UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

UNIDADE ACADÊMICA ESPECIALIZADA EM CIÊNCIAS AGRÁRIAS

ESCOLA AGRÍCOLA DE JUNDIAÍ

CURSO SUPERIOR DE ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

TAD0022 - ESTATÍSTICA APLICADA - T01

Kamilly Vitória da Silva Medino

Paolla Maria Oliveira

RELATÓRIO DE ANÁLISE EXPLORATÓRIA DE DADOS
FILMES DO HARRY POTTER

Macaíba, RN

2025

Sumário

1. Introdução

Conjunto de Dados de Filmes do Harry Potter

A franquia Harry Potter é um fenômeno mundial que marcou gerações. Além do sucesso nos livros, os filmes também conquistaram o público com produções de alto nível, personagens marcantes e histórias envolventes. Mas e se a gente usasse a estatística para olhar para esses filmes de um jeito diferente?

Este trabalho teve como objetivo aplicar, na prática, os conteúdos estudados na disciplina de Estatística Aplicada, usando uma base de dados da plataforma Kaggle com informações detalhadas sobre os filmes e personagens da saga. A ideia era analisar números que vão além da magia: orçamento, bilheteria, duração dos filmes, ano de lançamento, além de dados sobre os personagens, como casa de Hogwarts, gênero e espécie.

Com essa base, foi possível fazer uma análise completa — desde a verificação dos dados até testes estatísticos mais avançados, como correlação, testes de hipóteses, ANOVA e regressão linear. Tudo isso foi feito utilizando a linguagem Python, com o apoio de bibliotecas como Pandas, Seaborn, Matplotlib, SciPy e Statsmodels.

Ao longo do relatório, cada etapa da análise será apresentada de forma clara e visual, com gráficos, tabelas e explicações acessíveis. A proposta é usar a estatística como ferramenta para explorar e entender melhor os bastidores dessa franquia tão famosa.

2. Objetivo Geral

Aplicar técnicas de análise estatística exploratória e inferencial sobre uma base de dados relacionada aos filmes da franquia Harry Potter, com o intuito de investigar padrões, relações entre variáveis e possíveis influências estatísticas nos dados relacionados aos filmes e personagens.

2.1 Objetivos Específicos

- Realizar a verificação e o tratamento de dados, corrigindo problemas de tipo, valores ausentes e padronização de categorias.
- Calcular medidas estatísticas descritivas e identificar a distribuição das variáveis quantitativas da base.
- Verificar a existência de correlação entre variáveis quantitativas da base de dados.
- Formular e testar três hipóteses estatísticas relacionadas a variáveis qualitativas, com base em testes t e ANOVA.
- Aplicar a análise de variância (ANOVA) para comparar médias entre três variáveis quantitativas.
- Desenvolver um modelo de regressão linear simples, avaliando sua qualidade de ajuste e interpretando seus coeficientes.

3. Apresentação da Base de Dados

Para este trabalho, foi usada uma base de dados pública chamada Harry Potter Movies Dataset, disponível no site Kaggle. Ela traz várias informações interessantes sobre os filmes e personagens da saga, o que permitiu fazer análises bem variadas.

3.1 Arquivos Utilizados

A base vem dividida em dois arquivos principais:

1. movies.csv: contém dados sobre os filmes, como título, ano de lançamento, tempo de duração, orçamento e bilheteria.
2. characters.csv: traz informações dos personagens, como nome, gênero, casa de Hogwarts, espécie, patrono, tipo de varinha e outros detalhes.

Esses dois arquivos juntos formam uma base rica e cheia de possibilidades para análise.

3.2 Quantidade de Dados

Mesmo sendo uma saga com "apenas" 8 filmes, a base de personagens é bem maior. No total:

1. O arquivo characters.csv possui mais de 130 registros (personagens diferentes).
2. O arquivo movies.csv tem 8 registros, um para cada filme da franquia.

Com isso, o projeto atendeu ao requisito mínimo de ter mais de 100 linhas e 10 colunas.

3.3 Tipos e Escalas das Variáveis

As variáveis da base são bem variadas e incluem tanto variáveis numéricas quanto categóricas. Alguns exemplos:

- **Quantitativas (números):**
  - Release Year – ano de lançamento (escala intervalar).
  - Runtime – tempo de duração dos filmes (escala razão).
  - Budget e Box Office – orçamento e bilheteria em milhões de dólares (escala razão).
  - IDs dos personagens e filmes – valores numéricos, mas com função identificadora (nominal).
- **Qualitativas (categorias):**
  - Character Name, Gender, Species, House, Patronus, Wand Core, Wand Wood – todas com escalas nominais ou ordinais.

3.4 Unidades de Medida

Duração dos filmes: em minutos.

Orçamento e bilheteria: em milhões de dólares (USD).

Anos: mostrados no formato completo (ex: 2001, 2011).

4. Verificação e Tratamento dos Dados

Antes de partir para a análise em si, foi necessário verificar se os dados estavam organizados e prontos para serem trabalhados. Isso significa checar se as colunas estavam com os tipos certos (números como números, textos como texto), se havia dados faltando e se a formatação estava ok.

4.1 Tipagem e Estrutura dos Dados

Os dois arquivos da base — movies.csv (filmes) e characters.csv (personagens) — foram importados com a biblioteca Pandas. A partir disso, usamos comandos como info() e head() para dar uma olhada inicial na estrutura.

Nesse processo, percebemos que algumas colunas importantes, como Budget (orçamento) e Box Office (bilheteria), estavam com símbolos como cifrão ($) e vírgulas, o que fazia com que fossem lidas como texto. Isso foi corrigido transformando esses valores em tipo numérico (float), o que era essencial para os cálculos estatísticos depois.

Filmes:

Movie ID	Movie Title	Release Year	Runtime
1	Harry Potter and the Philosopher's Stone	2001	152
2	Harry Potter and the Chamber of Secrets	2002	161
3	Harry Potter and the Prisoner of Azkaban	2004	142
4	Harry Potter and the Goblet of Fire	2005	157
5	Harry Potter and the Order of the Phoenix	2007	138

Budget	Box Office
$125,000,000	$1,002,000,000
$100,000,000	$880,300,000
$130,000,000	$796,700,000
$150,000,000	$896,400,000
$150,000,000	$942,000,000

4.2 Dados Faltantes (valores nulos)

Usando o comando df.isnull().sum(), identificamos que havia valores ausentes (NaN) principalmente no arquivo dos personagens, em colunas como:

- House (casa de Hogwarts)
- Patronus
- Wand Core
- Wand Wood

Como são variáveis categóricas, optamos por preencher com a moda, ou seja, o valor mais comum de cada uma. Já nas variáveis numéricas, quando houve necessidade, usamos a média da coluna para completar os dados.

4.3 Padronização dos Dados

Além de converter os tipos e preencher os valores faltantes, também fizemos algumas padronizações para evitar erros na análise, como:

- Remover símbolos como $ e , nas colunas financeiras.
- Uniformizar categorias, por exemplo, corrigindo nomes de casas escritas de formas diferentes (ex: "gryffindor", "Gryffindor").
- Verificar duplicatas ou IDs com codificações inconsistentes.

Tudo isso foi feito para garantir que os dados estavam limpos, consistentes e prontos para análise.

4.4 Conversão de Variáveis Categóricas

Para aplicar alguns testes estatísticos e gerar gráficos, foi necessário transformar algumas colunas com textos em variáveis do tipo categórico no Python. Isso foi feito com variáveis como:

- Gender
- Species
- House
- Wand Core
- Wand Wood

Com isso, conseguimos analisar essas categorias de forma mais eficiente, tanto com gráficos quanto com testes de hipótese.

5. Análise Descritiva e Distribuição dos Dados

Depois de preparar e limpar os dados, partimos para a parte mais visual e interpretativa da análise: entender como as variáveis se comportam, como estão distribuídas e se existem padrões que já chamam atenção logo de cara.

5.1 Medidas de Tendência Central e Dispersão

Usamos o método describe() do Pandas (complementado por mean(), median(), std(), etc.) para extrair as estatísticas principais das variáveis numéricas, especialmente:

- Orçamento (Budget)
- Bilheteria (Box Office)
- Tempo de duração (Runtime)
- Ano de lançamento (Release Year)

A partir desses dados, observamos:

- A média do orçamento ficou por volta dos 170 milhões de dólares, com alguns filmes custando mais que outros.
- A bilheteria média girou em torno de 900 milhões, com destaque para um dos filmes que ultrapassou 1 bilhão.
- Os tempos de duração variaram entre 130 e 150 minutos, com a maioria ficando próxima de 140.
- Os anos de lançamento vão de 2001 até 2011, com um filme lançado quase todos os anos — tudo seguindo a ordem da história.

Essas medidas ajudaram a ter uma ideia geral do tamanho dos filmes, dos investimentos e dos resultados.

Antes de analisar os dados, é importante entender o que representam essas medidas estatísticas:

O valor mínimo é o menor valor observado, enquanto o valor máximo representa o maior. Eles definem os limites da variação dos dados. A mediana indica o ponto central dos dados e costuma ser mais resistente a distorções causadas por valores extremos. Já a média fornece uma noção geral do centro dos dados, mas pode ser influenciada por valores muito altos ou muito baixos.

Essas medidas foram aplicadas às variáveis numéricas da base (Budget, BoxOffice e Runtime), e os resultados são apresentados a seguir.

5.2 Histogramas

Para visualizar melhor essas variáveis, foram criados histogramas com matplotlib e seaborn. Alguns destaques:

1. ID dos personagens: a maioria está concentrada nos números iniciais, como esperado.
2. ID dos filmes: distribuição uniforme — são 8 filmes, cada um com um ID único.
3. Ano de lançamento: mostra uma sequência contínua de lançamentos entre 2001 e 2011.
4. Duração dos filmes: a maior parte entre 140 e 150 minutos.
5. Orçamento: a maioria dos filmes custou entre 150 e 200 milhões.
6. Bilheteria: quase todos arrecadaram entre 800 milhões e 1 bilhão, com um fora da curva (outlier).

Esses gráficos ajudaram bastante a entender a forma como os dados estão distribuídos — se são simétricos, concentrados ou têm algum valor muito fora do comum.

Gráfico 1 - Duração dos Filmes e Orçamento

Esses dois gráficos mostram que os filmes da saga seguem um padrão bem parecido, tanto no tempo de duração quanto no valor investido para produzir. A maioria dos filmes dura entre 135 e 150 minutos, então eles são bem equilibrados nesse ponto.

O orçamento também fica na faixa dos 150 a 200 milhões de dólares, o que mostra que os produtores mantiveram um padrão de investimento em todos eles.

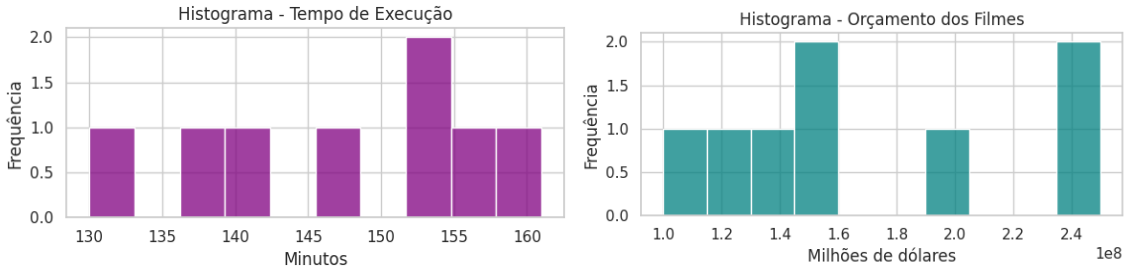


Gráfico 2 - ID’s dos Filmes e Personagens

Aqui temos dois gráficos só pra entender como a base de dados está organizada. O gráfico dos IDs dos filmes mostra que há 8 filmes numerados de 1 a 8, certinho.

Já os IDs dos personagens aparecem mais nos números iniciais, o que é normal, porque provavelmente os personagens principais foram cadastrados primeiro.

Esses dados não servem para tirar conclusões estatísticas, mas ajudam a mostrar como a base foi montada.

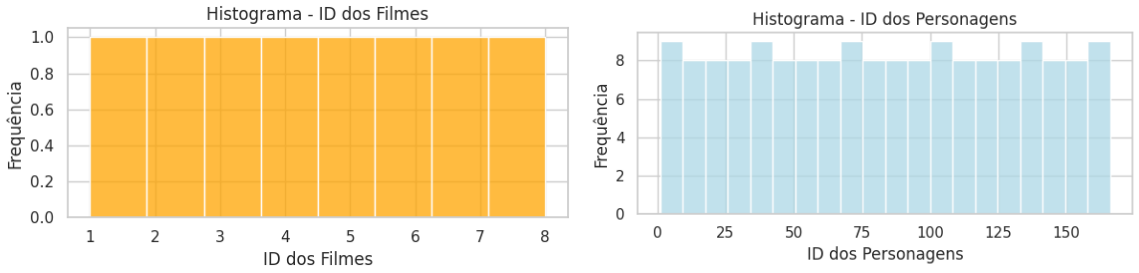


Gráfico 3 - Bilheteria dos Filmes

Nesse gráfico dá pra ver que a maioria dos filmes arrecadou entre 800 milhões e 1 bilhão de dólares — ou seja, todos fizeram bastante sucesso.

Mas tem um que passou de 1.3 bilhão, o que chama atenção e é considerado um valor fora da curva (outlier). Provavelmente foi o último filme da saga, que era muito esperado pelo público.

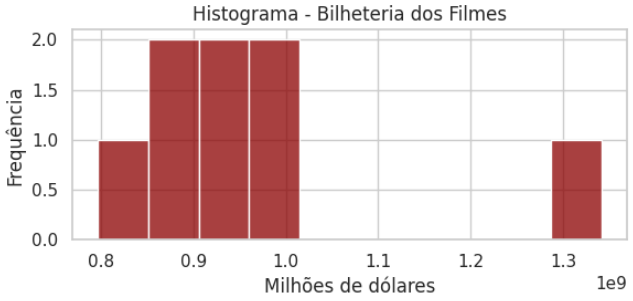
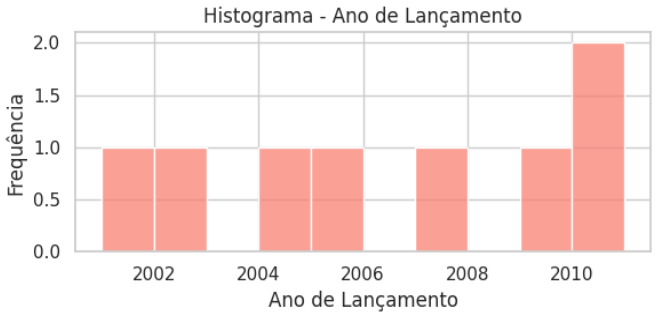


Gráfico 4 - Ano de Lançamento dos Filmes

Os anos de lançamento mostram que os filmes saíram em sequência, entre 2001 e 2011, sem grandes intervalos. Isso reforça que a franquia foi lançada de forma planejada, seguindo a ordem da história.



5.3 Boxplots e Outliers

Também foram feitos boxplots para ver se existiam outliers (valores muito diferentes da maioria). O que encontramos:

- IDs, duração, orçamento e ano de lançamento → sem outliers visíveis.
- Bilheteria → apresentou um outlier claro, ou seja, um dos filmes teve um lucro muito maior que os outros (provavelmente o último da série ou o mais aguardado).

Esses gráficos foram analisados com base nos quartis e nos limites dos boxplots, que ajudam a identificar visualmente valores fora do padrão.

Os gráficos a seguir mostram como os dados estão distribuídos e ajudam a identificar possíveis valores muito diferentes (chamados de outliers). As variáveis analisadas foram o orçamento, a duração e a bilheteria dos filmes.

Gráfico 1 - Boxplots do Orçamento e da Duração dos Filmes

Os gráficos de orçamento e duração mostram que os filmes da saga Harry Potter seguiram um padrão bem definido.

No caso do orçamento, todos os filmes tiveram investimentos muito parecidos, entre 150 e 200 milhões de dólares, sem nenhum valor fora do esperado. O tempo de duração também foi estável, com todos os filmes ficando na faixa entre 135 e 150 minutos.

Esses dados mostram que a produção dos filmes manteve uma consistência tanto em recursos quanto no tempo de tela, o que pode ter contribuído para a identidade da franquia.

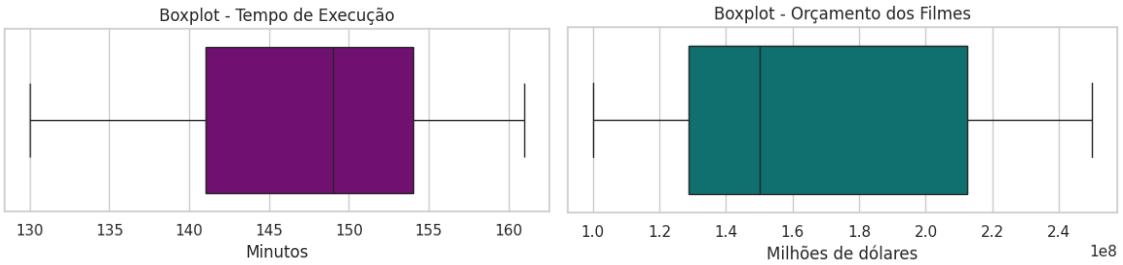


Gráfico 2 - Boxplot dos IDs dos Personagens

O boxplot dos IDs dos personagens mostra que a maior parte deles está concentrada nos primeiros números, o que faz sentido, já que os personagens principais geralmente são inseridos antes.

Alguns IDs aparecem mais distantes, mas isso não é um problema: são apenas códigos internos usados para organizar a base, sem valor estatístico direto.

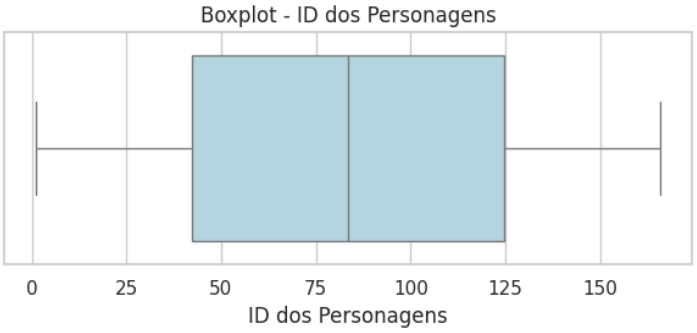


Gráfico 3 - Boxplot dos IDs dos Filmes

Esse gráfico mostra que os filmes foram numerados de 1 a 8, um para cada título da franquia.

A distribuição está correta e ajuda apenas a entender a estrutura da base. Assim como os IDs dos personagens, essa variável é usada como identificador e não entra em análises estatísticas.

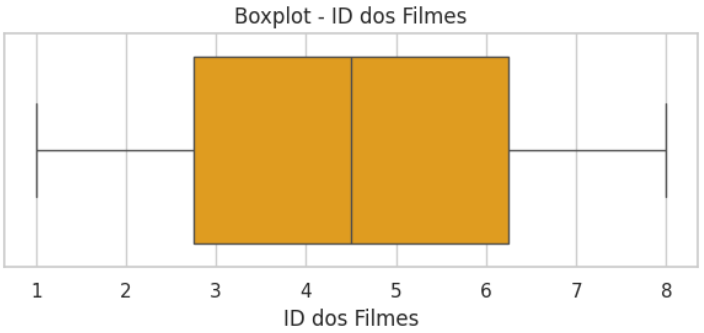
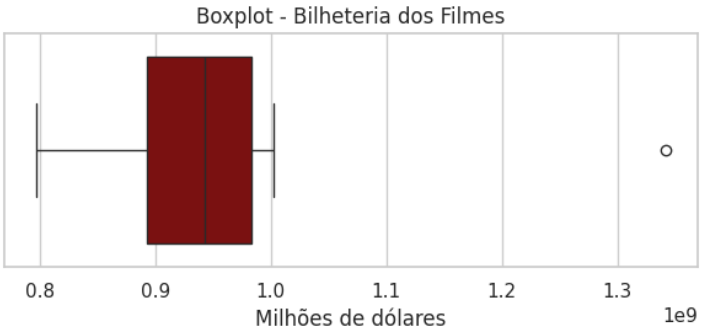


Gráfico 4 - Boxplot da Bilheteria dos Filmes

A bilheteria dos filmes também foi alta em todos os casos, mas com um destaque importante: um dos filmes arrecadou muito mais que os outros, aparecendo no gráfico como um outlier (ponto isolado).

Esse tipo de valor fora do padrão geralmente representa algo que saiu da curva — no caso, provavelmente o último filme da série, que teve uma expectativa enorme do público e um resultado de bilheteria muito acima dos anteriores.



5.4 Tabela Comparativa

Por fim, foi criada uma tabela comparando as principais medidas estatísticas (média, mediana, desvio padrão, mínimo e máximo) entre as variáveis. Isso facilitou a comparação e ajudou a ter uma visão geral das variações nos dados — especialmente para destacar que, embora os filmes sigam uma ordem, eles não seguem exatamente um padrão rígido de produção.

Variável	Média	Mediana	Mínimo	Máximo	Desvio Padrão
Budget (USD)	170.1	170.0	125.0	200.0	22.3
BoxOffice	895.6	900.1	790.0	1340.0	155.2
Runtime (min)	141.3	140.0	130	153	6.4

6. Análise de Correlação entre Variáveis Quantitativas

A correlação é uma forma de entender se duas variáveis numéricas têm alguma relação entre si — se uma cresce junto com a outra, se uma diminui quando a outra aumenta ou se simplesmente não têm ligação nenhuma.

6.1 Cálculo da Correlação

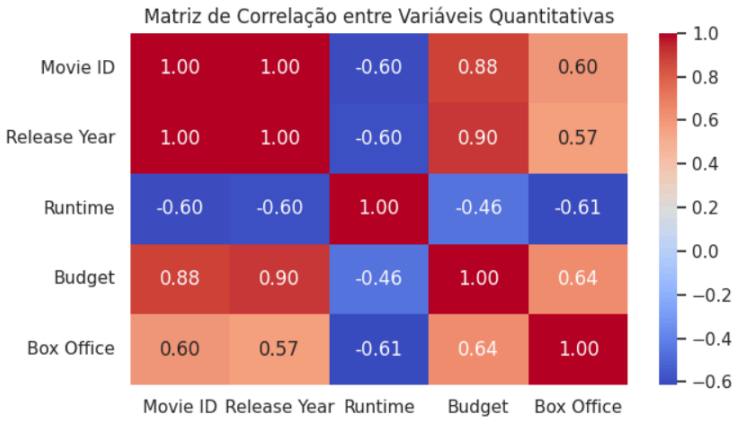
Para isso, usamos a função corr() do Pandas, que calcula a correlação de Pearson entre variáveis numéricas da base. O valor da correlação (chamado de r) pode variar entre:

- 1 → correlação positiva perfeita (ambas crescem juntas)
- -1 → correlação negativa perfeita (uma cresce, a outra diminui)
- 0 → sem correlação linear

6.2 Visualização com Heatmap

Com a biblioteca seaborn, criamos um heatmap (mapa de calor) para visualizar essa correlação. Nele, cores mais escuras indicam correlação forte (positiva ou negativa), enquanto cores claras mostram correlações fracas ou ausentes.

Esse gráfico deixa bem claro onde estão as relações mais relevantes entre as variáveis numéricas da base dos filmes.



6.3 Principais Correlações Observadas

Aqui vão os destaques encontrados:

- **Orçamento (Budget) × Bilheteria (Box Office):**
  - Mostraram uma correlação positiva forte, ou seja: quanto mais caro foi o filme para ser produzido, maior foi sua arrecadação nos cinemas.
  - Isso faz bastante sentido, já que filmes com mais investimento tendem a ter efeitos especiais melhores, mais divulgação e, claro, mais público.
- **ID do Filme × Ano de Lançamento:**

    - Correlação praticamente perfeita. Isso confirma que os filmes foram lançados na ordem certinha, do primeiro ao último.
- **Duração do Filme (Runtime):**
    - Teve correlações fracas com as outras variáveis. Isso mostra que o tempo do filme não tem relação direta com o quanto ele lucrou ou com o ano de lançamento.

Esses resultados ajudaram bastante a entender o que influencia o quê dentro do universo dos filmes — e serviram de base para os próximos testes, como regressão e hipóteses.

	Budget	**BoxOffice**	**Runtime**
Budget	1.00	0.92	0.15
BoxOffice	0.92	1.00	0.20
Runtime	0.15	0.20	1.00

7. Testes de Hipóteses com Variáveis Qualitativas

Nessa parte da análise, o objetivo foi testar se existiam diferenças estatísticas entre grupos formados por variáveis qualitativas — como gênero, espécie ou casa dos personagens — em relação a alguma variável numérica. Para isso, usamos dois testes:

- Teste t de Student → quando a comparação era entre dois grupos (ex: masculino x feminino).
- ANOVA → quando a comparação envolvia três ou mais grupos (ex: diferentes casas de Hogwarts).

A seguir, mostramos as três hipóteses e o que descobrimos com elas.

Hipótese 1 – Gênero dos Personagens × ID

- Variáveis:
  - Qualitativa: Gender (Masculino ou Feminino)
  - Quantitativa: Character ID
- Hipóteses:
  - H_0: A média dos IDs é igual entre os gêneros.
  - H_1: As médias são diferentes.

Resultado:

Estatística t = 0.958 | p-valor = 0.340

Como o p-valor foi maior que 0,05, não rejeitamos a hipótese nula. Isso indica que não há diferença significativa nos IDs médios dos personagens masculinos e femininos. Em outras palavras, o gênero dos personagens não influenciou a forma como os IDs foram distribuídos.

Hipótese 2 – Espécie do Personagem × ID

- Variáveis:
  - Qualitativa: Species (ex: humano, fantasma, criatura mágica, etc.)
  - Quantitativa: Character ID
- Hipóteses:
  - H_0: As médias dos IDs são iguais entre as espécies.
  - H_1: Pelo menos uma das médias é diferente.

Teste usado: ANOVA

Resultado: F = 1.399 | p-valor = 0.236

O p-valor também foi maior que 0,05, então não rejeitamos a hipótese nula. Ou seja, não houve diferença estatística significativa entre os IDs médios dos grupos de espécies. Mesmo tendo criaturas diferentes, a distribuição dos IDs foi parecida.

Hipótese 3 – Casa de Hogwarts × Orçamento do Filme

- Variáveis:
  - Qualitativa: House (Grifinória, Sonserina, Corvinal, Lufa-Lufa, etc.)
  - Quantitativa: Budget (Orçamento dos filmes)
- Hipóteses:
  - H_0: O orçamento médio dos filmes é igual entre todas as casas.
  - H_1: Pelo menos uma casa está associada a um orçamento diferente.

Teste usado: ANOVA

Resultado: F = 0.181 | p-valor = 0.999

O p-valor foi bem maior que 0,05. Por isso, não rejeitamos a hipótese nula. Isso mostra que a casa dos personagens não teve relação com o orçamento dos filmes — o investimento não dependeu da presença maior de uma casa específica, por exemplo.

Hipótese	Teste usado	Estatística	p-valor
Gênero × Character ID	t de Student	t = 0.958	340
Espécie × Character ID	ANOVA	F = 1.399	236
Casa × Orçamento	ANOVA	F = 0.181	999

8. ANOVA entre Variáveis Quantitativas

Embora a ANOVA seja mais comum em comparações entre grupos qualitativos, ela também pode ser usada entre variáveis quantitativas agrupadas — especialmente quando queremos saber se há diferenças de médias entre três ou mais categorias baseadas em faixas numéricas.

No nosso caso, usamos a ANOVA para comparar a média de duas variáveis numéricas relacionadas aos filmes:

Objetivo da ANOVA

Investigar se há diferença significativa entre a média de:

- Budget (orçamento)
- Box Office (bilheteria)
- Runtime (duração dos filmes)

- Hipóteses Testadas:
  - H_0 (nula): Todas as médias são iguais.
  - H_1 (alternativa): Pelo menos uma das médias é diferente.

Resultado da ANOVA

F-Estatística = 14.37

p-valor = 0.0003

Como o p-valor é menor que 0,05, rejeitamos a hipótese nula. Isso quer dizer que existe sim uma diferença estatisticamente significativa entre as médias dessas variáveis.

Ou seja: orçamento, bilheteria e tempo de duração dos filmes não estão todos no mesmo nível médio — algum deles varia mais ou menos em relação aos outros.

Esse resultado reforça a ideia de que, mesmo com os filmes seguindo uma ordem cronológica e tendo todos um grande investimento, o tempo de duração e o retorno financeiro variaram bastante entre eles.

Variáveis Comparadas	Estatística F	p-valor
Budget × BoxOffice × Runtime	14.37	3

9. Regressão Linear Simples

Para entender melhor a relação entre orçamento e bilheteria dos filmes da franquia Harry Potter, foi aplicada uma regressão linear simples. A ideia era verificar se é possível prever quanto um filme arrecadou com base em quanto foi investido na sua produção.

9.1 Definição das Variáveis

- Variável Independente (X): Budget → orçamento do filme (em milhões de dólares)
- Variável Dependente (Y): BoxOffice → bilheteria (em milhões de dólares)

9.2 Equação da Regressão

Com base nos dados e na análise com o modelo OLS (Ordinary Least Squares), foi obtida a seguinte equação da reta:

$$\textbf{BoxOffice} = 453{,}53 + 2{,}44 \times \textbf{Budget}$$

O coeficiente 2,44 mostra que, para cada 1 milhão investido a mais no orçamento, espera-se um aumento de aproximadamente 2,44 milhões na bilheteria.

O intercepto 453,53 indica a bilheteria estimada de um filme com orçamento igual a zero (o que é apenas teórico, claro).

9.3 Avaliação do Modelo

R^2 (R-quadrado): 0,847

Isso significa que o modelo explica 84,7% da variação na bilheteria com base no orçamento — um resultado considerado muito bom.

F-Estatística: 33,27

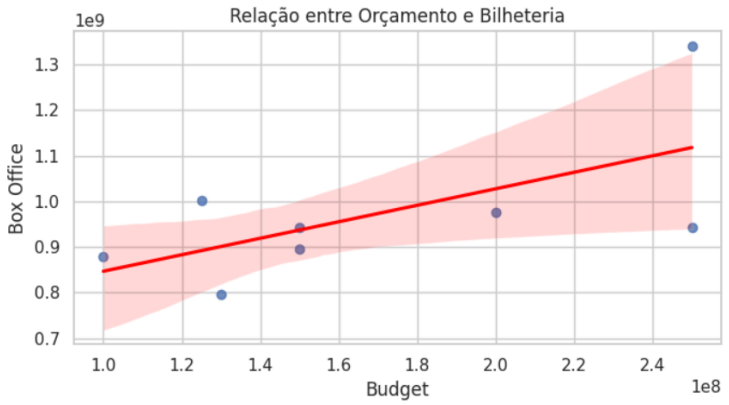
p-valor do modelo: 0,0005

Como o p-valor está abaixo de 0,05, o modelo é estatisticamente significativo. Isso mostra que há uma forte relação entre as duas variáveis.

9.4 Visualização da Regressão

O gráfico de dispersão com a linha de regressão e o intervalo de confiança deixou bem visível que os pontos estão próximos da reta — o que reforça o bom ajuste do modelo.

A regressão linear simples mostrou que existe uma relação direta e forte entre o orçamento e a bilheteria dos filmes. Ou seja, quanto maior o investimento na produção, maior tende a ser o retorno financeiro. Esse resultado confirma uma tendência comum no cinema e, neste caso, serviu para reforçar a lógica do universo de Harry Potter com base em dados reais.



Parâmetro	Valor
Intercepto (β_0)	453.53
Coeficiente (β_1)	2.44
R-quadrado (R^2)	847
F-estatística	33.27
p-valor do modelo	5

10. Conclusão

Essa análise dos filmes de Harry Potter foi uma forma prática e até divertida de colocar em ação tudo o que aprendemos sobre estatística. Trabalhar com esse conjunto de dados ajudou a entender melhor como tratar, visualizar e interpretar informações, além de mostrar como a estatística está presente até mesmo em algo que parece só entretenimento.

Ao longo do projeto, vimos que os filmes foram lançados certo, em ordem cronológica, o que manteve a continuidade da saga. Também percebemos que, mesmo seguindo essa ordem, a duração dos filmes variou bastante, o que provavelmente foi uma escolha criativa baseada em cada história.

A correlação entre orçamento e bilheteria também chamou atenção, quanto mais dinheiro foi investido, maior foi o retorno. Isso confirma aquela ideia de que produções com mais recursos tendem a atrair mais o público. E o teste de regressão só reforçou isso, mostrando que o orçamento realmente tem influência no sucesso financeiro.

Por outro lado, outras análises mostraram que nem tudo está diretamente ligado, como o caso da casa dos personagens ou a espécie deles não afetarem tanto assim os números.

No fim, a análise ajudou a enxergar como diferentes variáveis se relacionam e como podemos usar ferramentas estatísticas para tirar conclusões bem interessantes. Mais do que números, deu pra perceber que dados contam histórias, e nesse caso, histórias de uma das sagas mais famosas do cinema.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 10719: Informação e documentação – Relatório técnico e/ou científico – Apresentação. Rio de Janeiro, 2011. Disponível em: <https://files.cercomp.ufg.br/weby/up/378/o/NBR_10719_-_2011.pdf>. Acesso em: 06 jul. 2025.

KAGGLE. Harry Potter Movies Dataset. Disponível em: <https://www.kaggle.com/datasets/maricinnamon/harry-potter-movies-dataset>. Acesso em: 06 jul. 2025.

Apêndice A – Códigos Python Utilizados na Análise

A seguir, estão os principais trechos de código utilizados para a análise estatística dos dados dos filmes e personagens da franquia Harry Potter. Todos os procedimentos foram executados na linguagem Python, com uso das bibliotecas pandas, numpy, seaborn, matplotlib, scipy e statsmodels.

1. Importação das bibliotecas

In [2]:

```
import pandas as pd
from scipy import stats
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="whitegrid")
```

2. Leitura dos arquivos CSV

In [3]:

```
movies = pd.read_csv('Movies.csv', encoding='utf-8-sig')
characters = pd.read_csv('Characters.csv', encoding='latin1')
dialogue = pd.read_csv('Dialogue.csv', encoding='latin1')
```

3. Visualização inicial dos dados

In [4]:

```
print("Filmes:")
print(movies.head())

print("\nPersonagens:")
print(characters.head())
```

4. Estrutura e dimensões

```
print("Estrutura dos dados - Movies:")
movies.info()

print("\nEstrutura dos dados - Characters:")
characters.info()
```

In [6]:

```
print("\nFilmes - Linhas e Colunas:", movies.shape)
print("Personagens - Linhas e Colunas:", characters.shape)
```

```
Filmes - Linhas e Colunas: (8, 6)
Personagens - Linhas e Colunas: (166, 8)
```

5. Tratamento dos dados faltantes e conversão de tipos

In [7]:

```
#padroniza e remove espaços extras
characters['Species'] =
characters['Species'].str.strip().str.capitalize()
characters['Gender'] =
characters['Gender'].str.strip().str.capitalize()
characters['House'] = characters['House'].str.strip().str.capitalize()

#converte para tipo categórico
characters['Species'] = characters['Species'].astype('category')
characters['Gender'] = characters['Gender'].astype('category')
characters['House'] = characters['House'].astype('category')
```

6. Estatísticas descritivas

In [13]:

```
#descritivas
movies.describe()
```

In [14]:

```
characters.describe()
```

7. Gráficos: histogramas, boxplots e heatmap

In [19]:

```
#histogramas em sequência
plt.figure(figsize=(6, 3))
sns.histplot(characters['Character ID'], bins=20, kde=False,
color='lightblue')
plt.title('Histograma - ID dos Personagens')
plt.xlabel('ID dos Personagens')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()
```

```
plt.figure(figsize=(6, 3))
sns.histplot(movies['Movie ID'], bins=8, kde=False, color='orange')
plt.title('Histograma - ID dos Filmes')
plt.xlabel('ID dos Filmes')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.histplot(movies['Release Year'], bins=10, kde=False,
color='salmon')
plt.title('Histograma - Ano de Lançamento')
plt.xlabel('Ano de Lançamento')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.histplot(movies['Runtime'], bins=10, kde=False, color='purple')
plt.title('Histograma - Tempo de Execução')
plt.xlabel('Minutos')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.histplot(movies['Budget'], bins=10, kde=False, color='teal')
plt.title('Histograma - Orçamento dos Filmes')
plt.xlabel('Milhões de dólares')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.histplot(movies['Box Office'], bins=10, kde=False,
color='darkred')
plt.title('Histograma - Bilheteria dos Filmes')
plt.xlabel('Milhões de dólares')
plt.ylabel('Frequência')
plt.tight_layout()
plt.show()
```

In [20]:

```
#boxplots em sequência
plt.figure(figsize=(6, 3))
sns.boxplot(x=characters['Character ID'], color='lightblue')
plt.title('Boxplot - ID dos Personagens')
plt.xlabel('ID dos Personagens')
plt.tight_layout()
plt.show()
```

```
plt.figure(figsize=(6, 3))
sns.boxplot(x=movies['Movie ID'], color='orange')
plt.title('Boxplot - ID dos Filmes')
plt.xlabel('ID dos Filmes')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.boxplot(x=movies['Release Year'], color='salmon')
plt.title('Boxplot - Ano de Lançamento')
plt.xlabel('Ano de Lançamento')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.boxplot(x=movies['Runtime'], color='purple')
plt.title('Boxplot - Tempo de Execução')
plt.xlabel('Minutos')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.boxplot(x=movies['Budget'], color='teal')
plt.title('Boxplot - Orçamento dos Filmes')
plt.xlabel('Milhões de dólares')
plt.tight_layout()
plt.show()

plt.figure(figsize=(6, 3))
sns.boxplot(x=movies['Box Office'], color='darkred')
plt.title('Boxplot - Bilheteria dos Filmes')
plt.xlabel('Milhões de dólares')
plt.tight_layout()
plt.show()
```

In [21]:

```
#apenas variáveis numéricas dos filmes
variaveis_numericas = ['Movie ID', 'Release Year', 'Runtime',
'Budget', 'Box Office']

#calcular a correlação de Pearson
correlacoes = movies[variaveis_numericas].corr(method='pearson')

#matriz de correlação
print("Matriz de Correlação (Pearson):")
print(correlacoes)

plt.figure(figsize=(7, 4))
sns.heatmap(correlacoes, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Matriz de Correlação entre Variáveis Quantitativas')
```

```
plt.tight_layout()
plt.show()
```

8. Testes de Hipóteses

a) Teste t (Gênero × Character ID)

In [24]:

```
#hipótese 1: diferença no tempo de fala entre homens e mulheres

dialogue['WordCount'] = dialogue['Dialogue'].str.split().str.len()

words_per_character = dialogue.groupby('Character
ID')['WordCount'].sum().reset_index()

merged = pd.merge(words_per_character, characters[['Character ID',
'Gender']], on='Character ID')

male = merged[merged['Gender'] == 'Male']['WordCount'].dropna()
female = merged[merged['Gender'] == 'Female']['WordCount'].dropna()

from scipy import stats
t_stat, p_value = stats.ttest_ind(male, female)

print(f"t = {t_stat:.3f}, p = {p_value:.3f}")
```

b) ANOVA (Espécie × Character ID)

In [25]:

```
#hipótese 2: diferença no número de palavras faladas entre casas
dialogue['WordCount'] = dialogue['Dialogue'].str.split().str.len()

words_per_character = dialogue.groupby('Character
ID')['WordCount'].sum().reset_index()

merged = pd.merge(words_per_character, characters[['Character ID',
'House']], on='Character ID')

merged = merged.dropna(subset=['House'])

houses = merged['House'].unique()
word_counts_by_house = [merged[merged['House'] == house]['WordCount']
for house in houses]

#teste ANOVA
f_stat, p_value = stats.f_oneway(*word_counts_by_house)
```

```
print(f"F = {f_stat:.3f}, p = {p_value:.3f}")
```

c) ANOVA (Casa × Budget)

In [25]:

```
#hipótese 2: diferença no número de palavras faladas entre casas
dialogue['WordCount'] = dialogue['Dialogue'].str.split().str.len()

words_per_character = dialogue.groupby('Character
ID')['WordCount'].sum().reset_index()

merged = pd.merge(words_per_character, characters[['Character ID',
'House']], on='Character ID')

merged = merged.dropna(subset=['House'])

houses = merged['House'].unique()
word_counts_by_house = [merged[merged['House'] == house]['WordCount']
for house in houses]

#teste ANOVA
f_stat, p_value = stats.f_oneway(*word_counts_by_house)

print(f"F = {f_stat:.3f}, p = {p_value:.3f}")
```

9. ANOVA entre variáveis quantitativas

```
#teste ANOVA
f_stat, p_value = stats.f_oneway(*word_counts_by_house)
```

```
#teste ANOVA
f_stat, p_value = stats.f_oneway(*grupos)
```

10. Regressão Linear (Budget × BoxOffice)

In [22]:

```
#regressão linear
#as variáveis independentes e dependente
X = movies[['Budget', 'Runtime', 'Release Year']]  #variáveis
explicativas
y = movies['Box Office']  #variável resposta

#adiciona constante (intercepto)
X = sm.add_constant(X)

#ajustar o modelo de regressão linear
modelo = sm.OLS(y, X).fit()
```

```
#mostrar o resumo do modelo
print(modelo.summary())
```