Projeto Final:

Pipeline de ETL Integrado:

Uma análise do perfil dos conteúdos da renomada Netflix

Integrantes ENIAC :

Alessandra Machado
Caroline Cruz
Clara Maria
Dayane Lurdes
Elaine Castro
Maria Elacide
Tandara Jesus
Eduarda Lima

WOMAKERS® CODE

Sobre o projeto:

Esta é a documentação consolidada do projeto final do Bootcamp de Business Intelligence da WoMakersCode, do Squad ENIAC.

Temos como objetivo final deste relatório, aplicar técnicas e ferramentas de análise de dados para transformar informações brutas de streaming da Netflix em conhecimento estratégico sobre o conteúdo produzido.

A iniciativa busca não apenas gerar valor a partir dos dados, mas também fortalecer as competências do grupo em Power BI, Python, modelagem de dados, storytelling analítico e construção de relatórios executivos.

Girls™ in Tech

Introdução e contexto da escolha do dataset:

A indústria de streaming cresce rapidamente no mundo, com cada vez mais marcas buscando um espaço no mercado. Essa competitividade exige análises estratégicas do catálogo para orientar decisões de mercado.

A Netflix, como líder global do setor, expande continuamente seu portfólio, buscando equilibrar diversidade, escala e alcance internacional. Este projeto analisa o perfil dos conteúdos adicionados à plataforma entre 2016 e 2021, avaliando tipo, gênero, país de origem e evolução anual.

A partir dos dados brutos, desenvolvemos um pipeline completo, da extração à visualização, para gerar insights sobre prioridades de conteúdo a serem produzidos e tendências de mercado.

Problema central - (O que o projeto se propõe a resolver?)

Qual é o perfil de conteúdo que a Netflix mais prioriza e como esse conteúdo se relaciona com o volume de títulos lançados anualmente, pensando em tendências de mercado e possibilidades de investimento?

Objetivo principal:

Analisar e quantificar o catálogo da Netflix (2016-2021) para determinar o conteúdo que recebe maior prioridade (Filmes vs. Séries, gêneros e países de produção).

O objetivo é fornecer insights claros sobre a estratégia de conteúdo da plataforma, validar se a tendência observada deverá ser seguida em futuras produções e identificar padrões de crescimento que justifiquem próximos investimentos

Objetivos específicos (entregáveis):

- Processamento de dados (SQL & Python): Transformar dados brutos do catálogo Netflix em um conjunto de dados estruturado e pronto para análise.
- Análise (SQL & Power BI): Mapear e visualizar o crescimento anual do volume de lançamentos (entre 2016 e 2021), separando a evolução de Filmes vs. Séries.
- Identificação de padrões (SQL & Power BI):
  - Identificar e rankear os gêneros predominantes no catálogo.
  - Mapear a distribuição de conteúdo por país de origem (priorizando produções originais).
- Entrega de Insights (Power BI): Dashboard interativo (Power BI) que relaciona o perfil do catálogo com a evolução anual e forneça a base para a resposta final sobre a estratégia de expansão e mercado da Netflix

Relevância do tema:

O tema é altamente relevante por motivos estratégicos, analíticos e de mercado, especialmente no contexto de plataformas de streaming como a Netflix. É importante entender o cenário de algo tão presente no mercado e na vida das pessoas, além de ser algo em constante expansão e que podemos contribuir com estes insights.

Organização do projeto:

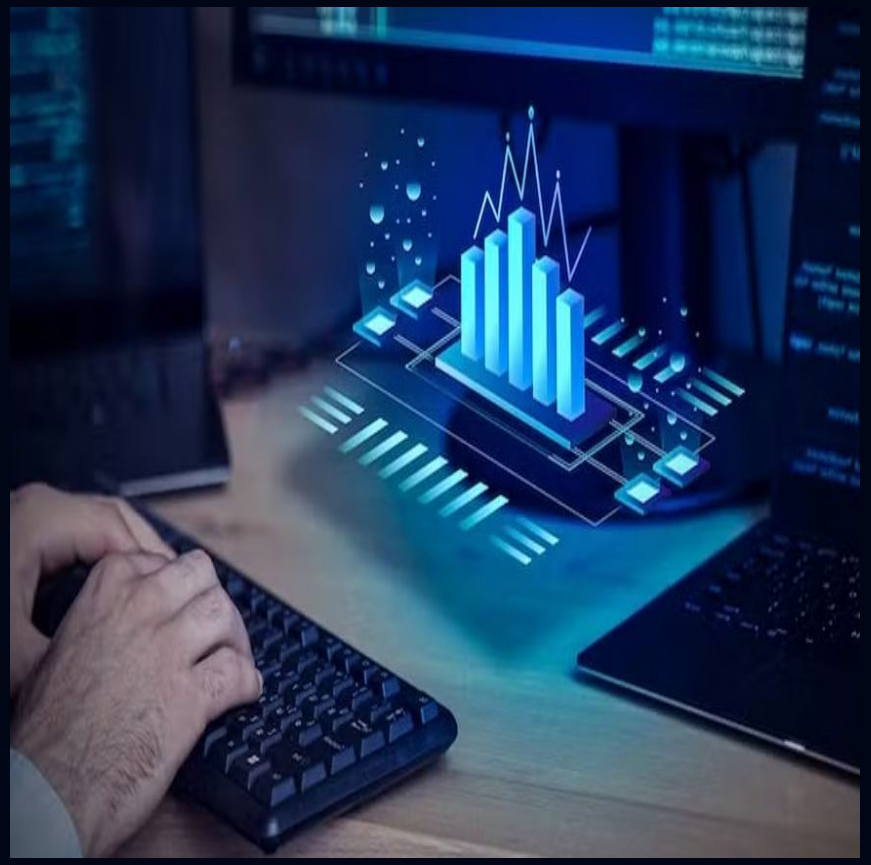
O trabalho foi desenvolvido de forma colaborativa com divisão de responsabilidades:

- **ETL (Google Colab):** limpeza, padronização e criação da camada gold
- **SQL/SQLite:** estruturação do banco e validação dos dados
- **Análise e Storytelling:** interpretação dos resultados e construção das hipóteses
- **Power BI:** dashboards, cálculos e visualizações
- **Documentação:** consolidação técnica e revisão final

Metodologia aplicada:

1) **Extração:** leitura e exploração inicial do dataset bruto
2) **Transformação:** tratamento de inconsistências, criação de colunas e normalização
3) **Carga:** armazenamento em SQLite para consumo no BI
4) **Visualização:** dashboards com métricas, tendências e indicadores

Planejamento do projeto



Base de dados

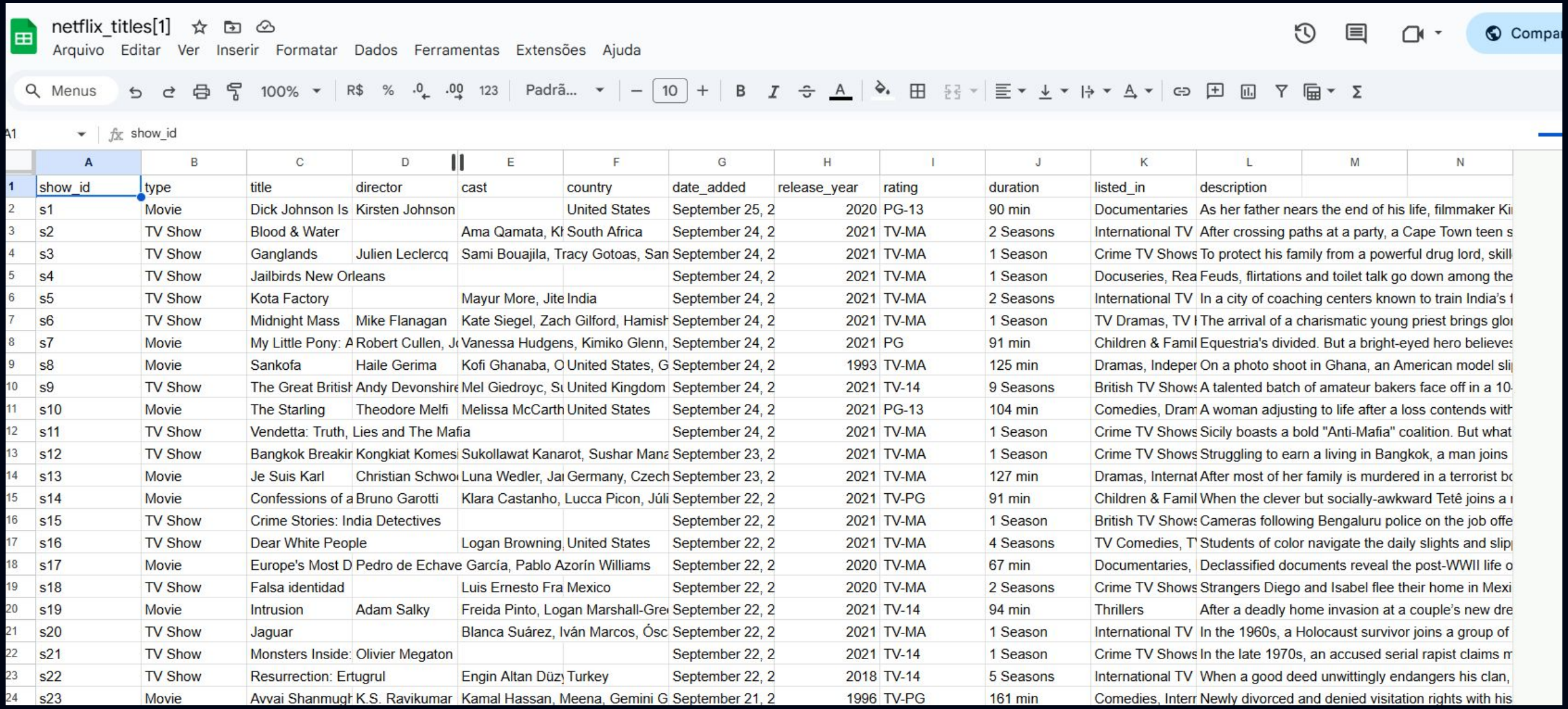
Formato dos dados: (CSV) dados tabulares simples.

Tamanho e complexidade:

- O dataset tem 8.807 linhas e 12 colunas principais (show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description.
- O tamanho do arquivo : 3,4 MB.

Escolha do dataset:

- Permite análises de perfil de conteúdo.
- Facilita avaliação de tendências ao longo do tempo.



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
2	s1	Movie	Dick Johnson Is	Kirsten Johnson		United States	September 25, 2	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmmaker Ki
3	s2	TV Show	Blood & Water		Ama Qamata, Kl	South Africa	September 24, 2	2021	TV-MA	2 Seasons	International TV	After crossing paths at a party, a Cape Town teen s
4	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, San		September 24, 2	2021	TV-MA	1 Season	Crime TV Shows	To protect his family from a powerful drug lord, skill
5	s4	TV Show	Jailbirds New Orleans				September 24, 2	2021	TV-MA	1 Season	Docuseries, Rea	Feuds, flirtations and toilet talk go down among the
6	s5	TV Show	Kota Factory		Mayur More, Jite	India	September 24, 2	2021	TV-MA	2 Seasons	International TV	In a city of coaching centers known to train India's f
7	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish		September 24, 2	2021	TV-MA	1 Season	TV Dramas, TV I	The arrival of a charismatic young priest brings glo
8	s7	Movie	My Little Pony: A	Robert Cullen, Jo	Vanessa Hudgens, Kimiko Glenn,		September 24, 2	2021	PG	91 min	Children & Famil	Equestria's divided. But a bright-eyed hero believes
9	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, O	United States, G	September 24, 2	1993	TV-MA	125 min	Dramas, Indeper	On a photo shoot in Ghana, an American model sli
10	s9	TV Show	The Great British	Andy Devonshire	Mel Giedroyc, Su	United Kingdom	September 24, 2	2021	TV-14	9 Seasons	British TV Shows	A talented batch of amateur bakers face off in a 10
11	s10	Movie	The Starling	Theodore Melfi	Melissa McCarth	United States	September 24, 2	2021	PG-13	104 min	Comedies, Dram	A woman adjusting to life after a loss contends with
12	s11	TV Show	Vendetta: Truth, Lies and The Mafia				September 24, 2	2021	TV-MA	1 Season	Crime TV Shows	Sicily boasts a bold "Anti-Mafia" coalition. But what
13	s12	TV Show	Bangkok Breakir	Kongkiat Komesi	Sukollawat Kanarot, Sushar Mana		September 23, 2	2021	TV-MA	1 Season	Crime TV Shows	Struggling to earn a living in Bangkok, a man joins
14	s13	Movie	Je Suis Karl	Christian Schwo	Luna Wedler, Jal	Germany, Czech	September 23, 2	2021	TV-MA	127 min	Dramas, Internat	After most of her family is murdered in a terrorist bo
15	s14	Movie	Confessions of a	Bruno Garotti	Klara Castanho, Lucca Picon, Júli		September 22, 2	2021	TV-PG	91 min	Children & Famil	When the clever but socially-awkward Tetê joins a r
16	s15	TV Show	Crime Stories: India Detectives				September 22, 2	2021	TV-MA	1 Season	British TV Shows	Cameras following Bengaluru police on the job offe
17	s16	TV Show	Dear White People		Logan Browning,	United States	September 22, 2	2021	TV-MA	4 Seasons	TV Comedies, T	Students of color navigate the daily slights and slip
18	s17	Movie	Europe's Most D	Pedro de Echave	García, Pablo Azorín Williams		September 22, 2	2020	TV-MA	67 min	Documentaries,	Declassified documents reveal the post-WWII life o
19	s18	TV Show	Falsa identidad		Luis Ernesto Fra	Mexico	September 22, 2	2020	TV-MA	2 Seasons	Crime TV Shows	Strangers Diego and Isabel flee their home in Mexi
20	s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall-Gre		September 22, 2	2021	TV-14	94 min	Thrillers	After a deadly home invasion at a couple's new dre
21	s20	TV Show	Jaguar		Blanca Suárez, Iván Marcos, Ósc		September 22, 2	2021	TV-MA	1 Season	International TV	In the 1960s, a Holocaust survivor joins a group of
22	s21	TV Show	Monsters Inside:	Olivier Megaton			September 22, 2	2021	TV-14	1 Season	Crime TV Shows	In the late 1970s, an accused serial rapist claims m
23	s22	TV Show	Resurrection: Ertugrul		Engin Altan Düz	Turkey	September 22, 2	2018	TV-14	5 Seasons	International TV	When a good deed unwittingly endangers his clan,
24	s23	Movie	Avvai Shanmugh	K.S. Ravikumar	Kamal Hassan, Meena, Gemini G		September 21, 2	1996	TV-PG	161 min	Comedies, Interr	Newly divorced and denied visitation rights with his

Planejamento do projeto

Fonte de dados

Nome do dataset: **Netflix Movies and TV Shows** (também chamado “netflix-shows”) por *shivamb* no Kaggle. Baselight

Questionamentos que permeiam a análise:

- Qual é a distribuição de **filmes vs séries** no Netflix ao longo dos anos?
- Quais são os gêneros mais frequentes e como eles variam por país?
- Quais gêneros são mais adicionados por usuários?
- Como podemos analisar a relevância de títulos por país e ano adicionado?

Desenvolvimento do pipeline

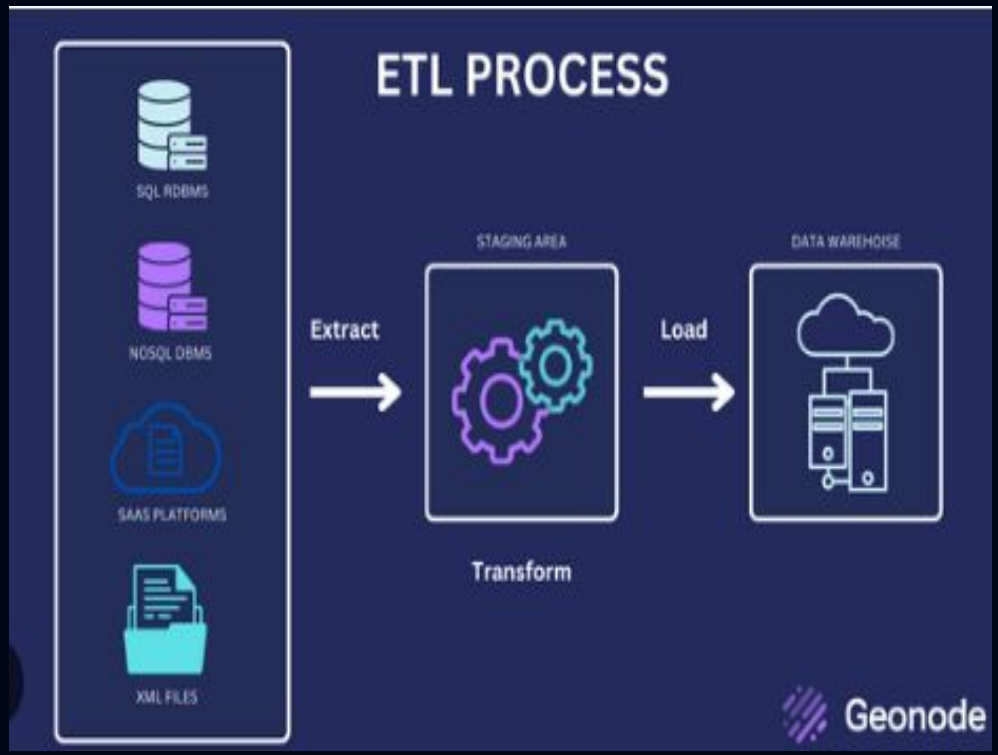
Fonte: *Kaggle* — “Netflix Movies and TV Shows” por **shivamb**:

Transformações necessárias: Limpeza, criação de tabelas, métricas, (camada gold), como o Google Collab.

Destino: banco de dados Sqlite e (Power BI)

Visualização:

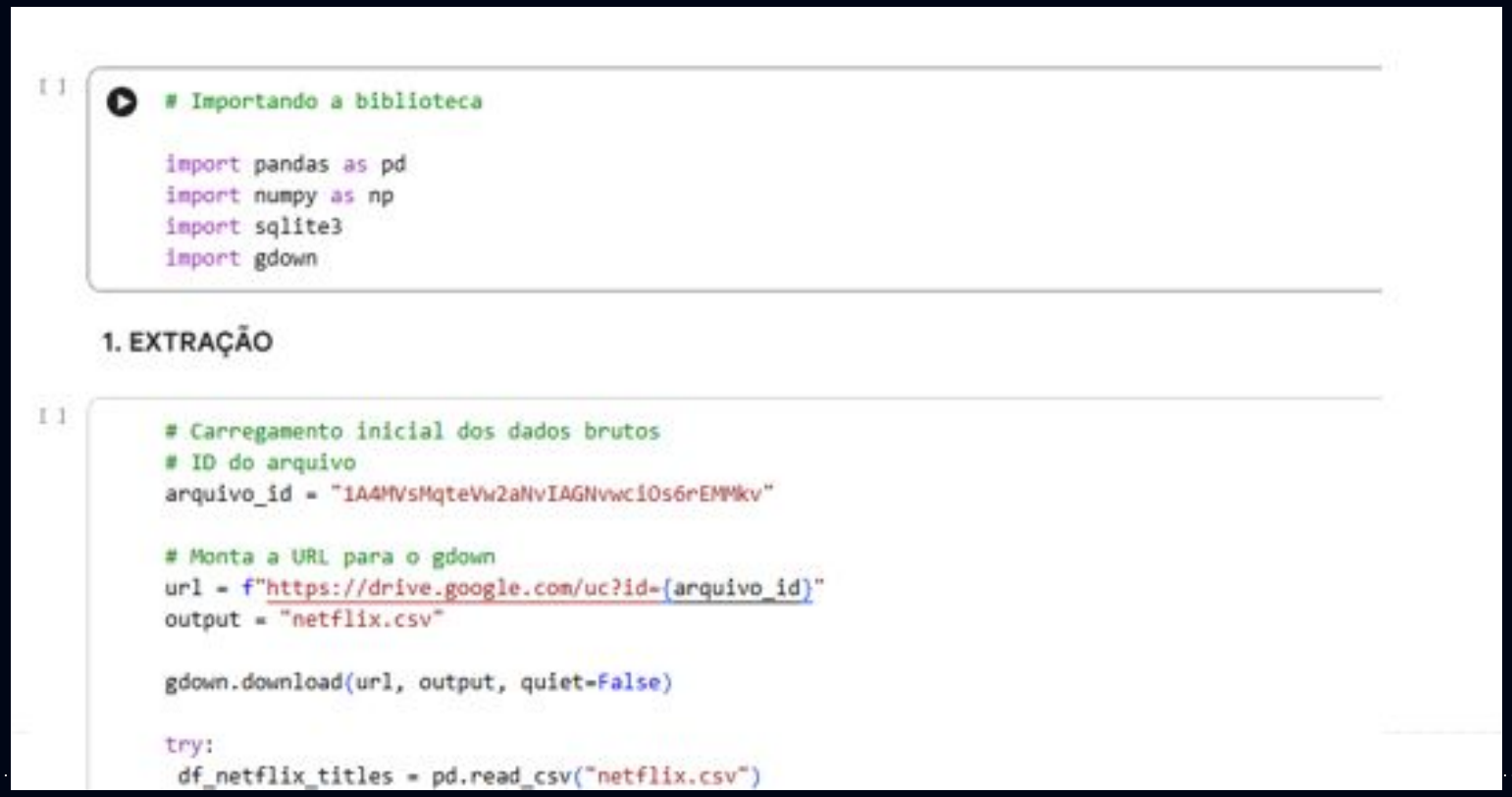
- Relatório, dashboard (Power BI)
- Gráfico de linha: lançamentos por Gênero.
- Gráfico de barras: distribuição de País x Ano.
- Cards: Total de títulos / Filmes x Series / Títulos anual.



Desenvolvimento do pipeline no Google Collab

Primeira etapa: EXTRAÇÃO

Você pode acompanhar a construção nesse link: https://colab.research.google.com/drive/1oWVQSI7Dt01RaCvzovo8kzl03x-4vdqs?usp=sharing - Fazer a leitura de arquivos locais (**CSV - Comma – Separated Vakeus**)



```
# Importando a biblioteca

import pandas as pd
import numpy as np
import sqlite3
import gdown
```

1. EXTRAÇÃO

```
# Carregamento inicial dos dados brutos
# ID do arquivo
arquivo_id = "1A4MVsMqteVw2aNvIAGNvwciOs6rEMMkv"

# Monta a URL para o gdown
url = f"https://drive.google.com/uc?id={arquivo_id}"
output = "netflix.csv"

gdown.download(url, output, quiet=False)

try:
 df_netflix_titles = pd.read_csv("netflix.csv")
```

Desenvolvimento do pipeline

Primeira etapa: EXTRAÇÃO REALIZADA

- # Importando a biblioteca
- # Carregamento inicial dos dados brutos
- # ID do arquivo
- # Monta a URL para o gdown
- # Exibição de tipos de colunas
- # print("\Contagem de valores nulos por Coluna:")
- # print("Visualizando linhas com dados faltantes:")
- # print("\Estatísticas")

Desenvolvimento do pipeline

Segunda etapa: TRANSFORMAÇÃO (Google Colab)

```
# Criada uma cópia para a camada silver

df_silver = df_netflix_titles.copy()
```

```
# Limpeza Inicial

# Removidas duplicatas
df_silver.drop_duplicates(inplace=True)
```

```
# Para validar se há duplicidades
numero_de_duplicatas_restantes = df_silver.duplicated().sum()
if numero_de_duplicatas_restantes == 0:
    print("✅ Sucesso! Todas as duplicatas foram removidas.")
else:
    print(f"⚠️ Atenção! Ainda existem {numero_de_duplicatas_restantes} duplicatas.")
```

```
✅ Sucesso! Todas as duplicatas foram removidas.
```

```
# Tratamento de nulos: Preencher com 'Desconhecido'
dicionario = {
```

Desenvolvimento do pipeline

Segunda etapa: TRANSFORMAÇÃO REALIZADAS

```
# Criada uma cópia para a camada silver
# Limpeza Inicial
 -Removidas duplicatas)
# Para validar se há duplicidades
# Tratamento de nulos: Preencher com 'Desconhecido'
# Para validar se há nulos
   -Lista das colunas que foram tratadas)
   -Série com a contagem de nulos nas colunas tratadas
# Removidos espaços em branco no início e fim de todas as colun
de texto
# Validar espaços em branco nos textos
  -Identificar as colunas de texto (object)
  -Checar se há valores que começam ou terminam com espaço

 #Padronizados nomes para Primeira Maiúscula

 colunas_titulos = ['title', 'director', 'cast']

  -Checa se o valor original é diferente do valor após a re-
aplicação do .str.title()

   -Se forem diferentes, o valor original (após a sua execução)
estava no formato correto.
# Dicionário de unificação dos países
   -Aplicação da unificação

# Dicionário de Classificação Etária para padrão Brasil
# Livre, 10+, 12+, 14+, 16+, 18+
# Verificar se a nova coluna 'rating_br'tem 6 classificações
```

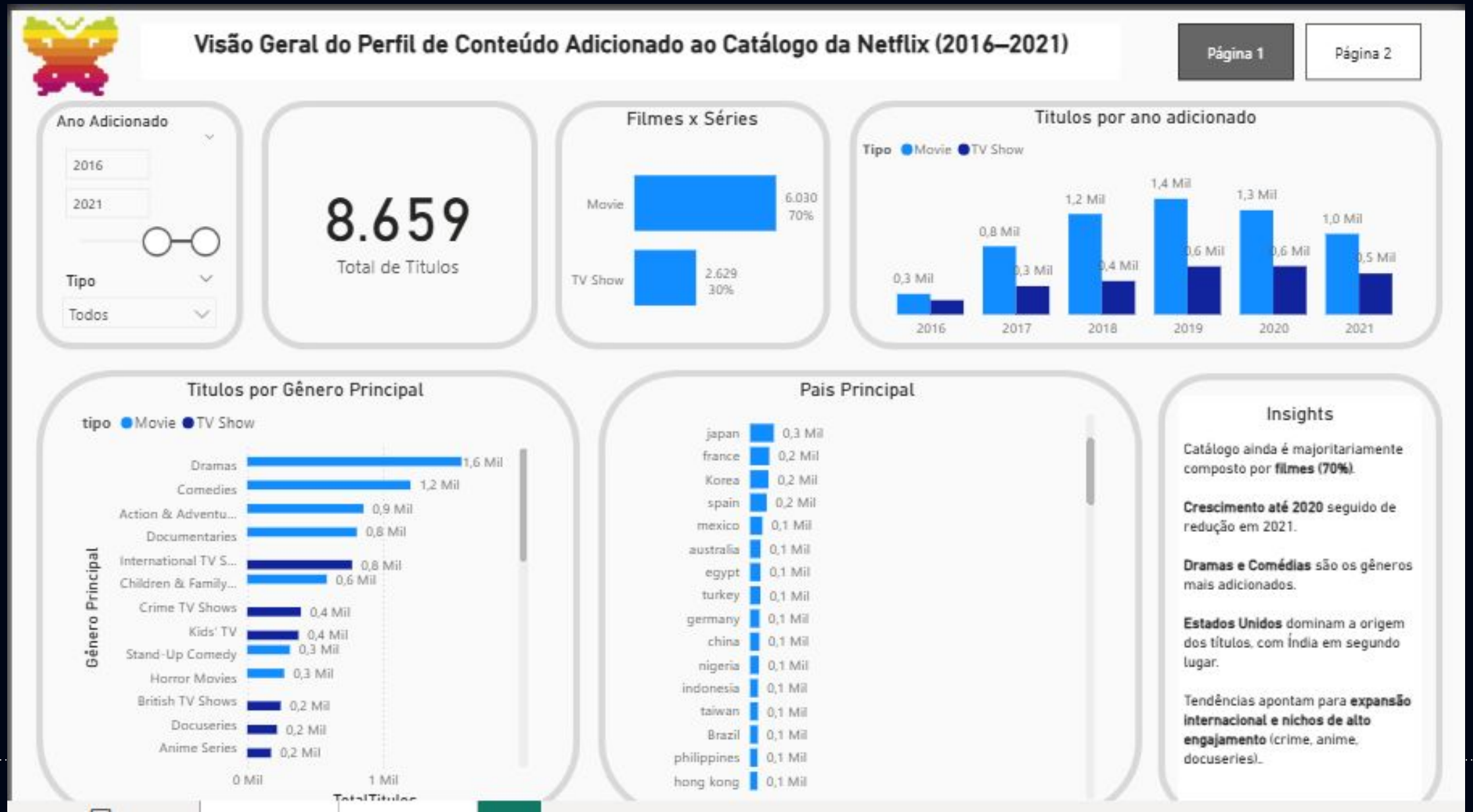
```
# Ajuste datas e tipos
# Convertidas date_added para datetime
# Checa o tipo de dado da coluna 'date_added'
# Criadas colunas de Ano e Mês de adição
# Calculado intervalo entre o lançamento e a adição na plataform
# Colunas para conferir (-Seleciona 5 linhas aleatórias
# Duração - Separar valor numérico da unidade (min vs seasons)
# Cria uma coluna de resumo da duração
# Converter para Inteiro INT
# Amostra visual (-Seleciona 5 linhas aleatórias para verificaçã
# Novas Colunas (- Gênero principal (primeiro da lista)
   - País principal (primeiro da lista
# Tabela dimensão para 'genero_principal'
  -Criação do ID único (chave primária)
  -Começa o ID em 1 para ser mais amigável

  -Chave Estrangeira

  -Adiciona a chave estrangeira na tabela principal (df_silver)

# Tabela de dimensão para 'pais_principal'

  -ID único (chave primária)

  -Começa o ID em 1

  -Chave Estrangeira
  -Adiciona a chave estrangeira na tabela principal (df_silver)
```

Desenvolvimento do pipeline

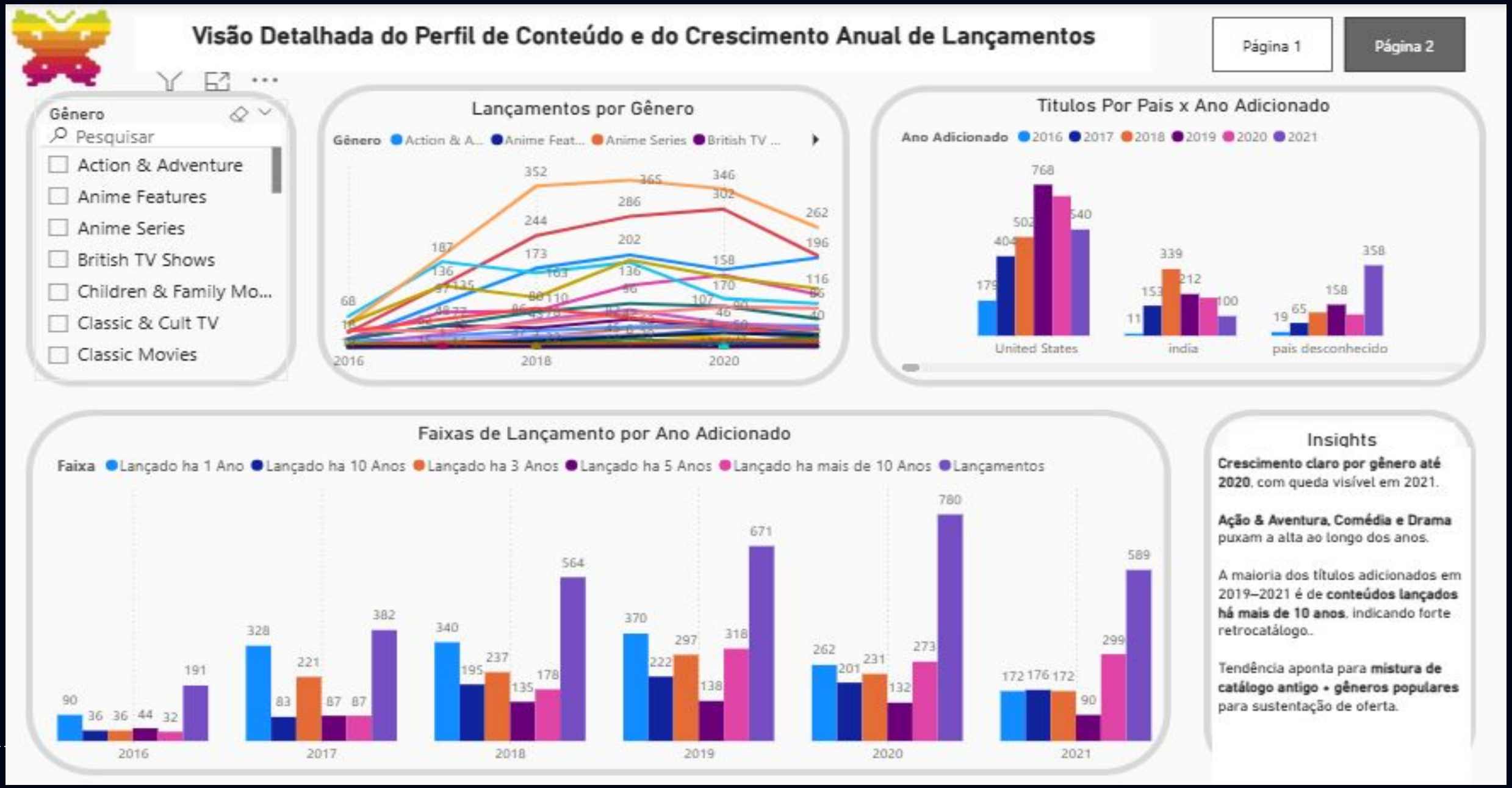
Terceira etapa: CARREGAMENTO

- Salvar os dados no banco de dados **banco SQLite** e armazenamento no banco de dados (DBeaver)
- Arquivo final tratado (camada gold) deve estar em um formato para ser consumido na **ferramenta de BI**

Dashboard desenvolvido
(Netflix_Dashboard_BI.pbix no GitHub)



Dashboard desenvolvido



Documentação - GitHub

Disponível neste link:
https://github.com/alessandramdsz/pipeline_ETL_projeto_final_ENIAC/blob/main/README.md

Documentamos todas as etapas primordiais para o desenvolvimento do projeto no GitHub.

As etapas para execução e visualização dos dados são:

1. Clonar o repositório
2. Preparar o ambiente
3. Instalar dependências
4. Executar o Notebook de ETL
5. Executar o banco de dados
6. Abrir o Dashboard no Power BI
7. Reproduzir o pipeline completo

Documentação - GitHub - Se preparando para executar

Para execução do projeto, deve-se confirmar se possui os seguintes programas/ferramentas instalados:

- Python 3.8
- Jupyter Notebook ou Google Colab
- SQLite3
- DBeaver (opcional)
- Power BI Desktop

Além de confirmar se está com as dependências instaladas (pip instal):

- Pandas (ler, limpar e tratar o *datasource* do projeto)
- Numpy (Criação de arrays)
- Matplotlib (visualizando e entendendo os dados)os dados
- Sqlite3 (Criar o banco, preencher com os dados tratados e fazer as consultas com o pipeline)

Documentação - GitHub - Executando o projeto

Recomenda-se a execução por meio do Google Colab:

- Acesse o Colab
- Faça upload do notebook disponível neste repositório
- Faça upload do arquivo CSV na pasta/content
- Execute todas as células na ordem

[45]
✓ 0s

```python
# CONEXÃO COM BANCO DE DADOS LOCAL (SQLite)
conn = sqlite3.connect('projeto_netflix.db')

# Salvando as tabelas no banco de dados
try:
    df_gold_titulos.to_sql('dim_titulos', conn, if_exists='replace', index=False)
    df_gold_generos.to_sql('fato_generos', conn, if_exists='replace', index=False)
    df_gold_paises.to_sql('fato_paises', conn, if_exists='replace', index=False)
    df_dim_generos.to_sql('dim_generos', conn, if_exists='replace', index=False)
    df_dim_paises.to_sql('dim_paises', conn, if_exists='replace', index=False)
    df_gold_generos.to_sql('ponte_generos_todos', conn, if_exists='replace', index=False)
    df_gold_paises.to_sql('ponte_paises_todos', conn, if_exists='replace', index=False)

    print("Carga realizada com sucesso! Tabelas criadas")

except Exception as e:
    print(f'Erro na carga: {e}')
finally:
    conn.close()

print('\n--- Pipeline Finalizado ---')
print("O arquivo 'projeto_netflix.db' está pronto para ser conectado ao Power Bi")
```

```
Carga realizada com sucesso! Tabelas criadas

--- Pipeline Finalizado ---
O arquivo 'projeto_netflix.db' está pronto para ser conectado ao Power Bi
```

Análise e conclusão estratégica

A análise de dashboard mostra o **crescimento anual de lançamentos**, destacando a evolução no volume de novos conteúdos ao longo dos anos. Observa-se que há uma tendência de **expansão, estabilidade até 2020 seguido de uma redução em 2021,** facilitando a compreensão do ritmo de crescimento e da eficiência das estratégias adotadas.

- **Tipo de conteúdo:** A plataforma prioriza filmes, que representam aproximadamente 70% do catálogo analisado, demonstrando foco em volume e diversidade de títulos de longa-metragem.
- **Gêneros predominantes:** Drama, Comédia e Ação formam o nicho principal do catálogo. Esta prioridade em gêneros de apelo global garante a ampla aceitação e escalabilidade em diferentes mercados internacionais.
- Os Estados Unidos lideram amplamente a **origem** dos títulos, seguidos pela Índia. Isso indica a manutenção da base ocidental e o crescimento estratégico em hubs regionais para o desenvolvimento de conteúdo local.

- Há um investimento crescente em **nichos de alto engajamento**, como crime, anime e séries, para reter públicos específicos.
- O alto volume de lançamentos anuais é mantido por meio do forte uso de **retrocatálogo** (conteúdos lançados há mais de 10 anos, mas adicionados à plataforma entre 2019 e 2021), indicando que a **mistura de conteúdo antigo e novos títulos é fundamental para sustentar a plataforma.**

Conclusão:

Como esse conteúdo se relaciona com o volume de títulos lançados anualmente, pensando em tendências de mercado e possibilidade de investimento?

A estratégia de conteúdo da Netflix se baseia em um equilíbrio entre custo, volume e diversidade, o que é importante para a competitividade no mercado de *streaming*. A plataforma prioriza o filmes de apelo global, sustentado pelo forte uso de retrocatálogo e por uma expansão geográfica estratégica.

Interpretação estratégica:

- Foco em conteúdos de **apelo global** para escalar presença internacional. Prioridade em **filmes** são essenciais para escalar a presença internacional e a aceitação.
- **Retrocatálogo** aumenta oferta rápido e com menor custo.
- Conteúdo **regional** que complementa a expansão em mercados locais.

Implicações para investimento:

- Investir em **gêneros globais** (drama/comédia/ação) tende a ter maior retorno.
- A aquisição de direitos de distribuição de **retrocatálogo** representa uma oportunidade de baixo custo e alto volume, fundamental para manter a oferta e a retenção.
- O investimento em mercados regionais e nichos de alto engajamento (crime, anime) indica potencial de crescimento futuro e de retenção de audiência.

Em suma, a Netflix prioriza filmes de apelo global, reforçados por retrocatálogo e expansão geográfica, mantendo o volume alto com uma estratégia de equilíbrio que maximiza o alcance global e o retorno sobre o investimento.

Obrigada!

Squad ENIAC

Alessandra Machado
Caroline Cruz
Clara Maria
Dayane Lurdes
Elaine Castro
Maria Elacide
Tandara Jesus

WOMAKERS® CODE