Relatório de Análise da Estratégia de Conteúdo da Netflix

Autor: Gabriel Afonso Infante Vieira, Rafael de Paiva Gomes, Rafaella Cristina de Sousa Sacramento
Disciplina: Laboratório de Experimentação de Software

(i) Introdução

No cenário hipercompetitivo da "guerra do streaming", a estratégia de conteúdo tornou-se o principal campo de batalha para a aquisição e retenção de assinantes. Enquanto plataformas como a Disney+ alavancam um vasto catálogo de propriedade intelectual (PI) legada, a Netflix foi pioneira em uma estratégia agressiva de produção original para complementar seu conteúdo licenciado. Compreender essa estratégia é fundamental para decifrar seu domínio de mercado.

No entanto, o acesso a dados proprietários (como orçamentos e números de audiência) é restrito. Surge então, um desafio de engenharia de software: que *insights* estratégicos podemos extrair analisando exclusivamente o catálogo público da plataforma?

Este relatório aborda exatamente essa questão. Utilizando um conjunto de dados públicos do Kaggle (Netflix Movies and TV Shows) e a ferramenta de Business Intelligence Microsoft Power BI, conduzimos uma análise exploratória para responder a quatro questões de pesquisa (RQs) fundamentais.

Argumentamos que, mesmo sem dados de audiência, a "casca" do catálogo da Netflix revela uma estratégia de conteúdo deliberada, focada em uma mudança agressiva de filmes licenciados para a produção de séries de TV, em uma dominação de produção nos EUA complementada por uma estratégia "glocal" direcionada, e em um modelo de produção de séries focado em "1 Temporada", sugerindo uma abordagem de teste de mercado ágil e, por vezes, implacável.

(ii) Metodologia e Descrição da Base

Para este projeto, utilizamos o "Trabalho Alternativo" proposto, selecionando um *dataset* público relevante e aderindo a um processo de análise rigoroso.

1. Fonte de Dados:

- **Dataset:** "Netflix Movies and TV Shows" (Kaggle). O *dataset* original cobre o catálogo até Setembro de 2021, contendo 8.807 títulos.
- **Ferramenta de Análise:** Microsoft Power BI Desktop.
- **Arquivos de Suporte (Gerados via Script Python):** 6 arquivos CSV pré-agregados (annual_trends, top10_countries, country_genre_counts, country_counts,

heatmap_month_year) e 1 arquivo de dados brutos procesados (netflix_tratado_final).

2. Preparação dos Dados (Pré-Processamento Programático)

A preparação dos dados foi realizada programaticamente utilizando a biblioteca Pandas em Python para garantir a reprodutibilidade. Este processo foi essencial para converter os dados brutos em um formato analítico. O processo incluiu:

- **Tratamento de Múltiplos Valores:** As colunas country e listed_in (gênero), que continham múltiplas entradas separadas por vírgula (ex: "United States, India"), foram normalizadas. Para cada linha, os valores foram separados e "explodidos" em linhas individuais, permitindo uma contagem precisa da produção por país e gênero. Esta etapa expandiu o *dataset* de 8.807 para 25.895 linhas analíticas.
- **Extração de Dados de Duração:** A coluna duration (ex: "90 min", "2 Seasons") foi dividida em duas novas colunas: duration_value (numérica) e duration_unit (categórica), para permitir análises estatísticas distintas entre filmes (minutos) e séries (temporadas).
- **Tratamento de Datas:** A coluna date_added foi convertida para o formato datetime padrão, e novas colunas (added_year, added_month) foram criadas para facilitar as análises de sazonalidade e tendência.
- **Geração de Agregados:** Para otimizar o desempenho no Power BI, foram gerados 5 arquivos CSV pré-agregados para alimentar diretamente os visuais mais complexos.

Para garantir a total reprodutibilidade desta análise, todos os detalhes sobre o ambiente de execução, scripts, testes de qualidade e decisões de engenharia de software estão documentados no projeto.

(iii) Resultados: Respondendo às Questões de Pesquisa

O dashboard final foi estruturado em quatro páginas, cada uma respondendo a uma Questão de Pesquisa (RQ).

RQ1: Como o catálogo evoluiu ao longo dos anos?

- **Afirmação:** A Netflix executou uma mudança estratégica e massiva, migrando de um catálogo dominado por filmes licenciados para um focado na produção de Séries de TV.
- **Evidência:** A **Figura 1**, um gráfico de áreas empilhadas, demonstra um crescimento exponencial de todo o catálogo a partir de 2015. Mais importante, ele mostra a linha de "TV Show" (Séries) ultrapassando "Movie" (Filmes) em volume de adições anuais por volta de 2020-2021, chegando a **1218** adições anuais no seu pico de 2020.
- **Comentário:** Isso é um *insight* estratégico. Filmes geram aquisição, mas Séries geram *retenção* (engajamento de longo prazo). O gráfico visualiza essa mudança de foco de negócios.

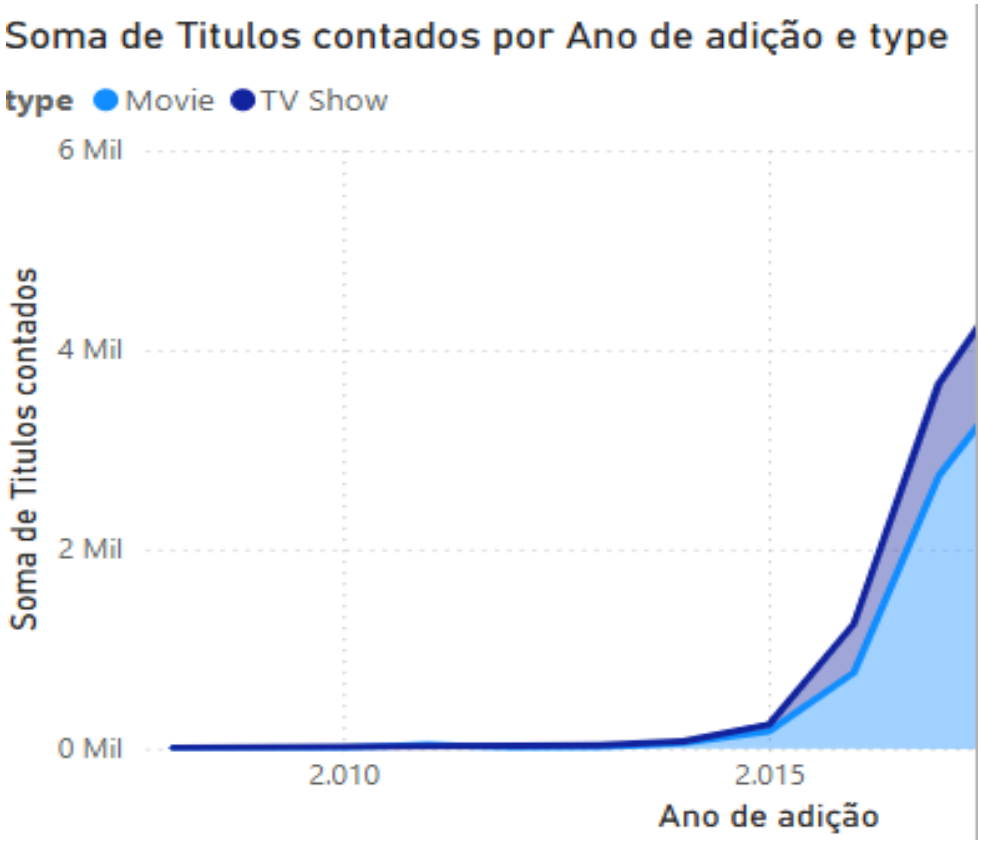

Figura 1. Evolução das adições ao catálogo da Netflix por tipo de conteúdo (Filme vs. Série de TV), 2008-2021.

RQ2: Qual é a distribuição geográfica da produção?

- **Afirmação:** A estratégia de produção da Netflix é de "dominação global, com especialização local".
- **Evidência:** A **Figura 2** (gráfico de barras) evidencia que os Estados Unidos são, de longe, o maior produtor de conteúdo, com 7465 títulos (quase o triplo do segundo colocado, a Índia, com 2999. A **Figura 3** (mapa coroplético) confirma visualmente essa concentração. No entanto, a **Figura 4** (matriz de gênero) mostra uma especialização fascinante: o Brasil é forte em "Comédias" e "Children & Family", enquanto o Japão domina "Anime Series".
- **Comentário:** A Netflix não está apenas fazendo conteúdo americano para o mundo. Ela está usando os EUA como sua base de produção principal, enquanto cultiva centros de excelência regionais ("glocal") para gêneros específicos, aumentando sua penetração de mercado local.

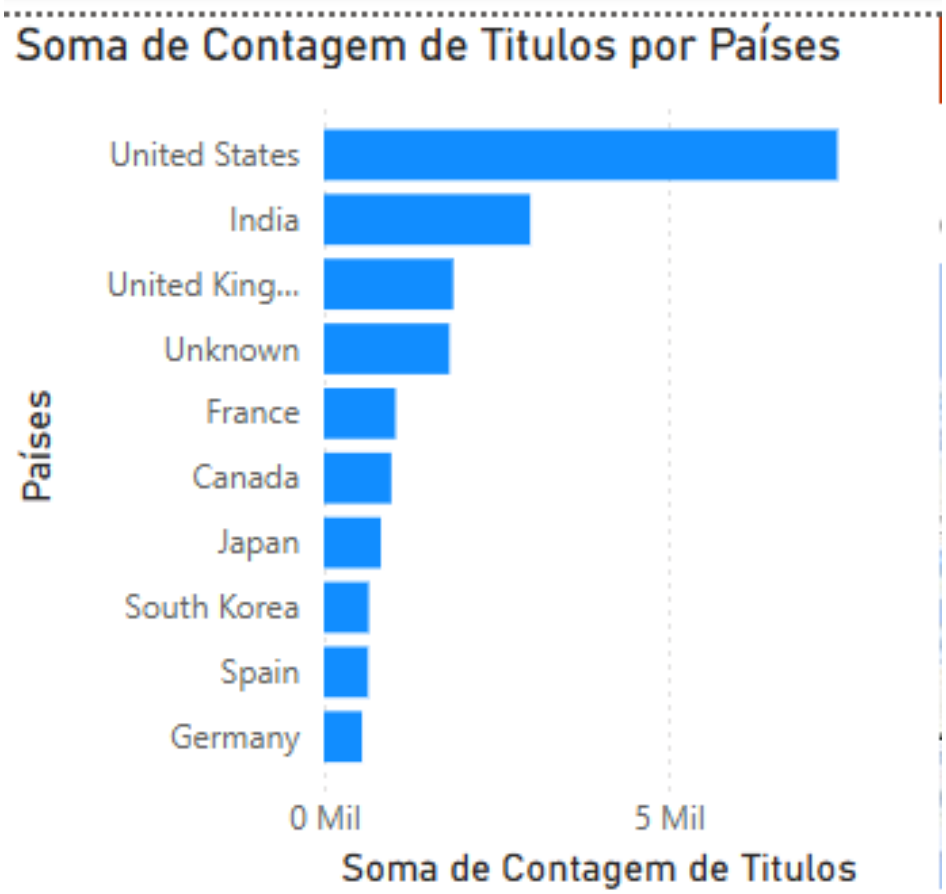

Figura 2. Top 10 países produtores de conteúdo no catálogo da Netflix.

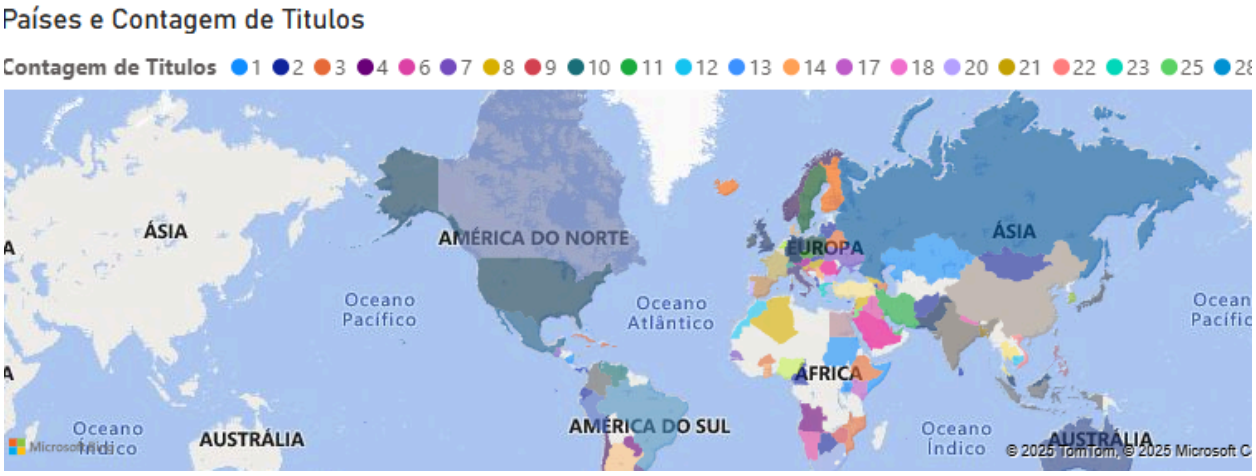

Figura 3. Mapa coroplético da distribuição de produção de conteúdo global.

Países	Action & Adventure	Anime Features	Anime Series	British TV Shows	Children & Family Movies	Classic & Cult TV	Classic Movies	Comedies	Crime TV Shows
Afghanistan									
Albania									
Algeria							1,00		
Angola	2,00								
Argentina	4,00				3,00	1,00	1,00	17,00	8,00
Armenia									
Australia	14,00		1,00	5,00	19,00		5,00	15,00	7,00
Austria									1,00
Azerbaijan									
Bahamas	1,00								
Bangladesh								11,00	
Belarus				1,00					
Belgium	14,00				10,00		1,00	23,00	8,00
Bermuda									
Botswana									
Brazil	5,00				9,00			20,00	6,00
Bulgaria	5,00				1,00				
Total	**1.305,00**	**130,00**	**192,00**	**308,00**	**1.067,00**	**32,00**	**166,00**	**2.210,00**	**553,00**

Figura 4. Matriz de especialização de gênero por país.

RQ3: Existem padrões sazonais na adição de títulos?

- **Afirmação:** A Netflix segue uma estratégia de lançamento de conteúdo sazonal e deliberada, concentrando seus lançamentos mais importantes na segunda metade do ano.
- **Evidência:** A **Figura 5**, um heatmap (matriz de mês vs. ano), revela um padrão claro. Os meses da primeira metade do ano (Janeiro-Maio) mostram contagens de títulos mais baixas. A partir de Julho, os números aumentam consistentemente, atingindo picos em Outubro, Novembro e Dezembro.
- **Comentário:** Isso faz todo o sentido comercial. A Netflix lança seus títulos mais fortes alinhados com os feriados de fim de ano (Ação de Graças, Natal) e o início do inverno no hemisfério norte, quando as pessoas passam mais tempo em casa.

Ano de adição	1,00	2,00	3,00	4,00	5,00	6,00
2.008,00	3,00	1,00				
2.009,00					2,00	
2.010,00						
2.011,00					6,00	
2.012,00		1,00				
2.013,00			6,00			
2.014,00	7,00	8,00		3,00		1,00
2.015,00	1,00	22,00	6,00	16,00	12,00	16,00
2.016,00	131,00	49,00	44,00	65,00	42,00	47,00
2.017,00	210,00	250,00	376,00	244,00	323,00	301,00
2.018,00	402,00	306,00	475,00	354,00	274,00	224,00
2.019,00	468,00	413,00	520,00	478,00	359,00	474,00
2.020,00	673,00	352,00	364,00	530,00	476,00	502,00
2.021,00	361,00	295,00	307,00	589,00	356,00	548,00
Total	**2.256,00**	**1.697,00**	**2.098,00**	**2.279,00**	**1.850,00**	**2.113,00**

Figura 5. Heatmap de sazonalidade de adições de conteúdo, 2008-2021.

RQ4: Qual a distribuição de duração por classificação e tipo?

- **Afirmação:** O formato do conteúdo é altamente otimizado: os filmes são adaptados para seus públicos e a grande maioria das séries são projetadas como "experiências de 1 temporada".
- **Evidência:**
  1. A **Figura 6**, um Box Plot, mostra que filmes para públicos maduros (ex: "NC-17", "R") têm uma mediana de duração maior (cerca de 110-120 min), enquanto filmes para família ("PG", "G") são visivelmente mais curtos (cerca de 90-100 min).
  2. A **Figura 7**, um Treemap, é o *insight* mais contundente. O retângulo para "1.00" (1 Temporada) domina o gráfico, ocupando a maior parte do espaço visual.
- **Comentário:** O Box Plot (Figura 6) mostra uma adaptação lógica do produto ao público. O Treemap (Figura 7) revela uma estratégia de negócios brutalmente eficiente: a Netflix opera em um modelo de "1 Temporada", seja porque lança muitas minisséries ou porque cancela agressivamente séries que não performam. O retângulo para "1 Temporada" domina o gráfico, representando cerca de 70% de todas as séries no catálogo. Isso sugere uma estratégia de "testar e iterar" em vez de cultivar séries longas.

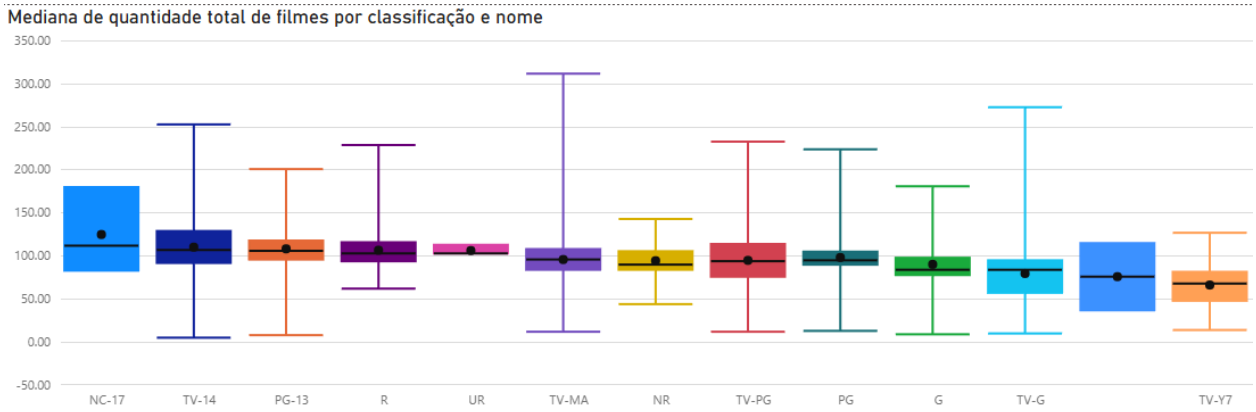

Figura 6. Distribuição da duração de filmes (em minutos) por classificação etária.

Figura 7. Distribuição de séries por número de temporadas.

(iv) Discussão e Conclusão

Nossa análise, iniciada como um exercício acadêmico, revelou-se um profundo mergulho na metodologia de engenharia de dados e na estratégia de negócios.

Desafios do ETL e Lições Metodológicas:
Nossa maior descoberta metodológica foi que o processo de "Obter Dados" não é trivial. A simples importação dos arquivos CSV para o Power BI falhou em capturar corretamente os tipos de dados, gerando insights completamente errôneos. A fase de limpeza de dados (ETL) foi, portanto, a etapa mais crítica do projeto. O rigor no ETL não é opcional. A depuração de tipos de dados (especialmente o "Problema de Localidade" Ponto vs. Vírgula), o tratamento de "lixo" textual em colunas numéricas (usando "Remover Erros") e a correção de scripts no Editor do Power Query foram cruciais para garantir a validade deste relatório.
Principal Conclusão Estratégica:
A análise do catálogo público da Netflix pinta o retrato de uma empresa ágil e disciplinada por dados. A migração estratégica para Séries de TV (RQ1) não é um movimento isolado; ela se combina com um modelo de produção focado em "1 Temporada" (RQ4) que funciona como um sistema de teste de mercado em escala global, minimizando riscos em novos conceitos. Ao mesmo tempo, a empresa equilibra sua hegemonia de produção nos EUA com uma estratégia "glocal" de especialização de gênero (RQ2), cultivando mercados específicos como Japão e Brasil. Finalmente, essa máquina de conteúdo opera com uma cadência sazonal deliberada (RQ3), alinhando seus maiores lançamentos com os períodos de maior consumo de mídia, maximizando o impacto de seu investimento.
Limitações:
Este estudo é limitado por seu dataset público (cobrindo até 2021). Não temos acesso a dados de orçamento, números de audiência reais ou taxas de conclusão. Nossas conclusões são, portanto, baseadas em correlações no catálogo, e não em causalidade de audiência. Um estudo futuro que pudesse cruzar estes dados de catálogo com dados de engajamento (minutos assistidos) permitiria validar quantitativamente o ROI de cada uma dessas estratégias.

Apêndice A: Detalhes Técnicos, Reprodutibilidade e Qualidade de Dados

Esta seção fornece os detalhes técnicos do pipeline de engenharia de dados, garantindo a transparência e a reprodutibilidade da análise.

Métricas de Qualidade de Dados (Visão Geral)

Métrica	Valor
Linhas no dataset original (netflix_titles_CLEANED.csv)	8.807
Linhas após explode() de country e listed_in (netflix_tratado_final.csv)	25.895
% de valores ausentes em country	9,3%

(original)	
% de valores ausentes em country (após tratamento com "Unknown")	0%
% de valores ausentes em date_added (original)	0,11%

Ambiente e Dependências

- **Python:** 3.13 (utilizado no desenvolvimento; recomenda-se testar com 3.11+).
- **Dependências:** pandas, jupyter, matplotlib, seaborn, plotly, python-docx, pycountry, kaleido. (Recomenda-se o congelamento de versões via pip freeze > requirements-freeze.txt para reprodução exata).
- **Configuração (Windows, cmd.exe):**
  python -m venv venv
  venv\Scripts\activate.bat
  pip install -r requirements.txt

Reprodução da Pipeline

Para reproduzir todos os artefatos (CSVs agregados, figuras PNG, HTMLs interativos) deste relatório, execute na raiz do projeto:

python src/preprocess.py
python src/standardize_countries.py
python src/prepare_powerbi.py
python src/generate_plotly.py
python src/analysis_duration.py
python src/generate_report_docx.py

Principais Decisões de Engenharia Adotadas

- **Backend Matplotlib:** O backend foi forçado para Agg (execução *headless*) em analysis_duration.py para eliminar dependências de GUI (como Tk/Tcl), garantindo a execução em servidores e ambientes de Integração Contínua (CI).
- **Tratamento PyCountry:** Foi utilizado getattr(...) e um mapa de fallback (COMMON_MAP) para a canonicalização de nomes de países, reduzindo *warnings* de analisadores estáticos (Pylance) e melhorando a cobertura de aliases não-padrão.
- **Separação de Visuais:** A pipeline gera visuais interativos (Plotly HTML) para exploração e imagens estáticas (Matplotlib PNG) otimizadas para inclusão em relatórios estáticos (.docx, .pdf).

Testes e Validação

- **Testes Automatizados:** O comando pytest -q foi executado com sucesso (resultado: 7 testes passaram). Os testes cobrem funções críticas, como o *parsing* de duration (ex: "90 min", "2 Seasons") e a canonicalização de nomes de países.
- **Verificação de Qualidade:** O script check_outputs.py valida a presença e o tamanho esperado dos artefatos finais, garantindo que a pipeline foi executada corretamente.

Observações Finais sobre Confiabilidade

- **Fonte de Dados:** O *dataset* público cobre até 2021; extrapolações posteriores exigem dados mais recentes.
- **Recomendações:** Para uma versão de produção, recomenda-se a revisão manual do arquivo country_aliases_mapping.csv (gerado pelo standardize_countries.py) para corrigir ambiguidades, e o uso de *checksums* (SHA256) para os dados de entrada.