📘 IMDb Movies Data Analysis - Detailed Project Report

📌 Project Title:

IMDb Movies Data Analysis using Python

🧾 Objective:

The objective of this project is to perform comprehensive exploratory data analysis (EDA) on a dataset of movies obtained from IMDb. The aim is to uncover trends and insights regarding movie popularity, genres, revenue, ratings, return on investment (ROI), and more. This analysis will be useful for stakeholders in the movie industry or researchers interested in entertainment data trends.

🔧 Libraries Used:

1. pandas

Used for data manipulation and analysis. It provides data structures like DataFrames and Series which are ideal for cleaning, transforming, and summarizing datasets.

2. numpy

A numerical computing library used to handle operations on arrays, manage missing values, and work with mathematical functions like isfinite.

3. matplotlib.pyplot

A core visualization library for plotting static graphs like line plots, bar charts, histograms, and pie charts.

4. plotly.express

Used to create interactive visualizations like treemaps and animated plots. It helps in building web-ready, responsive visualizations easily.

📊 Project Steps & Explanation:

🔹 Step 1: Importing Libraries

The project begins by importing all necessary libraries (pandas, numpy, matplotlib.pyplot, and plotly.express). These form the foundation for data analysis and visualization.

🔹 Step 2: Loading the Dataset

The dataset movies.csv is loaded using pandas.read_csv() into a DataFrame named df. We then view the first few rows and get an idea of the dataset structure.

🔹 Step 3: Initial Exploration

We check the shape of the dataset, list of columns, and count of missing values to understand data quality.

🔹 Step 4: Dropping Irrelevant Columns

Columns like id, homepage, tagline, overview, and others are dropped as they do not contribute to the core analysis.

🔹 Step 5: Handling Missing Values

Rows with missing genres or director values are removed. Other columns such as keywords and production_companies are filled with placeholder values (0).

🔹 Step 6: Data Cleaning

The popularity column is rounded to 2 decimal places for neatness.

🔹 Step 7: Feature Engineering

Two new columns are created:

- profit: Calculated as revenue - budget
- roi: Calculated as profit / budget

Infinite ROI values are replaced with NaN to avoid computational issues.

🔹 Step 8: Data Visualization - Histogram

A histogram is plotted for all numeric features to observe their distributions.

🔹 Step 9: ROI Trend Over Time

Using a line plot, we analyze the average ROI across each release year. This helps determine which years were most profitable on average.

🔹 Step 10: Popularity Trend Over Time

A similar line plot is created to show the total popularity by release year. It highlights how interest in movies has changed over time.

🔹 Step 11: Ratings Trend Over Time

The average rating (vote_average) is plotted year-wise to observe rating patterns.

🔹 Step 12: Popularity vs Rating Scatter Plot

A scatter plot shows whether there is any visible correlation between how popular a movie is and its average vote.

🔹 Step 13: Genre Analysis

Genres are split and exploded so each movie contributes to multiple genres where applicable. We then sum popularity by genre and visualize it with a horizontal bar chart.

🔹 Step 14: Monthly Popularity

The month is extracted from the release date, and we calculate total popularity per month. The result is plotted using a bar chart.

🔹 Step 15: Monthly Revenue

Just like popularity, total revenue is aggregated by release month and visualized.

🔹 Step 16: Top 5 Profitable Movies

Movies are grouped by title and their total profit is calculated. The top 5 are shown using a pie chart.

🧾 Final Outcome / Key Insights:

- **Profitability varies greatly by year**: Certain years had exceptionally high ROI.
- **Popularity has trended upward over time**, with some fluctuations.
- **Ratings have remained fairly consistent**, showing general viewer satisfaction.
- **Genres like Action and Adventure tend to be the most popular**.
- **Movies released in the middle of the year (May–July)** tend to earn higher revenue.
- **Some movies achieve very high profit margins**, despite not always being the most popular.

✅ Conclusion:

This analysis provides actionable insights into how movie success (popularity, revenue, ROI) is influenced by factors such as release time, genre, and year. It demonstrates how Python and data science tools can help decode entertainment trends effectively.

📌 Recommendations:

- Use more recent or comprehensive datasets for industry-grade results.
- Integrate IMDb ratings and review sentiment analysis for deeper insights.
- Explore predictive models for box office performance using machine learning.

Prepared by: Sanjana Singh
Course/Project Title: IMDb Data Analysis
Date: June 2025