1. ANÁLISE INICIAL DOS DADOS

A planilha de dados sobre filmes tem as seguintes características:

- Contém 999 observações e 16 variáveis.
- As variáveis Released_Year, Runtime e Gross estão no formato de texto e precisarão ser convertidas para o formato numérico.
- Não há linhas duplicadas no conjunto de dados.
- Algumas variáveis têm valores ausentes: Gross (169), Meta_Score (157) e Certificate (101).

O tratamento de dados ausentes é uma etapa crucial no pré-processamento e análise dos dados. Existem várias maneiras de lidar com dados ausentes, como removê-los ou substituí-los por valores estimados. No entanto, neste caso, nenhuma ação foi tomada ainda, pois a abordagem ideal deve ser decidida em conjunto pelos cientistas/analistas de dados, especialistas no assunto e de acordo com as políticas da organização.

Após analisar as estatísticas das variáveis com valores ausentes, algumas estratégias possíveis para lidar com esses dados são:

- **Variável Gross**: A média é de cerca de $68,1 milhões, mas o desvio padrão é alto ($109,8 milhões), indicando grande variação nos valores. Considerando essa variação e a presença de possíveis outliers, a mediana ($23,5 milhões) pode ser uma escolha mais robusta para imputação, pois é menos afetada por outliers.
- **Variável Meta_Score**: A média é 78 e a mediana é 79. Como esses valores são próximos, isso indica que a distribuição é simétrica. Portanto, substituir os valores ausentes pela média (78) ou mediana (79) é uma opção razoável, com a média sendo preferida pela simplicidade.
- **Variável Certificate**: Como é uma coluna categórica, preencher os valores ausentes com a categoria mais comum (moda) é uma abordagem adequada. Outra opção é criar uma nova categoria "Desconhecida".

2. ESTATÍSTICAS DESCRITIVAS DAS VARIÁVEIS NUMÉRICAS

	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
Contagem	998	999	999	842	999	830
Média	1991	122,9	7,9	78	271621	68,1 M
Desvio padrão	23	28,1	0,3	12,4	320913	109,8 M
Mínimo	1920	45	7,6	28	25088	1305
1° Quartil 25%	1976	103	7,7	70	55472	3,2 M
2° Quartil 50%	1999	119	7,9	79	138356	23,5 M
3° Quartil 75%	2009	137	8,1	87	373168	80,9 M
Máximo	2020	321	9,2	100	2,3 M	936,7 M

Figura 1 – Estatísticas descritivas das variáveis numéricas

As estatísticas descritivas fornecem um panorama geral dos dados. Analisando cada variável separadamente:

- **Released_Year**: Os anos de lançamento dos filmes variam de 1920 a 2020, com uma média em torno de 1991. Isso indica que a maioria dos filmes na amostra foi lançada nas últimas décadas do século XX e nas primeiras do século XXI. Esse padrão está de acordo com dados de sites especializados como IMDb (https://www.imdb.com/), Box Office Mojo (https://www.boxofficemojo.com/) e MPA (https://www.motionpictures.org/). O aumento na produção cinematográfica nesse período pode estar ligado a avanços tecnológicos, ao crescimento do mercado de streaming e à diversificação de gêneros e públicos. A grande variação nos dados, com um desvio padrão de 23 anos, reflete a inclusão de filmes tanto antigos quanto recentes.
- **Runtime**: A duração dos filmes varia de 45 a 321 minutos, com uma média de aproximadamente 123 minutos. Esse tempo médio parece equilibrar a narrativa e o ritmo da história, além de ser adequado para exibição nos cinemas, mantendo o interesse do público e reduzindo os custos de produção. Filmes mais longos na amostra podem ser associados a gêneros específicos, como dramas épicos ou documentários.
- **IMDB_Rating**: As avaliações variam de 7,6 a 9,2, com uma média de 7,9. Essa faixa relativamente estreita sugere que os filmes na amostra são bem avaliados, provavelmente porque foram escolhidos filmes populares ou aclamados pela crítica. A pontuação IMDb é dada pelos usuários registrados no site IMDb, que avaliam filmes, programas de TV e outras produções audiovisuais.
- **Meta_score**: As pontuações variam de 28 a 100, com uma média de 77 e um desvio padrão de 12. Isso também indica uma tendência para filmes de alta qualidade. O Meta Score é uma média de críticas de várias fontes respeitadas, como críticos de cinema profissionais e publicações de entretenimento, calculada pelo site Metacritic (https://www.metacritic.com/).
- **No_of_Votes**: O número de votos varia significativamente, de cerca de 25.088 a mais de 2,3 milhões. A média de aproximadamente 271.621 votos sugere que os filmes são amplamente vistos e discutidos.
- **Gross**: O faturamento dos filmes varia drasticamente, de pouco mais de mil dólares a quase 937 milhões de dólares. A média de cerca de 68 milhões de dólares indica uma mistura de grandes sucessos de bilheteria e filmes de menor escala.

3. BOXPLOTS PARA DETECÇÃO DE OUTLIERS

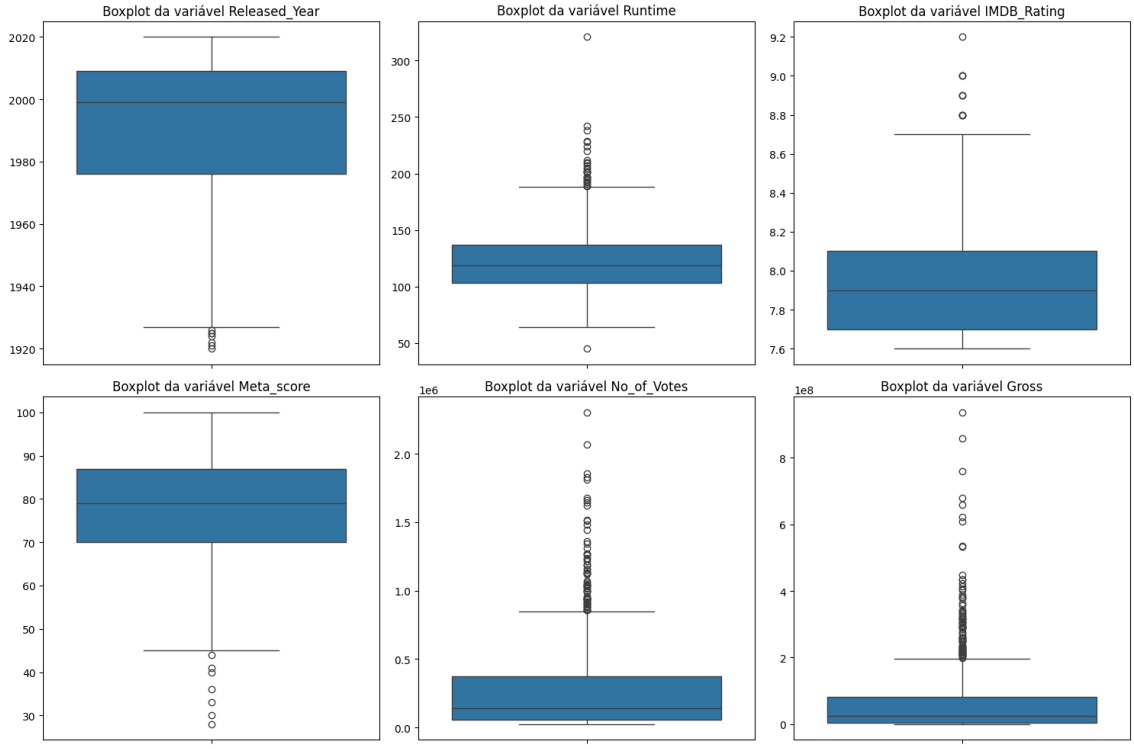

Figura 2 – Boxplots para detecção de outliers

Os boxplots são uma ferramenta útil para identificar outliers em cada uma das variáveis:

- **Released_Year**: Existem poucos outliers, o que sugere que a maioria dos filmes se concentram em décadas específicas.
- **Runtime**: Filmes muito longos (aqueles acima de 180 minutos) são considerados outliers. Isso pode ser explicado pela existência de gêneros específicos, como documentários ou filmes épicos.
- **IMDB_Rating**: Existem poucos outliers, indicando que a maioria dos filmes é consistentemente bem avaliada.
- **Meta_score**: Há outliers abaixo do valor de 40, possivelmente representando filmes que receberam críticas negativas apesar de serem populares.
- **No_of_Votes**: Existem muitos outliers com um número muito alto de votos, sugerindo que são filmes muito populares.
- **Gross**: Existem vários outliers com faturamentos muito altos, indicando grandes sucessos de bilheteria.

4. ANÁLISE DA DISTRIBUIÇÃO DAS VARIÁVEIS NUMÉRICAS

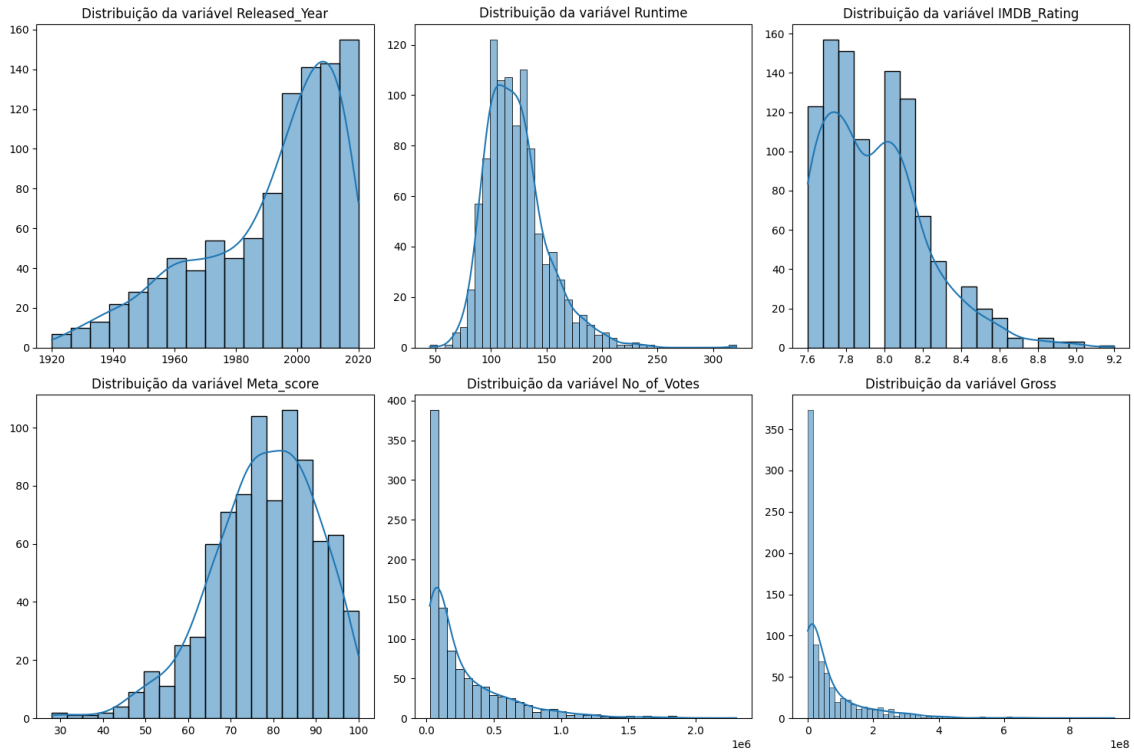

Figura 3 – Gráficos de distribuição das variáveis numéricas

Os gráficos de distribuição fornecem uma boa ideia de como as variáveis estão distribuídas:

- **Released_Year:** A distribuição é bimodal, com dois picos, um por volta das décadas de 1970-1980 e outro entre 2000-2010.
- **Runtime:** A distribuição é um pouco assimétrica à direita, com a maioria dos filmes tendo entre 90 e 150 minutos de duração.
- **IMDB_Rating:** A distribuição é aproximadamente normal, centrada em torno de 7,9, sugerindo que a maioria dos filmes são bem avaliados.
- **Meta_score:** A distribuição é mais assimétrica, com um viés negativo, indicando que a maioria dos filmes têm pontuações acima da média.
- **No_of_Votes:** A distribuição é altamente assimétrica à direita, com muitos filmes recebendo uma quantidade moderada de votos, mas alguns recebendo milhões.
- **Gross:** Semelhante ao número de votos, o faturamento é altamente assimétrica à direita, mostrando que a maioria dos filmes têm faturamentos modestos, enquanto alguns são grandes sucessos.

5. HISTOGRAMAS DAS DISTRIBUIÇÕES DAS VARIÁVEIS NUMÉRICAS

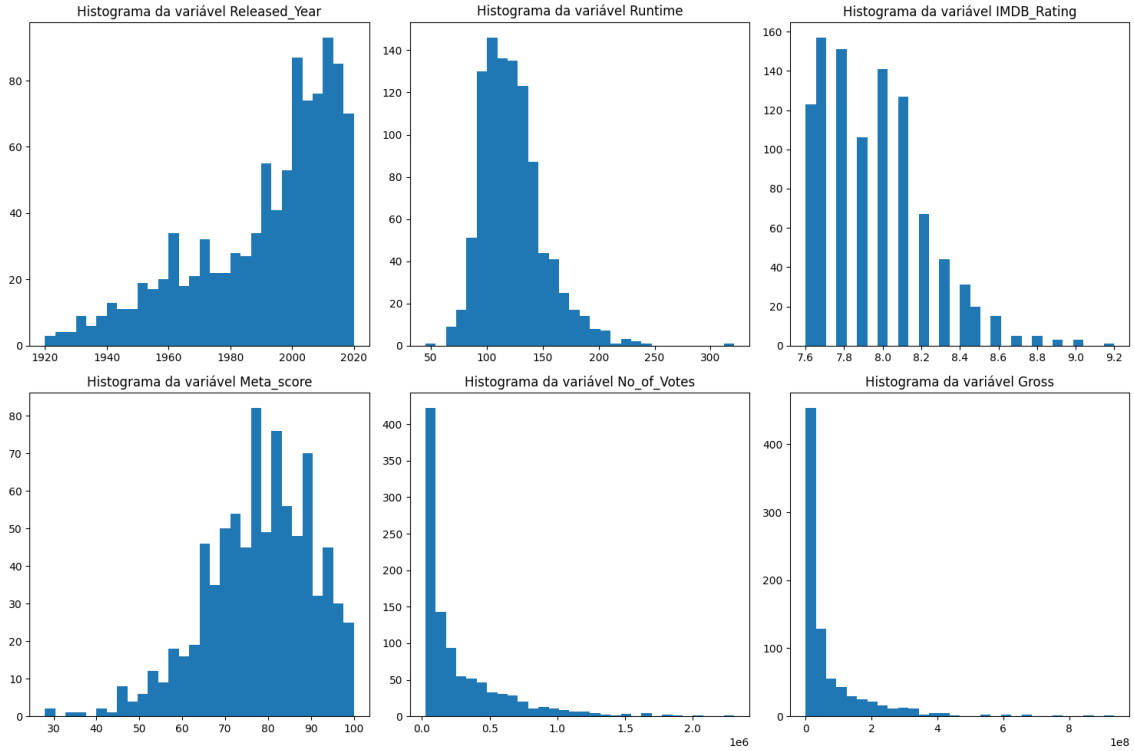

Figura 4 – Histogramas das distribuições das variáveis numéricas

Os histogramas confirmam as observações feitas nos gráficos de distribuição, mostrando visualmente a forma de cada distribuição e destacando a presença de outliers e assimetrias.

Análise detalhada dos resultados dos histogramas das distribuições:

1. **Variável Released_Year** O histograma da variável "Released_Year" mostra uma distribuição bimodal. Isso indica que houve dois períodos distintos com muitos lançamentos de filmes: um pico nas décadas de 1970-1980 e outro pico mais recente nos anos 2000-2010.

   Hipóteses:

   - **Evolução tecnológica:** A indústria cinematográfica passou por grandes avanços tecnológicos nessas décadas, como a introdução de efeitos especiais nos anos 70-80 e a transição para o cinema digital nos anos 2000.
   - **Popularidade e demanda:** A demanda por novos filmes pode ter aumentado devido ao crescimento da população e ao maior acesso a filmes por meio de novas plataformas, como o streaming.

2. **Variável Runtime** O histograma de "Runtime" mostra uma distribuição assimétrica à direita, com a maioria dos filmes tendo entre 90 e 150 minutos de duração.

   Hipóteses:

   - **Padrões industriais:** A maioria dos filmes é projetada para ter uma duração adequada para exibições em salas de cinema, geralmente entre 90 e 150 minutos, para maximizar o número de exibições por dia.
   - **Preferências do público:** O público geralmente prefere filmes que não sejam muito longos, pois podem se tornar cansativos. Filmes muito longos são frequentemente reservados para gêneros específicos, como épicos ou documentários.

3. **Variável IMDB_Rating** A distribuição de "IMDB_Rating" é aproximadamente normal, centrada em torno de 7,9, indicando que a maioria dos filmes são bem avaliados.

   Hipóteses:

   - **Seleção de filmes:** A amostra pode incluir filmes que já são populares ou bem recebidos, resultando em uma média alta.
   - **Tendência positiva:** O público pode ter uma tendência a avaliar filmes de forma mais favorável, especialmente aqueles que foram bem promovidos ou pertencem a franquias populares.

4. **Variável Meta_score** A distribuição de "Meta_score" é mais assimétrica, com um viés negativo. A maioria dos filmes tem pontuações acima da média.

   Hipóteses:

   - **Critérios de seleção:** Os filmes escolhidos para análise podem ser aqueles bem recebidos pela crítica, resultando em uma distribuição tendenciosa para pontuações altas.
   - **Críticas construtivas:** Críticos podem ser mais rigorosos em suas avaliações, mas ainda assim, filmes amplamente lançados tendem a ter uma produção de qualidade, garantindo pontuações melhores.

5. **Variável No_of_Votes** A distribuição de "No_of_Votes" é altamente assimétrica à direita. A maioria dos filmes recebe um número moderado de votos, mas alguns filmes recebem milhões de votos.

   Hipóteses:

   - **Popularidade:** Filmes extremamente populares, frequentemente de grandes franquias ou com grande marketing, tendem a receber muito mais votos do que filmes menos conhecidos.
   - **Acessibilidade:** A facilidade de votar em plataformas online como IMDb permite que filmes populares recebam uma enorme quantidade de votos.

6. **Variável Gross** A distribuição de "Gross" é similarmente assimétrica à direita, com a maioria dos filmes gerando faturamentos modestos, mas alguns sendo grandes sucessos de bilheteria.

   Hipóteses:

   - **Blockbusters:** Um pequeno número de filmes, geralmente blockbusters com grande orçamento, pode gerar faturamentos excepcionalmente altos.
   - **Distribuição de renda:** A maioria dos filmes não consegue atingir o mesmo nível de sucesso comercial, resultando em uma distribuição onde poucos filmes têm faturamentos muito altos.

6. MATRIZ DE CORRELAÇÃO ENTRE AS VARIÁVEIS NUMÉRICAS

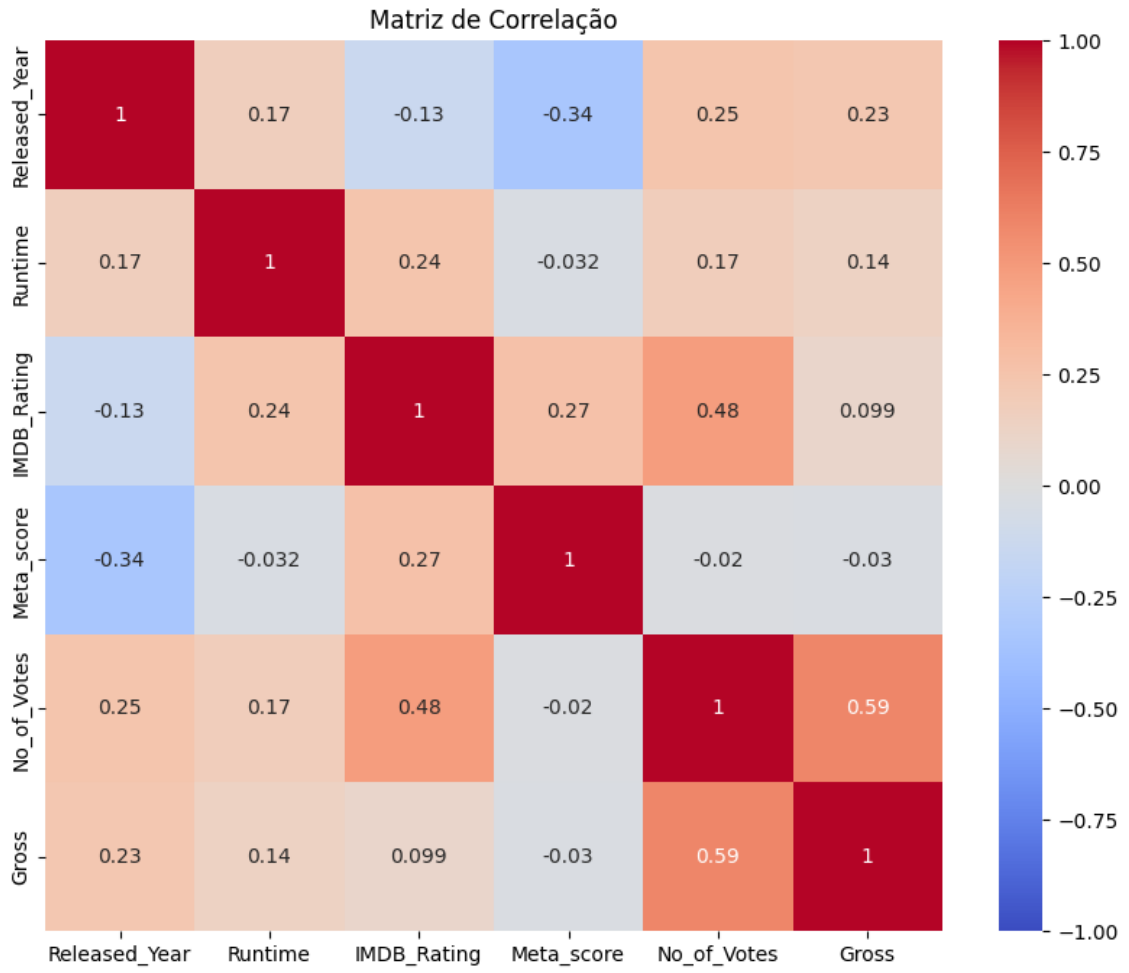

Figura 5 – Mapa de calor da matriz de correlação

A matriz de correlação mostra as relações lineares entre as variáveis numéricas do conjunto de dados.

Análise detalhada dos resultados da matriz:

1. **Relação entre No_of_Votes e Gross (0,59)**
    - **Descrição:** Há uma forte correlação positiva entre o número de votos e o faturamento. Isso sugere que filmes com mais votos tendem a gerar mais receita.
    - **Hipóteses:**
        - **Popularidade e marketing:** Filmes que atraem mais atenção e são amplamente vistos tendem a gerar maior faturamento devido ao maior alcance e visibilidade.
        - **Qualidade:** Filmes bem recebidos e com boas críticas tendem a atrair mais votos e espectadores, resultando em maior faturamento.

2. **Relação entre IMDB_Rating e Meta_score (0,27)**
    - **Descrição:** Há uma correlação positiva moderada entre as avaliações do IMDb e as pontuações dos críticos, indicando alguma concordância entre o público e os críticos, mas não uma correspondência perfeita.
    - **Hipóteses:**
        - **Diferenças de percepção:** O público e os críticos podem ter critérios diferentes para avaliar filmes. Os críticos podem focar mais em aspectos técnicos, enquanto o público pode ser influenciado por fatores emocionais ou de entretenimento.
        - **Diversidade de preferências:** A diversidade de preferências entre o público pode resultar em uma ampla gama de avaliações que nem sempre coincidem com as opiniões dos críticos.

3. **Relação entre Released_Year e Meta_score (-0,34)**
    - **Descrição:** Há uma correlação negativa entre o ano de lançamento e a pontuação do crítico, sugerindo que filmes mais antigos tendem a ter pontuações mais altas.
    - **Hipóteses:**
        - **Valorização com o tempo:** Filmes antigos podem ser vistos de forma mais favorável ao longo do tempo, seja por nostalgia ou reconhecimento de seu impacto histórico.
        - **Críticas contemporâneas:** Críticos contemporâneos podem ser mais rigorosos, ou os critérios de avaliação podem ter mudado ao longo do tempo.

4. **Relação entre No_of_Votes e IMDB_Rating (0,48)**
    - **Descrição:** Há uma correlação positiva moderada entre o número de votos e as avaliações do IMDb, sugerindo que filmes mais votados tendem a ter melhores avaliações.
    - **Hipóteses:**
        - **Efeito de popularidade:** Filmes mais populares têm maior probabilidade de atrair um grande número de votos positivos, criando um ciclo onde a popularidade aumenta a avaliação e vice-versa.
        - **Distribuição de votos:** Filmes com uma grande base de fãs podem receber mais votos positivos, influenciando a avaliação média.

7. ANÁLISE DAS VARIÁVEIS CATEGÓRICAS: GENRE E CERTIFICATE

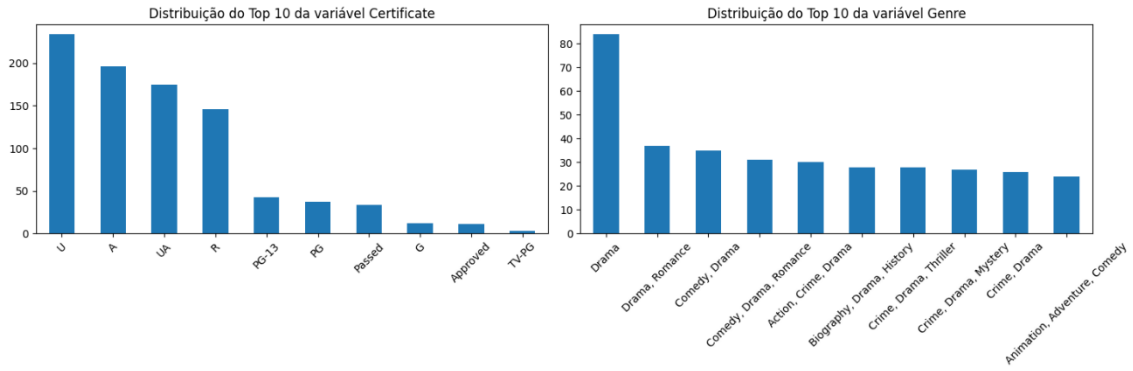

Figura 6 – distribuição das variáveis categóricas Genre e Certificate

Os gráficos de barras mostram a distribuição das seguintes variáveis categóricas:

- **Certificate:** A maioria dos filmes possui a classificação "U" (Universal), seguida pelas classificações "A" (Adulto) e "UA" (Universal Adulto). Isso indica que muitos filmes são apropriados para todas as idades, embora também haja uma quantidade significativa de filmes com restrições de idade.
- **Genre:** O drama é o gênero apresentado como sendo o mais comum, frequentemente combinado com outros gêneros como romance e comédia. Isso pode ocorrer porque dramas têm um apelo mais amplo (atraem um público mais variado) e são frequentemente premiados.

7.1. Análise da relação entre os gêneros mais comuns com notas do IMDb e faturamentos

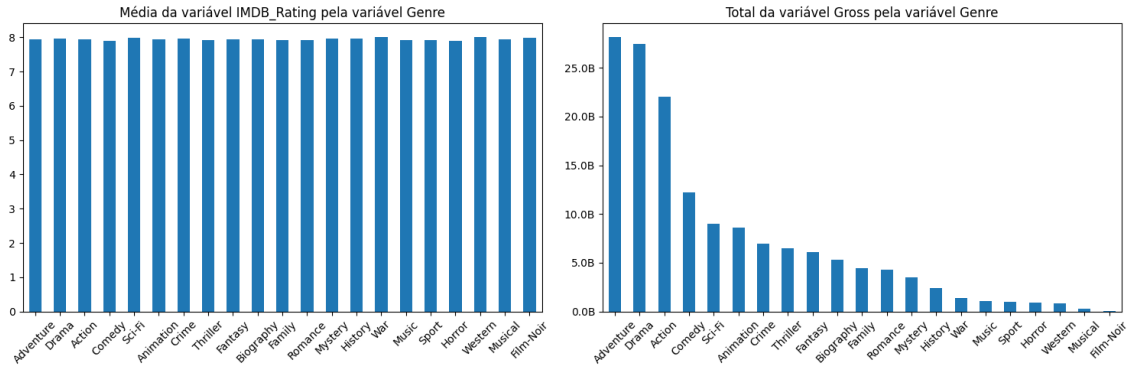

Figura 7 – Distribuição do gênero de acordo com a nota do IMDb e o faturamento

Análise detalhada dos resultados dos gráficos:

1. **Relação entre os gêneros e a média das notas do IMDb**
   - **Observações:**
     - Os gêneros War (Guerra) e Western (Faroeste) apresentam as maiores médias de avaliação, acima de 8.
     - Os gêneros Horror e Musical têm as menores médias, abaixo de 7,9.
   - **Hipóteses:**
     - **Expectativas do público:** Gêneros como War e Western podem atrair um público mais nichado, mas muito engajado, que tende a dar avaliações mais altas.
     - **Qualidade percebida:** Filmes de Horror podem variar muito em qualidade, resultando em avaliações mais baixas em média.
     - **Críticas e Premiações:** Gêneros como Drama, Biography (Biografia) e History (História) frequentemente recebem mais atenção de críticos e prêmios, influenciando avaliações mais altas.

2. **Relação entre os gêneros e a soma dos faturamentos**
   - **Observações:**
     - Os gêneros Adventure (Aventura) e Drama possuem os maiores faturamentos.
     - Os gêneros Musical e Film-Noir têm os menores faturamentos.
   - **Hipóteses:**
     - **Popularidade e apelo de massa:** Gêneros como Adventure, Drama, Action (Ação) e Comedy (Comédia) tendem a ter um apelo mais amplo, atraindo maiores audiências e, portanto, maiores faturamentos.
     - **Orçamentos e investimentos:** Filmes de gêneros populares geralmente têm maiores orçamentos de produção e marketing, o que pode levar a maiores faturamentos.
     - **Lançamentos e distribuição:** Filmes de gêneros menos populares ou nichados, como Musical e Film-Noir, podem ter lançamentos mais limitados e menos marketing, resultando em faturamentos mais baixos.

7.2. Análise da relação entre as classificações etárias com notas do IMDb e faturamentos

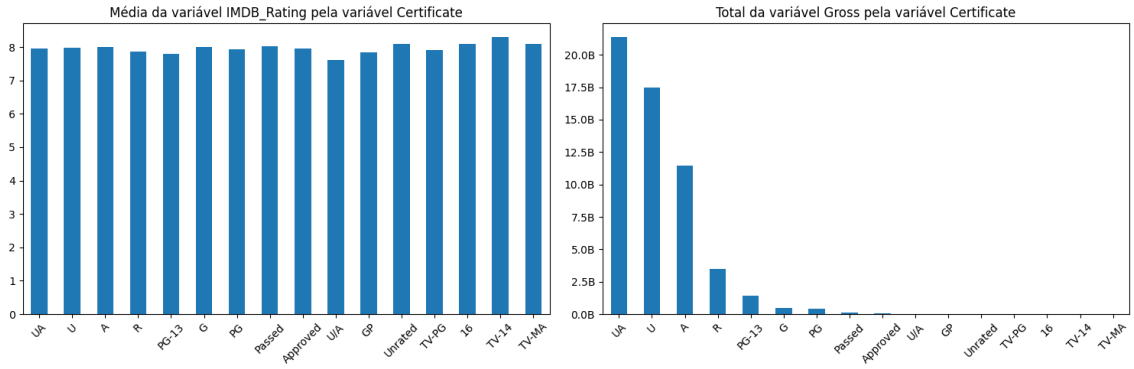

Figura 8 – Distribuição da classificação etária de acordo com a nota do IMDb e o faturamento

Análise detalhada dos resultados dos gráficos:

1. **Relação entre as classificações e a média das notas do IMDb**
   - **Observações:**
     - As classificações "Passed" (8), "Unrated" (8,1), "16" (8,1), "TV-14" (8,3) e "TV-MA" (8,1) têm as médias de avaliação mais altas.
     - As classificações "U/A" (7,6) e "PG-13" (7,8) têm médias de avaliação mais baixas.
   - **Hipóteses:**
     - **Público-alvo:** Filmes com classificação "R" tendem a ter conteúdo mais maduro, limitando seu público, mas podem atrair uma base de fãs mais engajada que valoriza a qualidade e a profundidade do conteúdo. Filmes com classificações mais amplas ("U" e "UA") atraem uma audiência maior e mais diversa, resultando em uma média de avaliação que pode refletir uma base mais ampla de opiniões.
     - **Qualidade percebida:** Filmes com classificações restritas (R, TV-MA) podem ter maior liberdade criativa, permitindo temas e narrativas que ressoam profundamente com os espectadores e resultam em avaliações mais altas. Filmes com classificações amplas precisam atender a um público mais geral, o que pode limitar a profundidade dos temas abordados.

2. **Relação entre as classificações e a soma dos faturamentos**
    - **Observações:**
        - As classificações "UA" e "U" possuem os maiores faturamentos, refletindo seu apelo de massa.
        - Classificações como "GP", "Unrated", "16", "TV-14" e "TV-MA" têm faturamentos muito baixos ou zero.
    - **Hipóteses:**
        - **Apelo de massa:** Filmes com classificação "UA" e "U" são destinados a audiências amplas e podem incluir grandes blockbusters que atraem grandes audiências, resultando em faturamentos altos. Filmes com classificação "R" e "PG-13", embora populares, têm um público mais restrito em comparação com os filmes "U" e "UA", resultando em faturamentos mais baixos.
        - **Distribuição e marketing:** Filmes com classificações como "TV-14", "TV-MA", "16" podem ter distribuição limitada, sendo destinados a plataformas específicas ou lançamentos limitados, resultando em faturamentos menores ou nulos. Filmes com classificações como "Passed" e "Approved" podem ter faturamentos acumulados ao longo de muitos anos, mas ainda assim, são inferiores aos lançamentos modernos devido à diferença no valor do dinheiro ao longo do tempo.

8. ANÁLISE DAS VARIÁVEIS CATEGÓRICAS: DIRECTOR E STAR#

8.1 Análise da relação entre os diretores com notas do IMDb e faturamentos

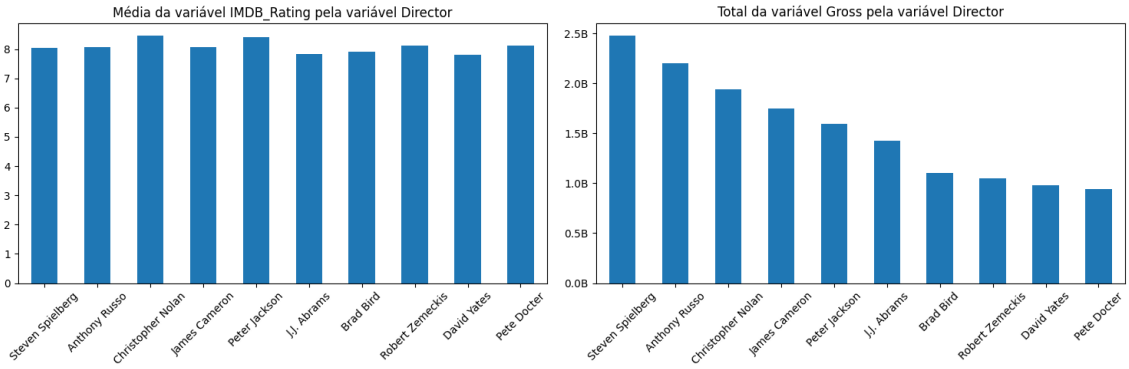

Figura 9 – Distribuição do diretor de acordo com a nota do IMDb e o faturamento

Análise detalhada dos resultados dos gráficos de acordo com artigos, matérias e entrevistas encontrados em sites especializados como IMDb, Rotten Tomatoes, Metacritic, Omelete, entre outros:

1. **Relação entre os diretores e a média das notas do IMDb**
    - **Observações:**
        - **Diretores com avaliações altas**: Christopher Nolan (8,5), Peter Jackson (8,4), e Pete Docter (8,1) estão entre os diretores com as médias mais altas.
        - **Diretores com avaliações mais baixas**: David Yates (7,8), J.J. Abrams (7,8), e Brad Bird (7,9) têm médias mais baixas.
    - **Hipóteses:**
        - **Estilo e qualidade consistente**: Diretores como Christopher Nolan e Peter Jackson são conhecidos por seus estilos únicos e narrativas complexas que agradam tanto críticos quanto o público, resultando em avaliações altas. Steven Spielberg é famoso por sua capacidade de contar histórias de maneira envolvente e por sua versatilidade em diversos gêneros, o que também contribui para altas avaliações.
        - **Engajamento do público**: Diretores como James Cameron e Anthony Russo dirigem grandes franquias e blockbusters que atraem muitos fãs dedicados, resultando em avaliações altas.
        - **Número de filmes**: Diretores com menos filmes, mas altamente avaliados, como Pete Docter, podem ter médias de avaliações mais altas devido à consistência na qualidade de seus trabalhos.

2. **Relação entre os diretores e a soma dos faturamentos**
    - **Observações:**
        - **Diretores com maiores faturamentos**: Steven Spielberg, Anthony Russo, e Christopher Nolan lideram em termos de faturamentos.
        - **Diretores com menores faturamentos**: Pete Docter e David Yates estão entre os que têm menores faturamentos.
    - **Hipóteses:**
        - **Estilo e qualidade consistente**: Diretores como Christopher Nolan e Peter Jackson são conhecidos por seus estilos únicos e narrativas complexas que agradam tanto críticos quanto o público, resultando em avaliações altas.
        - **Popularidade e apelo de massa**: Diretores como Steven Spielberg e Anthony Russo fazem filmes que atraem uma audiência ampla, resultando em faturamentos altíssimos. Spielberg, por exemplo, tem um histórico de sucessos de bilheteria ao longo de várias décadas. James Cameron é conhecido por blockbusters como "Avatar" e "Titanic", que geraram faturamentos globais enormes.
        - **Franquias e séries de sucesso**: Anthony Russo está envolvido com filmes da franquia "Marvel", que têm um apelo de massa global. Peter Jackson é famoso pelas trilogias "O Senhor dos Anéis" e "O Hobbit", que atraíram grandes audiências.
        - **Número de filmes e orçamentos**: David Yates dirigiu vários filmes da série "Harry Potter", que têm um apelo global, mas a soma dos faturamentos pode ser menor em comparação com diretores que têm um número maior de blockbusters.

8.2 Análise da relação entre os atores/atrizes com notas do IMDb e faturamentos

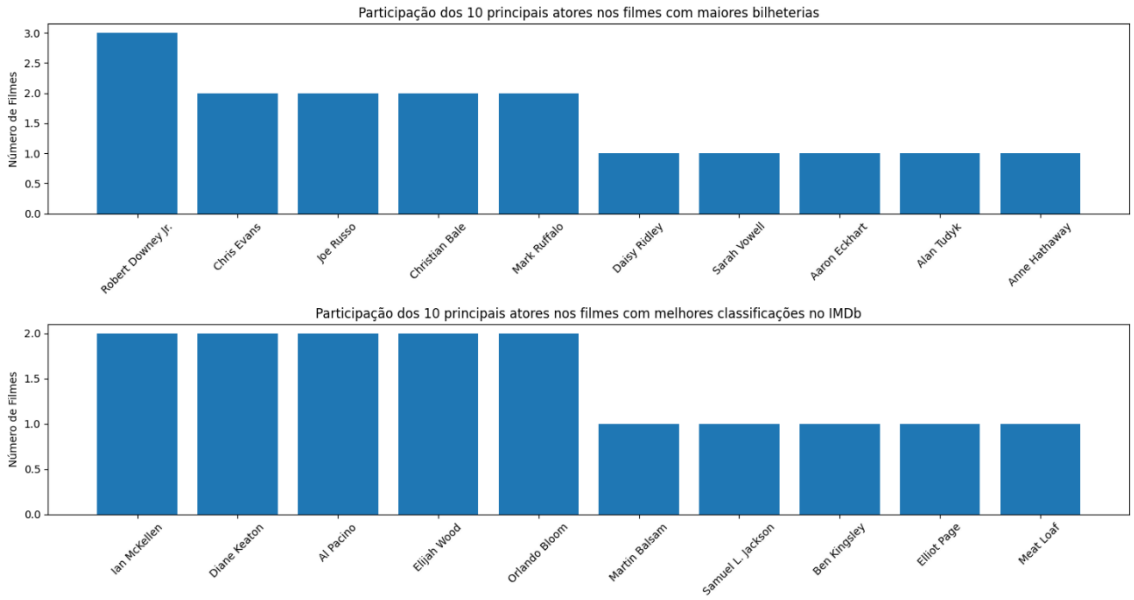

Figura 10 – Distribuição da participação dos atores/atrizes nos filmes com maiores bilheterias e melhores classificações no IMDb

Análise detalhada dos resultados dos gráficos de acordo com artigos, matérias e entrevistas encontrados em sites especializados como IMDb, Rotten Tomatoes, Metacritic, Omelete, entre outros:

1. **Participação dos 10 principais atores nos filmes com maiores bilheterias**
    - **Robert Downey Jr.**: Está no topo, principalmente devido ao seu papel significativo na franquia "Vingadores" da Marvel, conhecida por suas altas bilheterias.
    - **Chris Evans e Mark Ruffalo**: Também fazem parte do Universo Cinematográfico Marvel, contribuindo para suas altas participações.
    - **Joe Russo**: Como diretor e ator em alguns filmes de alta bilheteria, sua presença reflete o sucesso desses filmes.
    - **Christian Bale**: Conhecido por sua interpretação do Batman na trilogia de Christopher Nolan, filmes que tiveram grande sucesso de bilheteria.
    - **Outros atores como Daisy Ridley e Anne Hathaway**: Participaram de franquias de sucesso como "Star Wars" e filmes populares, respectivamente.

2. **Participação dos 10 principais atores nos filmes com melhores classificações no IMDb**
    - **Ian McKellen e Elijah Wood**: Conhecidos por seus papéis na franquia "O Senhor dos Anéis", altamente aclamada pela crítica.
    - **Diane Keaton e Al Pacino**: Estiveram em muitos filmes clássicos e aclamados pela crítica, como "O Poderoso Chefão".
    - **Orlando Bloom**: Também participou de "O Senhor dos Anéis" e "Piratas do Caribe", ambos bem recebidos pela crítica.
    - **Martin Balsam e Ben Kingsley**: Conhecidos por papéis em filmes clássicos e altamente aclamados pela crítica.
    - **Samuel L. Jackson e Elliot Page**: Têm uma carreira diversificada com participação em muitos filmes bem avaliados.

3. **Comparação entre os Gráficos**
    - **Franquias populares**: Atores que participam de franquias populares, especialmente de super-heróis ou fantasia (como MCU e "O Senhor dos Anéis"), aparecem frequentemente em ambos os gráficos.
    - **Diversidade de papéis**: Atores com uma carreira diversificada, abrangendo tanto blockbusters quanto filmes aclamados pela crítica, aparecem em ambos os gráficos.
    - **Impacto da Bilheteria vs. Crítica**: Alguns atores aparecem mais em filmes de alta bilheteria devido ao apelo de massa, enquanto outros são mais reconhecidos por suas performances em filmes criticamente aclamados.

9. ORIENTAÇÕES SOBRE O DESENVOLVIMENTO DE FILMES PARA O ESTÚDIO PPRODUCTIONS

Análise do comportamento das informações do banco de dados cinematográfico em relação ao mercado atual

1. **Gêneros populares**: A planilha mostra que os gêneros mais populares são Ação, Aventura, Drama e Thriller. Isso está em sintonia com as tendências atuais do mercado cinematográfico, onde esses gêneros continuam a atrair grandes audiências e gerar faturamentos significativos. Exemplos de grandes produções que fazem sucesso incluem "Despicable Me 4" e "Godzilla x Kong: The New Empire" (https://www.movieinsider.com/movies/genres/2024).
2. **Classificação e impacto**: Os dados indicam que filmes com classificações mais altas no IMDb tendem a ter maior faturamento e mais votos. Filmes bem avaliados geralmente atraem mais espectadores, resultando em maior engajamento e faturamento. A crítica e a avaliação do público são essenciais para o sucesso de um filme.
3. **Relação entre faturamento e número de votos**: Há uma forte correlação entre faturamento (Gross) e o número de votos (No_of_Votes) no IMDb, sugerindo que filmes que atraem mais atenção e engajamento tendem a gerar maior faturamento. Isso é evidente no mercado atual, onde filmes com grandes campanhas de marketing e forte presença nas redes sociais se destacam nas bilheterias.
4. **Tempo de execução**: Filmes com duração moderada, entre 100 e 120 minutos, são mais comuns e geralmente têm um desempenho melhor. Esse tempo de execução consegue manter a atenção do público sem se tornar cansativo, aumentando a satisfação do espectador e a probabilidade de recomendações boca a boca.
5. **Elenco e direção**: Atores e diretores renomados trazem uma reputação e prestígio que podem atrair tanto a crítica quanto o público. Por exemplo, um filme dirigido por Steven Spielberg ou estrelado por Chris Evans tende a gerar altas expectativas devido ao histórico de sucesso e qualidade. Investidores e estúdios têm mais confiança em financiar projetos com talentos reconhecidos, pois isso diminui o risco percebido e aumenta a confiança no retorno do investimento.

10. QUESTÕES PROPOSTAS

10.1. Qual filme você recomendaria para uma pessoa que você não conhece?

Para recomendar um filme a alguém desconhecido, a escolha de um filme seria baseada em critérios que indicam qualidade geral, popularidade e acessibilidade. Um bom ponto de partida seria selecionar um filme com alta pontuação no IMDb, um número significativo de votos, alto faturamento e uma boa nota no Meta Score, pois essas métricas geralmente refletem uma boa recepção tanto da crítica quanto do público.

As variáveis escolhidas foram:

1. **IMDB_Rating**:
    - **Motivo da escolha**: A pontuação no IMDB é uma métrica popular e amplamente utilizada que reflete a avaliação geral do público. Filmes com altas pontuações no IMDB são geralmente considerados de alta qualidade e agradam a uma ampla audiência.
    - **Relevância**: Indica a qualidade percebida do filme pela maioria dos espectadores.

2. **No_of_Votes**:
    - **Motivo da escolha**: O número de votos é uma métrica de popularidade que indica quantas pessoas avaliaram o filme. Um grande número de votos sugere que o filme foi amplamente visto e avaliado.
    - **Relevância**: Filmes com muitos votos são populares e têm uma avaliação mais robusta devido ao grande volume de feedback.

3. **Gross**:
    - **Motivo da escolha**: O faturamento é um indicador de sucesso comercial. Filmes com altos valores de faturamento geralmente são populares e têm ampla aceitação do público.
    - **Relevância**: Reflete a aceitação do filme no mercado e sua popularidade comercial.

4. **Meta_Score**:
    - **Motivo da escolha**: A pontuação no Metacritic é uma média ponderada das críticas de vários críticos de cinema. Incluí-la permite considerar a avaliação crítica profissional além das avaliações do público.
    - **Relevância**: Fornece uma perspectiva adicional sobre a qualidade do filme, baseada em opiniões de críticos profissionais.

Essas variáveis foram escolhidas porque oferecem uma visão equilibrada da qualidade e popularidade do filme, tanto do ponto de vista do público quanto da crítica. Elas são métricas objetivas que podem ser usadas para fazer recomendações de filmes de alta qualidade, independentemente do gosto pessoal do usuário.

Gosto pessoal e outras variáveis: Outras variáveis, como gênero, diretor e atores/atrizes, podem depender do gosto pessoal da pessoa para quem se está recomendando o filme:

- **Gênero**: Diferentes pessoas têm preferências por gêneros específicos, como ação, comédia, drama, etc.
- **Diretor**: Alguns espectadores têm diretores favoritos cujos trabalhos eles acompanham.
- **Atores/Atrizes**: A presença de certos atores ou atrizes pode influenciar a decisão de assistir a um filme.

Ao fazer uma recomendação personalizada, seria ideal considerar essas preferências específicas. No entanto, como a recomendação é para uma pessoa desconhecida, foca-se em métricas de qualidade e popularidade amplamente aceitas para garantir que a recomendação tenha uma alta probabilidade de agradar a uma ampla audiência.

Definindo Thresholds: São necessários thresholds (valores mínimos) para definir o valor mínimo que cada variável deve apresentar para que o filme seja recomendado. Esses thresholds podem ser definidos com base em quartis ou percentis para filtrar os filmes que se destacam. Nesse caso, foi utilizado o 75º percentil (3º quartil), que pode ser uma boa maneira de selecionar os melhores filmes.

De acordo com a Figura 1, os valores do 3º quartil das variáveis escolhidas são:

- IMDB_Rating: 8,1
- No_of_Votes: 373.168
- Gross: 80,9 M
- Meta_Score: 87

A partir das informações do banco de dados cinematográficos e de acordo com os valores definidos para os thresholds, algumas recomendações foram geradas:

Nome do filme	Ano de lançamento	Nota do IMDb	Meta Score	Gênero	Diretor
The Godfather	1972	9,2	100	Crime, Drama	Francis Ford Coppola
Pulp Fiction	1994	8,9	94	Crime, Drama	Quentin Tarantino
The Lord of the Rings: The Return of the King	2003	8,9	94	Action, Adventure, Drama	Peter Jackson
Schindler's List	1993	8,9	94	Biography, Drama, History	Steven Spielberg
The Lord of the Rings: The Fellowship of the Ring	2001	8,8	92	Action, Adventure, Drama	Peter Jackson
The Lord of the Rings: The Two Towers	2002	8,7	87	Action, Adventure, Drama	Peter Jackson
Saving Private Ryan	1998	8,6	91	Drama, War	Steven Spielberg
Star Wars	1977	8,6	90	Action, Adventure, Fantasy	George Lucas
The Lion King	1994	8,5	88	Animation, Adventure, Drama	Roger Allers
Back to the Future	1985	8,5	87	Adventure, Comedy, Sci-Fi	Robert Zemeckis
WALL·E	2008	8,4	95	Animation, Adventure, Family	Andrew Stanton
Apocalypse Now	1979	8,4	94	Drama, Mystery, War	Francis Ford Coppola
Spider-Man: Into the Spider-Verse	2018	8,4	87	Animation, Action, Adventure	Bob Persichetti
Toy Story	1995	8,3	95	Animation, Adventure, Comedy	John Lasseter
Toy Story 3	2010	8,2	92	Animation, Adventure, Comedy	Lee Unkrich
Up	2009	8,2	88	Animation, Adventure, Comedy	Pete Docter
Inside Out	2015	8,1	94	Animation, Adventure, Comedy	Pete Docter
Platoon	1986	8,1	92	Drama, War	Oliver Stone
Finding Nemo	2003	8,1	90	Animation, Adventure, Comedy	Andrew Stanton
The Truman Show	1998	8,1	90	Comedy, Drama	Peter Weir
Mad Max: Fury Road	2015	8,1	90	Action, Adventure, Sci-Fi	George Miller

Figura 11 – Recomendações de filmes

10.2. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

Em relação às variáveis numéricas:

A análise da matriz de correlação das variáveis numéricas revela os principais fatores que influenciam a expectativa de faturamento de um filme:

- **Número de Votos (No_of_Votes):** A correlação mais forte é com o número de votos, com um valor de 0,59. Isso indica que filmes com mais votos no IMDB tendem a ter maior faturamento.
- **Ano de Lançamento (Released_Year):** Existe uma correlação positiva de 0,23 entre o ano de lançamento e o faturamento, sugerindo que filmes mais recentes tendem a faturar mais.
- **Duração (Runtime):** A duração do filme tem uma correlação de 0,14 com o faturamento, mostrando uma relação positiva, embora não muito forte.
- **Avaliação no IMDB (IMDB_Rating):** A correlação entre a avaliação no IMDB e o faturamento é 0,1, indicando uma relação positiva, mas não tão forte quanto as outras.

Outros fatores, como a pontuação no Metacritic (Meta_score), não mostraram uma correlação significativa com o faturamento. Portanto, filmes com mais votos, lançados recentemente, com maior duração e boas avaliações no IMDB tendem a faturar mais.

Em relação às variáveis categóricas:

Para analisar as variáveis categóricas (Certificate, Genre, Director, Star1, Star2, Star3, Star4) e seu impacto no faturamento (Gross), foi utilizada a técnica de codificação one-hot para transformá-las em variáveis numéricas e calcular a correlação com o faturamento.

As variáveis categóricas que mostraram uma relação significativa com o faturamento são:

1. Gêneros de filmes:

- Action, Adventure, Sci-Fi (correlação de 0,312)
- Animation, Adventure, Comedy (correlação de 0,225)
- Action, Adventure, Fantasy (correlação de 0,221)

2. Diretores:

- Anthony Russo (correlação de 0,306)
- J.J. Abrams (correlação de 0,223)
- James Cameron (correlação de 0,200)

3. Principais estrelas:

- Joe Russo (correlação de 0,306)
- Mark Ruffalo (correlação de 0,301)
- John Boyega (correlação de 0,275)
- Daisy Ridley (correlação de 0,275)
- Chris Evans (correlação de 0,250)
- Robert Downey Jr. (correlação de 0,250)
- Zoe Saldana (correlação de 0,227)
- Domhnall Gleeson (correlação de 0,220)
- Sam Worthington (correlação de 0,219)
- Michelle Rodriguez (correlação de 0,219)
- Chris Hemsworth (correlação de 0,193)
- Sigourney Weaver (correlação de 0,193)

4. Classificação etária:

- UA (correlação de 0,284)

Esses resultados indicam que certos gêneros, diretores e estrelas estão positivamente relacionados com a alta expectativa de faturamento de um filme. Filmes de ação, aventura e ficção científica, dirigidos por nomes renomados como Anthony Russo, J.J. Abrams e James Cameron, e estrelados por atores famosos como Robert Downey Jr., Daisy Ridley e Mark Ruffalo, tendem a ter maior faturamento. Além disso, a classificação indicativa UA também está positivamente correlacionada com o faturamento.

10.3. Quais insights podem ser tirados com a coluna Overview? É possível inferir o gênero do filme a partir dessa coluna?

A coluna "Overview" oferece descrições textuais dos filmes, proporcionando uma visão geral que pode ser muito útil. Aqui estão alguns insights que podem ser extraídos:

Possíveis insights

1. **Palavras-chave frequentes:** Ao analisar as palavras mais comuns nas descrições, pode-se identificar temas recorrentes como "heroi", "batalha", "amor", etc., que ajudam a sugerir gêneros específicos.
2. **Tópicos relevantes:** Usando técnicas de processamento de linguagem natural (NLP), como a análise de tópicos, é possível descobrir os temas principais abordados nos filmes.
3. **Sentimento:** A análise de sentimento pode revelar se a descrição do filme tende a ser mais positiva, negativa ou neutra, o que pode estar relacionado a certos gêneros (por exemplo, comédias tendem a ter descrições mais positivas).
4. **Complexidade e Estilo:** A complexidade da linguagem e o estilo da descrição podem indicar o tipo de público-alvo e o gênero. Descrições complexas e detalhadas, por exemplo, podem estar associadas a dramas ou filmes históricos.

Inferência do gênero

Para inferir o gênero do filme a partir da coluna "Overview", é possível usar técnicas de aprendizado de máquina. Resumindo o processo:

1. **Pré-processamento do texto:**
    - **Limpeza do texto:** Remover pontuação, stop words, etc.
    - **Tokenização:** Dividir o texto em palavras ou n-grams.
    - **Vetorização:** Transformar o texto em vetores numéricos usando técnicas como TF-IDF ou embeddings.

2. **Modelo de classificação:**
    - **Treinamento de um modelo de classificação:** Usar as descrições como entradas e os gêneros como rótulos (ex.: Naive Bayes, SVM, ou redes neurais).
    - **Avaliação do modelo:** Garantir que ele pode prever corretamente os gêneros com base nas descrições.

3. **Validação e ajuste:**
    - **Validação cruzada:** Ajustar hiperparâmetros e evitar overfitting.
    - **Teste do modelo:** Usar um conjunto de dados separado para avaliar sua precisão.

Abordagem estruturada com NLP (Natural Language Processing)

1. **Pré-processamento do texto:**
    - **Limpeza do texto:** Remover pontuações, números, caracteres especiais e stop words.
    - **Tokenização:** Dividir o texto em unidades menores, como palavras ou n-grams.
    - **Lemmatização/Stemming:** Reduzir palavras às suas formas base ou raiz (ex.: "correndo" -> "correr").

2. **Vetorização:**
    - **Bag of Words (BoW):** Transformar o texto em um vetor de frequências de palavras.
    - **TF-IDF (Term Frequency-Inverse Document Frequency):** Ponderar as palavras pela frequência inversa dos documentos, dando mais importância a palavras menos comuns.
    - **Word embeddings:** Usar representações densas de palavras que capturam contexto, como Word2Vec, GloVe, ou embeddings gerados por modelos pré-treinados como BERT ou GPT.

3. **Extração de características:**
    - **N-grams:** Considerar sequências de palavras (bi-grams, tri-grams) para capturar contexto local.
    - **Features de sentimento:** Analisar a polaridade do texto para incluir características de sentimento.
    - **Análise de tópicos:** Usar técnicas como Latent Dirichlet Allocation (LDA) para identificar temas recorrentes no texto.

4. **Construção e treinamento do modelo:**
    - **Modelos de classificação tradicionais:**
        - **Naive Bayes:** Bom para texto devido à sua simplicidade e eficácia.
        - **Support Vector Machines (SVM):** Eficaz para classificação de texto.
        - **Random Forest:** Pode capturar relações não lineares entre características.
    - **Modelos de Deep Learning:**
        - **Redes Neurais Convolucionais (CNN):** Eficientes para capturar padrões locais no texto.
        - **Redes Neurais Recorrentes (RNN) e LSTM:** Adequadas para sequências de texto, capturando dependências de longo alcance.
        - **Transformers (BERT, GPT):** Modelos avançados que capturam contexto em múltiplas camadas de atenção.

5. **Validação e avaliação:**
    - **Validação cruzada:** Dividir os dados em partes para treinar e testar o modelo em diferentes subconjuntos.
    - **Métricas de avaliação:** Usar precisão, recall, F1-score e matriz de confusão para avaliar a performance do modelo.

6. **Implementação e melhoria contínua:**
    - **Ajuste de hiperparâmetros:** Usar técnicas como Grid Search ou Random Search para otimizar os hiperparâmetros do modelo.
    - **Feedback Loop:** Melhorar o modelo continuamente com novos dados e feedback.