Question 1:

a) The code printout looks as follows:

Solution to 1a
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) The code printout looks as follows:

Solution to 1b
Top 5 most similar user ids are: ['818752', '16272', '1037245', '2118461', '1707198']

c) The code printout looks as follows:

Solution to 1c
The user-id of the user with highest score in aux is 818752

Movie Ratings from dB and aux side by side:
Database: 03768: 1 Auxiliary: 03768: 1.5
Database: 01292: 2 Auxiliary: 01292: 2
Database: 06004: 1 Auxiliary: 06004: 1.5
Database: 03276: 2 Auxiliary: 03276: 2.5
Database: 14199: 4 Auxiliary: 14199: 3.5
Database: 03124: 3 Auxiliary: 03124: 3.5
Database: 10935: 1 Auxiliary: 10935: 2
Database: 02137: 2 Auxiliary: 02137: 1.5
Database: 11977: 2 Auxiliary: 11977: 2
Database: 17113: 3 Auxiliary: 17113: 4
Database: 16944: 2 Auxiliary: 16944: 2.5

Commenting on the similarity between the ratings of the user with user-id 818752:

The database user's movie ratings show a mixed degree of similarity to the ratings obtained in the auxiliary data. While some ratings are practically similar, such as those with absolute differences of 0 (Movie ids: 01292, 11977), indicating a great alignment, others have little variances, such as those with absolute differences of 0.5 (Movie ids: 03768, 06004, 03276, 14199, 03124, 02137, 16944) indicating minor changes. There are, however, instances of more significant dissimilarity, as indicated by absolute differences of 1 (Movie ids; 17113, 10935), indicating major discrepancies in how the user assessed particular films compared to the auxiliary data. Overall, the user's evaluations appear to be moderately aligned with the auxiliary data, with differences in preferences for specific movies contributing to both similarities and disparities.

d) The code printout looks as follows:

Solution to 1d
Difference between highest and second highest score is 0.010905589431324858
Working on checking whether or not we accept the candidate:

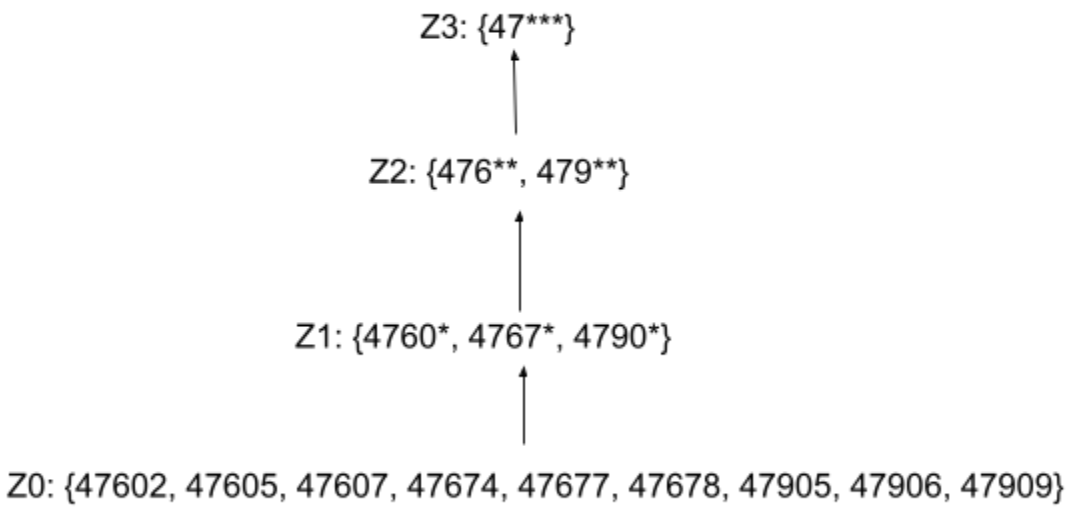
First, assuming that the value of gamma is 0.1,
Value of gamma*M is 0.01209421088715624
No, the difference is not greater than the threshold.
Therefore, we cannot accept the candidate

Now assuming gamma is 0.05
Value of gamma*M is 0.00604710544357812
Yes, the difference is greater than the threshold.
Therefore, we can accept the candidate

Question 2

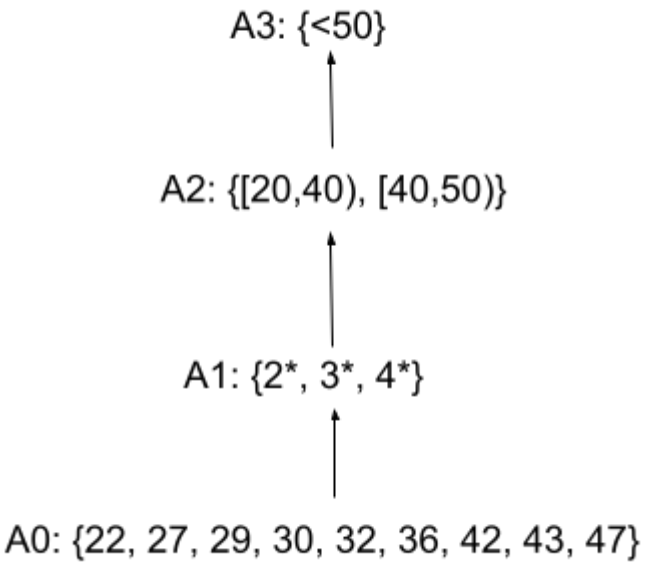
a) As we have been taught in class, Quasi-identifiers are attributes that, when combined, have the capacity to identify people even when direct identifiers such as names are removed. Individuals' sensitive or private information is stored in sensitive attributes. Based on that, in this table:
   i) Quasi Identifiers are Zip Code and Age. Both of these attributes can be used in combination with other attributes to identify individuals.
   ii) Sensitive attributes in this table are Salary and Disease. Both of them contain sensitive information about individuals.

b) To create a 3-anonymous, 3-diverse table, we can generalize the information to ensure that there is minimal data loss and maximum privacy is attained.
   For **Zip-Code**, the generalization hierarchy looks as follows:
   We can suppress the data and divide it into three equivalence classes as follows:
   - 4760*,4767*, 4790*

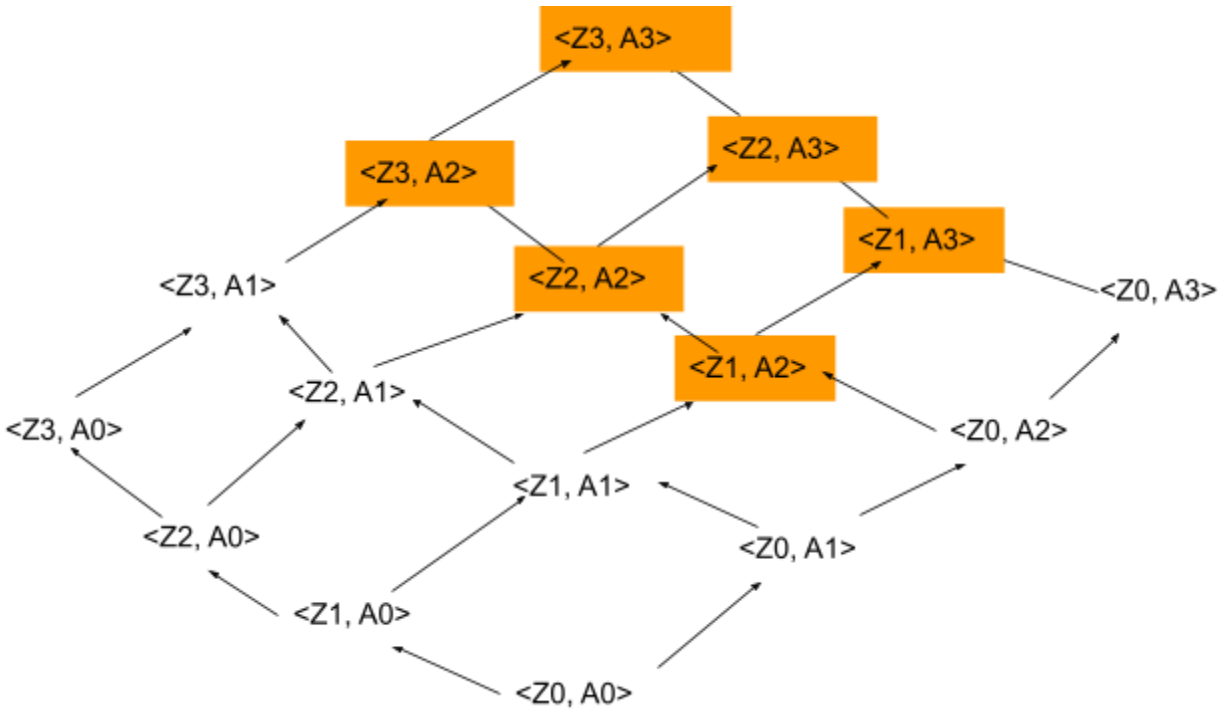


In the above diagram, I have sorted all the zip-code for Z0 for better suppression and better visibility.

For Age, the generalization hierarchy looks as follows:



Now, we will use the Incognito algorithm to draw the generalization lattice.



We obtain 3-anonymity at <Z1, A2> generalization. <Z1, A2> obtains 3 anonymity because the data in the table is separated into equivalence classes, each with at least three records. This also provides little data loss and maximum privacy. As a result of the generalization property, all of the orange nodes will provide 3-anonymity.
The final anonymized table looks as follows:

ID	Zip Code	Age	Salary	Disease
1	4760*	[20,40)	4K	Gastritis
2	4760*	[20,40)	7K	Pneumonia
3	4760*	[20,40)	10K	Pneumonia
4	4767*	[20,40)	3K	Gastric Ulcer
5	4767*	[20,40)	5K	Stomach Cancer
6	4767*	[20,40)	9K	Bronchitis
7	4790*	[40,50)	6K	Gastritis
8	4790*	[40,50)	8K	Bronchitis
9	4790*	[40,50)	11K	Stomach Cancer

The above table is sorted.
Each equivalence class has a minimum of three records in the table above. The three equivalence classes are as follows: -

- 4760*, [20, 40) >
- 4767*, [20, 40) >
- 4790*, [40, 60] >

As a result, the above table is 3-anonymous.
Furthermore, we can see that each class has 'well defined' values for both sensitive attributes- Salary and Disease. As a result, we may conclude that the presented table is 3-diverse.

The final 3-anonymous, 3-diverse, unsorted table looks 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*	[40,50)	6K	Gastritis
5	4760*	[20,40)	7K	Pneumonia
6	4790*	[40,50)	8K	Bronchitis

7	4767*	[20,40)	9K	Bronchitis
8	4760*	[20,40)	10K	Pneumonia
9	4790*	[40,50)	11K	Stomach Cancer

c) Computing t-closeness with respect to the Salary attribute as follows:
S- {3K,4K,5K,6K,7K,8K,9K,10K,11K}
The equivalence classes can be defined as follows:
P1: {3K, 5K, 9K } P2 = { 4K, 7K, 10K } P3 = { 6K, 8K, 11K }

Now let's transform equivalence class 1 to the original class S.
Taking LCM of 3, 9 to find out the amount of dirt that is to be moved. LCM(3,9) = 9. Therefore, 1/9th of the dirt from 3K will be moved to 6k bucket and 1/9th of the dirt from 3K bucket will be moved to 7K bucket.
Normalization of the distance = Maximum distance that can be moved = 11k-3k = 8K
Now, let's compute the distance of transformation:

- D[P1,Q] : transform P1 to Q
  - 3k ->6k, 3k ->7k cost: 1/9*(3+4)/8
  - 4k->8k,4k->9k cost: 1/9*(4+5)/8
  - 5k->10k,5k->11k cost: 1/9*(5+6)/8
  - Total cost: 1/9*27/8=0.375
- D[P2,Q] :
  - 4k ->3k, 4k ->5k cost: 1/9*(1+1)/8
  - 7k->6k,7k->8k cost: 1/9*(1+1)/8
  - 10k->9k,10k->11k cost: 1/9*(1+1)/8
  - Total Cost: 1/9*6/8 = 0.0833
- D[P3,Q] :
  - 6k ->3k, 6k ->4k cost: 1/9*(3+2)/8
  - 8k->5k,8k->7k cost: 1/9*(3+1)/8
  - 11k->9k,11k->10k cost: 1/9*(2+1)/8
  - Total Cost: 1/9*12/8 = 0.1667

Therefore, the t-value is the max(D[P1,Q], D[P2,Q], D[P3,Q])
= max(0.375, 0.0833, 0.1667)
=0.375

Hence, t-closeness for the Salary attribute is 0.375.