Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Trabalho Prático 2 -
Introdução a Banco de Dados

Carolina Penido Barcellos
2024024054
carolinabarcellos@ufmg.br

Gabrielly Xavier dos Santos
2024023724
gabyxsantosufmg@gmail.com

Lívia Caroline Rodrigues Pereira
2024024003
liviacaroline456@gmail.com

Conteúdo

1 Introdução

O presente documento apresenta um relatório desenvolvido como material de entrega ao Trabalho Prático II da matéria Introdução a Banco de Dados, cursado pela Universidade Federal de Minas Gerais. O objetivo do trabalho envolve entender os pontos negativos e positivos dos bancos de dados públicos disponibilizados por entidades governamentais, buscando melhorar a qualidade dos dados selecionados e compreender as informações que podem ser extraídas do banco, a partir de uma análise exploratória. O presente documento apresenta as seguintes informações:

- Informações iniciais sobre os dados, com o processo realizado antes da passagem para o gerenciador do banco de dados.
- Processo de engenharia reversa, com recuperação do esquema conceitual e produção de informações úteis sobre o banco.
- Análise exploratória dos dados com questões descritivas, estatísticas e de correlação
- Análise crítica da organização de dados
- Conclusão do trabalho

Vale ressaltar que a coleta e análise de dados públicos é resguardada pela Lei de Acesso à Informação (Lei nº 12.527/2011), ou LAI, que assegura a transparência de informações mantidas pelo poder público ou de entidades, incentivando a participação social na fiscalização dos dados, sendo uma importante ferramenta democrática, que auxilia na realização de fiscalizações mais efetivas, fomentando o direito de cobrança, por parte da população, sobre políticas públicas governamentais nos mais diversos assuntos da sociedade. Todo o conteúdo de código citado pela presente documentação pode ser encontrado no repositório presente no caminho:

https://github.com/Livia-CRPereira/analise_exploratoria_cinema_br_2024/tree/main

2 Seleção dos dados

2.1 Escolha e problemas iniciais

A primeira decisão a ser tomada para o início do trabalho foi o tema de foco que, no caso do presente documento, foi selecionado como cinema: a ideia inicial gira em torno de entender como o cinema e sua bilheteria variam ao longo do período de um ano (selecionado como 2024) e cobrar políticas do governo que favoreçam a valorização e acesso à cultura por parte de todos, principalmente a cultura nacional. Assim, foram selecionados bases de dados disponibilizadas pela Agência Nacional de Cinema (ANCINE), encontradas na página https://www.gov.br/ancine/pt-br/oca/dados-abertos.

Após uma cuidadosa análise, foram escolhidos os dados:

- Relatório de bilheteria diária de obras informadas pelas distribuidoras (todos os relatórios mensais do ano de 2024), disponibilizado pela Ancine.
- Salas de exibição e complexos registrados pela Ancine.

A decisão de separar os dados de bilheteria dos 12 meses de 2024 e da tabela com informações sobre salas e complexos cinematográficos para posterior análise foi feita com o objetivo de identificar padrões de público e possíveis relações entre as diversas variáveis, entendendo também a estrutura e as características detalhadas das instalações cinematográficas no território nacional, bem como a performance individual e coletiva de cada sala e complexo ao longo do ano. Para além, a utilização da base de avaliações e de elenco do imdb sobre as obras permite entender como qualidade técnica e público trocam relações e como a presença de determinados profissionais da sétima arte contribuem para o sucesso do filme. Essa abordagem permitiria uma análise multifacetada, cruzando informações de infraestrutura e qualidade com o desempenho de bilheteria, para conclusões sobre o setor.

Todavia, ao tentar carregar as tabelas para o banco de dados, percebeu-se inconsistência entre os tipos que estavam sendo identificados pelo banco e os tipos reais das colunas das tabelas e, além disso, algumas codificações errôneas nas tabelas, que não estavam possibilitando o carregamento. Assim, com a ajuda do modelo de linguagem avançado da Google, Gemini, os dados foram inicialmente identificados de forma a conseguirem ser acessados pelo gerenciador. Após o processamento inicial (de maneira que todos os meses foram únicos em apenas uma tabela, com a utilização de recursos do SQL que "simulam um loop"), os dados foram então devidamente convertidos para os tipos inicialmente esperados. Assim, na etapa pré-analítica, os scripts a serem rodados, em ordem, são: criacao_tabelas.sql, insercao_tabelas.sql e conversao_tipos_tabelas.sql

2.2 Tipos de acesso aos dados

Os dados provenientes da Ancine foram disponibilizados em formato CSV (Comma-Separated Values), um tipo de arquivo tabular amplamente compatível com bancos de dados relacionais (SQL), como PostgreSQL ou MySQL. Esses bancos seguem uma estrutura rígida, com tabelas bem definidas, chaves primárias e integridade referencial, o que facilita a realização de consultas estruturadas (Structured Query Language), especialmente em contextos onde há necessidade de cruzamento de múltiplas tabelas.

2.3 Gerenciador de Banco de Dados

Para o gerenciamento e análise do banco, foi escolhido o sistema gerenciador de banco de dados Postgre, com a contribuição da interface gráfica DBeaver para facilitar os processos de análise.

3 Qualidade dos Dados

3.1 Análise Inicial do Banco de Dados

Como primeira ação após a escolha das tabelas a serem utilizadas no trabalho, foi realizada uma rápida vizualização dos dados, onde foi identificado que as bases não respeitam à normalização fortemente sugerida pelos profissionais da área. Assim sendo, já no início do processo de engenharia reversa para a criação do esquema conceitual, percebeu-se uma necessidade de reorganizar os dados, de maneira a tornar o banco mais organizado, normalizado e de fácil compreensão.

3.2 Reestruturação

Ao tentar reestruturar o esquema conceitual, a partir da realização de uma engenharia reversa, buscando o entendimento completo das colunas de cada uma das tabelas, identificamos inconsistências de normalização. Os dados brutos foram fornecidos em duas estruturas principais, que podemos caracterizar como tabelas "planas" ou não normalizadas: BILHETERIAFILMES2024 e SALASCOMPLEXOS. Nestas estruturas, um único registro continha informações de múltiplas entidades (filme, distribuidora, complexo, sala, etc.). Seguem as inconsistências encontradas:

- Redundância Massiva de Dados: Informações como titulo-original do filme, razao-social-distribuidora e o endereço completo do complexo eram repetidas para cada registro de bilheteria semanal, consumindo espaço e criando inconsistências.
- Anomalias de Atualização: Para alterar um dado simples, como a razão social de uma distribuidora, seria necessário varrer e atualizar milhares de registros na tabela de bilheteria, um processo ineficiente e propenso a falhas.
- Anomalias de Inserção: Era impossível cadastrar uma nova sala ou um novo filme que ainda não tivesse um registro de bilheteria associado. A entidade "sala" não podia existir sem um fato de "bilheteria".
- Anomalias de Exclusão: Ao apagar o último registro de bilheteria de um determinado filme, todas as informações sobre aquele filme (seu CPB/ROE, título, etc.) poderiam ser permanentemente perdidas da base de dados.
-

Assim, apresentando atributos multivalorados (explicita e implicitamente), os valores não estavam nem na Primeira Forma Normal, exigindo um tratamento efetivo e divisão correta dos valores.

A fim de se identificar as entidades (futuras tabelas) e divisão de atributos correta, foi criado um esquema conceitual no modelo Entidade-Relacionamento, que correspondesse corretamente à lógica do contexto. As perguntas feitas durante a criação desse esquema e as conclusões chegadas podem ser encontradas no sql $perguntas - conceitual$, que apresenta consultas feitas a fim de se identificar relacionamentos corretos a serem modulados. O resultado da reestruturação dos dados pode ser obtida pelos comandos em $normalizao.sql$. Importante pontuar alguns aspectos identificados durante o processo:

- Serão criadas as entidades: Distribuidora, Protocolo, Seção, Filme, Sala, Complexo, Exibidor e Grupo Exibidor.
- Uma sala precisa de um complexo, assim como um complexo precisa de um exibidor, porém um exibidor pode não fazer parte de um grupo exibidor.
- Exibidores que não têm grupo complexo associado tem o nome-grupo-complexo na tabela SalasComplexos dado como 'NÃO PERTENCE A NENHUM GRUPO EXIBIDOR'.
- A tabela bilheteriafilmes2024 tem as colunas registro-exibidor e registro-grupo-exibidor e a tabela SalasComplexos tem as colunas registro-exibidor e nome-grupo-exibidor, não possuindo registro-grupo-exibidor. A criação da tabela Grupo Exibidor não poderá, assim, ser feita de forma direta. Será necessário fazer uma conexão pelo registro-exibidor para saber o registro do grupo.
- 6376 linhas da tabela bilheteriafilmes2024 não têm registro-sala, o que tornaria complexa a criação de linhas para 280 salas. Como, mesmo tirando essas linhas, mantemos 99,7% dos dados de bilheteria, como escolha de negócio vamos excluir esses dados.
- Cada seção (cada linha de bilheteriafilmes2024) não possui uma chave primária clara. Como decisão de negócio, vamos criar um id para cada seção, para garantir unicidade.

3.3 Esquema Conceitual

O esquema conceitual foi então montado de maneira a garantir a normalização dos dados e estruturação correta do sistema.

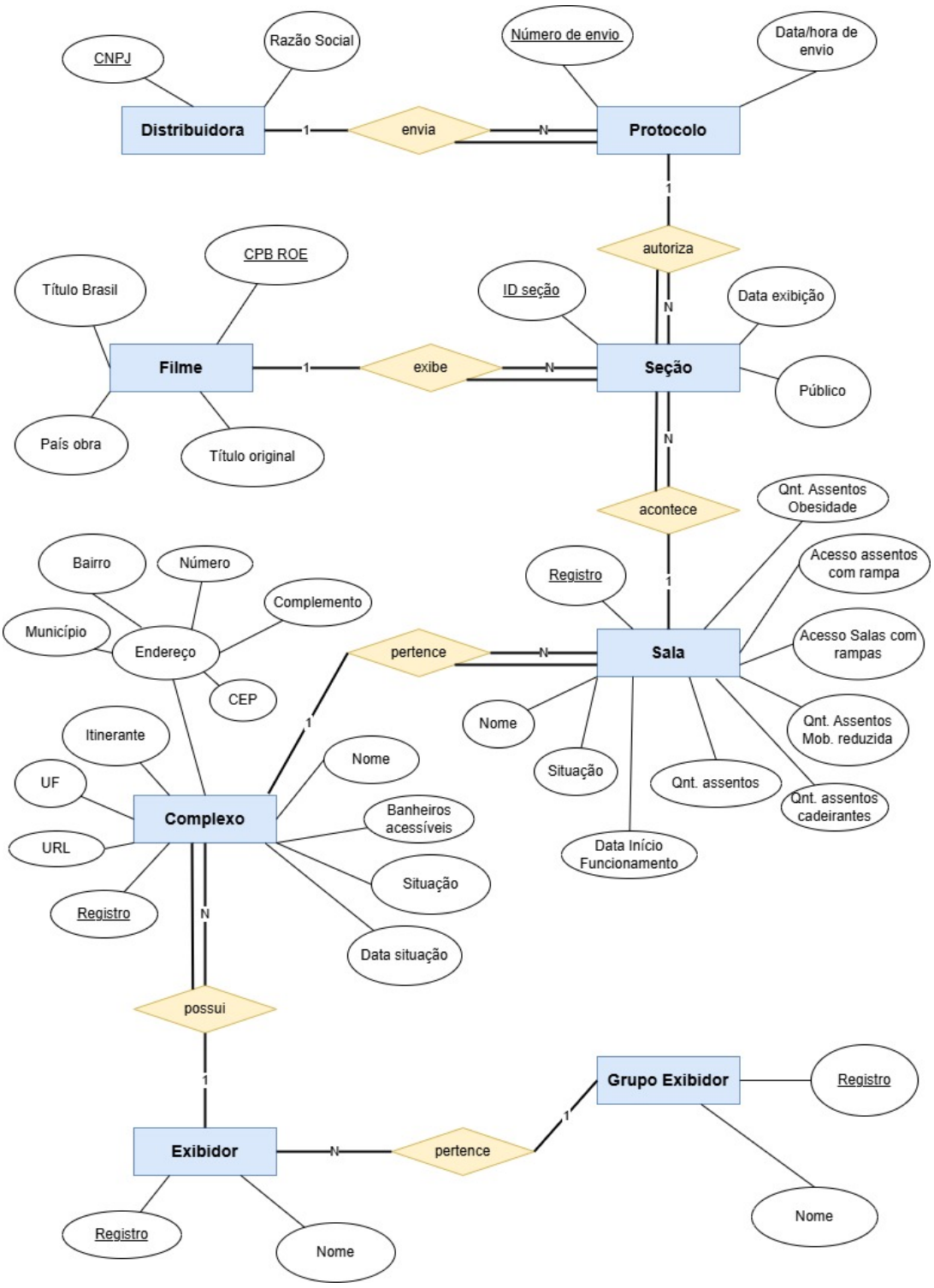

Figura 1: Modelo Conceitual Entidade-Relacionamento

3.4 Dicionário de Dados

O dicionário de dados será modulado de acordo com o modelo conceitual criado. A estrutura final do banco é essa, mantendo os valores nulos onde originalmente estão e apenas reformulando a estrutura para permitir uma análise efetiva.

Relação	Atributo	Tipo	Nulo?	Único?	Valores Possíveis	Restrições
Filme	cpb_roe	VARCHAR (20)	N	S	Código de registro da obra	Chave Primária (PK)
	titulo_original	VARCHAR (255)	N	N	Texto	-
	titulo_brasil	VARCHAR (255)	S	N	Texto	-
	pais_obra	VARCHAR (50)	S	N	Texto	-
Distribuidora	cnpj_distribuidora	VARCHAR (18)	N	S	Formato XX.XXX.XXX /XXXX-XX	Chave Primária (PK)
	razao_social _distribuidora	VARCHAR (150)	N	N	Texto	-
Protocolo	nr_protocolo_envio	BIGINT	N	S	Número inteiro sequencial	Chave Primária (PK)
	data_hora_envio _protocolo	TIMESTAMP	N	N	Data e hora no formato AAAA-MM-DD HH:MI:SS	-
	cnpj_distribuidora	VARCHAR (18)	N	N	CNPJ de uma distribuidora existente	Chave Estrangeira (FK) para Distribuidora
Seção	id_secao	BIGINT	N	S	Número inteiro sequencial	Chave Primária (PK), Auto-Incremento
	data_exibicao	DATE	N	N	Formato AAAA-MM-DD	-
	publico	INTEGER	N	N	Número de espectadores (inteiro)	-
	cpb_roe	VARCHAR (20)	N	N	CPB/ROE de um filme existente	Chave Estrangeira (FK) para Filme
	registro_sala	INTEGER	N	N	ID de uma sala existente	Chave Estrangeira (FK) para Sala
	nr_protocolo_envio	BIGINT	N	N	ID de um protocolo existente	Chave Estrangeira (FK) para Protocolo
Grupo Exibidor	registro_grupo _exibidor	INTEGER	N	S	Número inteiro sequencial	Chave Primária (PK)
	nome_grupo _exibidor	VARCHAR (150)	N	N	Texto	-
Exibidor	registro_exibidor	INTEGER	N	S	Número inteiro sequencial	Chave Primária (PK)
	nome_exibidor	VARCHAR (150)	N	N	Texto	-
	cnpj_exibidor	VARCHAR (18)	N	S	Formato XX.XXX.XXX /XXXX-XX	-
	registro_grupo _exibidor	INTEGER	S	N	ID de um grupo existente	Chave Estrangeira (FK) para Grupo Exibidor

Complexo	registro_complexo	INTEGER	N	S	Número inteiro sequencial	Chave Primária (PK)
	nome_complexo	VARCHAR (150)	N	N	Texto	-
	banheiros_acessiveis	VARCHAR (3)	S	N	'SIM', 'NÃO'	-
	situacao_complexo	VARCHAR (50)	S	N	'Em funcionamento', 'Fechado'	-
	data_situaca _complexo	DATE	S	N	Formato AAAA-MM-DD	-
	pagina_eletronica _complexo	VARCHAR (255)	S	N	URL de um website	-
	endereco_complexo	VARCHAR (255)	S	N	Texto	-
	numero_endereco _complexo	VARCHAR (50)	S	N	Texto	-
	complemento _complexo	VARCHAR (100)	S	N	Texto	-
	bairro_complexo	VARCHAR (100)	S	N	Texto	-
	municipio_complexo	VARCHAR (100)	S	N	Texto	-
	cep_complexo	VARCHAR (10)	S	N	Formato XX.XXX-XXX	-
	uf_complexo	VARCHAR (2)	S	N	Sigla de 2 letras de um estado	-
	complexo_itinerante	VARCHAR (3)	S	N	'SIM', 'NÃO'	-
	registro_exibidor	INTEGER	N	N	ID de um exibidor existente	Chave Estrangeira (FK) para Exibidor
Sala	registro_sala	INTEGER	N	S	Número inteiro sequencial	Chave Primária (PK)
	nome_sala	VARCHAR (150)	N	N	Texto	-
	situacao_sala	VARCHAR (50)	S	N	'Em funcionamento', 'Fechada'	-
	data_inicio _funcionamento	DATE	S	N	Formato AAAA-MM-DD	-
	assentos_sala	INTEGER	S	N	Número inteiro positivo	-
	assentos_cadeirantes	INTEGER	S	N	Número inteiro positivo	-
	assentos_mobilidade _reduzida	INTEGER	S	N	Número inteiro positivo	-
	assentos_obesidade	INTEGER	S	N	Número inteiro positivo	-
	acesso_assentos _com _rampa	VARCHAR (3)	S	N	'Sim', 'Não'	-
	acesso_sala _com _rampa	VARCHAR (3)	S	N	'Sim', 'Não'	-
	registro_complexo	INTEGER	N	N	ID de um complexo existente	Chave Estrangeira (FK) para Complexo

4 Metadados

Os dados a serem utilizados na análise foram obtidos pelo site de dados abertos do Gov, na aba da Agência Nacional do Cinema, ou ANCINE.

4.1 Bilheteria diária de obras informadas pelas distribuidoras

Fonte: https://dados.gov.br/dados/conjuntos-dados/relatorio-de-bilheteria-diaria-de-obras-informadas-pelas-distribuidoras

- **Data de obtenção:** 29/05/2025

- **Orgão produtor:** Agência Nacional do Cinema
- **Área técnica responsável:** Coordenação de Gestão das Informações Regulatórias
- **Data de referência (última atualização):** 26/05/2025
- **Atualização periódica:** Semanal
- **Limitações registradas:** Os dados brutos de bilheteria continham registros com o campo registro_sala nulo. Como a bilheteria não podia ser atribuída a um local específico, esses registros (correspondendo a menos de 1% do total) foram excluídos da base de dados final para garantir a integridade dos relacionamentos. Os dados de todos os anos foram reduzidos a aoenas 2024, unindo todas as tabelas dos meses em apenas 1. Na tabela original, não existe chave explícita para cada seção. Foi criada uma chave como índice (na ordem que aparece nas tabelas originais).
- **Cobertura:** Território da República Federativa do Brasil
- **Tipos de documentos disponíveis:** CSV; JSON; ODT; XML;

4.2 Salas de Exibição e Complexos Registrados na Ancine

Fonte: https://dados.gov.br/dados/conjuntos-dados/salas-de-exibicao-e-complexos-registrados-na-ancine

- **Data de obtenção:** 29/05/2025
- **Orgão produtor:** Agência Nacional do Cinema
- **Área técnica responsável:** Coordenação de Gestão das Informações Regulatórias
- **Data de referência (última atualização):** 01/05/2025
- **Atualização periódica:** Mensal
- **Limitações registradas:** Para o atributo registro_grupo_exibidor, a ausência de um grupo era representada pelo texto 'NÃO PERTENCE A NENHUM GRUPO EXIBIDOR' em vez de um valor nulo. Foi necessário um tratamento para converter essa string em NULL, padronizando a representação de ausência de valor. As colunas registro_grupo_exibidor e nome_grupo_exibidor estão em tabelas diferentes, de forma que foi necessário uní-las por um terceiro atributo, registro_exibidor.
- **Cobertura:** Território da República Federativa do Brasil
- **Tipos de documentos disponíveis:** CSV; JSON; ODT; XML;

5 Análise Crítica e Exploratória

5.1 Objetivos

O trabalho tem como objetivo realizar uma análise crítica e descritiva da qualidade das bases utilizadas, identificando problemas como valores nulos, inconsistências, desatualização e preenchimento esparso. Por fim, propõe-se a integrar os dados, analisando suas relações e identificando possíveis correlações que possam enriquecer a compreensão do tema analisado e estudado.

Objetivos específicos:

- Identificar a quantidade de nulos incoerentes em cada tabela.
- Calcular estatísticas básicas
- O quanto a acessibilidade faz diferença no público em um cinema.
- Possível existência de outliers na correlação de tamanho da sala e público.
- Sazonalidade do público.
- Correlação entre a quantidade de filmes lançados pela distribuidora e desempenho da distribuidora no ano.

5.1.1 Quantidade de Nulos

Visando observar o quão completas as nossas tabelas estão, elaboramos consultas em SQL e computamos os nulos ou valores faltantes em cada tabela:

- COMPLEXO:

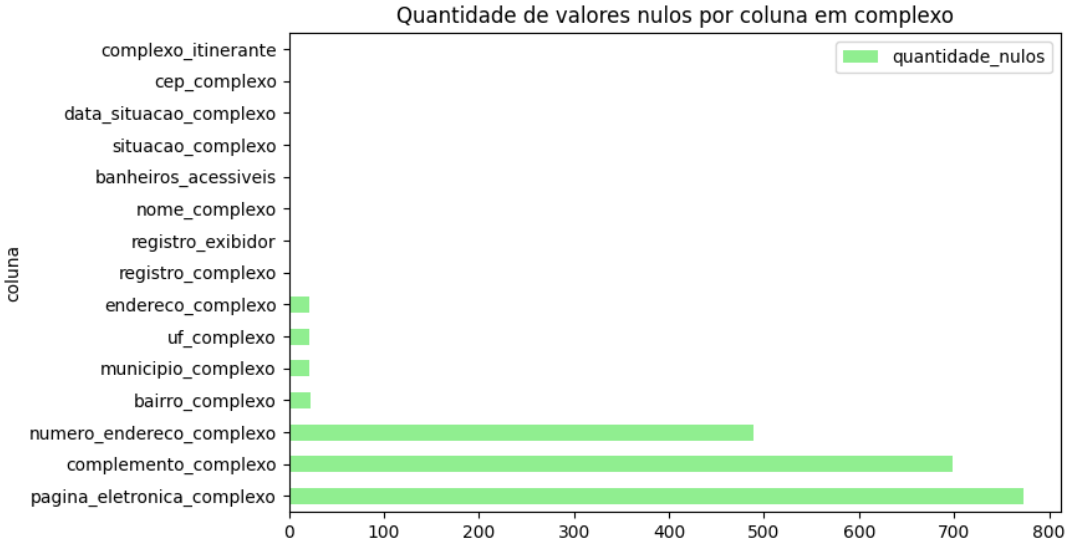

Figura 2: Quantidade de nulos em Complexo

Através dessa análise, foi observado que a coluna com maior quantidade de nulos foi a de pagina eletrônica, e isso nos instigou a olhar como os dados da coluna estavam. Diante disso, por meio de uma análise visual, notamos que também haviam campos imcompletos, os quais estavam apenas com HTTP://, apontando para mais uma incompletude da tabela. Ao consultar a quantidade de linhas incompletas, encontramos 193. Portanto, mais da metade das linhas da tabela de Complexo está sem a informação completa da página eletrônica.

- SALA:

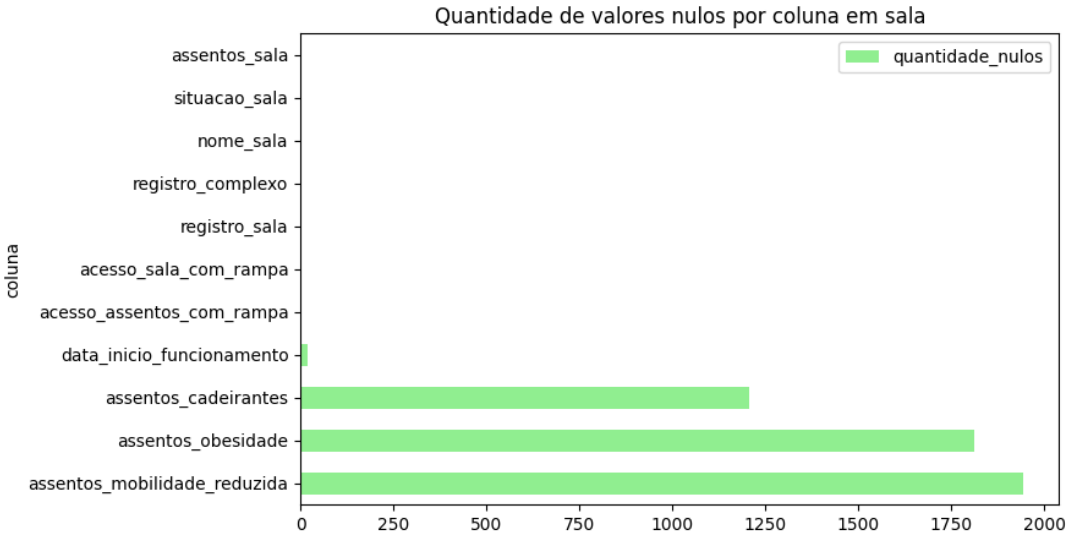

Figura 3: Quantidade de nulos em Sala

- EXIBIDOR:

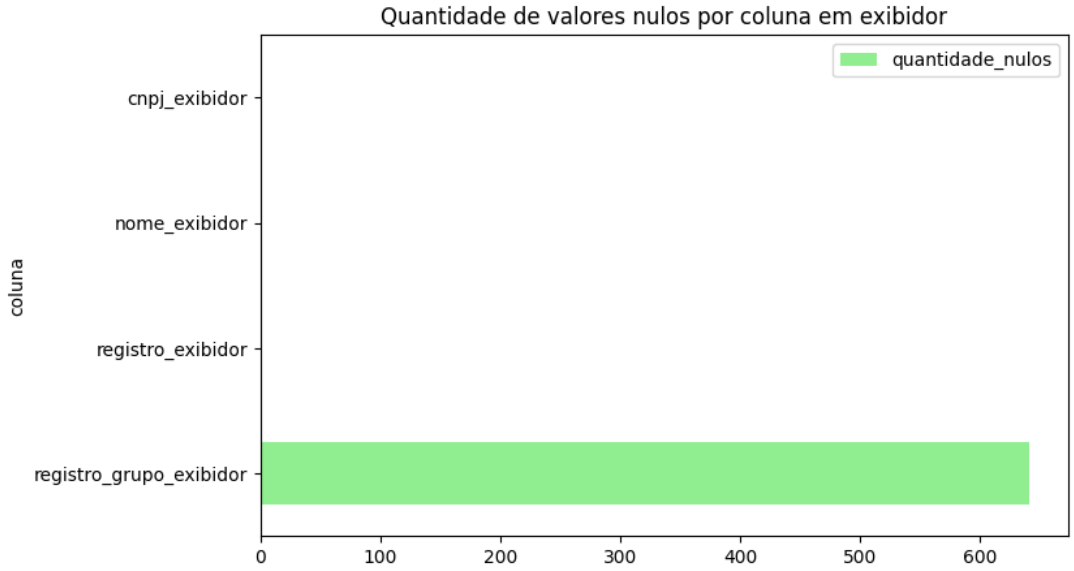

Figura 4: Quantidade de nulos em Exibidor

- FILME:

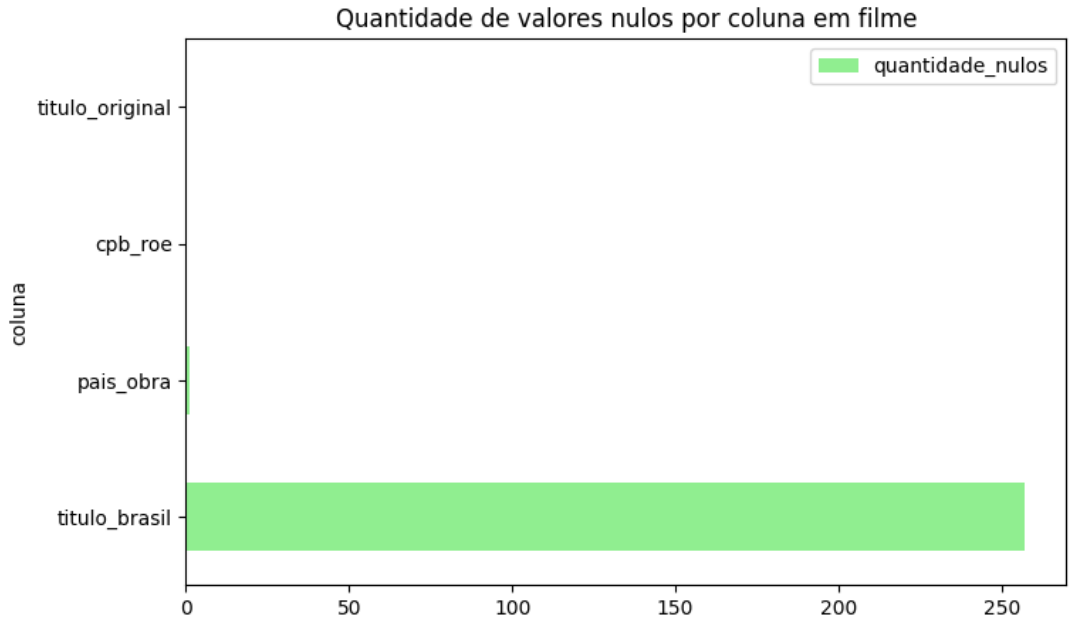

Figura 5: Quantidade de nulos em Filme

É notório que a única coluna com nulos discrepantes é a de Titulo Brasil, o que foi um ponto interessante. Analisando, notamos que tal fato ocorreu, pois todos os filmes brasileiros, ou seja, produzidos em nosso país, tem o nome citado apenas na coluna titulo original, ou seja, esses nulos na coluna titulo brasil dizem respeito a filmes brasileiros.

As demais tabelas não apresentam nenhum valor nulo, uma vez que ao criarmos adicionamos restrições que garantiram que isso não ocorreria.

5.1.2 Acessibilidade

Tendo em vista que obtemos dados que descrevem o quanto uma sala de cinema está preparada para receber e acomodar pessoas com algum tipo de deficiência, ou até dificuldade, nossa ideia foi analisar as condições do cinema no Brasil.

Ao realizar o JOIN das tabelas Sala, Complexo e Secao, conseguimos observar que, de todos os cinemas e salas que temos em todo o Brasil, pouquissímos possuem efetiva acessiblidade para a população.

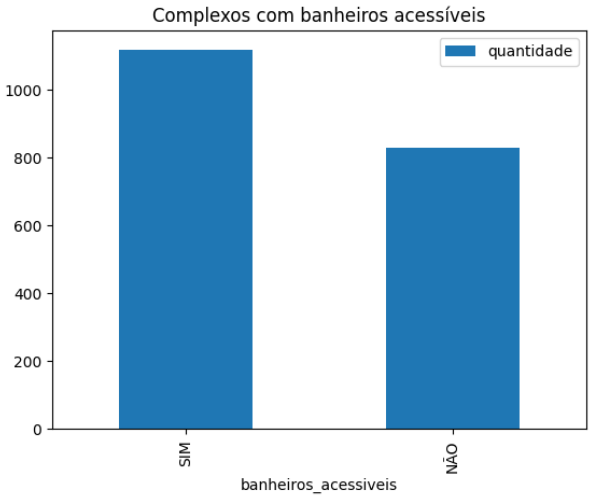

Figura 6: Complexos com banheiros acessíveis

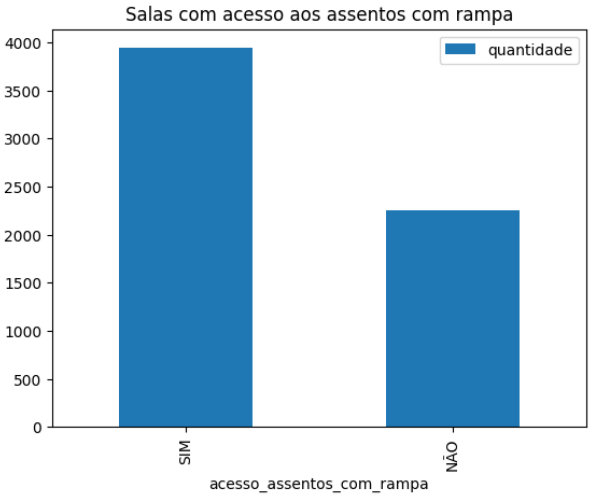

Figura 7: Salas com acesso aos assentos com rampa

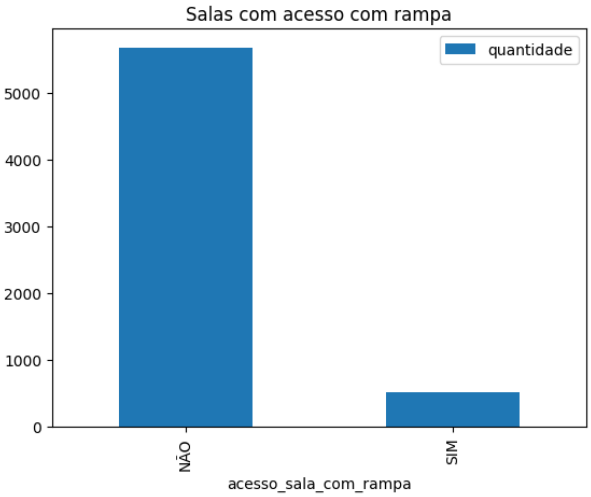

Figura 8: Salas com acesso com rampa

A partir dos três gráficos acima, observa-se que muitos complexos ainda não possuem banheiros acessíveis, assim como há salas sem acesso por rampa e sem acesso aos assentos com rampa. Também chama a atenção que há mais salas com acesso aos assentos por rampa do que salas com acesso por rampa. Pode ser que o critério para considerar uma sala com "acesso por rampa" seja mais restrito (por exemplo, entrada principal da sala), enquanto o "acesso aos assentos com rampa" contabilize rampas internas ou acessos específicos dentro da própria sala.

Assentos Obesidade X Quantidade de Salas

	assentos_obesidade	quantidade
0		1812
1	1	1152
2	2	1066
3	0	636
4	3	530
5	4	485
6	5	112
7	6	109
8	8	50
9	10	42

Figura 9: Gráfico Assentos Obesidade X Quantidade de Salas

Assentos Cadeirantes X Quantidade de salas

	assentos_cadeirantes	quantidade
0	4	1528
1	2	1219
2		1206
3	3	593
4	6	509
5	5	387
6	1	299
7	8	122
8	7	120
9	0	106

Figura 10: Gráfico Assentos Cadeirantes X Quantidade de Salas

Assentos Obesidade X Quantidade de Salas

	assentos_obesidade	quantidade
0		1812
1	1	1152
2	2	1066
3	0	636
4	3	530
5	4	485
6	5	112
7	6	109
8	8	50
9	10	42

Figura 11: Gráfico Assentos Mobilidade Reduzida X Quantidade de Salas

Os gráficos 9,10 e 11 mostram que a grande maioria das salas possui poucos assentos destinados a pessoas cadeirantes, com obesidade ou com mobilidade reduzida. Além disso, observa-se um número significativo de salas sem essa informação registrada (valores nulos).

Diante dessas apurações, é interessante pensar, quantos complexos contêm acessibilidade completa (dentro dos nossos paramêtros):

COMPLEXOS x ACESSIBILIDADE	
Total de complexos	2.087.293
Complexos com acessibilidade completa	10.533
Percentual de complexos com acessibilidade completa	0,50%

Figura 12: Tabela Complexos X Acessibilidade

Tal realidade trás à tona a seguinte questão, ainda que sejam poucos cinemas acessíveis, será que eles estao bem distribuídos no território brasileiro?

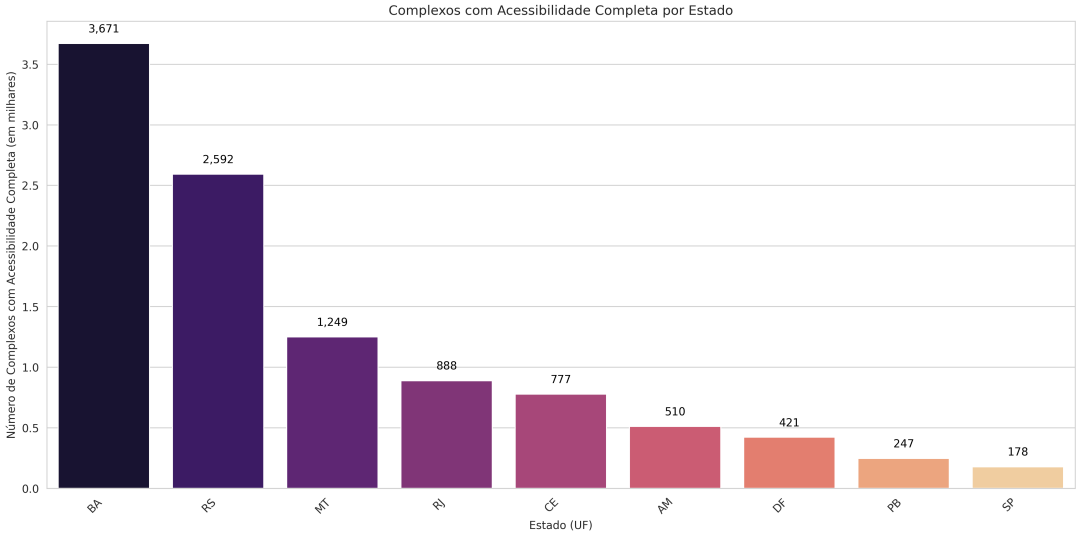

Figura 13: Gráfico Complexos acessívei por UF

As consultas feitas provam que não, afinal, dos 26 estados brasileiros, apenas 9 deles contêm certa quantidade de complexos que acomôdam uma gama maior de indíduos, com destaque para a Bahia.

A análise aprofundada dos dados de acessibilidade em complexos de cinema no Brasil revela um cenário desafiador e uma evidente lacuna na inclusão para pessoas com deficiência e mobilidade reduzida. Em suma, os dados revelam que o direito ao lazer e à cultura em salas de cinema ainda é um privilégio para poucos no Brasil. A ausência de acessibilidade universal na vasta maioria dos complexos, somada à sua concentração em poucas UFs, exige ações urgentes e coordenadas para promover um ambiente mais inclusivo e acessível para todos os cidadãos, independentemente de suas necessidades.

5.1.3 Existência de Outliers

Identificamos a existência dos seguintes outliers:

- **Outlier - Quantidade de Assentos por Sala**

Outliers - Quantidade de Assentos por Sala

	nome_sala	assentos_sala
0	SUPER CINE ESPAÇO DAS AMÉRICAS DRIVE IN	3000
1	CINE DRIVEIN	2000
2	SÃO LEO BOURBON 01	1693
3	CINE 9 DE ABRIL	1535
4	SÃO LUIZ	1219
5	CENTRO CULTURAL SESC LUIZ SEVERIANO RIBEIRO	1200
6	SÃO LUIZ	1181
7	CINE ALVORADA	1180
8	CINETEATRO SÃO LUIZ	1050
9	IPIRANGA I	1030
10	CINE MARROCOS	1012
11	ARCOIRIS MARROCOS	1012

Figura 14: Outlier - Quantidade de Assentos por Sala

A partir da tabela apresentada acima, observamos que algumas salas possuem uma quantidade de assentos significativamente superior à média, sendo, portanto, classificadas como outliers. Um exemplo é a primeira sala, que conta com 3.000 assentos, por se tratar de um cinema drive-in que funcionou temporariamente na cidade de São Paulo.

- **Outlier - Quantidade de Complexos por UF**

Outliers - Quantidade de Complexos por UF

	uf_complexo	count
0	SP	529
1	RJ	216
2	MG	188

Figura 15: Outlier - Quantidade de Complexos por UF

Pode-se afirmar que, conforme esperado, os estados mais populosos e economicamente desenvolvidos do país — São Paulo, Rio de Janeiro e Minas Gerais — concentram um número significativamente maior de complexos, sendo, portanto, identificados como outliers na distribuição.

- **Outlier — Público Total por Filme**

A partir dos dados apresentados na tabela abaixo, observa-se que o lançamento de alguns filmes registrou um público muito acima da média, que foi de 186.736 pessoas. Entre esses destaques estão Divertidamente 2, Meu Malvado Favorito 4 e Moana 2, títulos que se tornaram extremamente populares em 2024.

Outliers - Público Total por Filme

	titulo_original	publico_total
0	INSIDE OUT 2	22438049
1	DESPICABLE ME 4	7919873
2	MOANA 2	7916241
3	DEADPOOL & WOLVERINE	7450522
4	AINDA ESTOU AQUI	3043422
5	THE FORGE	3001175
6	IT ENDS WITH US	2954050
7	KINGDOM OF THE PLANET OF THE APES	2912326
8	VENOM: THE LAST DANCE	2603493
9	AQUAMAN AND THE LOST KINGDOM	2493653
10	MUFASA: THE LION KING	2356833
11	KUNG FU PANDA 4	2144705
12	GODZILLA X KONG: THE NEW EMPIRE	2140607
13	GLADIATOR II	2114898
14	JOKER: FOLIE À DEUX	2114355

Figura 16: Outlier - Público Total por Filme

- **Outlier - Público Total por Distribuidora**

Outliers -Público Total por Distribuidora

	razao_social_distribuidora	publico_total
0	THE WALT DISNEY COMPANY (BRASIL) LTDA.	48762520
1	WARNER BROS. (SOUTH) INC.	30385038
2	COLUMBIA TRISTAR FILMES DO BRASIL LTDA	17685261
3	PARAMOUNT PICTURES BRASIL DISTRIBUIDORA DE FILMES LTDA	8404587
4	SM DISTRIBUIDORA DE FILMES LTDA	7884111
5	DIAMOND FILMS DO BRASIL PRODUÇÃO E DISTRIBUIÇÃO AUDIOVISUAL LTDA.	4507384
6	WMIX DISTRIBUIDORA LTDA.	3801609
7	FREESPIRIT DISTRIBUIDORA DE FILMES LTDA.	1934258
8	H2O DISTRIBUIDORA DE FILMES LTDA	1754354
9	UNITED CINEMAS INTERNATIONAL BRASIL LTDA.	545233
10	SA DISTRIBUIDORA DE CONTEÚDO AUDIOVISUAL LTDA	474841
11	O2 PRODUÇÕES ARTÍSTICAS E CINEMATOGRÁFICAS LTDA.	403748
12	PROVIDENCE DISTRIBUIDORA DE FILMES LTDA - EPP	231933
13	ANTONIO FERNANDES FILMES LTDA	190304

Figura 17: Outlier - Público Total por Distribuidora

Algumas distribuidoras registraram um público muito acima da média, que foi de 1375425.9, como Walt Disney e Warner Bros., que são amplamente reconhecidas mundialmente. Por outro lado, algumas se destacaram pelo baixo público, como Providence Distribuidora e Antônio Fernandes Filmes.

- **Outlier - Total de Filmes por Distribuidora**

Outliers - Total de Filmes por Distribuidora

	razao_social_distribuidora	quantidade_filmes
0	WARNER BROS. (SOUTH) INC.	87
1	SM DISTRIBUIDORA DE FILMES LTDA	50
2	TAG CULTURAL DISTRIBUIDORA DE FILMES LTDA	44
3	PROVIDENCE DISTRIBUIDORA DE FILMES LTDA - EPP	37
4	DIAMOND FILMS DO BRASIL PRODUÇÃO E DISTRIBUIÇÃO AUDIOVISUAL LTDA.	36
5	THE WALT DISNEY COMPANY (BRASIL) LTDA.	36

Figura 18: Outlier - Total de Filmes por Distribuidora

Pode-se afirmar que algumas distribuidoras produziram um número de filmes muito acima da média, que foi de 7 filmes, especialmente aquelas amplamente conhecidas no mercado.

5.1.4 Sazonalidade do Público

Uma hipótese a mais que levantamos foi o fato de que, seria esperado que em meses de férias escolares, o público nos cinemas aumentasse, ou seja, em Janeiro, Julho e Dezembro.

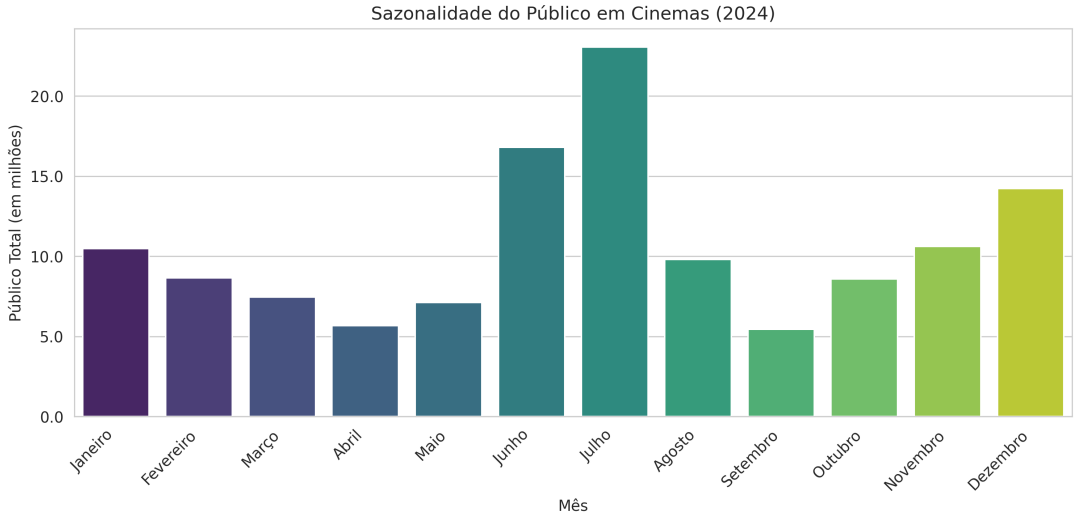

Figura 19: Público por mês de 2024

Entretanto, ao observar os resultados obtidos, é notório que nossa hipótese não foi totalmente satisfeita. Visando investigar os motivos para tais resultados, consultamos quais foram os filmes com mais público de 2024 (top 5) e quando suas respectivas seções foram exibidas:

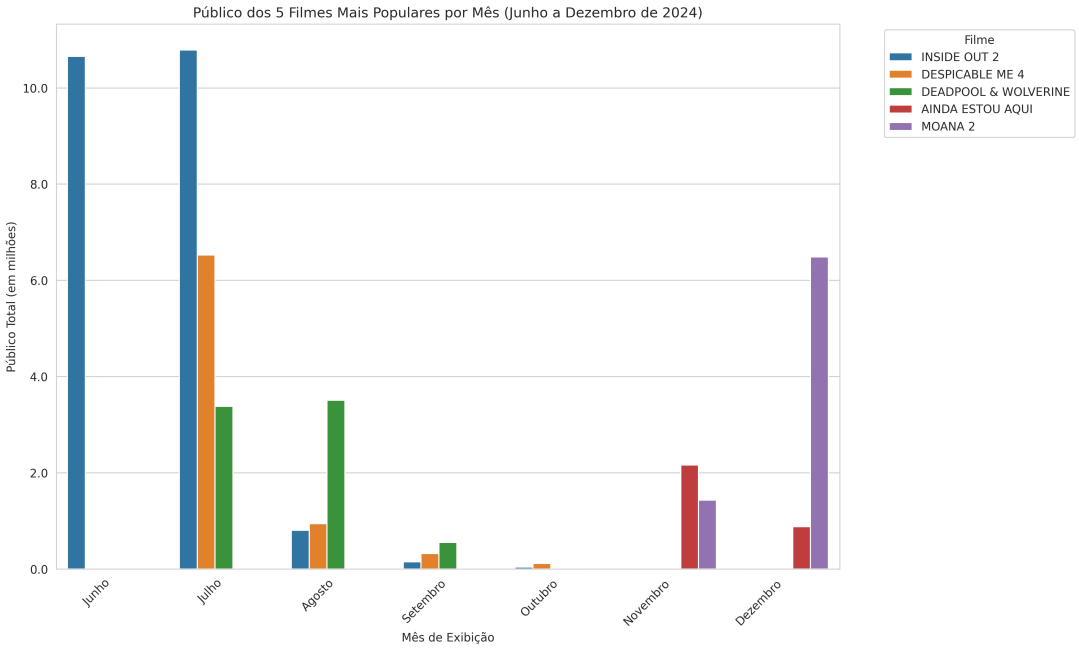

Figura 20: Seções dos 5 filmes mais populares de 2024

Por meio desse resultado, é possível observar que as principais seções dos princiapis filmes ocorreram justamente nos meses nos quais o publico nos cinemas foi bem maior (Junho, Julho, Novembro e Dezembro). Portanto, podemos concluir que a quantidade de público não é só orientada pelas férias escolares, mas, principalmente pelos filmes que estão sendo exibidos e sua popularidade.

5.1.5 Correlação

Correlação entre a quantidade de filmes lançados pela distribuidora e desempenho da distribuidora no ano, medido pelo público total.

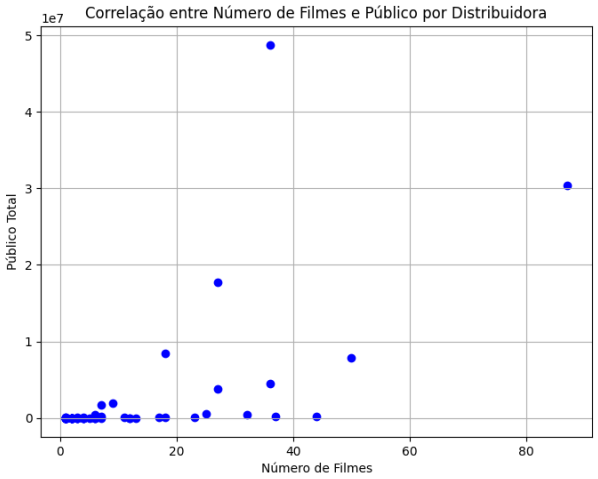

Figura 21: Correlação entre Número de Filmes e Público por Distribuidora

Identificamos alguns outliers e, antes de realizarmos a análise, retiramos esses valores.

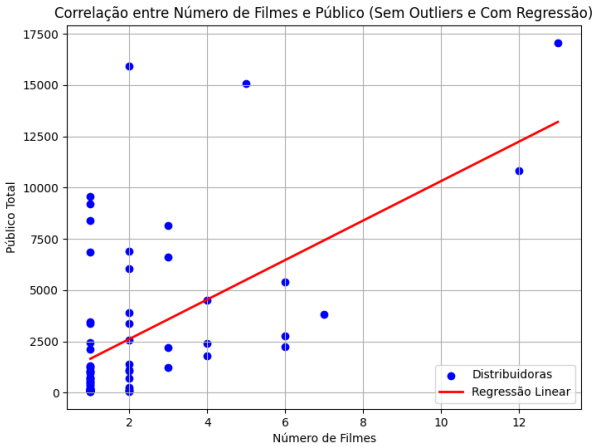

Figura 22: Correlação entre Número de Filmes e Público (Sem Outliers e Com Regressão)

Com base nas análises de correlação realizadas, observa-se uma correlação positiva, porém de intensidade moderada a fraca, entre o número de filmes e o público total por distribuidora. Isso indica que, de modo geral, distribuidoras que lançam mais filmes tendem a alcançar um público maior, mas essa relação não é tão expressiva.

O coeficiente de determinação calculado foi R^2 =0,31, o que significa que aproximadamente 31% da variação no público total pode ser explicada pela quantidade de filmes lançados. Isso demonstra que, embora exista uma associação, o modelo possui um ajuste limitado, sugerindo que outros fatores — como a popularidade dos filmes, estratégias de marketing, gêneros, distribuição geográfica, ou o número de sessões — exercem influência significativa sobre o tamanho do público.

Portanto, a quantidade de filmes é apenas um dos elementos que contribuem para o desempenho de público das distribuidoras, não sendo, isoladamente, um fator determinante.

5.2 Consultas extras

- **Tabela filme:**

Como já foi afirmado anteriormente, analisando a tabela filme, notamos que todos os filmes brasileiros, ou seja, produzidos em nosso país, tem o nome citado apenas na coluna titulo original, deixando portanto valores nulos na coluna titulo_brasil.

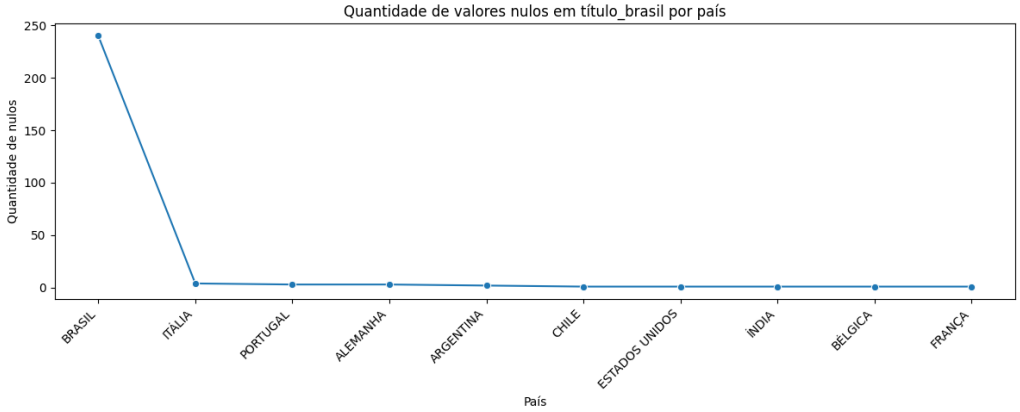

Figura 23: Quantidade de valores nulos em título_brasil por país

Claramente, os títulos que vem do Brasil, em grande parte, não possuem a coluna titulo_brasil completa, como já foi mencionado anteriormente. Quase todo filme que vem do Brasil não tem título_brasil, já que seu valor seria igual a de titulo_original. Há apenas um filmes brasileiro que possui titulo_brasil. Consulta a ser feita para descobrir qual é ele:

```
SELECT titulo_original, titulo_brasil
FROM public.filme
WHERE pais_obra = 'BRASIL'
AND titulo_brasil IS NOT NULL;
```

Essa consulta retorna o seguinte: MY PENGUIN FRIEND | MEU AMIGO PINGUIM

Pesquisando sobre o filme, descobrimos que filme My Penguin Friend é uma coprodução brasileira e estadunidense, com distribuição internacional e elenco/maker técnicos voltados a um público global, e por isso o título está em inglês.

Tratamento possível: Um tratamento possível para esse problema que pode ser feito aqui é copiar os valores de uma coluna para outra nos filmes brasileiros. Isso poderia ser feito com o comando SQL:

```
UPDATE public.filme
SET titulo_brasil = titulo_original
WHERE pais_obra = 'Brasil'
AND titulo_brasil IS NULL;
```

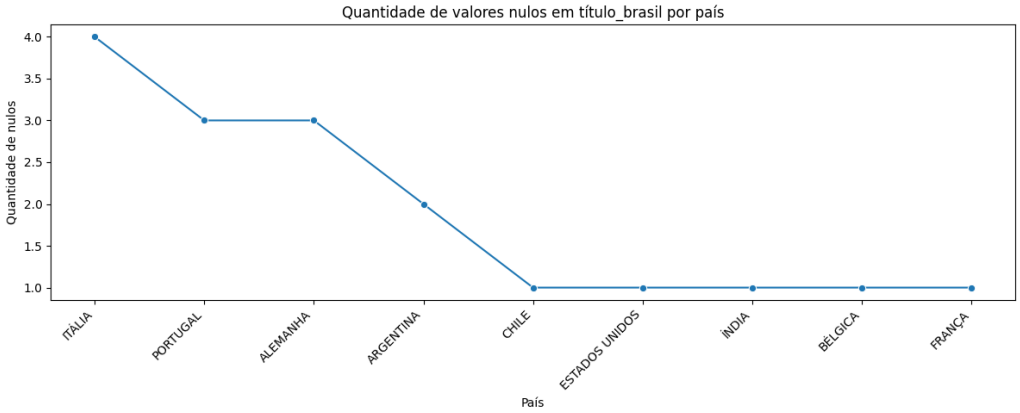

Figura 24: Quantidade de valores nulos em título_brasil por país após tratamento

- **Tabela complexo:**

Na tabela complexo, foram feitas análises sobre a situação dos complexos e sobre a quantidade de complexos itinerantes, que é bastante baixa.

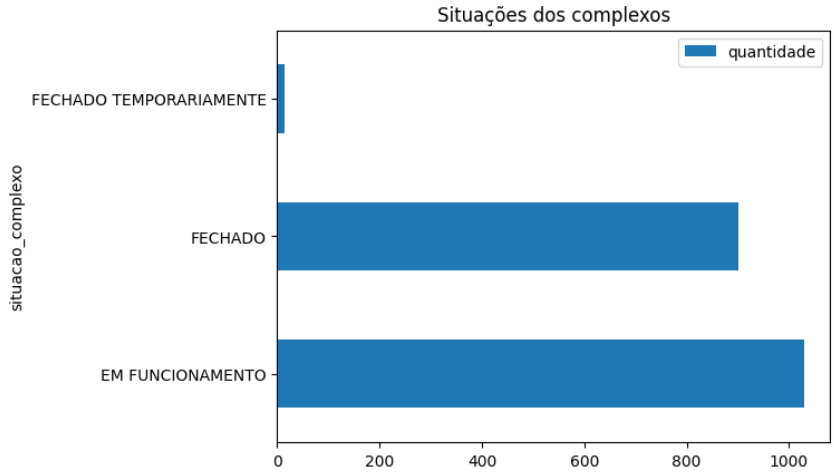

Figura 25: Gráfico - Situações dos Complexos

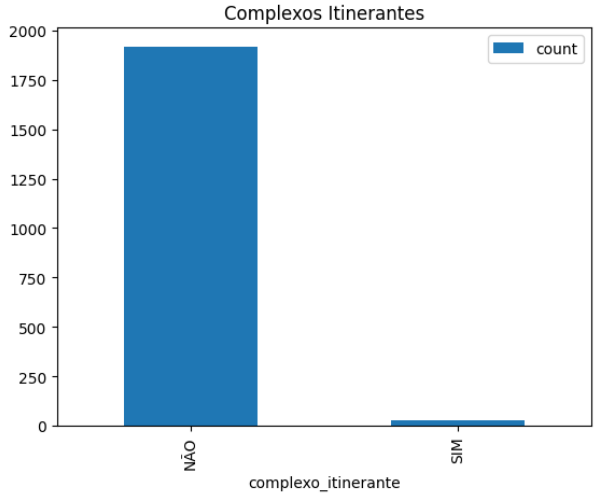

Figura 26: Gráfico - Complexos Itinerantes

6 Conclusão

Esse trabalho prático teve como objetivo a análise crítica de um banco de dados público, cujo tema foi escolhido de forma arbitrária. Por meio da aplicação da técnica de engenharia reversa, foi possível identificar e compreender os principais problemas presentes nos dados disponibilizados, permitindo, assim, uma reestruturação organizacional inicial da base. A partir desse processo, foi realizada uma análise exploratória dos dados, capaz de gerar conclusões numéricas e estatísticas relevantes sobre a própria base. Além disso, o desenvolvimento do trabalho proporcionou o aprimoramento de habilidades críticas na interpretação de informações, na identificação da necessidade de reestruturação dos dados e na realização de análises estatísticas. Por fim, destacou-se a importância de saber formular as perguntas corretas antes de partir para a elaboração de gráficos, consultas e interpretações, evidenciando que a etapa de questionamento é fundamental para uma análise de dados eficaz e bem fundamentada.

7 Referências bibliográficas

Chen, P. (1976) The entity-relationship model — toward a unified view of data. ACM Transactions on Database Systems 1(1) (March 1976), 9–36.

Codd, E. F. (1970) A Relational Model of Data for Large Shared Data Banks. Communications of the ACM 13(6):377-387.

Elmasri, R., Navathe, S. B. (2011). Sistemas de Banco de Dados (6ª ed.). São Paulo: Pearson Prentice Hall.