Big Data Analytics Project Report

IMDB Movies Data Analysis and Prediction

By

Priyanka Marathe (826133692)

Pooja Vaidya (826523588)

Introduction

The primary purpose of this project is to deep dive into the IMDB dataset and create data visualizations by analyzing various aspects of the dataset. The project also aims at demonstrating various python libraries that we learnt during the course and their implementations in real world dataset. As the amount of data in the entertainment industry is rising, the businesses in this industry are earning profits by using the results obtained from the analysis on previously generated data. As IMDB is one of the popular websites for rating movies or series, we thought it would be interesting to analyze its data.

This project consists of two parts:

1) Exploratory data analysis:

   The exploratory data analysis includes preprocessing the data and analyzing various columns in the dataset and finding the parameters that best determine the rating of the movie. This data analysis also computes some statistics and demonstrates some interesting insights about the features in the dataset and their correlation.

2) Predicting the rating of the movie :

   The prediction of rating includes determining appropriate features from the dataset and applying two machine learning algorithms to compare and find better accuracy scores. Considering the data shape and type, we selected Random forest and Decision Tree classifiers for making predictions by training and testing the dataset.

About the Data

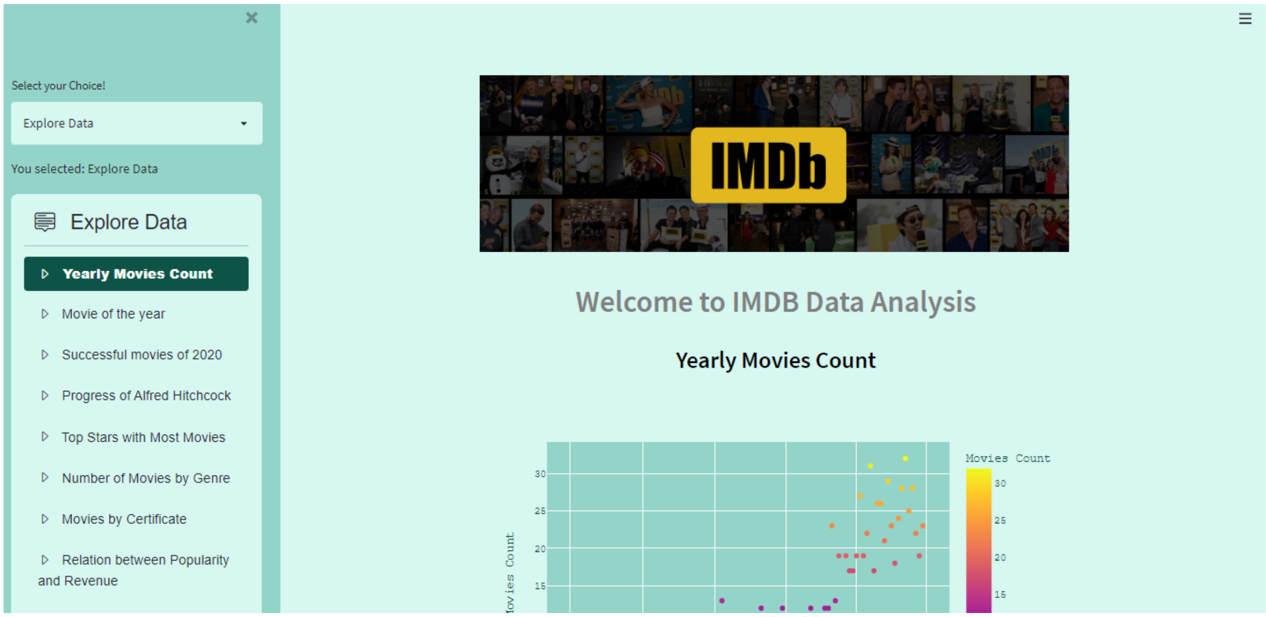
- The data used is the IMDB dataset from Kaggle.
- This dataset contains top 1000 movies and TV shows.
- The shape of the dataset:

  Rows: 1000

  Columns: 16
- Column Labels:
  - Poster_Link : Link of the poster that imdb using
  - Series_Title: Name of the movie
  - Released_Year: Year at which that movie released
  - Certificate: Certificate earned by that movie
  - Runtime : Total runtime of the movie
  - Genre: Genre of the movie
  - IMDB_Rating: Rating of the movie at IMDB site
  - Overview: mini story/ summary
  - Meta_score: Score earned by the movie
  - Director: Name of the Director
  - Star1: Name of the Star1
  - Star2: Name of the Star2
  - Star3: Name of the Star3
  - Star4: Name of the Star4
  - No_of_Votes: Total number of votes
  - Gross: Money earned by that movie

IMDB Exploratory Data Analysis Results

- **Glimpse of our project dashboard on streamlit (To highlight the navigation options)**



Select your Choice!

Explore Data

You selected: Explore Data

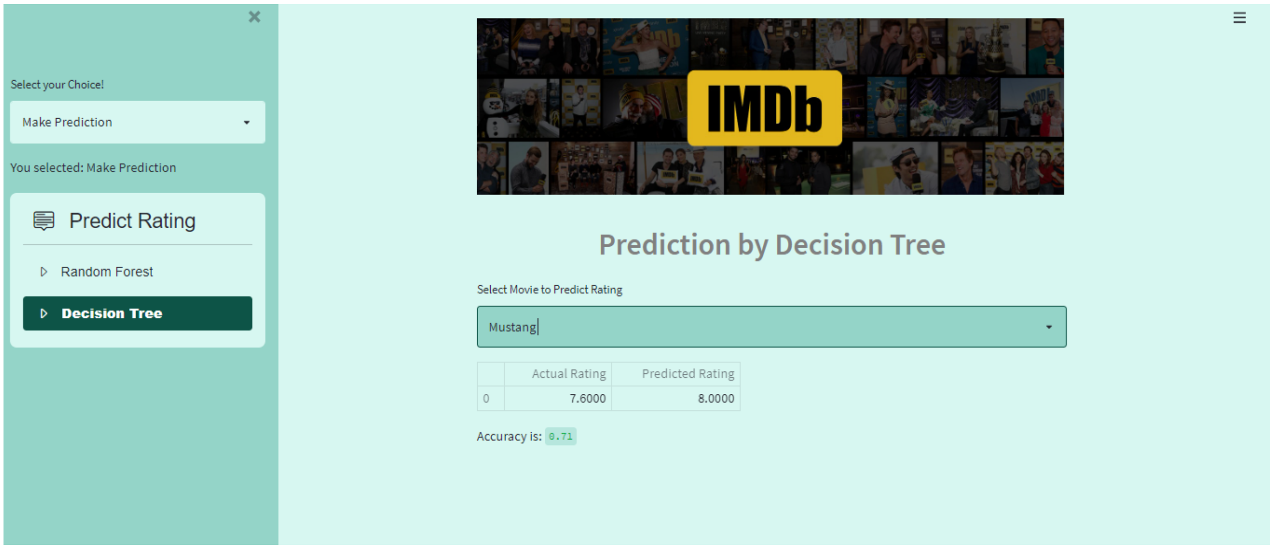
Explore Data

- Yearly Movies Count
- Movie of the year
- Successful movies of 2020
- Progress of Alfred Hitchcock
- Top Stars with Most Movies
- Number of Movies by Genre
- Movies by Certificate
- Relation between Popularity and Revenue

Welcome to IMDB Data Analysis

Yearly Movies Count

The above screenshot shows some of the navigation options for Explore Data.



Select your Choice!

Make Prediction

You selected: Make Prediction

Predict Rating

- Random Forest
- Decision Tree

Prediction by Decision Tree

Select Movie to Predict Rating

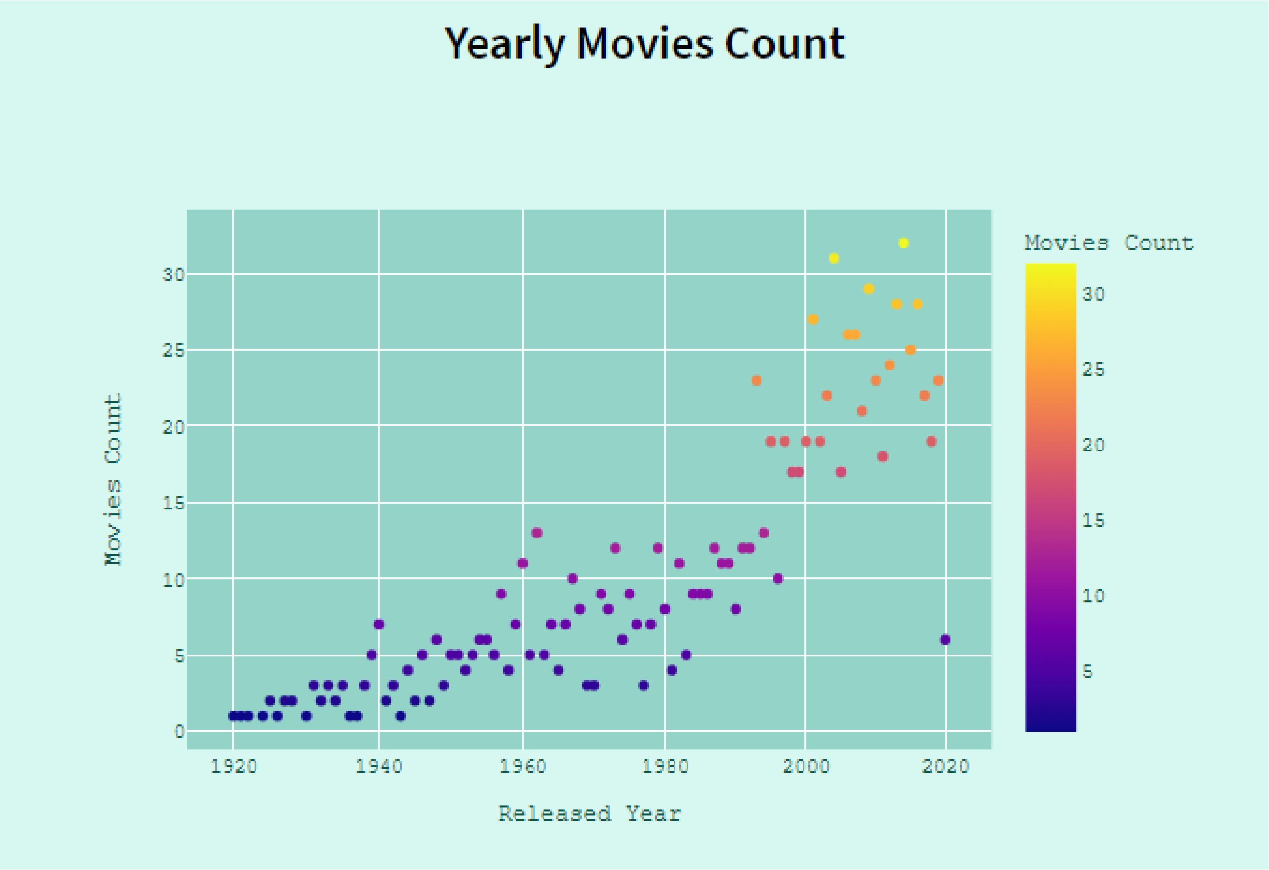
Mustang

	Actual Rating	Predicted Rating
0	7.6000	8.0000

Accuracy is: 0.71

The above screenshot shows the navigation between two classifiers for making predictions.

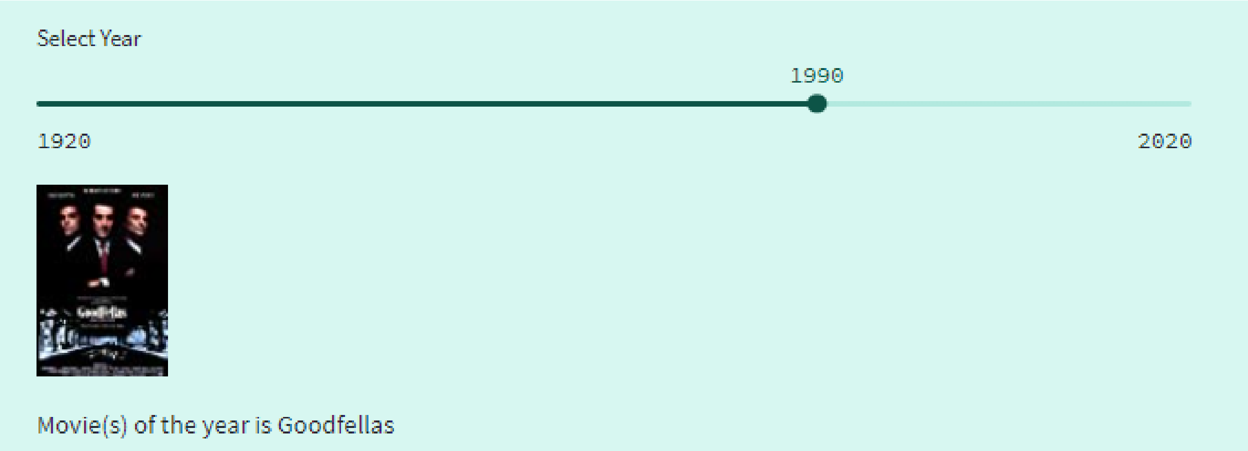
- **Yearly Movies Count**



Analysis:

1. Above figure shows the yearly couts of movies with the help of scatter plot.
2. Different color ranges are used to show the count from lowest to highest values.
3. From the above chart it can be observed that most movies are released from the year 2010 to 2020.
4. Whereas the number of movies are less in the years 1920 to 1940.

- **Movie(s) of the Year**

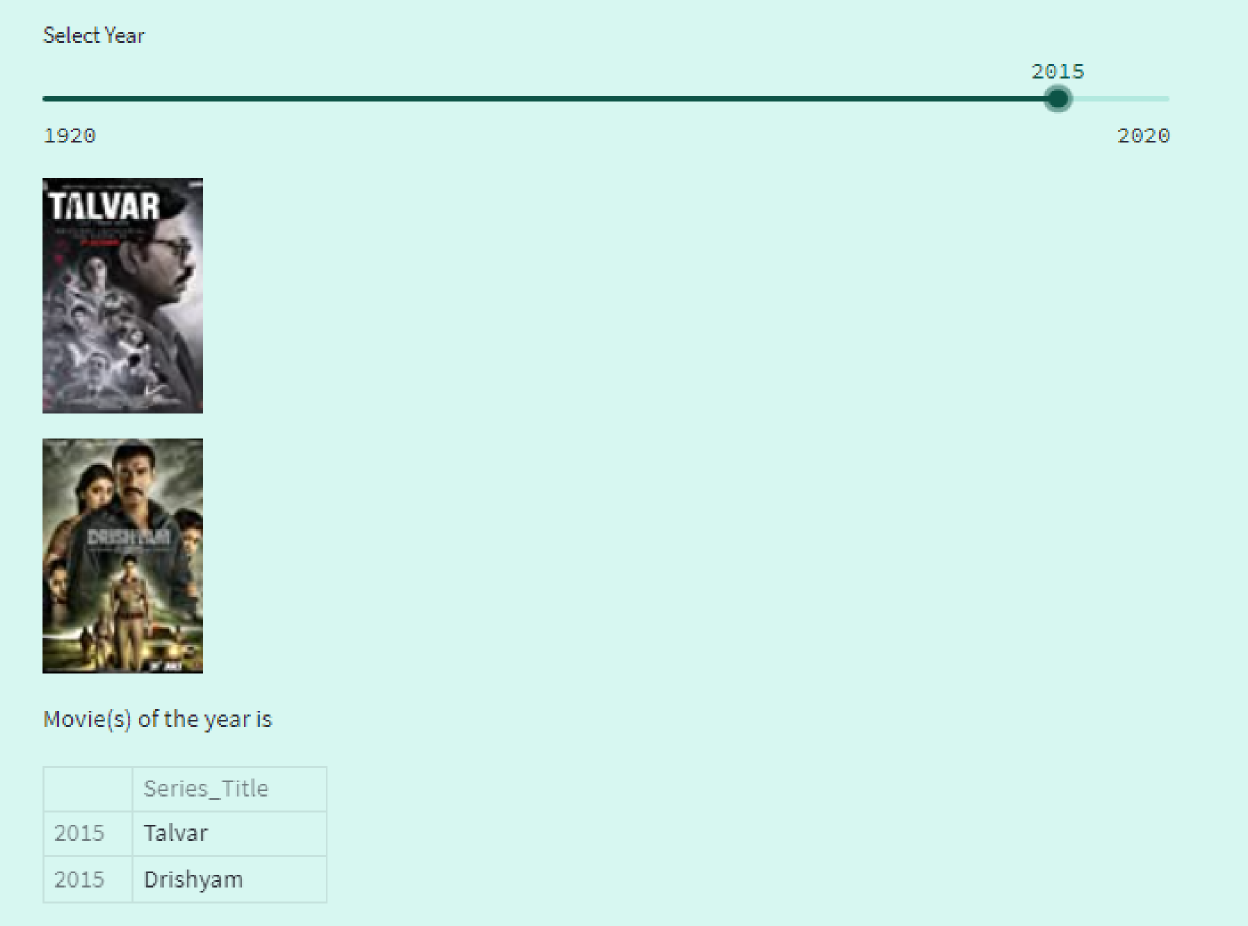


Select Year

1990

1920 2020

Movie(s) of the year is Goodfellas



Select Year

2015

1920 2020

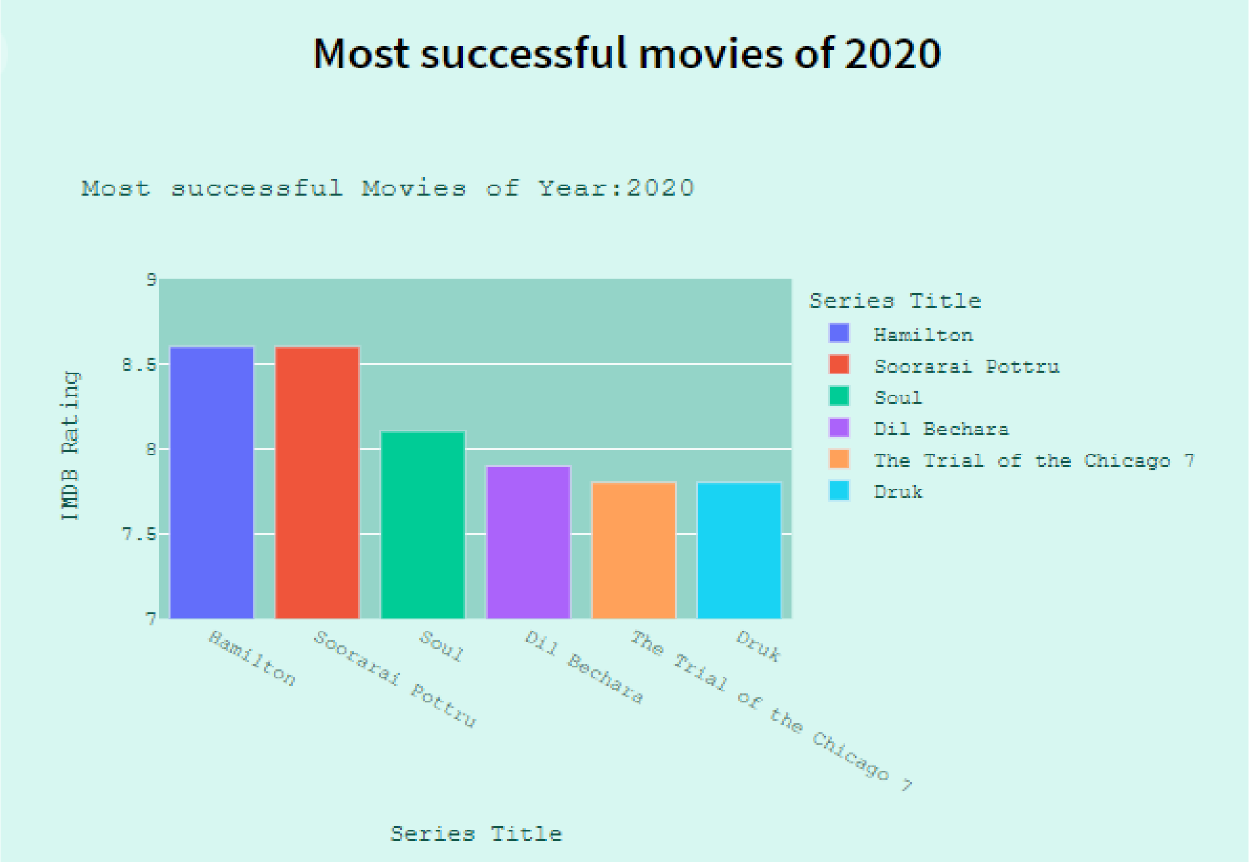
Movie(s) of the year is

	Series_Title
2015	Talvar
2015	Drishyam

Analysis:

1. The user can decide to watch the movie based on the highest rating.
2. This functionality helps to find the highest rated movies yearly.
3. Slider is used for the selection of the year and corresponding highest rated movies are shown.
4. Some of the years have more than one highest rated movie that can be seen in a form of table.

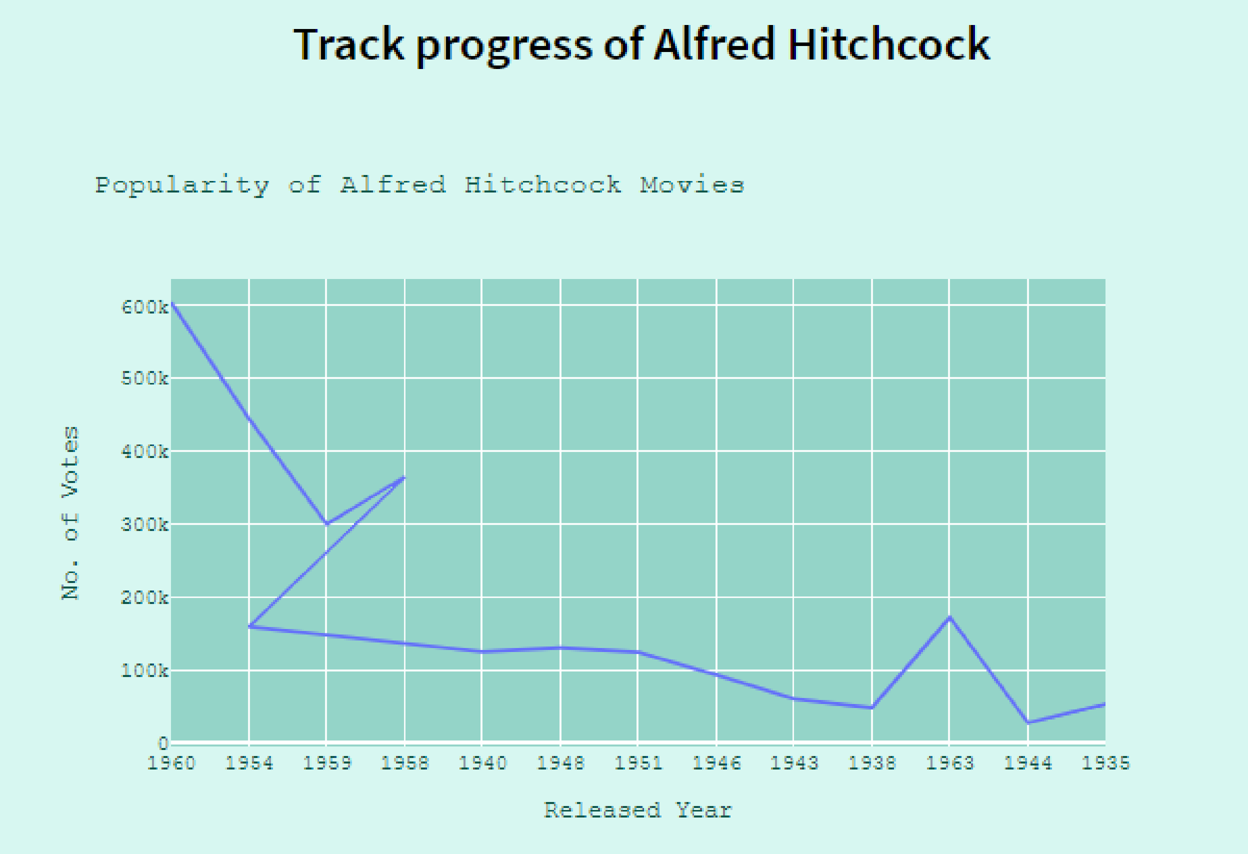
- **Successful Movies of 2020**



Analysis:

1. Top successful movies of the year 2020 are shown based on the IMDB_Rating.
2. The movies are shown with different colors so that they can be easily distinguishable.

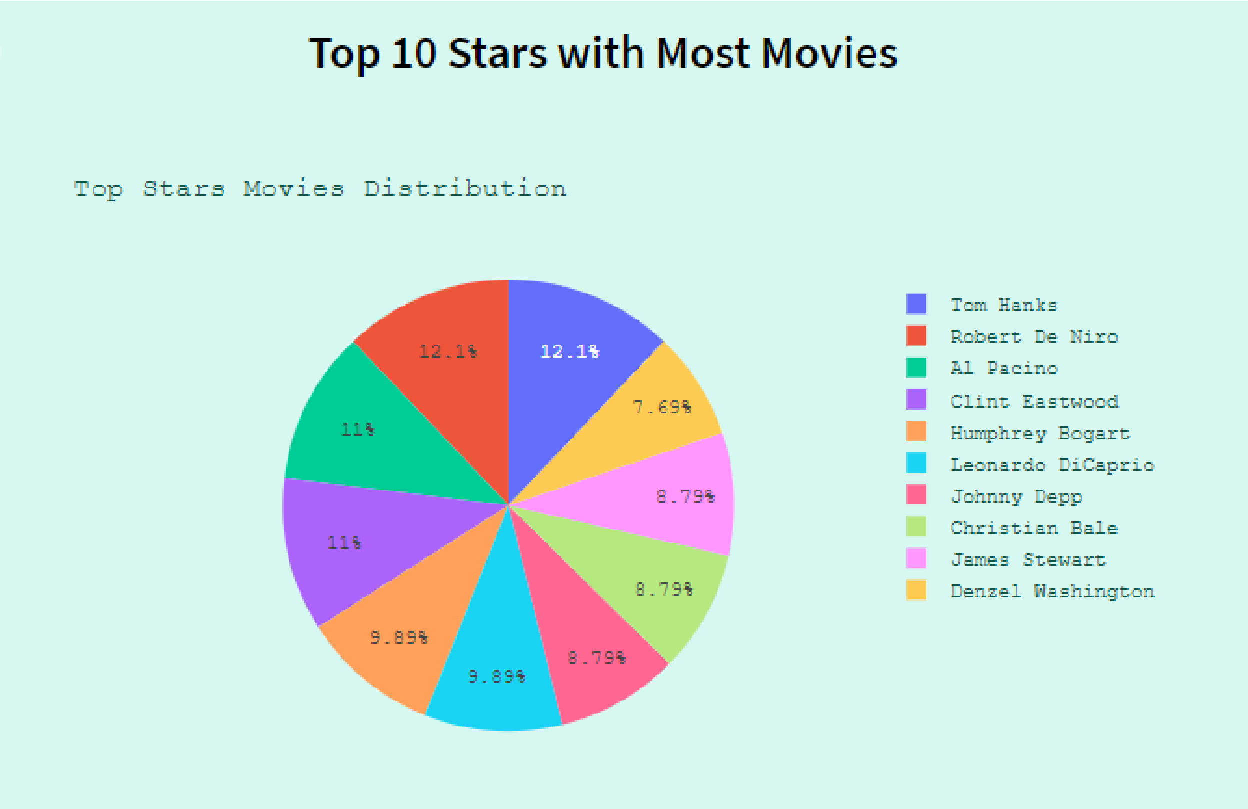
- **Track the Progress of the Director**



Analysis:

1. Director Alfred Hitchock has the most movies among all the directors hence we have chosen him for tracking the progress.
2. The line plot from plotly is used to show popularity gained by ‘Alfred Hitchock’ which is represented by number of votes.
3. From a plot it can be observed that in the years of 1960’s he gained the highest number of votes.
4. But In the years from 1935 to 1959 the number of votes were almost in the same range.
5. Scatter plot arranges X axis values better which helps to identify ranges rasily.

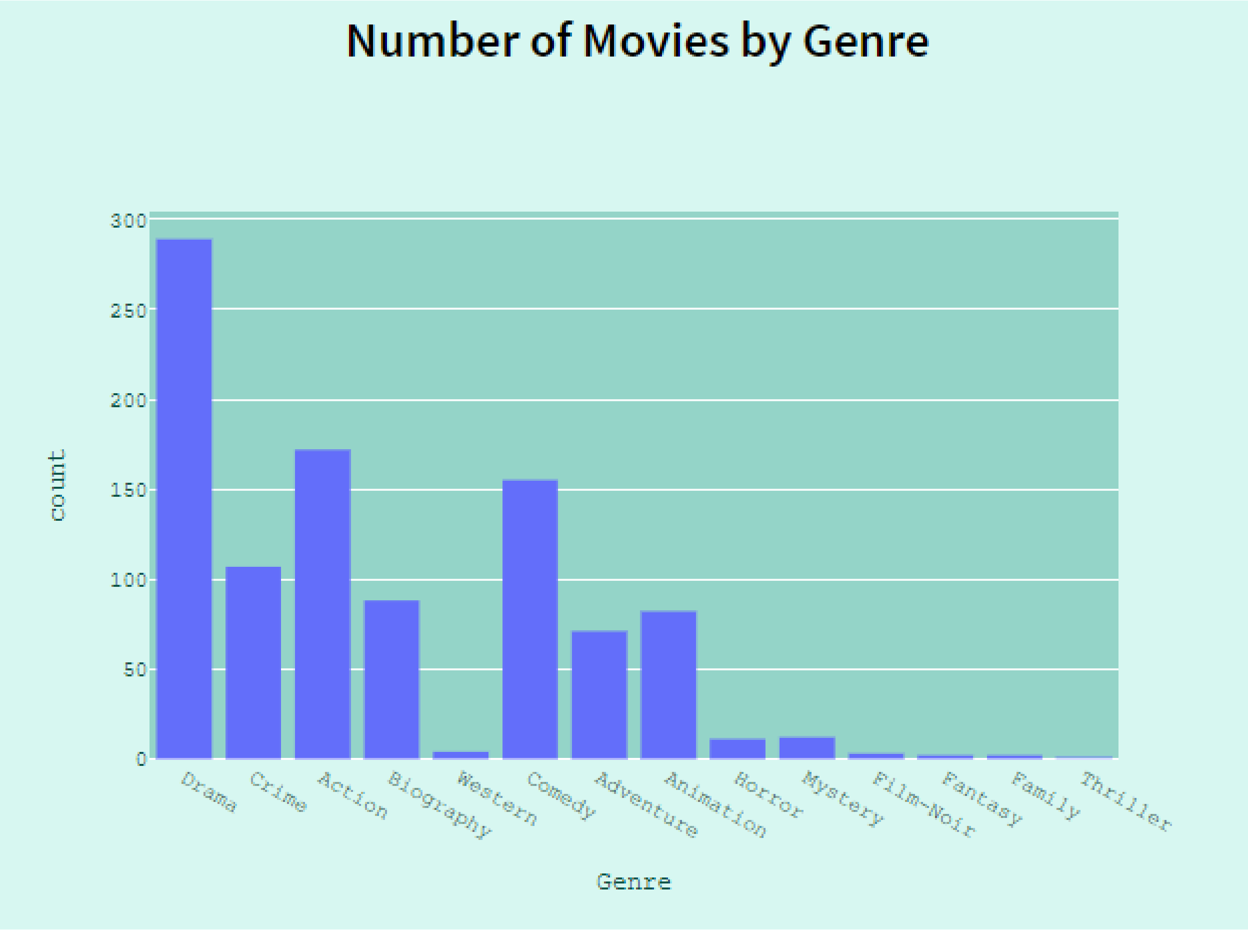
- **Top 10 Stars with most Movies**



Analysis:

1. Pie chart from plotly is used to show top 10 starts with the most number of movies
2. It can be observed that Tom Hanks and Robert De Nitro acted in 11 movies, which is represented by maximum percentage in a pie diagram.
3. In a dashboard you can see the actual values by hovering over the pie diagram.

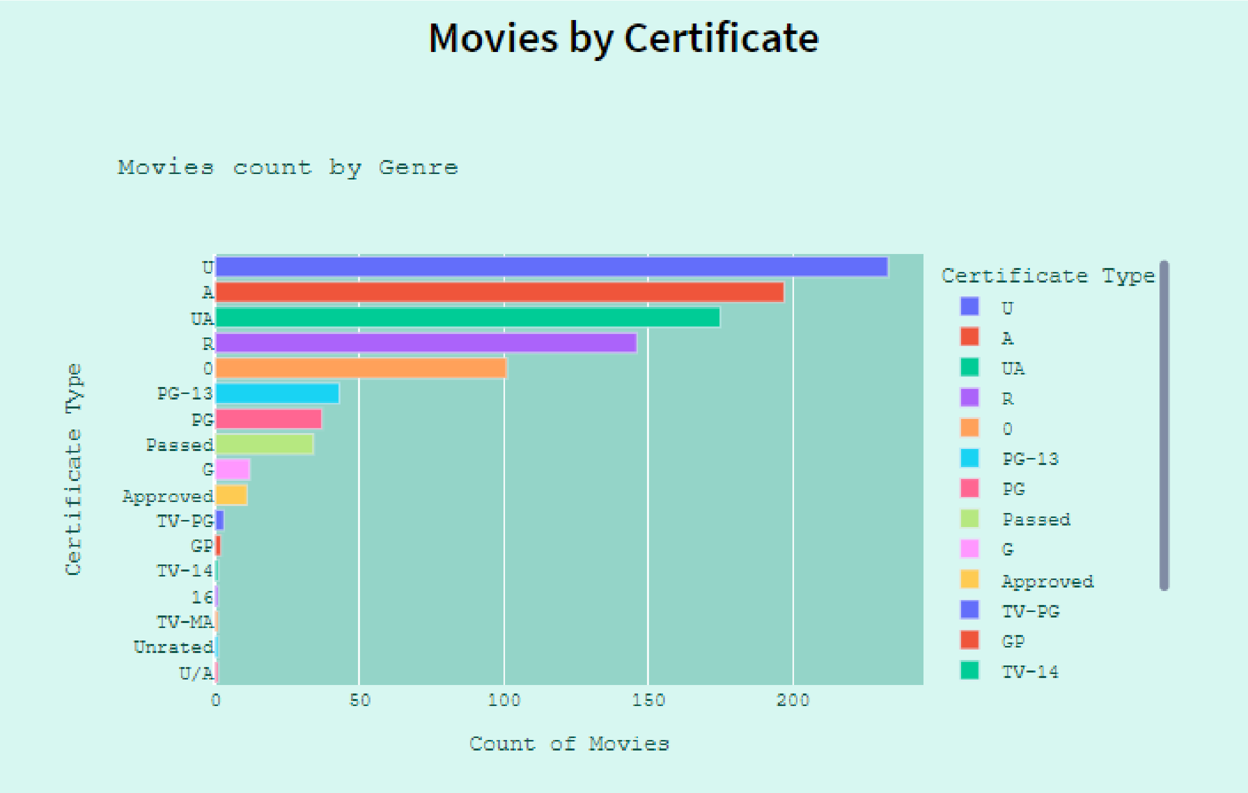
- **Number of Movies by Genre**



Analysis:

1. The above figure is an histogram which represents the number of movies by the genre.
2. As the genre column can have multiple genres associated with each movie, firstly we have separated the genres column.
3. From this histogram we can conclude that Drama, Action, and Comedy genres have the highest number of movies.
4. The least number of movies are from genres Thriller, Fantasy, and Family.

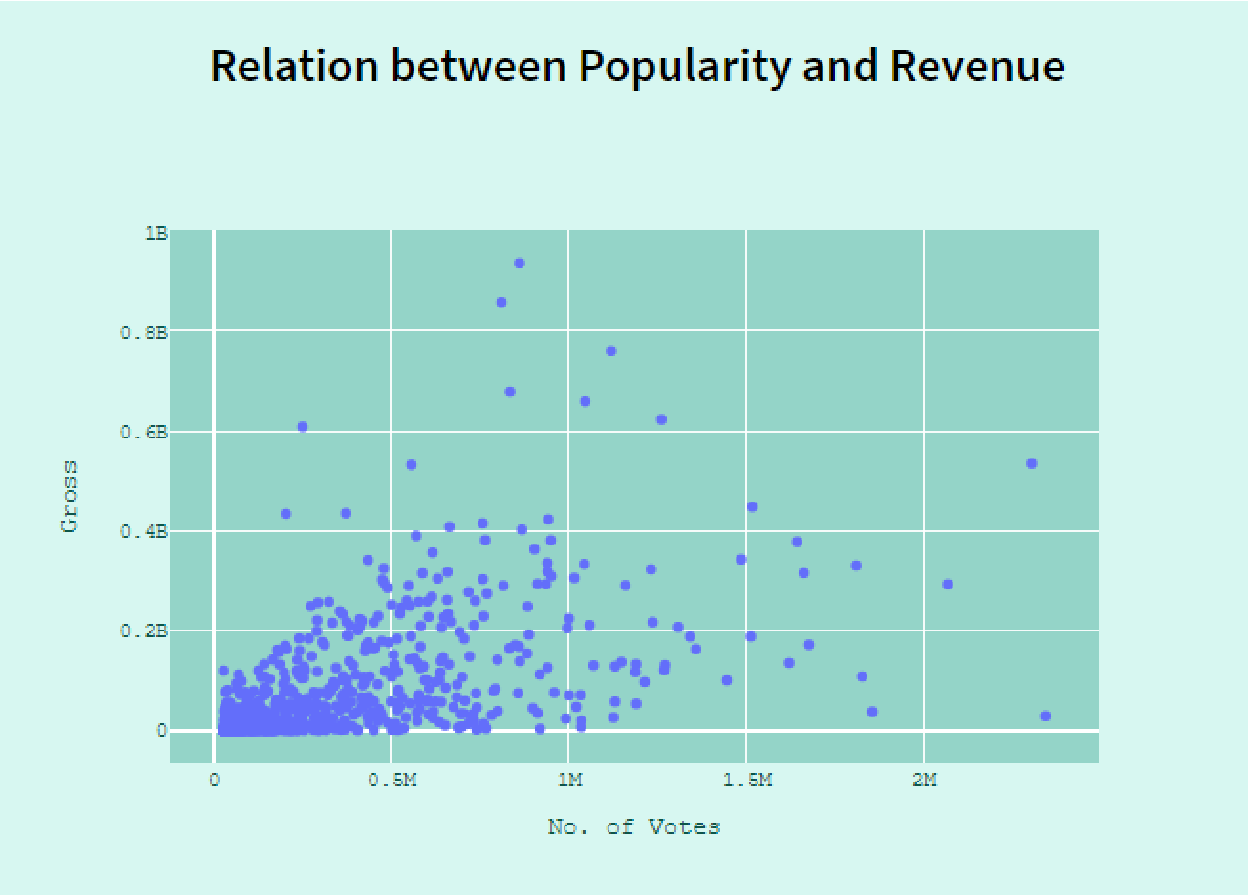
- **Number of Movies by Certificate**



Analysis:

1. The above diagram represents how many movies are there for each certificate in the form of the plotly bar chart.
2. From the diagram we can say that the ‘U’ and ‘A’ certificate has the highest number of movies and the least movies are of the ‘TV-MA ‘ category.
3. Different color values are used for different certificate types to make a chart more readable.

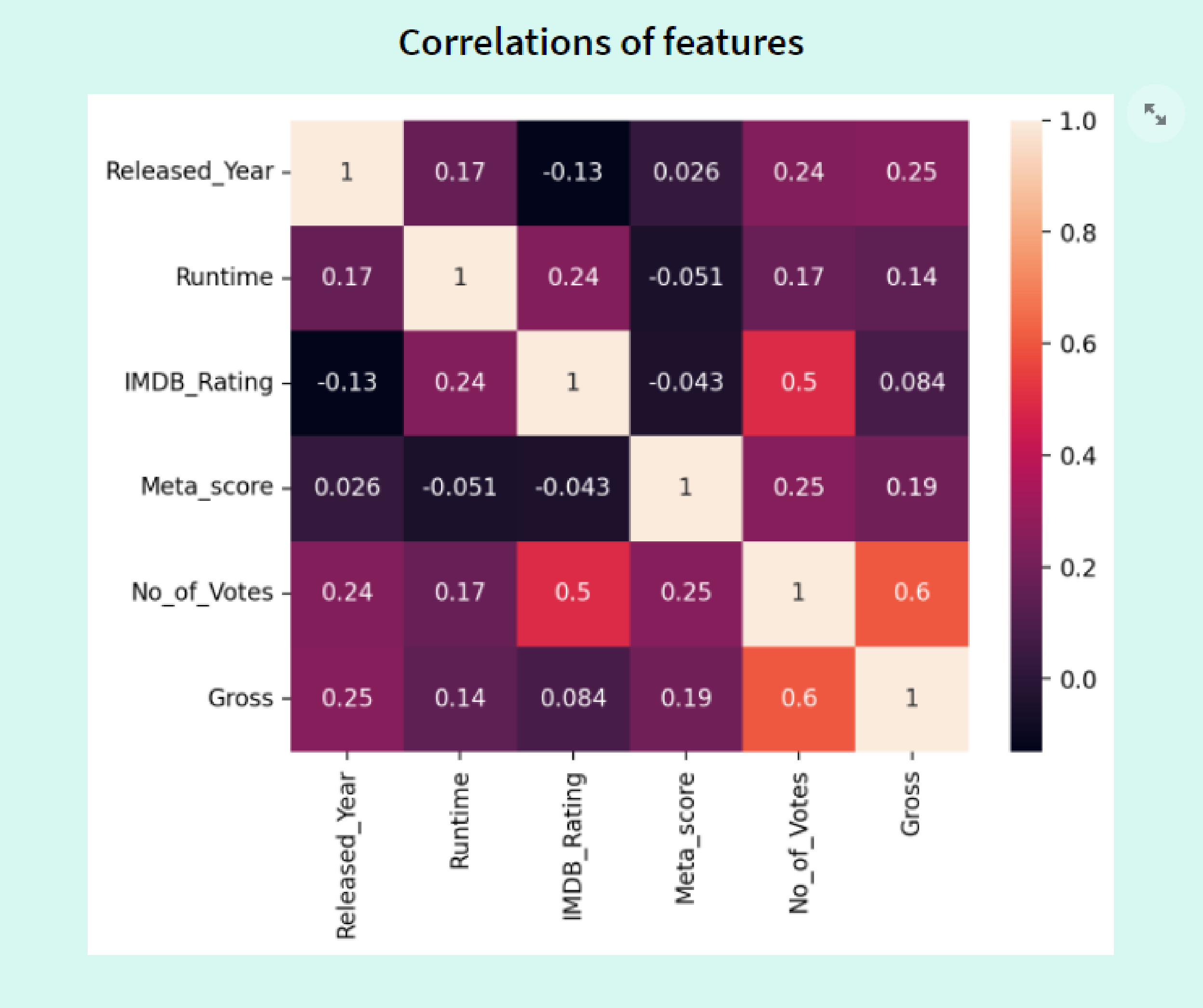
- **Relation between Popularity and Revenue**



Analysis:

1. Above scatter plot shows the relationship between the revenue generated by movies to its respective number of votes which show the popularity among the public.
2. From the above plot we can infer that there are many movies with the number of votes between 0 to half million and its corresponding revenue lies between 0 to 200 million.
3. As a summary we can say that the gross and popularity seems to be directly proportional with very few exceptions where less popular movies have gained huge profits and very popular movies did not perform well budget wise.

- **Correlation of Features**



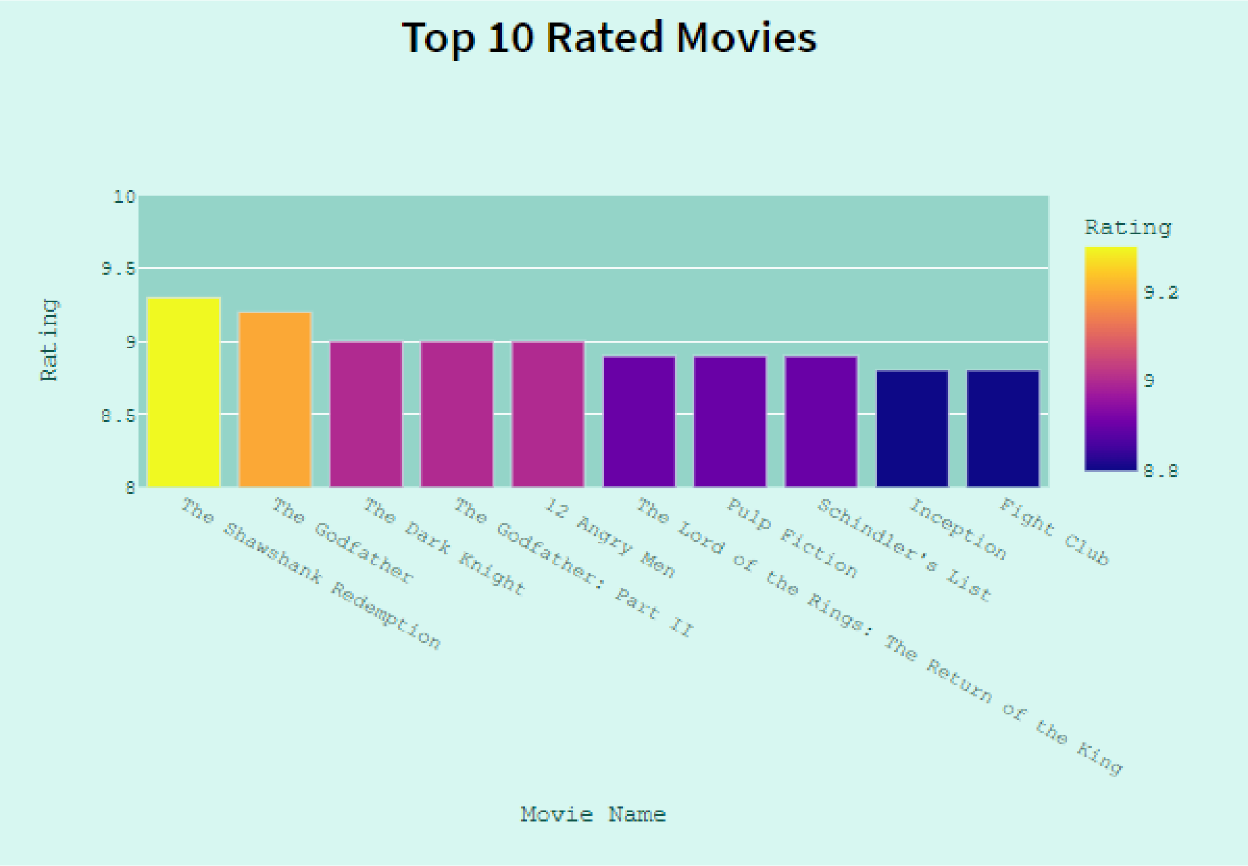
Analysis:

1. Above heat map shows correlation between the features of the dataset which have only numerical values.
2. The leading diagonal shows the correlation of the feature with itself, hence its value is 1.
3. The method used to find the correlation is Person method and generally Pearson correlation coefficient value nearing towards 1 indicates the presence of multicollinearity. Thus, we can say that features such as Gross and No. of votes have maximum positive correlation followed by No.of

votes and IMDB Rating. A strong positive correlation between two features indicates that both features grow in the same direction.

4. On the other hand, the features Released year and IMDB rating are negatively correlated as they have a negative most Pearson correlation coefficient. Hence their growth is in the opposite direction.
5. This correlation matrix is helpful in determining the most relevant features that will contribute to the rating prediction.

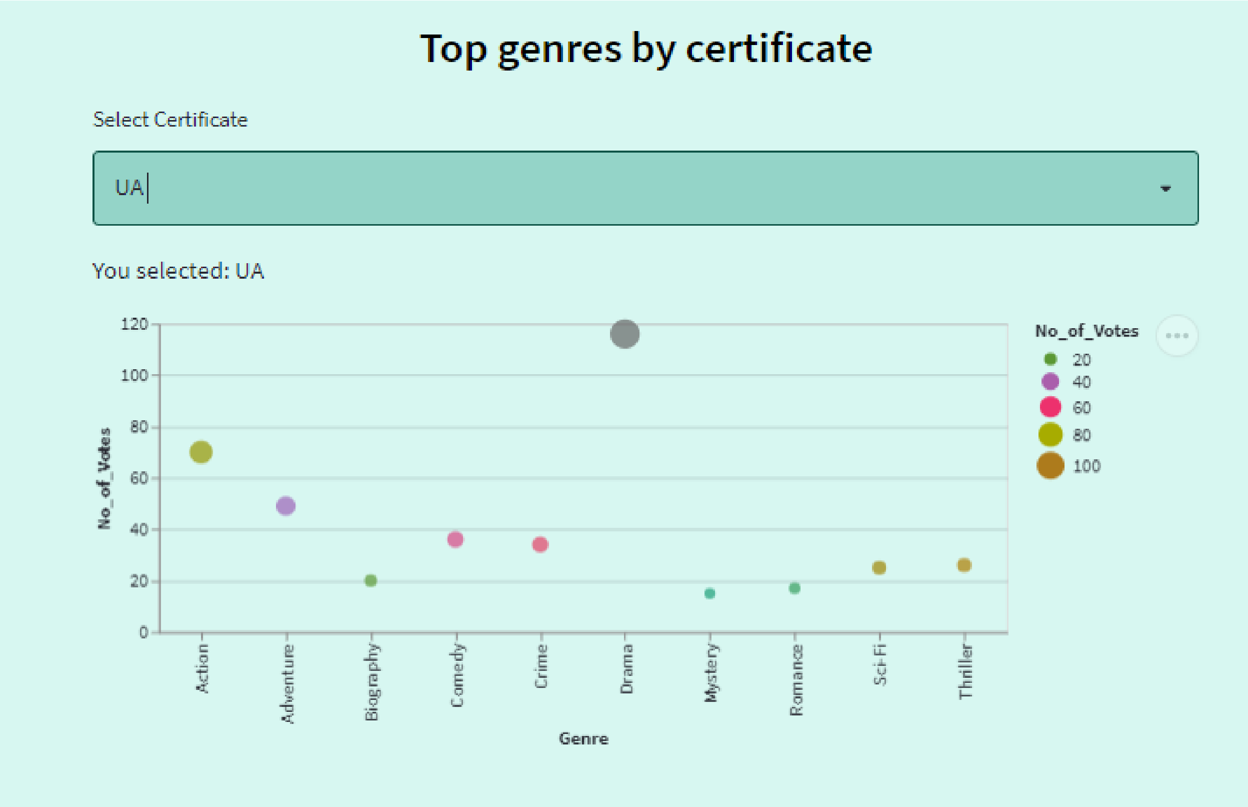
- **Top 10 Rated Movies**



Analysis:

1. The bar plot shows the top 10 movies based on the rating.
2. We can see that the range of rating for top 10 movies lies between 8.5 to 9.5.
3. The movie with the topmost rating is “The Shawshank Redemption”.

- **Top genres by Certificate**

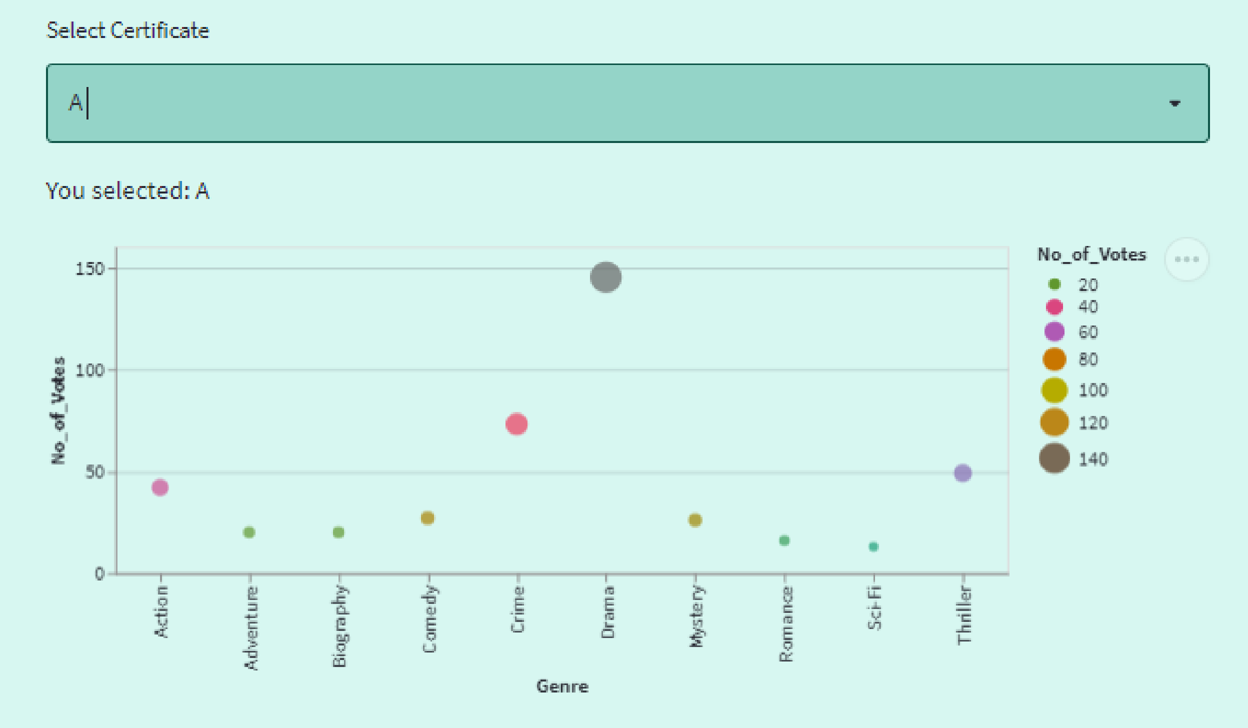


Top genres by certificate

Select Certificate

UA

You selected: UA



Select Certificate

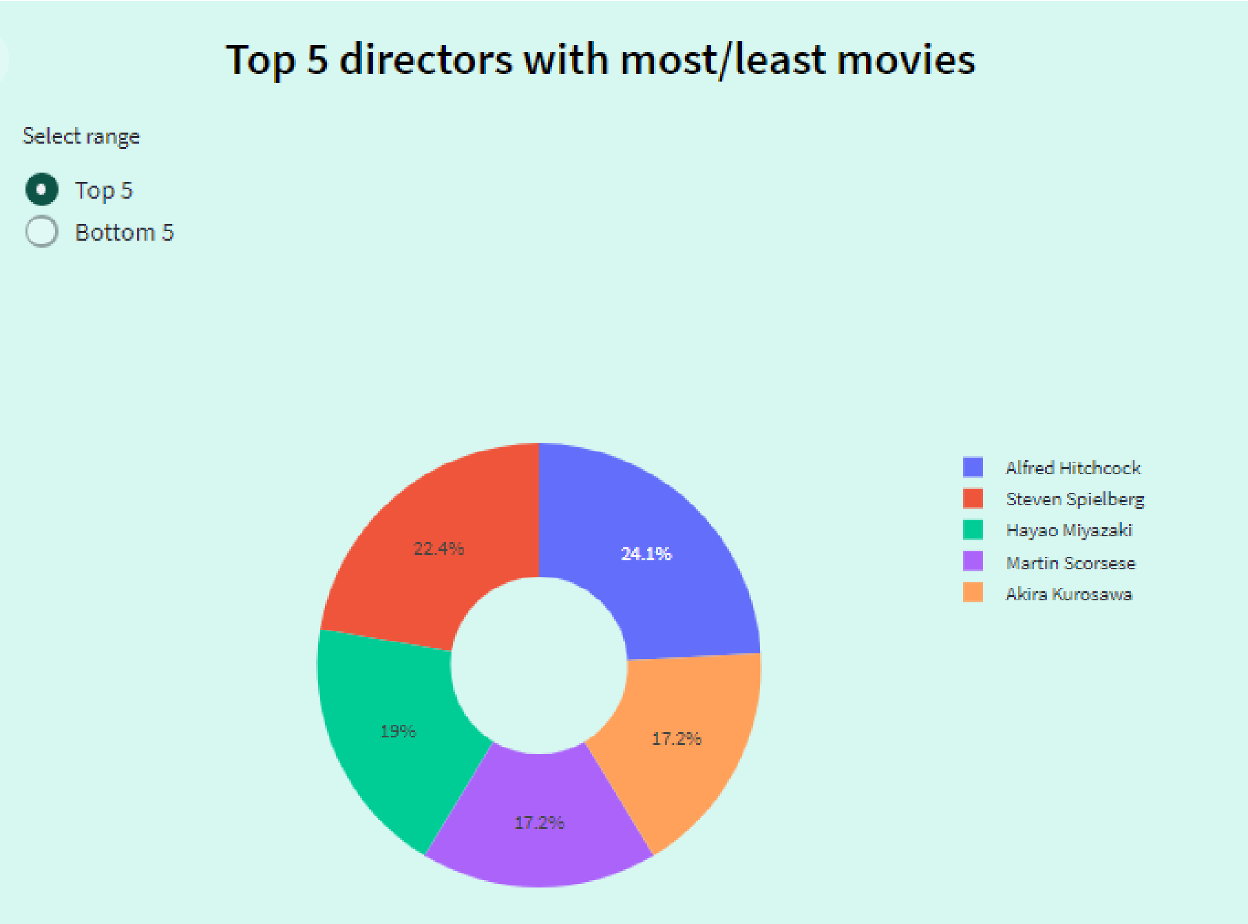
A

You selected: A

Analysis:

1. The above graph shows top genres based on certificate number of votes and number of votes for those genres.
2. The user can select the certificate for which he wants to view the top genres from the drop down menu.
3. This will help to choose a movie belonging to a particular genre for a particular target audience.

- **Top 5 directions with most/least movies**

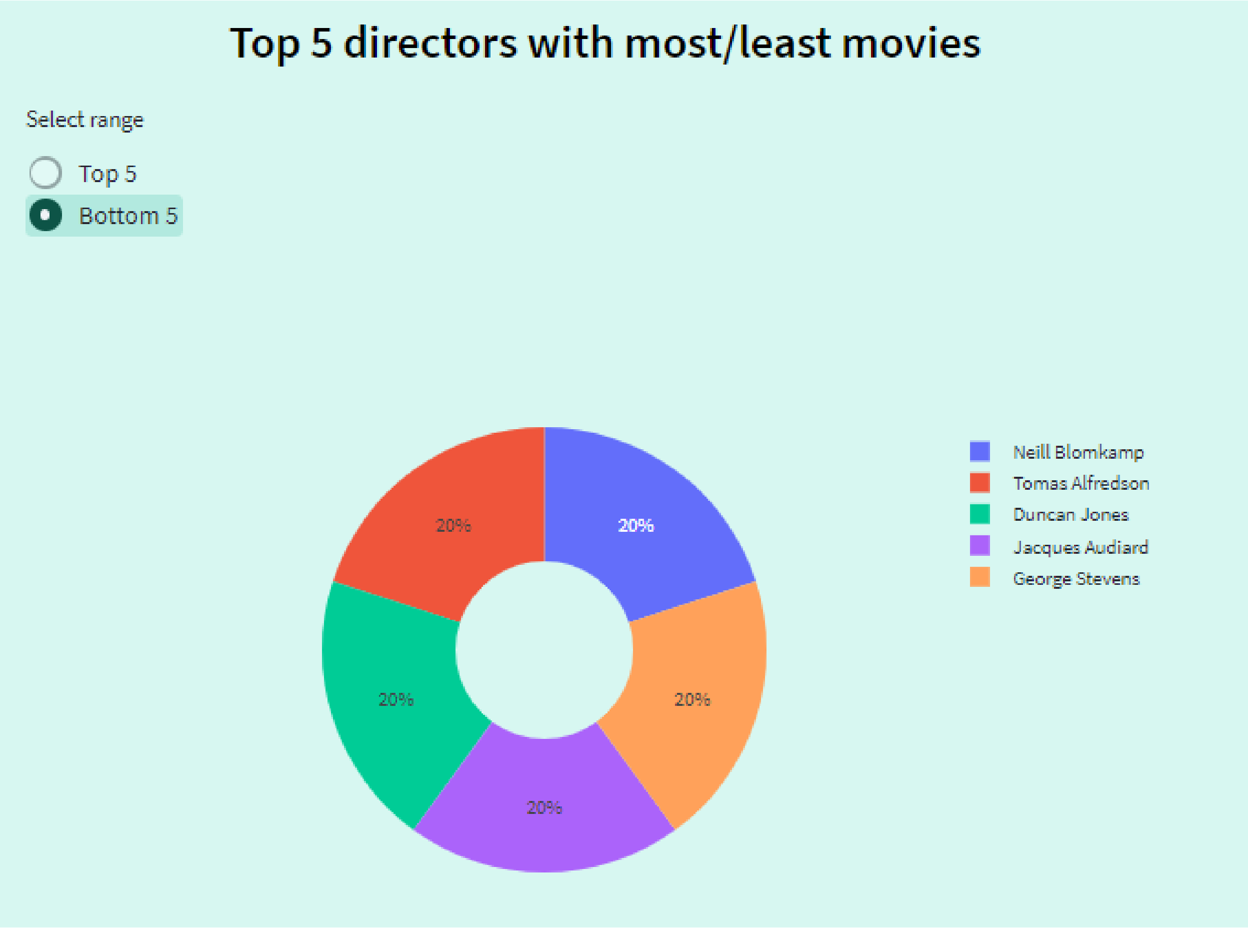


Top 5 directors with most/least movies

Select range

● Top 5

○ Bottom 5



Top 5 directors with most/least movies

Select range

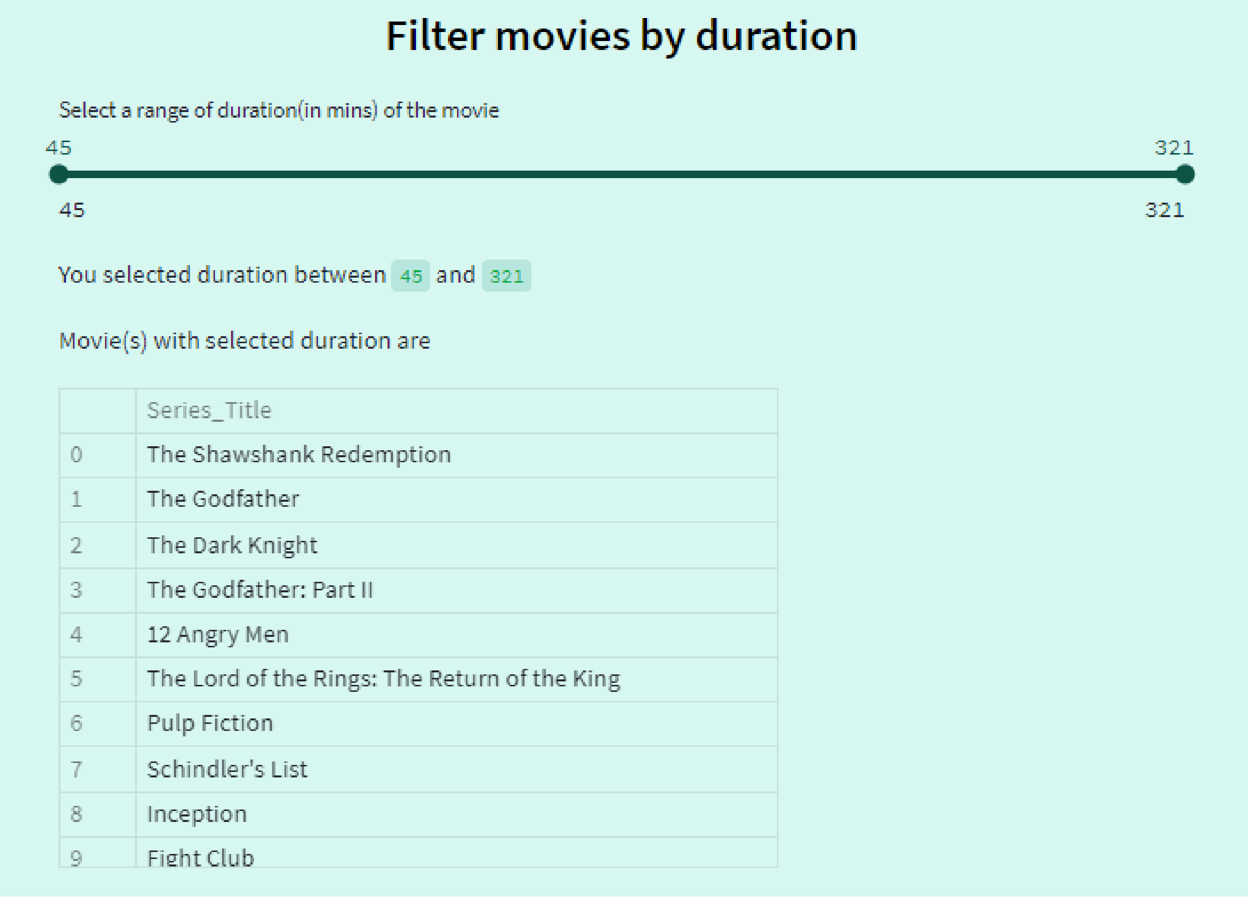
○ Top 5

● Bottom 5

Analysis:

1. This functionality will enable you to view the most active and least active Director based on the number of movies he produced.
2. The user can select whether he wants to view top 5 or the bottom 5 by clicking on the radio button.
3. The color in the donut chart represents the names of the director and the number on it represents the number of movies he has produced.

- **Filter movies by duration**



Filter movies by duration

Select a range of duration(in mins) of the movie

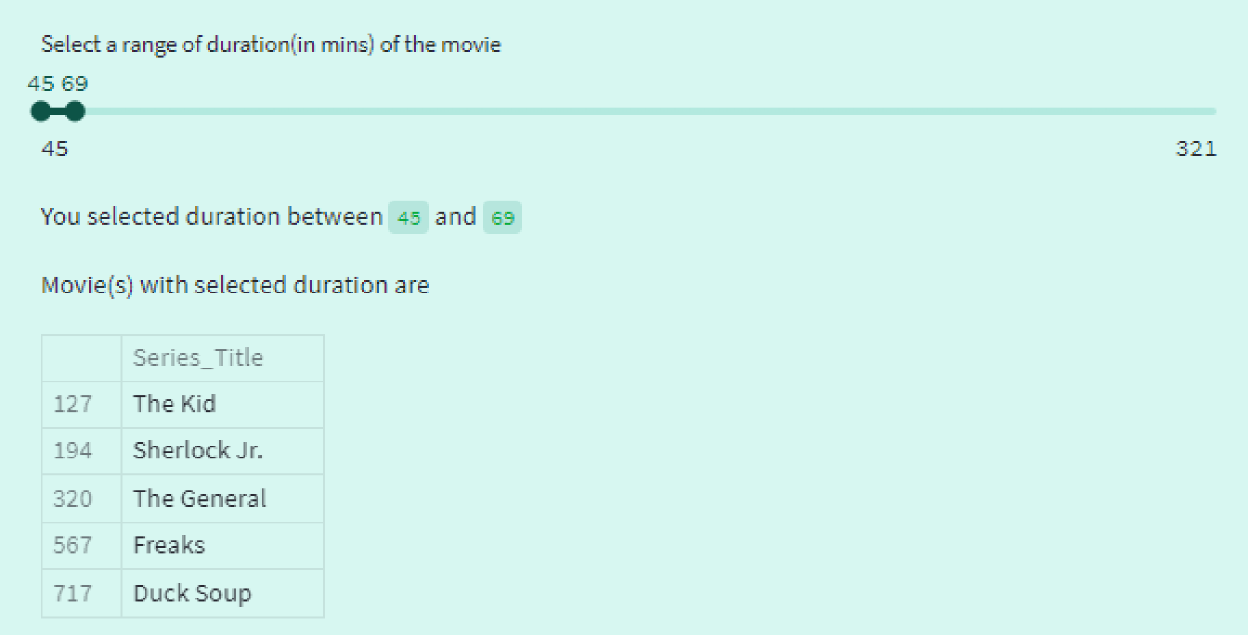
45 321

45 321

You selected duration between 45 and 321

Movie(s) with selected duration are

	Series_Title
0	The Shawshank Redemption
1	The Godfather
2	The Dark Knight
3	The Godfather: Part II
4	12 Angry Men
5	The Lord of the Rings: The Return of the King
6	Pulp Fiction
7	Schindler's List
8	Inception
9	Fight Club



Select a range of duration(in mins) of the movie

45 69

45 321

You selected duration between 45 and 69

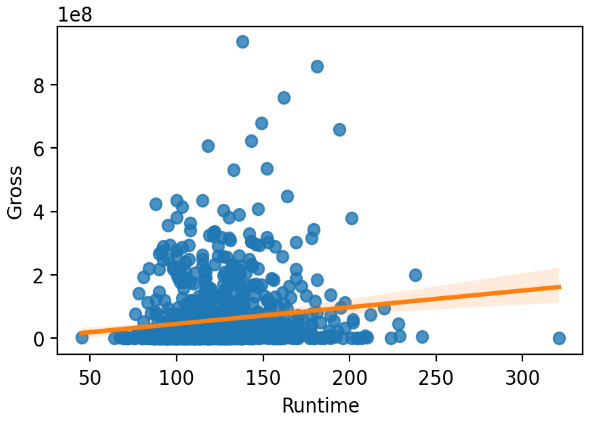
Movie(s) with selected duration are

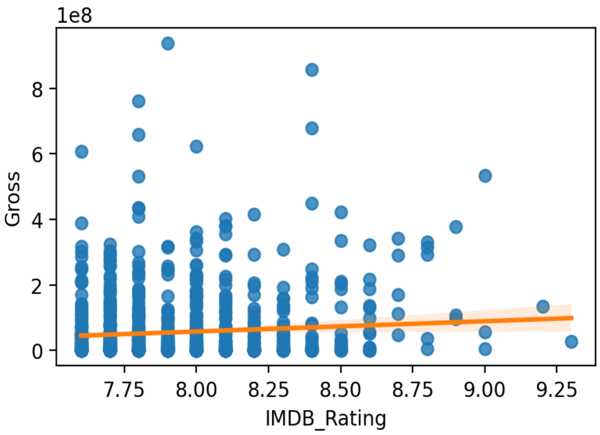
	Series_Title
127	The Kid
194	Sherlock Jr.
320	The General
567	Freaks
717	Duck Soup

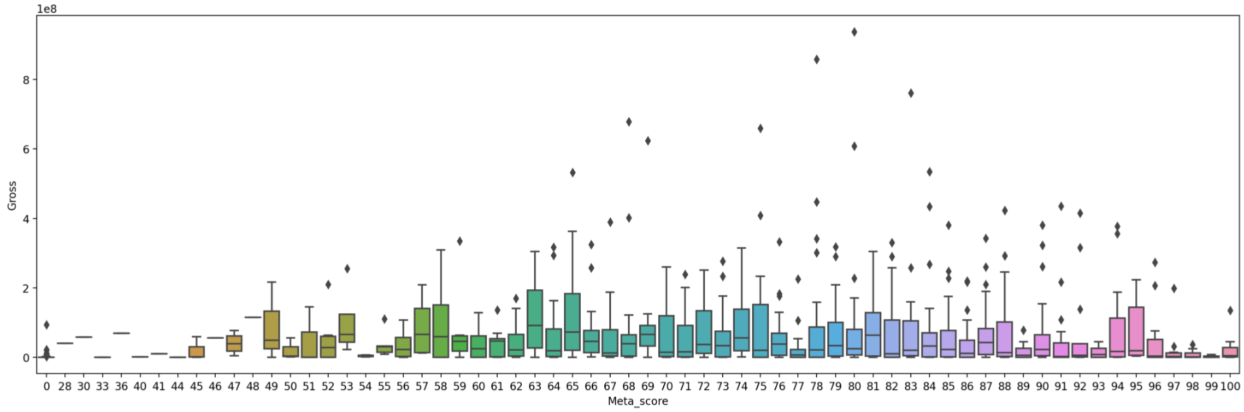
Analysis:

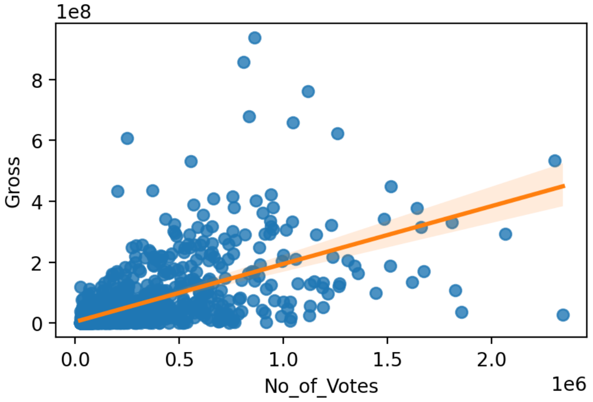
1. This functionality enables users to see the number of movies by duration (in minutes).
2. There is a slider to select the range of the duration.
3. As the user selects the duration, he can see the list of movies within that duration.
4. The first figure above shows a scrollable list of all movies between the duration of 45 minutes to 321 minutes.
5. The second figure shows a list of movies between 45 minutes to 69 minutes.

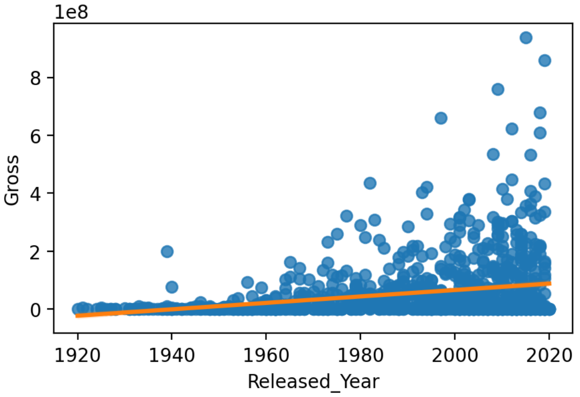
- **Tradeoff between Revenue and Features**







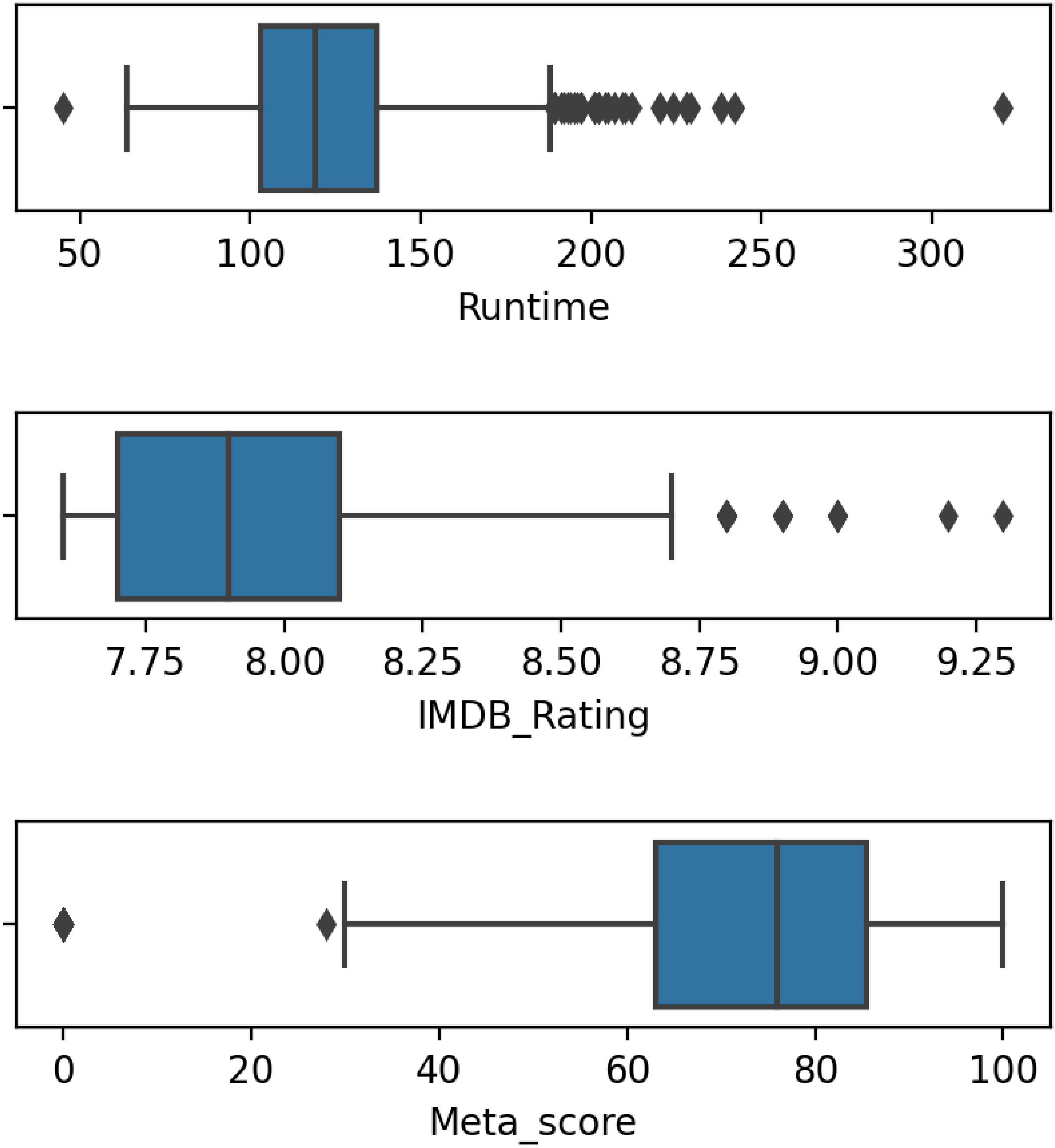


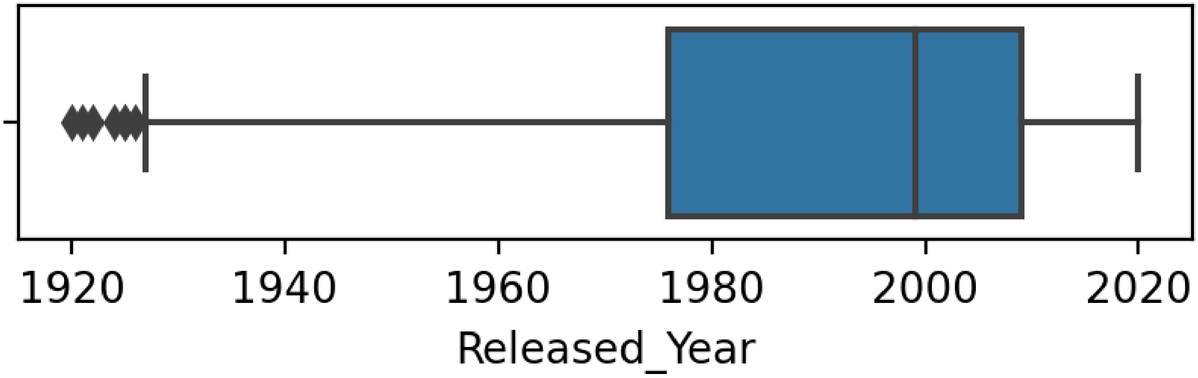


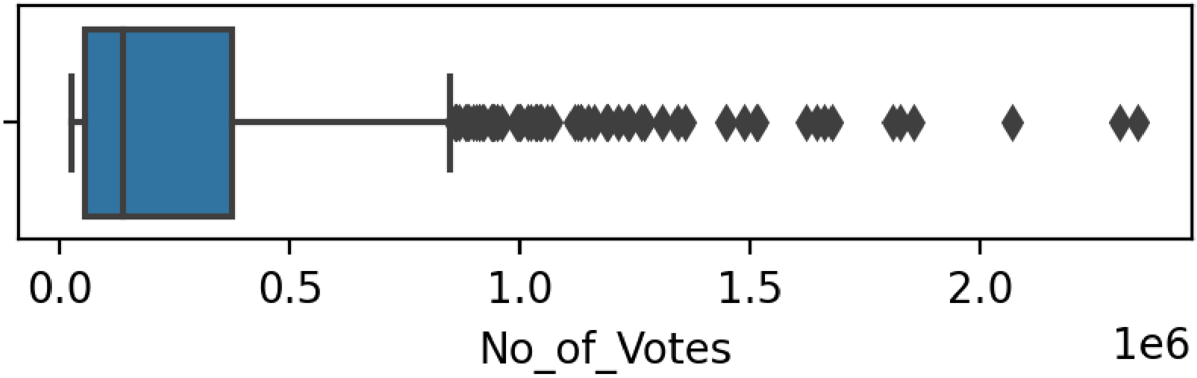
Analysis:

1. Above diagrams represent the relationship of revenue with the other features using regression plots in seaborn.
2. From the above graphs we can conclude that there is a positive correlation between the number of votes and a gross generated. Most of the movies with higher number of votes have higher revenue generated.
3. From the graph of Released_Year vs gross we can conclude that in each new year, the revenue of the movies released is increasing
4. The runtime doesn't seem to have much impact on the revenue it generates, so we cannot necessarily say that the movies of greater runtime will have higher revenue. Even though from the scatter plot we can conclude that the movie with the running time between 130-200 min can have higher revenue.
5. Reviewing boxplot of the Meta_Score vs revenue we can say that some of the movies which have the highest meta score have low gross, hence meta score has no positive correlation with gross.
6. IMDB Rating can be somewhat a predictor of the revenue generated for some movies. It cannot be stated surely because multiple movies with the same IMDB ratings have different values of gross generated.

- Box Plots


50
100
150
200
250
300
Runtime
7.75
8.00
8.25
8.50
8.75
9.00
9.25
IMDB_Rating
0
20
40
60
80
100
Meta_score





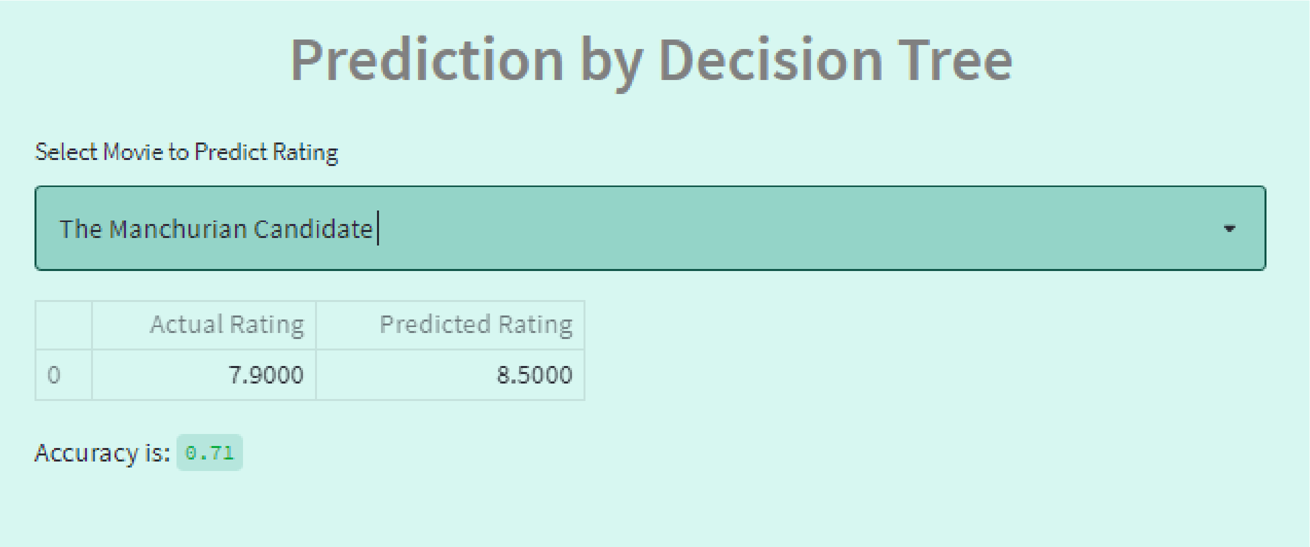
Analysis:

1. The above diagram shows boxplot for the features Runtime, IMDB_Rating, Meta_Score, released_year, and No_of_Votes.
2. Each boxplot gives information minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value.
3. By using a boxplot we are trying to compare values range of the different variables which will be used for the prediction in a later stage of a project and identifying outliers
4. For Runtime minimum value is around 70, median is 125, and maximum value is around 190. We can also observe that there are many outliers which are above maximum value 190.

5. For IMDB_Rating minimum value is 7.5, median is 7.85 and maximum value is around . We can also observe that some of the ratings are above maximum value.
6. For Meta_Score minimum value is around 30, median is 76, and maximum value is around 100.
7. For Released_Year minimum value is around 1979, and maximum value is around 2020.
8. In the features IMDB_rating, Runtime, and No_of_votes we can see many outliers which shows that the data is quite spreaded out.

IMDB_Rating Prediction

- In this project after the visualizations are done we are trying to predict the IMDB_Rating for the movies by using two machine learning algorithms namely Decision Tree Classifier and Random Forest.
- For both of the algorithms features(X) used to make predictions are: Meta_score , No_of_Votes, Gross , Runtime , and Released_Year
- Prediction Parameter (y): IMDB_Rating
- In this project we have created a dashboard which provides a drop down menu to select a movie for rating prediction.
- If we want to predict the rating of the entirely new movie, then we can add that movie into the test dataframe manually and predict the rating using a dashboard.



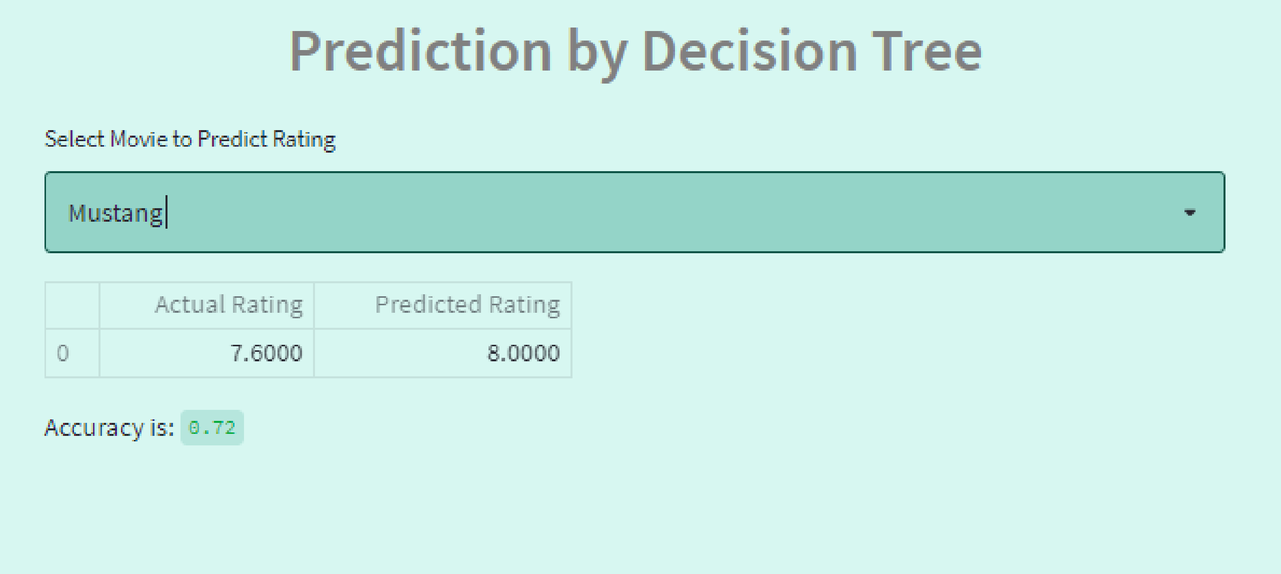
Prediction by Decision Tree

Select Movie to Predict Rating

The Manchurian Candidate

	Actual Rating	Predicted Rating
0	7.9000	8.5000

Accuracy is: 0.71



Prediction by Decision Tree

Select Movie to Predict Rating

Mustang

	Actual Rating	Predicted Rating
0	7.6000	8.0000

Accuracy is: `0.72`

Prediction By Decision Trees:

1. The decision trees the data is interpreted in a form of a tree. There are decision nodes and leaf nodes.
2. The decision nodes are used to split the data based on a condition. The leaf nodes are used to decide the class of a new data point.
3. At each point the decision tree tries to maximize the information gain and minimize entropy which is calculated using the gini index in our code. The mathematical representation for Gini Index is :

   $$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

   Where, C is the total number of classes and p(*i*) is the probability of picking the data point with the class i.

   Mathematical representation for Information Gain is:

   IG = G(parent) −G(children)

4. The decision tree classifier in sklearn uses ‘gini’ criteria by default for calculating entropy.

5. We have divided the IMDB_rating into bins so that the more precise values can be considered in the form of categorical data.
6. Next, we have split the data into 4 major categories - X_train, X_test, y_train, y_test by taking 70% data for training and 30% data for testing.
7. X_train and y_train are input and output features for the training dataset and X_testand y_test are the input and output features for the testing dataset.
8. Then we applied the classifier and fitted it for the training dataset.
9. Next step is to predict the values of the X_test and compare it to the y_test values.
10. We have calculated the accuracy and confusion matrix to evaluate the results.
11. It has been observed that the accuracy of prediction by using Decision Tree classifier is around 72%.

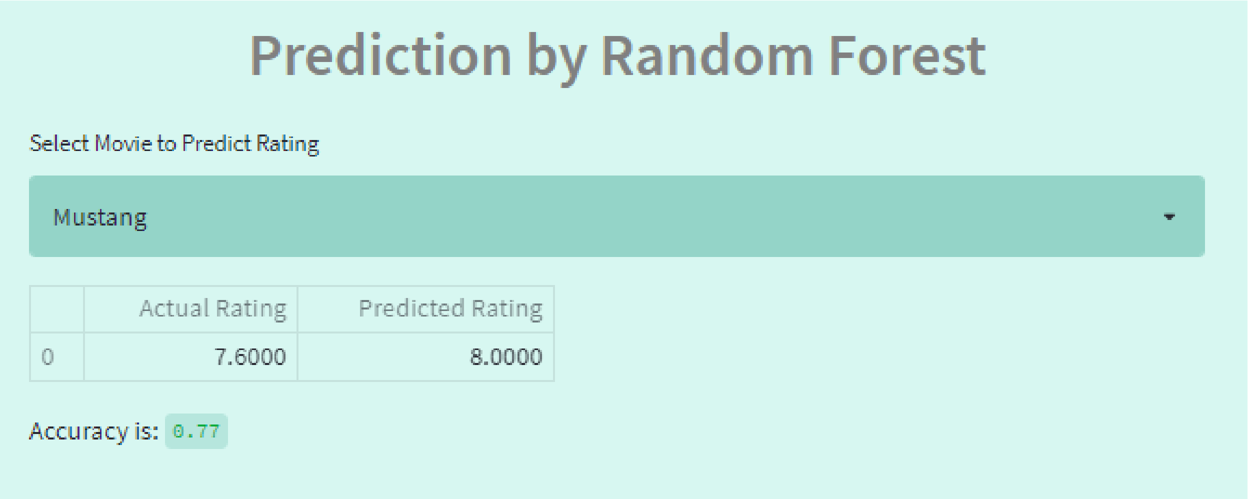
Advantages of using decision trees:

- Decision trees can be easily used for the classification as well as regression problems.
- Compared with other algorithms, pre-processing of datas in a decision tree requires less effort and does not require normalization of data.
- The decision tree has no assumptions about space distributions and classifier structure.

Disadvantages of using Decision Tree Classifier:

- A small change in the data tends to cause a big difference in the tree structure, which causes instability.
- Calculations involved can also become complex compared to other algorithms, and it takes a longer time to train the model.

Prediction By Random Forest:



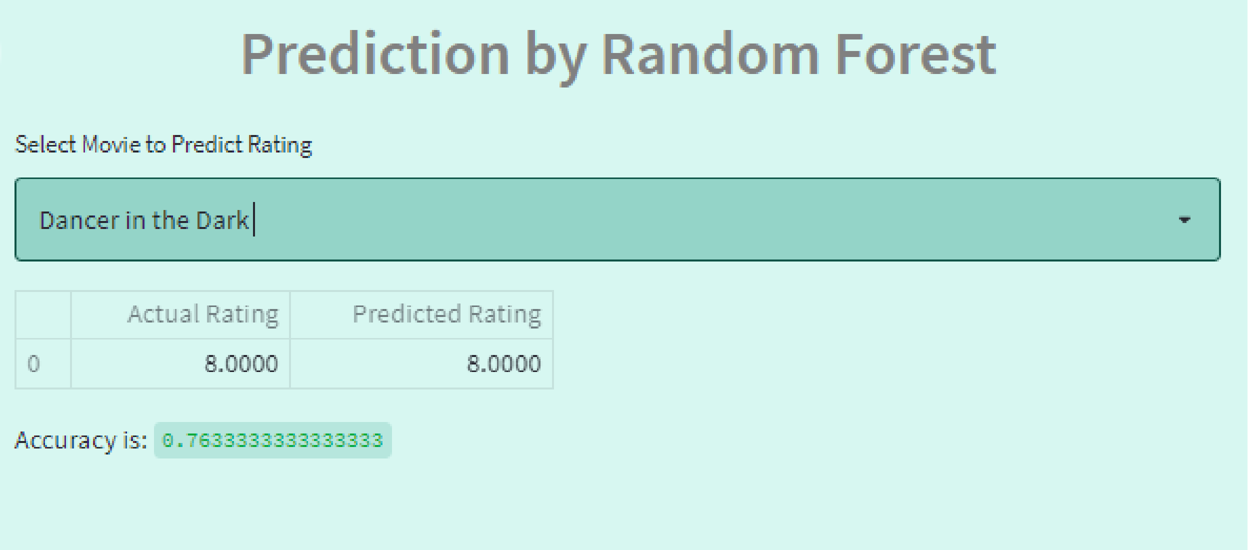
Prediction by Random Forest

Select Movie to Predict Rating

Mustang

	Actual Rating	Predicted Rating
0	7.6000	8.0000

Accuracy is: `0.77`



Prediction by Random Forest

Select Movie to Predict Rating

Dancer in the Dark

	Actual Rating	Predicted Rating
0	8.0000	8.0000

Accuracy is: `0.7633333333333333`

1. As Decision trees are highly sensitive to the training data, if we change only a small set of values in training data the entire decision tree changes. This might result in high variance and our model will fail to generalize.

2. To avoid these problems, in a random forest algorithm, we generate multiple decision trees and use majority voting/average to predict the value of a new point.

3. Random forest algorithm is performed in 2 steps, first is bootstrapping and second is aggregation together called bagging.

4. In bootstrapping multiple decision trees are generated by selecting random rows and random features for each tree from the given dataset. This ensures that we are not using the same data every time.

5. Random feature selection reduces the correlation between the trees which helps to improve accuracy of a decision tree.

6. To apply random forest from sklearn on our dataset we have divided IMDB_Ratings into bins.

7. While applying a random forest algorithm, we have given the n_estimators parameter as 200 which states the number of trees in a forest.

8. We have calculated the confusion matrix and accuracy score using predicted result and actual result to calculate how many predictions are correct.

Advantages of using Random Forest Classifier:

- Random Forest is based on the bagging algorithm and uses Ensemble Learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces overfitting problems in decision trees and also reduces the variance and therefore improves the accuracy.
- Random Forest is robust to outliers and can handle them automatically.
- The Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.
- Random Forest is less impacted by noise.

Disadvantages of using Random Forest:

- Random Forest creates a lot of trees and combines their outputs. By default, it creates 100 trees in the Python sklearn library. To do so, this algorithm requires much more computational power and resources.
- Random Forest requires much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes the decision on the majority of votes.

In conclusion, to predict IMDB_Rating, Random Forest algorithm (76%) works better than decision trees (72%).