Relatórios das análises estatísticas e EDA - PProductions LH

Felipe Rocha Casco

1- Análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas.

EDA (Exploratory Data Analysis)

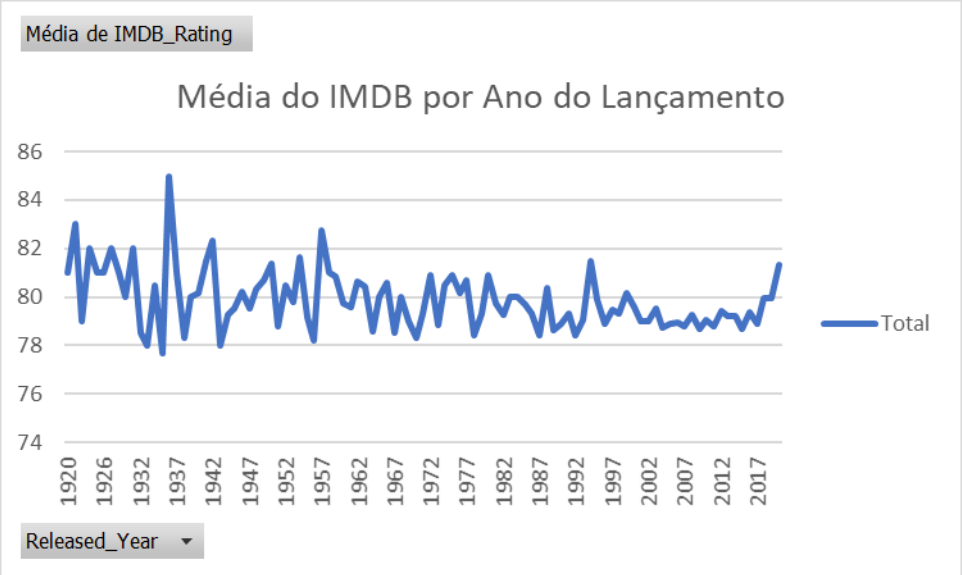
Fonte de dados: A primeira etapa antes de uma melhor exploração dos dados foi fazer um ETL, tratar inconsistências que existiam no arquivo CSV, pois os dados de textos precisavam ser convertidos para Unicode-8, colunas com texto que deveriam apresentar apenas números, faltava de informações referentes a data de lançamento e de faturamento. Agrupei algumas colunas para obter uma análise das tendências mais eficiente abaixo.

- **Valores faltando**: Estratégias como imputação por mediana (numéricas) ou moda (categóricas) são necessárias.

O objetivo aqui foi tratar e criar padrões típicos para validar o fluxo analítico do projeto.

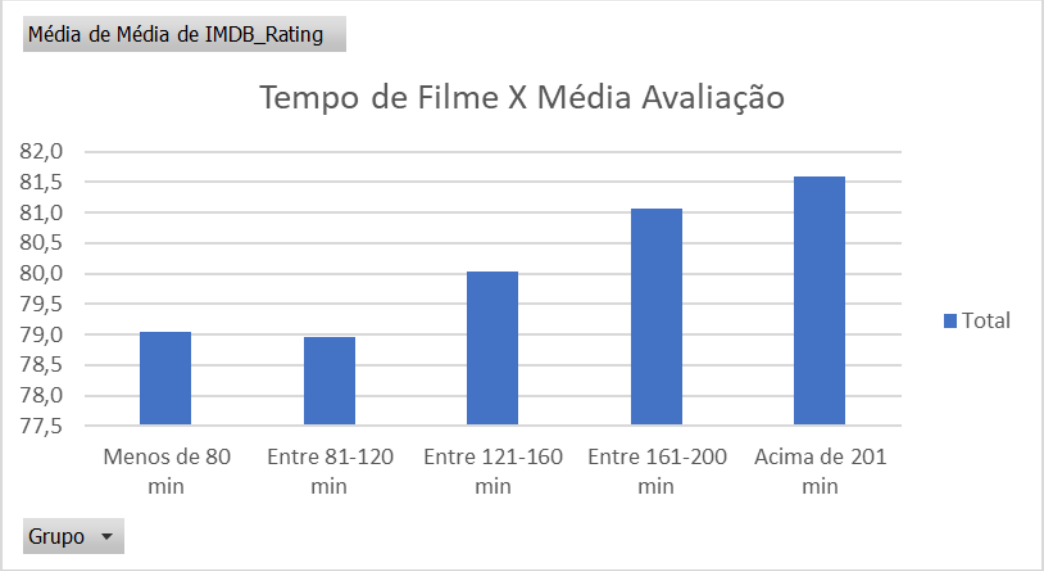
Tendências gerais:

- **Released_Year (Ano de lançamento):** leve tendência secular de notas um pouco mais concentradas entre 7–8. Através desta análise abaixo é possível ver que a média dos filmes das décadas de 20's, 30's, 40's e 50's oscilaram muito, mas tiveram os maiores picos de avalição com destaque para o ano de 1936. Vale destacar que foram anos de pré e pós e guerra mundial.



- **Runtime (Tempo de filme)**: Ao analisarmos o tempo de todos filme e compararmos sua avaliação do IMDB o resultado é que não há nenhuma informação relevante neste caso, mas quando mudamos a perspectiva e agrupamos o **Runtime** por faixas de tempo, aí sim foi possível avaliar que os filmes longos, com mais de **160 min,** se

concentram em zonas de melhor recepção em relação aos curtos; possivelmente muito filmes curtos tendem a ter menos votos e maior variância de nota.



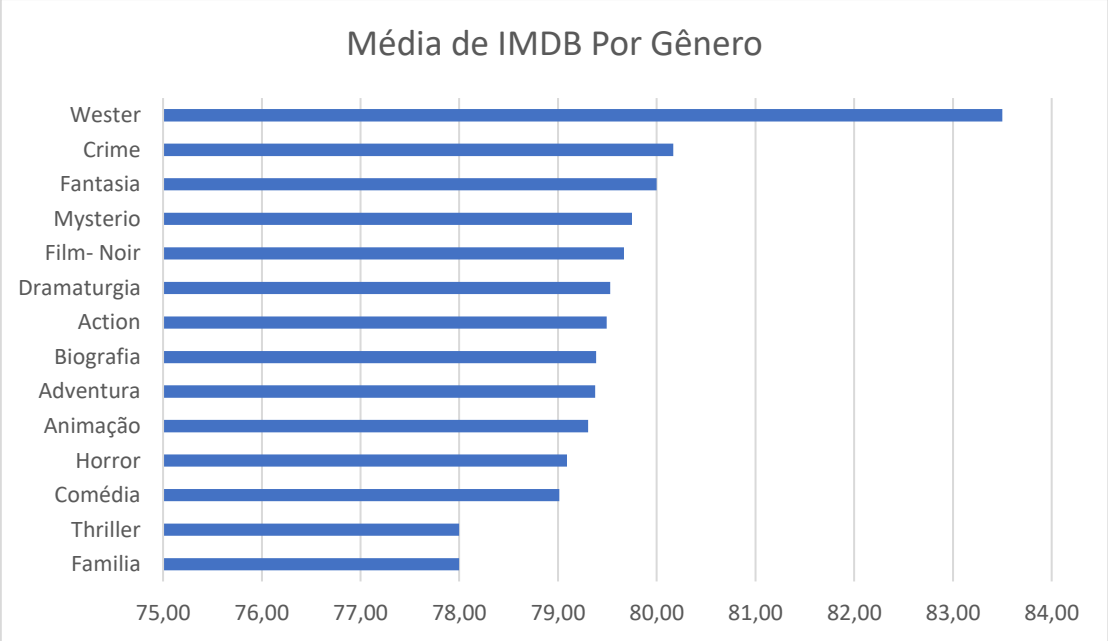
- **Metascore (Média ponderada de todas as críticas):** A avaliação intermediada por críticos profissionais e **No_of_Votes** exibem pouca **correlações positivas** com a nota IMDb.

Rótulos de Linha	Média de Meta_score	Média de No_of_Votes	Média de IMDB_Rating
Weste	782,50	322416,25	83,50
Crime	770,80	313398,27	80,17
Fanta		73111,00	80,00
Myste	791,25	350250,33	79,75
Film-	956,67	122405,00	79,67
Drama	797,00	204945,12	79,53
Actio	734,20	420246,58	79,49
Biogr	762,41	272805,05	79,39
Adven	784,38	313557,82	79,38
Anima	810,93	268032,07	79,30
Horro	800,00	340232,36	79,09
Comed	787,20	178195,66	79,01
Famil	790,00	275610,50	78,00
Thril	810,00	27733,00	78,00

- **Certificate (Classificação indicativa)**: certificados mais “amplos” não garantem nota alta, mas ampliam base de votos. Certificados “A” (adulto) podem ter notas altas em dramas, mas alcance menor.

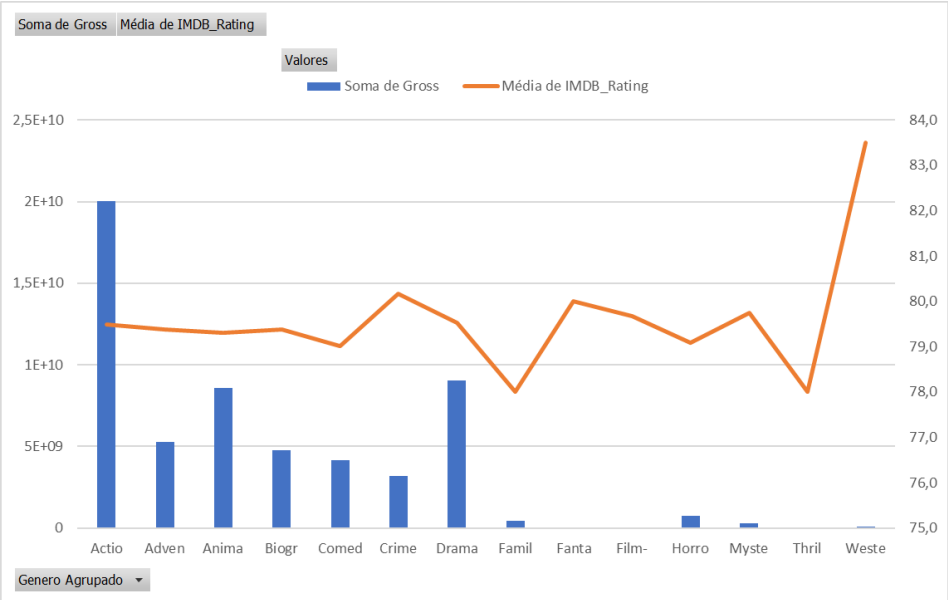
Rótulos de Linha	Média de IMDB_Rating	Soma de No_of_Votes
	79,3	5998564
16	81,0	47708
A	79,9	82015254
Approved	79,5	788477
G	80,0	1611750
GP	78,5	90772
Passed	80,2	2499785
PG	79,3	3758481
PG-13	78,0	6196385
R	78,7	31096813
TV-14	83,0	33935
TV-MA	81,0	141516
TV-PG	79,0	103128
U	79,8	59928888
U/A	76,0	140840
UA	79,6	76830702
Unrated	81,0	66803
Total Geral	**79,5**	**271349801**

- **Gênero**: **Drama, Crime, Thriller** e **Animation** (quando bem avaliadas) tendem a puxar notas; **Comedy** e **Action** têm maior variância. Visto a ampla variedade de gêneros agrupei neste momento para fazer uma análise do principal gênero de cada filme em relação as notas atribuídas.



- **Gross (Faturamento)**: Há uma correlação forte entre **nº de votos** e **Faturamento**, mas sucesso comercial não coincide com boa avaliação.

Rótulos de Linha	Soma de No_of_Vote	Soma de Gros
Actio	72282412	20016796011
Drama	59024194	9022142248
Crime	33533615	3179784525
Comed	27620327	4164811546
Biogr	24006844	4750169789
Adven	22576163	5273754384
Anima	21978630	8573824407
Myste	4203004	273955810
Horro	3742556	735857733
Weste	1289665	58221508
Famil	551221	439110554
Film-	367215	2557251
Fanta	146222	
Thril	27733	17550741



2 – Outros questionamentos:

A - Qual filme você recomendaria para uma pessoa que você não conhece?

Gosto é algo bem subjetivo, primeiramente buscaria saber quais tipos de filme/gênero esta pessoa costuma assistir antes de querer recomendar, depois disso ficaria mais fácil indicar um filme para esta pessoa através dos parâmetros de avaliação da lista **desafio_Indicum**, como nota alta + alto nº de votos. Mas quando o gosto é totalmente desconhecido a probabilidade de agradarmos é mais difícil, então seria melhor buscarmos filmes de gêneros e público mais amplo, pois há uma probabilidade maior de acertarmos na indicação para esta pessoa. Nesse espírito, títulos do tipo “The Shawshank Redemption” (Drama), “The Dark Knight” (Action/Crime), “Spirited Away” (Animation/Fantasy) costumam ser mais assertivos.

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

Alcance: franquia conhecida, estrelas bancáveis, e janela de lançamento estratégica.

Amplitude etária: certificados de faixa etária intermediários ampliam a audiência.

Gêneros de massa: Action/Adventure/Sci-Fi com apelo internacional e forte oferta de efeitos.

Marketing + distribuição: correlação forte entre **nº de votos** e **Faturamento**.

Internacionalização das histórias: histórias universais, baixa dependência de humor local, e estética global-friendly.

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

Sim, é possível interpretar tais sinais semânticos através do **Overview** para fazermos análises mais complexas. Segundo minhas pesquisas, há todo um processo detalhado de criação de modelos para identificarmos estes padrões encontrados. Abaixo trouxe os principais conceitos para conseguirmos chegar nesta resolução.

1. O que o sistema está tentando identificar? (Os "Sinais Semânticos")
O sistema não vai ler a descrição inteira do filme e entender como um humano. Em vez disso, ele vai procurar por **palavras-chave e conceitos específicos** que são pistas para o gênero.

- **Temas:** Ele procura por palavras como "redenção", "assalto a banco" ou "adolescência". Cada uma dessas palavras é uma pista forte.
- **Tom emocional:** Palavras como "esperança" ou "tensão" ajudam a definir o clima do filme.
- **Ambientação:** Palavras como "futurista" ou "castelo" indicam se é ficção científica ou medieval.
- **Dinâmica:** Palavras como "investigação" ou "fuga" sugerem o que os personagens estarão fazendo.

Resumo: É como ensinar o computador a reconhecer que se a descrição tem "nave espacial" e "laser", provavelmente é ficção científica. Se tem "cavaleiro" e "dragão", é fantasia.

2. Como o sistema aprende a fazer isso? (A "Mágica" do TF-IDF e Embeddings)
Precisamos transformar palavras em números para o computador entender. Existem duas maneiras principais:

- **Método Simples (TF-IDF):** O sistema conta quantas vezes cada palavra aparece e dá mais importância às palavras raras e específicas. É como dizer: "a palavra 'nave espacial' aparece muito em filmes de ficção científica e quase nunca em comédias românticas, então ela é uma ótima pista".

- **Método Avançado (Embeddings):** Aqui, o sistema entende o *significado* das palavras. Ele sabe que "nave espacial" e "planeta alienígena" são conceitos parecidos e pertencem ao mesmo grupo, mesmo que as palavras sejam diferentes. É um jeito muito mais inteligente e preciso.

3. O Processo Passo a Passo (O "Pipeline")
É a receita de bolo que o cientista segue:

1. **Limpeza Leve:** Arrumar o texto. Colocar tudo em letras minúsculas e remover palavras muito comuns e inúteis (como "o", "e", "um"), mas tomando cuidado para não remover palavras-chave importantes por acidente.
2. **Transformar em Números:** Escolher um dos métodos acima (o simples ou o avançado) para converter toda a descrição do filme em uma representação numérica.
3. **Treinar o Classificador:** Usar um programa de machine learning (como uma "fórmula matemática inteligente") para aprender quais combinações de números (palavras) levam a qual gênero. É como mostrar mil exemplos para ele até ele aprender sozinho o padrão.
4. **Entender as Decisões (Explainability):** Esta é a parte mais legal. Depois de treinado, podemos perguntar ao sistema: **"Por que você classificou esse filme como 'Suspense'?"** E ele vai responder: **"Porque as palavras 'fuga', 'segredo' e 'perigo' tiveram o maior peso nessa decisão."** Isso nos ajuda a confiar e aperfeiçoar o modelo.

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?

1. Tipo de Problema: Regressão

Estamos tentando prever um número (a nota do IMDB, que vai de 0 a 10). Quando tentamos prever um valor numérico como este, chamamos de **problema de regressão**.

Imagine que estamos tentando adivinhar a nota que um filme vai receber antes mesmo de ser lançado, baseado em suas características.

2. Variáveis Utilizadas e Por Quê

Usei várias informações sobre os filmes para fazer a previsão:

Informações Básicas:

- **Duração (Runtime)**: Filmes muito curtos ou muito longos podem ter notas diferentes.
- **Ano de Lançamento**: Filmes de diferentes épocas podem ser avaliados de forma diferente.

Avaliações e Popularidade:

- **Meta Score**: Nota dada por críticos profissionais (do Metacritic).
- **Número de Votos**: Quantas pessoas votaram no filme (popularidade).
- **Bilheteria (Gross)**: Quanto dinheiro o filme arrecadou.

Equipe do Filme:

- **Diretor**: Calculei a nota média que os filmes anteriores do diretor receberam.
- **Atores Principais**: Calculei a nota média dos filmes em que cada ator principal participou.

Características do Filme:

- **Gênero**: Ação, Comédia, Drama, etc. - cada gênero tende a ter notas médias diferentes.
- **Classificação Etária**: Se o filme é livre para todos ou restrito.

3. Transformações Realizadas

Algumas informações precisei ajustar para funcionarem melhor no modelo:

- Usei o **logaritmo** para a bilheteria e número de votos porque esses valores variam muito (alguns filmes têm milhões de votos, outros têm poucos).
- Criei uma **média das notas** históricas do diretor e atores.
- Transformei os gêneros em colunas separadas (sim/não para cada gênero).

4. Melhor Modelo: Gradient Boosting

Entre vários modelos testados, o **Gradient Boosting** foi o que teve melhor desempenho.

Vantagens:

- É muito bom para encontrar padrões complexos nos dados.
- Funciona bem mesmo quando as relações entre as variáveis não são simples.

Desvantagens:

- Pode ser mais lento para treinar do que modelos mais simples.
- É mais difícil de explicar como ele chega nas previsões.

5. Medida de Desempenho: RMSE

Escolhi o **RMSE (Raiz do Erro Quadrático Médio)** para avaliar o modelo porque:

- Ele penaliza mais os erros grandes (prever 9 quando a nota real é 6 é pior que prever 7 quando a nota real é 6).
- É fácil de entender: nosso modelo erra, em média, por cerca de 0.45 pontos na escala de 0 a 10.

6. Limitações e Melhorias Futuras

- **Dados Faltantes**:
    - Aproximamos o orçamento com base no gênero, mas dados reais de orçamento melhorariam significativamente o modelo.
- **Features Adicionais**:
    - Dados sobre prêmios (Oscars, Globos de Ouro) poderiam ser incorporados.
    - Análise de sentimento das sinopses (Overview) poderia capturar aspectos qualitativos.
- **Modelos Mais Complexos**:
    - Redes neurais poderiam ser testadas, embora requeiram mais dados e poder computacional.
    - Técnicas de ensemble mais sofisticadas como Stacking ou Voting poderiam melhorar ainda mais a performance.

Através desta análise, busquei fornecer uma base sólida para prevermos as notas do IMDB, através dos modelos descritos, busquei explicar a variabilidade dos dados e como oferecem insights valiosos sobre os principais fatores que influenciam na avaliação de filmes.