

# TP5 VERNAY

## 1. Exploitation du fichier des données

Comme on peut le voir nous avons à notre disposition plusieurs tables.

Par exemple, Author est la table des auteurs, Book celle des livres etc.

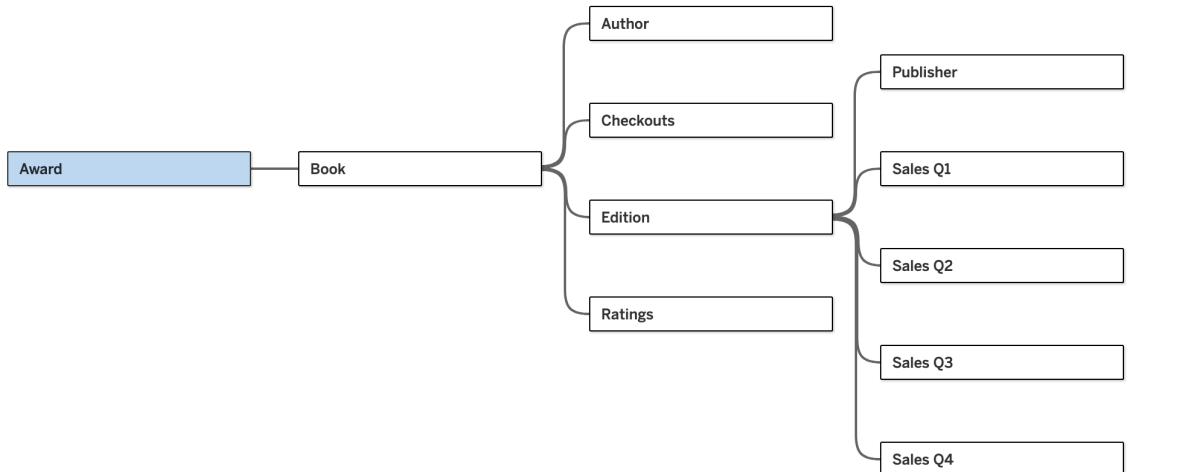
Chacune de ces tables contiennent des informations correspondant à son unité statistique. Par exemple, dans celle des auteurs, on a l'id qui est un identifiant unique, le nom, le prénom (variable string) l'anniversaire (date), le lieu de résidence (localisation) et enfin le nombre d'heures passées à écrire par jour (variable quantitative continue).

A travers ces tables, on a la possibilité de faire des jointures.

Par exemple, dans la table des livres, on trouve pour chaque livre son identifiant unique, son titre, et l'identifiant de l'auteur qui l'a écrit.

Ainsi la jointure se fait par cet attribut : on peut retrouver qui est l'auteur qui a écrit le livre et notant l'id auteur de la table livre et en allant le chercher dans la table auteur.

Tableau fait ce travail pour nous.



On se retrouve alors avec ce schéma de relations, où chaque liaison correspond à une jointure qu'on a mentionné précédemment.

## 2. Mise en place de la structure de données complètes

Les sources book et info peuvent soit être fusionnées, soit mises en relation.

On essaye d'abord de les mettre en relation.

The screenshot shows the Tableau Data Source interface. On the left, there's a dropdown menu "Book — Info". Below it, a section titled "How do relationships differ from joins? Learn more" contains fields for "Book" and "Info". The "Book" field has "Abc Book ID" and an operator "=". The "Info" field has "[BookID1]+[BookID2]". A red oval highlights this concatenation formula. To the right is a preview of the joined data, showing columns: BookID1, BookID2, Genre, Series ID, Volume Number, and Staff Comment. The data includes rows like MM (BookID1: 424, BookID2: Young Adult), NR (BookID1: 695, BookID2: Mystery), etc.

Un problème se pose : la clé de jointure (l'attribut qui nous permet de faire correspondre les deux tables ensemble) est séparée en deux champs dans la table info (dans les colonnes BookID1 et BookID2 : on a respectivement le début et la fin de l'identifiant, avec d'abord les lettres puis ensuite les chiffres).

Pour assurer la liaison, il faut se servir de la fonction Edit Calculation (en cliquant sur le champ entouré en rouge qui correspond à la table info). On concatène ensuite les deux colonnes pour assurer la jointure.

Pour la fusion, tableau ne propose pas d'option pour pouvoir l'opérer en faisant correspondre une colonne avec deux colonnes dans l'autre table. On doit créer manuellement un nouvel attribut pour que la jointure puisse se faire.

A screenshot of an Excel spreadsheet. The formula bar at the top shows "=CONCAT(A1;B1)". The main table has columns G, H, I, and J. Column G contains the formula "=[BookID1]&[BookID2]". Column H contains the concatenated values: MM424, NR695, AM124, AK974, AD222, AY135, BB194, BF889, BC244, BR858, BF374, BS284, CH391, CC830, and so on. The text "htly. But what, Lydan asks herself, does it mean to save something? Especially when the cost is...concer" is visible in the bottom-left corner. The text "nd, once an Australian Aboriginal penal colony, transformed into an eco-tourism destination thanks to" is also visible at the bottom.

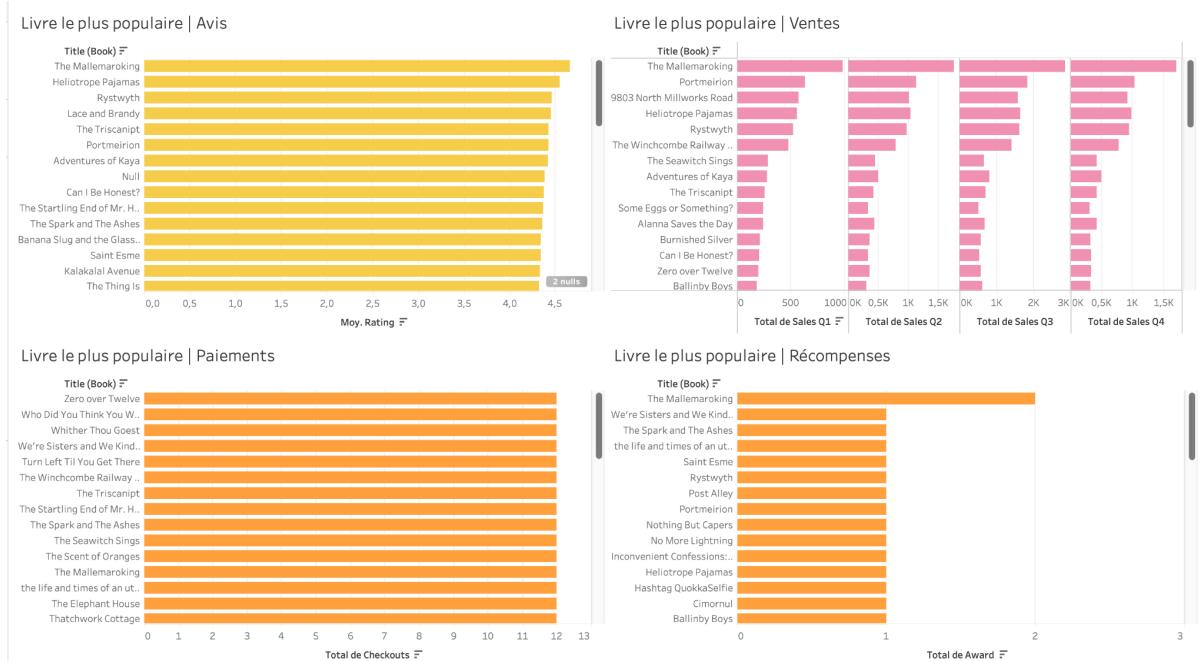
Union      11 fields 116 rows

Name	Union	BookID	Title	AuthID (Union)	BookID1	BookID2	Genre	SeriesID	Volume Number	Staff Comment
BB194	Ballinby Boys	AM329		null	null	null	null	null	null	null
NC652	Nothing But Capers	AS443		null	null	null	null	null	null	null
AD222	Alanna Saves the Day	BH149		null	null	null	null	null	null	null
PA169	Post Alley	BM856		null	null	null	null	null	null	null
TC188	Thatchwork Cottage	BM856		null	null	null	null	null	null	null
ZT703	Zero over Twelve	BM856		null	null	null	null	null	null	null
PP866	Portmeirion	BT132		null	null	null	null	null	null	null
RR774	Rystwyth	BT132		null	null	null	null	null	null	null
TM925	The Mallemaroking	BT132		null	null	null	null	null	null	null

Pour l'instant, l'union ne semble pas marcher.

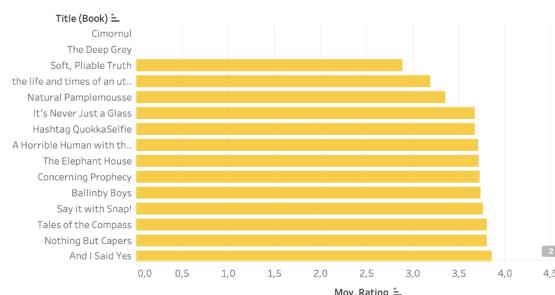
### 3. Analyses

- Quels livres sont les plus populaires ? Le moins populaire ? Est-ce basé sur les ventes, les avis, les paiements ou une autre mesure ?

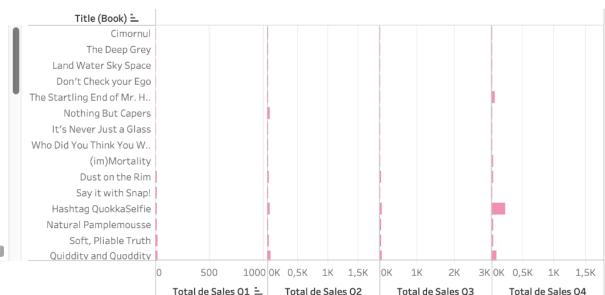


Commençons par chercher quel livre semble le plus populaire. On remarque que *The Mallemaroking* apparaît en tête des ventes sur tous les trimestres, il a les meilleures notes (avec une moyenne de 4,6/5 !) et le plus de récompenses.

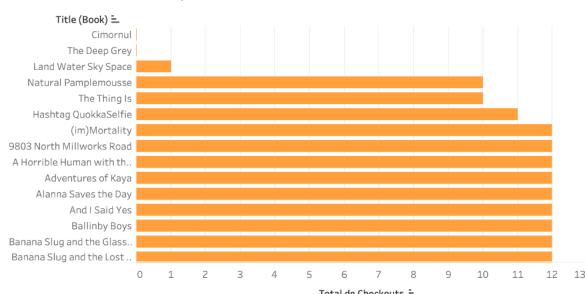
Livre le plus populaire | Avis



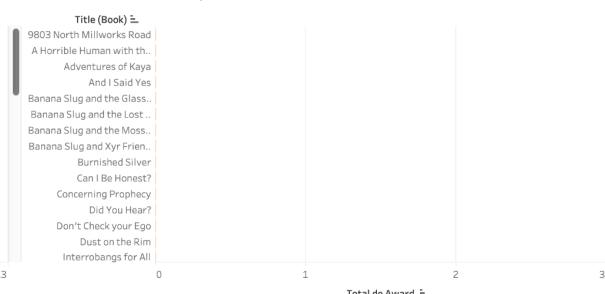
Livre le plus populaire | Ventes



Livre le plus populaire | Paiements



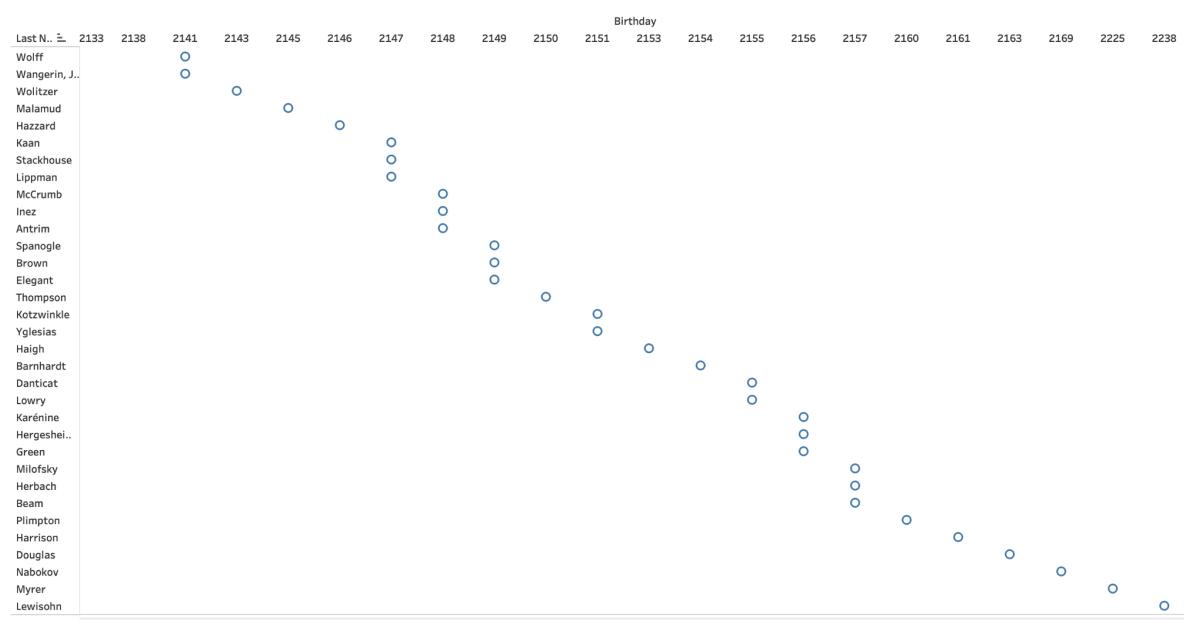
Livre le plus populaire | Récompenses



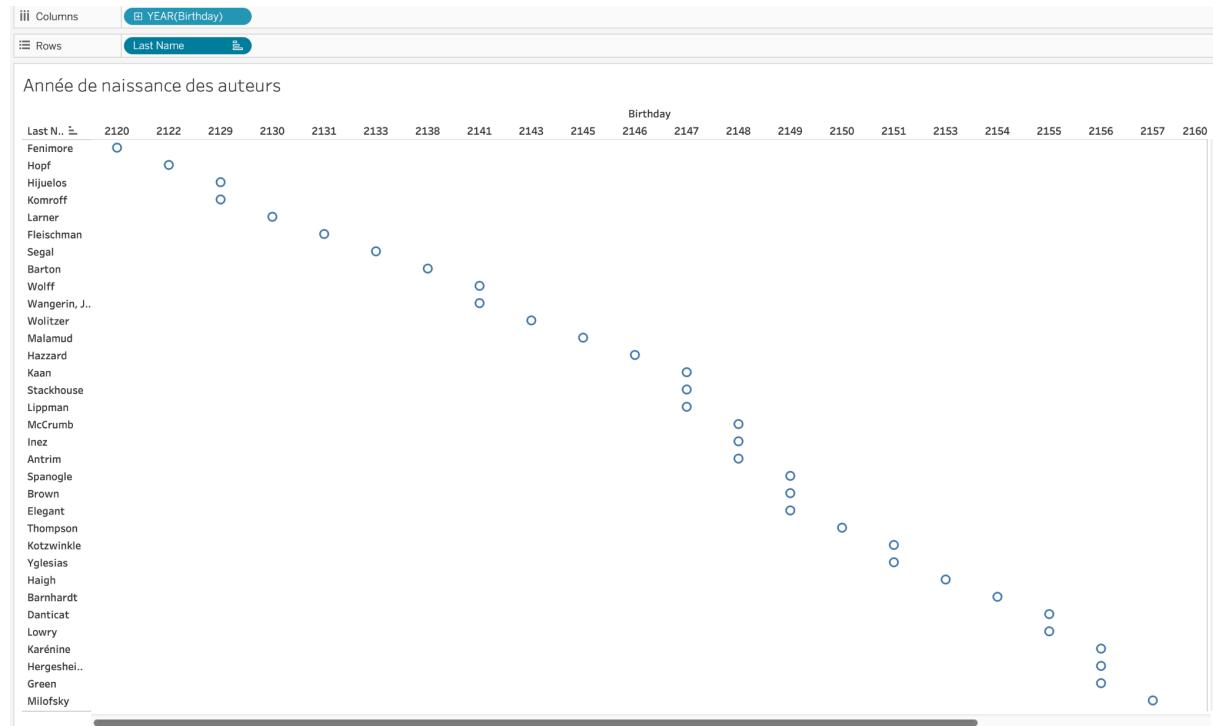
Pour celui qui est le moins populaire, les résultats sont plus contrastés. Celui qui a les moins bonnes notes est *Soft, Pliable Truth* moins de 3/5. On le retrouve également dans le bas des ventes par trimestre. Celui qui a le moins de paiements est *Land Water Sky Space*. (Concernant *Cimornul* et *The deep grey* qui sont tout en bas du classement, ce sont sûrement des valeurs manquantes. On les filtrera par la suite).

#### • Qui était le plus jeune premier auteur ? Qui était le plus âgé ?

Année de naissance des auteurs

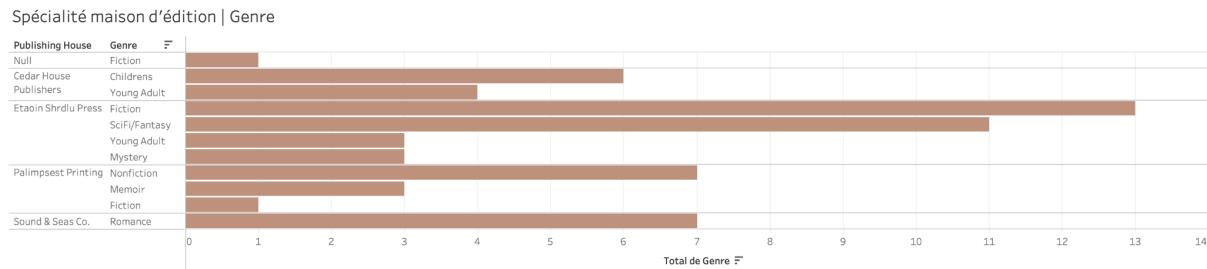


Pour obtenir ce résultat, on cherche l'auteur né le plus tard. C'est donc Bravig Lewisohn, né en 2238, qui est le plus jeune auteur.

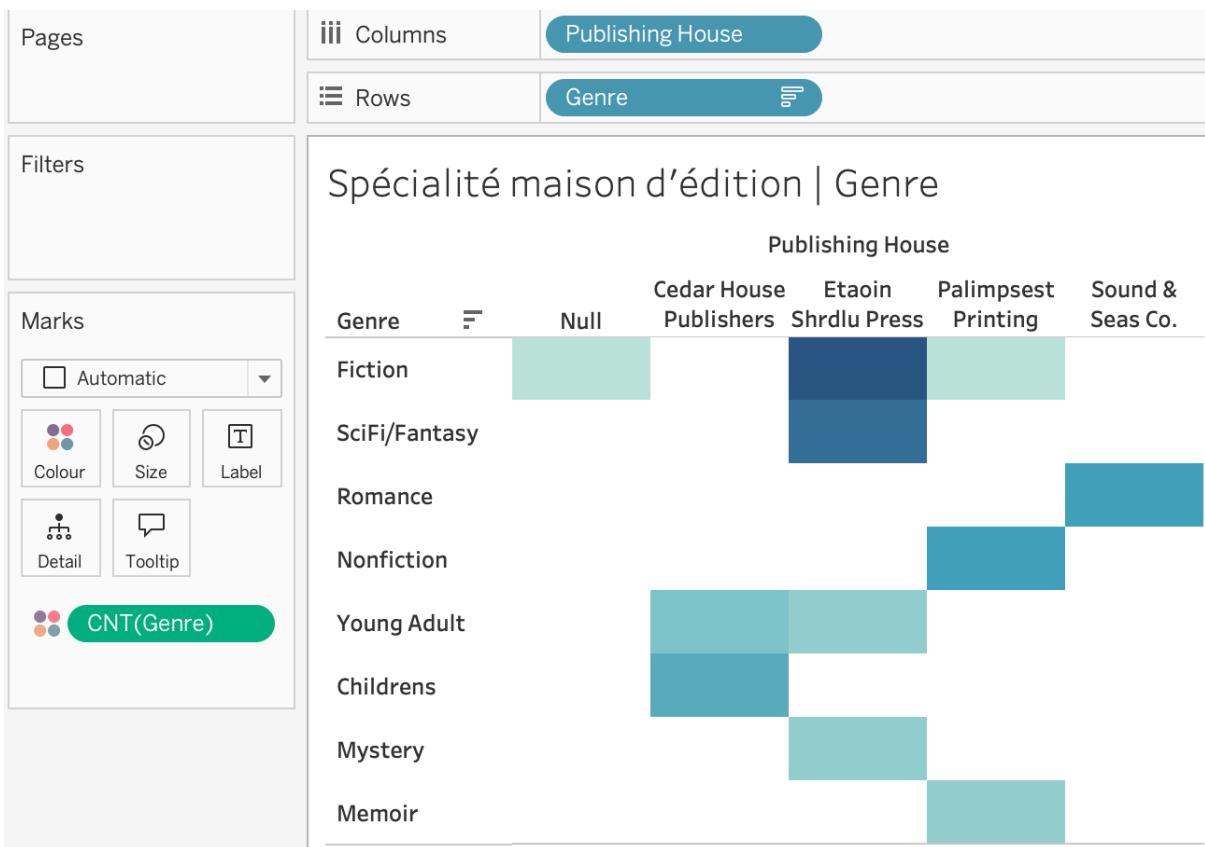


Pour le plus vieux, on procède en sens inverse. C'est Charles Fenimore, né en 2120.

- Certaines maisons d'édition semblent-elles spécialisées d'une manière ou d'une autre ?



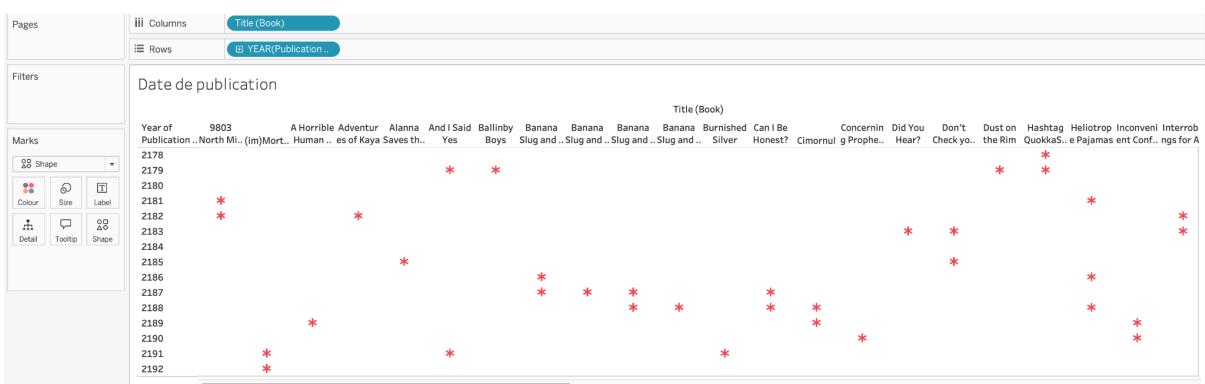
Comme on fait beaucoup de graphiques en barre, on montre aussi une autre possibilité avec une heatmap.



On voit que la maison d'édition Sound & Seas Co. ne publie que des romances. Etaoin Shrdlu Press fait plus dans la fiction et la fantaisie.

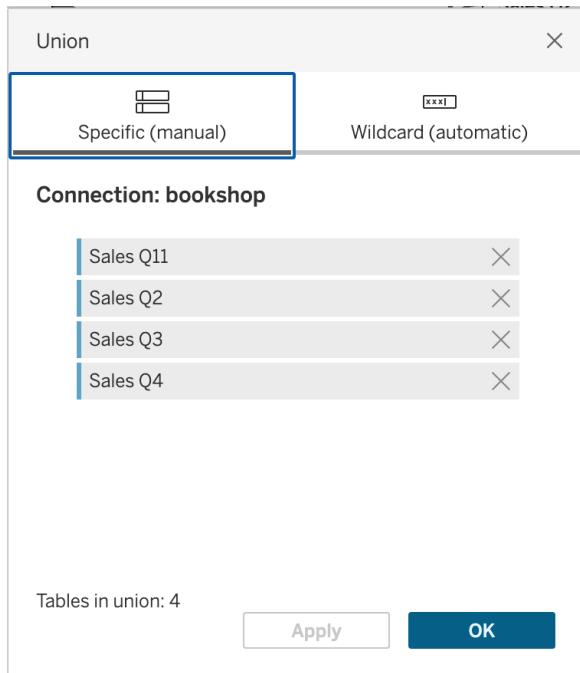
On met aussi le nombre d'auteurs affiliés à chaque maison d'édition pour porter un regard critique sur nos résultats. On voit par exemple que Sound & Seas Co. ne signe qu'un auteur. Donc on peut se demander si la romance est le genre de prédilection de la maison d'édition ou simplement celui de l'auteur unique qu'elle publie.

- Quelle a été la durée la plus longue entre les éditions d'un même livre ?



On voit que les deux points qui sont sur une même ligne et les plus éloignés sont ceux qui font référence aux publications de *And I Said Yes*, avec une publication 2179 et en 2191, ce qui fait un écart de 12 ans.

- Existe-t-il des tendances saisonnières pour les ventes ? Et les caisses ? Certains titres ou genres ont-ils des fluctuations saisonnières ?



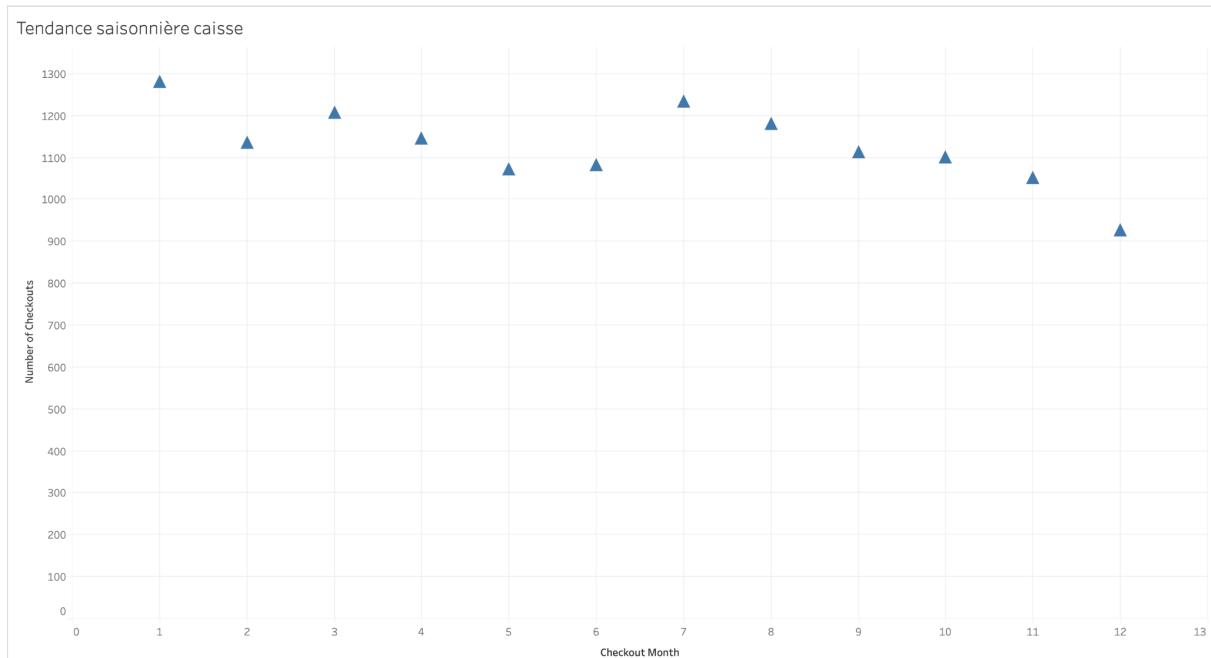
Pour cette question, il est utile de créer une union. En effet, les données de ventes sont séparées par trimestre dans des tables différentes, ce qui n'est pas pratique pour représenter des évolutions sur l'année. Ainsi, pour visualiser tous les mois côté à côté, on fait une union qui les réunit dans une et même table.

Ensuite, on peut faire des graphiques comme celui-ci :

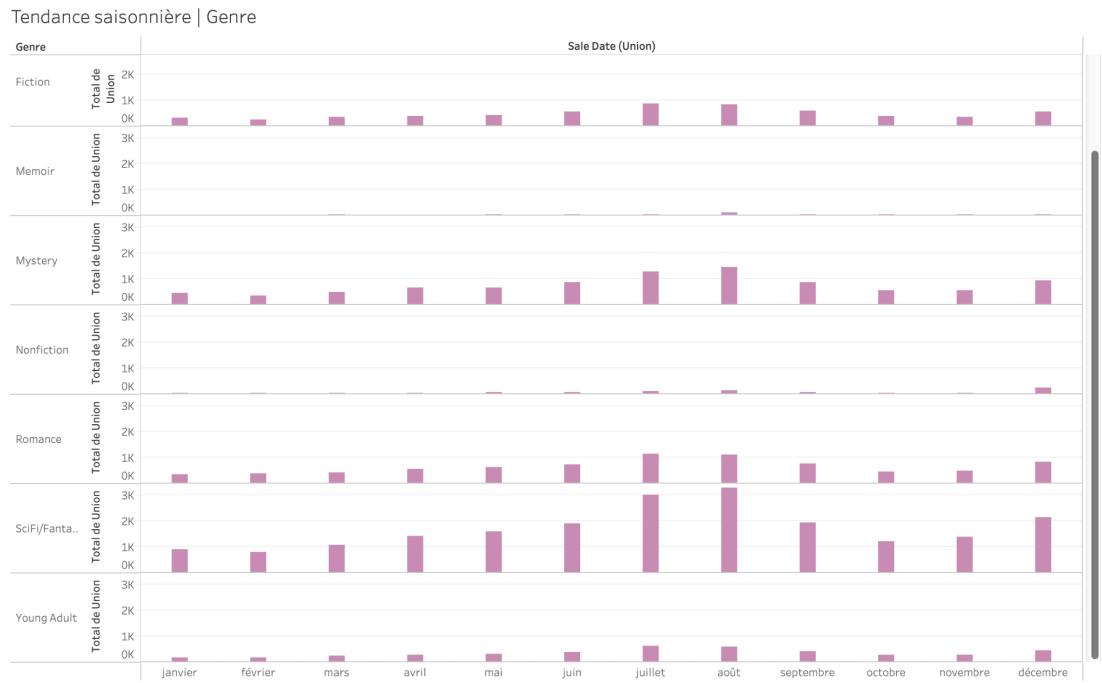


On remarque une forte hausse pour les mois de juillet et août. On peut faire l'hypothèse que les vacances sont une période idéale pour lire, et les lecteurs ont aussi plus de temps ; il leur faut donc des livres.

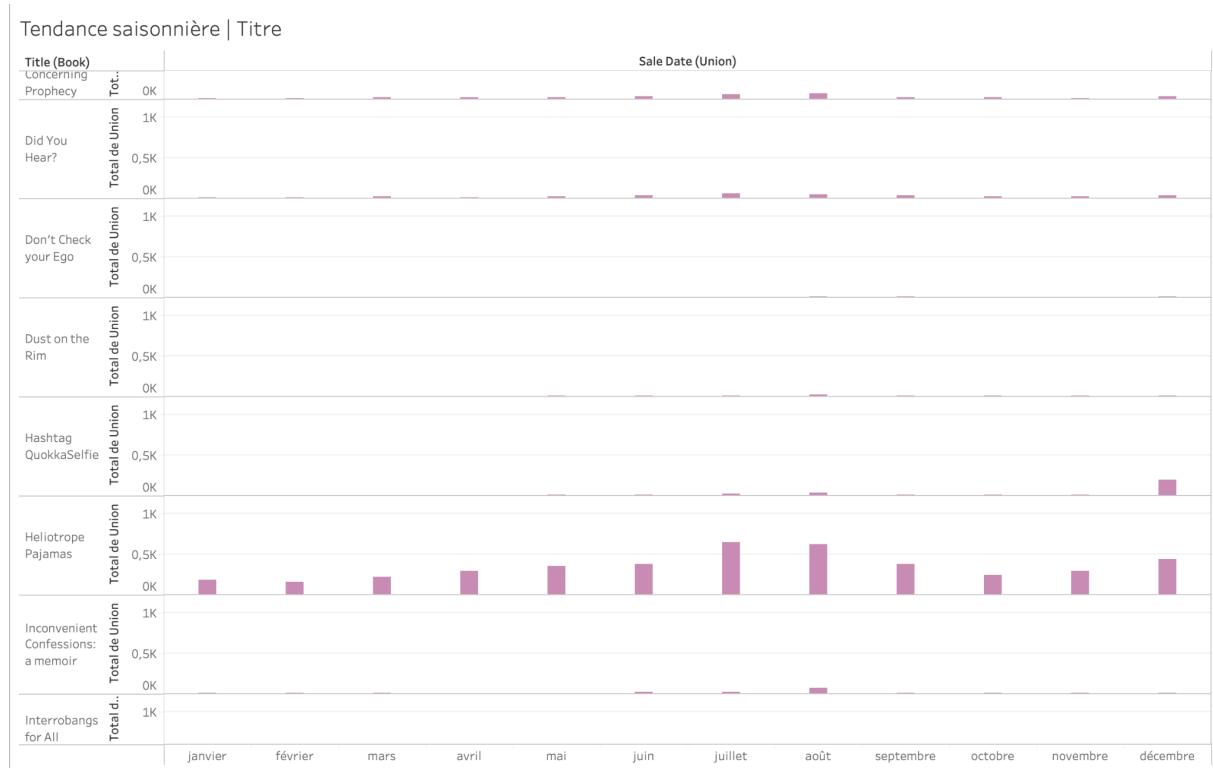
On remarque également une hausse logique en décembre, avec les cadeaux de Noël.



En ce qui concerne les caisses, on ne remarque pas de variations flagrantes ; à part peut-être une baisse progressive jusqu'à la fin de l'année.

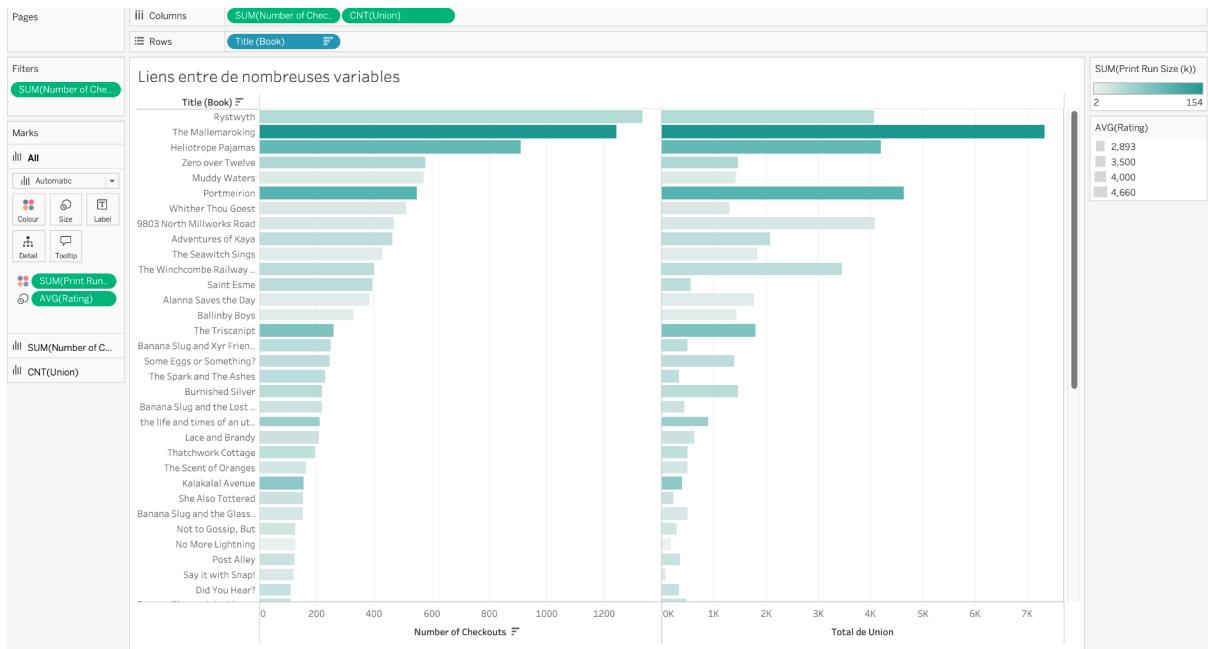


Pour ce qui est des variations saisonnières par genre, tous les genres suivent la tendance générale. Les livres de fantaisie/science-fiction représentent le plus gros des ventes, notamment en été (avec 3 000 ventes pour chacun de ces mois).



En ce qui concerne les titres, ils suivent aussi les variations globales (décrisées précédemment) pour la grande majorité. On remarque néanmoins le livre *Hashtag QuokkaSelfie* qui a le plus de ventes en décembre (et pas en juillet/août comme les autres).

- Existe-t-il des corrélations entre les paiements, la taille du tirage, les notes des critiques de livres et le volume des ventes ?



Cette visualisation est très riche en information pour répondre à la question, et nécessite des explications.

La taille d'une barre correspond à la note moyenne du livre. La couleur, à la taille du tirage. La longueur des barres de gauche représente les paiements (triés par ordre décroissant) ; et celle de droite représente le volume des ventes.

Commençons avec la corrélation entre les paiements et les autres variables. On compare donc la taille des barres de gauche à toutes les autres caractéristiques.

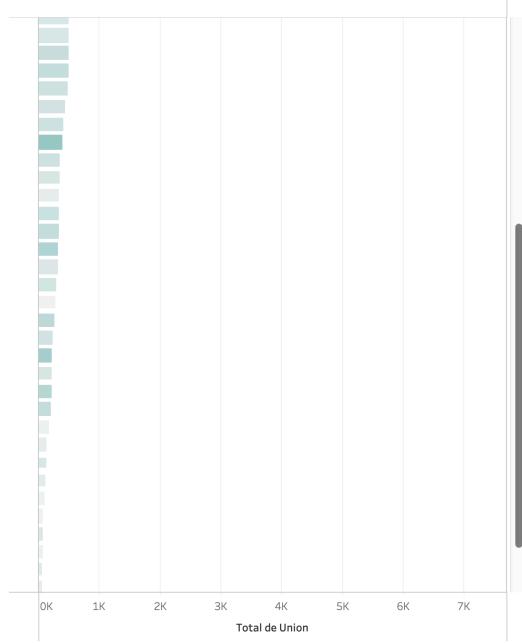
De manière générale, plus le nombre de paiements diminue, et moins le tirage est grand (la couleur est de plus en plus pâle). On constate tout de même un nombre d'exceptions notables ; où le nombre de paiements est dans la tranche haute, mais qui ont un tirage plutôt petit.

Pour la corrélation entre les paiements et le volume des ventes, la dynamique est la même : la tendance globale baisse mais il existe des exceptions.

Enfin, le lien entre paiement et note ne semble pas évident (on a des largeurs de barres assez disparates à travers le tri décroissant).

Ensuite pour le total des ventes.

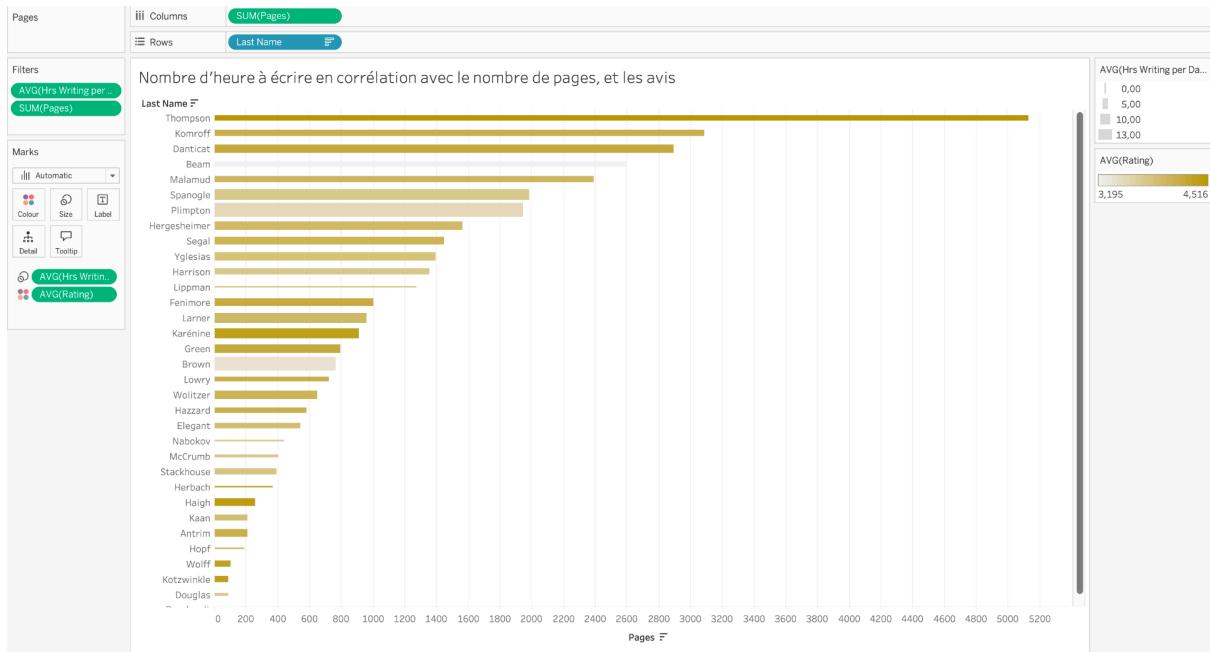
Quand le total des ventes est grand, le nombre de tirage l'est aussi (ce qui semble logique).



En triant cette fois par le volume des ventes, on voit que quand celui-ci baisse, la note aussi.

Enfin, pour la corrélation entre la taille des tirages et les notes, on ne remarque pas de tendance particulière à première vue.

- Les auteurs qui passent le plus de temps à écrire ont-ils les livres les plus réussis ? Ont-ils le nombre de pages le plus élevé ?



On a fait une heatmap combinée avec un barplot.

Ainsi, pour lire ce graphique, on fait attention aux éléments suivants : plus une bar est foncée en jaune, plus la moyenne des avis de rapprochent de 5 (sur 5), plus la bar est longue, plus le nombre de page est important, et enfin plus une bar est épaisse, et plus le nombre d'heure passées à écrire par jour est important.

Par exemple, le premier auteur Thompson ne passe pas particulièrement beaucoup de temps à écrire par jour (barre assez fine) mais il a de très bon avis (barre foncée) et le plus grand nombre de pages écrites (barre la plus longue).

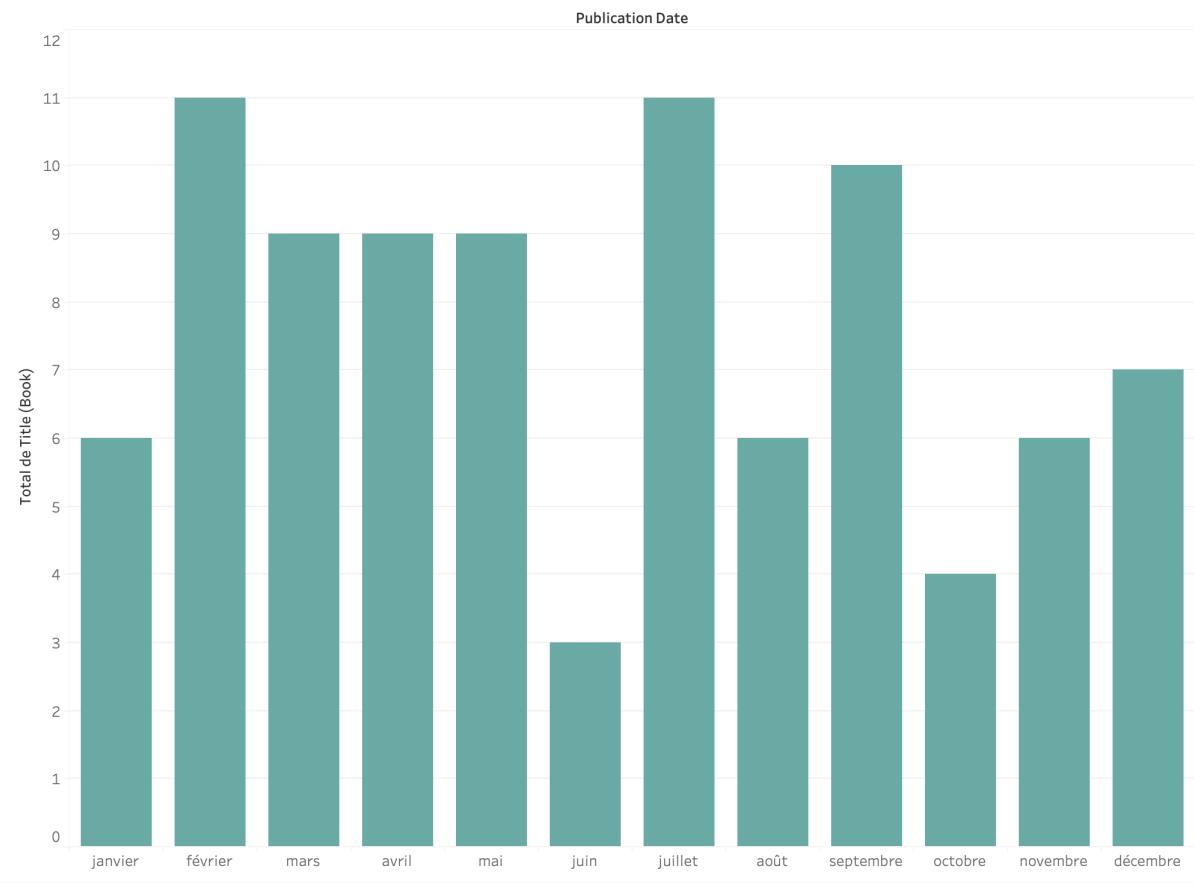
Avec cette grille de lecture, on ne remarque pas de corrélation flagrante entre le nombre d'heures par jour passées à écrire et les avis, ou même le nombre de pages.

On voit par exemple Beam qui passe beaucoup de temps à écrire, mais sa note moyenne est de 3. Quant à Brown, il passe beaucoup de temps à écrire mais il n'a pas beaucoup de pages.

Néanmoins, on remarque que les auteurs qui produisent moins de pages écrivent moins longtemps.

- Quand la plupart des livres sont-ils publiés ? Y a-t-il des anomalies ?

Publi. livre | Mois

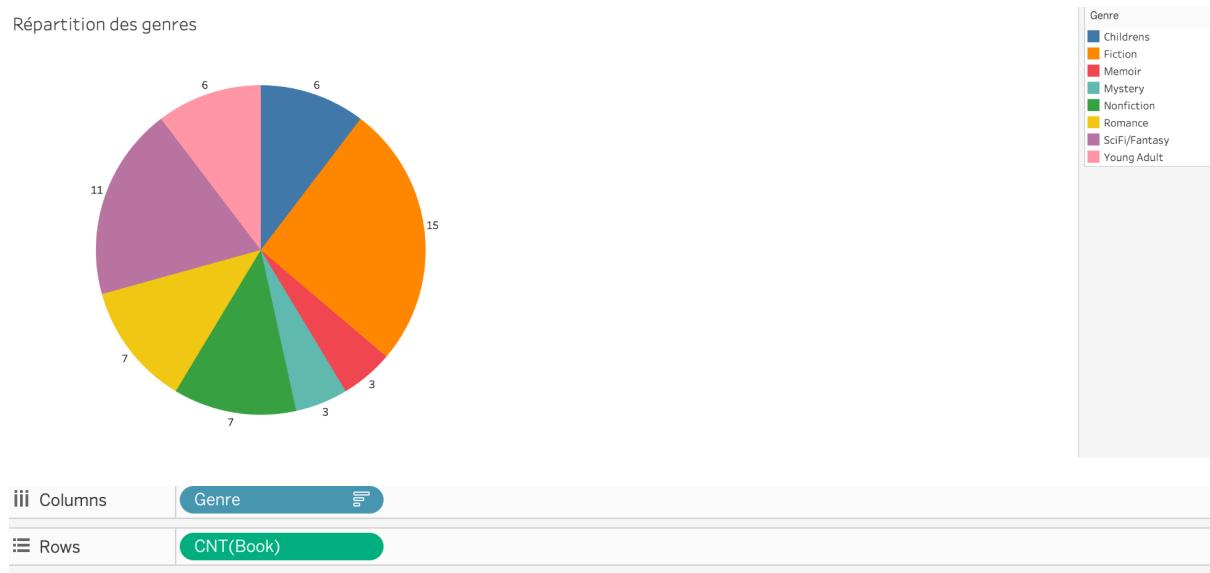


Les mois avec le plus de publications sont les mois de février et de juillet (11 publications chacun), avec septembre qui suit de près (avec 10 publications).

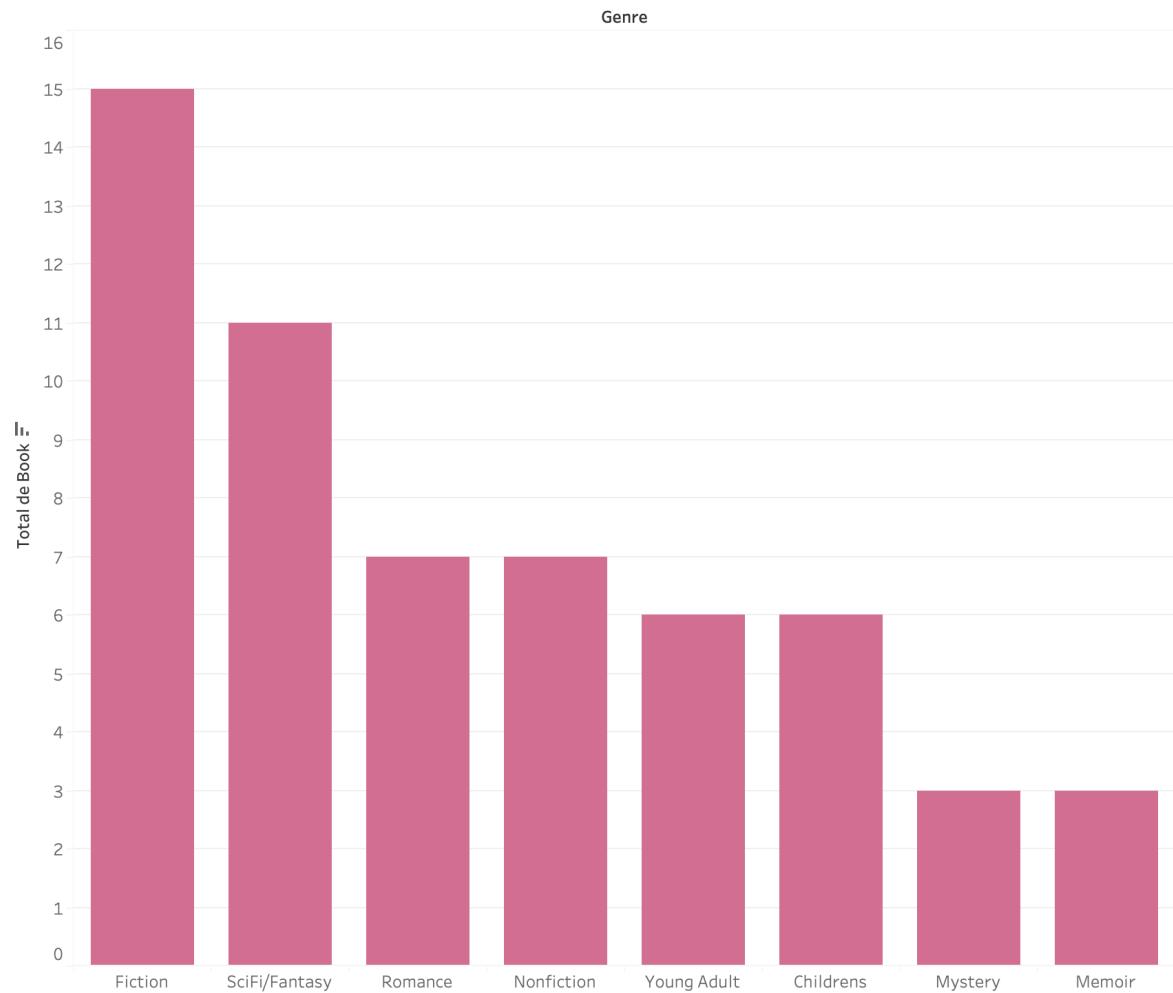
On a aussi remarqué une anomalie, où un livre n'avait pas de date de publication (champ Null). On l'a retiré du graphique.

- Existe-t-il des tendances en matière de genre, de format et de prix ?

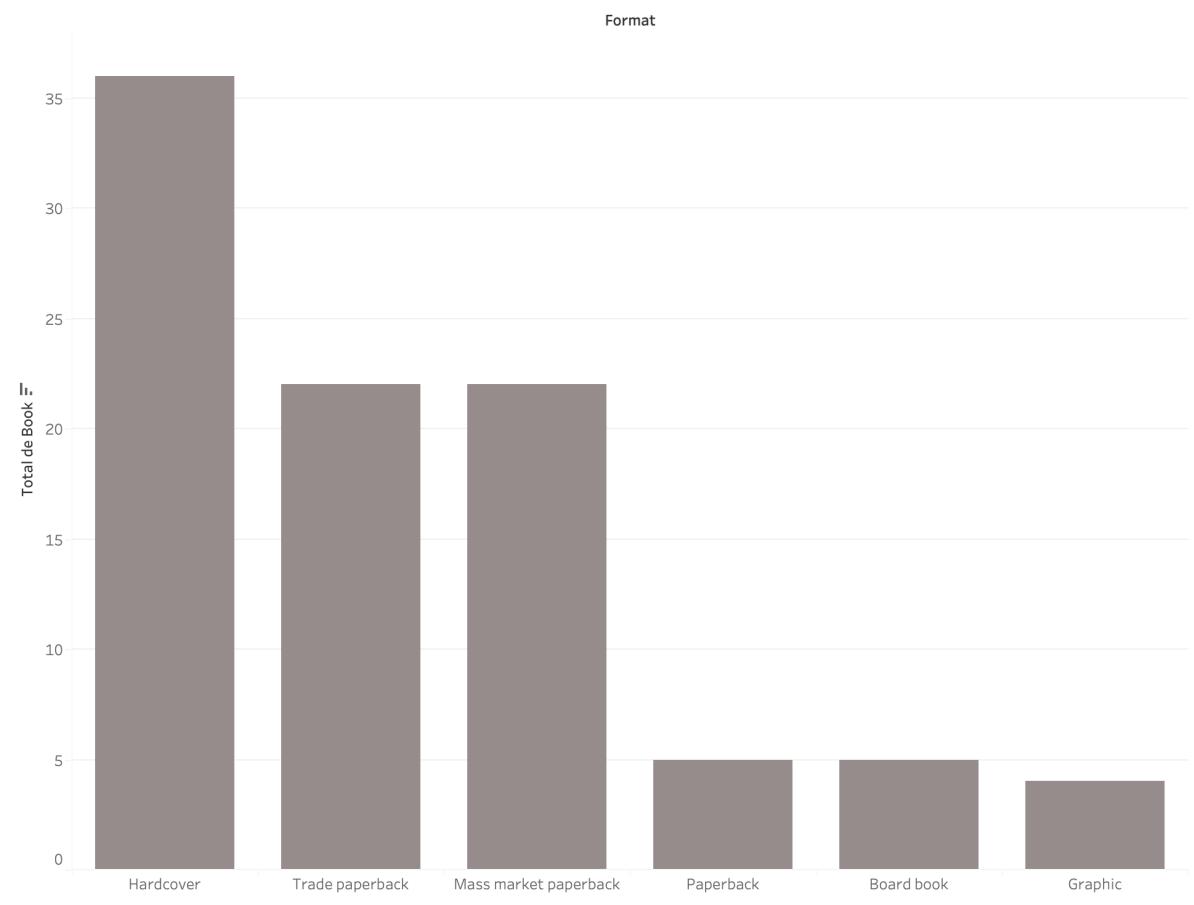
Répartition des genres



Répartition du genre des livres

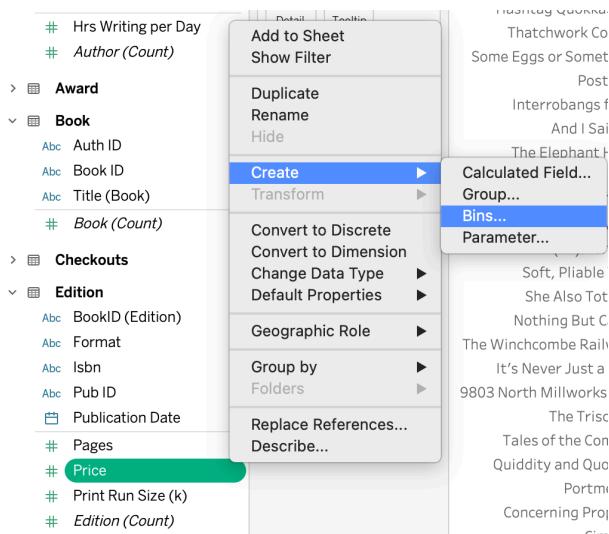


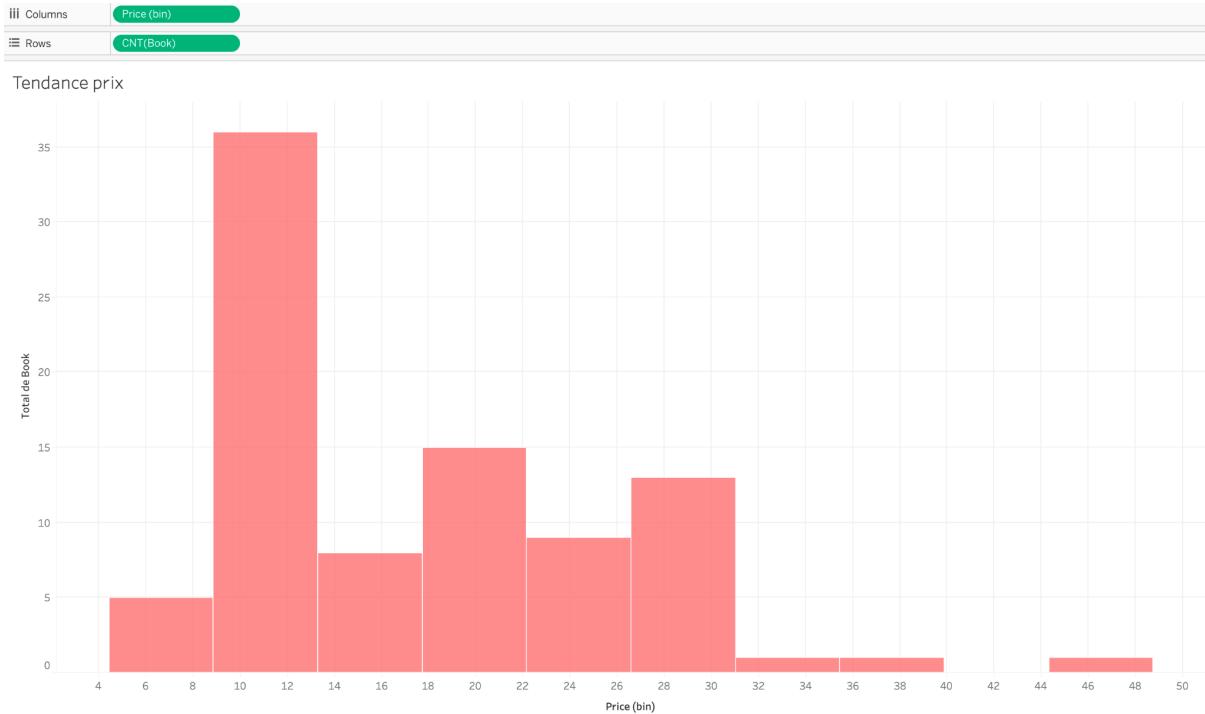
Tend. format



Pour les tendances de prix, comme c'est une donnée continue, on doit faire un histogramme.

Pour cela, on créer des bins :



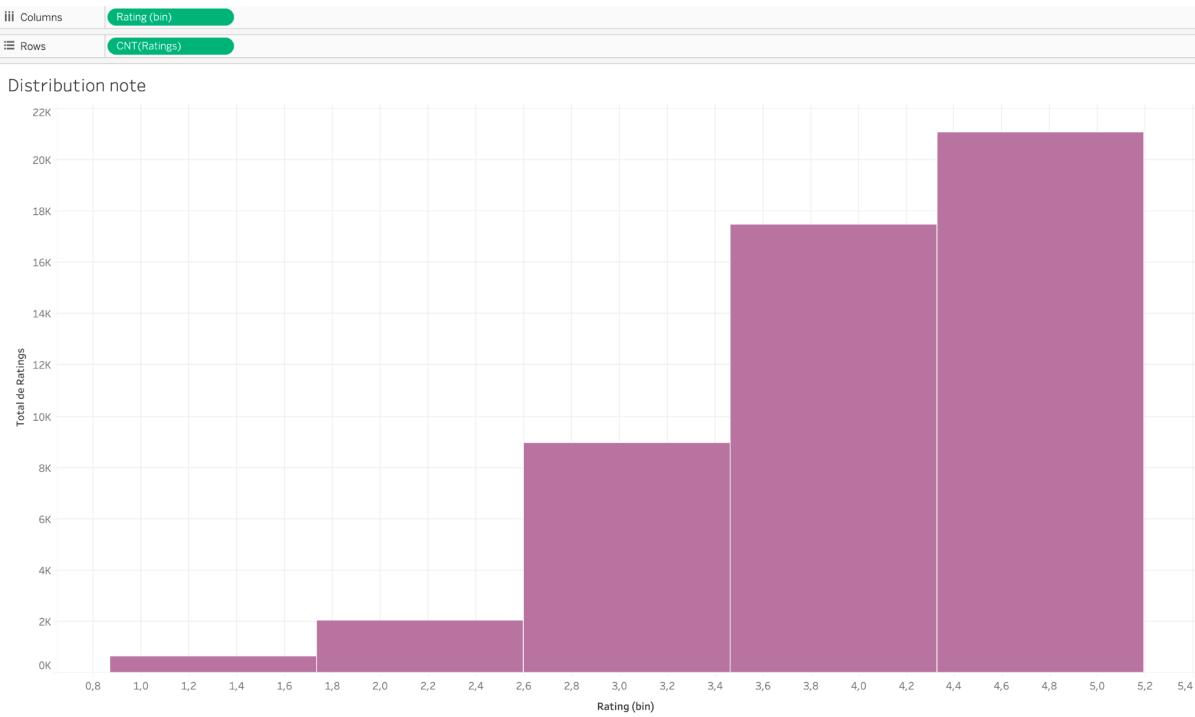


La fiction est le genre avec le plus de livres (15 livres) suivi de la science-fiction/fantaisie (11 livres).

Pour le format, c'est très clairement la couverture solide qui domine (hardcover).

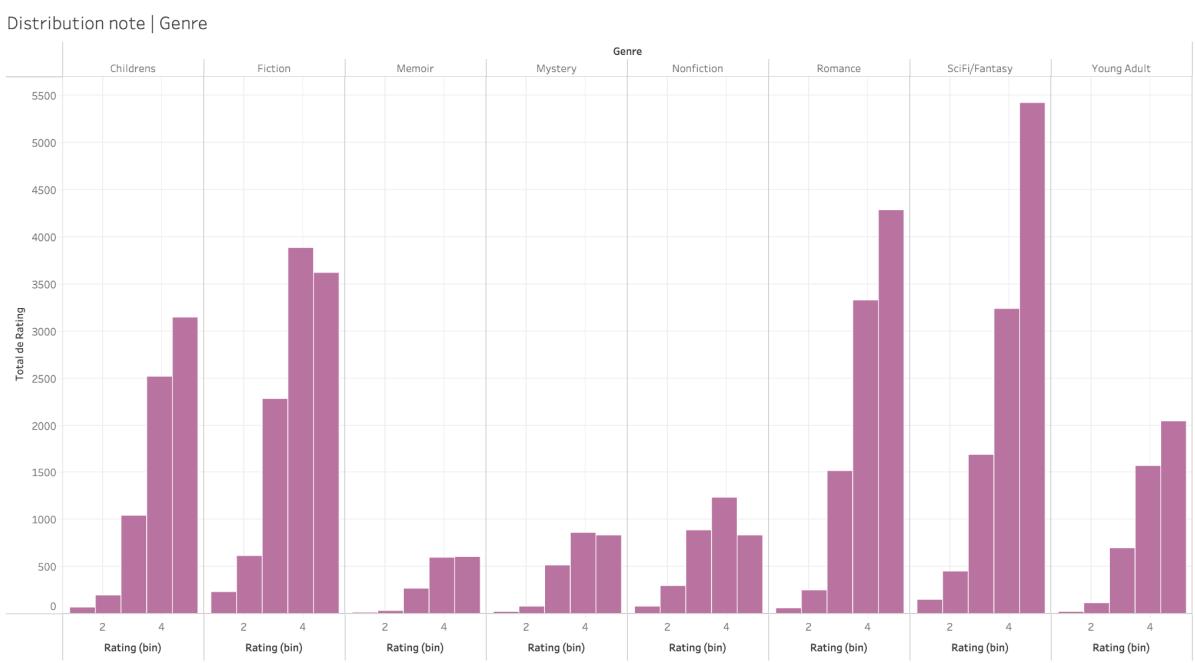
En termes de prix, on voit que beaucoup de livres se situent entre 9 et 13 euros. Très peu dépassent les 31 euros.

- Quelle sorte de distributions les notes ont-elles ? Ces distributions varient-elles selon le livre ? Par genre ? Semblent-ils s'aligner sur les récompenses ?



La distribution des notes montre que la majorité des avis sont positifs, si on considère toutes les notes, tous livres et tous genres confondus. Plus on considère une tranche de note élevée, plus il y a de notes dans ladite tranche (on voit l'histogramme “croissant”).

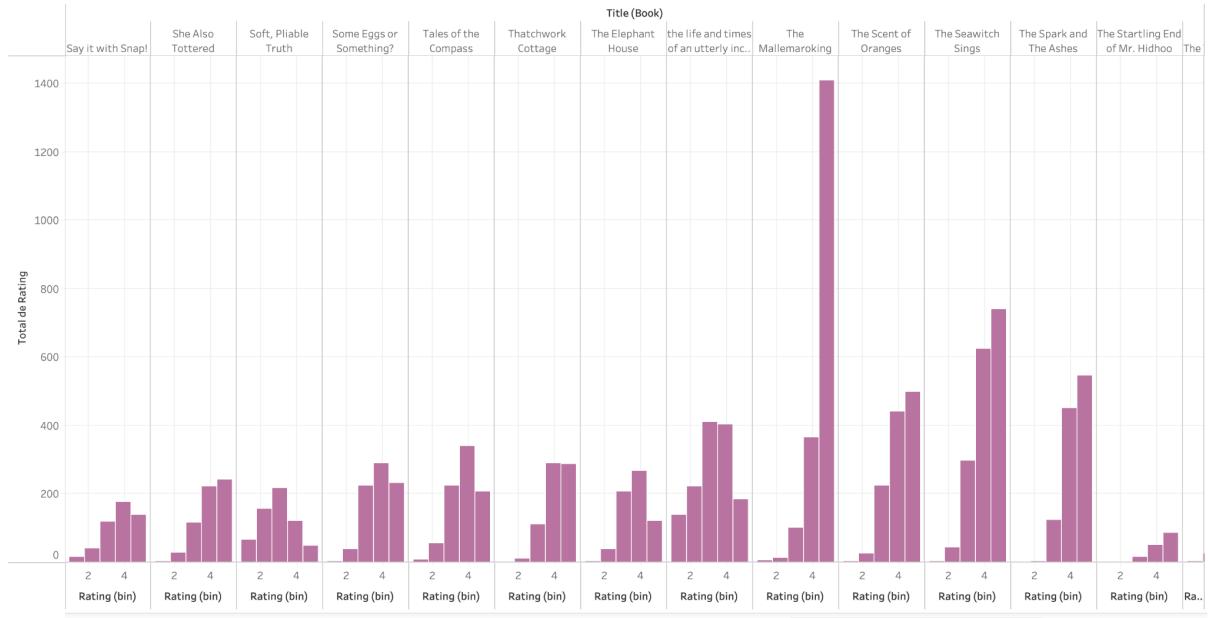
Essayons maintenant de voir en fonction des genres.



Selon les genres, la tendance reste à peu près la même : il y a un plus grand nombre de bonnes notes (4 ou 5 sur 5). Plus on considère une tranche haute, plus il y a de notes dans ladite tranche.

Cela dit, quelques catégories dérogent à cette règle, notamment la fiction, et la nonfiction, où les très bonnes notes se font un tout petit peu plus rares que les bonnes notes.

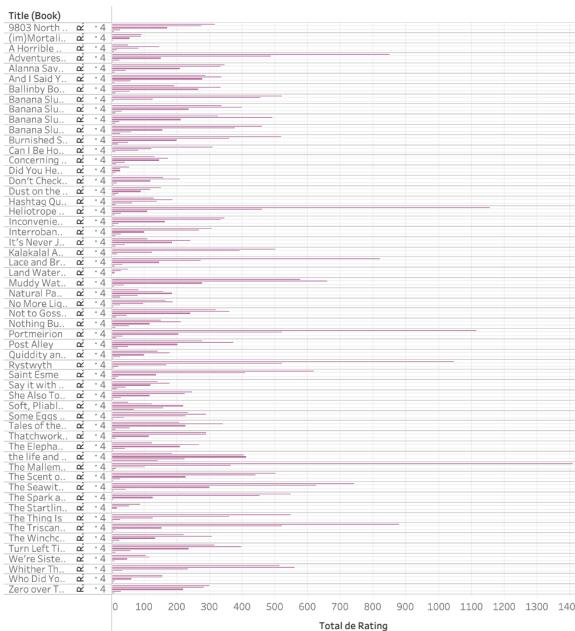
Distribution note | Livre



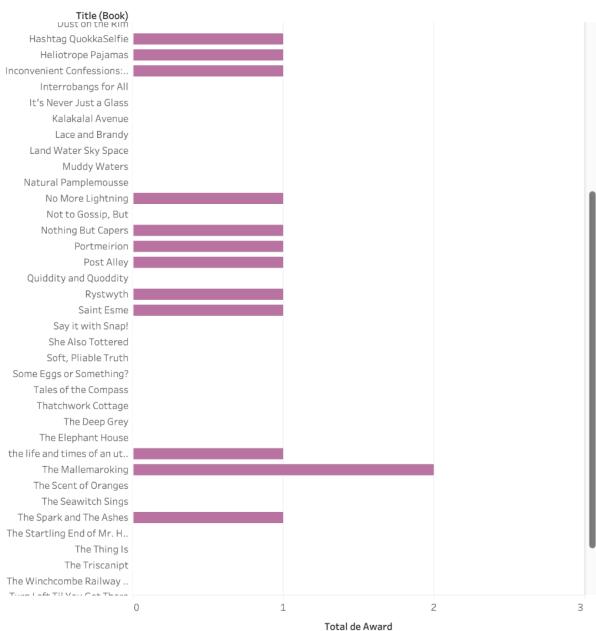
Pour ce qui est des différents livres, la distribution varie déjà plus ; sûrement en fonction de si le livre a plu ou non. On a par exemple *The Mallemaroking*, qui revient maintenant plusieurs fois comme étant le “meilleur” livre de notre base de données ; avec une très grande majorité de très bonnes notes.

A l’inverse, *The life and times of an utterly inconsequential person* n’a que peu de très bonnes notes ; mais plus de notes moyennes (en proportion). Pareil pour *Soft, Pliable Truth*.

Distribution note | Livre



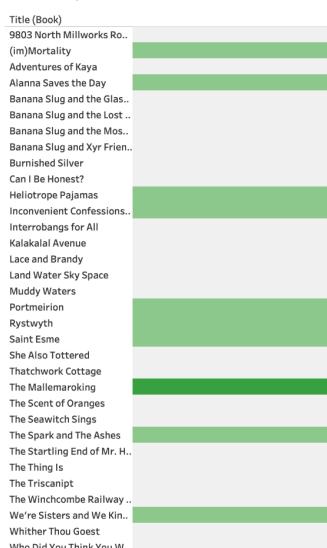
Distribution note | Récompense



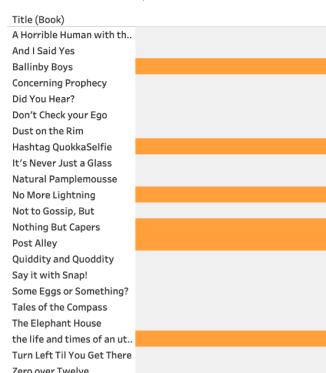
Maintenant on essaye de déterminer si cette distribution des notes est corrélée avec le nombre de récompenses du livre.

En mettant le nombre de récompenses et les distributions côté à côté, on voit qu'on ne repère pas bien, visuellement, si c'est corrélé. On préfère une heatmap ; celle-ci nous renseigne sur la question de savoir si la moyenne des notes pour un livre est corrélée avec son nombre de récompenses, et pas la répartition des notes. Mais c'est plus intelligible.

Note > 4 | Récompense



Note entre 3 et 4 | Récompense



Note < 3 | Récompense

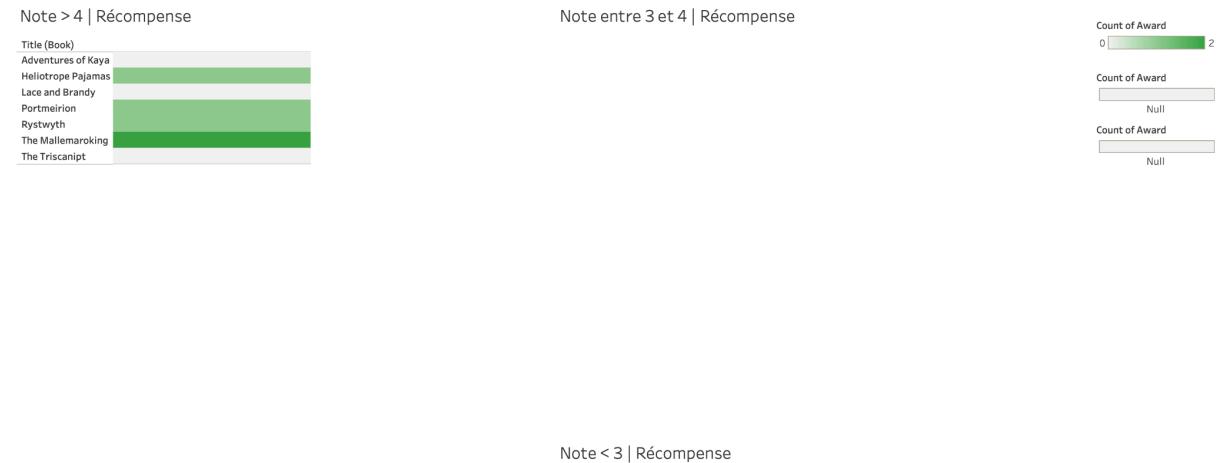


Ici, on met des filtres pour avoir uniquement les livres avec plus de 4 de moyenne dans une feuille, entre 3 et 4 dans un autre, et en dessous de 3 encore dans une autre feuille. La heat

map indique le nombre de récompenses. On met le tout dans un dashboard pour visualiser l'ensemble.

Ainsi, on peut voir dans la catégorie des livres qui ont eu des très bonnes notes, la proportion de ceux qui ont eu des récompenses.

Là, on voit que la proportion de récompenses dans les livres avec de très bonnes notes ( $> 5$ ) est à peu près la même que dans les livres avec de bonnes notes (entre 3 et 4).



On peut encore affiner notre visualisation en faisant un filtre progressif (les bornes  $> 5$ , entre 3 et 4, et  $< 3$  étant un peu arbitraires et générales).

On ajoute au dashboard un barplot des notes moyennes triées par ordre décroissant, puis on le définit comme filtre.

On peut ensuite sélectionner plusieurs livres, en partant du mieux noté, pour voir la progression du nombre de livres récompensés ou justement non récompensés.

- Comment calculez-vous le prix de vente, étant donné qu'il y a parfois, mais pas toujours, une remise accordée au moment de la vente ?

Pour cette question, on remarque d'abord qu'il peut y avoir plusieurs prix pour un même livre (en fonction des éditions) : on fait donc la moyenne de ces prix.

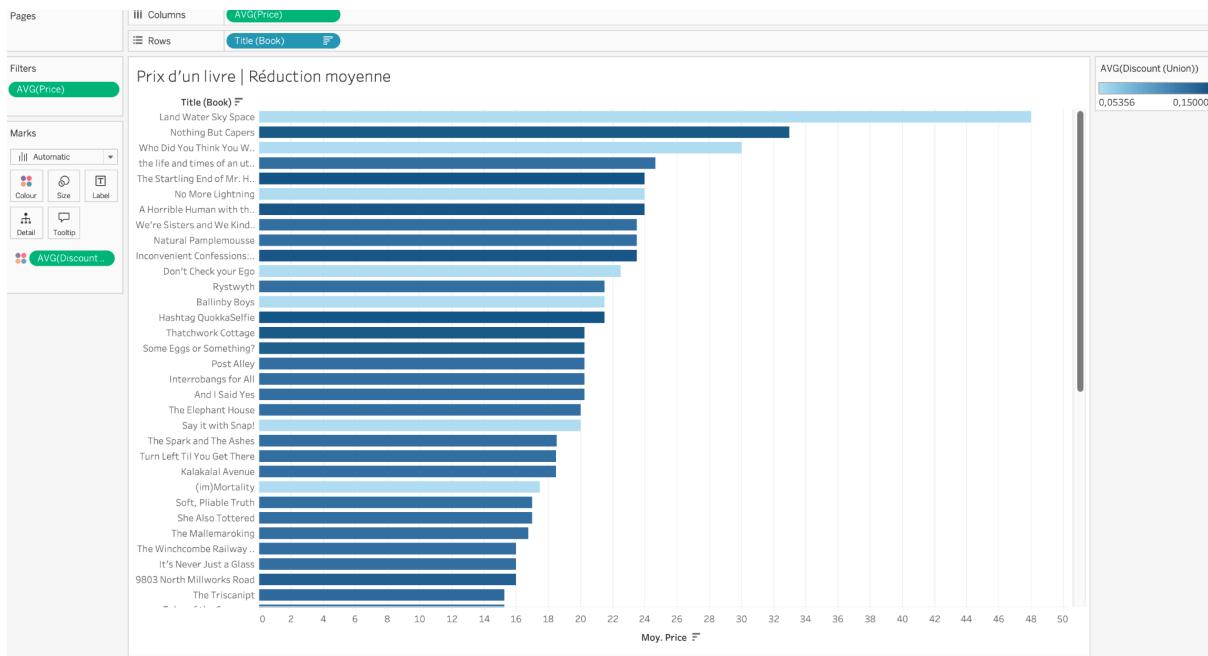


Ainsi, on peut considérer plusieurs choses. La somme dépensée pour le marketing du produit, (barplot de gauche) ; la moyenne des réductions (largeur des barres) ; et le nombre de ventes des livres déjà sorties (couleurs).

Avec ces visualisations, on ne voit pas de formule précise qui ressortirait pour fixer le prix d'un livre ; dont le succès peut être assez imprédictible, comme on le voit parfois avec certains livres avec beaucoup de tirages mais peu de vente etc.

Néanmoins, on pourrait fixer un prix un peu plus élevé pour les livres avec le plus de réduction en moyenne.

Considérons le graphique suivant :



On met un dégradé de couleur pour la réduction moyenne de chacun ; avec une couleur foncée signifiant une réduction moyenne plus élevée.

On constate que même en triant les livres par ordre décroissant de prix, la réduction moyenne reste assez disparate.

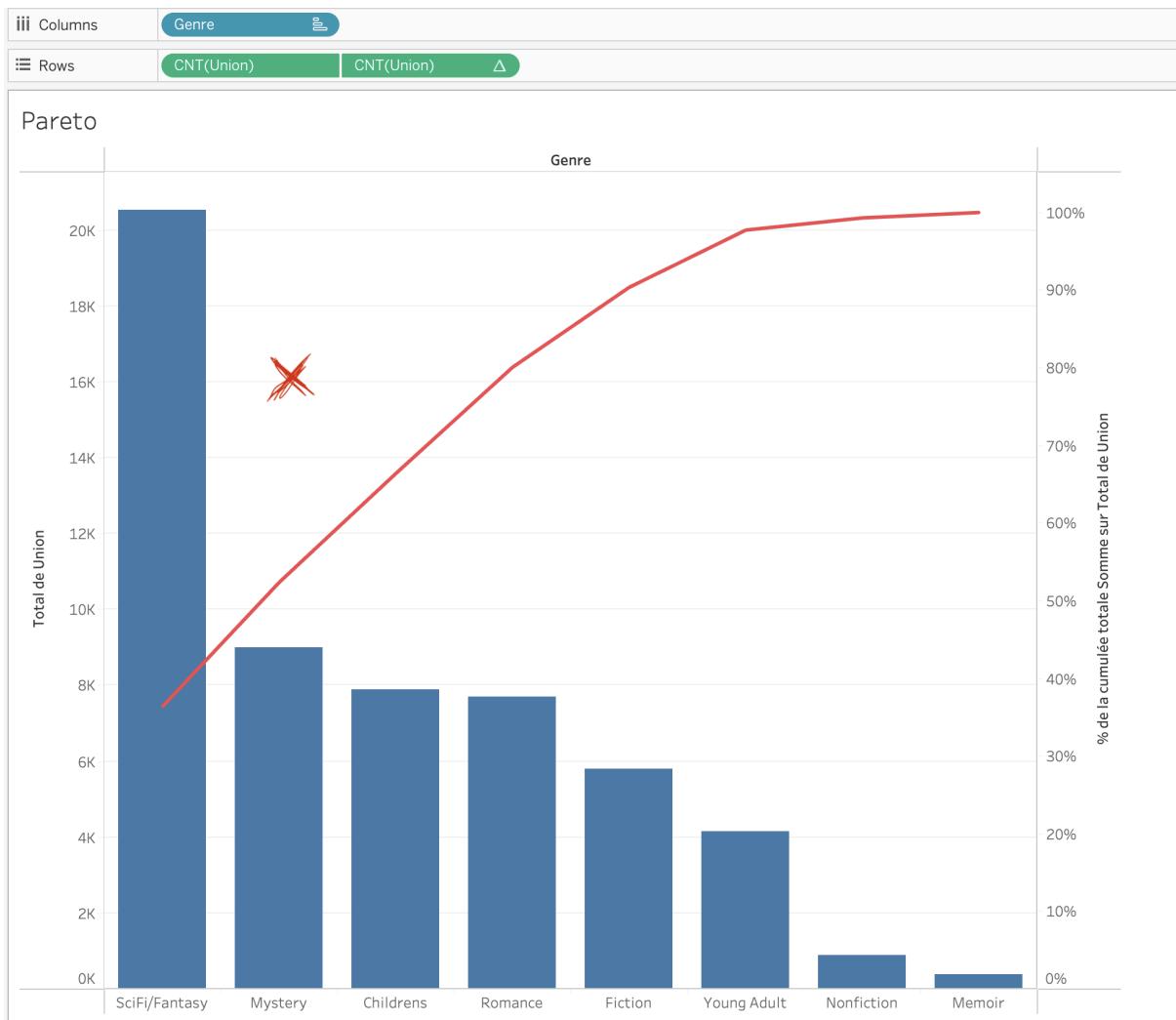
On pourrait faire en sorte d'avoir un dégradé ; où plus le livre est en réduction plus il est cher.

- Les ventes se rapprochent-elles du principe de Pareto ?

Le principe de Pareto énonce que 80% des conséquences sont dues à 20% de la totalité des causes. Autrement dit, un phénomène peut s'expliquer majoritairement par une petite portion de tout ce qui l'influence.

Pour étudier ce phénomène sur Tableau, on se sert de ce tutoriel :  
<https://help.tableau.com/current/pro/desktop/fr-fr/pareto.htm>

On le suit pour les ventes en fonction des genres :



Expliquons ce graphique. Nous avons trié les ventes par ordre décroissant, et avec leur effectif cumulé (les barres bleues) selon le genre. La courbe rouge correspond au pourcentage total de contribution aux ventes de chaque genre.

On a 8 genres, donc 20% de cette cause revient à dire environ 2 genres. On met une croix rouge là où la courbe est censée passer si le principe de Pareto est respecté (donc 80% de contribution aux ventes en ordonnée à 20% des causes en abscisses).

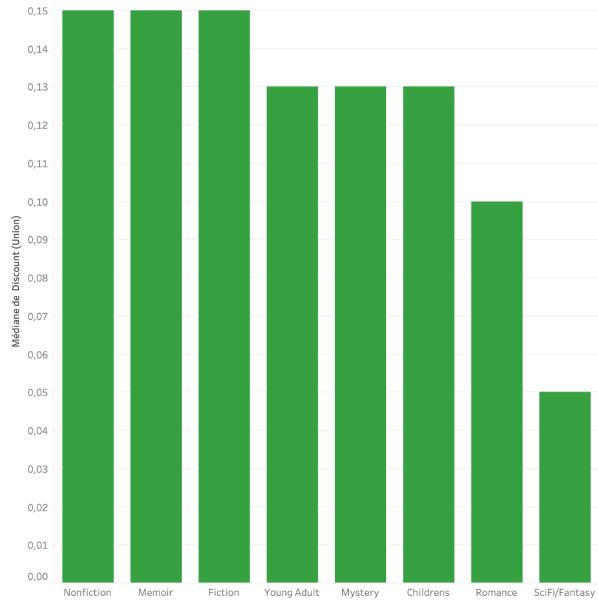
On voit qu'ici ce n'est pas le cas.

- Y a-t-il des modèles dans les remises ?

Tout d'abord, et comme on l'a vu précédemment, le prix ne semble pas influer sur les remises.

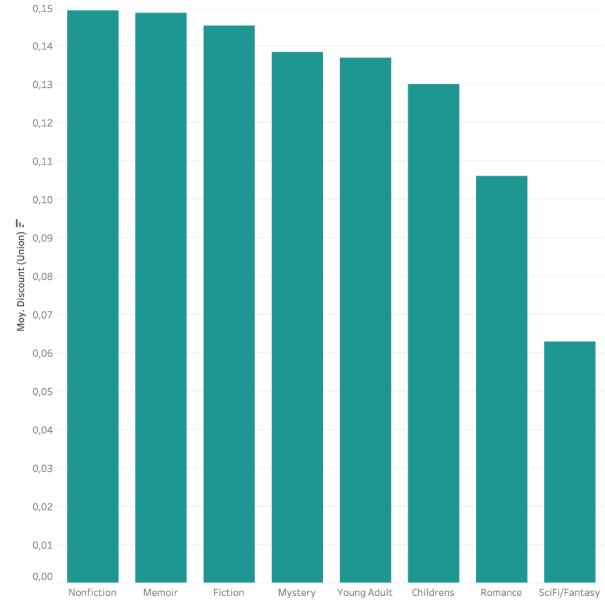
Tendance remise | Médiane

Genre



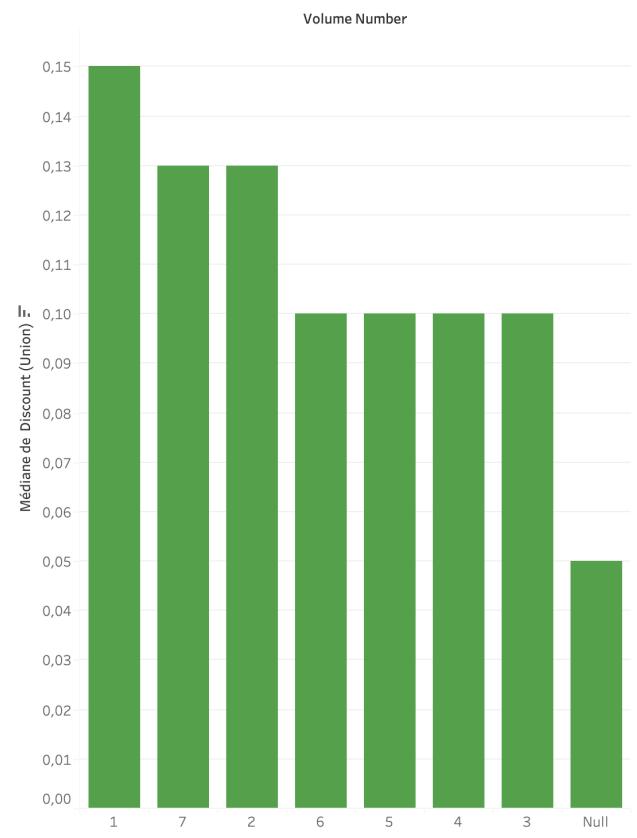
Tendance remise | Moyenne

Genre

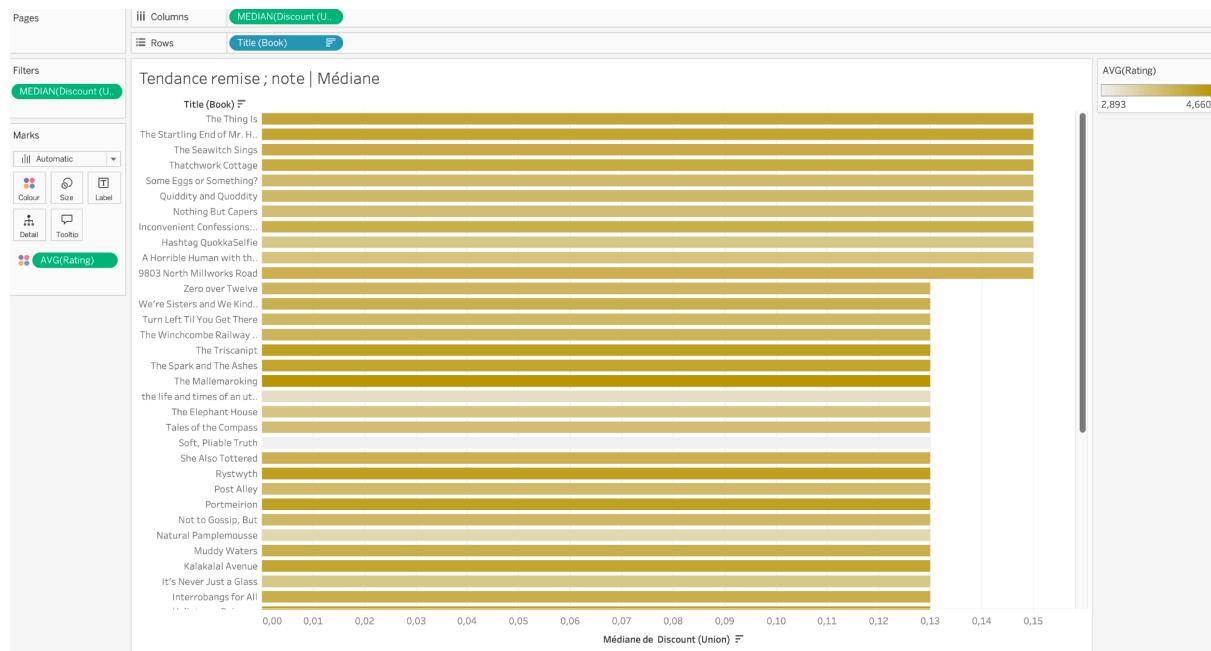


Pour les genres, on voit que le Scifi/fantaisie est beaucoup moins remisé que le reste des genres.

### Tendance remise ; numéro du volume | Médiane



On pourrait penser que les remises changent en fonction du volume de la série de livres. On voit que les premiers volumes sont les plus remisés (peut-être pour encourager les lecteurs à se lancer dans la série).



Pour finir, on fait en fonction des notes : aucune tendance ne semble se dégager.

On pourrait encore investiguer les remises avec d'autres variables (taille du tirage, format du livre etc.)

- Certaines tables en particulier semblent-elles contenir des données sales ?

On a effectivement rencontré pas mal de données manquantes pendant le travail.