Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name)	4. Saha, Soumyajit 5. Parashar, Tridev 6. 7.
1. Dubey, Naveen 2. Shetty, Rakshith 3. Semaleesan, Sneha	**GROUP NUMBER:** **9**
MSc in: Business Analytics	
Module Code: SMM 636	
Module Title: Machine Learning	
Lecturer: Dr. Rui Zhu	**Submission Date: 05 April 2023**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

Executive Summary

This study aims to utilize unsupervised learning techniques to identify key numerical attributes and similarities among the top 50 rated movies that contribute to their higher ratings on IMDB. Principal Component Analysis (PCA) and clustering methods have been employed to achieve this goal, and the findings have managerial implications that are discussed below.

Principal Component Analysis

Use of PCA in the dataset –

The dataset underwent dimensionality reduction using a technique such as Principal Component Analysis (PCA) to extract the crucial numerical factors that contribute to the success of movies, while eliminating extraneous information that does not aid in our findings. **NB – There are 3 missing values in the revenue columns. Therefore, the corresponding movie data for the same have been removed.**

Results and Interpretation –

Importance of Features	PC1	PC2	PC3	PC4
Standard deviation	1.314	1.0884	0.8864	0.5494
Proportion of Variance	0.432	0.2962	0.1964	0.07546
Cumulative Proportion	*0.432*	*0.7281*	*0.9245*	*1*

Table 1 – Illustrates the outputs of the PCA along with a view on the proportion of variance explained by each PC

In Table 1, the PCA outputs reveals that the first principal component (PC1) exhibits a standard deviation of 1.314, accounting for 43.2% of the total variance in the data. Similarly, the second principal component (PC2) demonstrates a standard deviation of 1.0884 and explains 29.62% of the total variance. The third and fourth principal components (PC3 and PC4) exhibit standard deviations of 0.8864 and 0.54940, respectively, and explain 19.64% and 7.54% of the total variance, respectively.

Further, drawing insights from the cumulative PVE plot (Figure 1), it can be deduced that a minimal set of 3 components is adequate to account for more than 90% of the variance in the movie dataset. This implies that the said 3 components effectively encapsulate a considerable proportion of the fundamental information present in the dataset.

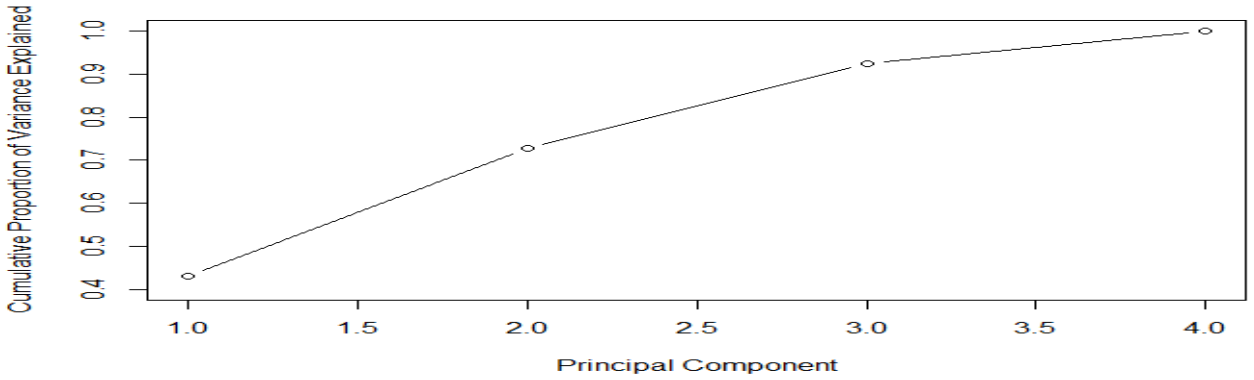

Figure 1 – Cumulative proportion of variance of principal components

Features	PC1	PC2	PC3	PC4
Runtime Minutes	0.274519	0.627404	0.708938	-0.16855
Rating	0.444094	0.472413	-0.67415	-0.35374
Votes	0.693758	-0.11582	0.002863	0.710829
Revenue Millions	0.496105	-0.60809	0.207177	-0.58411

Table 2 – Illustrates the loadings/coefficients of the principal components

Table 2 presents the coefficients or loadings assigned to each variable on each principal component, which offer valuable information on the degree of influence that variable has on the component. Additionally, the feature loadings' significance in identifying each principal component can be articulated through this representation.

$$\begin{aligned} \boldsymbol{PC1} &= \boldsymbol{0.27 * Runtime + 0.44 * Rating + 0.69 * Votes + 0.50 * Revenue} \\ \boldsymbol{PC2} &= \boldsymbol{0.63 * Runtime + 0.47 * Rating - 0.11 * Votes - 0.61 * Revenue} \\ \boldsymbol{PC3} &= \boldsymbol{0.71 * Runtime - 0.67 * Rating + 0.002 * Votes + 0.21 * Revenue} \\ \boldsymbol{PC4} &= \boldsymbol{-0.17 * Runtime - 0.35 * Rating + 0.72 * Votes - 0.58 * Revenue} \end{aligned}$$

As observed, the loadings can either contribute in a positive or negatively correlated manner towards the value of its respective principal component. E.g. – PC1 has a positive correlation with all the loadings mentioned, which infers the fact that a higher runtime/rating/votes/revenue will lead to a higher value on this principal component.

Managerial Implications –

The outputs of the principal component analysis can aid in identifying the significant drivers that impact the success of the movies, as evidenced by the principal components explaining most of the variance. Moreover, by examining the biplot (for PC1 and PC2) depicted in the appendix (moved to appendix to accommodate its size), valuable insights can be derived for developing additional marketing and promotional strategies for these movies, as follows –

- Movie titles 1, 2, and 9 exhibit significant positive scores on the first principal component, with the number of votes received by these movies being the primary driving factor. This is because the first principal component places the greatest emphasis on the number of votes received by the movies.
- Conversely, movie titles 3, 12, and 14 exhibit substantial positive scores on the second principal component, with the runtime minutes of these movies being the primary driving factor. This is because the second principal component places the greatest emphasis on the runtime minutes of the movies.

These insights, among others, could have pertinent implications for formulating marketing and promotional strategies for these movies.

Clustering

Use of clustering in the dataset

The dataset was subjected to clustering analysis in order to distinguish similarities and differences among its observations and attributes. The primary goal of this machine learning technique is to group attributes into distinct categories, where each group represents similar attributes for the observations within it and is distinct from other groups. Two of the most employed clustering algorithms, namely Hierarchical and K-means clustering, were utilized in this study. In evaluating the quality of clustering and determining the ideal number of clusters, we utilized the Silhouette score as a metric. This score measures the similarity of each data point to its own cluster compared to other clusters. A high Silhouette score indicates that the data point is well-

matched to its cluster, while a low score suggests that it might belong to a different cluster. Further, a positive sign signifies correct clustering and a negative sign signifies incorrect clustering.

Results and Interpretation –

- Hierarchical Clustering –

  It is an unsupervised technique that groups similar objects into nested clusters based on a tree-like structure called a dendrogram. It uses a pairwise dissimilarity measure and a linkage function to iteratively merge the two most similar clusters until all objects belong to one large cluster or a pre-specified number of clusters. This method is flexible and widely used to reveal the underlying structure of complex datasets and provide insights into the relationships between observations.

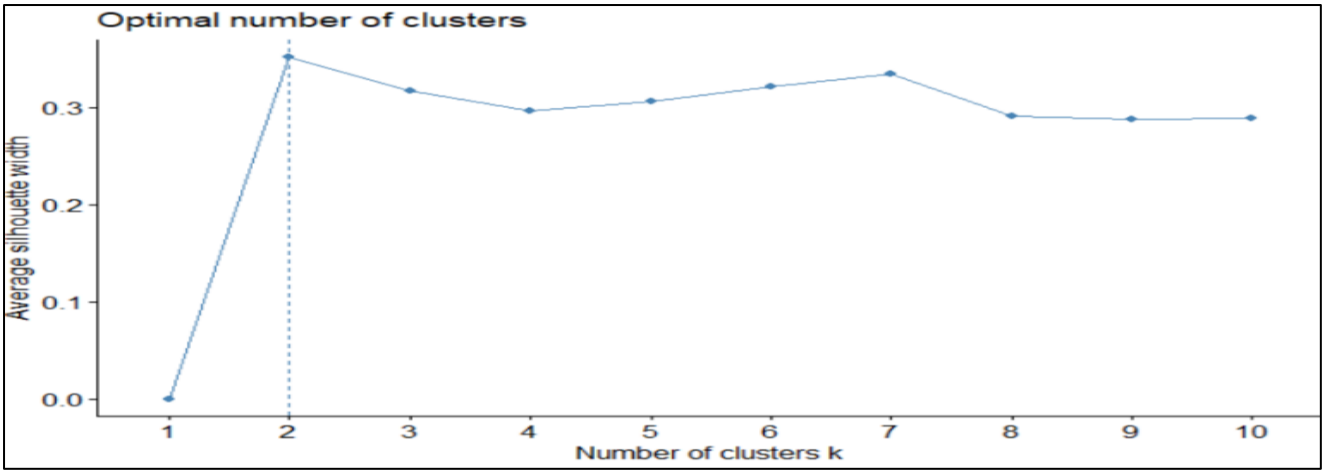

Figure 2 – Determining the optimal number of clusters in case of hierarchical clustering

Figure 2's results indicate that the optimal number of clusters for the dataset is 2 as per the hierarchical clustering analysis. However, selecting the clusters is not a simple process due to various reasons.

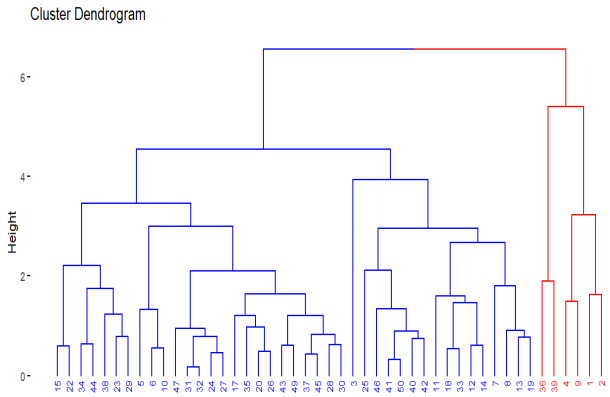

Figure 3 – Dendrogram with two clusters

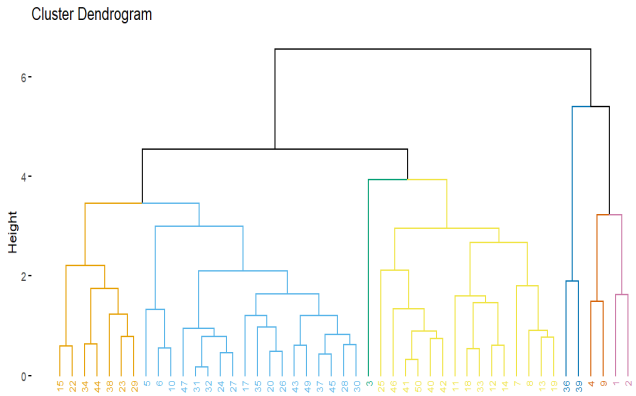

Figure 4 – Dendrogram with seven clusters

1. According to Figure 2, the maximum average Silhouette width is achieved with 2 clusters. However, creating only 2 clusters may result in multiple nested sub-clusters. To avoid this, we can consider the second-highest average Silhouette width, which occurs with 7 clusters. This approach allows for a more in-depth analysis of similarities and dissimilarities between the movies.
2. The dissimilarity between observations is reflected in the vertical height difference on the dendrogram. While most observations fuse before a height value of 2, some fuse after and form subclusters that fuse at even higher heights, as evident from the dendrogram in Figure 3. Additionally, there are multiple distinct

subclusters that can be identified through visual inspection which corroborates the idea that more than 2 clusters may exist.

Therefore, we examined the dendrogram for 7 clusters, shown in Figure 4. By observing the leaf for movie 3 (green coloured) in Figure 4, it is apparent that it fused with other observations at a very high height value near 4, indicating significant dissimilarity. This provides further support for assigning it to a separate cluster, thus justifying the choice of 7 clusters.

- K-means Clustering –

K-means clustering involves dividing a dataset into K distinct and non-overlapping clusters by assigning each observation to one of the K clusters based on their distance from the centroid of each cluster. This process is iterative and continues until no further improvement is possible. The steps for performing K-means clustering include choosing a value for K, randomly assigning each observation to an initial cluster, and repeating the process until the cluster assignments stop changing by computing the centroid of each cluster and assigning each observation to the closest centroid based on Euclidean distance to minimize the sum of squared distances.

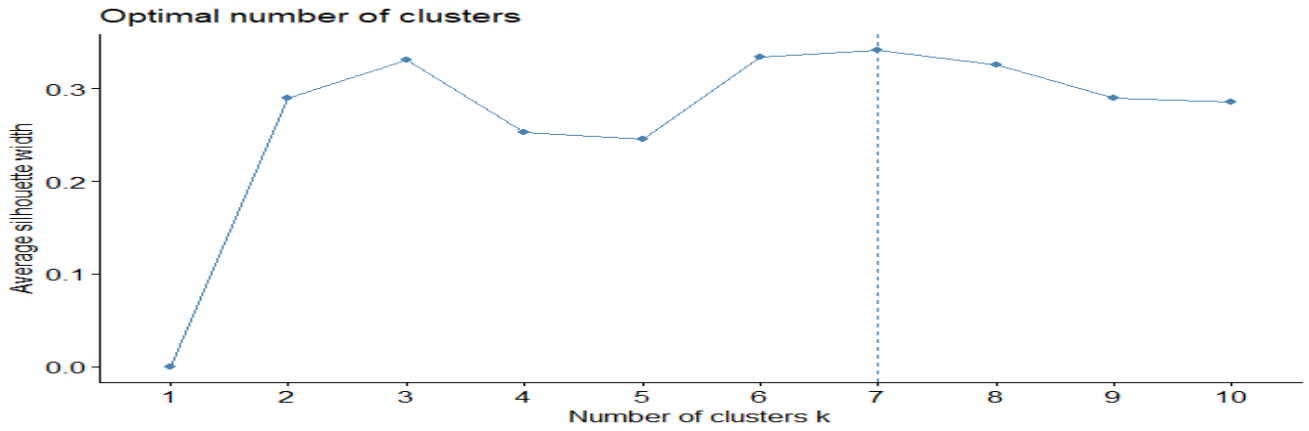

Figure 5 – Determining the optimal number of clusters in case of K-means clustering

According to the results displayed in Figure 5, the k-means clustering analysis suggests that the ideal number of clusters for the given dataset is 7. Here, we keep the optimal choice at 7 as the previous local maximum was at 3 but its silhouette value was lesser than 7 and after 7, the graph was monotonically decreasing and there were no local maximums again.

Now, since we have the ideal number of clusters derived from the methods of clustering, we could do quick check based on the average silhouette width (also known as the silhouette score) to determine the better fit clustering method for the dataset in consideration.

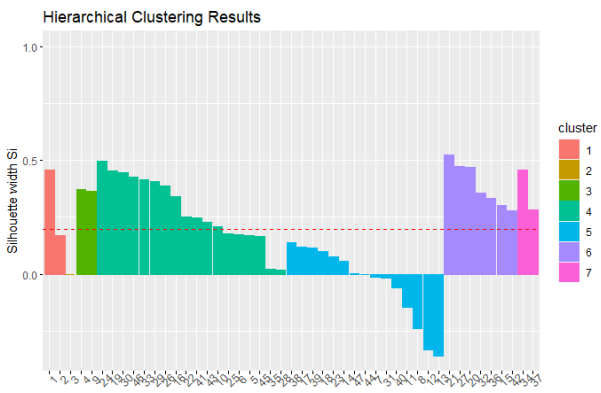

Figure 6 – Silhouette widths of hierarchical

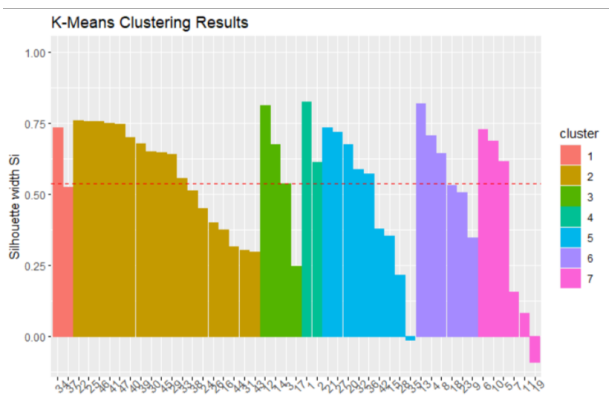

Figure 7 – Silhouette widths of K-means clustering

NB – In both figures 6 and 7, the average silhouette score can be determined from the red dotted line indicated in both the graphs.

Based on the observations from Figures 6 and 7, K-means clustering appears to be a superior method as it exhibits a higher silhouette score compared to hierarchical clustering, and the number of mis-clustered observations (with a silhouette score lower than 0) is lower for K-means, with 2 observations as compared to 7 observations in hierarchical clustering.

Managerial Implications –

Clustering movies based on their runtime, rating, votes, and revenue ($millions) can have significant managerial implications for the movie industry.

It can help in marketing and promotion by targeting specific segments of the audience who are more likely to watch and appreciate a particular type of movie. A marketing manager can use multiple combination of the numerical variables, which may have been used to devise the clusters, to devise prospective marketing strategies. Some possible combinations which may have been used to formulate the clusters are (purely from a business perspective and observed in the K-means clusters) –

1. Ratings higher than 8 but below 8.8 can be used to form a cluster (as seen in movie titles 12, 14, 3 and 17 in the k-means clusters)
2. Ratings higher than 8.8 can be used to form another cluster (as seen movie titles 1 and 2 in the k-means clusters)

Like the combinations shown above, there can be other useful combinations such as ratings and revenue, votes and revenue etc.

Overall, clustering can facilitate more informed decisions about movie production, marketing, and distribution, leading to potentially higher revenues and better audience satisfaction.

Appendix

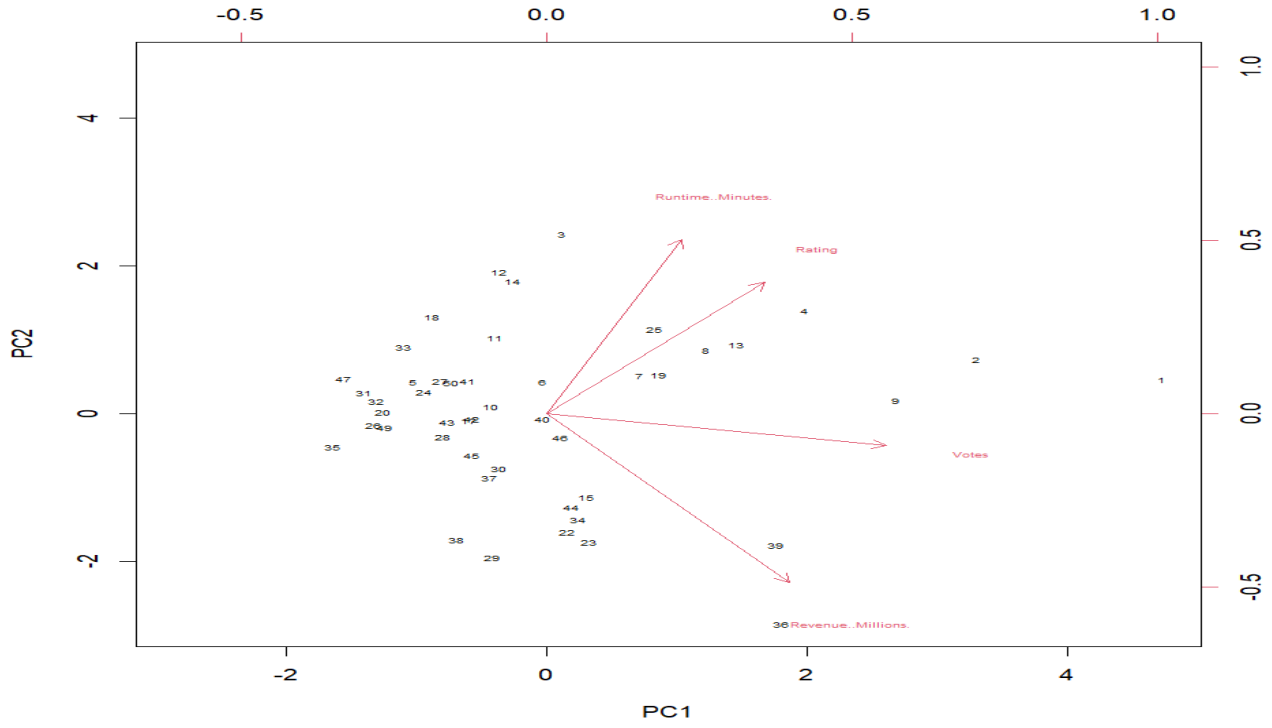

Figure – Biplot for Principal components 1 and 2