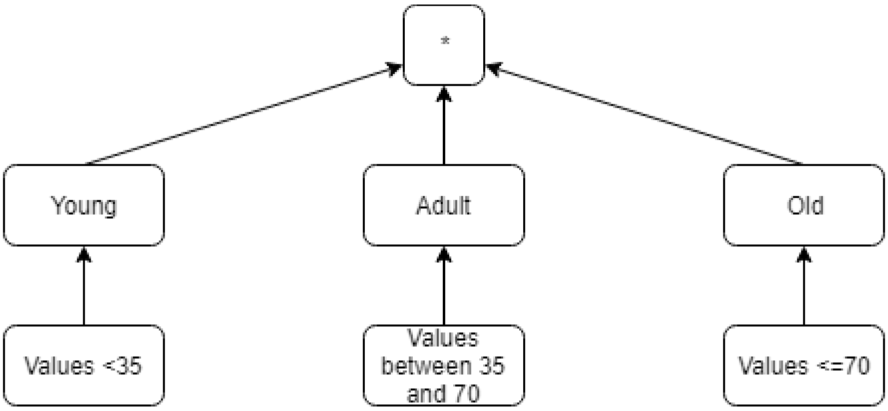
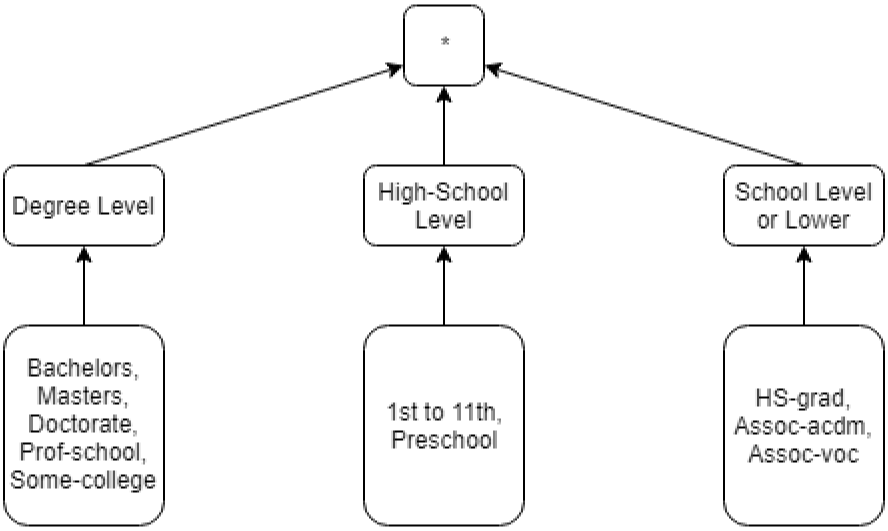
Part I

1.Define a hierarchy for each of the 4 attributes.

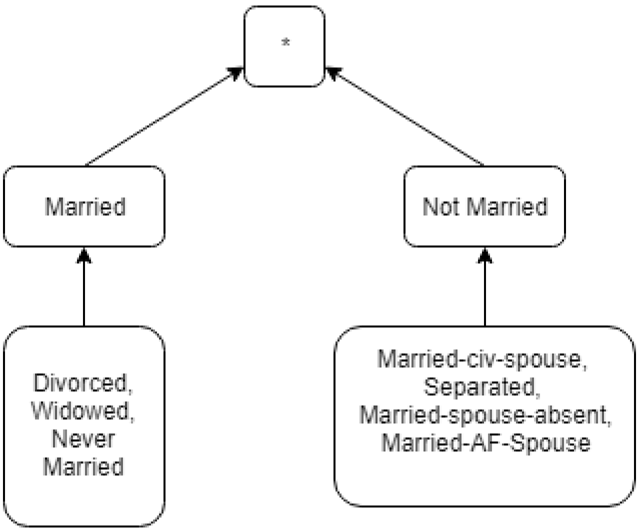
Hierarchy of Age:



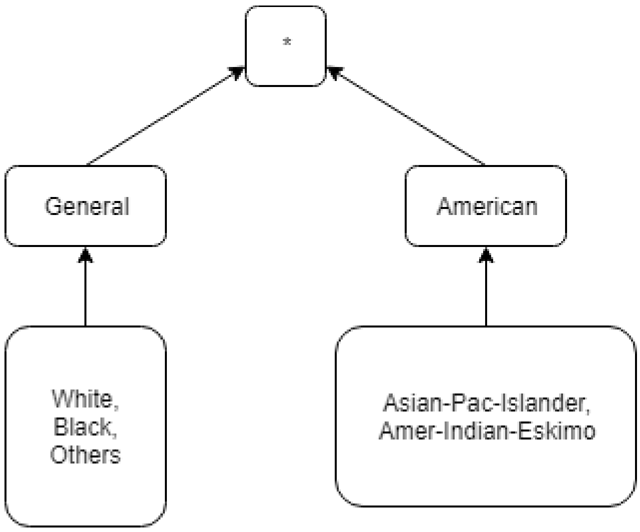
Hierarchy of Education:



Hierarchy of Marital Status:



Hierarchy of Race:



2. Write a program for the heuristic algorithm (which generalizes/suppresses the data while minimizing the utility loss). You can use any programming language you feel comfortable, e.g, Java, Python and C++.

The program for this problem of the assignment is commented, and there it is explained what have been done for obtaining the results.

Here there is the output, where it can be seen the hierarchies applied to the data and also the K on the right of the image.

	Age	Education	Marital-Status	Race	k
0	Adult	Degree Level	Married	American	300
1	Adult	Degree Level	Married	General	5397
2	Adult	Degree Level	Not Married	American	103
3	Adult	Degree Level	Not Married	General	2786
4	Adult	High School Level	Married	American	138
5	Adult	High School Level	Married	General	4371
6	Adult	High School Level	Not Married	American	92
7	Adult	High School Level	Not Married	General	2612
8	Adult	School Level or lower	Married	American	52
9	Adult	School Level or lower	Married	General	1322
10	Adult	School Level or lower	Not Married	American	24
11	Adult	School Level or lower	Not Married	General	686
12	Old	Degree Level	Married	American	2
13	Old	Degree Level	Married	General	153
14	Old	Degree Level	Not Married	American	9
15	Old	Degree Level	Not Married	General	88
16	Old	High School Level	Married	American	5
17	Old	High School Level	Married	General	111
18	Old	High School Level	Not Married	American	2
19	Old	High School Level	Not Married	General	109
20	Old	School Level or lower	Married	American	3
21	Old	School Level or lower	Married	General	74
22	Old	School Level or lower	Not Married	American	4
23	Old	School Level or lower	Not Married	General	69
24	Young	Degree Level	Married	American	112
25	Young	Degree Level	Married	General	1853
26	Young	Degree Level	Not Married	American	238
27	Young	Degree Level	Not Married	General	4317
28	Young	High School Level	Married	American	66
29	Young	High School Level	Married	General	1943
30	Young	High School Level	Not Married	American	136
31	Young	High School Level	Not Married	General	3365
32	Young	School Level or lower	Married	American	26
33	Young	School Level or lower	Married	General	514
34	Young	School Level or lower	Not Married	American	38
35	Young	School Level or lower	Not Married	General	1441

3. Calculate the distortion and precision of your algorithm

In order to calculate these values, I have followed the formulas of the slides and explained in class, applying my data.

Distorsion: $(\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2})\Big/_4 = 0.5$

Precision: $1 - \frac{\left(\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}\right)x32561}{32561x4} = 0.5$

Part II

Laplacian Mechanism

1. in case of Epsilon = *0.5, generate 1,000 results for the query over the original dataset, and generate 1,000 results for the query over each of three other datasets (removing a record with the oldest age; removing any record with age 26; removing any record with the youngest age).*

For doing this exercise, I have filtered the data and created a dataframe with only the rows with age over 24 years, in order to obtain the average of the ages from 25 years.

Then, I have obtained one group of 1000 tuples (the first ones of the dataframe obtained before), and another three which I have delete from them 1 row with the age 90 years, 1 row with the age 26 and 1 row with the age 25, respectively.

The results can be seen in the txt file, where can be found the print of the groups.

2. in each of the above 4 groups of 1,000 results, round each number to two decimal places, define a measure and utilize it to validate that each of the last 3 groups of results and the original results are 0:5-indistinguishable.

In this exercise I have rounded the values of the averages obtained from the 4 groups and rounded to 2 decimals, As can be seen in the images and in the .txt.

After that, I have defined a measure to check the indistinguibility, which is the following one:

$$log\left(\frac{average\ of\ group\ of\ 999}{average\ of\ original\ 1000\ results}\right) \leq \varepsilon$$

And I have obtained all the results below 0.5, and then, all of them are 0.5 indistinguisable as can be seen in the following image:

measureSetWithout25Ep05	float64	(1,)	[-0.000207]
measureSetWithout26Ep05	float64	(1,)	[-0.000207]
measureSetWithout90Ep05	float64	(1,)	[0.00051794]

3. repeat all the above for = 1, utilize the above measure to validate that each of the last 3 groups of results and the original results are 1-indistinguishable.

Using the same formulas and procedure as question 2, I have completed this question obtaining the following measures:

measureSetWithout25Ep1	float64	(1,)	[-0.00047653]
measureSetWithout26Ep1	float64	(1,)	[-0.00047653]
measureSetWithout90Ep1	float64	(1,)	[0.00119232]

4. define another measure and utilize it to justify that the distortion of the 4,000 results for epsilon *= 1 is less than that of* epsilon *= 0,5.*

We can see in the answers of the last 2 methods that the values of the measures with the epsilon 1 (exercise 3) are far from 0 than the ones with calculated with epsilon 0.5. This means that the measures of 0.5 shows that the data are more similar and then, whit less distortion than with the epsilon 1. This is because if the result were 0, the data will be completely the same, without distortion.

Another measure for calculating this could be counting the number of elements that are the same in the groups that are being measured and divide this number by the total number of elements. But in our case it is not useful for measuring, because in the 3 cases the numbers are the same (999/1000).

Nevertheless, it would be possible using the same formula but without the logarithm, as I did:

For epsilon 0.5:

measureSetWithout25Ep05	float64	(1,)	[0.99952381]
measureSetWithout26Ep05	float64	(1,)	[0.99952381]
measureSetWithout90Ep05	float64	(1,)	[1.00095374]

For epsilon 1:

measureSetWithout25Ep1	float64	(1,)	[0.99952358]
measureSetWithout26Ep1	float64	(1,)	[0.99976173]
measureSetWithout90Ep1	float64	(1,)	[1.00119303]

Exponential Mechanism

5. in case of = 0:5, generate 1,000 results for the query over the original dataset, and generate 1,000 results for the query over each of three other datasets (removing a record with the most frequent "Education"; removing any record with the second most frequent "Education"; removing any record with the least frequent "Education").

For doing this exercise, I have obtained one group of 1000 tuples (the first ones of the dataframe obtained before), and another three which I have delete from them 1 row with the most frequent education value (HS-grad), 1 row with the second most frequent education value and 1 row with the least frequent education value, respectively.

The results can be seen in the txt file, where can be found the print of the groups.

6. in each of the above 4 groups of 1,000 results, define a measure and utilize it to validate that each of the last 3 groups of results and the original results are 0:5-indistinguishable.

In this exercise I have obtained the most frequent values of the education in each group, and applied the exponential mechanism, using the following formula for measuring the indistinguibility, which is the following one:

$$\frac{\Pr[\mathcal{M}_E(x,u,\mathcal{R})=r]}{\Pr[\mathcal{M}_E(y,u,\mathcal{R})=r]} = \frac{\left(\frac{\exp(\frac{\varepsilon u(x,r)}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(x,r')}{2\Delta u})}\right)}{\left(\frac{\exp(\frac{\varepsilon u(y,r)}{2\Delta u})}{\sum_{r'\in\mathcal{R}}\exp(\frac{\varepsilon u(y,r')}{2\Delta u})}\right)}$$

With this formula, we obtain the epsilon, and we can check if it is 0.5 indistinguisable or not.

These are the values obtained from the measures:

epsilonOriginalAndWithoutLeastEp05	float64	1	1.0
epsilonOriginalAndWithoutMostEp05	float64	1	1.0000000000107223
epsilonOriginalAndWithoutSecondMostEp05	float64	1	0.9999999999916493

As we can see, these values are lower than $e^{0.5} = 1.64$, and then, accomplishes the indistinguibility.

7. repeat all the above for = 1, utilize the above measure to validate that each of the last 3 groups of results and the original results are 1-indistinguishable.

Using the same formulas and procedure as question 6, I have completed this question obtaining the following measures:

epsilonOriginalAndWithoutLeastEp1	float64	1	1.0
epsilonOriginalAndWithoutMostEp1	float64	1	1.0
epsilonOriginalAndWithoutSecondMostEp1	float64	1	1.0

As we can see, these values are lower than $e^1 = 2.71$, and then, accomplishes the indistinguibility.

8. define another measure and utilize it to justify that the distortion of the 4,000 results for epsilon = 1 is less than that of epsilon = 0:5.

Another way would be apply log to the division, comparing directly with the epsilon:

$$\ln(division) \leq \varepsilon$$

Using this method I had for epsilon 0.5:

epsilonOriginalAndWithoutLeastEp05	float64	1	0.0
epsilonOriginalAndWithoutMostEp05	float64	1	1.0722311927167353e-11
epsilonOriginalAndWithoutSecondMostEp05	float64	1	-8.350653502055443e-12

Using epsilon 1:

epsilonOriginalAndWithoutLeastEp1	float64	1	0.0
epsilonOriginalAndWithoutMostEp1	float64	1	0.0
epsilonOriginalAndWithoutSecondMostEp1	float64	1	0.0