1. **Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!**

   Essas são as nossas variáveis:

   Series_Title(string) – Nome do filme → É um dos primeiros aspectos que capta a atenção das pessoas e precisa ter um nome forte que condiz com o que é abordado no filme. Normalmente, nomes muito grandes tendem a ser cansativos para o consumidor, mas caso o filme apresente uma identidade visual cativante e uma presença relevante no universo audiovisual, o tamanho do nome do filme não apresentará tamanho peso na classificação de um bom filme

   Released_Year(int) - Ano de lançamento → Pode ser um fator determinante de audiência para algumas pessoas que possuem preconceitos com filmes mais antigos, mas caso sejam obras marcantes, podem transcender gerações.

   Certificate(acronyms) - Classificação etária → Ótimo como forma de proteger crianças de conteúdos não condizentes com suas idades, mas também, por muitas vezes ser uma classificação errônea e mal analisada, pode privar certos indivíduos de verem certas obras, diminuindo a audiência e a fama da obra para alguns grupos etários.

   Runtime(int) – Tempo de duração → Não acredito que possa impactar na qualidade e na relevância de um filme, pois são notórios os filmes vencedores do Oscar que possuem menos de duas horas de duração, como, por exemplo:

   - **Marty (1955)** – 90 minutos
     Annie Hall (1977) – 93 minutos
   - **Sunrise: A Song of Two Humans (1927)** – 94 minutos

- **Driving Miss Daisy (1989)** – 99 minutos
- **The Artist (2011)** – 100 minutos
- **The Broadway Melody (1929)** – 100 minutos
- **The Lost Weekend (1945)** – 101 minutos
- **Casablanca (1942)** – 102 minutos
- **The French Connection (1971)** – 104 minutos
- **It Happened One Night (1934)** – 105 minutos
- **Kramer vs. Kramer (1979)** – 105 minutos
- **Nomadland (2020)** – 107 minutos
- **On the Waterfront (1954)** – 108 minutos
- **In the Heat of the Night** – 109 minutos
- **All the King's Men (1949)** – 110 minutos
- **Cavalcade (1933)** – 110 minutos
- **CODA (2021)** – 111 minutos
- **Moonlight (2016)** – 111 minutos
- **Crash (2005)** – 112 minutos
- **Grand Hotel (1932)** – 112 minutos
- **An American in Paris (1951)** – 113 minutos
- **Chicago (2002)** – 113 minutos
- **Midnight Cowboy (1969)** – 113 minutos
- **Gigi (1958)** – 115 minutos

Genre(string) - Gênero → Utilizado para dividir filmes baseados em características em comum e ótimo para designar indivíduos que apreciam mais um gênero do que outro, podendo conhecer o fandom de um gênero de filmes e investir a audiência do seu filme nesse grupo. Não define se um filme é bom ou ruim somente pelo gênero, pois são outros aspectos que realizam esse trabalho, porém, infelizmente vemos que certos gêneros de filmes são pouco vistos em grandes premiações, como filmes de animação, de terror e de super-hérois.

IMDB_Rating(float com uma casa decimal) - Nota do IMDB → Representa a nota de avaliação dos filmes por usuários do site chamado IMDB. Pode representar em alguns pontos se o filme é bom ou ruim, mas não totalmente, pois nem toda a comunidade de críticos de filmes estão no site, nem todas as pessoas do mundo, então não podemos ter uma certeza da nota do filme baseado na população geral. Logo, essa nota é uma média ponderada de vários fatores. Mas é uma ótima medida de uso quando for escolher qual o próximo filme que irá assistir.

Overview(string) - *Overview* do filme → Um fator que, quando escrito de forma pertinente e persuasiva, pode atrair o público, mesmo o filme não sendo tão bom quanto pensávamos. Juntamente com o nome do filme, fazem uma ótima dupla, mas dificilmente definem com certeza a qualidade e relevância de um filme.

Meta_score(int) - Média ponderada de todas as críticas → Pode definir com maior precisão sobre a qualidade de um filme, porém sabemos que muitas vezes as notas dadas pelos renomados críticos de filmes não condizem com a opinião da massa, mas é um dos fatores mais relevantes, dados os citados, que classificariam se um filme é bom ou não.

Director(string) – Diretor → É possível que um mesmo diretor possa ter dirigido um filme definido como excelente e outro filme super rejeitado tanto pela crítica quanto pelo público. Dessa forma, por mais que esteja longe de ser uma regra que todos os filmes de um diretor sejam incríveis, o fato de um diretor anunciar um novo filme já faz com que a obra carregue uma certa fama, baseado nas características de direção e de temáticas do diretor e pressão de condizer com as expectativas geradas pelos outros filmes ou de quebrar as expectativas. Então não acredito ser um fator extremamente

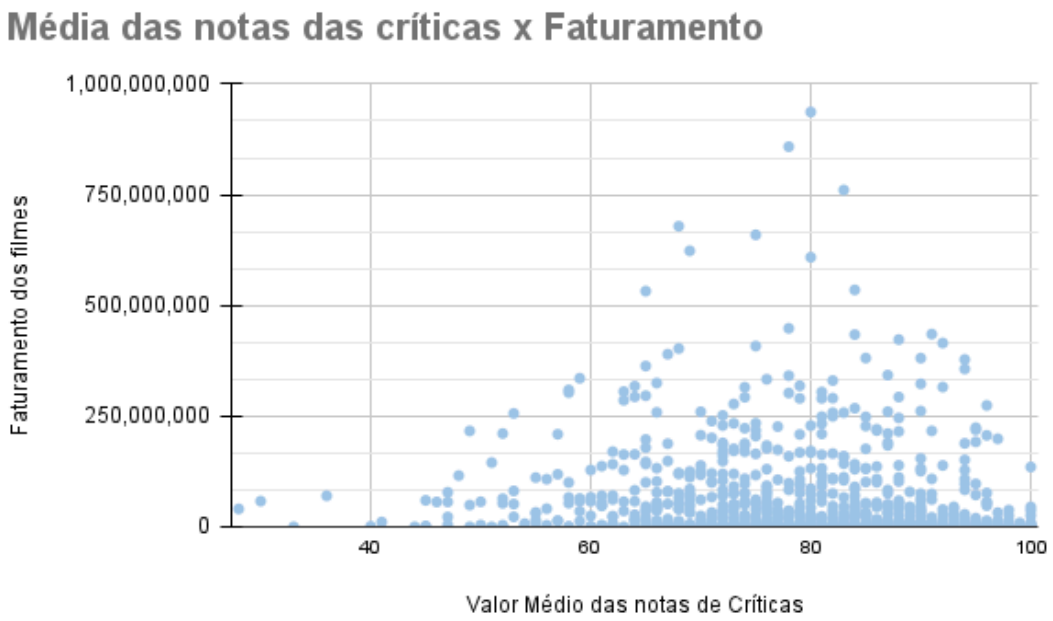
determinante para a qualidade de um filme, pois diretores são seres humanos e nem sempre estão bem para dirigir todos os seus filmes.

Star1(string) - Ator/atriz #1
Star2(string) - Ator/atriz #2
Star3 (string)- Ator/atriz #3
Star4(string)- Ator/atriz #4

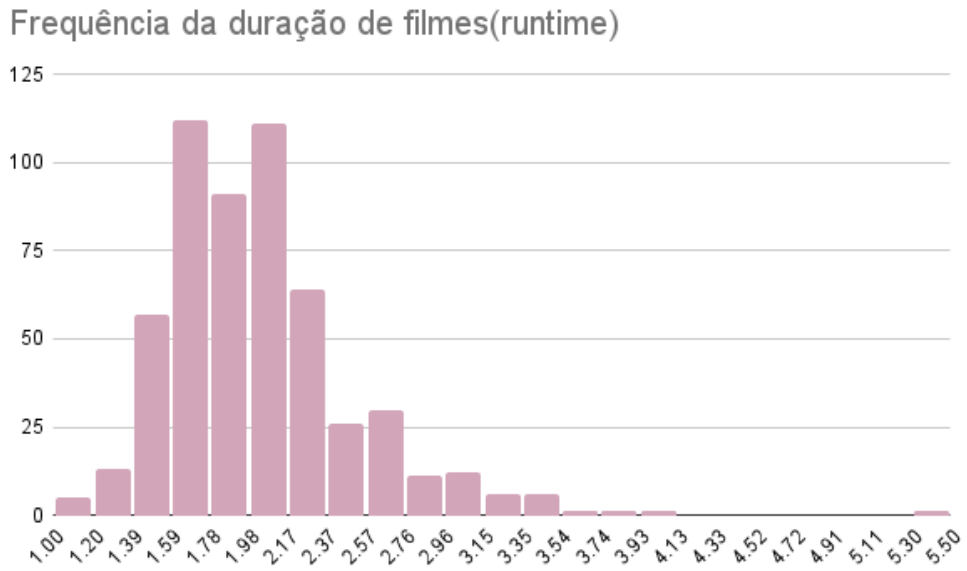
Da mesma forma que funciona com os diretores, não é porque temos um renomado ator em um filme que implica se o filme será digno de notas altas ou de uma nomeação em alguma premiação de renome, porém, atores e atrizes famosos e bons podem tanto aumentar a audiência do filme devido à visibilidade de cada um, quanto contribuir para a atuação de outros atores com poucas experiências, Além disso, caso um ator mundialmente conhecido e renomado esteja em um filme ruim, mas que possa lhe indicar alguma premiação por razões diversas, esse prêmio pode trazer visibilidade e maior audiência para o filme.

No_of_Votes(int) - Número de votos → este é outro fator que evidencia parcialmente se um filme é considerado bom ou ruim, pois mostra quantas pessoas que utilizam o site IMDB votaram no filme em questão, logo a análise deste é semelhante ao das notas do site, pois, por mais que não podemos considerar com certeza a classificação do filme baseada no número de votos, o valor da votação representa a visibilidade que certos filmes tem e quantas pessoas se importaram o bastante para comentar sobre a obra.

Gross(float) - Faturamento → este fator pode indicar, com certa força, que o filme é bom ou não, pois esse valor representa basicamente a bilheteria do filme, e sabemos que o filme é bom quando ele é visto mais de uma vez por várias pessoas ou quando uma abundância de pessoas paga para assistir à obra.



Neste primeiro gráfico que relaciona a média ponderada das notas dos críticos sobre os filmes com o faturamento de cada filme, evidencia que mesmo a crítica amando o filme, o faturamento do filme o qual a crítica considera melhor não é tão alto quanto o faturamento de filmes considerados medianos pelos críticos. Para observar isso, podemos olhar para o marco 80 e o 100 no eixo das abscissas. Sendo 100 as notas máximas, é visível que estes filmes não passam nem da metade do valor arrecadado por filmes que apresentam média 80, ou até mesmo médias bem menores. E foi escolhido o faturamento como dependente do valor das notas, porque é o público geral que gasta nos cinemas com filmes, comprando ingressos mais de uma vez para rever o filme, enquanto os críticos normalmente não agem dessa forma. Por isso essas duas variáveis foram escolhidas, pois uma representa opinião de estudiosos, enquanto a outra deixa explícita o amor do público pelo filme.



Neste histograma temos um gráfico que mostra a quantidade de filmes que pertencem a um intervalo de tempo de duração do filme, evidenciando que a maioria dos filmes não são nem muito curtos e nem muito longos, pois tendem a ser muito cansativos ou tão curtos que acabam não apresentando nenhuma profundidade de conteúdo ou que não permitem que o público desenvolva afeto pelos personagens

2. Responda também às seguintes perguntas:

a)Qual filme você recomendaria para uma pessoa que você não conhece?

Eu recomendaria o filme que possue o maior faturamento ou nota na IMDB, pois como foi visto em gráficos acima, a nota dada pelos críticos não representa, necessariamente, a massa popular,. Por isso, escolho baseado nessas duas variáveis para conseguir escolher um filme que foi “curtido” pela maior quantidade de pessoas, assim minha margem de erro seria uma das menores possíveis.

Logo, dentre esses 5 filmes dispostos em ordem decrescente de faturamento e com as notas IMDB expostas

- Star Wars: Episode VII - The Force Awakens → 7,9
- Avengers: Endgame → 8,4
- Avatar → 7,8

- Titanic → 7,8
- Incredibles 2 → 7,6

Eu iria escolher o Avengers: Endgame pelas seguintes razões:

- Uma dos maiores faturamentos;
- Dentre os de maior bilheteria é o que possue a maior avaliação no site, logo é muito provável que algum conhecido dessa pessoa provavelmente já tenha visto e já falou sobre o filme para ela(ou até ela mesma já viu), o que colocaria uma pulga atrás da orelha dela, e a deixaria tentada a assistir ou até mesmo a pesquisar.

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

- Quando o gasto para fazer o filme é muito alto, o que demonstra um esforço em entregar um filme bem feito e com efeitos visuais mais modernos e um marketing mais elaborado;
- Quando o filme é a continuação de um filme muito conhecido, ou quando é a continuação de uma saga muito relevante;
- Quando o filme apresenta atores muito renomados no elenco;
- Quando o filme representa um fechamento importante de um universo cinematográfico;
- Quando o filme é uma releitura de um livro muito conhecido e venerado;
- Quando o filme apresenta uma boa avaliação pela crítica;
- Muitas vezes o gênero do filme também está relacionado com o alto faturamento, pois existem gêneros de filmes com mais fãs

c)Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna?

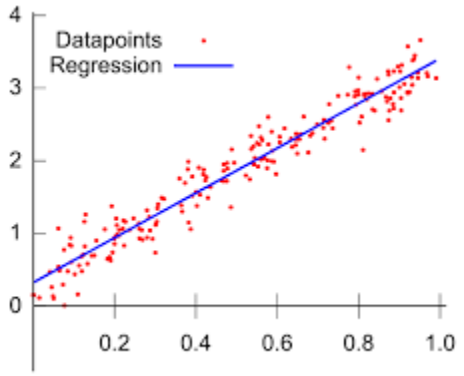
O gênero do filme, muitas vezes o rumo que o filme irá tomar, quem seriam os personagens principais do longa-metragem, os problemas que serão enfrentados pelos personagens principais, um resumo bem conciso do filme. Muitas vezes é possível, mas às vezes quem escreve o overview pode escrever de forma tendenciosa para encaixar o filme em uma categoria específica, aumentando a expectativa que o público pode ter quanto ao longa-metragem, mas muitas vezes decepcionando a massa caso o overview tenha aumentado muito os eventos que aconteceram no filme.

3. **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?**

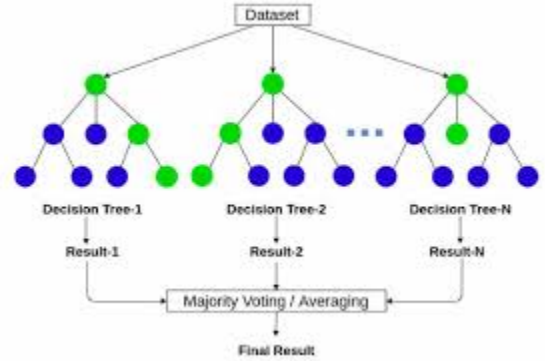
Para fazer a análise da nota IMDB a partir dos dados disponibilizados, logo isso se classifica como Regressão, pois prever um valor numérico, nesse caso entre 0 e 10, baseado em informações que iremos disponibilizar, assim como em um gráfico de dispersão com uma linha de inferência que prevê, baseado em uma previsão, como uma variável irá se comportar ao longo do tempo. Para essa análise, nem todas as variáveis poderão ser usadas, pois não seriam relevantes, como a do overview(seria útil para tentar descobrir o gênero do filme baseado em palavras-chave).

Sobre os modelos que vamos utilizar, temos:

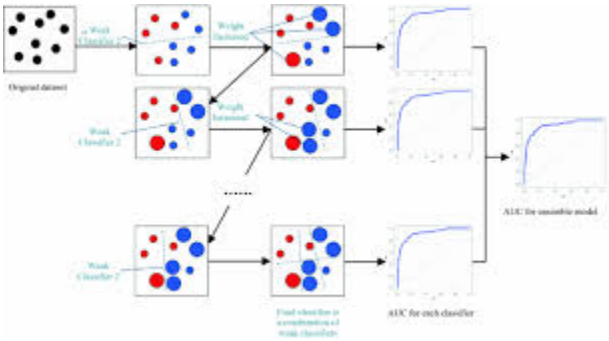
—> Regressão linear: baseadas em informações passadas anteriormente, a linha de regressão analisa e desenha um caminho linear que melhor representaria o comportamento analisado. Sobre os prós, temos que é um modelo simples de interpretar, mas (contras) caso tenhamos uma abundância de dados, pode ser um pouco lento e não há uma análise de outliers que apareceriam no gráfico.



— > Regressor Random Forest, esse modelo, conhecido como árvore de decisões, é baseado em uma tomada de decisões por meio de instruções específicas que guiam por um certo caminho os dados analisados para poder chegar em um resultado que representa a média das árvores. Seus prós são: consegue detectar relações de outliers e lida bem com dados numéricos e categóricos. Seus contras são: é relativamente difícil de interpretar e fica pesado quando tem muitos dados.



— > Gradient Boosting: é uma espécie com semelhanças ao random forest, mas as árvores são construídas uma a uma e a seguinte tenta corrigir a anterior. OS prós são: captura relações não lineares e possui alta perfomance. Sobre os contras, temos: é complexo de treinar a possue um risco de overfitting.



Sobre as medidas de performance, usaremos aquelas mais aconselhadas para a regressão, que é com o que estamos lidando nesse contexto

- **RMSE (Root Mean Squared Error)**: penaliza erros grandes, muito usado para notas de filmes e em esportes, onde a maior e a menor nota são canceladas para definir uma média.
- **MAE (Mean Absolute Error)**: penaliza erros linearmente, menos sensível a outliers.
- **R^2 (Coeficiente de Determinação)**: mostra quanto da variabilidade da nota conseguimos explicar com o modelo.

Foram utilizados como medidas de performance o **RMSE** como métrica principal (porque queremos prever a nota o mais próximo possível do valor real), e **R^2** como métrica complementar para entendermos o poder explicativo do modelo.

4. **Supondo um filme com as seguintes características:**

```
{'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'}
```

Qual seria a nota do IMDB?

Baseado no código desenvolvido pelo modelo, a nota prevista que temos é : 8.75