UNIVERSIDADE FEDERAL DO CEARÁ

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GUSTAVO CAMPELO DE SOUSA – 511817

PABLO KAUAN MARTINS TIMBÓ – 556012

RODOLFO RODRIGUES DE ARAÚJO – 508473

LUCAS EVANGELISTA DE CARVALHO - 510158

TRABALHO DE CIÊNCIA DOS DADOS

CRATÉUS

2025

SUMÁRIO

1 DATASET IMDB 2
1.1 Origem e características do dataset 2
1.1.1 Principais Dados Faltantes 2
1.1.2 Transformações 2
1.1.3 Questões 3
1.1.4 Limitações 5
2 DATASET PLAYSTATION GAMES INFO 2/15/2025 6
2.1 Origem e características do dataset 6
2.1.1 Principais Dados Faltantes 6
2.1.2 Transformações 7
2.1.3 Análise exploratória dos dados 7
2.1.4 Questões e visualização dos dados 8
2.1.5 Limitações 11
3 DATASET CÂMARA DOS DEPUTADOS 12
3.1 Origem e características do dataset 12
3.1.1 Principais Dados Faltantes 12
3.1.2 Questões e Visualizações de Dados 12
3.1.3 Limitações 15

1 DATASET IMDB

1.1 Origem e características do dataset

O conjunto de dados analisado foi extraído do **Internet Movie Database (IMDB)**, utilizando dois arquivos principais — movies.csv e ratings.csv — que, após serem unidos, totalizaram cerca de 11,9 milhões de registros. Foram abordados aspectos éticos relacionados ao uso dos dados, que são públicos e anonimizados, garantindo a ausência de informações pessoais e justificando sua utilização para fins educacionais.

No processamento dos dados, realizou-se a junção dos conjuntos por meio de uma chave única e a eliminação de colunas redundantes ou pouco relevantes. Para lidar com valores faltantes — especialmente expressivos nas colunas de avaliação e votos —, optou-se pela eliminação de registros críticos sem informação essencial e pelo uso de codificação one-hot para variáveis categóricas, como tipo de título e gênero. Além disso, foram aplicadas transformações para categorizar a duração dos filmes e agrupar anos de lançamento em intervalos.

A análise identificou padrões relevantes, como o crescimento exponencial de produções a partir dos anos 2000 e a predominância do gênero drama, além de correlações positivas entre nota média e número de votos. Apesar dos desafios relacionados à qualidade dos dados, como alta taxa de valores ausentes e distribuições assimétricas, o trabalho demonstrou a viabilidade de utilizar esse conjunto para aplicações futuras, como sistemas de recomendação, análise de tendências e modelos preditivos

1.1.1 Principais Dados Faltantes

Tabela 1 – Principais Dados Faltantes

Coluna	**NaN**
averageRating	10.3M (86.3%)
numVotes	10.3M (86.3%)
runtimeCategory	7.7M (64.6%)

1.1.2 Transformações

- **One-Hot Encoding**:
  - titleType → 11 features

  - genres → 28 features
  - startYear → intervalos de 5 anos
- **Categorização**:
  - runtimeMinutes → [Curto, Médio, Longo, Extenso]
  - isAdult → booleano

1.1.3 Questões

- **O que mudou nos gêneros quando tratado?**
  A aplicação do One-Hot Encoding na variável de gêneros transformou uma coluna categórica múltipla em 28 features binárias individuais, permitindo análise quantitativa dos gêneros cinematográficos.

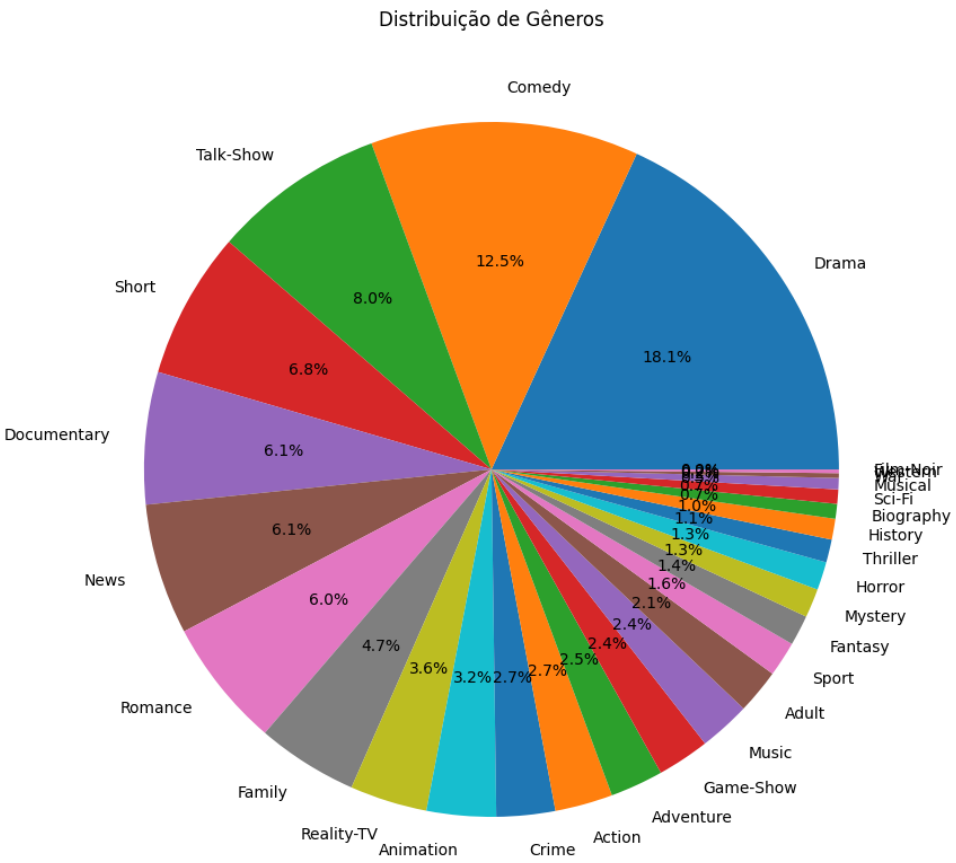


Figura 1 – Distribuição de Gêneros - Antes do Tratamento

-

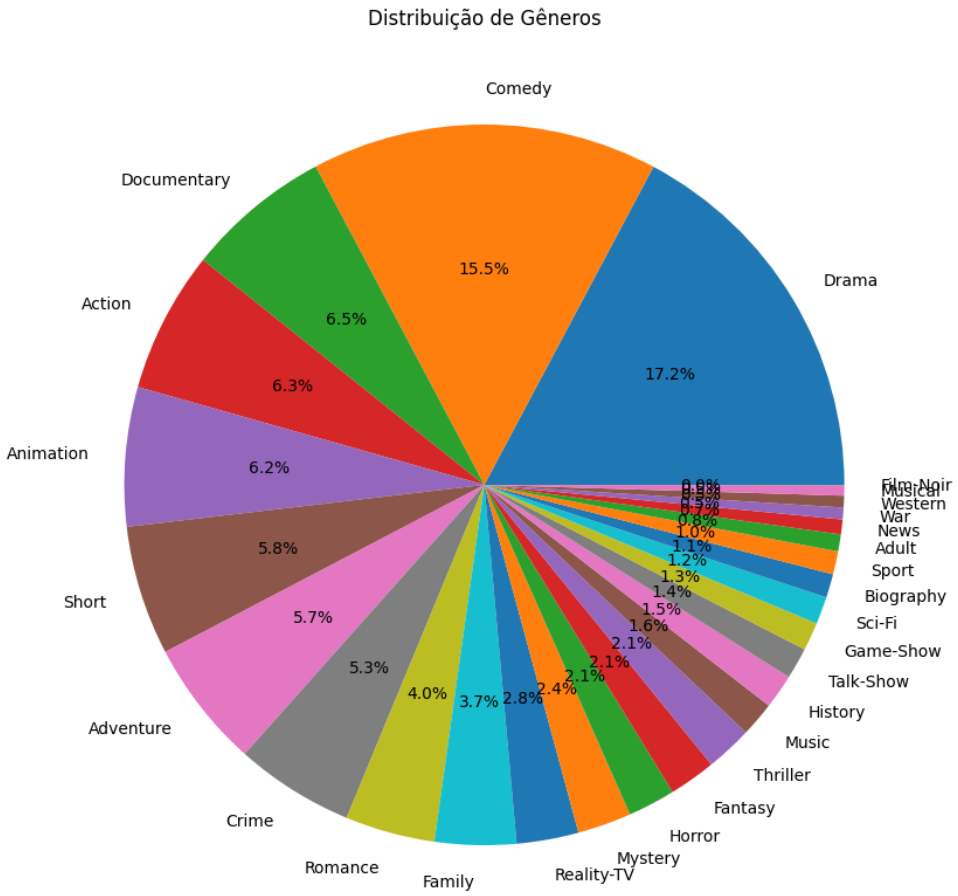
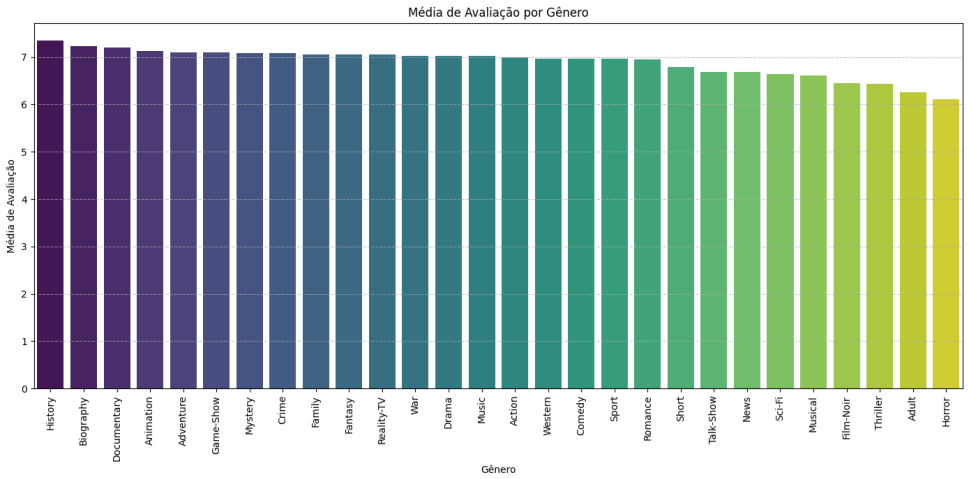


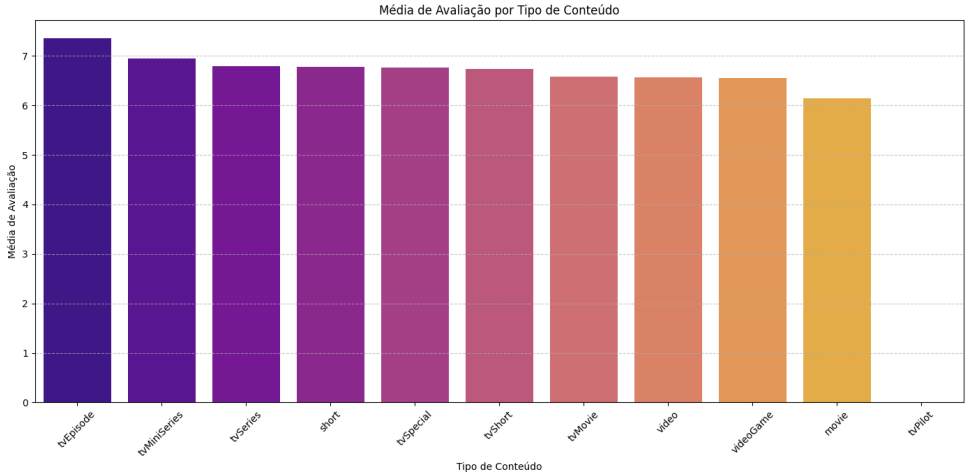
Figura 2 – Distribuição de Gêneros - Após o Tratamento

O gênero e o tipo influenciam na avaliação?

Sim, podemos ver nos respectivos gráficos, embora não tenha uma influência tão grande.



-



Houve algum tempo especifico de maior coleta ou influencia no dataset?

Sim, podemos ver nos respectivos gráficos, que o setor cineasta cresceu muito durante os anos após 2009 e muitos filmes foram lançados.

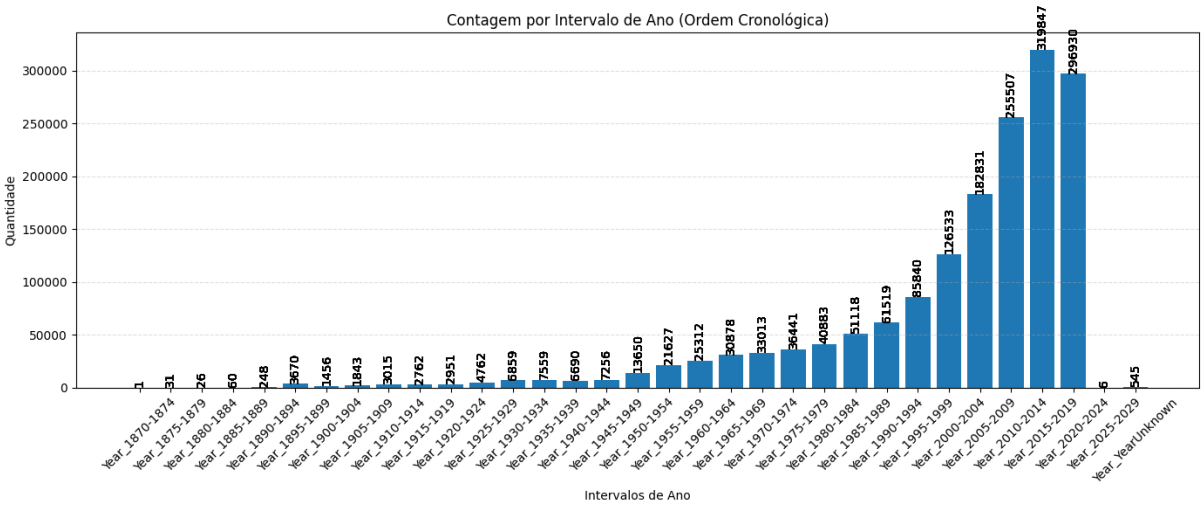


Figura 3 – Distribuição de Filmes pelo Ano

1.1.4 Limitações

Houve a limitação como o poder de processamento local insuficiente para realizar manipulações em um dataset tão grande como o do IMDB. Devido a isso, passou-se a utilizar um projeto no Google Colab.

2 DATASET PLAYSTATION GAMES INFO 2/15/2025

2.1 Origem e características do dataset

O conjunto de dados analisado foi extraído da plataforma **Kaggle** e tem como nome *PlayStation Games Info 2/15/2025*. Contém informações detalhadas sobre jogos de PlayStation, combinando dados oficiais da PlayStation Store com avaliações de críticos e de usuários do Metacritic. Tal *dataset* tinha, inicialmente, um arquivo chamado *game_details.csv*, do qual, após um tratamento realizado, foi obtido um arquivo chamado *games_tratado.csv*. Durante o processo, foram levados em consideração aspectos que facilitariam o tratamento deles, como o uso do *one-hot encoding* na coluna de plataformas para separar as *features* em uma melhor visualização das plataformas disponíveis. Foram abordados aspectos éticos relacionados ao uso dos dados, que são públicos e não contém informações sensíveis, justificando sua utilização para fins educacionais e a obtenção de uma noção sobre a indústria dos games dos consoles da Sony.

O *dataset* possui, originalmente 3526 linhas e 12 colunas. Inicialmente, foi obtido o tipo de cada dado e observado que boa parte não condizia com um formato adequado, e logo foram alterados, como a coluna de *highest_price*, que teve o símbolo do euro removido e assim convertida para *float64* e *release_date* para *datetime64[ns]*, *genre*, *publisher* e *category*, que passaram do tipo *object* para *category*. Houve também a conversão das colunas de avaliação dos jogos para o tipo *float64*.

2.1.1 Principais Dados Faltantes

Tabela 2 – Principais Dados Faltantes no dataset original

Coluna	**NaN**
highest_price	111 (3.14%)
publisher	55 (1.55%)
metacritic_score	2647 (72.23%)
metacritic_rating_count	2647 (72.23%)
metacritic_user_score	2646 (72.20%)
metacritic_user_rating_count	2646 (72.20%)
playstation_rating_count	969 (27.48%)

2.1.2 Transformações

Essa aplicação do one-hot encoding na coluna de plataforma trouxe novas colunas, o que permitiu gerar algumas visualizações, exibidas na subseção 2.1.3. A tabela 3 exibe as novas colunas com os valores faltantes correspondentes após um tratamento sobre nulos e a remoção de linhas duplicadas.

Tabela 3 – Valores faltantes após o one-hot encoding no novo dataset salvo

Coluna	**NaN**
metacritic_rating_count	2644
metacritic_score	2644
metacritic_user_rating_count	2643
metacritic_user_score	2643
playstation_rating_count	967
playstation_score	967
highest_price	108
publisher	55
genre	519
game_name	0
release_date	0
platform_PS Vita / PSP	0
platform_PS3	0
platform_PS3 / PS Vita	0
platform_PS3 / PS Vita / PSP	0
platform_PS3 / PSP	0
platform_PS4	0
platform_PS4 / PS Vita	0
platform_PS4 / PS3	0
platform_PS4 / PS3 / PS Vita	0
platform_PS5	0
platform_PS5 / PS4	0
platform_PS5 / PS4 / PS Vita	0
platform_PS5 / PS4 / PS3 / PS Vita	0
platform_PSP	0

2.1.3 Análise exploratória dos dados

Nessa etapa, obteve-se algumas informações importantes em relação aos dados, como:

Tabela 4 – Estatísticas Descritivas das Variáveis Numéricas

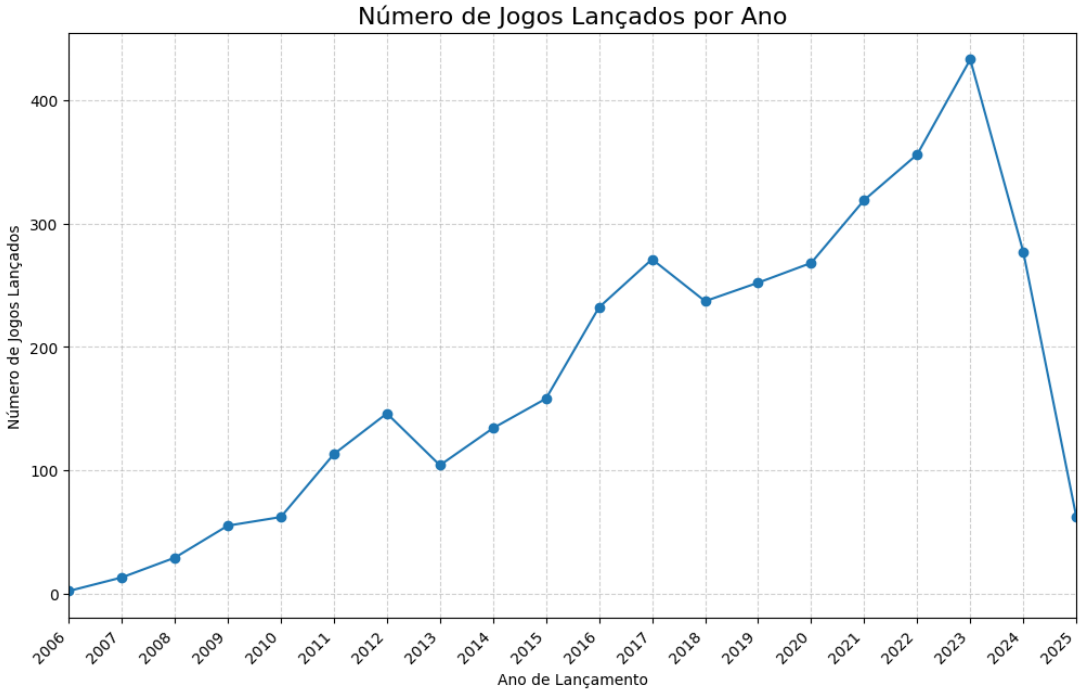
Variável	Média	Mediana	Valor máximo
highest_price	60.82	21.99	10000
metacritic_score	77.47	76.00	98.0
metacritic_rating_count	33.41	24.00	145.0
metacritic_user_score	6.65	7.40	9.5
metacritic_user_rating_count	1053.76	76.00	165959
playstation_score	4.36	4.42	5.0
playstation_rating_count	7398.34	527.00	854788

Tabela 5 – Estatísticas Descritivas das Variáveis Categóricas

Variável	Count	Valores Únicos	Mais Frequente	Frequência
game_name	3526	3270	The Legend of Heroes: Trails of Cold Steel	3
genre	3007	289	Action	373
publisher	3471	767	Sony Interactive Entertainment Europe	210

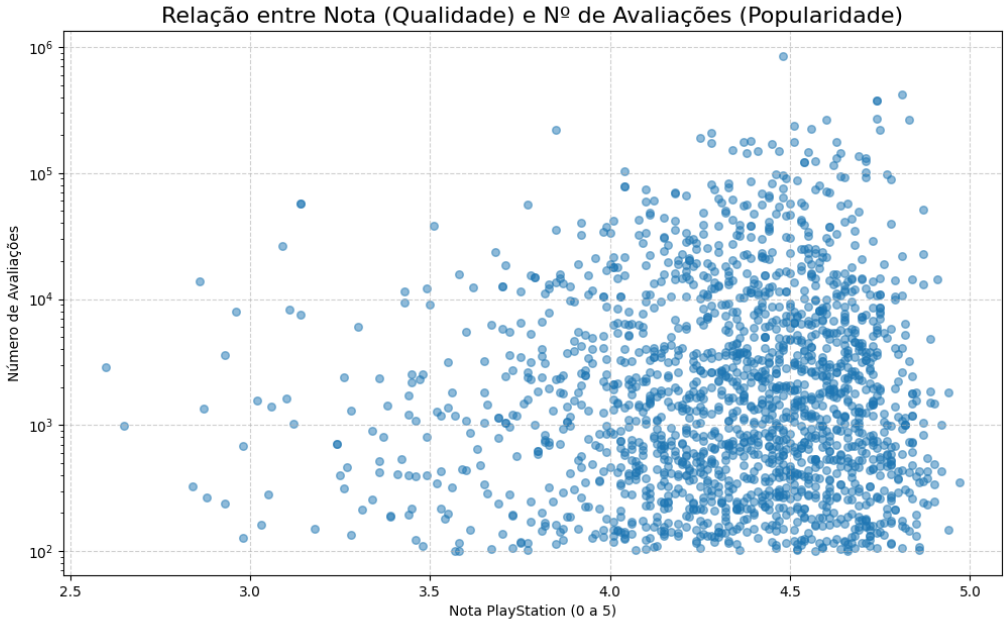
2.1.4 Questões e visualização dos dados

- **Como o número de jogos lançados nas plataformas variou ao longo dos anos?**
  Durante a geração do PS3 e, principalmente a do PS4 e PS5, a indústria aqueceu veementemente, chegando a lançar mais de 400 jogos por ano durante o ano de 2023. Entretanto, vale notar que boa parte dos jogos lançados a partir de 2021 são cross-gen, ou seja, lançados tanto para PS4 quanto para o PS5. Após 2023, houve uma queda vertiginosa e, devido a cortes de funcionários e lançamentos malquistos, as publishers optaram por um tempo de desenvolvimento mais longo, visando trazer jogos sem tantos bugs ou problemas.



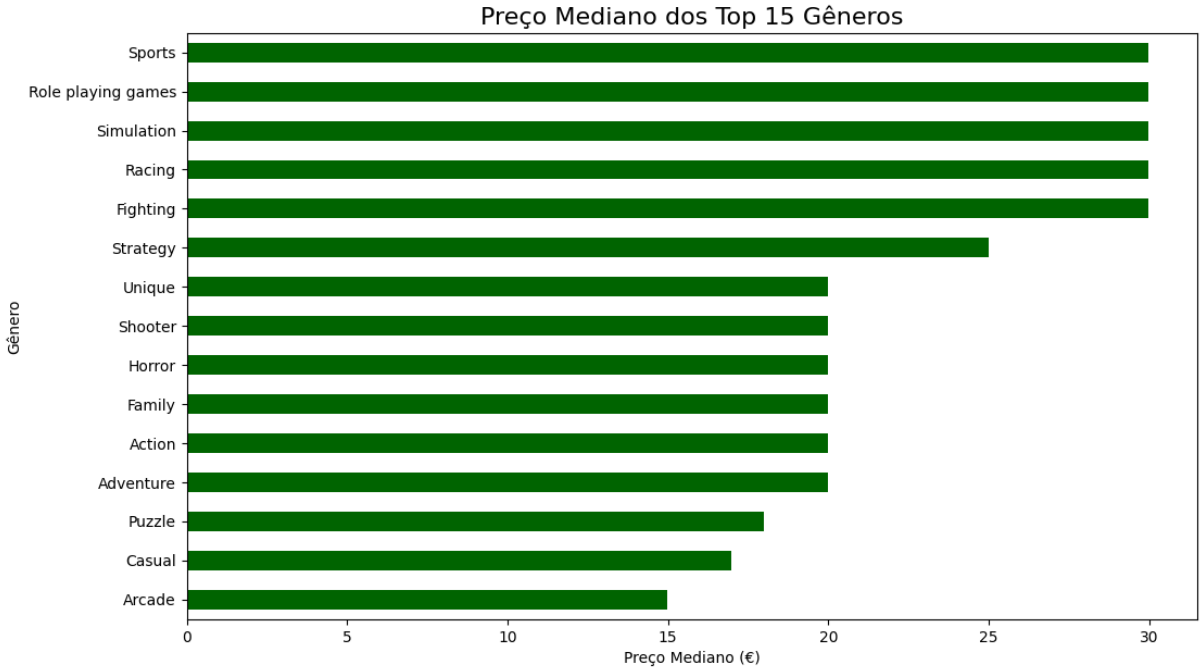
- **Jogos com notas mais altas (qualidade) são sempre os mais populares (em nº de avaliações)?**

Não, o gráfico de dispersão evidencia isso, mostrando que mesmo jogos com centenas de avaliações podem obter uma pontuação alta, e jogos com quase cem mil avaliações também podem receber notas próximas a 5, o que indica que, provavelmente, esses jogos possuem uma maior confiança, dado que mais pessoas o avaliaram. Nota-se que a maior concentração está entre 100 e 10000 avaliações, com notas acima de 4.0.



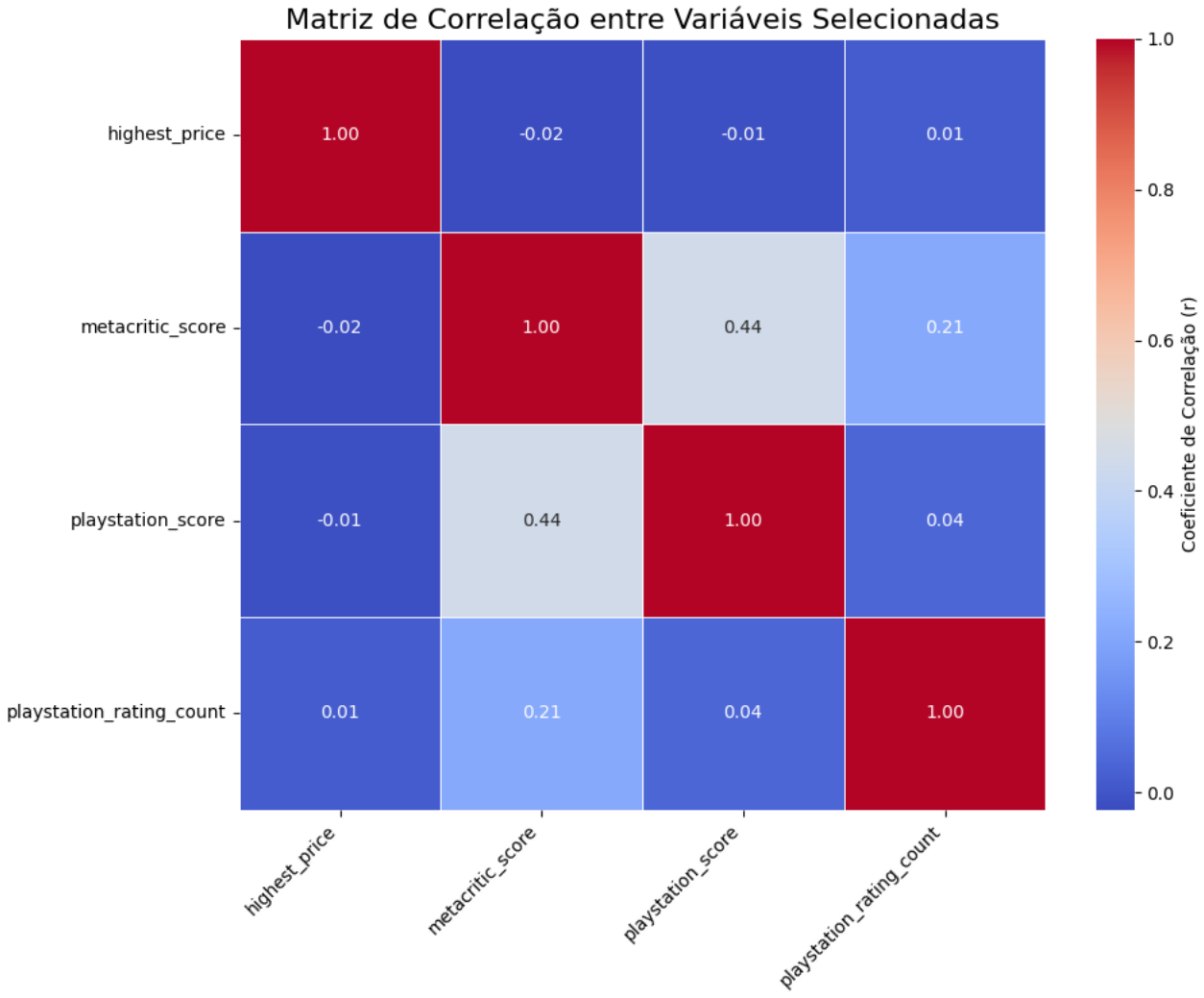
- **Existe algum gênero em que os jogos sejam mais caros do que em outros?**

Sim, os jogos de esportes, RPG e simulação se destacam nesse ponto. Um motivo curioso para isso é que se tratam exatamente de jogos lançados todos os anos, como os de futebol, basquete, Fórmula 1, UFC, entre outros, e dessa forma, sempre que há um lançamento, o preço é cheio, entre 50 a 60 euros, o que, de certa forma, enviesou para que fossem mais caros, mesmo que a média ainda se mostre com um preço relativamente baixo.



- **Existe uma correlação forte entre a nota da crítica (Metacritic), a nota do usuário (PlayStation) e o preço?**

  Não foi identificada uma correlação forte entre nenhuma das variáveis analisadas, devido a alguns fatores, como a ordem de grandeza ser diferente; por exemplo, o *metacritic_score* é uma avaliação de 0 a 100, enquanto a *playstation_score* é de 0 a 5. E não se identificou correlação visível entre a questão do preço e a taxa de avaliação dos jogos.



2.1.5 Limitações

Uma limitação evidente é a quantidade de valores vazios em algumas colunas, como a de *metacritic_score* e *playstation_score*, o que limita, de certa forma, aferições sobre a nota, considerado por muita gente, motivos para comprar os jogos.

3 DATASET CÂMARA DOS DEPUTADOS

3.1 Origem e características do dataset

O conjunto de dados analisado foi extraído da plataforma de **Dados Abertos da Câmara dos Deputados**. Ele contém informações detalhadas sobre parlamentares que já exerceram mandato na Câmara Federal, reunindo dados pessoais e públicos. Ao todo, o *dataset* possui **7868 registros**, distribuídos em **13 colunas**.

Durante o processamento, o primeiro passo consistiu em identificar os tipos de dados presentes para cada deputado, bem como valores ausentes. Em cada etapa da análise, verificou-se a existência de valores nulos, a fim de realizar os tratamentos necessários para garantir maior consistência ao conjunto de dados.

A análise inicial revelou uma predominância de parlamentares em faixas etárias mais avançadas, indicando que ainda há pouco espaço para a participação de pessoas mais jovens na política federal. Também se observou que a participação feminina permanece muito pequena quando comparada ao número de parlamentares do sexo masculino. Apesar da existência de dados faltantes, o conjunto analisado demonstra utilidade para estudos e formulação de políticas voltadas à ampliação e democratização da participação popular na vida pública.

3.1.1 Principais Dados Faltantes

Tabela 6 – Principais dados faltantes no dataset original

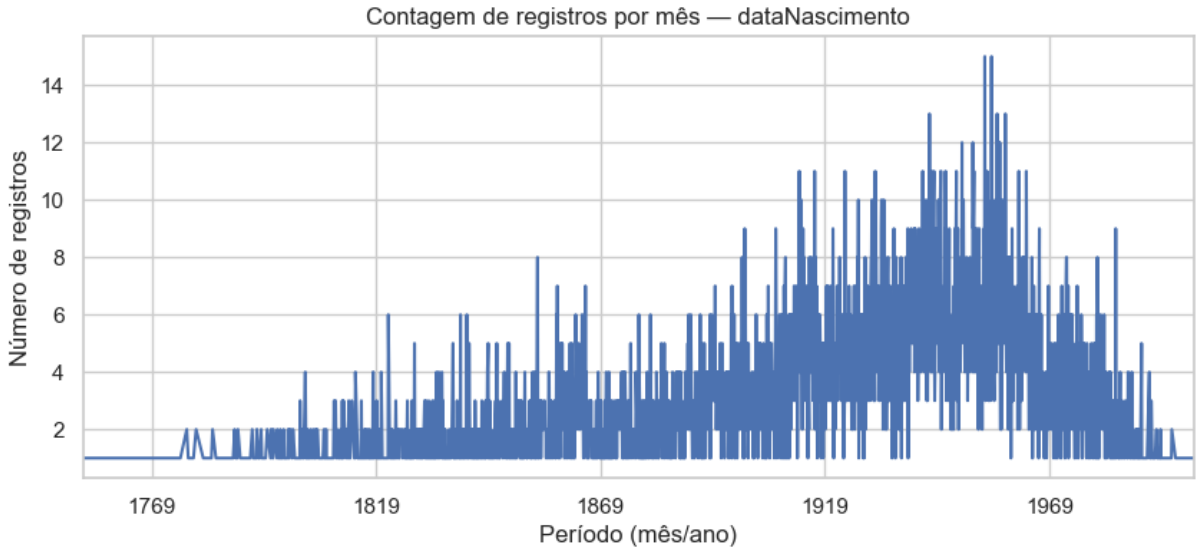
Coluna	**NaN**
CPF	7868 (100%)
municipioNascimento	1380 (17,53%)
dataNascimento	897 (11,40%)
dataFalecimento	3559 (45,23%)
ufNascimento	971 (12,34%)

3.1.2 Questões e Visualizações de Dados

- **Por que existe queda acentuada nos registros após os anos 1980?**
  Dados mais recentes aparecem menos porque parlamentares muito jovens ainda não chegaram ao cargo. A maioria dos parlamentares registrados ainda provém de gerações anteriores.

- **O pico de nascimentos entre 1930 e 1970 indica uma predominância geracional entre os parlamentares?**

  Sim. Grande parte dos parlamentares em exercício ou com registro nasceu nesse período, evidenciando forte concentração nas gerações pós-Guerra e durante o período militar.

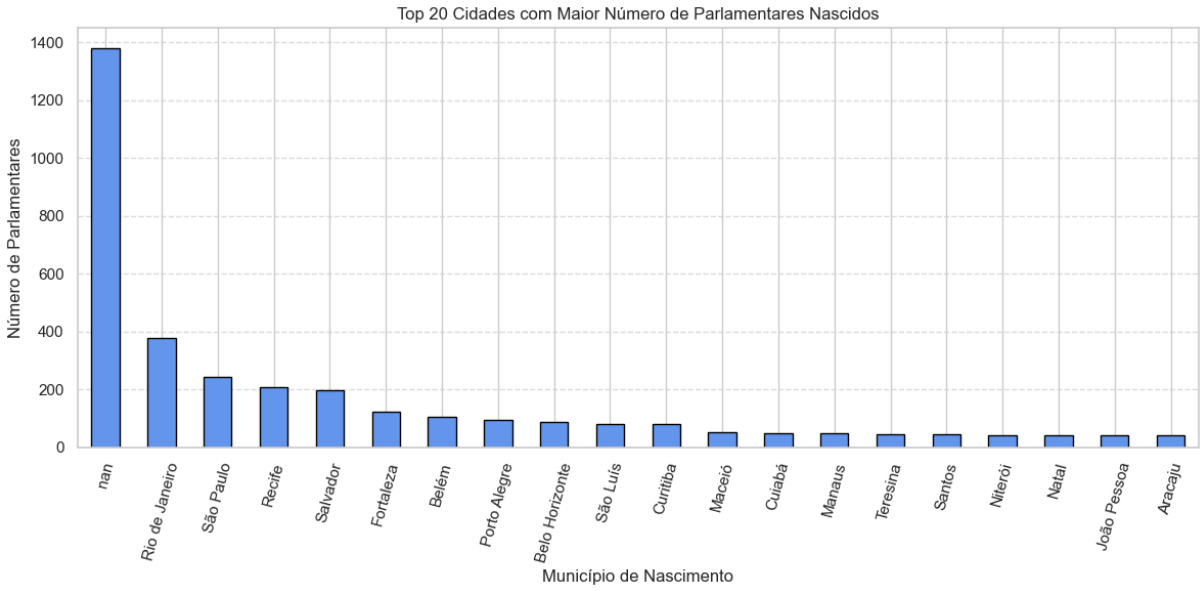


- **Existe relação entre o tamanho populacional da cidade e o número de parlamentares nascidos nela?**

  Cidades maiores tendem a produzir mais parlamentares. Entretanto, capitais como Manaus, Cuiabá e Curitiba apresentam números abaixo do esperado, contrariando parcialmente essa tendência.

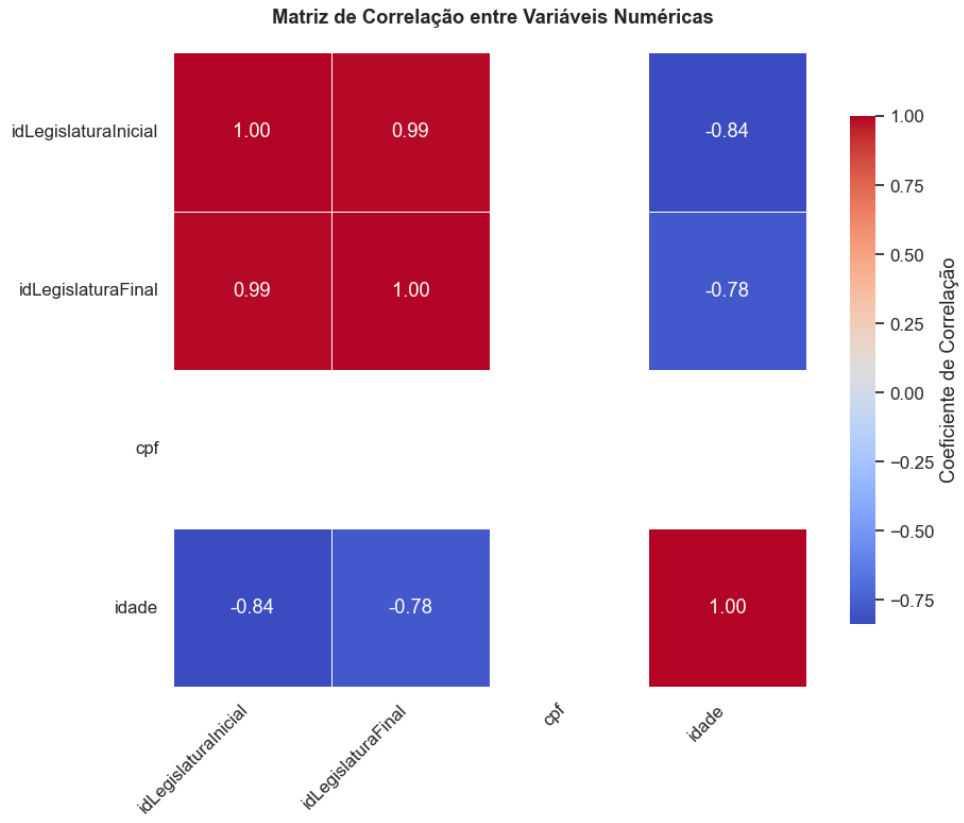
- **As capitais da região Norte estão menos representadas do que capitais de outras regiões?**

  Sim. As capitais do Norte aparecem de forma proporcionalmente menor. Fatores como menor acesso a estruturas partidárias, recursos educacionais e redes políticas podem explicar a quantidade reduzida de parlamentares nascidos nessas regiões.



- **Por que a idade tem forte correlação negativa com *idLegislaturaInicial* e *idLegislaturaFinal*?**

  Parlamentares mais idosos tendem a ter participado de legislaturas mais antigas. Como as legislaturas são sequenciais no tempo, quanto maior a idade do parlamentar, mais antigo tende a ser seu período de atuação registrado.



3.1.3 Limitações

Uma limitação que é possível citar é a falta de dados importantes, como a data de nascimento de alguns parlamentares e o município de nascimento, o que pode afetar durante a tentativa de visualização de dados referentes a essas colunas.