Animal Crossing - New Horizons

Reda Sarehane, Ismail Kadiri

28 avril 2022

"Animal Crossing - New Horizon" est un jeu vidéo de simulation de vie développé et édité par Nintendo sorti le 20 mars 2021. Nous avons choisi d'analyser les données tirées d'une base de données open-source développée par des fans contenant plusieurs données sur ce jeu.

1. Introduction

Le jeu de données est constitué de quatre tables de données : "critic.tsv", "user_reviews.tsv", "villagers.csv" et "items.csv". Nous avons choisi d'ajouter un jeu de données complémentaire : "acnh_villager_data.csv" pour nous aider dans notre analyse. Nous détaillerons ce choix dans une prochaine partie de ce rapport.

La table "critic.tsv" contient les critiques données par des professionnels. Ce jeu de données est constitué de 4 variables et 107 individus. Les variables donnent des informations sur la note donnée par les critiques, l'origine de la critique, un commentaire accompagnant la critique et la date de publication.

La table "user_reviews.tsv" contient les critiques données par des joueurs. Ce jeu de données est constitué de 4 variables et 2999 individus. Les variables donnent des informations sur la note donnée par les joueurs, le nom d'utilisateur du joueur, un commentaire accompagnant la critique et la date de publication.

La table "villagers.csv" contient des données sur les personnages non-jouables du jeu, les villageois. Ce jeu de données est constitué de 11 variables et 391 individus. Les variables donnent des informations sur les attributs de chaque personnage comme leur id, nom, genre ou encore leur personnalité. Certaines variables ne sont pas réellement utiles pour notre analyse, à l'instar du full id ou de l'url de l'illustration du villageois.

La table "items.csv" contient des données sur les objets du jeu. Ce jeu de données est constitué de 16 variables et 4565 individus. Les variables donnent des informations sur les attributs de chaque objet du jeu vidéo comme son id, nom, catégorie, prix de vente ou encore leur prix d'achat. Certaines variables sont également inutiles dans ce dataframe, par exemple le game-id ou l'url de l'illustration de l'item.

Compte tenu du jeu de données nous avons plusieurs analyses possibles. D'un côté, nous pouvons effectuer une analyse démographique des villageois. En essayant, par exemple, de prédire le villageois en se basant sur quelques variables de base. D'un autre côté, nous pouvons effectuer une analyse des retours sur le jeu. Nous pouvons, par exemple, essayer de prédire la note en fonction de leur commentaire.

Nous expliquerons dans le reste de ce rapport les démarches que l'on a suivies pour choisir la direction que nous comptons prendre pour nos analyses.

2. Jeu de donnée complémentaire

Afin d'ajouter plus de pertinence à l'analyse, nous avons décidé d'ajouter une information supplémentaire du ranking et du tier de chaque villageois. Ce jeu de donnée complémentaire provient

d'un utilisateur Kaggle qui a scrappé les données de rank et de tier à partir https://www.animalcrossingportal.com/games/new-horizons/guides/villager-popularity-list.php#/ pour tous les villageois du jeu. Il est important de noter que le rang et le tier sont des informations subjectives puisqu'elles sont déterminées par votes des utilisateurs du site. On s'est inspiré de son analyse https://www.kaggle.com/datasets/ampiiere/acnh-villager-popularity pour intégrer les données et pour les différentes idées d'analyses possibles.

3. Analyse exploratoire

1. Analyse des villageois

Nous nous sommes limités tout d'abord uniquement au jeu de données des villageois pour réaliser nos premières visualisations. Les features intéressantes propices à l'analyse sont le sexe, l'espèce et la personnalité de chaque villageois. On a pu ainsi de prime abord analyser ces 3 aspects:

- L'égalité des sexes est quasiment respectée ici (cf. Figure 5) avec légèrement plus de villageois mâles.
- Parmi les espèces les plus représentées on retrouve les chats, lapins, écureuils et de grenouilles (cf. Figure 6). Cependant, l'égalité des sexes n'est pas représentative quand on analyse les espèces (on a plus de chats femelles que mâles par exemple). Certaines espèces sont même uniquement représentées par un seul sexe (notamment les taureaux et lions mâles et les vaches femelles).
- Les personnalités sont-elles inérantes à chaque sexe. Les villageois mâles sont soit smug (chic), cranky (grincheux), jock (sportifs), ou lazy (paresseux) alors que les villageois femelles sont uchi (grande soeur), peppy (vives), snooty (arrogante) ou normal (normal). Une description des personnalités peut être trouvée sur https://animalcrossing.fandom.com/fr/wiki/Personnalité#Grande_s.C5.93ur.

2. Analyse de la popularité

Grâce au jeu de données supplémentaire concernant le rang et le tier des villageois, de nouveaux insights peuvent être déterminés conjointement aux données initiales sur les villageois. Nous pouvons visualiser les features utilisées précédemment en fonction du tier et/ou du rang. Nous obtenons les conclusions suivantes:

- En moyenne les villageois mâles sont légèrement moins bien classés par rapport aux villageois femelles (cf. Figure 8). On rappelle que plus un villageois a un petit rang, plus il est considéré comme populaire.
- Les espèces les plus populaires (cf. Figure 9) sont les octopus (pieuvre), les deer (cerf), les wolf (loup) et les cat (chat). Les moins populaires sont les pig (cochon) et les gorillas (gorilles).
- Les personnalités les plus populaires (cf. Figure 10) sont les uchi (grande soeur) et les normal (normal). Les moins populaires sont les snooty (arrogantes) et les cranky (grincheux), ce qui semble logique de prime abord.

3. Analyse des objets

Les analyses quantitatives sur les objets sont limitées. Nous avons analysé leurs valeur de vente et d'achat et on en a déduit quelques informations intéressantes. Il faut noter que les valeurs de vente et d'achat ont été transformé par la fonction logarithmique afin de pouvoir visualiser les différentes courbes:

- La valeur de vente des objets (transformée par la fonction log) peut-être approximée par une loi normale (cf. Figure 11). La médiane du prix des objets est de 390 clochettes (les clochettes étant la monnaie du jeu). 75% des objets du jeu ne dépassent pas 1000 clochettes. L'objet le plus cher coûte 300 000 clochettes (l'objet est la couronne royale) et le moins cher 5 clochettes (une branche d'arbre).
- La relation liant la valeur d'achat et de vente est linéaire (cf. Figure 12): la valeur d'achat d'un produit est 4 fois sa valeur de vente.
- Une grande majorité des objets sont soit des meubles et des photos (cf. Figure 13). Les catégories contenant le moins d'objets différents sont les fruits et les coquillages

4. Analyse des critiques

Afin d'analyser les critiques des joueurs et des médias nous avons appliqué des méthodes NLP (natural language processing), afin d'analyser des données textuelles. Pour ce faire nous avons utilisé la bibliothèque NLTK, natural language toolkit. Cette bibliothèque nous donne accès à une série de fonctions utiles à l'analyse et au traitement du langage naturel.

Nous avons d'abord fait une phase de data pré-processing afin de récupérer les informations clés du jeu de données. En premier lieu, nous avons vérifié la répartition des notes attribuées par les joueurs et les critiques. Nous pouvons alors remarquer la disparité des notes attribuées par les joueurs et les critiques. Pour les joueurs nous remarquons qu'il y a beaucoup plus de 0 et de 10 que le reste, alors que pour les critiques il y a une majorité de 90% (cf. Figure 1 en annexe).

En second lieu, nous avons essayé de tirer les mots les plus utilisés pour chaque note donnée. Pour ce faire, nous avons utilisé le principe de tokenization pour récupérer les mots utilisés dans chaque commentaire et nous avons analysé leur répartition. Nous pouvons remarquer la fréquence d'apparition des mots "one", "island", "game", "switch" et "play". Cela est expliqué par la disparité de la fréquence d'apparition des notes et le grand nombre de 0. Ce résultat nous semble logique, car le plus grand reproche des joueurs est le fait de ne pas pouvoir créer deux îles sur la même console. (cf. Figure 2 en annexe)

En troisième lieu, nous avons essayé de tirer les émotions transcrites par les commentaires en fonction de la note. Pour ce faire, nous avons utilisé une fonction de la bibliothèque nltk qui nous renvoie diverses informations sur les émotions positives ou négatives transmises par le texte donné. Nous pouvons remarquer que la répartition des mots est assez homogène. Cependant, les mots "one", "island" et "game" ressortent comme étant ceux avec le sentiment le plus négatif. (cf. figure 03 et figure 03 en annexe)

4. Difficultés et problèmes

La principale limite du jeu de données est le manque de données quantitatives. Les variables sont pour la plupart qualitatives (phrase type de chaque villageois, chanson préférée, nom des objets, source des objets…) ce qui limite les possibilités d'analyses possibles (d'où l'incorporation d'un jeu de données complémentaires). C'est aussi la principale raison pour laquelle on a décidé de s'orienter sur une analyse NLP (étudiée en LO17). Nous ne pouvons donc pas appliquer de la segmentation ou de la classification à priori. Nous sommes ouverts à toute proposition ou conseils concernant ce sujet là ou d'autres analyses.

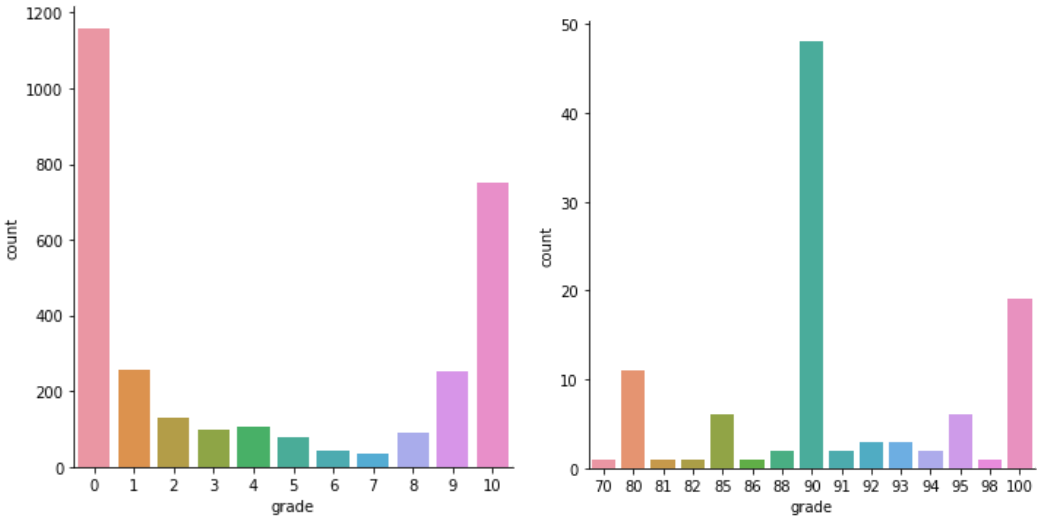

Figure 01 - A droite, répartition des notes données par les professionnels. A gauche, répartition des notes données par les joueurs.

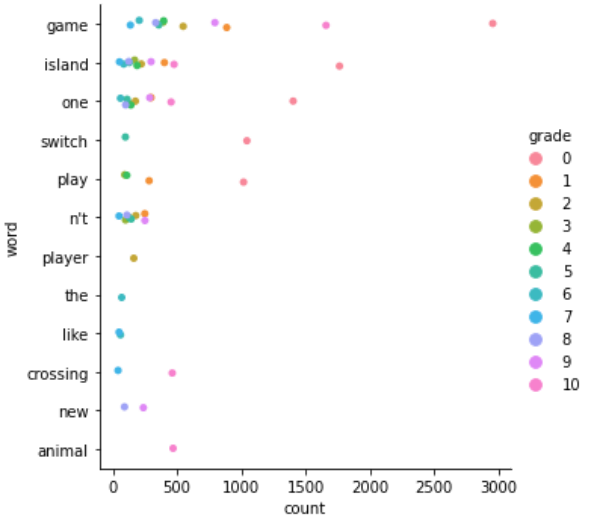

Figure 02 - Répartition des mots les plus utilisés dans les commentaires en fonction de leur notes.

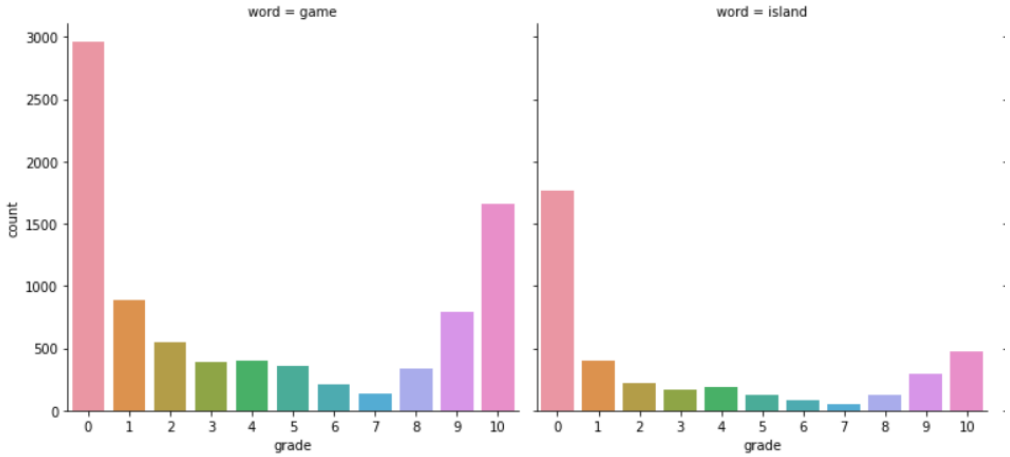

Figure 03 - Répartition des mots les plus utilisés dans les commentaires en fonction de leur notes pour chaque mot.

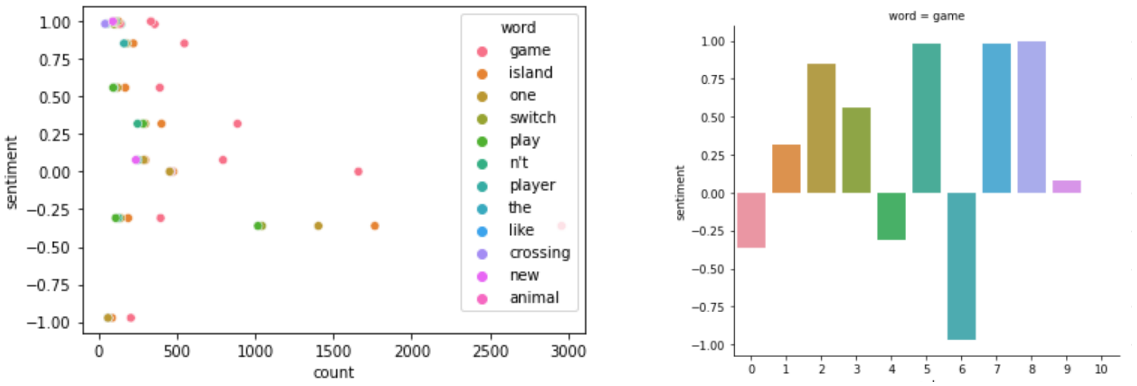

Figure 04 - Répartition des mots en fonction de la note de sentiments

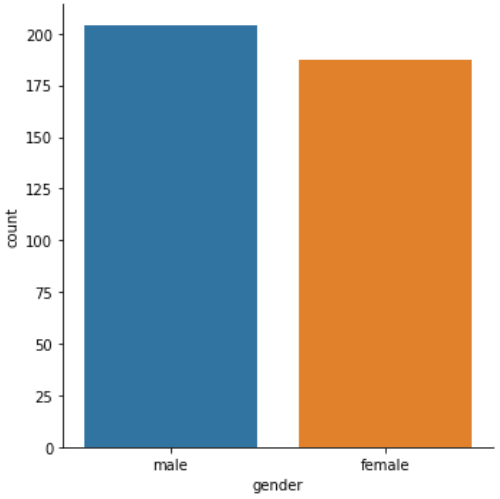

Figure 05 - Répartition des sexes des villageois

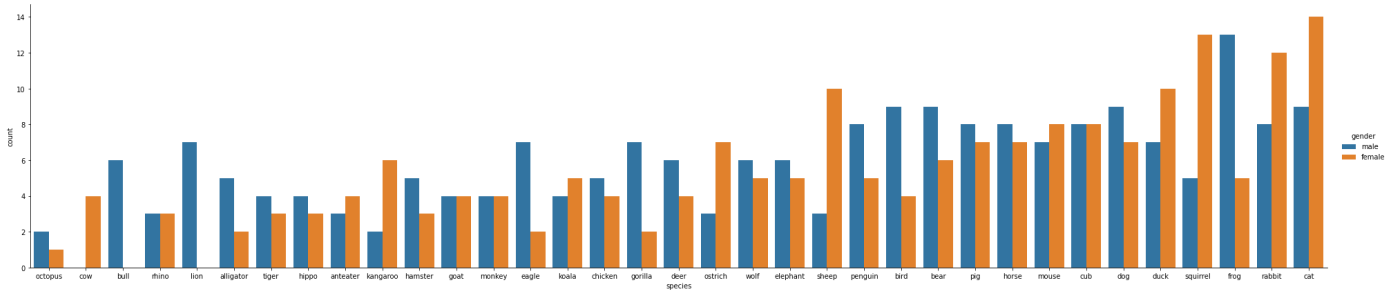

Figure 06 - Répartition des différences espèces par sexe

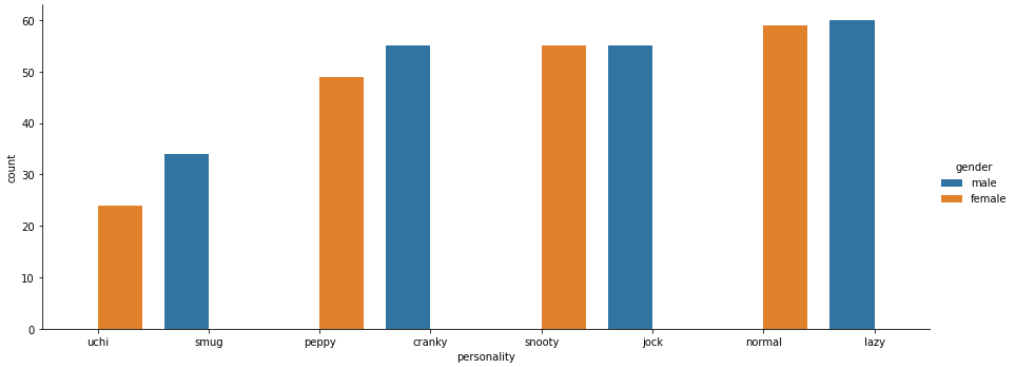

Figure 07 - Répartition des différences personnalités par sexe

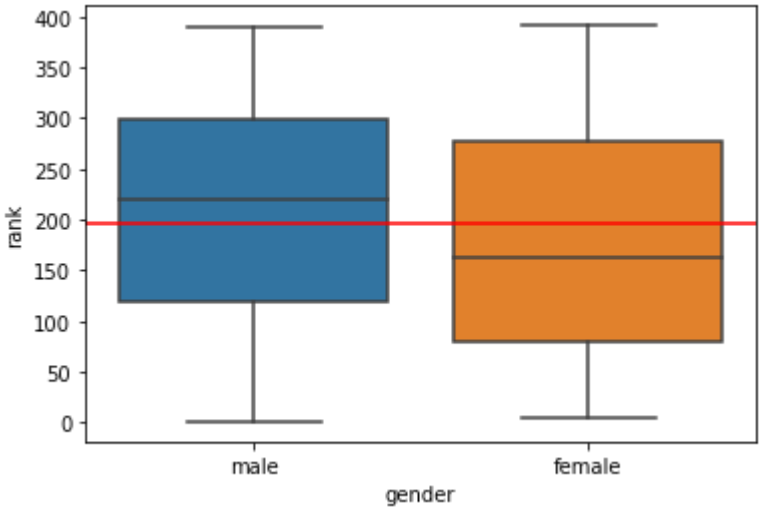

Figure 08 - Boxplots du rang en fonction du sexe des villageois

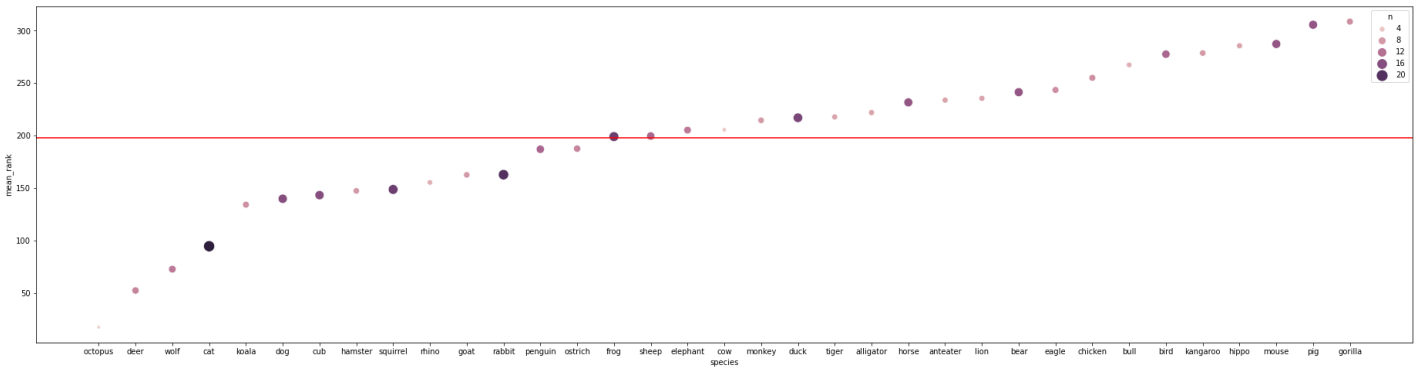

Figure 09 - Rang moyen des différentes espèces

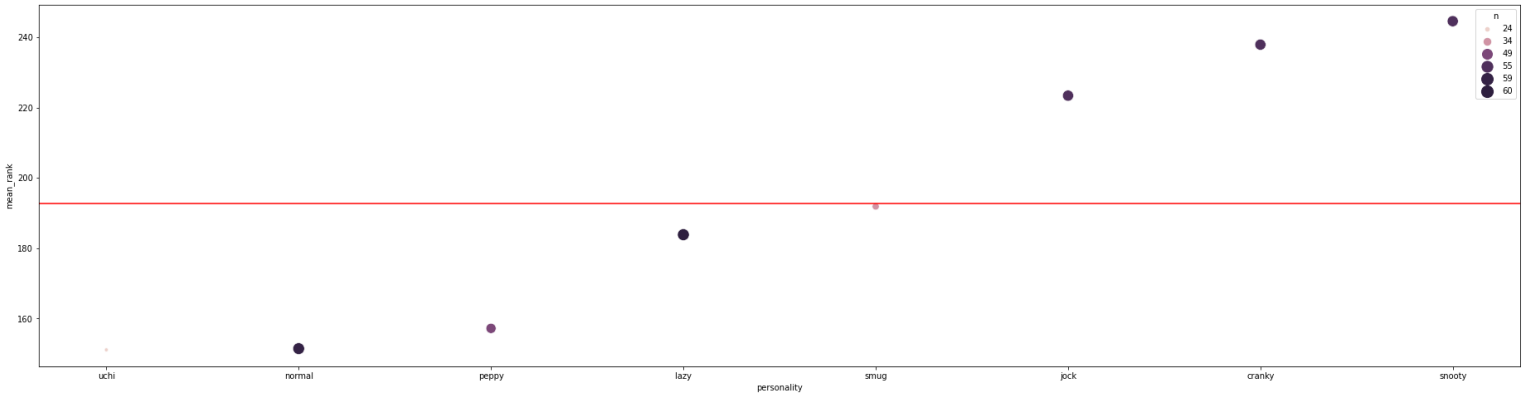

Figure 10 - Rang moyen des différentes personnalités

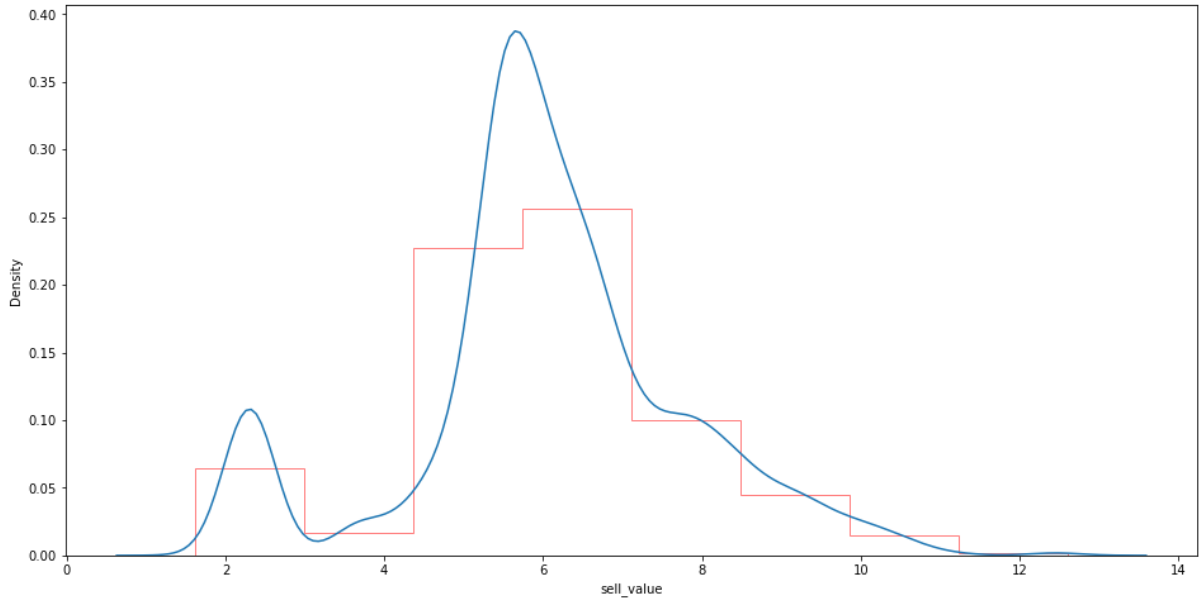

Figure 11 - Densité de la valeur de vente des objets

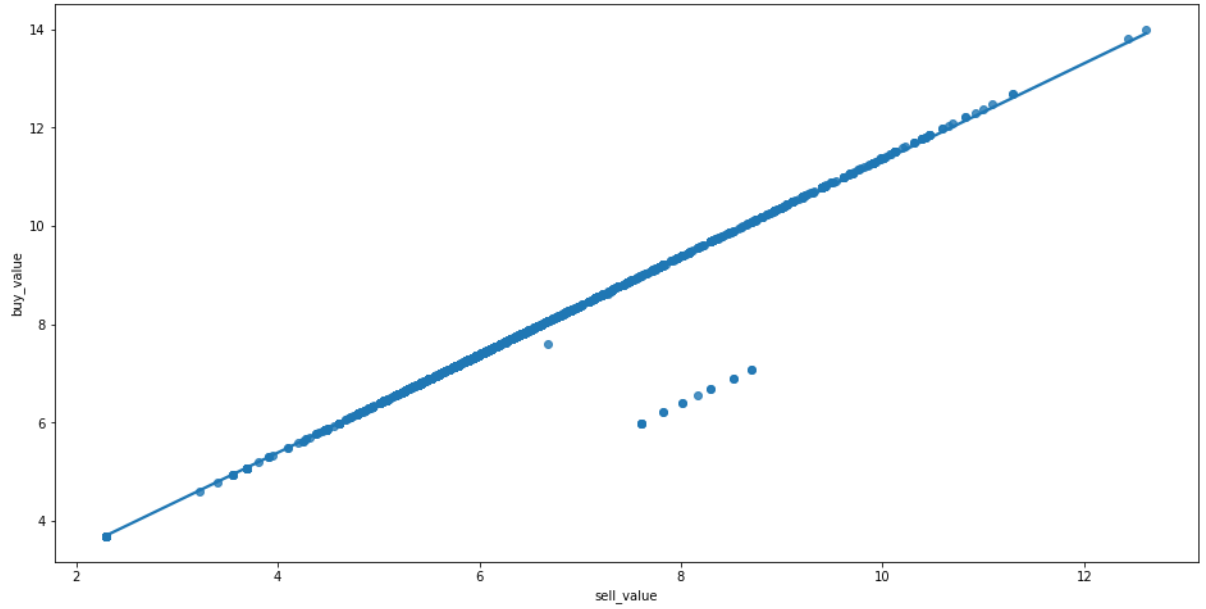

Figure 12 - Dispersion de la valeur d'achat en fonction de la valeur de vente des objets

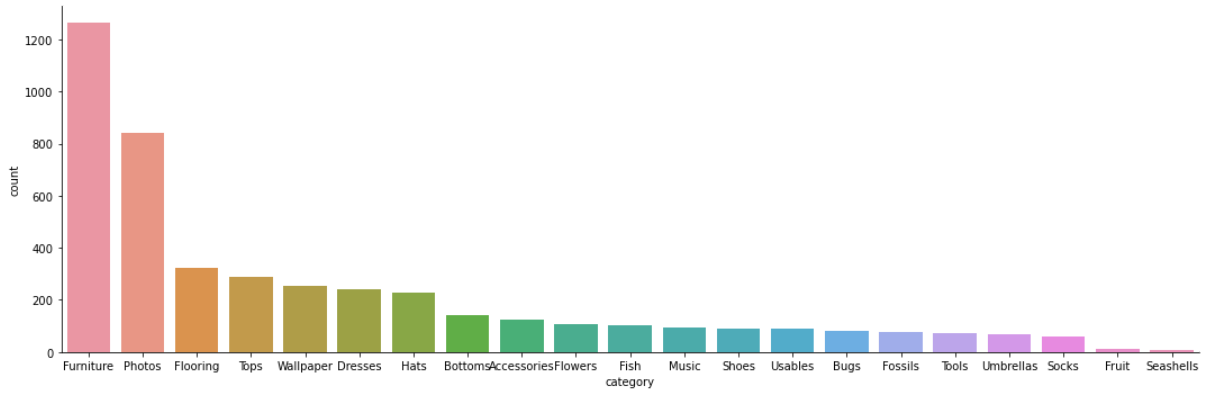

Figure 13 - Diagramme en boîtes des différentes catégories d'objets