كليـــة الهنـــدســـة
College of Engineering

جامعة عفت
EFFAT UNIVERSITY

Computer Science Department
Fall 24
Data Science

Data Science Project

Student's Name	ID	Section
Raghad Alamoudi	**S21107187**	**2**
Ehadaa AlMarhabi	**S19206078**	**2**
Manar Alharbi	**S21107392**	**1**

I. **Project Name:** Analysis of Viewer Preferences in the Harry Potter Franchise.

II. **Description:**

The project is based on datasets capturing various aspects of the "Harry Potter" franchise. The datasets collectively provide detailed information about spells, places, movies, characters, and chapters, as well as a data dictionary explaining the fields used. Here is a breakdown of the datasets:

- **Places Dataset** (74 entries, 3 columns): Lists iconic places in the franchise, categorized by location (e.g., Diagon Alley).
- **Movies Dataset** (8 entries, 6 columns): Details on the Harry Potter films, including titles, release years, runtimes, budgets, and box office earnings.
- **Characters Dataset** (166 entries, 8 columns): Provides data on characters such as their names, species, gender, Hogwarts house, patronizes, and wand details.
- **Chapters Dataset** (234 entries, 4 columns): Break down the films into individual chapters, linking them to the corresponding movie.
- **Data Dictionary** (31 entries, 3 columns): Explains the fields used across the datasets, aiding in the interpretation of the information.
- **Dialogue Dataset** (7,444 entries, 5 columns): Captures dialogue lines spoken by characters in specific chapters and places, providing context for key moments in the films.

Each dataset provides unique insights into different aspects of the Harry Potter world. This project aims to analyze these datasets for patterns and insights into audience preferences, focusing on the popularity of spells, places, and characters, as well as the financial and runtime metrics of the movies.

III. The Goal:

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. Specifically, the project aims to:

1. **Narrative Analysis:** Explore dialogues, chapters, and locations to identify significant plot points and the distribution of key events across the films.
2. **Character Insights:** Analyze character attributes, such as patronesses, compositions, and Hogwarts houses, to uncover patterns and unique traits.
3. **Magic and Worldbuilding:** Investigate the spells and iconic places in the "Harry Potter" universe to understand their roles in shaping the story's magical essence.
4. **Financial Performance:** Evaluate movie budgets, runtimes, and box office earnings to determine the financial success and trends across the franchise.
5. **Audience Engagement:** Use dialogues and character involvement to gauge the focus and appeal of specific characters or themes.
6. **Comprehensive Understanding:** Integrate insights from all datasets to present a holistic view of the franchise's storytelling and its reception by audiences.

This analysis aims to provide a comprehensive understanding of the "Harry Potter" universe, blending narrative elements with audience metrics to offer a holistic view of its success and legacy.

IV. Dataset Reference:

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Chapters.csv

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Characters.csv

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Data_Dictionary.csv

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Dialogue.csv

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Movies.csv

https://github.com/Srijita2002/Harry_Potter-Dataset-Analysis/blob/main/Places.csv

Part 1 - Proposal:

1. High-Level Statement of the Problem:

The "Harry Potter" franchise has become a cultural phenomenon, captivating audiences globally through its intricate storytelling, diverse characters, and magical elements. However, while the franchise has been analyzed extensively for its narrative and commercial success, there is limited quantitative analysis integrating multiple data dimensions, such as character dynamics, spell usage, and financial metrics, to understand what drives audience engagement and franchise success.

Research Question

What factors within the "Harry Potter" franchise contribute most significantly to its storytelling success and audience appeal?

Existing Research

Previous analyses have focused on literary critiques, fanbase studies, and box office trends. For instance, some researchers have studied how J.K. Rowling's narrative techniques fostered character development, while others have examined the financial success of the movies. However, there is still a lack of comprehensive quantitative research that integrates character, dialogue, and magical elements to explore storytelling success holistically.

2. Outcome Variable:

The primary outcome variable is audience engagement, operationalized as the box office earnings of each movie. This variable reflects how well each movie performed commercially, a critical indicator of audience appeal.

Description of the Outcome Variable

- Conceptual Relation to the Research Question: Box office earnings encapsulate audience interest and engagement with the franchise, directly linking to the success of storytelling, character dynamics, and worldbuilding.
- **Summary Statistics:**
  - Mean Box Office: $927,275,000
  - Median Box Office: $919,200,000
  - Minimum Box Office: $796,700,000
  - Maximum Box Office: $1,002,000,000

3. Predictor Variables: To model the outcome variable, I will use a set of predictors drawn from the datasets provided. These predictors span multiple dimensions of the franchise:

1. **Movie Runtime (Movies Dataset):** Reflects the depth of storytelling per film.
2. **Budget (Movies Dataset):** Indicates production scale.
3. **Number of Dialogues (Dialogue Dataset): Captures the narrative density of each movie.**
4. **Character Diversity (Characters Dataset):** Number of unique characters appearing in each movie.
5. **Character Gender Ratio (Characters Dataset):** Proportion of male to female characters.
6. **Number of Spells Used (Spells Dataset):** Represents the intensity of magical elements.
7. **Number of Unique Places Featured (Places Dataset):** Indicates the richness of the worldbuilding.
8. **Critical Plot Events per Chapter (Chapters Dataset):** Derived by counting significant chapters per movie.
9. **Dialogue Sentiment Analysis (Dialogue Dataset):** Aggregate sentiment score for dialogues in each movie.
10. **Hogwarts House Representation (Characters Dataset):** Number of key characters from each Hogwarts house.

Data Sources: All data will be derived from the provided datasets. Summary statistics and aggregation will be used to calculate some predictors (e.g., dialogue count, sentiment analysis, and house representation).

4. Definition of "Success":

Success within this project will be defined as the ability to extract actionable insights from the integrated datasets of the "Harry Potter" franchise, focusing on storytelling, character dynamics, and audience engagement. A successful project would include:

1. **Data Analysis and Insights:** Perform an in-depth analysis of the datasets to uncover key trends and patterns in the franchise. This includes identifying factors such as:
   - Which characters, spells, and locations are most impactful in shaping audience engagement?
   - How runtime, budget, and other movie-specific factors correlate with box office performance.
   - Relationships between dialogue sentiment and character development.
     Statistical analyses, machine learning models, and clustering techniques will be employed to uncover these patterns.
2. **Visualizations and Reporting:** Create compelling visualizations to communicate findings effectively. These might include:
   - Bar charts showing character diversity and representation by house.

- Heat maps of spell usage across movies or chapters.
- Regression plots showcasing relationships between predictors like runtime and box office earnings.

  A detailed, well-structured report will be prepared summarizing the insights, actionable recommendations, and implications for the franchise.

3. **Predictive Modelling:** Develop a predictive model to estimate box office performance based on factors such as dialogue density, character diversity, and magical elements. A model with high predictive accuracy (e.g., $R^2 > 0.80$) will demonstrate success in capturing the underlying relationships within the franchise data.
4. **Knowledge Contribution:** Prepare the findings for dissemination in relevant academic or industry contexts, such as conferences or journals in the fields of media studies, storytelling, or data science. The project will aim to contribute valuable insights to ongoing discussions about the impact of narrative elements on audience engagement.
5. **Practical Applications**

   Provide actionable recommendations for franchise stakeholders, including:
   - Optimizing character development and representation to appeal to broader audiences.
   - Balancing storytelling depth (e.g., runtime, chapter richness) with engagement.
   - Leveraging popular spells and iconic locations in future movies or marketing strategies.

     These insights will guide content creation and marketing strategies to better align with audience preferences, enhancing viewer satisfaction and franchise growth.

By achieving these goals, this project will demonstrate a high level of proficiency in data analysis, provide meaningful contributions to the understanding of narrative-driven engagement, and offer actionable insights for the continued success of the "Harry Potter" franchise.