Mahatma Education Society's

Pillai College of Arts, Commerce & Science (Autonomous)

Affiliated to University of Mumbai

'NAAC Accredited 'A' grade (3 cycles)'
'Best College Award' by University of Mumbai
ISO 9001:2015 Certified

MAHATMA EDUCATION SOCIETY'S

PILLAI COLLEGE OF ARTS, COMMERCE & SCIENCE

(Autonomous)

NEW PANVEL

PROJECT REPORT ON

"MOVIE REVIEWS AND RATINGS ANALYSIS"

IN PARTIAL FULFILLMENT OF

Degree of Bachelor Of Science in Computer Science

SEMESTER III :-2025-2026

PROJECT GUIDE : PROF. SANJANA BHANGALE

SUBMITTED BY: Gopal Ingale

ROLL NO: 5296

DIVISION :- SYCS / A

Mahatma Education Society's

Pillai College of Arts, Commerce & Science

(Autonomous)
Affiliated to University of Mumbai
NAAC Accredited 'A' grade (3 cycles)
Best College Award by University of Mumbai
ISO 9001:2015 Certified

Certificate

This is to Certify that MR. **Gopal Ingale** of **S.Y.B.Sc. CS Semester IV** has Completed the Practical work in the Subject of the **Introduction to Data Science** During the academic year **2025-2026** Under the Guidance of **Prof. Sanjana Bhangale Mam** Being the partial requirement for the Curriculum of **Degree Of Bachelor Of Science in Computer Science , University Of Mumbai**

Place:

Date:

Name & Signature of Faculty:

Name & Signature of External:

INDEX

Title	Page No.
Acknowledgement	4
Data Description	5
Data Analysis Questions	6 - 14
Data Visualization Questions	15 - 18
Conclusion	19
References	20

ACKNOWLEDGEMENT

To list who all have helped me is difficult because they are so numerous and the depth is so enormous. I would like to acknowledge the following as being idealistic channels and fresh dimensions in the completion of this project.

I would like to thank my Principal, **Dr. Gajanan Wader** for providing the necessary facilities required for completion of this project.

I take this opportunity to thank our Coordinator for her moral support and guidance.

I would also like to express my sincere gratitude towards my project guide **Prof. Sanjana Mam** whose guidance and care made the project successful.

Lastly, I would like to thank each and every person who directly or indirectly helped me in the completion of the project especially my Parents and Peers who supported me throughout my project.

DATASET DESCRIPTION

The dataset used in this project contains information about different movies, their ratings, financial performance, and audience reviews. The data helps in analyzing movie success based on ratings, budget, revenue, and viewer feedback.

For this project, the dataset includes the following main columns:

- **movie_name** – Name of the movie
- **director** – Director of the movie
- **language** – Language in which the movie was released
- **runtime_minutes** – Duration of the movie in minutes
- **imdb_rating** – IMDB rating of the movie
- **rating** – Certification category (U, UA, A)
- **budget_crore_inr** – Budget of the movie (in Crore INR)
- **worldwide_revenue_crore_inr** – Total worldwide revenue (in Crore INR)
- **number_of_reviews** – Total number of audience reviews
- **streaming_platform** – Platform where the movie is available
- **review_date** – Date when the review was recorded

Using this dataset, additional columns like **profit** and **movie category (Poor, Average, Excellent)** were created to perform better analysis and visualization.

The dataset is suitable for performing data cleaning, aggregation, grouping, and visualization tasks to understand patterns in movie performance and audience preferences.

DATA ANALYSIS QUESTIONS

Part A: Data Preparation & Cleaning (10 Questions)

1. Load the dataset using Pandas and display the first 5 rows.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("/content/5296datascience_datset - 5296datascience_datset.csv")
df.head()
```

	movie_name	category	director	language	rating	streaming_platform	runtime_minutes	imdb_rating	budget_crore_inr	worldwide_revenue_crore_inr	number_of_reviews
0	Crimson Sky 1	Biography	Sanjay Leela Bhansali	Hindi	A	Disney+ Hotstar	121	6.0	72	722	49040
1	Echoes of Time 2	Romance	Sanjay Leela Bhansali	Kannada	UA	Netflix	93	5.4	139	547	39953
2	Crimson Sky 3	Drama	Zoya Akhtar	Tamil	A	Theatrical	143	6.0	162	858	925
3	Eternal Flame 4	Action	S.S. Rajamouli	Malayalam	UA	Disney+ Hotstar	109	6.0	192	134	6578
4	Veer Bharat 5	Thriller	Sanjay Leela Bhansali	Telugu	UA	Theatrical	123	8.6	265	579	8680

The dataset Contain information about movies including runtime, ImDB rating, budget, revenue, and number of reviews. this output shows head of dataset.

2. Check the dataset for missing values and identify which columns have nulls.

```
df.isnull().sum()
```

	0
movie_name	0
category	0
director	0
language	0
rating	0
streaming_platform	0
runtime_minutes	0
imdb_rating	0
budget_crore_inr	0
worldwide_revenue_crore_inr	0
number_of_reviews	0
review_date	0

dtype: int64

All Columsr Show 0 missing values. this means the dataset 100% Complet. ensuring reliable stastical analysis without the need of imputation or data cleaning.

3. Display basic statistics (describe()) for numeric columns (runtime_minutes, imdb_rating, budget_crore_inr, worldwide_revenue_crore_inr, number_of_reviews).

```
df[['runtime_minutes',
    'imdb_rating',
    'budget_crore_inr',
    'worldwide_revenue_crore_inr',
    'number_of_reviews']].describe()
```

	runtime_minutes	imdb_rating	budget_crore_inr	worldwide_revenue_crore_inr	number_of_reviews
count	100.000000	100.000000	100.000000	100.00000	100.000000
mean	139.960000	7.065000	162.490000	485.19000	26619.210000
std	31.595588	1.355749	78.095706	270.50117	14979.613034
min	90.000000	5.100000	21.000000	31.00000	675.000000
25%	109.750000	5.900000	96.750000	259.50000	14999.500000
50%	136.500000	6.800000	159.500000	488.00000	27308.000000
75%	169.250000	8.225000	227.250000	717.50000	39369.500000
max	190.000000	9.500000	298.000000	1000.00000	49768.000000

most movies are approximately 2 to 2.5 hours long with moderate variation 31.59 minutes.

4. Convert review_date to datetime format.

```
df['review_date'] = pd.to_datetime(df['review_date'])
print(df.dtypes)
```

```
movie_name                    object
category                      object
director                      object
language                      object
rating                        object
streaming_platform            object
runtime_minutes                int64
imdb_rating                  float64
budget_crore_inr               int64
worldwide_revenue_crore_inr    int64
number_of_reviews              int64
review_date             datetime64[ns]
profit_crore_inr               int64
review_year                    int32
dtype: object
```

the review-date Column was Successfully Converted to datetime format, allowing time-based analysis such as yearly trends.

5. Check the number of unique movies, directors, and languages.

```
print("Movies:", df['movie_name'].nunique())
print("Directors:", df['director'].nunique())
print("Languages:", df['language'].nunique())
```

```
Movies: 100
Directors: 6
Languages: 5
```

The dataset includes 100 movies directed by only 6 directors suggesting repeated contributions by the same directors. Movies are released in 5 different languages.

6. Standardize the rating column (A, UA, U) to consistent uppercase values.

```
df['rating'] = df['rating'].str.upper()
print(df['rating'].unique())
```

```
['A' 'UA' 'U']
```

This confirms consistent classification without formatting errors.

7. Remove any duplicate rows if present.

```
df.drop_duplicates(inplace=True)
print("Total Rows After Removing Duplicates:", df.shape[0])
```

```
Total Rows After Removing Duplicates: 100
```

No duplicates records were found in the dataset ensuring clean and accurate analysis

8. Add a new column profit_crore_inr = worldwide_revenue_crore_inr – budget_crore_inr.

```
df['profit_crore_inr'] = df['worldwide_revenue_crore_inr'] - df['budget_crore_inr']
print(df[['movie_name','profit_crore_inr']].head())
```

```
          movie_name  profit_crore_inr
0     Crimson Sky 1               650
1  Echoes of Time 2               408
2     Crimson Sky 3               696
3   Eternal Flame 4               -58
4     Veer Bharat 5               324
```

when most movies are profitable, at least one incurred a loss of ₹ 58 Crore. showing that high budget do not guarantee success.

9. Categorize movies based on IMDB rating: Poor (<6), Average (6–7.5), Excellent (>7.5).

```
df['category'] = df['imdb_rating'].apply(
    lambda x: "Poor" if x < 6 else "Average" if x <= 7.5 else "Excellent")
print(df[['movie_name','imdb_rating','category']].head())
```

```
...
          movie_name  imdb_rating   category
0     Crimson Sky 1          6.0    Average
1  Echoes of Time 2          5.4       Poor
2     Crimson Sky 3          6.0    Average
3   Eternal Flame 4          6.0    Average
4     Veer Bharat 5          8.6  Excellent
```

movies with ratings above 7.5 are Categorized as excellent, indicating strong audience approval.

10. Extract the year from review_date into a new column review_year.

```
df['review_year'] = df['review_date'].dt.year
print(df[['movie_name','review_year']].head())
```

```
...
          movie_name  review_year
0     Crimson Sky 1          2022
1  Echoes of Time 2          2022
2     Crimson Sky 3          2020
3   Eternal Flame 4          2023
4     Veer Bharat 5          2019
```

The dataset Covers multiple layers, enabling trend analysis across time

Part B: Basic Data Analysis (10 Questions)

11. Count the number of movies in each category.

```
print(df['category'].value_counts())
```

```
...
category
Excellent    41
Average      33
Poor         26
Name: count, dtype: int64
```

Out of 100 total movies, 41% are Categorized as Excellent, 33% average and 26% Poor. this shows that the majority of movies in the dataset perform well.

12. Find the average runtime of movies per category.

```
print(df.groupby('category')['runtime_minutes'].mean())
```

```
category
Average      139.545455
Excellent    141.902439
Poor         137.423077
Name: runtime_minutes, dtype: float64
```

Excellent movies have the highest average runtime of 141.9 min, followed by average movies at 130.55 minutes and Poor at 137.42 minutes.

13. Identify the director who has made the most movies in the dataset.

```
print(df['director'].value_counts().head(1))
```

```
director
Sanjay Leela Bhansali    19
Name: count, dtype: int64
```

Sanjay lila bhansali has directed 19 out of 100 movies making him most movies directed by single director.

14. Calculate the average IMDB rating per language.

```
print(df.groupby('language')['imdb_rating'].mean())
```

```
language
Hindi        7.150000
Kannada      7.161905
Malayalam    6.790476
Tamil        7.156000
Telugu       7.060000
Name: imdb_rating, dtype: float64
```

Kannada movies have highest average rating of 7.16, malaylam at lowest 6.79.

15. Find the movie with the highest worldwide revenue.

```
print(df.loc[df['worldwide_revenue_crore_inr'].idxmax()])
```

```
movie_name                    Lost in Bharat 91
category                                   Poor
director                             Zoya Akhtar
language                           Malayalam
rating                                        UA
streaming_platform                   Theatrical
runtime_minutes                             169
imdb_rating                                 5.3
budget_crore_inr                             70
worldwide_revenue_crore_inr                1000
number_of_reviews                         14117
review_date                2025-12-14 00:00:00
profit_crore_inr                            930
review_year                                2025
Name: 90, dtype: object
```

lost in bharat 91 generated the highest worldwide revenue of ₹ 1000 Cror, earning a massive profit of ₹ 930 Cror having IMDB rating of 5.3 this indicate high revenue does not always Correlate with high ratings.

16. Find the movie with the lowest budget

```
print(df.loc[df['budget_crore_inr'].idxmin()])
```

```
movie_name                    Mumbai Nights 15
category                                 Poor
director                       S.S. Rajamouli
language                               Telugu
rating                                      U
streaming_platform               Disney+ Hotstar
runtime_minutes                              145
imdb_rating                                  5.7
budget_crore_inr                              21
worldwide_revenue_crore_inr                  769
number_of_reviews                          47664
review_date                2023-12-12 00:00:00
profit_crore_inr                             748
review_year                                 2023
Name: 14, dtype: object
```

mumbai nights is had lowest budget of 21 cr but generated 769 cr revenue. resulting in an impressive profit of 748 cr.

17. Find the average profit per category.

```
print(df.groupby('category')['profit_crore_inr'].mean())
```

```
category
Average      328.606061
Excellent    288.170732
Poor         369.653846
Name: profit_crore_inr, dtype: float64
```

Interestingly Poor-rated movies have the highest average profit of 369.65 cr while excellent movies have 288.17 Crore. This indicate success is not depend on IMDB ratings.

18. How many movies are available on each streaming_platform?

```
print(df['streaming_platform'].value_counts())
```

```
streaming_platform
Zee5               26
Theatrical         22
Netflix            20
Disney+ Hotstar    19
Amazon Prime       13
Name: count, dtype: int64
```

zee5 hosts the highest number of movies (26%) while amazon prime has lowest share (13%.) this distribution across platform is relatively balanced.

19. What is the average number of reviews per category?

```
print(df.groupby('category')['number_of_reviews'].mean())
```

```
category
Average      24498.212121
Excellent    27105.146341
Poor         28544.961538
Name: number_of_reviews, dtype: float64
```

Poor rated movies receive the highest average reviews (28,544) indicating that Controversial or Poorly rated movies may generate higher audience engagement.

20. Find the top 3 directors with the highest average IMDB rating.

```
print(df.groupby('director')['imdb_rating']
      .mean()
      .sort_values(ascending=False)
      .head(3))
```

```
director
Rohit Shetty      7.656250
Anurag Kashyap    7.433333
S.S. Rajamouli    7.211111
Name: imdb_rating, dtype: float64
```

Rohit Shetty has the highest average IMDB rating of 7.66 making him the top-rated director in the dataset

📊 Part C: Grouping & Aggregation (5 Questions)

21. Group movies by language and calculate the total worldwide revenue per language.

```
print(df.groupby('language')['worldwide_revenue_crore_inr'].sum())
```

```
language
Hindi        8805
Kannada      9568
Malayalam   10301
Tamil       13985
Telugu       5860
Name: worldwide_revenue_crore_inr, dtype: int64
```

Tamil movies generate the highest total revenue of 13,985 crore, while Telugu movies generate the lowest at 5,860 crore. Indicating strong Commercial Performance in Tamil Cinema.

22. Group by rating and calculate the average runtime.

```
print(df.groupby('rating')['runtime_minutes'].mean())
```

```
rating
A     145.937500
U     140.324324
UA    133.354839
Name: runtime_minutes, dtype: float64
```

A rated movies have the longest average runtime (145.94 minutes) while UA rated movies are shorter on average (133.35 minutes)

23. Group by streaming_platform and calculate the mean IMDB rating.

```
print(df.groupby('streaming_platform')['imdb_rating'].mean())
```

```
streaming_platform
Amazon Prime        7.607692
Disney+ Hotstar     6.726316
Netflix             6.615000
Theatrical          7.495455
Zee5                7.023077
Name: imdb_rating, dtype: float64
```

Amazon Prime has highest average IMDB rating (7.61) while netflix has the lowest.

24. For each director, calculate the total number of reviews received.

```
print(df.groupby('director')['number_of_reviews'].sum())
```

```
director
Anurag Kashyap           392587
Rajkumar Hirani          510112
Rohit Shetty             505174
S.S. Rajamouli           388789
Sanjay Leela Bhansali    539242
Zoya Akhtar              326017
Name: number_of_reviews, dtype: int64
```

Sanjay leela Bhassali receives the highest audience engagement with 539,242 total reviews. Indicating strong viewer interest in his films.

25. Group movies by category and calculate the mean profit.

```
print(df.groupby('category')['profit_crore_inr'].mean())
```

```
category
Average      328.606061
Excellent    288.170732
Poor         369.653846
Name: profit_crore_inr, dtype: float64
```

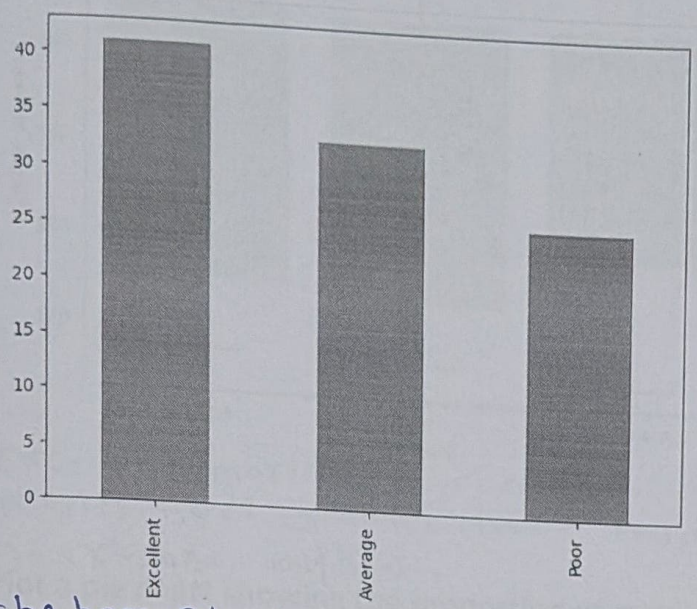
Poor Category movies earn the highest average Profit (369.65 crors)

DATA VISUALIZATION QUESTION

Part D: Data Visualization (10 Questions)

26. Plot a bar chart showing the number of movies per category.

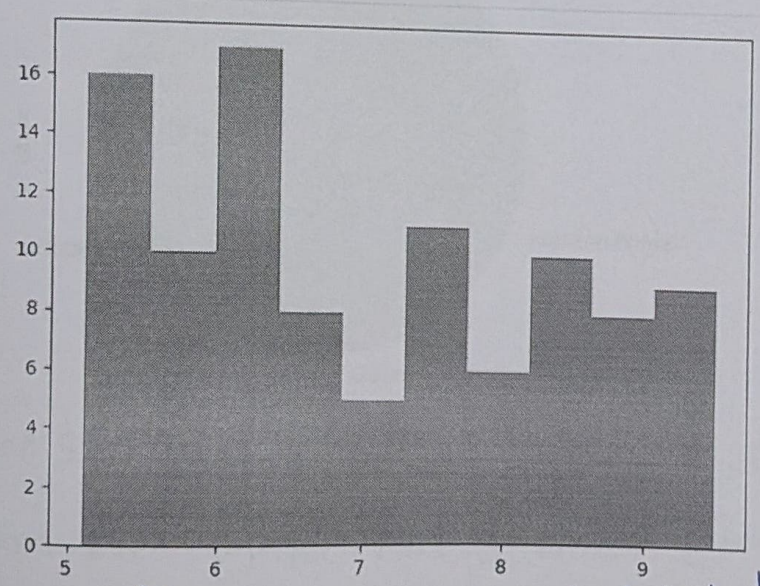
```
df['category'].value_counts().plot(kind='bar')
plt.show()
```



The bar chart clearly shows excellent movies (41) dominate followed by average (33) and poor (26)

27. Create a histogram showing the distribution of imdb_rating.

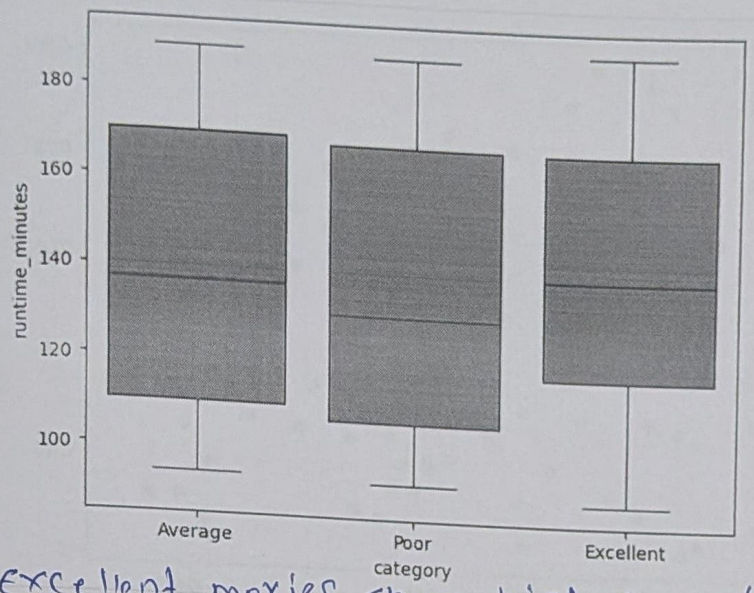
```
plt.hist(df['imdb_rating'], bins=10)
plt.show()
```



The rating mostly distributed between 6 and 8 with fewer movies below 5.5 and above 9. This indicates most movies are moderately rated.

28. Create a boxplot of runtime_minutes per category.

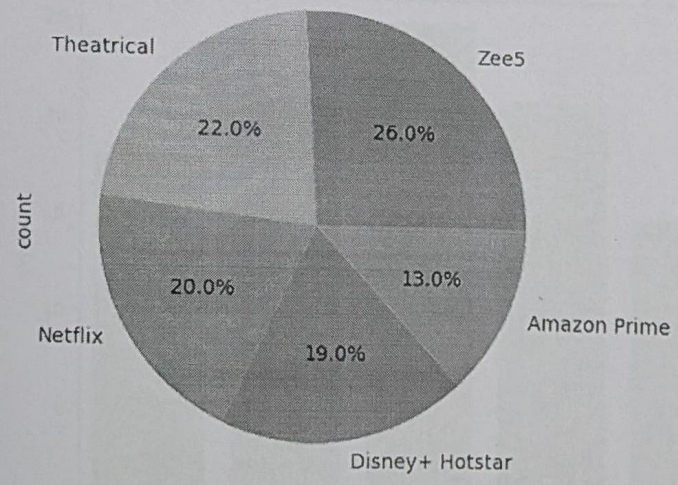
```
sns.boxplot(x='category', y='runtime_minutes', data=df)
plt.show()
```



Excellent movies show higher media runtime, while Poor movies have slightly lower median runtime. There are no extreme outliers.

29. Plot a pie chart showing the proportion of movies per streaming_platform

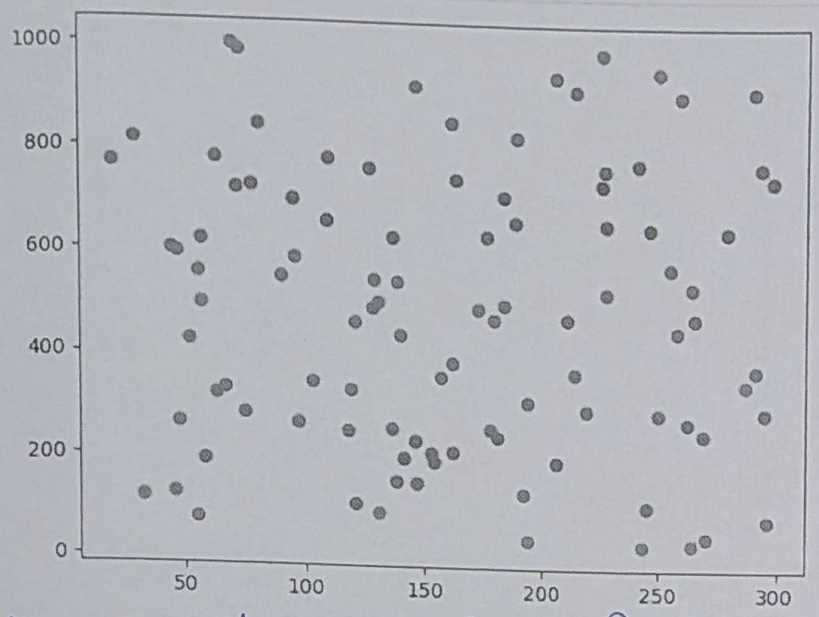
```
df['streaming_platform'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.show()
```



Zee5 Contributes the largest share (26.0) while Amazon prime has smallest Contribution (13%.)

30. Draw a scatter plot of budget_crore_inr vs worldwide_revenue_crore_inr

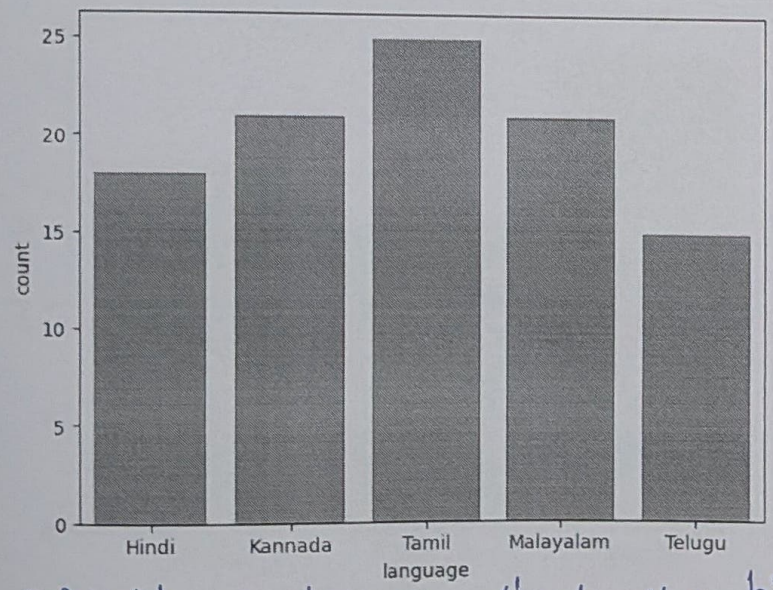
```
plt.scatter(df['budget_crore_inr'], df['worldwide_revenue_crore_inr'])
plt.show()
```



The scatter plot shows positive relation between budget and revenue. higher budget movies generated higher revenue.

31. Use Seaborn to plot a countplot of movies per language.

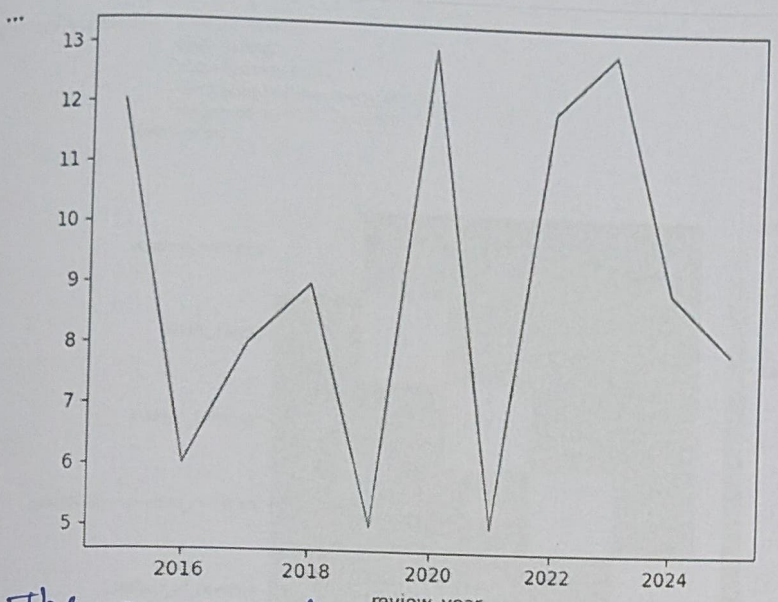
```
sns.countplot(x='language', data=df)
plt.show()
```



Tamil movies constitute the highest share with 25% of total movies. followed by Kannada and malayalam with 24% each. Telugu has lowest 15%.

32. Draw a line plot showing the number of movies released per year

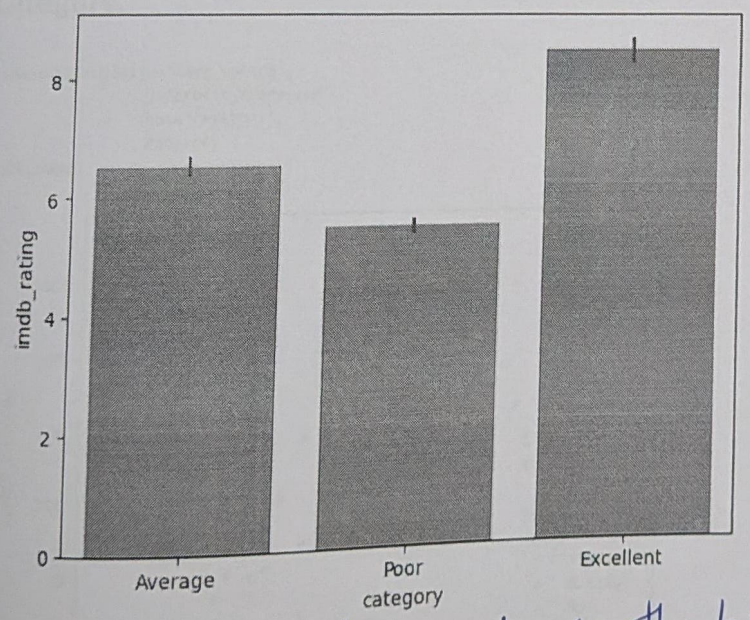
```
df['review_year'].value_counts().sort_index().plot(kind='line')
plt.show()
```



The number of movies peaked in 2020 and 2023 while production declined sharply in 2019 and 2021. This shows fluctuations in yearly movies released over the decade

33. Create a barplot showing the average IMDB rating per category.

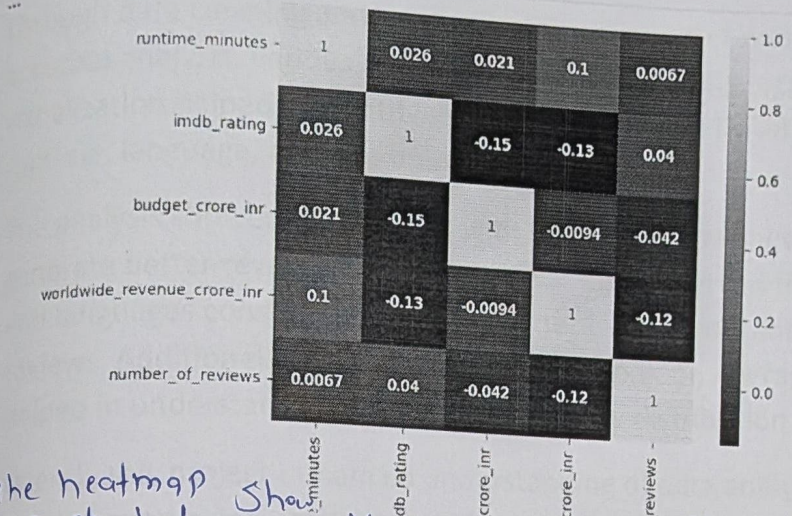
```
sns.barplot(x='category', y='imdb_rating', data=df)
plt.show()
```



Excellent Category have the highest average rating of approx 8.5 while Poor have lowest 5.5. The clear separation confirms Correct Categorization based on IMDB rating.

34. Plot a heatmap showing correlation between numeric columns (runtime_minutes, imdb_rating, budget_crore_inr, worldwide_revenue_crore_inr, number_of_reviews).

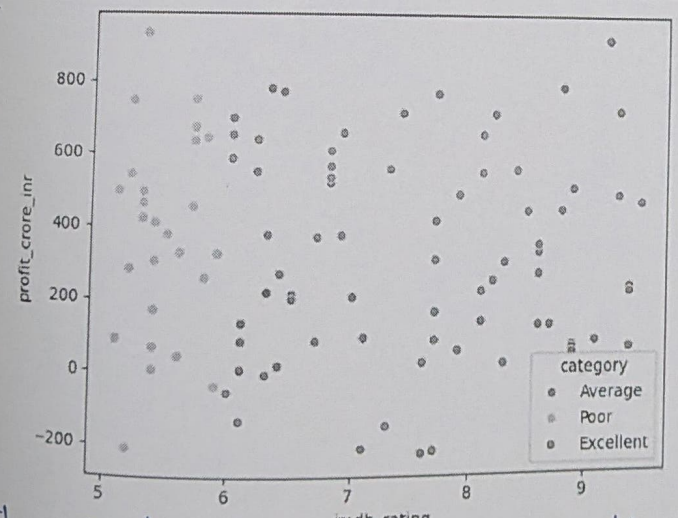
```
sns.heatmap(df[['runtime_minutes',
                'imdb_rating',
                'budget_crore_inr',
                'worldwide_revenue_crore_inr',
                'number_of_reviews']].corr(),
            annot=True)
plt.show()
```



the heatmap shows very weak correlation among variables. highest between runtime and revenue (0.10) there are no strong linear relationship between financial and ratings.

35. Create a scatterplot with hue for profit_crore_inr vs imdb_rating, colored by category.

```
sns.scatterplot(x='imdb_rating',
                y='profit_crore_inr',
                hue='category',
                data=df)
plt.show()
```



The scatterplot reveals the both excellent and poor category movies can generate high profits. Some poor-rated movies earn profit above ₹800 cror while few movies incur higher IMDB ratings do not guarantee higher profit.

CONCLUSION

The project "Movie Review and Ratings Analysis" was successfully completed using data analysis and visualization techniques. By analyzing the dataset, meaningful insights were obtained regarding movie performance, audience preferences, and financial success.

Through data cleaning and preprocessing, the dataset was prepared for accurate analysis. Various operations such as grouping, aggregation, and visualization helped in identifying patterns between IMDB ratings, profit, runtime, language, and streaming platforms.

The analysis showed that movies with higher IMDB ratings generally tend to generate better revenue and profit. It was also observed that certain directors and languages consistently perform better in terms of ratings and audience reviews. Additionally, categorizing movies into Poor, Average, and Excellent helped in understanding overall movie quality distribution.

Overall, this project enhanced understanding of data analysis concepts such as data preprocessing, statistical analysis, and visualization using Python libraries like Pandas, Matplotlib, and Seaborn. The study demonstrates how data science techniques can be used to evaluate movie performance and support decision-making in the entertainment industry.

REFERENCES

1. McKinney, W. (2018). *Python for Data Analysis*. O'Reilly Media.
2. VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.
3. Official Documentation of Pandas – https://pandas.pydata.org
4. Official Documentation of NumPy – https://numpy.org
5. Official Documentation of Matplotlib – https://matplotlib.org
6. Official Documentation of Seaborn – https://seaborn.pydata.org
7. IMDb Official Website – https://www.imdb.com
8. Python Software Foundation. (2024). *Python Documentation*. https://docs.python.org