A Case Study on Netflix TV Shows and Movies Dataset Analysis

Unsupervised Learning and Evolutionary Computation using R

Abhipsa Roy (4038397), Allen Mundackal (4051168), Namit Joshi (4039882), Pranjal Parmar (4011744), Saurabh Palve (4036021) and Tanmay Mhatre (4037759)

University of Paderborn, Germany
{abhipsa2, allen, namit, pranjalp, palve, tmhatre}@uni-paderborn.de

1 Introduction

The results and analysis from the case study for the course "Unsupervised Learning and Evolutionary Computation Using R" are presented in this report. This study's main goal is to use methods from data analysis and unsupervised learning to examine and evaluate a dataset of TV series and films that are accessible through Netflix.
The study is divided into three distinct parts:

1. Basic Data Analysis
2. Outlier Detection and Statistical Testing
3. Clustering Analysis

The focus of the study has been on using statistical insights and visualizations to make insightful findings. Tables and supporting figures are used to gather and show the results in a methodical manner.

2 Methodology

Using methods from unsupervised learning and evolutionary optimization in R, this work analyzes a Netflix dataset that includes details on TV series and films. Three sections make up the methodology, which reflects the organization of the analysis:

1. **Data Analysis:**
   Using R's `tidyverse` package, a simple exploratory data analysis was performed. To find the best action series, examine countries of production, and rank actors according to their IMDb ratings, key insights were obtained by modifying and filtering the dataset.

2. **Outlier Detection and Statistical Testing:**
   Visualization tools like box plots and histograms were used to examine the dataset for anomalies. The distribution of scores was evaluated using normality tests, and IMDb and TMDB ratings were compared using statistical tests. To look into actor-related data and find notable deviations, robust outlier analysis approaches were used.

3. **Clustering:**
   To find patterns and groupings in the dataset, clustering methods were used. Scaling numerical features and encoding categorical variables were preprocessing processes. Cluster formations were shown using dimensionality reduction techniques like Principal Component Analysis (PCA), and clustering performance was assessed subjectively.

A combination of statistical and visual techniques was used throughout the analysis to guarantee the reliability and robustness of the results. The main implementation tools were the R programming language and associated libraries.

3 Results and Discussion

3.1 Data Analysis

Task 1: Create a table that shows all action shows with an IMDb rating larger than 8.9. Show them in descending order.

id	title	release_year	imdb_score	genres
ts3371	Avatar: The Last Airbender	2005	9.3	['scifi', 'animation', 'action', 'family', 'fantasy']
ts32835	Hunter x Hunter	2011	9.0	['action', 'animation', 'comedy', 'fantasy']
ts20682	Attack on Titan	2013	9.0	['action', 'scifi', 'animation', 'horror', 'drama', 'fantasy']
ts222333	Arcane	2021	9.0	['scifi', 'action', 'drama', 'animation', 'fantasy']

Figure 1: Action shows with IMDb ratings greater than 8.9.

Task 2: Pick 2 different countries that are present in the production_countries column of the titles data table. For each chosen country, find the 2 best-rated movies according to the IMDb score column and show them in descending order. In case of ties, assign the ranks manually. Ignore any co-productions, that is any movies that have more than a single country involved in the production.

For this task, the two countries chosen were USA and Great Britain.

id	title	genres	production_countries	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
tm122434	Forrest Gump	['drama', 'romance']	['US']	tt0109830	8.8	2021343	63.449	8.478
tm155787	GoodFellas	['drama', 'crime']	['US']	tt0099685	8.7	1131681	50.387	8.463

Figure 2: Top 2 Movies Produced in the USA Based on IMDb Score

id	title	genres	production_countries	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_scor
tm853783	David Attenborough: A Life on Our Planet	['documentation']	['GB']	tt11989890	8.9	31625	15.935	8.5
tm188970	Bill Hicks: Revelations	['comedy', 'documentation']	['GB']	tt0152183	8.5	3098	3.784	8.1

Figure 3: Top 2 Movies Produced in Great Britain Based on IMDb Score

Task 3: Pick a movie genre. For the chosen genre, show the top 3 actors by the number of appearances, sorted discerningly.

name	appearances
Boman Irani	15
Fred Armisen	15
Kareena Kapoor Khan	14

Figure 4: Top 3 Actors in Comedy Movies by Number of Appearances

Task 4: For each actor, calculate their average IMDb score, and report on the top 3 best and worst actors.

name	average_imdb_score
Anna Gunn	9.5
Cricket Leigh	9.3
Jessie Flower	9.3

Figure 5: Top 3 Best Actors Based on Average IMDb Score

name	average_imdb_score
Abeer Mohammed	1.5
Ana Druzhynina	1.5
Derrik Sweeny	1.5

Figure 6: Top 3 Worst Actors Based on Average IMDb Score

3.2 Outlier Detection and Statistical Testing

Task 1: Pick two genres, and perform an analysis of their IMDb scores. Check for normality and whether there are any shows or movies that show up as outliers. Report your results both in text and using appropriate visualizations. Also consider if there are any other attributes in the data that you could use for a more robust analysis.

Shapiro-Wilk Normality Test for Drama
data: drama_data_imdb$imdb_score
W = 0.97453, p-value < 2.2e-16

Shapiro-Wilk Normality Test for Comedy
data: comedy_data_imdb$imdb_score
W = 0.98271, p-value = 9.099e-16

Thus IMDb score is not normally distributed. The boxplots support the numbers appropriately.

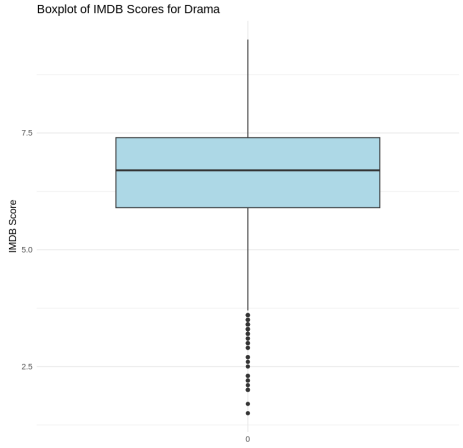

Figure 7: Boxplot of IMDB Scores for Drama

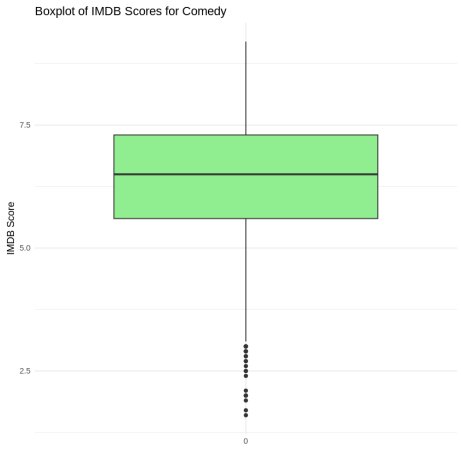

Figure 8: Boxplot of IMDB Scores for Comedy

The list of outliers for both Drama and Comedy genres fetched from the data is as follows:

1223	Day of the Dead: Bloodline	3.4
1300	An Imperfect Murder	3.2
1369	SPF-18	3.2
1380	Nothing to Lose	2.3
1490	A House of Blocks	2.3
1511	Frat Star	3.6
1528	B. A. Pass 2	2.2
1631	365 Days	3.3
1791	Cuties	3.4
1943	Kolaiyuthir Kaalam	2.6
1951	Drive	3.1
1984	Sin City	3.1
1988	Bulletproof 2	3.5
2069	Jinn	3.4
2087	Indoo Ki Jawani	3.0
2128	Shikara	3.3
2145	The App	2.7
2181	Kaali Khuhi	3.5
2182	Nothing to Lose 2	3.0
2193	90 ML	3.3

Figure 9: Drama Outliers

343	FRED 3: Camp Fred	2.0
487	Himmatwala	1.7
497	Grandmother's Farm	2.9
606	Richie Rich	3.0
691	Kyaa Kool Hain Hum 3	1.9
724	Santa Banta Pvt Ltd	2.7
885	Bonus Family	2.9
1348	Jiu Jitsu	2.9
1418	How High 2	3.0
1450	Until Dawn	2.4
1457	Me Against You: Mr. S's Vendetta	1.6
1473	Ni de coña	2.8
1495	Holiday on Mars	2.7
1621	Indoo Ki Jawani	3.0
1695	Luccas Neto in: Summer Camp	2.8
1732	Luccas Neto em: Uma Babá Muito Esquisita	3.0
1779	Luccas Neto in: Children's Day	2.5
2024	Thomas & Friends: All Engines Go!	2.0
2118	Sex: Unzipped	2.5
2158	He's Expecting	2.0

Figure 10: Comedy Outliers

Task 2: Try doing the same analysis, but using the TMDB scores instead. Is there a difference between the two websites? Confirm this by choosing an appropriate statistical test to determine if there is a statistically significant difference between their distributions. Remember to remove outliers if the statistical test you are using is affected by them.

Shapiro-Wilk Normality Test for Drama
data: drama_data_tmdb$tmdb_score
W = 0.97329, p-value < 2.2e-16

Shapiro-Wilk Normality Test for Comedy
data: comedy_data_tmdb$tmdb_score
W = 0.97605, p-value < 2.2e-16

Thus TMDB score is not normally distributed. The histogram and boxplot below show the acute difference between the IMDb and TMDb scores of the movies of Drama and Comedy genres.

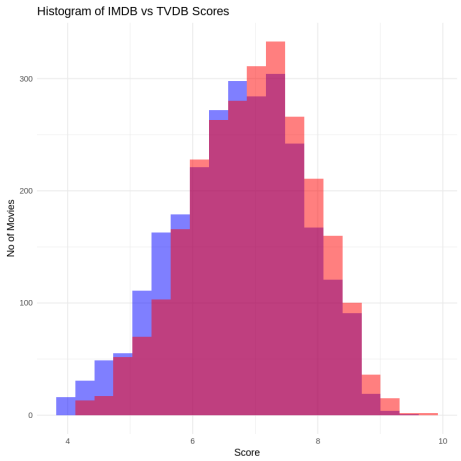

Figure 11: Histogram of IMDB vs TMDB Scores

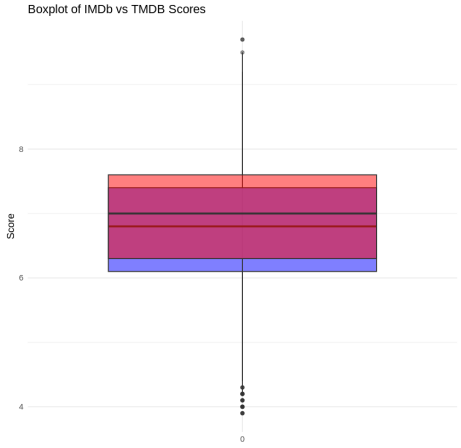

Figure 12: Boxplot of IMDb vs TMDB Scores

Task 3: Perform your own outlier analysis using the data for the actors. Feel free to examine the data in any way you think is appropriate, and report on any interesting findings.

name	num_appearances	avg_imdb_score
Anna Gunn	1	9.50
D.C. Young Fly	2	3.65
Grant S. Johnson	2	3.60
Joey Bragg	2	3.60
Paola Minaccioni	2	3.60
Turlough Convery	2	3.60
Abeer Sabry	1	3.60
Ahmed Safwat	1	3.60
Ahmed Taha	1	3.60
Albert P. Santos	1	3.60

Figure 13: Top Outliers according to IMDb score

The boxplot reveals that the average IMDb score of most performers falls between 5.5 and 7.5, with a few low scorers (3.6) and significant outliers like Anna Gunn (9.5). Due to exceptional performances or low ratings, actors with few appearances, like Anna Gunn, may have skewed averages. Consistent parts in low-rated content are highlighted by actors with low scores, such as D.C. Young Fly.

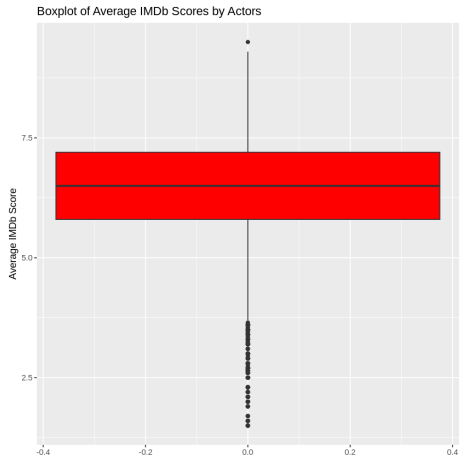

Figure 14: Boxplot of Average IMDb Scores by Actors

Task 4: For all the tasks in this part, try to also consider real-life context. For example, do the results of your analysis match what you would expect given the data? Can you find explanations for the outliers?

1. Distribution of IMDB and TMDB Scores

Expectation:

Most movies and TV shows will have scores close to the average (like 5-8 out of 10). Only a few will get really high (9+) or really low (below 3) scores. Popular or critically acclaimed movies should have higher ratings, while poorly received ones will have lower ratings.

Real-life Example:

A high-rated movie like *The Shawshank Redemption* (IMDB score: 9.3) is an outlier because it's a universally loved classic. A low-rated movie like *Disaster Movie* (IMDB score: 2.0) is an outlier

because it was widely criticized as one of the worst films ever made

2. Outliers in IMDB and TMDB Scores
Expectation:
Outliers happen when movies or shows are extremely well-received or heavily criticized. High outliers might be due to strong performances, storytelling, or fan followings. Low outliers could result from poor execution, niche audiences, or controversy.

Real-life Example:
A high outlier like *Avengers: Endgame* (IMDB: 8.4, TMDB: 8.3) reflects its massive popularity and box-office success. A low outlier like *Cats* (2019) (IMDB: 2.8, TMDB: 3.2) reflects how bad CGI and critical reviews led to terrible ratings.

3. Difference Between IMDB and TMDB Scores
Expectation:
IMDB and TMDB scores might differ because the audiences are different. IMDB has more general users, while TMDB could have more movie enthusiasts. IMDB scores might be more spread out because casual users tend to give extreme ratings (10 for favorites, 1 for disliked ones).

Real-life Example:
Joker (2019): IMDB score: 8.4, TMDB score: 8.3. Both scores are high because it was a critically and commercially successful movie. *Birdemic: Shock and Terror* (IMDB: 1.8, TMDB: 2.1): The scores are low on both platforms, but slightly higher on TMDB because niche movie enthusiasts might appreciate it as a "so bad it's good" cult classic.

4. Cultural or Contextual Influences
Expectation:
Cultural and contextual factors might affect ratings. Some genres, like Comedy, may have more mixed ratings because humor is subjective. Similarly, movies from non-English-speaking countries may have niche appeal but lower scores due to fewer reviews.

Real-life Example:
Parasite (2019): Despite being a non-English movie, it became a global hit, scoring 8.5 on IMDB, showing how quality transcends language barriers.
A niche Bollywood comedy might score high on TMDB (where specific fans rate it) but lower on IMDB due to a lack of international appeal.

3.3 Clustering

Task 1: Try to see if there are any attributes or combinations of attributes that give you clearly defined clusters. Do these clusters correspond to any attributes present in the data? For example, can you cluster movies based on their release dates?

For this task, the selected features for clustering algorithm were "runtime" and "release_year". These features were normalized and elbow method was practised to find out the value of k for k-means clustering.

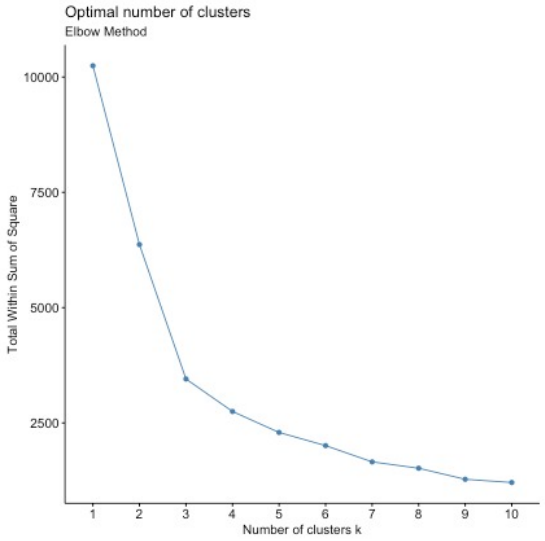

Figure 15: Elbow Plot for Task 1

As the elbow is noticed at K-value 3, there were three number of clusters. After applying K-means clustering, the following interactive plot was observed:

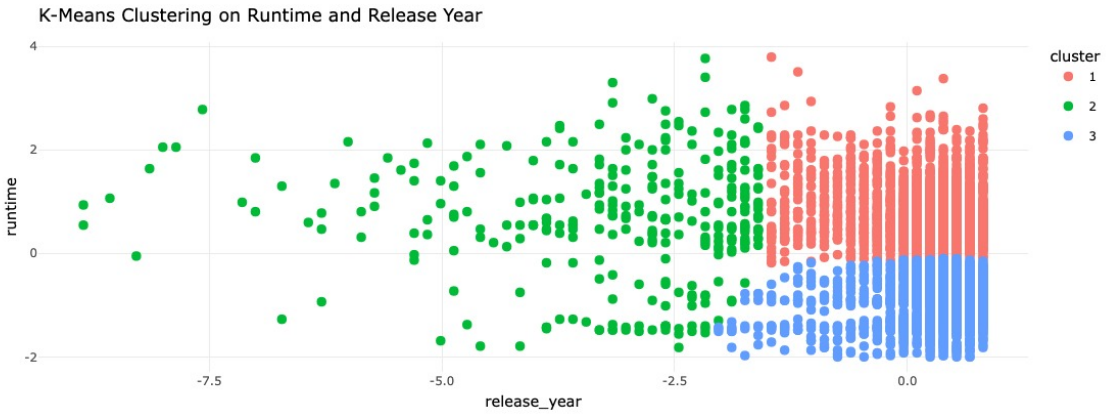

Figure 16: K-means Clustering for Task 1

From the clustering analysis visualized in the plot, here are some specific inferences:

Cluster 1 (Red): This cluster seems to represent movies released more recently (right side of the graph). These movies tend to have a more consistent and shorter runtime compared to others. It could indicate a trend where newer movies have standardized runtimes, possibly catering to audience preferences for shorter content.

Cluster 2 (Green): This cluster spans across a wide range of release years and runtimes. It appears to include older movies as well as some more recent ones. The broader spread suggests a diverse mix of runtime durations in different eras of filmmaking.

Cluster 3 (Blue): This cluster primarily focuses on more recent movies, similar to Cluster 1, but with noticeably shorter runtimes. These could represent modern movies or specific genres that favor brevity, such as short films or certain contemporary productions

There is a clear temporal distinction, with newer movies grouped in Clusters 1 and 3, highlighting potential changes in production trends over time. Older movies in Cluster 2 exhibit a wider variation in runtime, potentially reflecting greater experimentation in movie lengths in earlier eras. These clusters could be useful for segmenting movies based on release periods and runtime, enabling analysis of evolving trends in the film industry.

Task 2: Try to combine the data in the dataset in order to create novel attributes. For example, you could, for each movie, calculate the number of actors that are present in it (perhaps more recent movies have a larger cast size?).

To add new attributes, "actor_count", "engagement_score" and "genre_count" were selected. After handling the missing values, "actor_count" and "engagement_score" were selected for clustering. Then they were normalised and elbow method was practised. Like Task 1, k-value was observed as 3 here too. After performing clustering the following plot was observed:

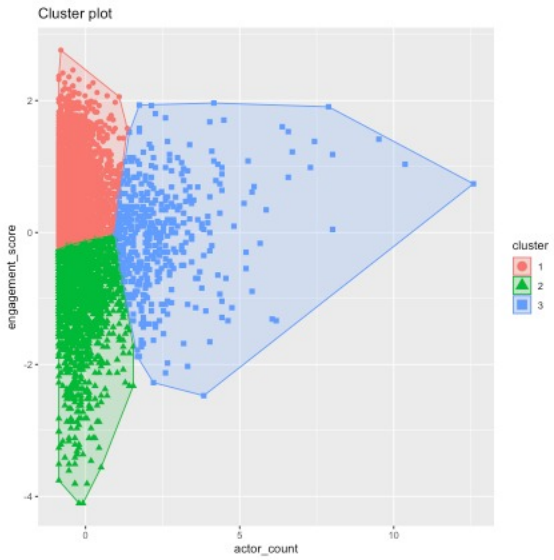

Figure 17: K-means Clustering for Task 2

It can be inferred from this plot that the higher the actor count, the higher is the average engagement score and the lower the actor count, more varied engagement score is observed.

Task 3: This dataset contains a large number of categorical features that cannot be used in a straightforward way by clustering algorithms such as k-means. Think of how you could still incorporate these attributes into your analysis. Some example techniques you could use include one-hot and multi-hot encoding, distance metrics that work with mixed data (such as Gower distance), dimensionality reduction techniques such as FAMD, or clustering algorithms adapted for mixed data (k-prototypes).

For Task 3, one hot encoding was performed on Genres, Countries and Age Certification and Genres seemed fit for clustering. It was normalised after using the elbow method, no definitve elbow was seen. Thus, the k-value was decided as 19 which is the total number of genres.

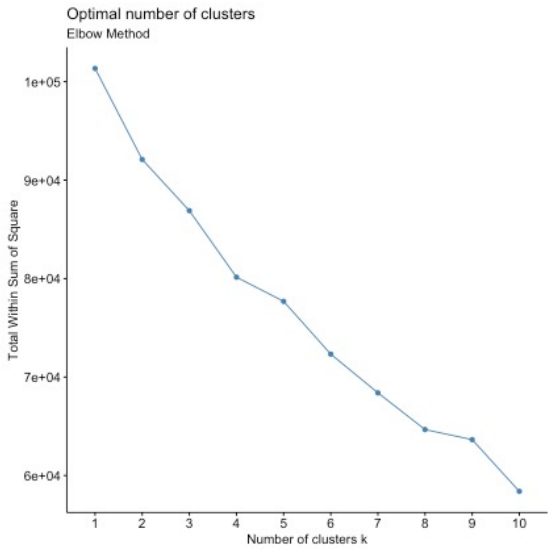

Figure 18: Elbow Plot for Task 3

After performing clustering, it was observed that using 19 clusters (corresponding to the number of genres) allows for potential genre-specific clusters.

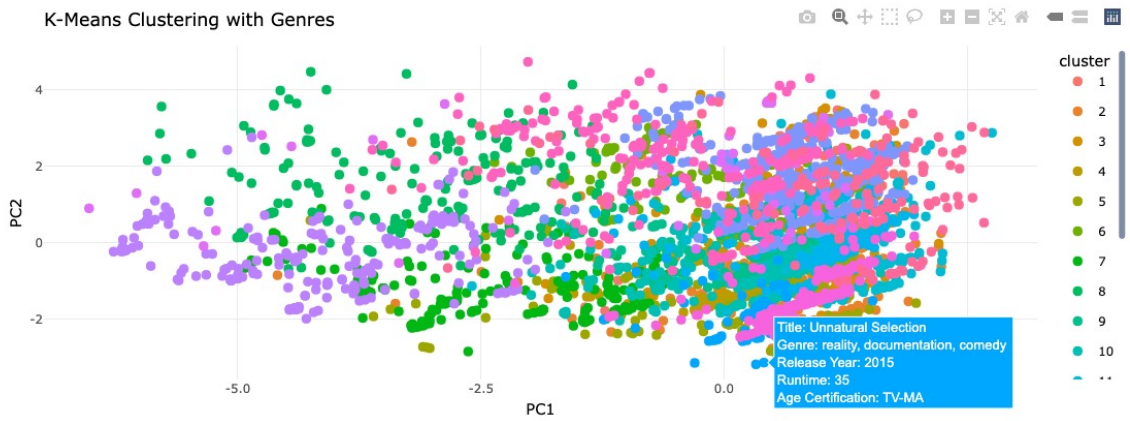

Figure 19: K-means Clustering for Task 3

However, the plot doesn't reveal distinct, well-separated clusters for each genre. This could indicate that:

Genres often overlap: Movies often fall into multiple genres, making it difficult to create pure genre-based clusters. Other factors play a role: Factors beyond genre, such as release year, runtime, and age certification, likely influence the clustering.

Relationship between Clusters: The relative positions of the clusters suggest potential relationships between them. For instance, clusters that are close to each other might contain movies with similar characteristics beyond their genres.

4 Conclusion

In this case study, we conducted a comprehensive analysis of Netflix's movie and TV show dataset, addressing tasks related to data analysis, outlier detection, and clustering.

Data Analysis:
Using `tidyverse` methods, we investigated important aspects of the dataset and found patterns in actor contributions, ratings, and genres. Finding the best-rated films by genre and nation, as well as the performers that appeared most frequently in a particular genre, were examples of specific insights. This preliminary research served as a strong basis for further investigation.

Outlier Detection:
We looked at IMDb and TMDb scores using statistical techniques and visualizations to find outliers within particular categories. High-scoring anomalies like *The Shawshank Redemption* and low-scoring examples like *Disaster Movie* were among the noteworthy findings. Subtle variations in audience preferences were found when comparing IMDb and TMDb scores, and statistical testing confirmed this finding. Furthermore, significant performers who consistently received above-average or below-average scores were identified through actor-level outlier analysis.

Clustering:
The clustering analysis revealed distinct patterns within the movie dataset. By experimenting with various attributes and encoding techniques, we identified clusters that reflect temporal trends, genre influences, and relationships between different movie characteristics. These findings offer valuable insights into the evolution of the film industry, audience preferences, and potential avenues for targeted recommendations and content strategies.

Final Thoughts:
This exercise showed how statistical methods and unsupervised learning may be applied practically to real-world datasets. Through the integration of computational techniques and domain knowledge, we were able to extract actionable insights and test our findings against expectations. The findings demonstrate the value of integrating quantitative analysis with contextual knowledge in decision-making processes and demonstrate the diversity of the Netflix library.

Statement of Contributions

Abhipsa Roy- Data Analysis and Outlier Detection and Statistical Testing.
Allen Mundackal- Data Analysis and Outlier Detection and Statistical Testing.
Namit Joshi- Clustering.
Pranjal Parmar- Outlier Detection and Statistical Testing.
Saurabh Palve- LaTeX Documentation and Clustering.
Tanmay Mhatre- Data Analysis and Documentation.