


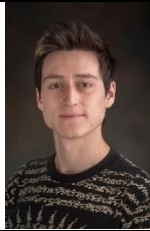




# Top 250 Movies Clustering

*Identifying Patterns among Successful Movies*

## Team

			
Avneet Kaur	Satyam Vatts	Negar Sadreddini	Ignacio Calderon de la Barca

## Table of Contents

<b><i>Introduction</i></b> .....	<b>3</b>
<b><i>Methods</i></b> .....	<b>3</b>
<b>Data Preparation</b> .....	<b>3</b>
<b>NLP</b> .....	<b>3</b>
<b>PCA</b> .....	<b>4</b>
<b>K-Means</b> .....	<b>4</b>
<b>Hierarchical Clustering</b> .....	<b>4</b>
<b><i>Results</i></b> .....	<b>5</b>
<b>NLP</b> .....	<b>5</b>
<b>PCA</b> .....	<b>5</b>
<b>K-Means</b> .....	<b>8</b>
<b>Hierarchical Clustering</b> .....	<b>8</b>
<b><i>Discussion</i></b> .....	<b>10</b>
<b>Cluster 0: Biographies &amp; Historical Movies</b> .....	<b>10</b>
<b>Cluster 1: R-Rated Crime Movies</b> .....	<b>11</b>
<b>Cluster 2: Superhero Action Movies</b> .....	<b>11</b>
<b>Cluster 3: Retro Movies</b> .....	<b>11</b>
<b>Cluster 4: International Movies</b> .....	<b>11</b>
<b>Cluster 5: Animated, Kids or Family Movies</b> .....	<b>11</b>
<b><i>Conclusion</i></b> .....	<b>12</b>
<b><i>Future Research</i></b> .....	<b>12</b>
<b><i>References</i></b> .....	<b>12</b>
<b><i>Appendix</i></b> .....	<b>13</b>
<b>Data Description</b> .....	<b>13</b>
<b>Data Distributions</b> .....	<b>13</b>

## Introduction

IMDB [\[1\]](#) is the largest and most accurate database for movie ratings and user reviews worldwide. It maintains a diverse set of cinematographic data ranging from actors & directors to production & marketing. The site provides an API to access this data [\[2\]](#) and particularly in our case we have gathered the top 250 site-ranked movies based on ratings and user reviews.

The aim of this project is to recognize patterns among these top-rated movies and identify similarities as well as differences among them. For this we applied unsupervised learning techniques to determine the type of natural clusters that exists within these movies.

## Methods

This section explains various techniques applied to preprocess the data and perform clustering.

### Data Preparation

The dataset consists of some common attributes related to movies such as genre, runtime, actors, plot etc. It has a wide variety of movies dating from 1921 to 2022, scrapped from IMDB website's updated list as of March 31, 2022.

### NLP

After the abandonment of complex rigid hand-written rules for language processing for Machine Learning algorithms, like k-mean clustering, the field of Natural Language Processing (NLP) has leaped into big advancements. In this present work, NLP technique has been used to derive insights from text data. A requirement for these type of algorithms to work is to preprocess the data with word embedding techniques. It is important to point that if a simple naive word encoding is used, there would be limitations in the sense that not all the words are worth of evaluation in text analysis. For example, there are many stop words in the English language that do not influence the overall sentiment or purpose of a text. Also, there is another type of unworthy characters in terms of text overall meaning that called punctuations. We identified both stop words and punctuations and removed them.

The next step implies vectorizing all the remaining words in the column Plot with TF-IDF [\[3\]](#) vectorizer. In this step, each word has their own specific column, and each row identifies if that particular word presence in the movie description (Plot) of that row. After vectorizing all the words, we extracted the TFIDF score for all the new columns and finally filtered out top 25 words based on TFIDF score. We removed the variable Plot and only included the new 25 column for the top 25 words.

At this stage, all the variables are in numerical type and the data can be used in analytical evaluations. But before staring clustering, we needed to check the dimension of the data.

## PCA

Before we start conducting any analysis, we must check the dimension of the data. If the number of columns is too high especially comparing with the number of rows, then we will face dimensionality problem. We checked the shape of the data and it had 250 rows and 95 columns. To address the problem of dimensionality, we implemented PCA technique.

In general, PCA reduces dimension of the data by producing new principal components as data variables [4]. Each principal component (PC) explains a portion of the main data with difference weights for each column. In this project, we chose the threshold of 80% for the variance explanation of the data and by checking the explained variance of each principal component, we realised that the first 25 PCs will reach to our defined goal.

We picked the top 25 PCs as our data to proceed the clustering techniques with. The first PC explains 11% variance of the data while the second and third PCs explain 9% and 8.5% respectively.

## K-Means

We have applied K-means clustering technique [5] to determine natural clusters among the movies. As a movie has different characteristics such as genre, rating, runtime etc., we cannot call it an outlier if it misses any one of the qualities that the other movies in that cluster have. We confirmed our hypothesis by implementing DBSCAN [6] which gave us the entire dataset as an outlier or 0 outliers. Thus, it confirmed that the existence of an outlier in such a dataset is not logical. Hence, we chose K-means as an optimal candidate for clustering rather than K-medoids (which is robust to outliers and generally used in data with noise).

Since the dataset after preprocessing consists of all numerical variables in the form of principal components obtained from PCA, the K-Means algorithm using Euclidean distances was taken as rational approach to form clusters.

The optimal number of clusters were determined using two popular approached as mentioned below:

- Elbow Method: Used to determine optimal K based on Inertia/WSS. We checked the number of clusters ranging from 2 to 10 as we didn't want large number of clusters.
- Silhouette Score: Based on the mean intra cluster and nearest cluster distance, where value ranges between 1 and -1.

## Hierarchical Clustering

The clusters formed by K-Means provided an intuition of further subdivision of the largest cluster. Since the objective of the project was to explore different unsupervised clustering techniques, we tried another approach by applying Agglomerative Hierarchical Clustering [7].

Using the same affinity of Euclidean distance, different linkage techniques – Ward, Complete, Average, and Single were compared based on Silhouette score metric to determine optimal approach for our dataset.

## Results

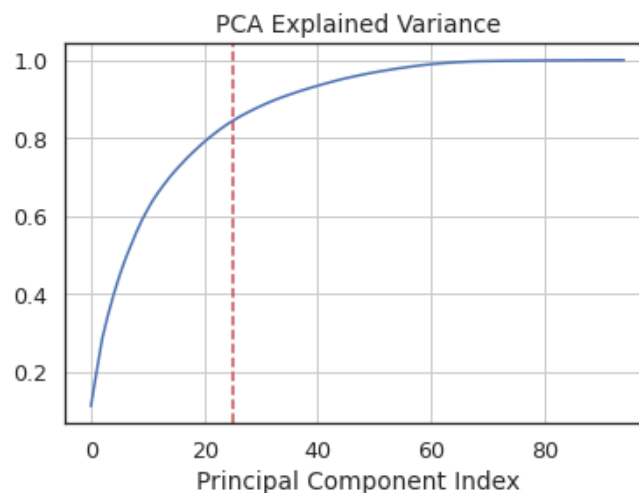
### NLP

One of the important non-numerical variables in our data is the column “Plot” that is a movie description. NLP technique (Tf-Idf Vectorization) was applied to convert it to meaningful and valuable numbers so that we perform further analysis. After the conversion of words to numbers, we selected top 25 most frequent words and only kept those words relating columns in the data. The table below, shows top 25 words with their weight (TFIDF scores) in the variable Plot:

Top 25 Words														
Word	Life	Find	young	Man	Help	War	Old	World	New	Year	Friend	Try	Family	Murder
TFIDF Score	3.06	2.75	2.74	2.57	2.24	2.16	2.04	1.93	1.87	1.83	1.77	1.76	1.75	1.70
Word	Son	Woman	City	Child	Work	American	struggle		Boy		Force	Wife	Live	
TFIDF Score	1.65	1.64	1.61	1.56	1.55	1.47	1.40		1.36		1.35	1.32	1.26	

### PCA

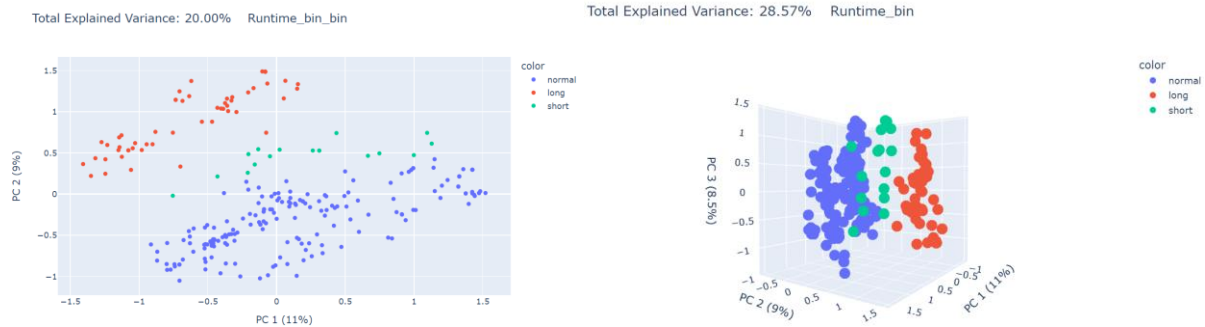
The PCA technique was performed with the threshold of 80% for variance explanation. The graph below, shows the cumulative explained variance of principal components:



Then, we randomly chose some of the initial features to check the spread of those categories among top 2 PCs in 2D plots and top 3PCs in 3D plots.

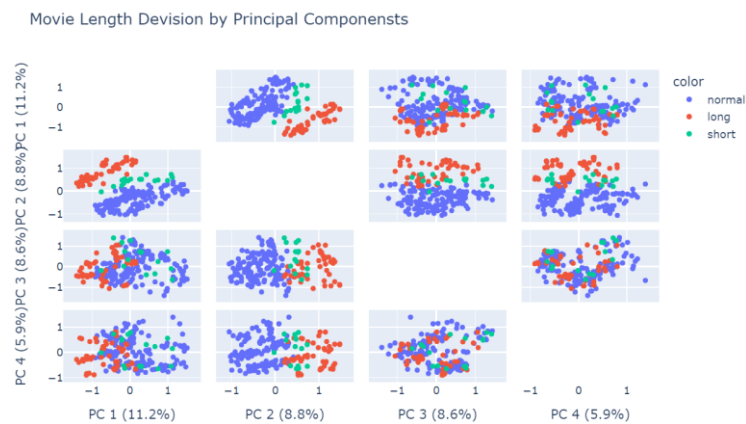
For example, for the categorical variable of length after binning it to 3 categories of short, normal, and long, the 2D and 3D plots of top principal components are shown below:

## Runtime Binned:

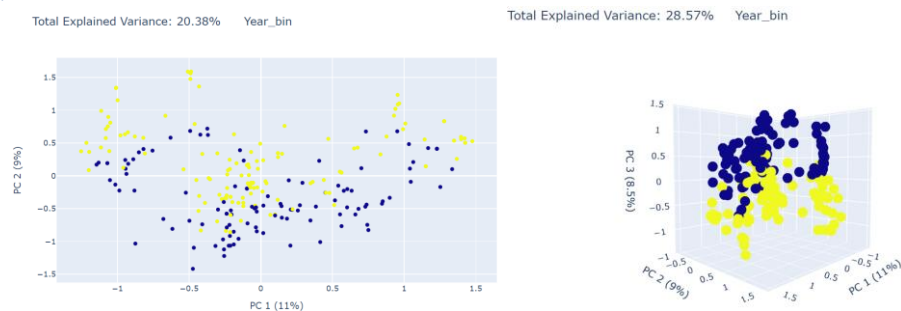


We observed that the different categories of movie lengths have been well-separated among top principal components.

The division of Runtime categories among the principal components are shown in the image below:



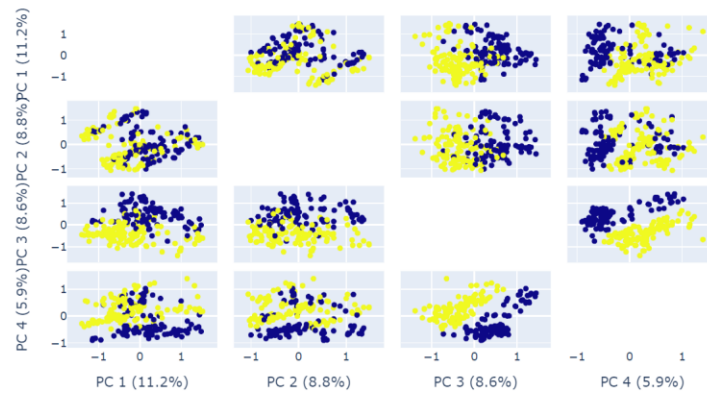
## Year Binned:



For this identifier, we saw that the first 2 components failed to split them while including the third component, there will be a considerable difference in terms of category division.

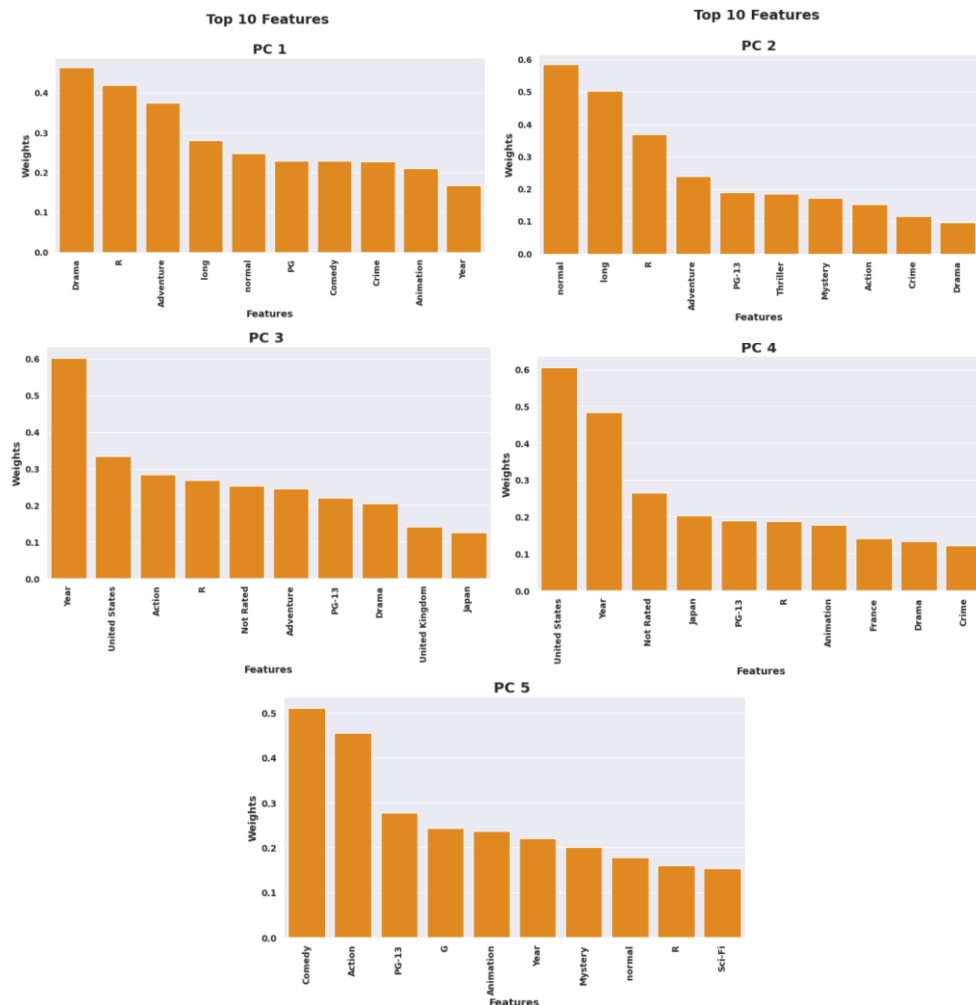
The division of Casted year categories among the first four principal components are shown in the image below:

Movie Casted Year Devision by Principal Componensts



We observed that the category of year casted is split best among the principal components of 3 and 4.

After checking the category split among the principal components, we looked deeply into the first 5 PCs to find the variable importance weigh in each of them. The part plots below, show the variable weight of those PCs respectively from PC1 to PC5:

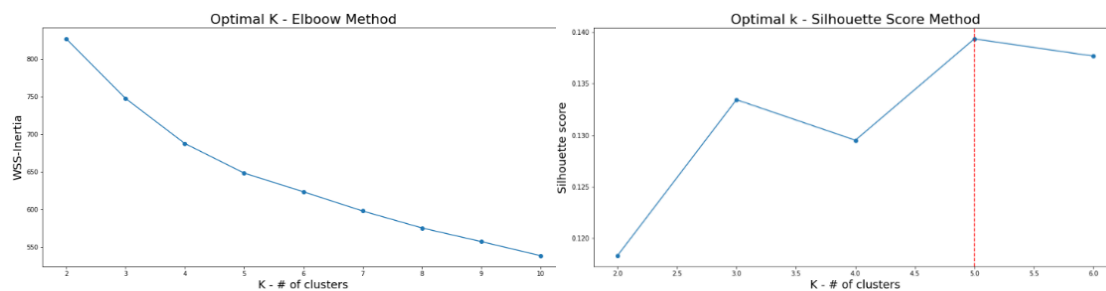


Looking at the first 5 principal component, we can see that each of them explains one feature of the IMDB data better than the other features. The table below, shows the main feature of each principal components shown in the bar plots above:

Principal Component	PC 1	PC 2	PC 3	PC 4	PC 5
Feature Identifier	Genre	Runtime (Movie Length)	Year	Country	Genre and Rating

## K-Means

The K-Means clustering analysis performed using Euclidian distances generated 5 distinct clusters based on the optimal K- value observed from Elbow Method and Silhouette Score.



Although, it was not quite clear whether the optimal k value is 4 or 5 from Elbow Method, but the Silhouette Score provided a clear indication of 5 clusters in the dataset.

The clusters developed are as follows:

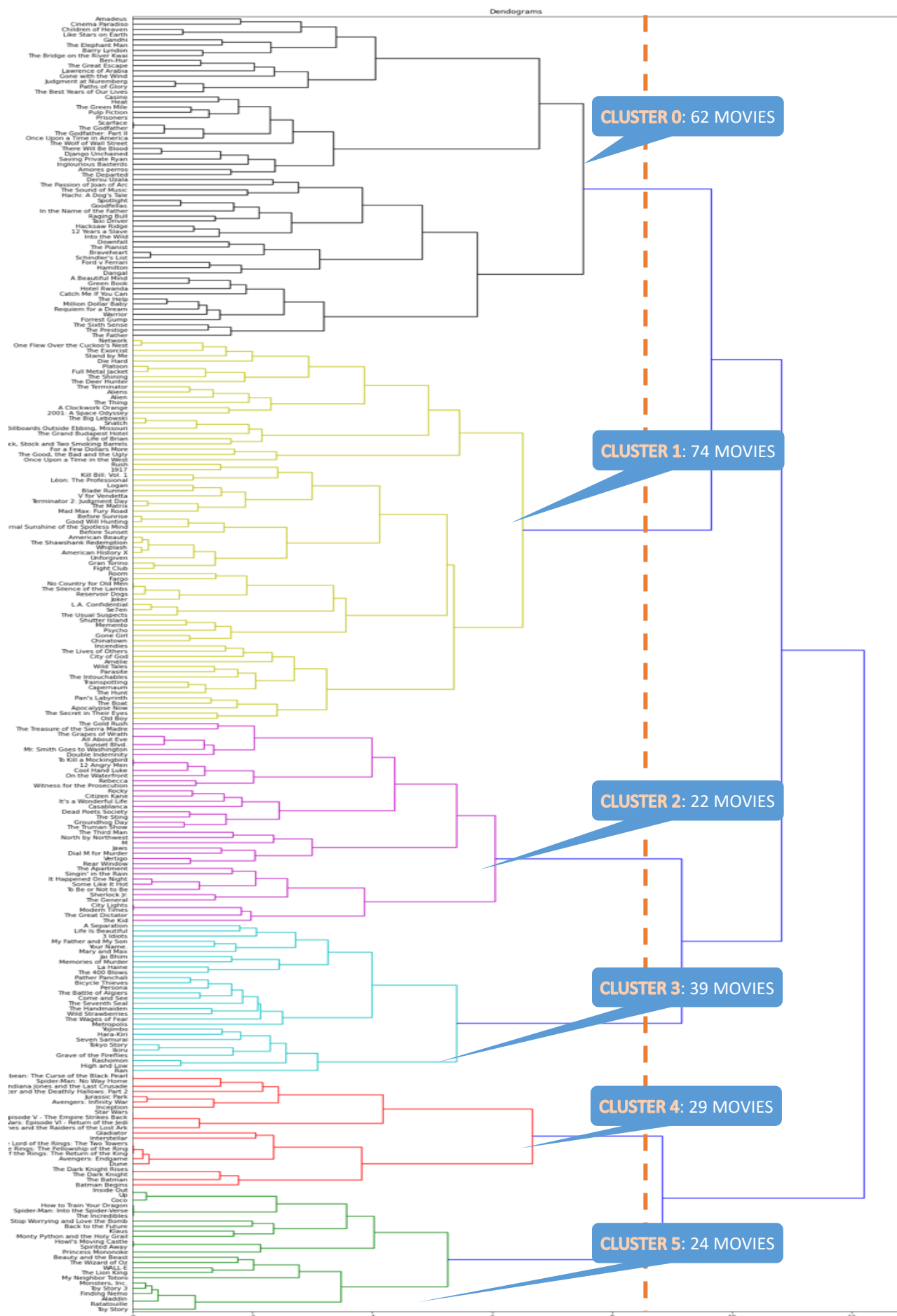
Cluster	0	1	2	3	4
Number of Movies	100	60	21	35	34
Proportions	40%	24%	8.4%	14%	13.6%

The largest cluster i.e., Cluster 0 contained 40% of the movies. This proportion was suspiciously high and required further analysis to figure out whether it can be broken down into further clusters. This led us to analyse the clusters through hierarchical clustering.

## Hierarchical Clustering

The Hierarchical clustering analysis was performed using Ward linkage & Euclidian affinity through Agglomerative approach. The dendrogram generated displayed 6 distinct clusters as shown below:





The comparison between K-Means and Agglomerative Clustering clearly describes the division of the largest cluster of K-Means into two distinctive clusters in Agglomerative Clustering.

		Hierarchical Agglomerative Clusters						
		0	1	2	3	4	5	All
K-Means Clusters	0	30	69			1		100
	1	12	2	4	39		3	60
	2						21	21
	3	14	1	18				35
	4	6	2			26		34
	All	62	74	22	39	29	24	250

Moreover, The Cluster 1 & 3 also get subdivided as shown in table above. This is based on various characteristics of the movies that separated them based on Hierarchical Structure. The ward linkage method used is similar to K-means algorithm as it analyses the variance of clusters thus, was an appropriate extension to drill down the results of K-Means.

## Discussion

The results of K-Means and hierarchical were utilized to draw conclusions regarding the 6 distinct clusters formed. The characteristics & appropriate labels of these clusters are described below. The word cloud image display frequent words in movie plots of corresponding cluster.

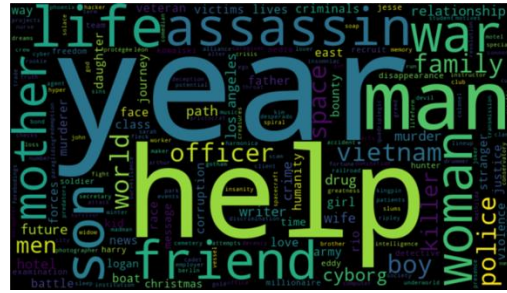
### Cluster 0: Biographies & Historical Movies

- Biographies: Wolf of the Wallstreet, Gandhi, Dangan Historical : Schindler's List, Hamilton
- Rated: Majorly R & PG
- Runtime: Dominantly Large (more than 150 mins)
- Actors: Robert De Niro, Al Pacino, Tom Hanks, Leonardo Di Caprio
- Directors: Martin Scorsese, Quentin Tarantino, Steven Spielberg.
- Plots: Mainly focused on World war, Battles, American Life



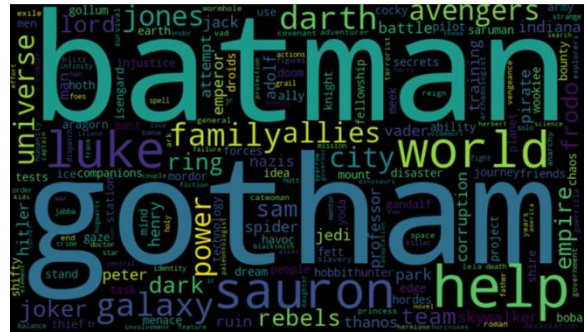
## Cluster 1: R-Rated Crime Movies

- Movies: Se7en, The Terminator, A Clockwork Orange, The Matrix
- Rated: R Rated
- Actors: Kevin Spacey, Clint Eastwood, Morgan Freeman
- Directors: Stanley Kubric, Sergio Leone.
- Plots: Mainly focused on Assassins, Murder, Investigations etc.



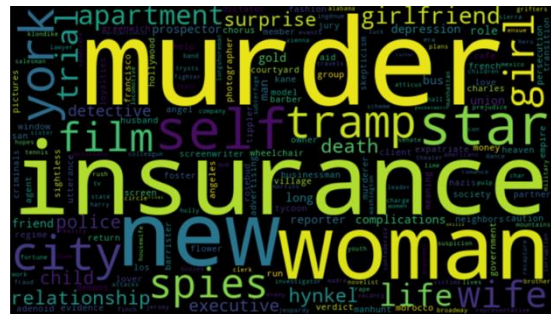
## Cluster 2: Superhero Action Movies

- Biographies: Batman Trilogy, Interstellar, Avengers-Endgame & Infinity War, Star Wars, Indiana Jones
- Rated: Majorly PG-13 & PG
- Year: Released after 1990
- Actors: Harrison Ford, Christian Bale
- Directors: Christopher Nolan, Steven Spielberg
- Plots: Mainly focused on superhero like Batman, Avengers, or Sci-Fi like Star Wars



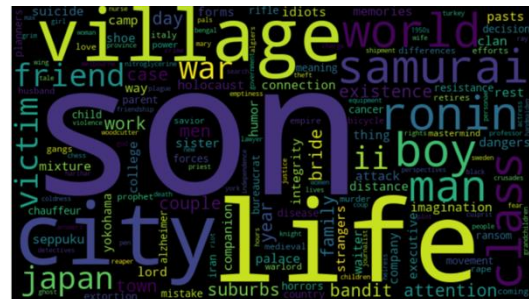
### Cluster 3: Retro Movies

- Movies: RomComs such as Vertigo, The Apartment, Modern Times, Rear Window
- Year: Released before 1990 thus, retro
- Rated: PG, Passed or Approved
- Actors: Charles Chaplin, James Stewart
- Directors: Alfred Hitchcock, Billy Wilder, Charles Chaplin
- Plots: Mainly focused on women, murders, relationships etc.



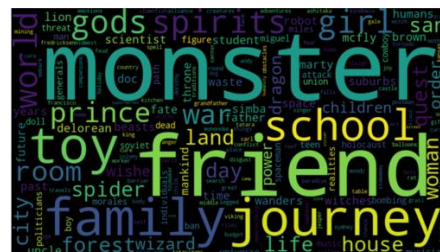
## Cluster 4: International Movies

- Movies: Seven Samurai, 3 Idiots, Yojimbo, Life is Beautiful, A Separation, Ran
- Rated: Majorly Not rated
- Country: Japan, France, India, Australia, Italy
- Actors: Toshiro Mifune, Tatsuya Nakadai
- Directors: Akira Kurosawa



### Cluster 5: Animated, Kids or Family Movies

- Movies: Toy Story 1 & 3, Spirited Away, Finding Nemo, Aladdin, Wall-E, The Incredibles, Beauty and the Beast
- Rated: Majorly PG & G
- Actors: Tom Hanks
- Directors: Hayao Miyazaki, Lee Unkrich



## Conclusion

Based on the analysis performed, the following conclusions are drawn:

- Hierarchical Clustering generated well-defined Clusters better than K-Means
- Crime, Drama, Action & Biography Movies were the dominating genre.
- Superhero Action & Sci-Fi Movies were more popular after 1990 potentially because of better VFX.
- Retro Movies were more popular for Comedy & Romance Dramas.
- International Movies were dominantly from Japan
- Kids & Family Movies were mostly Anime with Adventurous Plots

## Future Research

The study can be further extended by implementing the following techniques:

- Transformers like BERT are better pre-processing techniques for processing textual data like plot. It enables to extract contextual information as well. It can dramatically improve the clustering results and may provide more insights into data.
- The Cluster Labels obtained can be used to label top 250 movies and performed supervised learning techniques to utilize as recommendation systems in applications such as Netflix, Prime Video etc.
- The clustering analysis can be even extended to TV Shows, and other media available on IMDB to determine user preferences based on reviews by similar users. This can potentially work as user segmentation based on movie preferences.

## References

- [1] IMDB. IMDB API Documentation. url: <https://imdb-api.com/API>. Accessed: 2022-31-03.
- [2] IMDB. IMDB: Ratings Reviews and Where to Watch. url: [https://www.imdb.com/?ref\\_=nv\\_home](https://www.imdb.com/?ref_=nv_home). Accessed: 2022-15-04.
- [3] Akiko Aizawa. "An information-theoretic perspective of tf-idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45–65
- [4] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459
- [5] K Krishna and M Narasimha Murty. "Genetic K-means algorithm". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29.3 (1999), pp. 433–439.
- [6] Erich Schubert et al. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". In: *ACM Transactions on Database Systems (TODS)* 42.3 (2017), pp. 1–21
- [7] Athman Bouguettaya et al. "Efficient agglomerative hierarchical clustering". In: *Expert Systems with Applications* 42.5 (2015), pp. 2785–2797

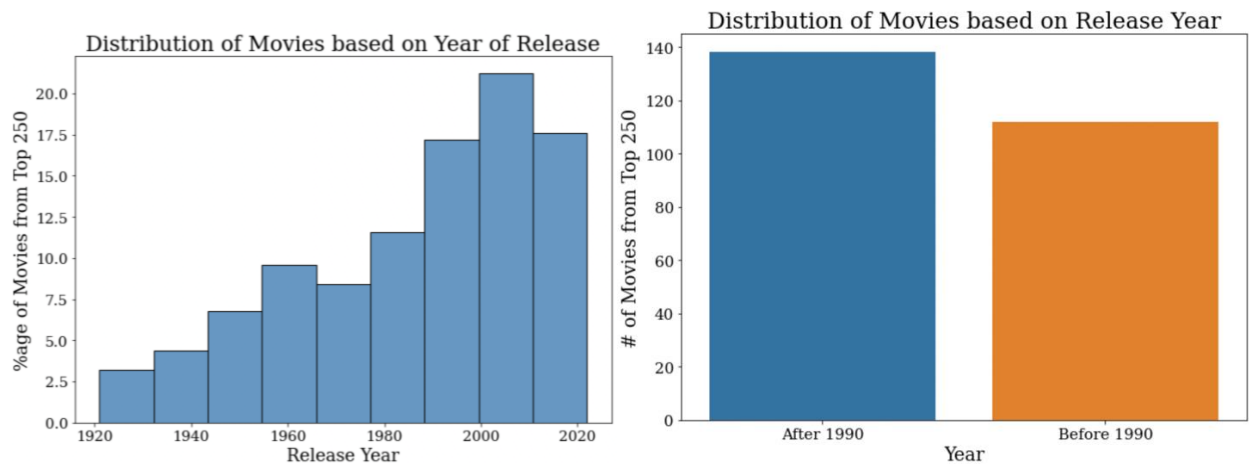
# Appendix

## Data Description

Variable	Type	Description
Title	Categorical-Nominal	Name of the movie
Year	Numerical-Continuous	The year movie casted
Rated	Categorical-Nominal	Classification of movies based on their suitability for audiences
Released	Date	The date movie was released
Runtime	Numerical-Continuous	Movie length
Genre	Categorical-Nominal	Movie genre
Director	Categorical-Nominal	Director of the movie
Writer	Categorical-Nominal	Writer of the movie
Actors	Categorical-Nominal	Actors starred in the movie
Plot	Text	Movie description
Language	Categorical-Nominal	Language of the movie
Country	Categorical-Nominal	Country of movie was casted
Awards	Categorical-Nominal	Number of winnings and nominations for Oscar and other film festivals
Poster	Text	Link of the poster image of the movie
Ratings	Json	Movie ratings on IMDB, Rotten Tomatoes, and Metacritic
Metascore	Numerical-Continuous	
IMDB-Rating	Numerical-Continuous	Movie rate on IMDB
IMDB-Votes	Numerical-Continuous	Number of votes for the movie on IMDB
IMDB-ID	ID	Movie ID on IMDB
TYPE	Categorical-Nominal	All the values are: "movie"
DVD	Date	The date movie DVD released
BoxOffice	Numerical-Continuous	The money movie made in USD
Production	NA for all values	NA for all values
Website	NA for all values	NA for all values
Response	Boolean	True for all the values

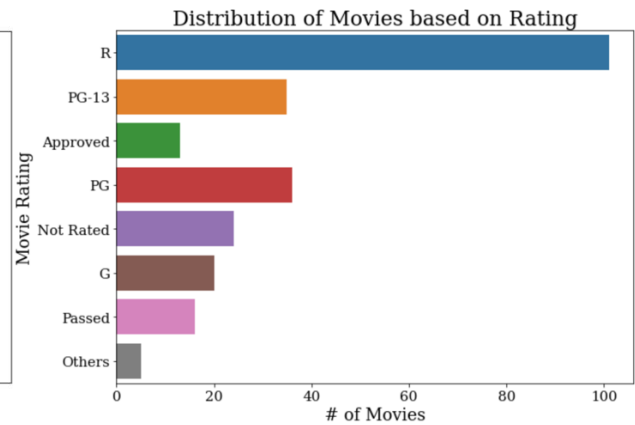
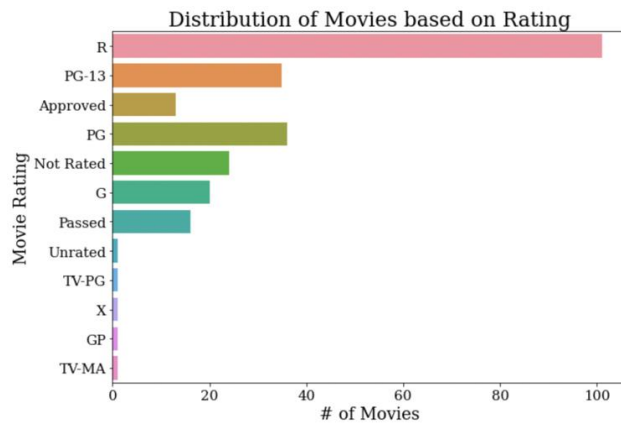
## Data Distributions

**Year:**

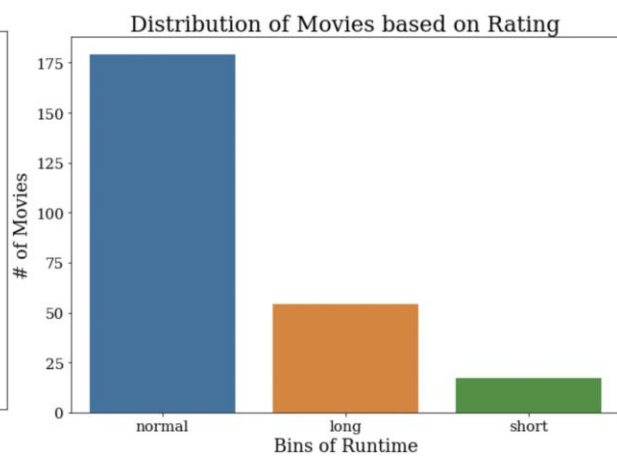
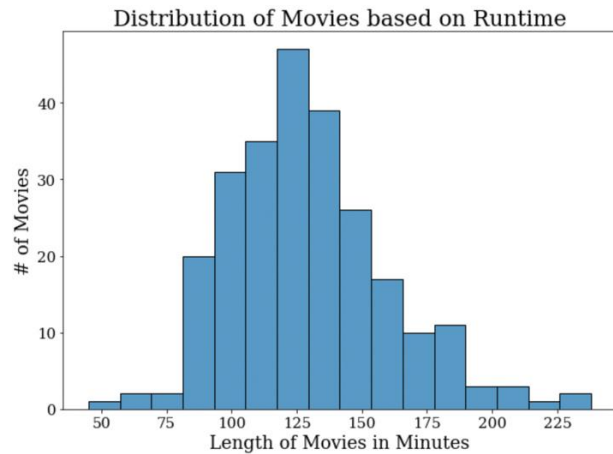




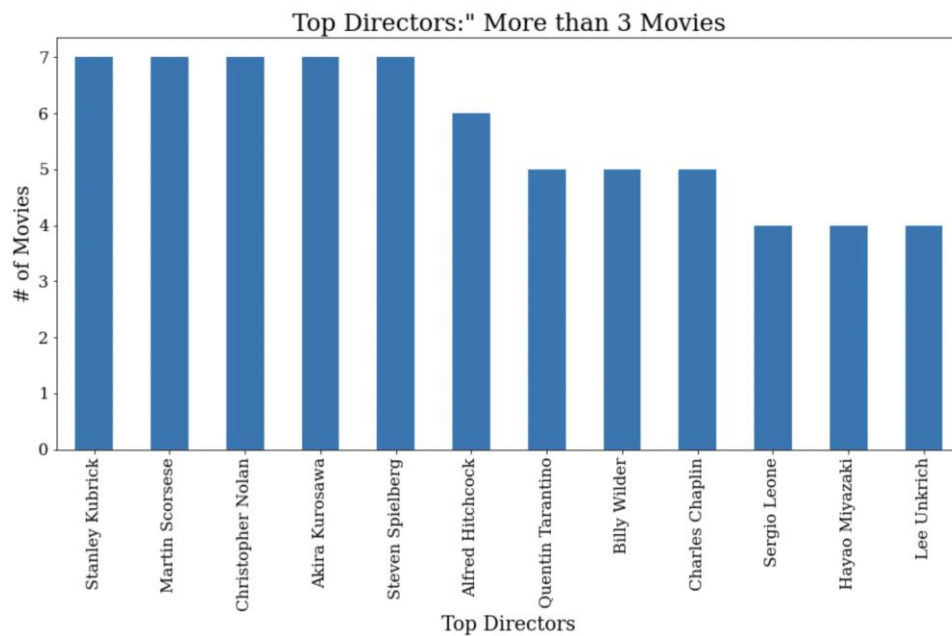
## Rated:



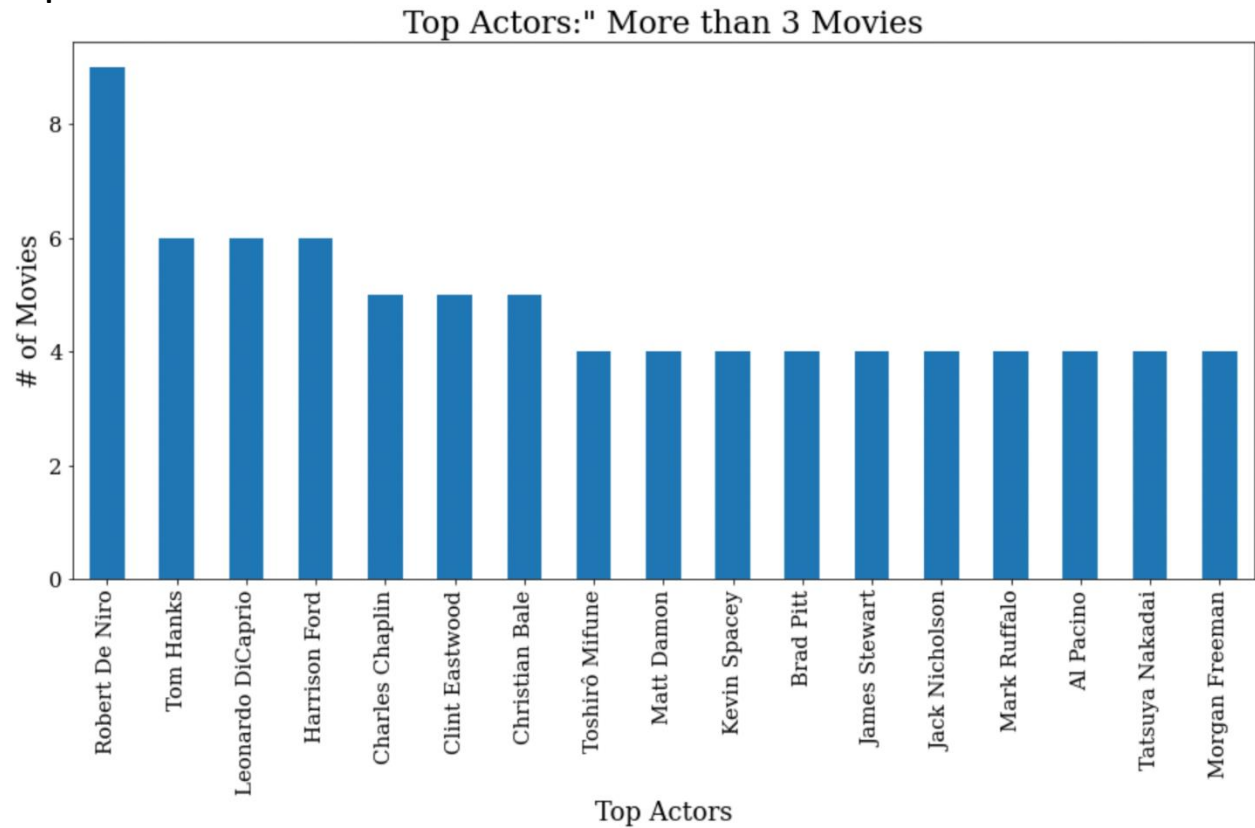
## Runtime:



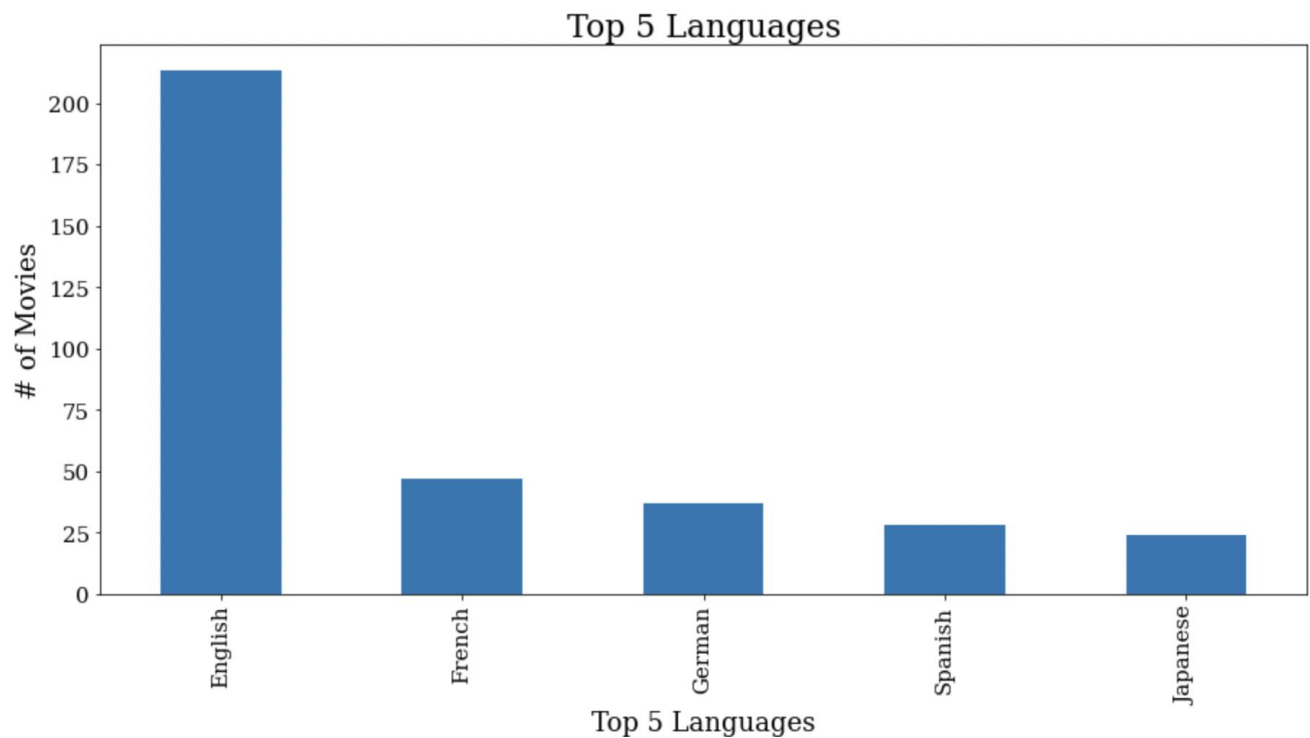
## Top Directors: With 4 or more movies



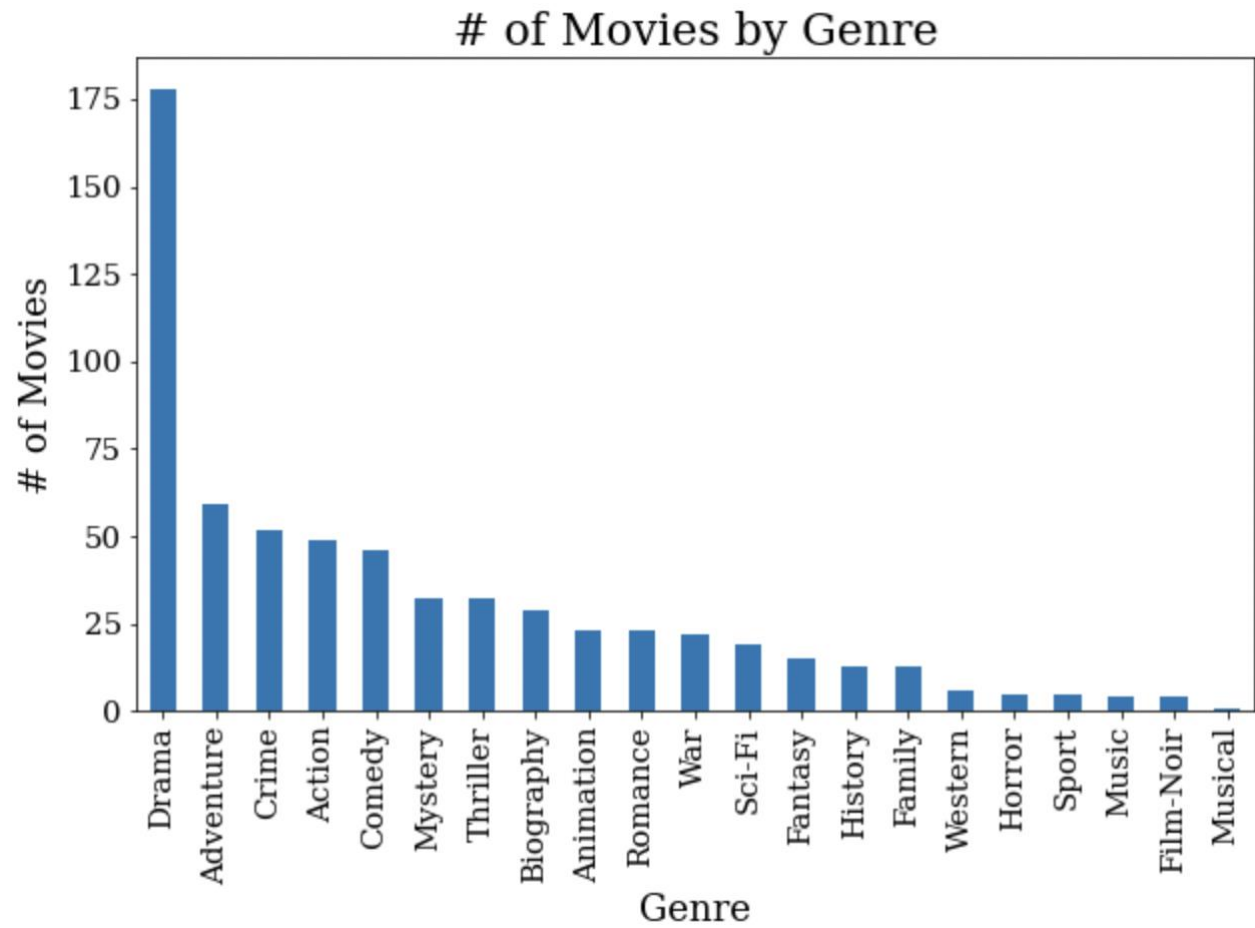
### Top Actors: With 4 or More Movies



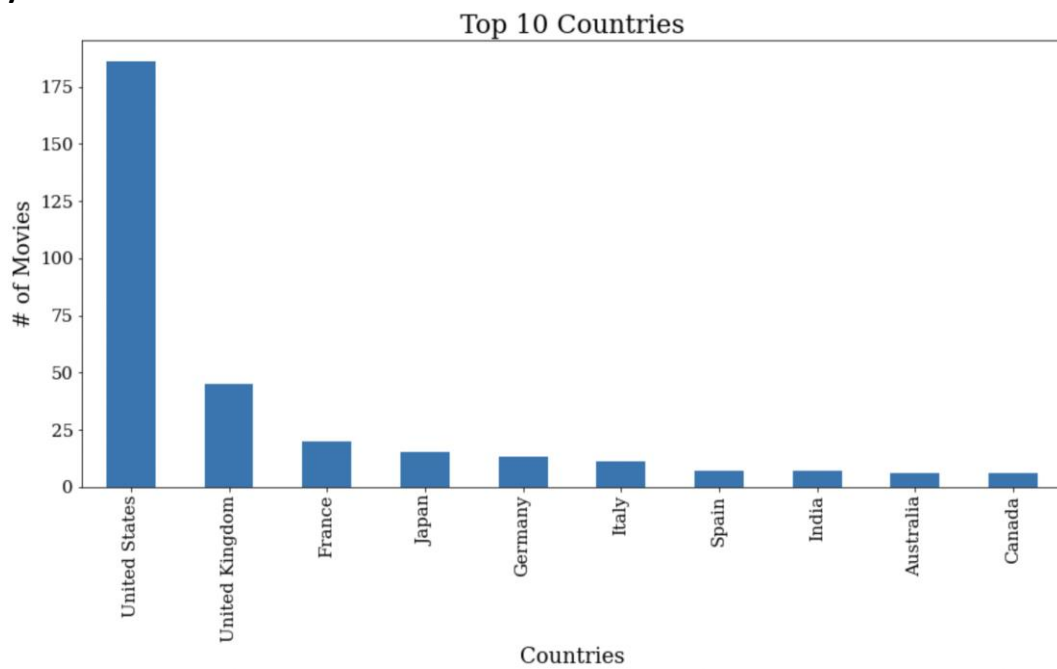
### Languages: Top 5



Genre:



Country:





**Top 25 words**

