Question 1)
a)
Weights for each movie are given as follows:

Movie: Weight
03124: 0.11770853187873027
14199: 0.11777100194311198
06315: 0.11773120804276736
07242: 0.11779380481558255
17113: 0.11781093740858115
10935: 0.11779095192271759
11977: 0.11781951351471258
03768: 0.12646568017465448
02137: 0.1264362816549587
06004: 0.12645979505076255
08191: 0.11782237367228886
15267: 0.12656614446987832
03276: 0.11784814790608385
16944: 0.11774540395660324
01292: 0.12650695217626995

b)
Number of users in the database are 44651

Highest score is 0.11335450355831592
Second highest score is 0.10292067949119368

c)
The user-id of the user with the highest score is 1664010

Movie ratings of the user from the database:

03124: 4
06315: 4
07242: 4
16944: 4
17113: 4
11977: 4
03276: 4
14199: 4
08191: 4
06004: 4
01292: 3

15267: 4
03768: 4
02137: 4

To verify the ratings between auxiliary db and user db, I have rounded off the value of the aux db to the nearest integer and checked if it is similar to the user's db rating.

Rating for movie 03124 is similar to movie rating in auxiliary db
Rating for movie 06315 is similar to movie rating in auxiliary db
Rating for movie 07242 is similar to movie rating in auxiliary db
Rating for movie 16944 is similar to movie rating in auxiliary db
Rating for movie 17113 is similar to movie rating in auxiliary db
Rating for movie 11977 is similar to movie rating in auxiliary db
Movie 03276 not found in auxiliary db
Rating for movie 14199 is similar to movie rating in auxiliary db
Rating for movie 08191 is similar to movie rating in auxiliary db
Rating for movie 06004 is similar to movie rating in auxiliary db
Rating for movie 01292 is similar to movie rating in auxiliary db
Rating for movie 15267 is similar to movie rating in auxiliary db
Rating for movie 03768 is similar to movie rating in auxiliary db
Movie 02137 not found in auxiliary db

d)
Value of eccentricity threshold is 0.012068344559199527
Difference between the highest and second-highest score is 0.010433824067122233

No, the difference is not greater than the eccentricity threshold.

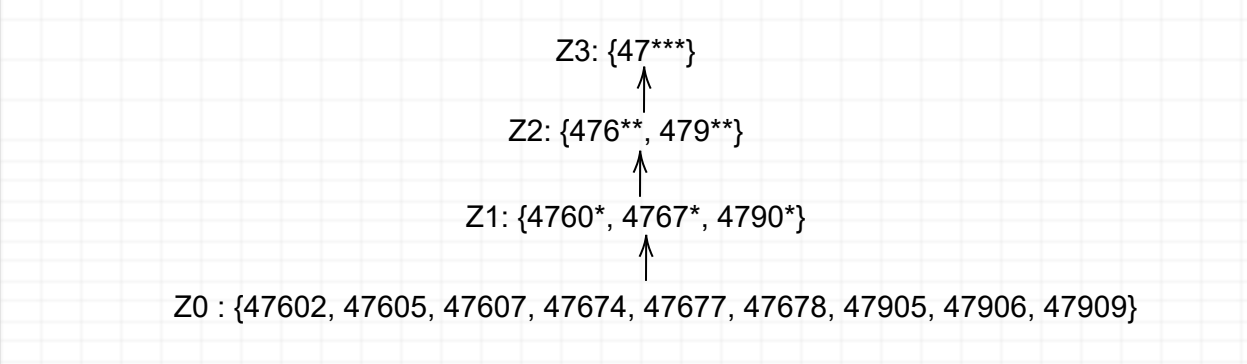
Question 2)

a)

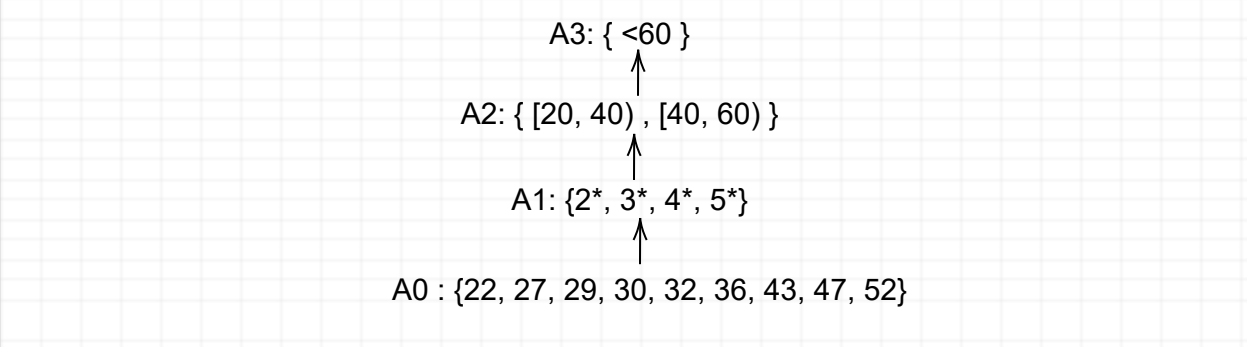
Since, ZIP code and Age would individually not help to conclude any information about the person, the quasi identifiers are **<ZIP code & Age>**. The sensitive attributes are **Salary, Disease** as they should be released for research purposes.

b)

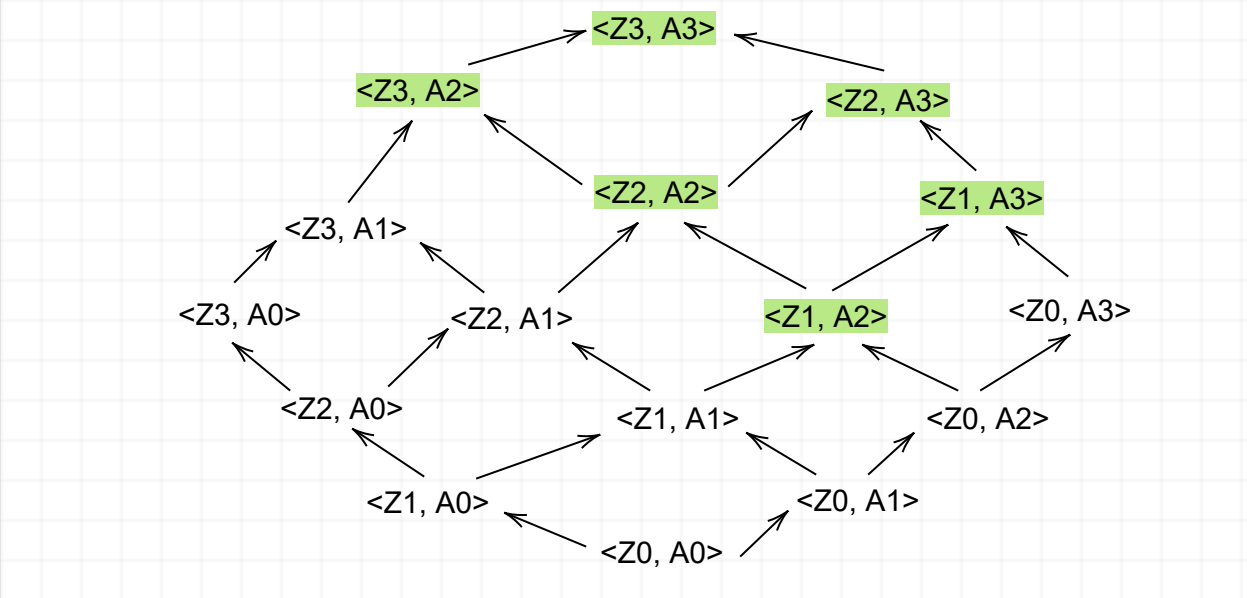
- To compose a 3-anonymous, 3-diverse table, we can generalize the information to make sure that there is minimum data loss and max privacy should be achieved.
- Lets consider ZIP code- to maintain min data loss, we can suppress the the data to divide it into three equivalence classes in the following way:
  - 4760*,4767*, 4790*
- Now, if we look at the Age column for the corresponding classes, we see that there is an overlap between ages.
  - For ex: Consider Ages for class 4760* : 22, 30, 32
  - If someone gets any of these values such as - <4760*,22> he/she can identify the particular record
  - Hence, we have to generalize the Age along with ZIP code.
- The ZIP code column can be generalized in the following way:



- The Age column can be generalized in the following way:



- Now the generalization lattice will be generated as follows:



- At <Z1, A2> generalization we achieve 3-anonymity. Hence, using the generalization property, all the nodes marked green will support 3-anonymity.
- <Z1, A2> achieves 3 anonymity because the data in the table will be divided into equivalence classes with each class having minimum three records.
- This generalization level also supports minimum data loss and maximum privacy
- If we divide the given data based on <Z1, A2>, we get three equivalence classes. The table can be described as follows:

ID	ZIP code	Age	Salary	Disease
1	4760*	[20, 40)	3K	Gastric ulcer
2	4760*	[20, 40)	5K	Stomach cancer
3	4760*	[20, 40)	9K	bronchitis

4	4767*	[20, 40)	4K	gastritis
5	4767*	[20, 40)	7K	flu
6	4767*	[20, 40)	10K	pneumonia
7	4790*	[40, 60)	6K	gastritis
8	4790*	[40, 60)	8K	bronchitis
9	4790*	[40, 60)	11K	Stomach cancer

- In the above table, each equivalence class has a minimum of 3 records. The three equivalence classes are:
  - < 4760*, [20, 40) >
  - < 4767*, [20, 40) >
  - < 4790*, [40, 60) >
- Hence, the above table is **3-anonymous**
- Additionally, we can observe that each class has 'well defined' values for both sensitive attributes- Salary, Disease
- Hence, we can conclude that the given table is **3-diverse**

- Based on the above generalization, the 3-anonymous, 3- diverse table will be as follows:

ID	ZIP code	Age	Salary	Disease
1	4767*	[20, 40)	3K	Gastric ulcer
2	4760*	[20, 40)	4K	gastritis
3	4767*	[20, 40)	5K	Stomach cancer
4	4790*	[40, 60)	6K	gastritis
5	4760*	[20, 40)	7K	flu
6	4790*	[40, 60)	8K	bronchitis
7	4767*	[20, 40)	9K	bronchitis
8	4760*	[20, 40)	10K	pneumonia
9	4790*	[40, 60)	11K	Stomach cancer

c)

i)

To compute t-closeness wrt Salary, let Salary ~ Q

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

Hence, for each equivalence class- P1, P2, P3

P1 = { 3K, 5K, 9K }

P2 = { 4K, 7K, 10K }

P3 = { 6K, 8K, 11K }

To transform P1 to Q, cost for each tranformation will be:

- 3K→4K, 6K : 1/9 * (1+3) / 8
- 5K→7K, 8K : 1/9 * (2+3) / 8
- 9K→10K, 11K : 1/9 * (1+2) / 8
  - Total Cost = 1/9 * (4+5+3) / 8 = 0.166

To transform P2 to Q, cost for each tranformation will be:

- 4K→3K, 5K : 1/9 * (1+1) / 8
- 7K→6K, 8K : 1/9 * (1+1) / 8
- 10K→9K, 11K : 1/9 * (1+1) / 8
  - Total Cost = 1/9 * (2+2+2) / 8 = 0.833

To transform P3 to Q, cost for each tranformation will be:

- 6K→3K, 4K : 1/9 * (3+2) / 8
- 8K→5K, 7K : 1/9 * (3+1) / 8
- 11K→9K, 10K : 1/9 * (2+1) / 8
  - Total Cost = 1/9 * (5+4+3) / 8 = 0.166

Therefore, to compute the t-closeness of table, we find the average of all the values:

Avg = (0.166 + 0.833 + 0.166) / 3 = 0.138

Hence, t-closeness for Salary is 0.138

ii)
Consider the given table computed in the solution above:

ID	ZIP code	Age	Salary	Disease
1	4760*	[20, 40)	3K	Gastric ulcer
2	4760*	[20, 40)	5K	Stomach cancer
3	4760*	[20, 40)	9K	bronchitis
4	4767*	[20, 40)	4K	gastritis
5	4767*	[20, 40)	7K	flu
6	4767*	[20, 40)	10K	pneumonia
7	4790*	[40, 60)	6K	gastritis
8	4790*	[40, 60)	8K	bronchitis
9	4790*	[40, 60)	11K	Stomach cancer

- For disease column, we can classify the diseases as follows:
  - Stomach illness: Gastric ulcer, Stomach cancer, gastritis
  - Lung illness: bronchitis, pneumonia
  - Viral infection: flu
- Based on the above conclusion, we can observe disease column values for each classes as follows:
  - Class I - Two stomach illnesses, one lung illness
  - Class II - One stomach illness, One viral infection, One lung illness
  - Class III - Two stomach illnesses, one lung illness
- Since there is no class with any disease value related to a particular illness/infection, we cannot perform similarity attack
- Thus, we can conclude that the above solution **will resolve** the similarity attack for the sensitive attribute disease.