SMM636 Machine Learning

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Group Coursework 2
IMDB Dataset

Group 8

Wenxu Tian
Linh Nguyen
Fan Xia
Hang Su
Soumya Ogoti

1. PCA

There are four numerical features in the IMDB dataset apart from the 'Year' namely 'Runtime..Minutes.', 'Rating', 'Votes', 'Revenue..Millions.'. While performing EDA, it was seen that 'Revenue..Millions.' column had 3 missing values. Those missing values were dropped before the PCA.

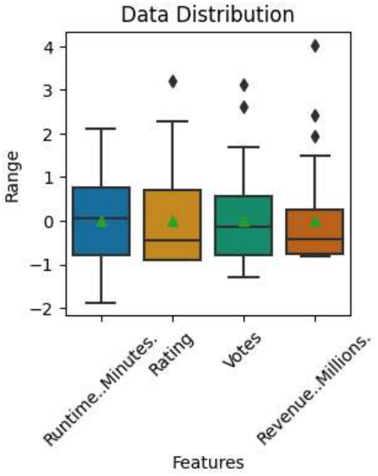


Fig 1. Distribution of standardised features

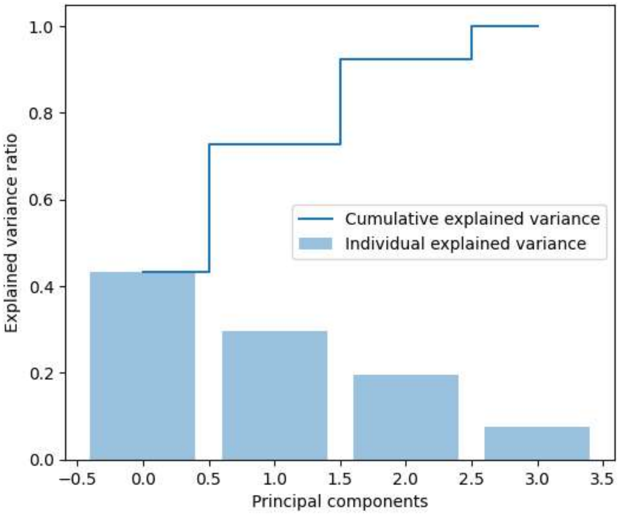


Fig 2. Explained variance of the Principal Components

The four numerical features were standardised by removing the mean and scaling to unit variance as shown in Figure 1. PCA was performed on this data with the number of components equal to that of numerical features initially. The variance explained by each principal component was observed to select those components that explain most of the variance in the data. Figure 2 shows the variance that can be attributed to each principal component as a bar plot and the cumulative explained variance as a step plot. The first three principal components (PCs) cover 92.5% of the variance. Table 1 shows the individual contribution and cumulative explained variance of the four PCs. The first three PCs were chosen to best represent the data by reducing the dimension from four to three for effective analysis.

	PC1	**PC2**	**PC3**	**PC4**
Individual explained variance	0.43	0.296	0.196	0.075
Cumulative explained variance	0.43	0.728	0.925	1

Table 1. Explained Variance of the four PCs

Table 2 shows the contribution of each feature towards the three PCs. The first PC accounts for 43% of the variance and the features that contribute the most towards this can be seen in this table. The absolute values of the weights show that the feature 'Votes' contributes the most, followed by 'Rating', 'Revenue..Millions.', 'Runtime..Minutes' towards PC1.

	Features	**Votes**	**Rating**	**Revenue..Millions.**	**Runtime..Minutes.**
PC1	**Absolute weight**	0.694	0.44	0.496	0.275
PC2		0.116	0.472	0.608	0.627
PC3		0.003	0.674	0.207	0.709

Table 2. Contribution of features towards PC1, PC2, PC3

To show how strongly the features influence the principal components, biplots are drawn with scaled PCA and loading plots. The coloured points in Figure 3 represent the loadings of each original feature which are the corresponding correlation coefficients between the feature and the principal component. The arrows from the origin to each of the coefficients show the directionality of the correlation. The grey dots are the scaled data using the principal components.

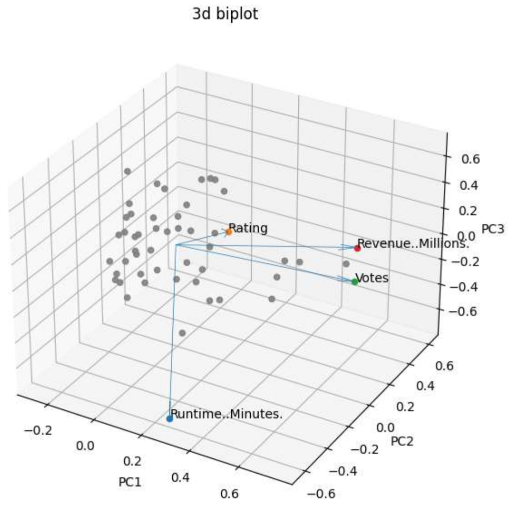


Fig 3. 3D Biplot

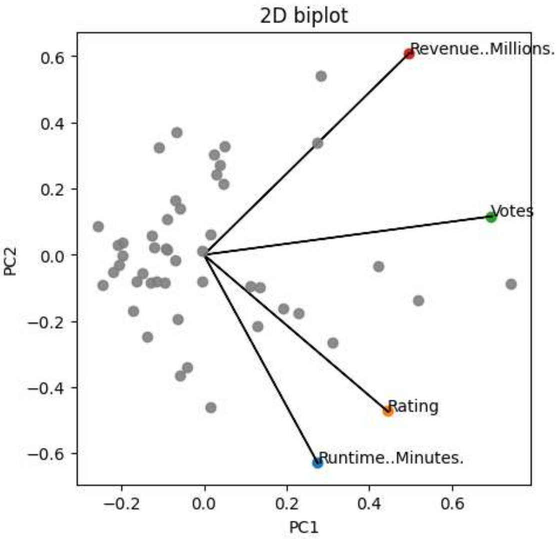


Fig 4. 2D Biplot of just the first two Principal components for easy visualisation.

2. Clustering

2.1. K-means clustering.

The K-means clustering method was selected for this case due to its suitability for unsupervised learning clustering problems. Since the dataset lacks labels, an unsupervised approach is necessary, and K-means is a widely used method for this type of problem. Furthermore, the simplicity of K-means and its ability to handle small datasets without requiring excessive computation time or memory make it a suitable choice. The algorithm provides interpretable results, which are particularly useful for small datasets, where understanding the underlying structure is essential. K-means groups data points into clusters based on similarities, with the centroids representing each cluster's "average" data point. This feature makes interpreting and explaining the clustering results easy, which is particularly valuable in many practical applications.

Implementation and findings:

After applying the Elbow method to the dataset, it was observed from Figure 5 that the Within Cluster Sum of Squares (WCSS) exhibit a linear decrease up to k = 3, after which the rate of decrease becomes marginally slower. While k = 3 could be considered the elbow point and, thus, the optimal number of clusters, this approach may not lead to the most optimal clustering outcomes. When the rate of decrease beyond k = 3 lacks clear differentiation, suboptimal clustering results may ensue. Appendix 1 contains a plot that displays the feature distributions in each cluster when k = 3. It can be inferred that the clusters do not exhibit well-defined boundaries.

As a solution to this challenge, Silhouette analysis can be utilized as an alternative method to determine the optimal number of clusters. Based on the silhouette scores depicted in Figure 6, it is observed that the K-means clustering method generates the highest silhouette score of 0.38 when k = 2, implying that this is the optimal number of clusters for the given dataset. This suggests that the clusters generated by K-means at k = 2 are more well-defined and separated, and the data points within each cluster are closely related to each other while being different from those in other clusters. Therefore, selecting k = 2 as the optimal number of clusters can produce a more accurate and reliable clustering outcome. Appendix 2 provides the detailed silhouette coefficient values of each cluster with different k.

Table 3 presents the feature values of Cluster 0 and Cluster 1 for ‘Runtime (Minutes)’, ‘Rating’, ‘Votes’, and ‘Revenue (Millions)’. Cluster 0 is characterized by negative values for all features, while Cluster 1 exhibits positive values for all features. This disparity in feature values suggests that the two clusters may represent distinct types or genres of movies, with Cluster 1 potentially reflecting more successful or popular movies. The negative feature values of Cluster 0 imply that these movies have shorter runtimes, lower ratings, fewer votes, and generate lower revenue compared to the movies in Cluster 1. The observation that Cluster 1 is associated with positive feature values across all dimensions underscores its potential as a higher-performing cluster.

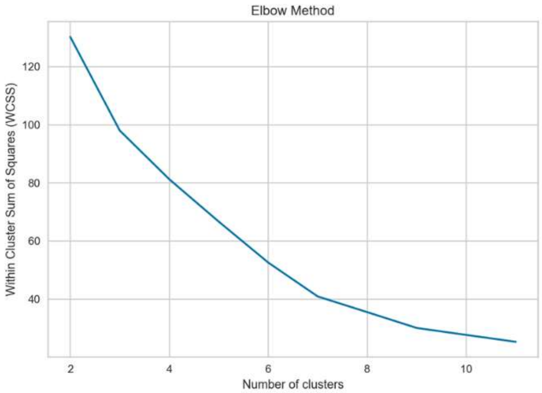


Fig 5. The plot of WCSS for each number of clusters

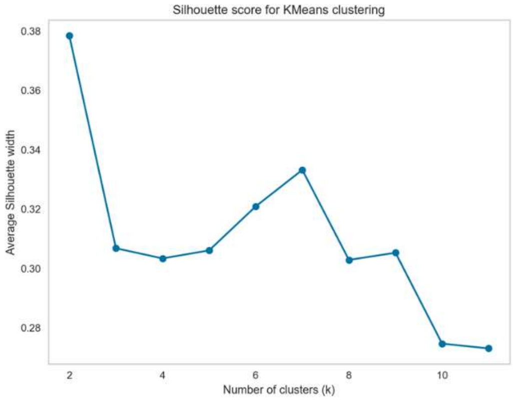


Fig 6. Silhouette score for K-means clustering.

Feature	**Cluster**	
	0	**1**
Runtime Minutes	-0.27	0.99
Rating	-0.19	0.69
Votes	-0.39	1.43
Revenue Millions	-0.28	1.02

Table 3. Feature Mean (scaled) when k=2

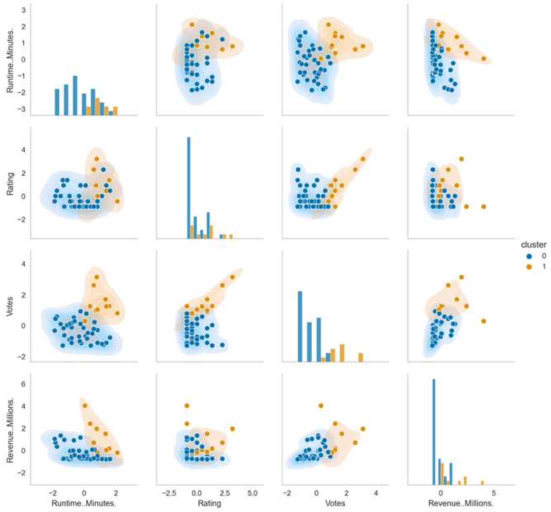


*Fig 7. Plots of feature distribution by clusters (*k = 2)

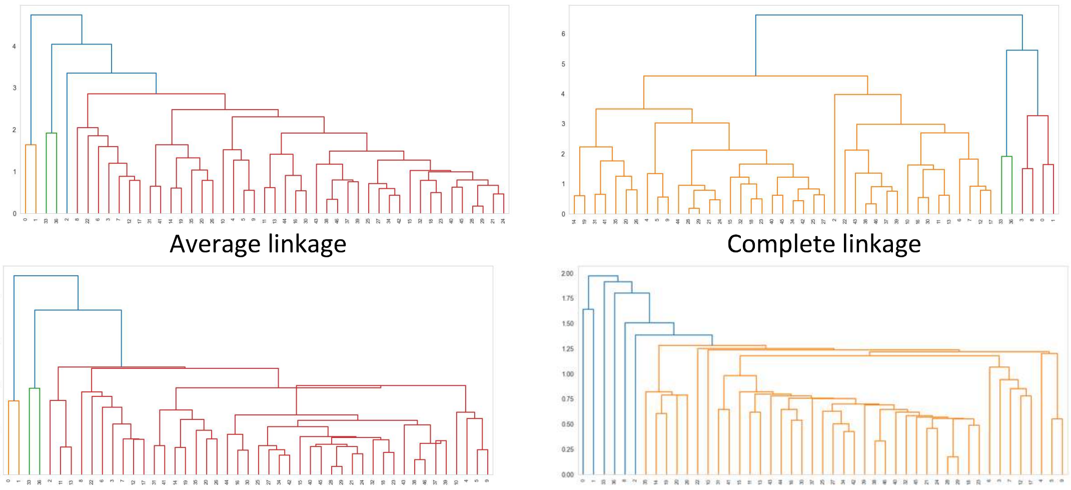
In Figure 7, the feature distributions of the two clusters are depicted, providing further insight into their characteristics. Notably, there are discernible relationships between certain features, such as the positive association between Votes and Revenue (Millions). Such findings are congruent with the idea that a higher number of votes means greater audience exposure and attention, ultimately translating to increased revenue for a movie. Additionally, it was seen that the number of observations in cluster 0 is higher than in cluster 1. The detailed numbers of each cluster's scaled mean for each feature can be found in Table 3.

2.2. Hierarchical clustering.

Another popular clustering approach for unsupervised learning clustering problems is hierarchical clustering. Hierarchical clustering can help to group similar data points into one cluster. The advantage of this approach is the hierarchical structure of the clusters can be visualized using dendrograms to identify natural clusters within the dataset. Another advantage of this approach is that hierarchical clustering can be used with different linkage methods like complete, single, average, centroid etc., allowing for customization to suit the specific needs of the classification. A difference between this clustering and K-means clustering is that the number of clusters does not need to be pre-defined in hierarchical clustering.

Implementation and findings:

Figure 8 shows the four different linkages in hierarchical clustering. The average linkage calculates the average distance between all pairs of points in the two clusters. The complete linkage calculates the distance between clusters as the maximum distance between any two points in the two clusters. The centroid linkage calculates the distance between clusters in hierarchical clustering. The single linkage is used to calculate the distance between clusters as the minimum distance between any two points in the two clusters. Based on the result of the hierarchical clustering with four different linkages, the complete linkage could be a good choice as its cluster sizes are more uniform compared to the other three linkages. After determining the linkages, the number of clusters would be chosen by identifying the level at which the dendrogram shows a significant jump in the distance between clusters. Finally, the 3 clusters could be a good choice.



Average linkage

Complete linkage

Centroid linkage

Single linkage

Fig 8. Different linkages in hierarchical clustering

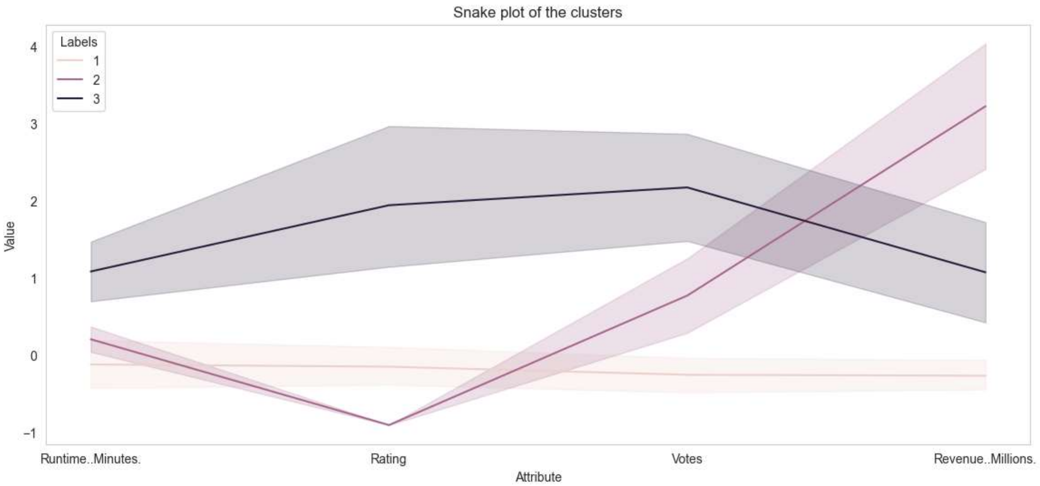


Fig 9. Snake plot of the clusters

Cluster	Runtime..Minutes.	Rating	Votes	Revenue..Millions.
1	-0.12	-0.15	-0.25	-0.26
2	0.21	-0.90	0.77	3.22
3	1.09	1.94	2.17	1.08

Table 4. Average scores for features in each cluster (3 clusters chosen, complete linkage)

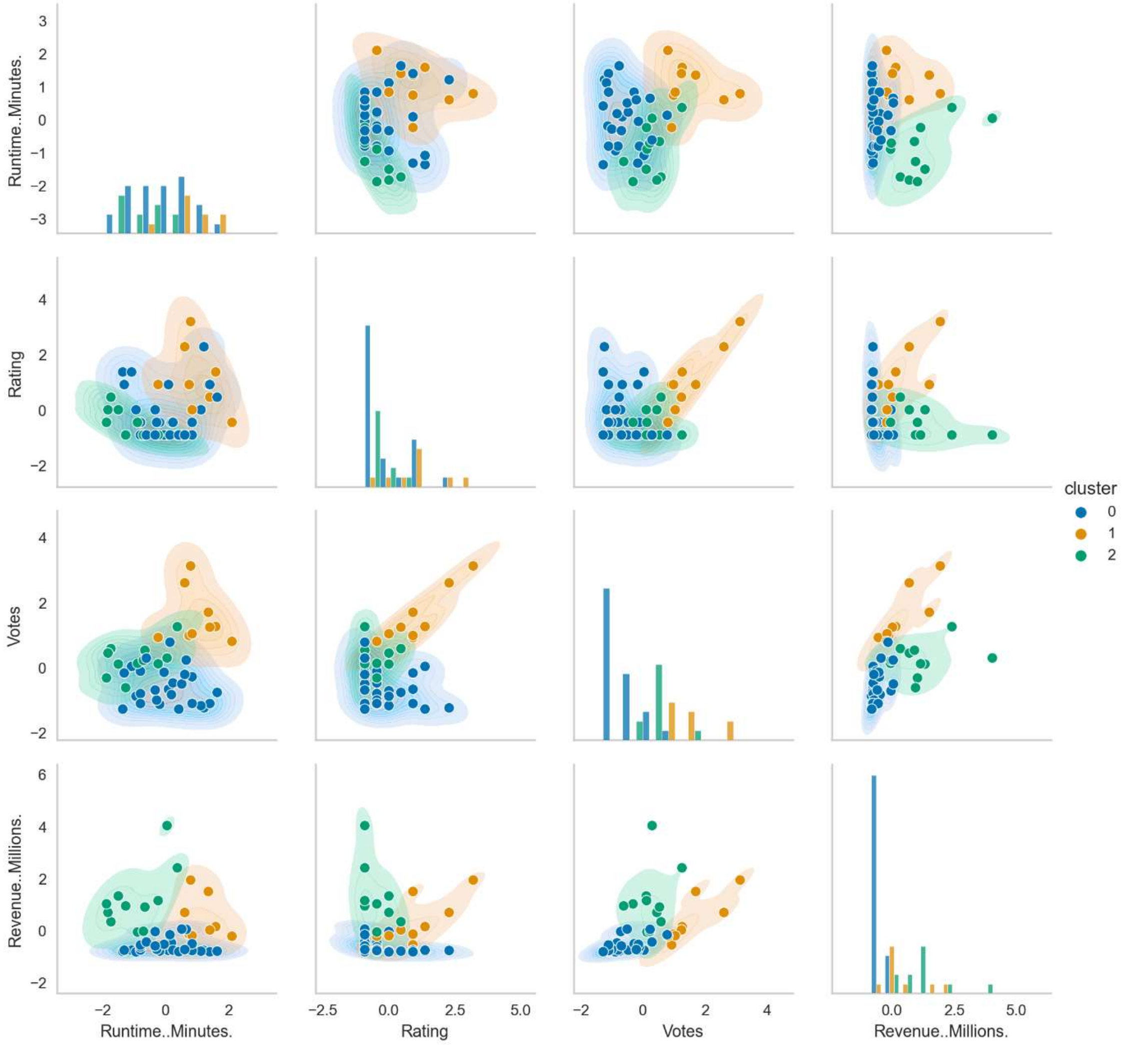
Figure 9 depicts the mean range of each feature within every cluster. The detailed scaled mean data can be found in Table 4. Cluster 1 is associated with the lowest scores across the three features of runtime, votes, and revenue, with all scores lying below the mean. This suggests that the movies in this cluster were the shortest in duration, received the least attention and generated the least revenue compared to the other two clusters. This cluster also was perceived to be of lower quality than cluster 3 with a lower rating.

Cluster 2 exhibits slightly above-average running times and votes, a below-average rating that is the lowest score among the three clusters. Surprisingly, this cluster generated the highest revenue. Meanwhile, cluster 3 has the longest run times, the highest rating, the highest number of votes and an above-average revenue but still not as high as Cluster 2. This indicates that cluster 3 includes movies which have the longest runtimes, the greatest attention, and the best reputation. The observation that both clusters 2 and 3 have higher votes and higher revenue can be explained by the positive correlation between the 'Votes' feature and the revenue generation of movies (as evidenced in the correlation matrix of all features in Appendix 3). Furthermore, the hierarchical clustering analysis highlights that the number of observations within Cluster 1 is significantly higher than the number in Clusters 2 and 3. The plot of feature distribution by clusters when the number of clusters was chosen as 3 as a result of hierarchical clustering can be found in Appendix 4.

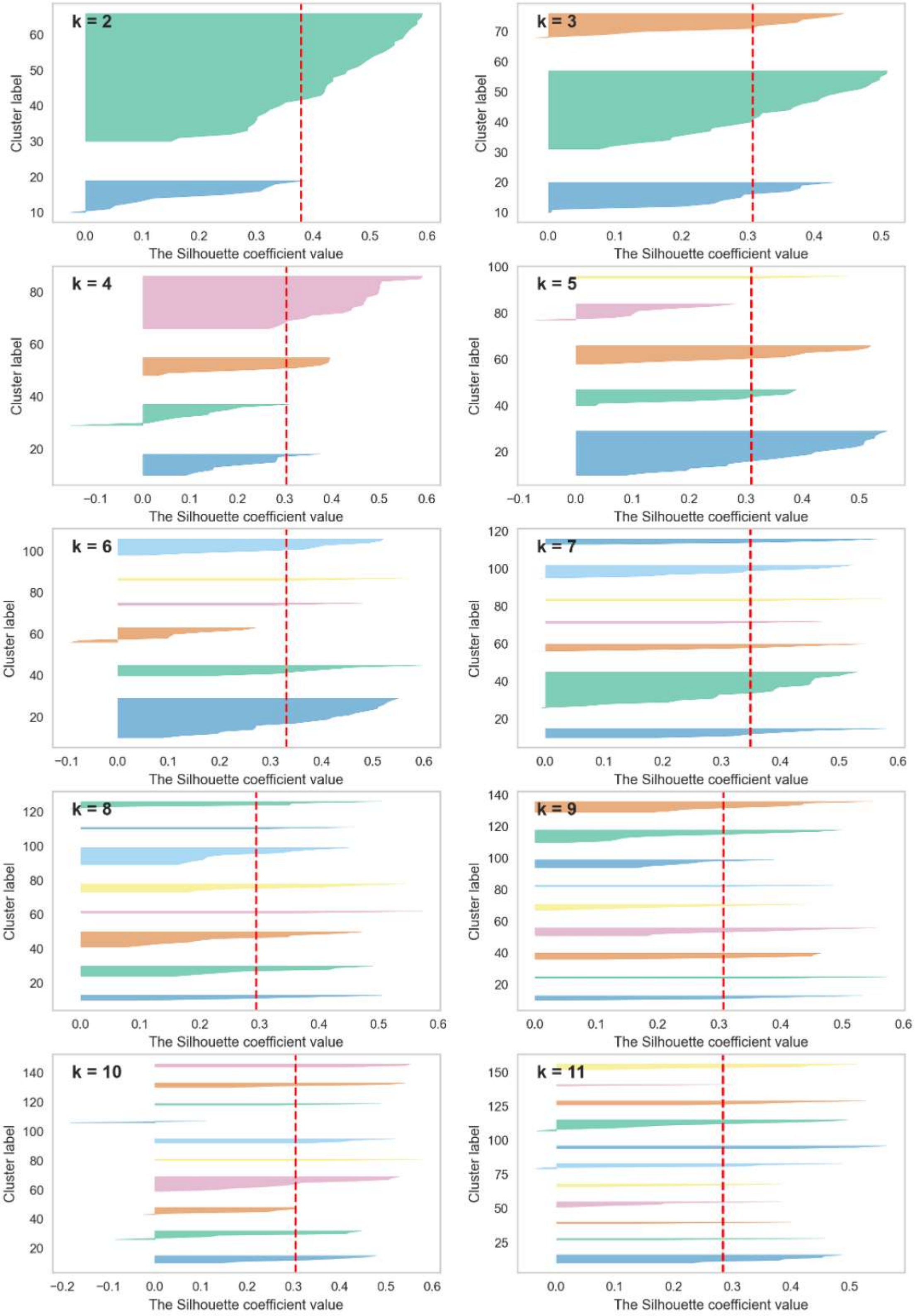
Overall, these findings provide valuable insights into the unique features of movies within each cluster, thereby allowing film production companies and marketers to make informed decisions on tailoring their approaches based on audience preferences and prevailing market trends. However, the selection between the clustering outcomes of K-means or hierarchical clustering analysis depends on users' interpretability and strategic preferences. Moreover, the integration of more extensive datasets and additional features into the analysis could potentially offer further insights and more robust clustering results. Such developments would significantly enhance the practical value of these findings for stakeholders in the film industry.

APPENDICES

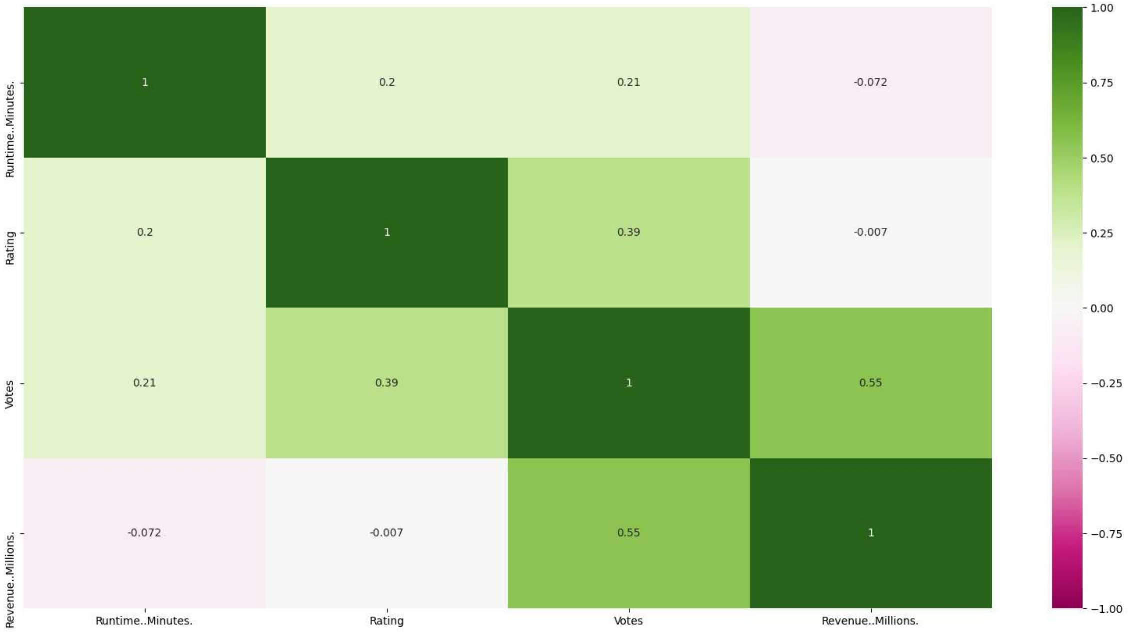
Appendix 1. Plots of feature distributions by clusters when k = 3 (K-means clustering)



Appendix 2. Silhouette plot for K-means Clustering.

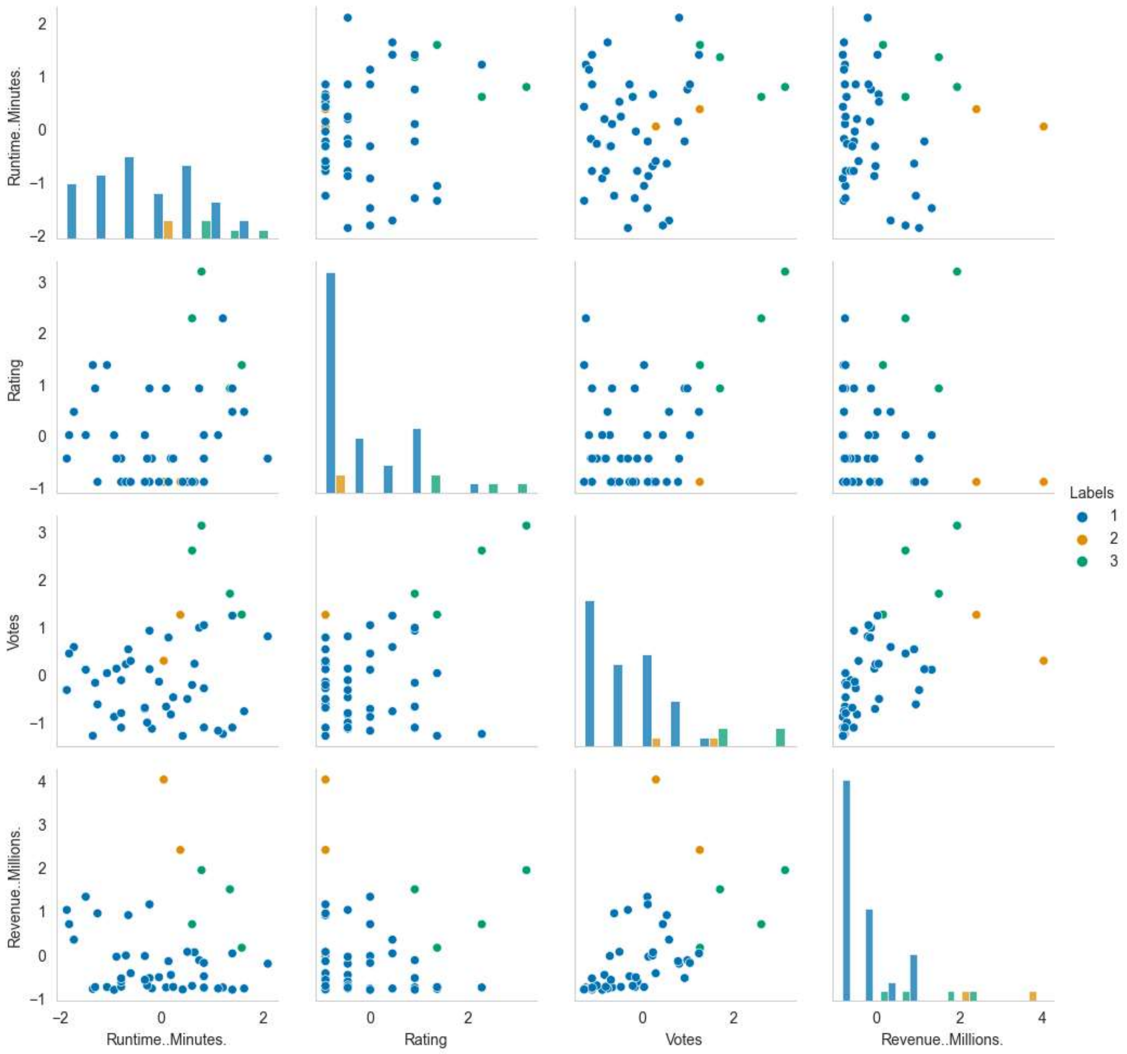


Appendix 3. Correlation Matrix of the 4 features



Appendix 4. Plots of feature distributions by clusters when there are 3 clusters.

(Hierarchical clustering)


Runtime..Minutes.
Rating
Votes
Revenue..Millions.
Labels