Segurança e Privacidade - Projeto 2

Miguel Marques 2017266263
Tiago Fernandes 2017242428

2020/2021

1 Preparação e Análise dos dados

Relativamente à construção do Dataset, utilizamos o site mockaroo.com para gerar os dados gerais como nomes, mail, género, endereço de IP, número de telemóvel, número da Segurança Social e número do cartão de crédito. Depois decidimos ser nós a criar as colunas aleatoriamente com as informações de trabalho, salário bruto anual, idade, número de filhos, cidade e signos. Deste modo, o nosso dataset original é constituido por 13 colunas e 1000 linhas.

Nas tabelas seguintes, agrupamos os dados para uma análise mais simples, sendo que os resultados apresentados são as médias.

Analysis by age

age	income_per_year	number_kids
23	50560.900000	1.850000
24	47311.242424	1.909091
25	51602.483871	1.677419
26	54566.285714	2.238095
27	39126.045455	1.863636
28	41129.684211	2.052632
29	47792.968750	2.000000
30	42328.960000	2.440000
31	36774.961538	1.615385
32	50397.382353	1.882353
33	47980.000000	2.235294
34	56580.277778	1.555556
35	46414.680000	2.080000
36	41603.200000	1.680000
37	50578.333333	2.000000
38	48759.714286	2.357143
39	39590.578947	2.000000
40	55183.111111	1.388889
41	49715.333333	2.233333
42	46109.742857	2.114286
43	47043.000000	1.931034
44	43582.954545	2.227273
45	47463.421053	2.473684
46	46249.833333	1.933333
47	49076.400000	1.866667
48	45500.000000	1.433333
49	49735.880000	1.840000
50	45680.444444	2.037037
51	43346.448276	1.689655
52	43579.454545	2.136364
53	45500.038462	1.730769
54	51984.969697	1.787879
55	44640.920000	2.120000
56	40173.035714	2.107143
57	42264.935484	1.774194
58	47649.861111	1.805556
59	47712.346154	1.692308
60	46152.840000	2.600000

Figure 1: Análise do dataset original

Analysis by gender

gender	income_per_year	number_kids
Female	46230.110220	2.02004
Male	46998.105788	1.89022

Analysis by city

city	income_per_year	number_kids
Aveiro	50495.992593	2.103704
Braga	45636.715517	1.965517
Castelo Branco	47734.728682	1.984496
Coimbra	48377.017241	1.965517
Faro	47572.396396	1.891892
Lisboa	46197.371901	1.983471
Porto	43783.720000	1.768000
Viseu	43477.285714	1.959184

Figure 2: Análise do dataset original

Analysis by job

job	income_per_year	number_kids
Accountant	50019.400000	1.776923
Doctor	80684.625000	1.947368
Engineer	50370.781250	1.968750
Lawyer	53983.263889	2.131944
Manager	49782.144828	2.027586
Teacher	27477.147651	1.885906
Waiter	15228.519737	1.934211

Analysis by signs

signs	income_per_year	number_kids
Aquarius	48025.326087	1.793478
Aries	47358.603960	1.881188
Cancer	47310.526316	2.039474
Capricorn	45997.771429	2.014286
Gemini	46777.067568	2.162162
Leo	46765.565789	2.289474
Libra	43984.583333	1.802083
Pisces	45042.026667	2.013333
Sagittarius	48642.576471	1.941176
Scorpio	46077.930233	1.837209
Taurus	46456.712329	1.821918
Virgo	46800.666667	1.979167

Figure 3: Análise do dataset original

2 Syntactic Models

2.1 Tipos de atributos

Atributos	Tipo de Atributos
Name	Identifying
email	Identifying
gender	Quasi-identifier
ip_address	Identifying
city	Quasi-identifier
phone	Identifying
job	Quasi-identifier
ssn	Identifying
credit_card	Identifying
income_per_year	Sensitive
age	Quasi-identifier
number_kids	Sensitive
signs	Quasi-identifier

Figure 4: Análise dos atributos do dataset

2.2 Hierarquia usada para os Quasi-Identifiers

2.2.1 Género

Aqui utilizamos sets com apenas 2 níveis, ou seja, ou mostra o género, ou não mostra.

2.2.2 Cidade

Aqui utilizamos sets com 4 níveis como podemos observar na tabela em baixo.

Level-0	Level-1	Level-2	Level-3
{Aveiro}	{Av, Br}	{Av, Br, CasB, Cbr}	*
{Braga}	{Av, Br}	{Av, Br, CasB, Cbr}	*
{Castelo Branco}	{CasB, Cbr}	{Av, Br, CasB, Cbr}	*
{Coimbra}	{CasB, Cbr}	{Av, Br, CasB, Cbr}	*
{Faro}	{Fr, Lx}	{Fr, Lx, Pt, Vs}	*
{Lisboa}	{Fr, Lx}	{Fr, Lx, Pt, Vs}	*
{Porto}	{Pt, Vs}	{Fr, Lx, Pt, Vs}	*
{Viseu}	{Pt, Vs}	{Fr, Lx, Pt, Vs}	*

Table 1: Níveis de hierarquia na coluna das cidades

2.2.3 Trabalho

Da mesma maneira que fizemos com a coluna das cidades, aqui também utilizamos uma hierarquia de sets como podemos ver na seguinte tabela.

Level-0	Level-1	Level-2	Level-3
{Accountant}	{Acc,Doc}	{Acc,Doc,Eng,Law}	*
{Doctor}	{Acc,Doc}	{Acc,Doc,Eng,Law}	*
{Engineer}	{Eng,Law}	{Acc,Doc,Eng,Law}	*
{Lawyer}	{Eng,Law}	{Acc,Doc,Eng,Law}	*
{Manager}	{Man,Tch}	{Man,Tch,Waiter}	*
{Teacher}	{Man,Tch}	{Man,Tch,Waiter}	*
{Waiter}	{Waiter}	{Man,Tch,Waiter}	*
{Viseu}	{Pt, Vs}	{Fr, Lx, Pt, Vs}	*

Table 2: Níveis de hierarquia na coluna dos trabalhos

2.2.4 Idade

Na idade decidimos utilizar intervalos, como a gama de idades varia entre 23 e 60 anos, decidimos utilizar 4 níveis em que o primeiro começa com 3 intervalos, convergindo num único intervalo no último nível, como podemos observar abaixo. É de notar que o nível 0 são os valores contínuos, ou seja, sem estar em intervalos.

Level-0	Level-1	Level-2	Level-3
	[23,38[	[23,53[	[23,61[
	[38,53[	[23,53[	[23,61[
	[53,61[	[53,61[	[23,61[

Table 3: Níveis de hierarquia na coluna das idades

2.2.5 Signos

No que toca aos signos decidimos aplicar sets como fizemos com os Trabalhos e as Cidades. Com um total de 4 níveis como podemos observar em baixo.

Level-0	Level-1	Level-2	Level-3
Aquarius	{Aq,Ar}	{Aq,Ar,Can,Cap}	*
Aries	{Aq,Ar}	{Aq,Ar,Can,Cap}	*
Cancer	{Can,Cap}	{Aq,Ar,Can,Cap}	*
Capricorn	{Can,Cap}	{Aq,Ar,Can,Cap}	*
Gemini	{Ge,Le}	{Li,Pi,Ge.Le}	*
Leo	{Ge,Le}	{Li,Pi,Ge.Le}	*
Libra	{Li,Pi}	{Li,Pi,Ge.Le}	*
Pisces	{Li,Pi}	{Li,Pi,Ge.Le}	*
Sagittarius	{Sa,Sco}	{Ta,Vi,Sa,Sco}	*
Scorpio	{Sa,Sco}	{Ta,Vi,Sa,Sco}	*
Taurus	{Ta,Vi}	{Ta,Vi,Sa,Sco}	*
Virgo	{Ta,Vi}	{Ta,Vi,Sa,Sco}	*

Table 4: Níveis de hierarquia na coluna dos signos

Nota: Os pesos dos atributos género, nascimento, cidade, emprego e signos estão todos a 0.5 e limitamos a supression a 10%.

2.2.6 Distinction and Separation

Quasi-identifier	Distinction	Separation
gender	0.2%	50.04985%
job	0.7%	85.74034%
city	0.8%	87.49009%
signs	1.2%	91.63163%
age	3.8%	97.35816%
gender, job	1.4%	92.86927%
gender, city	1.6%	93.75716%
gender, signs	2.4%	95.84184%
job, city	5.6%	98.1952%
gender, age	7.6%	98.68529%
job, signs	8.4%	98.81642%
city, signs	9.6%	98.93734%
job, age	26,00%	99.62763%
age, city	29.3%	99.65926%
age, signs	41.3%	99.79139%
gender, job, city	11.2%	99.1017%
gender, job, signs	16.8%	99.40961%
gender, city, signs	19.2%	99.47287%
gender, job, age	45.1%	99.81602%
gender, age, city	48,00%	99.82823%
job, city, signs	51.4%	99.84104%
gender, age, signs	62.1%	99.9005%
job, age, city	79.3%	99.94955%
age, city, signs	86.4%	99.97077%
job, age, signs	87.5%	99.97337%
gender, job, city, signs	69.5%	99.92192%
gender, job, age, city	88.8%	99.97598%
gender, age, city, signs	92.5%	99.98478%
gender, job, age, signs	95,00%	99.98999%
job, age, city, signs	97.8%	99.9954%
gender, job, age, city, signs	99.4%	99.9988%

Figure 5: Distiction and Separation

2.2.7 Privacy Risks

Measure	Value [%]
Lowest prosecutor risk	50,00%
Records affected by lowest risk	1.2%
Average prosecutor risk	99.4%
Highest prosecutor risk	100,00%
Records affected by highest risk	98.8%
Estimated prosecutor risk	100,00%
Estimated journalist risk	100,00%
Estimated marketer risk	99.4%
Sample uniques	98.8%
Population uniques	97.6144%
Population model	ZAYATZ
Quasi-identifiers	age, city, gender, job, signs

Figure 6: Risks Table

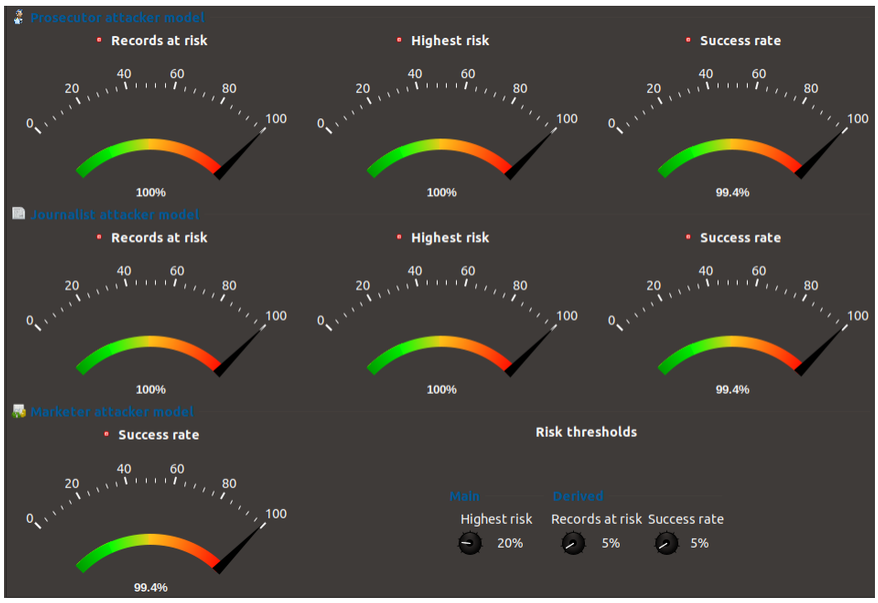

Figure 7: Risk Charts

2.3 [One] Usando o modelo de Privacidade L-Diversity, L=4 em todos os Sensitive

2.3.1 Distinction and Separation

Quasi-identifier	Distinction	Separation
job	0.20661%	49.49597%
city	0.20661%	49.96624%
gender	0.20661%	50.04636%
age	0.30992%	64.93735%
signs	0.30992%	66.70178%
job, city	0.41322%	74.70985%
gender, job	0.41322%	74.78634%
gender, city	0.41322%	75.00128%
job, age	0.61983%	82.1876%
gender, age	0.61983%	82.30362%
age, city	0.61983%	82.49058%
job, signs	0.61983%	83.14011%
gender, signs	0.61983%	83.36082%
city, signs	0.61983%	83.36852%
age, signs	0.92975%	88.22207%
gender, job, city	0.82645%	87.38366%
gender, job, age	1.23967%	91.03002%
job, age, city	1.23967%	91.07254%
gender, age, city	1.23967%	91.17083%
job, city, signs	1.23967%	91.57679%
gender, job, signs	1.23967%	91.58683%
gender, city, signs	1.23967%	91.6458%
job, age, signs	1.8595%	93.98305%
gender, age, signs	1.8595%	94.03732%
age, city, signs	1.8595%	94.11616%
gender, job, age, city	2.47934%	95.50753%
gender, job, city, signs	2.47934%	95.76713%
gender, job, age, signs	3.71901%	96.95189%
job, age, city, signs	3.71901%	96.96407%
gender, age, city, signs	3.71901%	97.02112%
gender, job, age, city, signs	7.02479%	98.44881%

Figure 8: Distiction and Separation

2.3.2 Privacy Risks

Measure	Value [%]
Lowest prosecutor risk	3.84615%
Records affected by lowest risk	2.68595%
Average prosecutor risk	7.02479%
Highest prosecutor risk	20,00%
Records affected by highest risk	0.51653%
Estimated prosecutor risk	20,00%
Estimated journalist risk	20,00%
Estimated marketer risk	7.02479%
Sample uniques	0,00%
Population uniques	0,00%
Population model	DANKAR
Quasi-identifiers	age, city, gender, job, signs

Figure 9: Risk Table

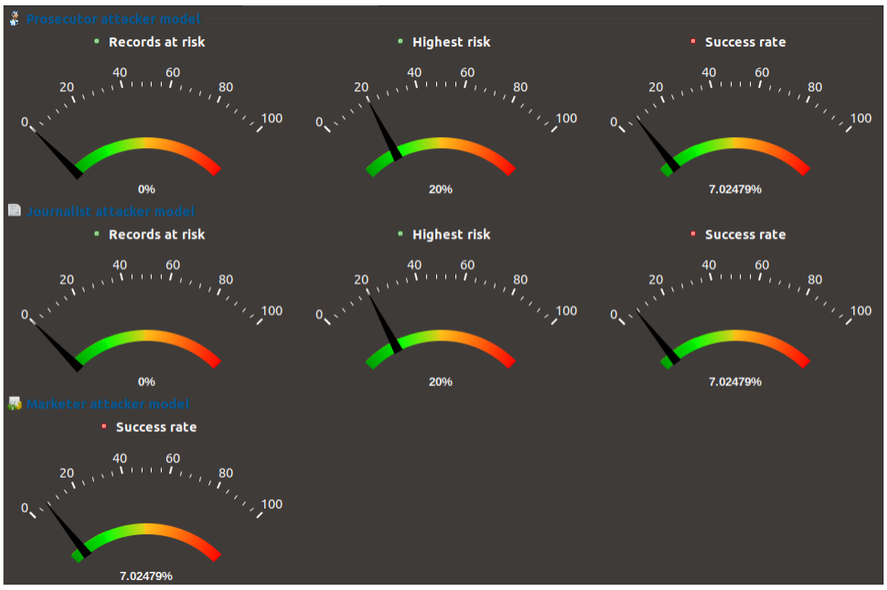

Figure 10: Risk Charts

2.3.3 Análises sob os dados transformados

Analysis by gender

gender	income_per_year	number_kids
*	51511.437500	2.093750
Female	45922.613497	1.997955
Male	46994.471816	1.901879

Analysis by city

city	income_per_year	number_kids
*	51511.437500	2.093750
{Aveiro, Braga, Castelo Branco, Coimbra}	47913.719828	2.002155
{Faro, Lisboa, Porto, Viseu}	45108.222222	1.902778

Analysis by age

age	income_per_year	number_kids
*	51511.437500	2.093750
[23, 38[	46651.083770	1.913613
[38, 53[	46998.856369	1.983740
[53, 61[	45176.119816	1.958525

Analysis by job

job	income_per_year	number_kids
*	51511.437500	2.093750
{Accountant, Doctor, Engineer, Lawyer}	59358.379439	1.940187
{Manager, Teacher, Waiter}	30507.568129	1.963048

Analysis by signs

signs	income_per_year	number_kids
*	51511.437500	2.093750
{Aquarius, Aries, Cancer, Capricorn}	47100.199377	1.925234
{Gemini, Leo, Libra, Pisces}	45027.945338	2.016077
{Sagittarius, Scorpio, Taurus, Virgo}	47153.735119	1.913690

Figure 11: Análises

Nota: Foram suprimidos 32 dos 1000 dados.

2.4 [Two] Usando o modelo de Privacidade K-Anonimity com K=10 e L-Diversity com L=8 no income-per-year e L=2 no number-kids

2.4.1 Distinction and Separation

Quasi-identifier	Distinction	Separation
city	0.1%	0,00%
job	0.2%	49.46627%
gender	0.2%	50.04985%
age	0.3%	65.12032%
signs	0.3%	66.71051%
job, city	0.2%	49.46627%
gender, city	0.2%	50.04985%
age, city	0.3%	65.12032%
city, signs	0.3%	66.71051%
gender, job	0.4%	74.78218%
job, age	0.6%	82.31471%
gender, age	0.6%	82.5039%
job, signs	0.6%	83.14454%
gender, signs	0.6%	83.3958%
age, signs	0.9%	88.36476%
gender, job, city	0.4%	74.78218%
job, age, city	0.6%	82.31471%
gender, age, city	0.6%	82.5039%
job, city, signs	0.6%	83.14454%
gender, city, signs	0.6%	83.3958%
age, city, signs	0.9%	88.36476%
gender, job, age	1.2%	91.15015%
gender, job, signs	1.2%	91.61321%
job, age, signs	1.8%	94.07628%
gender, age, signs	1.8%	94.15716%
gender, job, age, city	1.2%	91.15015%
gender, job, city, signs	1.2%	91.61321%
job, age, city, signs	1.8%	94.07628%
gender, age, city, signs	1.8%	94.15716%
gender, job, age, signs	3.6%	97.03904%
gender, job, age, city, signs	3.6%	97.03904%

Figure 12: Distiction and Separation

2.4.2 Privacy Risks

Measure	Value [%]
Lowest prosecutor risk	2,00%
Records affected by lowest risk	5,00%
Average prosecutor risk	3.6%
Highest prosecutor risk	6.25%
Records affected by highest risk	1.6%
Estimated prosecutor risk	6.25%
Estimated journalist risk	6.25%
Estimated marketer risk	3.6%
Sample uniques	0,00%
Population uniques	0,00%
Population model	DANKAR
Quasi-identifiers	age, city, gender, job, signs

Figure 13: Risk Table

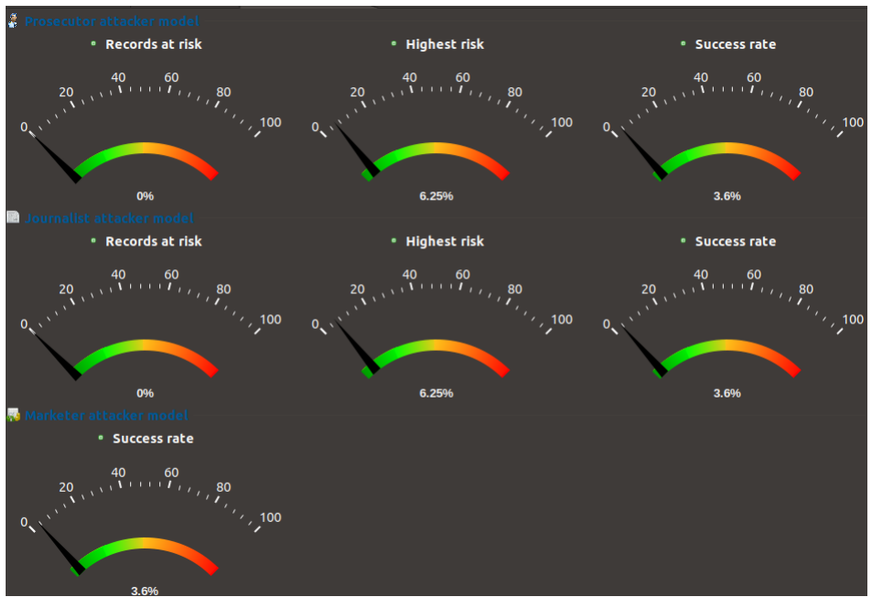

Figure 14: Risk Charts

2.4.3 Análises sob os dados transformados

Analysis by gender

gender	income_per_year	number_kids
Female	46230.110220	2.02004
Male	46998.105788	1.89022

Analysis by city

city	income_per_year	number_kids
{Aveiro, Braga, Castelo Branco, Coimbra, Faro, Lisboa, Porto, Viseu}	46614.876	1.955

Analysis by age

age	income_per_year	number_kids
[23, 38[	47023.931122	1.943878
[38, 53[	46619.502646	1.978836
[53, 61[	45910.100000	1.934783

Analysis by job

job	income_per_year	number_kids
{Accountant, Doctor, Engineer, Lawyer}	59544.467509	1.960289
{Manager, Teacher, Waiter}	30554.352018	1.948430

Analysis by signs

signs	income_per_year	number_kids
{Aquarius, Aries, Cancer, Capricorn}	47247.766962	1.920354
{Gemini, Leo, Libra, Pisces}	45533.825545	2.049844
{Sagittarius, Scorpio, Taurus, Virgo}	47004.485294	1.900000

Figure 15: Análises

2.5 Comparação da Distinction da Separation entre o Dataset original e os 2 datasets resultantes das transformações

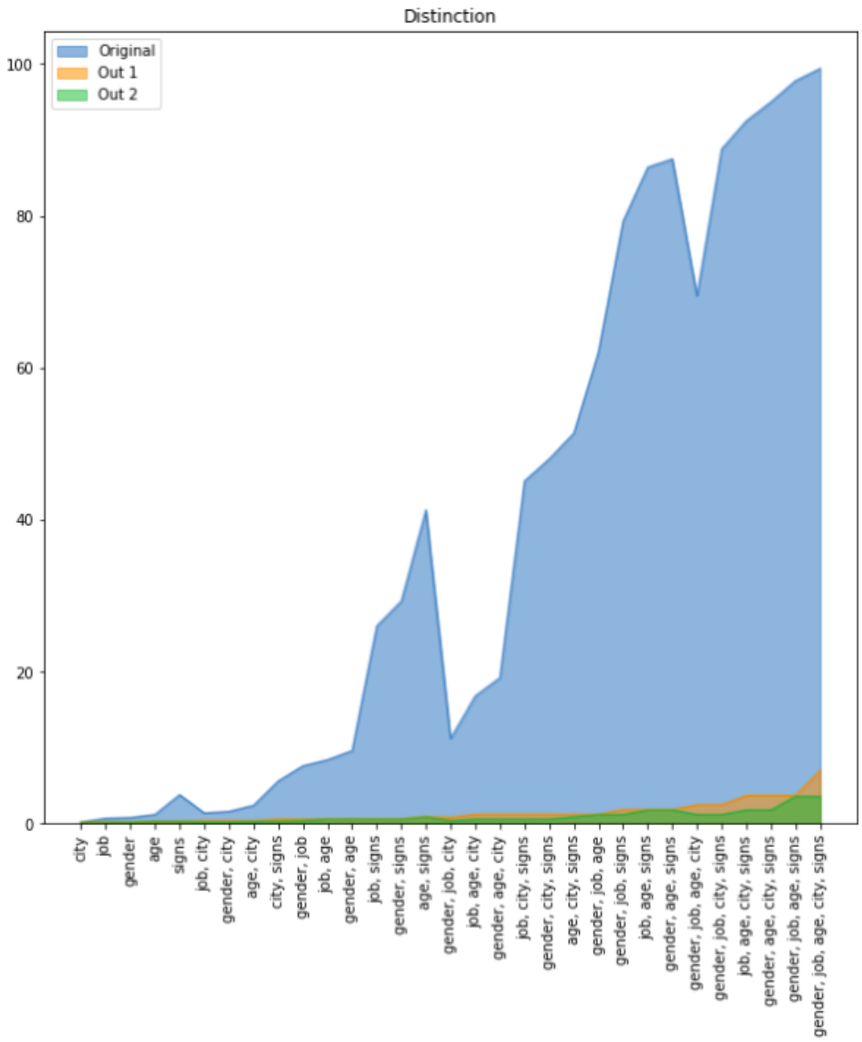

Figure 16: Comparação da distinction dos dados dos 3 datasets

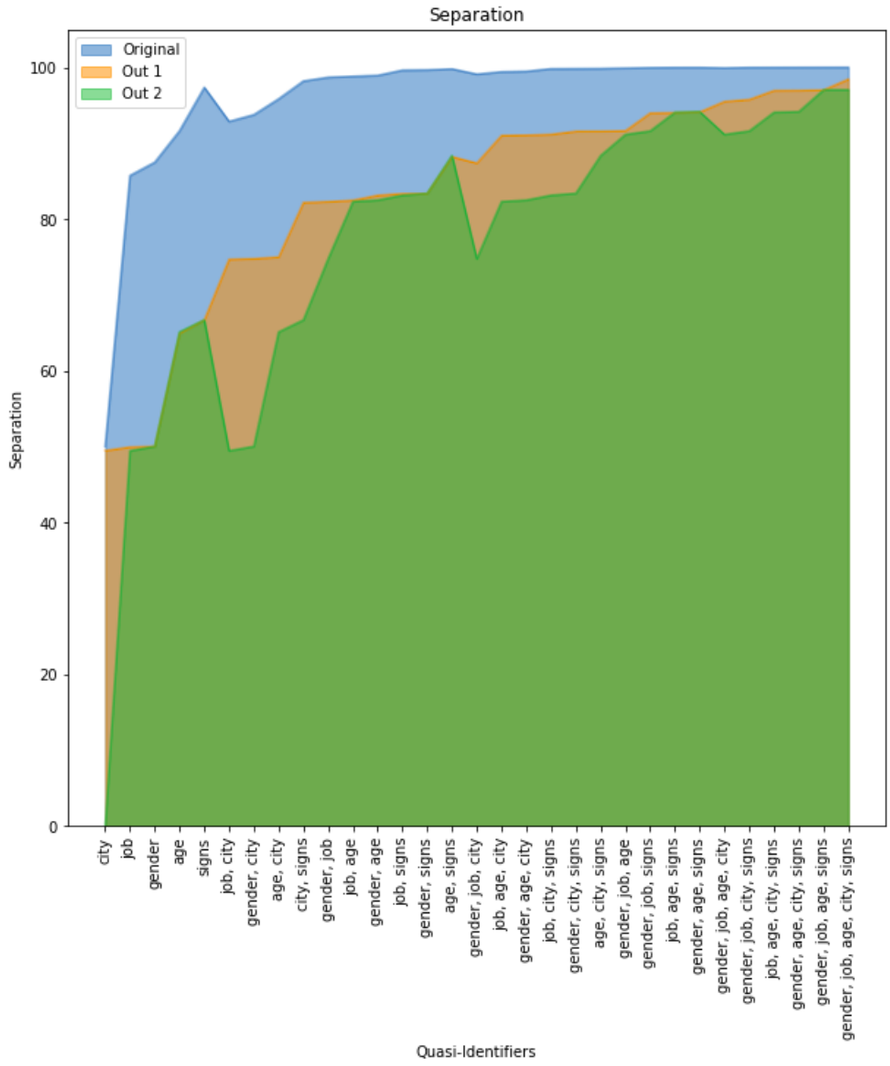

Figure 17: Comparação da Separation dos dados dos 3 datasets

2.6 Análise da Utilidade

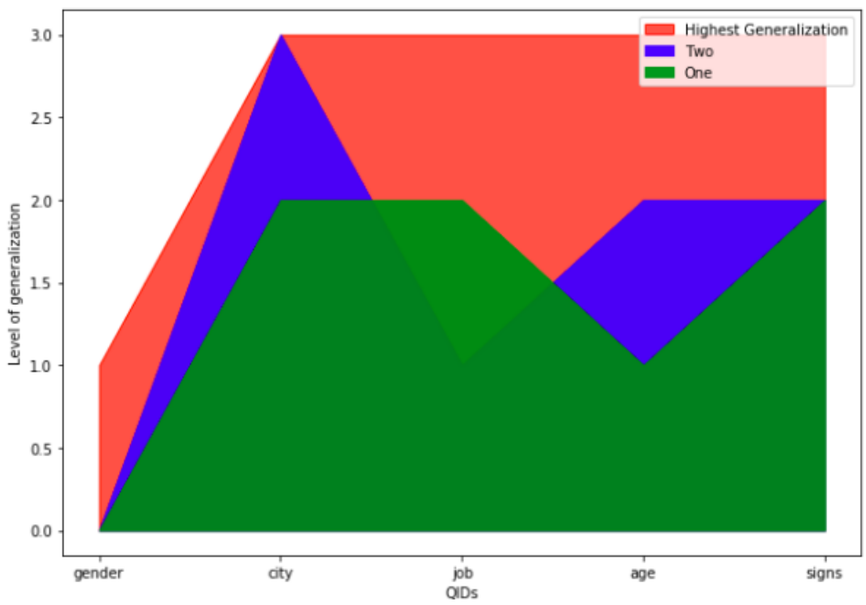

Figure 18: Comparação da Utilidade entre o dataset original e os 2 resultantes

Assim, a priori, vendo apenas pelo gráfico 18, podemos ver que o dataset resultante One tem maior utilidade que o Two, mas não nos podemos esquecer que esse mesmo dataset suprimiu 32 valores dos 1000 enquanto que o Two não suprimiu nenhum, por isso decidimos calcular a métrica Precision de ambos os dataset de modo a retirar conclusões finais.

Prec(One) = 0.51570

Prec(Two) = 0.46667

2.7 Conclusão

Durante a realização deste exercício, conseguimos compreender a teoria dos Modelos Sintáticos, ou seja, à medida que vamos aumentando a privacidade dos dados, vamos ao mesmo tempo diminuindo a sua utilidade. No dataset resultante One, aplicamos modelos sintáticos bastantes simples (4-Diversity em todos os atributos sensíveis) e concluimos que conseguimos aumentar bastante a segurança, mesmo que perdendo alguma utilidade. No dataset resultante Two, aumentamos a complexidade dos modelos Sintáticos, aplicando um 10-Anonymity juntamente com modelos de L-Diversity o que resultou numa excelente segurança, contudo, fez com que a utilidade dos dados diminuisse em relação ao dataset anterior. Deste modo, conseguimos perceber o *payoff* da privacidade dos dados.

3 Differential Privacy

3.1 Exercício 2.1

Para calcular a sensibilidade dos dados originais, recorremos à média e desvio padrão. Para tal, calculamos a média e o desvio padrão de todos os dados de cada uma das colunas com dados sensíveis. De seguida, retiramos um elemento e voltamos a calcular a média e os desvios padrões dos dados. No fim, verificamos qual era o maior valor absoluto, sendo que depois esses resultados são apresentados na tabela 5.

	Dataset Original
Income per year (average)	53.4306
Income per year (standard deviation)	55.7455
Age (average)	0.0188078
Age (standard deviation)	0.0104017
Number of kids (average)	0.00204705
Number of kids (standard deviation)	0.000812522

Table 5: Tabela com os valores de sensibilidade para o dataset original

3.2 Exercício 2.2

Para cada coluna e estatística (média e desvio padrão) dos dados originais, geramos 1000 valores aleatórios de uma distribuição de Laplace centrada em zero e escala (valor respetivo da tabela acima / epsilon). Os valores de epsilon usados foram os seguintes: 0.01, 0.2, ln(2) e ln(3). No fim, esses valores são adicionado a uma cópia da tabela com os dados originais, tendo assim aplicado o algoritmo.

3.3 Exercício 2.3

Aplicou-se o algoritmo explicado na secção acima às cópias dos dados. Na tabela 6 analisamos os dados da mesma forma no exercício 2.1. Podemos ver as diferenças para as várias colunas, aplicando os diferentes valores de epsilon. Nas figuras 19, 20 e 21 apresentamos graficamente os valores da tabela 6.

	Ipy Avg	**Ipy Std**	**Age Avg**	**Age Std**	**Nk Avg**	**Nk Std**
Ipy 0.01	67.4052	85.0813	—	—	—	—
Ipy 0.2	53.9457	58.1244	—	—	—	—
Ipy ln(2)	53.4563	55.8581	—	—	—	—
Ipy ln(3)	53.409	55.7415	—	—	—	—
Age 0.01	—	—	0.0275956	0.0186707	—	—
Age 0.2	—	—	0.0193263	0.0112802	—	—
Age ln(2)	—	—	0.0192863	0.0112562	—	—
Age ln(3)	—	—	0.0192863	0.0111734	—	—
Nk 0.01	—	—	—	—	0.00345045	0.00161498
Nk 0.2	—	—	—	—	0.00247147	0.00153305
Nk ln(2)	—	—	—	—	0.00245546	0.0016191
Nk ln(3)	—	—	—	—	0.00246847	0.00163075

Table 6: Tabela com os valores de sensibilidade após aplicar o algoritmo

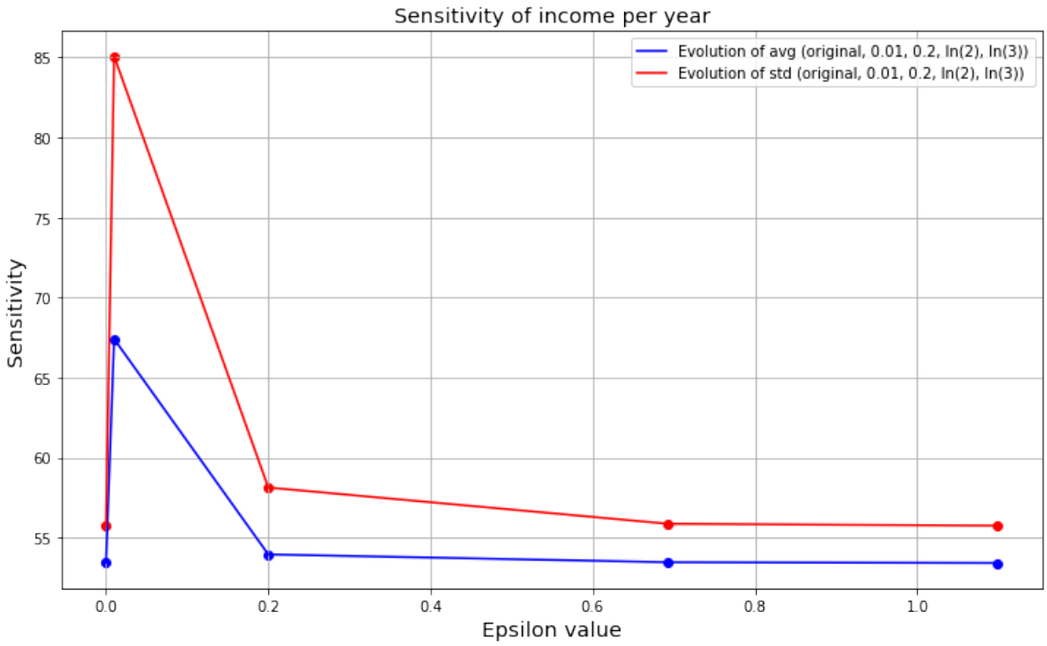

Figure 19: Análise da sensibilidade do salário anual bruto

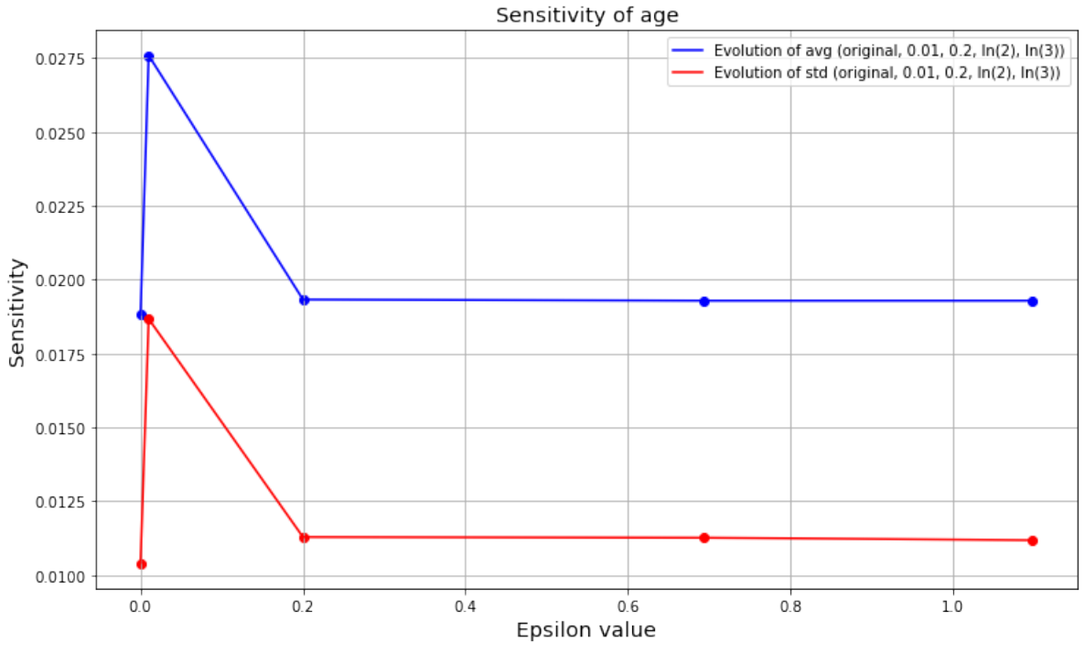

Figure 20: Análise da sensibilidade da idade

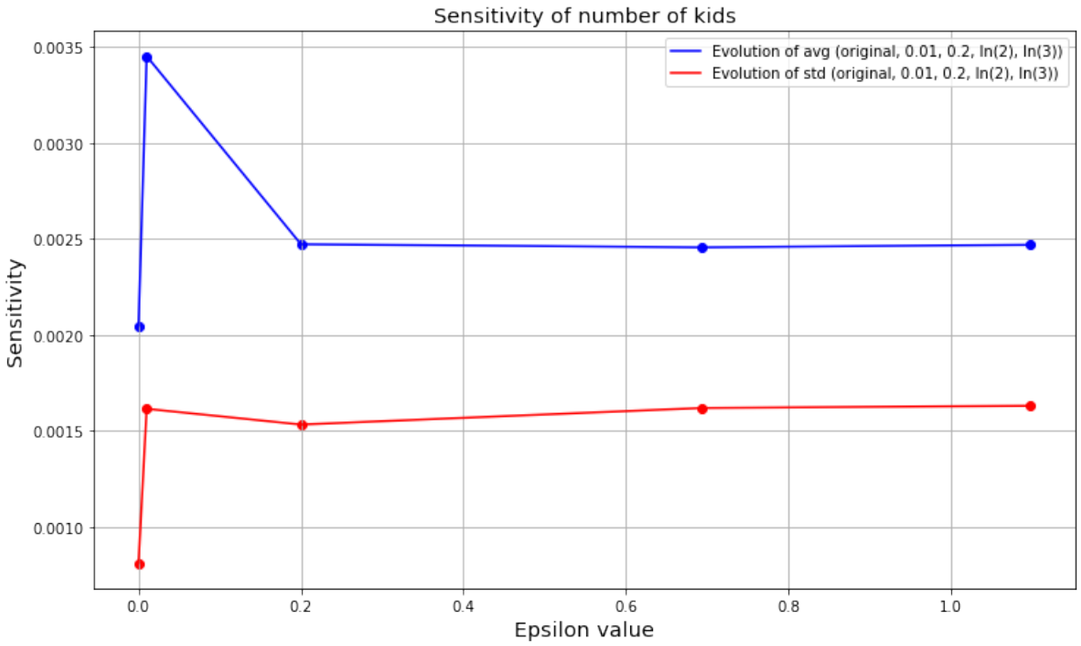

Figure 21: Análise da sensibilidade do número de filhos

3.4 Exercício 2.4

- Nos gráficos 22 e 23, não existe grande diferença dos valores originais para os alterados, o padrão que estes seguem é muito semelhante ao original (quando epsilon = 0.01, essa diferença é maior, enquanto que quando epsilon = ln(3), a diferença é mínima).

Figure 22: Comparação entre os valores médios de salário bruto anual agrupados por idade

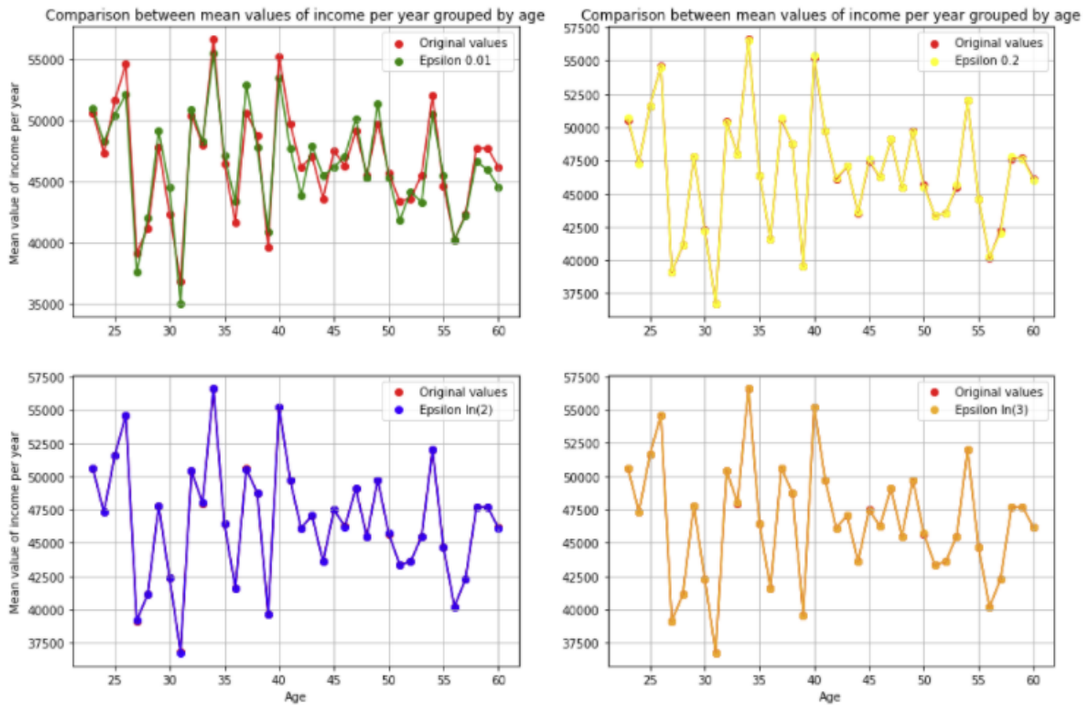

Figure 23: Comparação entre os valores médios de salário bruto anual agrupados por idades, separados por gráficos de diferentes valores de epsilon

- Nos gráficos 24 e 25, em essa diferença para os valores originais é notória, por norma, todos os dados são aleatórios.

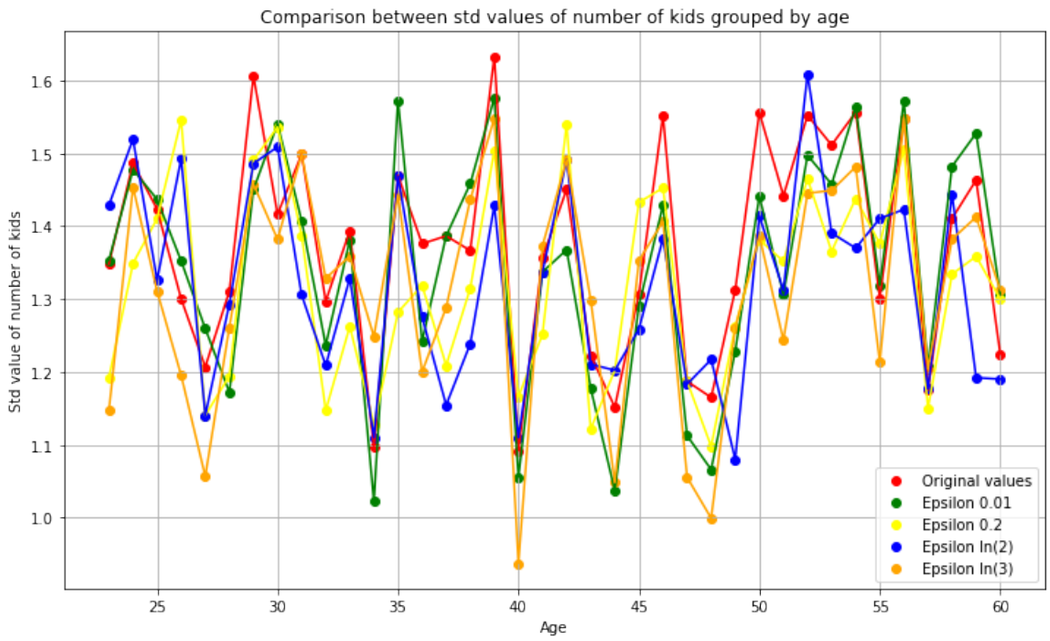

Figure 24: Comparação entre os valores de desvio padrão de número de filhos agrupados por idade

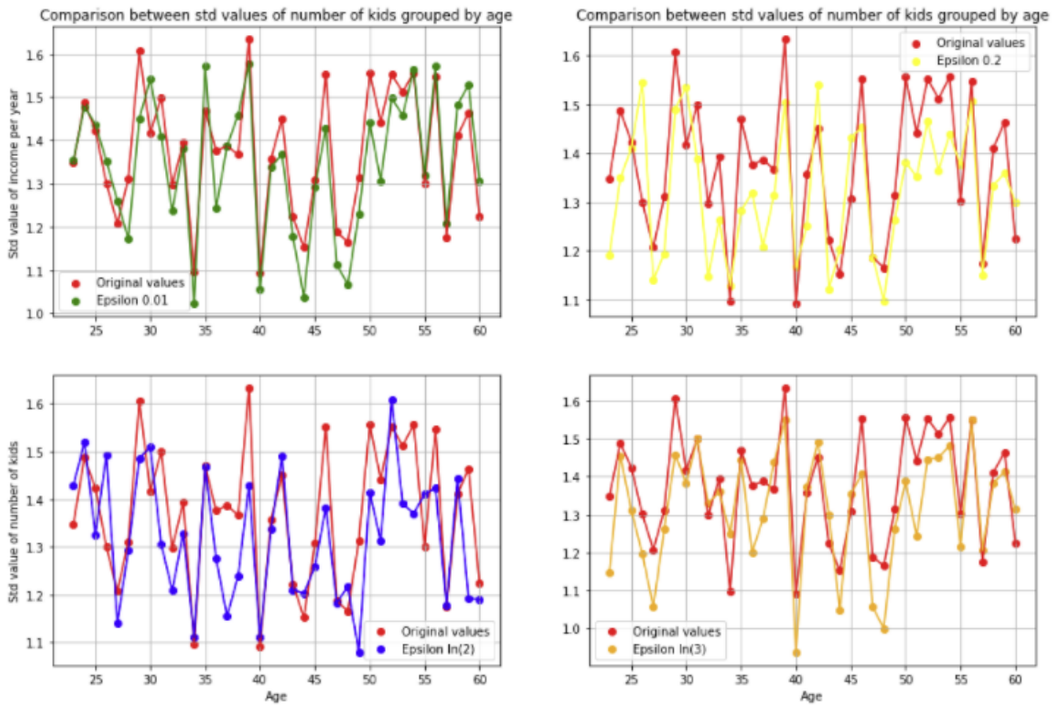

Figure 25: Comparação entre os valores de desvio padrão de número de filhos agrupados por idades, separados por gráficos de diferentes valores de epsilon

- Nos gráficos 26, foi repetida uma mesma query várias vezes, sempre aplicando a função explicada no exercício 2.2. Com isto, conseguimos compreender a importância da Differencial Privacy para fazer pesquisas nos dados e a diferença que faz os valores de epsilon aos dados.

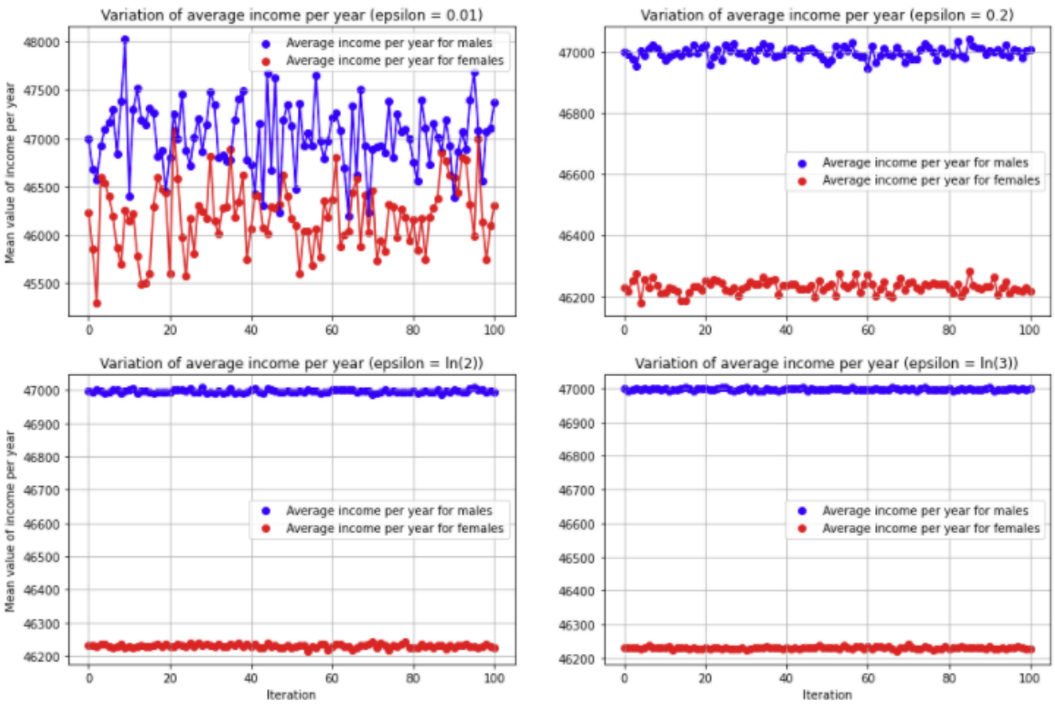

Figure 26: Comparação entre os valores de desvio padrão de número de filhos agrupados por idade separados por gráficos de diferentes valores de epsilon

NOTA: Neste ponto, fizemos uma grande análise com gráficos, pelo que apenas colocamos aqui alguns para exemplificar. Para ter acesso a todos, ver código-fonte.

3.5 Exercício 2.5

Vantagens:

- Ajusta os resultados das várias queries de modo a ser o mais preciso possível, garantindo a privacidade dos dados dos utilizadores
- É possível definir o nível de *noise* adicionado aos atributos sensíveis (valor de epsilon), aumentando ou diminuindo assim a privacidade dos dados

- Se quisermos fornecer o dataset a duas pessoas diferentes, fazendo variar entre eles o valor de epsilon, mesmo que estas tentem juntar-se para tentar recuperar os dados originais, torna-se impossível

Desvantagens:

- Funciona melhor para dados de baixa sensibilidade (por exemplo, se estivermos a procurar pela pessoa mais velha de uma equipa, este pode fornecer dado completamente errados)
- Custo computacional elevado (pois necessita de calcular a sensividade dos dados e de seguida aplicar o algoritmo)

Queries:

- Se estivermos a falar de queries de contadores, estas funcionam bem uma vez que a presença de mais ou menos um elemento apenas vai alterar o resultado ligeiramente.
- Somas, máximos, mínimos, médias podem ser um problema pois na presença de um elemento com um valor extremamente elevado vai implicar com que o resultado final se altere bastante uma vez que sejá necessário adicionar bastante *noise*.