**Réponse I :**

Discuter de 4 observations reliées à la variable « Note » avec des graphiques jugés pertinents.

Cette analyse va permettre de révéler quatre observations majeures concernant la variable **'Note'**. Elle constitue une base pour la construction d'un modèle prédictif et une exploration plus approfondie.

**## Observation 1** : Distribution des Notes L'histogramme montre que la majorité des films ont des notes comprises entre 5 et 8, reflétant une évaluation moyenne à bonne.

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquement

**## Observation 2** : Valeurs aberrantes Le boxplot montre quelques films ayant des notes extrêmes (9), ce qui justifie une exploration approfondie.

Une image contenant diagramme, texte, ligne, Parallèle

Description générée automatiquement

**## Observation 3** : Relation Votes-Notes Un scatterplot montre que les films ayant un nombre élevé de votes ont des notes stables (6-8), tandis que les films avec peu de votes présentent des variations plus importantes.

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

Description générée automatiquement

**## Observation 4** : Influence du Genre Les genres influencent les notes, avec des moyennes plus élevées pour Drama et Biography, et une variabilité plus importante pour Comedy

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

Description générée automatiquement

**Conclusion :**

L'analyse a mis en évidence des tendances générales et des facteurs influençant les notes des films, ce qui a mené à plusieurs observations intéressantes grâce à des visualisations pertinentes :

1-**Distribution générale des notes** :

* La majorité des films sont notés entre **5 et 8**, ce qui indique une tendance générale vers des évaluations moyennes à bonnes.
* Cette observation reflète potentiellement une échelle biaisée vers des appréciations positives modérées ou une sévérité limitée des utilisateurs dans leurs évaluations.

2- **Valeurs aberrantes** :

* Une petite proportion de films présente des notes **très basses (inférieures à 4)** ou **très élevées (supérieures à 9)**.
* Ces valeurs atypiques pourraient être liées à des films très polarisants, soit des chefs-d'œuvre acclamés, soit des productions particulièrement mal accueillies. Ces cas pourraient justifier une exploration plus approfondie pour identifier les facteurs qui contribuent à ces extrêmes.

3- **Impact du nombre de votes** :

* Les films ayant un grand nombre de votes montrent des notes plus **stables et concentrées autour de 6 à 8**.
* À l'inverse, les films ayant reçu peu de votes présentent une variabilité plus importante, avec des notes extrêmes fréquentes. Cela peut être attribué à une représentation biaisée des opinions ou à une base d'évaluateurs plus homogène.

4- **Influence des genres** :

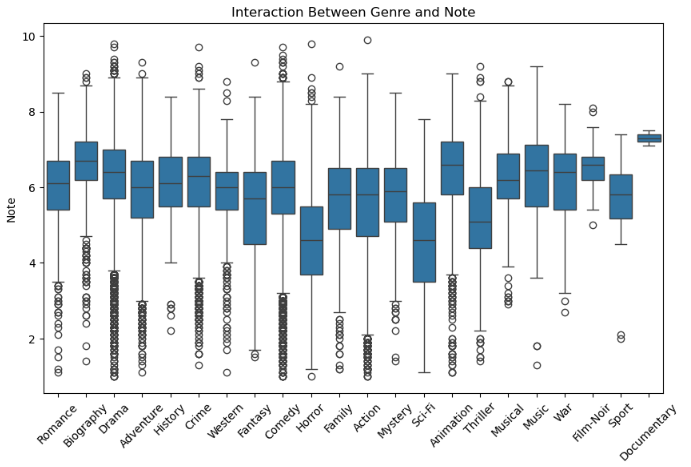
* L'analyse par genre révèle des différences significatives dans la répartition des notes.
  + Par exemple, les genres comme **"Drama"** et **"Biography"** tendent à obtenir des notes moyennes plus élevées, reflétant peut-être une préférence des utilisateurs pour des contenus plus sérieux ou émotionnels.
  + En revanche, des genres comme **"Comedy"** présentent une variabilité plus importante, ce qui pourrait être dû à des différences culturelles ou subjectives dans la perception de l'humour.

**Réponse 2 :**

L’objectif principal est de développer un modèle prédictif pour la variable numérique « Note » en explorant différentes variables explicatives présentes dans l’ensemble de données. Plusieurs variables ont été envisagées, notamment Genre, Pays et Langue, car elles pourraient influencer les notes attribuées par le public.

Voici les graphiques ainsi que quelque notes sur les différentes variables tester :

Notes Vs Genre



Il y a 22 genres uniques. La corrélation moyenne de la variable genre avec Note est de -0.00275.

Notes vs Pays

Une image contenant texte, croquis, dessin, art

Description générée automatiquement

Il y a 160 valeurs unique pour la variable Pays et sa corrélation avec Notes est de 0.00714.

Une image contenant ligne, typographie

Description générée automatiquement

Il y a 163 valeurs unique pour la variable Langue et sa correlation avec notes est de 0.00983.

La correlation des trois variables peuvent être visualiser avec les 3 graphique ci-bas :

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

Description générée automatiquement Une image contenant texte, capture d’écran, ligne, nombre

Description générée automatiquement

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

Description générée automatiquement

Nous pouvons donc observer que l’intervalle de confiance de Langue est de [-0.24 : 0.09], l’intervalle de confiance de pays est de [-0.195 : 0.08] et [-0.25 : 0.19].

Après avoir testé ces différentes variables, nous avons décidé de concentrer notre analyse sur Genre. Bien que d'autres variables comme Langue et Pays présentent des corrélations similaires avec Note, nous avons trouvé que Genre offrait une dimension intéressante en raison de son lien potentiel avec les préférences des spectateurs.

Nous allons donc faire 3 model en fonction de la variable genre pour prédire les notes donnée malgré qu’elle a la corrélation la plus faible, elle mérite tout de même d’être explorée.

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

Description générée automatiquement

Le boxplos de Genre et Note montrent une certaine variation entre les genres, ce qui suggère que Genre pourrait être significatif dans la prédiction.

Voici les 3 modele effectuer pour la prediction des notes selon le Genre :

**Modèle 1 : Suppression des lignes avec des valeurs manquantes**

Les lignes contenant des valeurs manquantes dans Genre ou Note ont été supprimées.

**Modèle 2 : Imputation et normalisation avec RobustScaler**

Les valeurs manquantes dans Note ont été imputées par la moyenne.

Une normalisation robuste a été appliquée pour minimiser l’impact des valeurs aberrantes.

**Modèle 3 : Imputation et standardisation avec StandardScaler**

Les valeurs manquantes dans Note ont également été imputées par la moyenne.

Une standardisation des données a été effectuée pour uniformiser l’échelle.

Voici les valeurs et graphiques des models :

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

Description générée automatiquement

Une image contenant capture d’écran, texte, Caractère coloré, Rectangle

Description générée automatiquement

Model Comparison:

Model 1: MSE = 1.3223, R2 = 0.1332

Model 2: MSE = 0.5099, R2 = 0.1280

Model 3: MSE = 0.8583, R2 = 0.1280

Comparaison des Modèles

Modèle 1 (Drop NaN) :

MSE : 1.3223

R² : 0.1332

Observations : Bien que ce modèle ait un R² légèrement supérieur, la suppression des données entraîne une perte d’informations qui augmente l’erreur moyenne quadratique.

Modèle 2 (RobustScaler) :

MSE : 0.5099

R² : 0.1280

Observations : Ce modèle offre la meilleure précision (faible MSE), grâce à une gestion efficace des valeurs aberrantes.

Modèle 3 (StandardScaler) :

MSE : 0.8583

R² : 0.1280

Observations : Ce modèle se situe entre les deux autres en termes de MSE, mais reste moins performant que le modèle RobustScaler.

Conclusion :

Nous pouvons conclure que le modèle 2 est le choix privilégie pour garantir des prédictions fiables dans des conditions réelles grâce a son MSE le plus faible. Il offre une meilleur précision tout en restant robuste face aux anomalies. C’est le choix à prendre quand il y a beaucoup de valeurs aberrantes ou une distribution non uniforme. Si la variable Genre ne contient pas beaucoup de valeurs manquantes ou peu de valeurs manquante, le modèle 1 devient l’option car il est le plus précis des 3, grâce a son R2, mais est le plus affecter aux valeurs manquantes.