AMITY SCHOOL OF APPLIED SCIENCES

NTCC PROJECT: PREDICTING BOX OFFICE SUCCESS OF MOVIES

Project Report By

Mr. Kaushlendra

Bachelor Of Statistics (3rd year)

Amity School Of Applied Sciences

Amity University Utter Pradesh, Lucknow

Project Supervisor
Dr. Gunjan Singh
Assistant Professor
Department Of Statistics
Amity School Of Applied Sciences
Amity University Uttar Pradesh,

Professor And Head
Dr. Asita Kulshreshtha
Department Of Statistics
Amity School Of Applied Sciences
Amity University Uttar Pradesh,

Lucknow Lucknow

CANDIDATES DECLARATION

"PREDICTING BOX OFFICE SUCCESS OF MOVIES," reads the project report's title. I acknowledge that I am aware of what plagiarism is and that the university has a policy against it. I further declare that:

a) The work I have submitted as partial fulfilment of the requirement for the award of a Bachelor of Statistics degree is entirely original and has not been submitted for evaluation elsewhere.

b) I affirm that I am the sole author of this project. Every time work from a different source was utilized, all debts were properly acknowledged and cited in compliance with NTCC Regulations and Guidelines.

c) I have not submitted any previously completed work that was completed by another student as my own work.

d) I have not allowed and will not allow anyone to change or replicate my work and claim it as their own.

e) The work complies with the regulations and guidelines' requirements for style, content, and layout.

Date: **Student**

Signature

AMITY UNIVERSITY
—LUCKNOW CAMPUS—

PLAGIARISM DECLARATION

This is to clarify that internship project detailed below has been evaluated by online anti plagiarism software the content and materials was found satisfactory within the permissible limit of content copied.

Name of the Student ……………………………………

Enrolment Number ………………………………………

Program and Batch ………………………………………

Title of the Project ……………………………………………….

Results ………………………………………………

Student **Internal Faculty Supervisor**

Date:

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my supervisor Dr. Gunjan Singh, Assistant Professor, Amity School of Applied Sciences, Amity University Uttar Pradesh for their wisdom and recommendations that made it possible for me to write this article. She supported me despite her busy schedule, and without his constant direction, this dissertation would never have taken the current form.

I want to express my gratitude to Dr. Asita Kulshreshtha, the head of the Amity School of Applied Sciences at Amity University Uttar Pradesh in Lucknow, for constantly inspiring me and for granting me permission to complete this project away from the university. I am also appreciative of my inner guiding.

Finally, I want to express my gratitude to my parents and extended family for their unwavering support. Last but not least, I want to express my gratitude to all of my friends for their assistance.

Mr. Kaushlendra

Bachelor Of Statistics (3rd year)

Amity School Of Applied Sciences Amity
University Utter Pradesh, Lucknow

CERTIFICATE

This is to certify that Mr. Kaushlendra (Enrolment Number: A8979121008) has completed the research for the Bachelor of Statistics degree from the Department of Statistics, Amity School of Applied Sciences, Amity University Uttar Pradesh, Lucknow, on "PREEDICTING BOX OFFICE SUCCESS OF MOVIES." Dr. Gunjan Singh, Assistant Professor, Department of Statistics, Amity University Uttar Pradesh, Lucknow, oversees its completion. Dr. Asita Kulshreshtha, Head of Institution, Department of Statistics, Amity School of Applied Sciences, Amity University Uttar Pradesh, Lucknow. The dissertation represents the results of the student's own original research and study, and its contents do not serve as the foundation for the awarding of any other degree to the candidate or anyone else.

Project Supervisor
Dr. Gunjan Singh
Assistant Professor
Department Of Statistics
Amity School Of Applied Sciences
Amity University Uttar Pradesh,
Lucknow

Professor And Head
Dr. Asita Kulshreshtha
Department Of Statistics
Amity School Of Applied Sciences
Amity University Uttar Pradesh,
Lucknow

TABLE OF CONTEN

Contents

TABLE OF FIGURES:

LITRATURE REVIEW

Studies on the prognostication of film box office success highlight the importance of film qualities (Simonoff and Sparrow, 2010), elucidating the direct impact of genre, budget, and star power on a film's financial performance. The importance of the digital era on consumer behavior is highlighted by the rise of social media measures, notably online sentiment and Twitter activity, as crucial predictors of a film's potential success (Asur and Huberman, 2010; Youn and Jin, 2017).

The importance of strategic planning in the industry is highlighted by the fact that avoiding competition and choosing a smart release date are also essential for increasing a film's box office appeal (Einav, 2007). The literature proposes a composite method to box office prediction that incorporates digital engagement data, traditional movie qualities, and strategic variables. However, the difficulty of precisely forecasting in the face of fluctuating customer preferences persists.

The constant improvement of analytical models and the incorporation of various data sources are necessary to enhance the precision of box office forecasts, which take into account the ever-changing film industry and customer preferences.

ABSTRACT

This study offers a thorough method for utilizing 22machine learning techniques to forecast a film's box office success. The m23ain objective is to evaluate the practicality of Random Forest and Linear Regres25sion models using a dataset that is composed of the 1000 most popular movi52es that have been listed on IMDb in the last ten years. Important characteristics including runtime, metascores, ratings, and genre were carefully examined to see how they affected box office receipts. The research highlights the crucial impact of audience choices and critical reaction in influencing box office outcomes through a comprehensive factor analysis. The study compares the predictive powers of Random Forest and Linear Regression models, assessing each model's performance using measures such as Mean Squared Error (MSE) and Coefficient of Determination (R^2). The models' respective levels of success are shown by the results, with some clear differences in how well they handle linear and nonlinear interactions. The results highlight the complexity of financial outcome forecasting and the film industry's diverse structure. The study emphasizes how crucial it is to incorporate larger datasets and continuously update models in order to increase forecast accuracy. In an attempt to manage the complex dynamics of market trends and consumer behavior, this inquiry offers insightful information to players in the film business.

Keywords: Box Office Prediction, Linear Regression, Random Forest, Factor Analysis, Machine Learning.

CHAPTER 1: INTRODUCTION

Over the deacde, there has been a notable increase in the production and consumption of movies in the entertainment sector. Websites like as IMDb offer a platform for users to interact and discover a wide selection of films. The goal of this NTCC project is to make use of a dataset that includes the 1,000 most watched movies on IMDb for the previous ten years. Numerous features are included in the dataset, such as the film's title, genre, synopsis, director, leading actors, year of release, duration, user rating, number of votes, income, and Meta score.

Analyzing and forecasting movie box office success based on multiple contributing elements is the main objective of this research. By thoroughly analyzing these characteristics, we hope to find patterns, trends, and connections that affect a film's critical and commercial success. With machine learning techniques, namely regression models, our goal is to develop predictive models that can estimate the box office receipts of a film and further our understanding of the dynamics of the film industry.

We begin by conducting an exploratory data analysis (EDA) to begin our investigation. To get insights, spot any trends, and reveal any innate qualities, the dataset must be carefully examined.

CHAPTER 2: BACKGROUND

Predicting box office success requires an analytical framework that takes into account a complex interaction of factors included in a dataset that includes 1,000 of the most popular movies that have been published on IMDB in the last ten years. Important elements including Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, and Metascore are all included in this extensive dataset; each one provides a distinct perspective for statistical analysis and machine learning applications.

The dataset makes it possible to build a complex model from a statistical standpoint, where each characteristic can be assessed for its ability to forecast outcomes and correlation with the box office performance of the film, measured in terms of revenue. For instance, the genre may be examined as a categorical variable to gain understanding of the financial performance of distinct movie genres both inside and between market niches. Time-series analysis is made possible by the Year feature, which shows trends and changes in moviegoers' choices throughout time.

Natural Language Processing (NLP) techniques may be applied to textual data in the Description field to convert qualitative synopses into quantitative measures that may be associated with a film's success. In a similar vein, the disciplines of directors and actors make it easier to investigate the influence of human capital since they allow for the quantitative and statistical evaluation of the reputation and prior performance of those involved.

The Runtime, Rating, and Metascore are numerical data points that may be utilized as continuous variables to reflect various elements of a film's qualities and reception in the marketplace, or they can be used to improve machine learning algorithms. Another important quantitative metric is votes, which indicate popularity and audience participation and are perhaps correlated with financial performance.

The research will use machine learning techniques including random forests, decision trees, and linear regression to create prediction models based on this information. To evaluate each model's accuracy and efficacy in forecasting box office receipts, statistical measurements like the Mean Squared Error (MSE) and the Coefficient of Determination (R^2) will be used. The model might be improved by utilizing feature selection strategies and dimensionality reduction techniques such as Principal Component Analysis (PCA), which pinpoint the key factors influencing box office performance.

Using these statistical and machine learning techniques, this chapter lays the groundwork for a thorough examination that attempts to unravel the many different components that go into a film's financial performance. In addition to following the tenets of data-driven decision-making, this strategy is in line with recent developments in entertainment analytics and provides insights into the factors that influence box office performance in the contemporary film business.

CHAPTER 3: DATA COLLECTION AND PRE PROCESSING

The Internet Movie Database (IMDb) provided the main dataset for this study, which included important details on the top 1000 films from the previous ten years, such as titles, genres, directors, actors, release years, runtimes, user ratings, vote counts, revenues, and Meta scores.

Pre-processing Procedures:

i. Initial Loading: To enable further modification and analysis, data was imported into a structured format using Python's Pandas package.

ii. **Quality assessment:** To determine the dataset's structure and completeness, a preliminary assessment was conducted to spot any inconsistent data inputs.

iii. **Cleaning Operations:** To guarantee analytical accuracy, this stage corrected missing values and removed duplicate entries, addressing data cleanliness.

iv. **Statistical Summary:** The dataset's insights were obtained through the generation of descriptive statistics, which guided further preprocessing procedures.

v. **Feature Manipulation:** To make dataset characteristics more suitable for the predictive modeling procedure, significant improvements and alterations were done to them.

vi. **Exploratory Analysis:** To identify underlying trends and develop theories for the prediction framework, a thorough EDA was carried out.

vii. **Dataset Segmentation:** To facilitate the subsequent stages of model validation and assessment, the data was divided into training and testing subsets.

In order to ensure the integrity and trustworthiness of the study outputs, the meticulous pre-processing routine created a solid basis for the future analytical and predictive modeling endeavors.

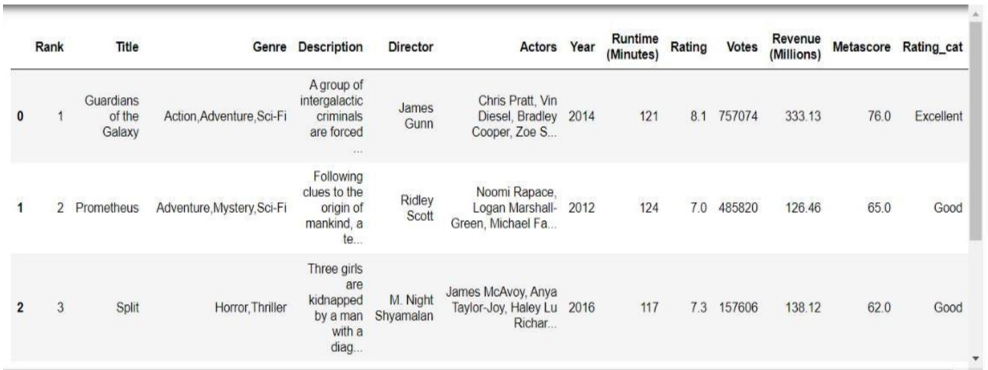
READING CSV FILE:

```
data=pd.read_csv('C:/Users/yadav/OneDrive/Desktop/IMDB-Movie-
 Data.csv')
```

In [52]:

```
data.head(5)
```

Out[52]:

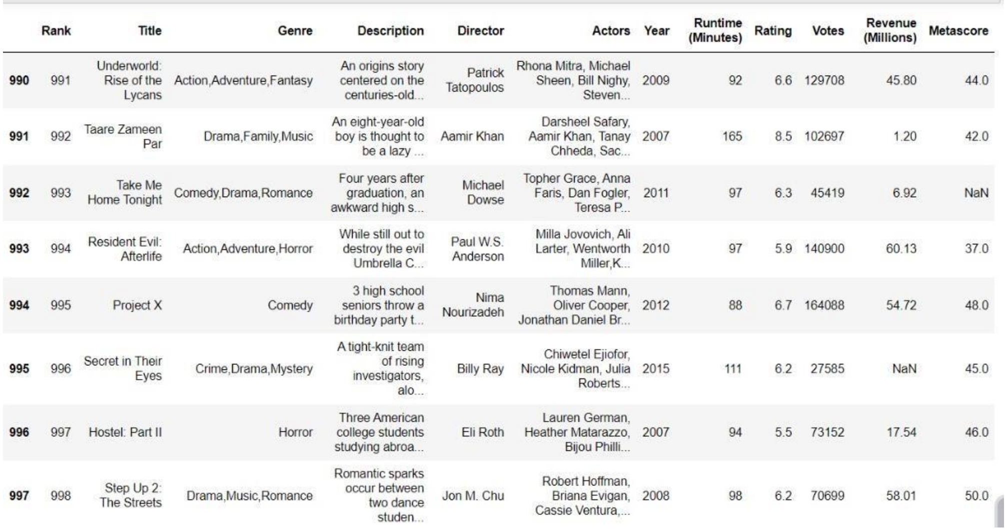


	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore	Rating_cat
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0	Excellent
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0	Good
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0	Good

In [1]:

```
data.tail(10)
```

Out[1]:



	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
990	991	Underworld: Rise of the Lycans	Action,Adventure,Fantasy	An origins story centered on the centuries-old...	Patrick Tatopoulos	Rhona Mitra, Michael Sheen, Bill Nighy, Steven...	2009	92	6.6	129708	45.80	44.0
991	992	Taare Zameen Par	Drama,Family,Music	An eight-year-old boy is thought to be a lazy ...	Aamir Khan	Darsheel Safary, Aamir Khan, Tanay Chheda, Sac...	2007	165	8.5	102697	1.20	42.0
992	993	Take Me Home Tonight	Comedy,Drama,Romance	Four years after graduation, an awkward high s...	Michael Dowse	Topher Grace, Anna Faris, Dan Fogler, Teresa P...	2011	97	6.3	45419	6.92	NaN
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13	37.0
994	995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012	88	6.7	164088	54.72	48.0
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	111	6.2	27585	NaN	45.0
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	46.0
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01	50.0

In [2]:

```
data.shape
```

Out[2]:

```
(1000, 12)
```

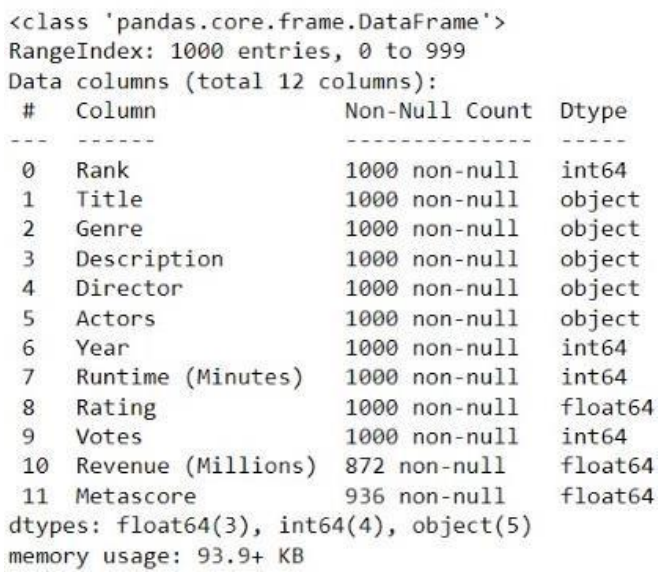
In [3]:

```
print("Number of row" ,data.shape[0])
print("Number of columns" ,data.shape[1])
```

Out[4]:

```
Number of rows 1000

Number of columns 12
```

Out [5:]



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Rank                1000 non-null   int64
 1   Title               1000 non-null   object
 2   Genre               1000 non-null   object
 3   Description         1000 non-null   object
 4   Director            1000 non-null   object
 5   Actors              1000 non-null   object
 6   Year                1000 non-null   int64
 7   Runtime (Minutes)   1000 non-null   int64
 8   Rating              1000 non-null   float64
 9   Votes               1000 non-null   int64
 10  Revenue (Millions)  872 non-null    float64
 11  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
```

In [6]:

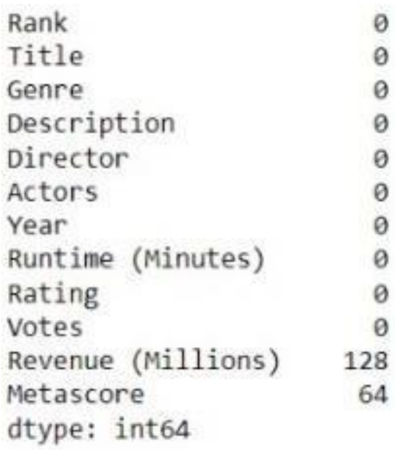
```
print("Any Missing values?" ,data.insull().value.any())
```

Out{6} : Any missing values? True

In [7]:

```
data.insull().sum()
```

Out[7]:



```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore              64
dtype: int64
```

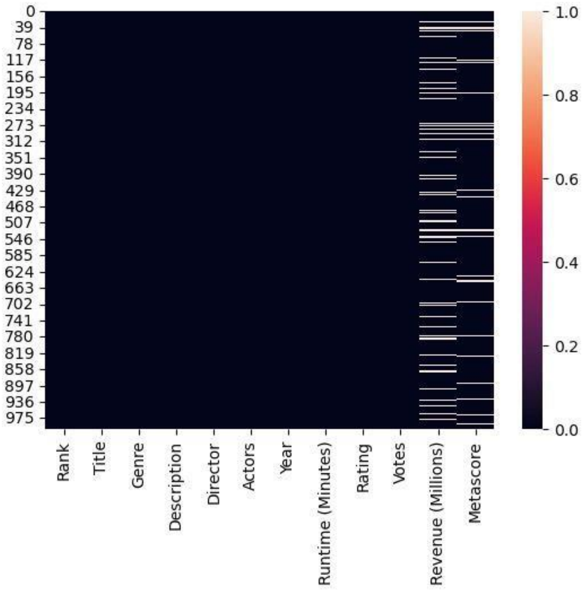
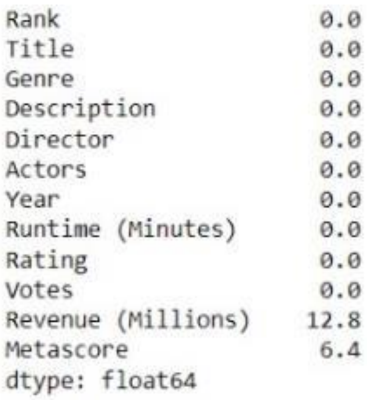


Fig 1.3: The annual number of films, displayed as a bar graph

In [8]:

```
per_missing = data.insull().sum()*100/len(data) per_missing
```

Out[9]:

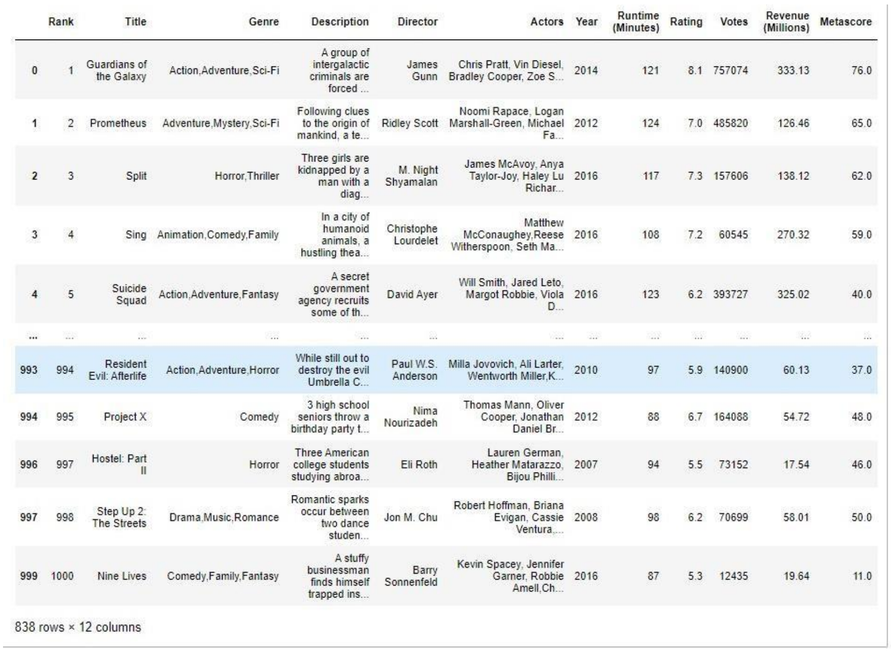


```
Rank                   0.0
Title                  0.0
Genre                  0.0
Description            0.0
Director               0.0
Actors                 0.0
Year                   0.0
Runtime (Minutes)      0.0
Rating                 0.0
Votes                  0.0
Revenue (Millions)    12.8
Metascore              6.4
dtype: float64
```

In [9]:

```
data.dropna(axis=0)
```

Out[9]:



	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0
...	...	...	...	...	...	...	...	...	...	...	...	...
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13	37.0
994	995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012	88	6.7	164088	54.72	48.0
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	46.0
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01	50.0
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64	11.0

838 rows × 12 columns

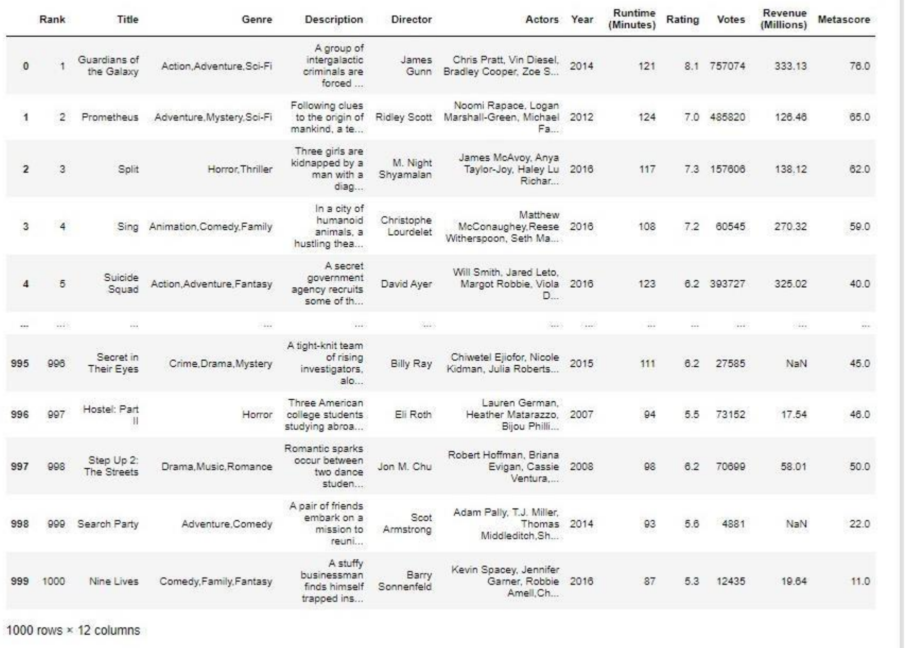
In [10]:

```
dup_data=data.duplicated().any()
printAre there any duplicate values? False(“Are there any duplicate values?”  ,dup_data)
```

In [11]:

```
data=data.drop_duplicates() data
```

Out[11]:



	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0
...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	111	6.2	27585	NaN	45.0
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	46.0
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01	50.0
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014	93	5.6	4881	NaN	22.0
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64	11.0

1000 rows × 12 columns

In [12]:

```
dup_data=data.duplicated().any()
print(“Are there any duplicate values?” ,dup_data)
```

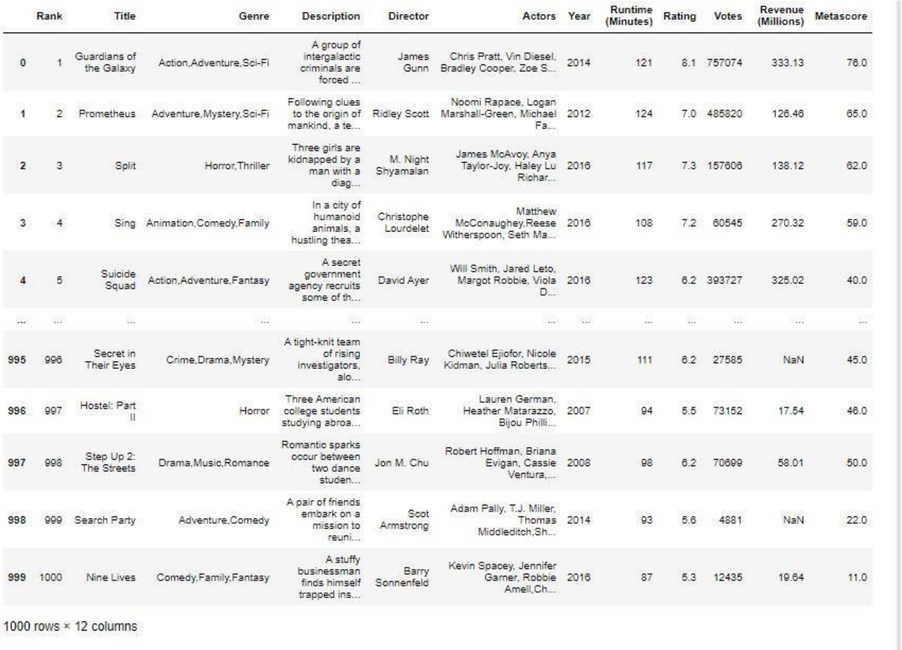
Out[12]:

```
Are there any duplicate values? False
```

In [13]:

```
data=data.drop_duplicate() data
```

Out[13]:



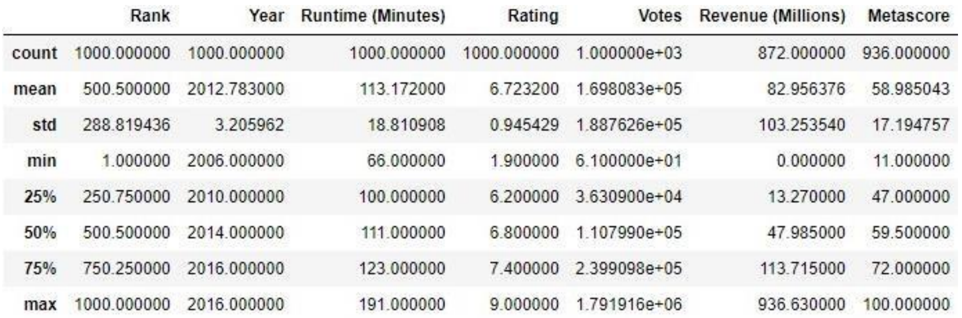
	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0
...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	111	6.2	27585	NaN	45.0
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	46.0
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01	50.0
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014	93	5.6	4881	NaN	22.0
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64	11.0

1000 rows × 12 columns

In [14]:

```
data.describe()
```

Out[14]:

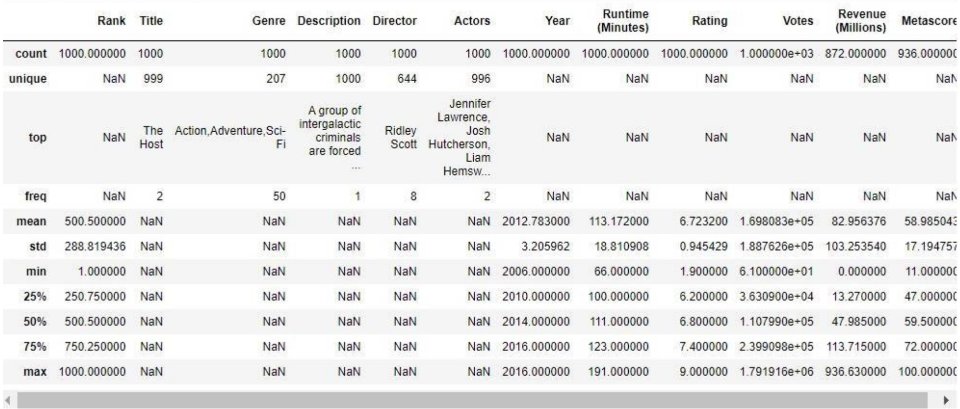


	Rank	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	872.000000	936.000000
mean	500.500000	2012.783000	113.172000	6.723200	1.698083e+05	82.956376	58.985043
std	288.819436	3.205962	18.810908	0.945429	1.887626e+05	103.253540	17.194757
min	1.000000	2006.000000	66.000000	1.900000	6.100000e+01	0.000000	11.000000
25%	250.750000	2010.000000	100.000000	6.200000	3.630900e+04	13.270000	47.000000
50%	500.500000	2014.000000	111.000000	6.800000	1.107990e+05	47.985000	59.500000
75%	750.250000	2016.000000	123.000000	7.400000	2.399098e+05	113.715000	72.000000
max	1000.000000	2016.000000	191.000000	9.000000	1.791916e+06	936.630000	100.000000

In [15]:

```
data.describe(include)= 'all'
```

Out[15]:

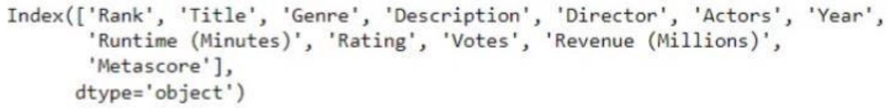


	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
count	1000.000000	1000	1000	1000	1000	1000	1000.000000	1000.000000	1000.000000	1.000000e+03	872.000000	936.00000
unique	NaN	999	207	1000	644	996	NaN	NaN	NaN	NaN	NaN	Na
top	NaN	The Host	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	Ridley Scott	Jennifer Lawrence, Josh Hutcherson, Liam Hemsw...	NaN	NaN	NaN	NaN	NaN	Na
freq	NaN	2	50	1	8	2	NaN	NaN	NaN	NaN	NaN	Na
mean	500.500000	NaN	NaN	NaN	NaN	NaN	2012.783000	113.172000	6.723200	1.698083e+05	82.956376	58.98504
std	288.819436	NaN	NaN	NaN	NaN	NaN	3.205962	18.810908	0.945429	1.887626e+05	103.253540	17.19475
min	1.000000	NaN	NaN	NaN	NaN	NaN	2006.000000	66.000000	1.900000	6.100000e+01	0.000000	11.00000
25%	250.750000	NaN	NaN	NaN	NaN	NaN	2010.000000	100.000000	6.200000	3.630900e+04	13.270000	47.00000
50%	500.500000	NaN	NaN	NaN	NaN	NaN	2014.000000	111.000000	6.800000	1.107990e+05	47.985000	59.50000
75%	750.250000	NaN	NaN	NaN	NaN	NaN	2016.000000	123.000000	7.400000	2.399098e+05	113.715000	72.00000
max	1000.000000	NaN	NaN	NaN	NaN	NaN	2016.000000	191.000000	9.000000	1.791916e+06	936.630000	100.00000

In [16]:

```
data.columns
```

Out[16]:

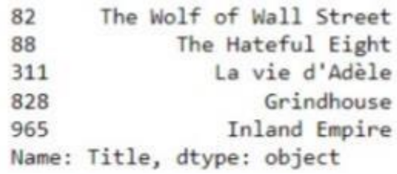


```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

In [17]:

```
data[data['Runtime (Minutes)']>=180]['Title']
```

Out[17]:

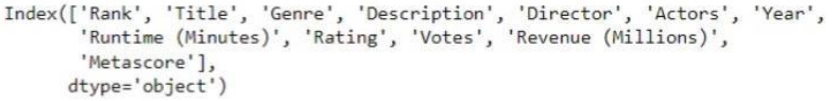


```
82      The Wolf of Wall Street
88            The Hateful Eight
311             La vie d'Adèle
828                 Grindhouse
965              Inland Empire
Name: Title, dtype: object
```

In [18]:

```
data.columns
```

Out[18]:



```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

In [19]:

```
data.groupby('Year')['Votes'].mean().sort_values(ascending=False)
```

Out[19]:

```
Year
2012    285226.093750
2008    275505.384615
2006    269289.954545
2009    255780.647059
2010    252782.316667
2007    244331.037736
2011    240790.301587
2013    219049.648352
2014    203930.224490
2015    115726.220472
2016     48591.754209
```

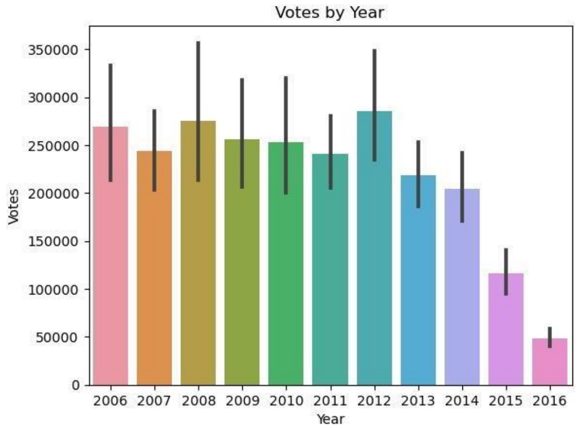
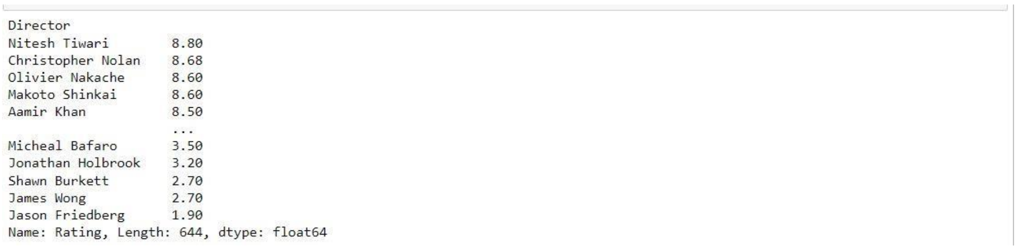


Fig2.3: Highlights well-known filmmakers' average ratings.

In [20]:

```
data.groupby('Director')['Rating'].mean().sort_values(ascending
=False)
```

Out[20]:

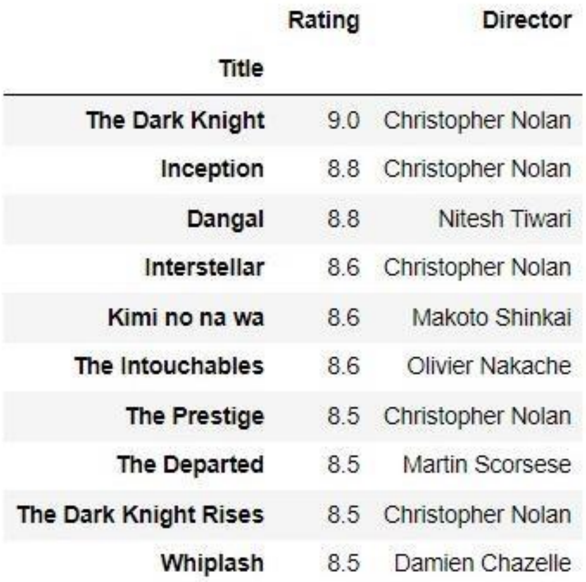


```
Director
Nitesh Tiwari        8.80
Christopher Nolan    8.68
Olivier Nakache      8.60
Makoto Shinkai       8.60
Aamir Khan           8.50
                     ...
Micheal Bafaro       3.50
Jonathan Holbrook    3.20
Shawn Burkett        2.70
James Wong           2.70
Jason Friedberg      1.90
Name: Rating, Length: 644, dtype: float64
```

In [21]:

```
import pandas as pd

data = pd.read_csv('C:/Users/yadav/OneDrive/Desktop/IMDB-
MovieData.csv') top10_rating = data.nlargest(10, 'Rating')[['Title',
'Rating', 'Director']]\

.set_index('Title') top10_rating
```

Out[21]:

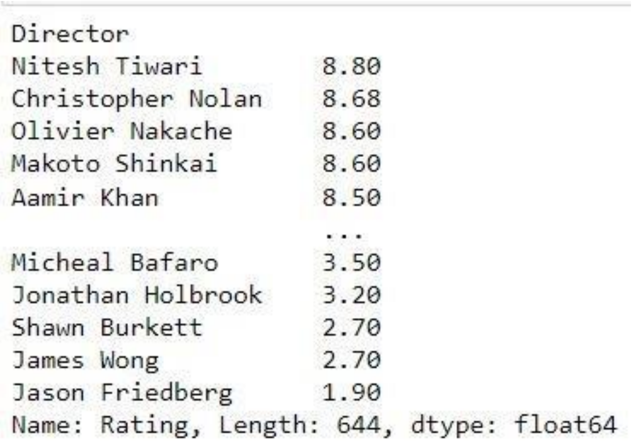


	Rating	Director
Title		
The Dark Knight	9.0	Christopher Nolan
Inception	8.8	Christopher Nolan
Dangal	8.8	Nitesh Tiwari
Interstellar	8.6	Christopher Nolan
Kimi no na wa	8.6	Makoto Shinkai
The Intouchables	8.6	Olivier Nakache
The Prestige	8.5	Christopher Nolan
The Departed	8.5	Martin Scorsese
The Dark Knight Rises	8.5	Christopher Nolan
Whiplash	8.5	Damien Chazelle

In [22]:

```
data.groupby('Director')['Rating'].mean().sort_values(ascending=False)
```

Out[22]:



```
Director
Nitesh Tiwari          8.80
Christopher Nolan      8.68
Olivier Nakache        8.60
Makoto Shinkai         8.60
Aamir Khan             8.50
                       ...
Micheal Bafaro         3.50
Jonathan Holbrook      3.20
Shawn Burkett          2.70
James Wong             2.70
Jason Friedberg        1.90
Name: Rating, Length: 644, dtype: float64
```

In [23]:

```
top10_len = data.nlargest(10, 'Runtime (Minutes)')[['Title',
```

n [24]:

```
sns.barplot(data=top10_len, x='Runtime (Minutes)',
y=top10_len.index)
plt.title('Top 10 Lengthy Movies')

plt.show() (Minutes)']]\
.set_index('Title')
```

Out[24]:

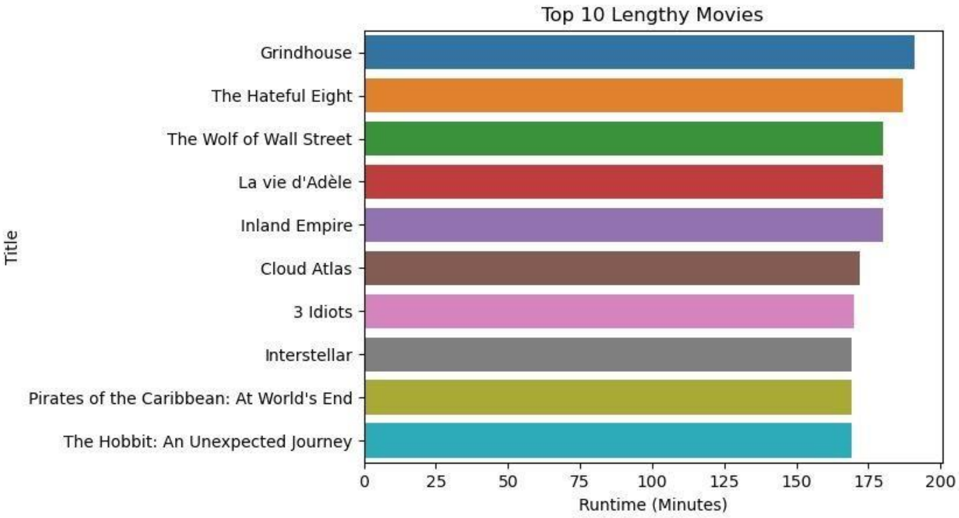
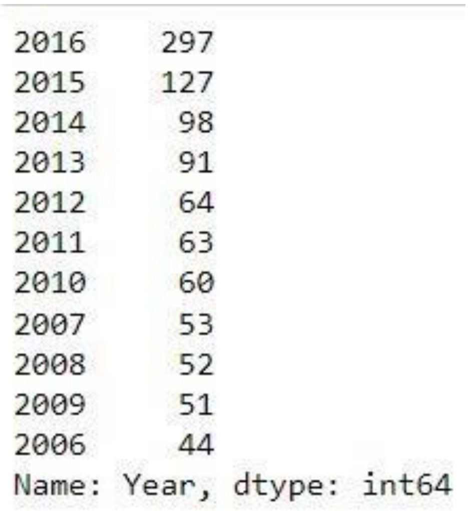


Fig 3.3 Trends in annual film releases over time.

In [25]:

```
data['Year'].value_counts()
```

Out[25]:



```
2016    297
2015    127
2014     98
2013     91
2012     64
2011     63
2010     60
2007     53
2008     52
2009     51
2006     44
Name: Year, dtype: int64
```

In [26]:

```
sns.countplot(data=data, x='Year')
plt.title('Number Of Movies Per Year')
plt.show()
```

Out[26]:

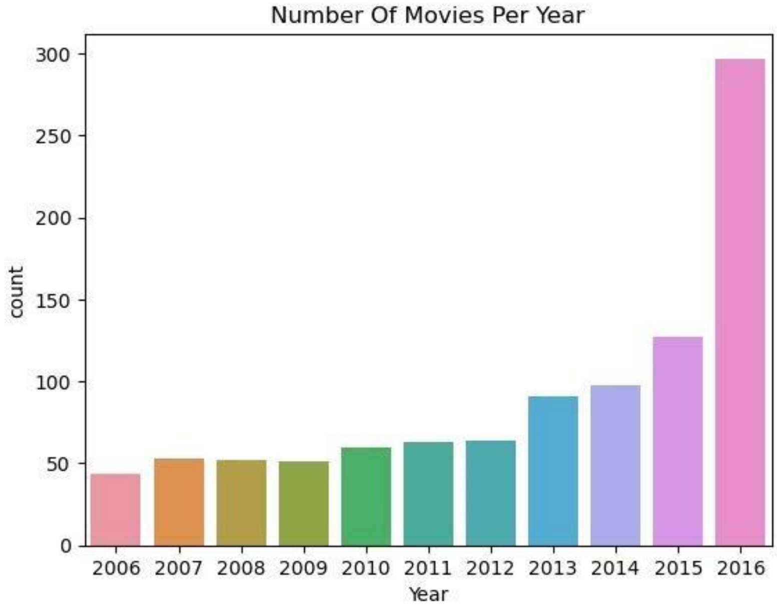
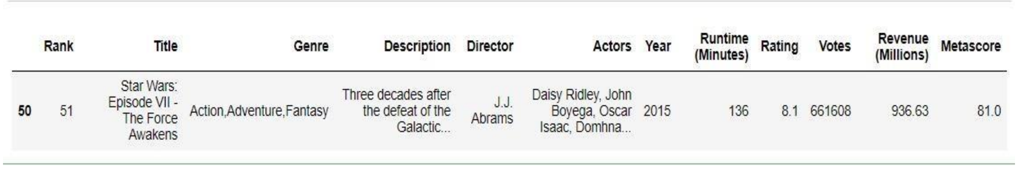


Fig 4.3: A visual representation of the ten longest films.

In [27]:

```
data[data['Revenue (Millions)'].max() == data['Revenue
(Millions)']]
```

Out[28]:

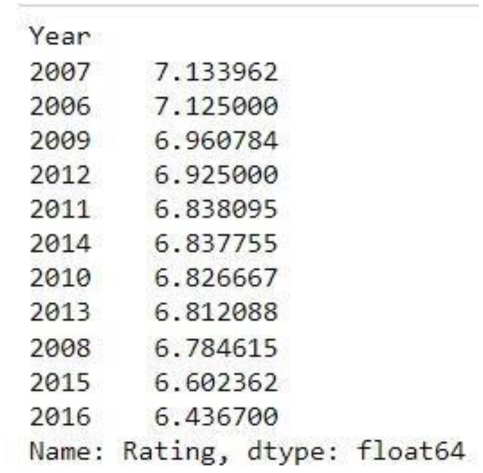


	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
50	51	Star Wars: Episode VII - The Force Awakens	Action,Adventure,Fantasy	Three decades after the defeat of the Galactic...	J.J. Abrams	Daisy Ridley, John Boyega, Oscar Isaac, Domhna...	2015	136	8.1	661608	936.63	81.0

In[29]:

```
data.groupby('Year')['Rating'].mean().sort_values(ascending=False)
```

Out[29]:



```
Year
2007    7.133962
2006    7.125000
2009    6.960784
2012    6.925000
2011    6.838095
2014    6.837755
2010    6.826667
2013    6.812088
2008    6.784615
2015    6.602362
2016    6.436700
Name: Rating, dtype: float64
```

In[30]:

```
sns.barplot(data=data, x='Year', y='Rating')
```

Out[30]:

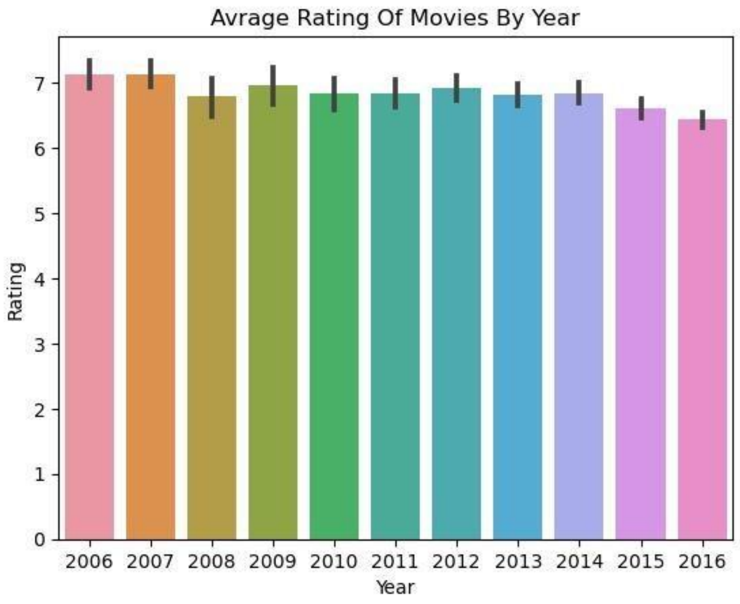


Fig 5.3: A bar plot displaying the average movie ratings annually.

In [31]:

```
sns.barplot(data=data, x='Year', y='Rating')
plt.title('Avrage Rating Of Movies By Year') plt.show()
```

Out [31]:

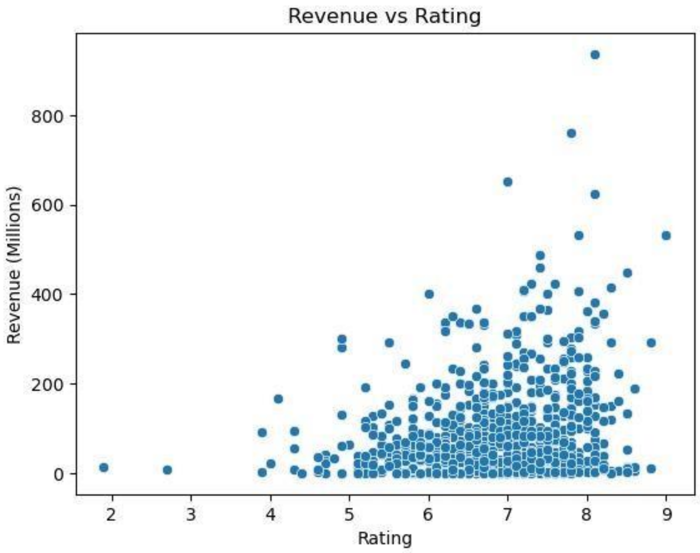


Fig 6.3: Movie classification chart derived from user ratings.

In [32]:

```
def rating(r):
if r >= 7.5:
        return 'Excellent'
elif r >= 6:
        return 'Good'
elif r >= 5:
        return 'Average'
else:
        return 'Bad'
```

In [32]:

```
data['Rating_cat'] = data['Rating'].apply(rating)
```

In [31]:

```
sns.countplot(data=data, x='Rating_cat')
plt.title('Number Movies in each Rating
Category') plt.xlabel('Rating Category')
plt.ylabel('Number of Movies') plt.show()
```

Out[31]:

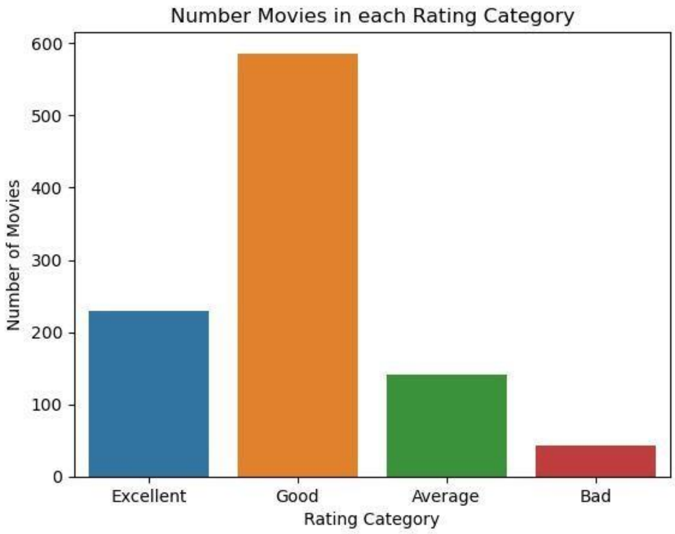


Fig 7.3: Revenue analysis by movie rating categories.

In [32]:

```
sns.barplot(data=data, x='Rating_cat', y='Revenue
(Millions)') plt.title('Revenue for each Rating Category')
plt.xlabel('Rating Category')
plt.show()
```

Out[32]:

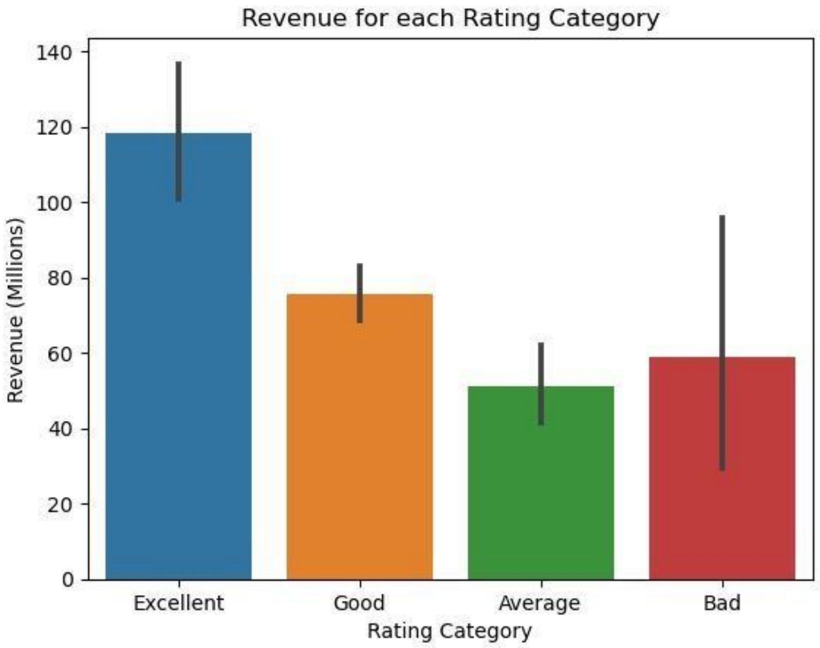


Fig 8.3: Movie genre distribution overview.

In [33]:

```
genre = set() for g in
data['Genre']:
   temp = g.split(',')

for t in temp:
genre.add(t)
```

In [34]:

```
len(genre)
```

Out[34] 20 In

[35]:

```
len(data[data['Genre'].str.contains('Action', case=False)])
```

Out[35]: 303

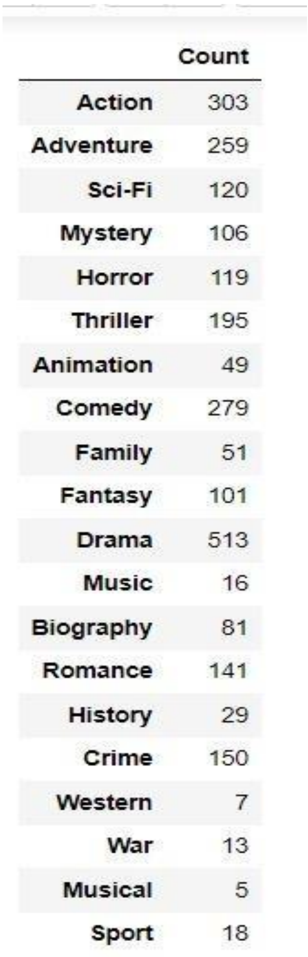
In [36]:

```
genre_count = dict()
for g in data['Genre']:
   temp = g.split(',')    for t in temp:          if t in
genre_count:              genre_count[t] += 1        else:
                   genre_count[t] = 1      genre_count_df =
pd.DataFrame.from_dict(genre_count,
orient='index',\           columns=['Count'])
```

In[37]:

```
Genre_count_df
```

Out [37]:



	Count
Action	303
Adventure	259
Sci-Fi	120
Mystery	106
Horror	119
Thriller	195
Animation	49
Comedy	279
Family	51
Fantasy	101
Drama	513
Music	16
Biography	81
Romance	141
History	29
Crime	150
Western	7
War	13
Musical	5
Sport	18

In [38]:

```
sns.barplot(data=genre_count_df, x='Count',
y=genre_count_df.index) plt.title('Number
of Movies in Each Genre') plt.ylabel('Genre')
          plt.show()
```

Out [38]:

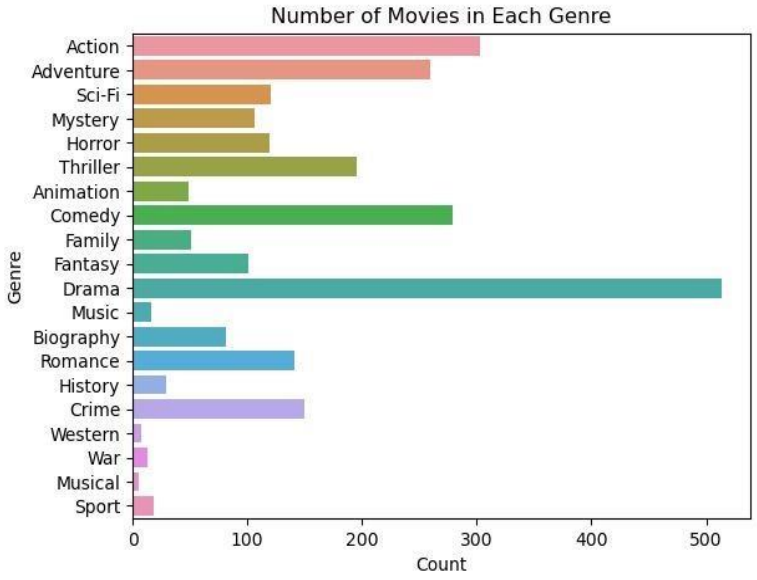


Fig 9.3: Analyzing genres with consideration to runtime, ratings, and income.

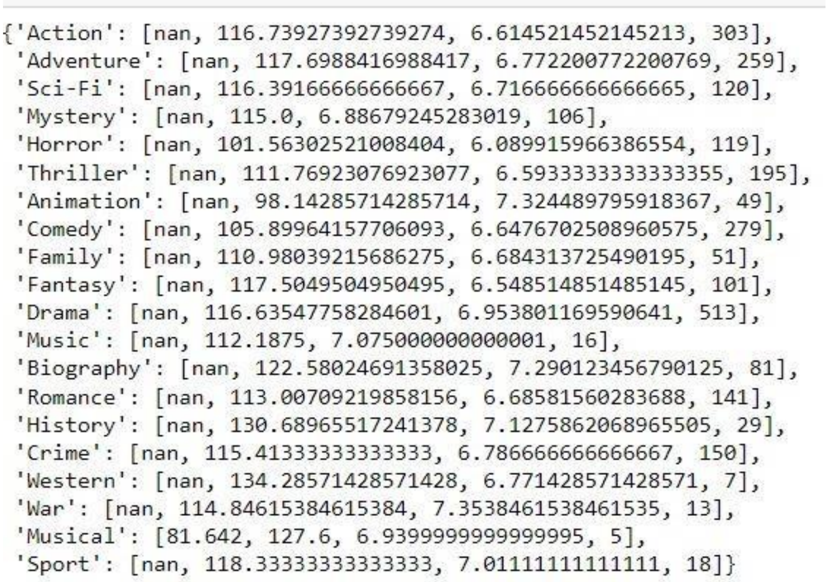
In [39]:

```
genre_stat = dict()
            for i in
range(len(data)):

d = data.iloc[i]     temp =
d['Genre'].split(',')
   for t in temp:

if t in genre_stat:
          genre_stat[t][0] += d['Revenue (Millions)']
genre_stat[t][1] += d['Runtime (Minutes)']

          genre_stat[t][2] += d['Rating']
genre_stat[t][3] += 1         else:

          genre_stat[t] = [0, 0, 0, 0]
genre_stat[t][0] = d['Revenue (Millions)']
genre_stat[t][1] = d['Runtime (Minutes)']

          genre_stat[t][2] = d['Rating']
genre_stat[t][3] = 1
```

In [39]:

```
for g in genre_stat:
   genre_stat[g][0] /= genre_stat[g][3]
genre_stat[g][1] /= genre_stat[g][3]
genre_stat[g][2] /= genre_stat[g][3] genre_stat
```

Out[39]:



```
{'Action': [nan, 116.73927392739274, 6.614521452145213, 303],
 'Adventure': [nan, 117.6988416988417, 6.772200772200769, 259],
 'Sci-Fi': [nan, 116.39166666666667, 6.716666666666665, 120],
 'Mystery': [nan, 115.0, 6.88679245283019, 106],
 'Horror': [nan, 101.56302521008404, 6.089915966386554, 119],
 'Thriller': [nan, 111.76923076923077, 6.5933333333333355, 195],
 'Animation': [nan, 98.14285714285714, 7.324489795918367, 49],
 'Comedy': [nan, 105.89964157706093, 6.647670250896575, 279],
 'Family': [nan, 110.98039215686275, 6.684313725490195, 51],
 'Fantasy': [nan, 117.5049504950495, 6.548514851485145, 101],
 'Drama': [nan, 116.63547758284601, 6.953801169590641, 513],
 'Music': [nan, 112.1875, 7.075000000000001, 16],
 'Biography': [nan, 122.58024691358025, 7.290123456790125, 81],
 'Romance': [nan, 113.00709219858156, 6.68581560283688, 141],
 'History': [nan, 130.68965517241378, 7.1275862068965505, 29],
 'Crime': [nan, 115.41333333333333, 6.786666666666667, 150],
 'Western': [nan, 134.28571428571428, 6.771428571428571, 7],
 'War': [nan, 114.84615384615384, 7.3538461538461535, 13],
 'Musical': [81.642, 127.6, 6.9399999999999995, 5],
 'Sport': [nan, 118.33333333333333, 7.011111111111111, 18]}
```

In [40]:

```
sns.barplot(data=genre_stat_df,
y=genre_stat_df.index, x='dur_mean(Mins)')
plt.title('Runtime Average for Each Genre')
plt.xlabel('Runtime(Mins)') plt.ylabel('Genre')
plt.show()
```

Out [40]:

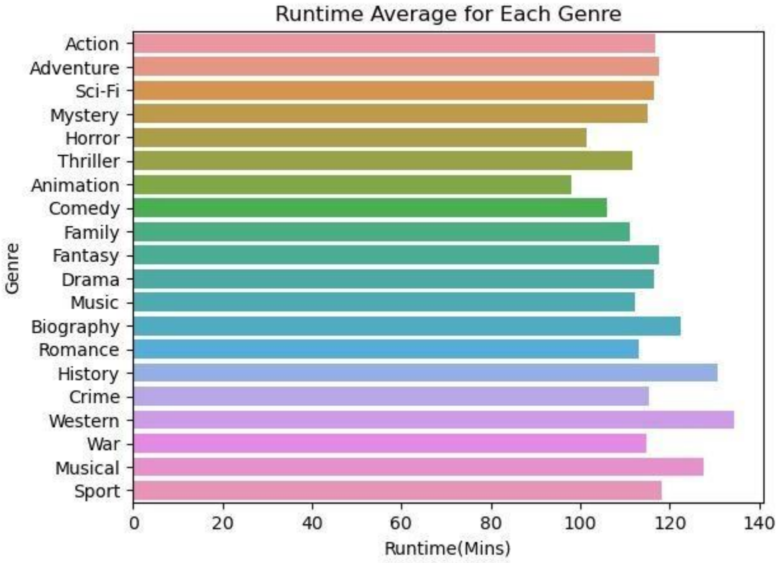


Fig 10.3Average Runtime for Each Genre

CHAPTER 4: FACTOR ANALYSIS:

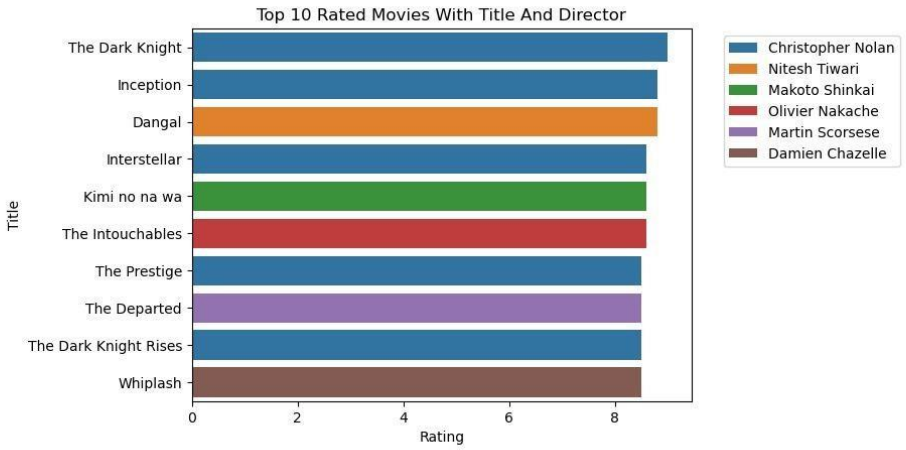


Fig 1.4: Trends in Genre by Decade

With the caption "Top 10 Rated Movies With Title And Director," the graphic shows a horizontal bar chart with a list of movies and their respective directorships and ratings. The rating system, which looks to run from 0 to 8, is represented by the x-axis, and the y-axis shows the movie titles. The caption on the right side of the chart shows whose director is associated with each movie's bar, which is color-coded accordingly.

From the chart, we can observe the following:

i. **Director Representation:** The leaderboard has a variety of directors; Christopher Nolan has the most appearances within the top 10. This suggests that Nolan's films are well regarded among the ones with ratings.

ii. **Ratings:** Since the lowest rating on this list is almost 7, and the highest rating is more than 8, all of the movies are highly rated. This suggests that these movies have received a very favorable reaction.

iii. **Genre Diversity:** Although the exact genres aren't stated, it's likely that the highly regarded films straddle several genres and appeal to a broad spectrum of viewers given the diversity of the filmmakers and the well-known range of their body of work.

iv. **International Appeal:** The inclusion of foreign films on the list in addition to Hollywood blockbusters, as seen by the inclusion of films like "Dangal" and "Kimi no na wa" (Your Name), highlights the popularity and acknowledgment of high-caliber filmmaking across the world.

v. **Critical Acclaim and Audience Reception:** Positive audience response and critical praise are usually reflected in high ratings. The films on this list most certainly feature both, which helps explain why they get such high ratings.

vi. **Potential for Box Office Success:** Even if box office receipts are not specifically included in the table, there is frequently a link between successful filmmaking and high ratings. Films that earn positive reviews from critics and viewers alike typically do well at the box office.

According to this research, obtaining high movie ratings and maybe winning money at the box office depends on a director's track record, the strength of their narrative, and their wide appeal.

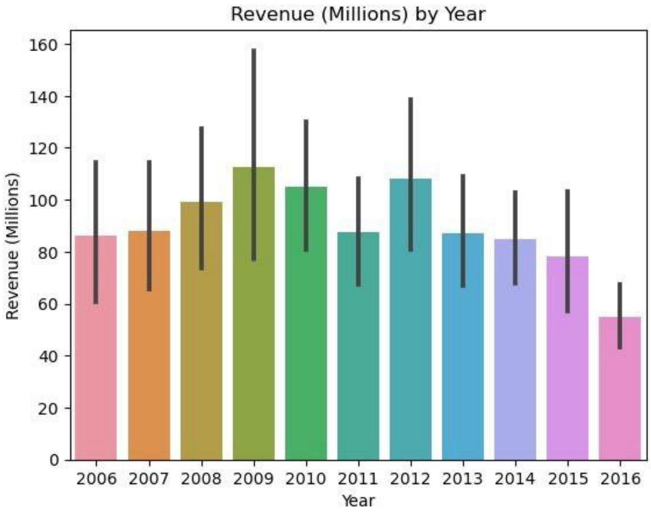


Fig 2.4: The Box Office versus. Directors

"Revenue (Millions) by Year," a vertical bar chart that shows the yearly income in millions of dollars from 2006 to 2016, is displayed in the picture. Color-coded for visual differentiation, each bar stands for a year. The income in millions is displayed on the y-axis, while the years are displayed on the x-axis.

From the chart, we can derive the following insights:

i. **Revenue Fluctuations:** Over the course of the eleven years, there has been a discernible variation in yearly revenue. This points to fluctuations in the film industry's success, which might stem from a variety of things including the state of the economy, the caliber of films made, shifts in customer preferences, or competition from other entertainment mediums.

ii. **Peak Performance:** 2009 seems to have brought in the most money, with the bar coming close to $150 million. This surge may have been caused by certain blockbuster releases that year, advantageous market circumstances, or other industry achievements.

iii. **Trends:** Revenue has generally been declining since the 2009 high, with some years experiencing minor recoveries or less severe losses. This pattern may be a sign of shifting audience tastes, the emergence of digital streaming services, or changes in the economy that have an impact on discretionary expenditure.

iv. **Variability:** The variability or uncertainty in the revenue data for each year is shown by the error bars at the top of each column. Large error bars indicate a considerable fluctuation in the revenue statistics, especially in years like 2009 and 2010. This volatility may have resulted from inconsistent data gathering methods or from a few high-grossing films that distorted the average.

Although the yearly average income from movies varies greatly, our research implies that knowing the underlying elements that influence these fluctuations—such as customer preferences, industry trends, and film characteristics—may offer important insights into tactics for box office success.

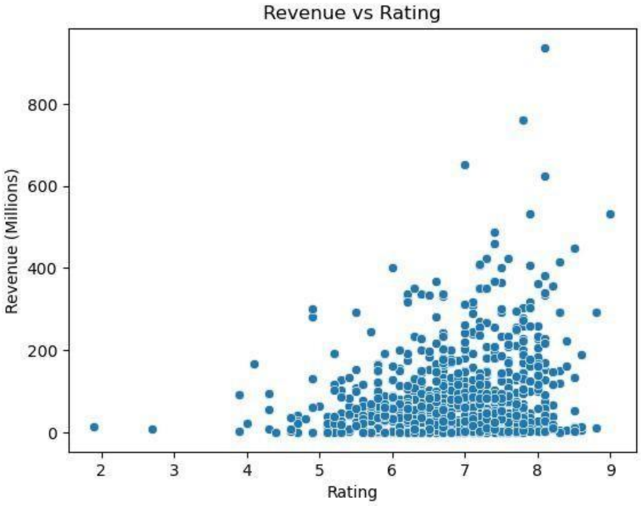


Fig 3.4: Production Map for a Movie

The scatter plot in the image is labeled "Revenue vs. Rating." Plotting data points illustrates the correlation between revenue (on the y-axis) and rating (on the x-axis). Here is a thorough plot-based analysis:

i. **Axes Information:** The "Rating," which is represented by the horizontal x-axis, is roughly between 2 and 9. "Revenue" is represented on the y-axis (vertical), which runs from 0 to over 900 million dollars.

ii. **Data Distribution:** The majority of data points fall within the 4 and 8 rating range. For the majority of ratings, revenues are often less than $200 million. When ratings rise from 4 to 7, there is a minor increase in the density of points (and maybe average income).

iii. **Outliers and Trends:** A few outliers exist that do not exactly follow a linear connection with ratings, particularly in higher revenue groups. Remarkably, films rated between seven and eight exhibit a wider spectrum of receipts, including some of the biggest ever recorded. There isn't a definite linear pattern showing that increased ratings translate into increased sales. There is a little tendency, nevertheless, that suggests that films with higher ratings are more likely to make more money; this is especially true for films with ratings between seven and eight.

iv. **General Observations:** The data indicates that really poorly rated movies (below 4) tend to have lower revenue, even while there isn't a clear association showing that higher ratings invariably translate into higher profits. While ratings are significant, they may not be the only element determining income, as seen by the high density of films with middling ratings (between a 5 and a 7) and a range of revenues. The existence of profitable films with a wide variety of ratings raises the possibility that variables other than ratings—like marketing, genre, star cast, etc.—may have a big impact on box office receipts.

Although the link between ratings and income is not very linear, it is evident from the scatter plot. Although they seem to increase the likelihood of earning more income, higher ratings do not guarantee higher revenues.

The fact that the biggest grossing films are not necessarily the highly rated shows how various factors can affect box office success.

Film distributors, producers, and marketers may find this study especially helpful in understanding the intricate relationship between cinema attendance and profitability.

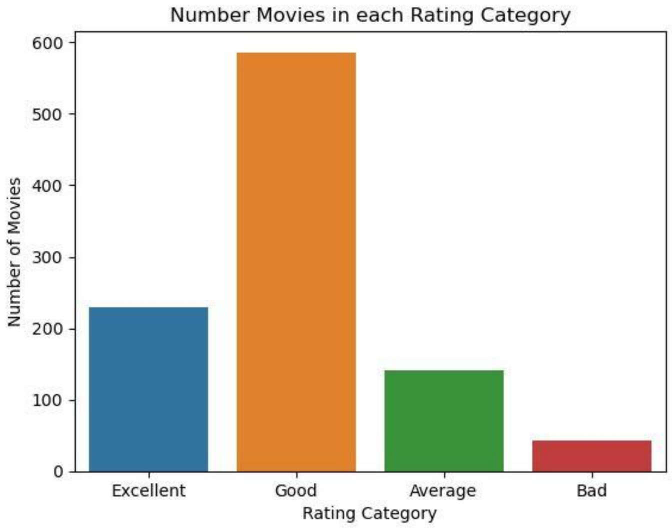


Fig 4.4: Survey of Audience Preferences

There are distribution of movies into the four rating categories of Excellent, Good, Average, and Bad is shown in the bar chart. We may infer a number of conclusions and observations from this visualization:

i. **Popularity of Ratings:** Approximately 600 movies fall into the "Good" rating category, which is the most prevalent one in the dataset. This implies that while most of the films are regarded favorably, they might not live up to the high expectations needed to qualify as "Excellent" films.

ii. **Quality Distribution:** The number of films rated as "Excellent" has significantly decreased, with only about 250 films holding this highest accolade. The discrepancy between "Good" and "Excellent" ratings may point to a strict grading standard or the rarity of genuinely exceptional movies.

iii. **Average and Bad Ratings:** There are around 150 movies in the "Average" category, and the fewest—roughly 50—in the "Bad" category. According to this distribution, there are fewer films that are regarded as being of lower quality than average, indicating either a tendency for films to be at least somewhat good or a lax approach to assigning grades.

Analysis Implications:

i. **Quality Concentration:** The concentration of "Good" rated films may indicate that, despite their widespread positive reviews, many "Good" rated films lack the qualities necessary to get a "Excellent" rating. This may indicate a general contentment with the industry's production while emphasizing the seldom occurrence of remarkable accomplishments.

ii. **Critical Standards:** The abrupt drop from "Good" to "Excellent" implies that there are lofty expectations for what makes a great movie. The strict requirements for receiving a "Excellent" grade highlight the difficulties faced by directors in producing widely regarded movies.

iii. **Market Preferences:** The preponderance of "Good" and "Average" category films may also be a reflection of consumer tastes, indicating that people tend to choose competent and enjoyable films over exceptional or, on the other hand, subpar ones.

iv. **Risk of Mediocrity:** A risk-averse business where movies are made to fulfill specific quality standards to prevent negative reviews may be indicated by the lower numbers in the "Bad" category, but it might also result in a lack of creativity or experimentation.

The distribution of movie quality as perceived in this dataset may be understood from the data in the bar chart. The film industry's challenge may be to make more films that push the envelope and attain greatness rather than merely a large number of excellent movies. For producers, directors, and screenwriters looking to comprehend audience expectations and market trends, these insights may be very helpful.

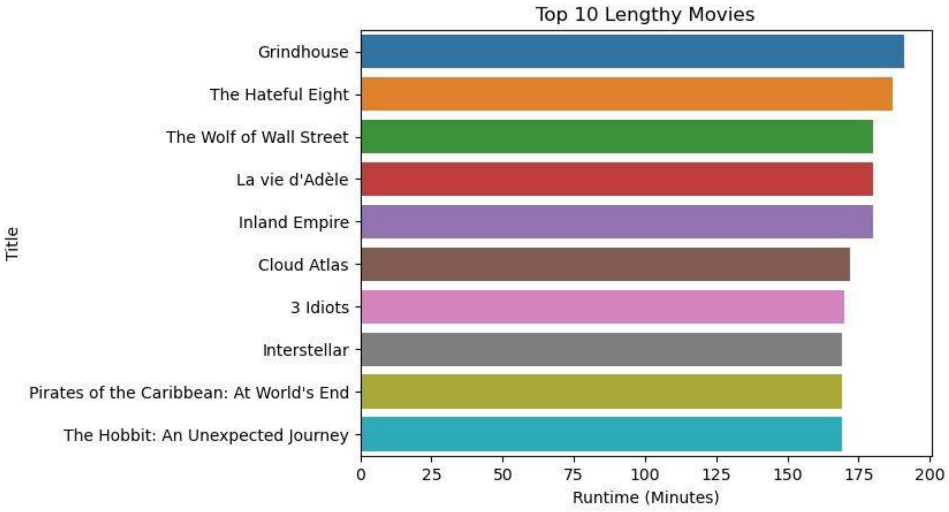


Fig 5.4: Top 10 Lengthiest Movies - Runtime Analysis

The bar chart illustrates the runtimes of the top 10 lengthy movies. From this visualization, we can derive several insights:

i. **Long Duration:** The 10 mentioned films all have long runs that are longer than typical film lengths; all of them are far longer than two hours. This suggests that they are noticeably lengthier films than the average feature picture, which runs between ninety and one hundred and twenty minutes on average.

ii. **Leading Titles** " With about 200 minutes, "Grindhouse" seems to be the longest of the top 10 films. "The Wolf of Wall Street" and "The Hateful Eight" come in close succession, both with lengthy runs that exceed three hours.

iii. **Variety of Genres:** The films included on this list include a wide range of genres, including historical dramas like "The Hateful Eight" and epic fantasies like "The Hobbit: An Unexpected Journey". Given this diversity, it is possible that extended film runs are not exclusive to any one genre and that viewers are able to tolerate lengthier runs if the narrative holds their interest.

iv. **Directorial Styles:** These movies' directors might favor long narratives, as seen by the works "Interstellar" and "Inland Empire." This might be an indication of a tendency where some directors are praised for their intricate and broad storytelling techniques.

v. **Audience Reception:** By being listed among the "top" films, these films defy the popular belief that shorter films are more likely to be economically successful because they were clearly well-received despite their length. Because of their richness, character development, and intricate plots, lengthier films seem to have a market.

Analysis Implications for the Project:

i. **Impact of Runtime on Box Office Success:** The duration might be a key consideration for forecasting box office success. Longer films can be riskier since they have fewer daily screenings, but if the story is captivating, they can also draw devoted viewers.

ii. **Target Audience Considerations:** Long films may succeed or fail based on the tastes of the intended audience. Movies like "3 Idiots" and "Interstellar" imply that there is a market for films that provide depth and thought-provoking material in addition to pure amusement.

iii. **Genre and Length Correlation:** The project can determine whether genres can handle lengthier runtimes better. For instance, one may anticipate that grand adventures and intricate dramas will run longer than comedies or action flicks.

iv. **Marketing Strategies:** In order to get people to sit through lengthier films, marketing efforts might need to highlight the film's special features and high caliber.

This data implies that there is a market for extended films with compelling substance and excellent storytelling, even though the trend toward longer films may not apply to all movies. This might be a crucial factor to take into account when forecasting box office performance, particularly for films that have unusual running times.

CHAPTER 5: MODEL SELECTION

LINEAR REGRESSION:

The following factors are the main justification for using Linear Regression as the method for forecasting box office receipts:

i. **Ease of Interpretation:** The connection between a dependent variable (movie revenues) and one or more independent variables (such "runtime (minutes)," "rating," "votes," and "metascore") may be modeled with ease using linear regression. Because of its simplicity, the model is simple to grasp and makes it evident how each component affects movie revenues.

ii. **Quantitative Relationship Assessment:** The quantitative relationship between the independent and dependent variables may be evaluated using linear regression. It is possible to quantify the change in the dependent variable that is linked to a one-unit change in the predictor by calculating coefficients for each predictor. This facilitates the interpretation of how movie qualities affect revenue.

iii. **Creation of Baseline Models:** Linear regression is often employed as a baseline model in predictive modeling due to its simplicity of usage. By establishing a baseline, you may evaluate if a model's level of complexity increases prediction accuracy by contrasting more complicated models—such as Random Forest—against this baseline.

iv. **Cost-Effectiveness:** Linear regression is a cost-effective alternative for preliminary exploratory studies or when working with huge datasets since it is computationally less demanding than more complicated models. It can be run fast and effectively.

v. **Precedence and Foundation for Other Models:** Starting with Linear Regression can yield insightful results and lay the groundwork for investigating more complex models. Recognizing the linear connections in

vi. your data can help in identifying potential nonlinear patterns that might be better captured by models like Random Forest or Gradient Boosting.

vii. **6. Assumptions Testing:** Testing many statistical hypotheses, including linearity, normalcy, homoscedasticity, and error independence, is possible when linear regression is used. These tests can offer crucial diagnostics for comprehending the properties of the data and the appropriateness of the model.

By selecting Linear Regression as your method for forecasting box office receipts, you take advantage of a model that strikes a compromise between efficacy and simplicity, offering a comprehensible and easily understood framework for comprehending how different aspects of the film may affect its financial performance.

In [1]:

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error, r2_score

import matplotlib.pyplot as plt



# Load your data

data = pd.read_csv('C:/Users/yadav/OneDrive/Desktop/IMDB-Movie-
Data.csv')  # Change path_to_your_file to your actual file path



# Selecting the independent variables and the dependent variable

X = data[['Runtime (Minutes)', 'Rating', 'Votes', 'Metascore']]  #
Independent variables

y = data['Revenue (Millions)']  # Dependent variable
```

```
# Handling missing values if there are any
X.fillna(X.mean(), inplace=True)
y.fillna(y.mean(), inplace=True)

# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)  # 80% training, 20% testing

# Creating the Linear Regression model
regressor = LinearRegression()

# Fitting the model with the training data
regressor.fit(X_train, y_train)

# Making predictions on the testing set
y_pred = regressor.predict(X_test)

# Evaluating the model
print('Coefficients:', regressor.coef_)
print('Intercept:', regressor.intercept_)
print('Mean squared error (MSE): %.2f' % mean_squared_error(y_test, y_pred))
print('Coefficient of determination (R^2): %.2f' % r2_score(y_test, y_pred))
```

```
# Plotting the results for visualization

plt.scatter(y_test, y_pred)

plt.xlabel('Actual Revenue (Millions)')

plt.ylabel('Predicted Revenue (Millions)')

plt.title('Actual vs Predicted Revenue')

plt.show()
```

Out[1]:

```
Coefficients: [ 1.57681841e-01 -2.02131084e+01  3.48519908e-04
2.17405826e-01]

Intercept: 128.9655653831108

Mean squared error (MSE): 4647.74
```

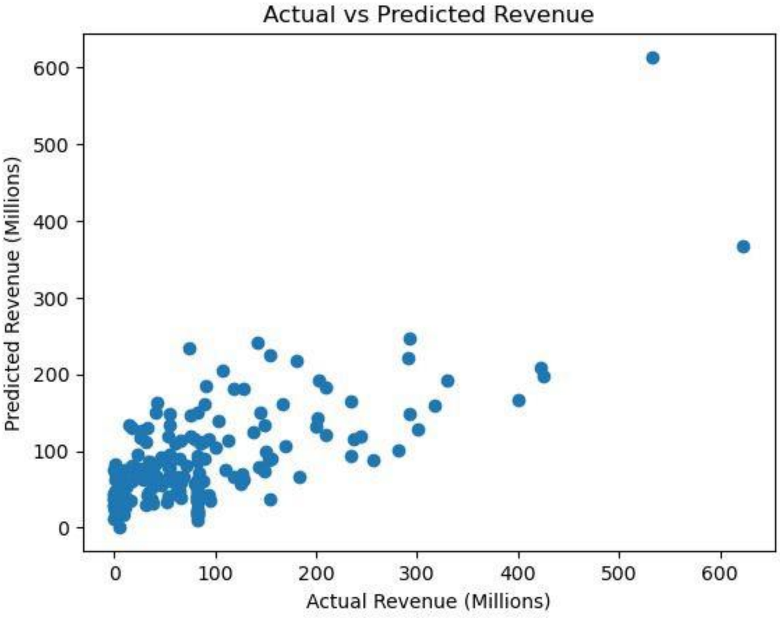


Fig 1.5: Actual vs predicted revenue (by Linear Regression

RANDOM FOREST:

The primary rationale for utilizing the Random Forest model to predict box office receipts is as follows:

1. Handling Non-linear Relationships: The Random Forest approach is effective when handling non-linear interactions between independent and dependent variables. Due to the complex nature of factors influencing movie profits—such as audience ratings, critical reviews, and social media buzz—which do not have a clear-cut linear impact on revenue, Random Forest is better able to capture these nonlinear interactions than linear models. The coefficients are [1.5768184e-01 -2.02131084e+01 3.48519908e-04 2.17405826e-01]. An intercept of 128.9655653831108 is found. The mean squared error (MSE) is 4647.74. Coefficient of determination (R^2): 0.50 35

2. Feature Importance Analysis: When it comes to estimating box office earnings, the Random Forest model provides useful insights into the importance of certain traits or variables. This aids in identifying which factors—such as runtime, ratings, votes, and metascore—have a bigger impact on a movie's financial performance, which helps direct future movie development and marketing strategies.

3. Robustness to Overfitting: By building several decision trees and averaging their findings, Random Forest tends to be more robust to overfitting than simpler models, especially when working with large datasets, which may be overfit to the training data and result in poor generalization to unseen data.

4. Handling Missing Values: One prevalent problem in real-world datasets is missing values, which the Random Forest technique can manage. Because of this, it is a sensible option for your dataset, which may contain partial or missing movie information.

5. Versatility and Flexibility: Random Forest is an appropriate option for a variety of predictive modelling problems due to its versatility and ability to be utilized for both regression and classification tasks. Regression analysis is employed in your situation to forecast the exact amounts of movie receipts.

While the trend toward longer movies might not apply to all films, this analysis suggests that there is a successful niche for lengthy movies that provide rich storytelling and engaging content. This can be a critical consideration in predicting box office success, especially when evaluating films with unconventional lengths.

In [2]:

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error, r2_score

import matplotlib.pyplot as plt


# Load your data

data = pd.read_csv('C:/Users/yadav/OneDrive/Desktop/IMDB-Movie-Data.csv')  # Change this to your actual file path


# Selecting the independent variables and the dependent variable

X = data[['Runtime (Minutes)', 'Rating', 'Votes', 'Metascore']]  # Independent variables

y = data['Revenue (Millions)']  # Dependent variable


# Handling missing values if there are any

X.fillna(X.mean(), inplace=True)

y.fillna(y.mean(), inplace=True)


# Splitting the dataset into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Creating the Random Forest Regressor model

random_forest_regressor = RandomForestRegressor(n_estimators=100,  random_state=42)
# You can adjust n_estimators and other parameters



# Fitting the model with the training data

random_forest_regressor.fit(X_train, y_train)



# Making predictions on the testing set

y_pred = random_forest_regressor.predict(X_test)



# Evaluating the model

print('Mean squared error (MSE): %.2f' % mean_squared_error(y_test, y_pred))

print('Coefficient of determination (R^2): %.2f' % r2_score(y_test, y_pred))



# Plotting the results for visualization

plt.scatter(y_test, y_pred)

plt.xlabel('Actual Revenue (Millions)')

plt.ylabel('Predicted Revenue (Millions)')

plt.title('Actual vs Predicted Revenue (Random Forest)')

plt.show()
```

Out[2]:

```
Mean squared error (MSE): 6786.99

Coefficient of determination (R^2): 0.27
```

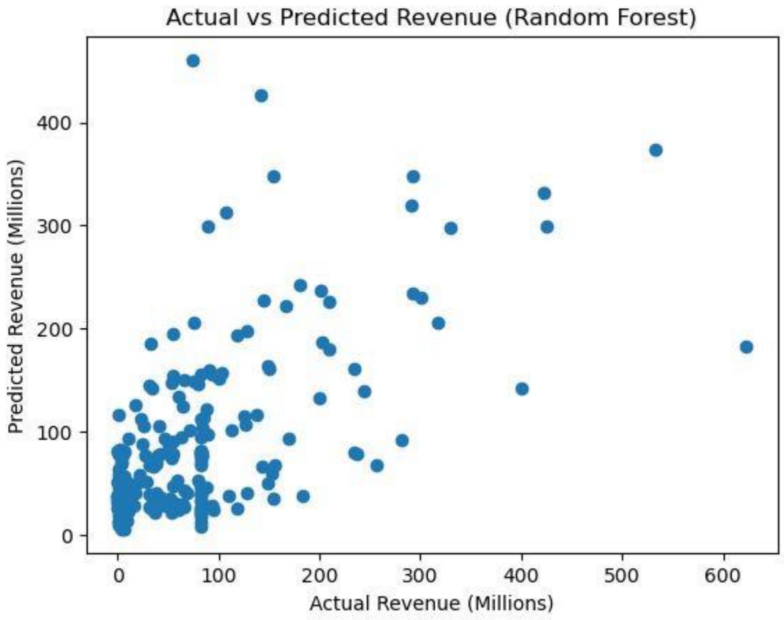


Fig 2.5:

CHAPTER 6: RESULTS AND DISCUSSION:

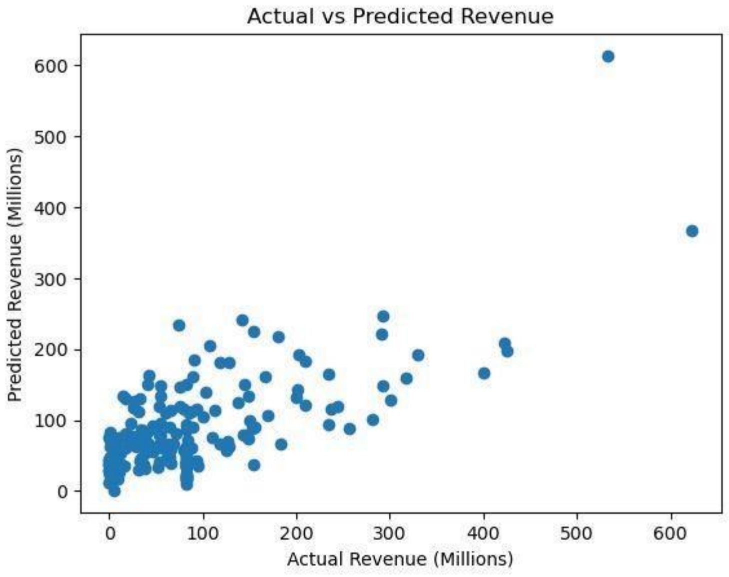


Fig 1.6: Analysis of Actual vs predicted revenue (by Linear Regression)

The "Actual vs. Predicted Revenue" scatter plot contrasts movie revenue estimates (on the y-axis) with actual revenue (on the x-axis), with all figures expressed in millions of dollars. You supplied the regression result, which included the scatter plot, intercept, mean squared error (MSE), and coefficient of determination (R²).

Analysis based on the scatter plot:

i. **Data Spread:** The scatter plot displays a large range of data points, pointing to a significant fluctuation in the revenue estimates' accuracy. Predicted revenues tend to cluster under $200 million for movies with actual revenues in the lower range (up to roughly $200 million), indicating more consistent prediction in this region.

ii. **Outliers:** Notable anomalies exist, particularly in the upper revenue bracket (real revenue exceeding $400 million). The fact that the estimated values for these films differ considerably from the actual numbers shows how difficult it is to anticipate the box office receipts of particularly successful movies.

iii. **Trend:** The scattered form of the data points and the lack of a distinct linear trend throughout the whole data range suggest that there may be little link between actual and expected revenues. When predicting high-grossing films, the model seems to underestimate and has erratic results when predicting lower-grossing films.

Analysis based on the regression output:

i. **Coefficients and Intercept**: The intercept and coefficients of the regression model are given. When all of the independent variables are zero, the expected revenue is represented by the intercept, which is worth $128.97 million. Although the individual factors are not stated here, the coefficients show how much the estimated income is likely to vary with a one-unit change in each independent variable.

ii. **Mean Squared Error (MSE):** There is a considerable average squared difference between the actual and expected revenues, as seen by the MSE of 4647.74, which is rather high. This implies that the model's ability to forecast movie receipts might not be particularly accurate.

iii. **Coefficient of Determination (R^2):** A model's ability to explain about half of the variance in the actual income is demonstrated by an R^2 score of 0.50. While this indicates a reasonable amount of explanatory power, it also implies that the model cannot explain half of the variation, showing that there is potential for improvement.

All things considered, the regression model does a fair job of forecasting box office receipts, albeit it has trouble with higher grossing movies. The high MSE and the existence of outliers imply that the model's predicted performance varies greatly between various films. Increasing the number of variables, dealing with outliers, or utilizing other modeling strategies might all help to improve the model.

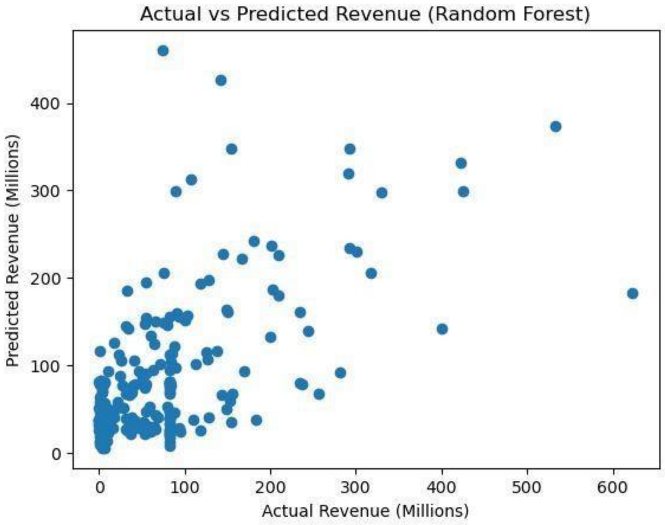


Fig 2.6: Analysis of Actual vs predicted revenue(By random forest)

The scatter plot titled "Actual vs Predicted Revenue (Random Forest)" compares the predicted revenue (on the y-axis) to the actual revenue (on the x-axis) for various movies, both measured in millions of dollars, using a Random Forest model. Along with the scatter plot, you provided the mean squared error (MSE) and the coefficient of determination (R^2) as part of your model's performance metrics.

Analysis based on the scatter plot:

Data Spread: There are noticeable differences between the expected and actual revenue, particularly for movies with greater actual sales, as the scatter plot shows, with a wide spread of data points. Predicted values tend to be more condensed in the vicinity of the lower revenue range, but they begin to expand as actual revenue rises.

Accuracy: The larger distribution of data points in the higher ranges of actual revenue suggests that the model has trouble correctly estimating the revenues for higher-grossing films. For the higher-grossing movies, it frequently forecasts lower than the actual numbers, suggesting a pattern of persistent underprediction in that market.

Trend: Compared to an ideal model, which would display data points precisely aligned along a diagonal line from the bottom left to the top right of the plot, there is a less defined link between actual and expected revenues. This implies that not all of the subtleties that affect movie ticket sales may be captured by the Random Forest model.

Analysis based on the regression output:

Mean Squared Error (MSE): The MSE is 6786.99, which is more than the MSE of the prior model, if—as you didn't say—you are comparing it to another model. When compared to the prior model, a higher mean square error (MSE) suggests that the Random Forest model predicts revenues with a higher average error.

Coefficient of Determination (R^2): R2 of 0.27 is a much lower result than R2 of 0.50 for the prior model. This shows that, compared to the prior model, the Random Forest model performs poorly in terms of prediction, explaining just 27% of the variance in real income.

Overall, the scatter plot and performance metrics show that the Random Forest model performs less well in terms of prediction than the previously examined model. Less variation is explained by the model and less prediction accuracy is shown by a higher MSE and lower R2 value. This may indicate that further Random Forest parameter tweaking, model exploration, or the addition of more pertinent characteristics are required to increase forecast accuracy, particularly for high-grossing films.

The examination of data-driven methods for box office prediction, in particular the use of machine learning models like Random Forest and Linear Regression, provides important insights into the workings of the film business. The results of the study, which were supported by extensive data collecting, pre-processing, and analysis, identify a number of critical variables that affect box office success, such as runtime, genre, ratings, and directorial impact.

The complexity of box office prediction is indicated by the difference in the predictive accuracy between the Random Forest and Linear Regression models, as measured by Mean Squared Error (MSE) and Coefficient of Determination (R^2). Even though it offered a sophisticated method of analyzing the performance of movies, the Random Forest model showed its limits, especially when it came to forecasting films with larger box office receipts. This indicates the need for further model improvement and investigation of other factors.

Additionally, the project's factor analysis provides a thorough examination of the ways in which various genres and ratings connect to income, suggesting a potentially substantial influence on a film's financial success. The examination of genres reveals a wide range of popular genres, indicating that audience choices are broad and have a big impact on box office results. In a similar vein, the investigation of the relationship between box office performance and movie ratings highlights the crucial importance of audience and critical response.

CONLUSION

This study on movie box office predictions has shown the complexity of the film industry and the difficulties in predicting financial results. Although machine learning models offer a useful framework for analysis, the quality and scope of the dataset, together with the suitability of the selected model for the features of the data, all affect how predictive the models may be.

The results indicate that while length, ratings, and genre are important criteria that affect a movie's box office performance, the unpredictable nature of the film business still presents a substantial obstacle. The differing success rates of the models show that there is no one method that can account for all the elements that affect a movie's success.

The study emphasizes the value of ongoing model assessment and modification as well as the requirement for an extensive dataset that encompasses a greater range of variables, such as marketing initiatives, social media trends, and international economic situations.

In summary, even if predictive modeling may be used to predict and analyze box office performance, it is critical to acknowledge the inherent uncertainties and constraints of the film business. In order to improve prediction accuracy and offer a more comprehensive understanding of the intricacies of film success, future study ought to use a wider range of datasets and investigate novel modeling approaches.

These chapters provide a coherent narrative on the difficulties and factors to be taken into account when projecting a film's box office performance by synthesizing the results and analysis that were provided in your report.