**Question1:**

Dimensions of the TF-IDF matrix: (4732, 16080)

**Question2:**

Confusion matrix:

[[ 72 2271]

[2305 84]]

By inspecting the contingency table, we found that the clustering result was actually pretty good, most of the points are clustered correctly.

5 measures for the K-means:

Homogeneity Score: 0.791

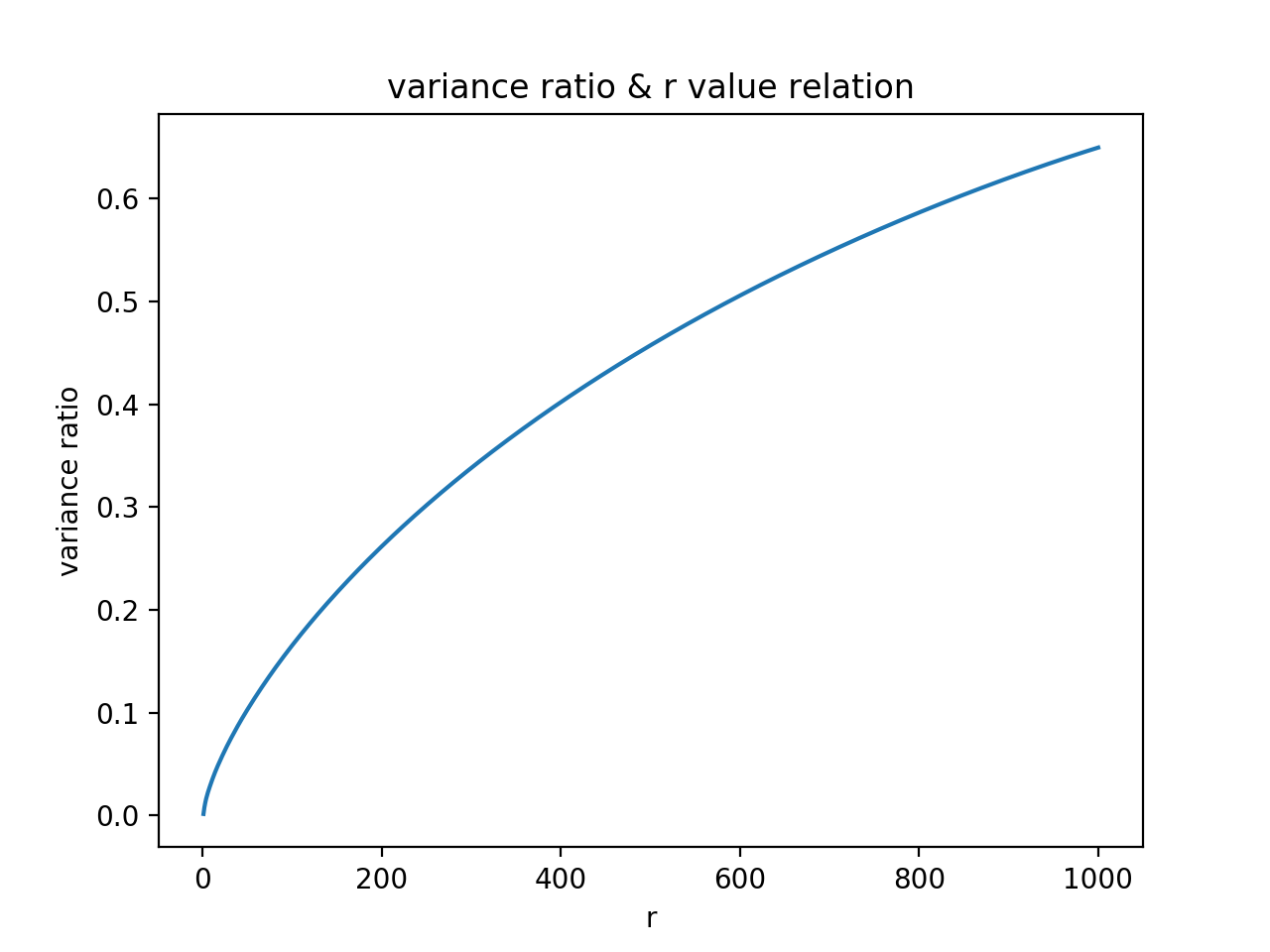
Completeness Score: 0.791

V-measure: 0.791

Adjusted Rand Score: 0.872

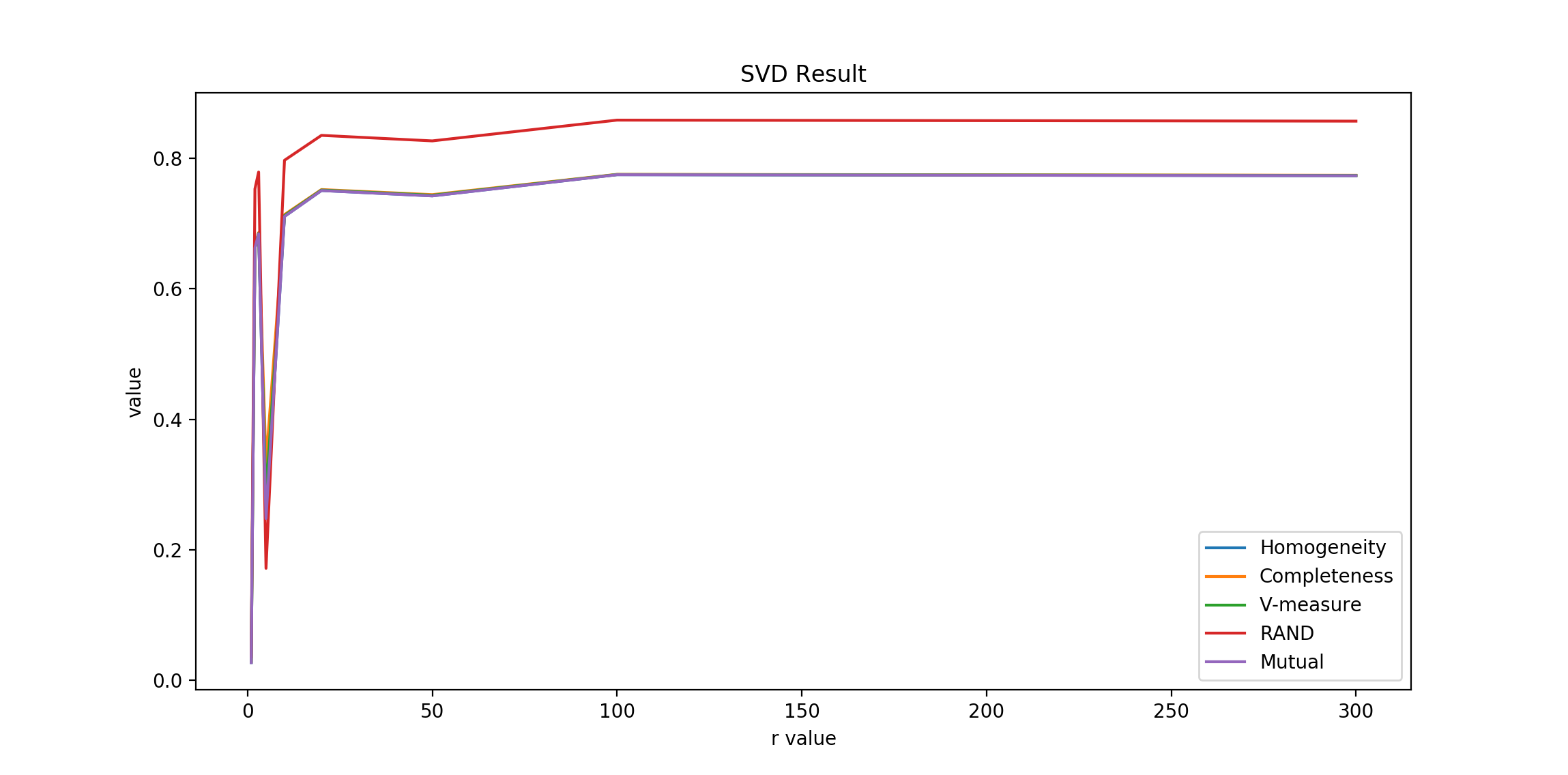
Adjusted Mutual Info Score: 0.791

**Question3:**

****

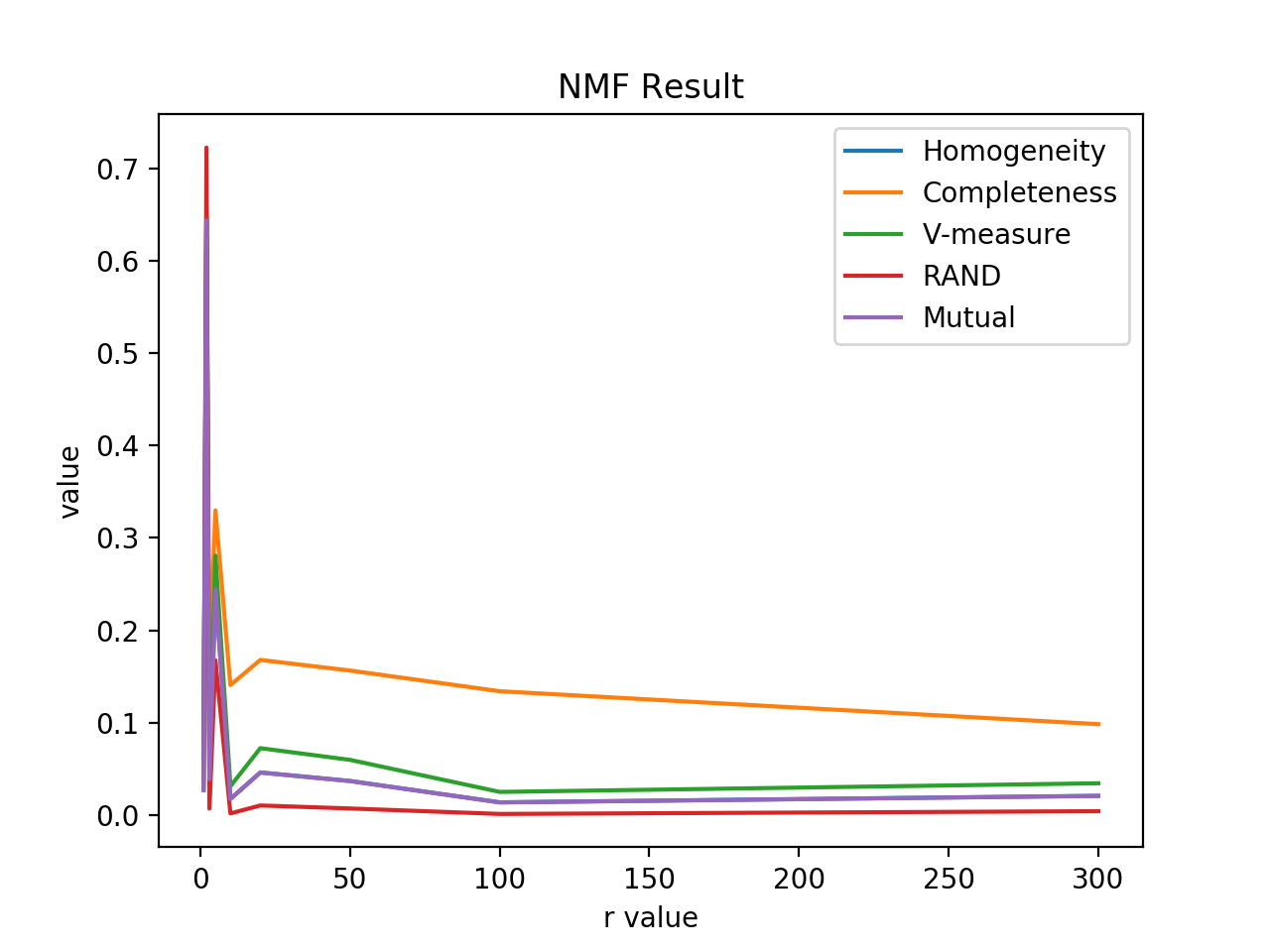
SVD

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Homogeneity | Completeness | V-measure | Rand Score | Mutual Info |
| r=1 | 0.791 | 0.791 | 0.791 | 0.872 | 0.791 |
| r=2 | 0.027 | 0.028 | 0.028 | 0.036 | 0.027 |
| r=3 | 0.644 | 0.651 | 0.647 | 0.722 | 0.644 |
| r=5 | 0.039 | 0.169 | 0.063 | 0.007 | 0.039 |
| r=10 | 0.244 | 0.330 | 0.281 | 0.168 | 0.244 |
| r=20 | 0.750 | 0.752 | 0.751 | 0.835 | 0.750 |
| r=50 | 0.742 | 0.744 | 0.743 | 0.826 | 0.742 |
| r=100 | 0.775 | 0.775 | 0.775 | 0.858 | 0.775 |
| r=300 | 0.773 | 0.774 | 0.773 | 0.857 | 0.773 |



NMF

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Homogeneity | Completeness | V-measure | Rand Score | Mutual Info |
| r=1 | 0.027 | 0.029 | 0.028 | 0.036 | 0.027 |
| r=2 | 0.662 | 0.666 | 0.664 | 0.753 | 0.662 |
| r=3 | 0.684 | 0.686 | 0.685 | 0.779 | 0.684 |
| r=5 | 0.248 | 0.332 | 0.284 | 0.172 | 0.248 |
| r=10 | 0.711 | 0.714 | 0.712 | 0.797 | 0.711 |
| r=20 | 0.046 | 0.168 | 0.072 | 0.010 | 0.046 |
| r=50 | 0.037 | 0.156 | 0.060 | 0.007 | 0.037 |
| r=100 | 0.014 | 0.134 | 0.025 | 0.001 | 0.014 |
| r=300 | 0.021 | 0.098 | 0.034 | 0.004 | 0.021 |



for the result above we can see that:

Best r for SVD: 100

Best r for NMF: 2

The non-monotomic behavior of the measures as r increases is because as r increases the information contained is more intact, which means that more information and noises will be included. This also means the data is less compressed. There should be a balance between information and noise, which will be our best r.

Confusion Matrix:

svd with r=1

[[1114 1229]

[ 685 1704]]

nmf with r=1

[[1232 1111]

[1705 684]]

svd with r=2

[[2095 248]

[ 65 2324]]

nmf with r=2

[[ 313 2030]

[2347 42]]

svd with r=3

[[ 200 2143]

[2311 78]]

nmf with r=3

[[2166 177]

[2389 0]]

svd with r=5

[[2342 1]

[1384 1005]]

nmf with r=5

[[ 1 2342]

[ 993 1396]]

svd with r=10

[[2139 204]

[ 50 2339]]

nmf with r=10

[[2262 81]

[2389 0]]

svd with r=20

[[2190 153]

[ 51 2338]]

nmf with r=20

[[2121 222]

[2387 2]]

svd with r=50

[[2176 167]

[ 48 2341]]

nmf with r=50

[[2162 181]

[2387 2]]

svd with r=100

[[ 113 2230]

[2328 61]]

nmf with r=100

[[2279 64]

[2389 0]]

svd with r=300

[[2226 117]

[ 59 2330]]

nmf with r=300

[[ 146 2197]

[ 13 2376]]

**Question4**:

==================================================

normal svd

==================================================

Confusion Matrix:

[[1396 947]

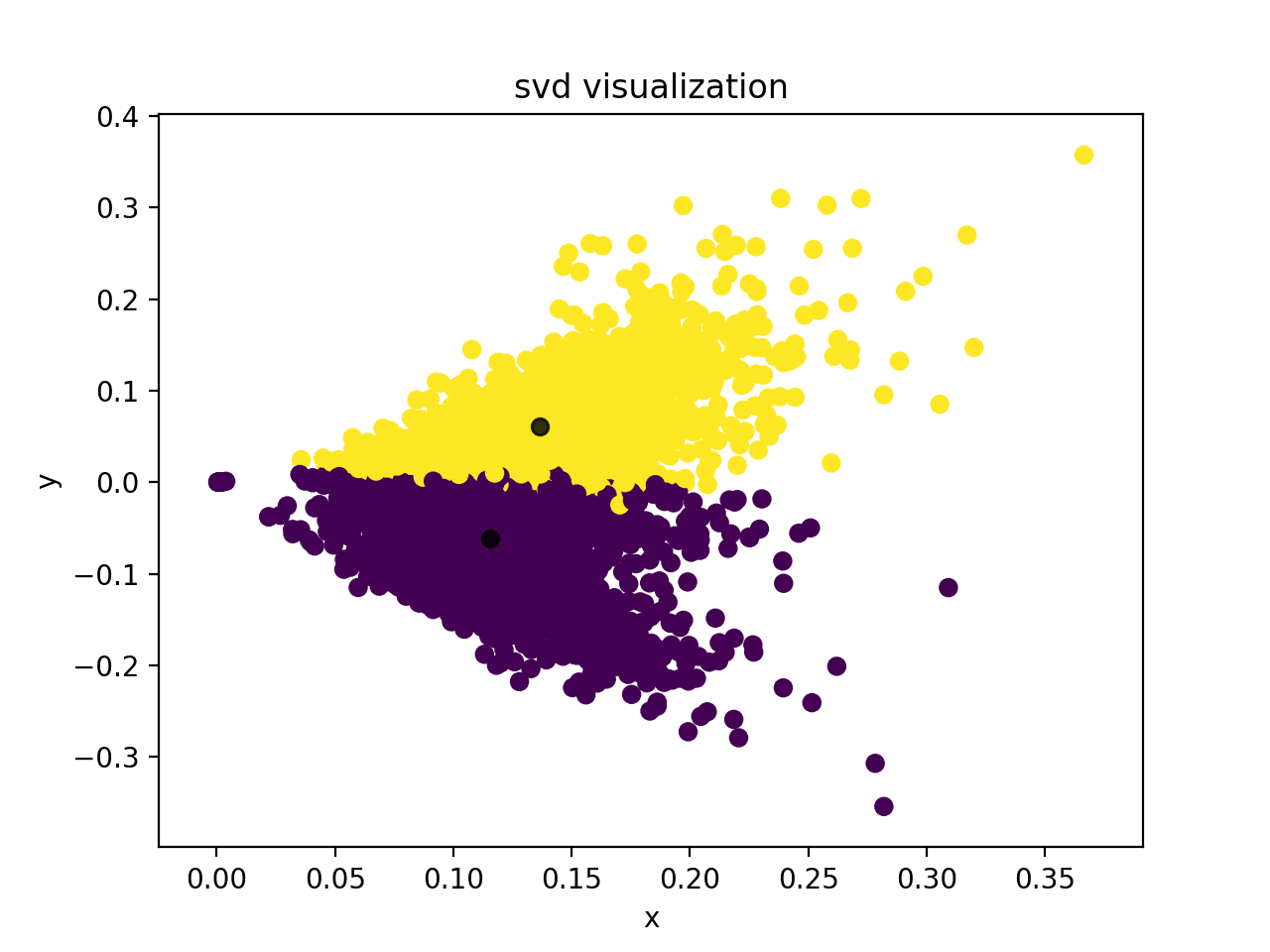
[ 119 2270]]

Homogeneity Score: 0.278

Completeness Score: 0.308

V-measure: 0.292

Adjusted Rand Score: 0.302

Adjusted Mutual Info Score: 0.278

==================================================

normal nmf

==================================================

Confusion Matrix:

[[ 139 2204]

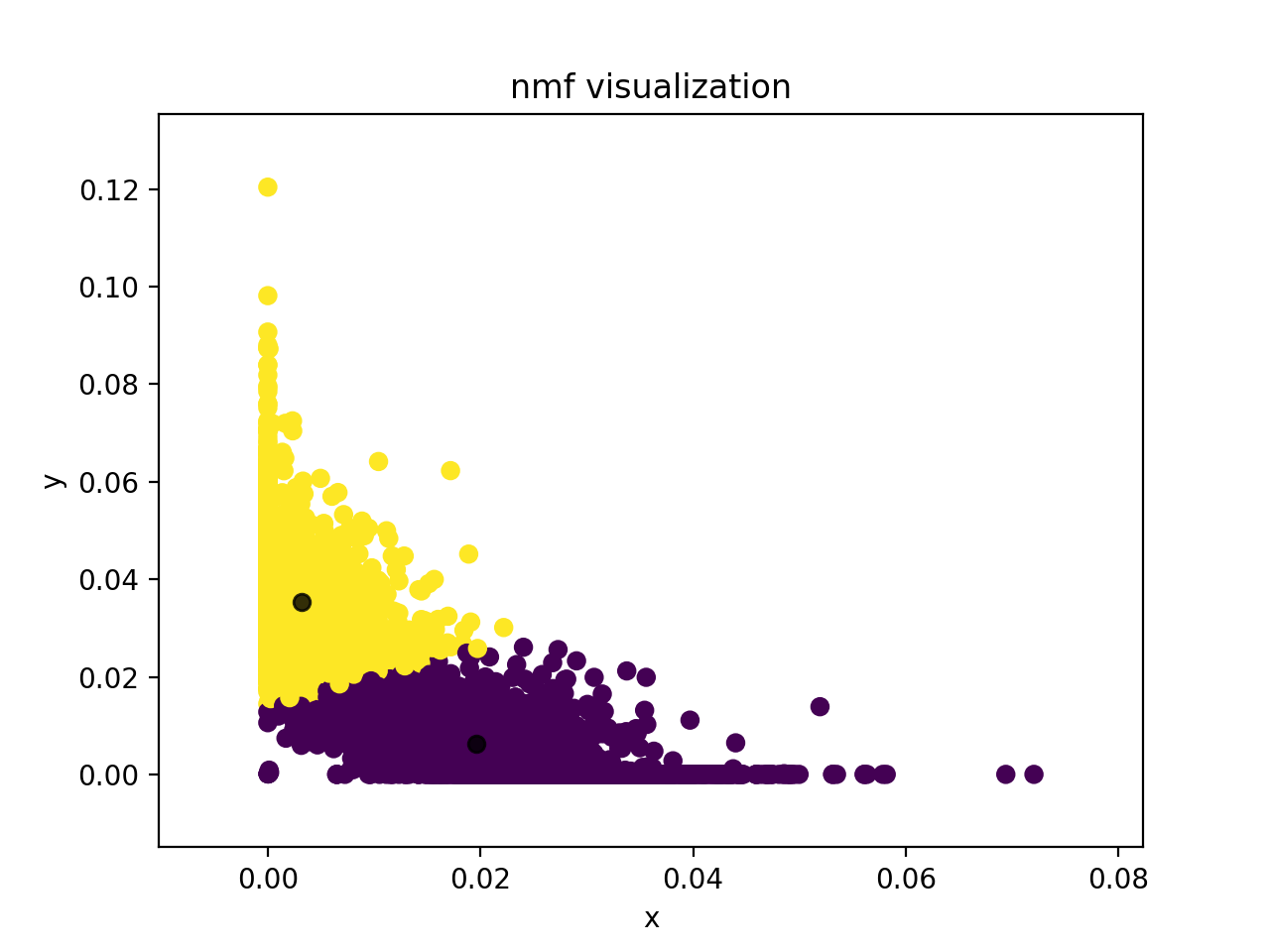
[2296 93]]

Homogeneity Score: 0.719

Completeness Score: 0.719

V-measure: 0.719

Adjusted Rand Score: 0.813

Adjusted Mutual Info Score: 0.719

==================================================

normalized svd

==================================================

Confusion Matrix:

[[1396 947]

[ 119 2270]]

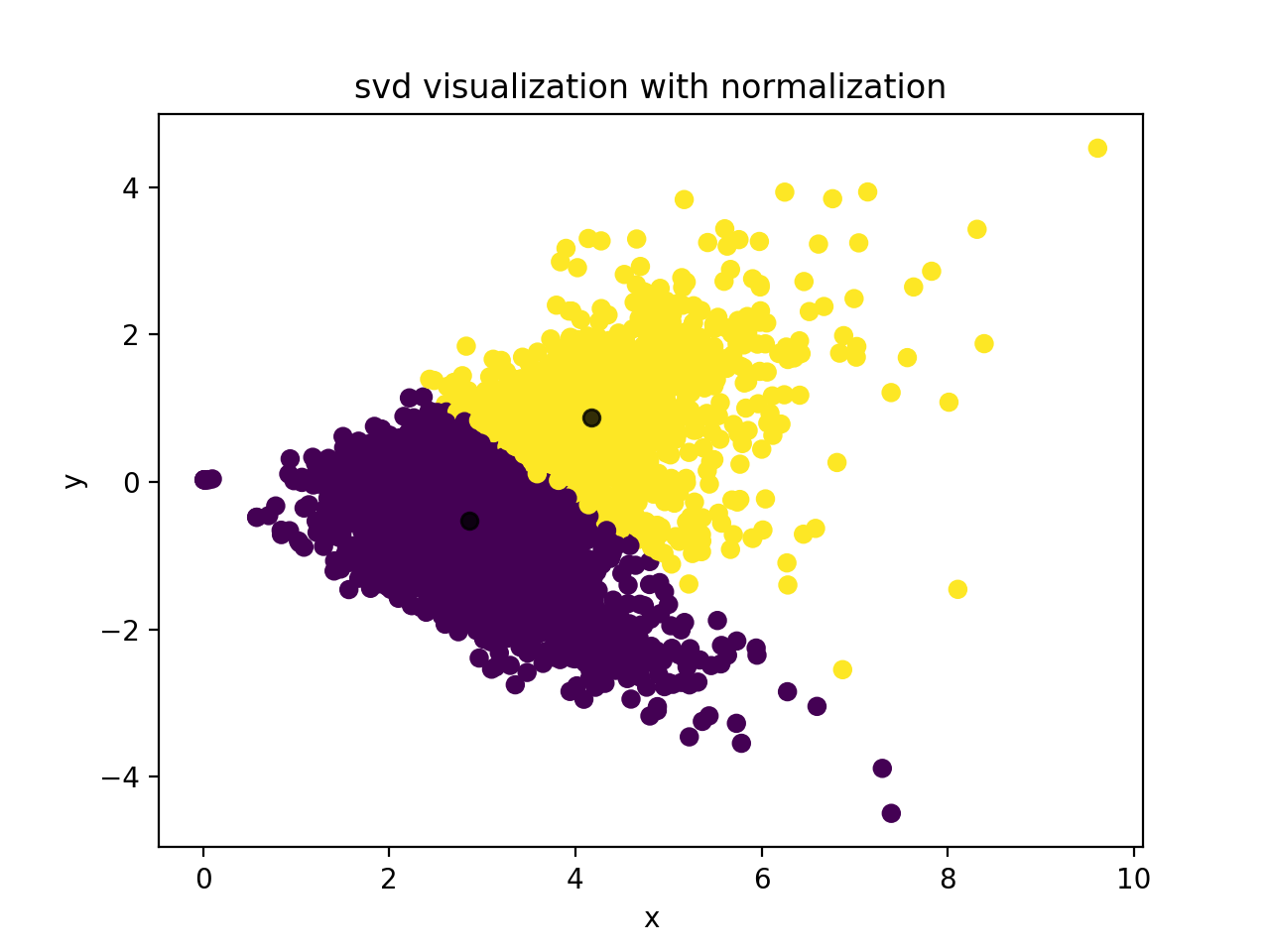
Homogeneity Score: 0.278

Completeness Score: 0.308

V-measure: 0.292

Adjusted Rand Score: 0.302

Adjusted Mutual Info Score: 0.278



==================================================

normalized nmf

==================================================

Confusion Matrix:

[[ 139 2204]

[2296 93]]

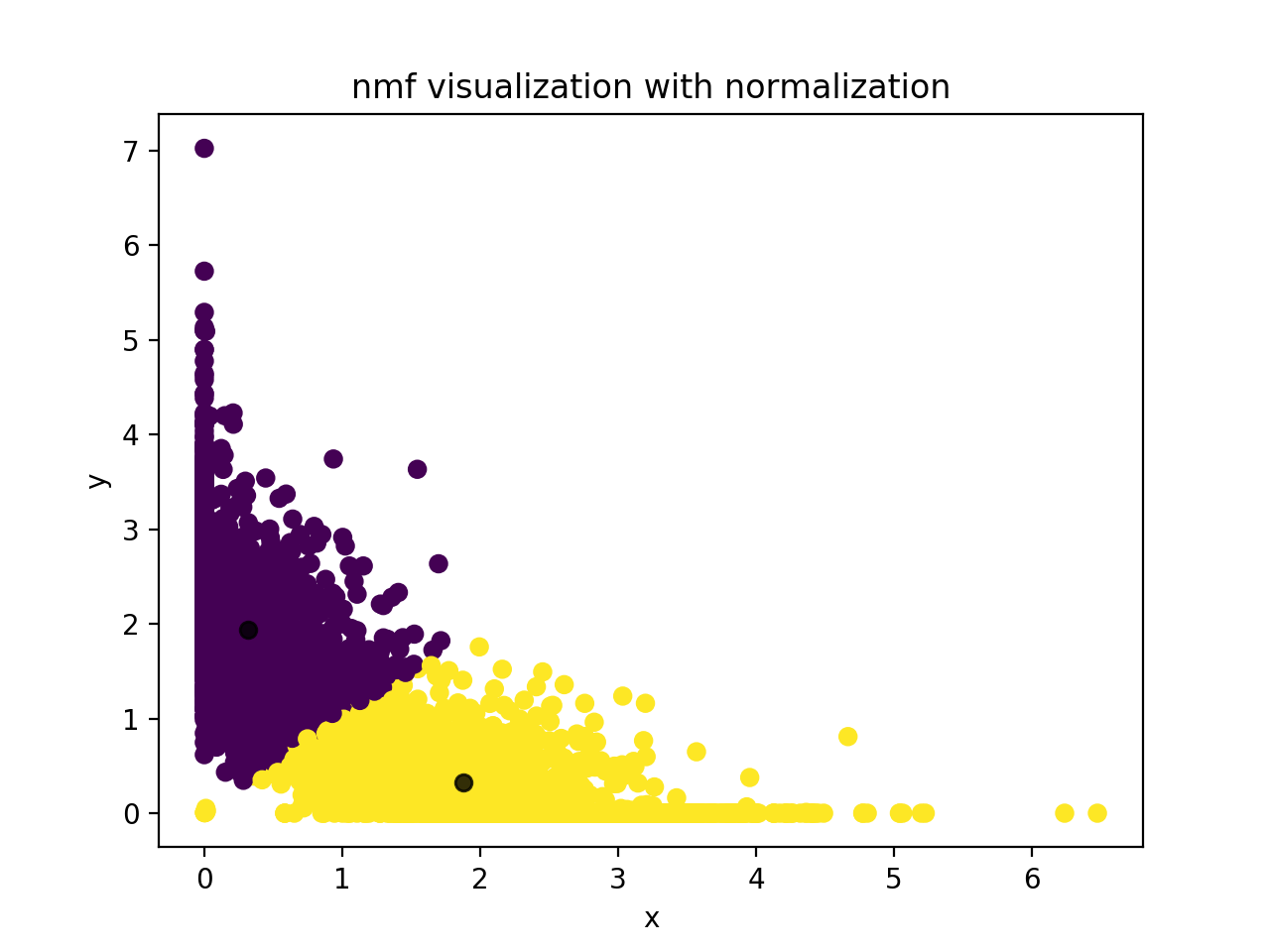
Homogeneity Score: 0.719

Completeness Score: 0.719

V-measure: 0.719

Adjusted Rand Score: 0.813

Adjusted Mutual Info Score: 0.719



==================================================

non-linear transformation for nmf

==================================================

Confusion Matrix:

[[ 301 2042]

[2347 42]]

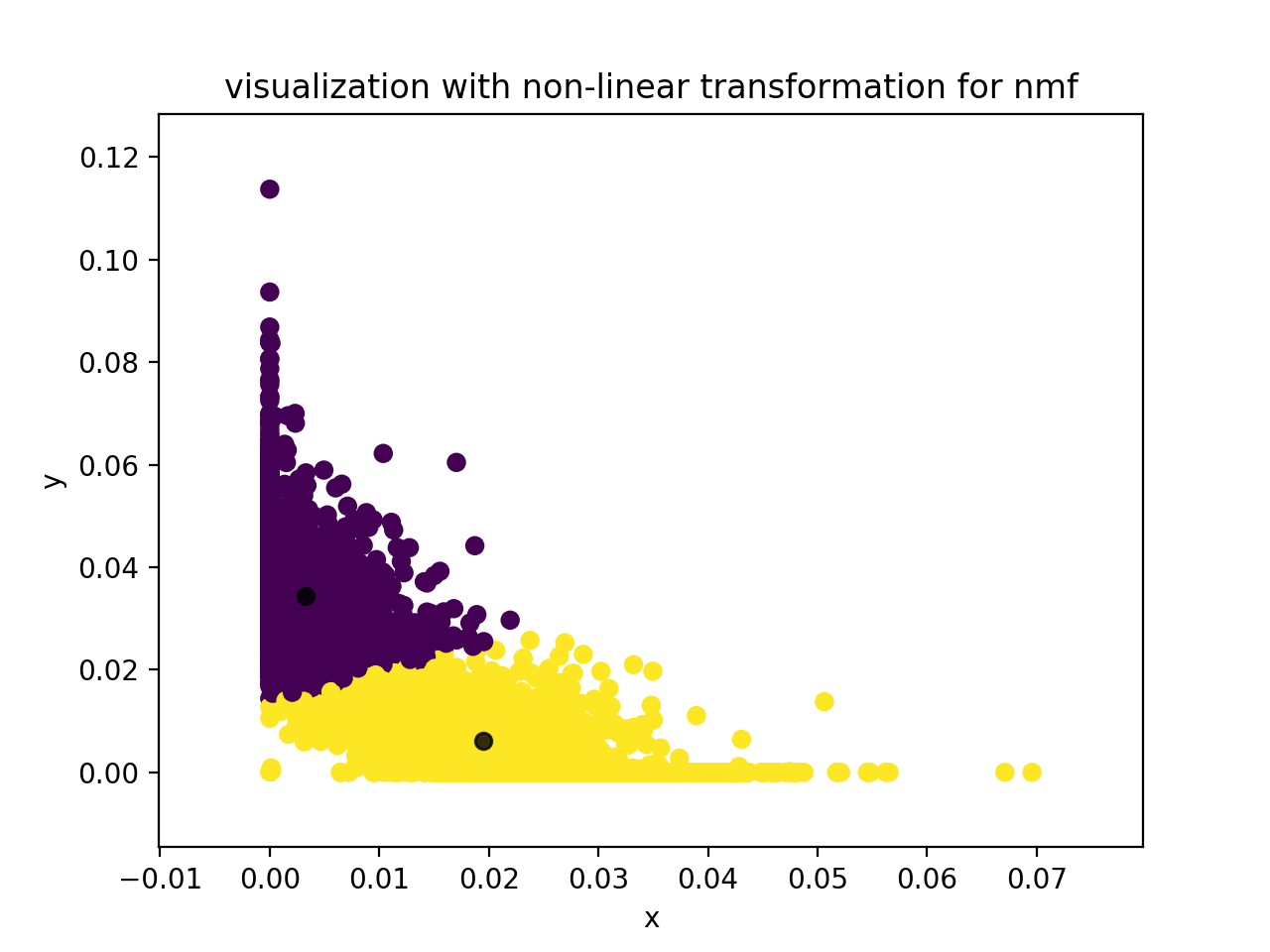
Homogeneity Score: 0.651

Completeness Score: 0.658

V-measure: 0.655

Adjusted Rand Score: 0.731

Adjusted Mutual Info Score: 0.651



==================================================

non-linear first and then normalization

==================================================

Confusion Matrix:

[[2225 118]

[ 101 2288]]

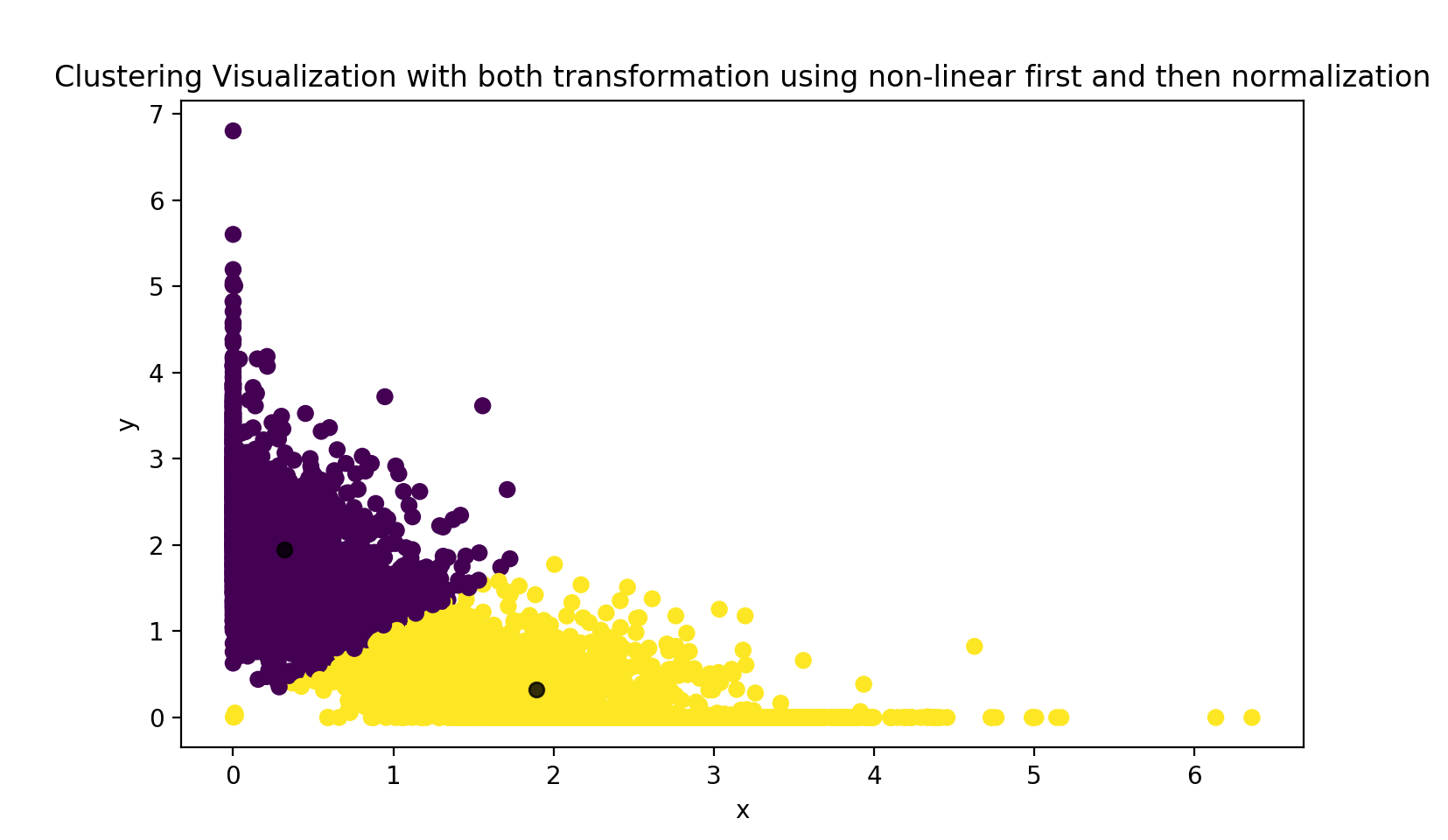
Homogeneity Score: 0.730

Completeness Score: 0.730

V-measure: 0.730

Adjusted Rand Score: 0.823

Adjusted Mutual Info Score: 0.730



==================================================

normalization first and then non-linear

==================================================

Confusion Matrix:

[[ 119 2224]

[2288 101]]

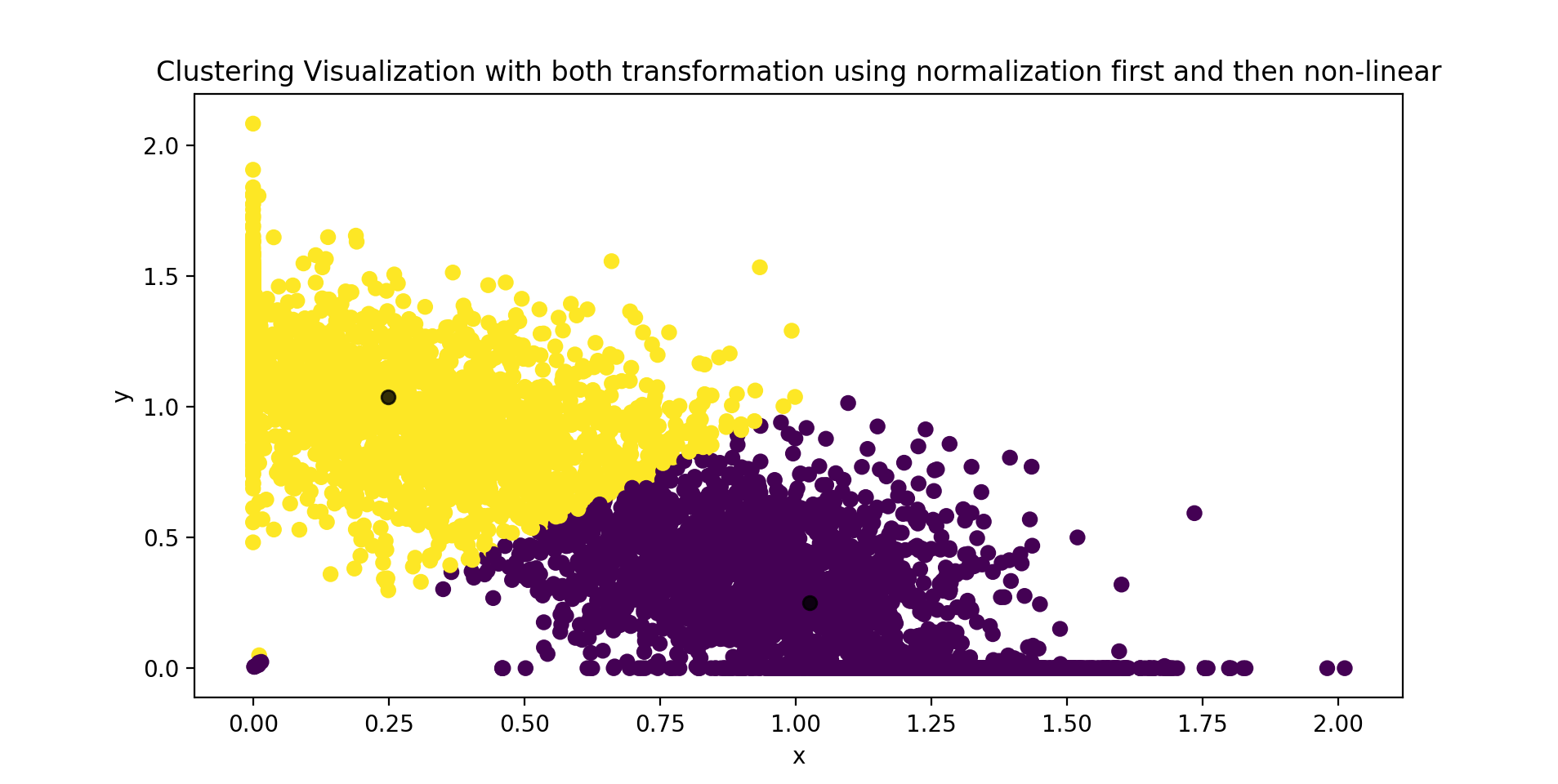
Homogeneity Score: 0.729

Completeness Score: 0.729

V-measure: 0.729

Adjusted Rand Score: 0.823

Adjusted Mutual Info Score: 0.729



Logarithm transformation will increase the clustering result because this transformation makes the data more organized, so that less noise is contained in the dataset. This fact makes clustering much easier.