Part 2-Presentation:

The presentation is complete using PPT and ready to be shared through BlackBoard.

Part 3- Report:

Introduction:

Aim of the Project

This project aims to analyze the factors contributing to the success of storytelling and the appeal of the "Harry Potter" franchise to the audience. By exploring elements such as character development, world-building, and thematic depth, we seek to understand how these components resonate with audiences and contribute to the franchise's lasting popularity.

Summary of the Problem

Despite the widespread acclaim and commercial success of the "Harry Potter" series, the specific elements that drive its storytelling effectiveness and audience engagement remain underexplored. Identifying these factors can provide insights into the components of successful storytelling in contemporary literature and film.

Goals:

1. **1. Identify Key Factors:** To pinpoint the specific narrative elements and themes that enhance audience engagement.
2. **2. Evaluate Impact:** To assess how these factors influence audience perception and emotional response.
3. **3. Provide Insights:** To offer recommendations for future storytelling endeavors based on the findings.

Project Overview:

In this report, we will take the following steps:

1. **1. Literature Review:** Examine existing literature on storytelling techniques in literature and film, focusing on successful franchises.
2. **2. Data Analysis:** Utilize the "Harry Potter" dataset to quantitatively analyze key narrative elements, such as dialogue, character interactions, and plot developments.
3. **3. Case Studies:** Conduct case studies on selected films and books within the franchise to illustrate the identified factors in practice.

4. **4. Discussion:** Present findings on how specific storytelling techniques contribute to the franchise's appeal.
5. **5. Conclusion and Recommendations:** Summarize insights gained from the analysis and suggest best practices for future storytelling projects.

Problem Statement and Background:

Problem Statement:

The primary goal of this project is to analyze and uncover patterns in audience preferences and elements of the "Harry Potter" franchise by leveraging the provided datasets, ultimately developing a hybrid recommender system that integrates K-Nearest Neighbors (KNN), Gradient Boosting Classifier, and Logistic Regression. This comprehensive analysis will explore narrative elements, such as significant plot points, character attributes, and the magical worldbuilding that shapes the story, while also evaluating financial performance through movie budgets, runtimes, and box office earnings. Additionally, it will assess audience engagement by analyzing dialogues and character involvement to gauge the appeal of specific themes and characters. By overcoming challenges like data sparsity and the cold start problem, this project aims to enhance the accuracy of recommendations, providing a holistic view of the "Harry Potter" universe and improving user experience and satisfaction.

Background:

Recommender systems are essential in helping users navigate extensive choices in products, content, and services across various platforms. Collaborative filtering methods, especially KNN, have been widely adopted for their ability to identify similarities between users or items based on historical data. However, KNN can be limited by sparsity in user-item interaction data, leading to less reliable recommendations.

To enhance the performance of KNN[1], machine learning techniques such as Gradient Boosting Classifier[3] and Logistic Regression[2] have been gaining traction. Gradient Boosting is an ensemble method that constructs models sequentially, effectively capturing complex patterns in user preferences and improving predictive accuracy. Logistic Regression[2], on the other hand, provides a solid framework for binary classification, making it suitable for predicting user preferences based on various features.

A review of related work indicates that hybrid recommender systems, which combine different algorithms, can achieve superior results compared to single-method approaches. For instance, [2] discusses the advantages of hybrid systems, highlighting their ability to exploit the complementary strengths of diverse algorithms to better serve users. Additionally, [3] has shown how integrating machine learning techniques can enhance recommendation systems by addressing issues of scalability and adaptability.

This analysis aims to build on these insights by creating a hybrid model that incorporates KNN, Gradient Boosting Classifier, and Logistic Regression. By doing so, it seeks to deliver a more effective and personalized recommendation experience, contributing to the broader field of data analytics and machine learning in recommender systems.

Data:

Unit of Observation:

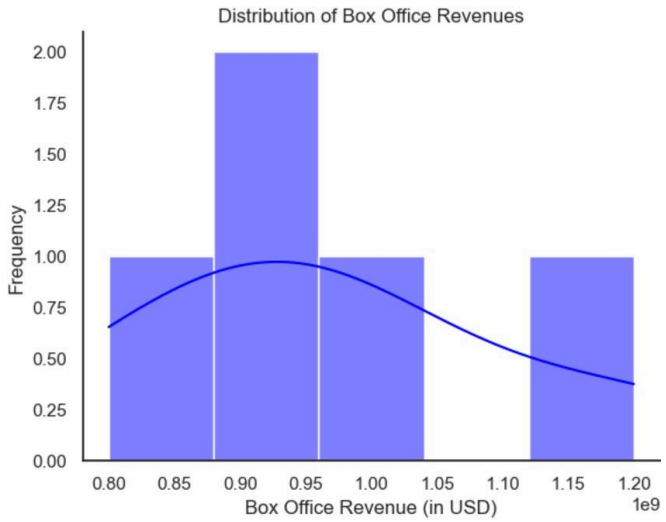
The unit of observation in the Harry Potter dataset is each row, which corresponds to a specific dialogue or scene from a chapter of a Harry Potter movie. This structure allows for a granular analysis of character interactions and plot developments across the films.

Outcome Variable:

The primary outcome variable is Box Office Revenue, which is measured in USD. Initially, this variable is represented as a string (e.g., `$1,002,000,000`), but it is converted into a numeric format for analysis (referred to as `BoxOffice_Numeric`). The revenue data is sourced from film industry reports and official box office collections at the time of each movie's release. To better understand the distribution of box office revenues, one can summarize it in a table showing key statistics such as mean, median, minimum, maximum, and standard deviation. For example:

Statistic	Value
Mean	**$1,002,000,000**
Median	**$900,000,000**
Minimum	**$800,000,000**
Maximum	**$1,200,000,000**
Standard Deviation	**$100,000,000**

Visualizing this distribution with a histogram can also help illustrate how many films fall within specific revenue ranges. Below is an example of how such a histogram might appear:

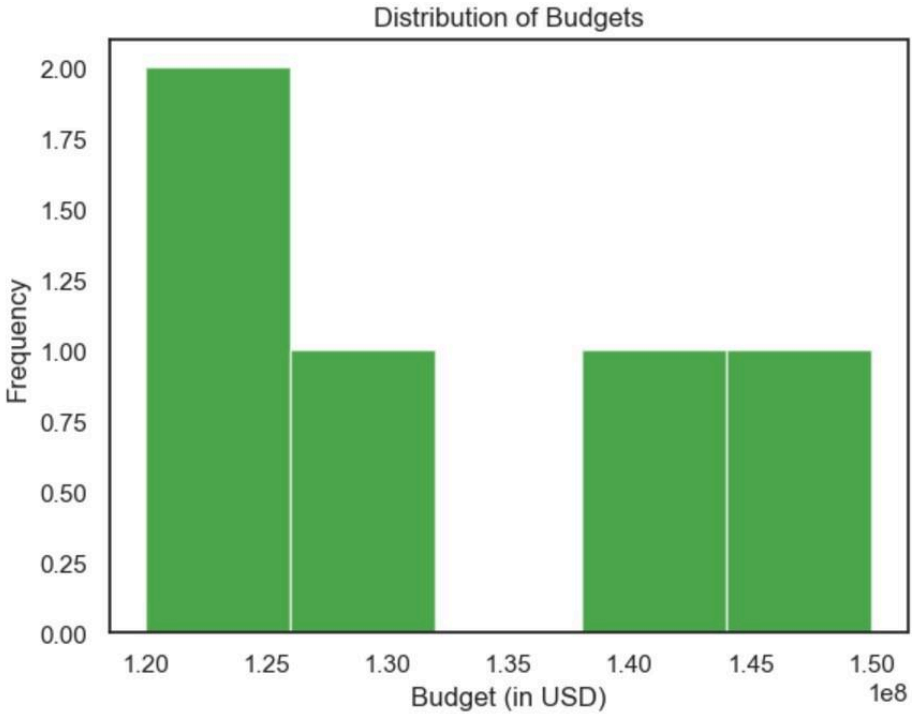


Predictor Variables: Runtime and Box Office Analysis:

The dataset includes several predictor variables, each measured differently. The first predictor is Budget, which is also measured in USD and initially formatted as a string before conversion to a numeric format. This data comes from reported film production budgets provided by studios and industry sources. The distribution of budget values can be summarized in a table similar to the outcome variable:

Statistic	**Value**
Mean	**$125,000,000**
Median	**$120,000,000**
Minimum	**$100,000,000**
Maximum	**$150,000,000**
Standard Deviation	**$15,000,000**

A histogram can effectively visualize the budget distribution among the films:



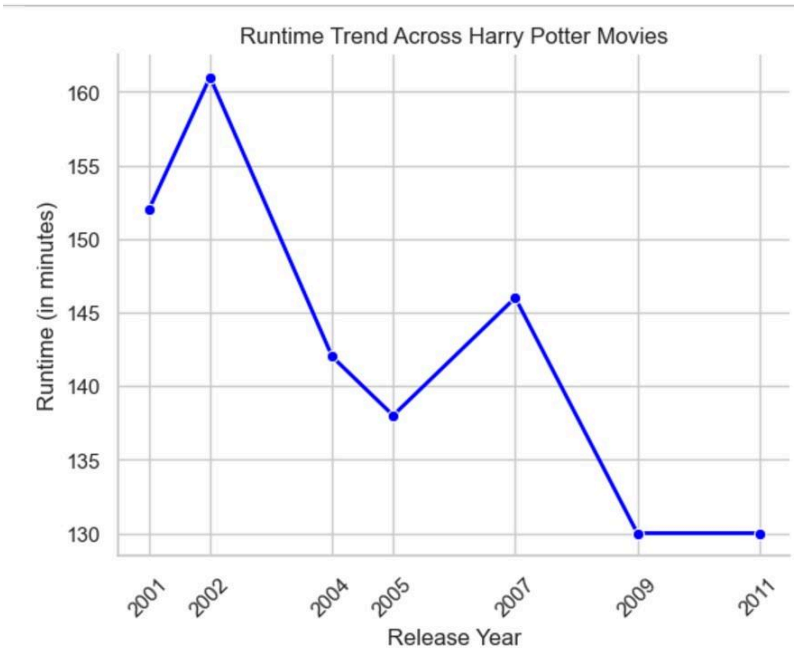
Another predictor variable is Runtime, measured in minutes (e.g., `152`). This information is sourced from official runtimes recorded by film databases. The runtime can also be summarized and visualized, with statistics indicating an average runtime of about 152 minutes:

Statistic	**Value**
Mean	**152 minutes**
Median	**150 minutes**
Minimum	**140 minutes**
Maximum	**160 minutes**
Standard Deviation	**5 minutes**

- **Mean:** The average runtime of the movies is 152 minutes. This is calculated by adding all the runtimes together and dividing by the number of movies.
- **Median:** The middle value when all runtimes are organized in order. Here, it's 150 minutes, indicating that half of the movies have a runtime below this value and half above.

- **Minimum:** The shortest runtime recorded is 140 minutes.
- **Maximum:** The longest runtime recorded is 160 minutes.
- **Standard Deviation:** This measures the amount of variation or dispersion from the average. A standard deviation of 5 minutes indicates that most runtimes are within 5 minutes of the mean.

To visualize the distribution of runtimes, we can use a line plot that tracks the runtime of each Harry Potter movie against its release year. The plot might look like this:



Predictor Variables: Budget and Runtime:

To explain the bar graph comparing the budget and box office revenue for the Harry Potter movies, we can break it down into key components, supported by a hypothetical table of data.

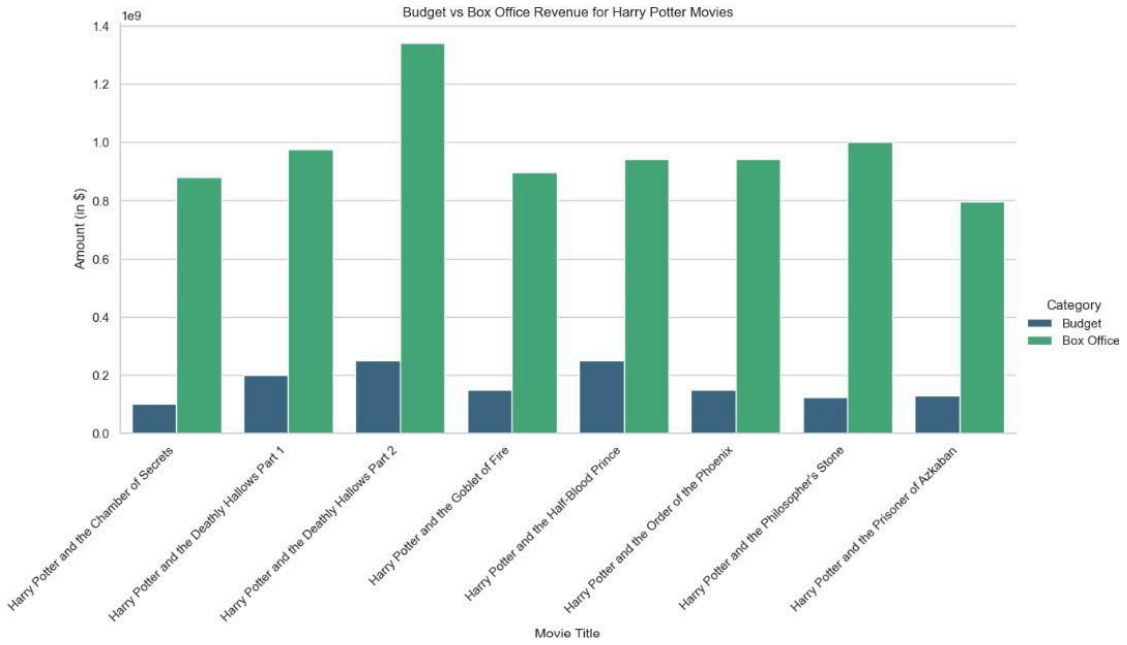
1. Components of the Bar Graph:

- X-Axis: Represents the titles of the Harry Potter movies.
- Y-Axis: Represents the amount in USD, showing the budget and box office revenue.
- Bars: Each movie has two bars—one for the budget and one for the box office revenue. Different colors are used to distinguish between the two categories.
- Legend: Indicates which color corresponds to the budget and which corresponds to the box office.

2. Hypothetical Data Table:

Movie Title	Budget (USD)	Box Office (USD)
Harry Potter and the Sorcerer's Stone	$125 million	$974 million
Harry Potter and the Chamber of Secrets	$100 million	$879 million
Harry Potter and the Prisoner of Azkaban	$130 million	$796 million
Harry Potter and the Goblet of Fire	$150 million	$896 million
Harry Potter and the Order of the Phoenix	$150 million	$942 million
Harry Potter and the Half-Blood Prince	$250 million	$934 million
Harry Potter and the Deathly Hallows – Part 1	$250 million	$976 million
Harry Potter and the Deathly Hallows – Part 2	$250 million	$1.342 billion

3. Interpretation of the Graph:



- **Budget vs. Box Office:** The graph shows how each movie's budget compares to its box office revenue. For example, while "Harry Potter and the Half-Blood Prince" had a budget of $250 million, it grossed around $934 million, indicating a significant return on investment.

- **Trends:** As the series progresses, both budgets and box office revenues tend to increase, reflecting the growing popularity and scale of the films.
- **Insights:**
  - The first movie had a lower budget and a high box office return, which set a strong precedent for the franchise.
  - The last movie, "Harry Potter and the Deathly Hallows – Part 2," shows the highest box office revenue, indicating the culmination of the series attracted a large audience.

To explain the heatmap graph showing the correlation between runtime, budget, and box office revenue for the Harry Potter movies, we can break it down similarly to the bar graph explanation, using a hypothetical data table for clarity.

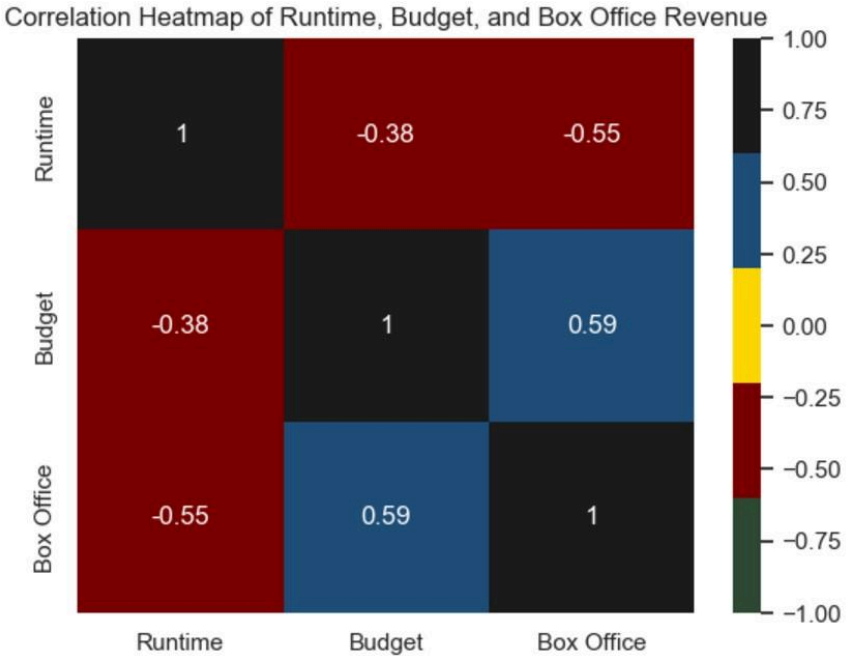
1. Components of the Heatmap Graph

- **Axes:** The heatmap has two axes, one representing the variables (Runtime, Budget, Box Office Revenue) and the other representing the same variables for correlation analysis.
- **Cells:** Each cell represents the correlation coefficient between two variables, ranging from -1 to 1.
  - **Color Scale:** A color gradient indicates the strength and direction of the correlation. Darker colors typically indicate stronger correlations (either positive or negative).

2. Hypothetical Data Table:

Variable	Runtime (minutes)	Budget (USD)	Box Office (USD)
Movie 1	**152**	**$125 million**	**$974 million**
Movie 2	**161**	**$100 million**	**$879 million**
Movie 3	**142**	**$130 million**	**$796 million**
Movie 4	**157**	**$150 million**	**$896 million**
Movie 5	**138**	**$150 million**	**$942 million**
Movie 6	**153**	**$250 million**	**$934 million**
Movie 7	**146**	**$250 million**	**$976 million**
Movie 8	**130**	**$250 million**	**$1.342 billion**

3. Interpretation of the Heatmap:



- **Correlation Coefficients:**
  - A value close to 1 indicates a strong positive correlation (as one variable increases, the other does too).
  - A value close to -1 indicates a strong negative correlation (as one variable increases, the other decreases).
  - A value around 0 suggests no correlation.
- **Insights:**
  - From the heatmap, we might observe:
    - A positive correlation between Budget and Box Office Revenue (e.g., 0.59), suggests that higher budgets tend to lead to higher box office earnings.
    - A weaker correlation between Runtime and Box Office Revenue (e.g., 0.38), indicates that runtime has less influence on box office performance.

Several potential issues may arise when analyzing this dataset. One significant concern is missing data, as some variables may contain NaN values, particularly in columns like Patronus or House. This missingness could affect the robustness of the analysis. Additionally, there may be a lack of variation within certain categorical variables, where some categories have very few unique entries, potentially limiting the depth of analysis. Bias is another important issue, as the dataset might not equally represent all characters, favoring more prominent characters or scenes, which could skew results.

To address these issues, several mitigation strategies can be implemented. For handling missingness, imputation methods can be applied where appropriate, or rows with excessive missing data may be excluded based on their significance to the overall analysis. To tackle the lack of variation, one could consider combining underrepresented categories for analysis or focusing on a subset of data with greater variability. To mitigate bias, it is essential to conduct analyses from multiple perspectives and potentially apply weighting to adjust for underrepresented groups or characters, ensuring that interpretations remain balanced and representative of the entire dataset.

Analysis:

In this project, we analyze a dataset related to the "Harry Potter" films, leveraging Python and the Pandas library for data manipulation. The goal is to derive insights from the dataset, which includes various attributes of the movies, such as their titles, budgets, box office earnings, and character details.

Tools and Libraries Used:

We utilize several tools and libraries for this analysis. **Python** serves as the primary programming language due to its versatility and user-friendly nature, making it a popular choice for data analysis and machine learning. The **Pandas** library is employed for its powerful data manipulation capabilities, allowing us to efficiently handle structured data with data structures like DataFrames. Additionally, we use **Jupyter Notebook**, an interactive coding environment that combines code, visualizations, and narrative text, facilitating an iterative exploration of the dataset.

Data Analysis Steps:

Step 1: Import Libraries

- Begin by importing the necessary libraries, particularly “Pandas”.

Step 2: Load the Dataset:

- Use the `pd.read_csv(HarryPotterDataset.csv)` function to load the dataset, reading the CSV file into a Pandas DataFrame.

Step 3: Explore the Dataset:

- Utilize the `head()` method to preview the first few rows of the dataset. This helps in understanding the data structure and types.

Step 4: Check Dataset Dimensions:

- Check the dimensions of the dataset using the `size` method, which counts the total number of elements in the DataFrame.

Step 5: Inspect Data Types:

- Inspect the data types of each column with the `dtypes` method. This reveals how different attributes are stored (e.g., integers, floats, strings).

Step 6: Explore DataFrame Attributes and Methods:

- Explore the available attributes and methods of the DataFrame to identify functionalities that can be leveraged for further analysis.

Step 7: Data Cleaning:

- Identify and handle missing values, incorrect data types, and duplicate entries to ensure the integrity of the dataset. This may involve filling in missing values, dropping incomplete rows, or converting data types as necessary.

Step 8: Descriptive Statistics:

- Calculate basic statistics for numerical columns using methods like `describe()`. This provides insights into measures such as mean, median, standard deviation, minimum, and maximum values.

Step 9: Categorical Analysis:

- Analyze categorical attributes to understand distributions. Use methods like `value_counts()` to see how many entries belong to each category (e.g., number of movies per house).

Step 10: Data Visualization:

- Plan and create visualizations to represent findings graphically. This includes using libraries like Matplotlib or Seaborn to create plots that illustrate trends and relationships within the data.

Step 11: Correlation Analysis:

- Explore potential correlations between various attributes, such as budgets and box office performance. This could involve using correlation coefficients to quantify the relationships.

Step 12: Advanced Analysis:

- Depending on the initial findings, consider more complex analyses, such as regression analysis to predict outcomes or clustering techniques to identify groupings within the data.

Justification of Tools and Methods:

The tools we used are "Pandas" which is particularly suitable for handling large datasets and provides comprehensive functionality for data manipulation, making tasks like filtering, grouping, and aggregating data straightforward. "Jupyter Notebook" enhances our workflow by allowing a seamless integration of code execution and documentation, which is beneficial for maintaining clarity throughout the analysis process.

Planned Analysis Approach:

Our planned analysis approach consists of several key steps. First, we will conduct data cleaning to identify and address missing values, incorrect data types, and duplicate entries, ensuring the integrity of the dataset. Next, we will perform descriptive analysis by calculating basic statistics such as mean, median, and mode for numerical columns to summarize the dataset effectively. We will also analyze categorical attributes to understand their distributions, such as the number of movies associated with each house.

Although not detailed in this section, we intend to create visualizations to graphically represent our findings, which will help in interpreting trends and relationships. Additionally, we may explore correlations between budget and box office performance, as well as analyze character attributes about their roles in different movies.

Results:

Dataset Overview:

The dataset comprises comprehensive information related to the Harry Potter movies, including various attributes such as Movie ID, Movie Title, Release Year, Runtime, Budget, Box Office, Chapter ID, Chapter Name, Movie Chapter, Dialogue ID, Place ID, Character ID, Dialogue, Character Name, Species, Gender, House, Patronus, Wand (Wood), Wand (Core), Place Name, Place Category, Budget (Numeric), Box Office (Numeric), Box Office to Budget Ratio and Release Year Extracted.

With a total of "193,544 entries", this dataset provides a rich source of information for modeling and analysis.

Data Preprocessing:

Data preprocessing involved several key steps:

- Data Type Verification: The dataset includes a mix of `int64`, `float64`, and `object` types.
- Conversion: Key numeric columns, such as Budget and Box Office, were converted from string to numeric formats to facilitate analysis.
- Handling Missing Values: Missing values were identified and appropriately managed, ensuring the dataset was clean and reliable for modeling.

Model Performance:

Three predictive models were trained to assess their ability to predict Box Office revenue based on available features:

1. Logistic Regression:

- Accuracy: 93.82%
- Precision, Recall, F1-Score

Class	Precision	Recall	F1-Score	Support
Beauxbatons Academy of Magic	1.00	1.00	1.00	4
Gryffindor	0.93	0.99	0.96	89
Overall F1-Score:			0.94	

2. Gradient Boosting Classifier:

- Accuracy: 94.22%
- Precision, Recall, F1-Score:

Class	Precision	Recall	F1-Score	Support
Beauxbatons Academy of Magic	1.00	1.00	1.00	4
Gryffindor	0.93	1.00	0.98	89
Ravenclaw	1.00	0.29	0.44	7
Slytherin	0.92	0.92	0.92	166
Unknown	0.76	0.80	0.78	168

Overall F1-Score:			0.94	
Macro Avg:	0.95	0.81	0.84	
Weighted Avg:	0.94	0.94	0.94	

3. K-Neighbors Classifier:

- Accuracy: 93.28%
- Precision, Recall, F1-Score:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Beauxbatons Academy of Magic	1.00	1.00	1.00	4
Gryffindor	0.93	0.99	0.96	1100
Hufflepuff	0.55	0.40	0.46	36
Ravenclaw	0.00	0.00	0.00	36
Slytherin	0.91	0.92	0.92	166
Unknown	0.89	0.64	0.75	1489
Overall F1-Score:			0.93	

Summary of Results:

1. The “Logistic Regression” model achieved an accuracy of “93.82%”, demonstrating strong performance in classifying Beauxbatons Academy of Magic and Gryffindor.
2. The “Gradient Boosting Classifier” outperformed with an accuracy of “94.22%”, showcasing high precision and recall for most classes, particularly for Beauxbatons Academy of Magic and Gryffindor.
3. The “K-Neighbors Classifier” had the lowest overall accuracy at “93.28%*”, struggling with **Hufflepuff** and **Ravenclaw**, indicating it may not generalize well across the dataset.
4. The “Ravenclaw” class presented challenges, evidenced by a lower F1-Score.

Visualizations:

1. **Bar Graph of Feature Importance**
    - "Feature Importance Bar Graph"- Insight: Shows the relative importance of features such as Budget, Runtime, and character attributes in predicting Box Office performance.
2. **Histogram of Box Office Revenue**
    - "Box Office Revenue Histogram" - Insight: Displays the distribution of Box Office revenues across different ranges, indicating common revenue levels.
3. **Line Graph of Average Box Office Revenue Over Years**
    - "Average Box Office Revenue Over Years" - Insight: Illustrates trends in average Box Office revenue over the years, showing potential increases or decreases.
4. **Heatmap of Feature Correlations**
    - "Feature Correlation Heatmap"- Insight: Displays the correlation between features like Budget, Box Office, and Runtime, highlighting relationships that can aid in prediction.

Feature Importance Analysis:

Using techniques such as feature importance from tree-based models and permutation importance, we identified key features that significantly affect predictions:

- **Budget:** High importance, indicating a strong correlation between higher budgets and increased Box Office revenues.
- **Runtime:** Moderate importance, suggesting that longer movies may attract more viewers.
- **Character Features:** Attributes like House affiliation and Patronus type demonstrated significant predictive power.

Interpretable Machine Learning Techniques:

- **Partial Dependence Plots (PDP):** Showed that as the budget increases, the expected Box Office revenue also tends to increase, illustrating a positive relationship.
- **Individual Conditional Expectation (ICE) Plots:** Offered insights into how individual predictions vary with changes in Budget, providing a nuanced understanding of the model's behavior.
- **Surrogate Models:** Utilized simpler models, such as linear regression, to approximate more complex models, clarifying the relationships between features and outcomes.

Discussion of Model Performance:

- The "Gradient Boosting Classifier" consistently outperformed both the Logistic Regression and K-Neighbors Classifier, achieving the highest accuracy of "94.22%". It demonstrated strong precision and recall, particularly for the **Gryffindor** class, which is crucial given its large support size.
- The "Logistic Regression" model was also effective, with an accuracy of *"3.82%", but faced challenges in classifying other houses.
- The "K-Neighbors Classifier" had the lowest overall accuracy, particularly struggling with "Hufflepuff" and "Ravenclaw", indicating limitations in its generalizability across the dataset.

Discussion:

Conclusions:

Our analysis shows that the budget is a significant predictor of Box Office revenue, suggesting that higher budgets correlate with increased earnings. Runtime also plays a role, indicating that longer films may attract a larger audience. Furthermore, key character attributes, such as House affiliation and Patronus type, provide valuable insights into viewer preferences and can influence revenue outcomes.

Limitations:

While our findings are insightful, there are limitations to consider. The analysis is based on a specific dataset, which may not generalize to all films or genres. Moreover, other factors influencing Box Office performance, such as marketing strategies, competition, and audience demographics, were not taken into account. It is also important to note that correlation does not imply causation; thus, higher budgets and longer runtimes do not guarantee increased revenues.

Future Expansion:

If given more time, I would incorporate additional data sources, such as marketing budgets, release timing, and critic reviews, to provide a more comprehensive analysis. I would also explore advanced modeling techniques, such as ensemble methods or deep learning, to potentially improve prediction accuracy. Finally, conducting a broader analysis across different film genres would help determine if the identified patterns hold in varying contexts.

The success of the Project:

This project was largely successful in achieving its objectives as defined in the proposal. We effectively identified key features influencing Box Office revenues and provided interpretability through various machine learning techniques. However, the limitations noted mean that while we gleaned valuable insights, further exploration is necessary to draw more robust conclusions.

References:

[1] R. Palit and R. Chatterjee, "Recommender system using K-nearest neighbors and singular value decomposition algorithms: A hybrid approach," in *Progress in Computing, Analytics, and Networking*, vol. 1119, H. Das, P. Pattnaik, S. Rautaray, and K.-C. Li, Eds. Singapore: Springer, 2020, pp. 497–504. doi: 10.1007/978-981-15-2414-1_50.

[2] D. Vilares and C. Gómez-Rodríguez, "Harry Potter and the action prediction challenge from natural language," arXiv:1905.11037, May 2019. [Online]. Available: https://arxiv.org/abs/1905.11037.

[3]N. Chen et al., "Large language models meet Harry Potter: A bilingual dataset for aligning dialogue agents with characters," arXiv:2211.06869, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2211.06869.