A DATA-DRIVEN ANALYSIS OF
IMDb RATINGS

Prepared in the partial fulfillment of the Summer Internship Program on Data Analysis

UNDER THE GUIDANCE OF

K.Narmada Mani, APSSDC

K. Meenakshi, APSSDC

SUBMITTED BY

Md. Nusrath Shariff	20551A4231
U. Dharani	20551A4252
K. Mahesh	20551A4229
Aisha Maryam	21551A4202

GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY, AP

AUGUST 2023

ACKNOLEDGEMENT

I would like to express my heartfelt gratitude to all those who have contributed to the successful completion of my summer internship project at **Andhra Pradesh Skill Development Corporation (APSSDC)**. This opportunity has been an enriching and transformative experience for me, and I am truly thankful for the support, guidance, and encouragement I have received along the way.

First and foremost, I extend my sincere regards to Mrs. Narmada Mani, Mrs. K. Meenakshi, my supervisors and mentors, for providing me with valuable insights, constant guidance, and unwavering support throughout the duration of the internship. Their expertise and encouragement have been instrumental in shaping the direction of this project.

I would like to thank the entire team at **Andhra Pradesh Skill Development Corporation (APSSDC)** for fostering a collaborative and innovative environment. The camaraderie, knowledge sharing, and feedback I received from my colleagues significantly contributed to the development and success of this project.

In conclusion, I am honored to have been a part of this internship program, and I look forward to leveraging the skills and knowledge gained to contribute positively to future endeavors.

Thank you.

Sincerely,

Mohammad Nusrath Shariff, 20551A4231
Uppala Dharani, 20551A4252
Kondumahanthi Mahesh, 20551A4229
Aisha Maryam, 21551A4202

TABLE OF CONTENTS

1. ABSTRACT

A DATA-DRIVEN ANALYSIS OF
IMDB RATINGS

In the dynamic landscape of the film industry, understanding the intricate dynamics that govern the interplay between critical acclaim, audience sentiment, and box office success is of paramount importance. This project undertakes a comprehensive exploration of these dimensions by analyzing the relationship between IMDb ratings and movie gross.

Through the lens of rigorous data analysis, this study unveils nuanced insights that challenge conventional assumptions. Contrary to prevailing notions, our findings reveal that while IMDb ratings hold significance, they alone do not exert a direct impact on box office performance. Instead, our analysis highlights the paramount significance of content in captivating audiences. Genres that resonate, narratives that intrigue, and performances that captivate appear to wield a more potent influence on cinematic triumph.

From data collection to statistical analyses, this project navigates the data analysis journey with the aid of pivotal libraries including Numpy, Pandas, Matplotlib.pyplot, and Seaborn. These tools empower us to efficiently process data, explore intricate relationships, and visualize patterns that shape our insights.

In conclusion, this project offers a comprehensive perspective on the relationship between IMDb ratings and box office earnings. By harnessing the power of data analysis, this exploration contributes to the evolving narrative of the film industry, offering valuable insights that guide decision-making, enhance content creation, and honor the enduring allure of the silver screen.

2. INTRODUCTION

OBJECTIVE:

- The objective is to undertake an expedition of discovery, driven by the principles of data analysis, and guided by the wealth of information bestowed by the dataset. As we traverse the data landscape, a central inquiry materializes: **Can the IMDb ratings, reflective of audience sentiment and endorsement, wield influence over a movie's box office earnings?**

SCOPE:

- This research primarily focuses on movies featured in the dataset and their corresponding IMDb ratings and box office gross. While the dataset offers a rich tapestry of cinematic details, we limit our analysis to the provided variables, namely the movie's title, release year, certificate, runtime, genre, IMDb rating, overview, director, cast, and gross. Additionally, we exclude external factors, such as marketing efforts or distribution strategies, to retain a clear focus on the connection between IMDb ratings and box office earnings. This study spans a diverse range of movies, providing a compelling overview of cinematic trends over the years while delving into the potential impact of audience sentiment on the commercial trajectory of films.

DATASET COMPONENTS:

- An evocative title that encapsulates a story's essence.
- The temporal context provided by the release year.
- The guiding certificate that determines its audience.
- The temporal rhythm dictated by the runtime.
- The genre that sets anticipations.
- The IMDb rating serving as a barometer of viewer appreciation.
- The concise overview that provides a glimpse into the narrative.
- The visionary directors who breathe life into the screen, the ensemble cast that embodies characters.
- The quantifiable gross that signifies its commercial trajectory.

AIM:

- Through diligent statistical analysis, visualization techniques, and careful exploration, our aim is to unearth revelations that illuminate the intricate interplay between critical reception and commercial triumph.

3. SYSTEM REQUIREMENTS

OPERATING SYSTEM	Microsoft Windows 11 Home Single Language
CPU	Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz, 2496 Mhz, 4 Core(s), 8 Logical Processor(s)
RAM	16GB
HARD DRIVE	250GB SSD
SOFTWARE	Jupyter Notebook, Microsoft Excel
PROGRAMMING LANGUAGE	Python

ADDITIONAL REQUIREMNETS: External Keyboard, External Mouse

4. ARCHITECTURE OF PROJECT

- **Data Collection:**
  - Obtain the dataset containing movie details, including IMDb ratings, box office gross, genres, release year, directors, cast, and other relevant information from online sources or databases.

- **Data Preprocessing:**
  - Clean and preprocess the dataset to handle missing values, duplicates, and any inconsistencies.
  - Convert data into appropriate formats for analysis, such as transforming dates, numerical values, and categorical variables.

- **Exploratory Data Analysis (EDA):**
  - Conduct a comprehensive exploration of the dataset to gain insights into its distribution, summary statistics, and characteristics.
  - Create visualizations to identify trends, correlations, and patterns between IMDb ratings, gross, and other relevant variables.
  - Explore the impact of genres, release years, and other factors on IMDb ratings and box office performance.

- **Data Visualization:**
  - Generate visual representations, such as scatter plots, bar charts, and heatmaps, to present key findings and trends effectively.
  - Create interactive dashboards for a user-friendly way of exploring and understanding the data.

- **Interpretation and Conclusion:**
  - Summarize the findings from the analysis, highlighting the significance of IMDb ratings in relation to box office success.
  - Draw conclusions on the impact of audience perception on movie earnings and discuss any implications for the film industry.

5. LIBRARIES

In the data analysis project aimed at exploring the relationship between IMDb ratings and movie gross, four powerful Python libraries, namely Numpy, Pandas, Matplotlib.pyplot, and Seaborn, serve as the backbone for processing, visualizing, and drawing insights from the dataset. As the project unfolds, these libraries work in unison to process the dataset, gain a comprehensive understanding of the data's characteristics, and provide meaningful insights into the dynamics governing cinematic achievements. By importing these libraries at the project's outset, the data analysis process is streamlined and poised for seamless manipulation and visualization, ultimately allowing for an in-depth exploration of the intriguing relationship between IMDb ratings and movie gross.

Numpy	•It is a fundamental library for numerical computing in Python. •Numpy is used to efficiently process and manipulate numeric data, laying the groundwork for further analysis.
Pandas	•Pandas is a versatile data manipulation library in Python. •Pandas is essential for loading, cleaning, and transforming the movie dataset into a structured format, making it a vital tool for exploratory data analysis.
Matplotlib.pyplot	•Matplotlib.pyplot is a widely used library for data visualization in Python. •Matplotlib.pyplot is employed to visualize the relationship between IMDb ratings and movie earnings. The visualizations help identify patterns and trends in the data, aiding in the interpretation of results.
Seaborn	•Seaborn is a statistical data visualization library built on top of Matplotlib. •Seaborn complements Matplotlib.pyplot by providing additional tools to explore relationships between variables and reveal insights into the dataset, by simplifing the process of creating complex statistical plots, such as box plots, violin plots, and heatmaps.

By leveraging these four libraries together, the IMDb ratings and movie gross analysis project benefits from efficient data handling, comprehensive data exploration, and compelling visualizations, ultimately facilitating a deeper understanding of the interplay between audience sentiments and movie financial success.

6. IMPLEMENTATION

IMPORTING LIBRARIES:

import numpy as np
import pandas as pd
import matplotlib.pyplot as mp
import seaborn as sb

➢ LOADING DATASET INTO DATAFRAME:

imdb = pd.read_csv("imdb.csv") ***#Dataset file Name imdb.csv***
imdb

PRINTING HEAD:

imdb.head() ***#Printing first 10 rows***

Output:

```
                                        Poster_Link  \
0  https://m.media-amazon.com/images/M/MV5BMDFkYT...
1  https://m.media-amazon.com/images/M/MV5BM2MyNj...
2  https://m.media-amazon.com/images/M/MV5BMTMxNT...
3  https://m.media-amazon.com/images/M/MV5BMWMwMG...
4  https://m.media-amazon.com/images/M/MV5BMWU4N2...

                Series_Title Released_Year Certificate  Runtime  \
0  The Shawshank Redemption          1994           A  142 min
1             The Godfather          1972           A  175 min
2           The Dark Knight          2008          UA  152 min
3    The Godfather: Part II          1974           A  202 min
4              12 Angry Men          1957           U   96 min

                   Genre  IMDB_Rating  \
0                  Drama          9.3
1           Crime, Drama          9.2
2  Action, Crime, Drama          9.0
3           Crime, Drama          9.0
4           Crime, Drama          9.0

                                            Overview  Meta_score  \
0  Two imprisoned men bond over a number of years...        80.0
1  An organized crime dynasty's aging patriarch t...       100.0
2  When the menace known as the Joker wreaks havo...        84.0
3  The early life and career of Vito Corleone in ...        90.0
4  A jury holdout attempts to prevent a miscarria...        96.0

                Director            Star1            Star2           Star3
\
0        Frank Darabont     Tim Robbins  Morgan Freeman     Bob Gunton

1  Francis Ford Coppola   Marlon Brando       Al Pacino     James Caan

2     Christopher Nolan  Christian Bale    Heath Ledger  Aaron Eckhart

3  Francis Ford Coppola       Al Pacino  Robert De Niro  Robert Duvall

4          Sidney Lumet     Henry Fonda     Lee J. Cobb  Martin Balsam


             Star4  No_of_Votes        Gross
0  William Sadler      2343110   28,341,469
1    Diane Keaton      1620367  134,966,411
2   Michael Caine      2303232  534,858,444
3    Diane Keaton      1129952   57,300,000
4    John Fiedler       689845    4,360,000
```

PRINTING TAIL:

imdb.head() ***#Printing last 10 rows***

Output:

```
                                            Poster_Link  \
995  https://m.media-amazon.com/images/M/MV5BNGEwMT...
996  https://m.media-amazon.com/images/M/MV5BODk3Yj...
997  https://m.media-amazon.com/images/M/MV5BM2U3Yz...
998  https://m.media-amazon.com/images/M/MV5BZTBmMj...
999  https://m.media-amazon.com/images/M/MV5BMTY5OD...

              Series_Title Released_Year Certificate  Runtime  \
995  Breakfast at Tiffany's          1961           A  115 min
996                  Giant          1956           G  201 min
997   From Here to Eternity          1953      Passed  118 min
998                Lifeboat          1944         NaN   97 min
999            The 39 Steps          1935         NaN   86 min

                        Genre  IMDB_Rating  \
995    Comedy, Drama, Romance          7.6
996            Drama, Western          7.6
997      Drama, Romance, War           7.6
998                Drama, War          7.6
999  Crime, Mystery, Thriller          7.6

                                              Overview  Meta_score  \
995  A young New York socialite becomes interested ...        76.0
996  Sprawling epic covering the life of a Texas ca...        84.0
997  In Hawaii in 1941, a private is cruelly punish...        85.0
998  Several survivors of a torpedoed merchant ship...        78.0
999  A man in London tries to help a counter-espion...        93.0

             Director               Star1               Star2
Star3  \
995     Blake Edwards     Audrey Hepburn     George Peppard   Patricia
Neal
996    George Stevens   Elizabeth Taylor        Rock Hudson       James
Dean
997    Fred Zinnemann     Burt Lancaster   Montgomery Clift    Deborah
Kerr
998  Alfred Hitchcock  Tallulah Bankhead        John Hodiak   Walter
Slezak
999  Alfred Hitchcock       Robert Donat  Madeleine Carroll  Lucie
Mannheim

               Star4  No_of_Votes        Gross
995     Buddy Ebsen        166544          NaN
996   Carroll Baker         34075          NaN
997      Donna Reed         43374   30,500,000
998  William Bendix         26471          NaN
999  Godfrey Tearle         51853          NaN
```

➢ CLEANING AND LOADING THE DATA:

imdb.info()

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Poster_Link    1000 non-null   object
 1   Series_Title   1000 non-null   object
 2   Released_Year  1000 non-null   object
 3   Certificate    899 non-null    object
 4   Runtime        1000 non-null   object
 5   Genre          1000 non-null   object
 6   IMDB_Rating    1000 non-null   float64
 7   Overview       1000 non-null   object
 8   Meta_score     843 non-null    float64
 9   Director       1000 non-null   object
 10  Star1          1000 non-null   object
 11  Star2          1000 non-null   object
 12  Star3          1000 non-null   object
 13  Star4          1000 non-null   object
 14  No_of_Votes    1000 non-null   int64
 15  Gross          831 non-null    object
dtypes: float64(2), int64(1), object(13)
memory usage: 125.1+ KB
```

```
imdb.shape
```

```
(1000, 16)
```

#The dataset contains total of 1000 rows and 16 columns.

imdb.columns ***#Printing columns***

Output:

```
Index(['Poster_Link', 'Series_Title', 'Released_Year', 'Certificate',
       'Runtime', 'Genre', 'IMDB_Rating', 'Overview', 'Meta_score',
'Director',
       'Star1', 'Star2', 'Star3', 'Star4', 'No_of_Votes', 'Gross'],
      dtype='object')
```

WORKING ON NULL VALUES AND DROPPING THEM:

Imdb.isnull().sum()

Output:

```
Poster_Link         0
Series_Title        0
Released_Year       0
Certificate       101
Runtime             0
Genre               0
IMDB_Rating         0
Overview            0
Meta_score        157
Director            0
Star1               0
Star2               0
Star3               0
Star4               0
No_of_Votes         0
Gross             169
dtype: int64
```

#There are toral 101 null values in "Certificate" column, 157 null values in "Meta_Score" and 169 in "Gross"

imdb[imdb['Gross'].isnull()].head() ***#null value in "gross" column of imdb dataset***

Output:

```
                                          Poster_Link
Series_Title  \
18  https://m.media-amazon.com/images/M/MV5BNjViNW...
Hamilton
20  https://m.media-amazon.com/images/M/MV5BOTc2ZT...          Soorarai
Pottru
30  https://m.media-amazon.com/images/M/MV5BYjBmYT...
Seppuku
32  https://m.media-amazon.com/images/M/MV5BZjc4ND...   It's a
Wonderful Life
46  https://m.media-amazon.com/images/M/MV5BZmY2Nj...           Hotaru
no haka

   Released_Year Certificate  Runtime                          Genre
IMDB_Rating  \
18          2020        PG-13   160 min   Biography, Drama, History
8.6
20          2020            U   153 min                      Drama
8.6
30          1962          NaN   133 min      Action, Drama, Mystery
8.6
32          1946           PG   130 min      Drama, Family, Fantasy
8.6
46          1988            U    89 min       Animation, Drama, War
8.5

                                             Overview  Meta_score  \
18  The real life of one of America's foremost fou...        90.0
20  Nedumaaran Rajangam "Maara" sets out to make t...         NaN
30  When a ronin requesting seppuku at a feudal lo...        85.0
32  An angel is sent from Heaven to help a despera...        89.0
46  A young boy and his little sister struggle to ...        94.0

            Director                 Star1             Star2
Star3  \
18      Thomas Kail   Lin-Manuel Miranda      Phillipa Soo    Leslie
Odom Jr.
20     Sudha Kongara                Suriya          Madhavan       Paresh
Rawal
30  Masaki Kobayashi      Tatsuya Nakadai    Akira Ishihama     Shima
Iwashita
32       Frank Capra        James Stewart        Donna Reed   Lionel
Barrymore
46     Isao Takahata      Tsutomu Tatsumi   Ayano Shiraishi    Akemi
Yamaguchi

                      Star4  No_of_Votes Gross
18  Renée Elise Goldsberry         55291   NaN
20        Aparna Balamurali         54995   NaN
30           Tetsurô Tanba         42004   NaN
32          Thomas Mitchell        405801   NaN
46        Yoshiko Shinohara        235231   NaN
```

imdb[imdb['Meta _ score’].isnull()].head() ***#null value in "meta score” cloumn of imdb dataset***

Output:

```
                                          Poster_Link  \
20  https://m.media-amazon.com/images/M/MV5BOTc2ZT...
54  https://m.media-amazon.com/images/M/MV5BNWJhMD...
55  https://m.media-amazon.com/images/M/MV5BY2FiMT...
57  https://m.media-amazon.com/images/M/MV5BMTQ4Mz...
65  https://m.media-amazon.com/images/M/MV5BMDhjZW...

                Series_Title Released_Year Certificate  Runtime  \
20           Soorarai Pottru          2020           U  153 min
54  Ayla: The Daughter of War          2017         NaN  125 min
55               Vikram Vedha          2017          UA  147 min
57                     Dangal          2016           U  161 min
65          Taare Zameen Par           2007           U  165 min

                        Genre  IMDB_Rating  \
20                      Drama          8.6
54  Biography, Drama, History          8.4
55       Action, Crime, Drama          8.4
57   Action, Biography, Drama          8.4
65             Drama, Family           8.4

                                            Overview  Meta_score  \
20  Nedumaaran Rajangam "Maara" sets out to make t...         NaN
54  In 1950, amid-st the ravages of the Korean War...         NaN
55  Vikram, a no-nonsense police officer, accompan...         NaN
57  Former wrestler Mahavir Singh Phogat and his t...         NaN
65  An eight-year-old boy is thought to be a lazy ...         NaN

         Director         Star1             Star2               Star3  \
20  Sudha Kongara        Suriya           Madhavan        Paresh Rawal
54     Can Ulkay     Erdem Can    Çetin Tekindor     Ismail Hacioglu
55       Gayatri       Pushkar           Madhavan    Vijay Sethupathi
57  Nitesh Tiwari    Aamir Khan     Sakshi Tanwar  Fatima Sana Shaikh
65     Aamir Khan   Amole Gupte   Darsheel Safary           Aamir Khan

                Star4  No_of_Votes        Gross
20  Aparna Balamurali        54995          NaN
54      Kyung-jin Lee        34112          NaN
55   Shraddha Srinath        28401          NaN
57     Sanya Malhotra       156479   12,391,761
65       Tisca Chopra       168895    1,223,869
```

imdb[imdb['Certificate'].isnull()].head ( ) ***#null value in "certificate" cloumn of imdb dataset***

Output:

```
                                          Poster_Link  \
30   https://m.media-amazon.com/images/M/MV5BYjBmYT...
54   https://m.media-amazon.com/images/M/MV5BNWJhMD...
77   https://m.media-amazon.com/images/M/MV5BOTI4NT...
92   https://m.media-amazon.com/images/M/MV5BNjAzMz...
121  https://m.media-amazon.com/images/M/MV5BZmM0NG...

                  Series_Title Released_Year Certificate  Runtime  \
30                     Seppuku          1962         NaN  133 min
54   Ayla: The Daughter of War          2017         NaN  125 min
77            Tengoku to jigoku          1963         NaN  143 min
92               Babam ve Oglum          2005         NaN  112 min
121                      Ikiru          1952         NaN  143 min

                         Genre  IMDB_Rating  \
30      Action, Drama, Mystery          8.6
54   Biography, Drama, History          8.4
77       Crime, Drama, Mystery          8.4
92               Drama, Family          8.3
121                      Drama          8.3

                                           Overview  Meta_score  \
30   When a ronin requesting seppuku at a feudal lo...        85.0
54   In 1950, amid-st the ravages of the Korean War...         NaN
77   An executive of a shoe company becomes a victi...         NaN
92   The family of a left-wing journalist is torn a...         NaN
121  A bureaucrat tries to find a meaning in his li...         NaN

             Director              Star1            Star2
Star3  \
30   Masaki Kobayashi   Tatsuya Nakadai   Akira Ishihama     Shima
Iwashita
54          Can Ulkay          Erdem Can   Çetin Tekindor    Ismail
Hacioglu
77      Akira Kurosawa    Toshirô Mifune      Yutaka Sada    Tatsuya
Nakadai
92         Çagan Irmak    Çetin Tekindor    Fikret Kuskan
Hümeyra
121     Akira Kurosawa   Takashi Shimura     Nobuo Kaneko   Shin'ichi
Himori

             Star4  No_of_Votes    Gross
30   Tetsurô Tanba         42004      NaN
54   Kyung-jin Lee         34112      NaN
77    Kyôko Kagawa         34357      NaN
92      Ege Tanman         78925      NaN
121   Haruo Tanaka         68463   55,240
```

#Here "Gross", "Meta_Score" and "Certificate" are important columns hence, we can't fill the null values in those columns. So dropping the null values is a reasonable approach here.

imdb.dropna(axis = 0, inplace = True)

imdb

imdb.shape

Output:

```
(714, 16)
```

Imdb.isnull().sum()

Output:

```
Poster_Link      0
Series_Title     0
Released_Year    0
Certificate      0
Runtime          0
Genre            0
IMDB_Rating      0
Overview         0
Meta_score       0
Director         0
Star1            0
Star2            0
Star3            0
Star4            0
No_of_Votes      0
Gross            0
dtype: int64
```

#Rows are decreased to 714 by removing null values.

CHECKING FOR DUPLICATES:

Imdb.duplicated().any()

Output:

```
False
```

Imdb.describe() ***#to show all mathematical analysis on numerical columns on dataset.***

```
       IMDB_Rating  Meta_score   No_of_Votes
count   714.000000  714.000000  7.140000e+02
mean      7.937115   77.158263  3.561348e+05
std       0.293278   12.401144  3.539011e+05
min       7.600000   28.000000  2.522900e+04
25%       7.700000   70.000000  9.600975e+04
50%       7.900000   78.000000  2.366025e+05
75%       8.100000   86.000000  5.077922e+05
max       9.300000  100.000000  2.343110e+06
```

CHANGING COLUMN NAMES:

Imdb.rename(columns={'Series_Title' : "Movie Name" , 'Meta score' : 'Audience _ rating ' }, inplace=True)

Imdb

Output:

```
                    Movie_Name Released_Year Certificate  Runtime  \
0    The Shawshank Redemption          1994           A  142 min
1                The Godfather          1972           A  175 min
2              The Dark Knight          2008          UA  152 min
3      The Godfather: Part II          1974           A  202 min
4                  12 Angry Men          1957           U   96 min

..                          ...           ...         ...     ...
990                Giù la testa          1971          PG  157 min
991             Kelly's Heroes           1970          GP  144 min
992          The Jungle Book             1967           U   78 min
994        A Hard Day's Night            1964           U   87 min
997     From Here to Eternity            1953      Passed  118 min

                            Genre  IMDB_Rating  \
0                           Drama          9.3
1                   Crime, Drama           9.2
2             Action, Crime, Drama          9.0
3                   Crime, Drama           9.0
4                   Crime, Drama           9.0
..                           ...           ...
990           Drama, War, Western           7.6
991         Adventure, Comedy, War          7.6
992  Animation, Adventure, Family           7.6
994         Comedy, Music, Musical          7.6
997            Drama, Romance, War          7.6

                                                Overview
Audience_rating  \
0    Two imprisoned men bond over a number of years...
80.0
1    An organized crime dynasty's aging patriarch t...
100.0
2    When the menace known as the Joker wreaks havo...
84.0
3    The early life and career of Vito Corleone in ...
90.0
4    A jury holdout attempts to prevent a miscarria...
96.0
..                                                 ...                 ..
.
990  A low-life bandit and an I.R.A. explosives exp...
77.0
991  A group of U.S. soldiers sneaks across enemy l...
50.0
992  Bagheera the Panther and Baloo the Bear have a...
65.0
994  Over two "typical" days in the life of The Bea...
96.0
997  In Hawaii in 1941, a private is cruelly punish...
85.0

                  Director            Star1              Star2
Star3  \
0          Frank Darabont     Tim Robbins     Morgan Freeman        Bob
Gunton
1    Francis Ford Coppola    Marlon Brando          Al Pacino
James Caan
2       Christopher Nolan   Christian Bale       Heath Ledger     Aaron
Eckhart
3    Francis Ford Coppola        Al Pacino     Robert De Niro     Robert
Duvall
4            Sidney Lumet      Henry Fonda        Lee J. Cobb     Martin
Balsam
..                    ...              ...                ...
...
990          Sergio Leone      Rod Steiger       James Coburn      Romolo
Valli
991        Brian G. Hutton   Clint Eastwood      Telly Savalas       Don
Rickles
992   Wolfgang Reitherman      Phil Harris    Sebastian Cabot       Louis
Prima
994         Richard Lester      John Lennon     Paul McCartney   George
Harrison
997         Fred Zinnemann   Burt Lancaster   Montgomery Clift
Deborah Kerr

                 Star4  No_of_Votes        Gross
0       William Sadler       2343110   28,341,469
1         Diane Keaton       1620367  134,966,411
2        Michael Caine       2303232  534,858,444
3         Diane Keaton       1129952   57,300,000
4         John Fiedler        689845    4,360,000
..                 ...           ...          ...
990        Maria Monti          30144      696,690
991  Carroll O'Connor           45338    1,378,435
992  Bruce Reitherman          166409  141,843,612
994         Ringo Starr         40351   13,780,024
997         Donna Reed          43374   30,500,000

[714 rows x 16 columns]
```

#As you have observed, "Gross" Column is in String, we need to convert it into integer by using apply() method

Def convert(s):

s=s.replace(“ ,”,””)

return int(s)

imdb[Gross‘]=imdb[‘Gross’].apply(convert) ***#To convert gross column into integer data***

imdb

Output:

```
                    Movie_Name Released_Year Certificate  Runtime  \
0    The Shawshank Redemption          1994           A  142 min
1               The Godfather          1972           A  175 min
2             The Dark Knight          2008          UA  152 min
3      The Godfather: Part II          1974           A  202 min
4                  12 Angry Men          1957           U   96 min
..                          ...           ...         ...      ...
990              Giù la testa          1971          PG  157 min
991             Kelly's Heroes          1970          GP  144 min
992          The Jungle Book          1967           U   78 min
994        A Hard Day's Night          1964           U   87 min
997     From Here to Eternity          1953      Passed  118 min

                            Genre  IMDB_Rating  \
0                           Drama          9.3
1                   Crime, Drama          9.2
2             Action, Crime, Drama          9.0
3                   Crime, Drama          9.0
4                   Crime, Drama          9.0
..                           ...          ...
990           Drama, War, Western          7.6
991        Adventure, Comedy, War          7.6
992  Animation, Adventure, Family          7.6
994        Comedy, Music, Musical          7.6
997           Drama, Romance, War          7.6

                                                  Overview
Audience_rating  \
0    Two imprisoned men bond over a number of years...
80.0
1    An organized crime dynasty's aging patriarch t...
100.0
2    When the menace known as the Joker wreaks havo...
84.0
3    The early life and career of Vito Corleone in ...
90.0
4    A jury holdout attempts to prevent a miscarria...
96.0
..                                                 ...                  ..
.
990  A low-life bandit and an I.R.A. explosives exp...
77.0
991  A group of U.S. soldiers sneaks across enemy l...
50.0
992  Bagheera the Panther and Baloo the Bear have a...
65.0
994  Over two "typical" days in the life of The Bea...
96.0
997  In Hawaii in 1941, a private is cruelly punish...
85.0

                  Director             Star1               Star2
Star3  \
0          Frank Darabont      Tim Robbins     Morgan Freeman        Bob
Gunton
1     Francis Ford Coppola    Marlon Brando          Al Pacino
James Caan
2        Christopher Nolan   Christian Bale       Heath Ledger     Aaron
Eckhart
3     Francis Ford Coppola        Al Pacino     Robert De Niro     Robert
Duvall
4             Sidney Lumet      Henry Fonda        Lee J. Cobb     Martin
Balsam
..                     ...              ...                ...
...
990           Sergio Leone      Rod Steiger       James Coburn      Romolo
Valli
991        Brian G. Hutton   Clint Eastwood      Telly Savalas       Don
Rickles
992    Wolfgang Reitherman      Phil Harris    Sebastian Cabot       Louis
Prima
994          Richard Lester      John Lennon     Paul McCartney   George
Harrison
997          Fred Zinnemann   Burt Lancaster   Montgomery Clift
Deborah Kerr

                 Star4  No_of_Votes      Gross
0       William Sadler       2343110   28341469
1         Diane Keaton       1620367  134966411
2        Michael Caine       2303232  534858444
3         Diane Keaton       1129952   57300000
4         John Fiedler        689845    4360000
..                 ...           ...        ...
990        Maria Monti         30144     696690
991  Carroll O'Connor          45338    1378435
992  Bruce Reitherman         166409  141843612
994        Ringo Starr         40351   13780024
997         Donna Reed         43374   30500000

[714 rows x 16 columns]
```

#The gross is changed and can be used in Visualization

➢ ANALYSIS AND VISUALIZATION:

- **Displaying number of movies released in each year for the last 10 years i.e., (2011-2019).**

Year = imdb.groupby(['Released_Year'])['Movie_Name'].count().tail(10)
Year

Output:

```
ye=imdb.groupby(['Released_Year'])['Movie_Name'].count().tail(10)

ye

Released_Year
2011    14
2012    14
2013    22
2014    28
2015    18
2016    20
2017    18
2018    12
2019    15
PG       1
Name: Movie_Name, dtype: int64
```

sb.barplot(data=ye.reset_index(), y="Released_Year", x="Movie_Name")

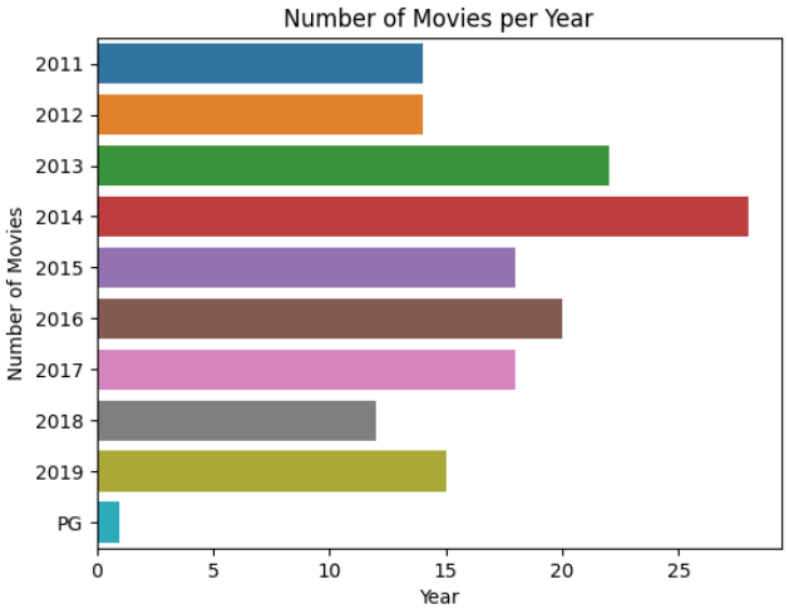
mp.xlabel('Year')

mp.ylabel('Number of Movies')

mp.title('Number of Movies per Year')

mp.show()

Output:



INSIGHTS:
The number of movies released per year shows an upward trend from 2011 to 2014.
The year 2014 had the highest number of movie releases, with a count of 28 movies. Sudden decrease in movies in 2015 is due to people exposure to new all kinds of cinema and decreases the remake rate The data suggests that the movie industry experienced growth and increased production over the years, with a peak in 2016.

- **CLASSIFY MOVIES BASED ON RATINGS.**

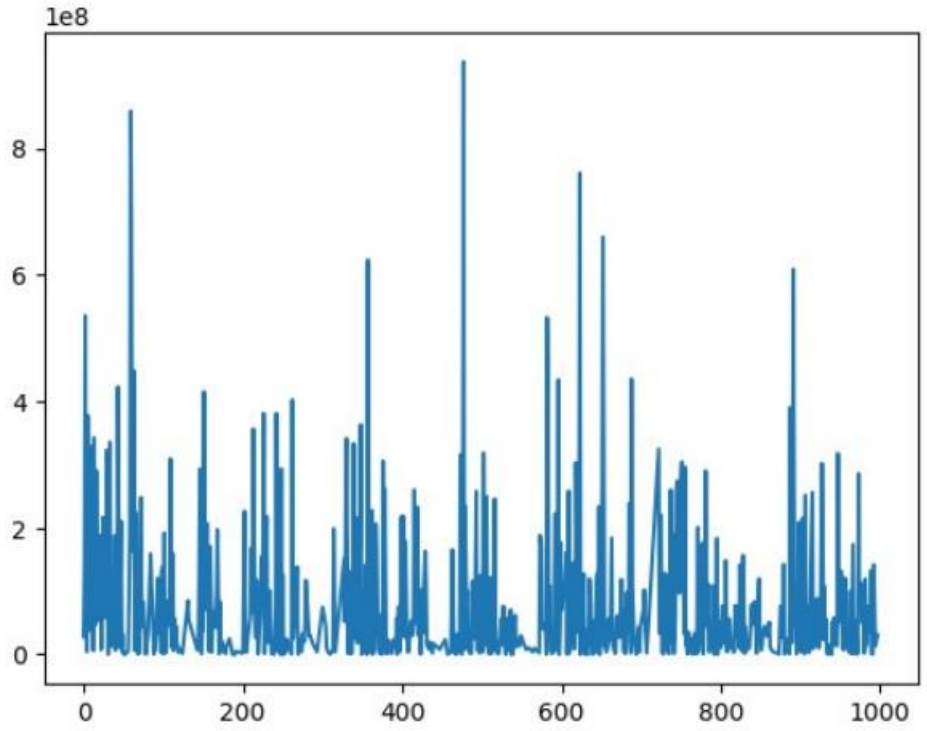
Def rating(r):
if r>=9.0
return "Classics"
elif r<9.0 and r>=8.0
return "Outstanding films"
elif r<8.0 and r>=7.0
return "Good films"
else:
return "Good Watch"
imdb['Categories']=imdb["IMDB_Rating"].apply(rating)

Output:

```
                  Movie_Name Released_Year Certificate  Runtime
Genre  \
0  The Shawshank Redemption          1994           A      142
Drama
1             The Godfather          1972           A      175  Crime,
Drama

   IMDB_Rating                                               Overview  \
0          9.3  Two imprisoned men bond over a number of years...
1          9.2  An organized crime dynasty's aging patriarch t...

   Audience_rating              Director          Star1
Star2  \
0             80.0        Frank Darabont    Tim Robbins  Morgan
Freeman
1            100.0  Francis Ford Coppola  Marlon Brando       Al
Pacino

        Star3           Star4  No_of_Votes      Gross categories
Categories
0  Bob Gunton  William Sadler      2343110   28341469   Classics
Classics
1  James Caan    Diane Keaton      1620367  134966411   Classics
Classics
```

- **REVENUE OF FILMS IN LAST 10 YEARS?**

Imdb.Gross.plat()
<Axes: >



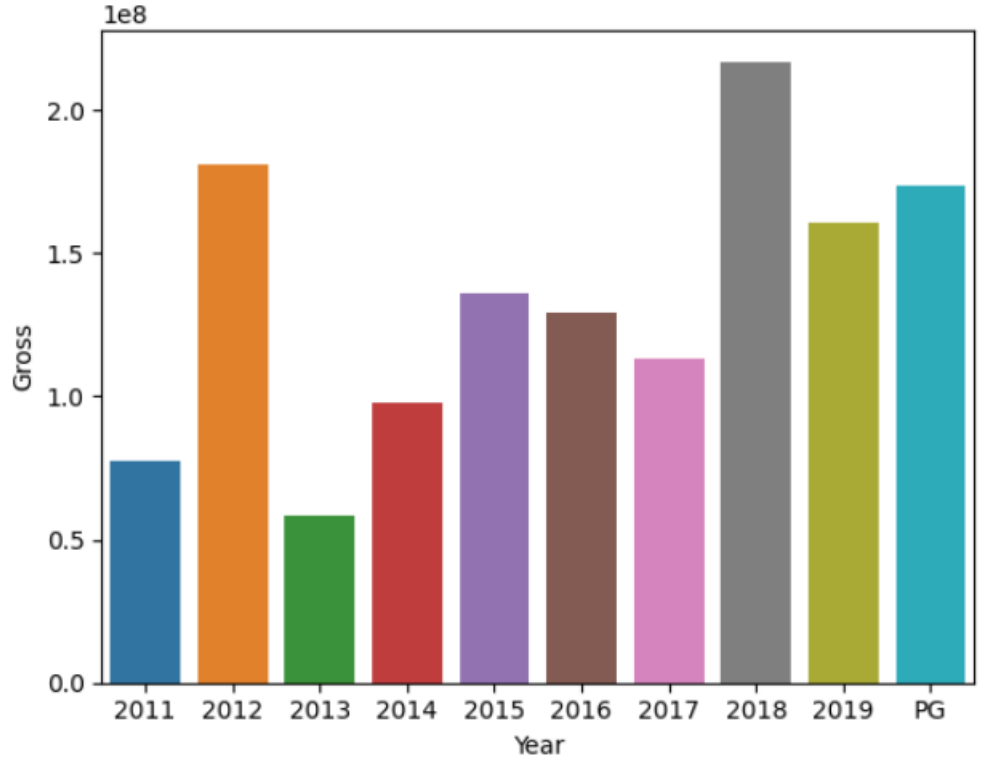
revenue = imdb.groupby([‘Released Year’])[‘Gross’].mean().tail(10)
revenue

Output:

```
Released_Year
2011    7.771727e+07
2012    1.811284e+08
2013    5.810114e+07
2014    9.799807e+07
2015    1.359031e+08
2016    1.290071e+08
2017    1.133773e+08
2018    2.168498e+08
2019    1.603896e+08
PG      1.738379e+08
Name: Gross, dtype: float64
```

sb.barplot(data=revenue.reset_index(),y=’Gross’,x=”Released_Year”)

Output:



INSIGHTS:

Year "**2018**" has **highest revenue so far** with more than $2B USD and revenue is increasing since 2013.

- **DOES RATING EFFECT THE REVENUE?**

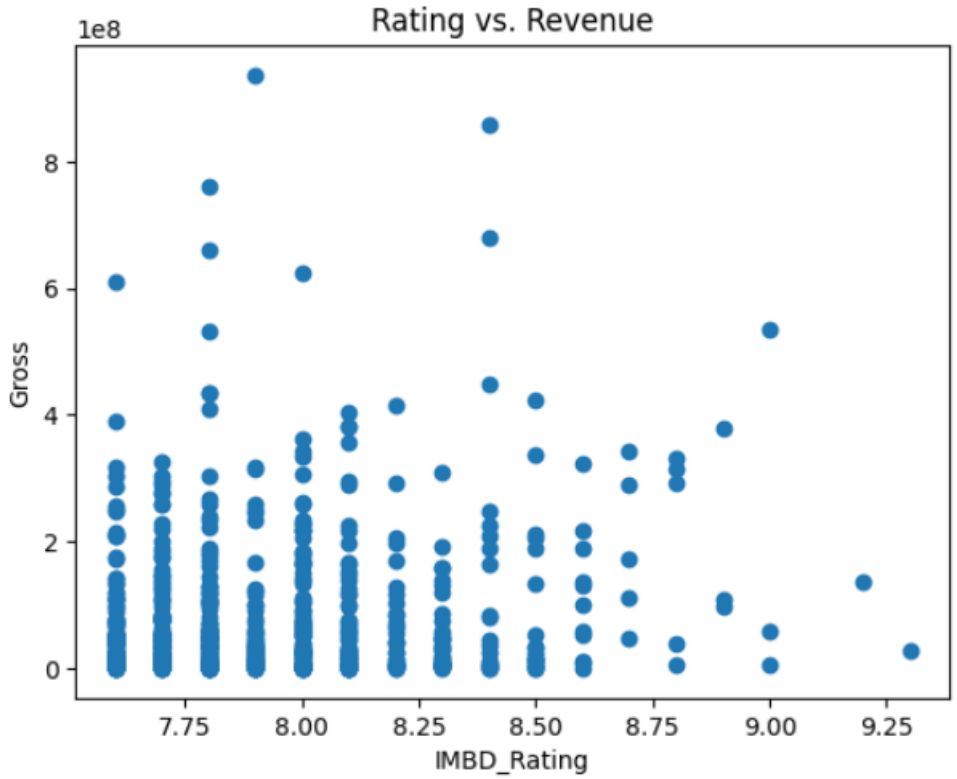
Imdb.columns
Index([‘Poster_Link’,’Movie_Name’,’Released_Year’,’Certificate’,’Runtime’,’Genre’, ’IMDB_Rating’,’Overview’,’Audience_rating’,’Director’,’Star1’,’Star2’,’Star3’,’Star4’, ’No_of_votes’,’Gross’,’categories’,’Categories’,’Cast’],dtype=’object’)
Imdb.IMDB_Rating.unique()

Output:

```
array([9.3, 9.2, 9. , 8.9, 8.8, 8.7, 8.6, 8.5, 8.4, 8.3, 8.2, 8.1,
8. ,
       7.9, 7.8, 7.7, 7.6])
```

mp.scatter(imdb[‘IMDB_Rating’], imdb[‘Gross’])
mp.xlabel(‘IMDB_Rating’)
mp.ylabel(‘Gross’)
mp.title(‘Rating vs Revenue’)
mp.show()

Output:



correlation = imdb[‘IMDB_Rating’].corr(imdb[‘Gross’])
correlation

Output:

```
0.1259915773854415
```

INSIGHTS:

A correlation coefficient of 0.125 suggests **a positive but weak correlation between the movie's rating and its revenue**. Hence, we can say that rating does not greatly affect the revenue of the film.

7. ADVANTAGES

The data analysis performed on the IMDb ratings and movie gross project offers several advantages that contribute to a deeper understanding of the relationship between audience perceptions and movie box office success. Some of the key advantages include:

- Exploring the relationship between imdb ratings and movie gross, it becomes possible to identify genres, directors, or actors that resonate more strongly with audiences, guiding future content creation.

Insights into Audience Preferences

- By examining the factors contributing to movie success, such as runtime, genres, or release years, the project sheds light on the complex dynamics that influence a movie's commercial trajectory.

Understanding Movie Success Factors

- The project's analysis can validate marketing strategies that lead to higher imdb ratings and, consequently, increased box office earnings.

Validation of Marketing Strategies

- The analysis helps identify high-performing movies that have achieved both critical acclaim and financial success. This knowledge allows filmmakers and investors to learn from successful patterns and replicate them in future projects.

Identifying High-Performing Movies

- By quantifying the impact of imdb ratings on box office earnings, filmmakers and studios can make informed choices regarding marketing strategies, release dates, and investment allocation.

Data-Driven Decision Making

- Studios and filmmakers who use data to understand audience sentiments can adapt to changing preferences, stay ahead of trends, and create content that resonates with viewers.

Competitive Advantage

- Understanding the correlation between imdb ratings and box office performance helps studios assess the potential success of a movie before its release, reducing the risk of box office failures.

Risk Mitigation

- Studios can allocate resources more efficiently based on the analysis, focusing on projects that align with audience preferences and have higher chances of commercial success.

Resource Optimization

- Sharing insights with industry professionals fosters collaboration and continuous improvement in the pursuit of cinematic excellence.

Valuable Industry Insights

8. CONCLUSION

- As we have taken the IMDb dataset for analysing top grosser films over past 100 Years.
- After completing the visualization and analysing, we have gained some insights from it:
  - We have over 29 films greater than 180 minutes(3hrs) and "Gone with the Wind" is the lengthiest Film over last 100 years.
  - In the last IO Years, over 160 films have released and 2014 has highest releases.
  - Coming to the Revenue, 2012 & 2018 have the highest revenue over $1.5B USD.
  - Over 10 directors have IMDb rating over 8.5 and "Frank Darabont" has highest rating 8.95.
  - **Review does affect the revenue of the film but not to the great extent. As we have seen, people go to the theatres if the film was good. Content attracts the audience.**

In summary, our analysis indicates that reviews have limited direct impact on box office outcomes. Instead, our findings emphasize the crucial role of captivating content in attracting audiences. Compelling genres, narratives, and performances appear to be more influential in driving box office success. **This underscores the significance of engaging storytelling and content quality in shaping audience preferences and cinematic triumph**.

9. REFERENCES

DATASET: *www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows*