Corey Capooci
HW1

Q1)
(a)

Movie ID	Corresponding Weight according to w()
03124	0.27103391082217876
14199	0.2711777534611824
06315	0.2710861246194568
07242	0.2712302590154106
17113	0.2712697082686535
10935	0.27122368998682783
11977	0.2712894554827877
03768	0.29119798994551205
02137	0.29113029735230445
06004	0.2911844389469680 5
08191	0.2712960412389864
15267	0.2914293175340726
03276	0.2713553886055061
16944	0.2711188119190368
01292	0.29129302224118986

(b)

Top 5 Similarity Scores		
	User ID	**Score**
1	1037245	2.782277557747669
2	2118461	2.5737639077919106
3	2602249	2.1998370845210404
4	16272	2.162017307044926
5	305344	2.1570261655310237

c)
User ID of the highest score is : **1037245**

The rating side by side is:

Side by side comparison of ratings of the most similar user and the auxiliary data		
Movie ID	Rating of User 1037245	Auxiliary Data Rating
03124	4	4
07242	1	2
17113	3	2
10935	2	4
11977	3	3

03276	3	3
14199	4	2
08191	2	2
06004	2	3
01292	2	2
03768	3	3
02137	2	1

The auxiliary database and user's 1037245 data are very similar. First off, the list of rated movies are exactly the same. In addition, the scores of 6 of the movies are the same. 4 of the scores are only off by 1. Only 2 of the scores are off by 2. No scores differ by greater than 2.

(d)
The difference between the highest and second highest scores is **0.2085136499557585**.

(a) If $\gamma = 0.1$ would you accept the top candidate
No I would not.

(b) If $\gamma = 0.05$ would you accept the top candidate?
Yes I would.

Q2

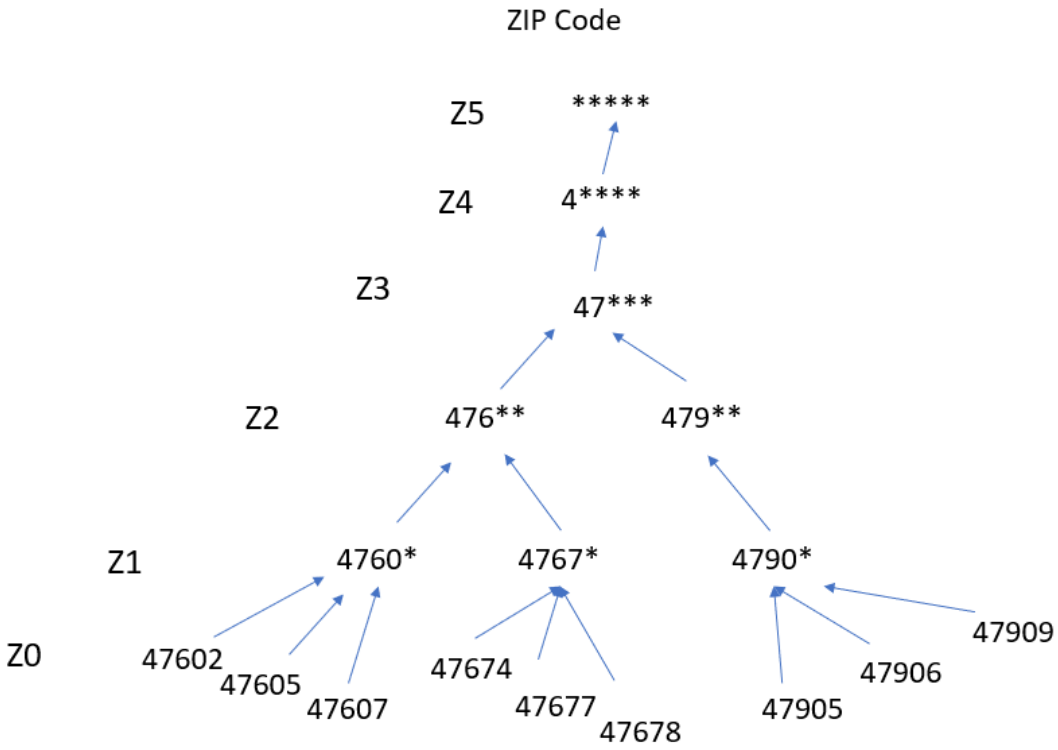
(a)

The quasi-identifiers are ZIP code, and Age.

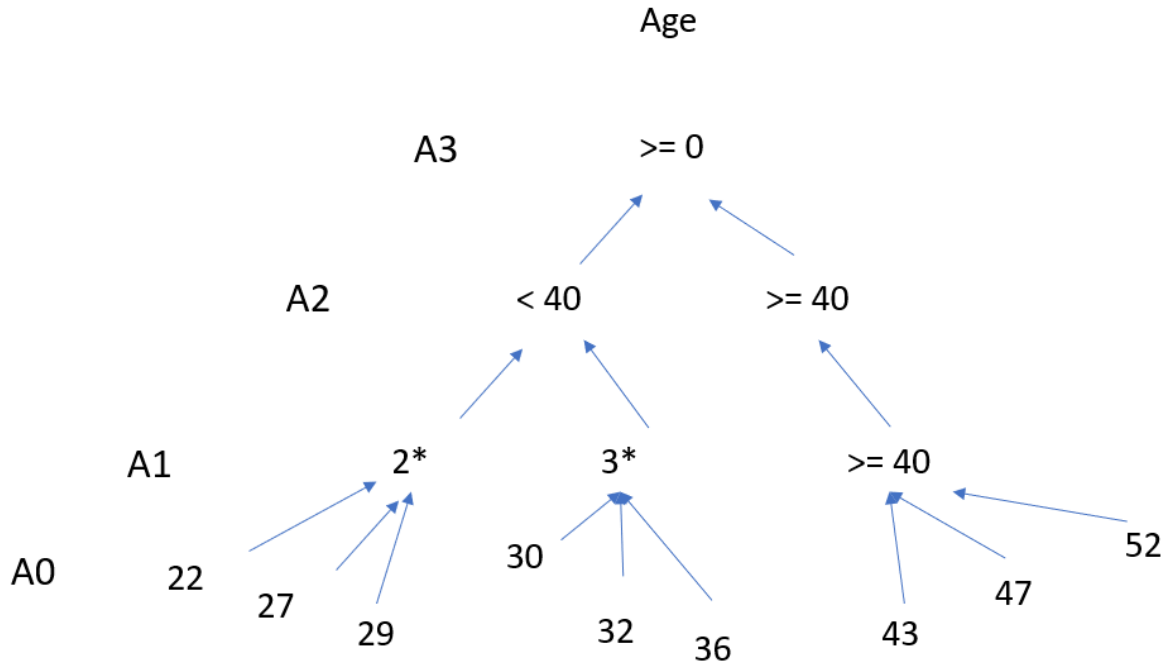
The sensitive attributes are Salary and Disease.

(b)

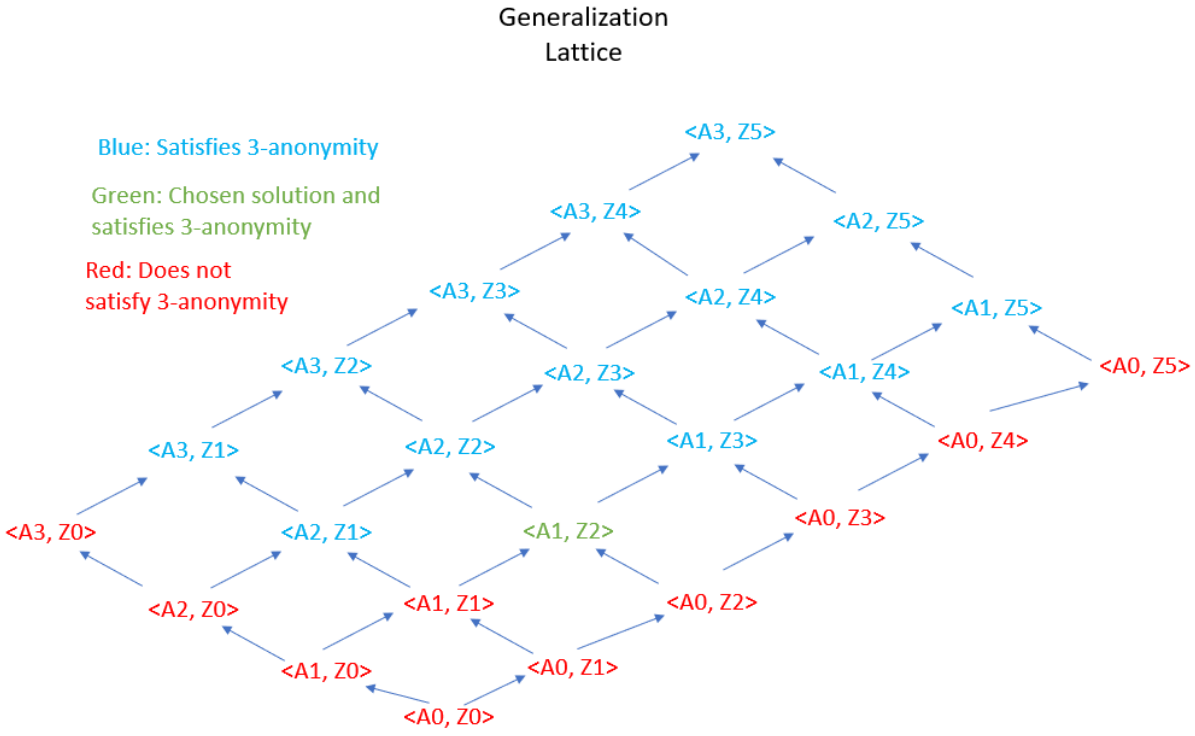
The generalized hierarchy for Zip Code is below. We are suppressing one digit at a time.



The generalization hierarchy for Age is below.



The Generalization lattice based on the generalization hierarchies above.



The final anonymized table is

ID	ZIP code	Age	Salary	Disease
1	476**	2*	3K	Gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	Stomach cancer
4	479**	>=40	6K	gastritis
5	476**	3*	7K	flu
6	479**	>=40	8K	bronchitis
7	476**	3*	9K	bronchitis
8	476**	3*	10K	pneumonia
9	479**	>=40	11K	Stomach cancer

(c)
Calculating t-closeness using the Earth Mover's Distance.

Q = {3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K}
P1 = {3K, 4K, 5K}
P2 = {7K, 9K, 10K}
P3 = {6K, 8K, 11K}

P1.
3K) (1/9) * (3+4)/8
4K) (1/9) * (4+5)/8
5K) (1/9) * (5+6)/8

(7/72) + (9/72) + (11/72) = **27/72 = 0.375**

P2.
7K) (1/9) * (3+4)/8
9K) (1/9) * (3+4)/8
10K) (1/9) * (2+1)/8

(7/72) + (7/72) + (3/72) = **17/72 = 0.2361**

P3.
6K) (1/9) * (3+2)/8
8K) (1/9) * (3+1)/8
11K) (1/9) * (2+1)/8

(5/72) + (4/72) + (3/72) = **12/72 = 0.167**

t is the smallest threshold that is no smaller than the any of the thresholds amongst all of the equivalence classes. Therefore, the value of t is 0.375 for these equivalence classes.

(d)
This solution does not resolve the similarity attack with respect to disease because if one knows a person in their 20s in a zip code resembling 476** that exists in the dataset, then this person has a salary less than or equal to 5K and has a stomach disease.