Introduction

This report explores unsupervised learning techniques applied to a dataset of 50 top-rated IMDb movies. The objective is to extract clustering insights using PCA and clustering methods. Additionally, the report will compare the manual analysis results with ChatGPT-4o results. We have conducted both data preprocessing and EDA prior to implementing the PCA and clustering tasks. Relevant observations are documented in the code notebook.

1. Principal Component Analysis

PCA is used to reduce dimensionality and uncover the most influential features contributing to the variability in the dataset.

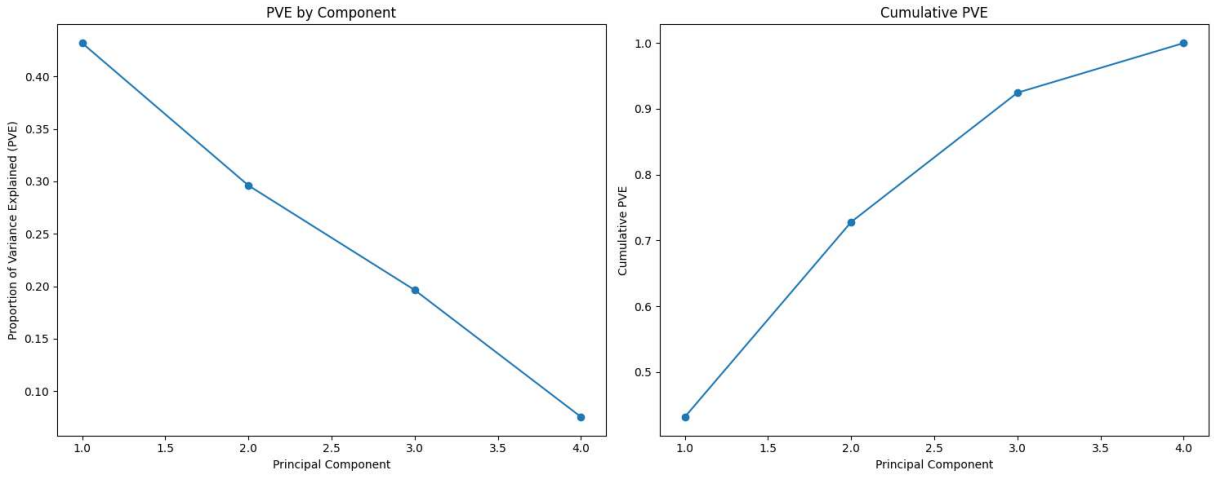


Figure 1

The elbow method was conducted to identify the optimal number of PCs by locating the point at which adding more components yields diminishing returns in explained variance. The elbow appears between PC2 and PC3, where the rate of increase in explained variance slows down. 3 PCs is a suitable choice, capturing over 92% of the total variance main drivers are Votes (0.69) and Revenue (0.50), it is likely to represent a "popularity" or "success" dimension. Movies with more votes and higher revenue score high on PC1, which is reflected in the 3D plot, where the PC1 axis extends in the direction of those gray points associated with highly voted and profitable movies.

	PC1	PC2	PC3
Runtime.. Minutes.	0.274519	0.627404	0.708938
Rating	0.444094	0.472413	-0.674149
Votes	0.693758	-0.115823	0.002863
Revenue.. Millions.	0.496105	-0.608091	0.207177

PC2 main drivers are Runtime (0.63) and Rating (0.47), while Revenue contributes

negatively (-0.61) shown in the opposite direction. PC2 might reflect a "quality vs commercial performance" dimension. Longer, higher-rated movies that are *not necessarily box office hits* may score higher here. In PC3, Runtime (0.71) contributes positively, while Rating (-0.67) contributes negatively, capturing a contrast between long but lower-rated films, or vice versa.

2. Clustering

Clustering is an unsupervised learning technique that identifies, and groups data points based on certain variables.

2.1 K- Means Clustering

This method begins by choosing an estimated number of clusters in the dataset, randomly placing centroids, and computing cluster means to converge to the true cluster continuous mean, thus minimizing within-cluster variance.

To determine the best number of clusters (K) for a given dataset, inertia which measures how compact a data point cluster is around the centroid, is used.

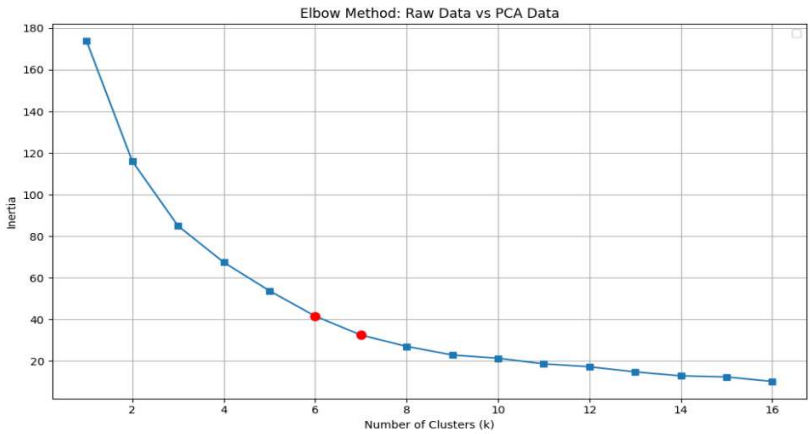


Figure 2

It is important to note that lower inertia improves clustering; however, the lowest inertia would be when clusters equal observations. Therefore, as seen in Figure 2, the elbow point is clusters 6-7, after which the drop in inertia becomes marginal.

Silhouette scores measure the clear separability of clusters, indicating that cluster 6 had clearer cluster separation. The average silhouette scores for 6 and 7 clusters were 0.376 and 0.371, respectively, due to the misclassification in 7 clusters.

2.2 Hierarchical Clustering

Hierarchical, as the name suggests, builds a tree-like structure using linkage methods between points. Complete linkage merges clusters with the farthest pair of points, whereas average linkage merges clusters based on the average distance between all points; these yielded the highest silhouette scores. Unlike the K-Means, there is no need to specify a number of clusters.

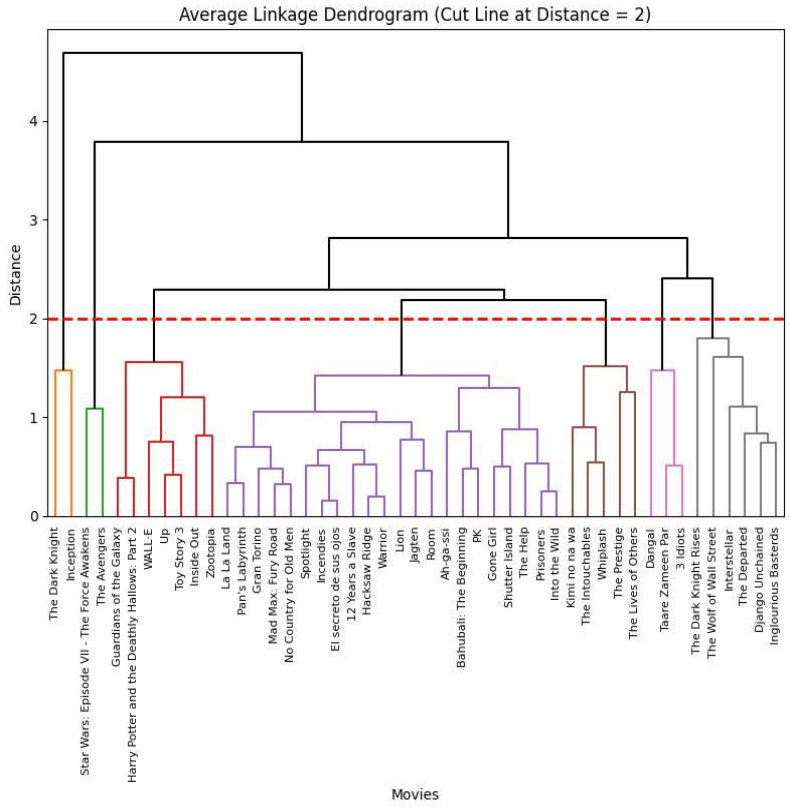


Figure 3

Complete, average, and single linkage methods were considered. The dendrograms illustrate how the single linkage isolated movies in their cluster. Whereas the complete and average linkage exhibited maximum silhouette scores of 0.344 and 0.378, respectively, for 7 clusters.

The figure illustrates the 7 clusters found by the average linkage method by specifying a cutoff distance of 2. A strong advantage of Hierarchical clustering is the ability to see the tree shaped dendrogram gives insight into within cluster similarities.

2.3. Comparative Analysis Hierarchical versus K-Means

The difference between the average silhouette score of Hierarchical and K-Means is very marginal, with Hierarchical scoring 0.007. Moreover, with the difference in clustering methodologies, this marginal difference is not informative.

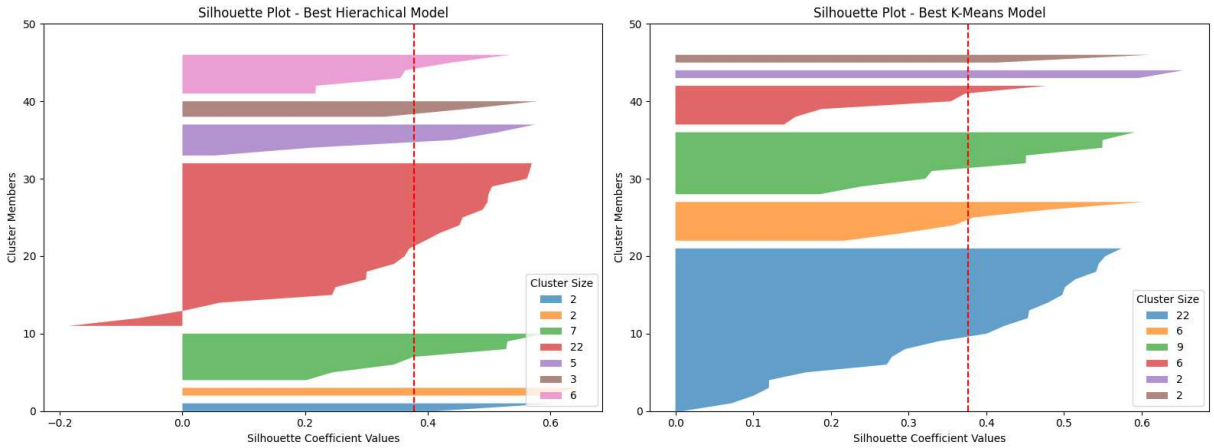


Figure 4

The silhouette graph above illustrates that K-means does not misclassify any points within the clusters, whereas the Hierarchical does. However, this misclassification falls under the same large cluster that K-Means identifies, whereas the smaller groups of clusters have a strong silhouette scoring in the hierarchical, indicating that the addition of the 7th cluster will give actionable insights.

2.4. Qualitative analysis

To enhance cluster interpretability, we used an interactive 3D plot to analyze the clusters, but it offered limited improvements. Therefore, we also created a data frame containing more details on the cluster. This helped us identify patterns by examining common factors. While not all clusters had a clear theme, the additional information improved the interpretability. Comparing the cluster results of k-means and hierarchical, hierarchical proved to have a slight edge on clarity of clusters. You can find clusters, titles in them, and their common characteristics in Appendixes [A]-[B].

Cluster	Description	Theme
1	Nolan's *Dark Knight* and *Inception*, which are his top 2 rated movies (IMDb, 2025)	Nolan's Top-Rated
2	*The Avengers* and *Star Wars: The Force Awakens*, both are major action blockbusters with large budgets and fanbases.	Large Budget BlockBusters
3	Pixar, Disney and other family adventure titles, easiest defined cluster.	Family Films
4	The largest cluster (22 titles) 11 countries of origin for movies, USA most represented, yet, it's the most country diversified, with majorly on drama.	International Drama Films
5	Focused on emotionally driven dramas centered around character transformation through human connection (e.g. *Kimi no wa*).	Character Growth Films
6	3 Bollywood films with Amir Khan as lead actor. (Outside of PK, included all of his movies in the dataset)	Bollywood Movies Starring Amir Khan
7	Six films equally split among Nolan, Tarantino, and Scorsese (all top 5 directors) with DiCaprio starring in half. Small yet diverse cluster.	High-profile director films

3. ChatGPT

During an exercise on the IMDB movie dataset involving dimensionality reduction and clustering, ChatGPT initially used PCA's cumulative explained variance to determine the optimal number of principal components. However, it misread the variance graph and prematurely concluded that two components accounted for most of the dataset's variance. As a result, it performed clustering with only two principal components. It was only when prompted to explore further components that ChatGPT added a third component, yielding more comprehensive insights.

To identify the optimal number of clusters, ChatGPT relied on the silhouette score—an established measure of clustering quality—and chose two clusters because they achieved the highest score. Although mathematically defensible, defaulting to two clusters oversimplified the data and missed opportunities for more nuanced segmentation. Another limitation was ChatGPT's exclusive use of K-means. It did not consider hierarchical clustering until explicitly prompted, at which point it found that three clusters provided a more revealing division of the IMDB data. Further prompting led ChatGPT to partition the dataset into five clusters, returning to the original data values to interpret each cluster's characteristics. This additional detail enriched the analysis and enhanced the overall understanding of the data.

While ChatGPT's approach produced results close to optimal, its findings are not entirely trustworthy without human oversight. Users should treat ChatGPT as a brainstorming and support tool rather than a definitive authority, verifying any methodology or conclusions it suggests.

4. Limitations and Conclusion

This analysis was limited by the data set size of 50 movies and the subjectivity of this dataset, where movie performance can be dependent on so many variables and nuances that create noise. Future analysis should consider more clustering methods, such as DBSCAN, and more thorough domain knowledge could shed light onto the clustering of certain segments.

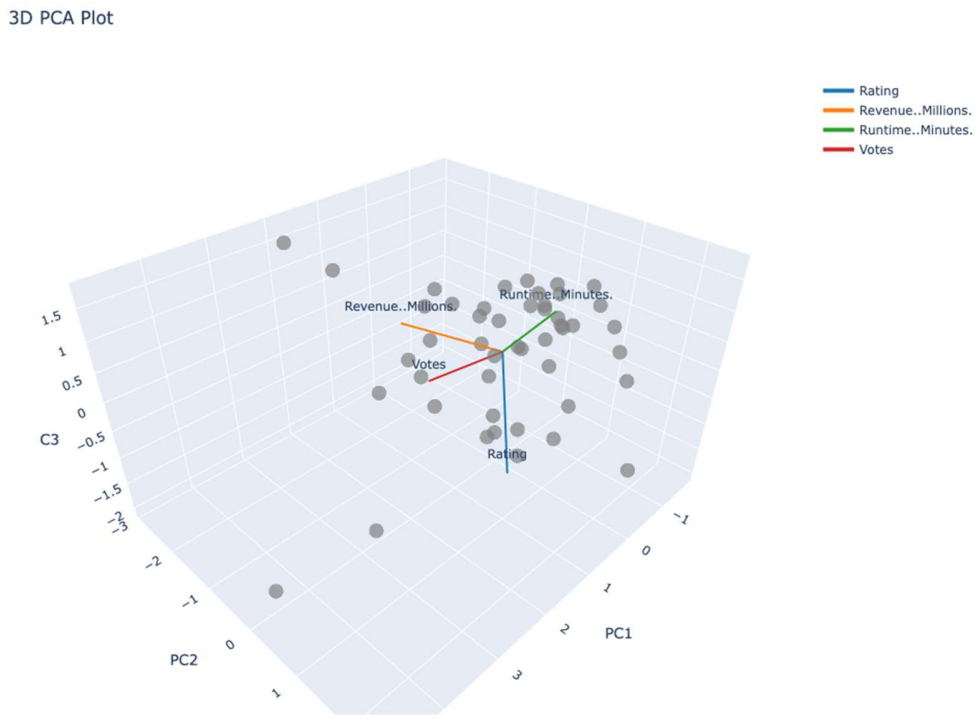
In conclusion, by considering multiple quantitative and qualitative analyses, we found that hierarchical clustering had the best silhouette score, and the 7 clusters made sense when looking at the domain knowledge as compared to K-means due to the additional segmentation of Bollywood movies.

References

- Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics, 28(1), p.100. doi:https://doi.org/10.2307/2346830.
- IMDb (2019). IMDb Top Rated Movies. [online] IMDb. Available at: https://www.imdb.com/chart/top/.
- IMDb. (n.d.). Best 50 Directors of the World. [online] Available at: https://www.imdb.com/list/ls071439230/.
- Naeem, A., Rehman, M., Anjum, M. and Asif, M. (2019). Development of an Efficient Hierarchical Clustering Analysis using an Agglomerative Clustering Algorithm. Current Science, 117(6), p.1045. doi:https://doi.org/10.18520/cs/v117/i6/1045-1053.

Appendices:

Appendix[A]: 3D PCA Plot



Appendix[B]: K-Means Clusters and Movies in them (with their most common characteristics)

Cluster ID	Number of Movies	Movie Titles	Most common characteristics and times they appear in cluster
0	22	Room, Spotlight, Warrior, Pan's Labyrinth, Incendies, El secreto de sus ojos, PK, Hacksaw Ridge, Lion, Prisoners, The Help, 12 Years a Slave, No Country for Old Men, Shutter Island, Ah-ga-ssi, Gone Girl, Jagten, Into the Wild, Bahubali: The Beginning, La La Land, Taare Zameen Par, 3 Idiots	Drama (Genre) — 21 Mystery (Genre) — 6 2016 (Year) — 4 Crime (Genre) — 4 Biography (Genre) — 4 2015 (Year) — 3
1	6	Interstellar, The Departed,	Drama (Genre) — 4

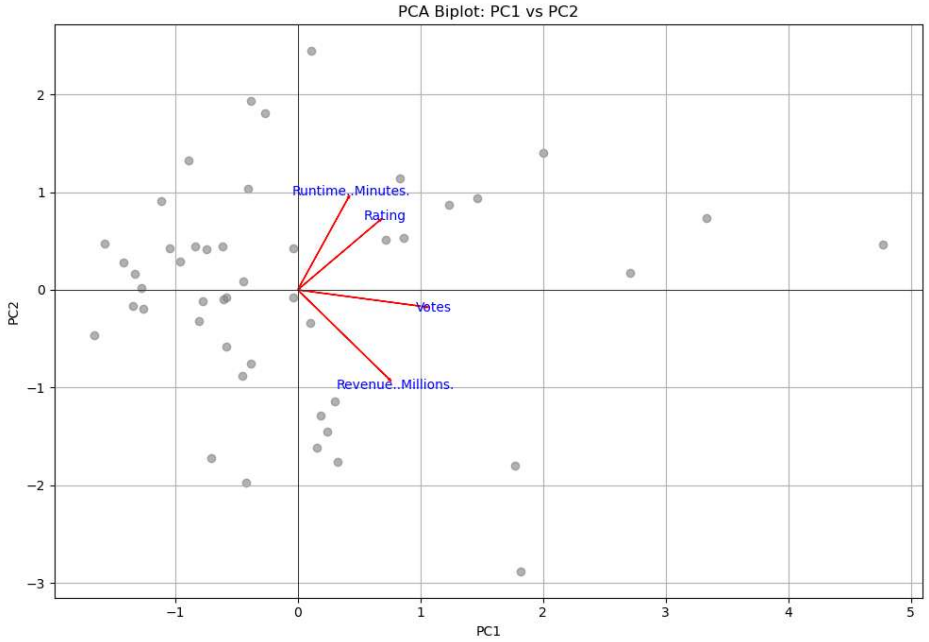
		Inglourious Basterds, The Dark Knight Rises, Django Unchained, The Wolf of Wall Street	Leonardo DiCaprio (Actor) — 3 2012 (Year) — 2 Christopher Nolan (Director) — 2 Martin Scorsese (Director) — 2 Quentin Tarantino (Director) — 2
2	9	Inside Out, Toy Story 3, Zootopia, Harry Potter and the Deathly Hallows: Part 2, Mad Max: Fury Road, WALL·E, Gran Torino, Guardians of the Galaxy, Up	Adventure (Genre) — 8 Animation (Genre) — 5 Comedy (Genre) — 4 2015 (Year) — 2 2008 (Year) — 2 Pete Docter (Director) — 2
3	6	Dangal, Whiplash, The Lives of Others, The Prestige, The Intouchables, Kimi no na wa	Drama (Genre) — 6 2016 (Year) — 2 2006 (Year) — 2 Biography (Genre) — 2
4	2	The Avengers, Star Wars: Episode VII - The Force Awakens	Action (Genre) — 2
5	2	Inception, The Dark Knight	Christopher Nolan (Director) — 2 Action (Genre) — 2

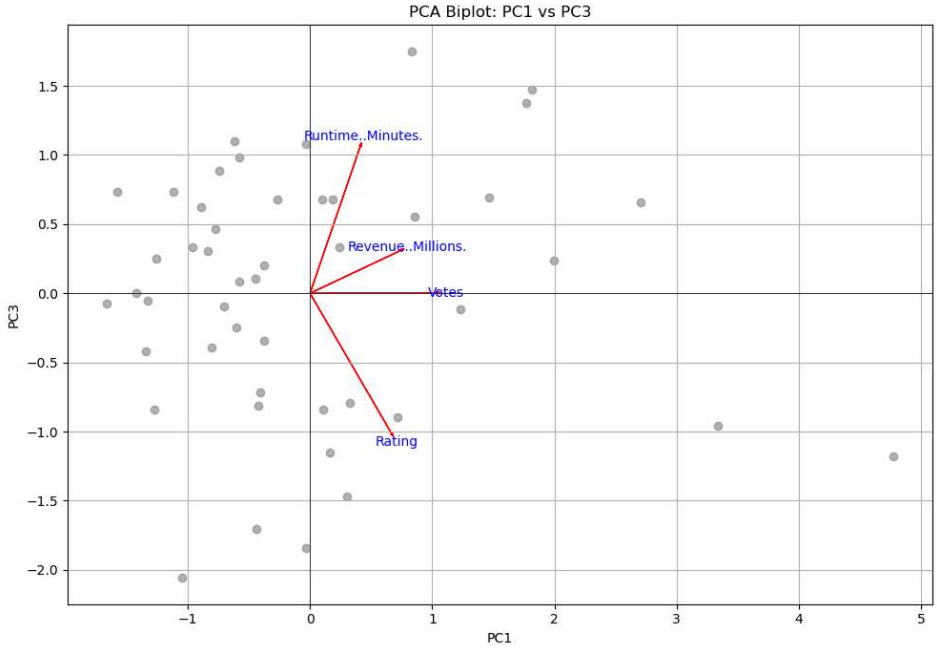
Appendix [D]: Hierarchical Clusters and the movies in them (with their most common characteristics)

Cluster ID	Number of Movies	Movie Titles	Most common characteristics and times they appear in cluster
1	2	Inception, The Dark Knight	Christopher Nolan (Director) — 2 Action (Genre) — 2
2	2	The Avengers, Star Wars: Episode VII - The Force Awakens	Action (Genre) — 2
3	7	Toy Story 3, Harry Potter and the Deathly Hallows: Part 2, Up, Guardians of the Galaxy, Zootopia, Inside Out, WALL·E	Adventure (Genre) — 7 Animation (Genre) — 5 Comedy (Genre) — 4 Pete Docter (Director) — 2
4	22	Gran Torino, Incendies, El secreto de	Drama (Genre) — 20

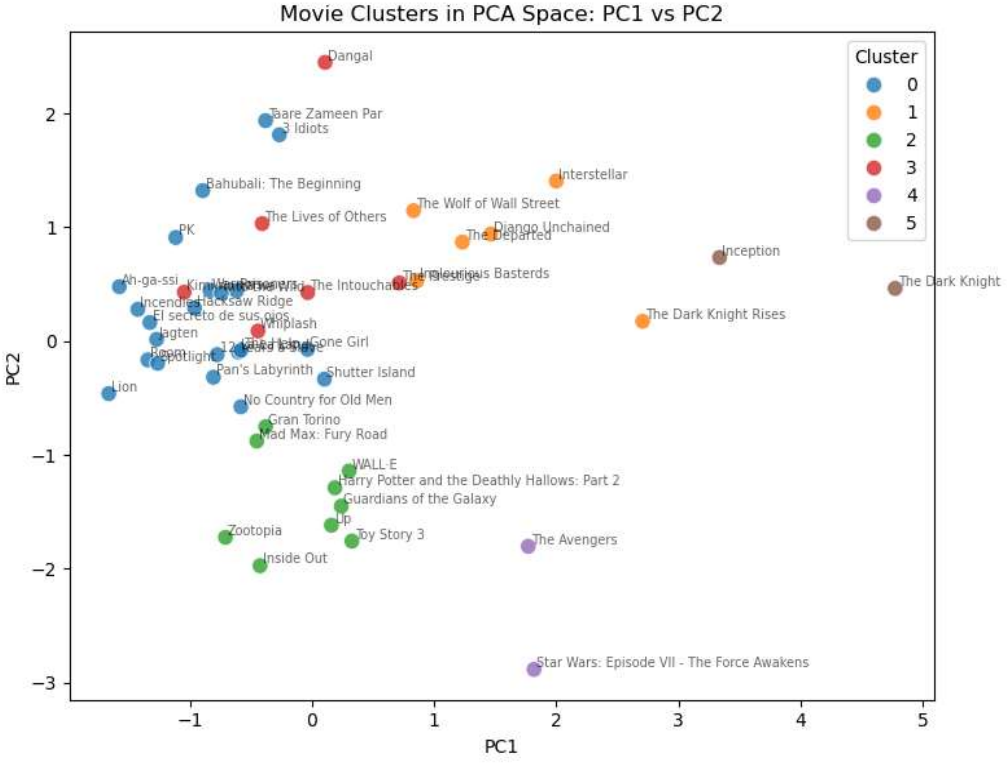
		sus ojos, PK, Mad Max: Fury Road, Gone Girl, Prisoners, The Help, 12 Years a Slave, No Country for Old Men, Shutter Island, Ah-ga-ssi, Lion, Pan's Labyrinth, Room, Spotlight, Warrior, La La Land, Bahubali: The Beginning, Into the Wild, Jagten, Hacksaw Ridge	Mystery (Genre) — 6 2015 (Year) — 4 2016 (Year) — 4 Crime (Genre) — 4 Biography (Genre) — 4
5	5	Whiplash, The Prestige, The Intouchables, Kimi no na wa, The Lives of Others	Drama (Genre) — 5 2006 (Year) — 2
6	3	3 Idiots, Taare Zameen Par, Dangal	Aamir Khan (Actor) — 3 Drama (Genre) — 3
7	6	Django Unchained, The Wolf of Wall Street, The Dark Knight Rises, The Departed, Interstellar, Inglourious Basterds	Drama (Genre) — 4 Leonardo DiCaprio (Actor) — 3 2012 (Year) — 2 Quentin Tarantino (Director) — 2 Martin Scorsese (Director) — 2 Christopher Nolan (Director) — 2

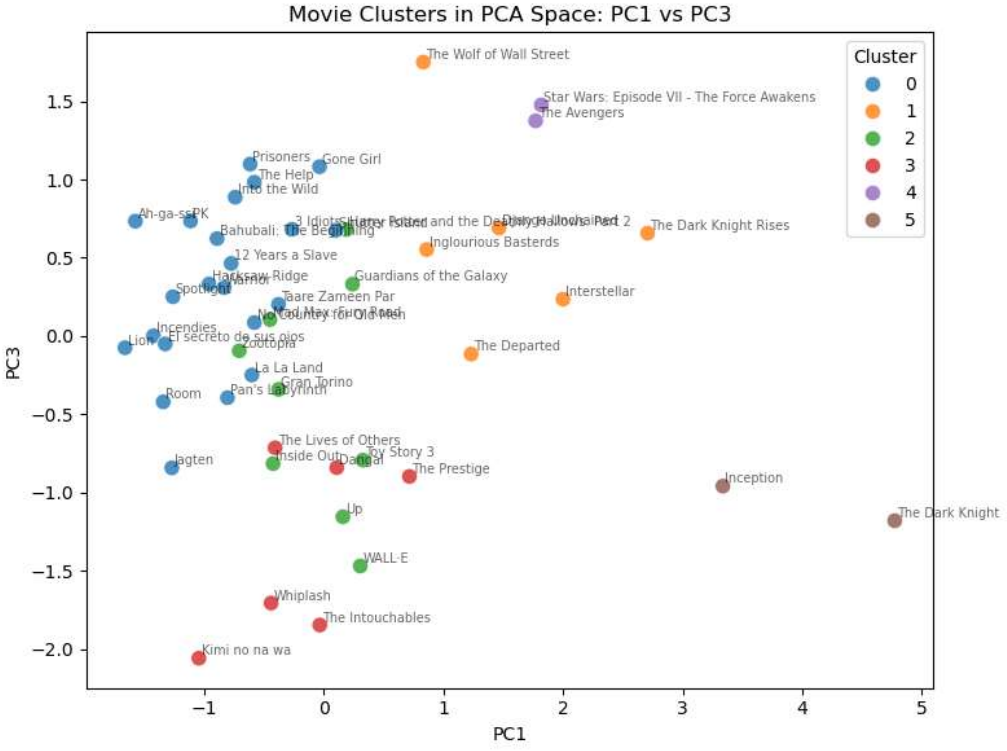
Appendix [E]: Plots for Loading PC1 vs 2 and PC1 vs 3 (3D interactive plot in code)



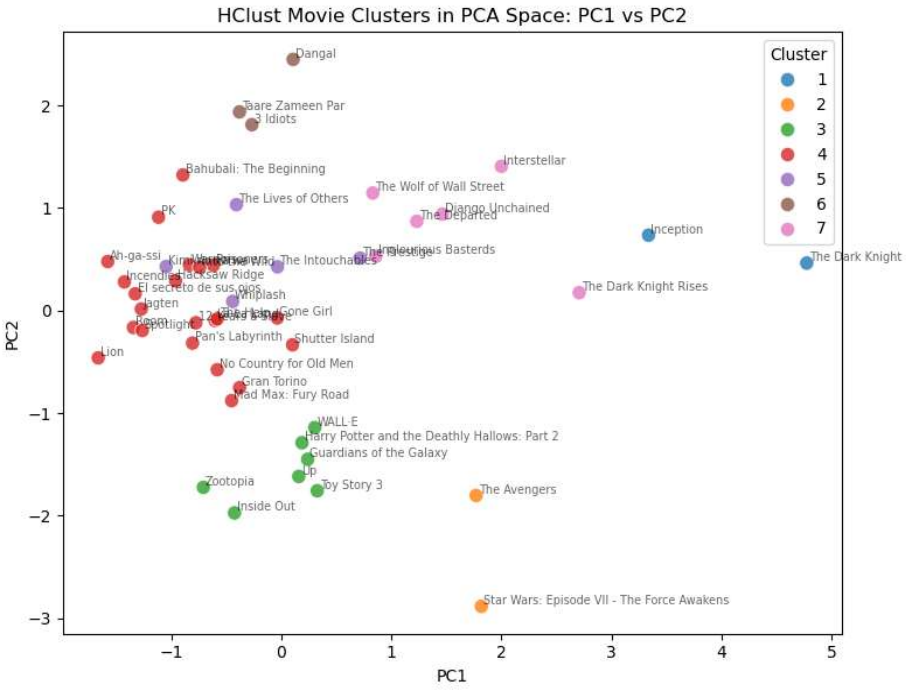


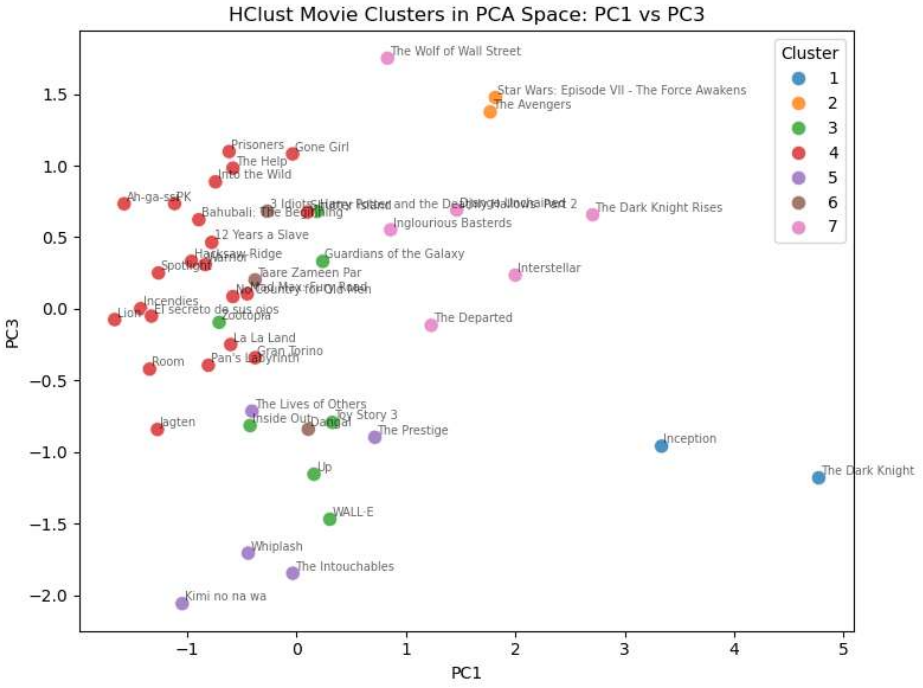
Appendix [F]: Plots for K-Means Clusters on PC1 vs 2 and PC1 vs 3 (3D interactive plot in code)





Appendix [G]: Plots for Hierarchical Clusters on PC1 vs 2 and PC2 vs 3 (3D interactive plot in code)





Appendix [H]: Table of Hierarchical versus Kmeans clustering movies

KMeans	0	1	2	3	4	5
HClust						
0	0	0	0	0	2	0
1	0	0	0	0	0	1
2	6	0	0	0	0	1
3	0	0	0	0	0	4
4	0	20	0	2	0	2
5	0	0	0	7	0	0
6	0	0	2	0	0	0