**Assignment 2**

**Rafael Gouveia e Andrei Fokin Teixeira**

**Segurança e Privacidade / Mestrado em Engenharia e Ciência de Dados**

**Estrutura do doc:**

**Parte 1:** Análise ARX (mascarando dados da base original)

**Parte 2:** Differential privacy (adicionando aleatoriedade às variáveis sensíveis)

**Parte 3:** Criação de dados sintéticos (adicionando “dados fake”)

**Extra:** Comparando análise de dados original com os dados privados (mais *insights*)

**Conclusões**

**Parte 1: Análise ARX**

Para a análise de privacidade foram testados mais de 20 modelos, mas alguns deles não tiveram resultados visíveis no ARX por não haver uma combinação estável de parâmetros (box amarelo em “explore results”), mas em 10 combinações houve resultados, apresentados no [Excel](https://docs.google.com/spreadsheets/d/1Rns4m26OldFESaVtyXU_vdWlVTiIUUCi_BuUFTz1_pI/edit#gid=84698325).

A seguir, um overview dos resultados do melhor modelo.

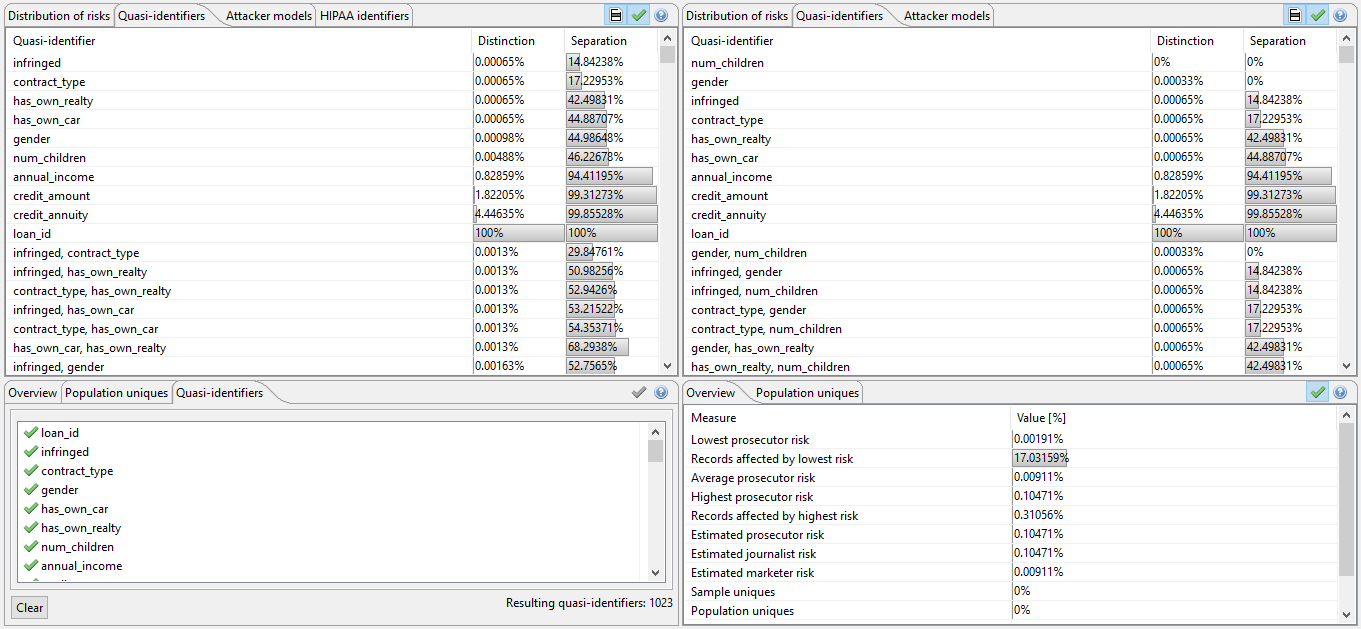
1. ARX é uma ferramenta de anonimização de dados. Para tanto, o primeiro passo é classificar as variáveis como sensíveis, quasi-identifiers, identifiers ou insensíveis. A seguir, uma lista das variáveis consideradas:

- identifiers: nenhum, pois nome e sobrenome estão em colunas separadas

- quasi-identifiers: “gender”, “has\_own\_car”, “has\_own\_realty”, “num\_children”, “income\_type”, “education”, “family\_status”, “housing\_type”, “age”, “occupation\_type”, “num\_family\_members”, “organization\_type”

- sensíveis: “infringed”, “annual\_income”, “credit\_amount”, “goods\_valuation”, “days\_employed”, “past\_avg\_amt\_credit”, “past\_loans\_approved”

- insensíveis: demais variáveis



2. Olhando para os riscos dos quasi-identifiers e sob a ótica das métricas de *distinction* e *separation*, temos a seguinte informação do ARX:

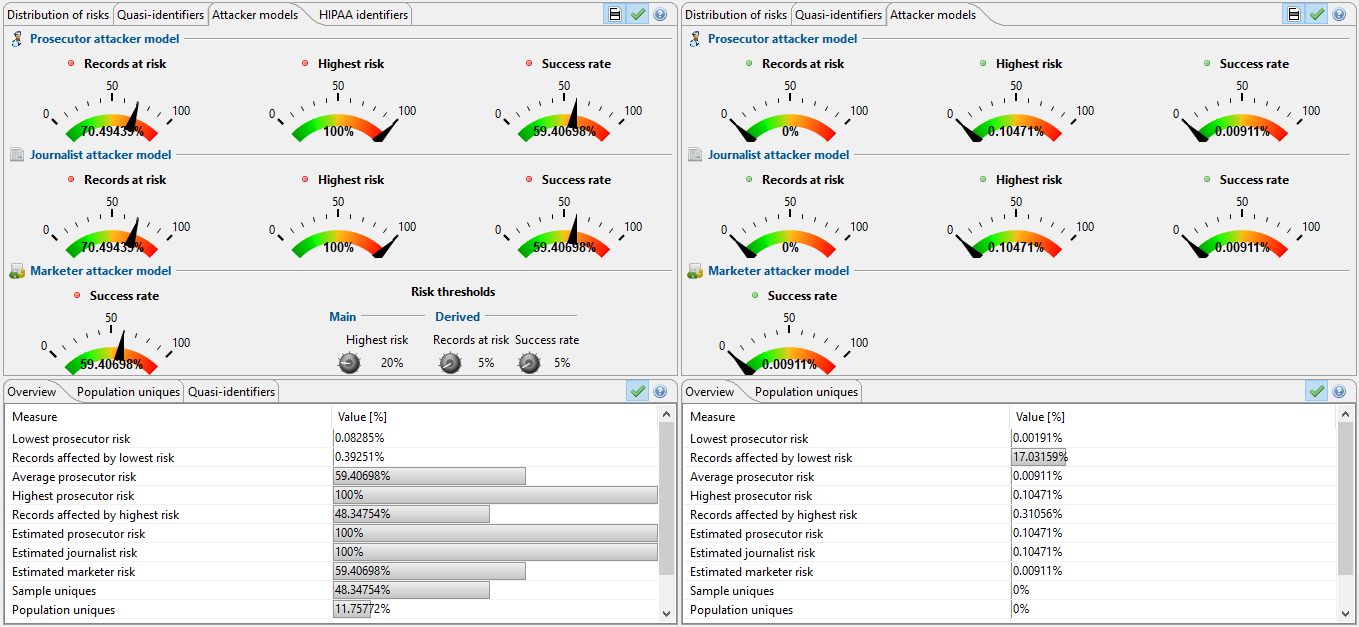
- apenas 17% dos valores são afetados pelo menor risco;

- por *distinction* os resultados são ainda melhores e com identificações quase nulas;

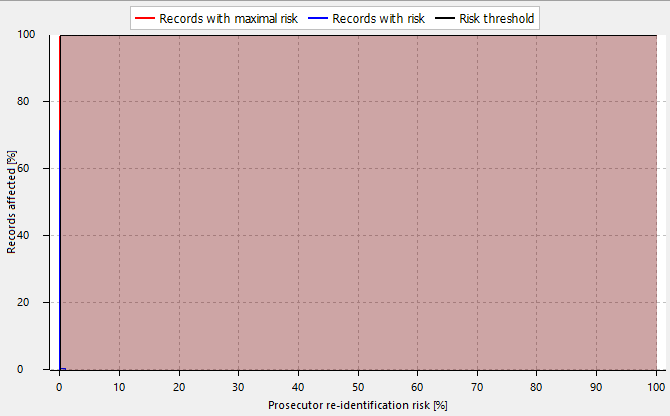
- mesmo por *separation*, há poucos valores na ordem dos 100% de identificação.

3.Riscos de ataque:

Todos eles, tanto o procurador, quanto o jornalista e o marqueteiro, baixaram a quase zero, mais precisamente: 0,00901%. Os valores afetados pelo risco mais baixo estão em 17,03% e pelo alto risco em 0,31%. Tais resultados advém do melhor modelo encontrado no grid search de variáveis alteradas nos testes.



O gráfico abaixo mostra como realmente com quase 0% de risco já se alcança a totalidade da base de dados. A linha azul gera um pico inicial que se aproxima dos 70% da base de dados, mas para um risco nulo. Em seguida, vai a zero rapidamente.



4. Número de hierarquias dos quasi-indentifiers:

- gender: 2

- has\_own\_car: 2

- has\_own\_realty: 2

- num\_children: 3

- income\_type: 3

- education: 3

- family\_status: 3

- housing\_type: 3

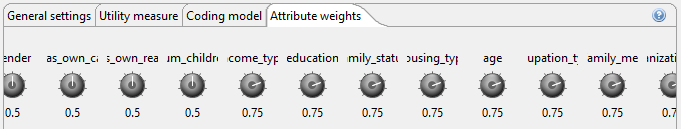
- age: 3

- occupation\_type: 3

- num\_family\_members: 3

- organization\_type: 3

Em relação aos pesos, aplicamos 50% a mais (passando para 75%) de peso nas variáveis que tiveram 3 níveis de hierarquia e mantivemos inalterado em 50% as com duas hierarquias. Decisão discricionária que resultou num ganho de 3 pontos percentuais no nível de anonimização mantendo o resto constante:



5. Requisitos de anonimização:

- limite de supressão: 0%

- utility measure usada: Loss, pois é a principal medida

- coding model: supressão = generalização = 50%

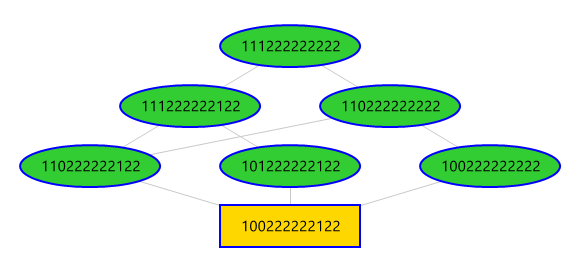
- pesos: apresentado anteriormente

6. Aplicando um modelo de privacidade (após muitas tentativas e erros):

- l-diversity = 2 para “infringed”

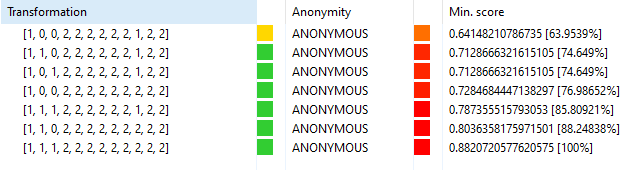
- l-diversity = 3 para demais variáveis sensíveis

A combinação campeã resultante é a que se segue, com muitas variáveis de três hierarquias sendo levadas até o terceiro nível:



7. Comparar privacidade vs. utilidade:

Após uma sete iterações buscando a melhor combinação de hierarquias dos quasi-identifiers, chegou-se a uma base de dados ainda anônima a nível de 63,95%. Logo, a utilidade deste modelo é de 36,05%, considerado nível laranja ainda, mas já fora da zona vermelha, o que indica haver informações disponíveis para uso.

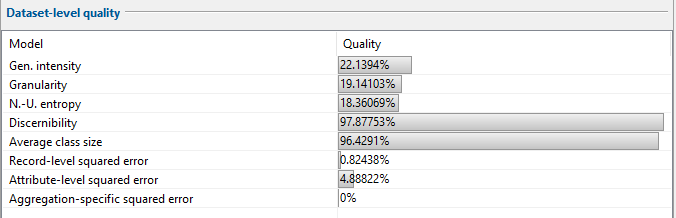


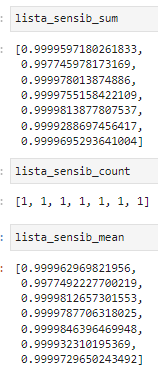
Com relação às métricas de utilidade, temos principalmente:

- intensidade de generalização (quanto mais próximo a 50% melhor): 22,13%

- disponibilidade (quanto mais próximo de 1 melhor): 97,87%

- average class size (quanto mais próximo de 1 melhor): 96,42%



**Parte 2: Differential privacy**

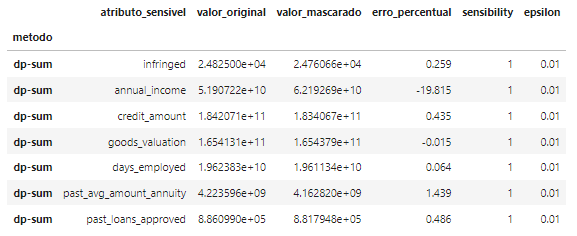
Neste tópico, o ponto importante é garantir que as variáveis sensíveis não tenham o verdadeiro valor em relação a agregados estatísticos como soma, contagem ou média. Se é possível adicionar um pequeno ruído a esses agregados, melhor para a privacidade.

Um único método foi aplicado, o de Laplace (manualmente) e para tal, as sensitividades foram calculadas para cada variável sensível:

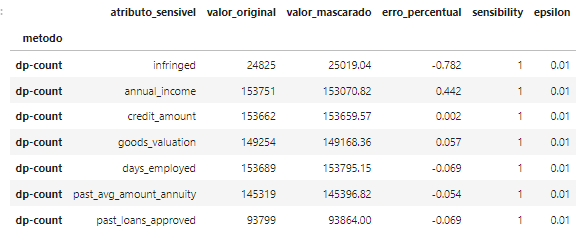
Para o count, é sempre igual a 1 porque todos os valores ficam com pesos iguais, não importa se o verdadeiro valor é o maior ou o menor do grupo de valores.

Ao lado, a sensitividade para SUM, COUNT e MEAN. Todas ficaram com 99,99%, porque alteramos apenas um dado em relação a um total de 15K.

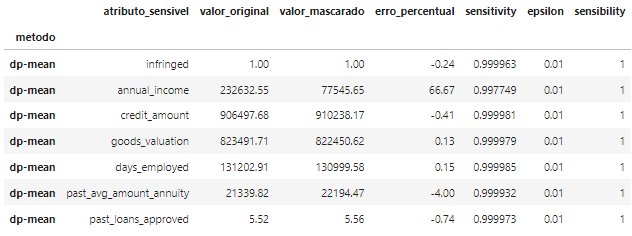
No sum, sensitivity não aparece no código de Laplace, então, por default, será igual a 1:



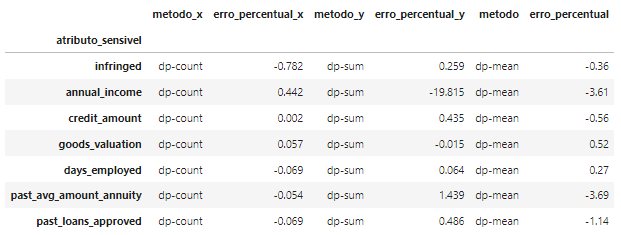
Resultados do count:



Resultados do mean:



Resumo:



Média e soma são os melhores porque geram maiores erros percentuais e quanto medimos o erro médio (em módulo), temos que o indicador da soma é o que gera mais privacidade:



Mais detalhes no documento “*ass\_02\_ex\_02.jpynb*”.

**Parte 3: Criação de dados sintéticos**

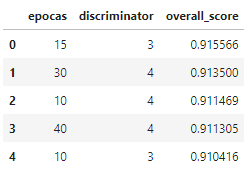
Nesta terceira parte, é adotada mais uma estratégia de anonimização de dados, que é criar novos dados de maneira randômica, mas respeitando a distribuição dos valores dos dados originais. Com essa técnica, melhoramos, por exemplo, tanto a difusão dos agregados da parte de cima quanto a descoberta de dados individuais, que são alguns deles fake.

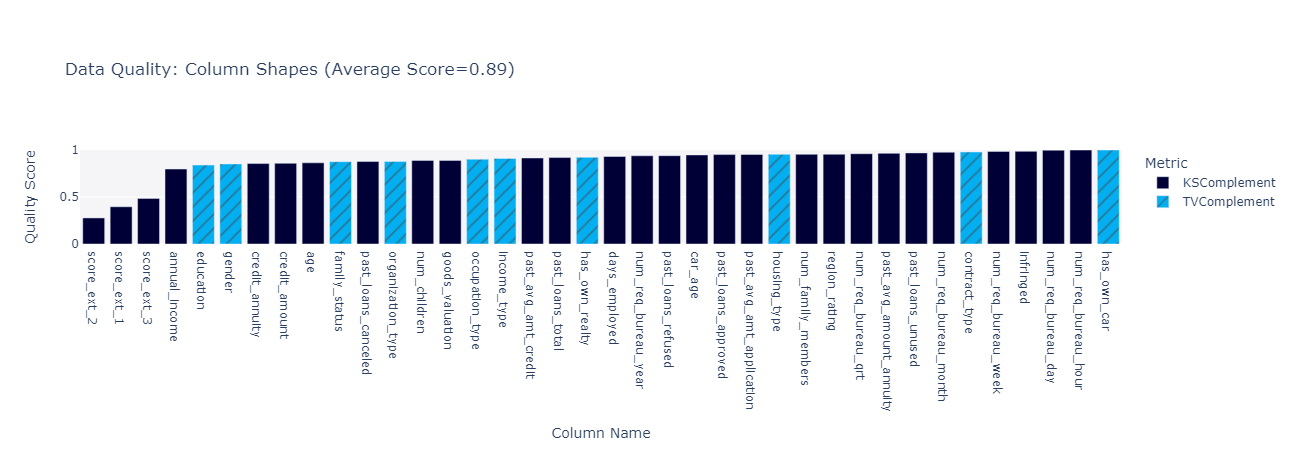
Para criá-los, vamos usar:

- epochs na faixa de 10 a 50, pulando de 5 em 5

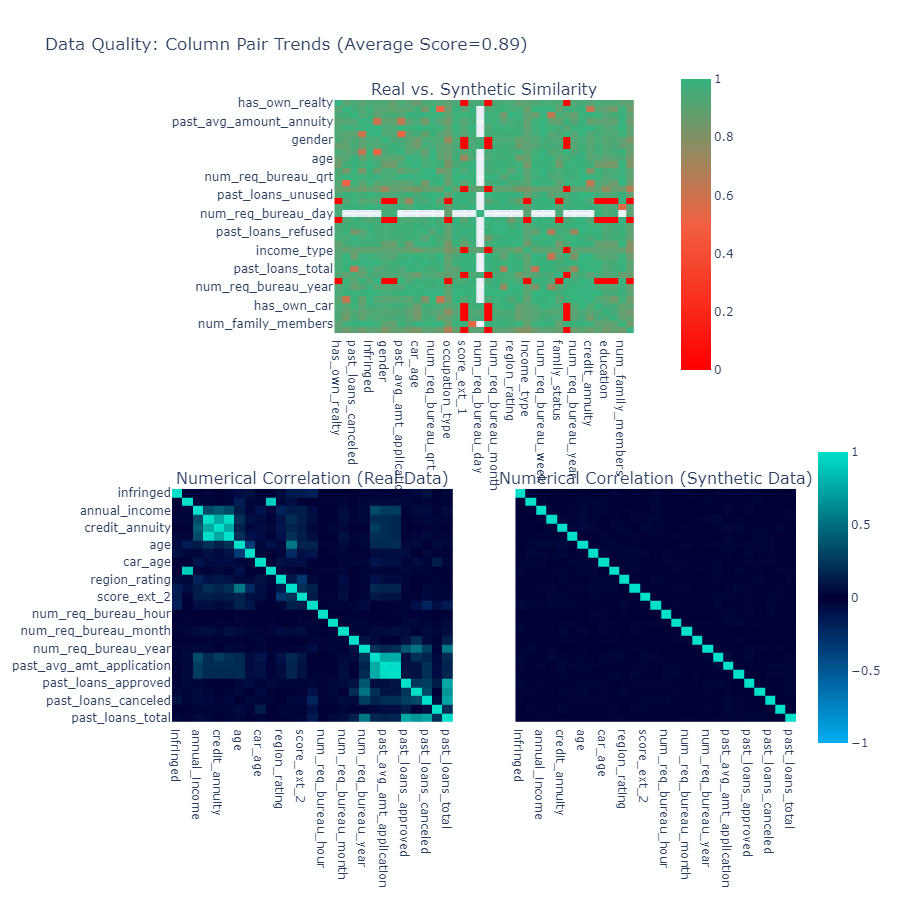
- discriminator na faixa de 1 a 5, de um em um

Os top 5 maiores escores observados foram e com o maior deles foi criada uma base de dados com 15K observações e o melhor modelo foi aquele com 15 épocas e 3 discriminators.

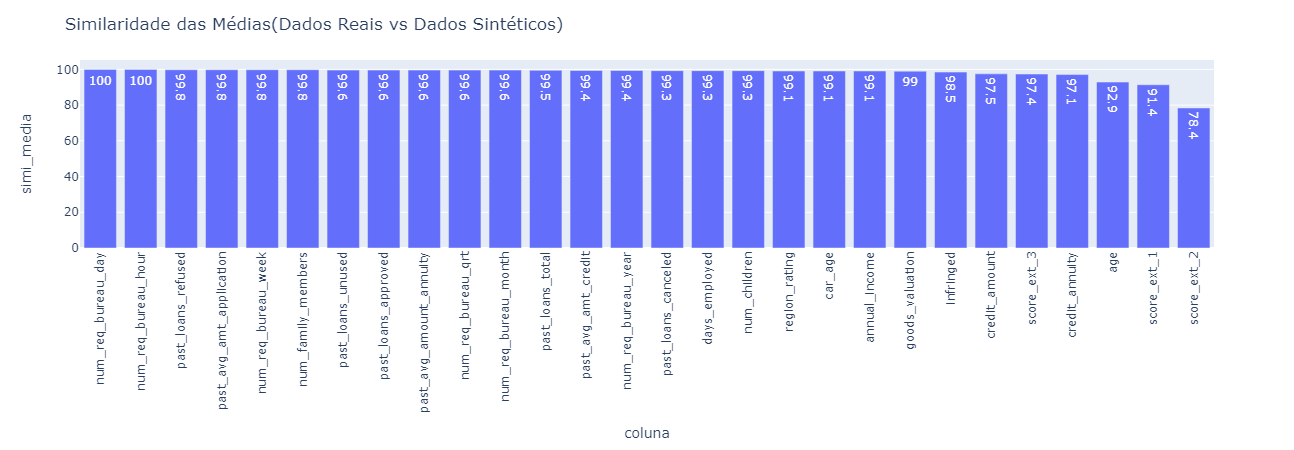
  
  
Os dados resultantes são bem próximos dos dados reais (em azul escuro quantitativas e em azul claro qualitativas): a média de similaridade geral é de 89%:



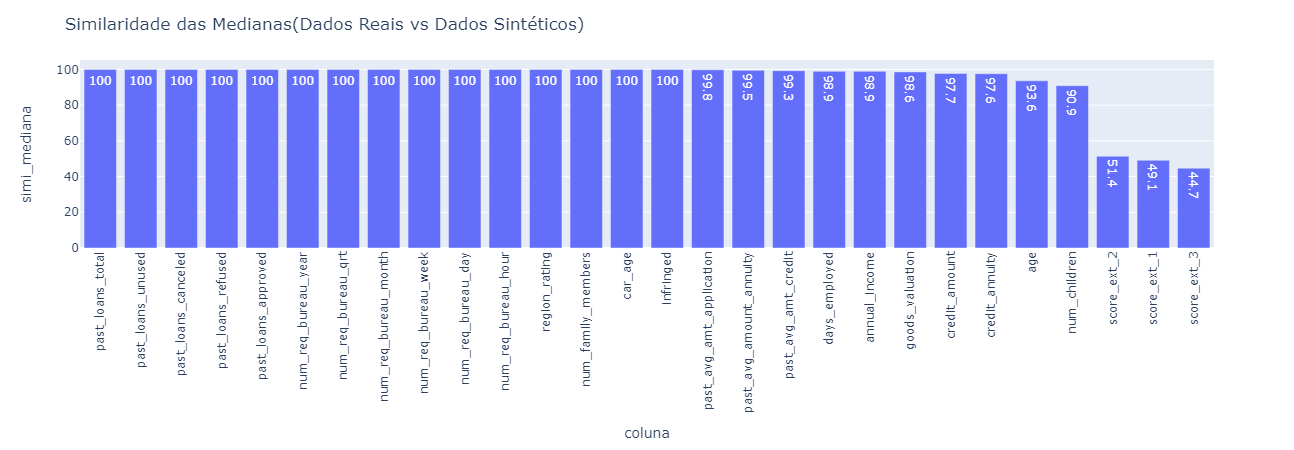
Outra maneira de visualizar a proximidade:



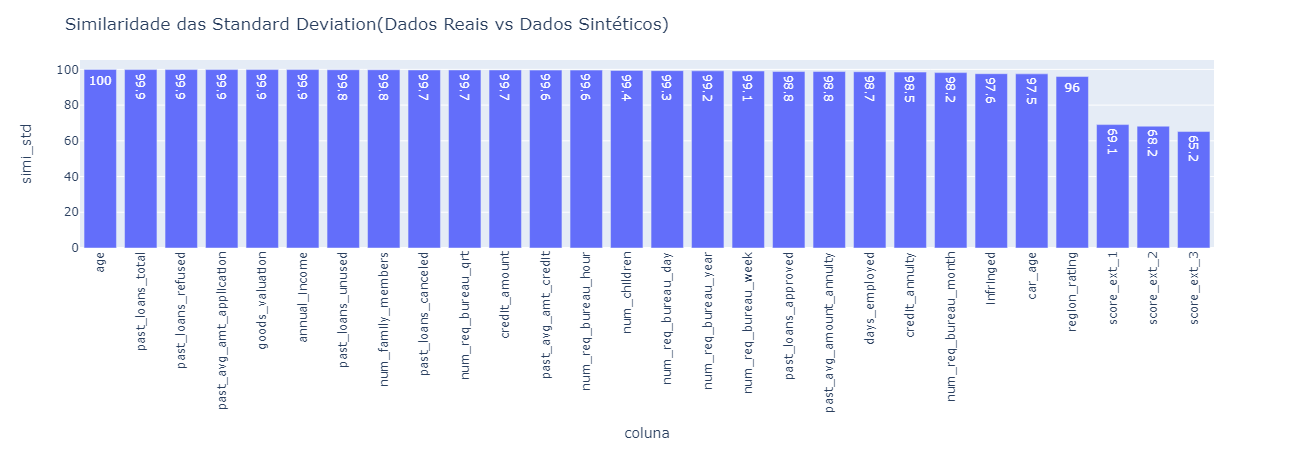
A mesma coisa pode ser dita das médias: elas são ainda mais parecidas, o que indica que podem até refletir o valor verdadeiro, mas não são os dados reais:



A mediana é ainda mais similar, com mais da metade das variáveis com valor idêntico (100% de similaridade):

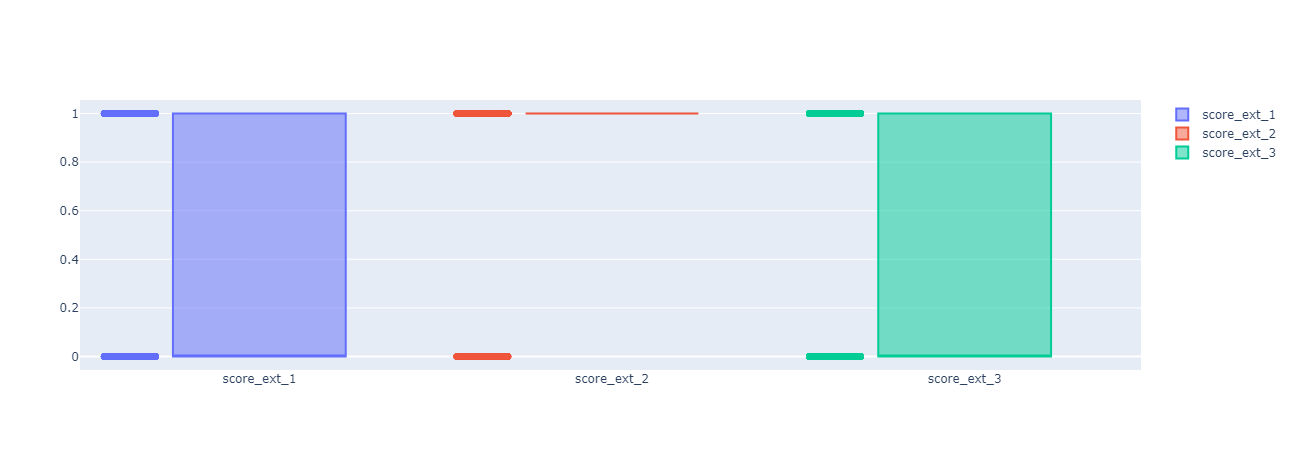


O desvio padrão também surpreende, com metade das variáveis sendo similares acima de 99,5%! Logo, a maneira como elas se distribuem é excelente. Isso ajuda, junto com a média, a manter a mesma distribuição normal dos dados originais.

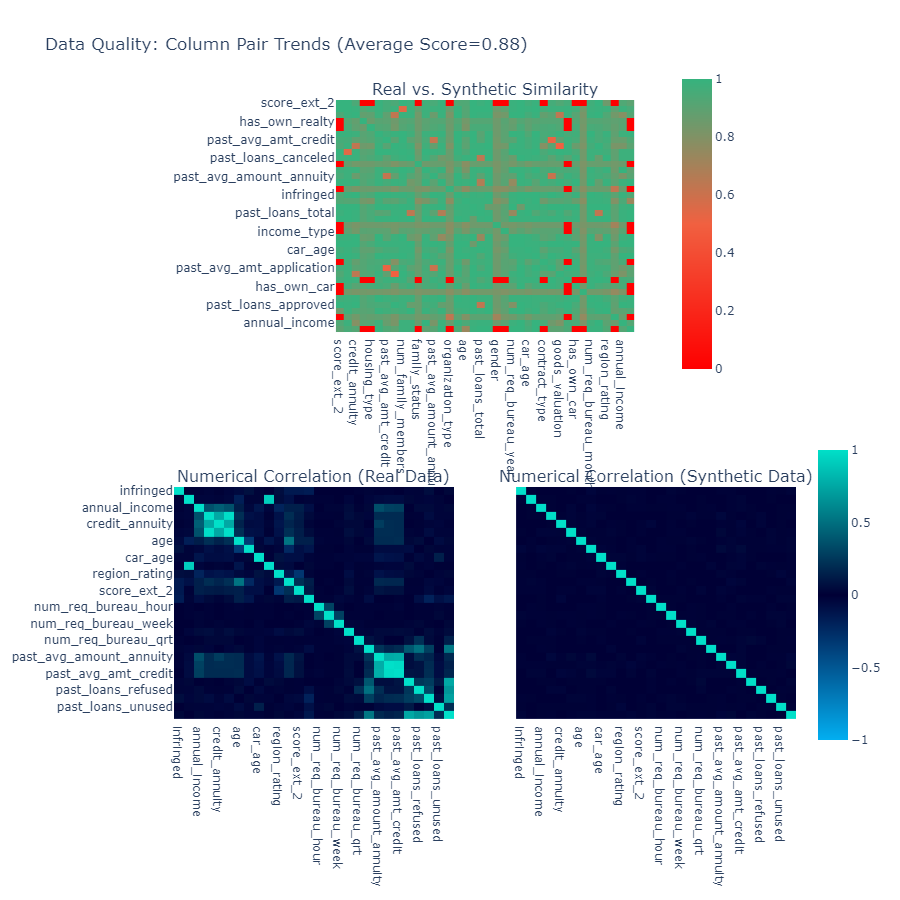


Nas três métricas, as três agregações de “score\_ext” tiveram uma performance de similaridade muito menor comparada às demais. Por isso, vale a pena entender o que aconteceu após a anonimização.

Em realidade, elas deixaram de ter valores contínuos entre 0 e 1 e passaram a se comportar de maneira binária (apenas zero ou apenas um). Portanto, a perda de similaridade está respaldado pelo ganho em anonimização. Melhor ainda, a distribuição entre zeros e uns ficou praticamente dividida meio a meio, conforme pode ser visto na imagem abaixo.



Por fim, olhando para a correlação interna de cada base de dados (correlação entre vars originais e depois correlação entre vars sintéticas), algo interessante acontece: ela vai quase a zero:



Mais detalhes no documento “*ass\_02\_ex\_03.jpynb*”.

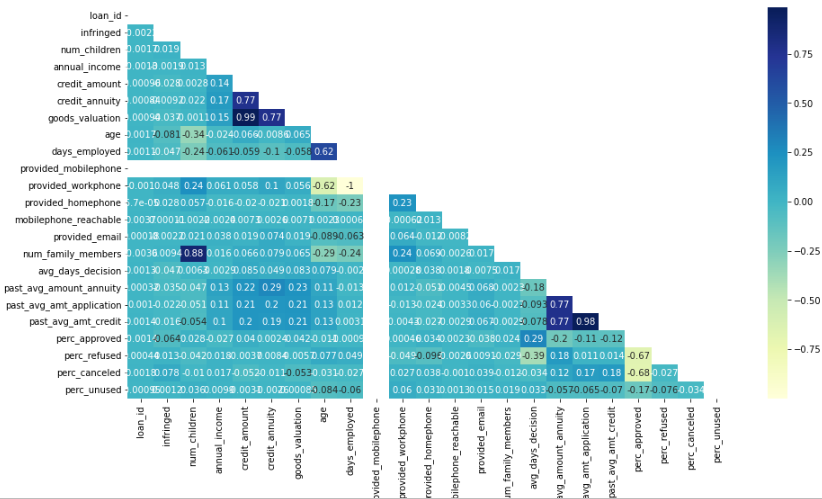
**EXTRA: Análise dos sintéticos**

Para finalizar a terceira parte, é feita uma análise de dados similar à feita inicialmente. A ideia é entender se as correlações das variáveis numéricas e a participação das variáveis qualitativas sobre “infringed” foram alteradas. E caso foram, de que maneira: para mais ou menos privacidade com os dados sintéticos. A hipótese inicial é que os sintéticos ajudam nesse objetivo.

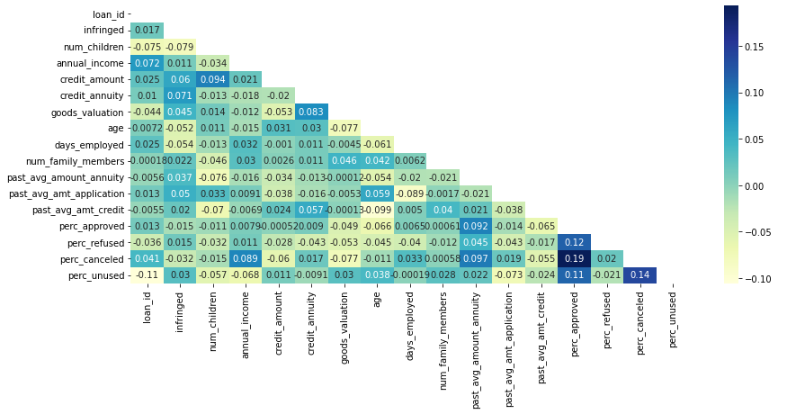
Variáveis quantitativas (Correlações):

- antes, as correlações estavam mais próximas de zero, mas havia muitos outliers com correlações acima de 70% entre variáveis

- depois, as correlações até aumentaram de maneira marginal. O mais importante é não há mais outliers, como pode se notar na escala à lateral



Na criação dos sintéticos “provided\_mobilephone”, “provided\_workphone, “provided\_homephone”, “provided\_email”, “first\_name” e “last\_name” foram retiradas por XXX.



Variáveis quantitativas (Regressão Linear - Causalidades):

- antes, quase todas as variáveis eram estatisticamente significantes, mas com pouco impacto

- depois, quase nenhuma é estatisticamente relevante e o impacto ficou ainda menor (inclusive, a constante foi a segunda mais importante para explicar “infringed”.

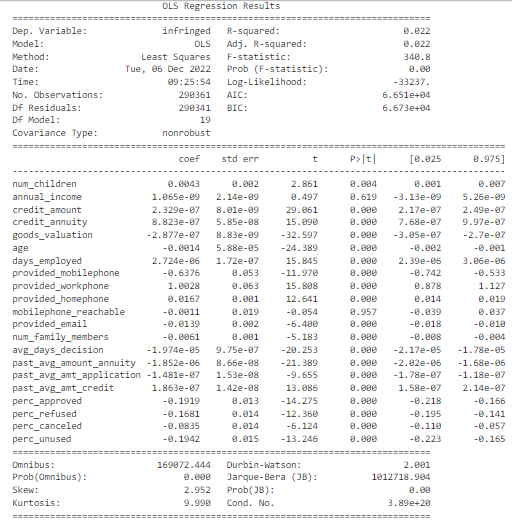
Regressão linear antes:

- valores de t quase sempre maiores a 3 (em módulo)

- valores de P>|t| quase sempre iguais a 0% (ideal é ser menos que 5%)

- coeficientes com valores muito pequenos

- R2 e R2 ajustados baixos e iguais a 2,2% (ideal é próximo a 100%)



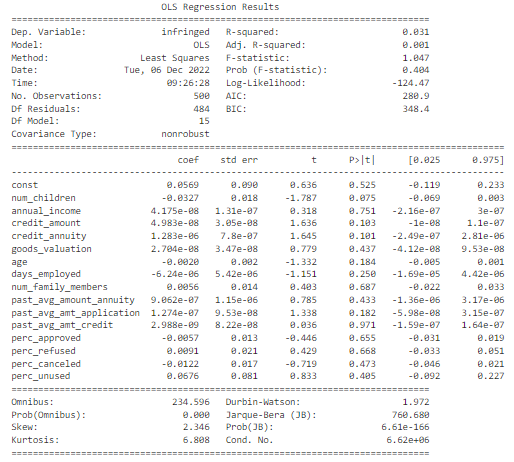
Regressão linear depois:

- valores de t quase sempre menores a 2 (em módulo)

- valores de P>|t| sempre maiores que 10% (ideal é ser menos que 5%)

- coeficientes ainda menores

- R2 = 3,1% (melhor, mas ainda muito ruim) e R2 ajustados baixos e iguais a 0,1%

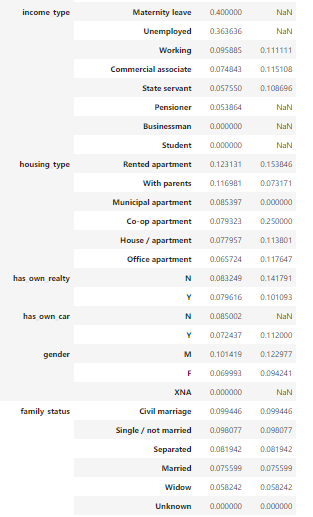


Variáveis qualitativas:

- antes, todas as variáveis tinham alguma relevância para explicar “infringed”

- depois, algumas deixaram de ter relevância (inclusive algumas mais ou menos sensíveis como “desempregado”, “mulher em licença maternidade”, “pessoas vivendo com os pais”...

Exemplo antes (coluna 2) e depois (coluna 3):



Mais detalhes no documento “*analise\_de\_dados\_sp.jpynb*”.