

MVP - Engenharia de dados

Engenharia de Dados para análise de Jogos de Tabuleiro

Objetivo:

Este projeto tem como objetivo explorar, analisar e aplicar princípios de engenharia de dados em um dataset contendo informações sobre avaliações de jogos de tabuleiro. A execução seguirá etapas essenciais, tais como: ingestão de dados, limpeza e transformação dos dados, modelagem, armazenamento e organização, disponibilização e, por fim, análise e elaboração de relatórios.

O desafio central consiste em transformar dados dispersos em conhecimento estruturado e extrair resultados e insights significativos por meio da resposta a questionamentos analíticos, de modo a identificar tendências, fatores de popularidade, padrões de comportamento e relações entre jogos e jogadores na comunidade.

Dataset

Coleta e Carga dos dados:

O conjunto de dados foi obtido na plataforma Kaggle

[fonte: <https://www.kaggle.com/datasets/andrewmvd/board-games>] e reúne informações extraídas do site BoardGameGeek (BGG), com informações detalhadas sobre jogos, avaliações, mecânicas, categorias e perfis de jogadores, uma das maiores comunidades online dedicadas a jogos de tabuleiro.

O dataset foi hospedado no GitHub, sendo o mesmo espelhado no Databricks para carregamento e leitura dos dados, conforme mostra o código no **Notebook-MVP-Eng-Dados-Tabuleiro**.



```

Just now (21s) 4: Coleta e Carga de Dados - Obtendo do workspace... Python ⚙️ ⚡ ⚪ ⚫ ⚬ ⚮
# Coleta e carga de dados

# URL do dataset
file_path = "/Workspace/Users/marciopug@gmail.com/mvp-engenharia-dados/bgg_dataset.csv"

# Ler direto com pandas
df_dados = pd.read_csv(file_path, delimiter=';')

# Converter para Spark DataFrame
df_tabuleiro = spark.createDataFrame(df_dados)

# Informações do dataset carregado
print("Dataset carregado com sucesso!")
print(f"Total de registros: {df_tabuleiro.count()}")
print(f"Total de colunas...: {len(df_tabuleiro.columns)}")
> See performance (1)
Optimize

df_dados: pandas.core.frame.DataFrame = [ID: float64, Name: object ... 12 more fields]
df_tabuleiro: pyspark.sql.connect.dataframe.DataFrame = [ID: double, Name: string ... 12 more fields]

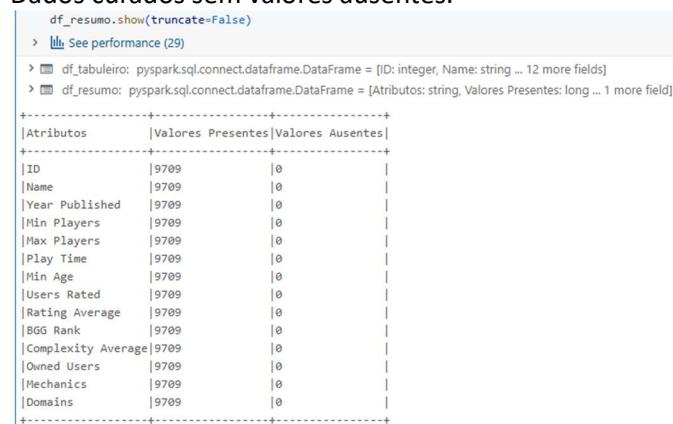
Dataset carregado com sucesso!
Total de registros: 20343
Total de colunas...: 14

```

As etapas de tratamentos de inconsistências estão descritas no código localizado no notebook:

Notebook-MVP-Eng-Dados-Tabuleiro.

Dados curados sem valores ausentes.

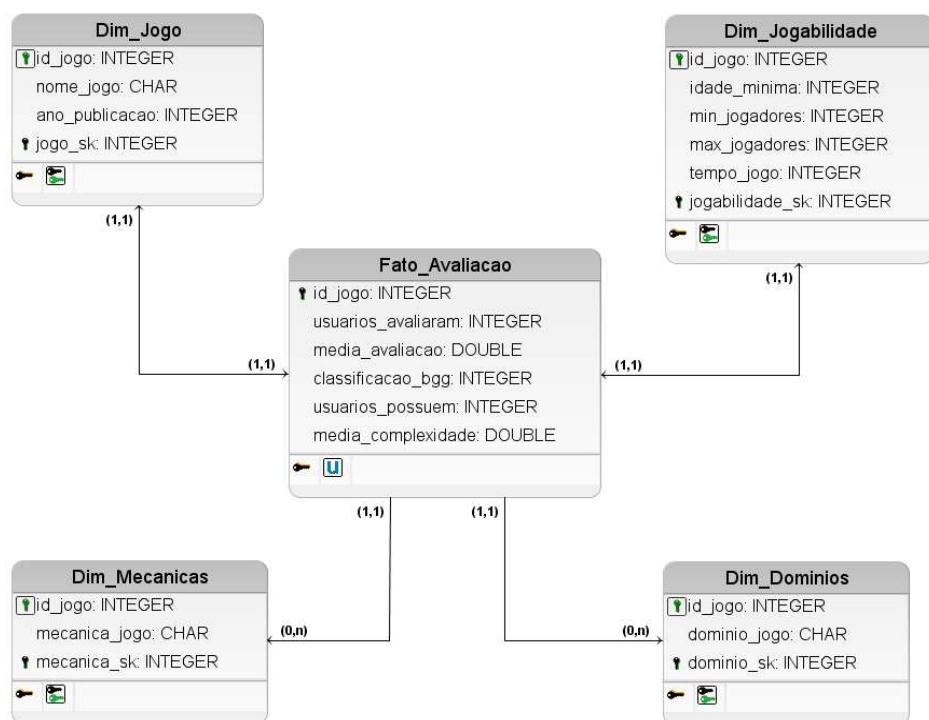
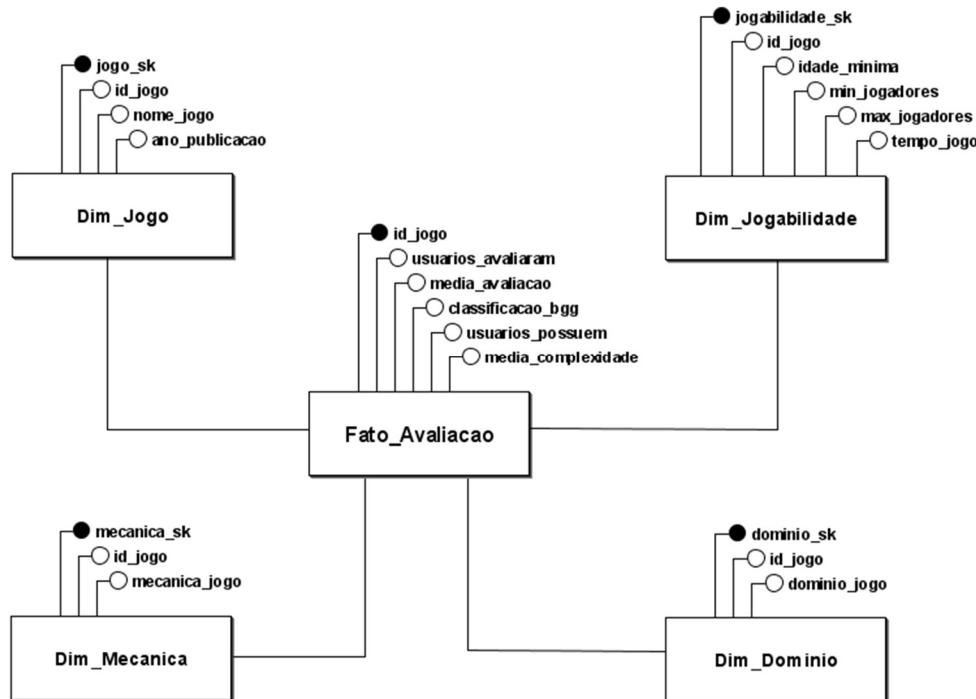


Atributos	Valores Presentes	Valores Ausentes
ID	9709	0
Name	9709	0
Year Published	9709	0
Min Players	9709	0
Max Players	9709	0
Play Time	9709	0
Min Age	9709	0
Users Rated	9709	0
Rating Average	9709	0
BGG Rank	9709	0
Complexity Average	9709	0
Owned Users	9709	0
Mechanics	9709	0
Domains	9709	0

Os atributos e seus respectivos dados foram modelados para alimentar o modelo estrela (tabela fato e tabelas dimensão) durante a criação das tabelas fato e dimensão, conforme descrito no código localizado no notebook: **Notebook-MVP-Eng-Dados-Tabuleiro**.

Modelagem

Através das informações do dataset, foi realizado a modelagem em estrela com tabelas com fato e dimensões. A seguir é mostrado o **modelo conceitual** e **lógico** do modelo estrela respectivamente.



Catálogo de Dados

The screenshot shows a data catalog interface with the following structure:

- Catalog**: Serverless Starter Warehouse Serverless 2XS
 - My organization
 - workspace (selected)
 - default
 - information_schema
 - tabuleiro (selected)
 - dim_dominios
 - dim_jogabilidade
 - dim_jogo
 - dim_mecanicas
 - fato_avaliacao

Catalog: workspace

Database ou Schema: tabuleiro

tabela dimensão:

dim_dominios; dim_jogabilidade; dim_jogo; dim_mecanicas

tabela fato: fato_avaliacao

Valores mínimo, máximo e valores únicos, antes dos atributos serem traduzidos.

```
Just now (<1s)
range_valores.show(truncate=False)
> See performance (1)

+-----+-----+-----+-----+
|Atributos |Mínimo|Máximo|Qtd Valores Únicos|
+-----+-----+-----+-----+
|ID |1 |322289|9709 |
|Name |- |- |9592 |
|Year Published |-3500 |2021 |162 |
|Min Players |0 |10 |10 |
|Max Players |0 |163 |36 |
|Play Time |0 |60000 |99 |
|Min Age |0 |21 |19 |
|Users Rated |30 |102214|2935 |
|Rating Average |1.43 |9.34 |522 |
|BGG Rank |1 |20344 |9709 |
|Complexity Average |0.0 |5.0 |377 |
|Owned Users |3 |155312|3821 |
|Mechanics |- |- |4589 |
|Domains |- |- |39 |
+-----+-----+-----+-----+
```

Descrição dos atributos:

tabuleiro				
TABELA	ATRIBUTOS	TIPO	VALORES	COMENTÁRIO
dim_jogo	id_jogo	int	1 a 3222289	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 3222289)
	nome_jogo	string	-	Nome completo e oficial do jogo de tabuleiro
	ano_publicacao	int	-3500 a 2021	Ano de publicação original do jogo (-3500 a 2021)
	jogo_sk	bigint	-	Chave substituta artificial para a dimensão jogo
dim_jogabilidade	id_jogo	int	1 a 3222289	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 3222289)
	min_jogadores	int	0 a 10	Número mínimo de jogadores necessários (0 a 10)
	max_jogadores	int	0 a 163	Número máximo de jogadores suportados (0 a 163)
	tempo_jogo	int	0 a 60000	Duração média da partida em minutos (0 a 60000)
	idade_minima	int	0 a 21	Idade mínima recomendada para jogar (0 a 21)
	jogabilidade_sk	bigint	-	Chave substituta artificial para a dimensão jogabilidade
dim_dominios	id_jogo	int	1 a 3222289	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 3222289)
	dominio_jogo	string	-	Categorias temáticas ou gêneros dos jogos.
	dominio_sk	bigint	-	Chave substituta artificial para a dimensão domínios
dim_mecanicas	id_jogo	int	1 a 3222289	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 3222289)
	mecanica_jogo	string	-	Regras e dinâmica de jogabilidade. Como o jogo funciona.
	mecanica_sk	bigint	-	Chave substituta artificial para a dimensão mecânicas
fato_avaliacao	id_jogo	int	1 a 3222289	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 3222289)
	usuarios_avaliaram	int	30 a 102214	Quantidade de usuários que avaliaram o jogo (30 a 102214)
	media_avaliacao	double	1.43 a 9.34	Nota média recebida (1.43 a 9.34)
	classificacao_bgg	int	1 a 20344	Posição no ranking da BoardGameGeek (1 a 20344)
	media_complexidade	double	0 a 5	Nível médio de dificuldade/complexidade (0 a 5)
	usuarios_posseum	int	3 a 155312	Quantidade de usuários que possuem o jogo (3 a 155312)

Catalog Explorer > workspace > tabuleiro >

dim_jogo

[Overview](#) [Sample Data](#) [Details](#) [Permissions](#) [Policies](#) [History](#) [Lineage](#) [Insights](#) [Quality](#)

Description

[AI generate](#) [Add](#)

Filter columns...

Column	Type	Comment	Tags	Column masking
id_jogo	int	ID original do jogo no dataset BoardGameGeek (BGG) (1 a 322289)	identificador	
nome_jogo	string	Nome completo e oficial do jogo de tabuleiro	descricao	
ano_publicacao	int	Ano de publicação original do jogo (-3500 a 2021)	temporal	
jogo_sk	bigint	Chave substituta artificial para a dimensão jogo	chave_substituta	

Catalog Explorer > workspace > tabuleiro >

dim_jogabilidade

[Overview](#) [Sample Data](#) [Details](#) [Permissions](#) [Policies](#) [History](#) [Lineage](#) [Insights](#) [Quality](#)

Description

[AI generate](#) [Add](#)

Filter columns...

Column	Type	Comment	Tags	Column masking
id_jogo	int	ID original do jogo no dataset BoardGameGeek (BGG)	identificador	
min_jogadores	int	Número mínimo de jogadores necessários (0 a 10)	jogabilidade_min	
max_jogadores	int	Número máximo de jogadores suportados (0 a 163)	jogabilidade_max	
tempo_jogo	int	Duração média da partida em minutos (0 a 60000)	duracao	
idade_minima	int	Idade mínima recomendada para jogar (0 a 21)	faixa_etaria	
jogabilidade_sk	bigint	Chave substituta artificial para a dimensão jogabilidade	chave_substituta	

Catalog Explorer > workspace > tabuleiro >

dim_dominios

[Overview](#) [Sample Data](#) [Details](#) [Permissions](#) [Policies](#) [History](#) [Lineage](#) [Insights](#) [Quality](#)

Description

[AI generate](#) [Add](#)

Filter columns...

Column	Type	Comment	Tags	Column masking
id_jogo	int	ID original do jogo no dataset BoardGameGeek (BGG)	identificador	
dominio_jogo	string	Categorias temáticas ou gêneros dos jogos. O que o jogo é (categoria/tema).	categoria	
dominio_sk	bigint	Chave substituta artificial para a dimensão domínios	chave_substituta	

Catalog Explorer > workspace > tabuleiro >

dim_mecanicas [Overview](#) [Sample Data](#) [Details](#) [Permissions](#) [Policies](#) [History](#) [Lineage](#) [Insights](#) [Quality](#)**Description** [AI generate](#) [Add](#) Filter columns...

Column	Type	Comment	Tags	Column masking
id_jogo	int	ID original do jogo no dataset BoardGameGeek (BGG)	identificador	
mecanica_jogo	string	Regras e dinâmica de jogabilidade. Como o jogo funciona (regras/dinâmica).	dinamica	
mecanica_sk	bigint	Chave substituta artificial para a dimensão mecânicas	chave_substituta	

Catalog Explorer > workspace > tabuleiro >

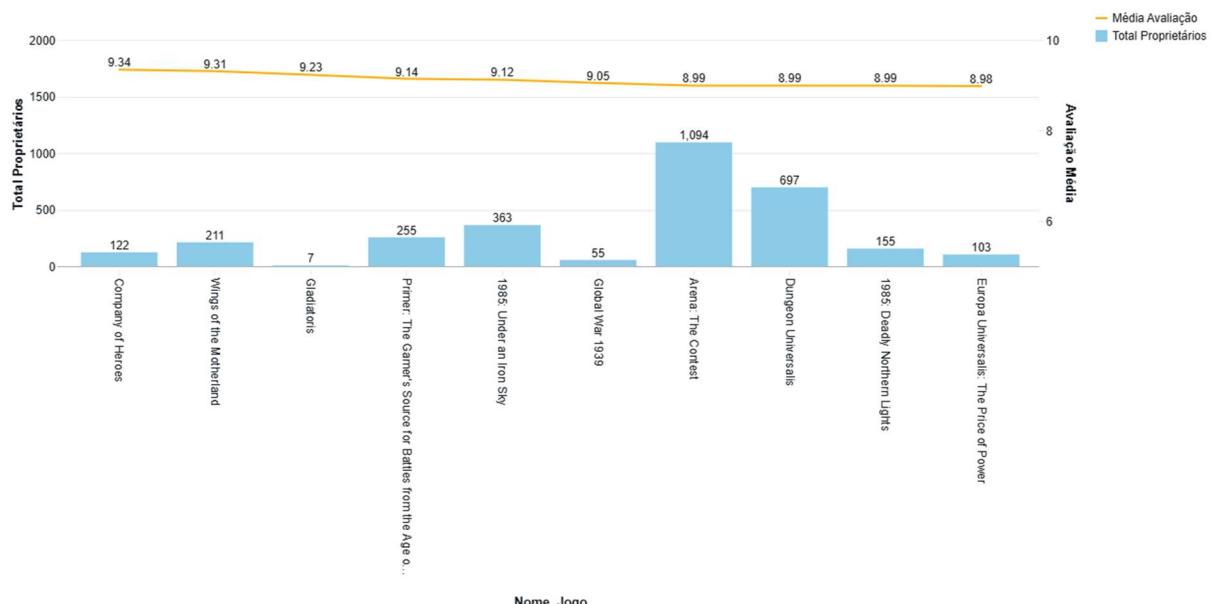
fato_avaliacao [Overview](#) [Sample Data](#) [Details](#) [Permissions](#) [Policies](#) [History](#) [Lineage](#) [Insights](#) [Quality](#)**Description** [AI generate](#) [Add](#) Filter columns...

Column	Type	Comment	Tags	Column masking
id_jogo	int	ID original do jogo no dataset BoardGameGeek (BGG)	identificador	
usuarios_avaliaram	int	Quantidade de usuários que avaliaram o jogo (30 a 102214)	engajamento	
media_avaliacao	double	Nota média recebida (1.43 a 9.34)	desempenho	
classificacao_bg	int	Posição no ranking da BoardGameGeek (1 a 20344)	ranking	
media_complexidade	double	Nível médio de dificuldade/complexidade (0 a 5)	nível	
usuarios_posuem	int	Quantidade de usuários que possuem o jogo (3 a 155312)	popularidade	

QUESTÕES ANALÍTICAS RESPONDIDAS.

1. Quais são os jogos (top 10) de tabuleiro mais bem avaliados e sua relação com a popularidade?

Table		Jogo vs Avaliação	Jogo vs Popularidade	Avaliação vs Popularidade	+
	Δ^B_c Nome_Jogo	Σ^2_3 Ano_Publicacao	1.2 Avaliacao_Media	Σ^2_3 Total_Avaliacoes	Σ^2_3 Total_Proprietarios
1	Company of Heroes	2020	9.34	47	122
2	Wings of the Motherland	2019	9.31	79	211
3	Gladiatoriis	2009	9.23	31	7
4	Primer: The Gamer's Source for Battles from the Age of Reas...	2013	9.14	58	255
5	1985: Under an Iron Sky	2018	9.12	90	363
6	Global War 1939	2011	9.05	34	55
7	Arena: The Contest	2019	8.99	600	1094
8	Dungeon Universalis	2019	8.99	454	697
9	1985: Deadly Northern Lights	2020	8.99	36	155
10	Europa Universalis: The Price of Power	2021	8.98	74	103

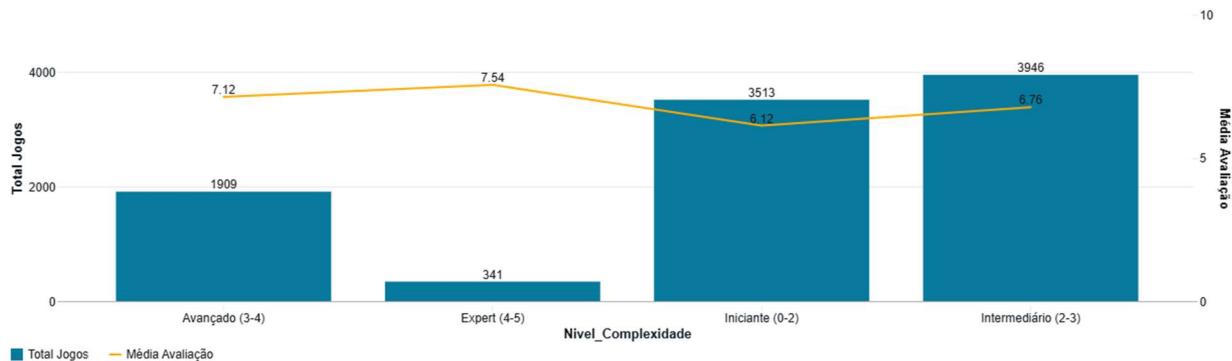


Comentário:

- Os jogos mais bem avaliados não são necessariamente os mais populares.
- A maioria dos jogos com notas altas é relativamente recente (2018–2021).
- Jogos com muitas avaliações tendem a ter notas ligeiramente menores, mas mais confiáveis estatisticamente.
- Jogos com poucas avaliações podem inflar notas, mas não representam consenso estatístico.
- Jogos com um volume maior de avaliações tendem a ter notas ligeiramente mais moderadas, porém estatisticamente mais confiáveis, refletindo um consenso mais amplo.

2. Como o nível de complexidade dos jogos influencia a avaliação média dos jogos?

Table		Complexidade vs Total Jogos	Complexidade vs Avaliação
	Δ^B_c Nivel_Complexidade	1.2 Media_Avaliacao	Σ^2_3 Total_Jogos
1	Iniciante (0-2)	6.12	3513
2	Intermediário (2-3)	6.76	3946
3	Avançado (3-4)	7.12	1909
4	Expert (4-5)	7.54	341



📌 Comentário:

- Jogos mais complexos (avançado e expert) tendem a receber avaliações mais altas. Há uma clara correlação positiva: quanto maior a complexidade, maior a média de avaliação.
- A maior parte dos jogos está concentrada nos níveis Iniciante e Intermediário (mais de 7.400 títulos).
- Jogos simples são abundantes, mas não tão bem avaliados.
- Jogos complexos são menos numerosos, mas recebem notas mais altas.

3. Quais as mecânicas de jogos que estão associadas às maiores avaliações e popularidade?

	Mecanica_Jogo	Mecânicas Jogo vs Avaliação		Mecânicas Jogo vs Popularidade	
		1.2 Avaliacao_Media	1.2 Popularidade	1.2 Total_Jogos	1.2 Popularidade
1	Legacy Game	7.87	280037	20	280037
2	Automatic Resource Grow...	7.82	228094	11	228094
3	Ownership	7.78	311224	27	311224
4	Delayed Purchase	7.77	258303	10	258303
5	Turn Order: Pass Order	7.75	176496	12	176496
6	Turn Order: Auction	7.74	146152	10	146152
7	Turn Order: Claim Action	7.7	474439	28	474439
8	Victory Points as a Resource	7.68	476511	36	476511
9	Command Cards	7.68	93200	10	93200
10	Market	7.66	322978	35	322978



📌 Comentário:

As mecânicas listadas têm sua avaliação próxima, sem grandes discrepâncias, o que indica homogeneidade na avaliação.

Não há mecânicas com notas muito baixas ou muito altas.

As mecânicas mais populares não são necessariamente as mais bem avaliadas.

Popularidade não garante avaliação máxima.

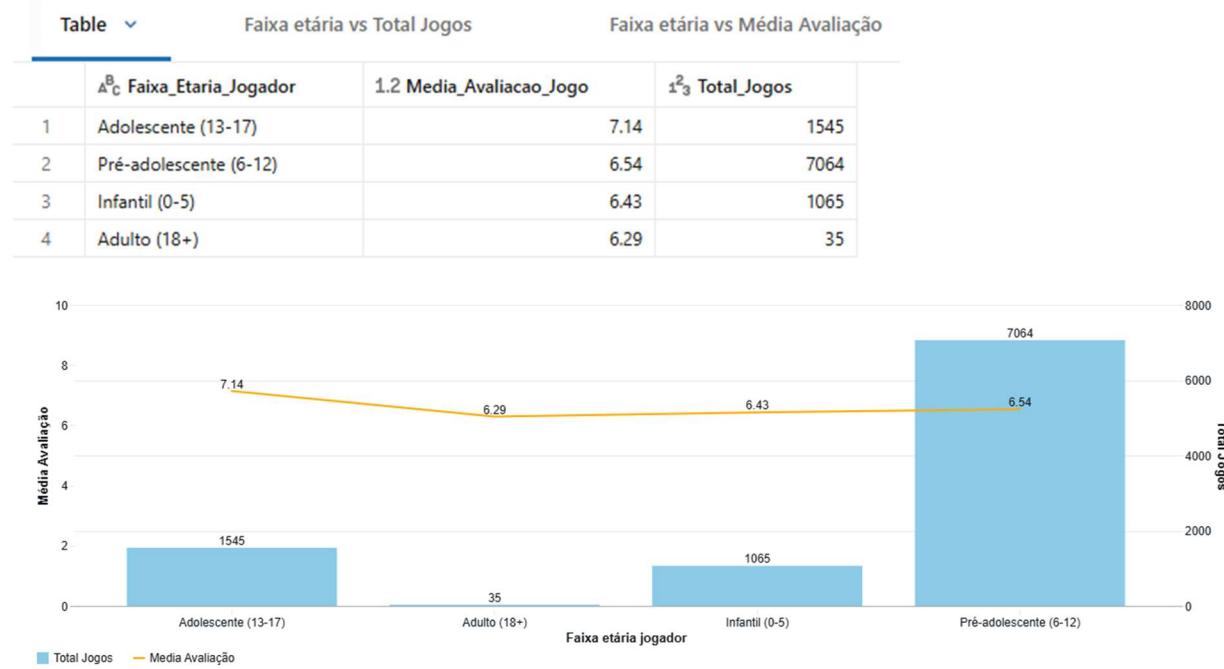
4. Qual a correlação entre a complexidade dos jogos e o tempo médio de uma partida?



Comentário:

Jogos com baixa e média complexidade são jogos rápidos, mas esses não alcançam notas tão altas. Têm acesso a uma grande oferta de jogos rápidos e simples, mas esses não alcançam notas tão altas. Jogos de muito alta complexidade são raros (apenas 341), mas recebem a melhor avaliação média. Jogadores tendem a valorizar mais jogos com alta complexidade, pode ser por ser mais desafiador e profundo, mesmo que exijam mais tempo.

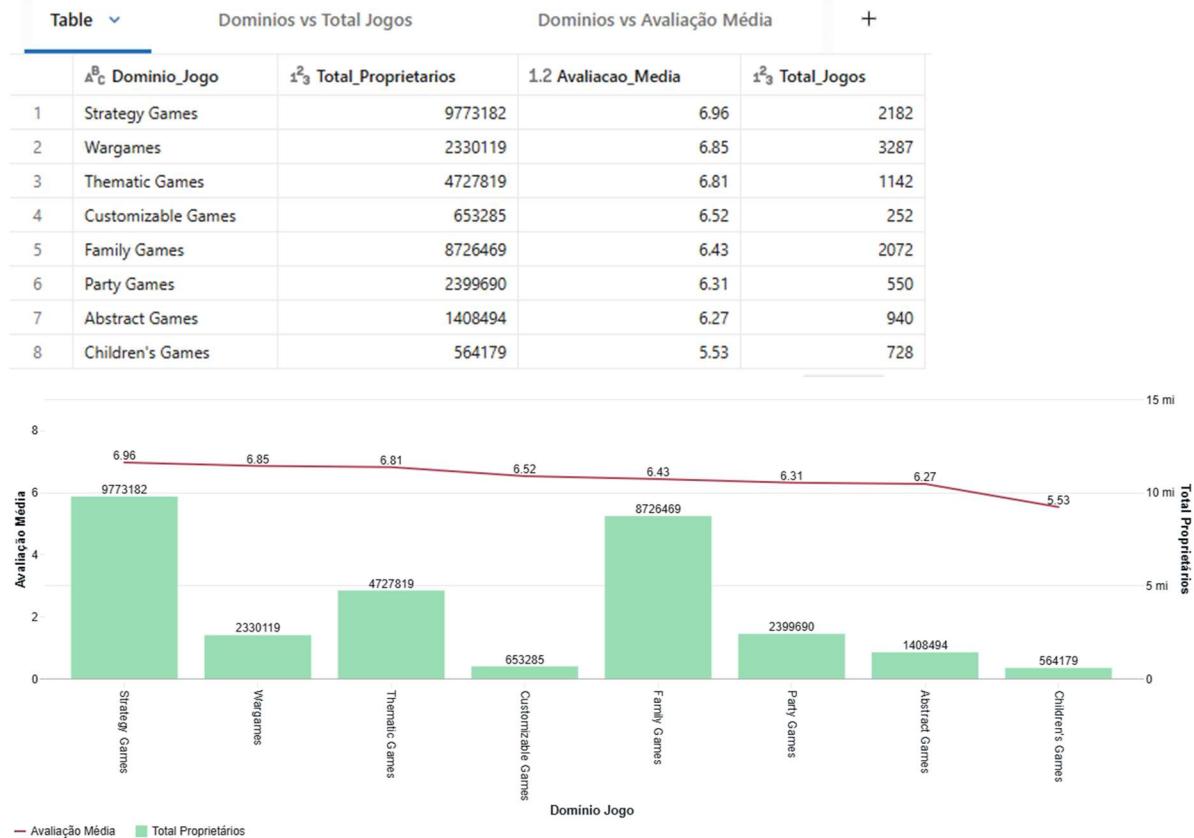
5. Como a faixa etária recomendada pelo jogo influencia a avaliação dos jogos?



Comentário:

Melhor avaliação entre adolescentes (13-17). Adolescentes são o público que mais valoriza os jogos disponíveis. Maior oferta de jogos para pré-adolescentes (6-12). Mostra que há pouquíssimos jogos para adultos (18+) e avaliação não é tão alta, embora seja a menor nota, ela está relativamente próxima das demais faixas.

6. Quais categorias (domínios) temáticas tem maior popularidade e mantêm alta avaliação?



💡 Comentário:

Strategy Games - tem alta popularidade (9,7 milhões de proprietários) e têm boa avaliação (6.96), equilibrando alcance e qualidade.

Family Games - também muito populares (8,7 milhões), mas com avaliação mediana (6.43), mostrando que volume não garante qualidade.

Wargames - maior número de títulos (3.287), avaliação sólida (6.85), mas público menos massivo.

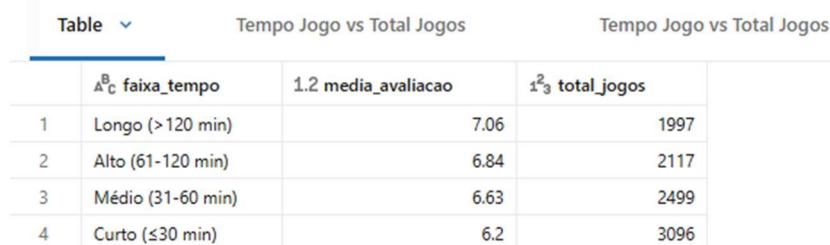
Thematic Games - boa avaliação (6.81) e popularidade intermediária, valorizados por públicos específicos.

Customizable Games - poucos títulos (252), mas comunidades fiéis (650 mil proprietários).

Party Games e Abstract Games - populares em contextos sociais ou conceituais, mas com notas abaixo de 6.5.

Children's Games - menor avaliação (5.53) e baixa popularidade, indicando espaço para inovação no segmento infantil.

7. Qual a relação entre tempo de jogo e satisfação dos jogadores?





📌 Comentário:

Jogos longos (>120 min) tendem a ter notas mais altas.

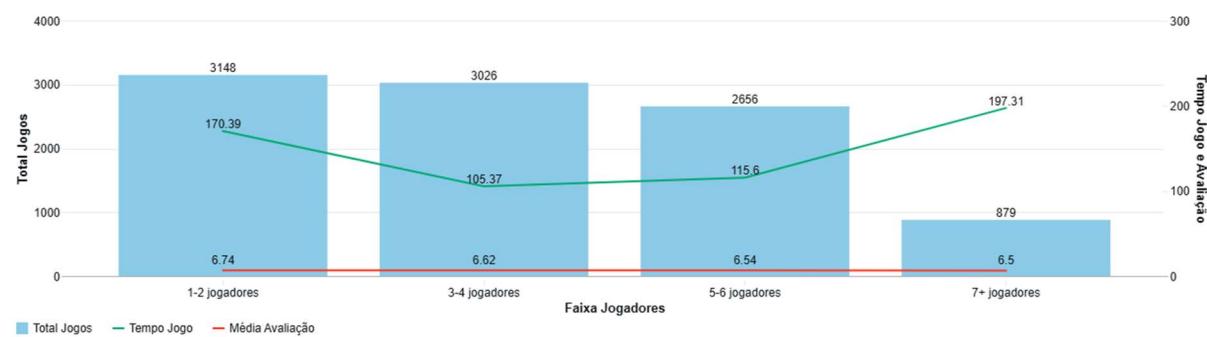
Jogos curtos (≤30 min) são abundantes, mas menos valorizados.

Satisfação cresce com maior investimento de tempo, sugerindo que jogadores associam duração a profundidade.

8. Qual é o perfil dos jogos (quantidade, avaliação e tempo) por faixa de jogadores?

Table ▾ Faixa Jogadores vs Avaliação vs Quantidade Jogos +

	A ^B C faixa_jogadores	1 ² 3 total_jogos	1.2 media_avaliacao	1.2 media_tempo_jogo
1	1-2 jogadores	3148	6.74	170.39
2	3-4 jogadores	3026	6.62	105.37
3	5-6 jogadores	2656	6.54	115.6
4	7+ jogadores	879	6.5	197.31



📌 Comentário:

Jogos para poucos jogadores são os mais bem avaliados.

Verificado que a concentram em faixas de até 4 jogadores, que somam mais de 6.000 títulos.

Grupos grandes (5+) têm menos opções, partidas longas e avaliações mais baixas.

🌟 Conclusão

Este projeto demonstra como a engenharia de dados pode ser aplicada em ambientes analíticos para transformar informações dispersas em conhecimento estruturado e açãoável.

Os resultados revelam que **popularidade não implica qualidade**: jogos amplamente jogados não são necessariamente os mais bem avaliados. Também foi possível identificar que **complexidade e duração** tendem a elevar as notas atribuídas pelos usuários, enquanto jogos curtos (≤ 30 minutos), apesar de numerosos, costumam receber avaliações mais modestas.

Além disso, surgem **lacunas relevantes para públicos específicos**, como adultos e grupos grandes, que contam com menor oferta de títulos e apresentam níveis inferiores de satisfação. As análises de categorias e mecânicas indicam a existência de nichos fiéis, mas mostram que isso **não garante**, por si só, avaliações consistentemente altas.

Em síntese, os achados reforçam que fatores como **complexidade, tempo de jogo e público-alvo** exercem influência significativa na percepção de qualidade dos jogos de tabuleiro, oferecendo insumos valiosos para decisões estratégicas.