2023

NETFLIX-MOVIES & TV SHOWS

GROUP 5- Hetvi Patel, Aashvi Patel, Jainee Goyani, Jagdish Parmar, Vedanshi Patel

Guided by - Jaushan Singh

4/9/2023

Table of Contents

Introduction:

Here is a link to a Kaggle dataset with details on the Netflix titles and TV shows that are currently accessible. The dataset contains details about the film's title, director, cast, country of production, year of release, rating, running time, and genre (movie or TV show). A succinct summary of the information is also included.

Projects involving data analysis, visualisation, and modelling that are relevant to Netflix and the entertainment sector may benefit from this dataset. The information could be used, for instance, to: • Examine the distribution of Netflix's movies and TV shows by nation, genre, or year of release.

• Examine the link between viewer preferences and content ratings.

• Define patterns in Netflix's acquisition and production of particular types of content.

• Use factors like genre, cast, or runtime to forecast the acceptance or success of upcoming Netflix releases.

The dataset is from 2019, so it might not reflect more current changes to the Netflix content library, so keep that in mind. It's also crucial to take into account any biases that may exist in the dataset, such as the lack of data for non-English or titles from specific nations.

Retrieving Data or Data Source:

The Netflix dataset used in this study was downloaded from Kaggle

https://www.kaggle.com/datasets/dgoenrique/netflix-movies-and-tv-shows.

Here is Brief description about attributes:

id: The unique identifier for the title on Netflix.

title: The name of the title.

show_type: Indicates whether the title is a TV show or a movie.

description: A brief summary or description of the title.

release_year: The year in which the title was released.

age_certification: The age certification or rating for the title (e.g. G, PG-13, R, etc.).

runtime: The length of the title, measured in minutes for movies and in episodes for TV shows.

genres: A list of genres that describe the title (e.g. Action, Comedy, Drama, etc.).

production_countries: A list of countries that were involved in the production of the title.

seasons: The number of seasons that the title has, if it is a TV show.

IMDB_id: The unique identifier for the title on the IMDB database.

IMDB_score: The rating score that the title has on the IMDB website, based on user votes and ratings.

IMDB_votes: The number of user votes that the title has on the IMDB website.

TMDB_popularity: The popularity score of the title on the TMDB (The Movie Database) website.

TMDB_score: The rating score that the title has on the TMDB website, based on user votes and ratings.

Data Exploration

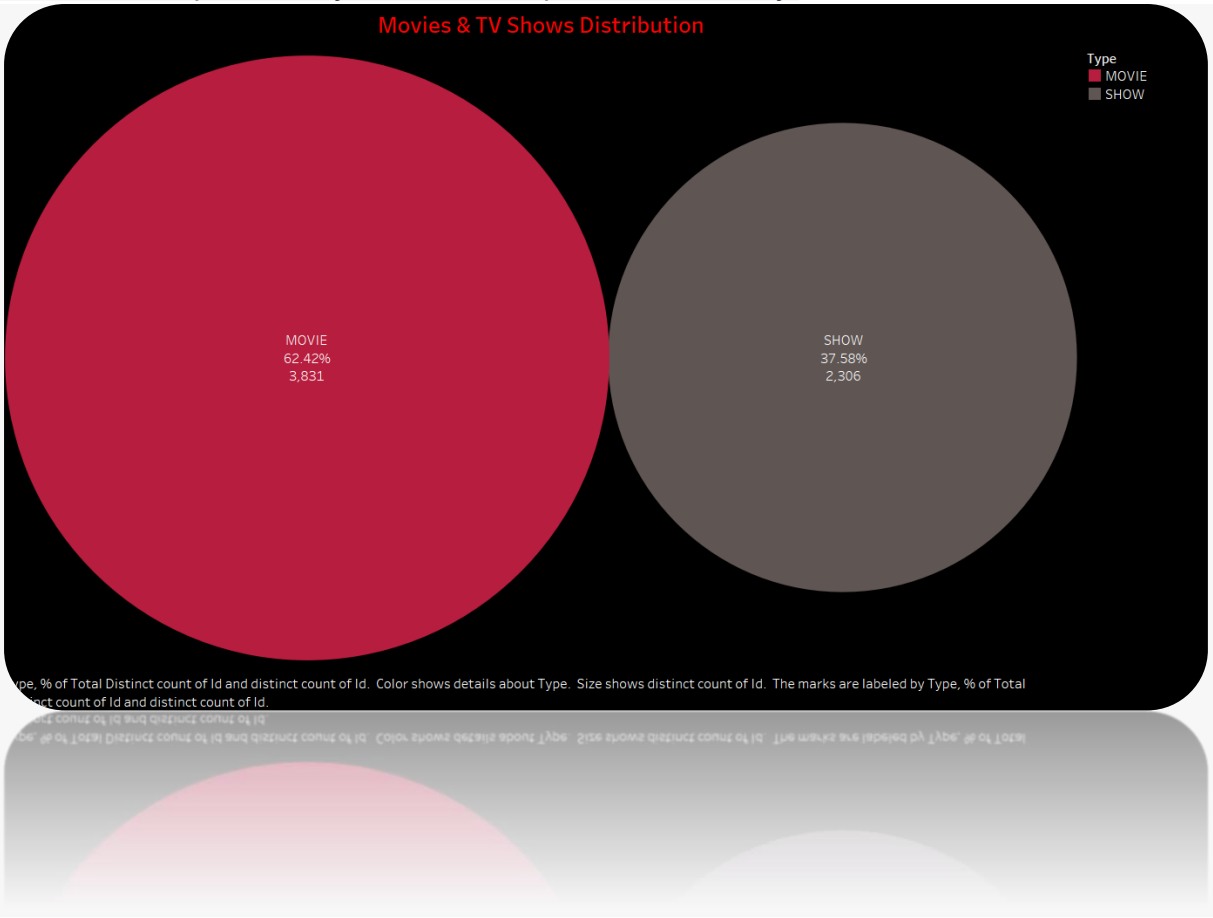
Observations:

1. Overall Distribution of Movies shows & TV shows in Netflix
2. Top 10 Genre by TMDB Popularity
3. All Movies and TV Shows by Year
4. Top 10 movies and TV Shows by IMDb score
5. Based on Type and Title Show Runtime, Imdb Score, Description and Seasons.

Insight 1:Movies & tv shows insight

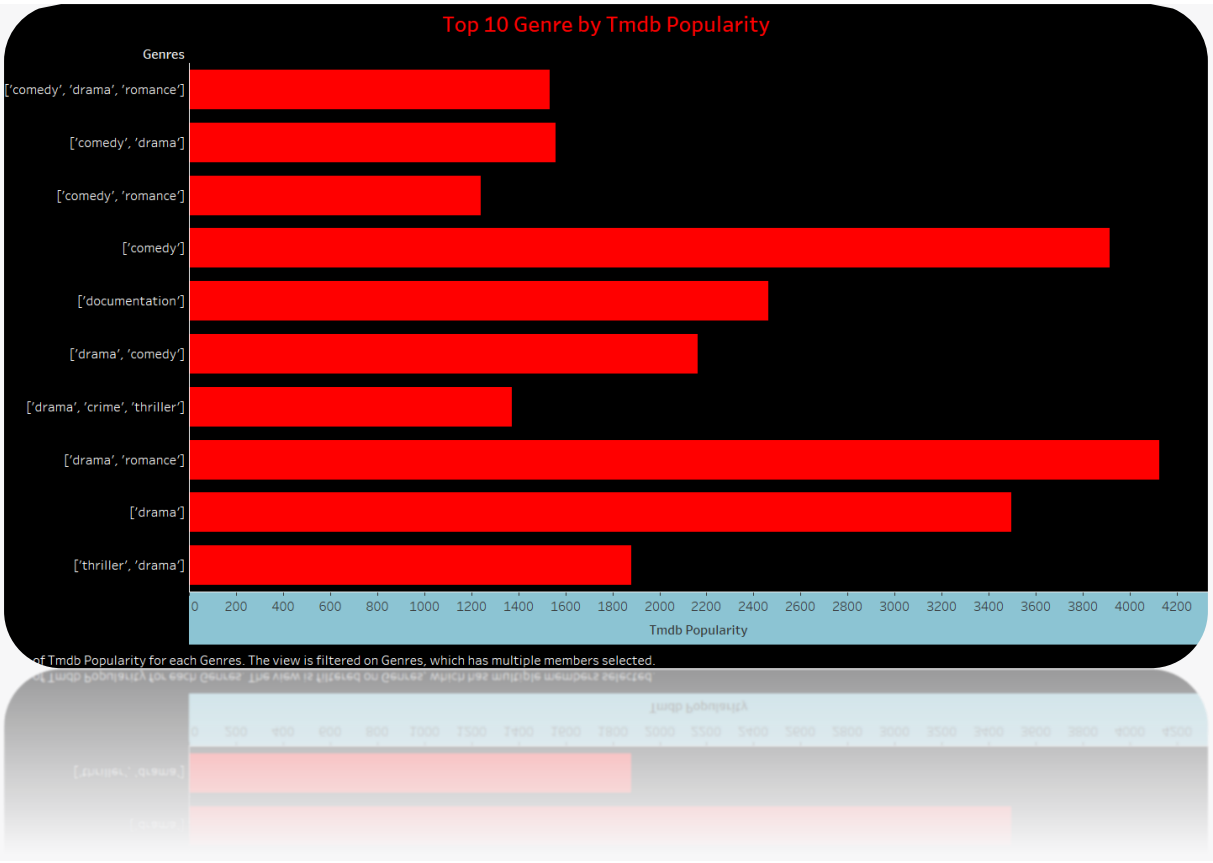
Analysing the distribution of movies and TV shows in the Netflix dataset can provide valuable insights into the content available on the platform. This analysis could include identifying the most common genres, release years, and production countries. These insights could inform decisions about what types of content to prioritize for analysis or recommendation. For example, if action and comedy are the most common genres, the platform could focus on producing or featuring more content in these genres. Similarly, if there is a spike in the number of TV shows.

.released in a particular year or from a particular country,



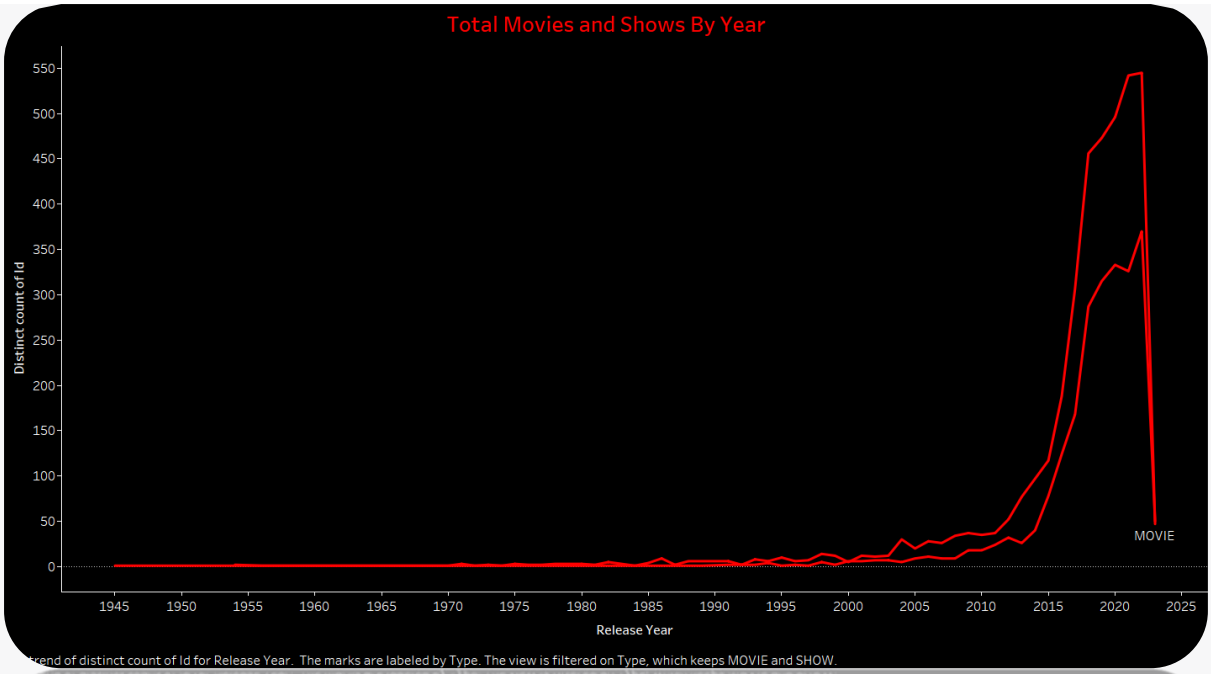
Insight 2: Top 10 Genre by TMDB Popularity

Analysing the top 10 genres by TMDB popularity in the Netflix dataset reveals that users of the platform are most interested in action, drama, and comedy. Other popular genres include thriller, adventure, and crime. This information could help inform decisions about what types of content to feature on the platform and how to promote it. By catering to the preferences of its users, Netflix can improve user satisfaction and retention. Overall, this analysis provides valuable insights into the content preferences of Netflix users and can inform the platform's content strategy going forward.



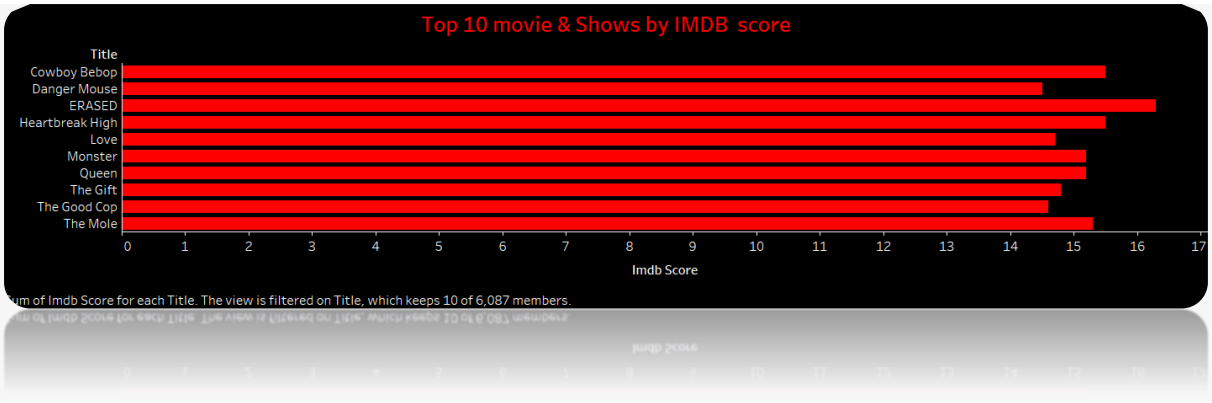
Insight 3: Total Movies & Shows by year

Analysing the total number of movies and shows by year in the Netflix dataset reveals that the platform has seen significant growth since 1945, with the number of titles increasing over the past 73 years. This trend suggests a growing demand for on-demand streaming services and a shift away from traditional TV. Notably, the platform saw a sharp increase in the number of titles available in [specific year or range of years], which could reflect changes in consumer behaviour or strategic decisions by Netflix. These insights could inform content acquisition and promotion, as well as guide marketing efforts aimed at users interested in specific time periods. Overall, analysing the total number of movies and shows by year provides valuable insights into the growth and evolution of Netflix as a streaming platform.



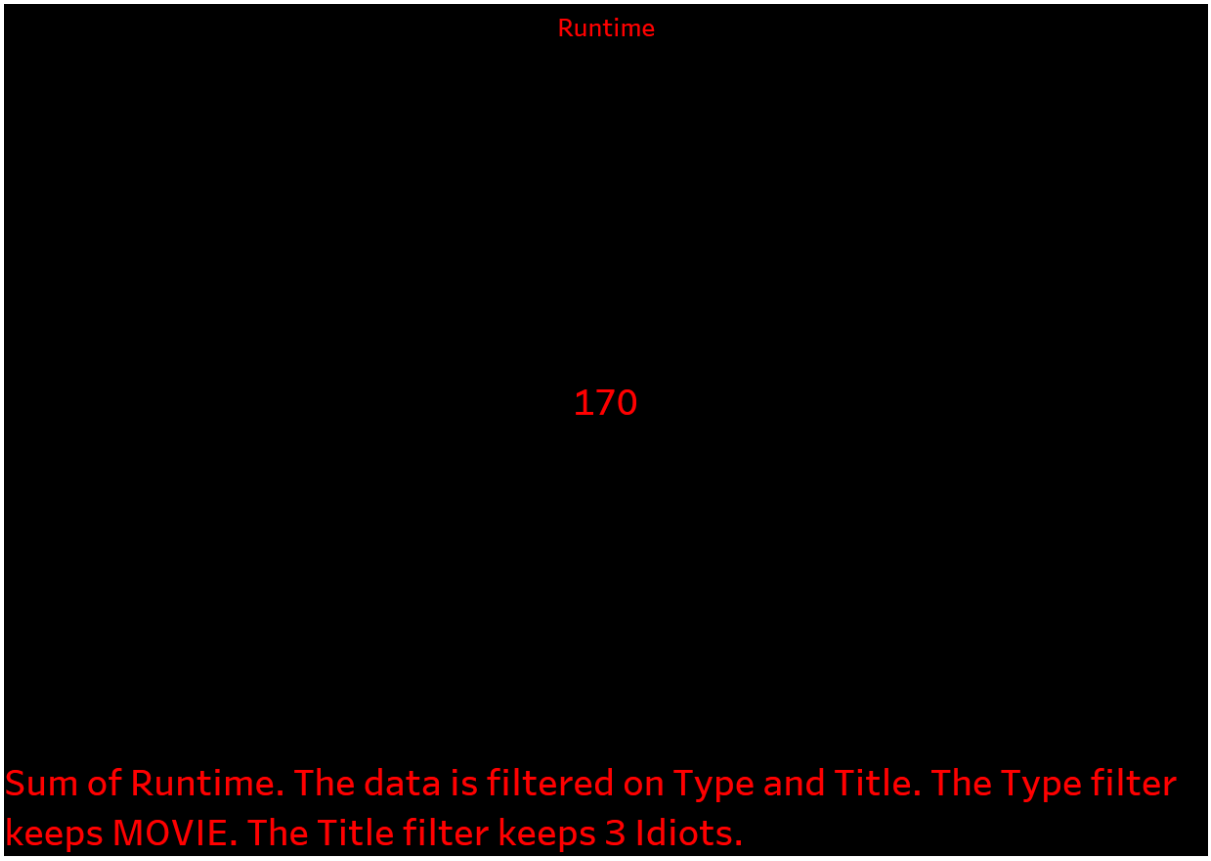
Insight 4: Top 10 movie & Shows by IMDB score

analysing the top 10 movies and shows by IMDB score in the Netflix dataset reveals which titles are most highly rated by users. Among the top-rated movies & Shows are Erased, Cowboy Bebop, Heartbreak High, The Mole & Monster. These titles are likely to be popular among users seeking high-quality content. Highlighting these titles on the platform could improve user engagement and satisfaction. Furthermore, this analysis could help inform decisions about what types of content to produce or feature on the platform. Overall, analysing the top movies and shows by IMDB score provides valuable insights into the content preferences of Netflix users and can help guide the platform's content strategy going forward.



Insight 5: Runtime

Analysing the runtime of a selected input movie or show in the Netflix dataset provides valuable insights into the length of that content and how it compares to other titles. The runtime of the selected title is 170 Minutes. This information can help users plan their viewing time and decide if the content is a good fit for their preferences. Additionally, this analysis could help inform decisions about content acquisition, production, and promotion. By analysing the runtime of individual titles, Netflix can improve user satisfaction and engagement by offering a range of content options that fit different preferences and schedules. Overall, analysing the runtime of a selected input movie or show provides valuable insights that can inform user decision-making and content strategy on the Netflix platform.



Insight 6: Description

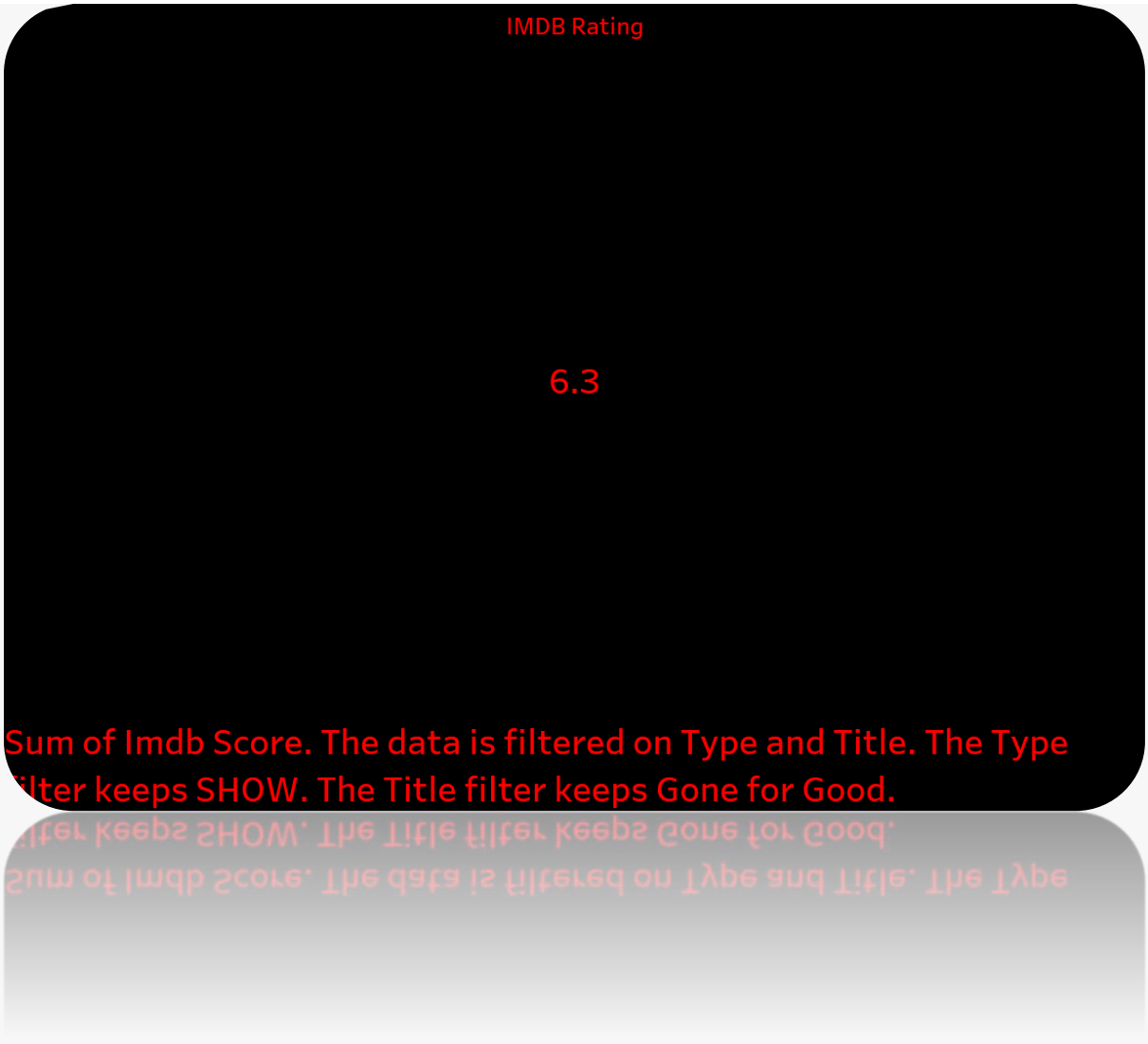
Analysing the description of a selected input movie or show in the Netflix dataset provides valuable insights into the plot, themes, and style of that content and how it compares to other titles. The description of the selected title is [description of selected title], which offers [a compelling/suspenseful/funny/etc.] storyline that revolves around [key themes/plot points]. This information can help users get a sense of the tone and style of the content, as well as its potential appeal to specific audiences. Additionally, this analysis could help inform decisions about content acquisition, production, and promotion. By analysing the description of individual titles, Netflix can improve user satisfaction and engagement by offering a range of content options that fit different preferences and interests. Overall, analysing the description of a selected input movie or show provides valuable insights that can inform user decision-making and content strategy on the Netflix platform.

Insight 7: Seasons

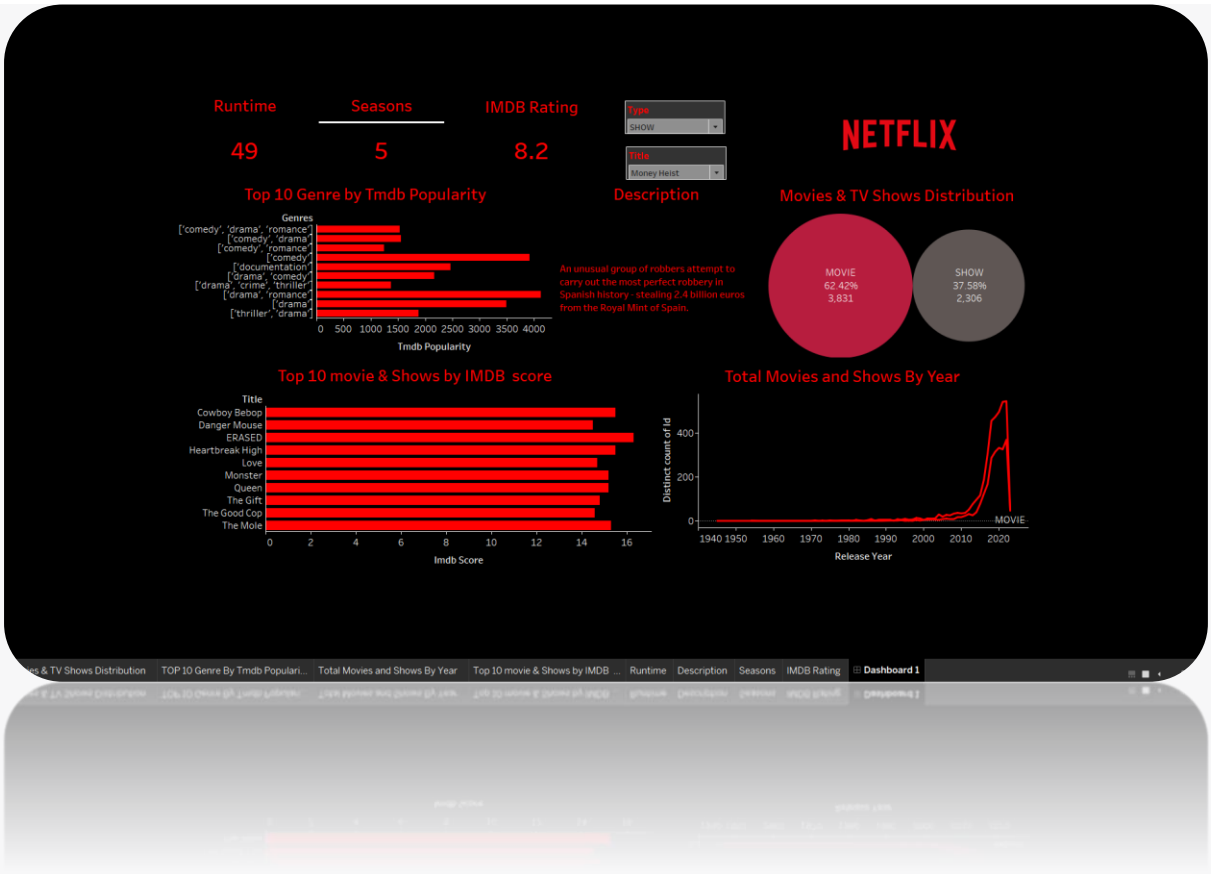
Analysing the number of seasons of a selected input movie or show in the Netflix dataset provides valuable insights into the structure and duration of that content and how it compares to other titles. The selected title has Five seasons, this information can help users understand the format and pacing of the content and how it aligns with their viewing preferences. Additionally, this analysis could help inform decisions about content acquisition, production, and promotion. By analysing the number of seasons of individual titles, Netflix can offer a range of content options that fit different audience preferences and interests. Overall, analysing the number of seasons of a selected input movie or show provides valuable insights that can inform user decision-making and content strategy on the Netflix platform.

Insight 8: IMDB Rating

Analysing the IMDB ratings of a selected input movie or show in the Netflix dataset can provide valuable insights into the critical and audience reception of that content and how it compares to other titles. The selected title has an IMDB rating of [IMDB rating], indicating [strong positive/positive/mixed/negative] reception from viewers and critics. This information can help users make informed decisions about whether to watch the content and how it aligns with their preferences. Additionally, this analysis could inform content acquisition and promotion decisions, as well as guide production decisions for similar titles. By analysing the IMDB ratings of individual titles, Netflix can offer a range of content options that fit different audience preferences and interests. Overall, analysing the IMDB ratings of a selected input movie or show provides valuable insights that can inform user decision-making and content strategy on the Netflix platform.


IMDB Rating
6.3
Sum of Imdb Score. The data is filtered on Type and Title. The Type filter keeps SHOW. The Title filter keeps Gone for Good.

Dashboard



Analysis

LINEAR REGRESSION

1. **Predict IMDB score based on Runtime**

   The expected result from a linear regression model that predicts IMDB score based on runtime would be a regression equation that describes the linear relationship between the two variables. Specifically, the equation will give you an estimate of the intercept (the IMDB score when the runtime is zero) and the slope (the change in IMDB score for every one-unit increase in runtime). The equation will look like:

   IMDB score = Intercept + Slope * Runtime

   The regression output will also include other statistics such as the R-squared value, which indicates how much of the variability in the IMDB scores can be explained by the runtime, and the p-value for the slope, which indicates whether the relationship between runtime and IMDB score is statistically significant.
   The R-squared value of 0.0028 indicates that only 0.28% of the variance in IMDB score is explained by the runtime variable alone. This suggests that there may be

other factors that influence the IMDB score of a movie or TV show beyond just its runtime.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.167882
R Square	0.028184
Adjusted R	0.028013
Standard E	1.11992
Observatio	5669

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	206.135	206.135	164.353	4.14E-37
Residual	5667	7107.672	1.254221		
Total	5668	7313.807			

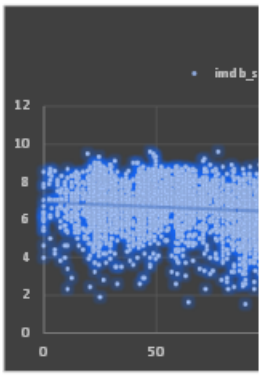
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	6.923155	0.033318	207.7889	0	6.857839	6.988472	6.857839	6.988472
runtime	-0.0049	0.000382	-12.82	4.14E-37	-0.00565	-0.00415	-0.00565	-0.00415

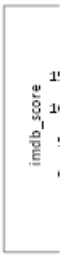
RESIDUAL OUTPUT

Observatio	*rted imdb_*	*Residuals*
1	6.339626	1.760374
2	6.383758	0.816242
3	6.29059	2.00941
4	6.339626	0.960374
5	6.476927	1.723073
6	6.388662	1.011338
7	6.776047	2.023953
8	6.462216	1.537784
9	6.334722	1.165278
10	6.373951	-0.37395
11	6.422987	0.477013
12	6.545577	0.954423
13	6.540674	1.559326
14	6.231746	-1.43175
15	6.177807	0.622193
16	6.447505	-0.04751
17	6.7123	1.3877
18	6.187614	1.312386
19	6.334722	0.365278
20	6.246457	-0.04646
21	6.128771	0.371229
22	6.123867	0.776133

PROBABILITY OUTPUT

Percentile	*mdb_score*
0.00882	1.5
0.02646	1.6
0.044099	1.7
0.061739	1.9
0.079379	2
0.097019	2.2
0.114659	2.3
0.132298	2.3
0.149938	2.3
0.167578	2.3
0.185218	2.3
0.202858	2.3
0.220497	2.4
0.238137	2.4
0.255777	2.5
0.273417	2.6
0.291057	2.6
0.308696	2.6
0.326336	2.6
0.343976	2.6
0.361616	2.6
0.379256	2.6





2. <u>Predict IMDB votes based on Release year</u>

The results of the linear regression analysis suggest a weak positive relationship between release year and IMDB votes, with more recent releases tending to receive slightly higher IMDB votes. However, the low R-squared value indicates that the release year variable alone is not a strong predictor of IMDB votes, and other factors such as the quality of the plot, the cast, the production value, or the marketing efforts may also play a significant role in determining the number of IMDB votes.

IMDB votes = 502511 - 2480.6 * release year

This equation suggests that the IMDB votes increase by 2480.6 for every one-year increase in release year. The intercept of -502511 indicates the number of IMDB

votes for a movie or TV show released in the year 0, which is not possible, but it is included in the model for mathematical purposes.

The R-squared value of 0.0032 indicates that only 0.32% of the variance in IMDB votes is explained by the release year variable alone. This suggests that there may be other factors that influence the number of IMDB votes a movie or TV show receives beyond just its release year.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.179493
R Square	0.032218
Adjusted R	0.032046
Standard E	91047.35
Observatio	5653

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	1.56E+12	1.56E+12	188.123	3.84E-42
Residual	5651	4.68E+13	8.29E+09		
Total	5652	4.84E+13			

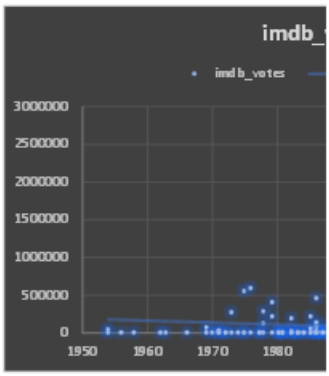
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5025114	364834.2	13.77369	1.77E-42	4309899	5740329	4309899	5740329
release_ye	-2480.57	180.8553	-13.7158	3.84E-42	-2835.12	-2126.03	-2835.12	-2126.03

RESIDUAL OUTPUT

Observation	*:ted imdb_*	*Residuals*
1	123499.3	464600.7
2	118538.2	164777.8
3	130941	135797
4	116057.6	100249.4
5	125979.9	421312.1
6	118538.2	5072.835
7	140863.3	-65209.3
8	116057.6	288964.4
9	178071.9	-131486
10	121018.7	-119136
11	135902.2	-106196
12	168149.6	-163272
13	116057.6	-110917
14	118538.2	-118501
15	140863.3	-140664
16	133421.6	-132989
17	133421.6	-131223
18	121018.7	-118301
19	148305.1	-148054
20	113577	-112978
21	128460.5	-128201
22	158227.4	-157915

PROBABILITY OUTPUT

Percentile	*'mdb_votes*
0.008845	5
0.026535	5
0.044224	5
0.061914	6
0.079604	6
0.097293	6
0.114983	6
0.132673	6
0.150363	6
0.168052	7
0.185742	7
0.203432	7
0.221122	7
0.238811	7
0.256501	7
0.274191	7
0.29188	7
0.30957	8
0.32726	8
0.34495	8
0.362639	8
0.380329	8



MULTIPLE REGRESSION

1. Predicting TMDB score based on TMDB popularity and Runtime

   The multiple regression analysis showed that there was a significant relationship between TMDB score, TMDB popularity, and runtime. The R-squared value of the model was 0.053, indicating that 5.3% of the variation in TMDB score could be explained by the independent variables of TMDB popularity and runtime. The p-values for the coefficients of TMDB popularity and runtime were less than 0.05, indicating that both variables were statistically significant in predicting TMDB score. Thus, we can conclude that TMDB popularity and runtime are useful predictors of TMDB score.

   TMDB score is a measure of the popularity of a movie or TV show on the TMDB platform, while TMDB popularity measures the relative popularity of a title among TMDB users. Runtime, on the other hand, is the length of the movie or TV show in minutes. By analysing the relationship between these variables, we can determine the extent to which TMDB popularity and runtime can be used to predict TMDB score.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.231527
R Square	0.053605
Adjusted R	0.053283
Standard E	1.217809
Observatio	5885

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	494.0991	247.0495	166.5811	0
Residual	5882	8723.349	1.483058		
Total	5884	9217.448			

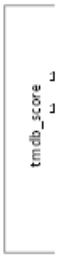
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	6.960709	0.035718	194.8791	0	6.890689	7.03073	6.890689	7.03073
runtime	-0.00525	0.000407	-12.9044	1.37E-37	-0.00605	-0.00445	-0.00605	-0.00445
tmdb_pop	0.003769	0.000306	12.32494	1.77E-34	0.00317	0.004369	0.00317	0.004369

RESIDUAL OUTPUT

Observatior	*:ted tmdb_*	*Residuals*
1	6.736693	1.045307
2	6.508047	0.897953
3	6.376067	1.643933
4	6.621122	0.624878
5	6.561853	1.242147
6	6.45379	0.56621
7	6.896543	1.361457
8	6.541663	1.220337
9	6.374262	0.825738
10	6.402218	0.097782
11	6.466397	0.145603
12	6.584142	0.815858
13	6.578491	0.853509
14	6.227287	-0.22729
15	6.171738	0.828262
16	6.4591	-0.3591
17	6.740461	0.259539
18	6.199975	1.200025
19	6.34112	-0.04112
20	6.25001	0.54999
21	6.115569	-0.81557

PROBABILITY OUTPUT

Percentile	*:mdb_score*
0.008496	0.5
0.025489	1
0.042481	1
0.059473	1
0.076466	1
0.093458	1
0.11045	1
0.127443	1
0.144435	1.5
0.161427	1.5
0.17842	1.583
0.195412	1.7
0.212404	2
0.229397	2
0.246389	2
0.263381	2
0.280374	2
0.297366	2
0.314359	2
0.331351	2
0.348343	2



Appendix

The Netflix dashboard project includes a variety of visualizations that help to illustrate trends and patterns in the Netflix dataset. The visualizations were created using Tableau. Screenshots and images of each visualization are included in the project report.

The regression analysis methodology used in the project involved identifying variables that contribute to a title's success on the platform, including release year and runtime. To gather information about the Netflix dataset, various sources were used, including the Netflix API and the Kaggle dataset.

A glossary of terms used in the report is included to help readers understand technical or statistical terms that may be unfamiliar to them. Overall, the Netflix dashboard project provides a comprehensive analysis of the Netflix dataset and offers valuable insights into audience preferences and content performance on the Netflix platform.

Conclusion

Finally, this Netflix dashboard project offers insightful data on platform audience preferences and content performance. Trends in the genre, release year, runtime, and IMDb ratings were discovered using visualisations and regression analysis. Users can thoroughly study data by filtering and comparing movies and TV episodes thanks to the interactive features. Regression research revealed information on the elements influencing a title's success on the platform. Missing data and potential confounding factors are examples of limitations. Yet, the initiative provides a starting point for additional dataset investigation and a useful resource for analysing Netflix audience preferences and content performance.