1. *(a) Tabulate the weights obtained for each movie.*

```
----------------------
Movie   Weight
----------------------
03276   0.117848147906
06004   0.126459795051
14199   0.117771001943
17113   0.117810937409
06315   0.117731208043
01292   0.126506952176
11977   0.117819513515
15267   0.12656614447
08191   0.117822373672
16944   0.117745403957
07242   0.117793804816
03768   0.126465680175
02137   0.126436281655
10935   0.117790951923
03124   0.117708531879
```

(b) How many users are present in the database? What is the highest score? What is the second highest score?

The number of users are: 44651
Highest score: 0.113354503558
Second highest score: 0.102920679491

(c) What is the user-id of the user with the highest score? Write out the ratings of this user from the database, and verify if they are similar to the ratings in the auxiliary information.

User with maximum score: 1664010

Ratings done by user 1664010

```
---------------------
Movie   Rating
----------------------
01292   3
02137   4
03124   4
03276   4
03768   4
06004   4
06315   4
07242   4
```

```
08191   4
11977   4
14199   4
15267   4
16944   4
17113   4
```

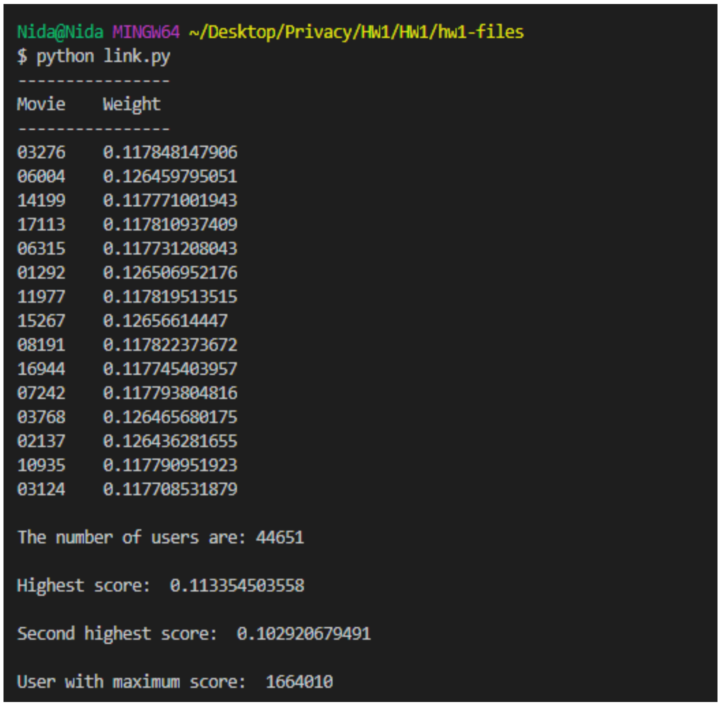
```
Comparing Aux and user ratings
----------------------------------------------
Movie   AUX Rating     User Rating
----------------------------------------------
01292   3.3            3
03124   3.5            4
03768   3.5            4
06004   3.9            4
06315   4.0            4
07242   3.9            4
08191   3.8            4
11977   4.2            4
14199   4.5            4
15267   4.2            4
16944   4.2            4
17113   4.2            4
```

Using the above table we can find out the ratings of the user are similar and comparable to ratings in aux.

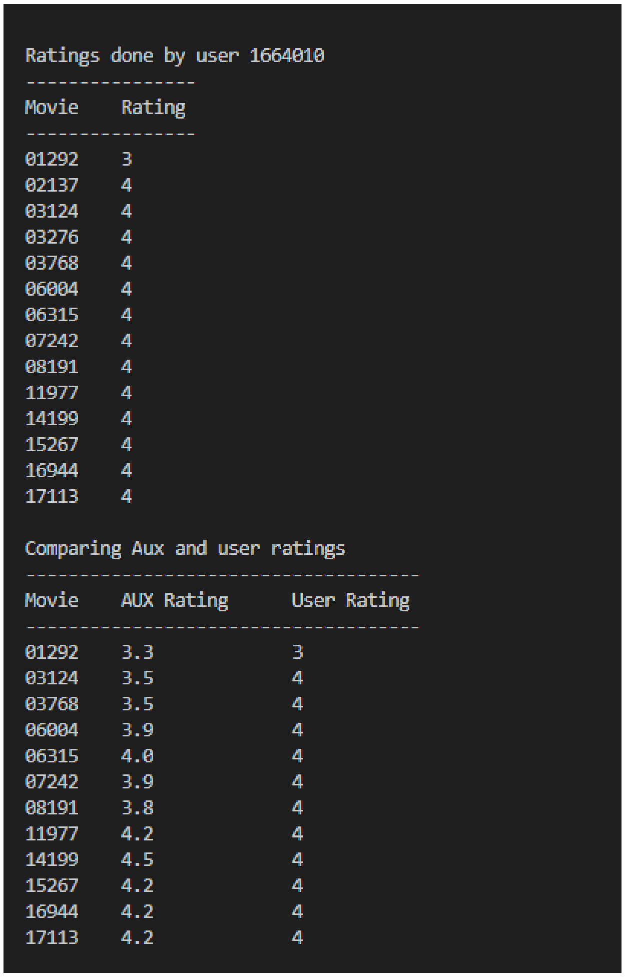
(d) What is the value of the eccentricity threshold? What is the difference between the highest and second highest score? Is it greater than the eccentricity metric?

The eccentricity with gamma value 0.1 is 0.0120683445592
Difference between the highest and second highest score 0.0104338240671
Difference between the highest and second highest score is lesser than the eccentricity metric

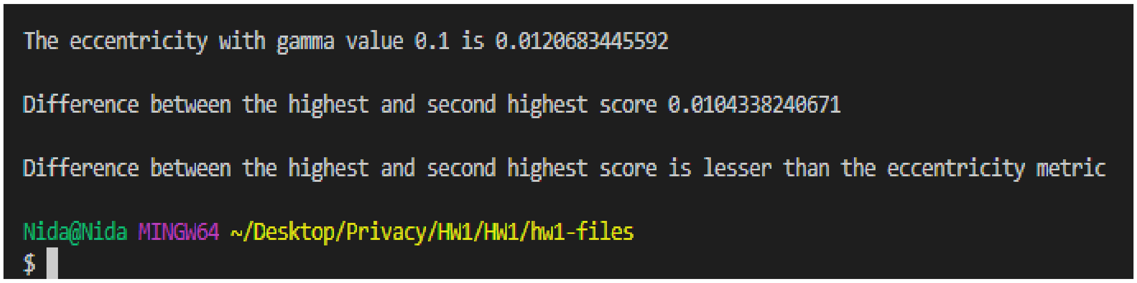
SCREENSHOTS:
P.T.O.



```
Nida@Nida MINGW64 ~/Desktop/Privacy/HW1/HW1/hw1-files
$ python link.py
----------------
Movie    Weight
----------------
03276    0.117848147906
06004    0.126459795051
14199    0.117771001943
17113    0.117810937409
06315    0.117731208043
01292    0.126506952176
11977    0.117819513515
15267    0.12656614447
08191    0.117822373672
16944    0.117745403957
07242    0.117793804816
03768    0.126465680175
02137    0.126436281655
10935    0.117790951923
03124    0.117708531879

The number of users are: 44651

Highest score:  0.113354503558

Second highest score:  0.102920679491

User with maximum score:  1664010
```



```
Ratings done by user 1664010
----------------
Movie    Rating
----------------
01292    3
02137    4
03124    4
03276    4
03768    4
06004    4
06315    4
07242    4
08191    4
11977    4
14199    4
15267    4
16944    4
17113    4

Comparing Aux and user ratings
-------------------------------------
Movie    AUX Rating       User Rating
-------------------------------------
01292    3.3              3
03124    3.5              4
03768    3.5              4
06004    3.9              4
06315    4.0              4
07242    3.9              4
08191    3.8              4
11977    4.2              4
14199    4.5              4
15267    4.2              4
16944    4.2              4
17113    4.2              4
```



```
The eccentricity with gamma value 0.1 is 0.0120683445592

Difference between the highest and second highest score 0.0104338240671

Difference between the highest and second highest score is lesser than the eccentricity metric

Nida@Nida MINGW64 ~/Desktop/Privacy/HW1/HW1/hw1-files
$
```

PROBLEM 2 :

a) QUASI- IDENTIFIERS :

① ZIP CODE

② AGE

SENSITIVE ATTRIBUTES :

① SALARY

② DISEASE

b) 3-ANONYMOUS 3-DIVERSE TABLE

FOR EQUIVALENCE CLASSES, LET US TAKE ZIP CODE FIRST:

Z2 : {476**, 479**}

Z1: {4767*, 4760*, 4790*}

Z0 : {47677, 47678, 47674, 47602, 47605, 47607, 47905, 47906, 47909}

TAKING AGE :

A1 : {<30, ≥ 30}

A0 : {29, 22, 27, 43, 30, 47, 36, 32, 52}

GENERALIZATION LATTICE

Z2 = {476**, 479**}

↑

Z1 = {47,67*, 4760*, 4790*}

↑

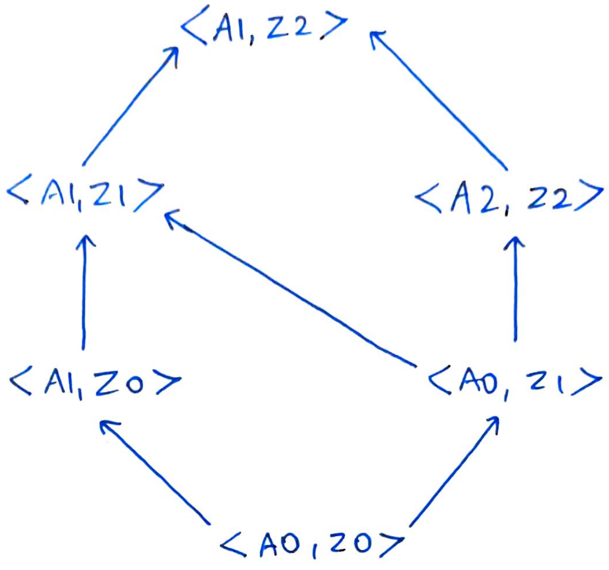
Z0 = {47677, ... , 47909}

ZIPCODE

A1 = {<30, ≥30}

↑

A0 = {29, ..., 52}



<A1, Z2> SATISFIES 3 ANONNYMITY

GENERALIZATION A1, Z2 SATISFIES THIS.

ZIP CODE	AGE	SALARY	DISEASE
476**	<30	3K	GASTRIC ULCER
476**	<30	4K	GASTRITIS
476**	<30	5K	STOMACH CANCER
476**	≥30	7K	FLU
476**	≥30	9K	BRONCHITIS
476**	≥30	10K	PNEUMONIA
479**	≥30	6K	GASTRITIS
479**	≥30	8K	BRONCHITIS
479**	≥30	11K	STOMACH CANCER

THIS TABLE IS DIVERSE ie) 3-DIVERSE AS EACH GROUP HAS 3 RECORDS FOR THE SENSITIVE ATTRIBUTES.

NOTE:

OTHER GENERALIZATIONS WITH A1: $\{<35, >35>$ OR A1: $\{\leq 30, >30>$ DID NOT YIELD 3-DIVERSE TABLE.

c) T-CLOSENESS:

FROM THE TABLE WE HAVE,

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

$P_1 = \{3K, 4K, 5K\}$

$P_2 = \{7K, 9K, 10K\}$

$P_3 = \{6K, 8K, 11K\}$

TRACE-DISTANCE:

$D[P_1, Q]$: TRANSFORM P_1 TO Q

- MOVE 1/9 PROBABILITY FOR EACH PAIR.
  - 3K → 6K, 3K → 7K

    COST: $\frac{1}{9}(3+4)/8$
  - 4K → 8K, 4K → 9K

    COST: $\frac{1}{9}(4+5)/8$
  - 5K → 10K, 5K → 11K

    COST: $\frac{1}{9}(5+6)/8$

TOTAL COST = $\frac{1}{9}(27)/8$

$= 0.375$

$D[P_2, Q]$: TRANSFORM P_2 TO Q

- MOVE 1/9 PROBABILITY FOR EACH PAIR:
  - 7K → 3K, 7K → 4K
    COST : $\frac{1}{9}(4+3)/8$
  - 9K → 5K, 9K → 6K
    COST : $\frac{1}{9}(4+3)/8$
  - 10K → 8K, 10K → 11K
    COST : $\frac{1}{9}(2+1)/8$

TOTAL COST = $\frac{1}{9}(17)/8$

$= 0.2361$

$D[P_3, Q]$: TRANSFORM P_3 TO Q

- MOVE 1/9 PROBABILITY FOR EACH PAIR:
  - 6K → 3K, 6K → 4K
    - COST : $\frac{1}{9}(3+2)/8$
  - 8K → 5K, 8K → 7K
    - COST : $\frac{1}{9}(3+1)/8$
  - 11K → 9K, 11K → 10K
    - COST : $\frac{1}{9}(2+1)/8$

$$\text{TOTAL COST} = \frac{1}{9}(12)/8$$

$$= 0.1667$$

$D[P_3, Q] = 0.1667 \Rightarrow P_3$ REVEALS LESS PRIVATE DATA

AVERAGE OF $D[P_1, Q], D[P_2, Q] \& D[P_3, Q] = \dfrac{0.375 + 0.2361 + 0.1667}{3}$

$$= 0.2592$$

THIS SOLUTION DOES NOT RESOLVE THE SIMILARITY ATTACK

IF I HAVE BACKGROUNG KNOWLEDGE THAT SOMEONE EARNS MORE THAN 10K AND HAS AN AGE MORE THAN 30, THEN THEY SUFFER FROM A STOMACH AILMENT.

ALTERNATIVELY, IF I JUST KNOW THAT SALARY >10K, THEN I CAN INFER THAT THE PERSON HAS CANCER OF STOMACH. ALSO RANGE OF 3K-5K TELLS ME THAT HE HAS A STOMACH DISEASE.

ALTERNATIVE SOLUTION:

$A1: \{\leq 40, > 40\}$

$A0: \{29, 22, 27, 30, 36, 32, 43, 47, 52\}$

WE CHANGE THE EQUIVALENCE CLASSES. FOR AGE.

SIMILARLY WE GET, GENERALIZATION LATTICE

AGAIN $\langle A1, Z2 \rangle$ SATISFIES 3-ANONYMITY

	ZIP CODE	AGE	SALARY	DISEASE
	476**	≤40	3K	GASTRIC ULCER
INTER CHANGE	476**	≤40	9K	BRONCHITIS
	476**	≤40	5K	STOMACH CANCER
	476**	≤40	4K	GASTRITIS
	476**	≤40	7K	FLU
	476**	≤40	10K	PNEUMONIA
	479**	>40	6K	GASTRITIS
	479**	>40	8K	BRONCHITIS
	479**	>40	11K	STOMACH CANCER

COMPUTING T-CLOSENESS :

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

$P_1 = \{3K, 5K, 9K\}$

$P_2 = \{4K, 7K, 10K\}$

$P_3 = \{6K, 8K, 11K\}$

$D[P_1, Q]$: TRANSFORM P_1 TO Q

- $3K \to 4K$, $3K \to 6K$ COST : $1/9\ (1+3)/8$
- $5K \to 7K$, $5K \to 8K$ COST : $1/9\ (2+3)/8$
- $9K \to 10K$, $9K \to 11K$ COST : $1/9\ (1+2)/8$

TOTAL COST = $\frac{1}{9}(12)/8 = 0.1667$

$D[P_2, Q]$: TRANSFORM P_2 TO Q

- 4K → 3K, 4K → 5K COST : $1/9\,(1+1)/18$
- 7K → 6K, 7K → 8K COST : $1/9\,(1+1)/18$
- 10K → 9K, 10K → 11K COST : $1/9\,(1+1)\,18$

TOTAL COST : $\frac{1}{9}(6)/18 = 0.0833$

$D[P_3, Q]$: TRANSFORM P_3 TO Q

- 6K → 3K, 6K → 4K COST : $1/9\,(3+2)/18$
- 8K → 5K, 8K → 7K COST : $1/9\,(3+1)/18$
- 11K → 9K, 11K → 10K COST : $1/9\,(2+1)/18$

TOTAL COST : $\frac{1}{9}(12)/18 = 0.1667$

$D[P_2, Q]$: 0.0833, P_2 REVEALS PRIVATE DATA

AVERAGE OF EQUIVALENCE CLASSES = $\frac{0.1667 + 0.0833 + 0.1667}{3}$

$= 0.1389$

THIS RESOLVES THE SIMILARITY ATTACK.

ANY RANGE OF SALARY I SAMPLE IN THE GROUP, I DO NOT KNOW IF THEY SUFFER DEFINITETIVELY FROM A RESPIRATORY OR GASTRIC ILLNESS.