

# Projeto 2: Análise Exploratória de Dados Netflix

## Introdução

Este projeto tem como objetivo desenvolver habilidades em análise exploratória de dados (EDA) com Python, utilizando um dataset real da Netflix contendo informações sobre filmes e séries.

**Duração estimada:** 1 semana (1 hora por dia)

**Ferramentas:** Python, Pandas, Jupyter/Colab

**Nível:** Iniciante intermediário (com experiência prévia em SQL/Python)

---

## Dataset

**Nome:** Netflix Movies and TV Shows

**Descrição:** Catálogo contendo filmes e séries da Netflix com informações como tipo de conteúdo, título, diretor, país de origem, data de adição, ano de lançamento, classificação indicativa, duração/temporadas e gêneros.

### Colunas principais:

- type: Movie ou TV Show
- title: Nome do filme/série
- director: Diretor(es)
- country: País(es) de origem
- date\_added: Data de adição ao catálogo
- release\_year: Ano de lançamento
- rating: Classificação etária (TV-MA, PG, etc.)
- duration: Duração em minutos (filmes) ou número de temporadas (séries)
- listed\_in: Gêneros

### Links de acesso:

- Opção 1: <https://www.kaggle.com/datasets/shivamb/netflix-shows>
- Opção 2: <https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies>

**Arquivo:** netflix\_titles.csv

---

# Objetivos de Aprendizado

Ao final deste projeto, você será capaz de:

- Carregar e explorar dados com Pandas
  - Identificar e tratar dados faltantes
  - Transformar colunas e extrair informações
  - Realizar análises agregadas e comparativas
  - Interpretar resultados e gerar insights
  - Documentar conclusões de forma clara e profissional
- 

## Pontos a Trabalhar (10 Tarefas)

### 1. Visão Geral do Dataset

**Objetivo:** Entender a estrutura e composição básica dos dados.

**Tarefas:**

- Carregar o CSV com `pandas.read_csv()`
- Exibir as 5 primeiras linhas: `head()`
- Verificar tipos de dados: `info()`
- Obter estatísticas descritivas: `describe(include='all')`
- Anotar as colunas disponíveis e seus tipos

**Saída esperada:**

- Entendimento da estrutura do dataset
  - Identificação de colunas numéricas, categóricas e com datas
  - Número total de linhas e colunas
- 

### 2. Limpeza Básica e Qualidade dos Dados

**Objetivo:** Avaliar a qualidade dos dados e decidir sobre tratamento de valores ausentes.

**Tarefas:**

- Contar valores nulos por coluna: `isnull().sum()`
- Verificar percentual de nulos em cada coluna
- Identificar colunas com mais de 20% de nulos
- Decidir estratégia: remover linhas, manter ou substituir com valor padrão

**Perguntas a responder:**

- Quais colunas têm valores faltantes?
- Em quais colunas o percentual é aceitável?
- Qual será a estratégia de tratamento?

**Saída esperada:**

- Dataset limpo e pronto para análise
  - Documentação da decisão tomada para cada coluna
-

### 3. Filmes vs Séries

**Objetivo:** Entender a proporção de conteúdo disponível na Netflix.

**Tarefas:**

- Contar registros por tipo: `value_counts()` na coluna `type`
- Calcular proporção percentual de cada tipo
- Criar uma visualização simples (ou apenas apresentar em texto)

**Perguntas a responder:**

- Quantos filmes existem no catálogo?
- Quantas séries existem?
- Qual é a proporção (percentual) de cada tipo?

**Saída esperada:**

- Contagem absoluta e percentual de filmes e séries
  - Insights sobre o mix de conteúdo da Netflix
- 

### 4. Análise por Ano de Lançamento

**Objetivo:** Identificar tendências históricas de produção de conteúdo.

**Tarefas:**

- Criar série com número de títulos por `release_year`
- Identificar o período com maior quantidade de lançamentos
- Comparar eras (antes de 2000, 2000-2010, 2010-2020, após 2020)

**Perguntas a responder:**

- Em qual década há mais conteúdo no catálogo?
- Houve um aumento significativo em lançamentos recentes?
- Qual foi o ano com mais títulos adicionados?

**Saída esperada:**

- Série temporal de lançamentos por ano
  - Identificação de períodos-chave de crescimento
- 

### 5. Países com Mais Títulos

**Objetivo:** Descobrir a distribuição geográfica do conteúdo.

**Desafio:** A coluna `country` contém múltiplos países em uma única célula (ex: "United States, United Kingdom").

**Tarefas:**

- Separar a coluna `country` em linhas individuais (usar `.str.split()` e `.explode()`)
- Contar frequência de cada país
- Listar os top 10 países com mais títulos

- Calcular percentual do conteúdo

**Perguntas a responder:**

- Qual país tem mais conteúdo na Netflix?
- Qual é a distribuição entre os 5 principais países?
- Qual percentual do catálogo vem dos EUA?

**Saída esperada:**

- Top 10 de países com contagem e percentual
  - Insights sobre concentração geográfica do conteúdo
- 

## 6. Análise de Duração: Filmes e Séries

**Objetivo:** Compreender características de duração/temporadas do conteúdo.

**Desafio:** A coluna `duration` tem formatos diferentes (ex: "90 min" para filmes, "2 Seasons" para séries).

**Tarefas:**

- Separar filmes e séries em subconjuntos
- Para filmes: extrair número de minutos de `duration` usando `.str.extract()` ou `.str.replace()`
- Para séries: extrair número de temporadas
- Calcular média, mediana, mínimo e máximo de duração para filmes
- Calcular média, mediana, mínimo e máximo de temporadas para séries

**Perguntas a responder:**

- Qual é a duração média de um filme?
- Qual é o número médio de temporadas de uma série?
- Há séries com muitas temporadas?

**Saída esperada:**

- Estatísticas de duração por tipo de conteúdo
  - Distribuição de filmes por faixa de duração
- 

## 7. Classificação Indicativa (Rating)

**Objetivo:** Analisar a distribuição de classificações etárias.

**Tarefas:**

- Contar registros por rating: `value_counts()`
- Calcular percentual de cada classificação
- Identificar a classificação mais comum

**Perguntas a responder:**

- Qual classificação indicativa é mais frequente?
- A Netflix tem mais conteúdo para adultos ou para famílias?

- Qual é a distribuição entre TV-MA, PG, TV-14, etc.?

**Saída esperada:**

- Distribuição de classificações indicativas
  - Insights sobre público-alvo do catálogo
- 

## 8. Gêneros Mais Comuns

**Objetivo:** Identificar preferências de gêneros no catálogo.

**Desafio:** A coluna `listed_in` contém múltiplos gêneros por título (ex: "Drama, Thriller, Crime").

**Tarefas:**

- Separar a coluna `listed_in` em gêneros individuais (usar `.strsplit()` e `.explode()`)
- Contar frequência de cada gênero
- Listar os top 10 gêneros
- (Bônus) Separar análise por tipo de conteúdo (filmes vs séries)

**Perguntas a responder:**

- Qual é o gênero mais comum na Netflix?
- Qual é a diferença de gêneros entre filmes e séries?
- Quais são os 5 principais gêneros?

**Saída esperada:**

- Top 10 de gêneros com contagem
  - Comparação de gêneros por tipo (filme vs série)
- 

## 9. Evolução de Conteúdo Recente

**Objetivo:** Avaliar tendências de crescimento de conteúdo nos últimos anos.

**Tarefas:**

- Filtrar títulos lançados a partir de 2015
- Contar número de títulos por ano a partir de 2015
- Comparar com período anterior (antes de 2015)
- Calcular taxa de crescimento anual

**Perguntas a responder:**

- O catálogo cresceu nos últimos anos?
- Em qual ano houve maior adição de conteúdo?
- Há tendência de crescimento contínuo?

**Saída esperada:**

- Série temporal de conteúdo recente
  - Análise de tendências de crescimento
-

## 10. Insight Final Escrito

**Objetivo:** Sintetizar conclusões e comunicar descobertas.

**Tarefas:**

- Revisar todas as análises anteriores
- Sintetizar as 3 a 5 descobertas mais importantes
- Escrever parágrafo de 6 a 10 linhas resumindo conclusões

**Perguntas a responder e incluir:**

- A Netflix tem mais filmes ou séries?
- De quais países vem a maioria do conteúdo?
- Como é a distribuição de classificações etárias?
- Quais gêneros dominam o catálogo?
- Qual é a principal característica do catálogo?

**Saída esperada:**

- Parágrafo conciso e bem estruturado com principais insights
- Pronto para apresentar ou compartilhar

---

## Metodologia e Boas Práticas

Estrutura do Código

### 1. Importações

```
import pandas as pd
import numpy as np
```

### 2. Carregar dados

```
df = pd.read_csv('netflix_titles.csv')
```

### 3. Explorar estrutura

```
df.head()
df.info()
df.isnull().sum()
```

### 4. Análises específicas

... código de cada tarefa

## 5. Documentar conclusões

... insights escritos

### Dicas Python/Pandas

- **Contar valores:** `df['coluna'].value_counts()`
- **Percentual:** `df['coluna'].value_counts(normalize=True) * 100`
- **Filtrar:** `df[df['coluna'] == 'valor']`
- **Extrair números:** `df['coluna'].str.extract(r'(\d+)')`
- **Separar múltiplos valores:** `df['coluna'].str.split(',').explode()`
- **Estatísticas:** `df['coluna'].describe()`
- **Agrupar:** `df.groupby('coluna')['outra_coluna'].agg(['count', 'mean'])`

### Documentação

Ao completar cada tarefa:

1. Execute o código
2. Anote o resultado (número ou padrão encontrado)
3. Responda as perguntas propostas
4. Passe para a próxima tarefa

---

## Cronograma Sugerido

Dia	Tarefa
1	Tarefas 1 e 2: Visão geral e limpeza
2	Tarefas 3 e 4: Filmes vs séries e evolução temporal
3	Tarefas 5 e 6: Países e duração
4	Tarefas 7 e 8: Classificações e gêneros
5	Tarefa 9: Conteúdo recente
6	Tarefa 10: Síntese e insights finais
7	Revisão, ajustes e documentação final

---

## Próximos Passos

1. Baixe o arquivo `netflix_titles.csv` do link do Kaggle
2. Carregue no Google Colab ou Jupyter Notebook
3. Execute as tarefas 1 e 2
4. Envie o código e resultados para feedback
5. Prossiga com as demais tarefas conforme orientação

---

## Referências e Recursos

- [Pandas Documentation](#)
- [Kaggle Datasets](#)
- [Python Data Analysis Guide](#)

---

**Última atualização:** Janeiro de 2026

**Tutor:** Analista de Dados - IA

**Status:** Projeto em andamento