Introdução a Banco de Dados

Trabalho 2

Letícia Scofield e Lorenzo Vagliano

Introdução

Nesse projeto, a fim de produzir um conteúdo de natureza extensionista, buscamos dados disponíveis publicamente que nos permitam fazer análises de utilidade pública.

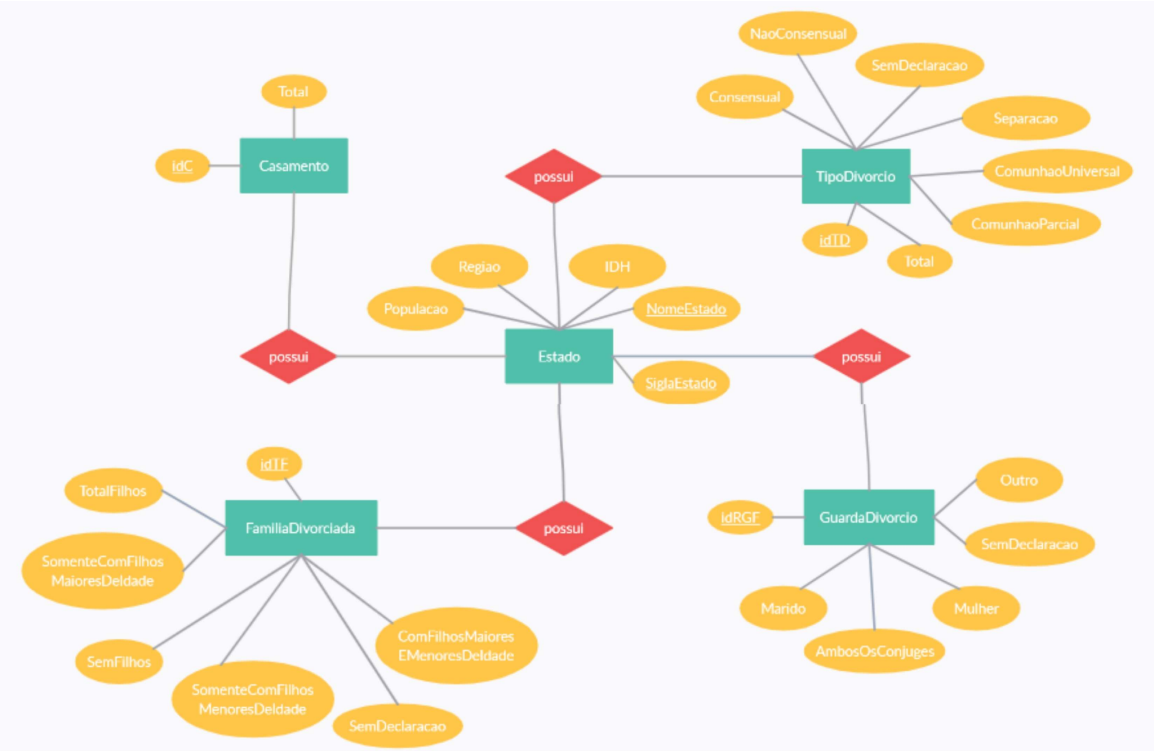
Para isso, buscamos conjuntos de dados acerca de Divórcios e Casamentos em cada Estado da Federação. Além disso, foram obtidas informações sobre cada Estado acerca de sua Região, População e IDH, permitindo assim uma análise composta da situação matrimonial em toda União.

Com a junção desses dados em um modelo Relacional, buscamos facilitar sua pesquisa e análise com o intuito de possibilitar a formação de hipóteses sobre as relações da situação de cada Estado ou Região com a frequência de tipos diferentes de divórcio e casamento.

Análise Exploratória

A partir dessa análise, buscaremos analisar a natureza dos dados obtidos por meio de esquemas gráficos e medidas que nos permitam avaliar o conjunto de dados.

Esquema Conceitual ER



Dicionário de Dados

RELAÇÃO	ATRIBUTO	TIPO/ LARGURA	NULO?	ÚNICO?	VALORES PERMITIDOS	COMPORTAMENTO
Estado	SiglaEstado	varchar(100)	N	S		PK
Estado	NomeEstado	varchar(100)	N	N		
Estado	Regiao	varchar(100)	N	N		
Estado	População	integer	S	N		
Estado	IDH	double precision	S	N		
TipoDivorcio	idTD	varchar(100)	N	S		PK
TipoDivorcio	SiglaEstado	varchar(100)	N	S		FK REFERENCES "Estado" ("SiglaEstado")
TipoDivorcio	Total	integer	S	N		
TipoDivorcio	Consensual	integer	S	N		
TipoDivorcio	NaoConsensual	integer	S	N		
TipoDivorcio	ComunhaoUniversal	integer	S	N		
TipoDivorcio	ComunhaoParcial	integer	S	N		
TipoDivorcio	Separacao	integer	S	N		
TipoDivorcio	SemDeclaracao	integer	S	N		
GuardaDivorcio	idRGF	varchar(100)	N	S		PK
GuardaDivorcio	SiglaEstado	varchar(100)	N	S		FK REFERENCES "Estado" ("SiglaEstado")
GuardaDivorcio	Marido	integer	S	N		
GuardaDivorcio	Mulher	integer	S	N		
GuardaDivorcio	AmbosOsConjuges	integer	S	N		
GuardaDivorcio	Outro	integer	S	N		
GuardaDivorcio	SemDeclaracao	integer	S	N		

FamiliaDivorciada	idTF	varchar(100)	N	S		PK
FamiliaDivorciada	SiglaEstado	varchar(100)	N	S		FK REFERENCES "Estado" ("SiglaEstado")
FamiliaDivorciada	SemFilhos	integer	S	N		
FamiliaDivorciada	SomenteComFilhosMaioresDeIdade	integer	S	N		
FamiliaDivorciada	SomenteComFilhosMenoresDeIdade	integer	S	N		
FamiliaDivorciada	ComFilhosMaioresEMenoresDeIdade	integer	S	N		
FamiliaDivorciada	SemDeclaracao	integer	S	N		
FamiliaDivorciada	TotalFilhos	integer	S	N		
Casamento	idC	varchar(100)	N	S		PK
Casamento	SiglaEstado	varchar(100)	N	S		FK REFERENCES "Estado" ("SiglaEstado")
Casamento	Total	integer	S	N		

Metadados

Casamentos

- Data de obtenção: 18/06/2023
- Órgão produtor: IBGE
- Data de referência: 2021
- Limitações registradas:
    - Dados desatualizados.
    - Alguns dados vazios.
    - Formatação incompatível com o SQL(espaços, vírgulas, etc).
    - Formatação de intuito estético no formato excel(título ocupando várias colunas, subtítulos, etc).
- Cobertura: Brasil

Divórcios

- Data de obtenção: 18/06/2023
- Órgão produtor: IBGE
- Data de referência: 2021
- Limitações registradas:

    - Dados desatualizados.
    - Alguns dados vazios.
    - Formatação incompatível com o SQL(espaços, vírgulas, etc).
    - Formatação de intuito estético no formato excel(título ocupando várias colunas, subtítulos, etc).
- Cobertura: Brasil

População

- Data de obtenção: 18/06/2023
- Órgão produtor: IBGE (Prévia do censo 2022)
- Data de referência: 2022
- Limitações registradas:
    - Ausência de uma tabela pronta, foi necessário criar manualmente com os dados da plataforma.
- Cobertura: Brasil

IDH

- Data de obtenção: 18/06/2023
- Órgão produtor: Pnud Brasil, Ipea e FJP
- Data de referência: 2021
- Limitações registradas:
    - Ausência de uma tabela pronta, foi necessário criar manualmente com os dados da plataforma.
- Cobertura: Brasil

Estatísticas

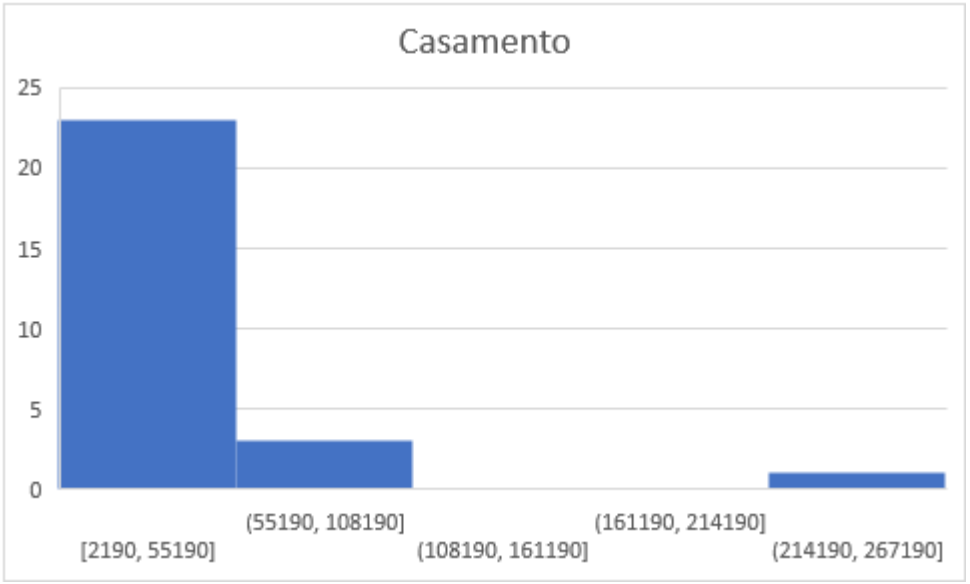
A fim de caracterizar nossos dados, algumas medidas estatísticas foram feitas. Todas elas foram feitas com os atributos “Total” de Casamento e TipoDivorcio, com a “População” e com o “IDH” dos estados.

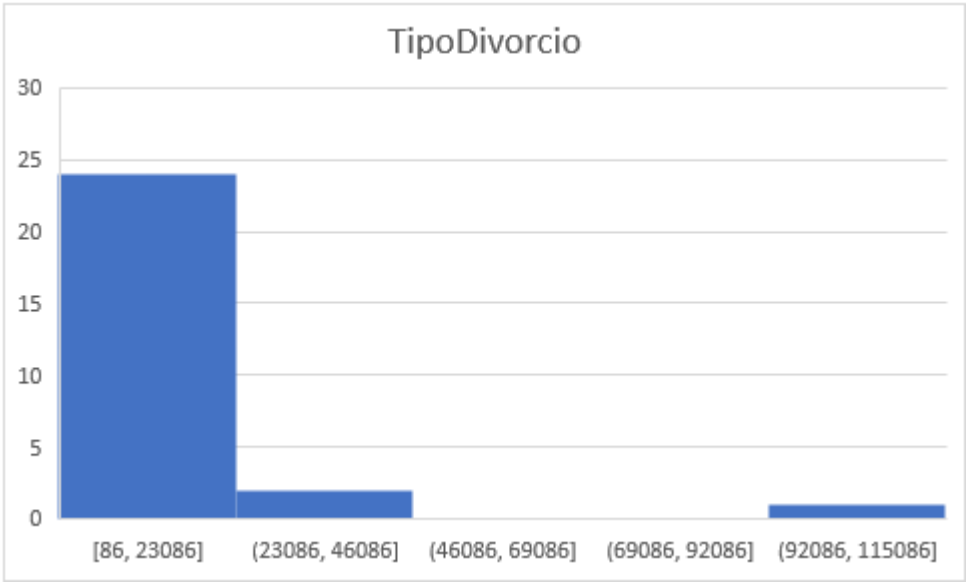
	Casamento	TipoDivorcio	População	IDH
Valor médio	34196.2963	11105.40741	7699396.259	0.730148148
Quantidade de Valores Distintos	27	27	27	24

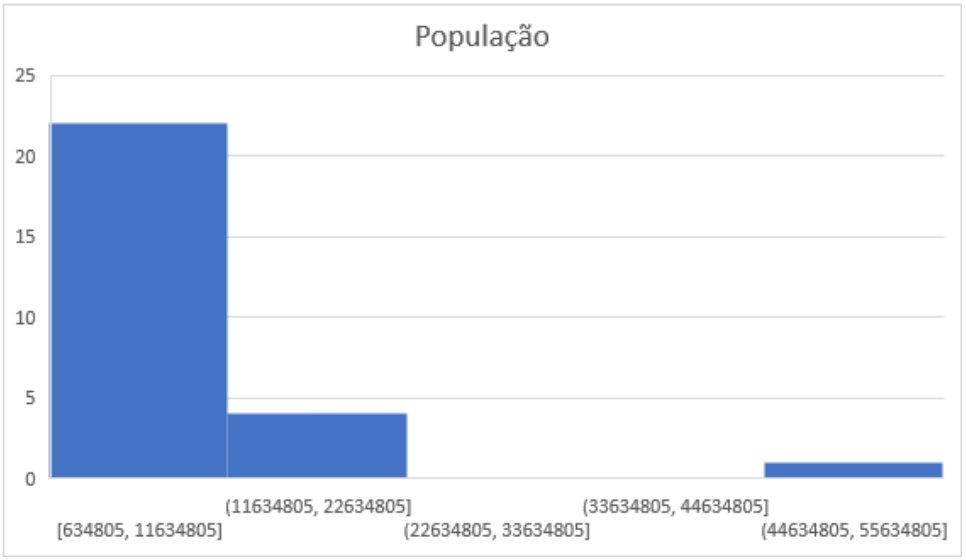
Com isso, podemos observar o alto número de valores distintos. Já que o maior número de observações é 27, a imensa maioria dos valores são distintos.

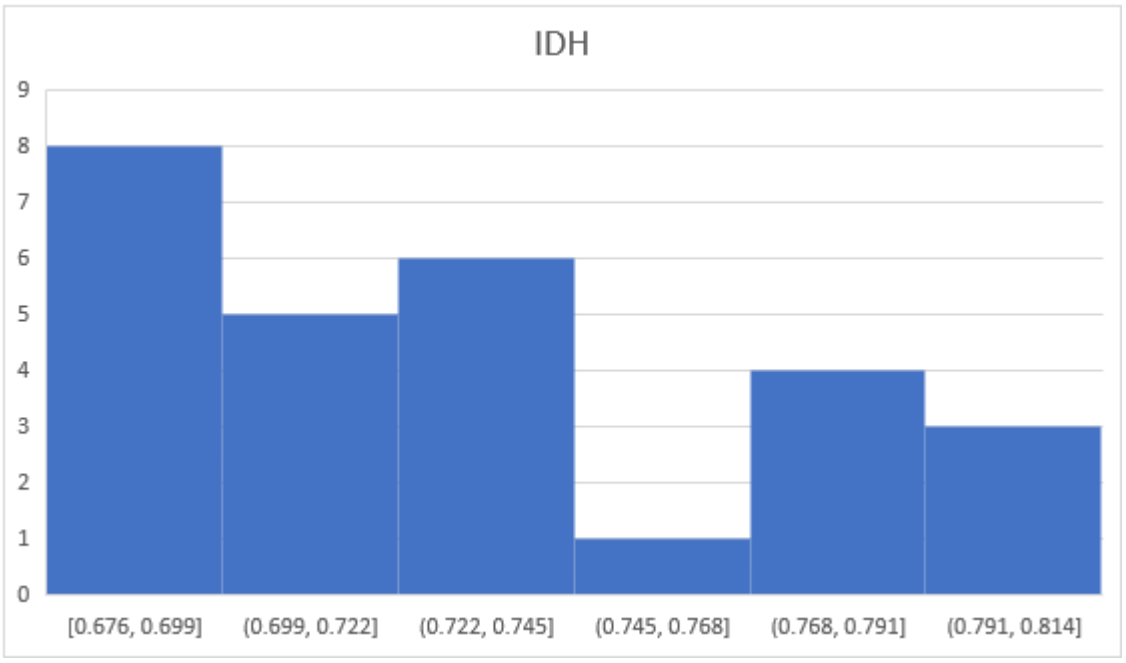
Para nenhum dos atributos foram encontrados valores determinados como *outliers* quando se leva em consideração a proporção entre o dado capturado e a população de cada Estado.

Por último, para determinar a distribuição de cada atributo, foram produzidos histogramas:


Casamento
25
20
15
10
5
0
[2190, 55190]
(55190, 108190]
(108190, 161190]
(161190, 214190]
(214190, 267190]


TipoDivorcio
30
25
20
15
10
5
0
[86, 23086]
(23086, 46086]
(46086, 69086]
(69086, 92086]
(92086, 115086]


População
25
20
15
10
5
0
[634805, 11634805]
(11634805, 22634805]
(22634805, 33634805]
(33634805, 44634805]
(44634805, 55634805]



Pelos histogramas já é possível determinar que os dados não seguem uma distribuição específica.

Porém, para confirmar essa hipótese pelo menos quanto à distribuição normal, que permitiria muitas análises, foi feito o teste de Shapiro-Wilk com cada atributo, utilizando um nível de significância de 0.05. Todos atributos falharam no teste, provando a diferença da distribuição normal.

Análise Crítica

Inicialmente, o próprio formato dos dados obtidos resultou em dificuldades na sua conversão para um banco de dados relacional.

As tabelas encontradas referentes a casamentos e divórcios não buscam representar a abstração de uma entidade ou de um relacionamento, mas sim uma coletânea de dados categorizados.

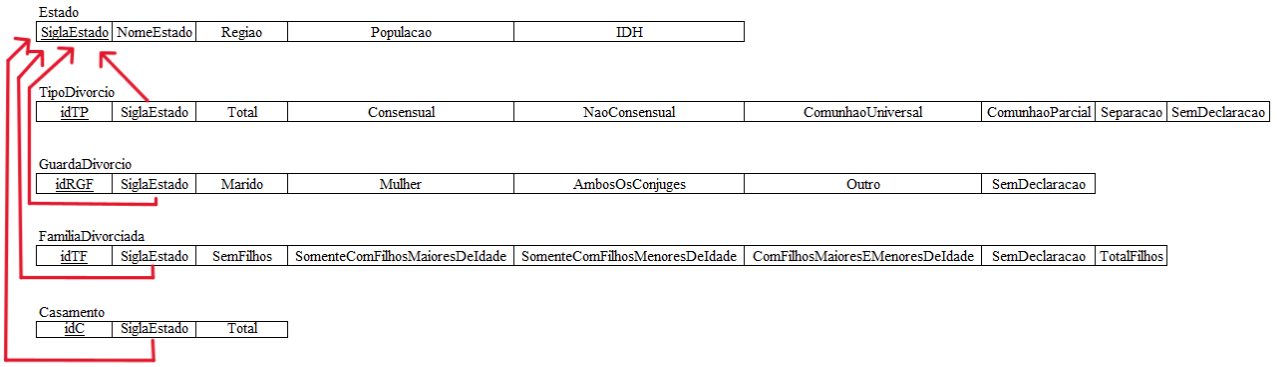
Mais especificamente, todos os atributos(colunas) dessas tabelas são sub categorias diferentes no que tange a tabela como um todo. Além disso, eles não registram ocorrências individuais, mas a quantidade de vezes essa categoria ocorreu, com a exceção do Estado pertencente e do ID artificial criado.

Por exemplo, em “TipoDivórcio” são registrados o número total de divórcios no atributo “Total” e o número de divórcios consensuais no atributo “Consensual”. Por conta dessa característica, foi difícil criar um critério para a divisão das tabelas a fim de produzir nossas próprias entidades.

Além disso, os dados mesclavam sua categorização regional entre Estado e Município, então foi necessário realizar uma extração manual dos dados estaduais. Também haviam vários atributos preenchidos esparsamente, sem uma determinação específica de valores nulos.

Por último, devido às diferentes fontes dos dados, alguns conjuntos de dados relacionados possuem datas de referência distintas, como a População, que foi retirada da prévia do censo de 2022, que era a referência mais próxima aos nossos outros dados (de 2021) em comparação com o censo de 2010.

Combinação e Integração dos Dados



Como pode ser observado no Esquema Relacional produzido, todos os dados possuem em comum o estado de sua origem. Dessa maneira, para relacioná-los, foi criada a entidade Estado, cuja sigla é chave estrangeira de todas as outras entidades.

À entidade Estado, também foram adicionados atributos referentes a sua Região, população e IDH para que uma análise acerca dos divórcios e casamentos possa ser feita levando esses fatores em consideração.

Dados Utilizados

Repositório com os dados utilizados em formato CSV: https://github.com/Lorenzovagliano/TP2ibd

No repositório também foi disponibilizado um **script SQL** para gerar o banco de dados criado. Ele já terá os dados registrados também.