Rotten Tomatoes Top Movies Rating and Technical Analysis

Venkat Goutham

Introduction:

With the constant evolution of cinema and changing consumer preferences, finding an enjoyable movie to watch and unwind with can prove to be a difficult task. Luckily enough, the rise of online reviews has led to sites like Rotten Tomatoes becoming significant players in the field of entertainment. Rotten Tomatoes has garnered a reputable name for movie recommendations amongst critics and general audiences. This project will break down the most highly reviewed movies and genres featured on the site along with analyzing other findings.

Problem Statement:

The dataset examined consists of various metrics regarding the Rotten Tomatoes site ranging from cast and crew information all the way to box office revenues and acclaim level. Additionally, it contains over 900+ highly ranked movies. By breaking down critic and audience perspectives through columns such as "critic_score" and "people_score" along with others, one can see which movies ranked highly. Through doing so, I hope to identify various genres and movies that can serve as high-quality entertainment to those seeking it out. Finding a great movie to watch should no more be a herculean task!

Dataset:

The dataset observed for this project comes from Kaggle. The link to the data is provided below:

https://www.kaggle.com/datasets/thedevastator/rotten-tomatoes-top-movies-ratings-and-technical

There are a total of 1,610 observations with 26 attributes.

Attributes used are:

- **#:** This attribute displays the id of the review.
- **title:** This attribute displays the movie title.
- **year:** This attribute displays the year that the movie was released.
- **synopsis:** This attribute displays a brief synopsis of the movie.
- **critic_score:** This attribute displays the critic score (0-100).
- **people_score:** This attribute displays the viewer score (0-100).
- **consensus:** This attribute displays a summary of reviews for the movie.
- **total_reviews:** This attribute displays the total number of reviews for the movie.
- **total_ratings:** This attribute displays the total number of ratings for the movie.
- **type:** This attribute displays the type of movie.
- **rating:** This attribute displays the MPAA rating of the movie.
- **genre:** This attribute displays the genre of the movie.
- **original_language:** This attribute displays the original language of the movie.
- **director:** This attribute displays the director of the movie.
- **producer:** This attribute displays the producer of the movie.
- **writer:** This attribute displays the writer of the movie.
- **release_date_(theaters):** This attribute displays the release date of the movie.
- **release_date_(streaming):** This attribute displays the streaming release date of the movie.
- **box_office_(gross_usa):** This attribute displays the USA box office gross of the movie.
- **runtime:** This attribute displays the runtime of the movie.
- **production_co:** This attribute displays the production company of the movie.
- **sound_mix:** This attribute displays the sound mix of the movie.
- **aspect_ratio:** This attribute displays the aspect ratio of the movie
- **view_the_collection:** This attribute displays the collection of the movie (franchises etc).
- **crew:** This attribute displays the crew of the movie.
- **link:** This attribute displays the link to the review.

Language:

We will use Python to conduct and complete our data analysis for this project.

Implementation:

Our dataset consisted of 1,610 records before pre-processing and cleaning. Provided below are screenshots of the data head, tail, and description.

1.) Data Head

```
data.head()
```

	Unnamed: 0	title	year	synopsis	critic_score	people_score	consensus	total_reviews	total_ratings	type	...	release_date_(theaters)	release_date_(streaming)	box_of
0	0	Black Panther	2018	After the death of his father, T'Challa return...	96	79.0	Black Panther elevates superhero cinema to thr...	519	50,000+	Action & Adventure	...	Feb 16, 2018 wide	May 2, 2018	
1	1	Avengers: Endgame	2019	Adrift in space with no food or water, Tony St...	94	90.0	Exciting, entertaining, and emotionally impact...	538	50,000+	Action & Adventure	...	Apr 26, 2019 wide	Jul 30, 2019	
2	2	Mission: Impossible -- Fallout	2018	Ethan Hunt and the IMF team join forces with C...	97	88.0	Fast, sleek, and fun, Mission: Impossible - Fa...	433	10,000+	Action & Adventure	...	Jul 27, 2018 wide	Nov 20, 2018	
3	3	Mad Max: Fury Road	2015	Years after the collapse of civilization, the ...	97	86.0	With exhilarating action and a surprising amou...	427	100,000+	Action & Adventure	...	May 15, 2015 wide	Aug 10, 2016	
4	4	Spider-Man: Into the Spider-Verse	2018	Bitten by a radioactive spider in the subway, ...	97	93.0	Spider-Man: Into the Spider-Verse matches bold...	387	10,000+	Action & Adventure	...	Dec 14, 2018 wide	Mar 7, 2019	

box_office_(gross_usa)	runtime	production_co	sound_mix	aspect_ratio	view_the_collection	crew	link
$700.2M	2h 14m	Walt Disney Pictures	DTS, Dolby Atmos	Scope (2.35:1)	Marvel Cinematic Universe	Chadwick Boseman, Michael B. Jordan, Lupita Ny...	http://www.rottentomatoes.com/m/black_panther_...
$858.4M	3h 1m	Marvel Studios, Walt Disney Pictures	Dolby Atmos, DTS, Dolby Digital, SDDS	Scope (2.35:1)	Marvel Cinematic Universe	Robert Downey Jr., Chris Evans, Mark Ruffalo, ...	http://www.rottentomatoes.com/m/avengers_endgame
$220.1M	2h 27m	Bad Robot, Tom Cruise	DTS, Dolby Atmos, Dolby Digital	Scope (2.35:1)	NaN	Tom Cruise, Henry Cavill, Ving Rhames, Simon P...	http://www.rottentomatoes.com/m/mission_imposs...
$153.6M	2h	Kennedy Miller Mitchell, Village Roadshow Pict...	Dolby Atmos	Scope (2.35:1)	NaN	Tom Hardy, Charlize Theron, Nicholas Hoult, Hu...	http://www.rottentomatoes.com/m/mad_max_fury_road
$190.2M	1h 57m	Lord Miller, Sony Pictures Animation, Pascal P...	Dolby Atmos, DTS, Dolby Digital, SDDS	Scope (2.35:1)	NaN	Shameik Moore, Hailee Steinfeld, Mahershala Al...	http://www.rottentomatoes.com/m/spider_man_int...

2.) Data tail

```
data.tail()
```

	Unnamed: 0	title	year	synopsis	critic_score	people_score	consensus	total_reviews	total_ratings	type	...	release_date_(theaters)	release_date_(streaming)	box_
1605	1605	Priest	2011	In a society ravaged by centuries of war betwe...	15	46.0	Sleek and stylish, but those qualities are was...	101	50,000+	Western	...	May 13, 2011 wide	Apr 16, 2012	
1606	1606	September Dawn	2006	In 1857 Capt. Alexander Fancher leads a wagon ...	16	49.0	With its jarring editing, dull love story, and...	55	5,000+	Western	...	Jun 22, 2007 wide	Jan 1, 2008	
1607	1607	American Outlaws	2001	After the Civil War ends, Confederate soldiers...	14	68.0	With corny dialogue, revisionist history, anac...	103	25,000+	Western	...	Aug 17, 2001 wide	Dec 1, 2017	
1608	1608	Jonah Hex	2010	Having cheated death, gunslinger and bounty hu...	12	20.0	Josh Brolin gives it his best shot, but he can...	152	100,000+	Western	...	Jun 18, 2010 wide	Mar 30, 2012	
1609	1609	Texas Rangers	2001	Texas, 1875. In a land without justice, where ...	2	29.0	As far as westerns go, Texas Rangers is strict...	51	5,000+	Western	...	Nov 30, 2001 wide	Oct 8, 2016	

box_office_(gross_usa)	runtime	production_co	sound_mix	aspect_ratio	view_the_collection	crew	link
$29.1M	1h 27m	Michael De Luca, Stars Road Entertainment	SDDS, Dolby Digital	NaN	NaN	Paul Bettany, Karl Urban, Cam Gigandet, Maggie...	http://www.rottentomatoes.com/m/10009274-priest
$1.1M	1h 50m	Voice Pictures Inc., September Dawn LLC	NaN	NaN	NaN	Jon Voight, Trent Ford, Tamara Hope, Jon Gries...	http://www.rottentomatoes.com/m/september_dawn
$13.3M	1h 33m	Morgan Creek Productions	Dolby Stereo, Dolby A, SDDS, DTS, Surround, Do...	Flat (1.85:1)	NaN	Colin Farrell, Scott Caan, Ali Larter, Gabriel...	http://www.rottentomatoes.com/m/american_outlaws
$10.5M	1h 21m	Mad Chance, Weed Road Pictures	NaN	NaN	NaN	Josh Brolin, John Malkovich, Megan Fox, Michae...	http://www.rottentomatoes.com/m/jonah_hex
$623.4K	1h 30m	Greisman Productions, Price Entertainment, Lar...	Dolby Stereo, Dolby Digital, Dolby A, Surround...	Scope (2.35:1)	NaN	James Van Der Beek, Dylan McDermott, Usher Ray...	http://www.rottentomatoes.com/m/1111103-texas_...

3.) Data description

```python
data.describe()
```

	Unnamed: 0	year	critic_score	people_score	total_reviews
count	1610.000000	1610.000000	1610.000000	1609.000000	1610.000000
mean	804.500000	1991.745963	92.693789	83.405221	143.652174
std	464.911282	28.054120	11.621759	11.263792	118.137144
min	0.000000	1919.000000	2.000000	10.000000	39.000000
25%	402.250000	1969.000000	92.000000	80.000000	56.000000
50%	804.500000	2005.000000	96.000000	87.000000	90.000000
75%	1206.750000	2014.000000	98.000000	91.000000	205.750000
max	1609.000000	2020.000000	100.000000	98.000000	561.000000

The describe function additionally provides us with key metrics regarding the dataset such as count, mean, standard deviation (std), minimum (min), 25%, 50%, 75%, and maximum (max) values.

Dataset Preprocessing:

The first step taken within this dataset was to identify missing values as they can often lead to incomplete results and erroneous interpretations.

```python
print(data.isnull().sum()) #to check count of missing values
```

```
Unnamed: 0                    0
title                         0
year                          0
synopsis                      8
critic_score                  0
people_score                  1
consensus                    17
total_reviews                 0
total_ratings                 0
type                          0
rating                      471
genre                         7
original_language            40
director                      1
producer                    120
writer                      344
release_date_(theaters)     507
release_date_(streaming)     15
box_office_(gross_usa)      508
runtime                       7
production_co               123
sound_mix                   685
aspect_ratio                946
view_the_collection        1432
crew                          0
link                          0
```

The next step is to drop columns with a majority or a significant of their entries missing. For instance, “sound_mix”, “aspect_ratio”, “view_the_collection”, “release_date_(theatres)”, “box_office_(gross_usa)”, “writer”, and “rating” all have a significant amount of missing data that could lead to inconsistent findings later on, so for the project, I chose to drop them.

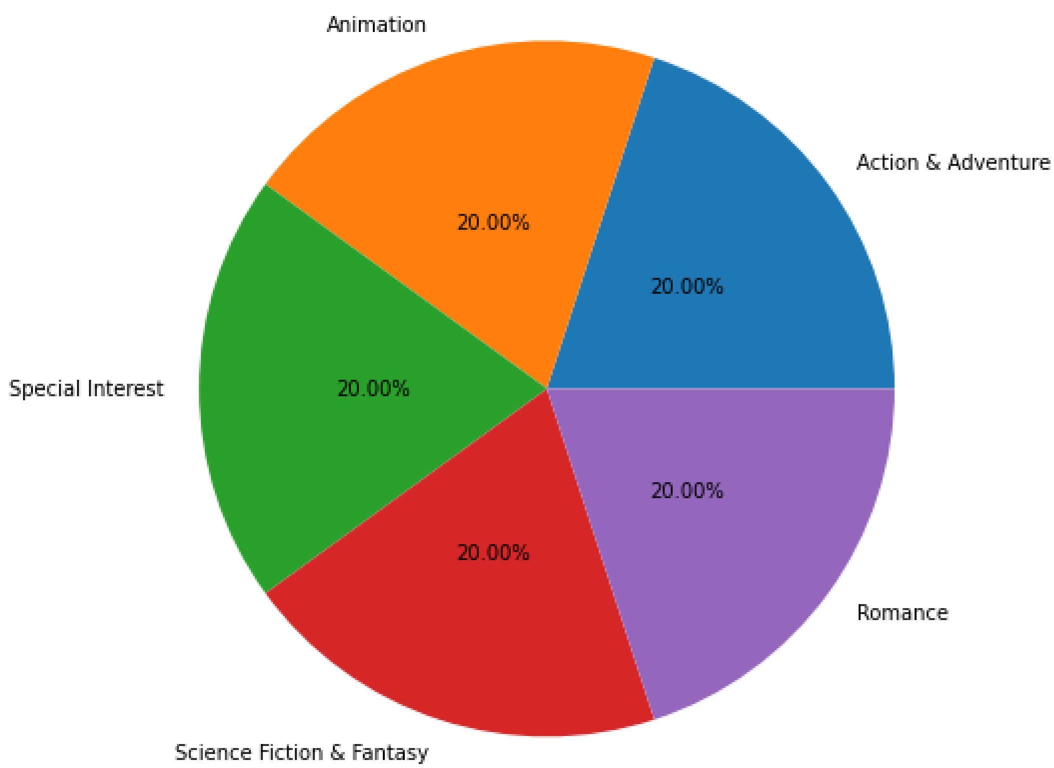
```
filtered_data = data.drop(['rating','writer',
                           'release_date_(theaters)','box_office_(gross_usa)',
                           'sound_mix','aspect_ratio','view_the_collection'], axis=1)
filtered_data = filtered_data.dropna(axis=0, how='all')
filtered_data.shape
filtered_data.isnull().all(axis=0)
```

```
Unnamed: 0                   False
title                        False
year                         False
synopsis                     False
critic_score                 False
people_score                 False
consensus                    False
total_reviews                False
total_ratings                False
type                         False
genre                        False
original_language            False
director                     False
producer                     False
release_date_(streaming)     False
runtime                      False
production_co                False
crew                         False
link                         False
```

We can see that the cleaning process went smoothly as there are no longer any insignificant values prohibiting us from obtaining key insights. After all the cleaning, we are left with the shape of the dataframe as 1,610 records with 19 attributes. Now, I will move to analyzing the dataset and configuring visuals to represent our findings which in turn, will help resolve the problem statement.

Data Analysis

Figure 1: The 5 most highly-ranked genres



Findings:

- We can see a uniform distribution in the results of the pie chart, this indicates that the dataset is set up with minimal bias as it covers a good percentage of all genres equally.
- Narrowing down movie options into genres can be helpful for those deciding on which film to watch. We can see that the top 5 highly rated genres belong to Animation, Action & Adventure, Special Interest, Romance, Science Fiction & Fantasy. I recommend movies in

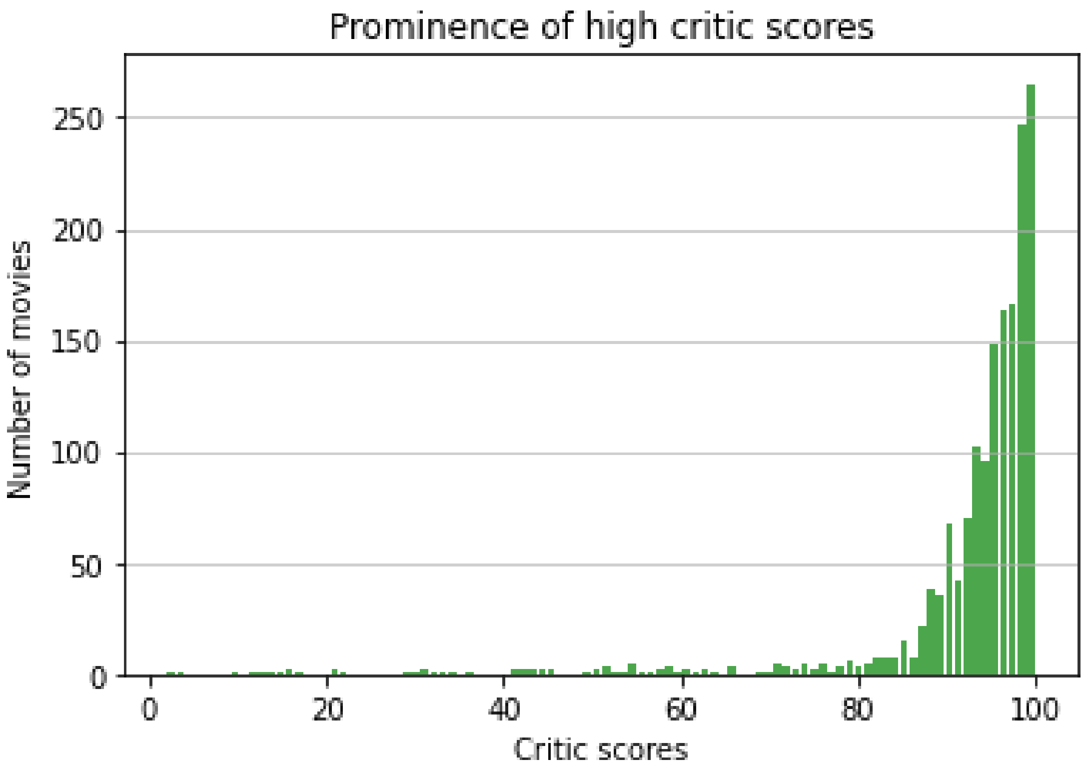
these genres for those individuals seeking out entertainment as they seem to be acclaimed in comparison to other genres.

Figure 2: Wordcloud of the review consensus

Findings:

- Here we can see the most prominent elements of a positive consensus displayed through Rotten Tomatoes. This is a sign of a good dataset with plenty of options for users to watch as it includes appealing features such as “funny”, “classic”, “entertaining, and “smart” along with a host of other words. Performance and story also show us how much the critics value acting prowess and a well-structured story. Therefore, the movies on this list are great to watch for both critics and audiences alike!

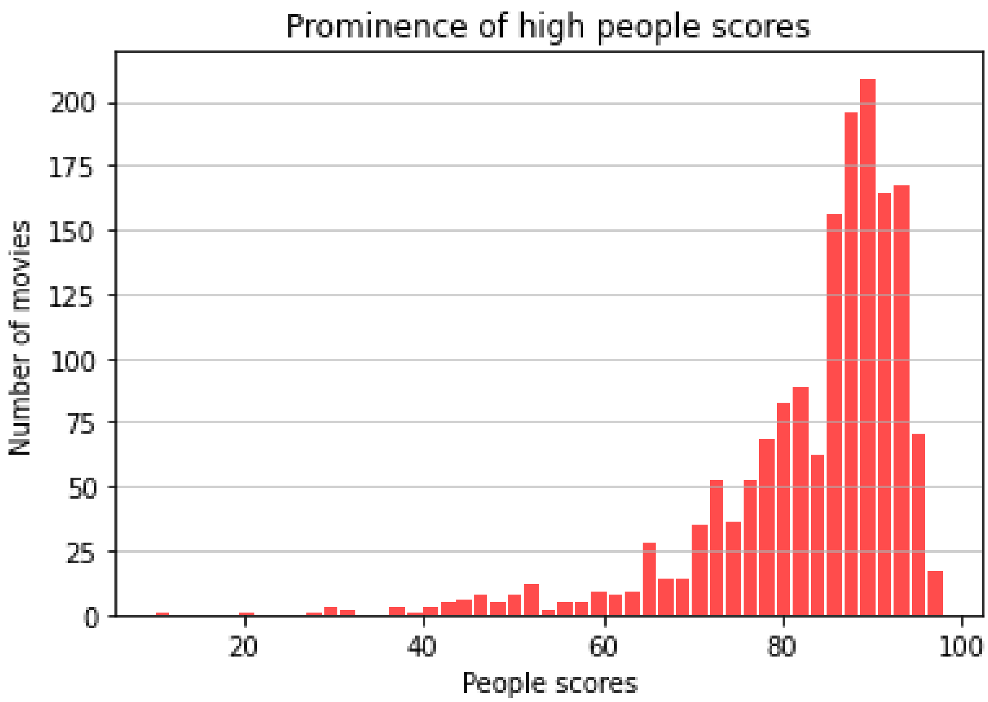
Figure 3: Prominence of high critic scores



Findings:

- The majority of ratings for the movies in the dataset are on the higher end (75+), with only a few movie entries receiving bad to below-average reviews. This left-skewed distribution is an indicator of a quality dataset as we can see that even if a user were to pick out a movie at random, there would be a good chance that the film selected is critically acclaimed to some extent. This is an excellent advantage as it makes the movie selection process much easier for confused or overwhelmed individuals.

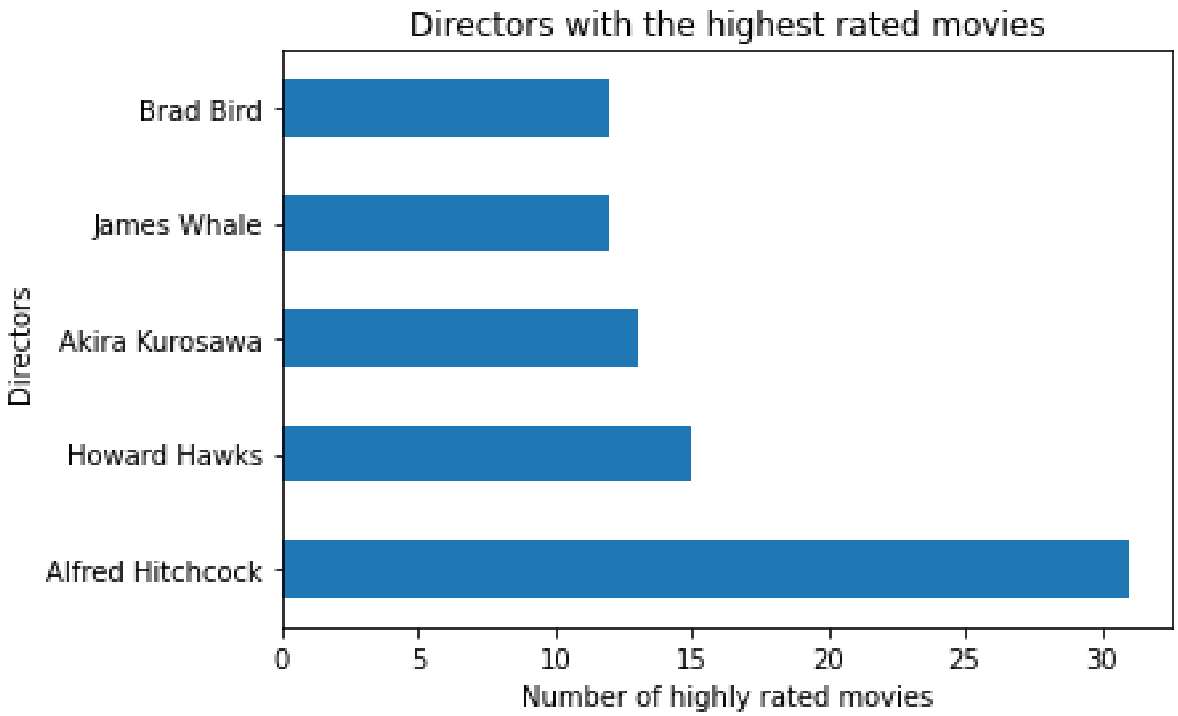
Figure 4: Prominence of high people scores



Findings:

- The majority of ratings for the movies in the dataset (from an audience perspective) are still on the higher end (75+). We can see a clear difference between audience mindsets and critic mindsets with the number of almost-perfect scores. Critics tended to give out scores near 100 more frequently than the audiences did, this could mean that different movies impressed different sects of viewers but regardless, this left-skewed distribution (similar to figure 3) is yet another indicator of a quality dataset. Not everyone watching movies will be approaching them from a critical POV, so it is a good sign to know that the general audiences, our intended group of individuals, are also highly enjoying the movies on this list as well.

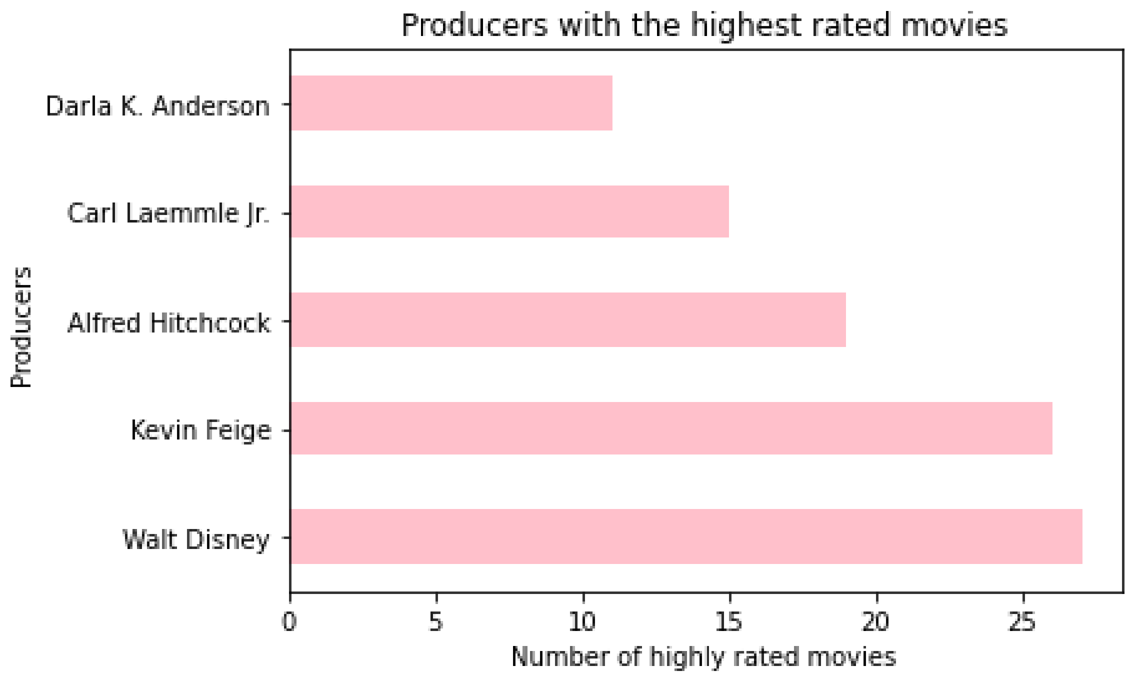
Figure 5: Directors with the highest-rated movies



Findings:

- The top 5 directors with the highest-rated movies in the dataset are featured in this visual. Alfred Hitchcock takes the lead with 30 critically acclaimed movies whereas Brad Bird has the least amount of high-rated movies (concerning this visual specifically). This is still an impressive feat as the 4 directors after Alfred Hitchcock is hovering around the same number of movies. The most impressive part is Hitchcock having double the amount of second place! I suggest any of these five directors' films to watch when faced with too many options, as we can see that they are highly revered in their respective crafts.

Figure 6: Producers with the highest-rated movies



Findings:

- The top 5 producers with the highest-rated movies in the dataset are featured in this visual. I chose to analyze producers for this dataset since sometimes directors may not have established names (in figure 5, the top 5 directors are highly established with well-known bodies of work) and instead, looking at the production house behind a movie can prove to be just as insightful concerning product quality. Walt Disney has the highest movie count with Kevin Feige leading second; both bring over 25 critically acclaimed films in this dataset alone. These five producers are most reliable with the content they put out into the entertainment field. When faced with a debut or unfamiliar director, I suggest looking to see if any of these five producers are associated with said film. These producers can provide a wide body of films to watch when faced with too many options, as we can see that they are highly revered in their crafts.

<u>Algorithm Selection - Naive Bayes Classifier</u>

For selecting the algorithm, I chose to implement a Naive Bayes Classifier. I chose this classifier as I was interested in the analysis of the textual data. My classifier uses a count vectorizer which results in an input dataframe that contains a lot of features. Naive Bayes classifiers perform well with data that has a large feature set, thus I chose this as our method to classify our text data. I used the review text data as input for the model to predict the rating of movie reviews from a critic's POV. To simplify the problem and ensure a readable and more applicable result from our model, we preprocessed the data in the following way: review ratings of 95 or higher were labeled as positive, and the rest were labeled as negative reviews. The high threshold of 95 might seem severe at first glance but this is reasonable since the dataset contains highly acclaimed movies for the most part. As a result, most of the reviews will be positive, to begin with anyways, but by creating a high threshold, we can see the best of the best.

```
predictor_data = filtered_data
predictor_data = predictor_data.dropna(axis=0, how='any')
def label_review(row):
    if row['critic_score'] >= 95:
        return 'positive'
    else:
        return 'negative'


# Add the label column to the DataFrame
predictor_data['review_class'] = predictor_data.apply(label_review, axis=1)
predictor_data.head()
```

We then built a pipeline that count vectorized (tokenized) the text data, applied a TFIDF transformation, and applied a Multinomial Naive Bayes Classifier with text data as input to predict if a review was positive or negative. To limit over and underfitting, the input data fed into the model followed a training testing split of 0.8 and 0.2 respectively. Additionally, to set our model parameters with the optimal values to prevent over and underfitting, we used GridSearchCV to find the optimal hyperparameter value for the alpha value.

```
def create_nb_classifier(df, test_size=0.2):
    pipeline = Pipeline([
        ('vectorizer', CountVectorizer()),
        ('tfidf', TfidfTransformer()),
        ('classifier', MultinomialNB())
    ])

    X = df['consensus']
    y = df['review_class']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size)
    # Grid Search for Optimal Hyperparameters
    param_grid = {
    'classifier__alpha': [0.1, 1.0, 10.0]
    }
    grid_search = GridSearchCV(pipeline, param_grid, cv=5)
    grid_search.fit(X_train, y_train)
    classifier = grid_search.best_estimator_
    return classifier, X_test, y_test
```

The Naive Bayes classifier is the algorithm recommendation for this data because it is optimal with text data and results in high accuracy and performance.

Performance Evaluation

We used the outputted predicted values of the model to measure accuracy against the actual values. We also printed a classification report that displayed the precision, recall, f1-score, and support for both positively classified and negatively classified examples.

```python
vals = create_nb_classifier(nbdata)
classifier = vals[0]
new_reviews = vals[1]
testval = vals[2]
predictions = classifier.predict(new_reviews)
accuracy = accuracy_score(testval, predictions)
print(f'Accuracy: {accuracy:.2f}')
print(classification_report(testval, predictions))
```

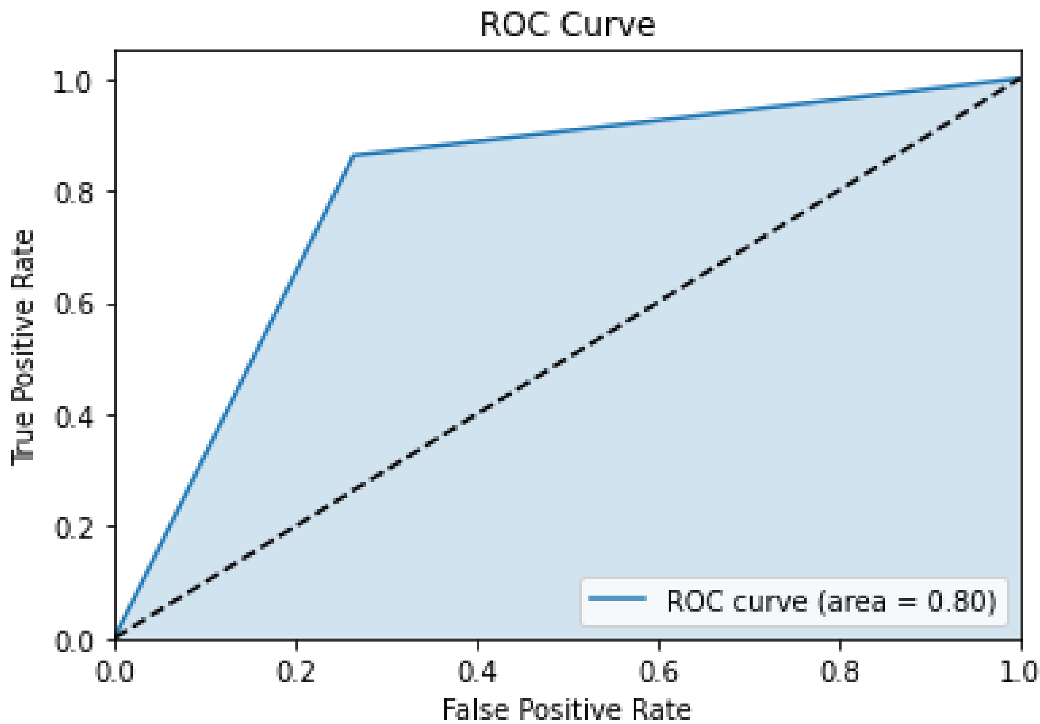
```
Accuracy: 0.87
              precision    recall  f1-score   support

    negative       0.92      0.76      0.83       119
    positive       0.84      0.95      0.89       154

    accuracy                           0.87       273
   macro avg       0.88      0.86      0.86       273
weighted avg       0.87      0.87      0.87       273
```

As the above data shows, the model performs well with an overall accuracy of .87. The precision of the model is also quite high for both classes. The model performs better for positive classes than negative based on the recall and f1-score. This is to be expected since the positive class has a higher distribution (as told by the support values), thus the model is better at classifying positive reviews accurately.

To further analyze the performance of our model, we used a ROC curve. An AUC (Area Under the Curve) of .5 indicates a model that performs randomly whereas an AUC of .1 indicates a perfect model. Based on the figure below, we can determine that the AUC of our ROC curve indicates a well-performing model with an AUC value of .80.



Conclusion & Recommendations

- One large positive was the frequency of positive reviews in the total dataset (high critic and audience scores as shown in Figures 3 & 4). This concludes that a lot of the movie options presented in the dataset do have the viability to them.
- Based on the visualizations and data analysis completed in this project, we recommend that viewers choose options within these 5 highly rated genres: Animation, Action & Adventure, Special Interest, Romance, Science Fiction & Fantasy. As shown in Figure 1, we can see the acclaim that these genres have and therefore, provide a wide range of options for viewers to watch while relaxing.
- Figure 2 also provides key insights behind the most prominent elements of a positive consensus displayed through Rotten Tomatoes. The wordcloud includes appealing movie traits such as “funny”, “classic”, “entertaining, and “smart” giving viewers a complete

package of a film. Additionally, performance and story hold weight in the cloud as well. All these features ensure that a majority of the films in the dataset are widely appealing and acclaimed highly.

- Another factor to consider when choosing the movie is the crew behind it, and oftentimes, the captain of the ship is the Director. Based on our Figure 5 findings, we recommend viewers watch films from the top 5 most acclaimed directors in the dataset (in descending order): Alfred Hitchcock, Howard Hawks, Akira Kurosawa, James Whale, and Brad Bird. These established directors will provide some great options to enjoy.
- Lastly, the producer also plays a key part in a film's success. It can be beneficial to know which producers are reliable with movie content when faced with an unfamiliar director or body of work. Figure 6 provides viewers with the 5 most reliable producers in the dataset (in descending order): Walt Disney, Kevin Feige, Alfred Hitchcock, Carl Laemmle Jr, and Darla K. Anderson. These established directors will provide some great options to enjoy.
- Based on the Naive Bayes classifier built on the text data, we can also conclude that review text data is a great predictor of whether or not a review will be classified as positive or negative concerning the review scale. Although this may seem obvious, this is could prove useful in the future if we are given an unclassified review text and need to determine if it is positive or negative.
- Overall, with these findings, we believe that stressed-out viewers looking to unwind can narrow down their movie options with little to no worries.