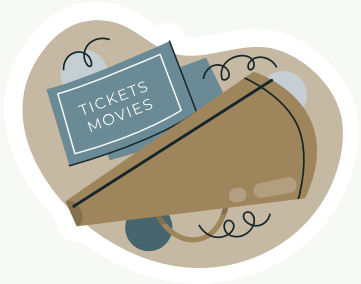
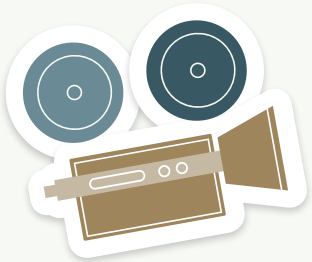
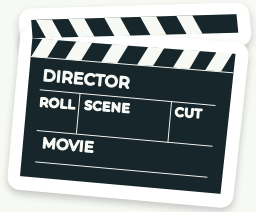
Rotten Tomatoes Movie Reviews analysis

Group 3

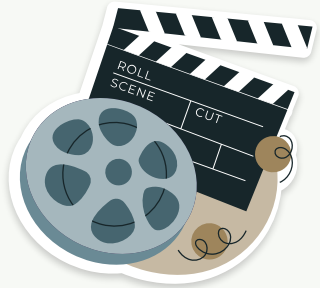
Table of contents



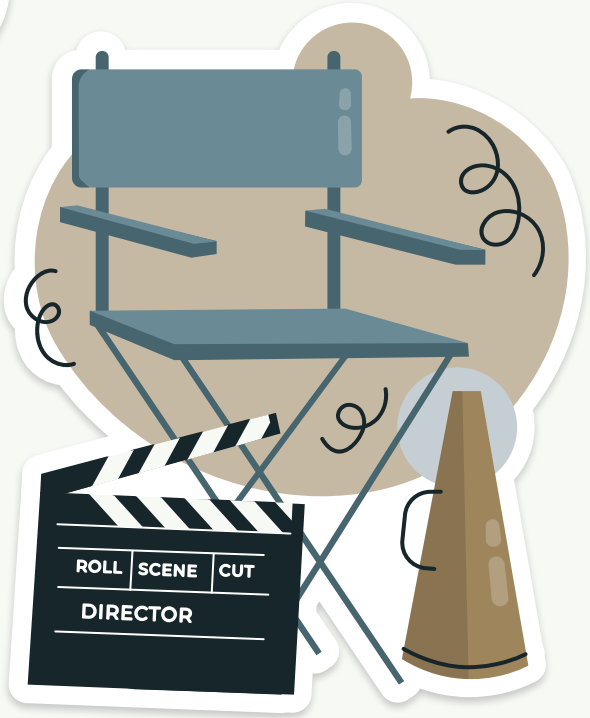




Objective



Introduction	Initial data exploration and preprocessing
Data Accuracy	Compare the Rotten Tomatoes movies dataset and summarize the result
Systematic Differences	Systematic differences between critics and audience reviews
Revenue Prediction	Predict popular movies' revenue
Detail Analysis	Detailed analysis of 3 movies



00

Introduction

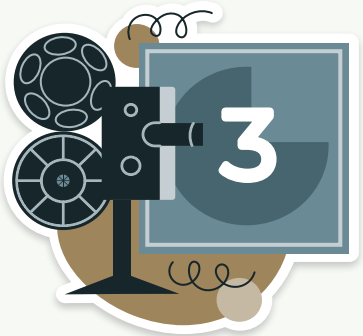
Data Exploration

We were provided with two kaggle datasets:

1) https://www.kaggle.com/datasets/bwandowando/rotten-tomatoes-9800-movie-critic-and-user-reviews
2) https://www.kaggle.com/datasets/kalilurrahman/top-box-office-revenue-data-english-movies?select=boxofficemojoustop1000.tsv

Dataset 1: Contains rotten tomatoes reviews for audience and critics along with a compilation of reviewers' scores and sentiments for each movie

Dataset 2: Box office review dataset that provides the 1000 movies with the highest revenue of all time

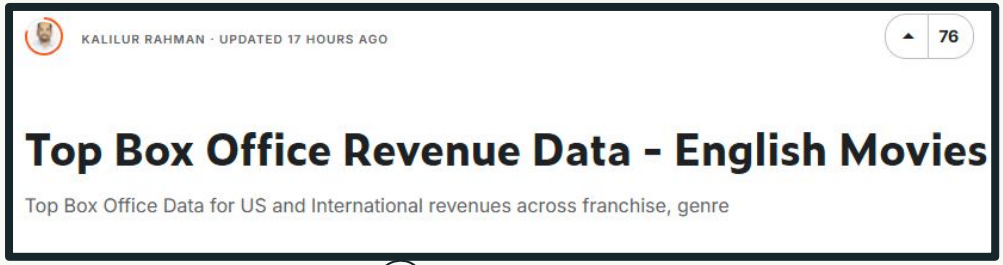


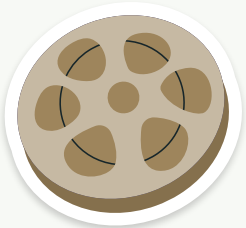
Data Exploration Cont.

Dataset 1: Updated 5 months ago

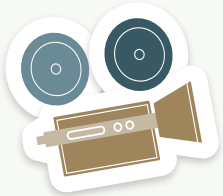
- The two datasets provided are updated fairly regularly; therefore, they would have more recent data.
- Due to the update rate of dataset 2, there could be some discrepancies such as new movie releases that did not get accounted for in dataset 1.

Dataset 2: Updated 17 hours ago (update daily)

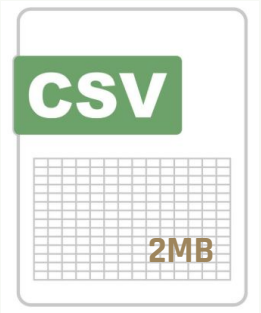




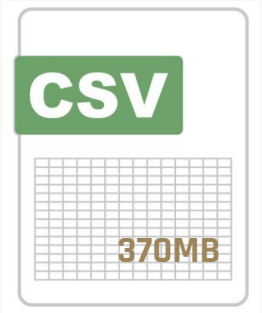
Data Retrieval



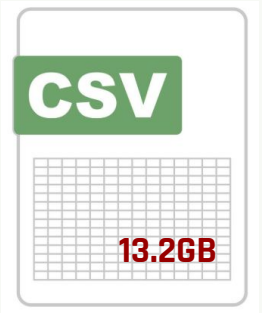
Movies.csv



Critic_reviews.csv



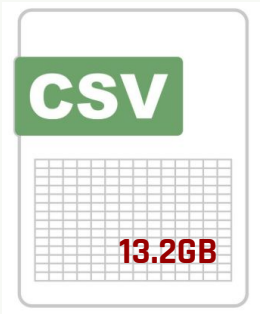
Audience_reviews.csv



By extracting the first dataset, we could see that one file, audience_reviews.csv, is very large. To do any type of analysis we needed to find a way to transform this data file into something that is more manageable.

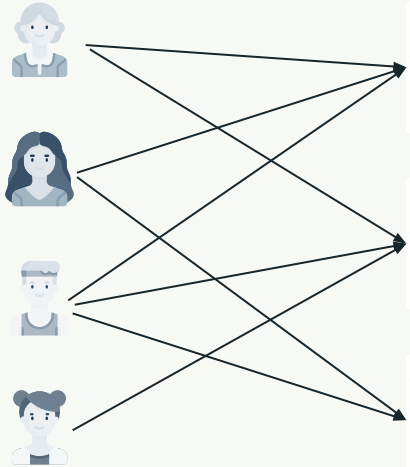
Data Manipulation

Audience_reviews.csv

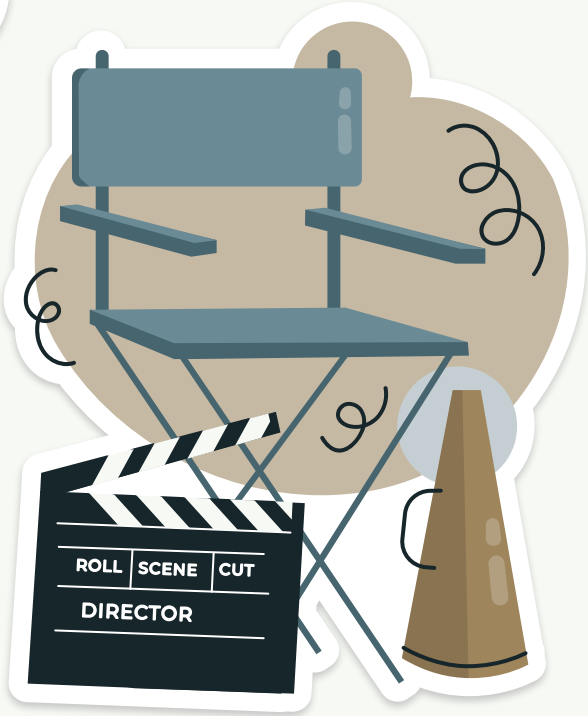


Batch every 10,000 rows

Give batch to Each Worker thread



Workers write to corresponding csv based on movie id



01

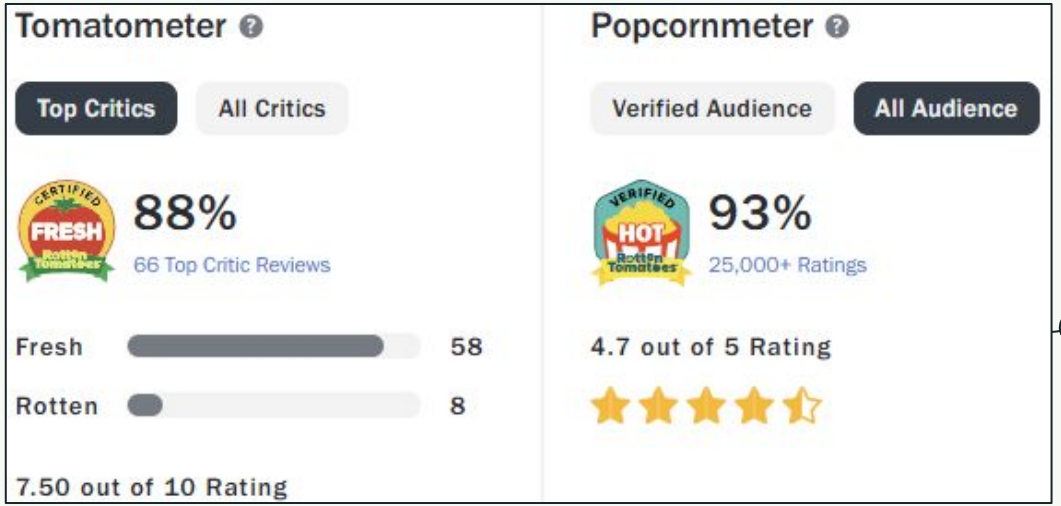
Data Accuracy

Raw Score Analysis

How does Rotten Tomatoes rate critic and audience scores?

Critic Scores: A binary rating, whether the movie is Fresh or Rotten, then calculate the average.

Audience Score: Rating out of 5. Users can rate in 0.5 increments from 0.5 to 5.0.



$$\text{Critic Score Percentage} = \frac{(\sum(isFresh = true))}{N\ (total\ reviews)}\ x\ 100$$

$$\text{Example Critic Score} = \frac{58}{66} x100\ = 87.88\%$$

$$\text{Audience Score Percentage} = \frac{(\sum audience\ rating)}{N\ (total\ reviews)} \Big/ _5\ x\ 100$$

$$\text{Example Audience Score} = {}^{4.7}\Big/_5\ x\ 100 = 94\%$$

Note: Due to rounding, 93% of Rotten Tomatoes is more accurate, and is what we used for our analysis

Raw Data Normalization

Using the data from movies.csv, we identified the maximum and minimum audience and user scores corresponding to positive and negative sentiments. Based on this analysis, scores between 0 and 59 indicate a negative sentiment, while scores from 60 to 100 resulted in a positive sentiment.

critic_sentiment	max	min
negative	59.0	0.0
positive	100.0	60.0

audience_sentiment	max	min
negative	59.0	0.0
positive	100.0	60.0

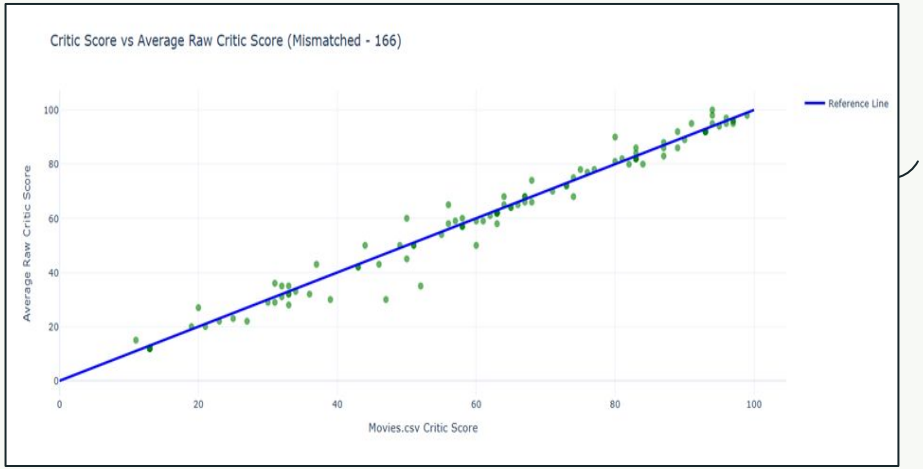
movieId	avg_audience_score	avg_audience_sentiment	avg_critic_score	avg_critic_sentiment
00112b83-23bd-44ce-8ae3-f27194f651d6	89	positive	98	positive
0052d083-dc96-3f49-9413-7235c9d68c07	63	positive	48	negative
...	...	...	...	...

Critic Score Analysis

There were two analysis done in this section of the Critic Scores:

1) **How many reviews in the movies.csv file are missing from the raw data, and how many reviews in the raw data are absent from the movies.csv file?**
   - From the results, we observe that all the movies in the raw data are present in the movies.csv dataset. However, there are 495 movies in the raw data that do not appear in movies.csv.
2) **Among the movies present in both datasets, how many reviews have differing scores?**
   - There are 166 movies that have different scores between the raw and movies.csv datasets.

```
Critic score reviews in Movies.csv that does NOT show up in raw data: 0
Critic score in raw data that does NOT show up in Movies.csv: 495
```

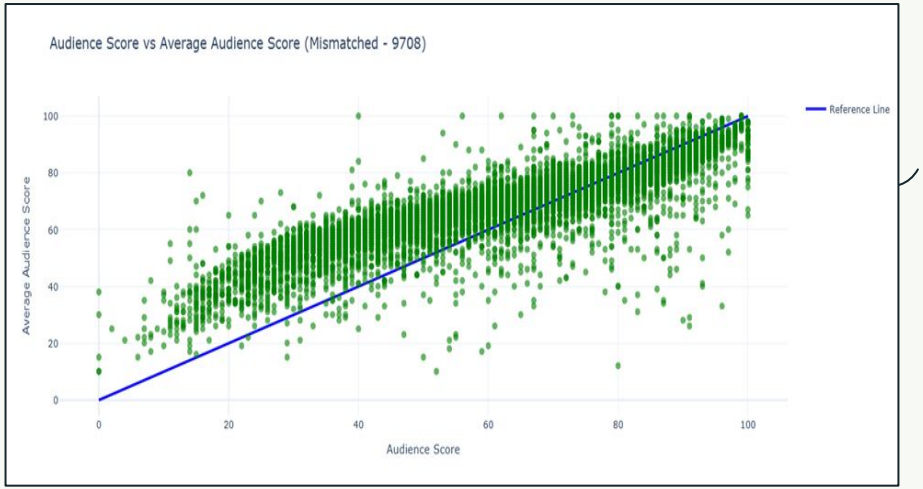


Audience Score Analysis

There were two analysis done in this section of the Audience Scores:

1) **How many reviews in the movies.csv file are missing from the raw data, and how many reviews in the raw data are absent from the movies.csv file?**
   - There are 13 movies present in the movies.csv file but missing from the raw data, and 280 movies found in the raw data that are absent from the movies.csv file.
2) **Among the movies present in both datasets, how many reviews have differing scores?**
   - Among the movies present in both datasets, 9,708 movies have differing scores.

```
Audience score reviews in Movies.csv that does NOT show up in raw data: 13
Audience score in raw data that does NOT show up in Movies.csv: 280
```

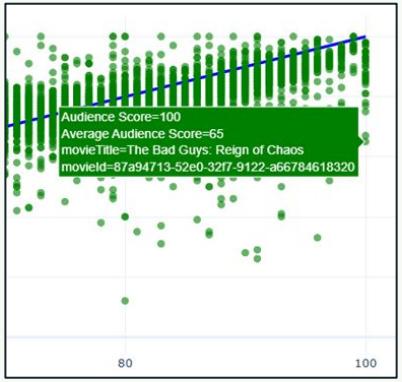


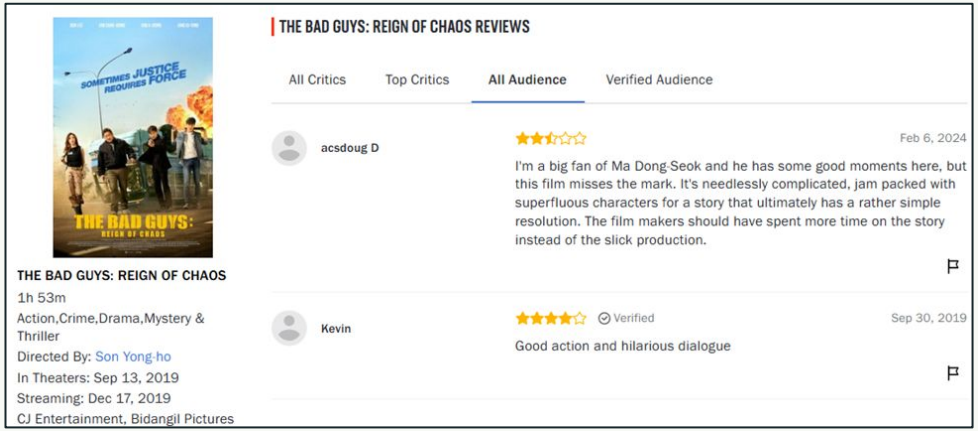
Audience Score Analysis Cont.

Audience Score Mismatch Comparison:

We compared a movie that appears in both the raw data and the movies.csv file by selecting one with a perfect rating in the movies.csv file but not in the raw data. The movie chosen for this comparison is *The Bad Guys: Reign of Chaos*.

Upon looking up the movie on Rotten Tomatoes, we found that there are only two audience reviews for it. The average and normalized review score for the movie is 65%, which matches the score in our raw data, not the movies.csv file.





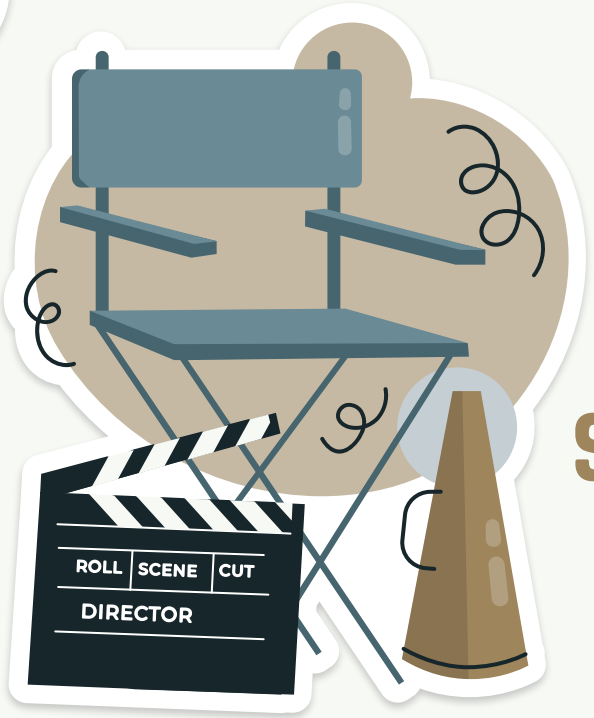
Results from our Analysis

What did we find?

- We found that there is definitely a mismatch between the movies.csv data and the raw data retrieved from this author.
- Upon further analysis, we discovered that every time the author updates the data repository, they do not update the movies.csv file to match the most up-to-date data on Rotten Tomatoes.

Conclusion

- To ensure the analysis remains accurate and up-to-date for the top movies revenue dataset, we decided to use data from the raw data instead of the outdated movies.csv file.

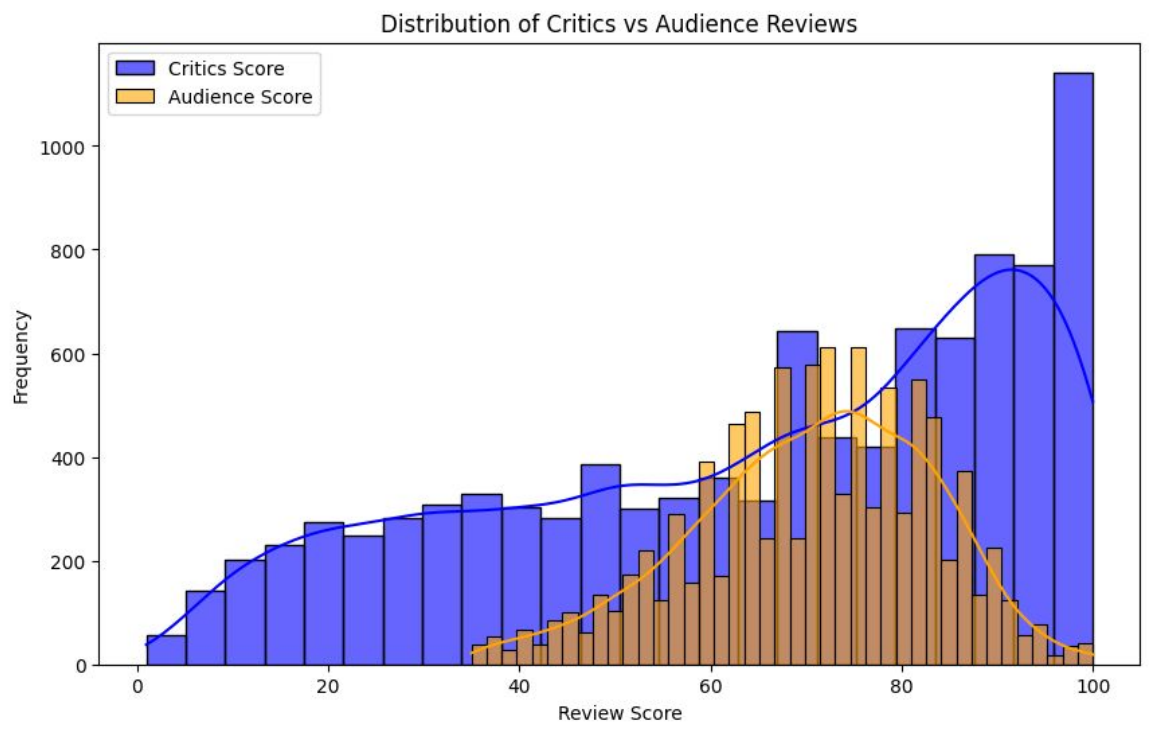


02

Systematic Differences

Systematic Differences

1. Bias in data collection
2. Measurement Inconsistency
3. External Influences

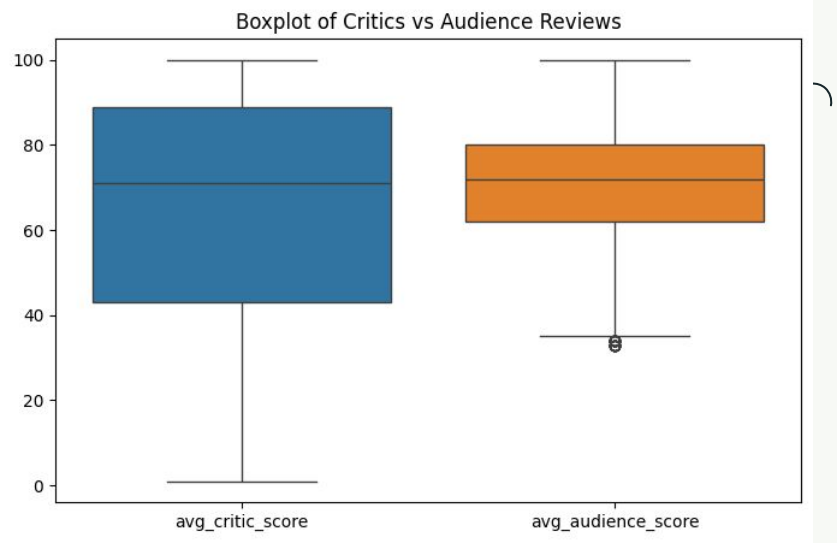


Audience vs Critic

There are distinct differences between score distributions.

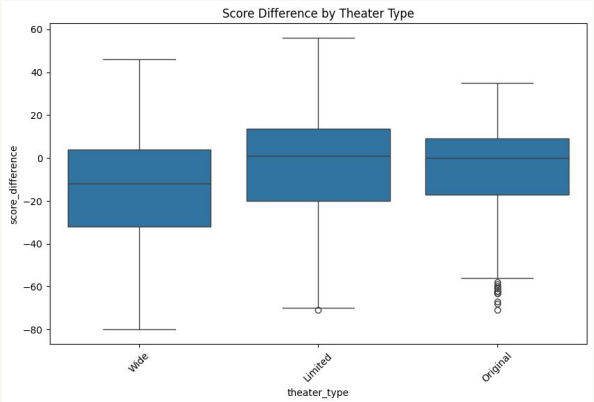
Critics - larger range & distribution below mean

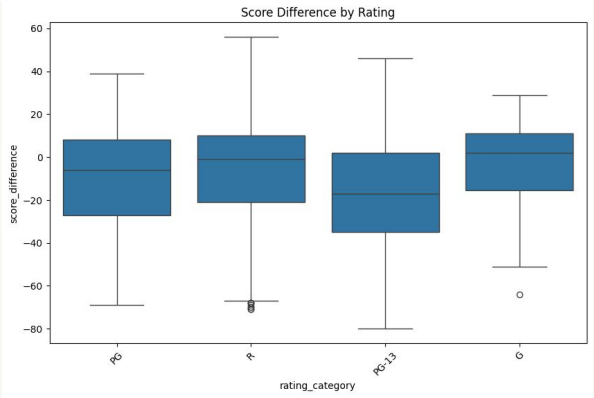
Audience - similar mean & fewer rating below mean



Correlation Coefficient : 0.54 suggests some relationship
T-statistic value : -18 suggests large negative difference between groups
P-value value : 9.0e-74 indicates that this difference is is significant

Assessing Associations





Score Difference : Positive means Critics scored higher

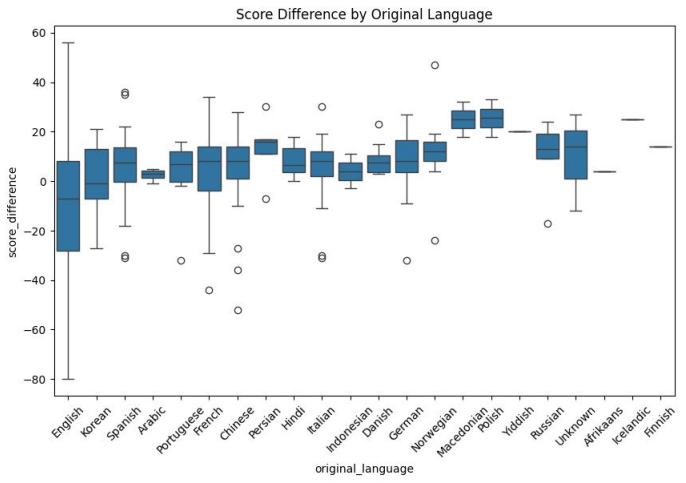
Theater Type :

- Results suggest Limited and Original releases were scored higher by critics
  - Anova test confirms finding

Rating Category :

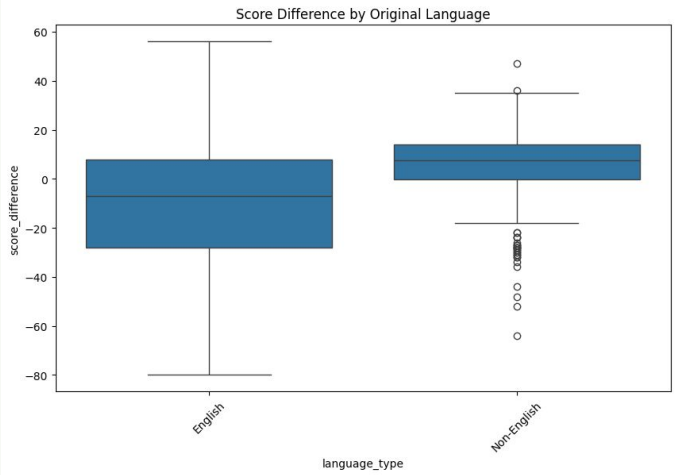
- Results suggest statistically significant difference. With critics favoring extremes (R,G).

Assessing Associations



Original Language :

Results suggested marginal but statistically significant difference



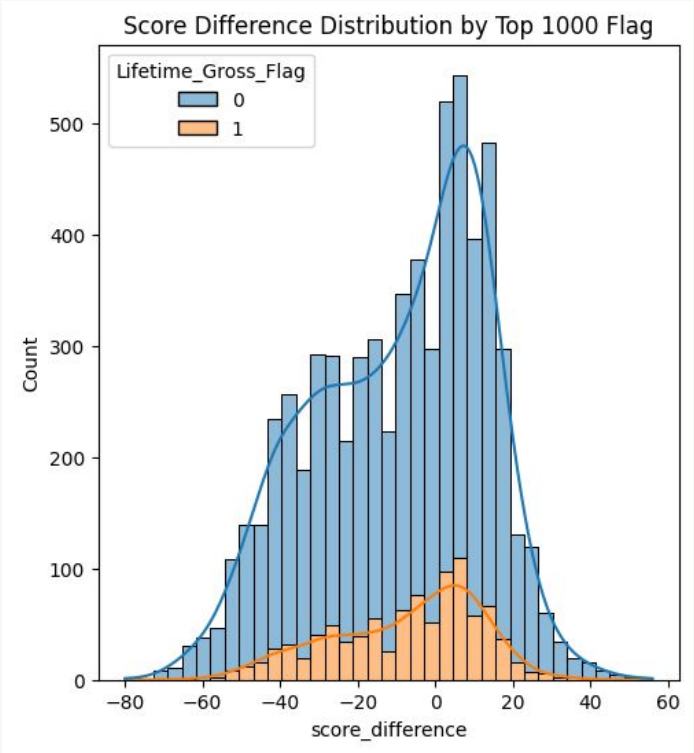
English vs Other :

After binning - Results showed statistically significant and moderate differences

Scores Vs Revenue

Shows comparison of Top Grossing (1) Vs non-Top Grossing(0).

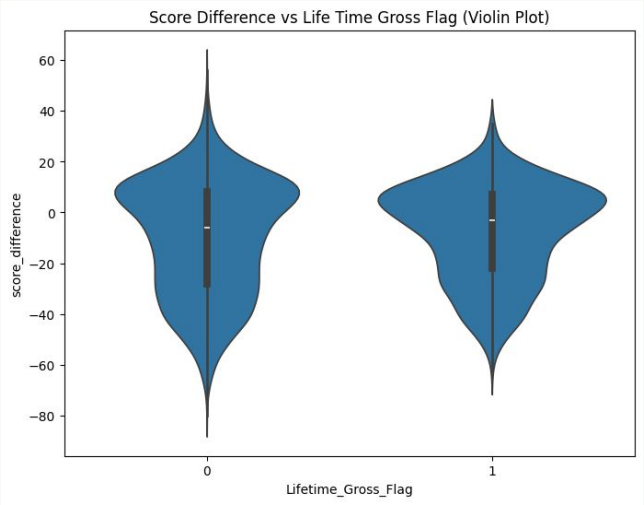
- Large gap Top Revenue(TR) per movie
- Critics appear to favor TR movies



Scores Vs Revenue

Violin plot

- Top-Grossing(right) - Tight distribution ~10 suggests critics favor TR movies
- There is a statistically significant difference in score_difference between the different categories.



Anova Results : F-statistic: 8.24 P-value: 0.00409

Systematic Differences

Critics Favor

Top-Grossing

Non-English

R/G Rated

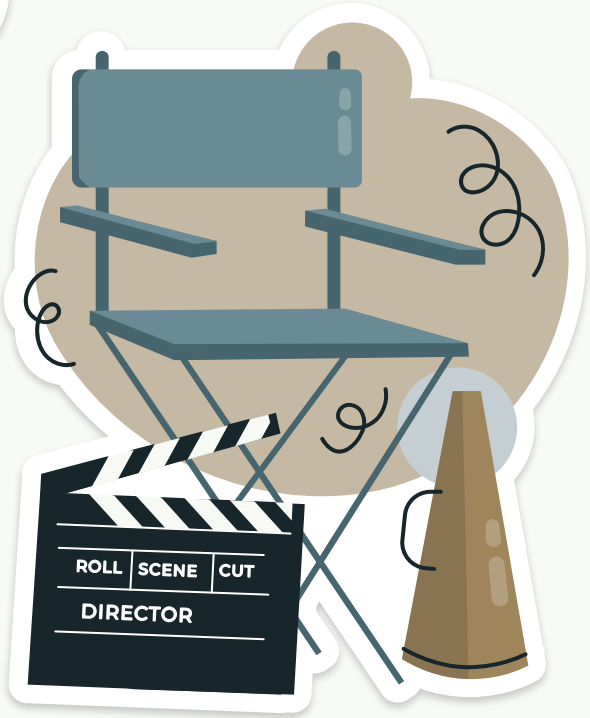
Limited/Original

What does this mean?

Potential Bias

Corporate Influence

Systematic Differences

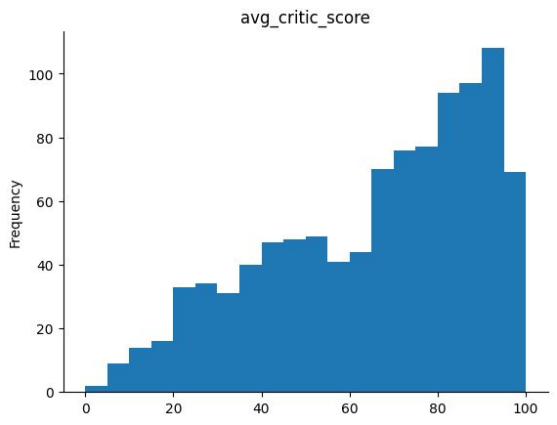


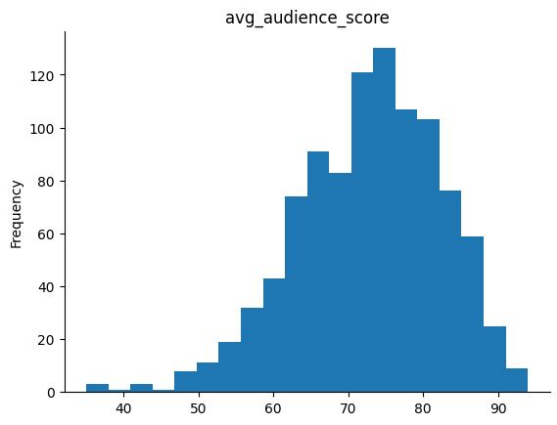
03

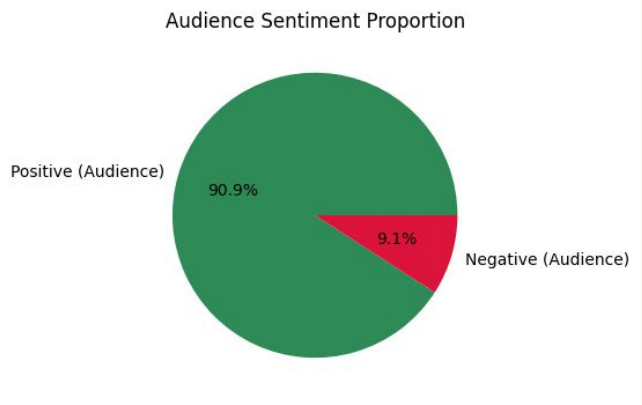
Revenue Prediction

3. Can reviews predict movie revenues?

Understanding Critic and Audience Scores:

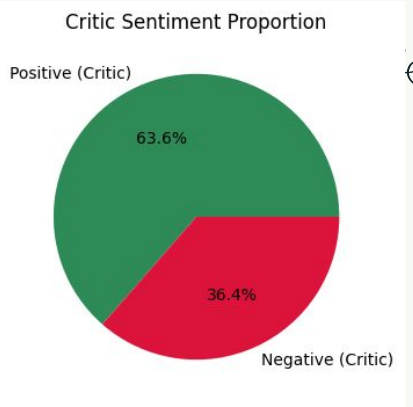






- peak frequency around 80-100
- Fewer movies- 0-40 score range

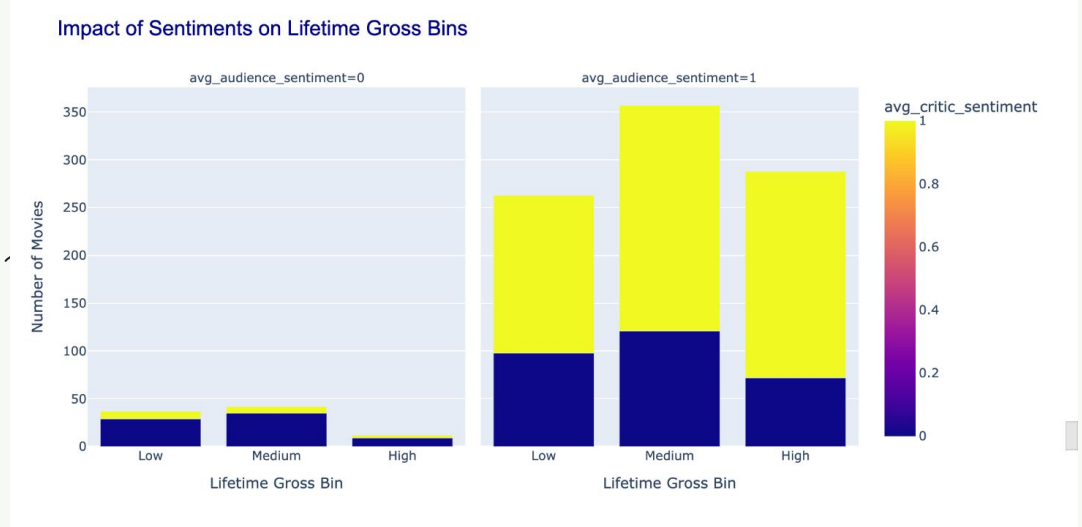
- common audience scores- 70 to 80
- Ratings below 50 are very rare

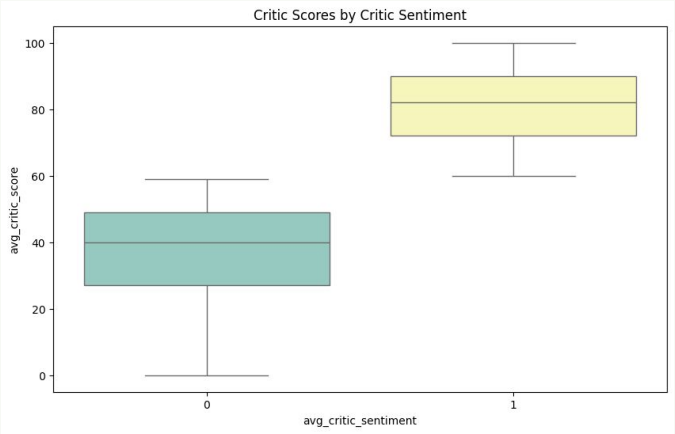


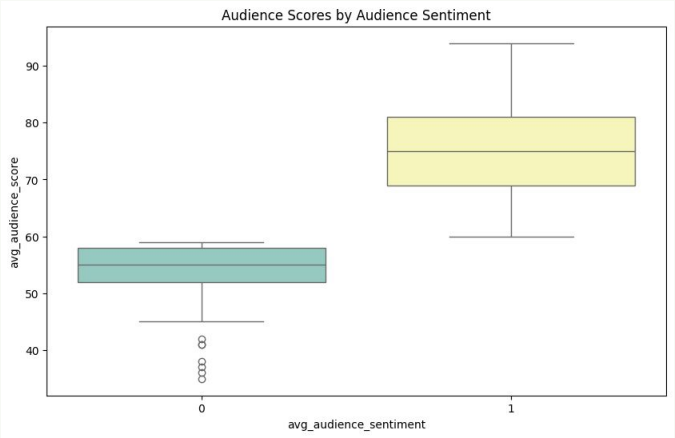
Unlike the right-skewed distribution of critic scores, audience scores are more centered. Audience scores are less extreme compared to critic scores.

Critics have a higher proportion of negative sentiment (36.44%) compared to audiences (9.11%). Audiences overwhelmingly favor movies, with over 90% of reviews being positive.

Insights on Scores Vs Sentiments







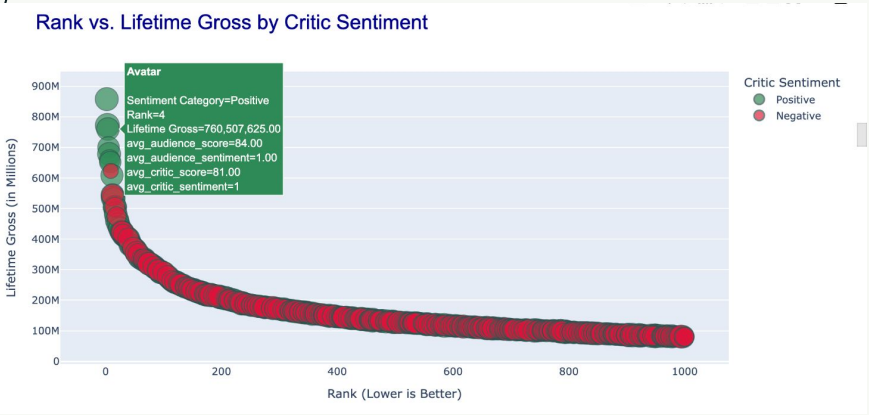
High and Medium Gross Bins: Positive sentiment - higher grossing categories, while negative sentiment - lower box-office earnings.

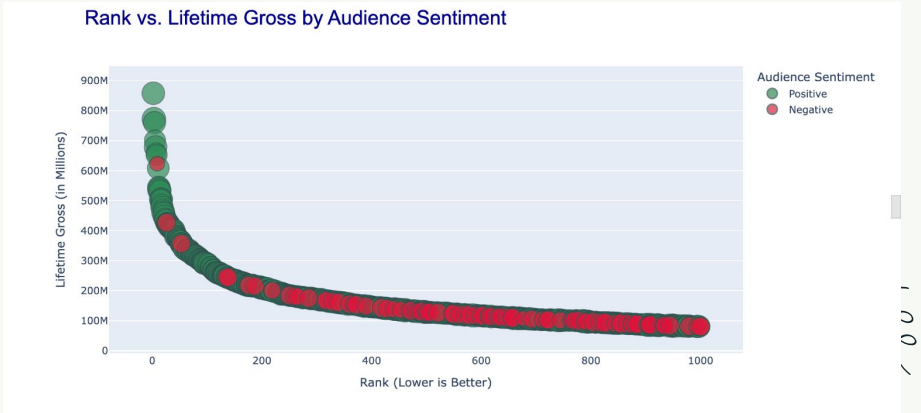
Low Gross Bin Outliers: Low Gross Bin - more negatively received by both critics and audiences.

Comparison Between Critic Sentiments: Positive critic sentiment - high critic scores. Negative critic sentiment correlates with lower scores but exhibits more variability.

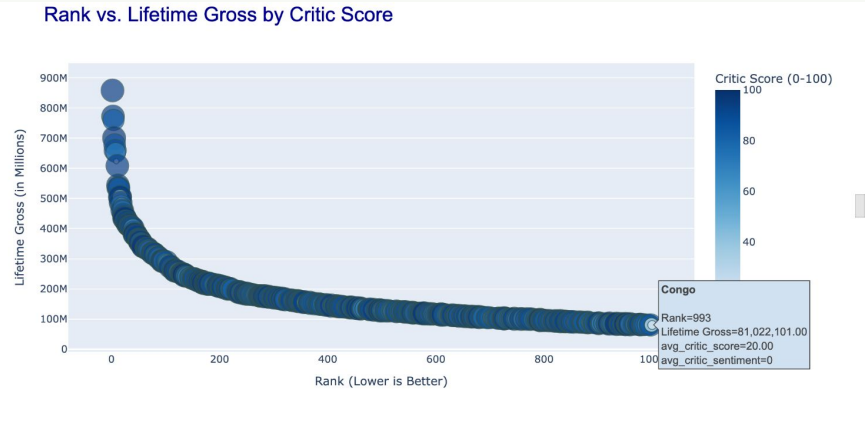
Comparison Between Audience Sentiments: Positive Sentiment -higher audience scores. Negative Sentiment cluster tightly around the lower scores.

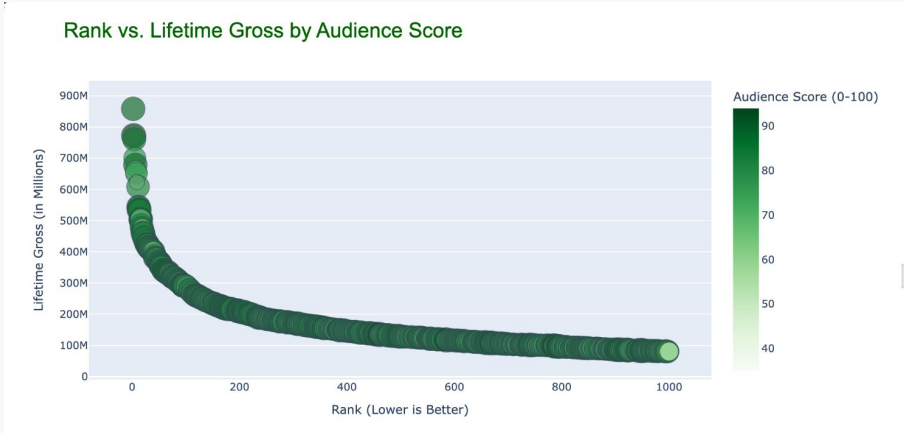
Do Rank and Revenue change with sentiment and scores?





As rank decreases, revenue increases. More positive sentiment and high scores can be seen to generate higher revenues.





Building a Predictive Model

MODEL	ACCURACY
Decision Tree	0.81
Random Forest	0.78
Naive Bayes	0.72
K-Nearest Neighbour	0.80
NN	**0.83**

Strengths of Decision Tree:

The DT model has competitive test accuracy (81%) and is easier to interpret compared to the neural network.

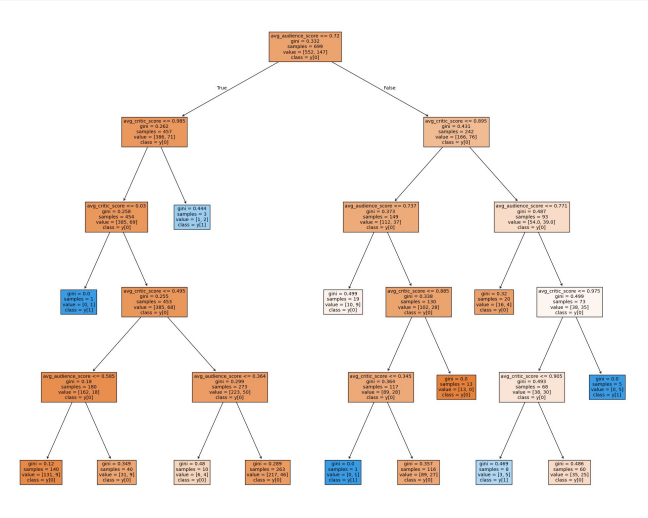
This makes it a good choice if explainability is critical.

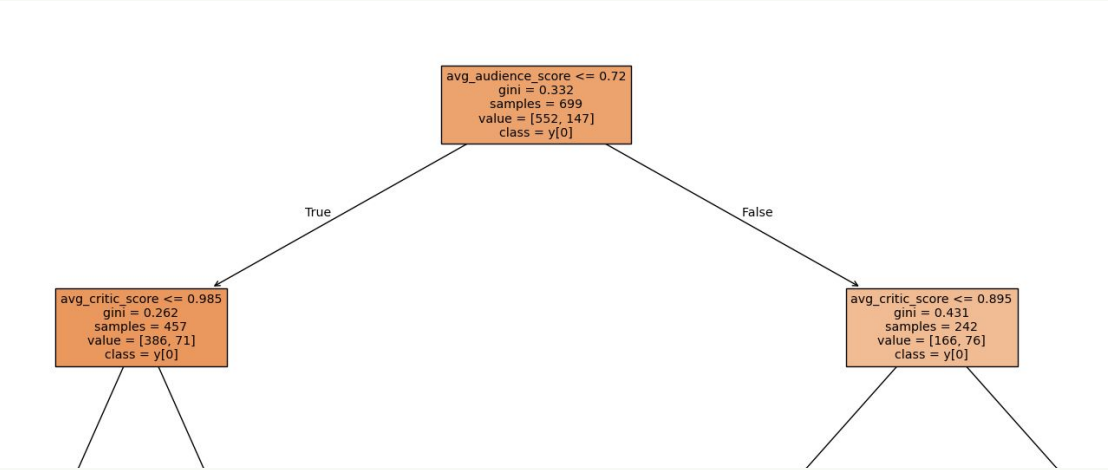
Neural Network as the Best Model:

Train Accuracy: 0.7897 (78.97%)
Test Accuracy: 0.8350 (83.50%)

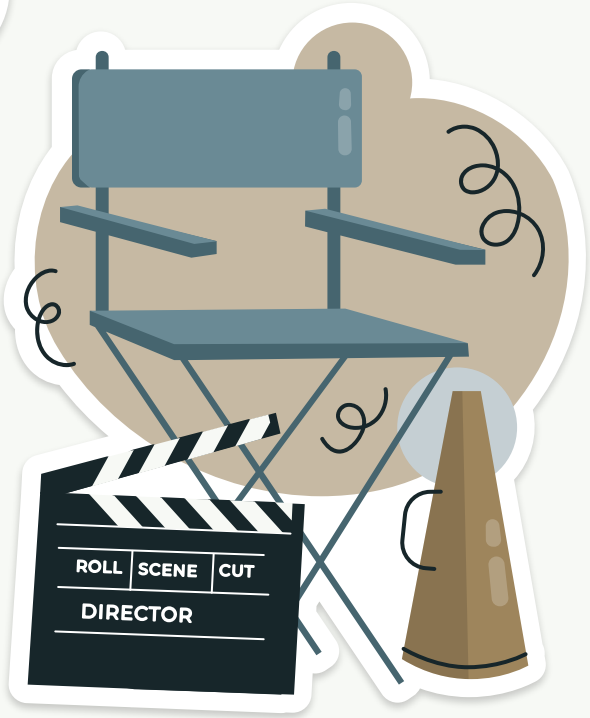
Achieves the highest test accuracy among all models, indicating strong generalization to unseen data.

Interpreting the Decision Tree





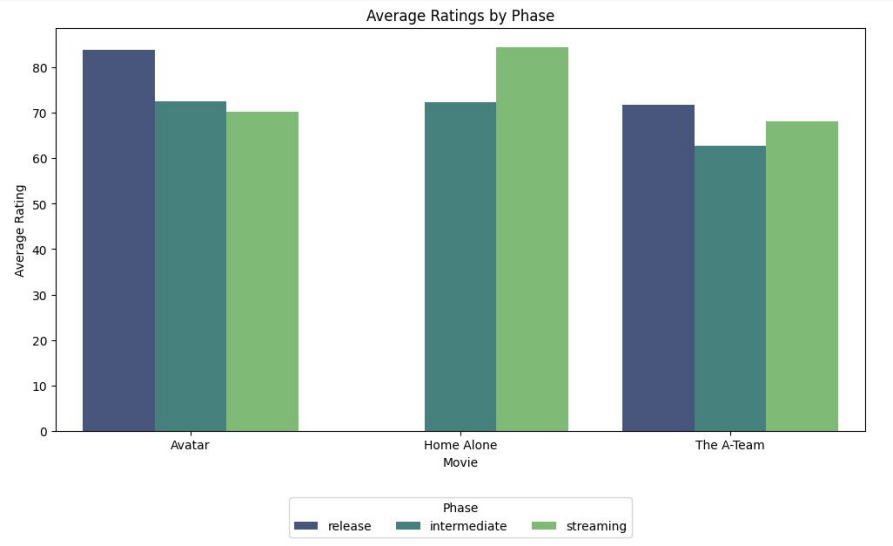
- a combination of audience and critic scores can effectively separate movies into revenue categories.
- For example, movies with high critic scores (> 0.895) or audience scores (> 0.737) are consistently associated with **higher revenues.**

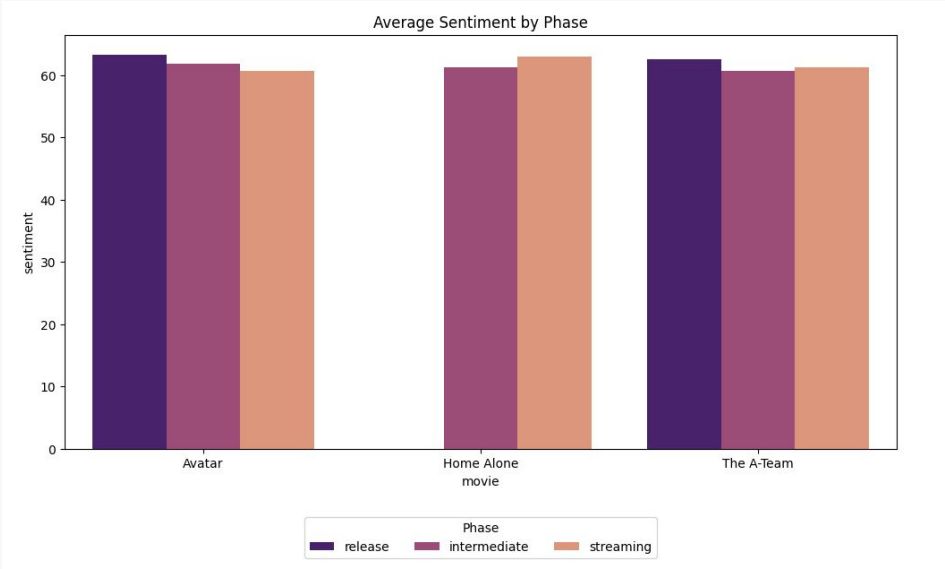


04

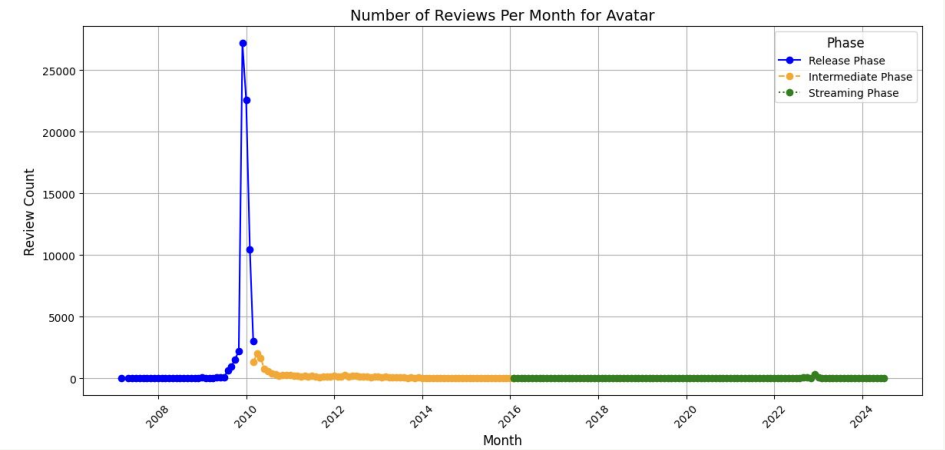
Detailed Analysis

Insights from Textual Audience Reviews





Release Phase: Covers the first three months after the movie's theatrical debut, reflecting initial audience excitement.
Intermediate Phase: Occurs between theatrical release and home media availability, such as DVD or digital sales.
Streaming Phase: Begins when the movie becomes accessible on streaming platforms, showing long-term audience interest.

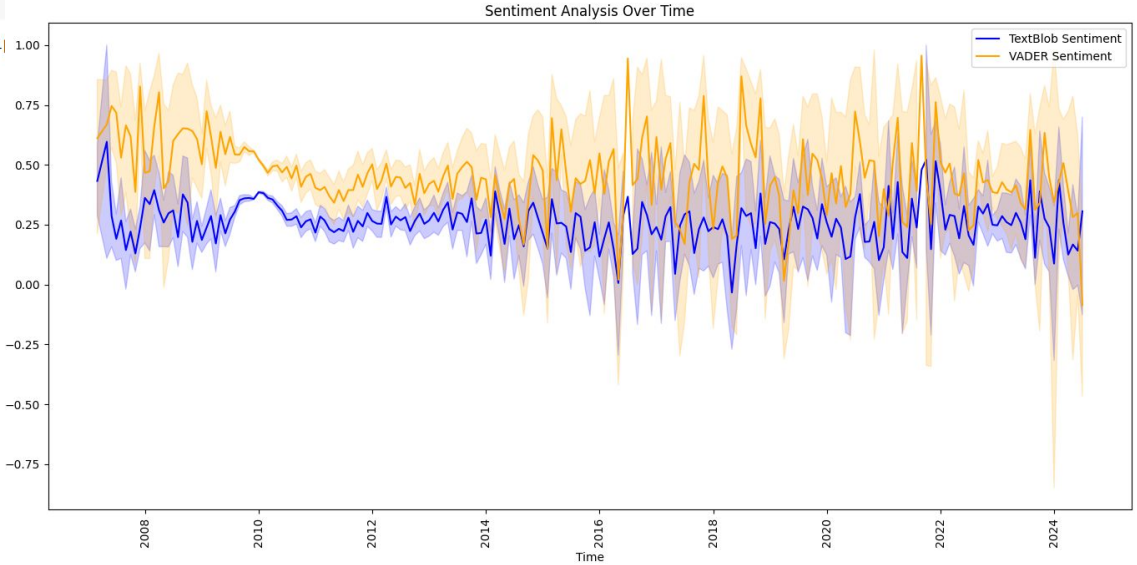


AVATAR

- Animation big thing at release in 2009 (3d)
- Streaming around 2017 shows dip in sentiment - animation not be justified on TV vs as in theatre
- Average Sentiment: 73.55

```
2  model.wv.most_similar('effect',topn=20)
```

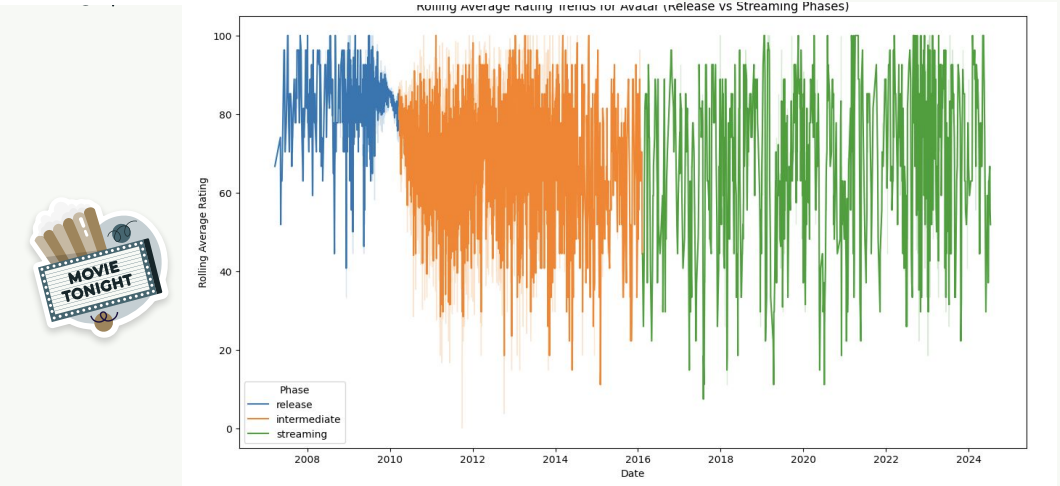
```
/usr/local/lib/python3.10/dist-packages/ipykernel/i
  and should_run_async(code)
[('fx', 0.8153723478317261),
 ('graphic', 0.7815039157867432),
 ('efect', 0.748035728931427),
 ('movieth', 0.7437546849250793),
 ('affect', 0.7248343825340271),
 ('cinematographi', 0.7090364694595337),
 ('chart', 0.7052009105682373),
 ('artwork', 0.704243540763855),
 ('specialeffect', 0.6990286111831665),
 ('totali', 0.6960344910621643),
 ('sfx', 0.6943899989128113),
 ('eyepop', 0.6790892481803894),
 ('beati', 0.6779509782791138),
 ('astonishingli', 0.6751843094825745),
 ('wellmad', 0.6748132705688477),
 ('stateoftheart', 0.6731354594230652),
 ('cgi', 0.6715654730796814),
 ('grafic', 0.6710447669029236),
 ('beuti', 0.6673539280891418),
 ('altogeth', 0.6594464182853699)]
```

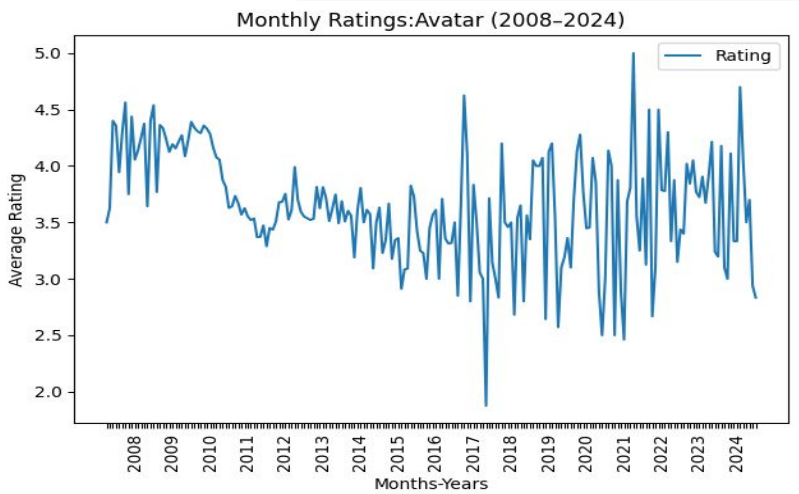


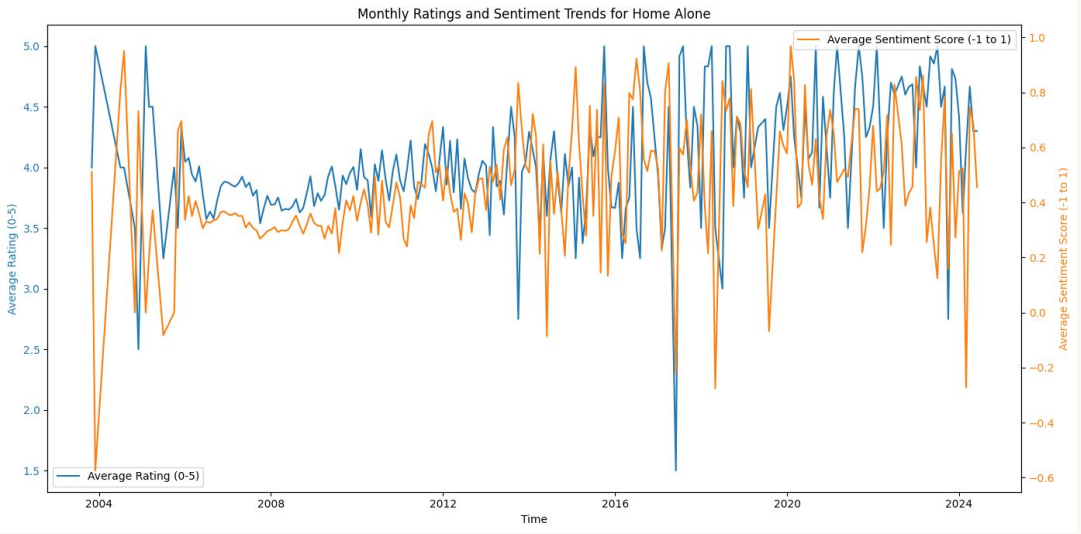
	count
movi	63653
film	29651
3d	28298
stori	24537
see	23782
avatar	17335
like	16368
ameron	16308
good	15753
visual	15737
effect	15508
great	15368
one	14591
best	13424
amaz	12826

Avatar

- Phase wise shift in user reviews : top words
- Releasing of franchise impacts original movie reviews as well
- Release of second part - Avatar: The Way of Water in 2022





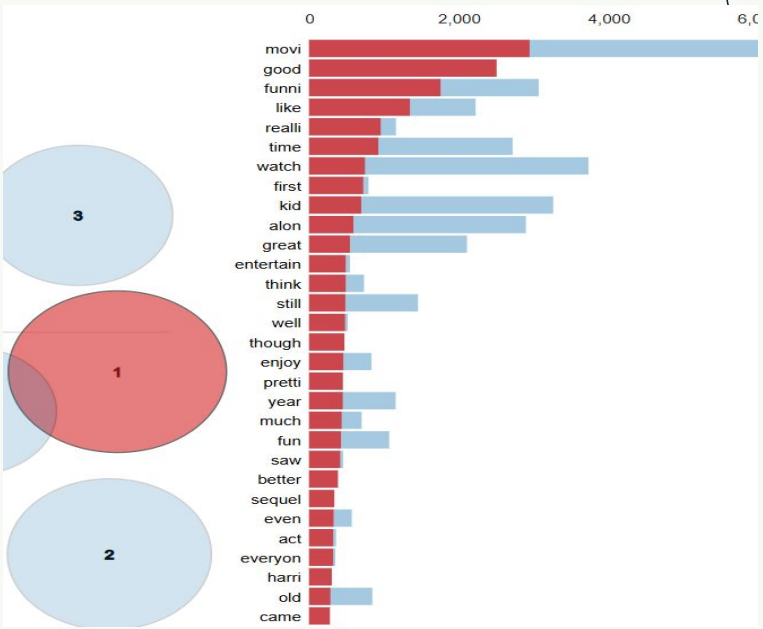


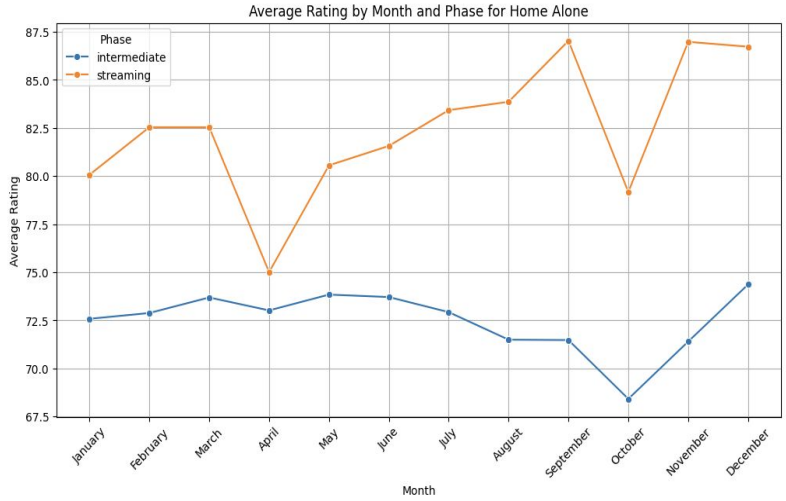
Home Alone

- Seasonal movie, sentiment peaks around December
- Re-releasing or creating spin-offs during Christmas might amplify engagement.
- Average Sentiment: 60.93

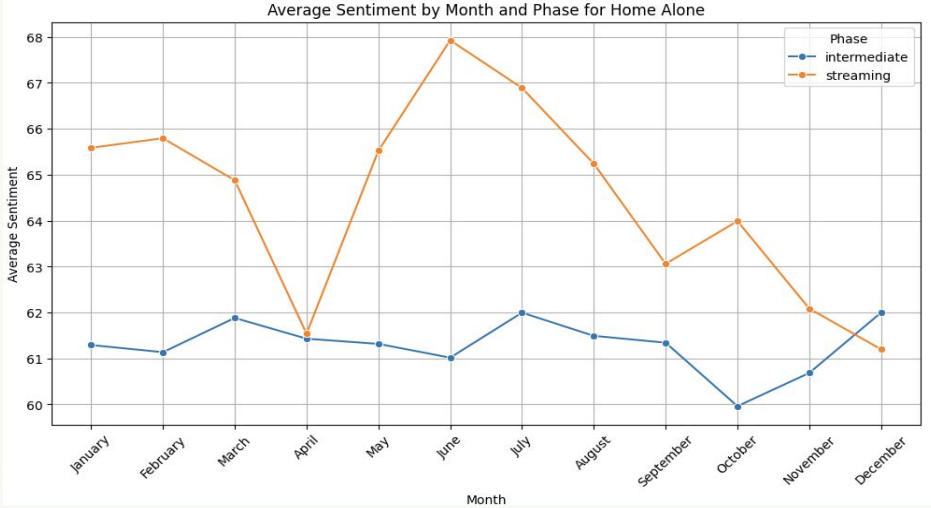
```
1
2   model.wv.most_similar('bad',topn=10)
```

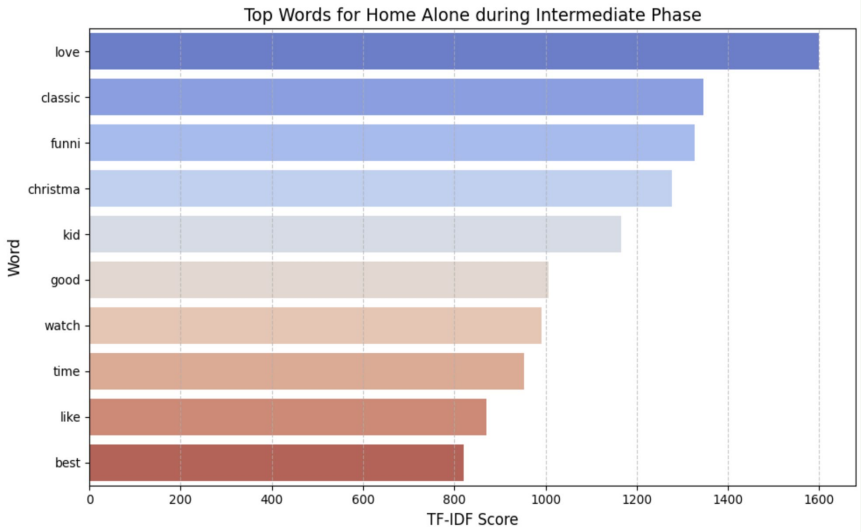
```
'usr/local/lib/python3.10/dist-packages/ip
  and should_run_async(code)
[('crap', 0.7281978130340576),
 ('beat', 0.7255004048347473),
 ('stupid', 0.694123387336731),
 ('win', 0.6937059760093689),
 ('thirsti', 0.689080536365509),
 ('kick', 0.6824852228164673),
 ('freak', 0.6620721220970154),
 ('big', 0.6566953659057617),
 ('sorri', 0.6564332842826843),
 ('grown', 0.6512394547462463)]
```

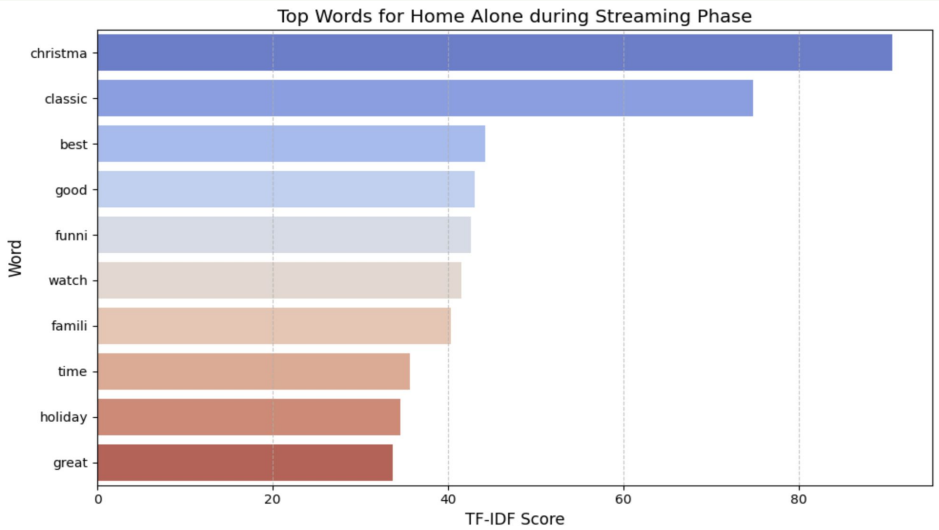



Average Rating by Month and Phase for Home Alone
Phase
intermediate
streaming
Average Rating
87.5
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
January
February
March
April
May
June
July
August
September
October
November
December
Month

HOME ALONE

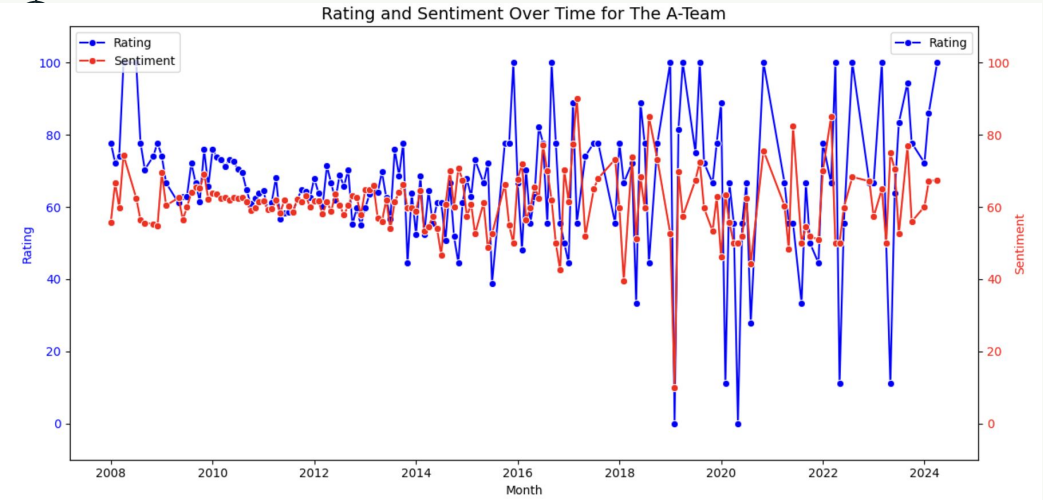

Average Sentiment by Month and Phase for Home Alone
Phase
intermediate
streaming
Average Sentiment
68
67
66
65
64
63
62
61
60
January
February
March
April
May
June
July
August
September
October
November
December
Month


Top Words for Home Alone during Intermediate Phase
Word
love
classic
funni
christma
kid
good
watch
time
like
best
0
200
400
600
800
1000
1200
1400
1600
TF-IDF Score


Top Words for Home Alone during Streaming Phase
Word
christma
classic
best
good
funni
watch
famili
time
holiday
great
0
20
40
60
80
TF-IDF Score

The A-Team

- Ahead of its time, appreciated majorly only after streaming
- Average Sentiment: 82.44, surprisingly higher compared to other movies release score
- “Great”,“good”, “Liam”, “charact”, “love”,

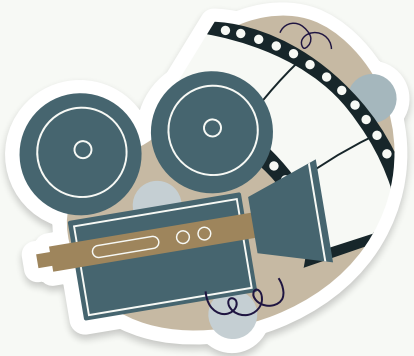



```
1
2   model.wv.most_similar('good',topn=10)
```

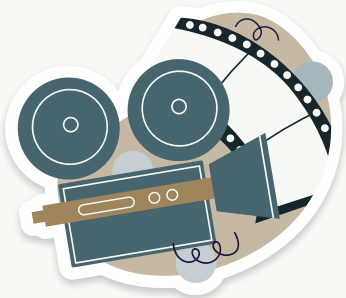
```
/usr/local/lib/python3.10/dist-packages/ipyk
  and should_run_async(code)
[('great', 0.8173941969871521),
 ('decent', 0.8167936205863953),
 ('alright', 0.7750751972198486),
 ('solid', 0.7662924528121948),
 ('sweet', 0.7333602905273438),
 ('wouldv', 0.7299250960350037),
 ('besid', 0.7269318699836731),
 ('okay', 0.7222349643707275),
 ('defin', 0.7198705673217773),
 ('awsom', 0.7189032435417175)]
```

Conclusion

What did we find and what are our challenges?



Final Thoughts



Conclusion

- There are systematic differences in scores
- Reviews by Critics and Audience successfully predict films' revenues.
- Reviews exhibit patterns influenced by the movie's lifecycle phases
- Textual analysis reveals themes and sentiments

Future Improvement Suggestions

- Additional revenue data to analyze full review dataset
- Additional sources of audience and critic reviews such as IMDB and Metacritic
- Audience sentiment towards actor(s) in highly coverage movies

Thank You!

Questions?