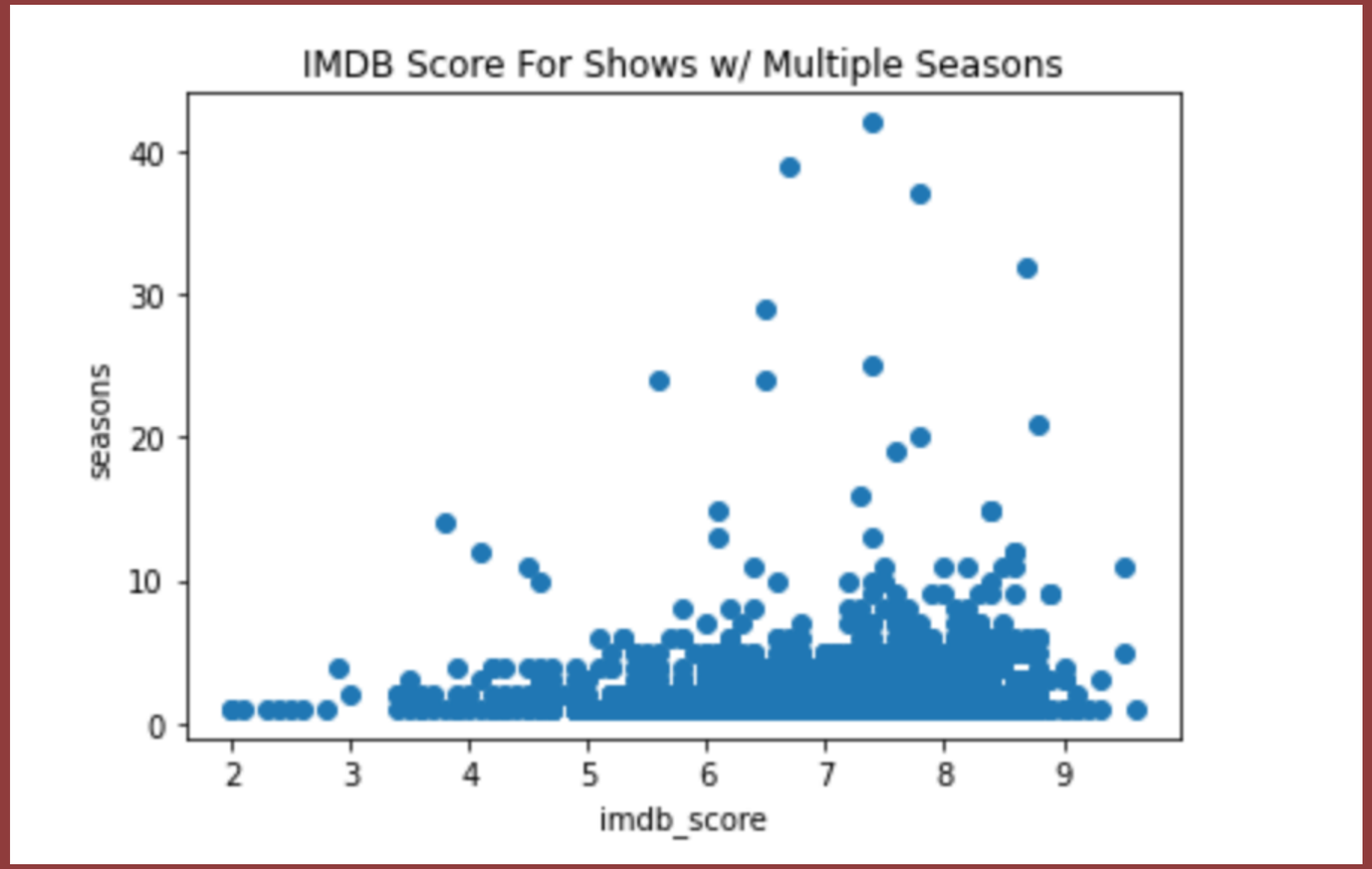
DATA
VISUALS & PREDICTIVE MODELING

SCATTER PLOT

- Correlation of Seasons and Ratings: The scatter plot examines the relationship between the number of seasons of a show and its IMDb score, indicating that shows with higher ratings tend to have more seasons.

- Concentration of Higher Scores: A dense cluster of points shows that a majority of shows with multiple seasons have IMDb scores in the range of 6 to 8.

- Longevity vs. Quality: There's a visible trend where shows with the highest number of seasons do not necessarily have the highest IMDb scores, suggesting that factors other than just rating quality may contribute to a show’s longevity.

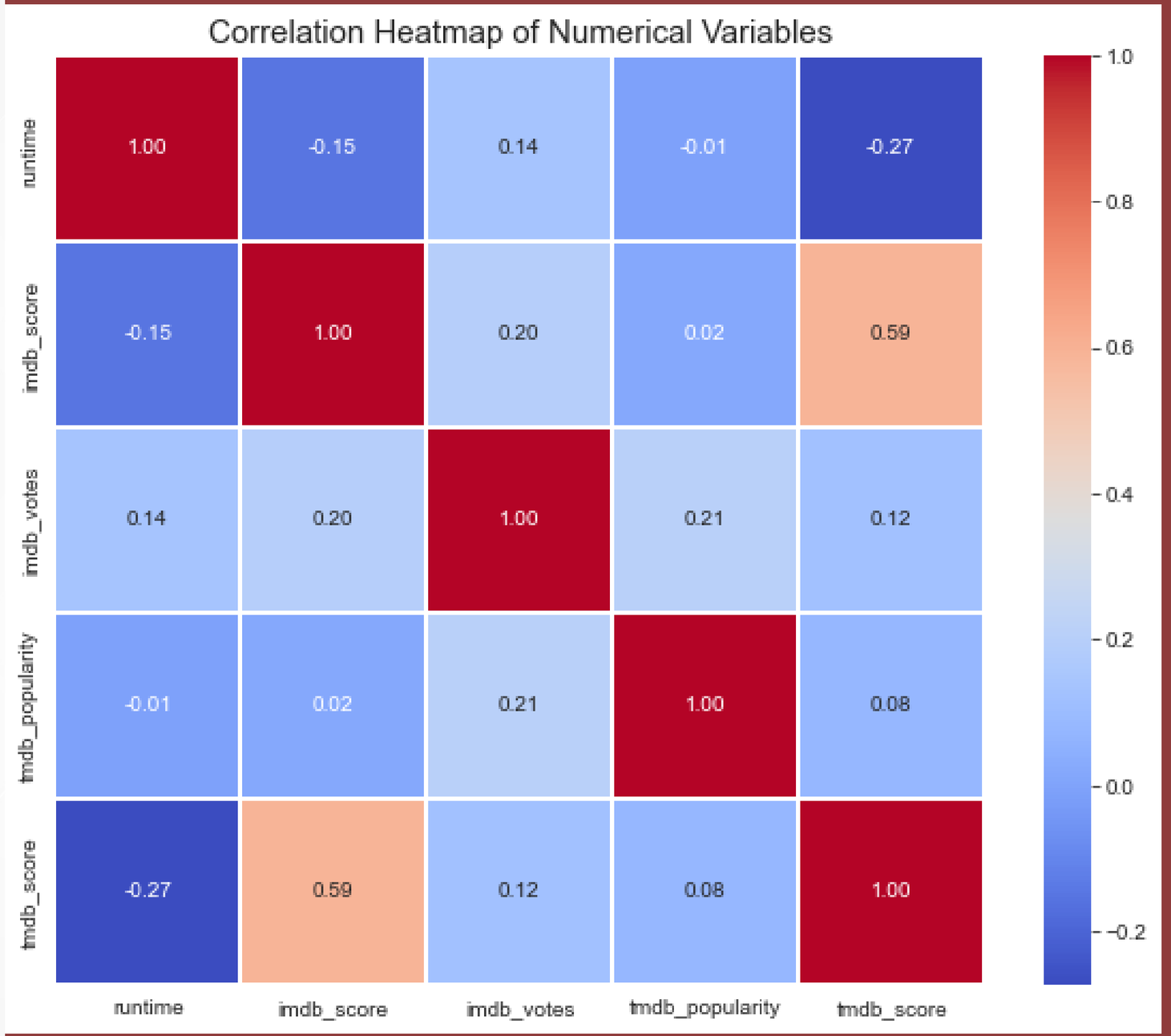


HEATMAP

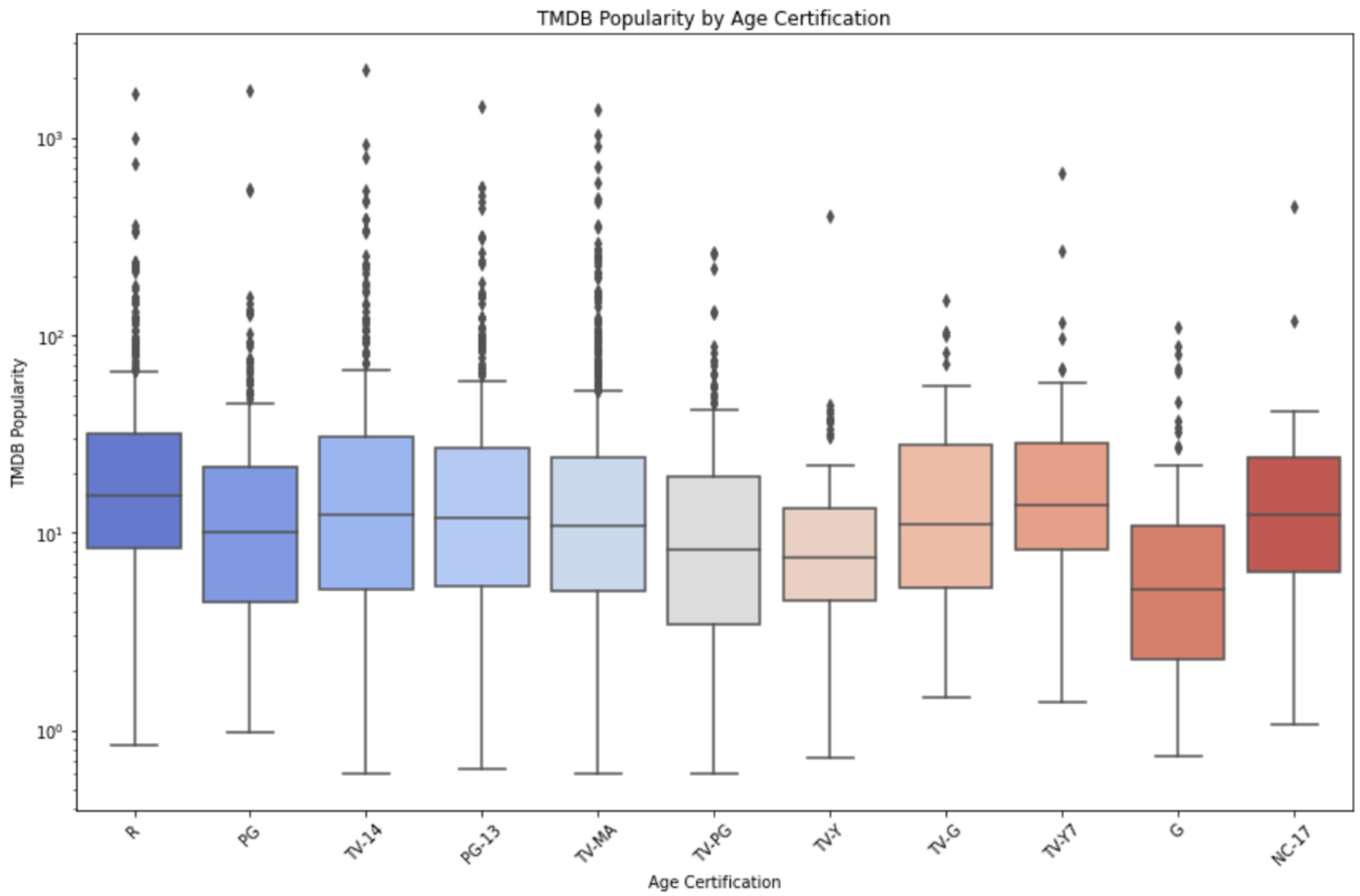
- Analytical Foundation: Constructed to discern correlations among Netflix metrics, the heatmap pinpoints how variables like ratings and runtime interlink, aiding predictive model accuracy.

- Consistent Ratings Correlation (0.59): The data shows titles with high IMDb scores typically also enjoy high TMDb scores,confirming cross-platform rating consistency.

- Runtime Influence (-0.15 to -0.27): The negative correlation with runtime suggests titles beyond a certain length tend to see a dip in ratings, a crucial consideration for content curation



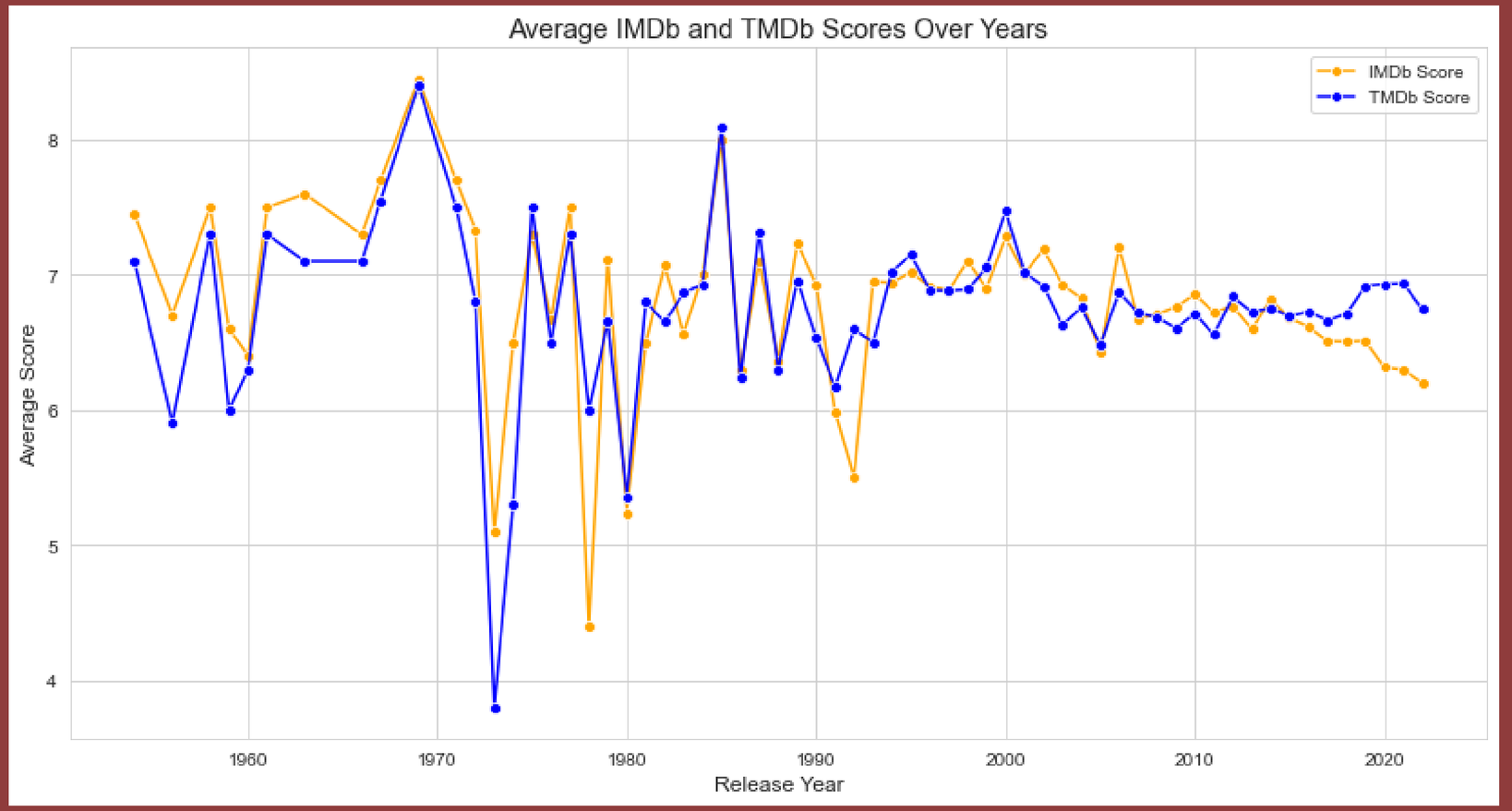
BOXPLOT



- Popularity Distribution: The box plot categorizes TMDb popularity by age certification, showing variability in content popularity with a notably wide range for 'R' rated titles.

- Age Certification Impact: Certifications 'R', 'PG', and 'TV-14' exhibit higher median popularity scores, suggesting these categories may align well with the preferences of TMDb's audience demographic.

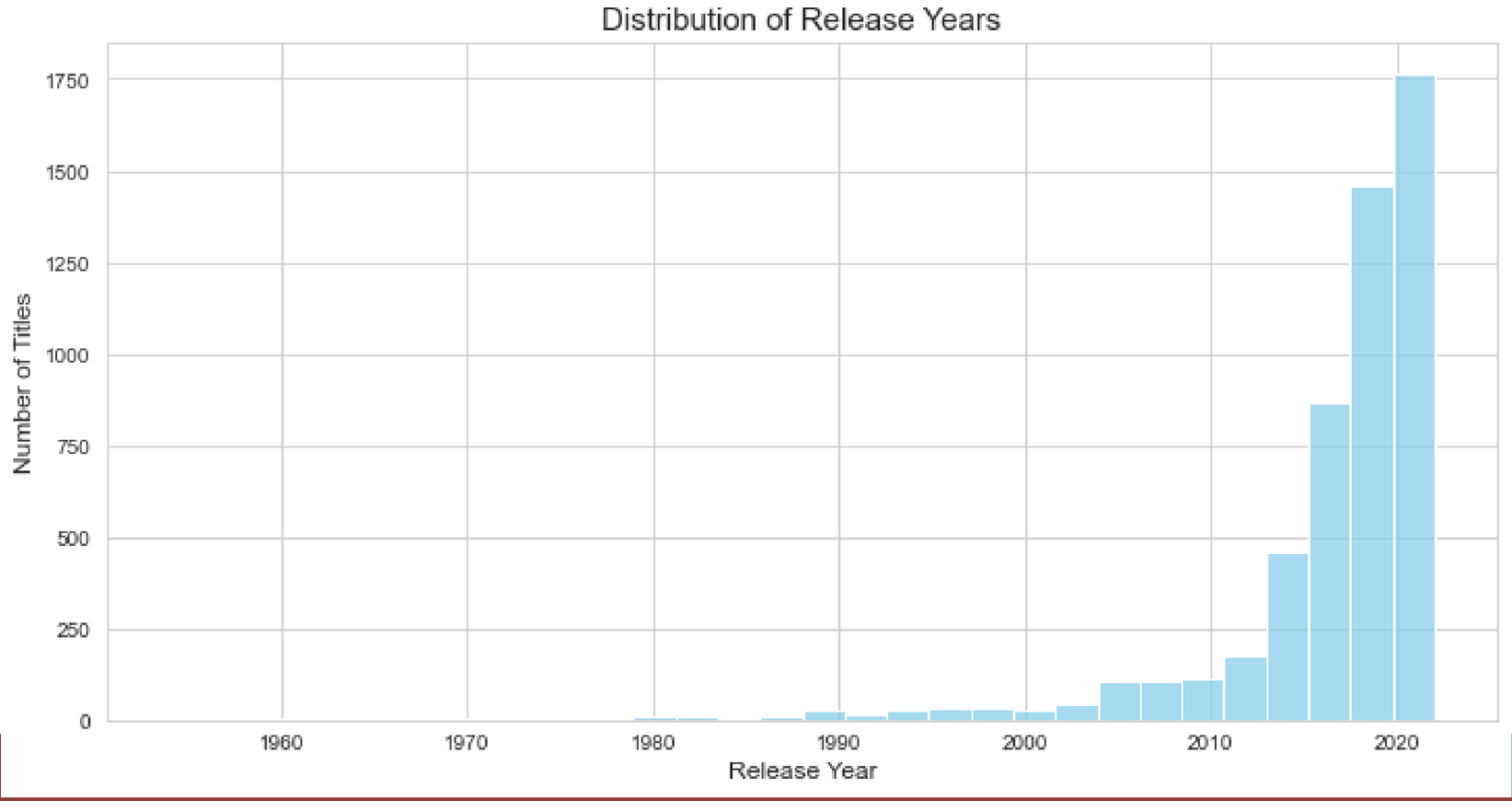
- Outliers and Range: Several age categories display outliers indicating titles with exceptional popularity, with 'R' and 'TV-MA' showing particularly high ranges, pointing to the existence of both highly popular and less popular titles within these ratings.



- Trend Analysis Over Time: The graph illustrates the trend of average IMDb and TMDb scores from 1960 to 2020, showing score fluctuations over different release years.

- Score Consistency Between Platforms: Both IMDb and TMDb scores follow a similar trend over the years, with no significant divergence, highlighting a consistent audience perception across platforms.

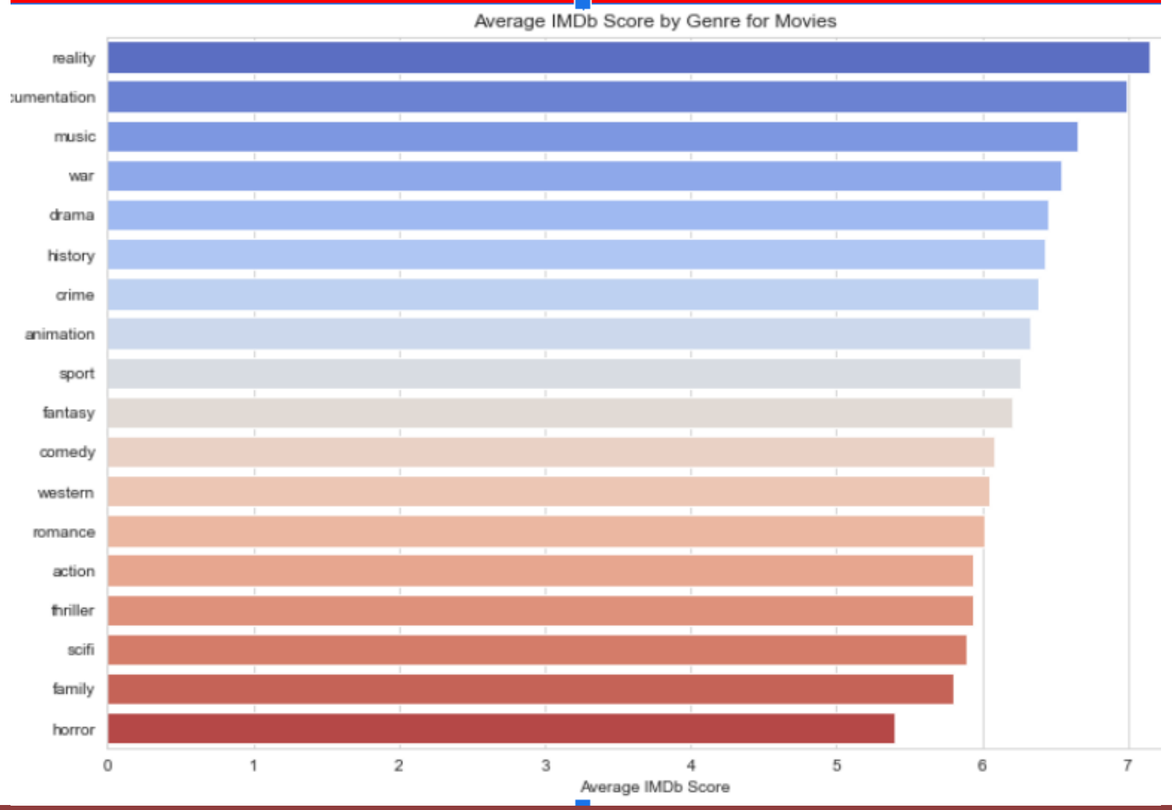
- Stability in Quality Ratings: Despite variability in specific years, the graph depicts a relatively stable average score range (around 6 to 8), suggesting a consistent quality of titles offered by Netflix.

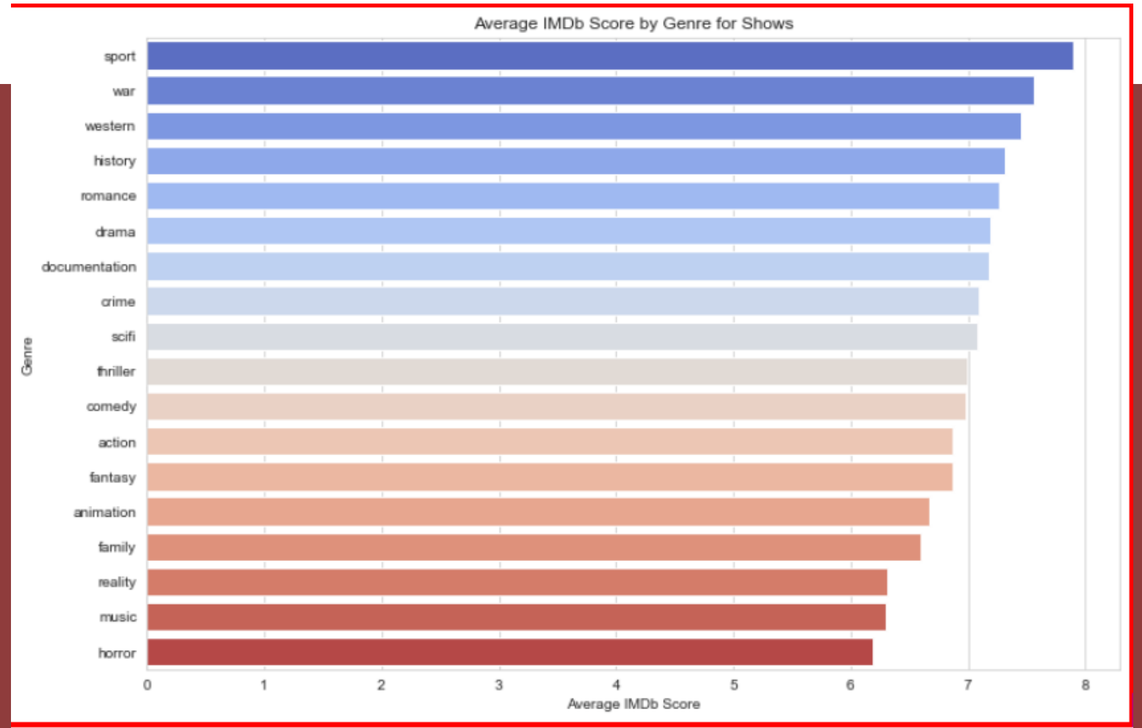


- Recent Content Surge: There is a marked increase in content production with the majority of shows/movies released after the year 2000, highlighting a surge in availability of newer content.

- Peak Release Period: The data peaks around 2020, indicating a potential strategic focus on producing content during this period, possibly to meet growing consumer demand for streaming content.

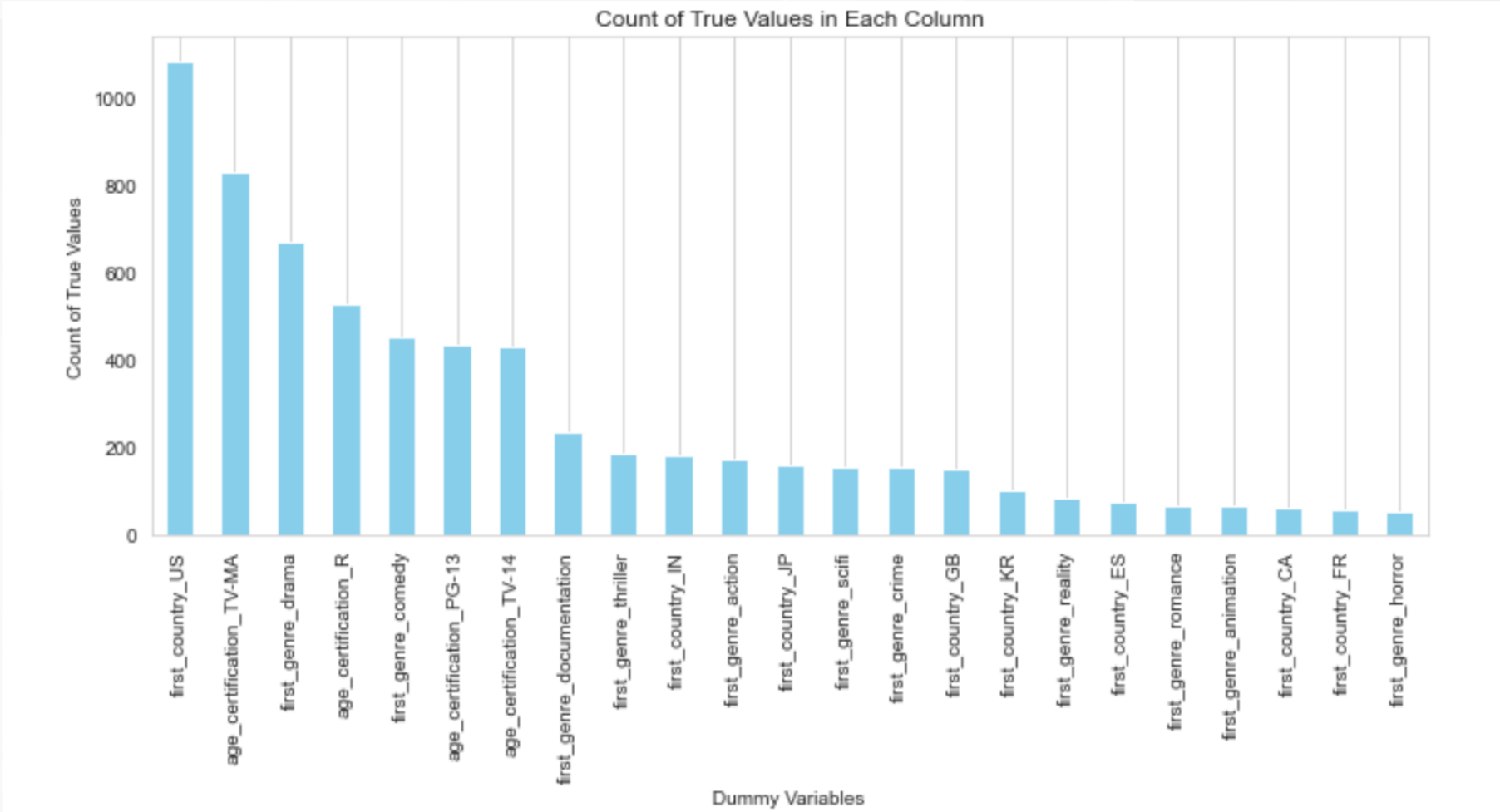
- Historical Content Rarity: Fewer titles are from the period before the year 2000, suggesting either a strategic collection focusing on modern content or a natural trend of digital availability for more recent titles.


Average IMDb Score by Genre for Movies
reality
umentation
music
war
drama
history
crime
animation
sport
fantasy
comedy
western
romance
action
thriller
scifi
family
horror
0
1
2
3
4
5
6
7
Average IMDb Score


Average IMDb Score by Genre for Shows
Genre
sport
war
western
history
romance
drama
documentation
crime
scifi
thriller
comedy
action
fantasy
animation
family
reality
music
horror
0
1
2
3
4
5
6
7
8
Average IMDb Score

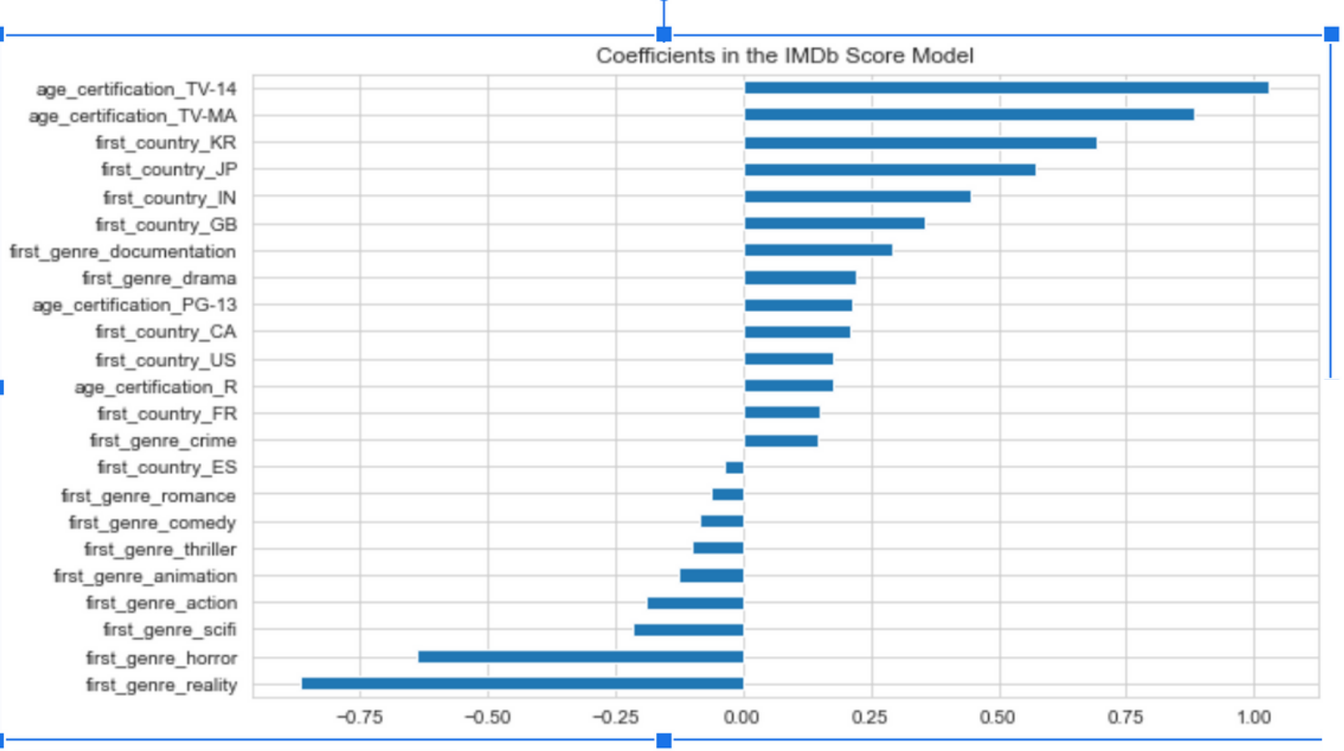
PREDICTIVE MODELING

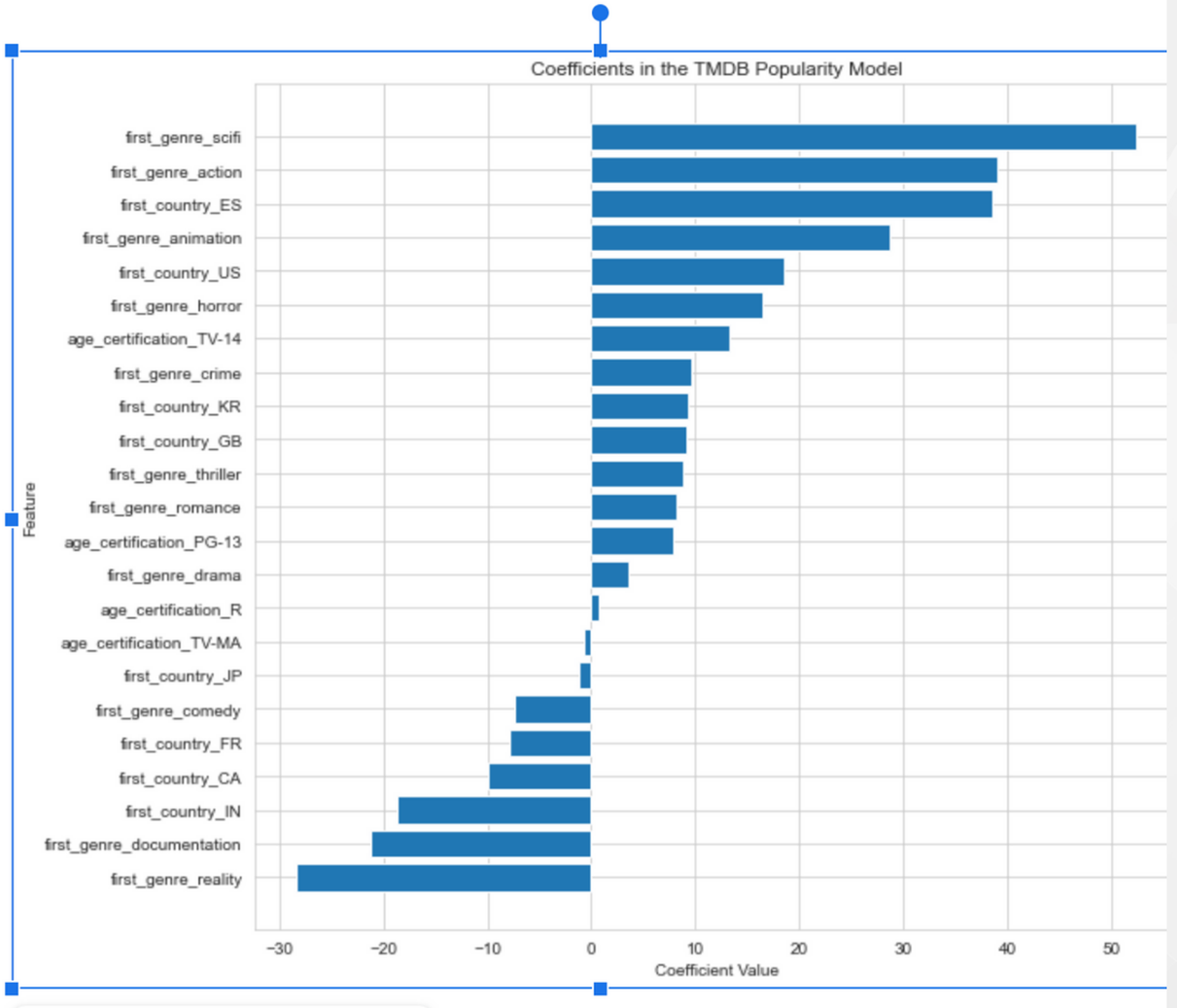
Linear Regression



IMDb Score Model - MSE: 1.0156862368548183, R-squared: 0.1378528368624028

TMDB Popularity Model - MSE: 12792.740931333932, R-squared: 0.011559028196384658


Coefficients in the IMDb Score Model
age_certification_TV-14
age_certification_TV-MA
first_country_KR
first_country_JP
first_country_IN
first_country_GB
first_genre_documentation
first_genre_drama
age_certification_PG-13
first_country_CA
first_country_US
age_certification_R
first_country_FR
first_genre_crime
first_country_ES
first_genre_romance
first_genre_comedy
first_genre_thriller
first_genre_animation
first_genre_action
first_genre_scifi
first_genre_horror
first_genre_reality
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Coefficients in the TMDB Popularity Model
Feature
first_genre_scifi
first_genre_action
first_country_ES
first_genre_animation
first_country_US
first_genre_horror
age_certification_TV-14
first_genre_crime
first_country_KR
first_country_GB
first_genre_thriller
first_genre_romance
age_certification_PG-13
first_genre_drama
age_certification_R
age_certification_TV-MA
first_country_JP
first_genre_comedy
first_country_FR
first_country_CA
first_country_IN
first_genre_documentation
first_genre_reality
−30
−20
−10
0
10
20
30
40
50
Coefficient Value

PREDICTIVE MODELING

Hyperparameter Tuning using Cross Validation

- **Best Parameters for IMDB Score:**

Test MSE: 1.0216249459079372
Test R-squared: 0.13281186950746382

- **Best Parameters for TMDB Popularity: Test MSE for TMDB Popularity: 13144.578723571822**
Test R-squared for TMDB Popularity: -0.015625989552629838

- Decision tree models underperform for IMDb scores and TMDB popularity, explaining little variance and showing weak predictive capabilities.
- The negative R-squared for the TMDB model highlights its unsuitability, possibly due to unrepresented influential factors.
- Enhancing model performance may require different algorithms, further feature engineering, or additional data to better grasp the factors affecting scores and popularity.

PREDICTIVE MODELING

Decision Trees

- Decision Tree Regressor - IMDb Score - MSE: 1.1253489787331514, R-squared: 0.04476757255374641 5

- Decision Tree Regressor - TMDb Popularity - MSE: 14229.502394442565, R-squared: -0.09945345180832033

- Based on these metrics, it appears that the decision tree models are not providing a better fit for the data than the linear regression models, at least not with the default settings or possibly the features that are being used.

- This could be due to various reasons, such as overfitting, where the model learns noise in the training data rather than true patterns. Decision trees are prone to overfitting.