

Caracterização do Dataset – Netflix Titles

- **O que é o dataset?**

O dataset “**Netflix Titles**” reúne informações sobre produções disponíveis na plataforma de streaming **Netflix**.

Ele contém dados sobre **filmes e séries** — incluindo título, tipo, ano de lançamento, classificação indicativa, diretores, elenco, países de origem e gêneros.

O conjunto é amplamente utilizado em projetos de **engenharia de software**, **análise de dados** e **aprendizado de máquina** para estudar padrões de consumo, tendências de produção e análise de conteúdo audiovisual.

- **De onde ele veio?**

O dataset foi obtido a partir do site oficial do **Tableau Public**, na seção de **amostras de dados (Sample Data)**:

➤ <https://public.tableau.com/app/learn/sample-data>

- **Fonte:** Tableau Public
- **Instituição:** Salesforce / Tableau Software
- **Finalidade:** Base de dados pública para fins **educacionais e de demonstração** em ferramentas de visualização e análise de dados.
- **Licença:** Uso livre para fins acadêmicos, de estudo e experimentação.
- **Coleta:** Dados inspirados em informações reais sobre o catálogo da Netflix, mas adaptados para uso didático e ilustrativo.

- **O que ele contém?**

O arquivo contém **5 planilhas** que se relacionam entre si por meio da chave `show_id`, representando diferentes aspectos de cada título.

Planilha	Linhas	Colunas	Descrição
netflix_titles	6.236	9	Informações principais dos títulos (nome, tipo, ano, duração etc.)
netflix_titles_directors	4.852	2	Nome dos diretores associados a cada produção
netflix_titles_countries	7.179	2	País(es) de origem de cada produção
netflix_titles_cast	44.311	2	Lista de atores e atrizes de cada título
netflix_titles_category	13.670	2	Gêneros e categorias de cada título

Número total de instâncias: aproximadamente **76 mil registros** somando todas as planilhas.

Número total de atributos (colunas): 17, distribuídos entre as tabelas.

Tipos de dados principais: texto (string), numérico (int), data (datetime).

- **Qual é o objetivo do dataset?**

O objetivo do dataset é **analisar o catálogo da Netflix**, permitindo:

- Explorar **tendências de produção** (por tipo, país ou gênero).
- Avaliar **padrões de lançamento ao longo dos anos**.
- Medir a **diversidade de conteúdo** na plataforma.
- Apoiar **modelos de recomendação e análises de popularidade**.
- Servir como base para **experimentos de medição e mineração de dados** em projetos de Engenharia de Software.

- **Como os dados estão estruturados?**

O arquivo está no formato **XLSX (Excel)**, mas cada aba representa uma tabela relacional que pode ser facilmente convertida em **CSV** ou **SQL**.

A estrutura é centrada na tabela principal **netflix_titles**, que se relaciona com as demais através da coluna **show_id**.

Colunas principais:

- **show_id:** Identificador único de cada título

- title: Nome do filme ou série
- type: Tipo do conteúdo (“Movie” ou “TV Show”)
- release_year: Ano de lançamento
- duration_minutes: Duração em minutos (para filmes)
- duration_seasons: Número de temporadas (para séries)
- rating: Classificação indicativa
- date_added: Data de entrada no catálogo
- description: Breve resumo
- director, cast, country, listed_in: Informações complementares ligadas por tabelas auxiliares

- **Visão Geral do Dataset**

O arquivo contém **5 planilhas**, todas relacionadas a informações sobre produções disponíveis na Netflix.

O conjunto de dados está de forma **relacional**, com a coluna `show_id` funcionando como chave para relacionar as tabelas.

01. Planilha `netflix_titles`

- **Linhas:** 6.236
- **Colunas:** 9
- **Colunas:**
 - duration_minutes – Duração em minutos (para filmes).
 - duration_seasons – Número de temporadas (para séries).
 - type – Tipo do título (“Movie” ou “TV Show”).
 - title – Nome da produção.
 - date_added – Data de adição ao catálogo.
 - release_year – Ano de lançamento.
 - rating – Classificação indicativa.
 - description – Breve sinopse.

- show_id – Identificador único do título.

Função: É a tabela principal, contendo informações gerais sobre cada produção.

02. Planilha netflix_titles_directors

- **Linhas:** 4.852
- **Colunas:** 2
- **Colunas:**
 - director – Nome do diretor.
 - show_id – Identificador do título.

Função: Associa diretores aos títulos.

Pode haver títulos sem diretores registrados ou múltiplos diretores para o mesmo show_id.

03. Planilha netflix_titles_countries

- **Linhas:** 7.179
- **Colunas:** 2
- **Colunas:**
 - country – País de produção.
 - show_id – Identificador do título.

Função: Relaciona os países onde as produções foram feitas.

Um título pode ter múltiplos países associados.

04. Planilha netflix_titles_cast

- **Linhas:** 44.311
- **Colunas:** 2
- **Colunas:**
 - cast – Nome de um ator/atriz.
 - show_id – Identificador do título.

Função: Lista o elenco das produções.

Há várias linhas para o mesmo título, indicando múltiplos atores.

05. Planilha `netflix_titles_category`

- **Linhas:** 13.670
- **Colunas:** 2
- **Colunas:**
 - `listed_in` – Categoria ou gênero.
 - `show_id` – Identificador do título.

Função: Define as categorias ou gêneros (por exemplo: “Drama”, “Comedy”, “Documentary”).

Resumo Relacional

Planilha	Chave Primária	Chave Estrangeira	Relação com <code>netflix_titles</code>
<code>netflix_titles</code>	<code>show_id</code>	—	—
<code>netflix_titles_directors</code>	—	<code>show_id</code>	1:N
<code>netflix_titles_countries</code>	—	<code>show_id</code>	1:N
<code>netflix_titles_cast</code>	—	<code>show_id</code>	1:N
<code>netflix_titles_category</code>	—	<code>show_id</code>	1:N