UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

LUCAS ESCOBAR
RODRIGO DE CASTRO MICHELASSI

Relatório - Privacidade em Saúde

São Paulo
2025

Sumário

1 Introdução

A Lei Geral de Proteção aos Dados (LGPD), aprovada no Brasil em 2018, declara que todos indivíduos têm direito sobre o armazenamento e processamento dos seus dados. Todavia, o Sistema Único de Saúde (SUS) disponibiliza dados sobre seus pacientes online, de maneira acessível a qualquer interessado. Entre os dados disponibilizados, informações sensíveis são capazes de tornar qualquer indivíduo, dono desses dados, identificável. Dessa forma, a maneira que os dados dos indivíduos são disponibilizados violam seus direitos individuais, garantidos pela LGPD.

Nesse projeto, exploramos os dados disponibilizados pelo datasus, referente ao estado do Rio de Janeiro, a fim de encontrar falhas na garantia da privacidade dos pacientes do SUS. Além disso, aplicamos técnicas como K-Anonymity e Differential Privacy, para tentar, em partes, propor um dataset mais seguro e tornar os pacientes do SUS menos facilmente identificáveis.

Assim, ao aplicar essas técnicas, obtivemos dados com muito mais segurança de privacidade, além de chegar a conclusões práticas relacionadas aos trade-offs de funcionalidade e qualidade dos dados e a qualidade da anonimização, além de ser capaz de selecionar qual modelo de anonimização é mais interessante para aplicação em dados de saúde.

2 Análise e tratamento dos dados

2.1 Tratamento do dataset

Ao obter os dados do datasus, nos deparamos com uma diversidade grande de colunas, das quais muitos nomes não dizem muito sobre os dados que são representados. A fim de fazer uma análise mais profunda dos dados, a documentação das colunas foi de extrema importância, e está destacada no jupyter notebook entregue.

Nesse contexto, percebemos que diversas colunas são documentadas como "Zerado", ou seja, carregamos uma grande quantidade de dados que não serão úteis para o projeto, pois não transmitem nenhuma informação. Com isso, uma decisão inicial foi limpar o dataframe obtido, removendo todas colunas zeradas.

Além disso, encontramos incoerência nos dados de idade e data de nascimento, visto que as idades eram limitadas entre 0 e 99 anos, porém haviam pacientes nascidos antes de 1924, logo nenhum desses pacientes estaria na faixa correta de idade. Como decisão de projeto, todos pacientes nascidos antes de 1924 foram removidos.

2.2 Análise dos dados

No jupyter notebook entregue, há uma análise completa dos dados, juntamente a comentários detalhados do que foi observado. Nesse relatório, iremos documentar apenas as informações mais importantes e impactantes observadas.

Primeiramente, trazendo uma análise sobre a distribuição de pacientes por município, conseguimos observar na figura 1 que há um grande número de pacientes no dataset, porém o número de CEPs é uma porção muito inferior. Esse dado nos leva a concluir, a princípio que há uma relação forte de pessoas por CEP.

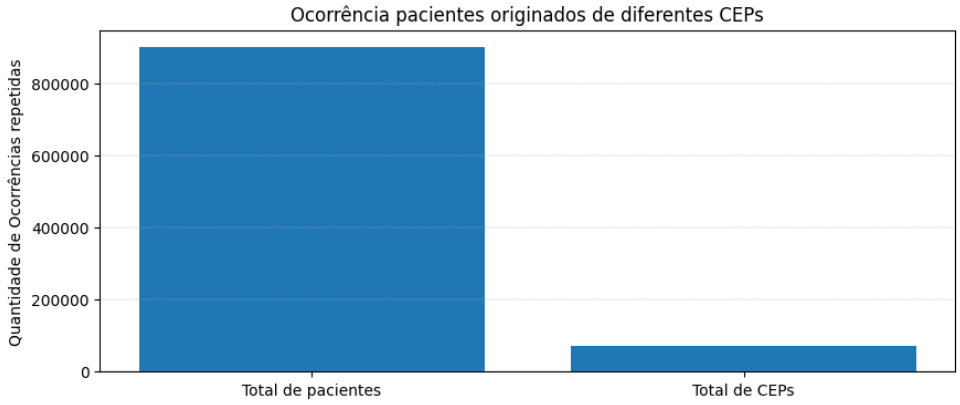

Figura 1 – Análise da proporção de pacientes e CEPs de residência

Por outro lado, ao analisarmos a figura 2, conseguimos ver, no eixo Y, a quantidade de CEPs que possuem um número de pacientes, nas faixas destacadas pelo eixo X. Assim, observe que grande parte dos CEPs cadastrados possuem um número inferior a 10 pacientes. Se destrincharmos mais ainda esse gráfico, veremos um grande número de CEPs com apenas 1 paciente.

Essa análise nos leva a perceber problemas sérios de privacidade, envolvendo a fácil identificação dos indivíduos presentes no dataset. Como visto em aula, se soubermos apenas CEP, gênero e idade de um indivíduo, sua identificação pode ser fácil e, além de todos esses dados estarem presentes no dataset, há ainda a presença de informações sensíveis, como a enfermidade que o paciente foi diagnosticado.

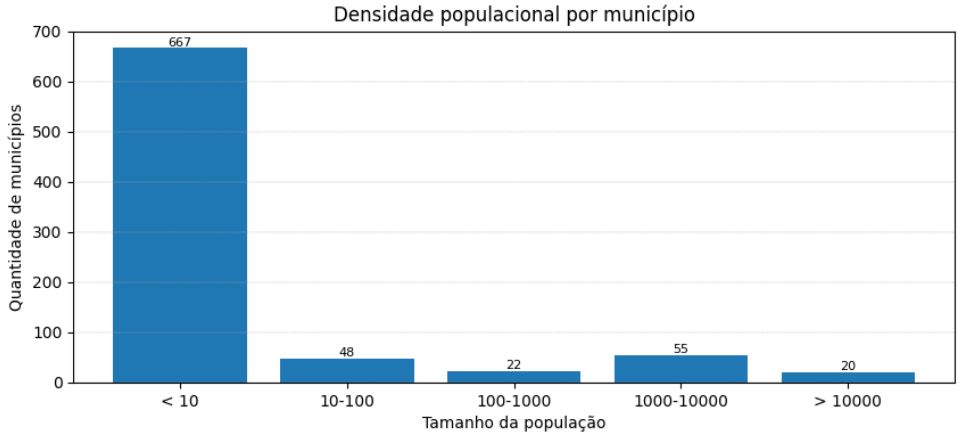

Figura 2 – Análise do tamanho da população por CEP

Ao analisarmos também a presença de CEPs no dataset, há uma presença de 90 CEPs que seguem prefixos iguais (XY*), como evidenciado na figura 3. Esse dado revela ainda mais sobre a especificidade de se ter o CEP de um indivíduo, ainda mais se

considerarmos novamente a figura 2, na qual é possível observar que muitos CEPs possuem um pequeno número de pacientes.

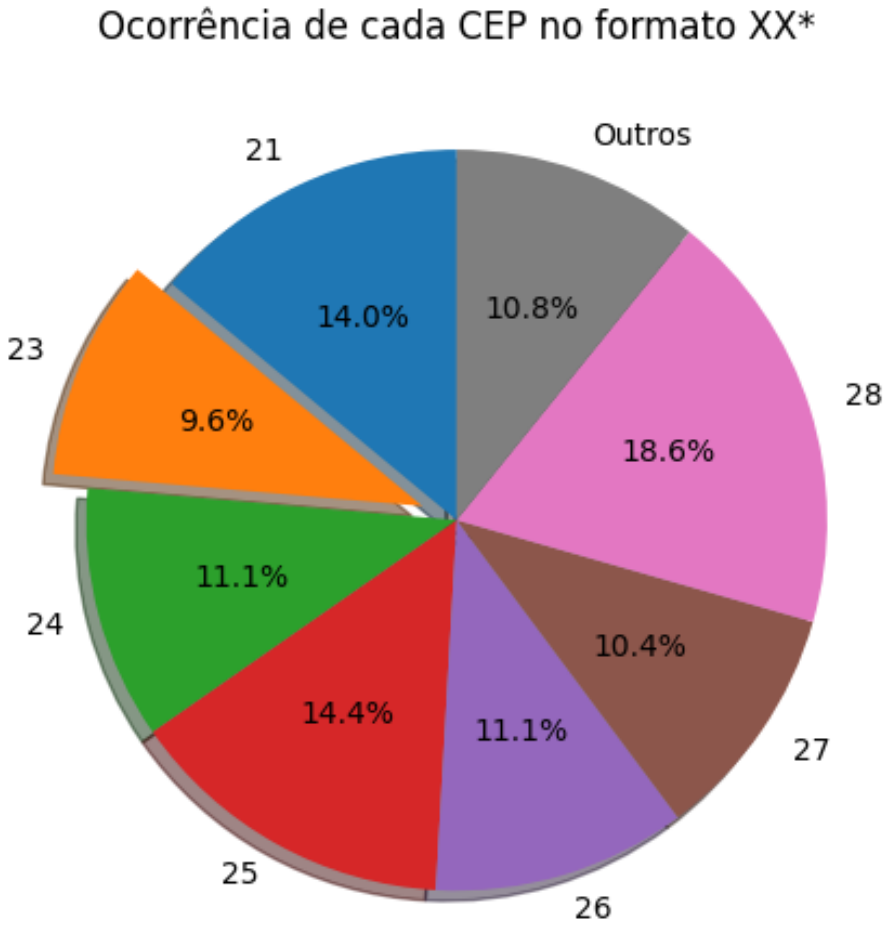

Figura 3 – Proporção de cada CEP no dataset

No jupyter notebook entregue, outros dados interessantes são explorados, porém não estaremos tratando sobre esses arquivos aqui, pois pouco revelam sobre dados de privacidade semelhantes aos que iremos explorar a anonimização. Ainda assim, a leitura do arquivo faz sentido, pois revela outros problemas de privacidade presentes nos dados do Datasus.

3 Differential Privacy

3.1 Dados e problemas de privacidade

Para aplicar Differential Privacy nos dados, escolhemos seguir diretamente os exemplos propostos no enunciado, que dizem respeito ao agrupamento de diagnósticos de doença (CID-10) por município, separado por gênero, e somando o valor total da consulta. Note que, com essa seleção, nosso objetivo foi anonimizar a soma dos valores de consulta por município, assim como a quantidade de pacientes que contribuíram para esses gastos. Uma outra análise interessante seria aplicar differential privacy no número de pacientes diagnosticados com certa doença.

Inicialmente, para entender o problema que estávamos lidando, resolvemos separar os dados nesse formato manualmente, usando a biblioteca Pandas, e na figura 4 já nos deparamos com alguns problemas. Note que, para alguns municípios, conseguimos ver que há apenas um paciente, de um sexo específico, que obteve um diagnóstico especifico, e seu tratamento foi de um valor específico (em dólares).

```
df_grouped = df.groupby(['MUNIC_RES', 'DIAG_PRINC', 'SEXO']).agg(
    VALOR_USD=('US_TOT', 'sum'),
    OCORRENCIAS=('US_TOT', 'count')
).reset_index()
df_grouped.head()
```

	MUNIC_RES	DIAG_PRINC	SEXO	VALOR_USD	OCORRENCIAS
0	110002	D171	Feminino	82.63	1
1	110002	Q251	Masculino	2,047.78	1
2	110004	A319	Masculino	408.39	1
3	110004	N179	Masculino	344.87	1
4	110004	Q899	Masculino	88.97	1

Figura 4 – Separação dos dados em um dataframe

Esses dados problemáticos foram expostos também visualmente por meio de gráficos, que serão expostos na seção de comparação de resultados.

Para esse projeto, mantivemos a implementação de Differential Privacy no nível mais simples, sem a definição de public partitions. Ao aplicarmos Differential Privacy nos dados originais, além da anonimização dos dados, conseguimos organizar esses similarmente a figura 4.

Adicionalmente, na figura 5, podemos observar um histograma que representa a distribuição dos dados referentes a aparição de diversos valores clínicos. Note que a maior parte das consultas informadas nos dados tem um custo mais baixo, e a minoria, representada pela cauda a direita, são consultas de custo mais elevado.

Dessa forma, não só se torna importante anonimizar dados de consultas com valores mais elevados, como também se revela a escolha do parâmetro max_value. Esse parâmetro define a contribuição máxima que um dado pode sofrer, sua sensibilidade, durante a aplicação de differential privacy. Para esse projeto, escolhemos $max\_value = 400$, pois a maior parte dos dados está presente em uma região inferior a essa. Por outro lado, se quisermos preservar ainda mais a integridade dos dados, sacrificando um pouco de anonimização, deveríamos escolher um valor menor.

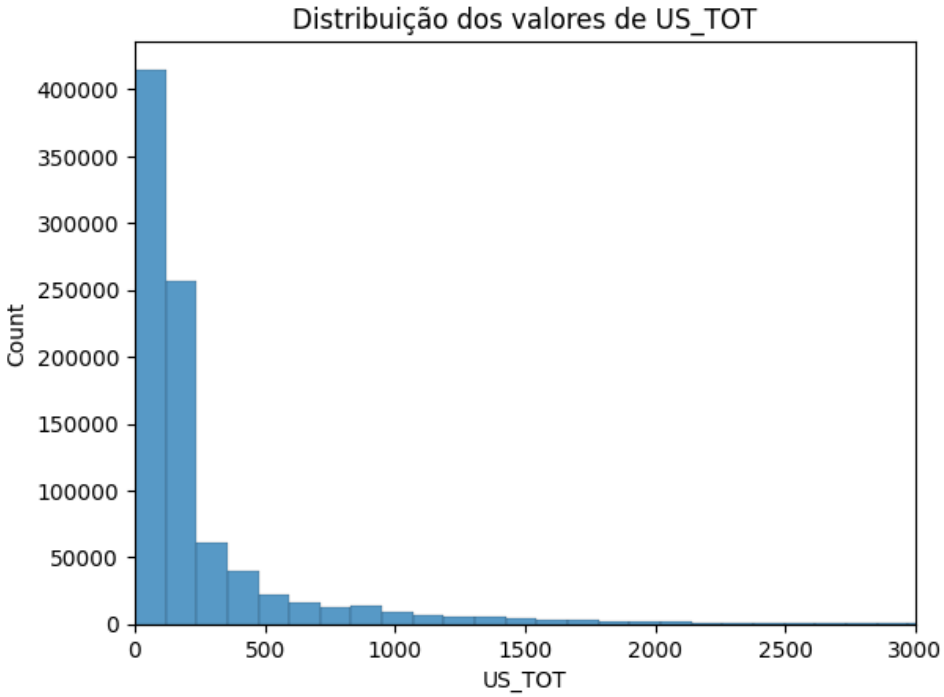

Figura 5 – Distribuição de gastos clínicos no dataset

Para escolher os outros parâmetros, além de realizar testes, seguimos as anotações propostas nos slides. Escolhemos $\delta = \frac{1}{n}$ e um privacy budget $\varepsilon = 1$, para equilibrar, em partes, a perda de informações do dataset.

Ao aplicar esses métodos nos dados originais e computar as ocorrências para os dados após aplicação de differential privacy, obtivemos a comparação da figura 6, na qual a tabela de cima representa os dados com differential privacy, e a tabela de baixo é a mesma vista anteriormente, dos dados originais, agrupados da forma proposta.

Além disso, muitos campos de valor ficaram negativos, e como decisão de projeto, apenas decidimos transformar todos em 0.

	MUNIC_RES	DIAG_PRINC	SEXO	VALOR_USD_DP	OCORRENCIAS_DP
3190	330480	O200	Feminino	0.00	47.94
2761	330455	N906	Feminino	0.00	87.42
1351	330023	F208	Feminino	0.00	49.07
1403	330040	F239	Masculino	0.00	56.61
1454	330040	F238	Masculino	0.00	44.22

	MUNIC_RES	DIAG_PRINC	SEXO	VALOR_USD	OCORRENCIAS
82524	330455	F001	Masculino	0.00	15
27307	330170	F54	Masculino	0.00	1
27306	330170	F53	Feminino	0.00	1
72117	330411	G40	Masculino	0.00	1
27305	330170	F505	Feminino	0.00	1

Figura 6 – Comparação dos dados após aplicar differential privacy

3.3 Resultados e Comparações

Uma primeira comparação feita foi interpretar como os valores gastos anonimizados se comportaram em relação aos dados originais. Uma boa forma de analisar isso seria observar o comportamento nas bordas, ou seja, entre os maiores valores e entre os menores valores. As figuras 7 e 8 demonstram que a aplicação de Differential Privacy teve o resultado esperado: os dados com valores muito altos, outliers, foram diminuídos, se equilibrando ao perfil do dataset. Já os dados com valores mais baixos foram aumentados, no geral, tirando um pouco do viés do dataset em relação aos baixos valores.

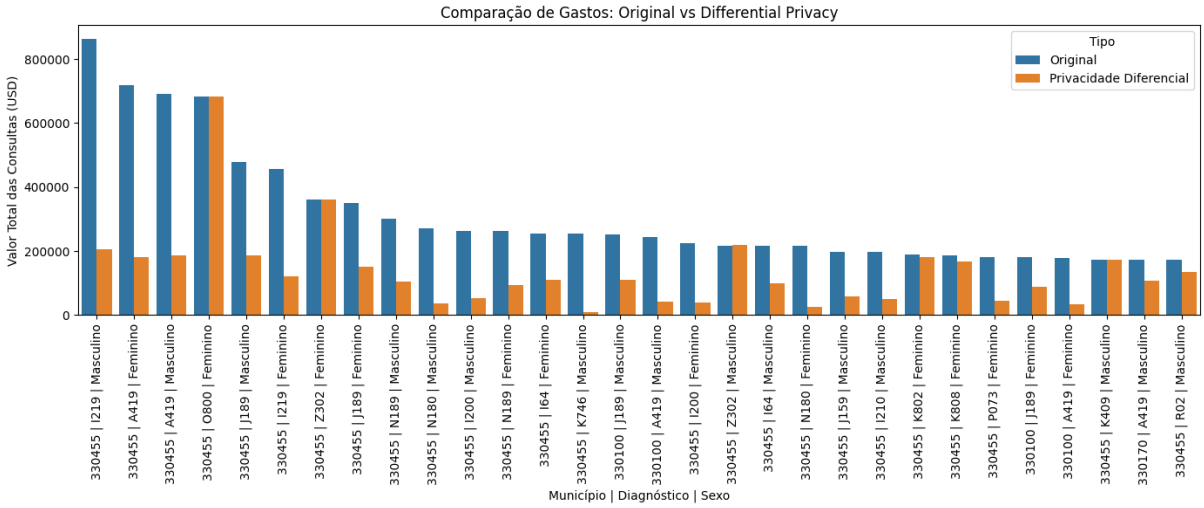

Figura 7 – Comparação de gastos altos após aplicar DP

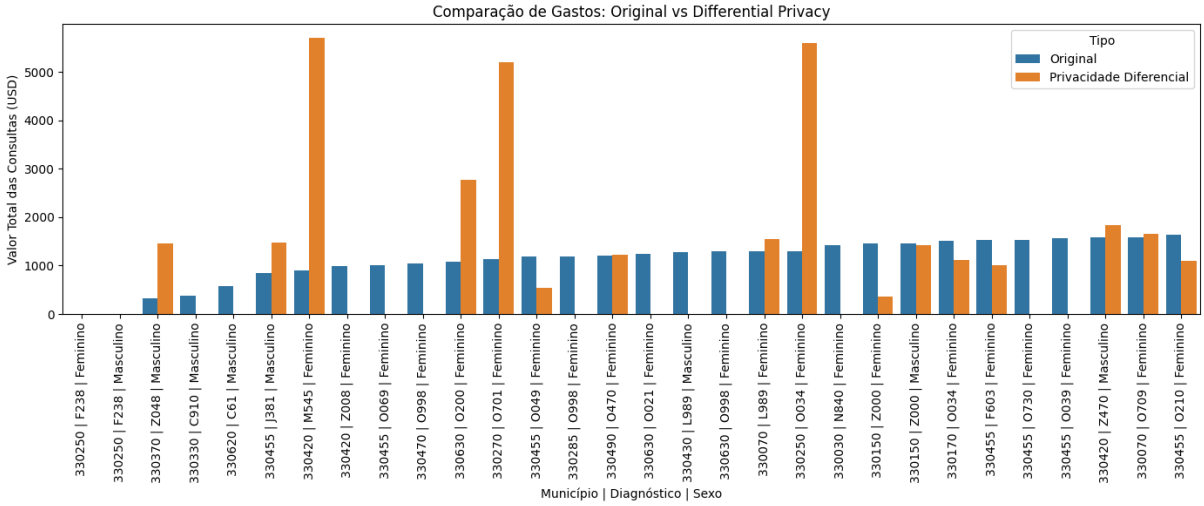

Figura 8 – Comparação de gastos baixos após aplicar DP

Além disso, comparamos a aplicação de differential privacy em outros contextos, para garantir que alcançamos a anonimização dos dados desejada.

Primeiramente, para os maiores gastos por município, de acordo com o sexo, podemos ver nas figuras 9 e 10, o padrão estatístico dos dados seguem semelhantes. Todavia, os valores são mais baixos, o que demonstra que a aplicação de differential privacy teve resultados positivos normalização de outliers.

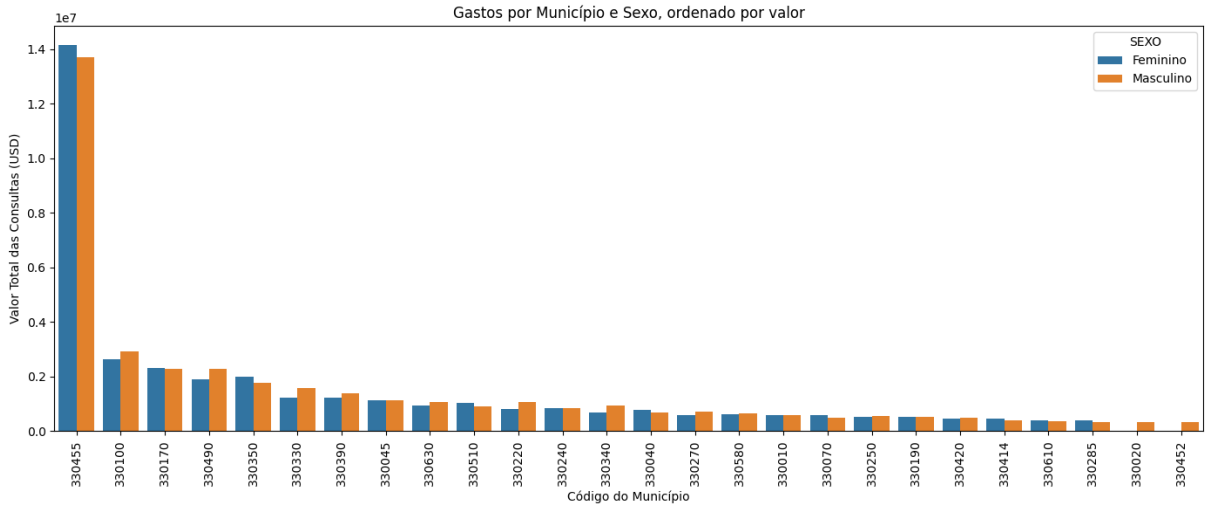

Figura 9 – Gastos por município, sem DP

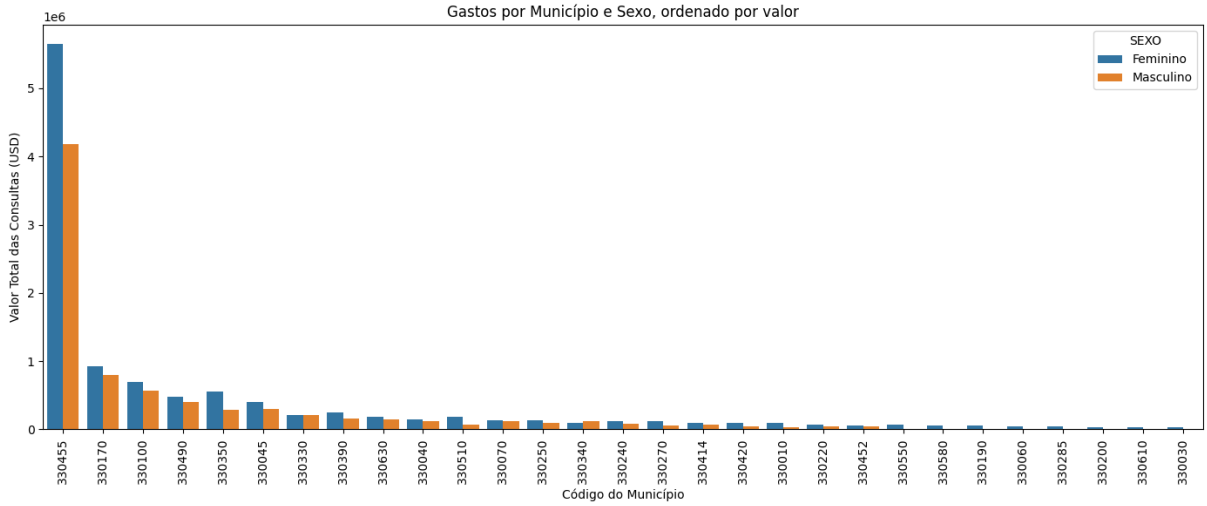

Figura 10 – Gastos por município, com DP

Além disso, fizemos uma comparação também da quantidade de pacientes por município que contribuíram para o valor clínico. Para essa comparação, ordenamos do menor para o maior, e podemos observar, na figura 11, que há muitos valores que sofreram contribuições de apenas um paciente. Dessa forma, na figura 12, após aplicarmos differential privacy, obtivemos um dataset mais adequado em termos de privacidade, onde os dados estão anonimizados. Nesse caso, os dados estão em valores bem altos, porém ao ajustar os parâmetros poderíamos obter uma anonimização mais fiel aos dados originais.

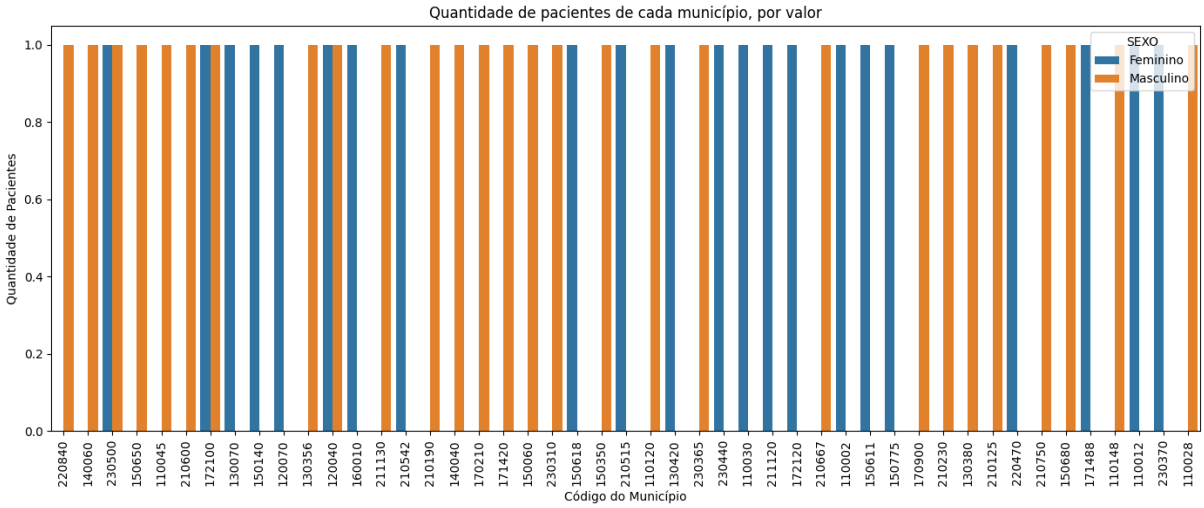

Figura 11 – Quantidade de Pacientes por valor arrecadado sem DP

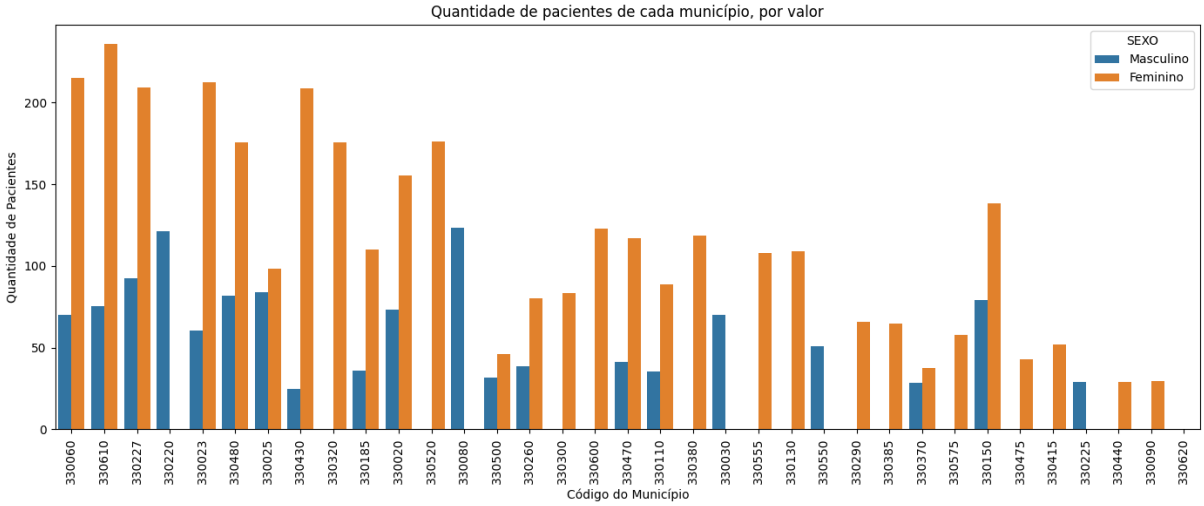

Figura 12 – Quantidade de Pacientes por valor arrecadado com DP

4 K-Anonymity

Para K-Anonymity, o propósito da anonimização segue de maneira diferente. Aqui, queremos agrupar dados em grupos, com pelo menos k dados correspondentes.

Para agrupar os dados correspondentes nesses grupos, selecionamos as colunas data de nascimento e CEP, na qual criamos grupos de data de nascimento com um alcance de 20 anos, e o CEP representado apenas por seu prefixo, com os dois primeiros dígitos.

Além disso, decidimos não trabalhar exatamente com CID-10, isso pois essa coluna está associada com um diagnóstico médico, que é um dado extremamente sensível. Se agrupássemos pessoas que possuem o mesmo CID-10, e alguma delas fosse exposta, teríamos o problema de revelar uma informação pessoal de forte valor, o que não é nossa ideia. Além disso, para pessoas com diagnósticos mais raros, sua informação estaria totalmente comprometida. Esse vazamento é conhecido como "grupo viesado".

Ao selecionarmos essas métricas para aplicar K-Anonymization, obtemos que 1326 grupos ainda violavam as condições estipuladas, ou seja, não puderam ser encaixados em grupo algum com pelo menos 3 dados. Assim, como uma decisão de projeto, removemos esses dados do dataset para fazer uma análise dos resultados obtidos.

4.1 Resultados

A figura 13 demonstra a distribuição dos grupos formados, após aplicar k-anonymity.

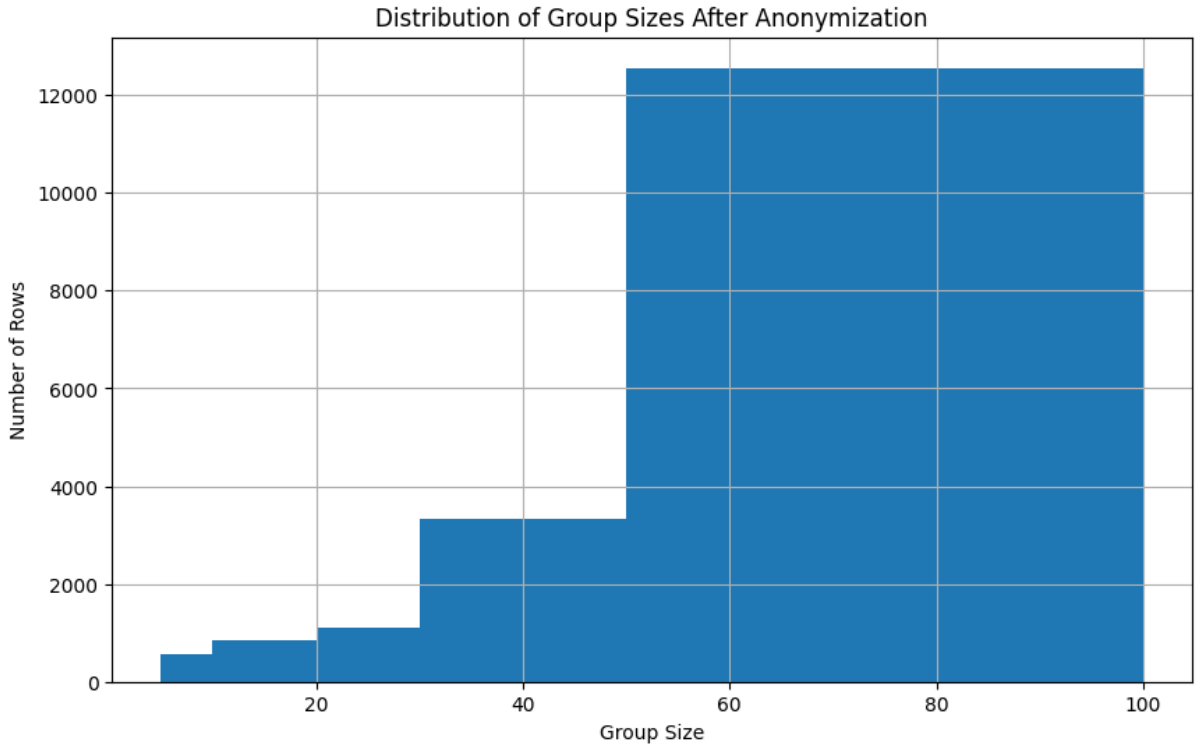

Figura 13 – Tamanho dos grupos, após k-anonymity

Ao analisarmos mais profundamente a distribuição dos grupos formados após a aplicação do k-anonymity, observamos que a maioria das classes de equivalência — ou seja, os conjuntos de registros com os mesmos quasi-identifiers — têm tamanhos relativamente pequenos. Isso é evidenciado na Figura 14, onde a distribuição está concentrada nos menores tamanhos de grupo, indicando que há muitas classes pequenas e poucas classes grandes.

Essa observação complementa o que foi mostrado na Figura 13, já que nela é possível ver que a maior parte dos registros pertence a grupos com tamanhos entre 50 e 100, refletindo a fragmentação da base mesmo após a anonimização. Isso poderia refletir um problema nos dados, porém, dado que escolhemos um valor pequeno para k, o resultado está acima do esperado.

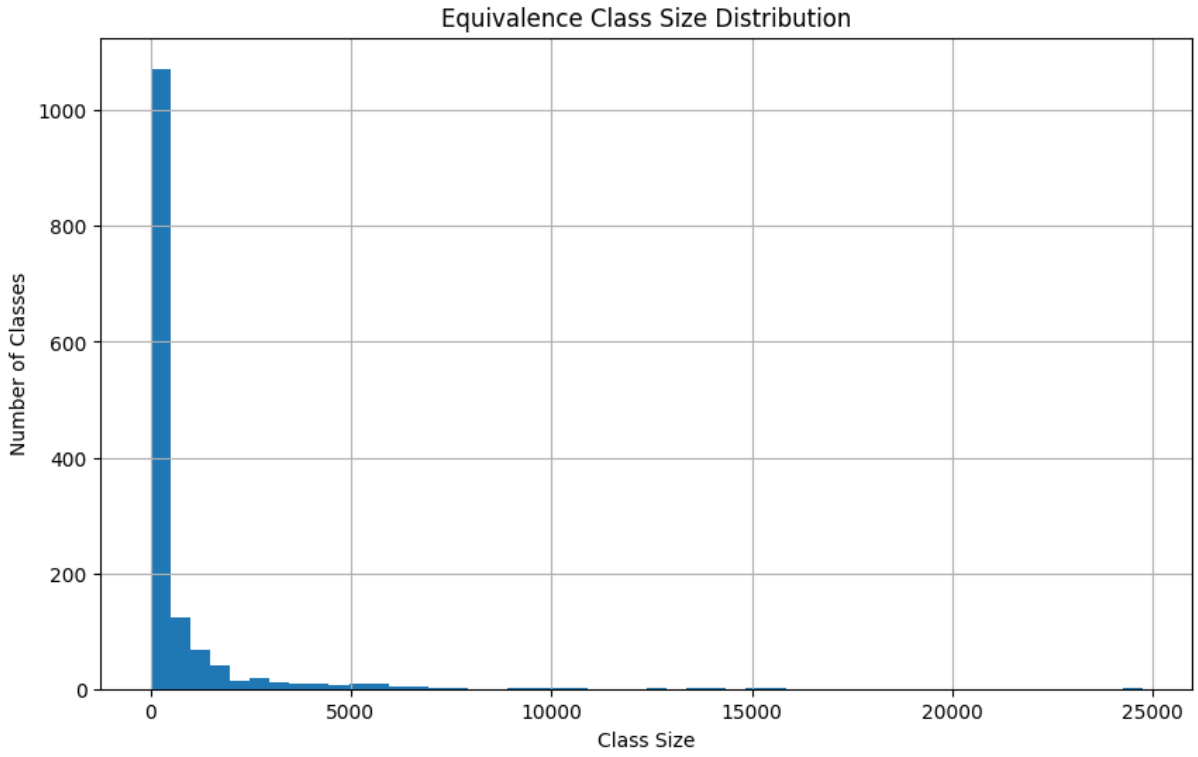

Figura 14 – Enter Caption

Após a aplicação desse método, realizamos uma análise da perda de informação no dataset anonimizado. Dessa forma, obtemos que, após a anonimização, conseguimos eliminar cerca de $99,93\%$ dos CEPs únicos do dataset, e cerca de $99,98\%$ dos aniversários únicos do dataset. Podemos lembrar que, ainda assim, não há garantia de anonimização, porém, ao procurarmos por pessoas que nasceram em uma data específica, e residem em um CEP específico, encontraremos essas pessoas e pelo menos $k = 3$ outras, com as mesmas informações no dataset. Na prática, o tamanho dos grupos obtidos são ainda muito maiores que k, obtendo um resultado melhor que o esperado.

5 Comparação de métodos de anonimização

No geral, a aplicação de ambos os métodos de anonimização foram suficientes para que obtenhamos dados anonimizados.

Em K-Anonymity, separamos os dados em classes de equivalência, com pelo menos $k = 3$ dados que agregavam os mesmos dois primeiros dígitos do CEP e pessoas em uma faixa de idade de um intervalo de 20 anos. Note que, embora tenhamos um k pequeno, na prática os grupos gerados foram muito grandes, o que permitiu que nossa anonimização tivesse mais sucesso.

Por outro lado, seria necessário aplicar k-anonymity em outros dados do dataset para garantir a anonimização total dos indivíduos. Note que, embora esses dados estejam anonimizados, qualquer correlação com outros dados públicos ainda poderia resultar na exposição desses usuários. Por exemplo, se tivermos alguma informação adicional que pode ser ligada ao prefixo do CEP, ainda haveria risco de privacidade

De forma oposta, a aplicação de differential privacy, que tem como objetivo anonimizar os indivíduos a partir da não identificação da entrada e saída de novos indivíduos no dataset parece ter o efeito esperado. Nas comparações feitas nas figuras 11 e 12, isso fica ainda mais evidente, visto que antes os indivíduos eram claramente identificáveis pelo seu CEP, e após a anonimização, conseguimos um grande número de indivíduos que compartilham o mesmo CEP.

Apesar de differential privacy inserir "dados falsos"no dataset, podemos concluir que esse método possui resultados melhores no mundo real pois, por um lado, não se perde informação sobre o CEP (por exemplo), e mantém as propriedades estatísticas do dataset, em média. Além disso, é possível modificar os parâmetros, como o privacy budget ε e δ, assim como fazer uma análise estatística rápida dos dados, e gerar um resultado garantidamente próximo dos seus objetivos, ou seja, manter qualidade dos dados, anonimizar mais os dados ou um equilíbrio entre ambos os objetivos.

6 Considerações finais

Concluímos que o processo de anonimização dos dados é extremamente difícil, ainda mais ao trabalhar com bases de dados muito grandes, com diversos campos.

As preocupações vão além de apenas aplicar uma técnica ou excluir campos específicos dos dados, mas também fazer análises estatísticas de suas distribuições e como esses dados podem ser interpretados em conjunto com outros.

Todavia, atualmente há métodos que, matematicamente, conseguem trabalhar uma garantia forte de anonimização de dados de usuários, ao mesmo tempo que podemos manter qualidade nos dados aplicados. De ambos os métodos aplicados, há um valor de anonimização de ambos e obtivemos bons resultados, dentro do esperado, para ambos, todavia, como visto em aula, Differential Privacy ainda possui uma garantia matemática de anonimização dos dados, o que o torna preferível para anonimizar o dataset.