Name: Neelesh Bhalkikar

Enrolment Number: EN12025086453

Batch Code: DA444S44

Mentor: *Sharayoo Dixit*

Capstone Project: Predictive Modeling of Movie Success

Problem Statement:

The film industry invests millions in movie production, yet predicting a movie’s success remains uncertain. This project aims to analyze TMDb movie data to uncover patterns related to revenue, ratings, and audience preferences. The goal is to develop accurate predictive models and uncover key factors that influence a movie's commercial and critical success. Through structured data analysis, the aim is to deliver insights that support smarter decisions in film production and marketing strategy.

Business Objective:

- Advanced Reporting and Insight Generation using SQL & Excel
- Exploratory Data Analysis & Visualization
- Understanding Audience Engagement and Production Company Contributions
- Statistical Analysis & Hypothesis Testing
- Predictive Modeling & Insight-Driven Decision Making

Data Description

Column Name	Description
title	Title of the movie
cast	Movie cast including the names of main actors and their genders
crew	All members of movie crew including their names, genders, job types, etc.
budget	Budget of the movie
genres	Genre of the movie (Action, Comedy, Horror, etc.)
homepage	Link of the homepage of the movie
id	Unique ID of the movie
keywords	Keywords describing the movie's plot
original_language	Original language of the movie
overview	Overview of the movie
popularity	Popularity score of the movie
production companies	Names of production companies
production_countries	Countries of production of the movie
release_date	Release date of the movie
revenue	Revenue earned by the movie
runtime	Duration of the movie
status	Statusof the movie (released or not)
tagline	Tagline ofthe movie
vote_average	Average vote for the movie
vote_count	Number of votes for the movie

Business Objective

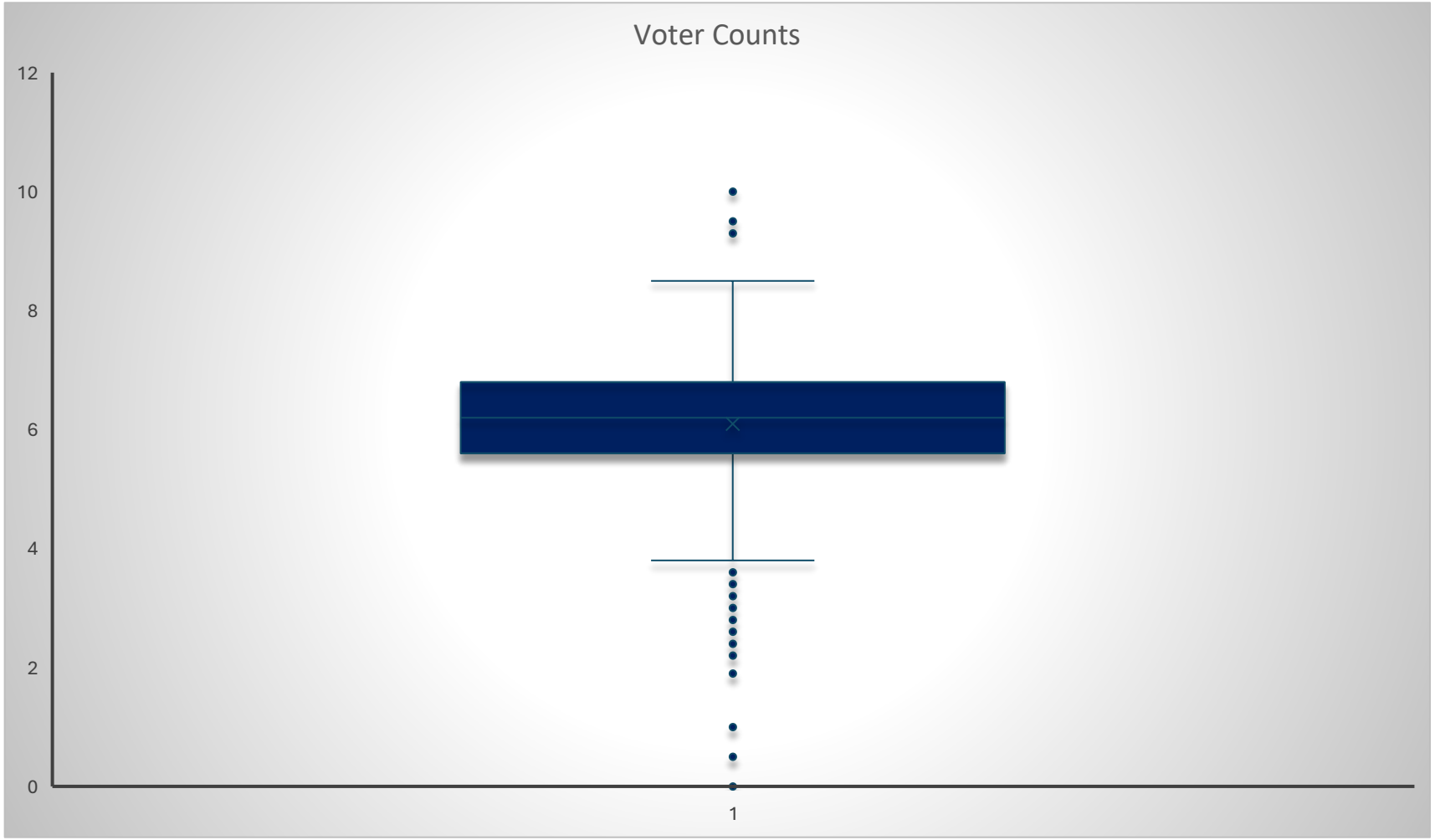
- Business Objective 1.1: Extract deeper insights using advanced SQL operations
- Business Objective 1.2: Develop interactive dashboards and trend visualizations using Excel
- Business Objective 2.1: Uncover temporal and financial trends
- Business Objective 2.2: Understand audience engagement metrics
- Business Objective 3.1: Assess audience voting patterns and distributions
- Business Objective 3.2: Identify key contributors and their market share in movie production
- Business Objective 4.1: Quantify performance differences based on movie characteristics
- Business Objective 4.2: Validate overall budget and rating assumptions
- Business Objective 5.1: Build and optimize predictive models for movie success
- Business Objective 5.2: Evaluate model performance and extract business insights

Tools and Techniques Used

- Excel
  - It is used to get a complete understanding and quick visualisation of the data by creating graphs and by using a dashboard.
- SQL
  - Displaying the necessary information based on the category and question needed.
- Python
  - For cleaning the data, joining the tables, Statical analysis and Machine Learning
- Tableau
  - To create advanced interactive visuals and analyse trends and patterns based on the visuas.

Insights

1. Data Preprocessing & Quality Assurance



Excel

Observations:

- The majority of the movies appear to have been made in the United States of America
- The skewness of the Voter Counts is towards the left, and the mean is 6.

1.2 Develop interactive dashboards and trend visualizations using Excel

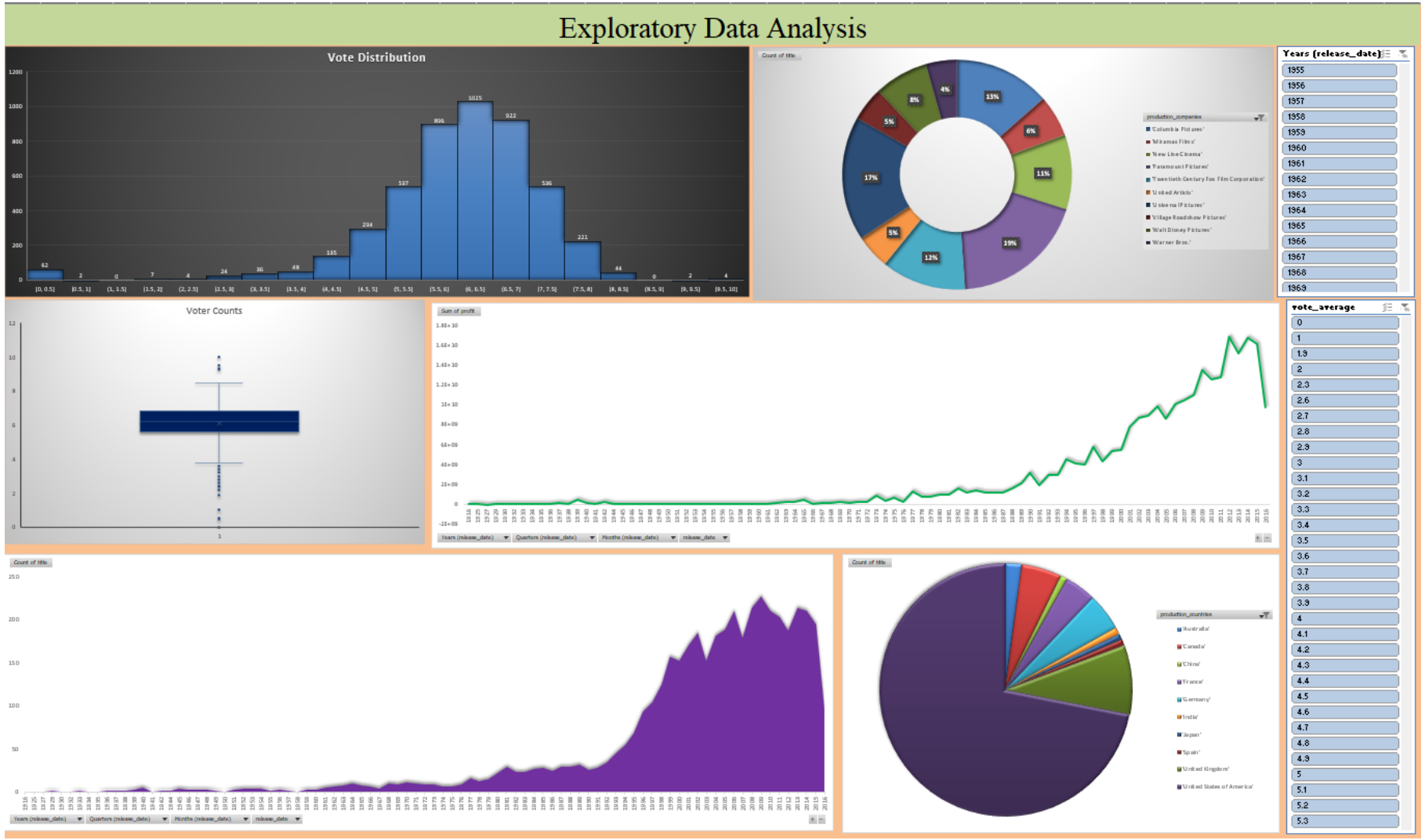
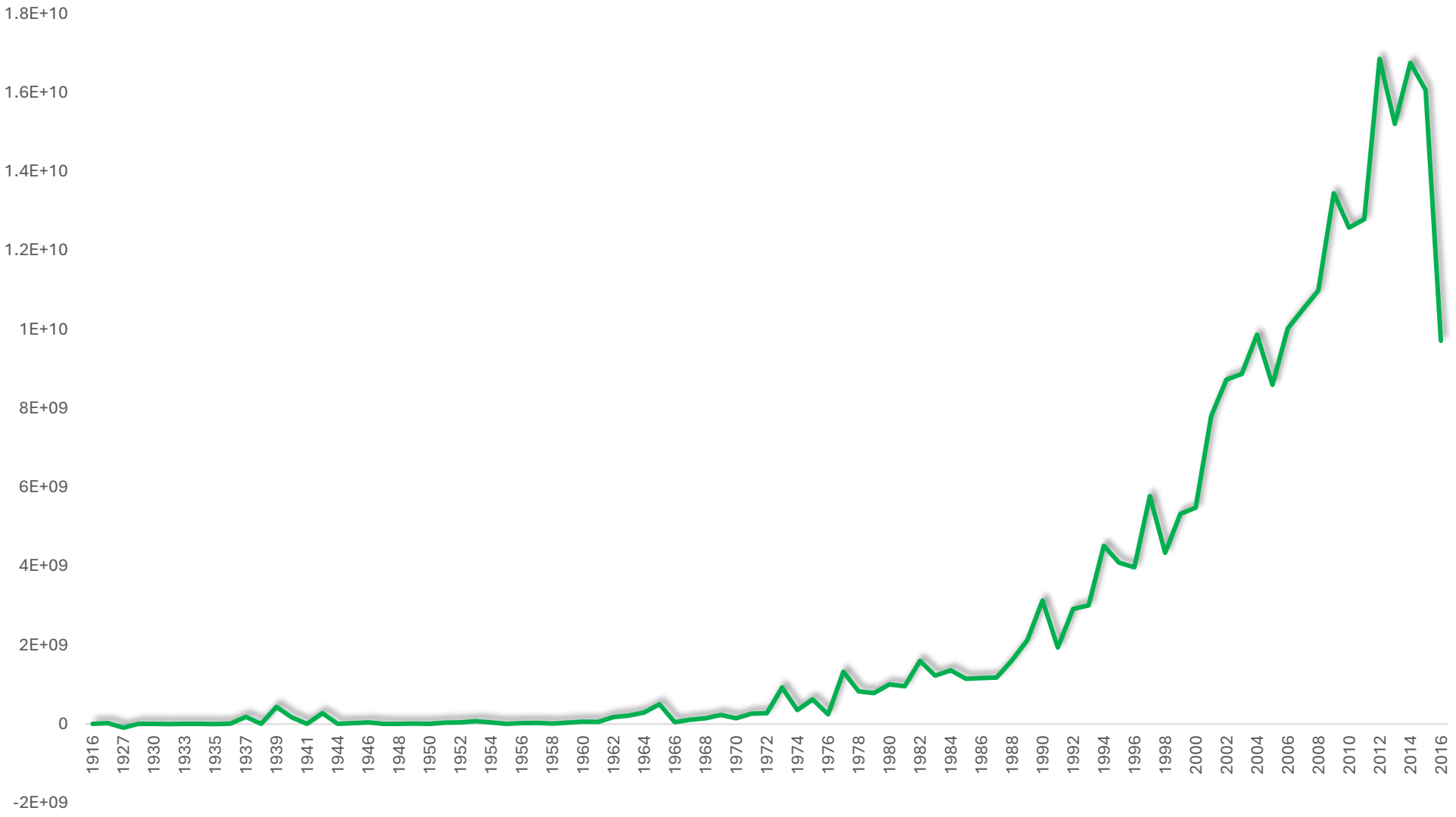


Tableau Public Dashboard

Observations:

- The United States of America is one of the biggest markets for the movies compared to any other country.
- The voters find majority of the movies to be in the range of 4.5 to 8, which would be considered as the average score for the movies.

2.1: Uncover temporal and financial trends



Excel

Observations:

- Over the years, the budget invested in movies and the profits earned by the movies has been increasing.
- The Total numbers of movies also have been increasing over the years.

2.2: Understand audience engagement metrics

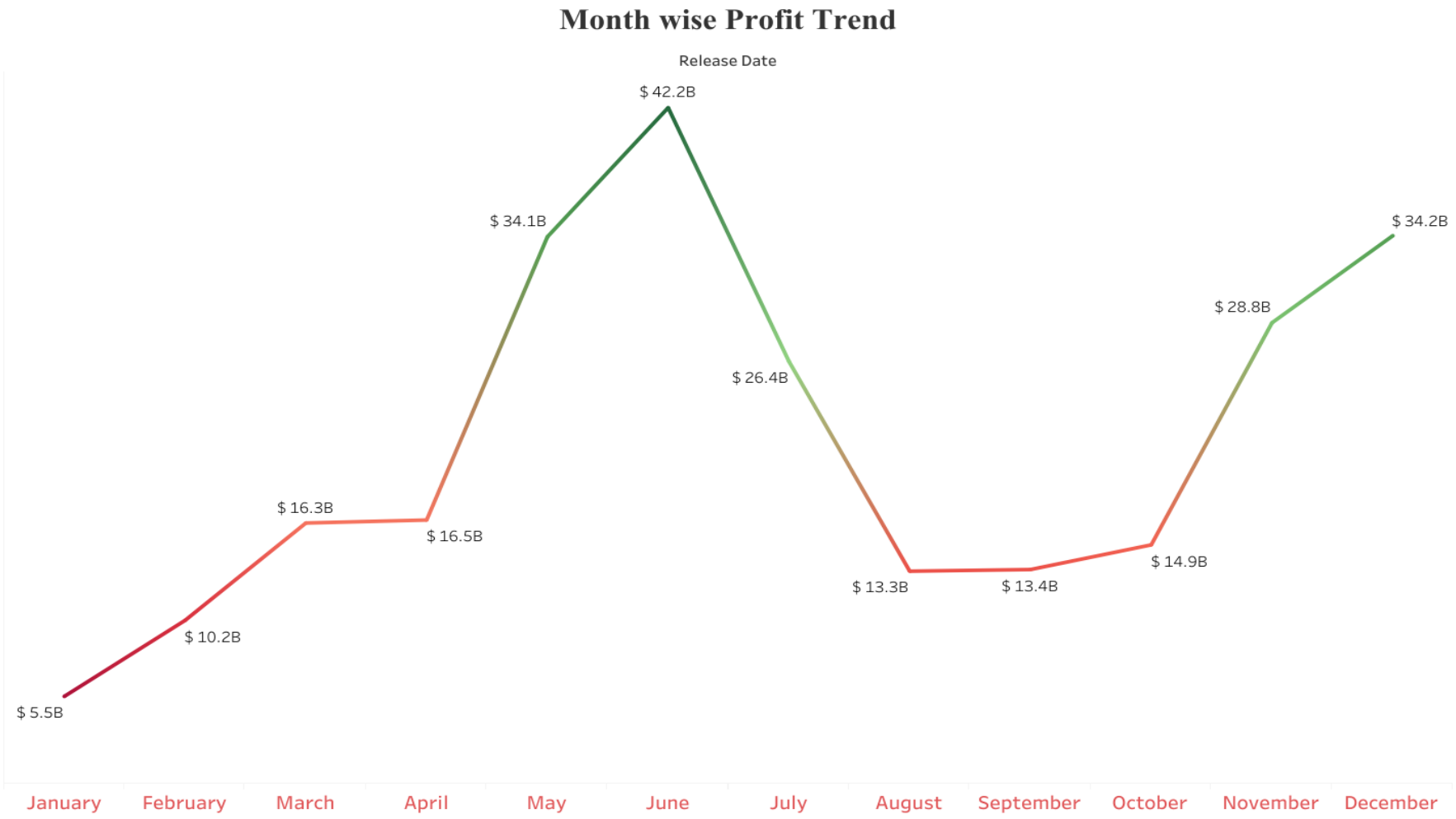


Tableau Public: Months Data

Observations:

- **The rise in popularity and profits for movies in the month of May and June is due to the beginning of the summer holidays for most schools and universities.**
- **The sharp drop in the profits is due to the people going to vacations in summer and not being able to watch as many movies as they would during the other months.**

3.2: Validate overall budget and rating assumptions

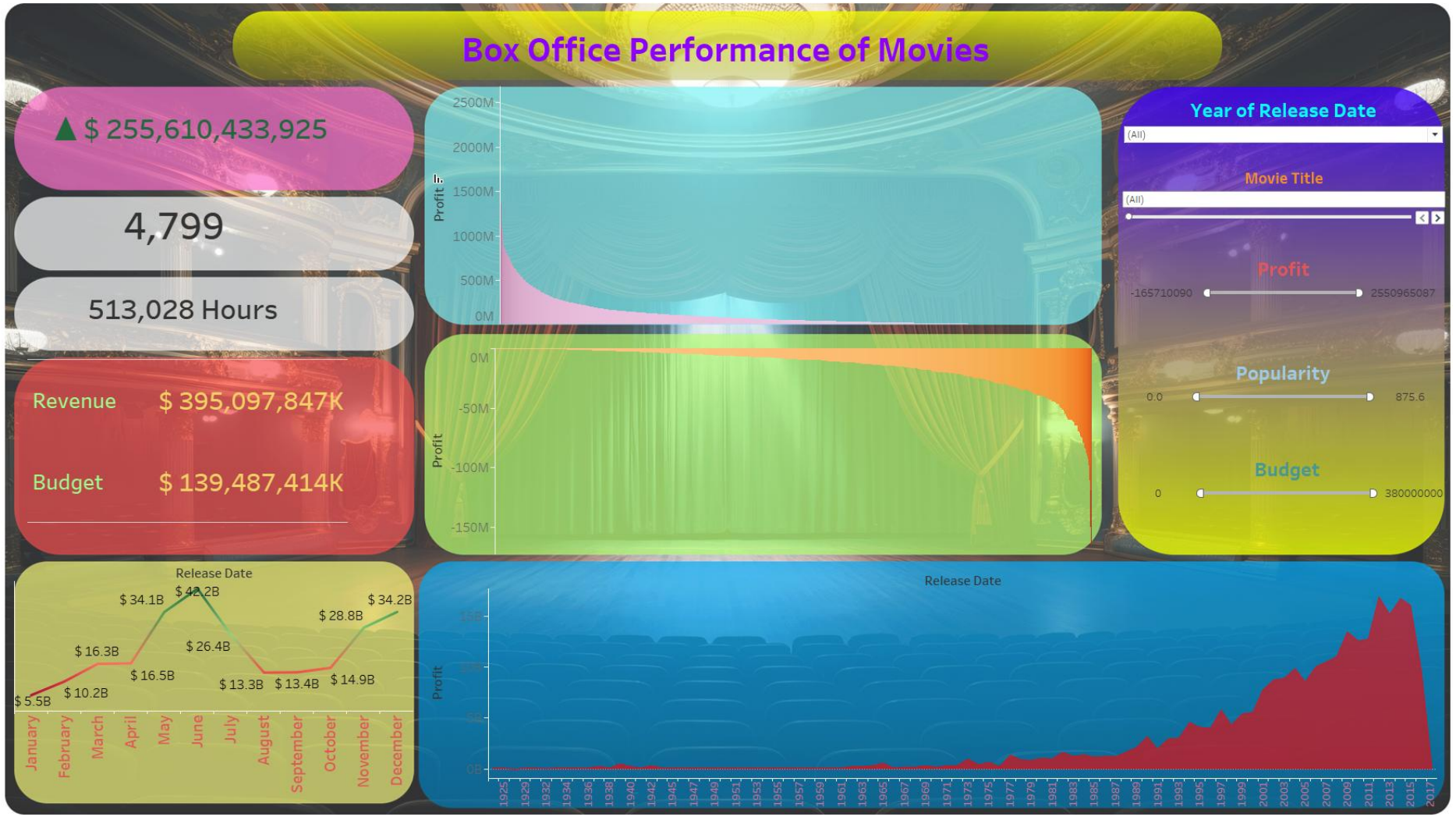


Tableau Public Dashboard

Observations:

- **This Dashboard describes the trend of movies over the years and displays the runtime, profit, revenue and budgets of movies and years.**
- **Before 1987, the total number of movies that were released were less and as a result the profits were not high.**

Year wise details of Movies

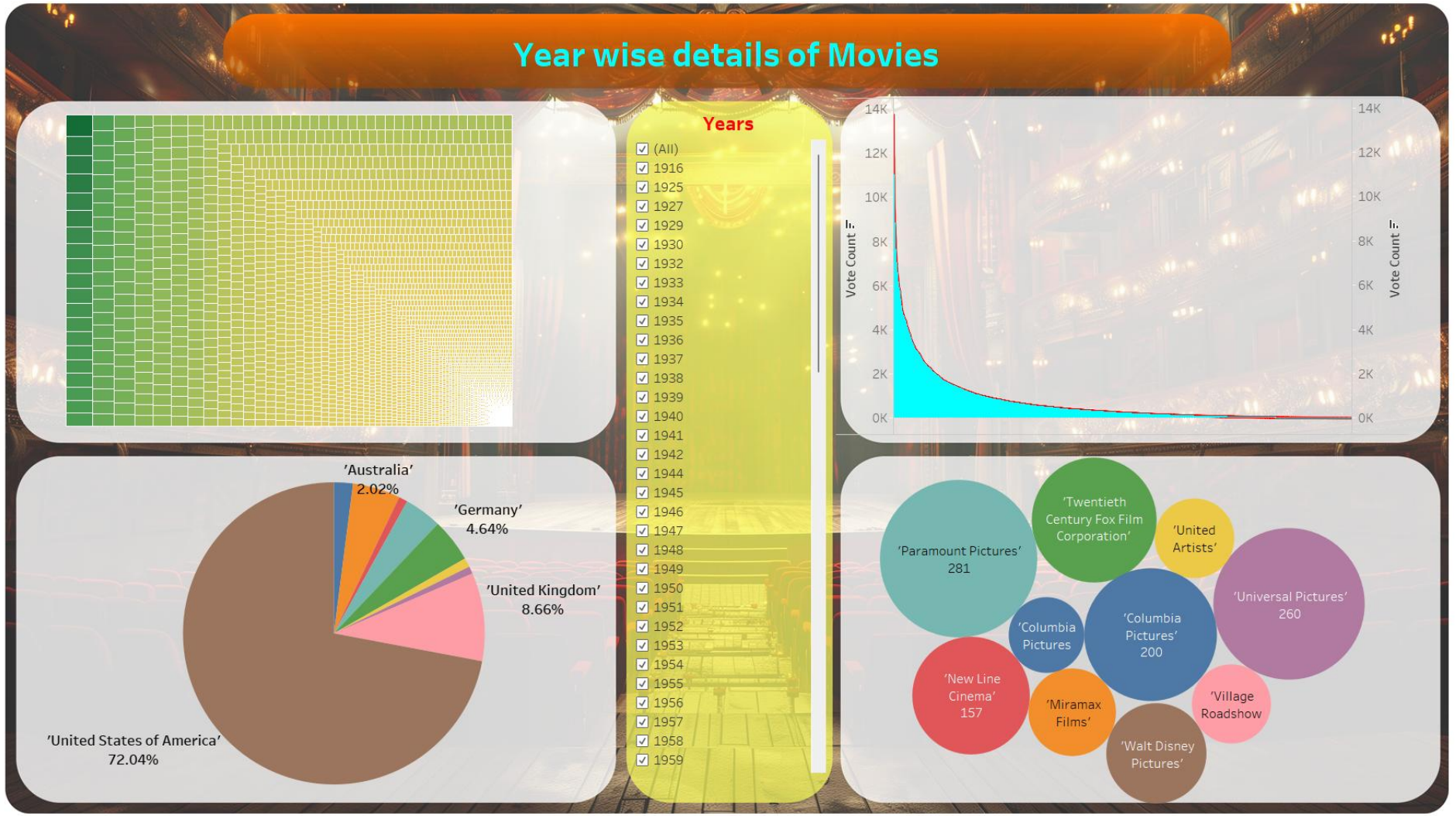
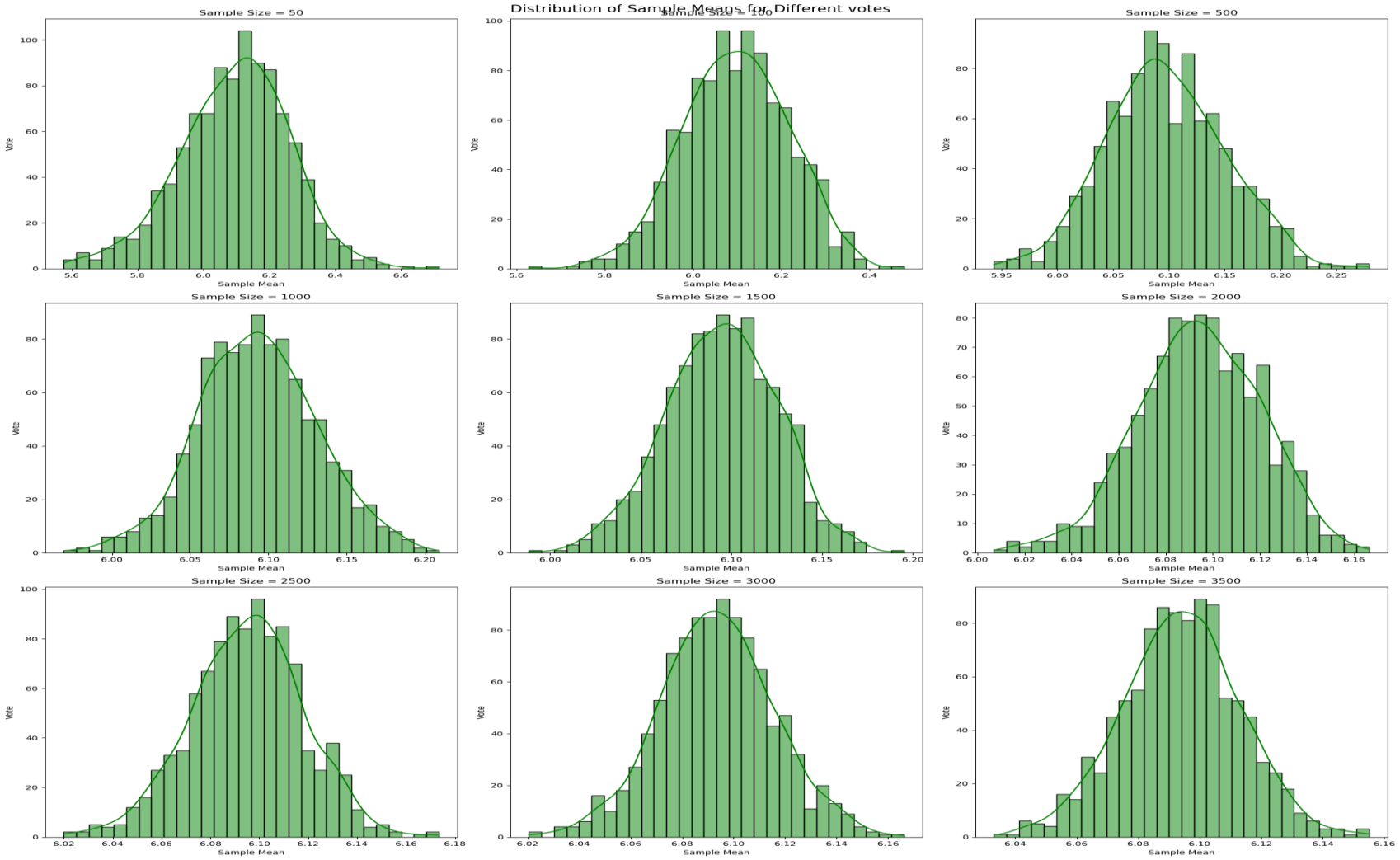


Tableau Public Dashboard

Observations:

- Over the years, there have been very few companies that produced movies, as a result, the numbers of movies made by each company was limited.
- Another major factor in ease of watching movies has been the distribution of movies in other countries which results in increasing the size of the audience and increasing the reach of the movies.

4.2: Validate overall budget and rating assumptions



Python

Observations:

- Based on the sample size taken, it can be observed that the data maintains a normal distribution.
- Based on the box plot in the previous slide, the distribution also follows the pattern of having the median score as ≈ 6

Summary

- The project utilized TMDb datasets to explore movie performance through data preprocessing, statistical analysis, machine learning, and visual reporting.
- A wide range of tools including Python, SQL, Excel, and Tableau were used to extract, clean, merge, visualize, and analyze the data.
- Key problems addressed include movie rating and revenue prediction, genre classification, and clustering of top-performing films using classification, regression, and clustering algorithms.
- Advanced interactive Tableau Dashboards, Excel dashboards and machine learning algorithms helped uncover trends by genre, director, release period, and audience engagement metrics.

Conclusion

- Budget, genre, and popularity emerged as the most significant predictors for both revenue and user ratings.
- Predictive models built using machine learning provided high accuracy in classifying movie ratings and estimating revenue outcomes.
- Clustering and statistical tests revealed strong performance patterns among certain directors, genres, and production strategies.
- The integration of technical tools and business analysis led to actionable insights that can guide film production, marketing, and investment decisions.

Business Implementation

- Invest more in movies with higher budgets and runtime, especially in popular genres like Action and Adventure.
- Focus marketing on high-popularity clusters and high vote-count segments.
- Use cast size and popularity to guide genre-specific promotions.

Business Implementation of the Model

Movie Name	Budget (USD)	Revenue (USD)	Profit (USD)
Avengers: Endgame (2019)	$356,000,000	$2,799,000,000	$1,749,000,000
Avatar: The Way of Water (2022)	$250,000,000	$2,320,250,281	$1,695,250,281
Avengers: Infinity War (2018)	$321,000,000	$2,048,359,754	$1,239,859,754
Spider-Man: No Way Home (2021)	$200,000,000	$1,921,206,586	$1,421,206,586
Top Gun: Maverick (2022)	$170,000,000	$1,495,696,292	$1,069,446,292
Mission: Impossible - Fallout (2018)	$178,000,000	$791,658,205	$345,158,205
Fast & Furious Presents: Hobbs & Shaw (2019)	$200,000,000	$760,732,926	$260,732,926 $
No Time to Die (2021)	$250,000,000	$774,153,007	$149,153,007
John Wick: Chapter 4 (2023)	$100,000,000	$440,146,694	$190,146,694 $
Dune: Part Two (2024)	$120,000,000	$714,711,520	$414,711,520
RRR (2022)	$72,000,000	$160,000,000	$40,000,000

Limitation and Future Scope

- Lack Story telling
- Creating inconsistencies
- Poor Directorial choices
- Improper promotion
- Inappropriate Release timings
- Controversies

THANK YOU