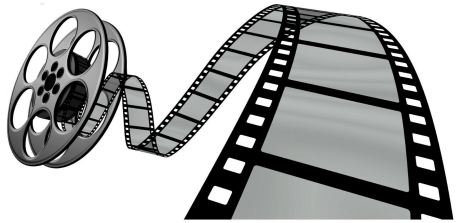


IMDB Movie Data Set Analysis: Clustering

Sergio Navia, Erik Vela, Jason Wo

Dataset

- IMDB Movies Dataset
- https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Our Dataset consisted of the top 1000 movies from IMDb. It includes information on these movies based on several different categories such as rating, runtime, box office gross, year of release and many others.

Variables of the Dataset

- Poster Link
  - A link to a poster of the movie provided by Amazon that IMDB uses.
- Series Title
  - The title of the movie.
- Release Year
  - The year the movie was released according to IMDB.
- Certificate
  - The movie rating decided by the country the movie was originally released in.
- Runtime
  - Duration of the movie in minutes.
- Genre
  - The genre the film was categorized in.
- IMDB Rating
  - A rating of the movie on the IMDB website.
- Overview
  - A description of the movie. Sometimes a summary.

Variables of the Dataset

- Meta Score
  - a weighted average of reviews from top critics and publications for a given movie calculated by the Metacritic website. https://www.metacritic.com/about-us/
- Director
  - Name of the person who directed the movie.
- Star 1, Star 2, Star 3, Star 4
  - The names of the main cast of the movie.
- Number of Votes
  - According to the IMDB, “IMDb registered users can cast a vote (from 1 to 10) on every released title in the database. Individual votes are then aggregated and summarized as a single IMDb rating, visible on the title's main page.”
  - This variable is the total amount of these votes for a particular movie.
- Gross
  - Money earned at the box office for every release of a movie. (Certain movies are released in theaters more than once.)

Variables of the dataset

The following variables are being used in the clustering analysis:

- Released_Year:
  - The year the movie was released.
- Runtime:
  - The duration/length of the movie (minutes).
- IMDB_Rating
  - Rating of the movie on the IMDB site with a scale 1-10.
- Meta Score
  - A weighted average of reviews from top critics on the movie.
- Gross
  - Gross earning of the movie, or the amount of money that the movie make(in US dollar).

Data Preparation

This Dataset initially was not usable, it contained many categorical variables, N/A terms, and missing or obviously incorrect values.

- Our first step in removing the data was to remove all N/A and non usable entries in the dataset
- Feature engineered Genre variable from single multiclass variable into several binary variables
- We then removed the units of measurement attached to some of the numerical values (e.g. minutes in runtime)
- Finally, we removed the column that contained the names of the movies and saved them as the names of the rows
- After all the data cleaning, we were left with 713 rows and 5 columns to do analysis on.
- In order to account for how large Gross can be and affecting the clustering, we standardized the dataset.

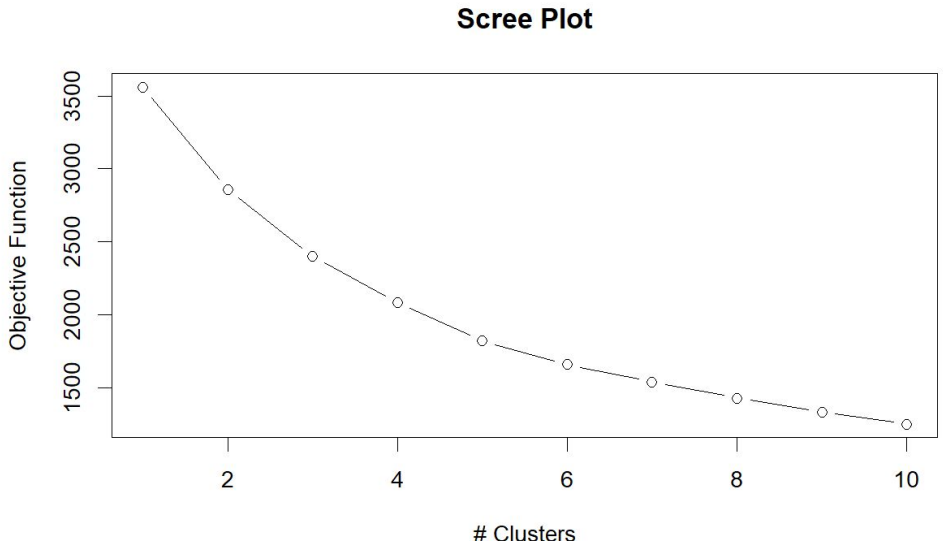
Data Preparation

- We could have turned certificate into numerical values by giving each rating of the movie a number from 1 - n, where n is the total number of ratings a movie can be given.
- But as soon as we tried numbering the ratings, we noticed that different countries had different rating scales compared to the U.S.

```
Released_Year      Runtime        IMDB_Rating      Meta_score         Gross                Action          Adventure       Biography        Animation
Min.   :1930   Min.   : 72.0   Min.   :7.600   Min.   : 28.00   Min.   :     1305   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:1987   1st Qu.:104.0   1st Qu.:7.700   1st Qu.: 70.00   1st Qu.:  6153939   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
Median :2001   Median :120.0   Median :7.900   Median : 78.00   Median : 34700291   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
Mean   :1996   Mean   :123.7   Mean   :7.938   Mean   : 77.16   Mean   : 78379891   Mean   :0.1964   Mean   :0.2272   Mean   :0.1234   Mean   :0.08836
3rd Qu.:2010   3rd Qu.:136.0   3rd Qu.:8.100   3rd Qu.: 86.00   3rd Qu.:102308889   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
Max.   :2019   Max.   :238.0   Max.   :9.300   Max.   :100.00   Max.   :936662225   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
     Crime            Comedy           Drama           History            SciFi           Romance          Western           Fantasy
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.0000   Median :0.0000   Median :1.0000   Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
Mean   :0.1992   Mean   :0.2258   Mean   :0.6999   Mean   :0.05189   Mean   :0.07854   Mean   :0.1234   Mean   :0.02244   Mean   :0.07714
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
    Thriller           War             Mystery            Music            Horror            Family            Sport            FilmNoir
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.000000
Mean   :0.1388   Mean   :0.04067   Mean   :0.09818   Mean   :0.04909   Mean   :0.02525   Mean   :0.06031   Mean   :0.02384   Mean   :0.008415
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
    Musical
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.01543
3rd Qu.:0.00000
Max.   :1.00000
```

Determining number of cluster (Scree Plot)

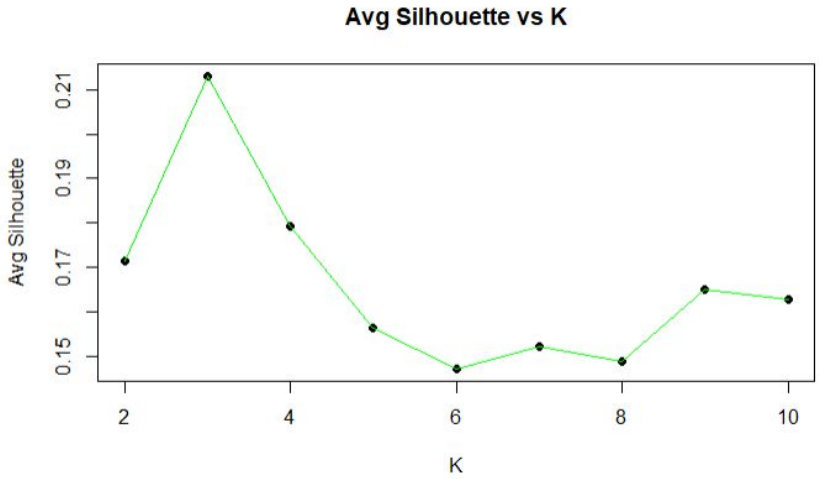
- k=2 to 6 all seems to be a pretty good option.
- k=3 seems to be the best here.
- Hard to be certain and objective
- We will use the Silhouette Statistic to help our decision.



Determining number of cluster (Silhouette Statistic)

- The average Silhouette score measure how well the object in each cluster matches the current clustering result.
- Higher average Silhouette score indicates better performance of clustering.
- Silhouette statistic also suggest k=3 to be the optimal number of cluster.

```
[1] "The average silhouette scores for K = 2 is 0.1713"  "The average silhouette scores for K = 3 is 0.2131"
[3] "The average silhouette scores for K = 4 is 0.1793"  "The average silhouette scores for K = 5 is 0.1563"
[5] "The average silhouette scores for K = 6 is 0.1471"  "The average silhouette scores for K = 7 is 0.152"
[7] "The average silhouette scores for K = 8 is 0.1488"  "The average silhouette scores for K = 9 is 0.165"
[9] "The average silhouette scores for K = 10 is 0.1628"
```



Stats about K Means Clusters

1

Released_Year	Runtime	IMDB_Rating	Meta_score	Gross
Min. :1930	Min. : 82.0	Min. :7.60	Min. : 58.00	Min. : 10177
1st Qu.:1960	1st Qu.:108.5	1st Qu.:7.90	1st Qu.: 80.00	1st Qu.: 5446437
Median :1973	Median :128.0	Median :8.10	Median : 87.00	Median : 23383987
Mean :1973	Mean :133.4	Mean :8.12	Mean : 86.37	Mean : 35248603
3rd Qu.:1987	3rd Qu.:151.5	3rd Qu.:8.30	3rd Qu.: 94.00	3rd Qu.: 46597035
Max. :2019	Max. :238.0	Max. :9.30	Max. :100.00	Max. :232906145

2

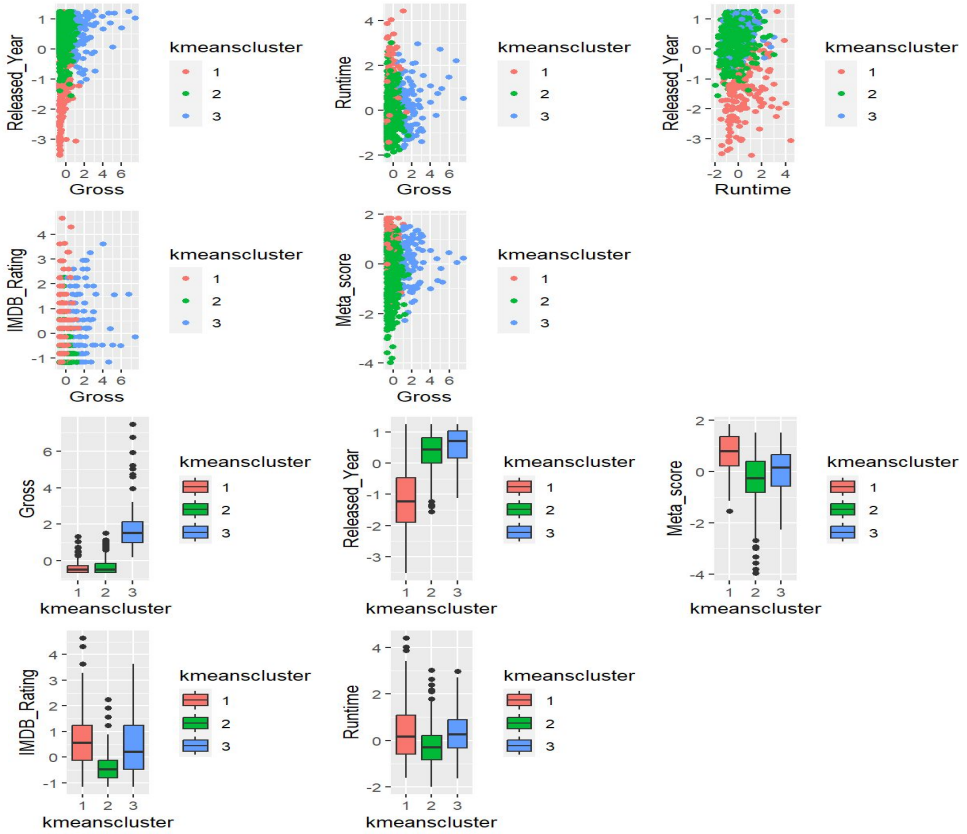
Released_Year	Runtime	IMDB_Rating	Meta_score	Gross
Min. :1967	Min. : 72.0	Min. :7.60	Min. :28.00	Min. : 1305
1st Qu.:1996	1st Qu.:102.0	1st Qu.:7.70	1st Qu.:67.00	1st Qu.: 4195902
Median :2004	Median :116.0	Median :7.80	Median :74.00	Median : 23345098
Mean :2003	Mean :117.5	Mean :7.82	Mean :73.08	Mean : 41379821
3rd Qu.:2011	3rd Qu.:129.0	3rd Qu.:7.90	3rd Qu.:82.00	3rd Qu.: 59852210
Max. :2019	Max. :202.0	Max. :8.60	Max. :96.00	Max. :251513985

3

Released_Year	Runtime	IMDB_Rating	Meta_score	Gross
Min. :1975	Min. : 81.0	Min. :7.600	Min. :49.00	Min. :100125643
1st Qu.:1999	1st Qu.:115.0	1st Qu.:7.800	1st Qu.:70.00	1st Qu.:191407502
Median :2009	Median :130.5	Median :8.000	Median :79.00	Median :252659101
Mean :2006	Mean :131.6	Mean :8.088	Mean :77.81	Mean :286723905
3rd Qu.:2015	3rd Qu.:147.0	3rd Qu.:8.300	3rd Qu.:85.25	3rd Qu.:323203039
Max. :2019	Max. :201.0	Max. :9.000	Max. :96.00	Max. :936662225

K-means clustering with k=3

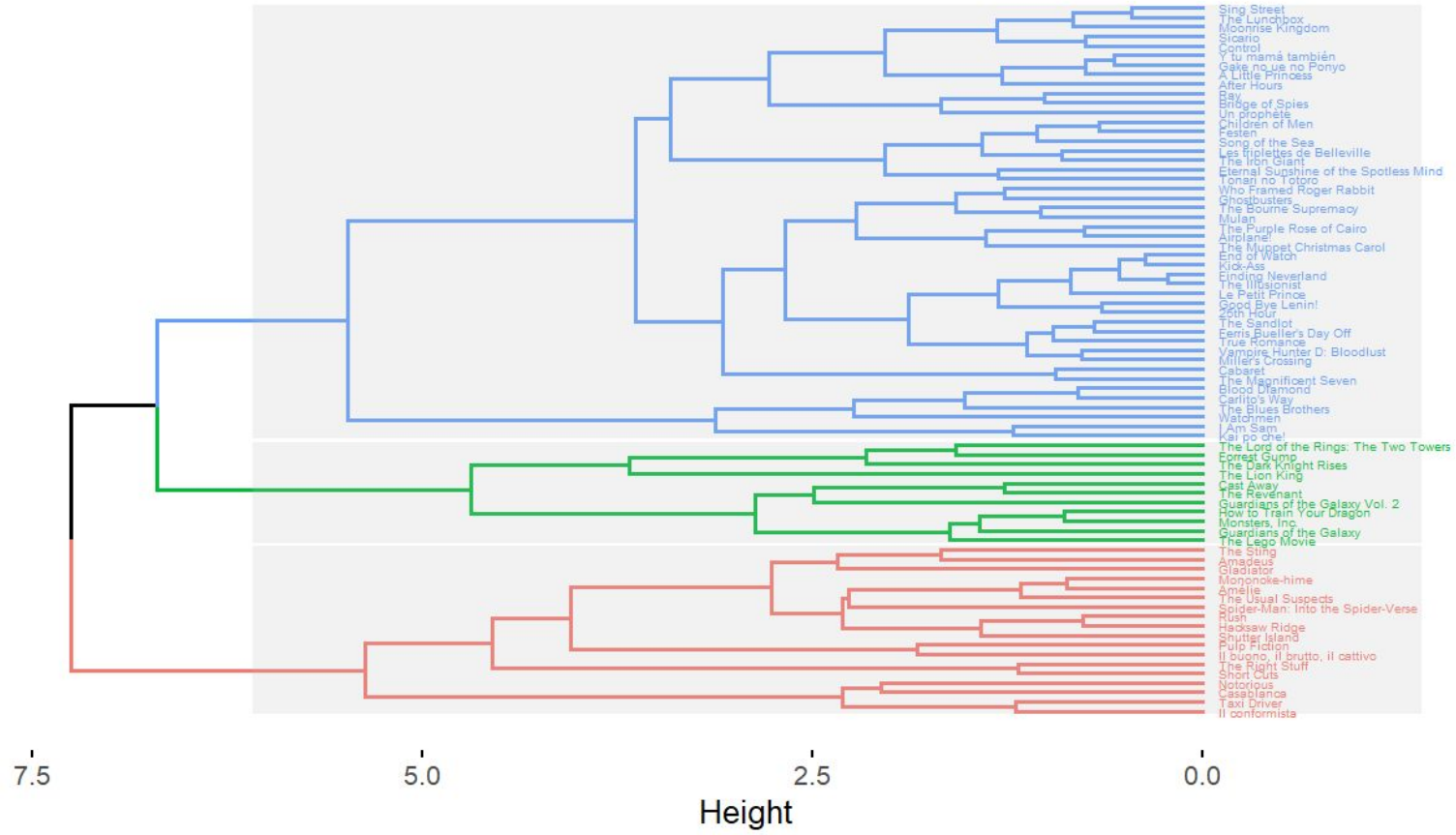
- Cluster 1 contains movies that are
  - Older
  - Higher rating
  - Low gross income
  - Long runtime
  - 74% of this cluster is Drama
- Cluster 2 contains movies that are
  - Relatively new
  - Lower rating
  - Low gross income
  - short runtime
  - 76% of this cluster is Drama
- Cluster 3 contains movies that are
  - Newer
  - High gross income
  - 66% of this cluster is Adventure



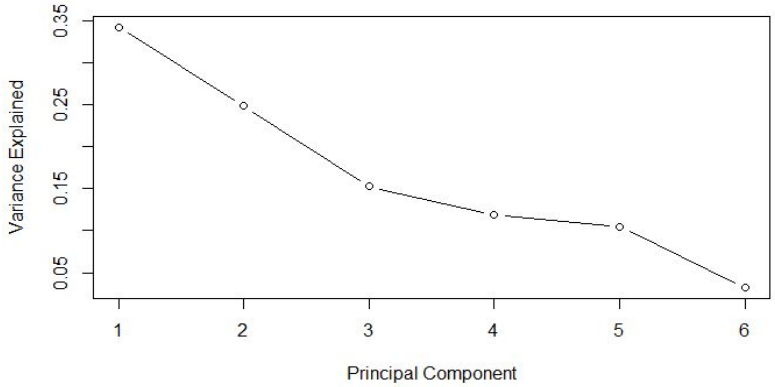
Movies in K Means Clusters

1	2	3
The Shawshank Redemption	La vita è bella	The Dark Knight
The Godfather	Whiplash	The Lord of the Rings: The Return of the King
The Godfather: Part II	The Intouchables	Inception
12 Angry Men	The Prestige	The Lord of the Rings: The Fellowship of the Ring
Pulp Fiction	American History X	Forrest Gump
Schindler's List	Léon	The Lord of the Rings: The Two Towers
Fight Club	Capharnaüm	The Matrix
Il buono, il brutto, il cattivo	Kimi no na wa.	Star Wars: Episode V - The Empire Strikes Back
Goodfellas	3 Idiots	Interstellar
One Flew Over the Cuckoo's Nest	Oldeuboi	Saving Private Ryan

Cluster Dendrogram



Principal Component Analysis



VarianceExplained <dbl>	SummationOfVarExplained <dbl>
0.34296273	0.3429627
0.24874222	0.5917049
0.15284221	0.7445472
0.11866335	0.8632105
0.10469427	0.9679048
0.03209523	1.0000000

Scree plot demonstrates that only 3 or 4 principal components are needed to explain a relatively large proportion of the variance in the dataset based on the 5 numerical variables

```
Standard deviations (1, .., p=5):
[1] 1.2395715 1.1568047 0.9382903 0.7972365 0.7805708

Rotation (n x k) = (5 x 5):
                     PC1        PC2         PC3         PC4        PC5
Released_Year  0.4880631  0.4417639 -0.25208172 -0.62759340  0.3304852
Runtime       -0.3010805  0.5112006  0.65381430  0.10421520  0.4579198
IMDB_Rating   -0.5993749  0.2606866 -0.02676844 -0.54696280 -0.5224054
Meta_score    -0.5584336 -0.1774663 -0.53837548 -0.02427944  0.6051611
Gross         -0.0068283  0.6663834 -0.46735101  0.54360641 -0.2048453
```

- Fairly New
- Low Ratings
- Shorter Runtime

- High Grossing
- Long Runtime
- Newer

- Long Runtime
- Low Grossing
- Low Rating
- Older

Top 10 Movies by PC Value

1

1	The Butterfly Effect
2	Seven Pounds
3	I Am Sam
4	Flipped
5	Jeux d'enfants
6	Saw
7	Kai po che!
8	Fear and Loathing in Las Vegas
9	Gifted
10	Tropa de Elite

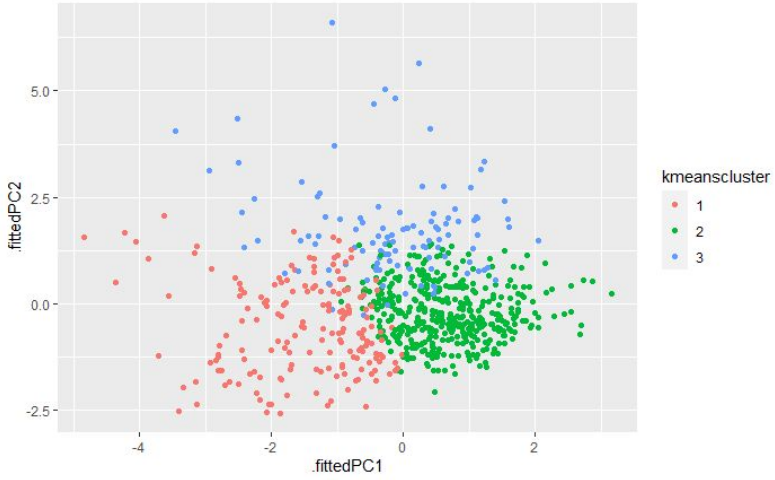
2

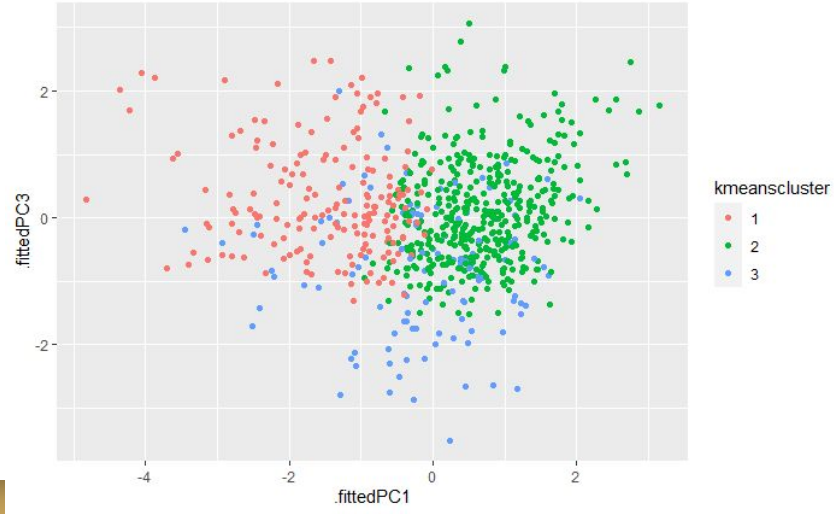
1	Avengers: Endgame
2	Star Wars: Episode VII - The Force Awakens
3	Avengers: Infinity War
4	Avatar
5	Titanic
6	The Dark Knight
7	The Avengers
8	The Lord of the Rings: The Return of the King
9	The Dark Knight Rises
10	Rogue One

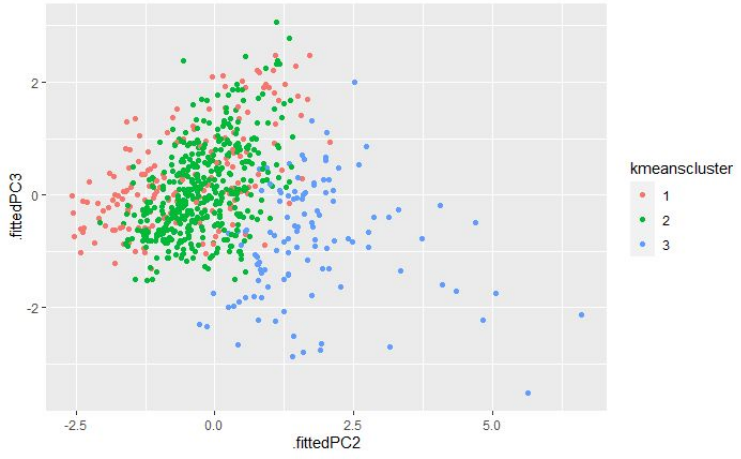
3

1	Bound by Honor
2	Kal Ho Naa Ho
3	Doctor Zhivago
4	Lagaan: Once Upon a Time in India
5	I Am Sam
6	Kelly's Heroes
7	Veer-Zaara
8	Malcolm X
9	My Name Is Khan
10	Dogville

K Means Clusters compared to Principal Components







Conclusion

- Data standardization was extremely important in this dataset, given the differing magnitudes of values based on variables/units
- By feature engineering the “Genre” variable so that it was more accessible, it dominated the clustering process
- Clustering based on only numerical values proved more effective, given that our K-Means clusters were well separated based on PCA