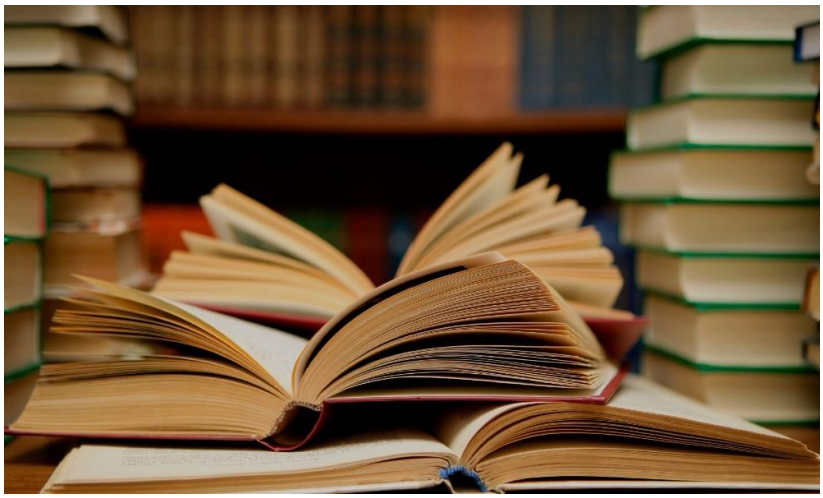
RAPPORT STA101



Utilisation de l'analyse de données pour étudier les propriétés du dataset de Goodreads et mettre en évidence des relations pertinentes entre les variables.

Sarah BITAN & Shmuel BITAN
Seconde session

2023-2024 SEMESTRE 2

PRESENTATION

Bien que les livres ne soient pas une invention récente, leur nature a été profondément transformée par l'évolution numérique. Dans un monde où l'abondance de livres rend impossible la lecture de tous les ouvrages disponibles, des listes de lecture et des recommandations se révèlent indispensables. C'est ici qu'intervient Goodreads, un réseau social détenu par Amazon, reconnu comme "le plus grand site au monde pour les lecteurs et les recommandations de livres" (Goodreads, 2022). Ce site nous permet de partager nos avis, de créer des listes de lecture personnalisées, et de découvrir de nouveaux ouvrages, tout en suivant nos lectures et en consultant les évaluations d'autres lecteurs.

En tant que lecteurs assidus, nous nous sommes souvent demandé s'il existe un moyen de guider nos choix subjectifs en nous basant sur des faits concrets. Quels genres sont particulièrement bien notés ? Faut-il privilégier une certaine époque de publication ? Quel rôle jouent les récompenses littéraires dans l'appréciation des livres ?

Pour répondre à ces questions, nous avons utilisé un ensemble de données disponible publiquement la liste "Best Books Ever" de Goodreads, qui comprend plus de 50 000 livres basés sur les votes de la communauté Goodreads. Ce dataset fournit des informations sur les évaluations et les critiques, ainsi que des données clés sur chaque livre.

L'objectif de ce projet est d'explorer les données de Goodreads pour mieux comprendre les relations entre différentes variables du dataset. En analysant les évaluations agrégées, le nombre de pages, la popularité des auteurs, et la distribution des langues, nous pourrons dégager des tendances et des modèles dans l'industrie du livre et les préférences des lecteurs. Ce type d'analyse peut nous aider à faire des choix éclairés sur les livres à lire ou à recommander, et fournir des informations précieuses aux éditeurs, auteurs, et libraires.

Dans ce contexte, l'utilisation de techniques comme le K-Means nous permet d'identifier des motifs et des relations au sein des données, en regroupant des éléments similaires en clusters. Cette approche nous permet de segmenter la liste en sous-groupes cohérents, facilitant ainsi l'exploration de ces vastes collections de livres.

Nous allons vous présenter nos différentes variables, dont 8 quantitatives, 2 de type date et 16 qualitatives. :

- **Series** : La série à laquelle le livre appartient, le cas échéant.
- **Author** : L'auteur du livre.
- **Rating** : La note moyenne du livre sur GoodReads.
- **Description** : Une brève description du livre.
- **Language** : La langue dans laquelle le livre est écrit.
- **ISBN** : Le numéro ISBN du livre.
- **Genres** : Les genres auxquels le livre appartient.
- **Characters** : Les personnages présents dans le livre.
- **BookFormat** : Le format du livre (par exemple, broché ou livre électronique). **Edition** : L'édition du livre.
- **Pages** : Le nombre de pages du livre.
- **Publisher** : L'éditeur du livre.
- **PublishDate** : La date de publication du livre.
- **FirstPublishDate** : La date de première publication du livre.
- **Awards** : Les récompenses que le livre a reçues.
- **NumAwards** : Le nombre de récompenses que le livre a reçues (variable crée dans le cadre de notre étude).
- **NumRatings** : Le nombre de notes que le livre a reçues.
- **RatingsByStars**: La répartition des notes du livre par étoiles.
- **LikedPercent** : Le pourcentage de lecteurs qui ont aimé le livre.
- **Setting** : Le cadre du livre.
- **coverImg** : L'image de couverture du livre.
- **BbeScore** : Le score du livre sur la liste des meilleurs livres de GoodReads. **BbeVotes** : Le nombre de votes pour le livre sur la liste des meilleurs livres de GoodReads.
- **Price** : Le prix du livre.
- **Month** : Le mois de publication du livre (variable crée dans le cadre de notre étude).

I. ANALYSE DU JEU DE DONNEES

A – NETTOYAGE DES DONNEES

Nous avons commencé par nettoyer notre jeu de données de 52478 lignes et 25 colonnes. Nous vérifions la présence de doublons en se basant sur les variables ‘title’ et ‘author’, puis nous les supprimons. Une fois la suppression des doublons effectué, il reste 52390 livres.

Après avoir calculé les statistiques descriptives de nos données, telles que les valeurs minimales, maximales, moyennes, etc., nous avons conclu qu'il n'y a pas de valeurs aberrantes. Par exemple, il n'y a pas de notes inférieures à 0 ou de prix négatifs, on peut alors passer à notre analyse de données.

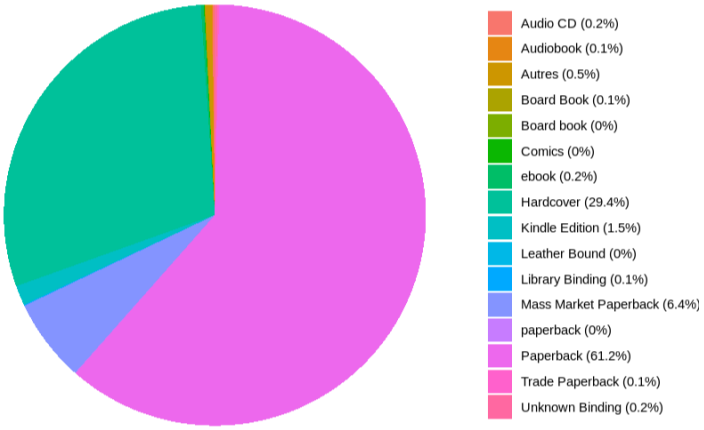
B – ANALYSE UNIVARIEE

Pour réaliser une analyse univariée de notre jeu de données, nous allons examiner la dispersion de chaque variable et les afficher de manière différente pour les variables qualitatives et quantitatives. Tout d'abord, nous identifierons les variables qualitatives, qui décrivent des catégories (par exemple, les genres littéraires), et les variables quantitatives, qui représentent des quantités ou des mesures (par exemple, le nombre de pages). Pour les variables qualitatives, nous calculerons les fréquences et les proportions pour chaque catégorie, puis les afficherons sous forme de tableaux de fréquences ou de graphiques en barres. Pour les variables quantitatives, nous calculerons les statistiques descriptives telles que la moyenne, la médiane, l'écart-type, le minimum et le maximum, et afficherons la distribution des données à l'aide d'histogrammes ou de densités. Cette approche nous permettra de mieux comprendre la répartition et les caractéristiques de chaque variable dans notre jeu de données.

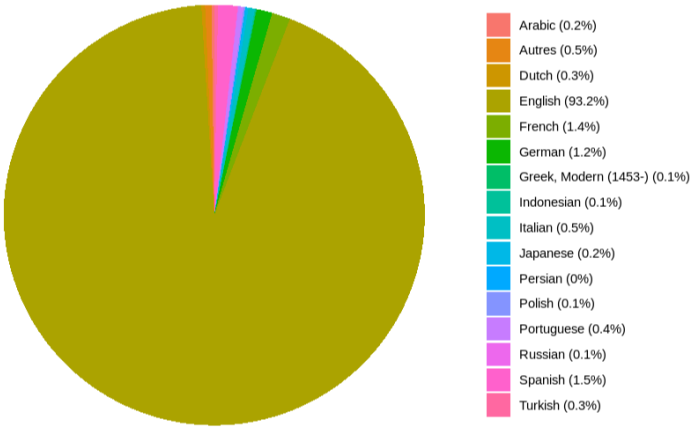
Pour poursuivre notre analyse univariée, nous allons maintenant intégrer les graphiques en camembert que nous avons réalisés pour les variables qualitatives `bookFormat, language et Edition`. Ces graphiques nous permettent de visualiser la répartition des différentes catégories au sein de ces variables.

1 ANALYSE DES VARIABLES : REPRESENTATION A L’AIDE DE CAMEMBERT

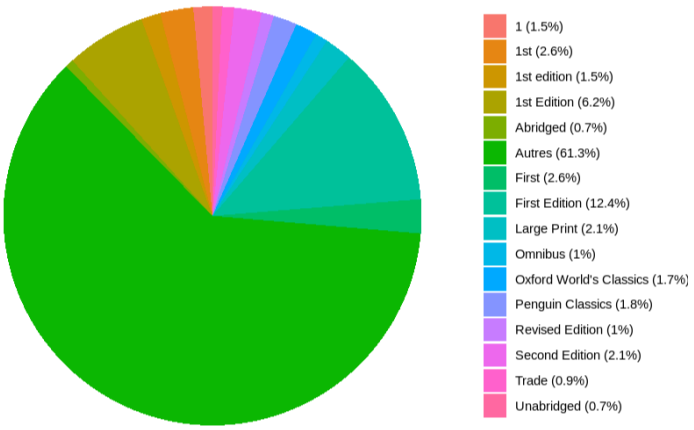
Distribution de bookFormat



Distribution de language



Distribution de edition



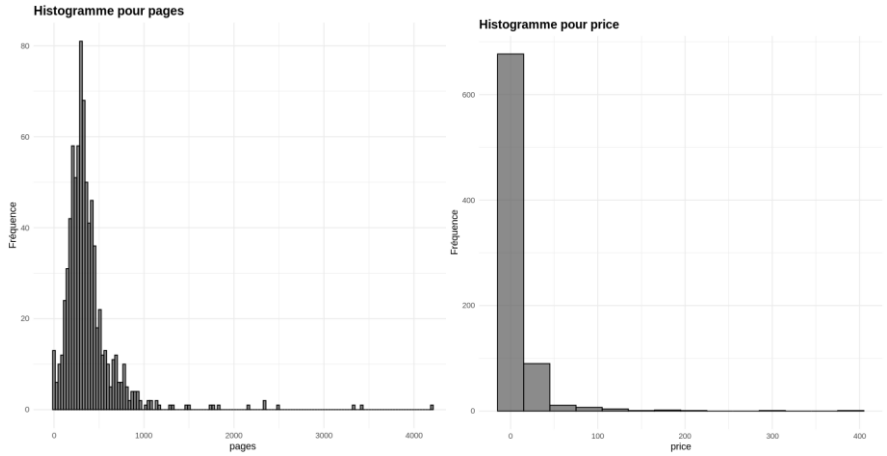
Les résultats des camemberts montrent une forte concentration des catégories "Autres" pour plusieurs variables, ce qui indique une distribution très dispersée parmi de nombreuses petites catégories. Pour la variable bookFormat, la majorité des livres sont en format "Paperback" (69,2%), suivis par "Hardcover" (21,5%) et "Mass Market Paperback" (6,6%), tandis que les autres formats sont nettement moins représentés.

La variable language montre une domination écrasante de l'anglais, qui représente 92,8% des livres, tandis que l'espagnol, le français, et l'allemand suivent à distance avec environ 1,6% chacun. Les autres langues représentent des proportions encore plus faibles.

Enfin, la variable Edition, bien que la majorité des éditions soit regroupée dans "Autres" (75,9%), certaines éditions spécifiques comme "First Edition" (6%) et "1st Edition" (3,5%) se démarquent légèrement.

Nous allons également réaliser une analyse univariée sur les variables quantitatives en créant des histogrammes. Cela nous permettra d'observer la répartition des valeurs de chaque variable.

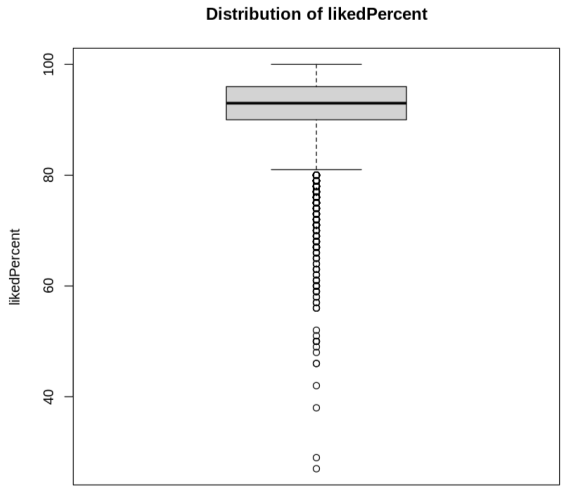
2 ANALYSE DES VARIABLES : REPRESENTATION SOUS FORME D'HISTOGRAMME

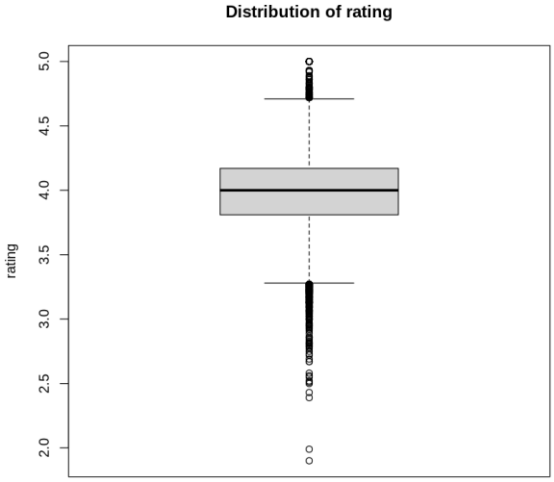


L'histogramme de la variable `pages` montre que la majorité des livres du dataset ont entre 0 et 1000 pages, avec un pic notable autour de 500 pages. Cela suggère que les livres de longueur moyenne sont les plus courants, ce qui pourrait refléter des normes éditoriales ou des préférences des lecteurs pour des livres d'une taille modérée. En ce qui concerne la variable `Price`, l'histogramme indique que la plupart des livres sont vendus entre 0 et 25 dollars, avec une concentration marquée dans cette fourchette de prix. Les livres dont le prix dépasse 50 dollars sont beaucoup moins fréquents, ce qui pourrait signifier une préférence des consommateurs pour des livres plus abordables ou une stratégie des éditeurs visant à maintenir des prix accessibles.

3 ANALYSE DES VARIABLES : REPRESENTATION A L'AIDE DE BOXPLOT

Pour compléter notre analyse univariée sur les variables qualitatives, nous avons décidé de visualiser certaines variables quantitatives à l'aide de boxplots, qui offrent une meilleure lisibilité et permettent de voir clairement la répartition des données, notamment en présence de valeurs extrêmes. Nous avons ainsi créé des boxplots pour les variables rating, likedPercent et num_awards. Ces visualisations permettent de mieux comprendre la distribution des données, en mettant en évidence la médiane, les quartiles et les valeurs aberrantes, et elles nous aident à identifier d'éventuelles particularités ou asymétries dans les caractéristiques des livres présents dans notre dataset goodreads_data.





En analysant les boxplots des variables `likedPercent, rating` et `num_awards`, nous observons des tendances distinctes qui méritent attention. Pour la variable `likedPercent`, les valeurs s'étendent de 80 à 100, avec une médiane située à 95. Cette distribution montre une forte concentration de données entre 95 et 100, suggérant que la plupart des livres sont largement appréciés par les lecteurs. Cette concentration élevée pourrait indiquer une tendance générale des utilisateurs de Goodreads à évaluer positivement les livres qu'ils lisent, ou bien un biais vers la notation élevée sur cette plateforme.

En ce qui concerne la variable `rating`, les valeurs varient entre 3,25 et 4,75, avec une médiane à 4. La distribution montre une concentration significative de points autour des valeurs extrêmes, entre 4,75 et 5, ainsi qu'entre 3,25 et 2,25. Cette répartition pourrait suggérer que les livres évalués sur Goodreads tendent à polariser les lecteurs, avec une forte proportion de notes très élevées ou relativement basses, tandis que les notes moyennes sont moins fréquentes. Cette polarisation pourrait refléter des préférences marquées chez les utilisateurs, où les livres populaires obtiennent des notes très élevées, et les moins appréciés, des notes plus basses.

4 CONCLUSION

L'analyse univariée a révélé des tendances importantes dans notre dataset. La majorité des livres sont en anglais, et les formats physiques comme le "Paperback" dominent largement. Les livres tendent à avoir entre 0 et 1000 pages, avec un pic autour de 500 pages, et sont principalement vendus à des prix abordables, entre 0 et 25 dollars. Ces résultats montrent une préférence pour des livres de taille modérée, en format physique, et abordables, ce qui semble correspondre aux attentes générales des consommateurs.

C – ANALYSE BIVARIEE

Pour approfondir notre compréhension des relations entre les variables, nous allons réaliser une analyse bivariée. Cette approche nous permettra **d'examiner** les liens et les influences réciproques entre deux variables à la fois. Nous commencerons par créer une matrice de corrélation pour les variables quantitatives, qui nous fournira une vue d'ensemble des indices de corrélation entre chaque paire de variables.

I MATRICE DE CORRELATION

Une matrice de corrélation montre la force et la direction des relations linéaires entre les variables quantitatives.

En examinant cette matrice de corrélation, nous pouvons identifier des corrélations significatives entre certaines variables :

- **`Rating` et `likedPercent`** : La corrélation est de 0,84, indiquant une forte corrélation positive. Cela suggère que les livres ayant de meilleures évaluations sont également plus susceptibles d'être aimés par les utilisateurs.
- **`bbeScore` et `num_ratings`** : La corrélation est de 0,71, montrant également une forte corrélation positive. Cela signifie que les livres avec des scores de Best Book Ever (BBE) plus élevés tendent à avoir plus de notations.
- **`numRatings` et `likedPercent`** : La corrélation est très faible, proche de 0, ce qui indique qu'il n'y a pas de relation linéaire significative entre le nombre de notations et le pourcentage de likes.
- **`Price` et `bbeVotes`** : La corrélation est également très faible, proche de 0, ce qui montre peu de relation entre le prix des livres et le nombre de votes pour le Best Book Ever.
- **`num_awards et Price`** : La corrélation négative de -0,03 entre le nombre de récompenses et le prix signifie que, dans notre étude, il existe une très faible relation inverse entre ces deux variables. En d'autres termes, lorsque le nombre de récompenses augmente, le prix tend légèrement à diminuer, il n'y a pas de lien significatif entre le nombre de récompenses qu'un livre reçoit et son prix.
- **`num_awards et numRatings`** : Le coefficient de corrélation de 0,31 entre le nombre de notations (`numRatings`) et le nombre de récompenses (`num_awards`) indique une relation positive mais faible. Cela signifie que bien que les livres récompensés aient tendance à recevoir plus de notations, d'autres facteurs influencent probablement davantage l'engagement des lecteurs sur Goodreads.

En résumé, cette analyse de corrélation nous permet d'identifier les relations importantes entre les variables et d'explorer plus en détail les interactions potentielles. Les corrélations fortes comme celle entre `avg_rating` et `likedPercent` ou entre `bbeScore` et `num_ratings` indiquent des liens significatifs qui méritent une investigation plus approfondie. D'autre part, les corrélations faibles nous indiquent quelles paires de variables n'ont pas de relation linéaire notable.

Pour approfondir notre analyse, nous allons explorer des problématiques spécifiques qui nous permettront d'observer les liens entre les variables et de tirer des informations pertinentes. Nous commencerons par découvrir quels sont les livres les mieux notés.

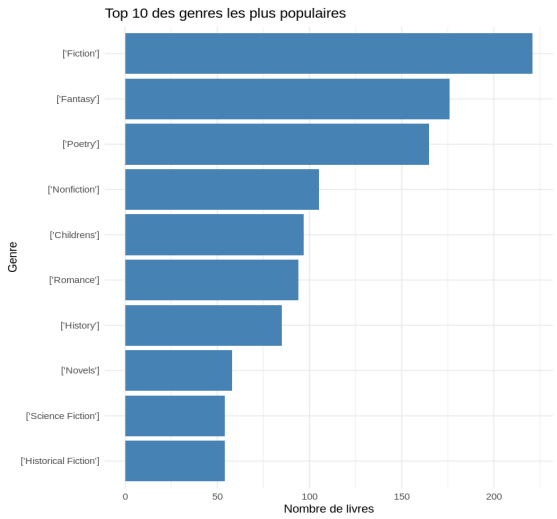
2 QUESTIONS PERTINENTES

Quels sont les variables impactant la réception de prix littéraire ?

D’après la matrice de corrélation, l’attribution d’un prix littéraire dépend principalement du nombre de votes ainsi que des variables bbe_score et bbe_vote, indiquant que la popularité et la reconnaissance publique jouent un rôle crucial dans la réception de récompenses littéraires

Quel genre est le plus populaire en termes de nombre de livres ?

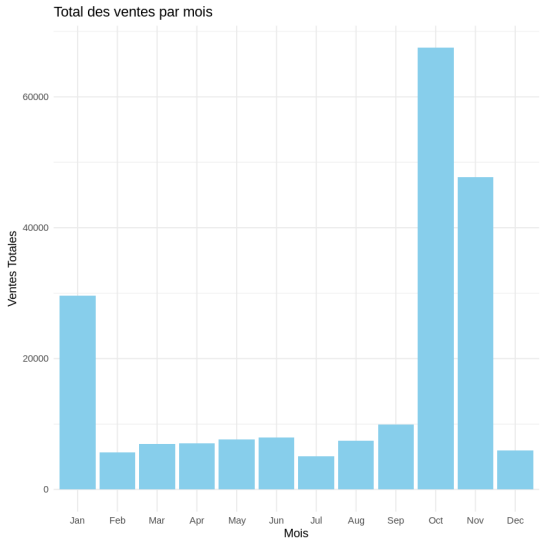
Nous allons maintenant examiner la répartition de la note moyenne des livres par genre. Pour cela, nous afficherons les 10 genres les plus populaires en termes de nombre de livres et nous analyserons leurs notes moyennes.



Les genres de livres les plus populaires en termes de nombre de livres dans notre dataset sont les suivants : fiction, fantasy, poetry, non-fiction, children, romance, history, novel, science-fiction, et historical fiction.

Quel mois génère le plus de ventes totales de livres ?

Pour explorer l'impact des mois de publication sur les ventes totales de livres, nous avons réalisé une analyse détaillée en examinant la relation entre le mois de publication et le nombre de livres vendues. Nous avons converti les dates de publication (`firstPublishDate`) au format Date et extrait les mois correspondants. Après avoir nettoyé les données, nous avons agrégé les ventes totales par mois et représenté ces informations à l'aide d'un barplot.



Pour finir, nous allons examiner la relation entre le mois de publication et le nombre total de ventes de livres afin de déterminer quel mois génère le plus de revenus. Nous avons réalisé un barplot montrant les ventes totales par mois. Il est notable qu'en octobre, la majorité des ventes ont été réalisées, suivies par les mois de novembre, janvier, et septembre. On observe que le nombre de livres vendus est beaucoup plus faible mais assez similaire pour les mois de février, mars, avril, mai, juin, août, septembre, et décembre. Le mois le moins performant en termes de ventes est juillet.

Une hypothèse pour les ventes élevées en janvier pourrait être la période de soldes, ce qui incite les consommateurs à acheter davantage. Bien que des soldes puissent également avoir lieu en juillet, il est possible que les vacances d'été influencent le comportement d'achat. Pendant cette période, les gens pourraient lire moins ou profiter autrement de leurs vacances, ce qui pourrait expliquer la baisse relative des ventes durant ce mois.

3 CONCLUSION

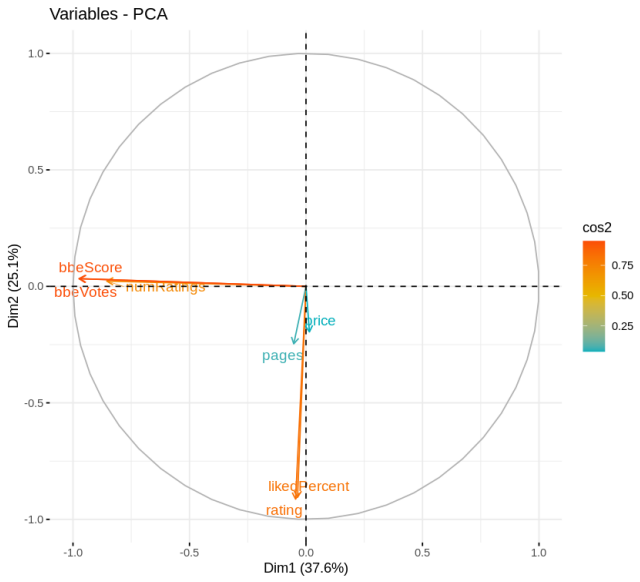
L'analyse bivariée a mis en évidence des relations significatives entre certaines variables. Notamment, une forte corrélation entre la note moyenne (rating) et le pourcentage de "liked" (likedPercent) indique que les livres bien notés sont généralement plus appréciés. En revanche, la relation entre le prix des livres et leur popularité (mesurée par les votes BBE) est faible, ce qui suggère que le prix n'est pas un facteur déterminant dans la popularité d'un livre. On constate aussi une faible corrélation entre num_rewards et rating ce qui indique que bien que certains livres récompensés soient bien notés, il n’y a pas de lien direct entre la note et le fait qu’un livre soit primé

II. APPORT D'ANALYSE EN COMPOSANTES PRINCIPALES ACP NORMEE

Nous avons entrepris une Analyse en Composantes Principales (ACP) dans le but de mieux comprendre les relations entre les variables de notre jeu de données et de réduire sa dimensionnalité tout en conservant l'essentiel de l'information. Cette technique nous permet d'identifier les dimensions principales, ou axes, qui capturent le plus de variance dans les données.

A - ANALYSE DU CERCLE DE CORRELATIONS DES VARIABLES ET PREMIERES CONCLUSIONS

L'ACP transforme nos variables originales en un nouvel ensemble de variables appelées composantes principales, chacune représentant une combinaison linéaire des variables d'origine. Les deux premières composantes principales (ou axes) sont les plus importantes car elles capturent la plus grande partie de la variance totale des données, soit plus de 60% dans notre cas. Ces axes sont interprétés comme suit :



Axe 1 (Popularité et mode du livre) : Cet axe capte la plus grande part de la variance. Il est fortement corrélé avec les variables bbe_score, bbe_vote, et num_ratings, qui sont des mesures de la popularité des livres. Cela signifie que les livres situés le long de cet axe sont principalement caractérisés par leur nombre de votes, leur score dans la liste des meilleurs livres (BBE), et le nombre total de notes reçues. Cet axe reflète donc les aspects quantitatifs de la popularité d'un livre.

Axe 2 (Bonne notation du livre) : Ce second axe, qui capte une part significative mais moindre de la variance, est fortement corrélé avec les variables liked_percent, rating, Price, et pages. Ces variables sont liées à la perception subjective de la qualité d'un livre (comme le pourcentage de lecteurs ayant aimé le livre) ainsi qu'à certaines caractéristiques spécifiques telles que le prix et le nombre de pages. Cet axe peut être

interprété comme capturant des aspects liés à la qualité perçue et aux caractéristiques physiques des livres.

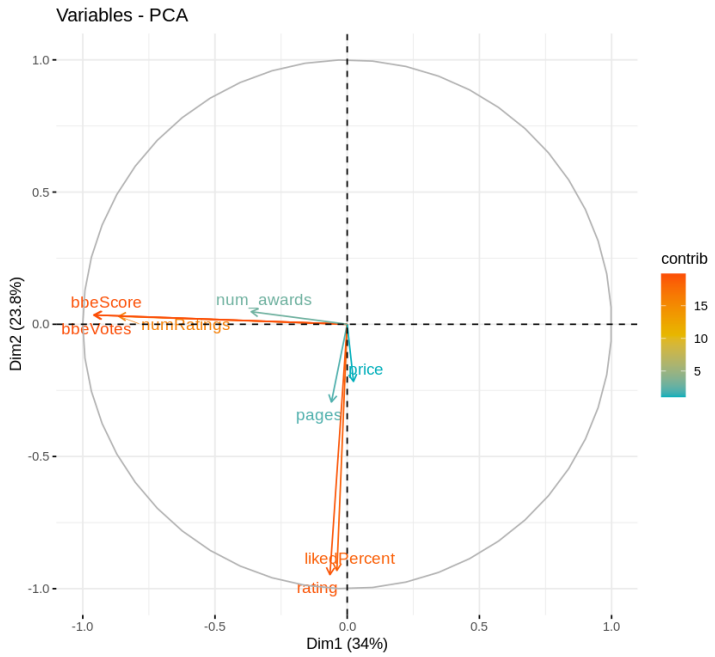
B - IMPORTANCE DES AXES ET METRIQUES (COS2 EST CONTR)

Dans notre analyse :

Les variables liked_percent et rating sont proche du cercle de représentation sur l'axe 2, ce qui signifie qu'elles sont bien représentées par cet axe et qu'elles influencent significativement la deuxième dimension.

De même, les variables bbe_score et bbe_vote ont un cos2 élevé avec l'axe 1, indiquant qu'elles sont bien capturées par la première dimension et qu'elles jouent un rôle clé dans la variance expliquée par cet axe.

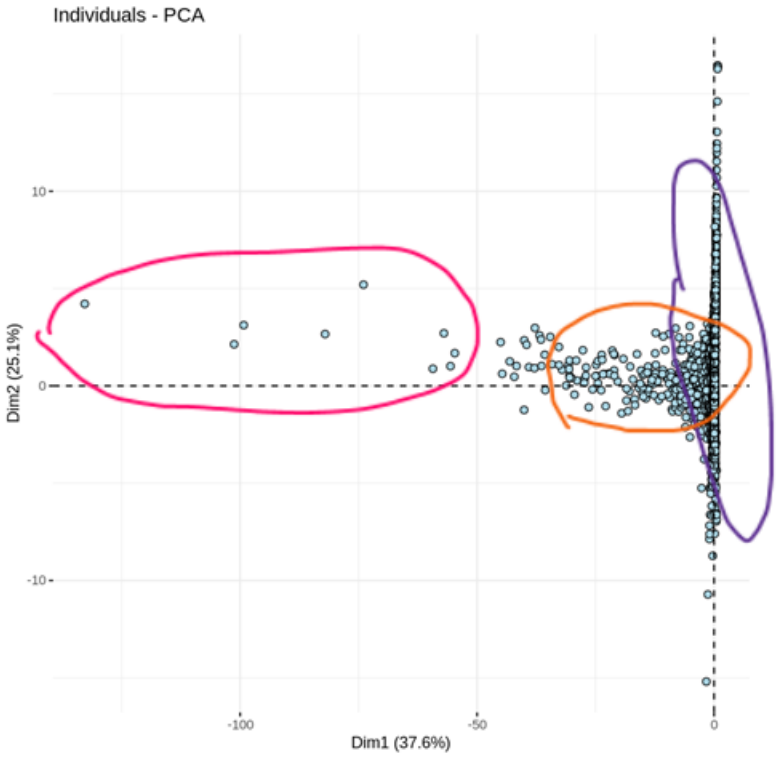
Nous pouvons également ajuster le paramètre `cos2` pour afficher les contributions des variables aux axes principaux.



En observant les résultats, on constate que les contributions sont assez similaires aux valeurs du `cos2`. Plus précisément, les variables `rating` et `likedPercent` contribuent de manière significative à l'axe 1, tandis que `bbeScore` et `bbeVotes` jouent un rôle important sur l'axe 2. Ces contributions renforcent l'idée que l'axe 1 est principalement influencé par les mesures de la qualité perçue, tandis que l'axe 2 est davantage lié à la popularité des livres.

C - ANALYSE DU NUAGE DES LIVRES

Pour visualiser l'impact des deux premières dimensions, nous avons créé un nuage de points des livres dans l'espace défini par l'axe 1 et l'axe 2. Cette représentation graphique permet d'observer comment les livres se distribuent en fonction des deux axes principaux, on discerne 3 groupes de points :



Groupe lié eu Rating : Ce groupe comprend les livres dont les points sont fortement alignés le long de l'axe 2. Les livres dans ce groupe partagent des caractéristiques communes influencées principalement par la variable `rating`. La concentration des points indique que ces livres sont perçus de manière relativement homogène en termes de qualité, avec des évaluations généralement élevées ou basses. Le fait que ces livres soient regroupés de manière serrée le long de l'axe 2 suggère une corrélation forte entre leurs évaluations (`rating`) et leur position sur cet axe.

Groupe lié à la popularité et au Rating : Un autre groupe important est situé autour du croisement des axes 1 et 2. Les livres dans ce groupe sont influencés à la fois par des facteurs de popularité (capturés par des variables telles que `bbe_score`, `bbe_votes`, et `num_ratings`) et par leur évaluation (`rating`). Ce regroupement suggère que ces livres sont à la fois populaires et bien notés, ce qui pourrait indiquer qu'ils sont souvent recommandés ou largement diffusés parmi les lecteurs. La position centrale de ces points reflète leur importance sur les deux axes, signifiant que ces livres sont influencés de manière équilibrée par leur qualité perçue et leur popularité.

Groupe lié à la popularité : Enfin, un troisième groupe est constitué de livres dont les points sont dispersés, mais qui se situent en grande partie le long de l'axe 1, avec moins d'alignement sur l'axe 2. Ce groupe est principalement influencé par des indicateurs de

popularité tels que le nombre de votes (`bbe_votes`) et le score de popularité (`bbe_score`). Les livres dans ce groupe ont peut-être des évaluations plus variées, mais leur popularité reste un facteur dominant, ce qui explique leur position plus éloignée le long de l'axe 1. Ce groupe montre que certains livres, bien que populaires, peuvent présenter des variations plus larges en termes de qualité perçue

D – CONCLUSION DE L'ACP

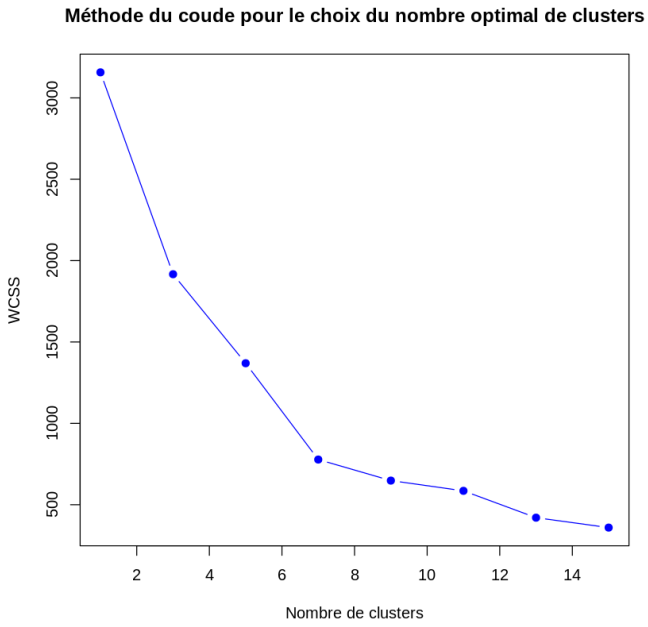
En résumé, l'ACP a non seulement permis de réduire la complexité de nos données, mais elle a également mis en évidence des axes d'interprétation clairs. L'axe 1 est dominé par des indicateurs de popularité, tandis que l'axe 2 est davantage associé à des critères de qualité perçue et de caractéristiques spécifiques. Ces résultats ont permis de mettre en évidence 3 groupes que nous essayerons d'affiner à travers le reste de notre étude.

III. REALISATION DU K-MEANS

A - METHODE DU COUDE

Nous allons réaliser un k-means sur notre jeu de données pour identifier des groupes ou des segments de livres ayant des caractéristiques similaires. Le k-means est une méthode de clustering qui permet de partitionner les observations en k clusters, de manière à minimiser la variance intra-cluster et maximiser la variance inter-cluster. L'ultime objectif de cette analyse est de découvrir des patterns cachés dans les données, ce qui nous permettra de mieux comprendre les tendances et les structures sous-jacentes. En segmentant les livres, nous pourrons obtenir des insights utiles pour des recommandations personnalisées, l'optimisation des collections de livres, et l'amélioration de l'expérience utilisateur sur la plateforme Goodreads.

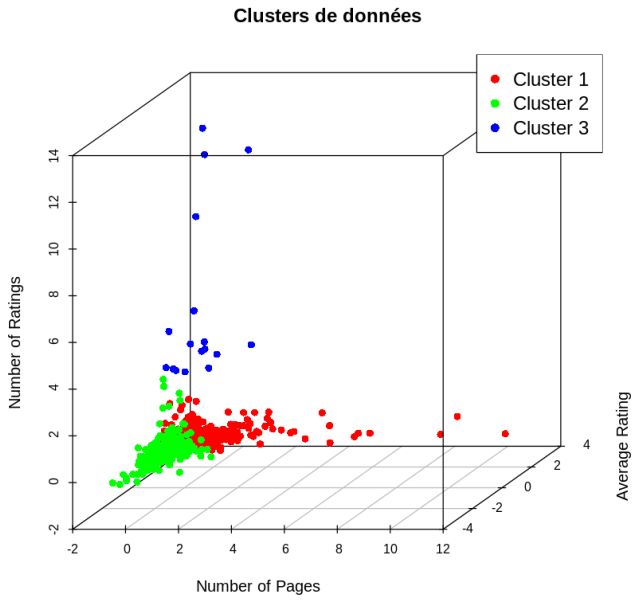
Pour commencer, nous allons tracer une courbe de la méthode du coude (Elbow Method) afin de déterminer le nombre optimal de clusters pour notre analyse. Cette méthode implique de représenter la somme des distances quadratiques moyennes (inertie) entre les points de données et leurs centres de cluster respectifs pour différents nombres de clusters. En identifiant le point où l'ajout de clusters supplémentaires n'entraîne plus une réduction significative de l'inertie, nous pourrons déterminer le nombre optimal de clusters à utiliser.



Dans notre cas, ce point d'inflexion se situe à 3, ce qui nous permet de conclure que le nombre optimal de clusters est 3.

Nous allons maintenant initialiser le nombre de clusters à 3 et appliquer l'algorithme K-means. Ensuite, nous représenterons nos clusters sous forme de visualisation en 3D.

Le graphique 3D est créé en utilisant trois variables (`pages`, `rating`, `numRatings`).



B - MISE EN EVIDENCE DE CLUSTERS

Nous avons aussi calculé les statistiques descriptives pour chaque cluster afin de mieux comprendre les caractéristiques de chaque groupe de livres. Voici les résultats :

Le **Cluster 1**, qui comprend 26 926 livres, est représentatif des **amateurs de livres évalués hautement**. Ce groupe se distingue par une note moyenne impressionnante de 4.27 et un pourcentage de "liked" exceptionnel de 95.6%. Les genres dominants dans ce cluster incluent la fiction, la romance, et la fantasy, avec une présence notable de livres destinés aux jeunes adultes et aux lecteurs contemporains. L'anglais est la langue prédominante, avec une proportion significative d'autres langues comme l'arabe, l'espagnol, et l'allemand. Les lecteurs de ce cluster privilégient des ouvrages particulièrement appréciés et bien notés dans des genres spécifiques, témoignant d'une préférence pour les livres ayant reçu des critiques très positives.

Le **Cluster 2**, plus restreint avec ses 75 livres, caractérise les **chercheurs de diversité littéraire**. Ce groupe présente une note moyenne de 4.10 et un pourcentage de "liked" de 92.1%. Les genres les plus fréquents incluent la fiction, le young adult, les romans, et les classiques, ainsi que des genres comme la fantasy et la littérature. Ce cluster est exclusivement anglophone, ce qui suggère que ses lecteurs recherchent une variété de genres tout en étant très sélectifs quant à la qualité des livres qu'ils choisissent. Ils

montrent un intérêt marqué pour des genres diversifiés tout en se concentrant sur des titres en anglais.

Enfin, le **Cluster 3**, regroupant 25 389 livres, illustre les **explorateurs de genres et de langues**. Ce groupe est caractérisé par une note moyenne de 3.76 et un pourcentage de "liked" de 88.7%. Les genres dominants incluent la fiction et la romance, avec une forte concentration dans le genre contemporain. L'anglais est la langue majoritaire, mais on trouve aussi des livres en arabe, en espagnol, et en français. Les lecteurs de ce cluster sont ouverts à une grande variété de genres et de langues, explorant une large gamme de styles littéraires et linguistiques malgré des évaluations moyennes un peu plus faibles.

Ces observations permettent de mieux comprendre les différents profils de lecteurs en fonction des genres, des langues, et des évaluations des livres dans chaque cluster, offrant ainsi des perspectives utiles sur les préférences littéraires et les caractéristiques des ouvrages appréciés.

C – APPROFONDISSEMENT : REALISATION D'UN K-MEANS SPECIFIQUE AU GENRE DE LIVRE

En complément de notre analyse générale avec le K-Means clustering, nous allons réaliser un K-Means spécifique aux genres des livres. Cette démarche vise à approfondir notre compréhension des livres en se concentrant particulièrement sur les genres. Le K-Means général nous a permis de regrouper les livres en fonction de plusieurs caractéristiques globales. Cependant, en appliquant le K-Means spécifiquement aux genres, nous pouvons isoler les regroupements de livres selon leurs genres littéraires prédominants. Cette approche nous offre une segmentation plus fine et détaillée, permettant de voir comment les genres influencent les regroupements de livres.

Pour analyser la distribution des livres au sein des clusters obtenus par le clustering K-Means, nous avons d'abord transformé les genres littéraires en variables binaires, puis appliqué l'algorithme de K-Means pour segmenter les livres en cinq clusters distincts. Afin de mieux comprendre et visualiser ces clusters, nous avons réduit les dimensions des données à trois dimensions en utilisant la technique de t-SNE. Cela nous a permis de créer une visualisation en 3D interactive des clusters.

Les résultats montrent une variation significative du nombre de livres entre les différents clusters. Le Cluster 2 contient le plus grand nombre de livres avec 10 892 titres, tandis que le Cluster 1 en compte 2 666. Les autres clusters présentent des tailles intermédiaires, avec des nombres de livres variant de 2 653 à 3 289. Cette distribution hétérogène reflète des différences notables dans la composition des livres au sein de chaque cluster.

En observant le nuage de points des clusters issus du K-means, qui met en relation les genres, les notes et le nombre de notes, plusieurs tendances se dégagent. Le Cluster 1 est celui qui a reçu le plus grand nombre de notes, avec des notes majoritairement élevées, comprises entre 3 et 5, bien qu'un point soit situé à une note de 2. Dans le Cluster 2, les notes sont principalement concentrées entre 3 et 4, et on remarque qu'à mesure que le nombre de notes augmente, les notes tendent à se rapprocher de 4. Pour le Cluster 3, on observe relativement peu de notes, dont certaines sont très basses, avec une note inférieure à 2 sur 5. Cependant, il y a également des notes élevées pour un nombre restreint de livres, et lorsque le nombre de notes augmente, celles-ci se situent généralement autour de 4. Dans le Cluster 4, quelques points sont concentrés autour d'une note de 4, attribués à un petit nombre de livres. Enfin, le Cluster 5 présente également quelques points avec des notes proches de 4, mais avec un nombre encore plus faible de notes assignées.

Cette analyse révèle des différences marquées dans les caractéristiques des livres au sein de chaque cluster, tant en termes de nombre de titres que de répartition des notes.

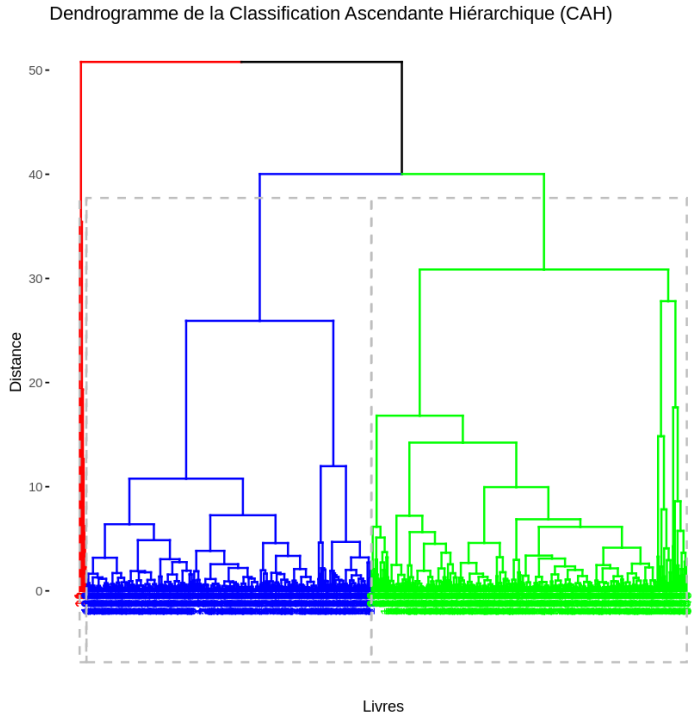
D – CONCLUSION DU K-MEANS

L'application du K-Means a révélé trois clusters principaux de livres, chacun ayant des caractéristiques distinctes. Le premier cluster regroupe les livres très bien notés et populaires, le second contient des livres avec des notes moyennes et une diversité de genres, tandis que le troisième cluster se compose de livres avec des notes plus basses et une moindre diversité linguistique. Cette segmentation montre que le succès des livres varie en fonction de leur popularité et de leur qualité perçue. Notre perception de l'ACP normée est donc conforme aux conclusions du K-MEANS.

IV. APPORT DE LA CAH

Nous avons réalisé une Classification Ascendante Hiérarchique (CAH) sur un échantillon notre jeu de données Goodreads (1000 données) afin de différencier les clusters de livres. Après avoir défini trois clusters avec K-Means, nous avons adapté notre CAH pour visualiser et identifier distinctement ces trois clusters. Les clusters ont été caractérisés en fonction de plusieurs variables telles que la note moyenne, le nombre de pages, le nombre de notes, le pourcentage de 'liked', le score et les votes BBE, ainsi que le prix moyen.

A - ETUDE DES CLUSTERS



Voici les résultats détaillés pour chaque cluster :

Le Cluster 1, que nous pouvons appeler "High Ratings", comprend 9 livres. Ces livres ont une note moyenne de 4.25 et comptent en moyenne 637.89 pages. Ils sont particulièrement populaires, avec un nombre moyen de notes atteignant 1,544,295.11 et un pourcentage moyen de 'liked' de 94.67%. Les livres de ce cluster ont également un score BBE moyen de 763,755.67 et recueillent en moyenne 8,011 votes BBE. Le prix moyen des livres de ce cluster est de 8.99€. Les genres les plus répandus dans ce cluster sont 'Fiction', 'Adventure', 'Classics', 'Young Adult', 'Audiobook' et 'Fantasy', tandis que la langue prédominante est l'anglais.

Le Cluster 2, que nous pouvons appeler "Mixed Ratings", est le plus grand avec 413 livres. Ces livres ont une note moyenne de 4.22 et comptent en moyenne 434.93 pages. Le nombre moyen de notes est de 48,494.36, avec un pourcentage moyen de 'liked' de

95.66%. Les livres de ce cluster ont un score BBE moyen de 4,069 et recueillent en moyenne 50.46 votes BBE. Le prix moyen des livres de ce cluster est de 15.88€. Les genres les plus répandus dans ce cluster sont 'Fiction', 'Fantasy', 'Romance', 'Young Adult' et 'Adventure'. La langue prédominante est également l'anglais, suivie de l'italien, l'espagnol, le portugais, le français et l'allemand.

Le Cluster 3, que nous pouvons appeler "Low Ratings", comprend 373 livres. Ces livres ont une note moyenne de 3.83 et comptent en moyenne 326.24 pages. Le nombre moyen de notes est de 24,565.86, avec un pourcentage moyen de 'liked' de 89.84%. Les livres de ce cluster ont un score BBE moyen de 935.51 et recueillent en moyenne 12.71 votes BBE. Le prix moyen des livres de ce cluster est de 6.21€. Les genres les plus répandus dans ce cluster sont 'Fiction', 'Fantasy', 'Romance', 'Young Adult' et 'Paranormal'. La langue prédominante est l'anglais, suivie de l'italien, l'espagnol, le français, l'arabe et le néerlandais.

Les résultats de la CAH montrent une segmentation claire des livres en trois clusters distincts basés sur leurs caractéristiques et popularité. Cluster 1 représente les livres les plus populaires et les mieux notés, principalement en anglais et dans des genres variés comme 'Fiction' et 'Adventure'. Cluster 2 est un mélange de livres avec des notes modérées et une grande diversité linguistique, tandis que Cluster 3 comprend des livres avec des notes plus basses et une diversité linguistique moindre. Cette classification permet de mieux comprendre la distribution et les caractéristiques des livres dans chaque cluster.

B – CONCLUSION DE LA CAH

La CAH a confirmé les résultats obtenus par le K-Means en identifiant trois groupes similaires. Le premier groupe rassemble les livres les plus populaires et mieux notés, principalement en anglais. Le second groupe est plus diversifié en termes de langues et de genres, tandis que le troisième regroupe des livres avec des notes plus faibles et une moindre diversité linguistique. Cette classification hiérarchique a consolidé notre compréhension des segments de marché existants.

V. CONCLUSION

Cette étude nous a permis de comprendre les facteurs influençant le succès des livres vendus en ligne à partir d'un ensemble de données provenant de Goodreads. Notre démarche a suivi plusieurs étapes, allant de la préparation et du nettoyage des données jusqu'à des analyses univariées, bivariées et multivariées, une analyse en composante principale normée ACP, une étude de K-MEANS et enfin une Classification Ascendante Hiérarchique CAH.

L'analyse univariée a mis en lumière certaines tendances notables, telles que la prédominance des livres en anglais et la préférence des lecteurs pour les formats physiques comme le livre broché, par rapport aux formats numériques. Les variables quantitatives ont révélé que la majorité des livres se situent entre 0 et 1000 pages, avec des prix généralement compris entre 0 et 25 dollars.

L'analyse bivariée a permis d'identifier des relations significatives entre différentes variables. Par exemple, une forte corrélation a été observée entre la note moyenne (avg_rating) et le pourcentage de lecteurs ayant aimé le livre (likedPercent), ce qui suggère que les livres mieux notés sont en général plus appréciés par les utilisateurs. De plus, la relation positive entre le score sur la liste des meilleurs livres de Goodreads (bbeScore) et le nombre de notes reçues (num_ratings) montre l'importance de la visibilité et de la popularité d'un livre pour son succès.

L'analyse en composantes principales (ACP) a permis de réduire la dimensionnalité des données tout en conservant l'essentiel de l'information. L'ACP a fait apparaître trois groupes distincts, chacun capturant différentes dimensions clés influençant le succès des livres : la popularité quantitative, la qualité perçue par les lecteurs, et les caractéristiques physiques des livres. Cette structuration a été confirmée par le K-Means et la Classification Ascendante Hiérarchique (CAH), qui ont également révélé trois groupes principaux dans les données.

Notre étude démontre que le succès des livres en ligne est multifactoriel, influencé à la fois par la visibilité (nombre de notes, popularité) et par la qualité perçue par les lecteurs (note moyenne, pourcentage de likes). Bien que les caractéristiques physiques telles que le nombre de pages et le prix jouent également un rôle, leur influence est moins déterminante. La convergence des résultats obtenus par l'ACP, le K-Means et la CAH, qui révèlent tous trois des groupes similaires, souligne l'importance de ces dimensions dans la compréhension des dynamiques du marché en ligne. Ces résultats fournissent des insights précieux pour les éditeurs et les auteurs souhaitant adapter leurs stratégies pour maximiser le succès de leurs ouvrages sur des plateformes comme Goodreads.