## **Analyse des Données de Livres : Démarche et Résultats**

### **1. Introduction**

L'objectif de ce projet est d'analyser des données de livres en ligne en utilisant des techniques de web scraping, de nettoyage des données, et de visualisation. Les étapes suivantes décrivent la méthodologie suivie pour obtenir et analyser ces données.

### **2. Importation des Bibliothèques**

Les bibliothèques suivantes ont été importées pour la manipulation des données, le scraping, et la visualisation :

* **pandas** et **numpy** pour la manipulation des données.
* **matplotlib** et **seaborn** pour la visualisation des données.
* **requests** pour le scraping des pages web.
* **BeautifulSoup** pour parser le contenu HTML des pages web.

### **3. Scraping des Données**

Pour obtenir les données des livres, les étapes suivantes ont été suivies :

* Utilisation de requests.get() pour télécharger le contenu des pages web contenant des informations sur les livres.
* Utilisation de BeautifulSoup pour parser le contenu HTML et extraire les informations pertinentes telles que le titre, la catégorie, et la disponibilité des livres.

Les données récupérées grâce au web scraping ont été stockées sous forme de dataframes, facilitant ainsi leur manipulation et analyse.

### **4. Connexion à MySQL Workbench et Déploiement sur Azure**

Les dataframes résultants du web scraping ont été connectés à MySQL Workbench pour un stockage structuré et une gestion efficace des données. De plus, le projet a été déployé sur Azure pour garantir l'accessibilité et la scalabilité de l'application. Les étapes suivantes ont été réalisées :

* Exportation des dataframes vers une base de données MySQL.
* Configuration et gestion des bases de données via MySQL Workbench.
* Déploiement des services et de l'application sur la plateforme Azure pour une accessibilité globale et une robustesse accrue.

### **5. Nettoyage et Préparation des Données**

Les données extraites ont été nettoyées et préparées pour l'analyse :

* **Suppression d'une colonne** : Une colonne inutile ou redondante a été supprimée pour simplifier l'analyse.  
  python  
  Copier le code  
  df = df.drop(columns=['unnecessary\_column'])
* **Transformation du type d'une colonne** : Conversion des types de données pour assurer la cohérence et faciliter les opérations analytiques.  
  python  
  Copier le code  
  df['price'] = df['price'].astype(float)
* **Renommage d'une colonne** : Pour clarifier les noms de colonnes et améliorer la lisibilité.  
  python  
  Copier le code  
  df = df.rename(columns={'old\_column\_name': 'new\_column\_name'})

### **6. Analyse des Données**

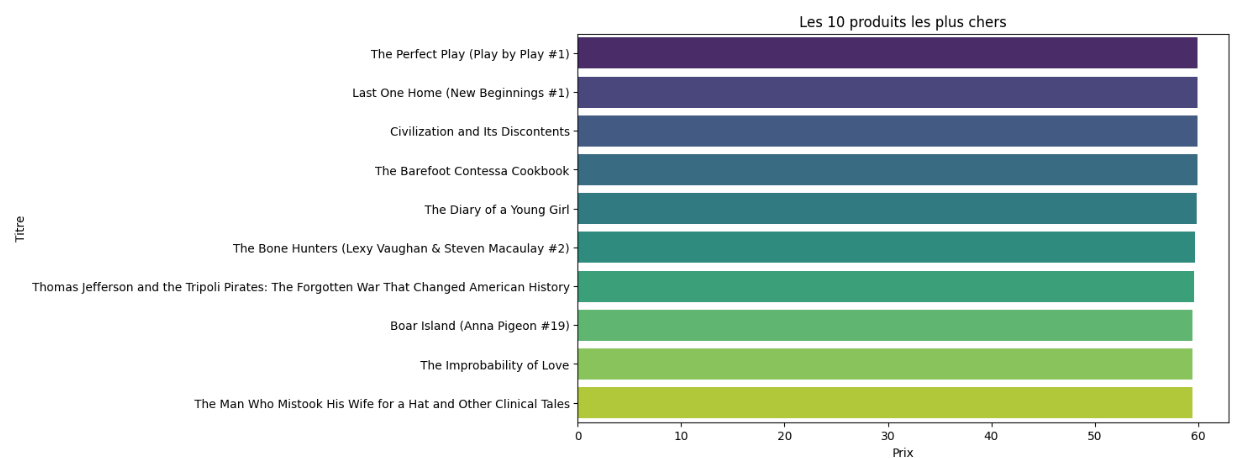
Les données ont été analysées en utilisant les méthodes suivantes :

* **Forme des données** : Utilisation de .shape pour connaître la taille du dataframe, c'est-à-dire le nombre de lignes et de colonnes. Cela aide à comprendre l'ampleur des données collectées.  
  python  
    
  df.shape
* **Description statistique** : Utilisation de .describe() pour obtenir des statistiques descriptives telles que la moyenne, l'écart-type, le minimum et le maximum pour les colonnes numériques. Cela permet de résumer rapidement les caractéristiques principales des données.  
  python  
    
  df.describe()
* **Comptage de valeurs** : Utilisation de .value\_counts() pour compter la fréquence des valeurs dans une colonne donnée. Cela est particulièrement utile pour analyser des colonnes catégorielles.  
  python  
    
  df['category'].value\_counts()
* **Calcul des moyennes** : Utilisation de .mean() pour calculer les moyennes des colonnes numériques. Cela aide à comprendre les tendances centrales des données.  
  python  
    
  df['price'].mean()

### **7. Visualisation des Données**

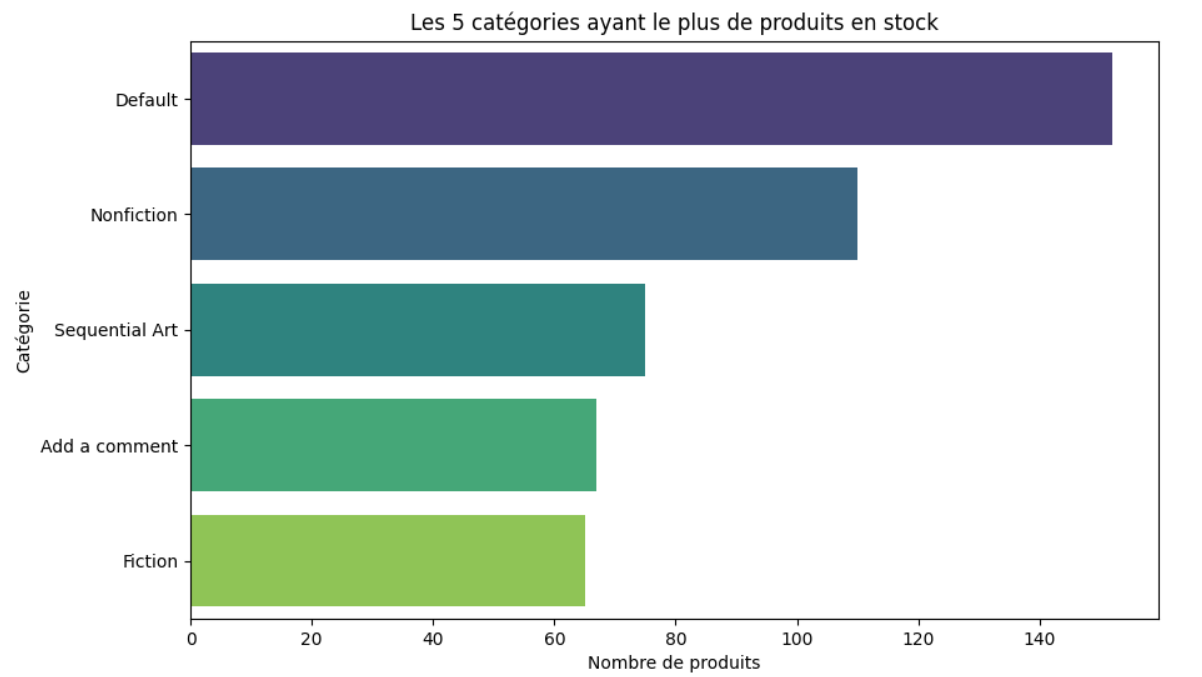
Les données ont été visualisées pour mieux comprendre les tendances et les distributions. Les visualisations suivantes ont été réalisées :

#### **Les 10 produits les plus chers**



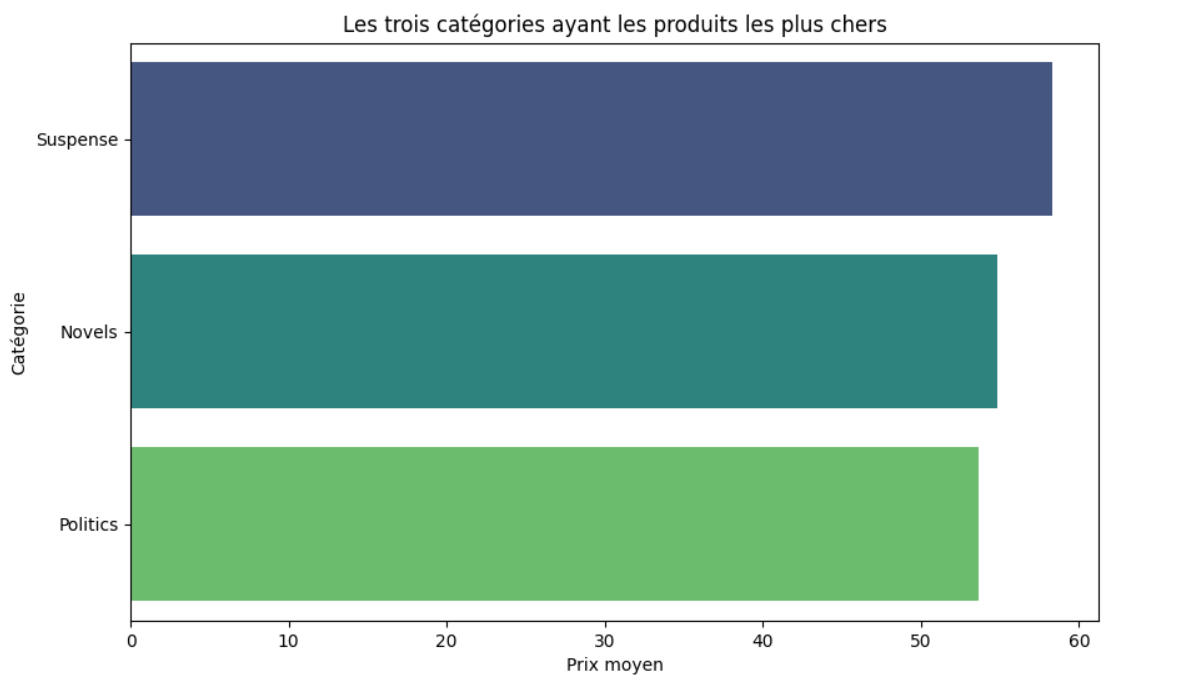
**Interprétation** : Cette visualisation montre les dix produits les plus chers. Par exemple, "The Man Who Mistook His Wife for a Hat and Other Clinical Tales" est le produit le plus cher avec un prix supérieur à 60 unités. Il est crucial de surveiller ces produits pour comprendre les tendances de prix et ajuster les stratégies de tarification en conséquence.

#### **Les 5 catégories ayant le plus de produits en stock**



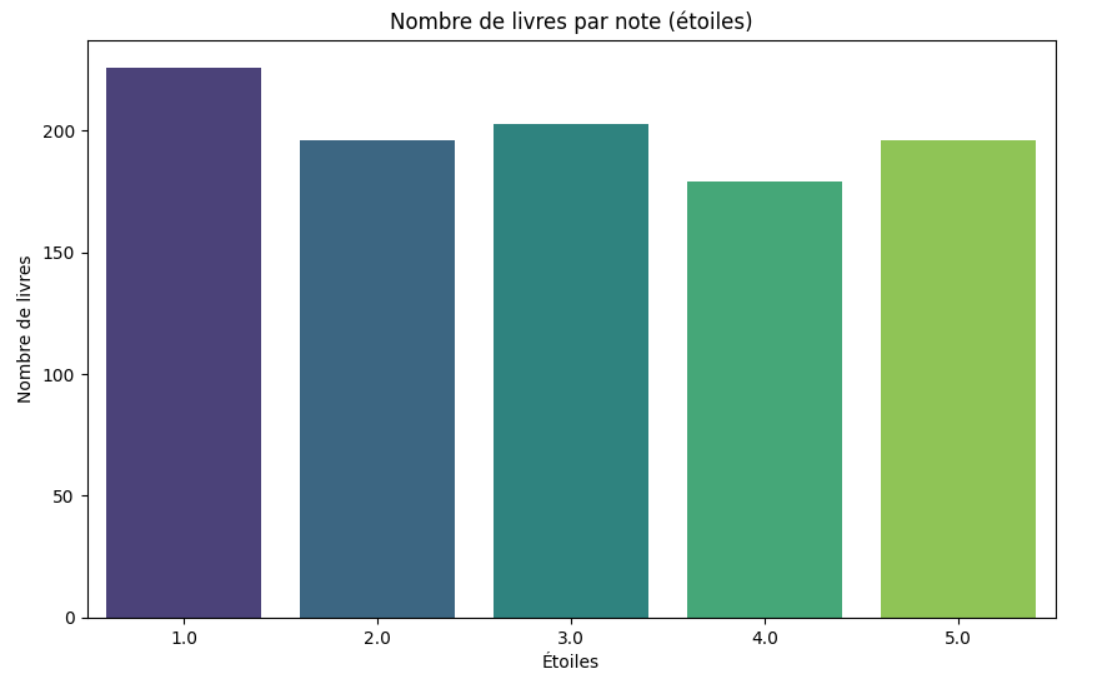
**Interprétation** : Cette visualisation indique que la catégorie "Default" a le plus de produits en stock, suivie de "Nonfiction" et "Sequential Art". Cela peut aider à optimiser l'inventaire et à identifier les catégories les plus populaires. La catégorie "Add a comment" semble également significative, indiquant peut-être un besoin de reclassification ou de nettoyage des données.

#### **Les trois catégories ayant les produits les plus chers**



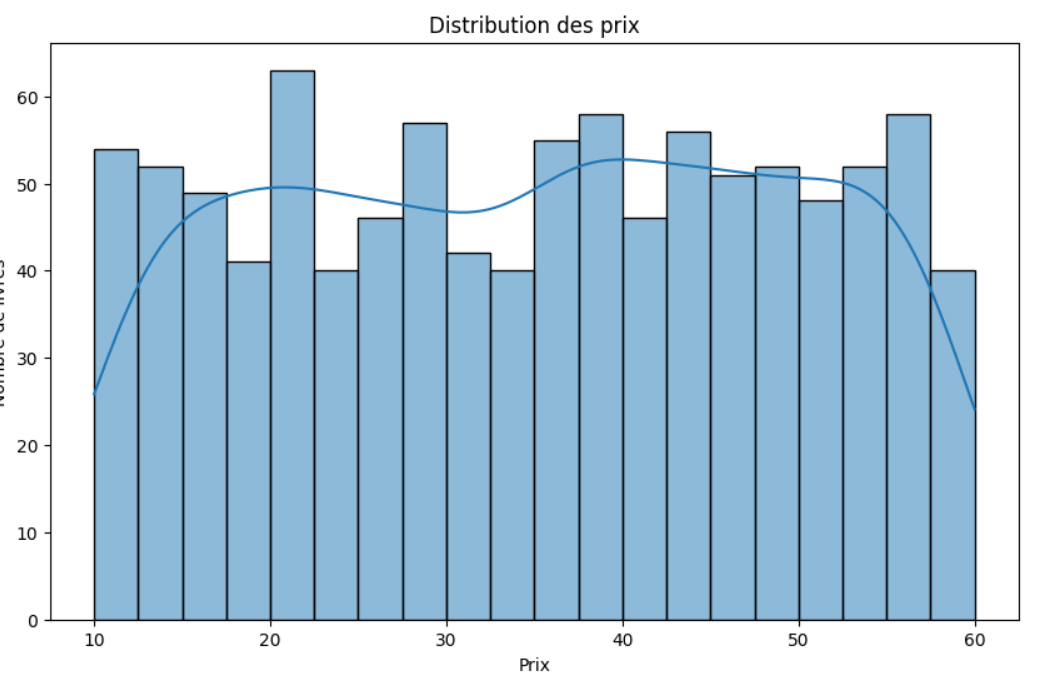
**Interprétation** : Cette visualisation montre que les catégories "Suspense", "Novels", et "Politics" contiennent les produits les plus chers, avec des prix moyens approchant ou dépassant 50 unités. Connaître ces catégories peut aider à cibler des campagnes marketing spécifiques et à maximiser les revenus.

#### **Nombre de livres par note (étoiles)**



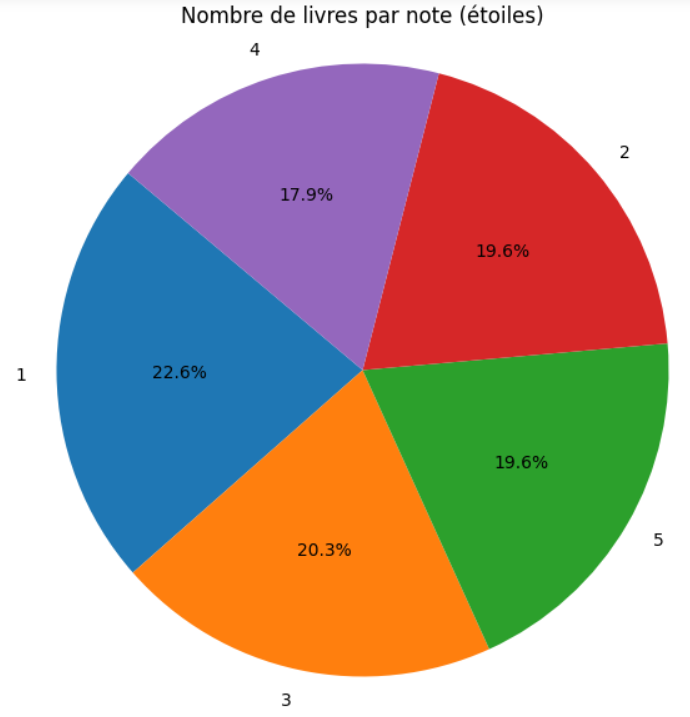
**Interprétation** : Cette visualisation représente la répartition des livres par note. Par exemple, la majorité des livres ont une note de 1 étoile, ce qui indique peut-être des problèmes de qualité ou des attentes clients non satisfaites. Cela peut être utilisé pour comprendre la satisfaction des clients et améliorer la qualité des produits proposés.

#### **Distribution des prix**



**Interprétation** : Cette visualisation montre une distribution relativement uniforme des prix des livres, avec un léger pic autour de 20 et 40 unités. Comprendre la répartition des prix aide à fixer des prix compétitifs et attractifs.

#### **Nombre de livres par note (étoiles) - Diagramme en camembert**



**Interprétation** : Ce diagramme en camembert montre la répartition des livres par note. Par exemple, 22,6% des livres ont une note de 1 étoile, ce qui est la proportion la plus élevée. Cela offre une vue d'ensemble rapide des notes des livres et peut guider des actions pour améliorer la satisfaction des clients.

### **8. Résultats et Conclusions**

L'analyse a permis de :

* Identifier les catégories de livres les plus populaires.
* Comprendre la disponibilité des livres par catégorie.
* Fournir des visualisations claires et informatives pour communiquer les résultats.

### **9. Conseils pour la Gestion du Site**

1. **Surveillance des Produits les Plus Chers** : Étant donné que les produits les plus chers, comme "The Man Who Mistook His Wife for a Hat and Other Clinical Tales", peuvent générer des revenus significatifs, il est crucial de les surveiller de près et d'ajuster les stratégies de tarification en fonction de la demande et des tendances du marché.
2. **Optimisation de l'Inventaire** : Les catégories ayant le plus de produits en stock, comme "Default" et "Nonfiction", doivent être régulièrement analysées pour s'assurer que l'inventaire est aligné avec la demande des clients. Les catégories les moins populaires devraient être réévaluées pour éviter des stocks excédentaires.
3. **Stratégies Marketing Ciblées** : Les catégories avec les produits les plus chers, comme "Suspense" et "Novels", peuvent bénéficier de campagnes marketing spécifiques pour attirer les clients disposés à payer plus.
4. **Amélioration de la Qualité des Produits** : Analyser les notes des livres, où une proportion significative a une note de 1 étoile, peut aider à identifier les produits de mauvaise qualité et à améliorer l'offre. Les produits avec des notes élevées devraient être mis en avant pour attirer de nouveaux clients.
5. **Gestion des Prix** : Comprendre la distribution des prix aide à fixer des prix compétitifs et attractifs. Il est important de trouver un équilibre entre maximiser les revenus et offrir des prix justes aux clients.

Le déploiement sur Azure et la connexion à MySQL Workbench ont également assuré une gestion efficace des données et une accessibilité améliorée pour les utilisateurs finaux.