**Movie Success Rate Prediction System**

**Abstract:**

The film business is a billion-dollar business, and extensive measure of data identified with motion pictures is accessible over the web. In this system we are analyzing the dataset for predicting the success of the movies. For doing this the analysis of the dataset is done in which the chronicled information of every segment, for example, actor, actress, director, music that impacts the achievement or disappointment of a motion picture is given weight age and after that dependent on different parameters we are predicting whether the movie will be a flop, average or superhit. Certain algorithms are used that can help to predict whether the movies will be a flop, average, or superhit. In this model we focus on the attribute selection for predicting success of the movies. A comparative analysis is to be performed so as to find the accurate results among the algorithms used. Few parameters that are important for predicting success of a movie are gross, genres, release date, star powers of actors, actress, directors, and budget etc. In the dataset there are 28 parameters. The task is to find out most relevant parameters

**Introduction:**

Film industry largely affects our general public. Distinctive assortments of motion picture are discharged for the current year. Hollywood is seen as the most prepared film industry of the world, and the best similarly as film industry net gain however Indian film is seen as the best film industry in regards to the amount of motion pictures made and the amount of tickets sold.

In spite of the fact that motion pictures are giving different diverse classes however what makes a film effective? These days, as film industry is becoming excessively quick, there are substantial measure of data accessible on the web, which can be commonly utilized for data examination. Motion picture achievement forecast is an exceptionally confused undertaking to do. The meaning of achievement of a film is relative, a few motion pictures are called effective dependent on its overall gross salary, and a few motion pictures may not.

Different motion picture database utilizes clicks, audits, web journals, star throws, remarks to anticipate yet connected four data mining procedures to the dataset. The data mining methods that are utilized are Decision Tree regressor, Lasso Regressor, Random Forest, Support vector Regression (SVR).

**Literature survey:**

In the past many researchers have tried to identify features that can be used to predict the success of a movie and have computed correlations between those variables

* K Meenakshi published paper which makes use of k-means clustering and decision tree algorithms and makes use of dataset consisting of 30000 records and the results are calculated from success rate ranging from flop to hit. Decision trees are complex especially in preparing decision trees with large branches, are complex and time consuming. Determining the splitting criterion for each node in a decision tree is complicated task that require more expertise and experience. Decision trees examine only a single field at a time, this may not correspond well with the actual distribution of records in the decision space.
* Vr, Nithin & Babu Pb, Sarath adopted many applications of machine learning such as linear regression and logistic regression. The accuracy of their model using linear regression was 51%. While using logistic regression, they obtained about 42.2% accuracy. Sequel of movies cannot not be predicted from this model. Predicting only on the basis of one attribute that is gross revenue but success of movie should be relative and therefore cannot be predicted by using only one attribute.
* Suhaas Prasad tried to propose a system capable of predicting movie ratings based on user rating histories. A typical filtering strategy is the k-nearest neighbor (KNN) technique in which the user-item inclination is dictated by looking at the evaluations of comparative users or items.The information for the user's appraisals and interpersonal organizations have been given by Flixster, a social motion picture stage where users can rate and audit films with their companions. For the baseline item-based KNN the RMSE comes out to be approximately 0.989. Flixster users may not represent one’s true social graphs, better to consider Facebook users.

**Models:**

The Random Forest Classifier is a part of the scikit-learn library in Python, which is widely used for machine learning tasks. Below is comprehensive information about the Random Forest Classifier in Python:

What is Random Forest Classifier?

Definition: RandomForestClassifier is an ensemble learning method used for classification tasks. It fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Key Parameters:

n\_estimators: The number of trees in the forest (default=100).

criterion: The function to measure the quality of a split (default='gini' for Gini impurity or 'entropy' for information gain).

max\_depth: The maximum depth of the tree.

min\_samples\_split: The minimum number of samples required to split an internal node.

min\_samples\_leaf: The minimum number of samples required to be at a leaf node.

max\_features: The number of features to consider when looking for the best split.

bootstrap: Whether bootstrap samples are used when building trees.

random\_state: Seed of the random number generator for reproducibility.

n\_jobs: The number of jobs to run in parallel for both fit and predict.

Methods:

fit(X, y): Build a forest of trees from the training set (X, y).

predict(X): Predict class for X.

predict\_proba(X): Predict class probabilities for X.

score(X, y): Returns the mean accuracy on the given test data and labels.

feature\_importances\_: The impurity-based feature importances.

Example Usage:

python

Copy code

from sklearn.ensemble import RandomForestClassifier

# Initialize RandomForestClassifier

rf\_classifier = RandomForestClassifier(n\_estimators=100, criterion='gini', max\_depth=None, min\_samples\_split=2, min\_samples\_leaf=1, max\_features='auto', bootstrap=True, random\_state=42)

# Train the classifier

rf\_classifier.fit(X\_train, y\_train)

# Make predictions

y\_pred = rf\_classifier.predict(X\_test)

# Calculate accuracy

accuracy = rf\_classifier.score(X\_test, y\_test)

print("Accuracy:", accuracy)

Advantages:

Handles large datasets with high dimensionality well.

Reduces overfitting compared to individual decision trees.

Provides feature importance scores.

Disadvantages:

Less interpretable compared to single decision trees.

Slower to train compared to single decision trees.

Not suitable for very small datasets.

**Methodology:**

1. **Data attainment**

Dataset is basically taken from Kaggle. This dataset

contains 839 records and 28 parameters which includes

historical data of each component such as actors, actress,

budget, production cast, genres etc.

Few parameters are

* Title
* Genre
* Description
* Director
* Actors
* Year
* Runtime (Minutes)
* Rating
* Votes
* Revenue (Millions)
* Metascore
* Action
* Adventure
* Animation
* Biography

1. **Data Exploration**

Data exploration is the preliminary step in the data analysis process, aimed at understanding the structure and characteristics of a dataset. It involves techniques such as descriptive statistics, data visualization, and identifying patterns and anomalies within the data. By exploring the data, we gain insights that inform subsequent analysis and modeling decisions.

1. **Data processing**

Data processing refers to the manipulation and transformation of raw data into a more useful format for analysis, visualization, or other applications. This process involves several steps, including:

* Data Collection: Gathering raw data from various sources such as databases, files, sensors, or web APIs.
* Data Cleaning: Identifying and correcting errors, inconsistencies, or missing values in the dataset. This may involve tasks like removing duplicates, filling in missing values, and standardizing formats.
* Data Transformation: Converting data into a more suitable format for analysis. This may include scaling numerical values, encoding categorical variables, or extracting features from text or images.
* Data Integration: Combining data from multiple sources into a single dataset. This may involve merging datasets based on common identifiers or keys.
* Data Reduction: Reducing the size of the dataset while preserving its important characteristics. This may involve techniques such as feature selection or dimensionality reduction.
* Data Aggregation: Combining individual data points into summary statistics or aggregates. This is often done to analyze data at a higher level of granularity.
* Data Normalization: Scaling data to ensure that different variables have comparable ranges. This is important for certain analysis techniques like clustering or neural networks.
* Data Validation: Checking the quality and integrity of the processed data to ensure that it is accurate and reliable for analysis.

1. **Feature Extraction**

In this phase we will select that attributes out of 28

parameters which can be used to predict success of movie

accurately. We will be using both pre-released as well as

post released attributes. We take into consideration scores

from various movie database websites for a particular

movie. We also consider twitter reviews, likes, retweet count

for a particular movie. The date is likewise a huge factor in

the matter of film industry. Movies released before any

occasion seems to have a greater chance to be fruitful

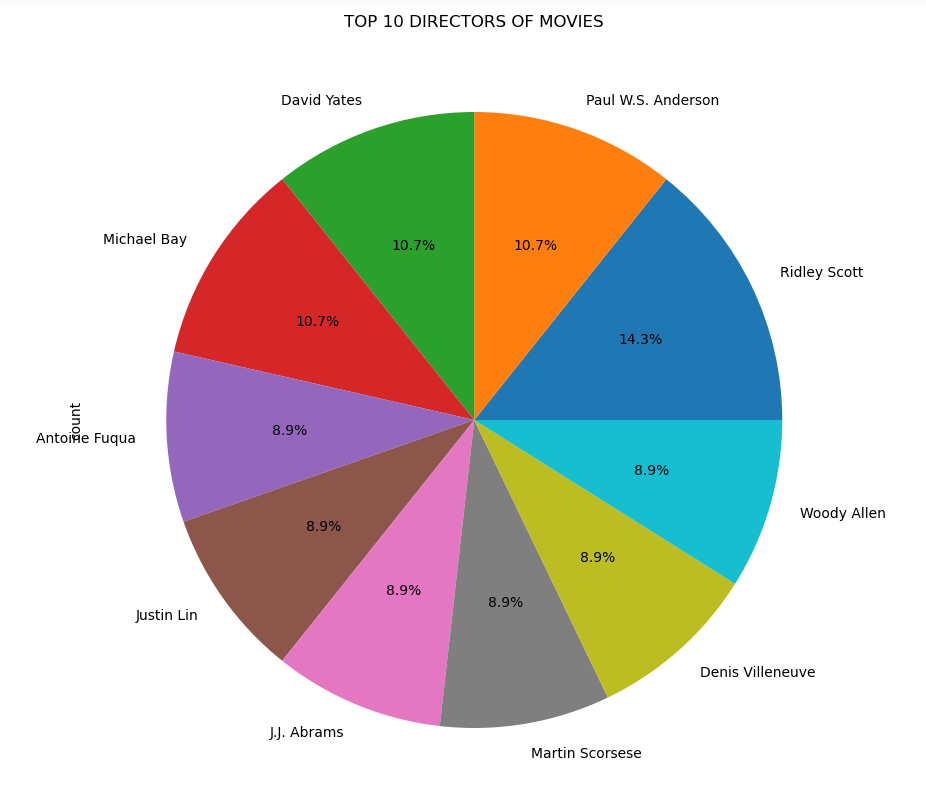
dsdarYour text here

**ANALYSIS:**

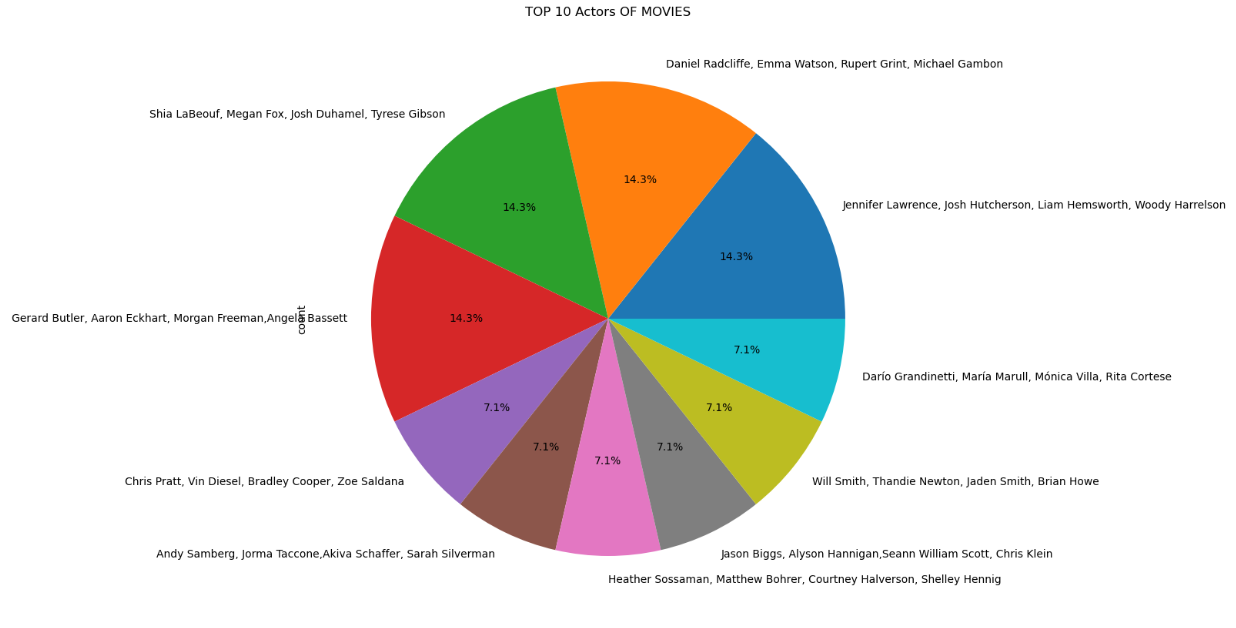
We have also performed an analysis on some attributes such as calculating top 10 directors of movies[fig.1] and top 10 actors of movies [fig.2].we performed year analysis like ratings based on year [fig.3] ,votes based on year [fig.4],revenue based on year [fig.5], Meta score based on year [fig.6] and performed rating analysis such as MOVIES WITH MEDIUM RATING , METASCORE[fig.7], MOVIES WITH MEDIUM RATING , VOTES[fig.8], MOVIES WITH MEDIUM RATING , REVENUE[fig.9], MOVIES WITH HIGH RATING , METASCORE[fig.10], MOVIES WITH HIGH RATING ,VOTES[fig.11], MOVIES WITH HIGH RATING ,REVENUE[fig.12], meta score[fig.13].We have also

generated a heat map which determines the correlation

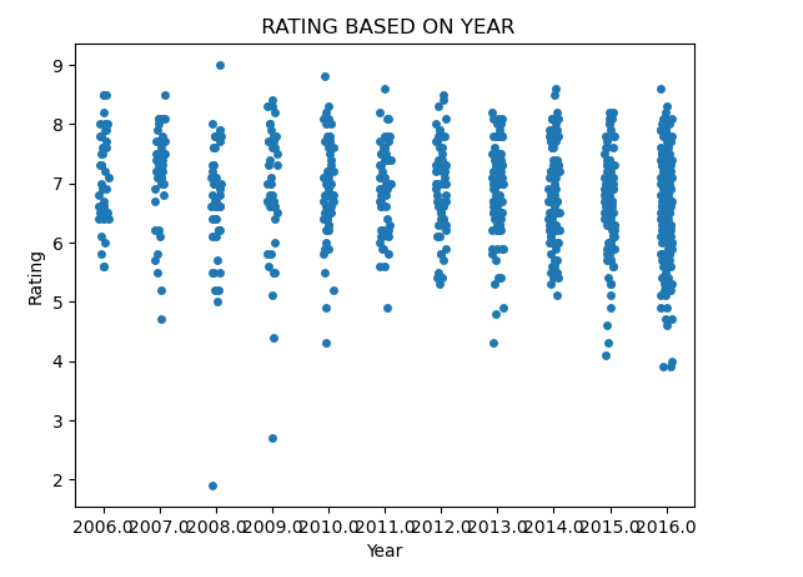
between various attributes used in the model[fig.14].



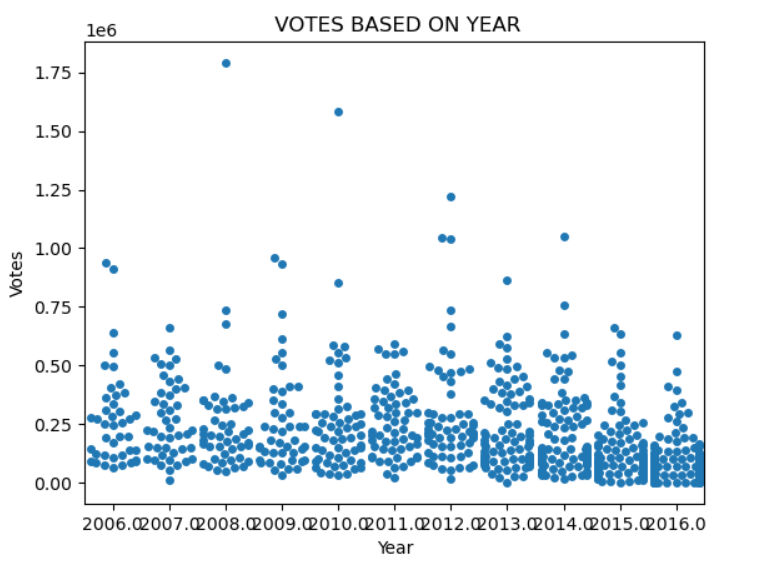
[Fig.1]



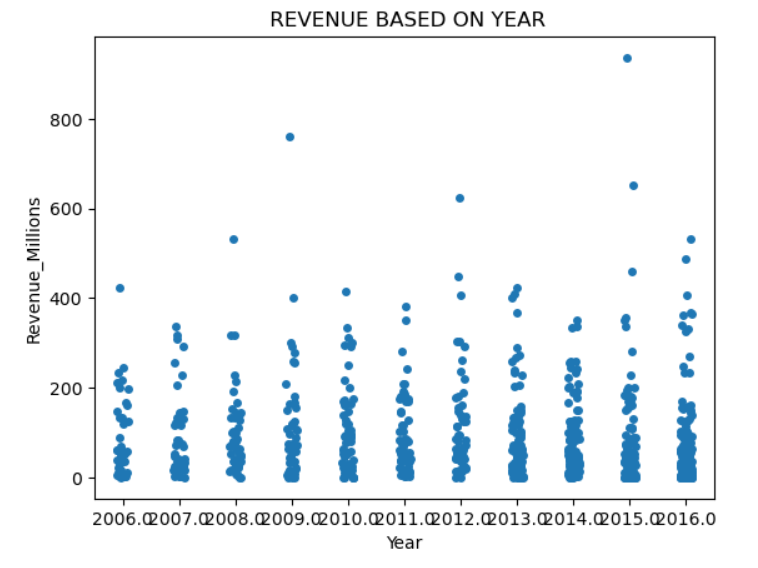
[Fig.2]



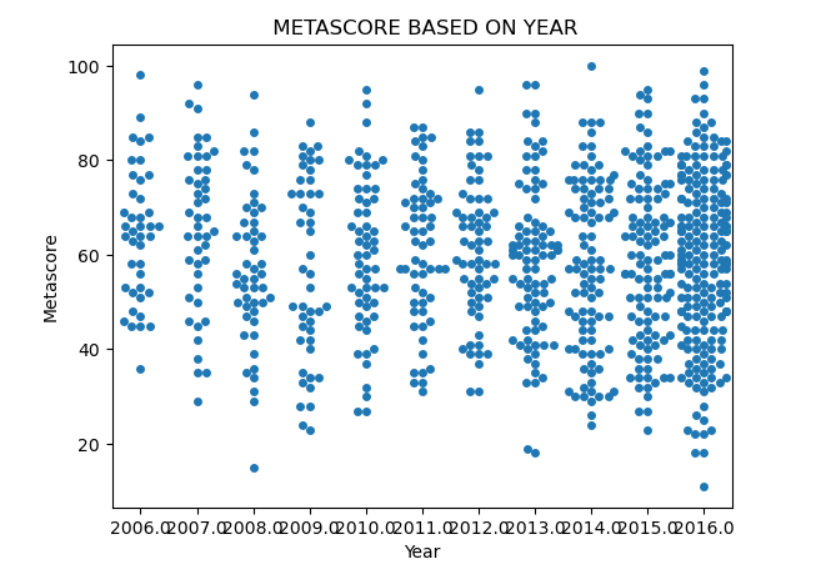
[Fig.3]



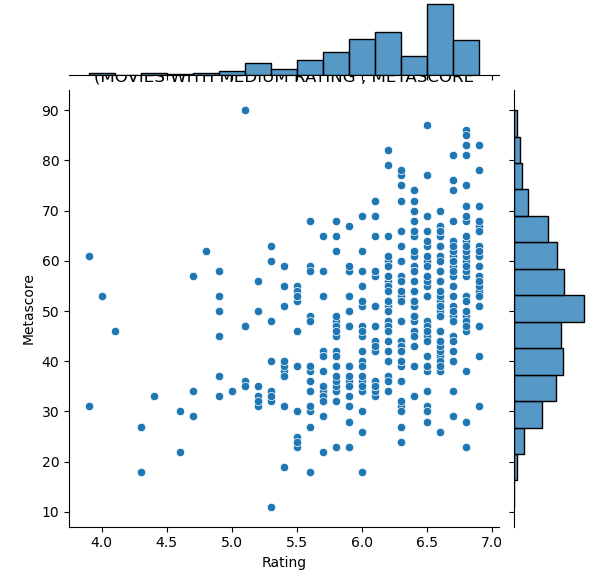
[Fig.4]



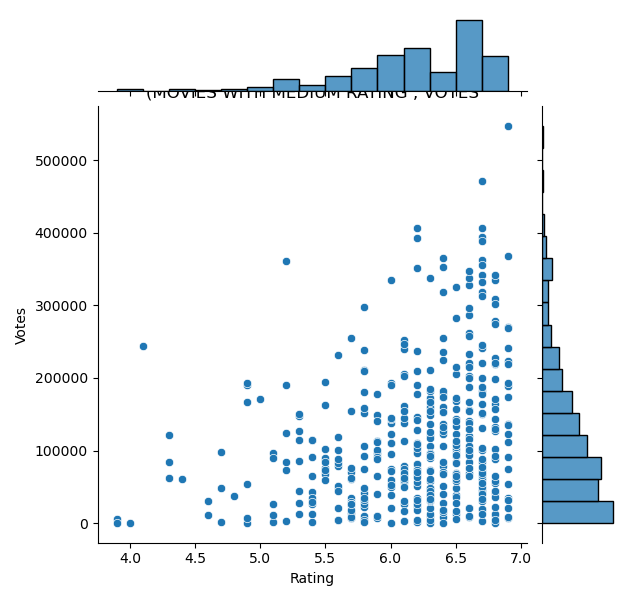
[Fig.5]



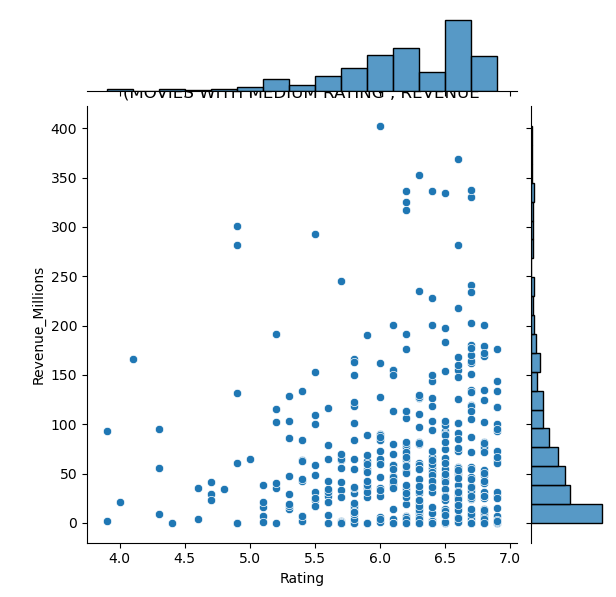
[Fig.6]



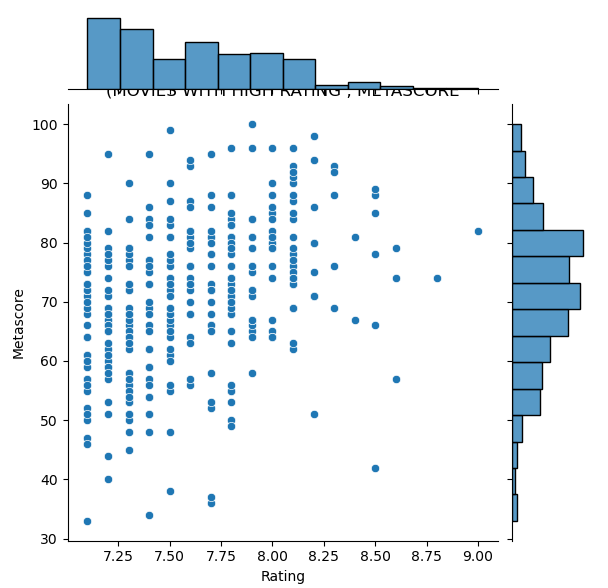
[Fig.7]



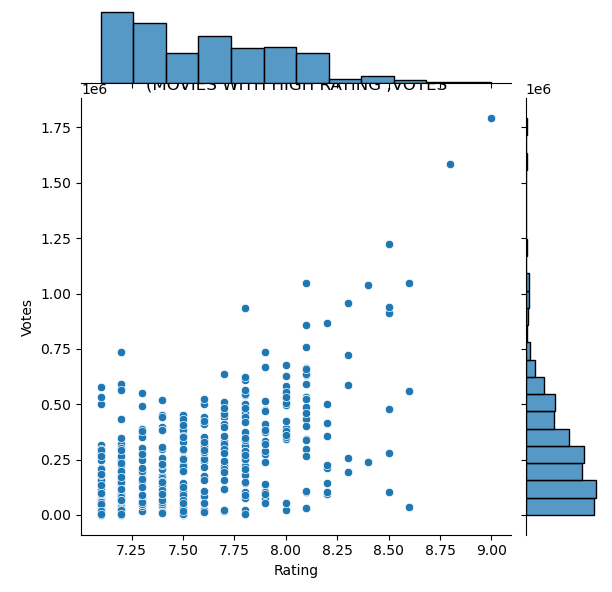
[Fig.8]



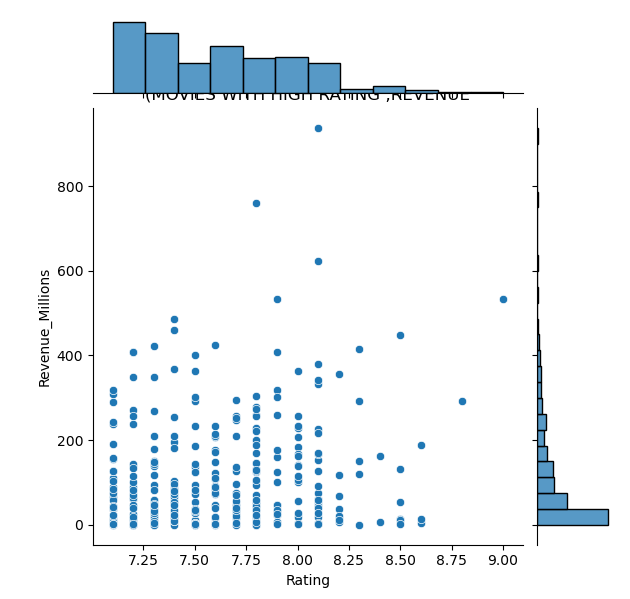
[Fig.9]



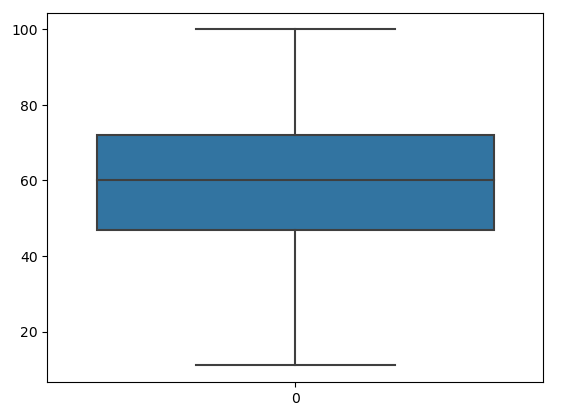
[Fig.10]



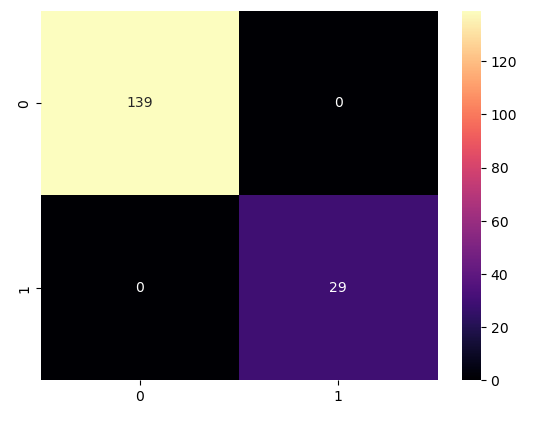
[Fig.11]



[Fig.12]



[Fig.13]



[Fig.14]

**Advantages**

 **Strategic Decision Making**: Predicting the success rate of movies can assist movie studios, producers, and distributors in making informed decisions about investment, marketing strategies, release dates, and distribution channels.

 **Resource Allocation**: By predicting the success rate of movies, stakeholders can allocate resources more effectively, such as budget allocation for production, marketing campaigns, and talent acquisition.

 **Risk Mitigation**: Predictive models can help in identifying potential risks associated with movie projects, allowing stakeholders to mitigate these risks through adjustments in production, marketing, or distribution strategies

**Disadvantages**

 **Complexity of Prediction**: Predicting the success rate of movies is challenging due to the multifaceted nature of the movie industry, including factors like changing consumer preferences, competition, and unpredictable market trends.

 **Data Availability and Quality**: Access to comprehensive and accurate data for building predictive models may be limited, especially for new or niche genres, making it difficult to create reliable predictions.

 **Subjectivity in Success Definition**: Defining and measuring the success of a movie can be subjective and may vary depending on factors like box office revenue, critical acclaim, audience ratings, or awards, making it challenging to create a standardized prediction model.