Relatório Indicium Lighthouse 2025

Esse documento contém o relatório do EDA, análise estatística e avaliação de modelo de ML e entrega final para o programa Lighthouse 2025 da Indicium.

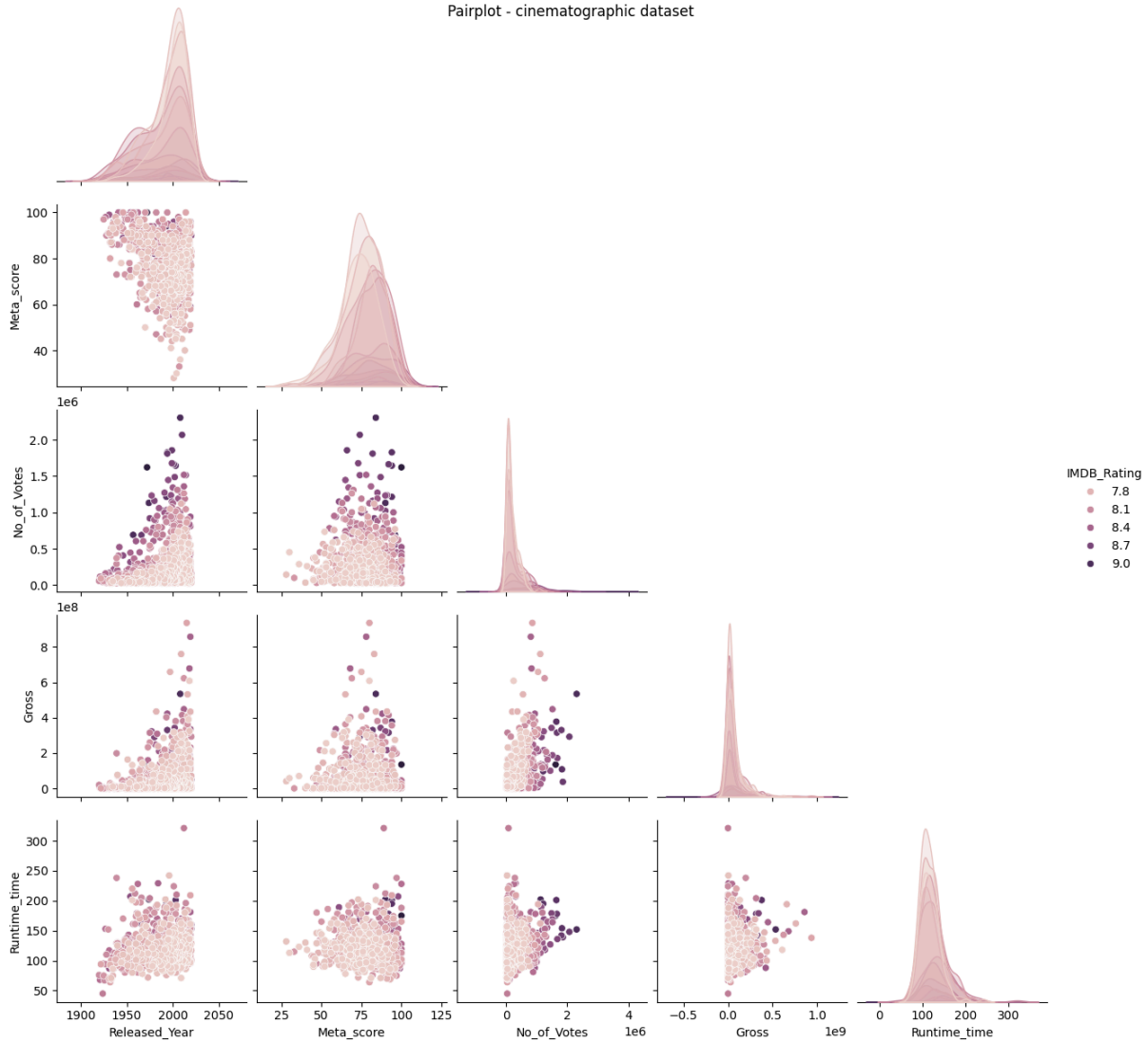
Observação:

1. Os comentários e observações feitas via markdown no código dos .ipynb foram feitas enquanto estava produzindo todo o projeto.

Análise Exploratória dos Dados

Antes de começar com a análise, precisei fazer algum pré-processamento das features para conseguir seguir com a análise de maneira mais fácil, e também facilitar o uso do dataset depois nos modelos, realizei as seguintes transformações:

1. Certificate, converti NaN para “Unrated”, transformei a coluna para “category”, converti GP para PG e U/A para UA.
2. Runtime, separei em duas novas features, Runtime_time e Runtime_type, porém Runtime_type não foi usada pois era constante em todos os registros.
3. Director transformei para “category”.
4. Gross era object (string), removi as vírgulas da string para converter em float. Para lidar com os valores NaN primeiro preenchi pela mediana agrupada por diretores dos registros, depois pela mediana do gênero e após isso pela mediana geral.
5. Released_Year, excluí o único registro divergente com “PG”

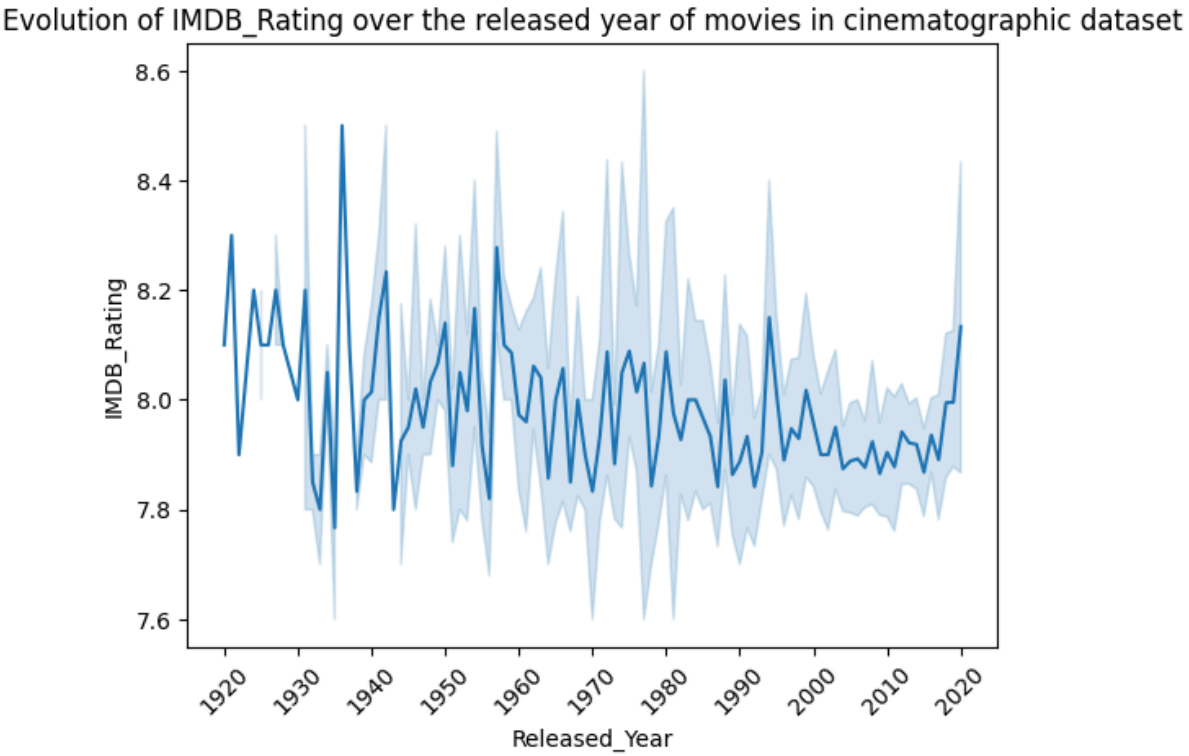


A partir do pairplot podemos visualizar detalhes sobre correlação entre as variáveis numéricas nessa etapa do EDA.

Parece existir uma concentração de filmes entre 50 e aproximadamente 90 de Meta_score, e os que passam de 90 tendem a ter maior IMDB_Rating.

A maioria dos filmes tem até 2 milhões de votos, porém a concentração maior é até 500 mil votos, e os que fogem dessa concentração para mais tendem a ter o maior IMDB_Rating.

Sobre o Runtime_time, filmes com maior duração tendem a ter notas mais altas de IMDB_Rating, essa tendência fica mais notável a partir de 200 min.



Sobre a evolução do IMDB_Rating ao passar das décadas vemos que há uma variação durante os anos, mas tende a permanecer a mesma média de IMDB_rating, porém se observarmos o intervalo de confiança gerado no plot, vemos que existe uma maior variação dentre os anos, principalmente entre 1975 e 1980.

- O desvio padrão em todo o intervalo é de 0.272

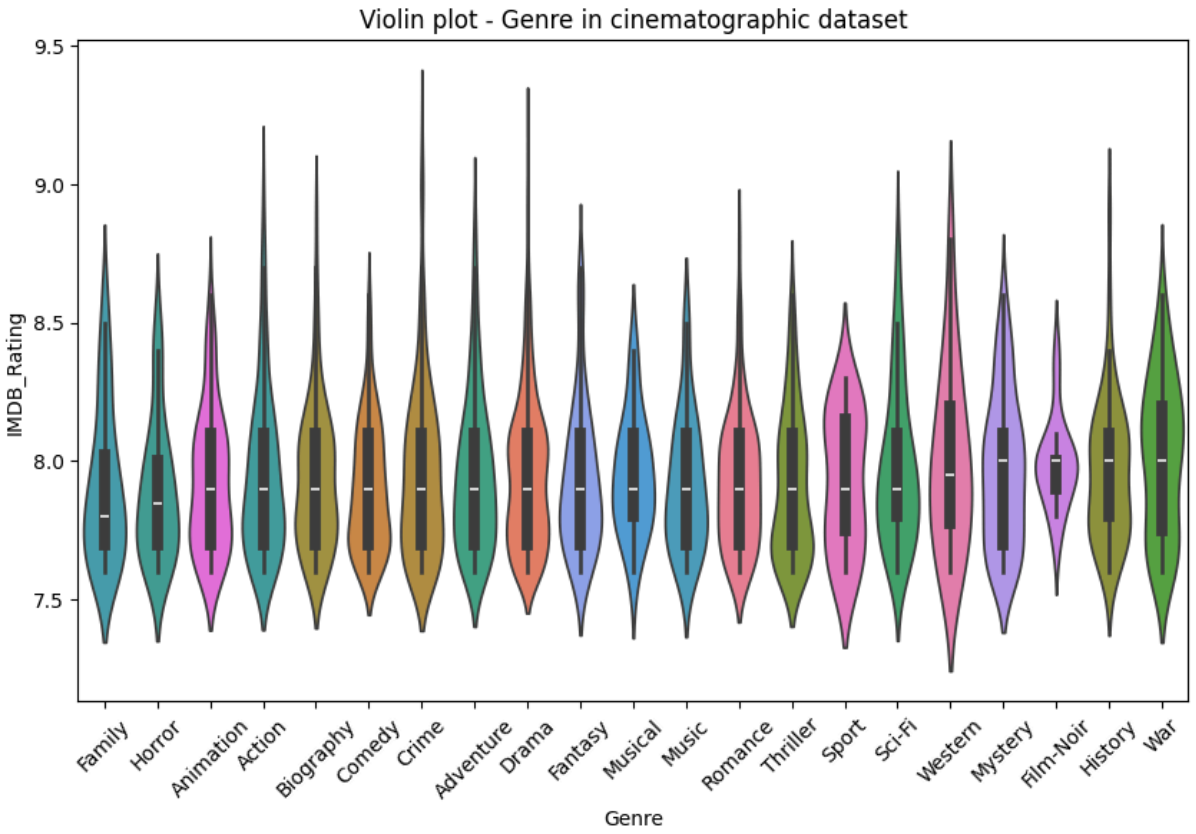
Testes ANOVA e Violin Plots

Nessa etapa realizei alguns testes ANOVA e visualizei a distribuição usando Violin Plot de variáveis categóricas. Foi utilizado 0.05 como nível de significância.

Como existem mais de 2000 atores diferentes e mais de 500 diretores (alta cardinalidade), não pude criar uma visualização da distribuição de IMDB_Rating por ator ou por diretor.

Portanto irei utilizar como estratégia de amostragem, obter até 14 atores/diretores entre o dataset para visualizar essa distribuição, essa amostragem será parametrizada de acordo com a quantidade de filmes produzidos (para diretores) e quantidade de filmes feitos (para atores), e realizar o teste ANOVA para testar as hipóteses. Fora isso, também testei categoricamente para o dataset inteiro (sem amostragem) porém não tive resultados diferentes.

Gêneros

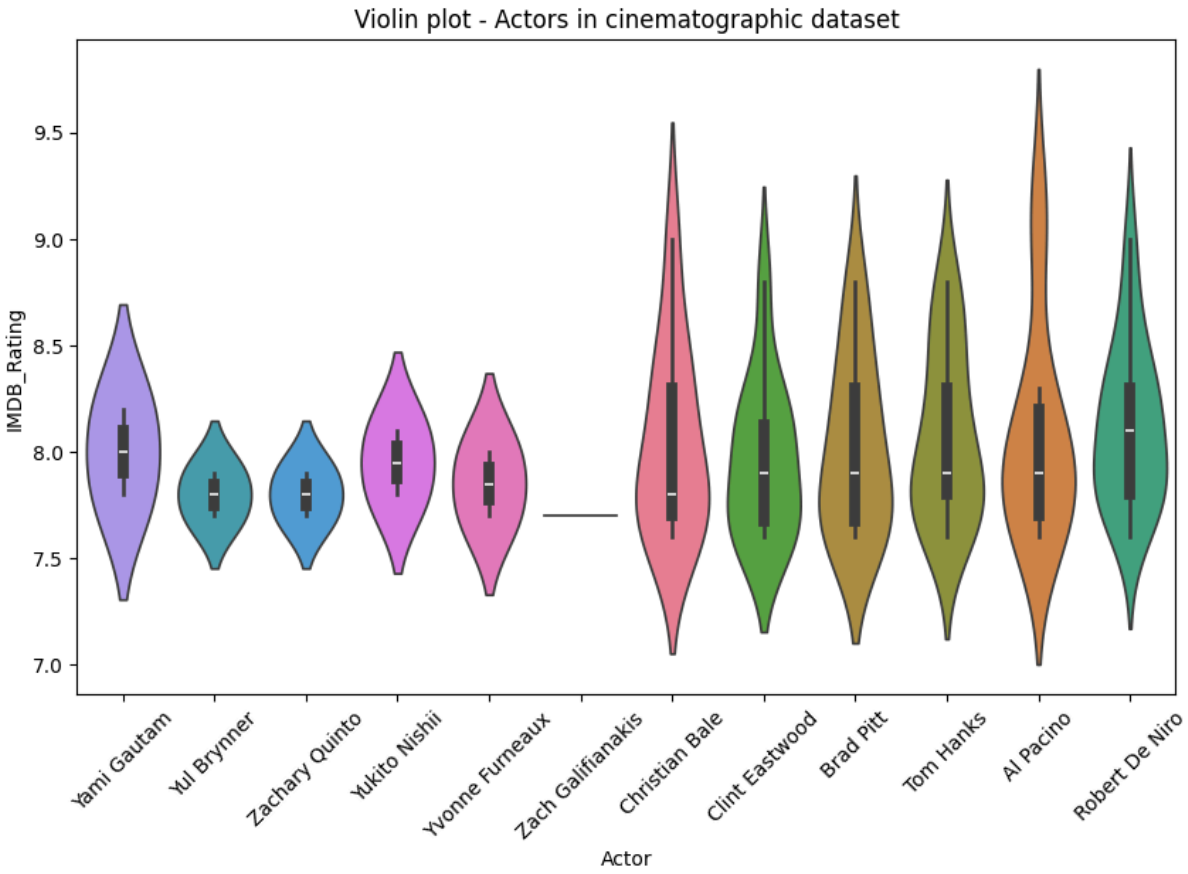


Como são relativamente poucos gêneros foi possível utilizar todos os gêneros em uma única visualização e observei o seguinte:

- Os ratings são muito semelhantes para a maioria dos gêneros, entre 7.5 e 8.3
- Poucos gêneros, como o de War, Mystery e Film-Noir parecem ter medianas um pouco mais altas, maior que 8.0
- Horror e Family possuem ter as medianas mais baixas, algo próximo de 7.7
- Em relação a distribuição é visto que a maioria são parecidas, apenas Film-Noir que é mais agrupada próximo a mediana.

-H0: Não há diferença entre as médias de IMDB_Rating entre os grupos de gênero.
-H1: Há uma diferença entre as médias.
Obtive um resultado com p-value de 0.45, muito maior que o nível de significância, levando a não rejeitar a hipótese nula, significando que provavelmente não há diferenças.

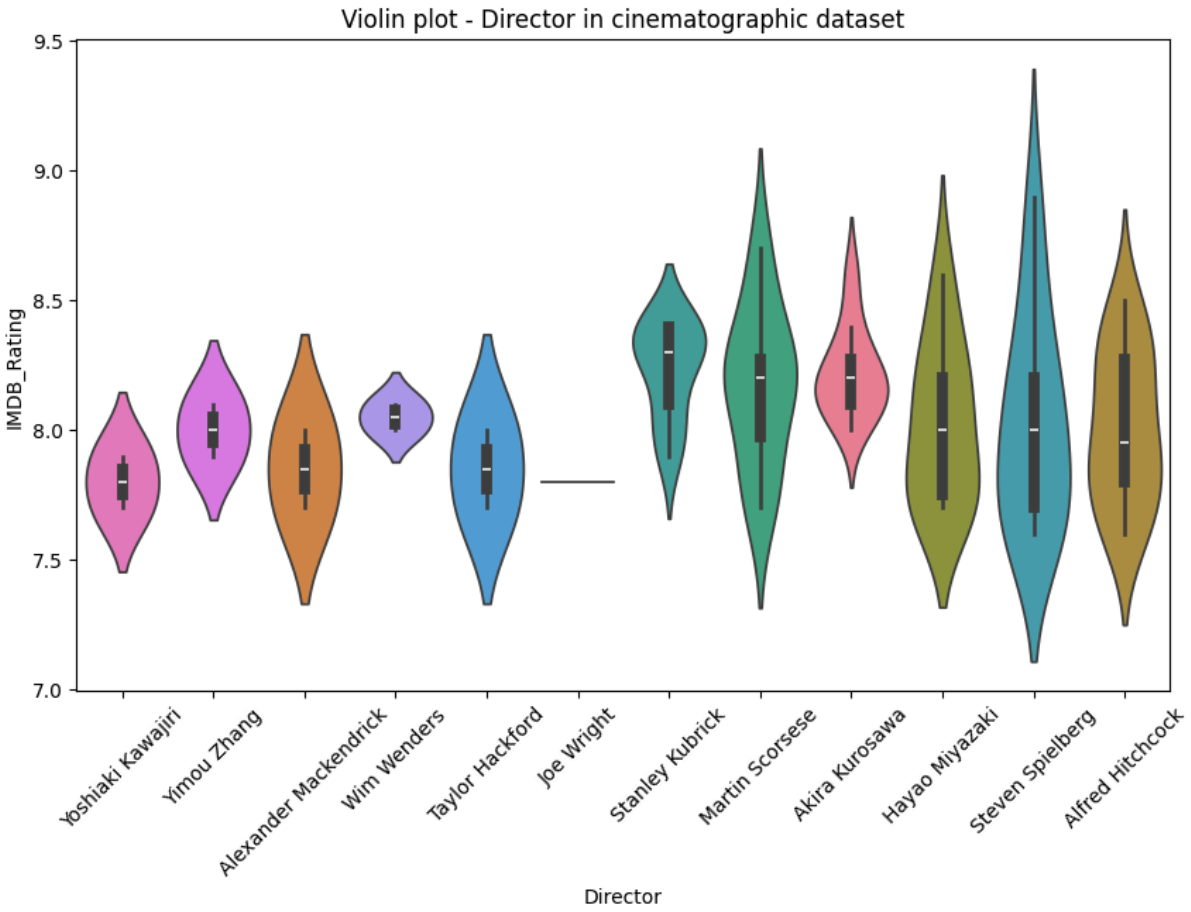
Atores (Star1, Star2, Star3, Star4)



- H0: Não há diferença entre as médias de IMDB_Rating entre os atores.
- H1: Há uma diferença entre as médias

Obtive um resultado com p-value de 0.94, levando a não rejeitar a hipótese nula.
O teste ANOVA não dá indícios de que as médias de IMDB_Rating sejam diferentes entre os atores.

Diretores

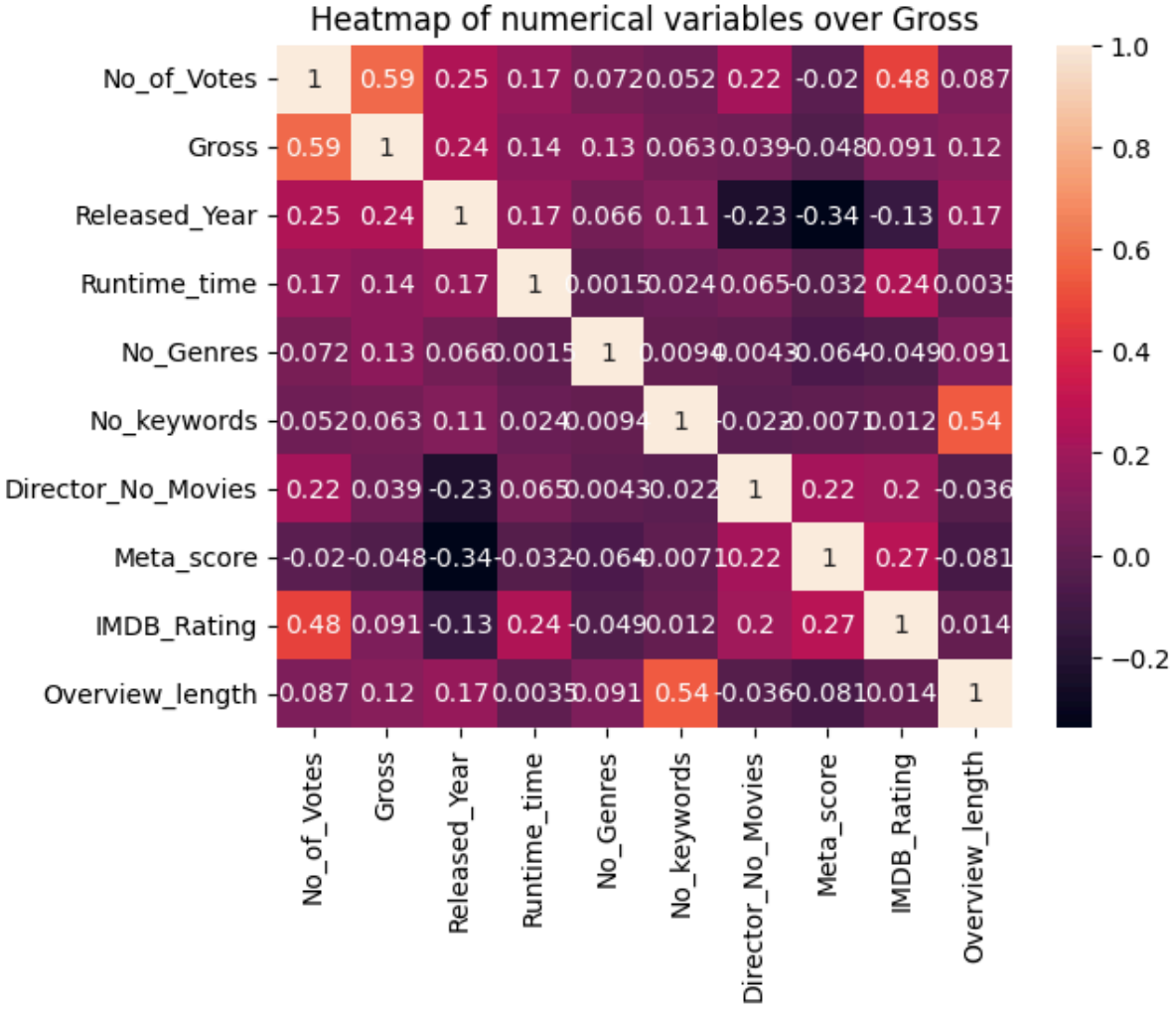


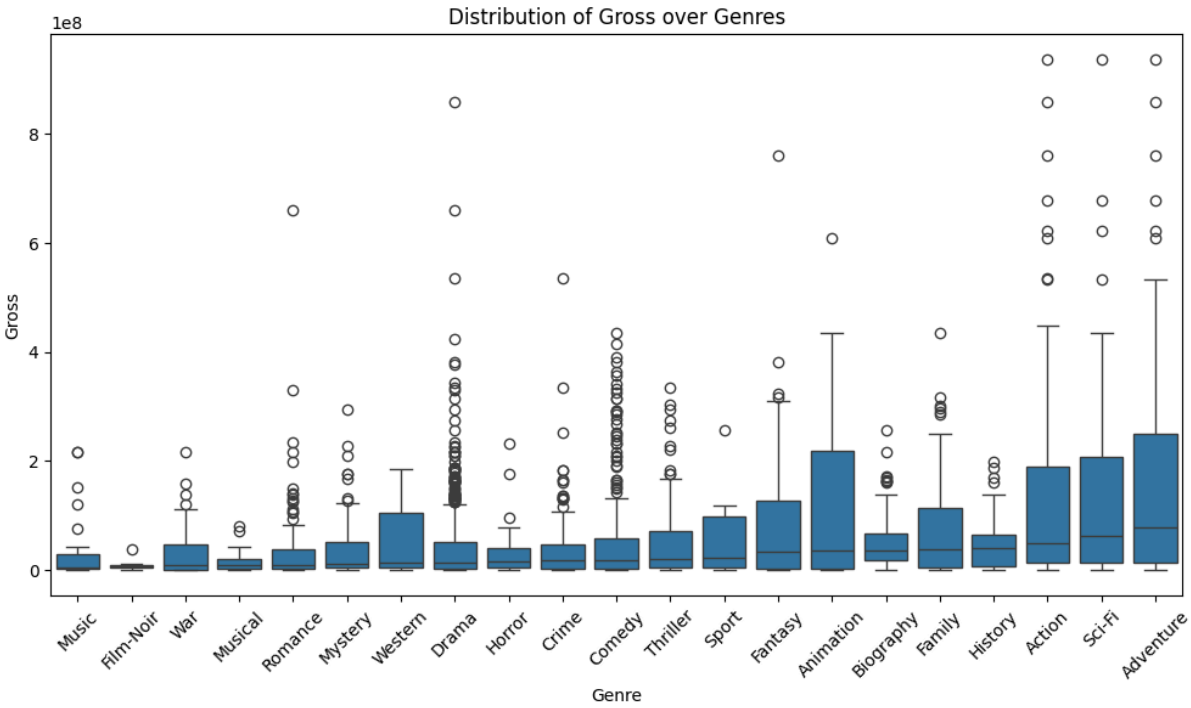
H0: Não há diferença entre as médias de IMDB_Rating entre os diretores.
H1: Há uma diferença entre as médias
Obtive um resultado com p-value de 0.08, levando a não rejeitar a hipótese nula.
O teste ANOVA não dá indícios de que as médias de IMDB_Rating sejam diferentes entre os diretores.

Respondendo as perguntas de entrega

A) Eu recomendaria The Dark Knight, é o filme mais recente (dos anos 2000 até hoje no dataset) com maior rating no IMDB e maior número de votos, o tornando mais popular.
B) As features No_of_Votes, Released_Year, Runtime, No_Genres e Overview_length são as que têm maior correlação positiva com o faturament(Gross), IMDB_Rating também porém em uma escala menor.
   Ao analisar a distribuição dos Gêneros com Gross utilizando Boxplot percebe-se também que os gêneros Fantasy, Drama, Action, Sci-Fi, Animation e Adventure alcançam faturamento (Gross) maior, sendo Adventure o gênero com maior mediana de Gross.

Usando o teste ANOVA novamente, é confirmado que essas diferenças entre os gêneros são estatisticamente significativas, com um p-value de 7.21 rejeita-se a hipótese nula.





C) A coluna overview parece ser uma descrição/sinopse do filme, isso sendo verdade é possível tirar algumas informações importantes. Apesar disso possui algumas limitações, Overview Length tem um valor médio de 25.01 com 7.7 de desvio padrão, mostra uma variância relativamente alta, e muitos filmes tem mais de um gênero associado, o que torna essa inferência mais complexa.

Com isso concluo que é com certeza possível tirar insights sobre a coluna Overview, como fiz para obter palavras chaves, porém inferir o gênero seria um problema de classificação mais complexo não determinístico por ter X gêneros associados a um filme.

Explique como você faria a previsão da nota do imdb a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Primeiramente, esse é um problema de regressão, estamos prevendo o valor de uma variável numérica (IMDB_Rating) a partir das variáveis independentes. Considerando que o dataset tem valores limitados de IMDB_Rating, por causa de algum filtro ou da origem do dataset, esse poderia também ser um problema de classificação ao criar categorías de baixo, médio ou alto rating IMDB, mas irei continuar como um problema de regressão.

Para fazer a previsão irei utilizar os dados e aplicar transformações e feature engineering para obter o máximo de informação possível dos dados, por exemplo:

- A partir do Overview pude utilizar um processo de NLP (NLTK + CountVectorizer) para transformar o em embeddings e conseguir as 300 keywords e com isso ter uma nova feature de No_keywords no dataset.

- Com Overview também criei a feature Overview_length, que é a quantidade total de palavras do Overview.

- Transformei Released_Year -> Released_Year_Group, pegando a divisão inteira por 10 e multiplicando por 10 para agrupar o ano de lançamento em agrupamentos de 10 em 10 anos.

- Separei o Runtime para Runtime_time e Runtime_category, como só temos filmes com o Runtime em minutos, posso descartar essa feature, com Runtime_time obtive o Runtime_category, com filmes short, medium ou long duration.

- Com Genre irei aplicar uma transformação a depender do algoritmo que estiver usando, como Count ou Target Encoding, mas também criei uma feature No_Genres que conta o número de gêneros no filme.

- Director também irei precisar aplicar uma transformação, como o Label Encodding por exemplo, mas também utilizei para criar a feature Director_No_Movies que conta quantos filmes o diretor produziu.

Quanto às métricas, R^2 mostra quanto da variabilidade da variável dependente (IMDB_Rating) é explicada pelas variáveis independentes do modelo, e RMSE (Raiz do erro médio ao quadrado) que é um valor importante para mostrar a diferença média entre o valor previsto e o valor real, essa métrica vem na mesma ordem da variável dependente, ou seja, na mesma ordem de IMDB_Rating, por isso é importante de se avaliar ela.

A seguir uma tabela comparativa dos modelos e quais transformações apliquei em cada.

Algoritmo	OneHot Encoder	Target Encoder	Multilabel Binarizer	Ordinal Encoder	RMSE	R2
GradientBoostingRegressor + Randomized Search + Cross Validation	Certificate, Director		Actors, Genre		0.0216	0.47
LassoRegressor + Cross Validation	Certificate, Director		Actors, Genre		0.0230	0.40
LassoRegressor + Cross Validation	Certificate	Genre, Director			0.0237	0.36
LassoRegressor + Cross Validation			Actors, Genre	Certificate e Director	0.0239	0.35

LassoRegressor + Cross Validation		Genre, Director		Certificate	0.0243	0.32
DecisionTree Regressor + Randomized Search + CrossValidation		Certificate, Director	Actors, Genre		0.0256	0.25
LassoRegressor + Randomized Search + Cross Validation			Actors, Genre	Certificate e Director	0.0272	0.15

A partir dessas avaliações o melhor modelo foi GradientBoostingtRegressor um modelo de Ensemble Learning, utilizando Cross Validation com 5 Folds para o Cross Validation, alcançando um R2 de 0.47 no teste e 0.74 no treino, um sinal de overfitting moderado, e 0.216 de RMSE.

Um filme com as características descritas no desafio teria 8.85 de IMDB Rating.