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

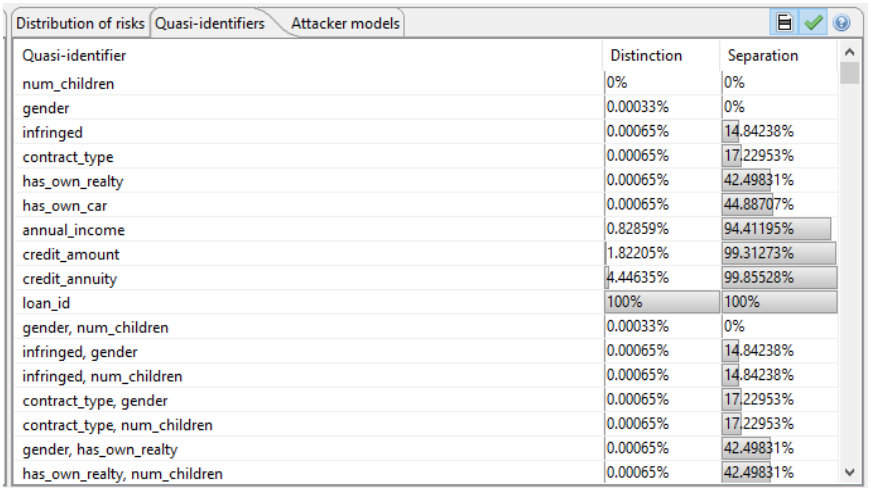
Parte 1: Análise ARX

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.

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

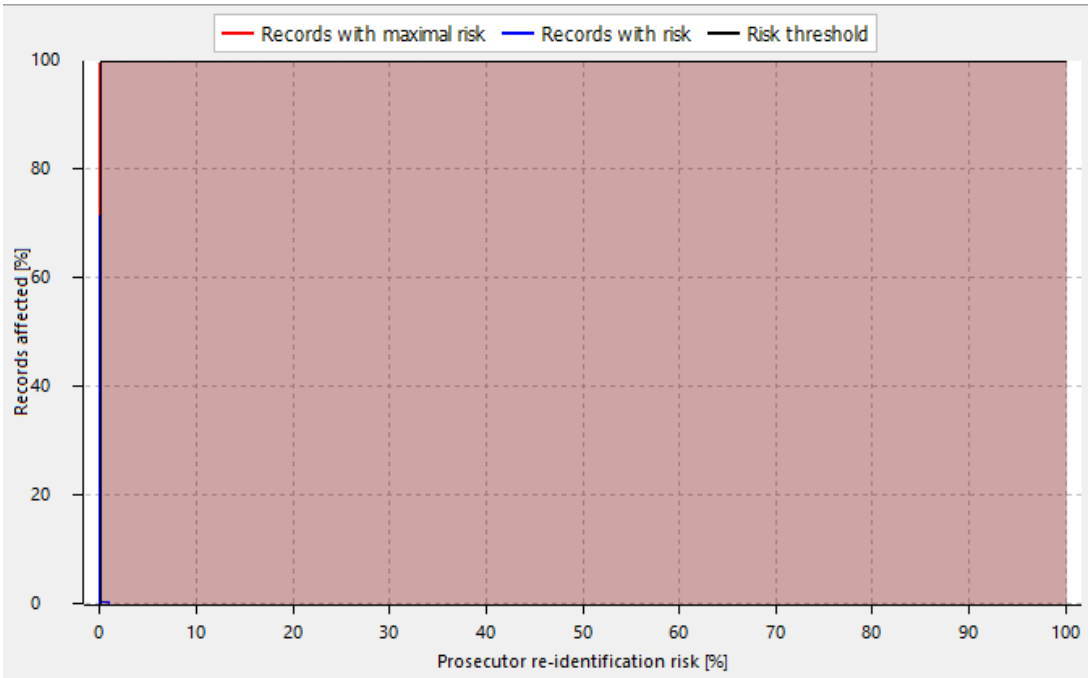

Distribution of risks | Quasi-identifiers | Attacker models

Quasi-identifier	Distinction	Separation
num_children	0%	0%
gender	0.00033%	0%
infringed	0.00065%	14.84238%
contract_type	0.00065%	17.22953%
has_own_realty	0.00065%	42.49831%
has_own_car	0.00065%	44.88707%
annual_income	0.82859%	94.41195%
credit_amount	1.82205%	99.31273%
credit_annuity	4.44635%	99.85528%
loan_id	100%	100%
gender, num_children	0.00033%	0%
infringed, gender	0.00065%	14.84238%
infringed, num_children	0.00065%	14.84238%
contract_type, gender	0.00065%	17.22953%
contract_type, num_children	0.00065%	17.22953%
gender, has_own_realty	0.00065%	42.49831%
has_own_realty, num_children	0.00065%	42.49831%

100% de identificação.

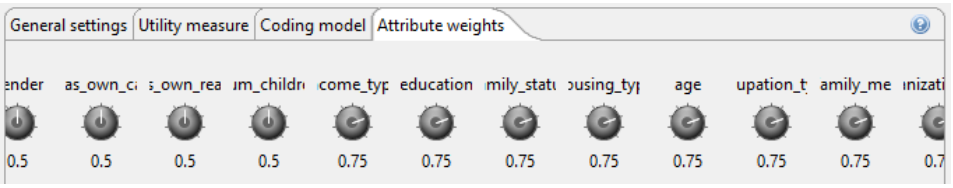
3.Riscos de ataque:
Todos eles, tanto o procurador, quanto o jornalista e o marqueteiro, baixaram a quase zero, mais precisamente: 0,00901%. Os valores afetados pelo risco mais baixo estão em 17,03% e pelo alto risco em 0,31%. Tais resultados advém do melhor modelo encontrado no grid search de variáveis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

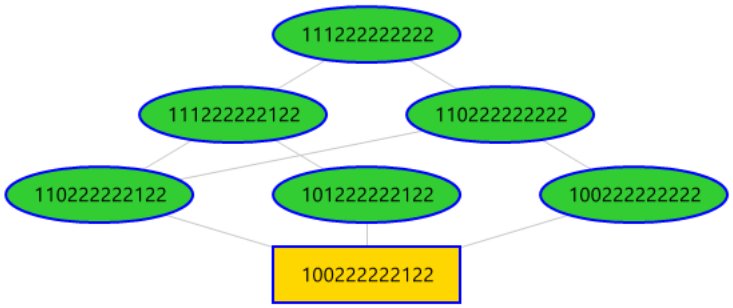
Em relação aos pesos, aplicamos 50% a mais (passando para 75%) de peso nas variáveis que tiveram 3 níveis de hierarquia e mantivemos inalterado em 50% as com duas hierarquias. Decisão discricionária que resultou num ganho de 3 pontos percentuais no nível de anonimização mantendo o resto constante:



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.

Transformation	Anonymity	Min. score
[1, 0, 0, 2, 2, 2, 2, 2, 1, 2, 2]	ANONYMOUS	0.64148210786735 [63.9539%]
[1, 1, 0, 2, 2, 2, 2, 2, 1, 2, 2]	ANONYMOUS	0.7128666321615105 [74.649%]
[1, 0, 1, 2, 2, 2, 2, 2, 1, 2, 2]	ANONYMOUS	0.7128666321615105 [74.649%]
[1, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2]	ANONYMOUS	0.7284684447138297 [76.98652%]
[1, 1, 1, 2, 2, 2, 2, 2, 1, 2, 2]	ANONYMOUS	0.787355515793053 [85.80921%]
[1, 1, 0, 2, 2, 2, 2, 2, 2, 2, 2]	ANONYMOUS	0.8036358175971501 [88.24838%]
[1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2]	ANONYMOUS	0.8820720577620575 [100%]

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%

Dataset-level quality

Model	Quality
Gen. intensity	22.1394%
Granularity	19.14103%
N.-U. entropy	18.36069%
Discernibility	97.87753%
Average class size	96.4291%
Record-level squared error	0.82438%
Attribute-level squared error	4.88822%
Aggregation-specific squared error	0%

<u>Parte 2</u>: 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:

$$sensitividade = max\,|\,dados - dados\_menos\_1\_elemento\,|$$

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.

```
: lista_sensib_sum
: [0.9999597180261833,
   0.997745978173169,
   0.999978013874886,
   0.9999755158422109,
   0.9999813877807537,
   0.9999288697456417,
   0.9999695293641004]
: lista_sensib_count
: [1, 1, 1, 1, 1, 1, 1]
: lista_sensib_mean
: [0.999962969821956,
   0.9977492227700219,
   0.9999812657301553,
   0.9999787706318025,
   0.9999846396469948,
   0.999932310195369,
   0.9999729650243492]
```

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

metodo	atributo_sensivel	valor_original	valor_mascarado	erro_percentual	sensibility	epsilon
dp-sum	infringed	2.482500e+04	2.476066e+04	0.259	1	0.01
dp-sum	annual_income	5.190722e+10	6.219269e+10	-19.815	1	0.01
dp-sum	credit_amount	1.842071e+11	1.834067e+11	0.435	1	0.01
dp-sum	goods_valuation	1.654131e+11	1.654379e+11	-0.015	1	0.01
dp-sum	days_employed	1.962383e+10	1.961134e+10	0.064	1	0.01
dp-sum	past_avg_amount_annuity	4.223596e+09	4.162820e+09	1.439	1	0.01
dp-sum	past_loans_approved	8.860990e+05	8.817948e+05	0.486	1	0.01

Resultados do count:

metodo	atributo_sensivel	valor_original	valor_mascarado	erro_percentual	sensibility	epsilon
dp-count	infringed	24825	25019.04	-0.782	1	0.01
dp-count	annual_income	153751	153070.82	0.442	1	0.01
dp-count	credit_amount	153662	153659.57	0.002	1	0.01
dp-count	goods_valuation	149254	149168.36	0.057	1	0.01
dp-count	days_employed	153689	153795.15	-0.069	1	0.01
dp-count	past_avg_amount_annuity	145319	145396.82	-0.054	1	0.01
dp-count	past_loans_approved	93799	93864.00	-0.069	1	0.01

Resultados do mean:

metodo	atributo_sensivel	valor_original	valor_mascarado	erro_percentual	sensitivity	epsilon	sensibility
dp-mean	infringed	1.00	1.00	-0.24	0.999963	0.01	1
dp-mean	annual_income	232632.55	77545.65	66.67	0.997749	0.01	1
dp-mean	credit_amount	906497.68	910238.17	-0.41	0.999981	0.01	1
dp-mean	goods_valuation	823491.71	822450.62	0.13	0.999979	0.01	1
dp-mean	days_employed	131202.91	130999.58	0.15	0.999985	0.01	1
dp-mean	past_avg_amount_annuity	21339.82	22194.47	-4.00	0.999932	0.01	1
dp-mean	past_loans_approved	5.52	5.56	-0.74	0.999973	0.01	1

Resumo:

atributo_sensivel	metodo_x	erro_percentual_x	metodo_y	erro_percentual_y	metodo	erro_percentual
infringed	dp-count	-0.782	dp-sum	0.259	dp-mean	-0.36
annual_income	dp-count	0.442	dp-sum	-19.815	dp-mean	-3.61
credit_amount	dp-count	0.002	dp-sum	0.435	dp-mean	-0.56
goods_valuation	dp-count	0.057	dp-sum	-0.015	dp-mean	0.52
days_employed	dp-count	-0.069	dp-sum	0.064	dp-mean	0.27
past_avg_amount_annuity	dp-count	-0.054	dp-sum	1.439	dp-mean	-3.69
past_loans_approved	dp-count	-0.069	dp-sum	0.486	dp-mean	-1.14

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:

```
Count: 0.21
Sum: 3.22
Mean: 1.45
```

Mais detalhes no documento "*ass_02_ex_02.jpynb*".

Parte 3: Criação de dados sintéticos

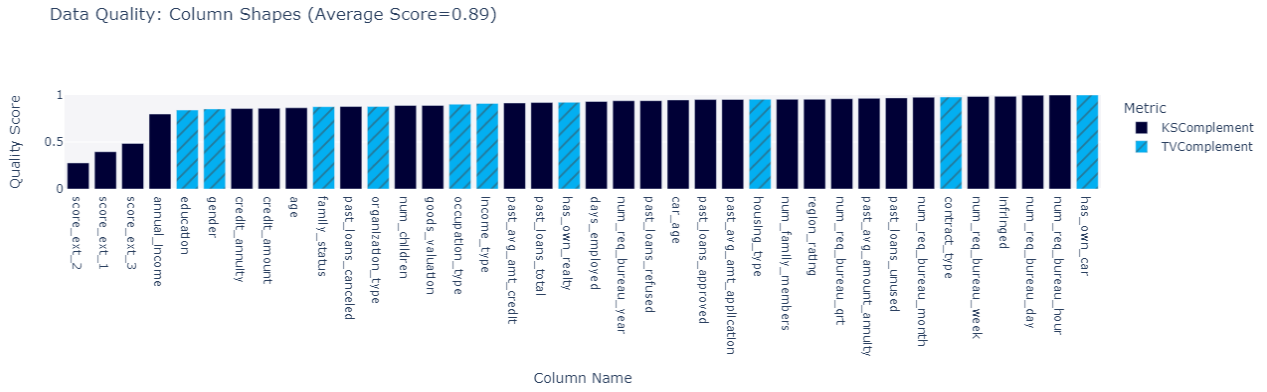
Nesta terceira parte, é adotada mais uma estratégia de anonimização de dados, que é criar novos dados de maneira randômica, mas respeitando a distribuição dos valores dos dados originais. Com essa técnica, melhoramos, por exemplo, tanto a difusão dos agregados da parte de cima quanto a descoberta de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

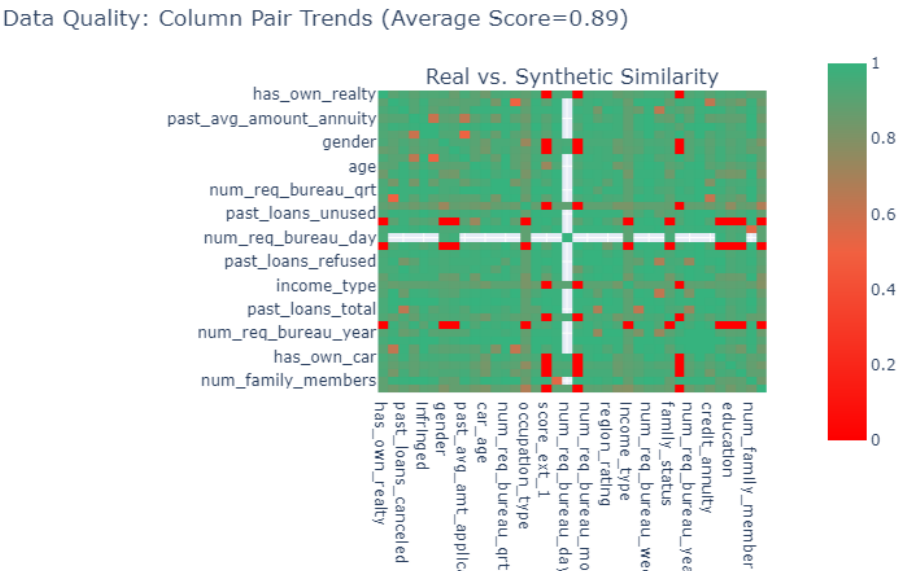
Os top 5 maiores escores observados foram e com o maior deles foi criada uma base de dados com 15K observações e o melhor modelo foi aquele com 15 épocas e 3 discriminators.

	epocas	discriminator	overall_score
0	15	3	0.915566
1	30	4	0.913500
2	10	4	0.911469
3	40	4	0.911305
4	10	3	0.910416

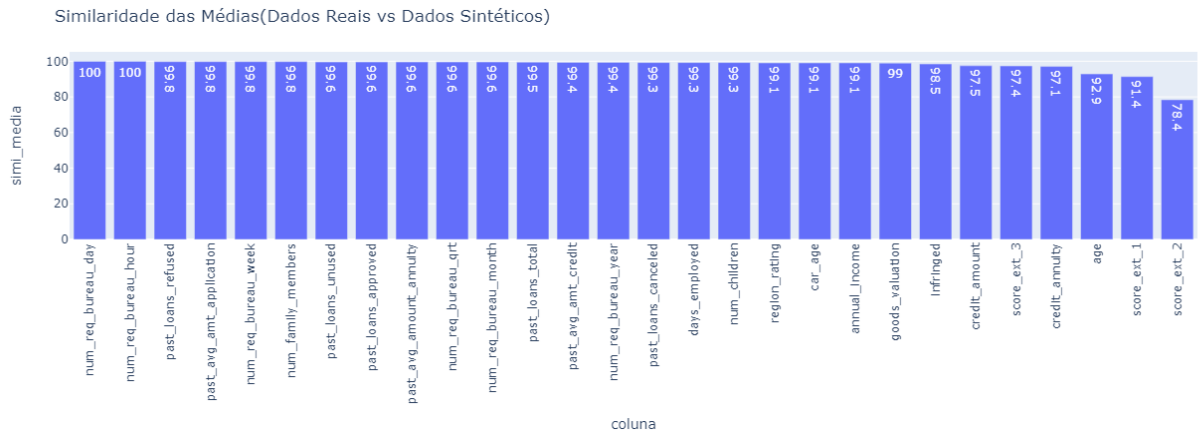
Os dados resultantes são bem próximos dos dados reais (em azul escuro quantitativas e em azul claro qualitativas): a média de similaridade geral é de 89%:



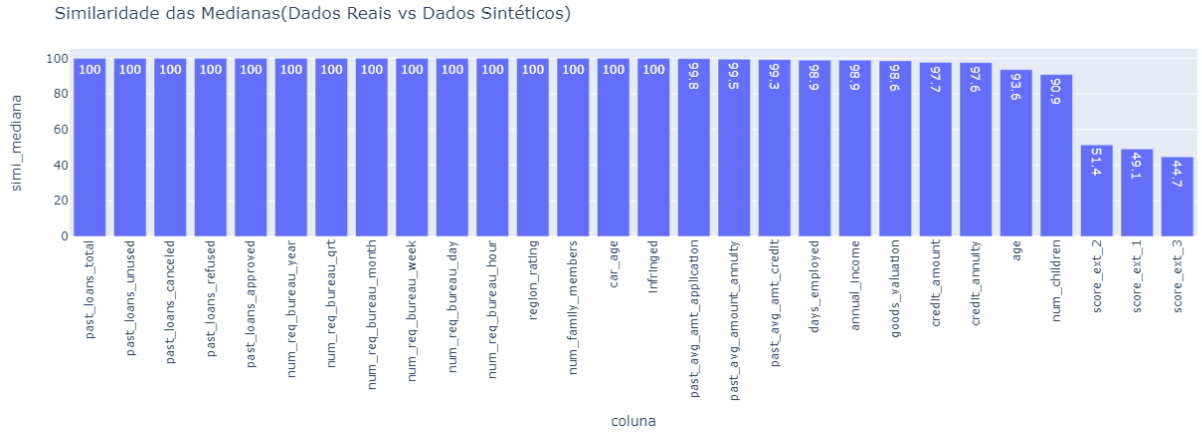
Outra maneira de visualizar a proximidade:



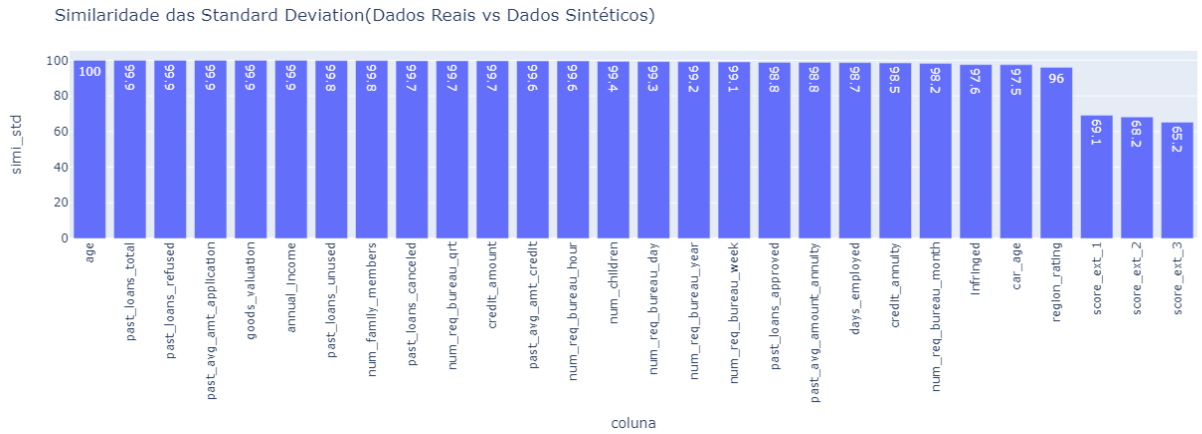
A mesma coisa pode ser dita das <u>médias</u>: elas são ainda mais parecidas, o que indica que podem até refletir o valor verdadeiro, mas não são os dados reais:



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



O <u>desvio padrão</u> 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.

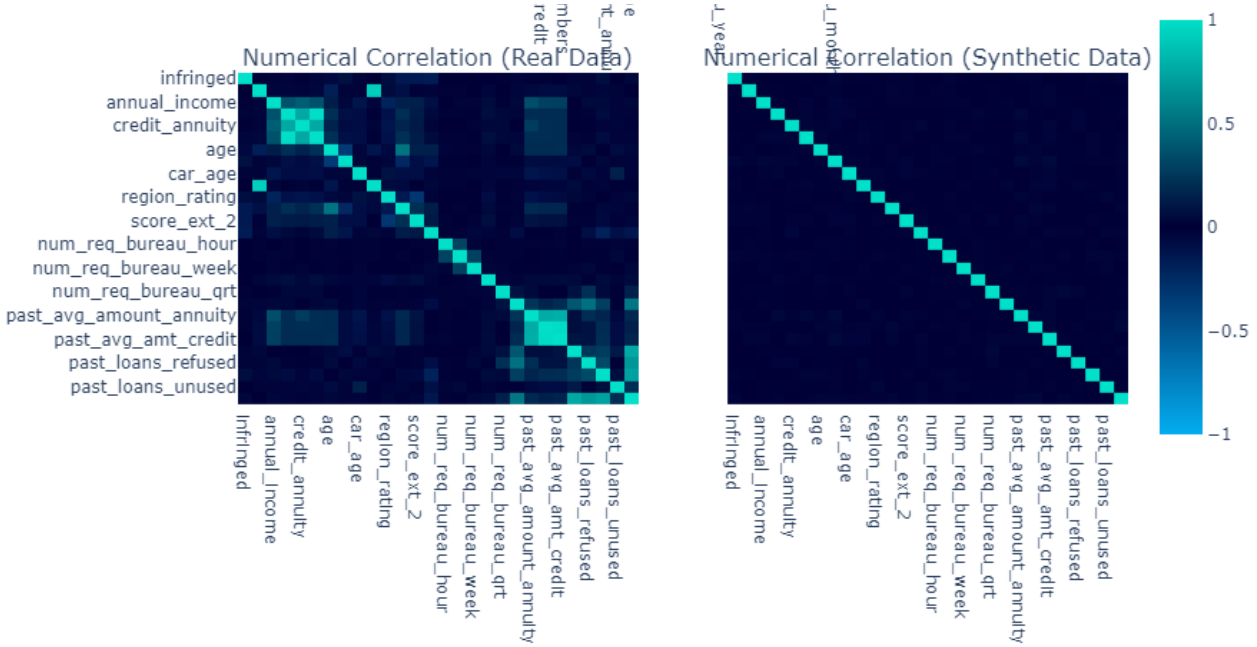


Nas três métricas, as três agregações de “score_ext” tiveram uma performance de similaridade muito menor comparada às demais. Por isso, vale a pena entender o que aconteceu após a anonimização.

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 <u>correlação interna de cada base de dados</u> (correlação entre vars originais e depois correlação entre vars sintéticas), algo interessante acontece: ela vai quase a zero:



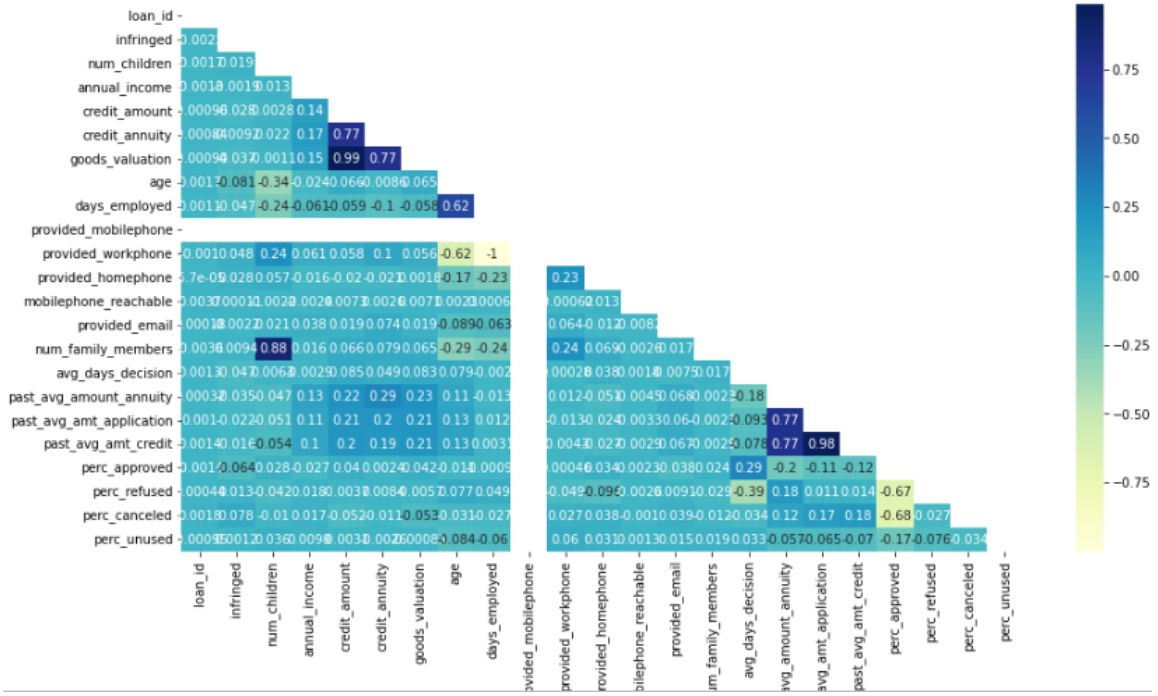
Mais detalhes no documento “*ass_02_ex_03.jpynb*”.

<u>EXTRA:</u> Análise dos sintéticos

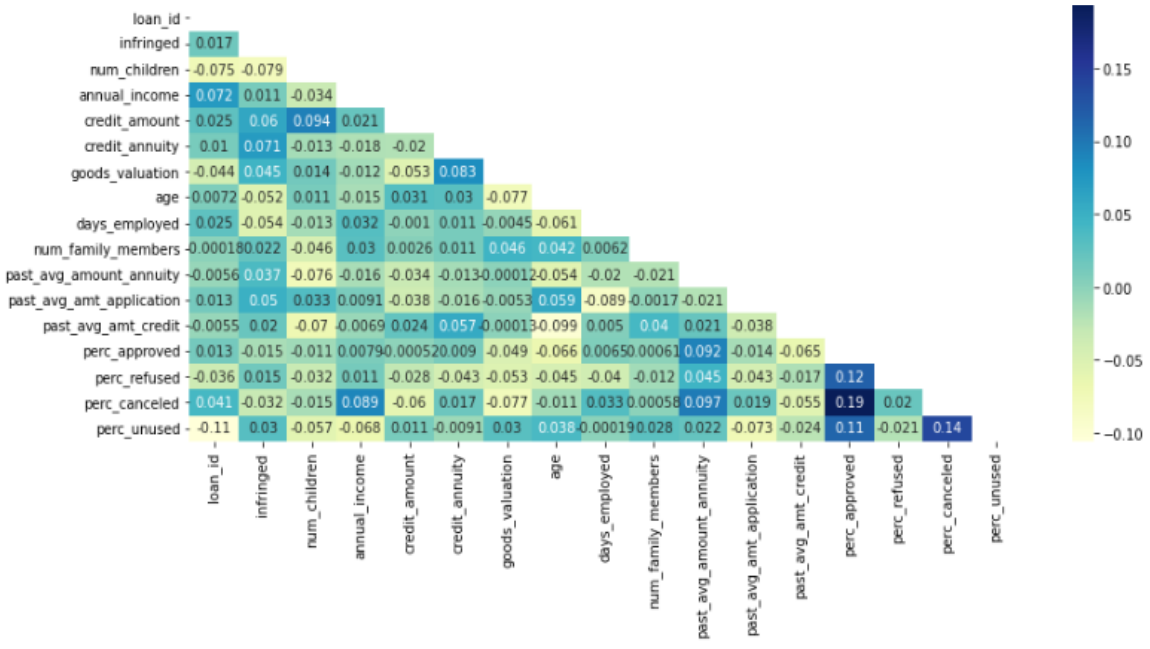
Para finalizar a terceira parte, é feita uma análise de dados similar à feita inicialmente. A ideia é entender se as correlações das variáveis numéricas e a participação das variáveis qualitativas sobre “infringed” foram alteradas. E caso foram, de que maneira: para mais ou menos privacidade com os dados 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
- R^2 e R^2 ajustados baixos e iguais a 2,2% (ideal é próximo a 100%)

OLS Regression Results

Dep. Variable:	infringed	R-squared:	0.022
Model:	OLS	Adj. R-squared:	0.022
Method:	Least Squares	F-statistic:	340.8
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	09:25:54	Log-Likelihood:	-33237.
No. Observations:	290361	AIC:	6.651e+04
Df Residuals:	290341	BIC:	6.673e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
num_children	0.0043	0.002	2.861	0.004	0.001	0.007
annual_income	1.065e-09	2.14e-09	0.497	0.619	-3.13e-09	5.26e-09
credit_amount	2.329e-07	8.01e-09	29.061	0.000	2.17e-07	2.49e-07
credit_annuity	8.823e-07	5.85e-08	15.090	0.000	7.68e-07	9.97e-07
goods_valuation	-2.877e-07	8.83e-09	-32.597	0.000	-3.05e-07	-2.7e-07
age	-0.0014	5.88e-05	-24.389	0.000	-0.002	-0.001
days_employed	2.724e-06	1.72e-07	15.845	0.000	2.39e-06	3.06e-06
provided_mobilephone	-0.6376	0.053	-11.970	0.000	-0.742	-0.533
provided_workphone	1.0028	0.063	15.808	0.000	0.878	1.127
provided_homephone	0.0167	0.001	12.641	0.000	0.014	0.019
mobilephone_reachable	-0.0011	0.019	-0.054	0.957	-0.039	0.037
provided_email	-0.0139	0.002	-6.400	0.000	-0.018	-0.010
num_family_members	-0.0061	0.001	-5.183	0.000	-0.008	-0.004
avg_days_decision	-1.974e-05	9.75e-07	-20.253	0.000	-2.17e-05	-1.78e-05
past_avg_amount_annuity	-1.852e-06	8.66e-08	-21.389	0.000	-2.02e-06	-1.68e-06
past_avg_amt_application	-1.481e-07	1.53e-08	-9.655	0.000	-1.78e-07	-1.18e-07
past_avg_amt_credit	1.863e-07	1.42e-08	13.086	0.000	1.58e-07	2.14e-07
perc_approved	-0.1919	0.013	-14.275	0.000	-0.218	-0.166
perc_refused	-0.1681	0.014	-12.360	0.000	-0.195	-0.141
perc_canceled	-0.0835	0.014	-6.124	0.000	-0.110	-0.057
perc_unused	-0.1942	0.015	-13.246	0.000	-0.223	-0.165

Omnibus:	169072.444	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1012718.904
Skew:	2.952	Prob(JB):	0.00
Kurtosis:	9.990	Cond. No.	3.89e+20

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
- R^2 = 3,1% (melhor, mas ainda muito ruim) e R^2 ajustados baixos e iguais a 0,1%

OLS Regression Results

Dep. Variable:	infringed	R-squared:	0.031
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.047
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.404
Time:	09:26:28	Log-Likelihood:	-124.47
No. Observations:	500	AIC:	280.9
Df Residuals:	484	BIC:	348.4
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0569	0.090	0.636	0.525	-0.119	0.233
num_children	-0.0327	0.018	-1.787	0.075	-0.069	0.003
annual_income	4.175e-08	1.31e-07	0.318	0.751	-2.16e-07	3e-07
credit_amount	4.983e-08	3.05e-08	1.636	0.103	-1e-08	1.1e-07
credit_annuity	1.283e-06	7.8e-07	1.645	0.101	-2.49e-07	2.81e-06
goods_valuation	2.704e-08	3.47e-08	0.779	0.437	-4.12e-08	9.53e-08
age	-0.0020	0.002	-1.332	0.184	-0.005	0.001
days_employed	-6.24e-06	5.42e-06	-1.151	0.250	-1.69e-05	4.42e-06
num_family_members	0.0056	0.014	0.403	0.687	-0.022	0.033
past_avg_amount_annuity	9.062e-07	1.15e-06	0.785	0.433	-1.36e-06	3.17e-06
past_avg_amt_application	1.274e-07	9.53e-08	1.338	0.182	-5.98e-08	3.15e-07
past_avg_amt_credit	2.988e-09	8.22e-08	0.036	0.971	-1.59e-07	1.64e-07
perc_approved	-0.0057	0.013	-0.446	0.655	-0.031	0.019
perc_refused	0.0091	0.021	0.429	0.668	-0.033	0.051
perc_canceled	-0.0122	0.017	-0.719	0.473	-0.046	0.021
perc_unused	0.0676	0.081	0.833	0.405	-0.092	0.227

Omnibus:	234.596	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	760.680
Skew:	2.346	Prob(JB):	6.61e-166
Kurtosis:	6.808	Cond. No.	6.62e+06

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):

income type	Maternity leave	0.400000	NaN
	Unemployed	0.363636	NaN
	Working	0.095885	0.111111
	Commercial associate	0.074843	0.115108
	State servant	0.057550	0.108696
	Pensioner	0.053864	NaN
	Businessman	0.000000	NaN
	Student	0.000000	NaN
housing type	Rented apartment	0.123131	0.153846
	With parents	0.116981	0.073171
	Municipal apartment	0.085397	0.000000
	Co-op apartment	0.079323	0.250000
	House / apartment	0.077957	0.113801
	Office apartment	0.065724	0.117647
has own realty	N	0.083249	0.141791
	Y	0.079616	0.101093
has own car	N	0.085002	NaN
	Y	0.072437	0.112000
gender	M	0.101419	0.122977
	F	0.069993	0.094241
	XNA	0.000000	NaN
family status	Civil marriage	0.099446	0.099446
	Single / not married	0.098077	0.098077
	Separated	0.081942	0.081942
	Married	0.075599	0.075599
	Widow	0.058242	0.058242
	Unknown	0.000000	0.000000

Mais detalhes no documento “*analise_de_dados_sp.jpynb*”.