Pipeline dos Dados

RAW ZONE

- **Característica**: armazenar dados em estado bruto
- **Formato de objetos e fontes***: arquivos .csv e .json, e dados armazenados separados por fonte
- **Quantitativos**:
  - 'csv' IMDB = 244.543 registros únicos
  - 'json' TMDB = 44.641 registros únicos
- **Observações**:
  - 'csv' fornecido com todos os gêneros de filmes disponíveis
  - 'json' extraído (via API) já com filtro para manter apenas filmes do gênero Crime e/ou Guerra

TRUSTED ZONE

- **Característica**: armazenar tratados, desde a RawZone
- **Formato de objetos e fontes**: arquivos 'parquet' e dados armazenados ainda separados por fonte
- **Quantitativos**:
  - 'parquet' IMDB = 29.128 registros únicos
  - 'parquet' TMDB = 44.641 registros
- **Observações**:
  - dados do IMDB foram filtrados para gêneros - manter apenas registros que contém Crime e/ou Guerra
  - dados TMDB tratados sem mais filtros

REFINED ZONE

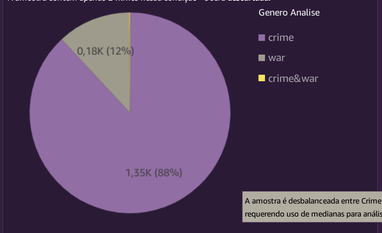
- **Característica**: armazenar dados tratados e modelados (multi-dimensionalmente), desde a TrustedZone
- **Formato de objetos e fontes**: arquivos 'parquet' e fontes combinadas, dados com *'JOIN'* realizado
- **Quantitativos**:
  - 'parquet' IMDB e TMDB (combinados) = **1.528 registros únicos**
- **Observações**:
  - IMDB têm os dados base, e dados TMDB cumprem papel de enriquecimento
  - filtros de *'NOT NULL'* em campos primordiais às análises

*Fontes dos dados: 1) IMDB - Internet Movie Database (https://www.imdb.com) | 2) TMDB - The Movie Database (https://www.themoviedb.org/)

Verificação de consistência da amostra (estatística descritiva básica)

Proporção de Filmes por Gênero na Amostra (Crime, War e Crime&War)

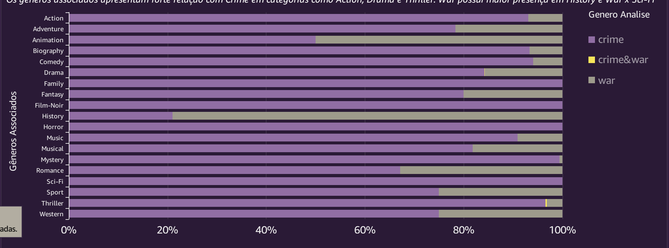
*'**Crime&War**' são filmes que contém os dois gêneros, simultaneamente.*
*A amostra contém apenas 2 filmes nessa condição - e será **descartada**.*



A amostra é desbalanceada entre Crime e War, requerendo uso de medianas para análises equiparadas.

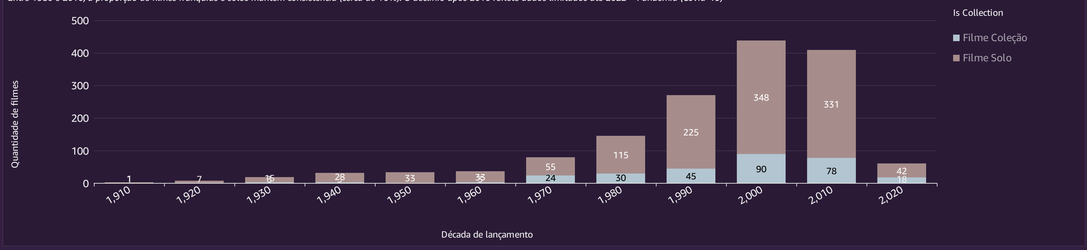
Distribuição Percentual de Gêneros Associados nos Filmes de Crime e Guerra

Os gêneros associados apresentam forte relação com Crime em categorias como Action, Drama e Thriller. War possui maior presença em History e War x Sci-Fi



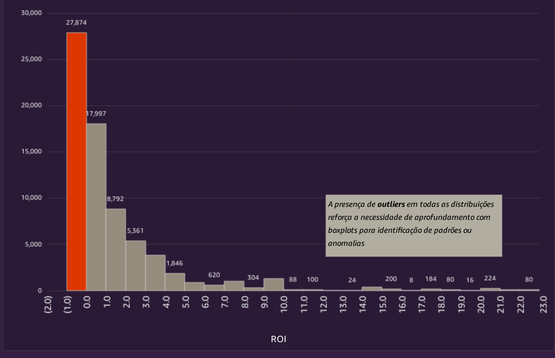
Histórico de lançamento: franquias x filmes únicos

Entre 1980 e 2010, a proporção de filmes franquias e solos mantém consistência (cerca de 16%). O declínio após 2010 reflete dados limitados até 2022 + Pandemia (Covid-19)



Distribuição Geral do ROI por Todos os Gêneros

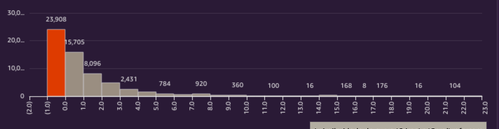
A maior concentração de ROI está entre -2 e 2, com picos de exceções na faixa negativa (-2) e positiva (1). Essa distribuição reflete um cenário balanceado, porém com outliers expressivos em valores extremos.



*A presença de **outliers** em todas as distribuições reforça a necessidade de aprofundamento com boxplots para identificação de padrões ou anomalias*

Distribuição do ROI nos Filmes do Gênero Crime

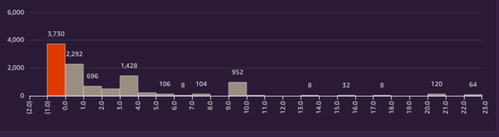
O ROI do gênero Crime segue o padrão geral, com maior concentração entre -2 e 2 e picos destacados em -2 e 1, indicando maior variabilidade



A similaridade da curva 'Crime' e 'Geral' reforçam como a amostra está desbalanceada - War tem outro formato de picos e grandeza

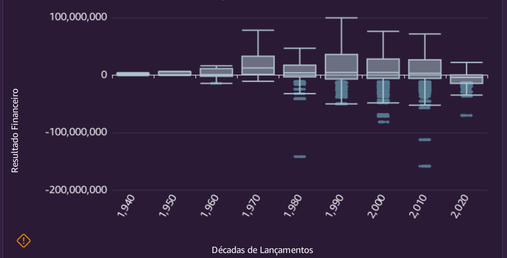
Distribuição do ROI nos Filmes do Gênero War

A distribuição do gênero War concentra-se entre -1 e 1, com picos isolados que sugerem necessidade de análise mais detalhada



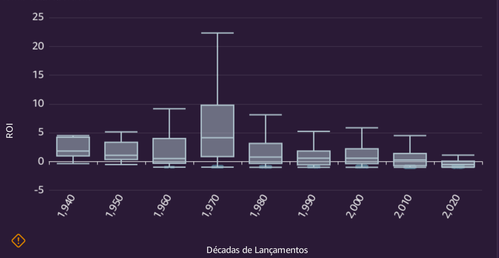
Resultado Financeiro: Estabilidade e Outliers ao Longo das Décadas

O resultado financeiro mostra consistência geral ao longo das décadas, com outliers inferiores significativos surgindo desde os anos 1970, destacando variabilidade em alguns períodos



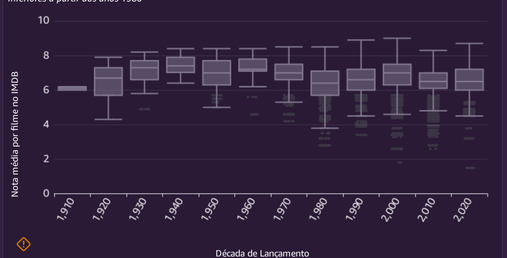
ROI: Fenômeno Isolado na Década de 1970

Apesar da estabilidade geral no ROI, a década de 1970 apresenta um fenômeno isolado que eleva o valor máximo de forma desproporcional



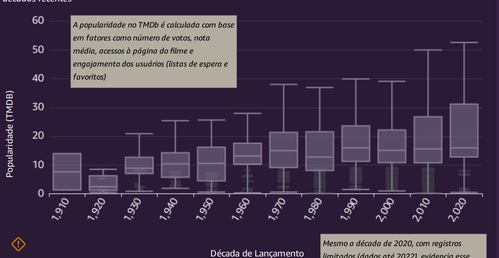
Avaliações: Nota Média no IMDB por Década

As notas médias dos filmes no IMDB apresentam estabilidade ao longo das décadas, com aumento de outliers inferiores a partir dos anos 1980



Avaliações: Popularidade no TMDB por Década

A popularidade dos filmes mostra crescimento consistente, com ampliação na dispersão e outliers inferiores em décadas recentes



A popularidade no TMDb é calculada com base em fatores como número de votos, nota média, acessos à página do filme e engajamento dos usuários (listas de espera e favoritos)

Mesmo a década de 2020, com registros limitados (dados até 2022), evidencia esse crescimento, possivelmente influenciado pelo comportamento durante a pandemia de Covid-19

Conclusões sobre a Consistência Estatística + Tratamentos Complementares Aplicados

Conclusões

- Os dados da *RefinedZone*, como estão, podem gerar interpretações enviesadas e conclusões imprecisas. Há uma necessidade evidente de tratamentos adicionais para assegurar maior robustez analítica.
- Massas de outliers foram identificadas, e sua análise separada, por meio de clustering ou outros métodos, poderia fornecer insights úteis. Entretanto, essas métricas devem ser excluídas das análises centrais para evitar distorções.
- Para cumprir nosso objetivo de gerar análises históricas consistentes que forneçam padrões e insights estratégicos, optaremos por descartar outliers, reduzindo os riscos associados às métricas financeiras e de avaliação do público.

Tratamentos Complementares

- Filtro Temporal: focaremos em produções lançadas entre 1980 e 2020, evitando picos ou vales (extremos) de dados, e que não reflitam o contexto cultural e histórico atual.
- Aplicação de Filtros Estatísticos: adotaremos apenas os dados entre os quartis Q1 e Q3 dos boxplots das métricas de Média de Notas, ROI e Resultado Financeiro, assegurando uma base mais estável para análise, desprezando outliers (especialmente os inferiores que não interessam como objetivo da análise).