**Ensemble de données EDA Top 50 des livres les plus vendus d’Amazon de 2009 à 2019** 

Dans ce travail, une analyse exploratoire des données a été réalisée sur l’ensemble de données des 50 meilleurs livres les plus vendus d’Amazon entre 2009 et 2019.

**CARACTÉRISTIQUES :**

1. Nom - Nom du livre

2. Auteur - L’auteur du Livre

3. Note de l’utilisateur - Amazon User Rating

4. Avis - Nombre d’avis écrits sur amazon

5. Prix - Le prix du livre (Au 13/10/2020)

6. Année - L’année(s) où elle s’est classée sur le best-seller

7. Genre - Qu’il s’agisse de fiction ou de non-fiction

**RÉPONSE AUX QUESTIONS SUIVANTES SONT DONNÉES :**

1. Quels livres d’auteur reçoivent la note moyenne la plus élevée (meilleurs auteurs).

2. Quel auteur a écrit le plus de best-sellers (meilleurs auteurs).

3. Quel livre a le plus de critiques (meilleurs livres).

4. Quels genres deviennent des best-sellers plus souvent.

5. Les genres sont-ils différents selon l’évaluation.

Vous allez me suivre étape par étape pour mieux comprendre mon analyse sur **R studio**. C’est parti !!

1. **DESCRIPTION DES DONNÉES ET NETTOYAGE DES DONNÉES**

Dans ce bloc, une analyse exploratoire des données sera réalisée, les types de données, les valeurs manquantes, les doublons, les mesures de la tendance centrale, ainsi que la relation entre les variables seront décrits

D’abord il faut choisir les bonnes librairies et importer les données.

Une image contenant texte, capture d’écran, affichage, Police

Le contenu généré par l’IA peut être incorrect.



Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Les données contiennent 547 enregistrements et 7 colonnes.

Les données contiennent 3 colonnes de catégorie et 4 colonnes contiennent des valeurs numériques. Convertissons le genre en type de données de catégorie.



**Valeurs manquantes :**

Calculons le pourcentage de valeurs vides et remplies pour toutes les colonnes.

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, Police, capture d’écran

Le contenu généré par l’IA peut être incorrect.

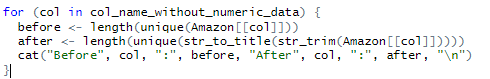
Les données n’ont pas de valeurs manquantes, donc aucune autre transformation n’est nécessaire

**Doublons :**

**Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.**

Vérifiez si les colonnes contiennent des mots dans différents cas et des espaces supplémentaires.



Une image contenant texte, Police, blanc, algèbre

Le contenu généré par l’IA peut être incorrect.

En effet, il y a des erreurs de frappe dans la colonne Nom. Corrigeons-les.



On remarque qu’il y a des erreurs de saisie sur JK Rowling et RR martin. Nous devons les corriger



**Colonne Year**

De plus, il est à noter qu’il y a des lignes en double dans les données, mais avec des années différentes. Supprimons les doublons. Laissons les données sans années pour une analyse ultérieure.



**Name**

Les enregistrements sont dupliqués par prix. Apparemment, dans différentes années, le livre coûte différemment, ce qui est assez logique en tenant compte de l’inflation, de la demande, etc. Laissons seulement les dernières entrées.



Ainsi, les données contiennent 350 livres différents écrits par 247 auteurs. Tous les livres sont présentés dans deux catégories Non Fiction et Fiction.

1. **Données catégorielles**

Dans cette section, nous analyserons des données catégorielles et répondrons aux questions suivantes :

Quels livres d’auteur reçoivent la note moyenne la plus élevée (meilleurs auteurs).

Quel auteur a écrit le plus de best-sellers (meilleurs auteurs).

Quels genres deviennent des best-sellers plus souvent.

Quel livre a le plus de critiques (meilleurs livres).

Créons de nouveaux dataframes basés sur la note moyenne des auteurs, le nombre de livres écrits par les auteurs, le nombre de critiques pour les livres et le nombre de livres par genre. Nous visualisons les informations reçues.

Une image contenant texte, Police, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.

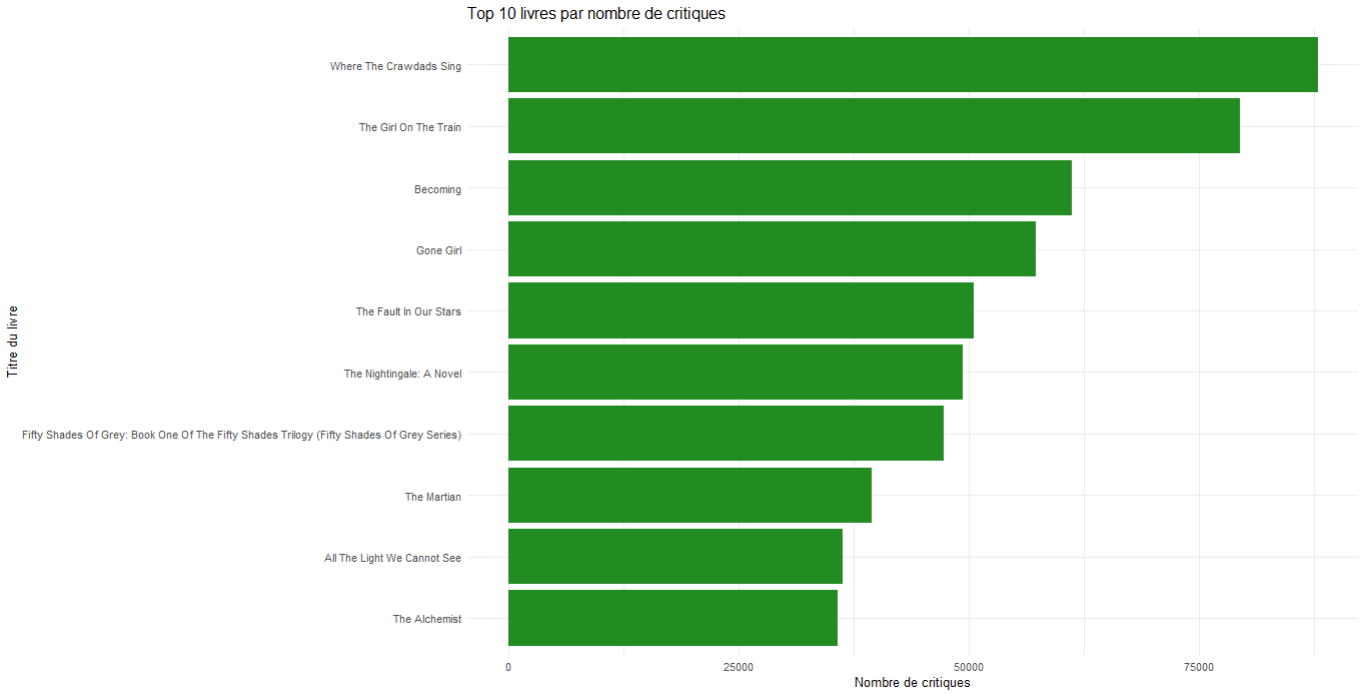
Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, nombre

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, affichage

Le contenu généré par l’IA peut être incorrect.

En analysant les données catégorielles, il est établi :

Les 13 auteurs suivants ont la meilleure note : Nathan W. Pyle, Patrick Thorpe, Eric Carle, Emily Winfield Martin, Chip Gaines, Jill Twiss, Rush Limbaugh, Sherri Duskey Rinker, Alice Schertle, Pete Souza, Sarah Young, Lin-Manuel Miranda, Bill Martin Jr., Dav Pilkey. La note moyenne pour leurs œuvres était de 4,9. Lors de l’achat d’un nouveau livre, vous devriez faire attention à ces auteurs.

Auteurs qui ont écrit plus de best-sellers : Jeff Kinney - 12 livres, Rick Riordan - 10 livres, J.K. Rowling - 8 livres, Stephenie Meyer - 7 livres, Dav Pilkey - 6 livres, Bill O’Reilly - 6 livres, John Grisham - 5 livres, E L James - 5 livres, Suzanne Collins - 5 livres, Charlaine Harris - 4 livres. Ces auteurs ont toujours quelque chose à lire.

Les livres avec le plus d’avis : Where The Crawdads Sing - 87841 avis, La fille du train - 79446 avis, Becoming - 61133 avis, Gone Girl - 57271 avis, The Fault In Our Stars - 50482 avis. Ça vaut vraiment la peine de lire le livre Where The Crawdads Sing, ce n’est pas pour rien qu’il fait le plus parler.

La non-fiction a plus de chances de devenir un best-seller. Plus tard, nous découvrirons comment les utilisateurs se rapportent à ces genres.

1. **Donnée numérique (Box plot)**

**Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.**

**Une image contenant texte, diagramme, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.**

**Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.**

**Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.**

En analysant les données numériques, il est établi :

Évaluation des utilisateurs :

* Les données ne sont pas distribuées normalement. Une asymétrie est observée.
* Les notes moyennes et médianes pour les livres sont de 4,6.
* Il y a des valeurs aberrantes dans les données. Il y a un petit nombre de livres dans les données en dessous de la note de 4,1.

Avis:

* Les données ne sont pas distribuées normalement. Une asymétrie est observée.
* Les données ont une large plage.
* Il y a des valeurs aberrantes dans les données. Il y a un petit nombre de livres recevant des critiques bien au-dessus du 75e centile.

Prix :

* Les données ne sont pas distribuées normalement. Une asymétrie est observée.
* Il y a des livres qui coûtent beaucoup plus cher que la moyenne, ainsi que des livres avec un coût de 0, ce qui est étrange. Il y a des livres qui coûtent beaucoup plus cher que moyenne, ainsi que des livres avec un coût de 0, ce qui est étrange. Soit le livre est donné gratuitement, soit cette erreur.

**Regardons la corrélation dans les données numériques**

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, Police, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, nombre, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, nombre, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Tracé, diagramme

Le contenu généré par l’IA peut être incorrect.

Sur la base de la matrice de corrélation construite ainsi que des visualisations construites, on peut voir que les données ne contiennent aucune relation linéaire positive ou négative entre la notation, les avis et le prix des livres.

**IV.HYPOTHÈSES DE TEST**

Dans cet article, testons l’hypothèse suivante : « Les genres diffèrent-ils en termes de notation ? ».

Nous testerons l’hypothèse selon l’algorithme suivant :

* La distribution a été vérifiée pour la normalité à l’aide du « test de normalité de Shapiro Wilk ».
* Deux échantillons ont été formés. Des livres liés au genre Non Fiction et des livres liés au genre Fiction.
* Des tests ont été effectués afin d’identifier les différences statistiques entre les groupes d’étude.

Formulons les hypothèses nulles et alternatives.

H0 - Il n’y a pas de différences entre les genres

H1 - Il y a des différences entre les genres

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Résultat :



Formons deux échantillons pour les tests.



Réalisons des tests. Puisque les données dans les échantillons ne sont pas normalement distribuées, nous utiliserons le test non paramétrique de Mann-Whitney.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Suite aux tests, des différences statistiquement significatives ont été obtenues entre les groupes comparés. Sur la base des résultats obtenus, on peut dire que les utilisateurs évaluent les livres différemment selon le genre, et en fonction des valeurs médianes, les lecteurs donnent la préférence aux œuvres de fiction.

**Conclusion**

Au cours de l’analyse, il a été établi quels auteurs reçoivent les meilleures notes des lecteurs, quels auteurs ont écrit le plus de best-sellers, quels livres reçoivent le plus d’avis des lecteurs. En outre, il a été constaté que la littérature non romanesque devient plus souvent un best-seller, mais les utilisateurs aiment aussi davantage la fiction, ce qui est confirmé par les résultats statistiquement significatifs obtenus lors des tests.