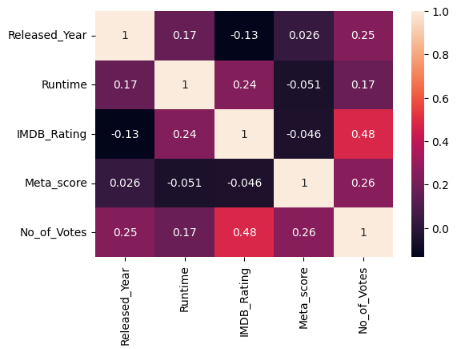
1. ANÁLISE EXPLORATÓRIA – EDA

Para iniciar a análise, foi feita uma matriz de correlação para entender a relação das variáveis



As correlações que mais se destacaram foram:

[“IMDB\_Rating”] e [“No\_of\_Votes”]: Reforça que um filme com nota alta tende a receber mais votos.

[“IMDB\_Rating”] e [“Meta\_score”]: Há uma certa discordância entre a opinião dos críticos e a do público.

[“Meta\_score”] e [“No\_of\_Votes”]: Indica que a nota dos críticos tem certa relação com a quantidade de votos. Ou seja, há relação com a popularidade, mas ainda traz o questionamento que um filme muito popular pode ter uma meta score relativamente baixa e vice-versa.

[“IMDB\_Rating”] e [“Runtime”]: A nota do filme está ligada ao tempo de tela, tornando Runtime uma variável importante para a predição.

[“No\_of\_Votes”] e [“Released\_Year”]: Nos leva a querer analisar se filmes mais antigos ainda recebem grandes quantidades de votos.

Já na familiarização com os dados, houve algumas inconsistências que foram resolvidas:

- Todos os registros e a coluna de [‘Unnamed: 0’], foram removidos por não apresentar relação com os dados, sendo apenas um identificador para os registros.

- A coluna [‘Released\_Year’] possuía um único registro ‘PG’ então foi feita a remoção e substituição pelo ano em que o filme em questão foi lançado.

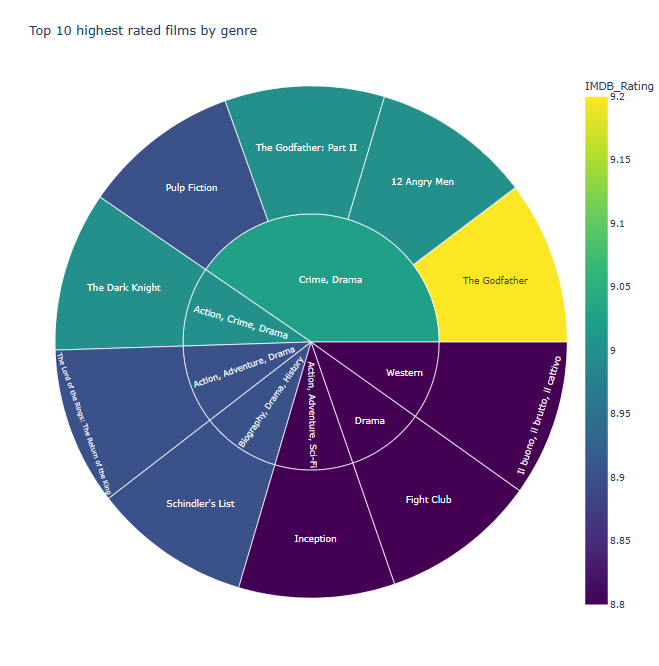
- Em [‘Runtime’], o tempo do filme estava acompanhado a ‘min’. O tipo da coluna passou a ser inteiro e o tempo ficou representado apenas em números.

- A coluna Gross possuía 177 dados faltantes e teve que ser tratada com a média para entender quais variáveis poderiam estar relacionadas com seus valores. \*

- Os valores nulos na coluna [‘Certificate’] foram substituídos por ’Unrated’, valor já existente e que indica que o filme não passou pelo teste para obter certificação.

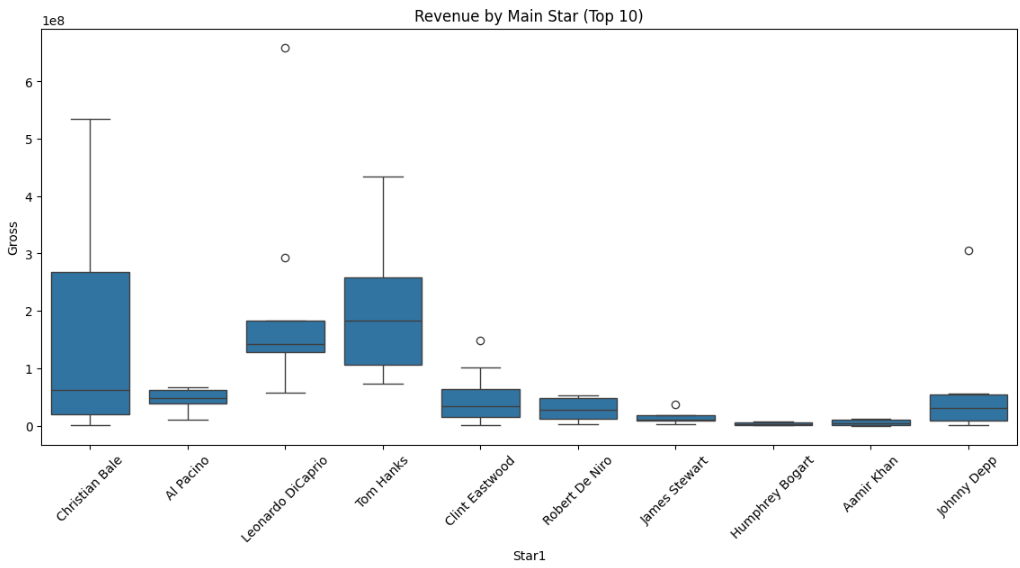
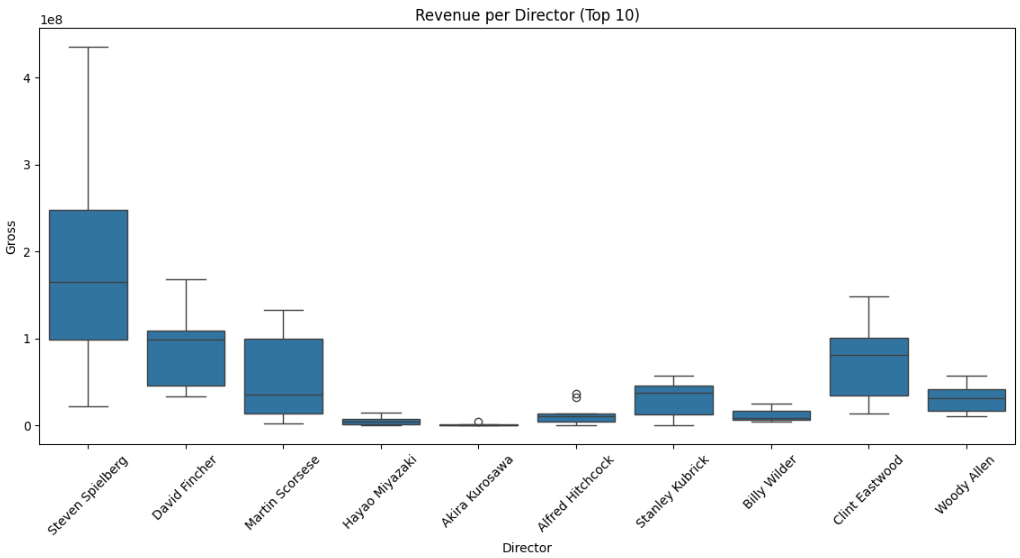
1. INDICAÇÃO DE FILME

Um filme interessante para recomendar a alguém desconhecido é “**The Lord of the Rings: The Return of The King**”. Considerando que não há dados a respeito da pessoa (como gênero, idade ou preferências), o filme conta com uma nota alta (8.9) e certificado universal, adequado para qualquer público. Apesar de ter sido lançado há alguns anos, possui muita popularidade até mesmo nos dias de hoje. Além disso, o filme ainda é considerado atemporal por sua qualidade artística e possui um elenco com atores brilhantes.



* 1. FATORES RELACIONADOS COM A ALTA EXPECTATIVA DE FATURAMENTO

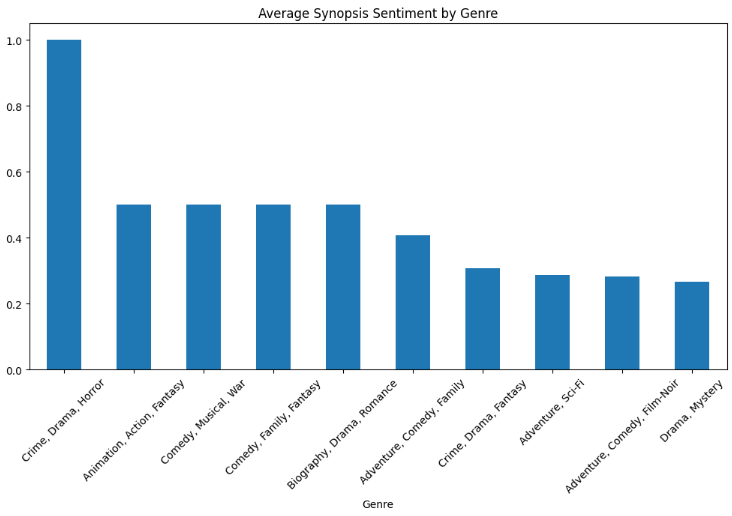
Utilizando o gráfico Boxplot, os dados de faturamento foram relacionados com todas as outras variáveis do conjunto de dados. Com isso, observou-se que o gênero e a classificação indicativa do filme **não variam drasticamente**. Porém, filmes de animação apresentam um faturamento muito mais alto que o restante dos gêneros, enquanto a classificação Universal Adulto (U/A) também comanda os faturamentos. Sendo assim, **os fatores mais importantes para investimento** são o elenco do filme: diretores e estrelas. A análise mostra que a mediana de faturamento dos diretores é superior às das outras variáveis, sem presença de outliers. Além disso, o elenco também se mostrou bastante influente em relação ao faturamento.

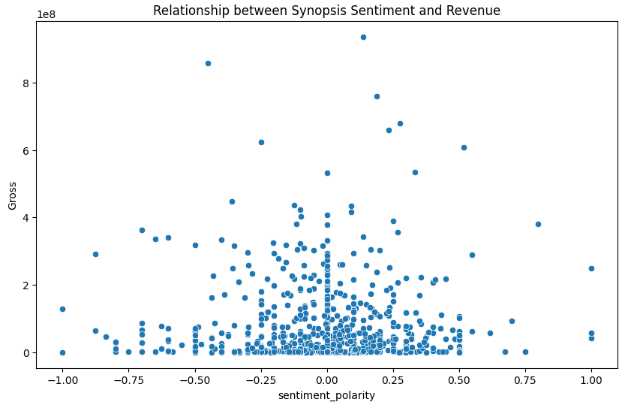


* 1. INSIGHTS DA COLUNA OVERVIEW

É possível extrair diversos insights com a coluna Overview. O tom emocional da descrição por si só não influencia significativamente o faturamento. Porém, em relação ao gênero, o sentimento varia bastante entre eles, indicando que os dados da coluna estão diretamente ligados às características do gênero.

Como exemplo, o gênero de terror tende a ter uma descrição mais negativa/neutra com palavras como “segredo”, “crime” e “conflito”. Com isso, é possível indicar a que gênero o filme pertence a depender do tom mais positivo e empolgante para gêneros como ação, comédia, animação e aventura ou tons negativos para filmes melancólicos: drama, mistério e thriller. Assim, é possível descobrir qual o gênero do filme a partir do tom predominante em sua descrição e as análises evidenciam que a Overview está diretamente alinhada ao gênero das obras.





1. PREVISÃO IMDB\_RATING

Sendo o objetivo prever a nota do filme (IMDB\_Rating), a variável-alvo é automaticamente numérica contínua, sendo um problema de regressão. O projeto foi iniciado com uma análise comparativa de 5 modelos preditivos simples para identificar a melhor performance. O resultado foi o seguinte:

1. GradientBoosting
2. RandomForest
3. Ridge

Com o modelo com a melhor performance, realizei uma otimização de parâmetros utilizando Grid Search e treinei o modelo com os dados de teste. Mesmo com a otimização, o modelo ainda não atingia o esperado, então enriqueci a base com features significativas (principalmente para as variáveis categóricas). Após isso, submeti os dois melhores modelos a outra rodade de treinamento, com novos modelos. Nessa etapa, o modelo de RandomForest se mostrou como o modelo com a melhor performance, mas ainda passaria por uma análise de overfitting que foi identificada posteriormente. Após um refinamento e testes, o modelo de **XGBoost** mostrou-se o mais ideal para os dados. As vantagens que encontrei ao utilizar esse modelo foram: rapidez no processamento e alta precisão em filmes que aparentam notas elevadas. Porém, o modelo requer um treinamento mais demorado e possui muitos parâmetros a serem ajustados, tornando-se um ponto negativo quando aplicado ao cenário do desafio (além da tendência ao overfitting). Além disso, nem todas as variáveis foram utilizadas: para um filme com atores novos por exemplo, o modelo teria muita dificuldade em aprender padrões. Ou seja, o modelo deve aprender padrões que podem ser aplicados a novos dados, então as colunas de Título, Diretor e Elenco não foram utilizadas!

1. EXEMPLO DE PREDIÇÃO

A nota prevista para o filme com essas características é 8.5/10:

{'Series\_Title': 'The Shawshank Redemption',

 'Released\_Year': '1994',

 'Certificate': 'A',

 'Runtime': '142 min',

 'Genre': 'Drama',

 'Overview': 'Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.',

 'Meta\_score': 80.0,

 'Director': 'Frank Darabont',

 'Star1': 'Tim Robbins',

 'Star2': 'Morgan Freeman',

 'Star3': 'Bob Gunton',

 'Star4': 'William Sadler',

 'No\_of\_Votes': 2343110,

 'Gross': '28,341,469'}