**Kushagra Rastogi**

**304640248**

**ECE 219**

**Project 2: Clustering**

**QUESTION 1:**

Shape of TF-IDF matrix: (7882, 27768)

**QUESTION 2:**

|  |  |  |
| --- | --- | --- |
| Contingency Table | | |
|  | Cluster 1 (Predicted) | Cluster 2 (Predicted) |
| Class 1 (True) | 4 | 3899 |
| Class 2 (True) | 1718 | 2261 |

**QUESTION 3:**

Homogeneity score: 0.2535958928926043

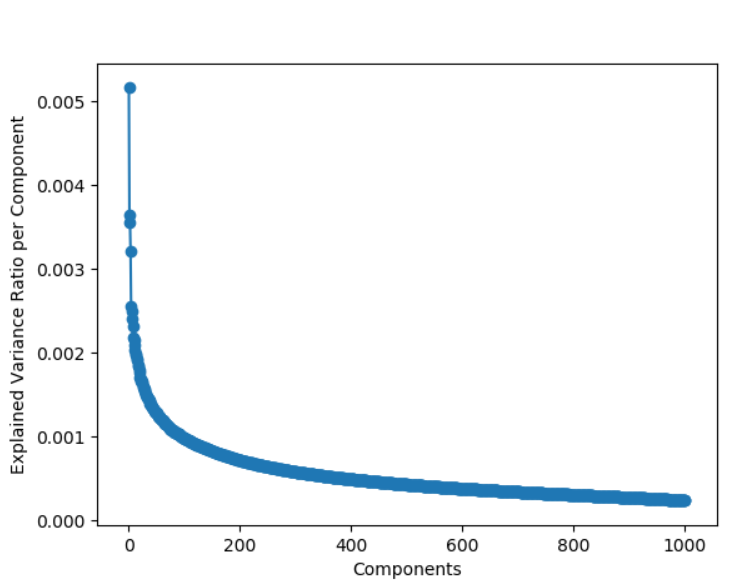
Completeness score: 0.334815748824373

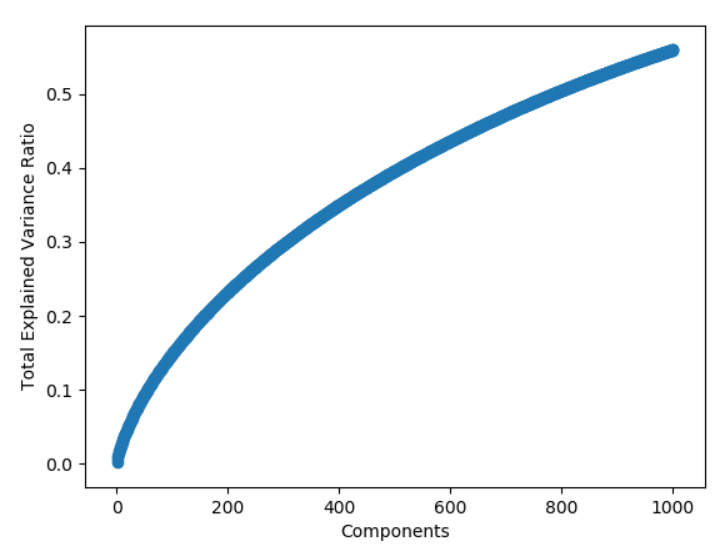
V-measure score: 0.28860033608397917

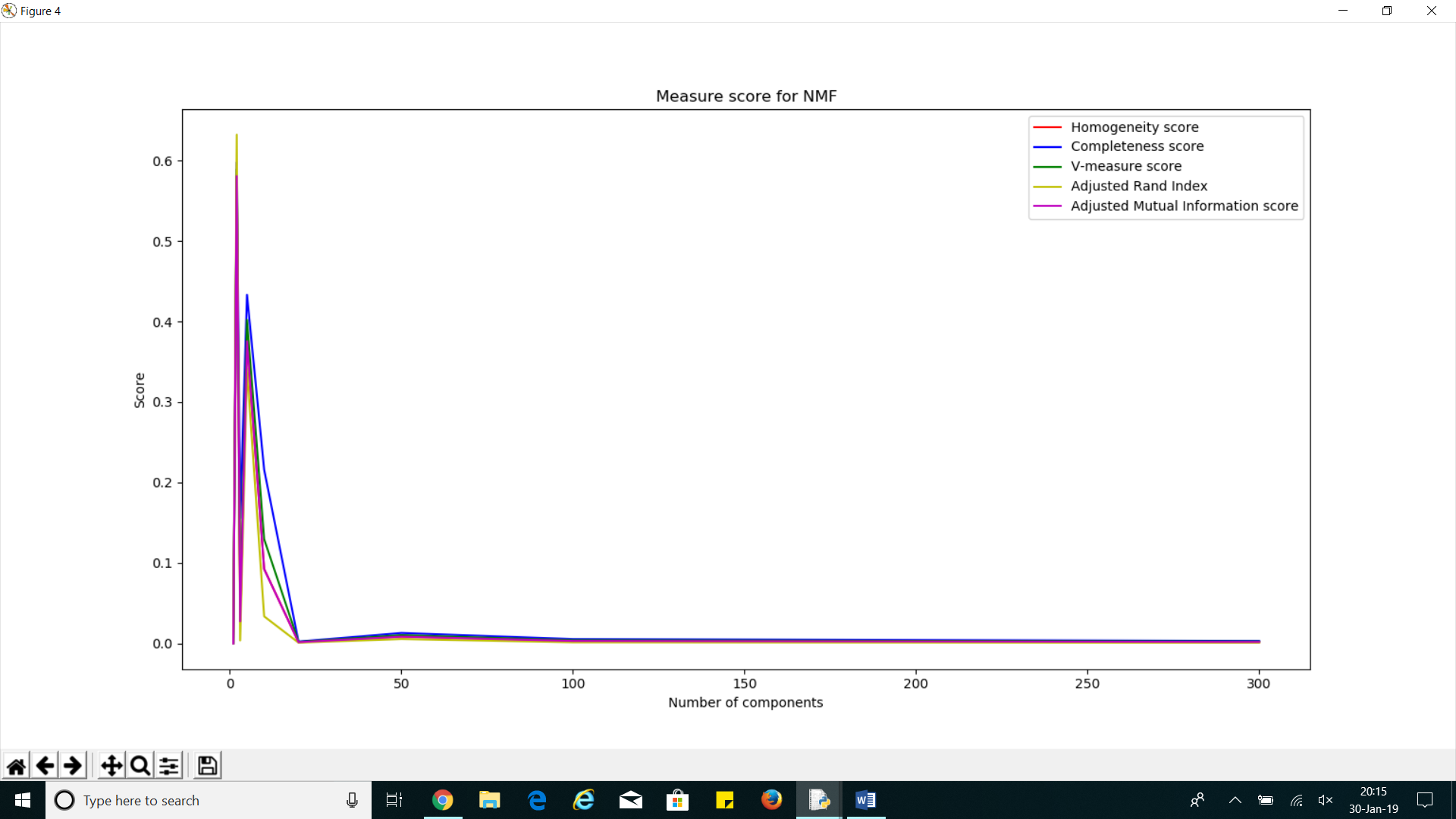
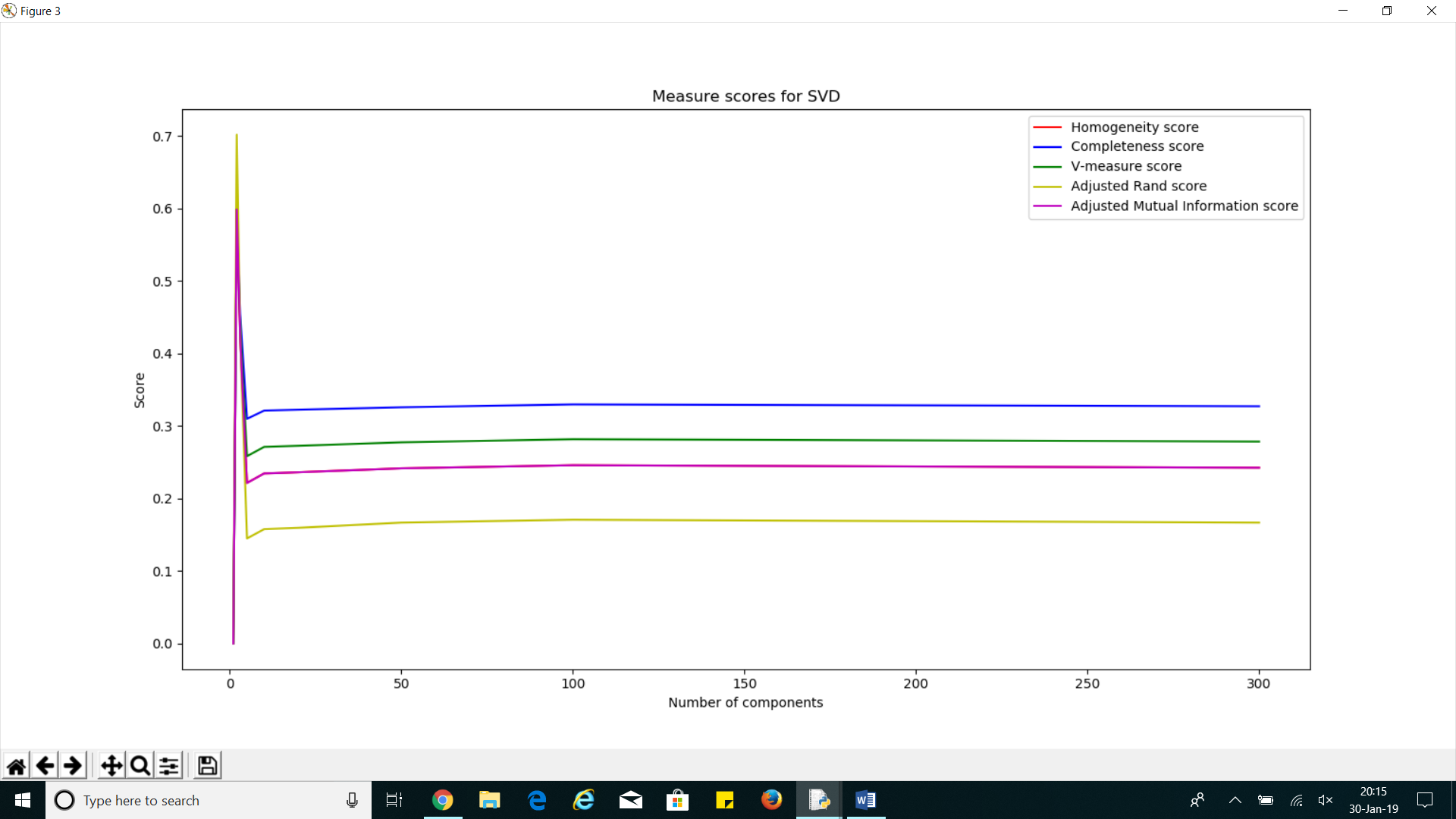
Adjusted Rand Index: 0.18076179588914554

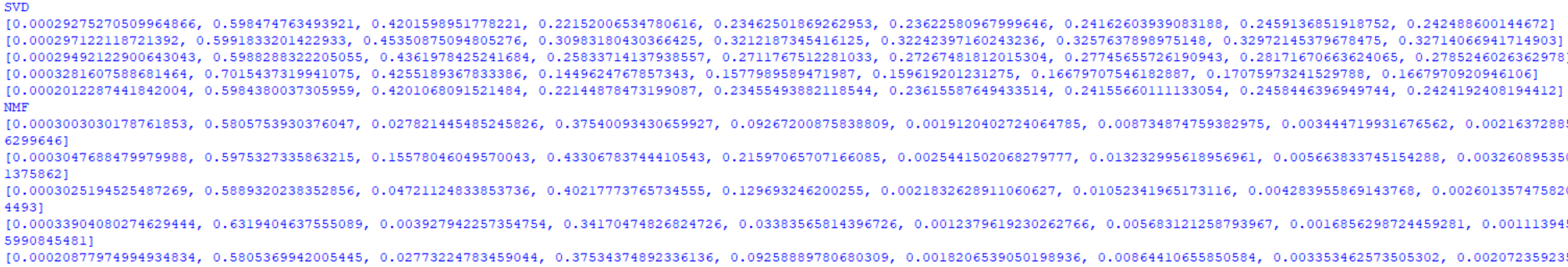
Adjusted mutual information score: 0.25352755133060884

**QUESTION 4:**





**QUESTION 5:**



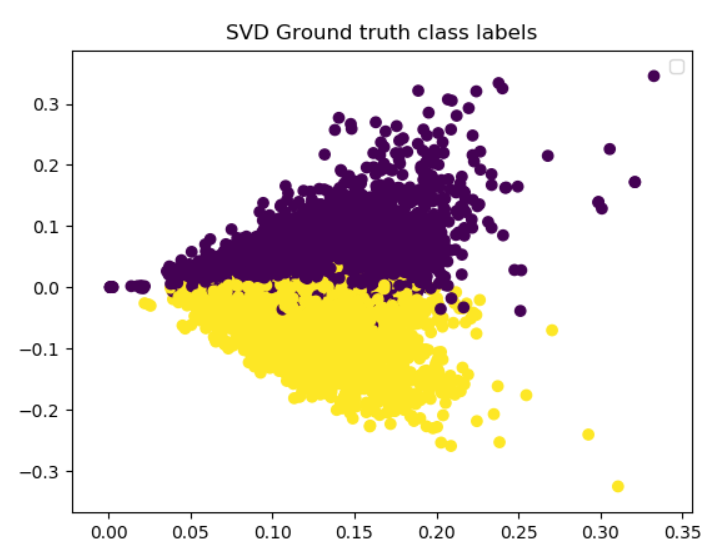
The first row of each dimensionality reduction algorithm corresponds to homogeneity score. The second row corresponds to completeness score. The third row corresponds to V-measure score. The fourth row corresponds to adjusted Rand Index. The fifth row corresponds to adjusted mutual information score.

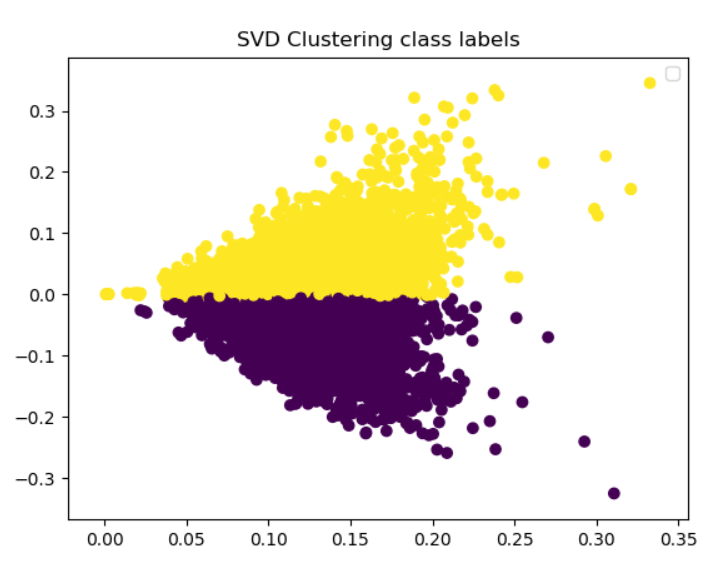
As it can be seen from the two graphs, some values of give the best scores for all the measures across the board. In the case of SVD, gave the best result for all measure scores. This was also the case for NMF. Thus, was chosen for both SVD and NMF.

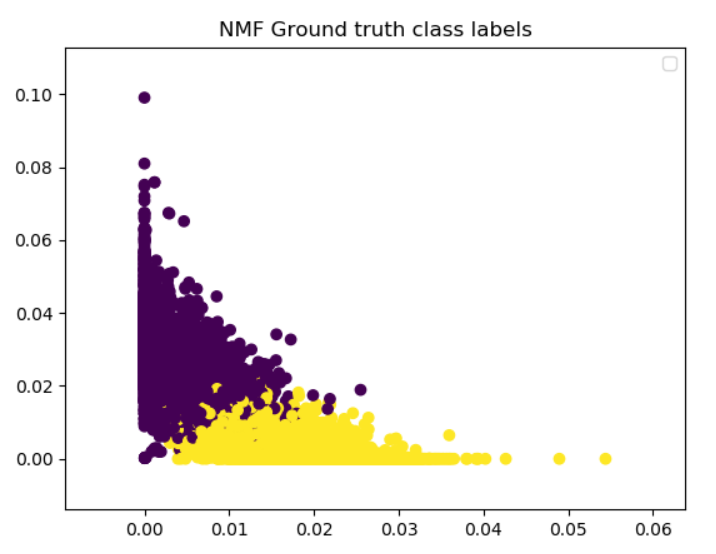
**QUESTION 6:**

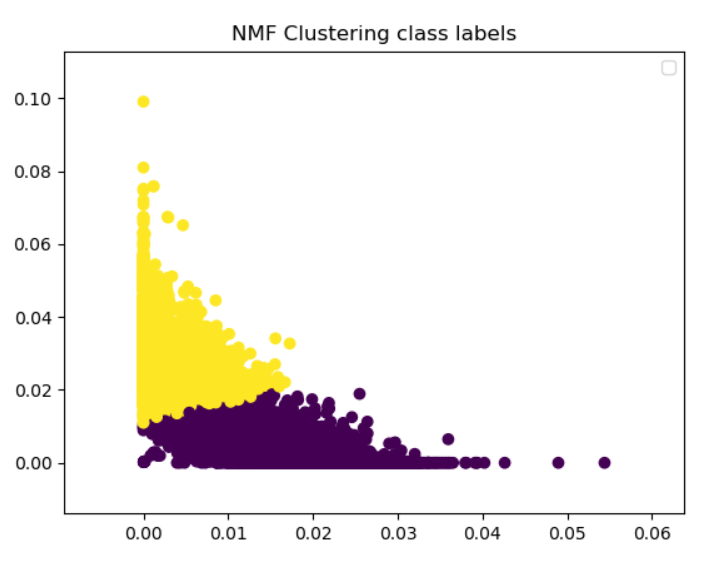
There is a non-monotonic behavior in the measures as increases. All the measures follow the same pattern: they initially peak, then fall down and eventually plateau off. As the number of components increases, the dimensions in which k-means needs to perform clustering increases. It is a well-known fact that k-means suffers from the curse of dimensionality because the Euclidean distance is not a good metric in high dimensions since the ratio between the nearest and farthest points approaches 1. This means that points in high dimensions are essentially equidistant from each other which makes it hard to perform clustering. As a result, increasing the number of components after the “Elbow point” adds no new information. Since no new information is given to the k-means algorithm, the measures remain constant and do not change.

**QUESTION 7:**

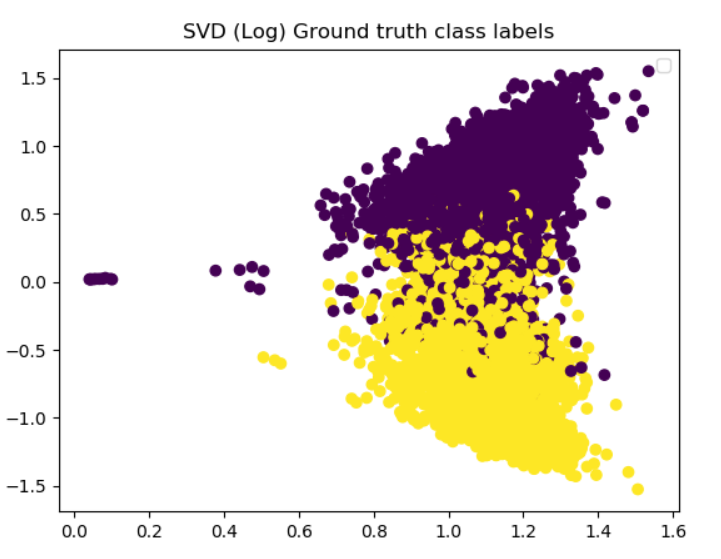


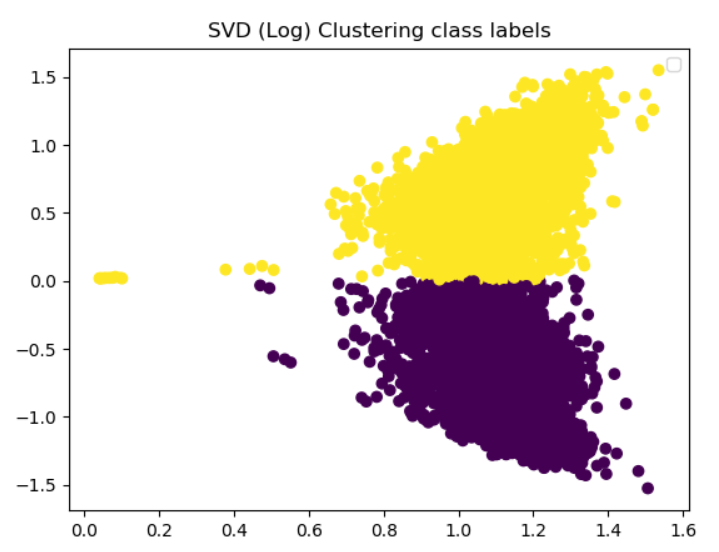


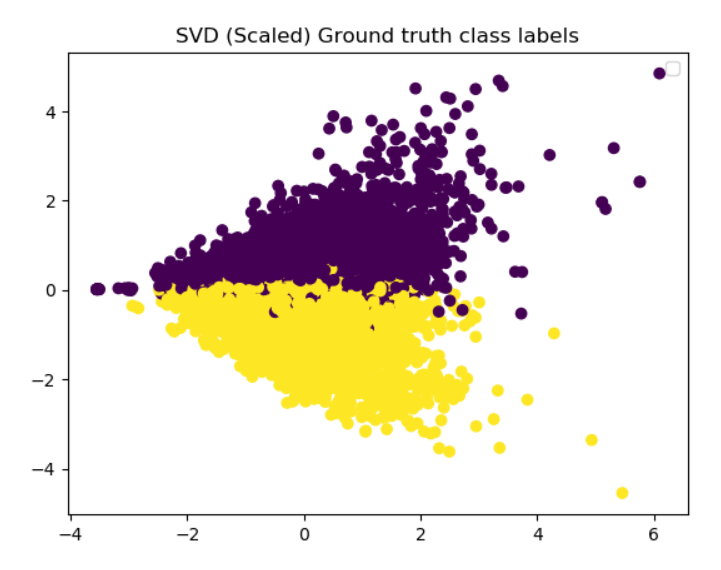


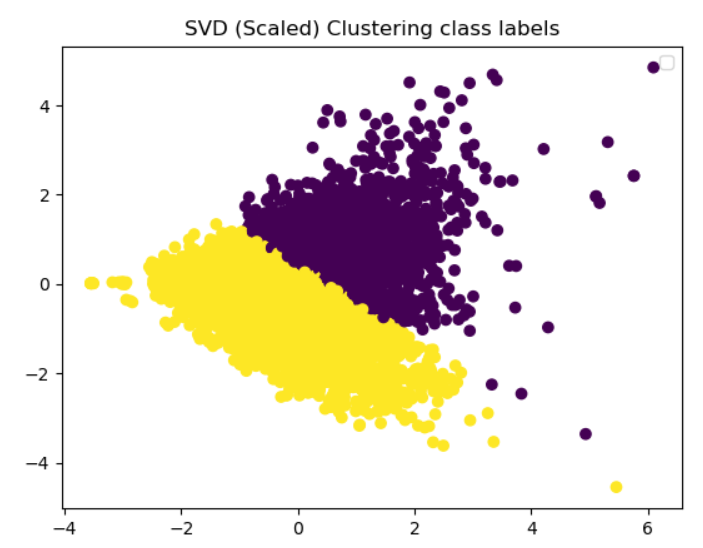


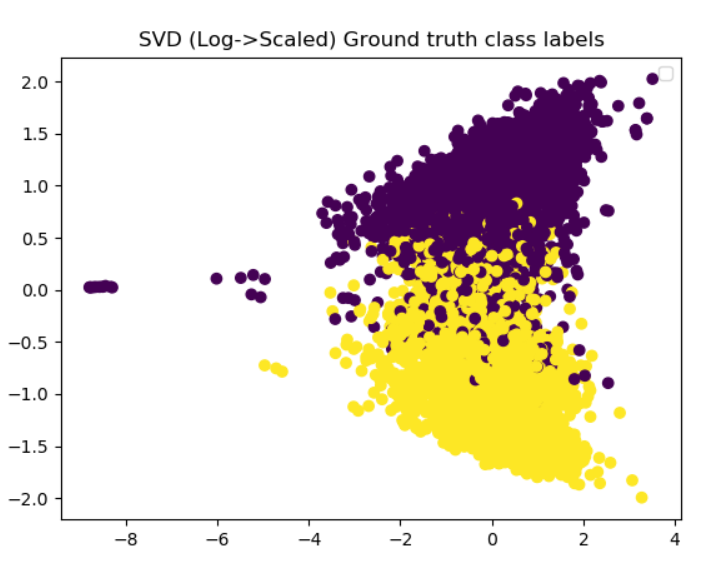
**QUESTION 8:**

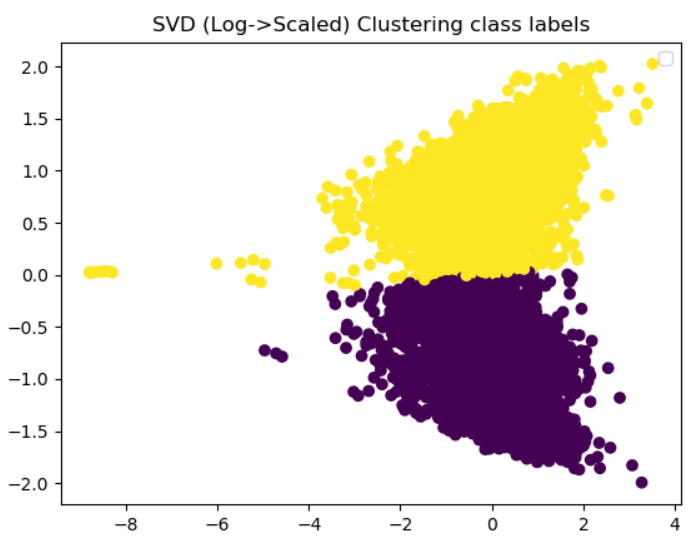






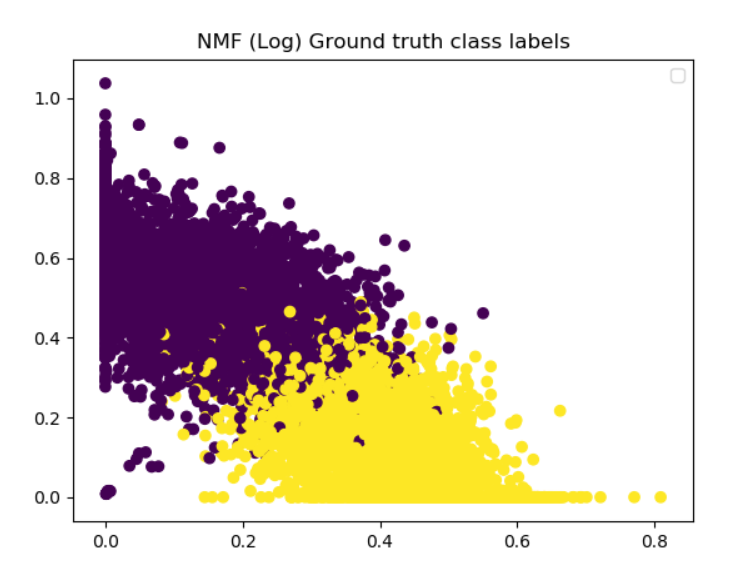




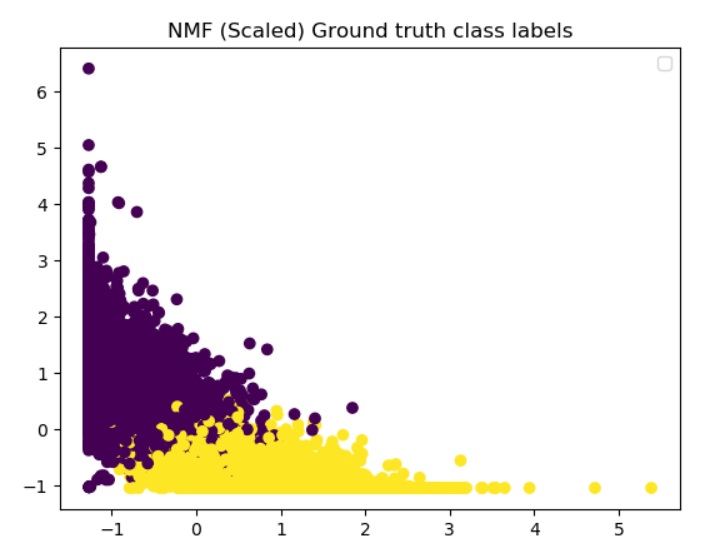


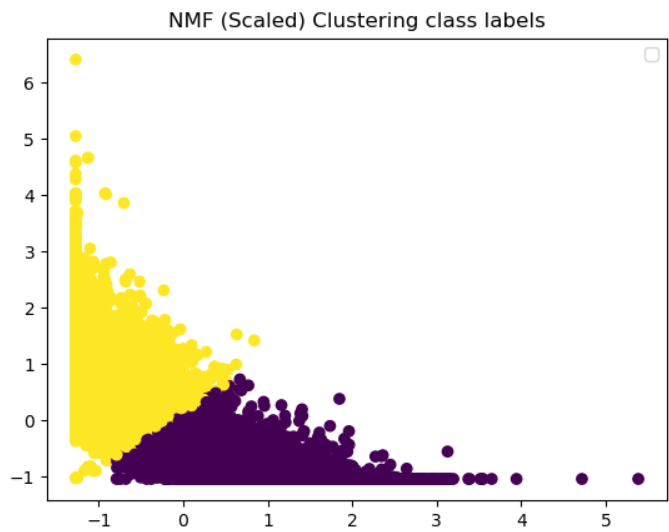


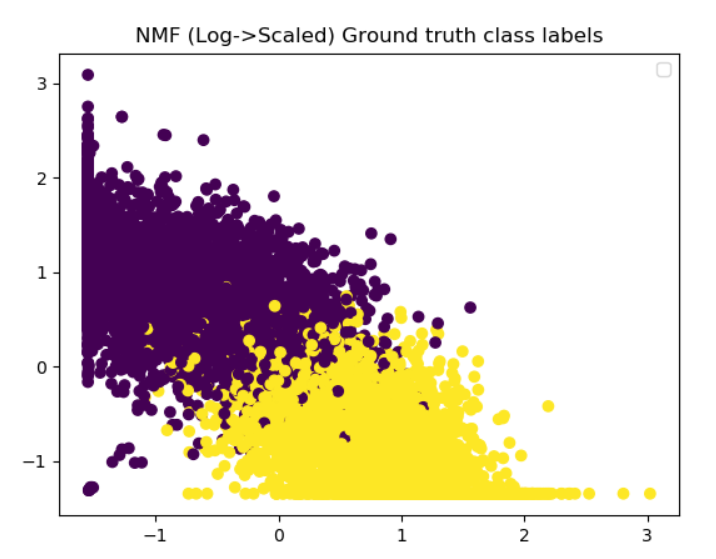


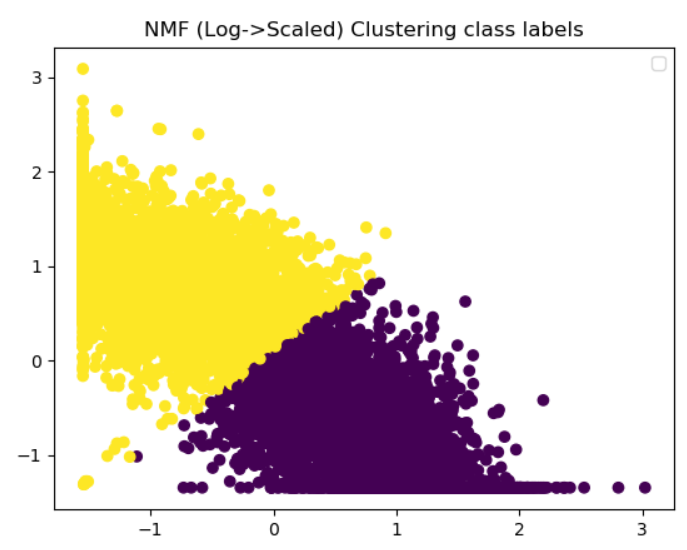


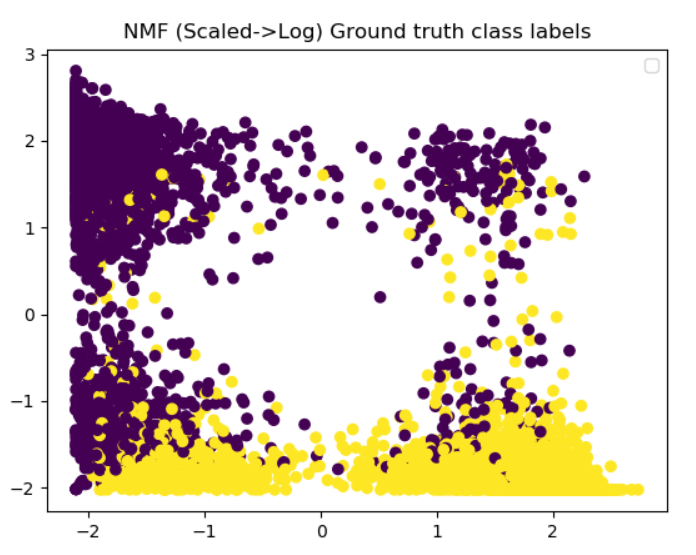


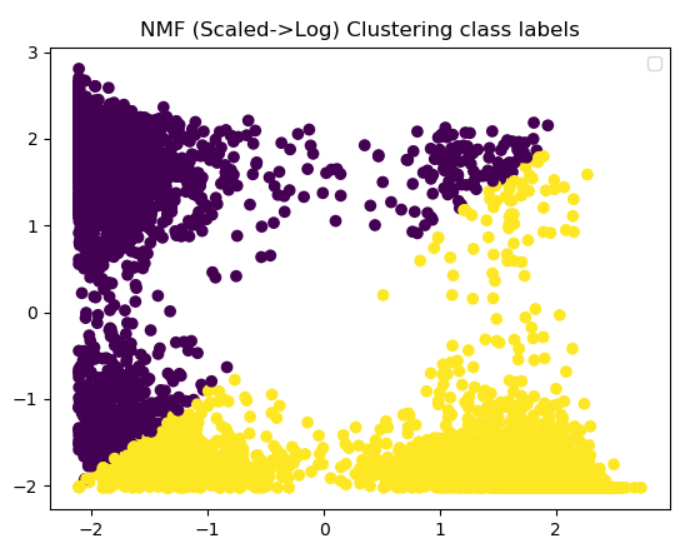








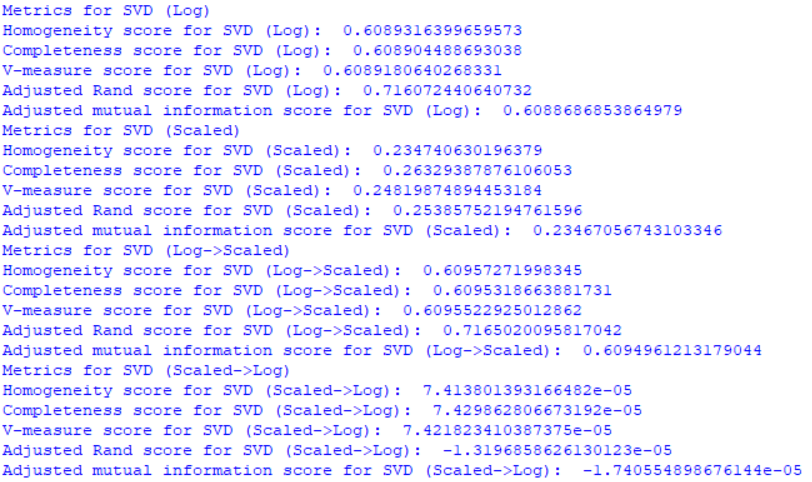


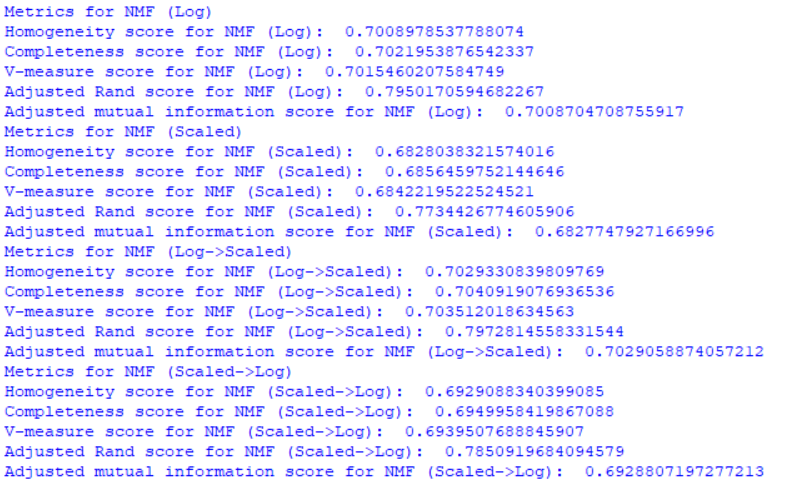


**QUESTION 9:**

The logarithm transformation may improve the clustering results because it increases the distance between the data points. Since the points are packed closely together in the dimension-reduced space, for . This means that the k-means algorithm can find a better, more discriminative decision boundary because the projections are further apart from each other.

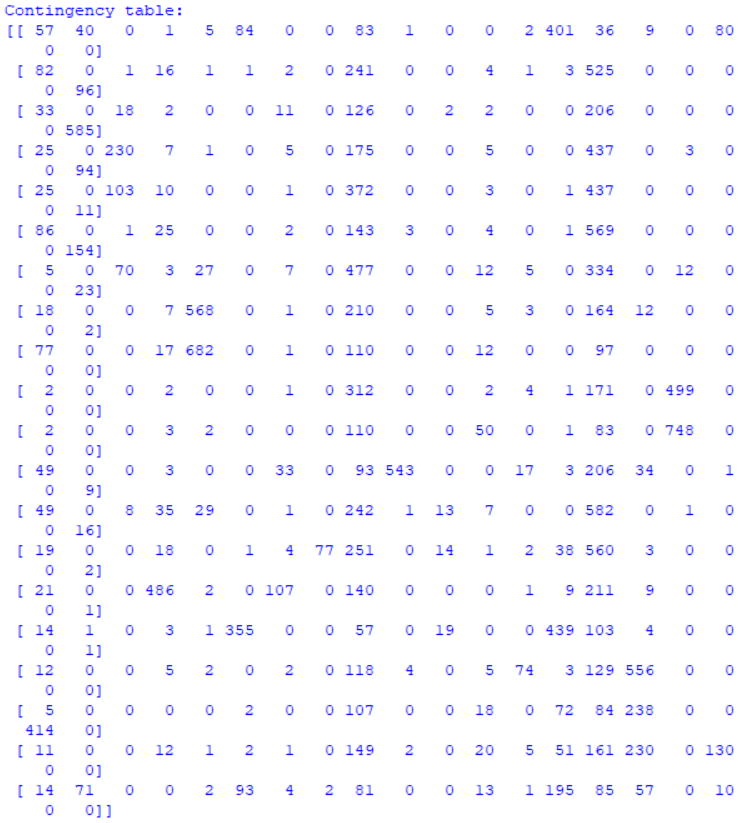
**QUESTION 10:**



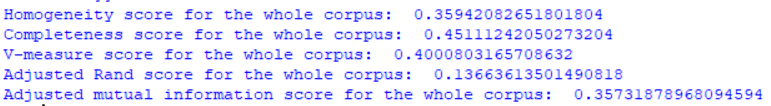


The best combinations were SVD (Log-> Scaled) and NMF (Log->Scaled).

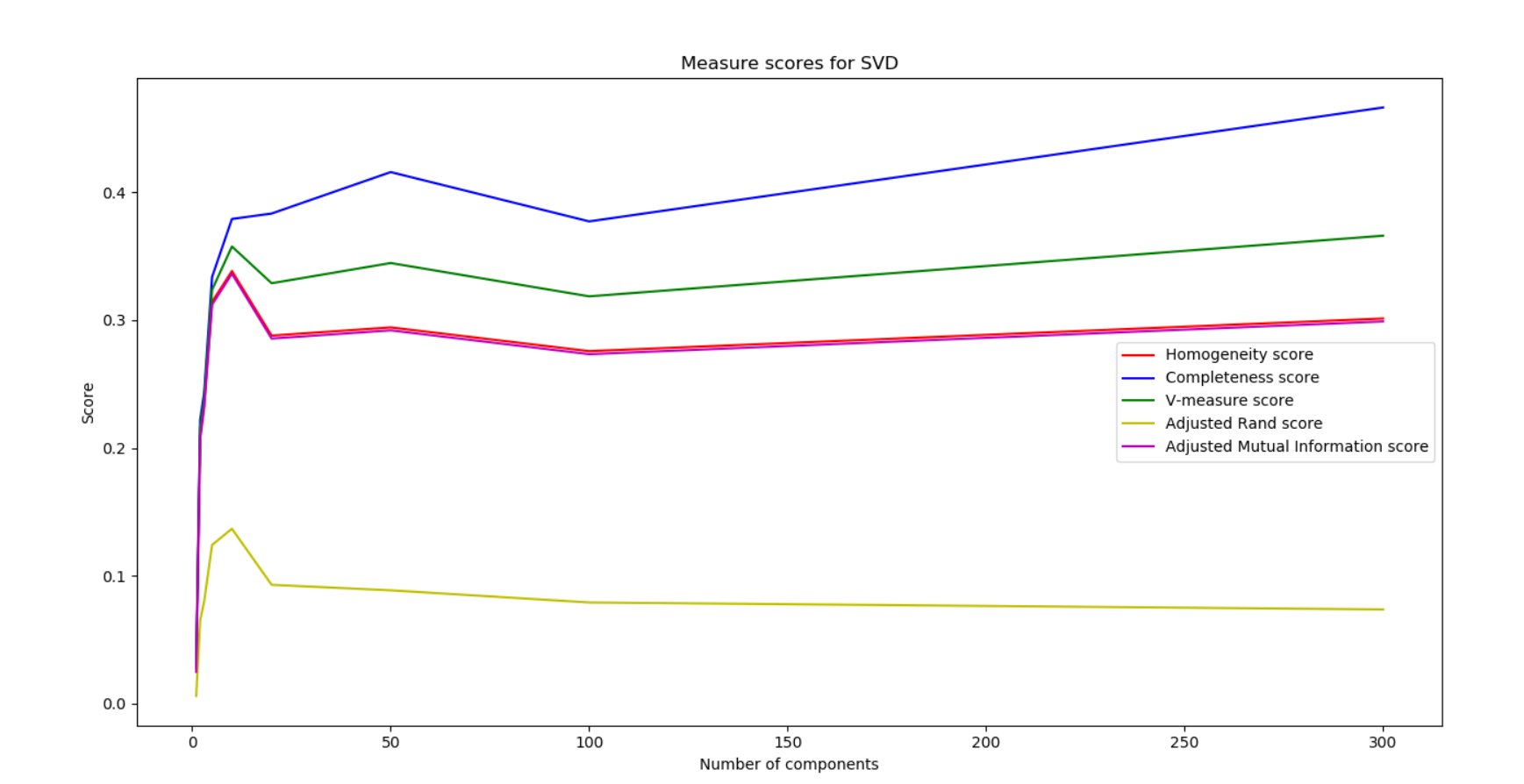
**QUESTION 11:**

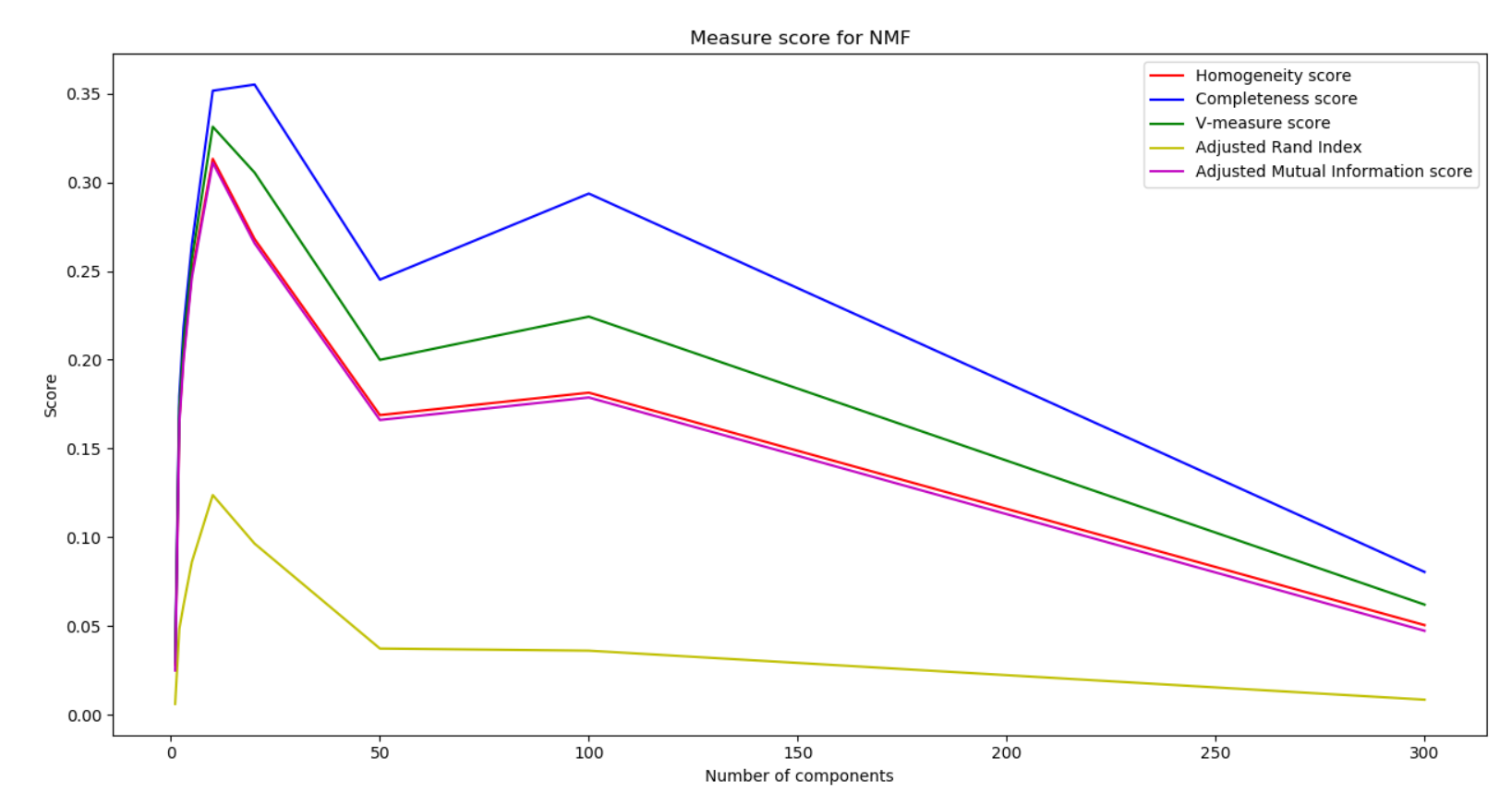


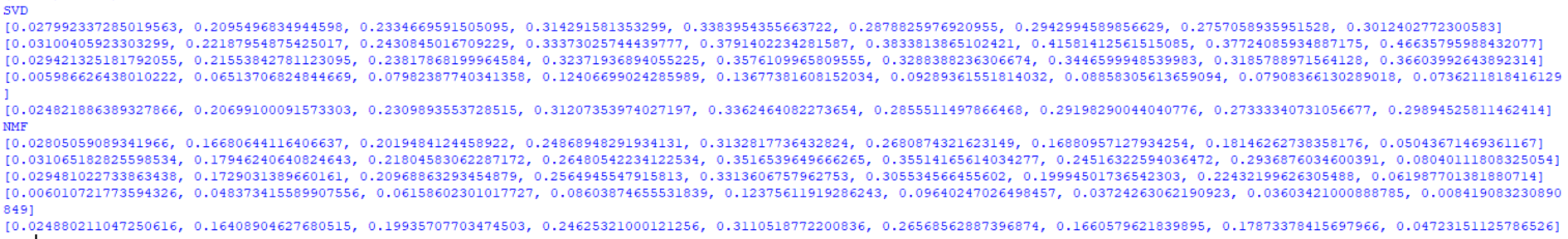
The rows of the contingency table correspond to the true class labels going from 1 to 20. The columns of the contingency table correspond to the predicted cluster labels going from 1 to 20.



**QUESTION 12:**



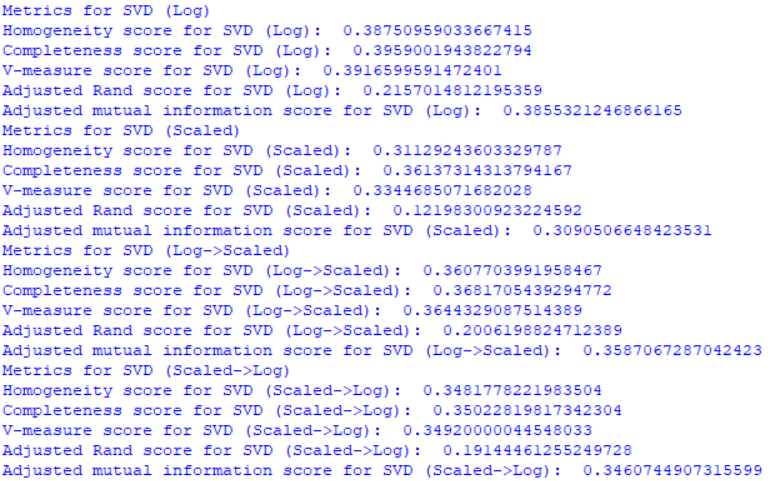




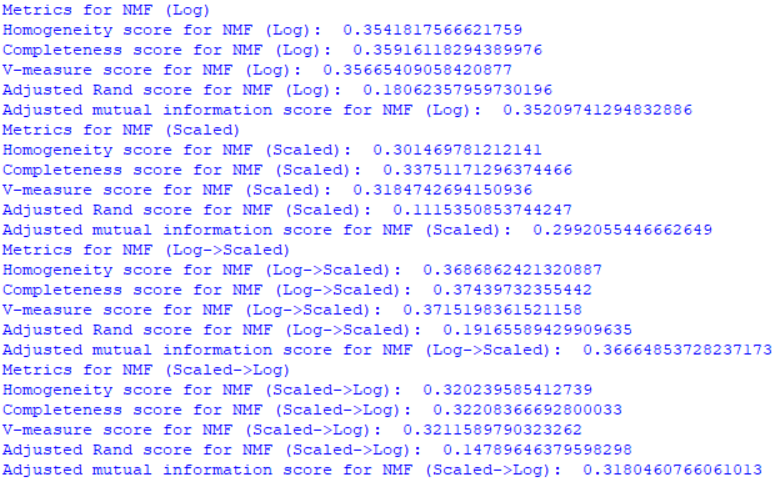
The first row of each dimensionality reduction algorithm corresponds to homogeneity score. The second row corresponds to completeness score. The third row corresponds to V-measure score. The fourth row corresponds to adjusted Rand Index. The fifth row corresponds to adjusted mutual information score.

From the above, it can be seen that the measures reach their highest point when or for SVD and NMF. Thus, I chose these two values and ran k-means using these values.

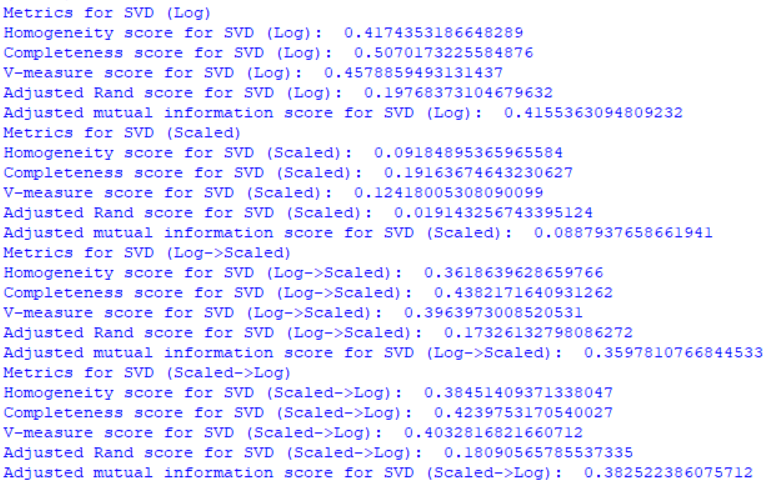
Metrics for when SVD was used with :



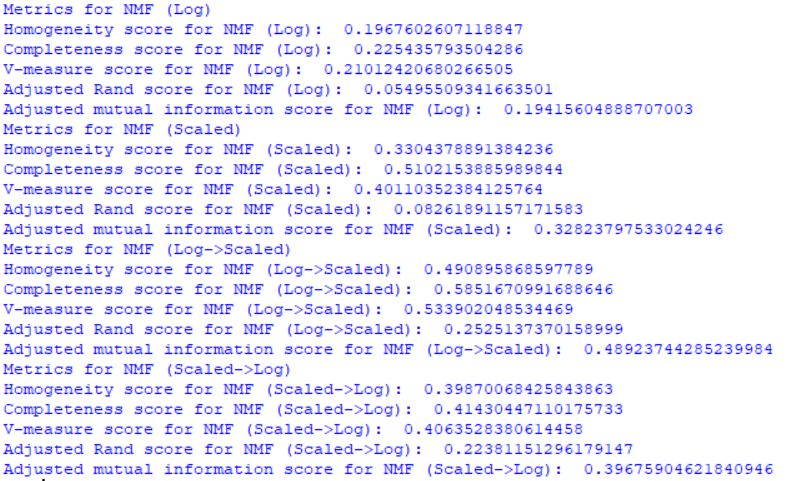
Metrics for when NMF was used with :



Metrics for when SVD was used with :



Metrics for when NMF was used with :



It can be seen that the results for are better than those of for both SVD and NMF.

The best combination is NMF () with logarithm transformation followed by scaling.

The second-best combination is SVD () with logarithm transformation. The best combination was better than the second-best combination by an average of 19.07% in all the measure scores.

To get good clustering results, the following combinations are desirable:

1. NMF with logarithm transformation
2. SVD with logarithm transformation
3. NMF with logarithm transformation followed by scaling
4. SVD with logarithm transformation followed by scaling.

The following combinations are not desirable:

1. SVD with scaling