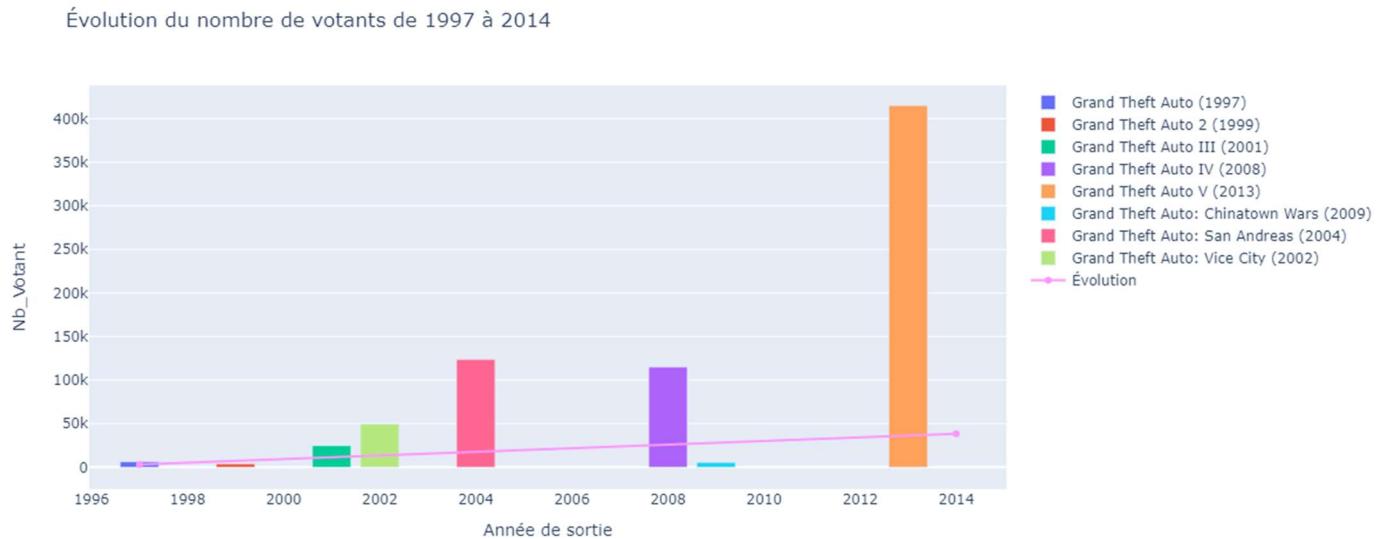
La Franchise Grand Theft Auto, bat des records, car c'est la deuxième franchise la plus vendus dans le monde juste après Minecraft.

6.2 Évolution des avis au fil du temps

La figure ci-dessous, montre de manière générale que le nombre de votants évolue de manière croissante au fil des années. On peut penser que la démocratisation de l'accès à internet a engendré ce résultat.

Cependant, on constate, que pour GTA Chinatown Wars le nombre de votant est plus petit que GTA San Andreas sortie 5 ans auparavant.

GTA V est véritablement au-dessus de tous les jeux avec 400k de nombre de votants

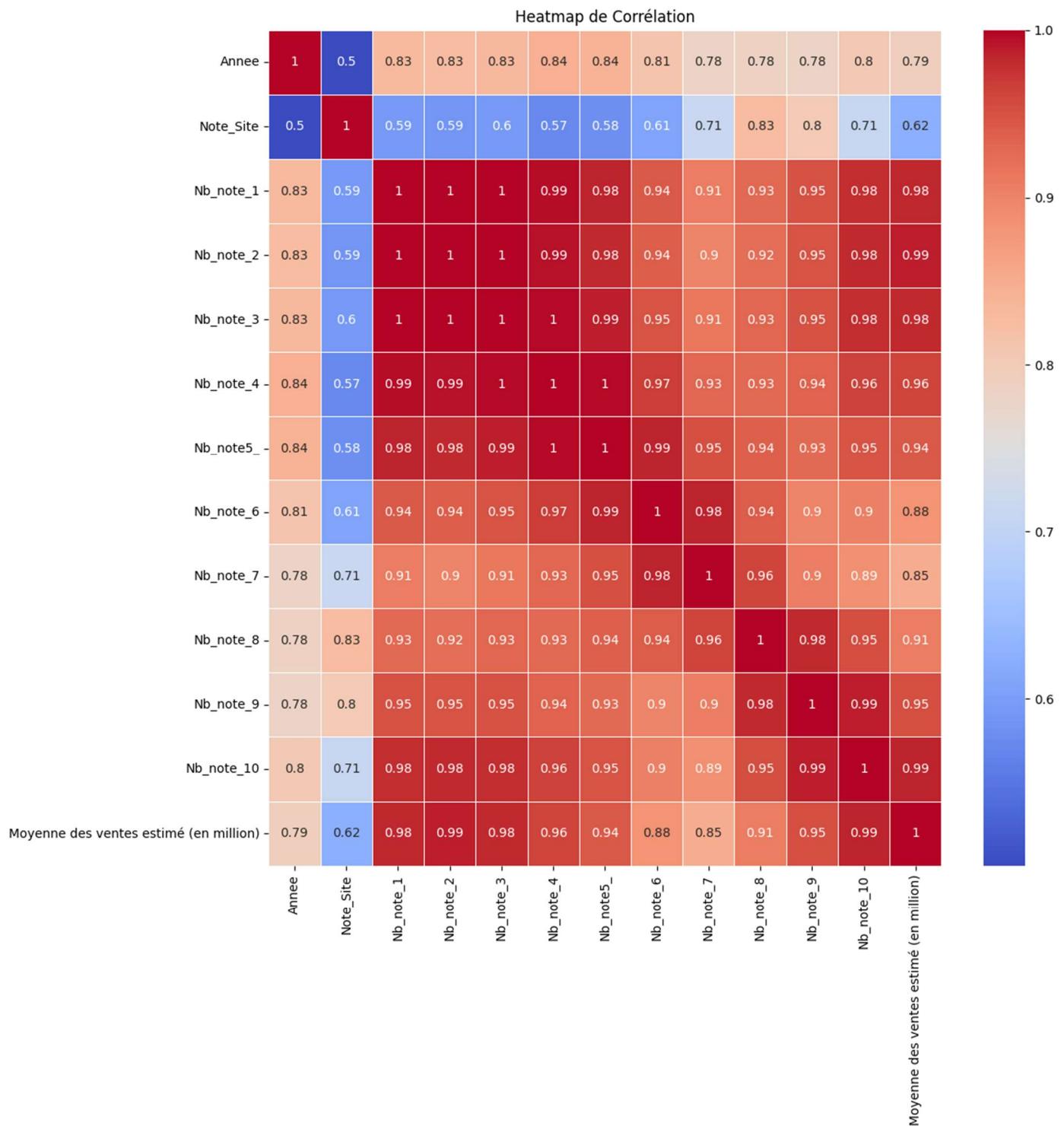


Grâce au Pie Chart suivants, nous constatons bien, que chaque jeu n'ont pas toujours fait l'unanimité. Même si au cumul, la note moyenne est comprise entre 7 et 8 pour l'ensemble des jeux. GTA V et Chinatown Wars sort du lot, pas pour les mêmes raisons. GTA V ressort avec une majorité de note à 10 à 23% (best of the best) tandis que GTA Chinatown Wars ressort avec une note de 6 pour 18.4% des votants.



6.3 Corrélation entre les variables

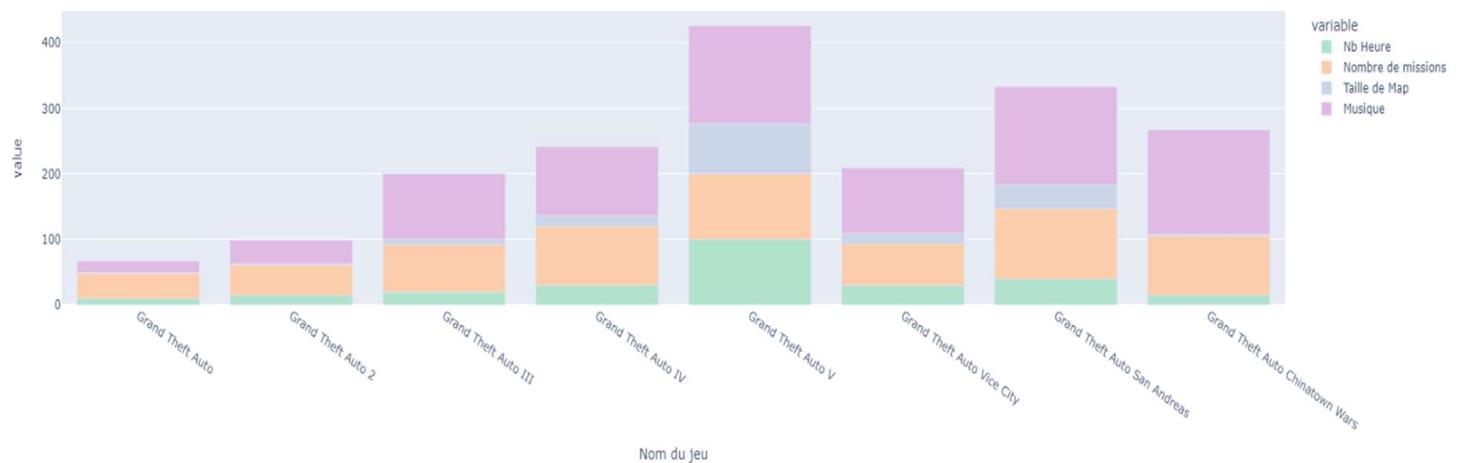
Cette Heatmap de corrélation, nous permet de mettre en évidence une corrélation très forte entre les notations des joueurs et la moyenne des ventes.



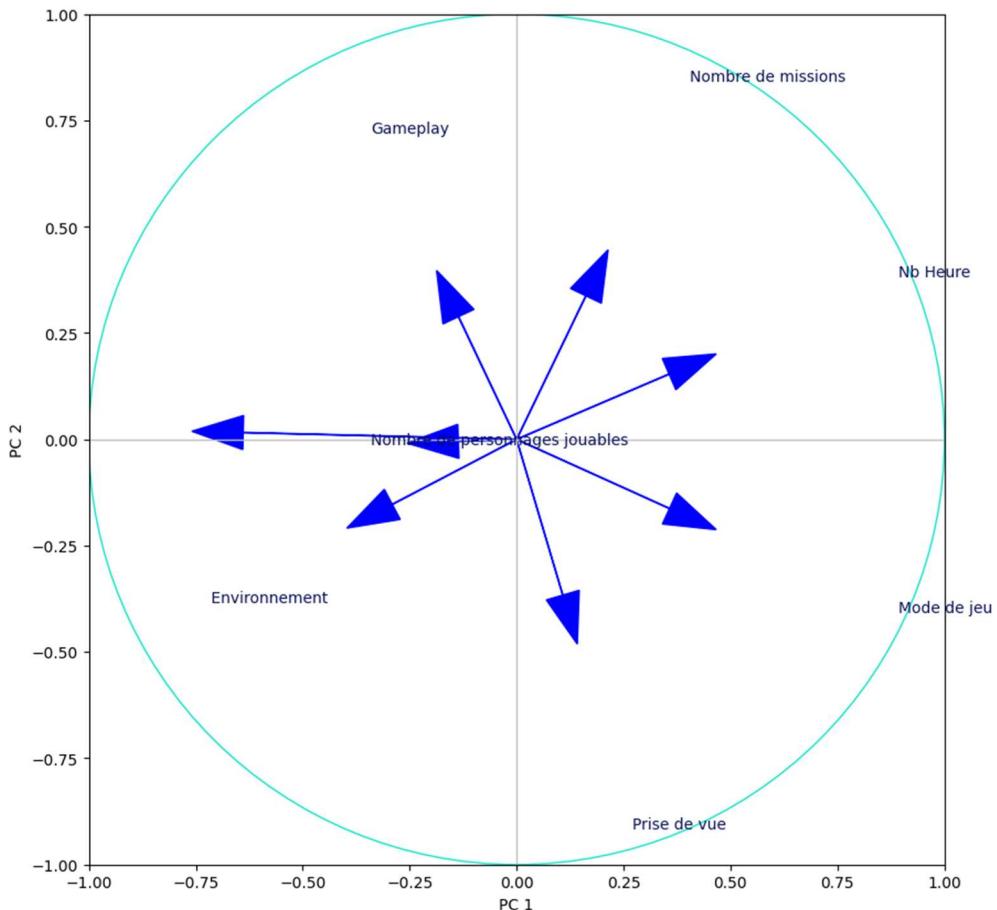
Avec ce graphique sur les caractéristiques, nous remarquons rapidement les différences entre les différents jeux en termes de taille de map, la durée de jeu possible, le nombre de missions et notamment le nombre de Sound Track disponible dans le jeu. On constate que GTA V et GTA San Andreas ont finalement qu'une seule différence dans ce graphique est la taille de la carte de jeu. (GTA San Andreas : 36km2 versus GTA V : 75.8 km2)

Grand Theft Auto V est le seul de la franchise à être multi-joueur.

Caractéristiques des jeux Grand Theft Auto (GTA)



Grâce à ce cercle de corrélation, nous constatons que la taille de la map et la durée de jeu et le nombre de missions ont une influence significative sur la variable cible qui est la vente.
Une standardisation des données a été nécessaire.



Pour conclure sur la visualisation, les notes des joueurs et certaine caractéristique du jeu influencent significativement le nombre d'unités vendues.

Avis personnel : il est possible que rockstar joue sur ces points sur son prochain GTA car, selon les rumeurs, la taille de la map dépasserai les 100km² et un nouveau personnage cette fois-ci féminin verrai le jour.

La taille de la map, le nombre de personnage incarné, ainsi que la durée de jeu, sont des éléments importants pour les joueurs. Cela influence leurs avis comme on a pu le voir dans l'analyse de sentiment.

Nous n'avons pas suffisamment de données pour faire de la modélisation de prédition de vente, comme nous allons l'expliquer dans la prochaine partie.

7. MODELISATION

7.1 Modèle de prédition

Il n'existe quasiment aucune donnée sur GTA VI. Il nous est donc impossible de prédire de manière réaliste le nombre de ventes du jeu. Les modèles de prédictions s'utilisent la plupart du temps (plutôt toujours) à partir d'une base de données solides et assez vaste. De plus, il existe un nombre incalculable de facteurs qui influencent le nombre de ventes d'un jeu. De plus, l'absence de données rend impossible la tâche.

7.2 Autre supposition

Il existe une seule information concernant GTA VI. C'est son trailer publié sur YouTube. Ainsi le plus judicieux à faire est donc de scrapper les réactions du trailer sur YouTube et de réaliser un sentiment Analysis qui permettrait d'avoir une idée concernant l'engouement

7.3 Méthodologie du Scrapping

Sur YouTube il faut défiler vers le bas de la page pour pouvoir charger les autres commentaires. Ainsi BeautifulSoup n'est pas la méthode appropriée à notre problème. Il fallait donc utiliser Sélénum encore une fois.

Cette fois-ci dans le code Sélénum la fonction PAGE.DOWN a été le plus importante. C'est celle-ci qui permet de charger les autres commentaires de YouTube.



Une fois tous les commentaires récupérés et le DataFrame construit il faut filtrer les commentaires dont la langue est différente de l'anglais.

En effet, la plupart des commentaires sont en anglais. Garder les commentaires qui sont dans une autre langue fausserait le sentiment Analysis qui va être analysé. C'est pourquoi la fonction « detect » de la librairie « langdetect » a été nécessaire durant cette étape.

CONCLUSION

Projet/Résultat

Ce jeu nous mettant le joueur dans la peau d'un criminel remplissant diverses missions aux quatre coins des villes où se déroulent le jeu.

Depuis 10 ans GTA V continue à se vendre dans le mode multijoueur en ligne. Les joueurs dépensent de l'argent réel contre des objets dans jeu.

De par la liberté que rencontre les joueurs, la violence exprimée ouvertement, les possibilités d'exploration de la carte, plus d'une centaine de missions qui font de cette série une série culte. Il génère un engouement important avec plus de 400 millions d'unités d'exemplaires vendu dans le monde.

Un attendu du nouveau opus très fort, avec plus de 90 millions de vues, en une journée, de la Bande annonce officiel de GTA VI (BFM tech & Co le 06/12/2023)

Perspective :

Il serait intéressant de pouvoir faire une comparaison avec d'autre série de jeu d'aventure, pour aller plus loin sur l'aspect psychologique des jeux.

De découvrir qui sont les joueurs, quels sont leurs profils ?

Quelles sont leurs comportements d'achat ?

La catégorie socioprofessionnelle a-t-elle un impact sur le style de jeu acheté ?

Le genre du joueur a-t-il un impact ? Est-ce vraiment deux marchés différents en 2024 ?

Difficultés rencontrées lors du projet

Les principales difficultés rencontrées ont été celle du webscraping et l'utilisation de Vader. De par les multiples essais avec la librairie BeautifulSoup sans comprendre pourquoi le Scrapping ne fonctionnait pas, pour ensuite apprendre et utiliser Sélénum.

La plus grande partie du webscraping c'est terminé fin janvier. Ce n'est qu'à partir de là que nous nous sommes rendus à la suite des étapes. Notre mentor, nous a fait remarquer qu'il nous manquait encore des données. Nous avons continué à alimenter nos tables quasiment jusqu'au 13 février.

Lisez ci-dessous le détail des difficultés

Premier Scrapping :

- ✖ La pagination devait être pris en compte dans le code pour récupérer les informations des jeux dans les autres pages
- ✖ Dans certains jeux, des informations n'étaient pas présentes.
- ✖ Il fallait donc penser à l'appel du Try/Except afin d'éviter que le code s'arrête sur une erreur, car l'information est introuvable

Deuxième Scrapping :

- ✖ Penser à coder l'attente pour attendre que les cookies apparaissent dans la page et les accepter
- ✖ Les ' href ' n'étaient pas en tant que 'text' entre les crochets. Il fallait donc penser à faire appel à la fonction get_attribute pour récupérer les liens qui sont écrit au sein d'une classe
- ✖ La balise qui mène vers les commentaires du jeu est exactement la même que celle qui mène vers les images du jeu. Ce sont des balises jumelles. Il fallait donc penser à simuler le bon Click().
- ✖ Les commentaires sont répartis sur plusieurs pages et il faut cliquer sur chacun des commentaires si on veut les voir en entier. Il fallait donc penser à définir une fonction qui récupère à la fois tous les liens qui mènent chacun vers un commentaire entier et ensuite récupère tout le commentaire et autres informations se trouvant à côté.
- ✖ Pour aller au commentaire suivant, il faut faire un retour en arrière. Cependant, le retour en arrière mène à chaque fois vers la page 1 des commentaires (hypothétiquement un bug du site). C'est pourquoi une deuxième fonction a été définie qui consiste à cliquer sur chaque page de la pagination pour finalement revenir à l'étape du Scrapping du commentaire où on s'était arrêté précédemment.
- ✖ Chaque retour arrière mène vers la page 1 des commentaires. C'est pourquoi il a été plus judicieux de récupérer à chaque page tous les liens menant vers un commentaire. De cette façon, cela évite de scrapper les mêmes commentaires. Cela a été un moyen efficace pour contourner la problématique liée au bug du site.
- ✖ Les timesleep ont été préconisés par rapport aux fonctions WaitUntil, car les balises des pages suivantes chargent et remplacent celles des pages précédentes. Le WaitUntil était incapable d'identifier cela ce qui avait pour conséquence de scrapper les mêmes commentaires.
- ✖ Le manque de timesleep provoquait des erreurs dans le code, car les pages n'avaient pas terminé de charger.
- ✖ La structure du code source HTML sur chacun des jeux GTA n'est pas forcément toujours la même. C'est pourquoi il a été nécessaire de réajuster le code en fonction du jeu GTA qu'on scrappe.

Et si c'était à refaire ?

En effet, il est toujours possible de faire mieux après coup.

Notamment en fixant une problématique claire et précise de ce que nous souhaitons démontrer dès le départ et pas l'affiner au fil des trouvailles ou des blocages.

Cela aurait pu nous faire gagner un peu de temps (une semaine peut être)

Avoir une séparation des tâches entre nous, plus efficace, nous pourrions se dire.

En vérité, nous avons fait de notre mieux tout au long de ce projet, en conciliant la formation et les connaissances à acquérir et à mettre en pratique. En se confrontant aux différentes problématiques en autonomie et surtout le plus important, la réussite de trouver les solutions, par nous-même.

Le projet n'aurait pas vu le jour si le webscrapping n'avait pas été terminé, ce qui a été difficile avec Selenium. Etant donnée que les sites internet se protège de plus en plus. C'est à partir de ce gros chantier, que nous avons pu voir ce qui nous manquait comme éléments pour poursuivre et faire un plan de nos actions à faire respectifs.

La mise en commun de toutes nos compétences techniques et transverses ont permis de vous livrer ce dossier.

Nous remercions @Yaniv pour son accompagnement et @DataScientest pour nous avoir permis de développer toutes ces compétences techniques qui nous serons très utiles pour notre prochain métier d'avenir.

