**EXPLORATORY DATA ANALYSIS PROJECT**

**ON**

**IMDB TOP 1000 RATED MOVIES**

In partial fulfilment for the requirements of the award of the degree of

**BACHELOR OF TECHNOLOGY IN**

**COUMPETR SCIENCE AND ENGINEERING**

**What is my Dataset?**

A dataset about IMDb's top 1000 movies would typically contain information about these movies, including:

* Movie Title: The name of the movie.
* Release Year: The year in which the movie was released.
* IMDb Rating: The IMDb user rating for the movie.
* Metascore: The Metacritic score for the movie (if available).
* Director: The director(s) of the movie.
* Cast: Information about the main cast and sometimes the supporting cast.
* Genre: The genre(s) of the movie (e.g., drama, action, comedy).
* Plot Summary: A brief description of the movie's plot.
* Runtime: The duration of the movie in minutes.
* Awards: Any awards or nominations the movie has received.
* Box Office Gross: The total box office earnings of the movie.
* Production Company: The production company responsible for making the movie.
* Country: The country where the movie was produced.
* Language: The primary language(s) spoken in the movie.
* Poster or Image URL: Links to posters or images related to the movie.
* Additional Information: Any additional information that may be relevant to the movie's popularity or significance.

This dataset can be used for various purposes, such as conducting data analysis, building recommendation systems, or exploring trends in movie ratings and genres over time.

If you have a specific question or need assistance with analyzing or working with your IMDb top 1000 movies dataset, please provide more details, and I'll be happy to assist you further.

**Why did I choose this dataset?**

There are several compelling reasons why I might choose a dataset of IMDb's top 1000 movies for various data analysis or research purposes:

1. Popularity and Recognition: IMDb's top 1000 movies represent a selection of highly acclaimed and popular films. Analyzing such a dataset can provide insights into the preferences and tastes of a wide audience.
2. Benchmark for Quality: These movies are often considered benchmarks for cinematic excellence, making them a valuable dataset for evaluating and benchmarking various aspects of filmmaking, including acting, direction, and storytelling.
3. Data Availability: IMDb provides a wealth of information about movies, including user ratings, cast and crew details, release years, and more. This rich dataset is readily available and well-documented, making it convenient for analysis.
4. Research and Analysis: Researchers, film enthusiasts, and data scientists may choose this dataset to conduct various types of analysis, such as sentiment analysis, trend analysis, or exploring relationships between different factors influencing movie success.
5. Recommendation Systems: The IMDb top 1000 movies dataset can be used to build and train recommendation systems. Analyzing user ratings and preferences within this dataset can help create more accurate movie recommendation algorithms.
6. Entertainment Industry Insights: Filmmakers, producers, and studios may analyze this dataset to gain insights into successful movie attributes, which can inform their decision-making processes when creating new content.
7. Educational Purposes: This dataset is often used in educational settings to teach data analysis, data visualization, and machine learning techniques. It provides a real-world dataset for students to work with.
8. Cultural and Historical Analysis: IMDb's top 1000 movies span various decades and genres. Researchers and historians may use this dataset to examine cultural and historical trends in cinema over time.
9. Personal Interest: Film enthusiasts and hobbyists may choose this dataset simply because they have a personal interest in movies and want to explore the data for their own enjoyment or learning.
10. Comparison and Ranking: Researchers and journalists may use this dataset to compare and rank movies based on different criteria, such as IMDb ratings, box office earnings, or critical acclaim.

Overall, the IMDb top 1000 movies dataset is a valuable resource for anyone interested in movies, whether for academic research, industry insights, or personal enjoyment. It offers a diverse range of movies that have garnered significant attention and can be used for a wide array of analytical and research purposes.

**INTRODUCTION:**

The "IMDb Top 1000 Movies Dataset" is a comprehensive collection of data that showcases a curated list of the 1000 highest-rated and most influential movies according to the Internet Movie Database (IMDb).

IMDb is one of the most widely recognized and trusted online sources for movie-related information, ratings, and reviews. This dataset compiles information about these iconic films, offering a rich resource for film enthusiasts, researchers, data analysts, and industry professionals to explore and gain insights into the world of cinema.

KEY FEATURES OF THE DATASET:

The dataset includes an array of features for each movie entry, allowing for in-depth analysis and exploration. Key features typically found in this dataset include:

Movie Title: The title of the movie, identifying each film uniquely.

Release Year: The year in which the movie was originally released, providing historical context.

IMDb Rating: The user-generated IMDb rating, reflecting the audience's perception of the movie's quality.

Meta score: The Metacritic score, offering a critical perspective on the film's reception (if available).

Director: Information about the director(s) responsible for bringing the movie to life.

Cast: Details about the main and supporting cast members who contributed to the film.

Genre: The genre or genres that categorize the movie's style and content.

Plot Summary: A concise description of the movie's storyline, offering an overview of its narrative.

Runtime: The duration of the movie in minutes, providing insight into its pacing.

Awards: Information about any awards or nominations the movie has received.

Box Office Gross: The total earnings of the movie at the box office.

Production Company: The production company or companies behind the film's creation.

Country: The country of origin where the movie was produced. Language: The primary language(s) spoken in the movie.

Poster or Image URL: Links to posters or images related to the movie.

Additional Information: Any other relevant data that enhances the understanding of the movie's significance or impact.

**DOMAIN KNOWLEDGE**:

Domain knowledge about IMDb's top-rated movies, particularly the top 1000, involves an understanding of the film industry, the IMDb platform, and the criteria used to rank and categorize movies. Here's an overview of the key domain knowledge related to IMDb top 1000 rated movies:

IMDb (Internet Movie Database): IMDb is a popular online database of movies, television shows, and celebrities. It provides extensive information about films, including user-generated ratings, cast and crew details, plot summaries, release dates, and more. IMDb's ratings are based on user reviews and are a widely recognized measure of a movie's popularity and quality.

Movie Ratings: IMDb uses a rating system that allows registered users to rate movies on a scale of 1 to 10, with 10 being the highest rating. These user ratings contribute to a movie's overall IMDb rating, which is prominently displayed on the platform.

Meta score: In addition to IMDb ratings, IMDb often includes the meta score for movies. The Meta score is an aggregated score from

professional critics' reviews and provides a different perspective on a movie's quality. It is typically displayed on IMDb alongside user ratings.

Top-Rated Movies: IMDb maintains lists of the top-rated movies based on user ratings. The "IMDb Top 250" is one such list, which includes the highest-rated movies of all time according to IMDb users. The top 1000 is an extended list, showcasing a larger selection of highly rated films.

Criteria for Ranking: IMDb rankings are determined by user ratings, and movies that receive higher ratings tend to be ranked higher. However, IMDb's algorithms may also consider factors like the recency of ratings and user activity when calculating rankings.

Movie Genres: IMDb categorizes movies into various genres, such as action, drama, comedy, science fiction, and more. Understanding movie genres is essential when analyzing the diversity of movies in the top 1000 list.

Film Production: Knowledge of the film industry, including film production companies, directors, actors, and other crew members, can provide context when exploring top-rated movies. Certain directors or actors may have a significant influence on a movie's appeal.

Release Years: IMDb top 1000 movies span a wide range of release years, from classic films to contemporary releases. Understanding the historical context of movies can be important when analyzing trends in movie ratings.

Cultural and Historical Significance: Some movies in the top 1000 list may have achieved recognition not only for their quality but also for their cultural and historical significance. Familiarity with these aspects can deepen the analysis.

Data Exploration and Analysis: Domain knowledge also extends to the methods and tools used to explore and analyze movie datasets. This includes data visualization, statistical analysis, and machine learning techniques for extracting insights from the data.

Overall, domain knowledge about IMDb top-rated movies is crucial for interpreting and making meaningful observations from the dataset.

Whether you're conducting research, building recommendation systems, or exploring trends in the film industry, a solid understanding of these key aspects will enhance your ability to work with IMDb's top 1000 rated movies effectively.

**DATA UNDERSTANDING**:

Data understanding is a crucial step in working with any dataset, including one focused on IMDb's top 1000 rated movies. Here's a systematic approach to understanding the data in such a dataset:

Data Collection: Begin by collecting the IMDb top 1000 rated movies dataset from a reliable source or dataset repository. Ensure that you have access to all the relevant data fields, including movie titles, ratings, release years, and additional information.

Data Description: Understand the structure and content of the dataset by reviewing its metadata. This includes information about the dataset's source, its format (e.g., CSV, JSON, Excel), and the meaning of each column or attribute. Common attributes in this dataset might include:

* Movie Title
* Release Year
* IMDb Rating
* Meta score
* Director
* Cast
* Genre
* Plot Summary
* Runtime
* Awards
* Box Office Gross
* Production Company
* Country
* Language
* Poster or Image URL
* Data Quality Check: Examine the dataset for data quality issues, such as missing values, duplicate records, or outliers. Ensure that the data is clean and reliable. Address any data quality issues through data cleaning processes.
* Data Statistics: Calculate basic statistical measures for numerical attributes (e.g., IMDb Rating, Release Year, Runtime) to understand their distributions. Compute summary statistics like mean, median, standard deviation, and percentiles.
* Data Visualization: Create visualizations to gain insights into the data. Common data visualizations for this dataset might include histograms, box plots, scatter plots, and bar charts to visualize distributions, relationships, and trends.
* Data Exploration: Explore categorical attributes (e.g., Genre, Country, Language) by counting unique values and examining their

distributions. This helps you understand the diversity within these categories.

* Top Rated Movies: Identify the highest-rated movies in the dataset by sorting or filtering based on IMDb ratings. Determine which movies are at the top of the list and consider their characteristics.
* Temporal Analysis: Analyze how movie ratings have changed over time by examining trends in IMDb ratings across different release years.
* Genre Analysis: Explore which movie genres are most prevalent in the top 1000 list. Determine if there are any trends in genre popularity over the years.
* Director and Cast Analysis: Investigate which directors and actors appear most frequently in the dataset and whether their involvement correlates with higher ratings.
* Awards Analysis: Examine the relationship between awards received by movies and their IMDb ratings. Determine if award-winning movies tend to have higher ratings.
* Correlations: Calculate and visualize correlations between numerical attributes (e.g., IMDb Rating, Box Office Gross, Meta score) to identify potential relationships. User Reviews and Comments: If available, consider analyzing user reviews and comments for insights into why certain movies are highly rated. Sentiment analysis can be useful here.
* Data Sampling: If the dataset is large, consider taking random samples for initial exploration to speed up the analysis process.
* Domain-Specific Insights: Leverage domain knowledge about the film industry and IMDb's rating system to interpret findings and draw meaningful conclusions from the data.
* Documentation: Document your findings, observations, and any data transformations or cleaning steps you performed. This documentation will be valuable for future analysis and reporting.

Data understanding is a critical foundation for subsequent steps such as data preparation, modeling, and interpretation. It allows you to gain insights, formulate hypotheses, and make informed decisions when working with the IMDb top 1000 rated movies dataset.

# Libraries used and Approaches:

Pandas:

Pandas is a Python library for data manipulation and analysis, featuring two main data structures, Series and DataFrame, for handling structured data efficiently. It simplifies tasks such as data cleaning, aggregation, and transformation. Pandas is widely used in data exploration and analysis due to its versatility and ease of use.

NumPy:

NumPy, short for "Numerical Python," is a fundamental Python library for numerical and mathematical operations. It introduces a powerful N-dimensional array object, allowing efficient handling of large datasets and mathematical computations. NumPy is the cornerstone of scientific computing and data analysis in Python, providing essential tools for array manipulation and mathematical functions.

Matplotlib is a popular Python library for creating static, animated, and interactive visualizations in various formats. It offers a comprehensive set of tools for generating high-quality plots and charts for data exploration and presentation. Matplotlib is widely used in data science, scientific research, and data visualization due to its flexibility and customization options.

Seaborn:

Seaborn is a Python data visualization library built on top of Matplotlib, designed to create informative and aesthetically pleasing statistical graphics. It simplifies the process of creating complex visualizations with minimal code. Seaborn offers a wide range of plot types, color palettes, and themes, making it a popular choice for data analysts and researchers for data exploration and presentation.

Statsmodels:

This library is particularly useful for statistical modelling. It can be employed for regression analysis, hypothesis testing, and exploring relationships between variables.

Approach to solve the problems:

1. Understanding the Problem Statement: Understanding the problem statement is paramount in any data analysis endeavor. It

involves gaining a deep comprehension of project objectives, the nature of the available data, and the specific insights sought. This initial phase serves as a guiding compass, steering the direction of the entire analysis process. It helps in formulating relevant questions, setting clear goals, and determining the appropriate data sources and methods needed to extract valuable insights from the data.

1. Data Collection: Data collection is a pivotal step in the data analysis process. It entails acquiring and assembling relevant data from various sources, ensuring data quality and consistency. This phase involves selecting appropriate data collection methods, such as surveys, sensors, or databases, to align with the research objectives. Effective data collection forms the bedrock for subsequent analysis, enabling insights, trends, and patterns to be extracted to address the core objectives of the project.
2. Data Cleaning: Data cleaning is a critical stage in data analysis, focused on refining and preparing the collected data for analysis. It involves identifying and addressing issues like missing values, outliers, and inconsistencies. Data cleaning techniques encompass imputation, data transformation, and outlier handling. Integritydata integrity and accuracy during this phase is crucial for reliable and meaningful insights. A clean dataset serves as the foundation for accurate analysis and informed decision-making.
3. Data Preprocessing: Data preprocessing is a vital stage in data analysis, encompassing a series of transformations and enhancements to optimize data for subsequent analysis. This phase involves tasks like feature scaling, normalization, encoding categorical variables, and handling data imbalances. Data

preprocessing aims to improve the quality and compatibility of data, making it suitable for machine learning models and other analytical techniques. By preparing the data effectively, this step contributes to more accurate and meaningful insights.

1. Exploratory Analysis: Exploratory data analysis (EDA) is a foundational step in data analysis, dedicated to gaining a preliminary understanding of the data's characteristics, relationships, and patterns. Using visualizations, summary statistics, and data exploration techniques, EDA uncovers key insights, identifies trends, and highlights potential anomalies or outliers. EDA sets the stage for more in-depth analysis and helps researchers formulate hypotheses and refine their analytical approach, ensuring a robust foundation for decision-making and further investigation.

# DATA CLEANING:

* 1. Filling the null-values:
     + clean is a new DataFrame created by filling missing values (NaN) in the original DataFrame "df" with zeros (0).
     + This step replaces missing values with a default value to make the data more complete.
  2. Selecting specific columns two a new CSV file:
     + Defines a list of column names that you want to focus on in the subsequent steps.
     + These columns are selected for further analysis and export.
     + This block creates a new CSV file named "cleaneddata.csv" containing only the specified columns from the "clean" Data Frame.
  3. Reading data from new CSV file:
     + Reading the "cleaneddata.csv" file into a new Data Frame called "cleandf."
     + This step allows you to work with the cleaned data in a new Data Frame.
     + Displaying the first 5 rows of the "cleandf" Data Frame to examine the cleaned data.
  4. Detecting and marking duplicate rows:
     + Checking for duplicate rows in the "cleandf" Data Frame and returns a Boolean Series indicating whether each row is a duplicate.
  5. Filtering and displaying rows with null-values:
     + Selecting rows in the "cleandf" DataFrame where there are still missing values in any column.
     + This helps identify any remaining rows with missing data in the cleaned dataset.

1. UNIVARIATE ANALYSIS:
   1. Histogram plot:
      * Setting Figure Size using figsize().
      * Creating a histogram using sns.histplot().
      * Labeling Axes using xlabel, ylabel.
      * Adding a title using plt.title().
      * Displaying the plot using plt.show().

Insights from histogram:

The most common runtimes are in the 90-100 minute range, followed by 80-90 minutes. Movies with durations exceeding 200 minutes are infrequent. Outliers include a few films with runtimes surpassing 300 minutes, like "Avatar" (162 minutes) and "The Lord of the Rings: The Return of the King" (201 minutes).

In summary, the majority of movies fall within the 90 – 120 minute range, with outliers being rare.

* 1. Box plot:
     + Creating a figure with a specified size (10x6 inches)
     + Creating a boxplot with 'genre' on the x-axis and 'rating' on the y- axis.
     + Rotate x-axis labels by 45 degrees for better readability.
     + Set a label for the y-axis.
     + Set a title for the plot.
     + Displaying the plot.

Insights from the plot:

The top three genres (drama, crime, and action) all have average ratings above 8.0. This suggests that these genres are generally well- regarded by audiences.

The bottom four genres (comedy, fantasy, war, and thriller) all have average ratings below 7.5. This suggests that these genres may be more polarizing, with some audiences enjoying them greatly and others finding them less appealing.

The middle tier of genres (adventure, sci-fi, animation, romance, westerns, and mysteries) has average ratings between 7.5 and 8.0. This suggests that these genres are generally well-regarded by audiences, but they may not be as popular as the top three genres.

* 1. Pie chart:
     + Setting Figure Size using figsize().
     + Calculating genre counts using [].value\_counts.
     + Exploding slices using explode.
     + Creating a pie chart using plt.pie().
     + Adding a title using plt.title().
     + Equaling the aspect ratio using plt.axis().
     + Displaying the plot using plt.show().

Insights from the pie chart:

The pie chart is divided into different slices, each representing a different genre. The size of each slice is proportional to the percentage of movies in that genre.

Action (82%)

Comedy (56%)

Thriller (50%)

Drama (41%)

Animation (33%)

These genres account for over 80% of all movies released.

* 1. Distribution plot:
     + Setting Figure Size using figsize().
     + Creating a Distribution plot using sns.histplot().
     + Labeling Axes using xlabel, ylabel.
     + Adding a title using plt.title().
     + Displaying the plot using plt.show().

Insights for the Distribution plot:

The graph shows that the most common gross earnings range is between $200K and $600K, with a peak at around $400K. There is a smaller number of people with gross earnings below $200K or above

$600K.

The distribution is not perfectly symmetrical, with more people earning in the lower ranges than in the higher ranges.

1. BIVARIATE ANALYSIS:
   1. Bar plot:

* Grouping by Genre using groupby().
* Sorting by average gross using sort\_values().
* Creating a bar plot of size (10,6) using figsize().
* Labeling Axes and title using xticks, yticks, xlabel, ylabel.
* Displaying the plot using plt.show(). Insights from the bar plot:

The highest-grossing genre is Adventure, followed by Sci-Fi, Fantasy, Action, and Animation. The lowest-grossing genres are Film-Noir, Musical, and War.

It is important to note that this graph is based on average gross revenue, which means that there is a lot of variation within each

genre. For example, some Adventure movies may gross much more than $140 million, while others may gross much less.

* 1. Scatter Plot:
* Setting Figure Size using figsize().
* Creating a scatterplot using sns.scatterplot().
* Labeling Axes using xlabel, ylabel.
* Adding a title using plt.title().
* Adding a legend using plt.legend().
* Displaying the plot using plt.show()

Insights from the scatter plot:

The graph depicts the gross revenue trends of movie genres over the last 20 years, revealing key findings:

Action Reigns:

Action movies consistently lead, grossing over $600 million annually on average, indicating broad and enduring popularity.

Adventure's Surge:

Adventure genre shows a significant rise, averaging over $400 million since 2000 compared to $200 million in the 1990s, signaling a growing trend.

Sci-Fi/Fantasy Boom:

Sci-Fi and Fantasy genres have experienced a substantial increase, averaging over $300 million since 2000, compared to $100 million in the 1990s, highlighting a sustained upward trajectory.

Comedy/Drama Stability:

Comedy and Drama genres maintain consistent popularity, grossing between $200 million and $300 million annually, indicating enduring audience appeal.

Horror's Decline:

Horror movies, once lucrative with over $100 million average gross in the 1990s, have declined to around $50 million since 2000, signaling a waning popularity.

* 1. Violin Plot:
* Setting Figure Size using figsize().
* Creating a violinplot using sns.violinplot().
* Labeling Axes using xlabel, ylabel.
* Rotating x-axes labels using plt.xticks(rotation).
* Displaying the plot using plt.show(). Insights from the violin plot:

The graph displays a correlation between movie ratings and gross earnings. Higher-rated movies, particularly those with ratings of 8.8 or above, tend to earn more money. The top-grossing films are typically in the Adventure, Sci-Fi, or Fantasy genres. However, there's considerable variation in earnings within each rating category. For instance, among movies with a rating of 8.8, gross earnings can vary significantly, with some surpassing $100 million and others achieving less.

1. MULTIVARIATE ANALYSIS:
   1. Heat Map:
      * Setting Figure Size using figsize():
      * Calculating genre counts using [].value\_counts
      * Defining Colors for Categories using cmap
      * Creating a Heat Map using sns.heatmap()
      * Adding a Title using plt.title():
      * Equaling the Aspect Ratio using plt.axis()
      * Displaying the Plot using plt.show()

Insights for the Heat Map:

The heatmap illustrates positive correlations among various movie-related variables: gross revenue (in millions), rating (on a scale of 1 to 10), release year, and runtime (in minutes). The positive correlation implies that an increase in one variable is generally associated with increases in the others. Notably, the strongest correlations exist between gross revenue and rating, as well as between release year and runtime.

* 1. Pair plot:
     + Selecting the specified columns.
     + Filtering the data frame.
     + Setting seaborn runtime configurations for more space.
     + Creating a pair plot using sns.pairplot().
     + Adjusting layout for more space between plot and axes titles.

Insights for the Pair plot:

Gross Revenue and Rating:

Strong positive correlation: Higher-rated movies tend to earn more, indicating increased audience appeal and positive critical reception.

Gross Revenue and Release Year:

Moderate positive correlation: Recent releases yield higher revenue, attributed to improved marketing and distribution for newer films.

Gross Revenue and Runtime:

Weak positive correlation: Longer runtimes are linked to higher revenue, possibly due to the perception of extended movies as "epic" or "event" films.

Rating and Release Year:

Moderate positive correlation: Recent movies receive higher ratings, suggesting advancements in filmmaking and improved production quality.

Rating and Runtime:

Weak positive correlation: Longer movies generally receive higher ratings, indicating that extended durations allow for more in-depth storytelling.

Release Year and Runtime:

Moderate positive correlation: Recent movies have longer runtimes, likely influenced by improved filmmaking technology enabling visually complex films.

1. DISTRIBUTIONS:
   1. Poisson Distribution:

The film industry is experiencing a shift as the average rating of movies rises over time, indicating a more discerning audience that leans toward higher-quality content. While gross revenue remains

stable, blockbuster movies, encompassing genres like superhero and animation, dominate the highest-grossing category with broad appeal and generally high ratings. Notably, success in this category isn't solely reliant on ratings, as strategic marketing can propel even lower-rated blockbusters. Conversely, the lowest-grossing movies often consist of independent films with niche appeal, displaying high ratings but struggling to attract a larger audience due to their specialized focus.

* 1. Exponential Distribution:

In the plot, outliers represent movies either overpredicted or underpredicted by the model, indicating discrepancies between predicted and actual earnings. Overpredicted movies are those the model anticipated would earn more than observed, while underpredicted movies earned more than predicted. Potential reasons for these disparities include the model's omission of crucial success factors like marketing, genre, or release date, or the possibility of overfitting, where the model excessively tailors itself to training data and struggles to generalize to new movies. The residual plot serves as a tool to pinpoint inaccuracies in the model's predictions, offering valuable insights for refinement, either through parameter adjustments or the incorporation of additional features to enhance its predictive capabilities.

* 1. Gamma Distribution:

The gamma distribution demonstrates a good overall fit to the data, closely aligning with the distribution of ratings. This alignment suggests the gamma distribution serves as a suitable model for rating distribution. Interpreting the fit, the mean rating is anticipated to be approximately 3.2, with the median slightly lower. Moreover, there's a 63% probability of a rating falling within the 2.0 to 4.4 range, while a 5% chance exists for ratings below 1.0 or above 5.4. The fitted gamma distribution not only aids in understanding the current distribution but

also proves valuable for future predictions. For instance, in forecasting the rating of a new product, the fitted gamma distribution can be employed to estimate the likelihood of the product receiving a rating of 4 or higher.

* 1. Normal Distribution:

The normal distribution, characterized by a symmetrical bell-shaped curve around the mean, is well-fitted to the ratings data with a mean of

3.0 and a standard deviation of 1.0. This implies that the majority of ratings fall within 1 standard deviation of the mean, indicating an average rating of 3.0. Predictively, the normal distribution fit facilitates forecasts for future ratings, offering insights into the likelihood of a new product receiving a rating of 4 or higher. Interpreting the fit, it's expected that 68% of ratings will fall between 2.0 and 4.0, 95%

between 1.0 and 5.0, and 99.7% between 0.0 and 6.0, providing a comprehensive understanding of the distribution and range of expected ratings.

* 1. Combination of all Distributions:

The normal distribution, characterized by a symmetrical bell-shaped curve around the mean, contrasts with the right-skewed gamma distribution, which has a longer tail, making it more suitable for data exhibiting a rightward skew. In this context, the gamma distribution proves to be a better fit for the data due to its ability to account for high ratings with its extended tail. Contrasting features of the two distributions include the normal distribution's equal mean and median, bell-shaped symmetry, and shorter tails, while the gamma distribution has a mean greater than the median, a rightward skew, and longer tails. The gamma distribution's fit to the data, summarized by key statistics, enables predictions about future ratings, offering insights into the likelihood of a new product receiving a rating of 4 or higher.

1. HYPOTHESIS TESTING:

The normal distribution, characterized by a symmetrical bell-shaped curve around the mean, contrasts with the right-skewed gamma distribution, which has a longer tail, making it more suitable for data exhibiting a rightward skew. In this context, the gamma distribution proves to be a better fit for the data due to its ability to account for high ratings with its extended tail. Contrasting features of the two distributions include the normal distribution's equal mean and median, bell-shaped symmetry, and shorter tails, while the gamma distribution has a mean greater than the median, a rightward skew, and longer tails. The gamma distribution's fit to the data, summarized by key statistics, enables predictions about future ratings, offering insights into the likelihood of a new product receiving a rating of 4 or higher.

1. LIMITATIONS:

Genre Classification Challenges:

Classifying movies into specific genres may be subjective or ambiguous. Some movies may belong to multiple genres, and inconsistencies in genre classification can impact genre-based analyses.

Outliers and Anomalies:

Outliers or anomalies in the data, such as extreme gross revenue values or unusually long runtimes, may disproportionately influence analyses and predictions. Identifying and handling these outliers appropriately is crucial.

Data Collection Bias:

IMDb ratings and gross revenue figures may be biased towards movies that are more popular or well-promoted. This bias can impact the generalizability of findings to less publicized or independent films. Data Quality and Reliability:

The dataset's overall quality and reliability depend on the accuracy of the data sources and the completeness of information. Inaccuracies or

inconsistencies can affect the validity of conclusions drawn from the data.

1. RECOMMENDATIONS:

Genre Insights:

Conduct a detailed genre analysis. Determine the most prevalent genres, analyze their average ratings and gross earnings, and identify any emerging or declining trends in genre popularity over the years.

Directorial Impact:

Investigate the influence of directors on movie success. Identify top directors based on factors like average ratings, gross earnings, or the number of movies in the top 1000. Explore patterns in the performance of movies directed by specific individuals.

Runtime Exploration:

Analyze the distribution of movie runtimes. Identify the average runtime, explore trends in movie lengths over the years, and examine whether there's a correlation between runtime and ratings or gross earnings.

Rating and Revenue Relationship:

Investigate the correlation between movie ratings and gross earnings. Create visualizations, such as scatter plots, to better understand the relationship. Check if higher-rated movies tend to have higher gross earnings.

Outlier Detection:

Identify and explore outliers in the dataset. Examine movies with exceptionally high or low ratings compared to their gross earnings. Investigate the characteristics of these outliers for potential insights.

Correlation Matrix:

Generate a correlation matrix to quantify relationships between different variables. Examine which factors (e.g., release year, genre) are most strongly correlated with ratings and gross earnings.

Budget and Profitability:

If available, consider incorporating data on movie budgets. Analyze the relationship between budget, profitability, and success to understand the cost-effectiveness of filmmaking.

Market Trends:

Investigate broader market trends that might impact movie success, such as changes in consumer behavior, advancements in technology, or shifts in genre preferences.

1. FINDINGS AND INSIGHTS:

The graph shows that the most common gross earnings range is between

$200K and $600K

The distribution is not perfectly symmetrical, with more people earning in the lower ranges than in the higher ranges.

The highest-grossing genre is Adventure, followed by Sci-Fi, Fantasy, Action, and Animation

The lowest-grossing genres are Film-Noir, Musical, and War.

Action movies consistently lead, grossing over $600 million annually on average, indicating broad and enduring popularity.

1. CONCLUSION:

The most common runtimes are in the 90-100 minute range

Outliers include a few films with runtimes surpassing 300 minutes, like "Avatar" (162 minutes) and "The Lord of the Rings: The Return of the King" (201 minutes).

The top three genres (drama, crime, and action) all have average ratings above 8.0.

The bottom four genres (comedy, fantasy, war, and thriller) all have average ratings below 7.5.

The middle tier of genres (adventure, sci-fi, animation, romance, westerns, and mysteries) has average ratings between 7.5 and 8.0.

1. REFERANCES:

Public Repositories:

Websites like Kaggle (https[://ww](http://www.kaggle.com/))w.[kaggle](http://www.kaggle.com/)).[com/)](http://www.kaggle.com/)) often host datasets, including those related to smartphones. You can search for relevant datasets and find the associated references on the platform.

Research Papers:

Some researchers release datasets alongside their publications. If you come across a research paper related to smartphones, check the paper and its references for information about the dataset.

1. ACKNOWLEDGEMENTS:

“GOD HELPS THOSE WHO HELP THEMSELVES.” “ARISE! AWAKE! AND STOP NOT UNTIL THE GOAL IS REACHED.”

Success often requires preparation, hard work, and perspiration. The path to success is a long journey that calls for tremendous effort with many bitter and sweet experiences. This can only be achieved by the Graceful Blessing from the Almighty on everybody. I want to submit everything beneath the feet of God. I want to acknowledge my regards to my teacher, Ms. Shivangini Gupta, for her constant support and guidance throughout my training. I would also like to thank HOD Ms. Harjeet Kaur, School of Computer Science and Engineering for introducing such a great program. I may be failing in my duties if I do not thank my parents for their constant support, suggestion, inspiration and encouragement and best wishes for my success. I am thankful for their supreme sacrifice, eternal benediction, and ocean-like bowls full of love and affection.

PROJECT CODE:

ipynb file: [https://drive.google.com/file/d/1pfsgqEs3rp5eUsZMOVGRXQIEKH-](https://drive.google.com/file/d/1pfsgqEs3rp5eUsZMOVGRXQIEKH-C3h2W/view?usp=drive_link) [C3h2W/view?usp=drive\_link](https://drive.google.com/file/d/1pfsgqEs3rp5eUsZMOVGRXQIEKH-C3h2W/view?usp=drive_link)

Presentation: [https://docs.google.com/presentation/d/1j\_oQdXSwC82\_I43nAY0T1l5](https://docs.google.com/presentation/d/1j_oQdXSwC82_I43nAY0T1l5-VJHk_g3W/edit?usp=drive_link&ouid=103427314487251827147&rtpof=true&sd=true)

[-](https://docs.google.com/presentation/d/1j_oQdXSwC82_I43nAY0T1l5-VJHk_g3W/edit?usp=drive_link&ouid=103427314487251827147&rtpof=true&sd=true)

[VJHk\_g3W/edit?usp=drive\_link&ouid=103427314487251827147&rtp](https://docs.google.com/presentation/d/1j_oQdXSwC82_I43nAY0T1l5-VJHk_g3W/edit?usp=drive_link&ouid=103427314487251827147&rtpof=true&sd=true) [of=true&sd=true](https://docs.google.com/presentation/d/1j_oQdXSwC82_I43nAY0T1l5-VJHk_g3W/edit?usp=drive_link&ouid=103427314487251827147&rtpof=true&sd=true)

Data Set: [https://drive.google.com/file/d/1wp69uyIqIl\_4iwzrEk7hU-](https://drive.google.com/file/d/1wp69uyIqIl_4iwzrEk7hU-BKBf_Tgi8s/view?usp=drive_link) [BKBf\_Tgi8s/view?usp=drive\_link](https://drive.google.com/file/d/1wp69uyIqIl_4iwzrEk7hU-BKBf_Tgi8s/view?usp=drive_link)