The objective is to know what we can conclude about the success of a movie based on its features. Thus, we need to know how we can judge a successful movie. From the given dataset, success measures can be based upon the popularity, generated revenues compared to the movie's budget or the movie's rating.

1. Preparation of data

- Generating a subset of only released movies to study only movies that have been released to have valid results of movies that have generated revues and been watched and thus rated

- Removing unimportant variables that will not help us in the analysis of successful movies

- Creating a subset of the data with only success indicator variables ("budget", "popularity", "revenue", "vote_average") to know what movies are most successful

- Normalize the data using min_max method

2. K-means Clustering

Different number of clusters were tested to know the number of clusters that will be used. After plotting the total within cluster sum of squares against the number of clusters as shown in figure 1, 6 clusters were found appropriate since increasing the no. of clusters above 6 will not cause a significant improvement in the total within cluster sum of squares.

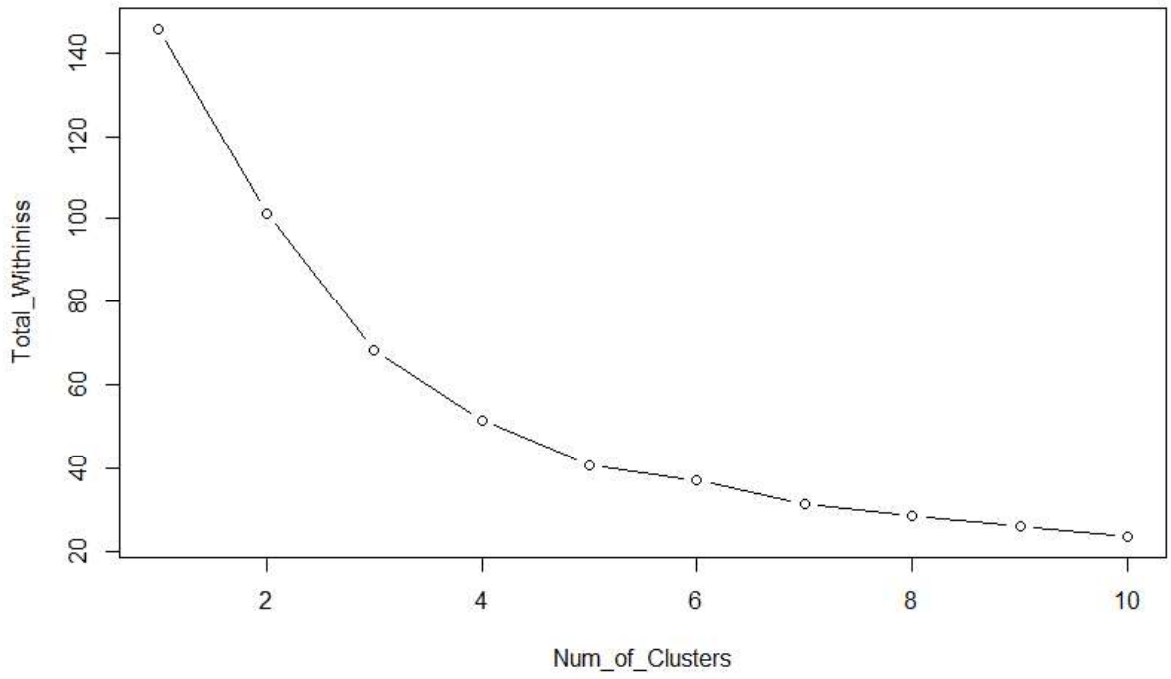


Figure 1: K-means Clustering

- Analyzing the clusters obtained, we found that:

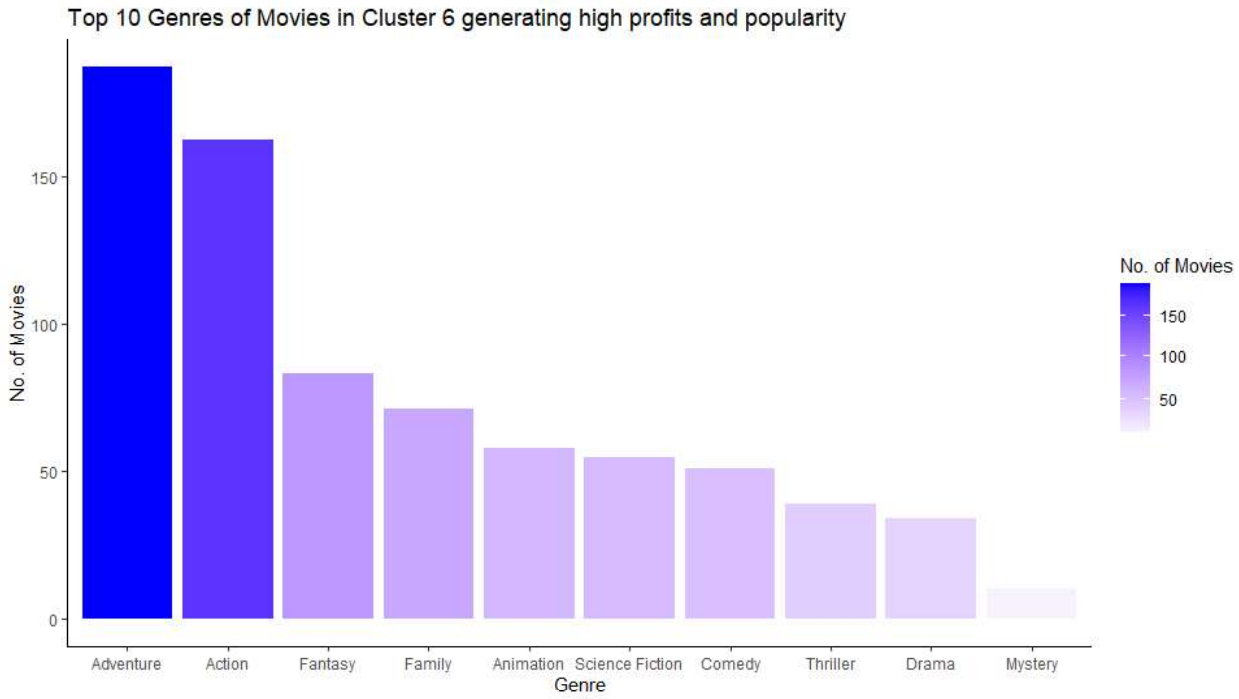
Cluster	1	2	3	4	5	6	7
Mean Budget	9,238,534	11,786,452	116,223	49,967,218	12,534,087	163,060,538	72,201,284
Mean Popularity	17.4	6.40	0.246	51.12	12.11	86.99	27.5
Mean Revenue	28,624,642	12,682,847	123,286	228,467,427	27,648,358	581,649,838	136,551,434
Mean Vote Average	7.03	4.45	0.282	7.10	5.89	6.55	5.83
Mean Profit	19,386,108	896,395	7,063	178,500,209	15,114,271	418,589,300	64,350,150
No. of Movies	1369	535	74	403	1578	223	613

From the above clusters of movies, it has been obvious that cluster 6 contains the movies generating the highest mean profits and having the highest mean popularity while cluster 4 contains movies having the highest mean vote average (rating). Thus, we will further investigate the movies in cluster 6 and cluster 1 to get an insight into the common features of these successful movies.

3. Analyzing Clusters

3.1 Analyzing Movies in ***cluster 6***:

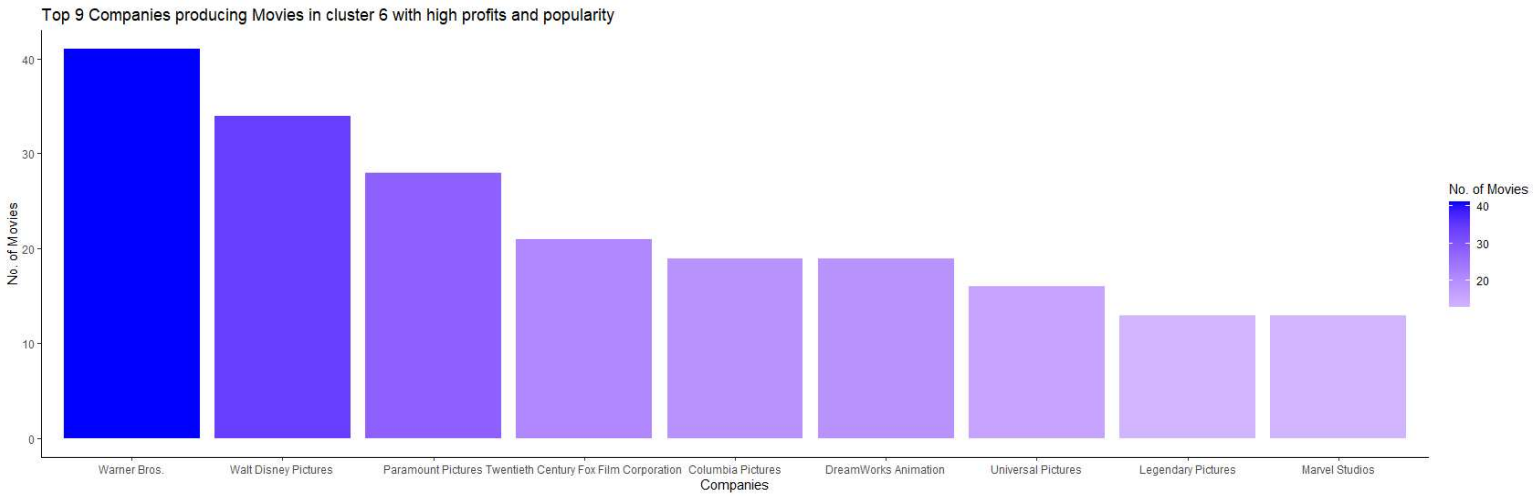
- The **Top 5 genres** generating movies with **high profits and high popularity** are:
    - Adventure
    - Action
    - Fantasy
    - Family
    - Animation



187 movies are Adventure (84% of movies in cluster 6), 162 are Action (73%), 83 movies are Fantasy (37%), 71 movies are Family (32%) and 58 movies are Animation (26%)

- The **top 5 companies** generating movies with **high profits and high popularity** are:

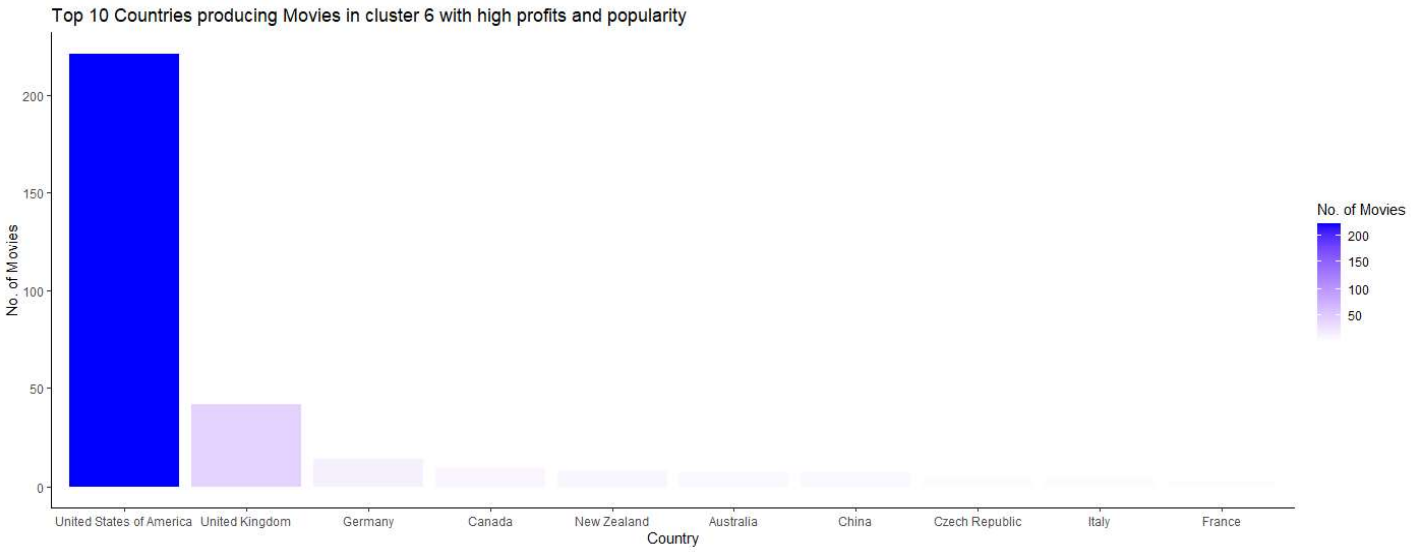
  - Warner Bros.
  - Walt Disney Pictures
  - Paramount Pictures
  - Twentieth Century Fox Film Corporation
  - Columbia Pictures



41 movies are generated by Warner Bros. (18% of movies in cluster 6), 34 movies are generated by Walt Disney Pictures (15%), 28 movies are generated by Paramount Pictures (13%), 21 movies are generated by Twentieth Century Fox Film Corporation (9%) and 19 movies are generated by Columbia Pictures (8.5%)

- the **top 2 countries** generating movies with **high profits and high popularity** are:

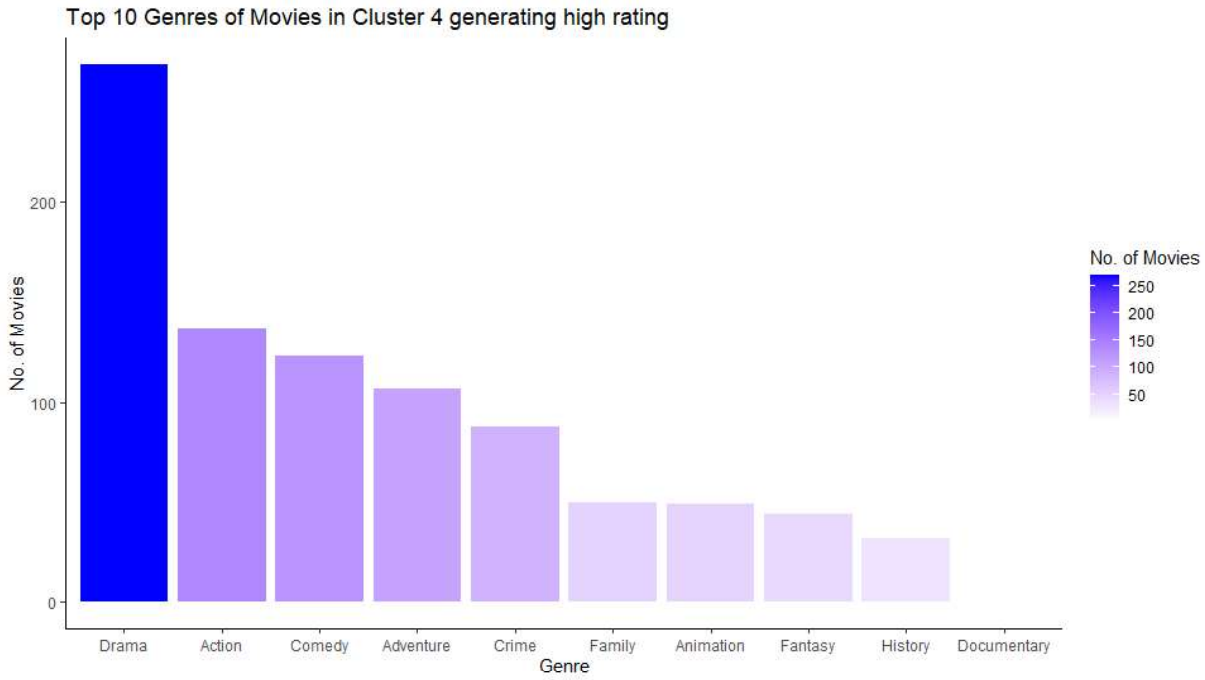
  - United States of America
  - United Kingdom



221 movies are generated by United States of America (99% of movies in cluster 6) and 42 movies are generated by United Kingdom (19%).

3.2 Analyzing movies in cluster 4:

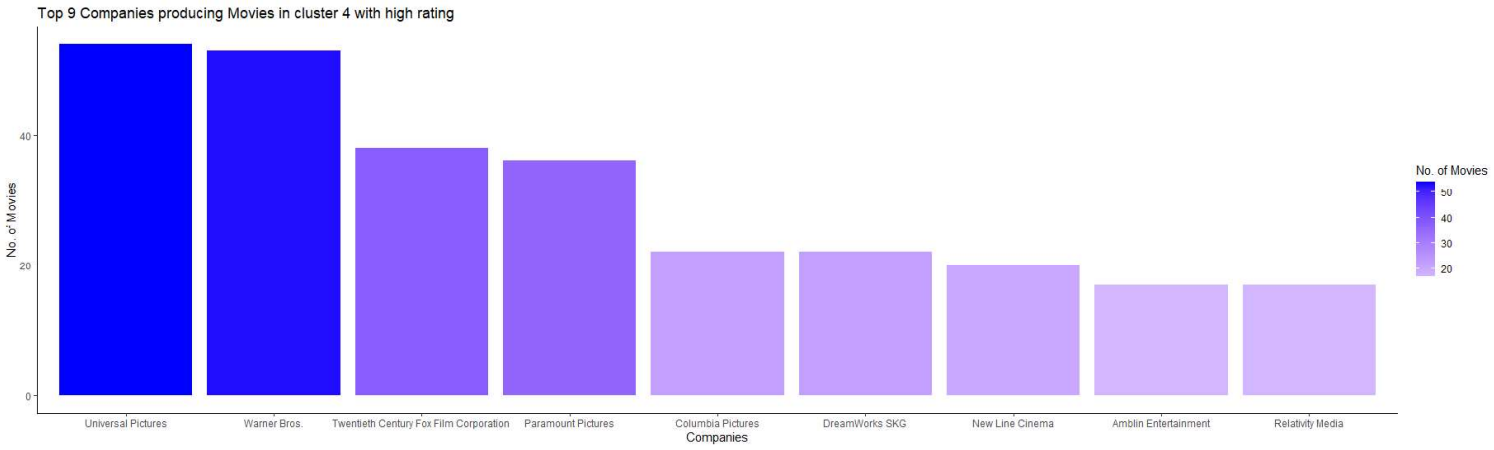
- The **top 5 genres** generating movies with **high rating**:
    - Drama
    - Action
    - Comedy
    - Thriller
    - Adventure



269 movies are Drama (67% of movies in cluster 4), 137 are Action (34%), 123 movies are Comedy (31%), 115 movies are Thriller (29%) and 107 movies are Adventure (27%)

- The **top 5 companies** generating movies with **high rating:**

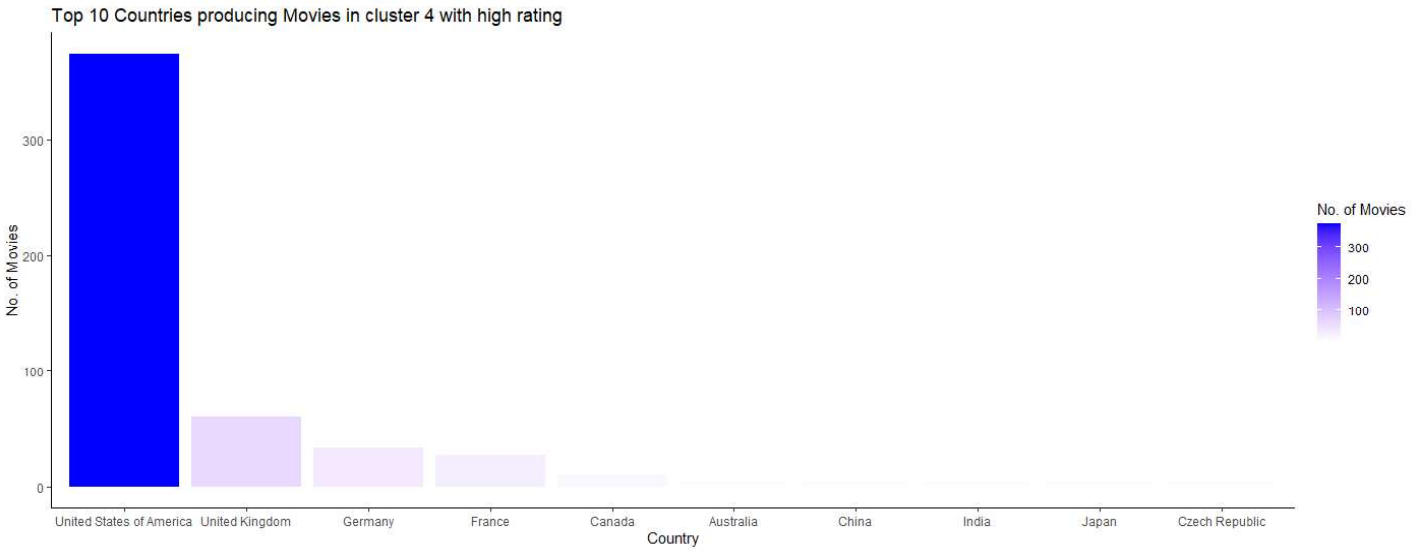
  - Universal Pictures
  - Warner Bros.
  - Twentieth Century Fox Film Corporation
  - Paramount Pictures
  - Columbia Pictures



54 movies are generated by Universal Pictures (13% of movies in cluster 4), 53 movies are generated by Warner Bros. (13%), 38 movies are generated by Twentieth Century Fox Film Corporation (9%), 36 movies are generated by Paramount Pictures (8.9%), and 22 movies are generated by Columbia Pictures (5.5%)

- The **top 2 countries** generating movies with **high rating**:

  - United States of America
  - United Kingdom



375 movies are generated by United States of America (93% of movies in cluster 4) and 61 movies are generated by United Kingdom (15%).

4. Insights

From the above analysis, the following insights can be drawn:

1. We can deduct whether a movie will be successful or not before it is released based on its features. If a movie's genre is adventure, action, fantasy, family and/or animation, it is more likely that this movie will generate high profits and it will be popular and If a movie's genre is Drama, action, comedy, thriller and/or adventure, it is more likely that this movie will be highly rated.

2. Top companies like Warner Bros., Walt Disney Pictures, Paramount Pictures, Twentieth, Century Fox Film Corporation and Columbia Pictures are more likely to produce movies that will generate high profits and will be popular

3. Top companies like Universal Pictures, Warner Bros, Twentieth, Century Fox Film Corporation, Paramount Pictures and Columbia Pictures are more likely to produce movies that will be highly rated

4. Movies that are generated by United States of America and/or United Kingdom are usually popular, highly rated and more likely to generate high profits