EDA - Análise Exploratória de Dados

Disciplina: Tópicos Avançados em Inteligência Artificial
Aluno: Diego Santos Seabra
Matrícula: 0040251
Tema: Jogos da Steam

Dataset

Para esta análise exploratória de dados, foi-se utilizado um dataset público, disponível no Kaggle, cujo conteúdo apresenta dados de jogos da plataforma Steam (Steam é uma plataforma de venda de jogos digitais). O dataset foi lido através do arquivo 'games.csv'.

Objetivo

O principal objetivo com este EDA era o de comparar os jogos indie com os jogos AAA.

Descrição das Variáveis

- Metacritic (Nota do jogo na plataforma metacritic)
- Genres (Gêneros associados com este jogo)
- Indie (Se o jogo é indie ou não >> boleano)
- Presence (Número total de 'artigos' em redes sociais, ex.: Reddit)
- Platform (Plataformas em que o jogo foi lançado)
- OriginalCost (Custo original do jogo, ou preço de venda)
- Players (Modo de jogo >> single player, multiplayer, etc)
- Controller (Se o jogo pode ser jogado com um gamepad ou não >> boleano)
- Languages (Idiomas em que o jogo pode ser jogado)
- ESRB (Classificação ESRB)
- Achievements (Número de conquistas disponíveis em um jogo)

Limpeza dos Dados

Antes de se realizar o EDA, foi-se necessária a limpeza dos dados do dataset. Algumas colunas apresentavam dados 'NaN' e outra colunas não eram úteis para a análise em questão.

Estas colunas em questão foram removidas:

- RawgID
- SteamURL
- RatingsBreakdown
- Soundtrack
- Franchise
- DiscountedCost
- Publisher
- Graphics
- Storage
- Memory
- Tags
- Description

As outras colunas do dataset foram mantidas.

In [58]:
```
# Imports
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import norm
```

In [59]:
```
# Lê o arquivo csv
df = pd.read_csv('games.csv')

# Remove os NaN (para evitar erros)
df.dropna(inplace=True)
```

In [114…
```
# Se necessário, cria uma amostra com um pedaço do dataset
# smp = df.sample(500)
# Aqui estarei usando o dataset por inteiro
smp = df
```

In [61]:
```
smp.head()
```

Out[61]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	OriginalC
0	1	Counter-Strike: Global Offensive	83.0	Action, Free to Play	0.0	1009588.0	PC, Xbox 360, PlayStation 3	2012-08-21	Free to
1	2	Destiny 2	82.0	Action, Adventure, Free to Play	0.0	1007425.0	PlayStation 5, Web, Xbox Series X, PC, Xbox On...	2017-09-06	Free To
4	5	Sea of Thieves	68.0	Action, Adventure	0.0	777456.0	PC, Xbox One	2018-03-20	$3
6	7	Tom Clancy's Rainbow Six Siege	75.0	Action	0.0	1001424.0	PlayStation 4, PC, Xbox One	2015-12-01	$1
7	8	Rocket League	86.0	Action, Indie, Racing, Sports	1.0	1006678.0	Linux, macOS, PC, PlayStation 4, Xbox One, Nin...	2015-07-07	$1

Alguns valores, como o 'OriginalCost' possuíam o '$' em sua string e foi necessária a sua remoção

In [62]:

```
# Corrigindo os valores da coluna 'OriginalCost'
smp['OriginalCost'] = smp['OriginalCost'].str.replace('$','')
costTerms = ['Free', 'free', 'Demo', 'demo', 'Party', 'party', 'try', 'Try
costContainer = smp['OriginalCost'].str.contains('|'.join(costTerms), na=Fa
smp.loc[costContainer, 'OriginalCost'] = '0.00'
```

Também foram criadas novas colunas no dataset, afim de quantificar:

- Número de Idiomas
- Número de Plataformas
- Número de Gêneros
- Número de Modos de Jogo

In [63]:
```
# Cálculos dos números de categorias
def countQte(x):
    return len(x.split(','))

# Calculando a quantidade de idiomas
smp['nLanguages'] = smp['Languages'].apply(lambda x: countQte(x))
# Calculando a quantidade de plataformas
smp['nPlatforms'] = smp['Platform'].apply(lambda x: countQte(x))
# Calculando a quantidade de gêneros
smp['nGenres'] = smp['Genres'].apply(lambda x: countQte(x))
# Calculando a quantidade de modos de jogo
smp['nPlayers'] = smp['Players'].apply(lambda x: countQte(x))
```

Foram empregadas mais algumas correções nas colunas 'Indie' e 'OriginalCost', para acertar o tipo de dados das mesmas

In [64]:
```
# Correção dos tipos das colunas
smp = smp.astype({'Indie': bool, 'OriginalCost': float})
```

Correlação de Dados

Uma das informações mais interessantes entre as variáveis de um dataset é a correlação entre essas variáveis. Na estatística, usamos a correlação para identificar a interdependência entre duas ou mais variáveis e portanto, foi aplicada abaixo.

In [65]:
```
# Cria um dataframe para a correlação de dados (faz o drop do id pois não
corrData = smp.corr()
corrData = corrData.drop('id', 0)
corrData = corrData.drop('id', 1)
corrData
```

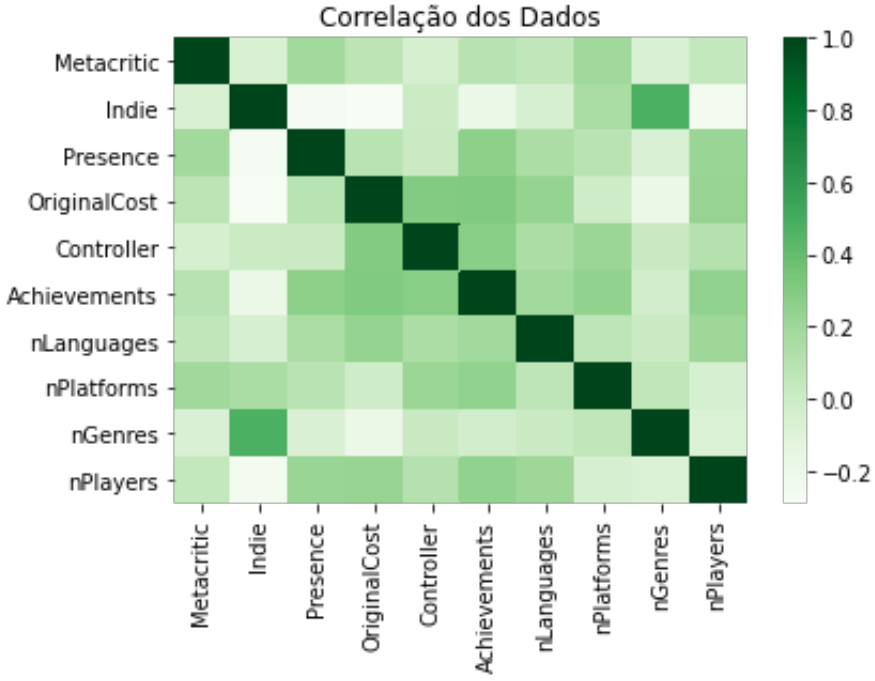
Out[65]:

	Metacritic	Indie	Presence	OriginalCost	Controller	Achievements	nl
Metacritic	1.000000	-0.058499	0.189542	0.078827	-0.039166	0.101042	
Indie	-0.058499	1.000000	-0.261363	-0.289052	0.009061	-0.173863	
Presence	0.189542	-0.261363	1.000000	0.097730	0.028020	0.269758	
OriginalCost	0.078827	-0.289052	0.097730	1.000000	0.302628	0.307741	
Controller	-0.039166	0.009061	0.028020	0.302628	1.000000	0.284224	
Achievements	0.101042	-0.173863	0.269758	0.307741	0.284224	1.000000	
nLanguages	0.064553	-0.046311	0.153851	0.234652	0.151194	0.189371	
nPlatforms	0.197477	0.154950	0.091183	-0.007082	0.219157	0.252577	
nGenres	-0.063375	0.485173	-0.063038	-0.177916	0.030955	-0.020249	
nPlayers	0.047073	-0.253658	0.221454	0.232563	0.116139	0.253695	

Demonstrando a correlação em um heatmap, temos:

In [66]:

```
# Criando a matriz de correlação dos dados
sns.heatmap(corrData, cmap='Greens')
plt.title('Correlação dos Dados')
plt.show()
```



Após uma primeira análise à este heatmap, é possível perceber que algumas variáveis parecem estarem correlacionadas, sendo elas:

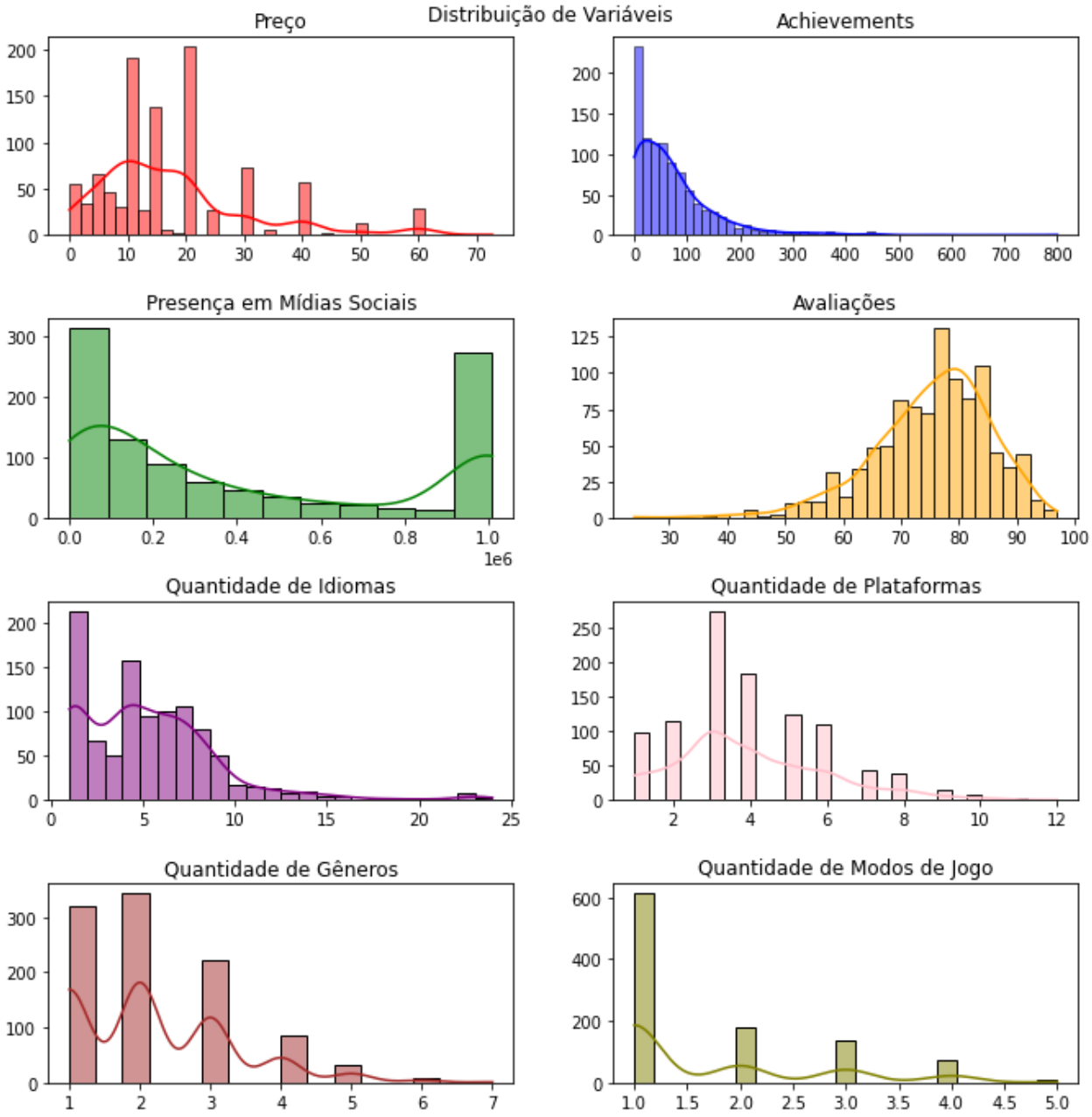
- Jogos indies *vs.* número de gêneros
- Custo do jogo *vs.* número de achievements
- Nota no metacritic *vs.* número de plataformas
- Custo do jogo *vs.* se o jogo permite ser jogado com um controle
- Nota no metacritic *vs.* presença nas redes sociais

Distribuição das Variáveis

É interessante, também, saber a distribuição (em forma de histograma) das variáveis deste dataset e, portanto:

In [67]:

```
# Distribuições de Variáveis
fig, ax = plt.subplots(4,2,figsize=(10,10))
fig.tight_layout(pad=3.0)
# plt.subplots_adjust(bottom=0.1,top=0.92,left=0.1)
fig.suptitle('Distribuição de Variáveis')
sns.histplot(data=smp['OriginalCost'], ax=ax[0][0], kde=True, color='red')
ax[0][0].set(title='Preço',ylabel='',xlabel='')
sns.histplot(data=smp['Achievements'], ax=ax[0][1], kde=True, color='blue'
ax[0][1].set(title='Achievements',ylabel='',xlabel='')
sns.histplot(data=smp['Presence'], ax=ax[1][0], kde=True, color='green')
ax[1][0].set(title='Presença em Mídias Sociais',ylabel='',xlabel='')
sns.histplot(data=smp['Metacritic'], ax=ax[1][1], kde=True, color='orange'
ax[1][1].set(title='Avaliações',ylabel='',xlabel='')
sns.histplot(data=smp['nLanguages'], ax=ax[2][0], kde=True, color='purple'
ax[2][0].set(title='Quantidade de Idiomas',ylabel='',xlabel='')
sns.histplot(data=smp['nPlatforms'], ax=ax[2][1], kde=True, color='pink')
ax[2][1].set(title='Quantidade de Plataformas',ylabel='',xlabel='')
sns.histplot(data=smp['nGenres'], ax=ax[3][0], kde=True, color='brown')
ax[3][0].set(title='Quantidade de Gêneros',ylabel='',xlabel='')
sns.histplot(data=smp['nPlayers'], ax=ax[3][1], kde=True, color='olive')
ax[3][1].set(title='Quantidade de Modos de Jogo',ylabel='',xlabel='')
plt.show()
```



Analisando os gráficos acima, é possível perceber que:

- A maioria dos jogos são disponibilizados para 3 plataformas
- A maioria dos jogos possuem nota média entre 75% e 85%
- O preço dos jogos varia bastante, porém a maioria dos jogos tem preços entre $10 e$ 20
- A maioria dos jogos não possuem achievements e aqueles que possuem normalmente possuem em pouca quantidade
- O número de idiomas disponíveis tende a variar entre 1 e 5 (em sua maioria)

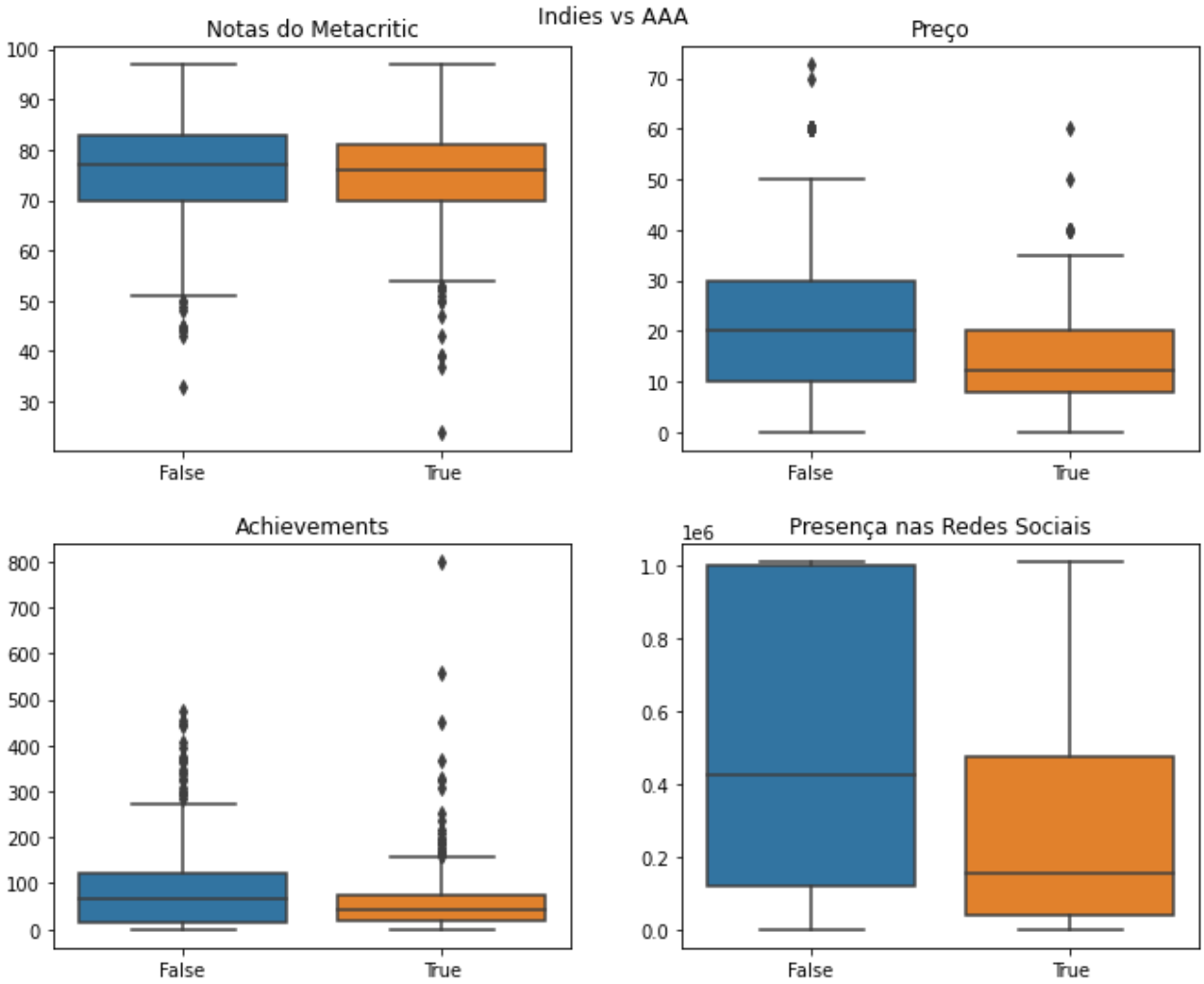
Comparação entre jogos indies e AAA

Nesta seção, tenta-se responder algumas perguntas comparando os jogos entre as categorias 'Indie' e 'AAA':

- Jogos indies são mais caros que os AAA?
- Qual das duas categorias possui melhores avaliações no Metacritic?
- Qual categoria possui mais achievements?
- Qual categoria possui maior presença nas redes sociais?

In [68]:

```
# Comparação Indie vs AAA
fig, ax = plt.subplots(2,2,figsize=(10,8))
fig.tight_layout(pad=3.0)
# plt.subplots_adjust(bottom=0.1,top=0.92,left=0.1)
fig.suptitle('Indies vs AAA')
sns.boxplot(data=smp, x='Indie', y='Metacritic', ax=ax[0][0])
ax[0][0].set(title='Notas do Metacritic',ylabel='',xlabel='')
sns.boxplot(data=smp, x='Indie', y='OriginalCost', ax=ax[0][1])
ax[0][1].set(title='Preço',ylabel='',xlabel='')
sns.boxplot(data=smp, x='Indie', y='Achievements', ax=ax[1][0])
ax[1][0].set(title='Achievements',ylabel='',xlabel='')
sns.boxplot(data=smp, x='Indie', y='Presence', ax=ax[1][1])
ax[1][1].set(title='Presença nas Redes Sociais',ylabel='',xlabel='')
plt.show()
```



Dos gráficos apresentados acima, pode-se perceber que:

- As avaliações entre jogos indie e jogos AAA são praticamente iguais
- Jogos indie em um geral custam menos que jogos AAA
- Jogos indie possuem menos achievements (em média) que os rivais AAA, porém possuem mais outliers, ou seja, alguns jogos indie possuem muitos achievements
- A maioria dos jogos AAA possuem maior presença em redes sociais. Isto pode ser devido ao grande investimento com *marketing* que as grandes corporações aplicam para adquirir mais jogadores
- Existe um jogo Indie (outlier) que possui a menor nota do dataset
- Existe um jogo AAA (outlier) que possui o maior preço do dataset

Uma avaliação interessante pode ser feita aqui:

Jogos indie, mesmo não possuindo grandes presenças em redes sociais (provavelmente com pouco investimento em marketing) e sendo normalmente mais baratos (em comparação com os AAA) conseguiram uma média de avaliações praticamente igual aos do AAA (com exceção dos outliers indie que foram extremamente mal avaliados)

Avaliações

- Quais são os piores jogos? Quais os melhores?
- Qual o jogo mais bem avaliado? Qual o pior? E os indies?
- Qual o jogo mais caro?

Jogo mais bem avaliado

In [69]:
```
# Jogo mais bem avaliado
smp.sort_values(by='Metacritic', ascending=False).head(1)
```

Out[69]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	OriginalCos
10	11	Grand Theft Auto V	97.0	Action, Adventure	False	1004912.0	PlayStation 5, PC, PlayStation 4, PlayStation ...	2013-09-17	29.9

Jogo com pior avaliação

In [70]:
```
# Jogo com pior avaliação
smp.sort_values(by='Metacritic', ascending=True).head(1)
```

Out[70]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	Original(
21089	21090	Postal III	24.0	Action	True	152640.0	PC	2012-02-17	1

Top 5 jogos mais bem avaliados

In [71]:
```
# Top 5 jogos mais bem avaliados
smp.sort_values(by='Metacritic', ascending=False).head(5)
```

Out[71]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDa
10	11	Grand Theft Auto V	97.0	Action, Adventure	False	1004912.0	PlayStation 5, PC, PlayStation 4, PlayStation ...	2013-09-
20083	20084	Grand Theft Auto V	97.0	Free to Play, Indie, Massively Multiplayer, St...	True	1004912.0	PlayStation 5, PC, PlayStation 4, PlayStation ...	2013-09-
15	16	Red Dead Redemption 2	97.0	Action, Adventure	False	1004105.0	PC, Xbox One, PlayStation 4	2018-10-2
753	754	BioShock	96.0	Action, RPG	False	1002857.0	Xbox One, PC, PlayStation 4, PlayStation 3, ma...	2007-08-
146	147	Portal 2	95.0	Action, Adventure	False	1006285.0	PlayStation 3, PC, Xbox 360, Linux, macOS	2011-04-

Top 5 jogos mais caros

In [72]:

```
# Top 5 jogos mais caros
smp.sort_values(by='OriginalCost', ascending=False).head(5)
```

Out[72]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	Or
6323	6324	Sleeping Dogs	81.0	Action, Adventure	False	1001518.0	Xbox 360, PC, PlayStation 3	2012-08-14	
103	104	Need for Speed Heat	73.0	Action, Adventure, Racing, Sports	False	667277.0	Xbox One, PlayStation 4, PC	2019-11-08	
64	65	DOOM Eternal	88.0	Action	False	31723.0	PC, Nintendo Switch, PlayStation 4, Xbox One	2020-03-20	
164	165	Call of Duty: Black Ops II	74.0	Action	False	1001917.0	PlayStation 3, PC, Xbox 360, Xbox One, Wii U	2012-11-13	
51	52	Call of Duty: Black Ops III	73.0	Action, Adventure	False	1002053.0	PC, PlayStation 4, Xbox 360, Xbox One, PlaySta...	2015-11-06	

Top 5 jogos com mais achievements

In [80]:

```
# Top 5 jogos com mais achievements
smp.sort_values(by='Achievements', ascending=False).head(5)
```

Out[80]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate
2026	2027	One Finger Death Punch 2	81.0	Action, Casual, Indie	1.0	115644.0	PC, Xbox One, Nintendo Switch	2019-04-14
189	190	The Binding of Isaac: Rebirth	86.0	Action	1.0	970552.0	PC, iOS, Nintendo 3DS, macOS, Linux, Wii U, PS...	2014-11-04
206	207	Neverwinter	74.0	Adventure, Free to Play, Massively Multiplayer...	0.0	1001196.0	PC, Xbox One, PlayStation 4	2013-06-20
13	14	Warframe	70.0	Action, Free to Play	0.0	1009230.0	Xbox Series X, PlayStation 5, PC, Nintendo Swi...	2013-03-25
506	507	Plague Inc: Evolved	80.0	Casual, Indie, Simulation, Strategy	1.0	393630.0	Nintendo Switch, Linux, macOS, PC, Xbox One, P...	2015-09-18

Top 5 jogos com mais presença em redes sociais

In [82]:

```
# Top 5 jogos com mais achievements
smp.sort_values(by='Presence', ascending=False).head(5)
```

Out[82]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	Origina
0	1	Counter-Strike: Global Offensive	83.0	Action, Free to Play	0.0	1009588.0	PC, Xbox 360, PlayStation 3	2012-08-21	
469	470	Mass Effect 2	94.0	RPG	0.0	1009410.0	PC, PlayStation 3, Xbox 360	2010-01-26	
13	14	Warframe	70.0	Action, Free to Play	0.0	1009230.0	Xbox Series X, PlayStation 5, PC, Nintendo Swi...	2013-03-25	
486	487	Mass Effect	89.0	Action, RPG	0.0	1009056.0	PC, PlayStation 3, Xbox 360	2007-11-16	
14	15	Monster Hunter: World	89.0	Action	0.0	1008752.0	PlayStation 4, PC, Xbox One	2018-01-26	

In [73]:
```
# Separa os jogos indie em um outro dataframe
indies = smp.loc[smp['Indie'] == True]
```

Top 5 jogos indies mais caros

In [74]:
```
# Top 5 jogos indies mais caros
indies.sort_values(by='OriginalCost', ascending=False).head(5)
```

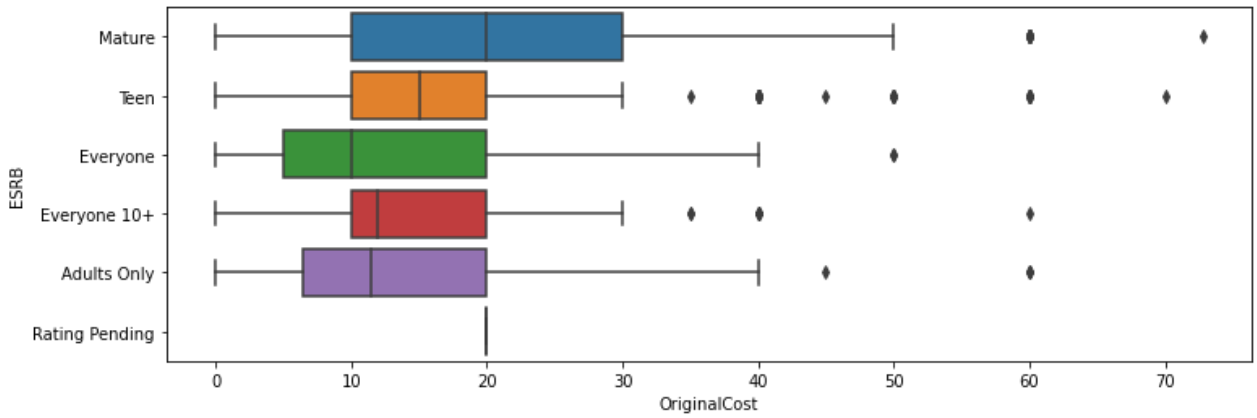
Out[74]:

	id	Name	Metacritic	Genres	Indie	Presence	Platform	ReleaseDate	C
526	527	We Happy Few	64.0	Action, Adventure, Indie	True	612944.0	PC, PlayStation 4, Xbox One	2018-08-09	
988	989	Disintegration	63.0	Action, Adventure	True	539.0	PlayStation 4, Xbox One, PC	2020-06-16	
922	923	Indivisible	79.0	Action, Indie, RPG	True	77300.0	PC, macOS, Xbox One, PlayStation 4, Linux, Nin...	2019-10-07	
47	48	Disco Elysium	91.0	RPG	True	52433.0	PC	2019-10-14	
355	356	Divinity: Original Sin - Enhanced Edition	94.0	Adventure, Indie, RPG, Strategy	True	180413.0	macOS, Linux, Xbox One, PlayStation 4, PC	2015-10-27	

Distribuição dos Preços por ESRB Rating

In [75]:

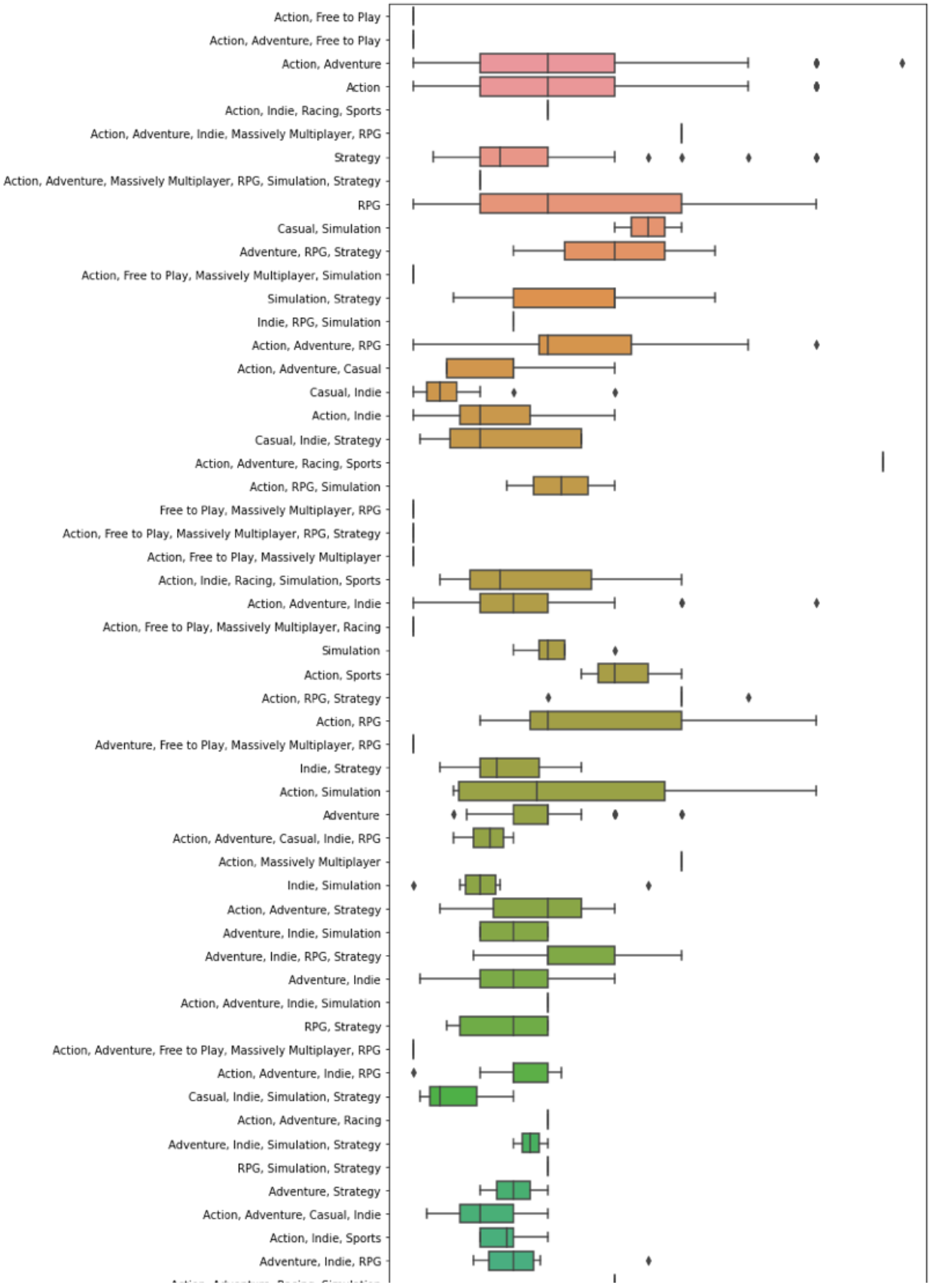
```
# ESRB
plt.figure(figsize=(12,4))
sns.boxplot(data=smp, x='OriginalCost', y='ESRB')
plt.show()
```



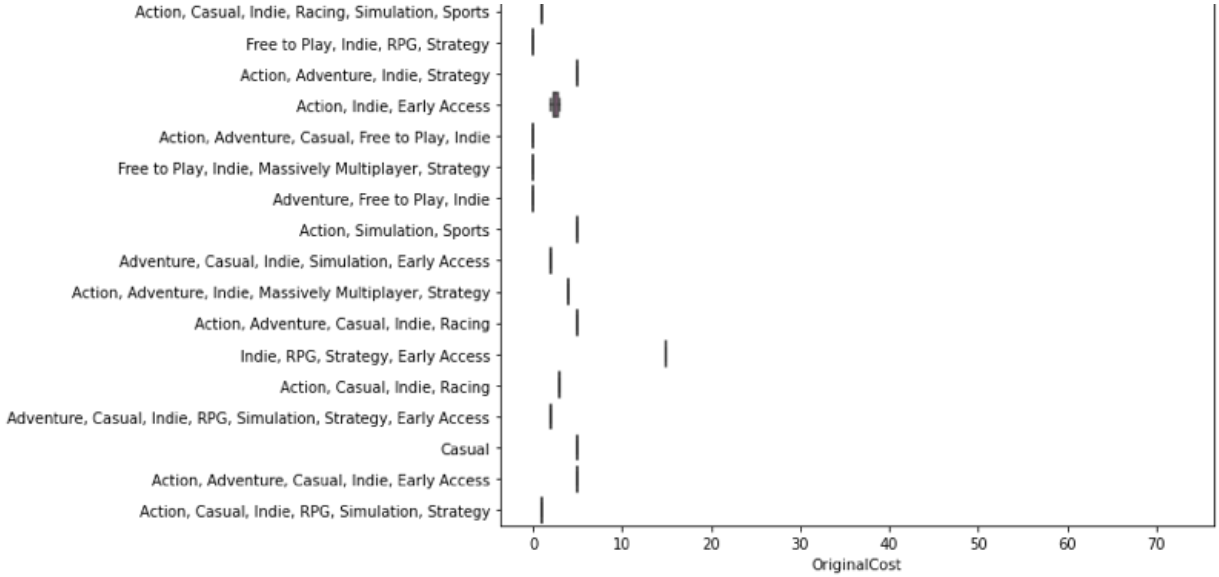
Distribuição dos Preços em cada Gênero

In [152…

```
# Distribuição dos Preços em cada Gênero
plt.figure(figsize=(12,40))
sns.boxplot(data=smp, x='OriginalCost', y='Genres')
plt.tight_layout()
plt.show()
```




Genres
Action, Adventure, Racing, Simulation
Adventure, RPG
Action, Indie, Simulation
Action, Adventure, Racing, Sports, Strategy
Indie, Simulation, Strategy
Action, Strategy
Racing, Simulation, Sports, Strategy
Action, Adventure, Indie, RPG, Early Access
Action, Simulation, Strategy
Racing, Simulation, Sports
Action, Indie, RPG
Indie, RPG, Strategy
Racing, Sports
Casual, Indie, Simulation
Adventure, Casual, Indie
Indie
Action, Indie, RPG, Strategy
Action, Casual
Casual, Strategy
Action, Racing
Adventure, Casual
Action, Indie, Strategy
Adventure, Indie, Strategy
Sports
Simulation, Sports
Indie, RPG
Action, Massively Multiplayer, Racing
Action, Casual, Indie
Action, Indie, Simulation, Strategy
Racing
Action, Adventure, Casual, Indie, Strategy
Action, Adventure, Racing, Simulation, Strategy
Action, Adventure, Casual, RPG
Free to Play, Massively Multiplayer, Simulation, Strategy
Adventure, Casual, Indie, RPG, Simulation, Strategy
Action, Free to Play, Massively Multiplayer, RPG
Action, Free to Play, Indie, Strategy
Action, Adventure, Casual, Indie, RPG, Simulation, Early Access
Adventure, Simulation
Adventure, Casual, RPG
Action, Racing, Simulation
Indie, RPG, Simulation, Strategy
Adventure, Casual, Indie, Simulation
Action, Indie, Racing, Strategy
Action, Casual, Simulation, Strategy
Action, Adventure, Casual, Indie, Simulation, Strategy
Indie, Casual
Free to Play, Strategy
Action, Adventure, Indie, RPG, Strategy
Action, Adventure, Indie, RPG, Simulation
Action, Casual, Indie, Early Access
Action, Casual, Indie, Strategy
Action, Adventure, Indie, RPG, Strategy, Early Access
Action, Strategy, Indie, Sports
Adventure, Casual, Indie, RPG
Action, Free to Play, Indie
Action, Adventure, Indie, Early Access
Adventure, Casual, Indie, RPG, Strategy
Casual, Indie, RPG, Strategy
Action, Free to Play, Indie, Early Access
Action, Casual, Indie, Simulation, Sports
Action, Adventure, Casual, Indie, Simulation
Action, Indie, Racing



Variação do Preço no Ano de 2020

In [113…

```
# Variação do preço no ano de 2020
jogos2020 = smp.loc[smp['ReleaseDate'] >= '2020-01-01']
jogos2020 = jogos2020.sort_values(by='ReleaseDate', ascending=True)
fig, ax = plt.subplots(figsize=(10,4))
sns.lineplot(data=jogos2020, x='ReleaseDate', y='OriginalCost', ax=ax)
plt.ylabel('Preço')
plt.xlabel('Data')
plt.xticks(rotation=30)
plt.title('Variação do Preço no Ano de 2020')
plt.show()
```

In [ ]: