Cross Platform Movie Recommendation System

Milestone 1 Progress Report

Project Objective

This project aims to develop a cross-platform recommendation application that helps users discover new movies and TV shows across different streaming services. If a user enjoys a particular show or movie on one platform, the app will suggest similar content available on another platform. To make this possible, the application will incorporate a recommendation model that analyzes user preferences and content similarities

Alongside the recommendation engine, the project will feature a simple, user-friendly interface that allows users to input their favorite content, select their preferred streaming services, and receive personalized recommendations. The goal is to make the app visually appealing, easy to navigate, and intuitive for all users.

Data Sources

To ensure a comprehensive analysis, three primary datasets have been sourced and integrated into the system:

1. **Netflix Titles** (netflix_titles.csv) - Contains 8,807 records (Click Here to Access)
2. **Amazon Prime Titles** (amazon_prime_titles.csv) - Contains 9,668 records (Click Here to Access)
3. **Hulu Titles** (hulu_titles.csv) - Contains 3,073 records (Click Here to Access)

All datasets share an identical structure, including key attributes such as show_id, type (Movie/TV Show), director, cast, country, date added, release year, rating, duration, listed genres, and description. This uniformity ensures efficient data merging and analysis.

Libraries and Frameworks Used

To build the content analysis tool, a structured technical stack has been adopted, ensuring scalability and efficiency in processing and visualization.

- **Primary Programming Language**: Python
- **Data Processing**: Pandas and NumPy
- **Visualization**: Matplotlib and Seaborn

Data Preprocessing Implementation

Data Cleaning Functions:

- Standardization of date formats to ensure uniformity across platforms
- Handling of missing values using appropriate imputation techniques
- Normalization of content types (Movies vs. TV Shows) for consistent classification
- Standardization of rating categories to align across platforms
- Cleaning and unifying country names to facilitate geographical analysis
- Normalization of genre listings to ensure uniform analysis

Data Integration:

- Merging datasets with designated platform identifiers
- Standardizing column formats across platforms to facilitate seamless comparison

Exploratory Data Analysis (EDA)

1. **Basic Platform Statistics**:
   - Size of content libraries for each platform
   - Distribution of Movies vs. TV Shows across platforms
   - Trends in release years for available content
   - Rating distributions across different streaming services
2. **Content Analysis**:
   - Genre distribution breakdown to identify platform-specific trends
   - Age analysis of available content based on release years
   - Country of origin breakdown to assess international diversity
   - Examination of rating classifications and their distribution
   - Content exclusivity analysis to determine the proportion of unique titles per platform
3. **Visualization Implementation**:
   - Overview of platform content distribution through detailed bar charts and pie charts
   - Representation of content type (Movie vs. TV Show) distributions
   - Genre-specific trends and patterns using categorical plots
   - Age-wise distribution analysis using histograms
   - Rating breakdown using comparative visualizations
   - Venn diagrams illustrating content overlap among platforms
   - Geographical distribution maps highlighting content origins

Key Findings

1. **Library Sizes**:
   - Amazon Prime has the largest catalog with **9,668 titles**
   - Netflix follows with **8,807 titles**
   - Hulu has the smallest library with **3,073 titles**
2. **Content Type Distribution**:
   - Netflix: **69.62% movies, 30.38% TV shows**
   - Amazon Prime: **80.82% movies, 19.18% TV shows**
   - Hulu: **48.29% movies, 51.71% TV shows** (largest share of TV shows among the three)
3. **Average Content Age**:
   - Netflix: **10.82 years**
   - Amazon Prime: **16.66 years** (older content compared to other platforms)
   - Hulu: **12.43 years**
4. **Content Exclusivity**:
   - Netflix: **93.44% exclusive content**
   - Amazon Prime: **94.12% exclusive content**
   - Hulu: **86.66% exclusive content**

Visualizations

Library Size and Content Type

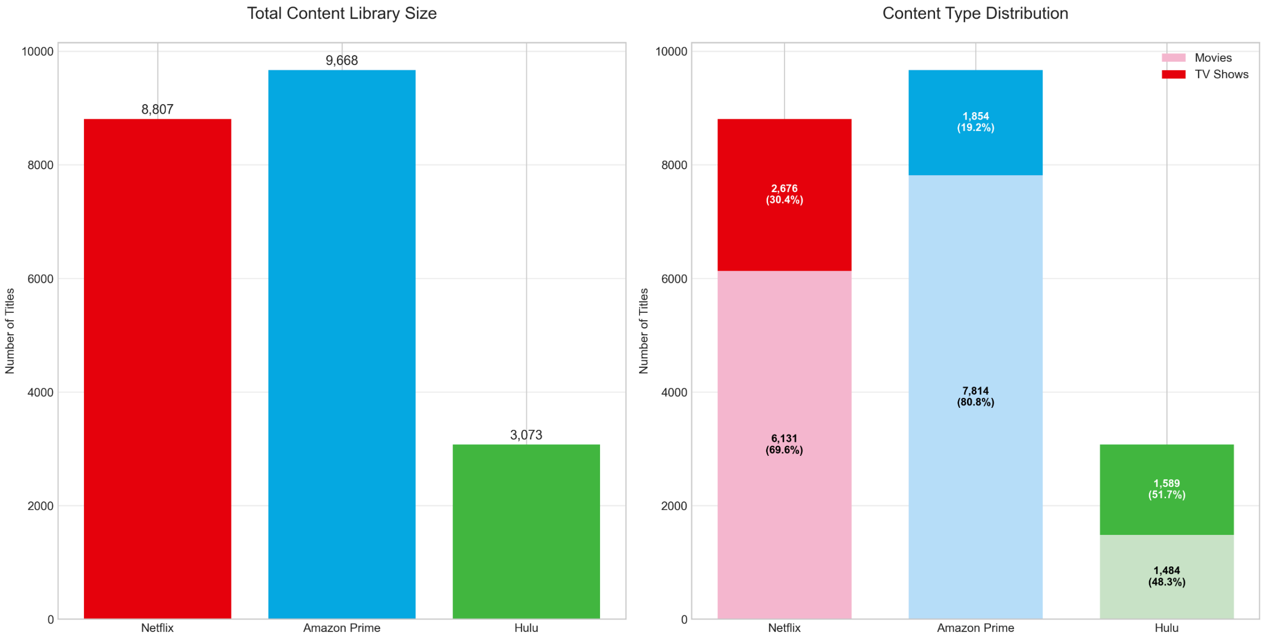
- Amazon Prime has the largest content library with 9,668 titles, making it the most extensive platform. However, much of this content consists of older titles.
- Netflix follows with 8,807 titles, offering a well-curated mix of TV shows and movies while maintaining exclusivity.
- Hulu has a significantly smaller library, with 3,073 titles, but compensates with a strong focus on TV shows.

Content Type Breakdown:

- Amazon Prime is primarily movie-focused, with 80.8% movies and 19.2% TV shows.
- Hulu is the only one where TV shows outnumber movies, with 51.7% TV shows and 48.3% movies.
- Netflix maintains a more balanced ratio but still favors movies, with 69.6% movies and 30.4% TV shows.

Strategic Implications:

- Amazon Prime prioritizes quantity, acquiring a vast number of movies, including many older titles.
- Hulu focuses on TV shows, positioning itself as the primary choice for serialized content.
- Netflix offers a more balanced mix, making it adaptable to various audience preferences.



Exclusivity and Overlap

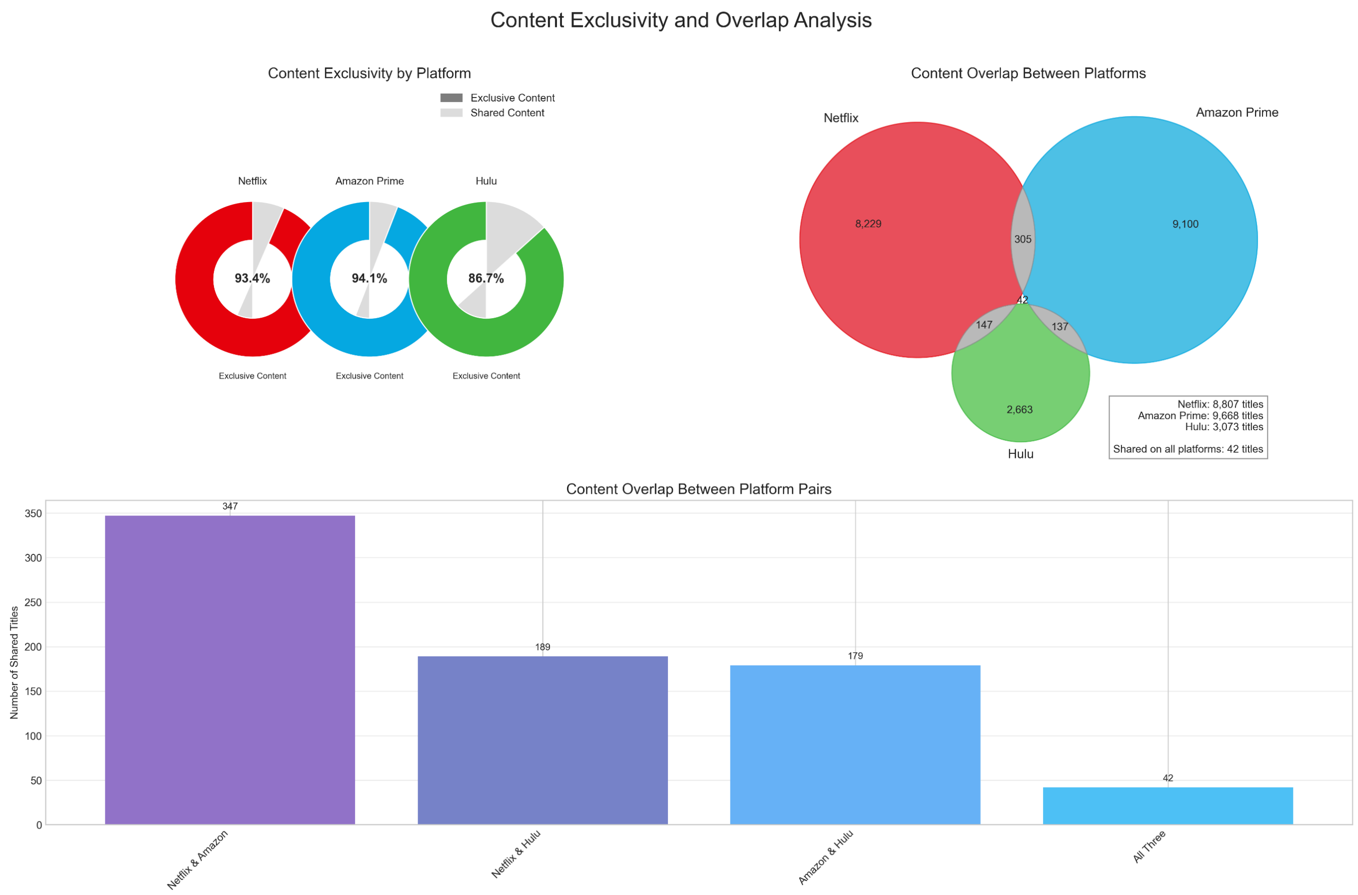
- Netflix: 93.4% exclusive content
- Amazon Prime: 94.1% exclusive content
- Hulu: 86.7% exclusive content

There is minimal overlap between platforms:

- Only 42 titles are shared across all three platforms.
- Netflix and Amazon Prime share the most content, with 347 overlapping titles.
- Other platform pairs share relatively few titles, ranging from 147 to 305 titles.

Strategic Implications:

- Each platform has developed a strong identity by prioritizing exclusive content.
- Consumers who want a more diverse selection of content are required to subscribe to multiple services.
- Netflix and Amazon Prime have some shared content, but the overlap remains low overall.

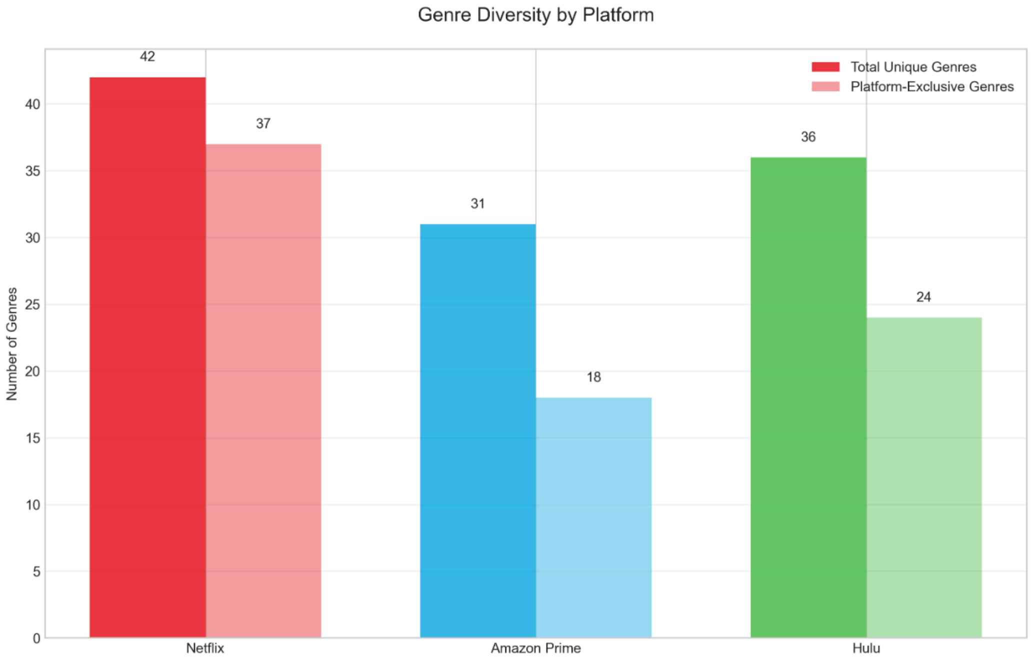


Genre Distribution

- Drama is the most common genre, with 7,021 titles across all platforms.
- Netflix has the highest genre diversity, featuring 42 unique genres.
- Netflix has the most platform-exclusive genres, with 37 genres not found on Amazon Prime or Hulu.
- Comedy and Action consistently rank among the top genres across all platforms.

Strategic Implications:

- Netflix's broad genre coverage allows it to cater to a wide range of audiences.
- Amazon Prime and Hulu focus on a smaller selection of genres, aligning with their content acquisition strategies.

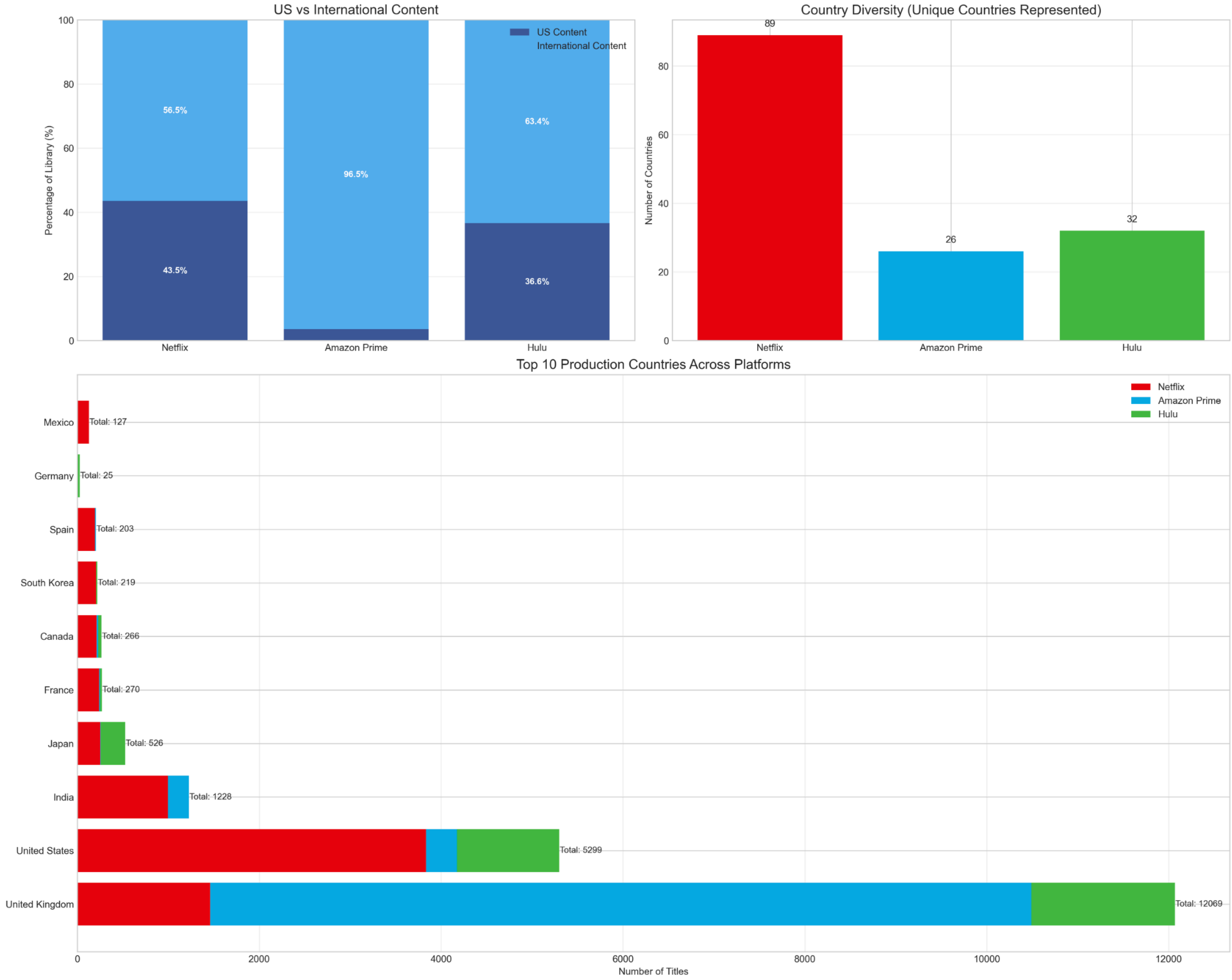


Content Origin

- Netflix features content from the highest number of countries, with 89 countries represented.
- The United Kingdom is the largest source of content overall.
- Amazon Prime has the highest percentage of international content, with 96.5% of titles originating outside the United States.
- Netflix has a more balanced US vs. international split, with 43.5% US-based content and 56.5% international content.

Strategic Implications:

- Netflix's global content strategy allows it to appeal to international audiences.
- Amazon Prime's heavy international focus suggests a strong investment in foreign markets.
- Hulu, with a lower percentage of international content, is primarily focused on US audiences.

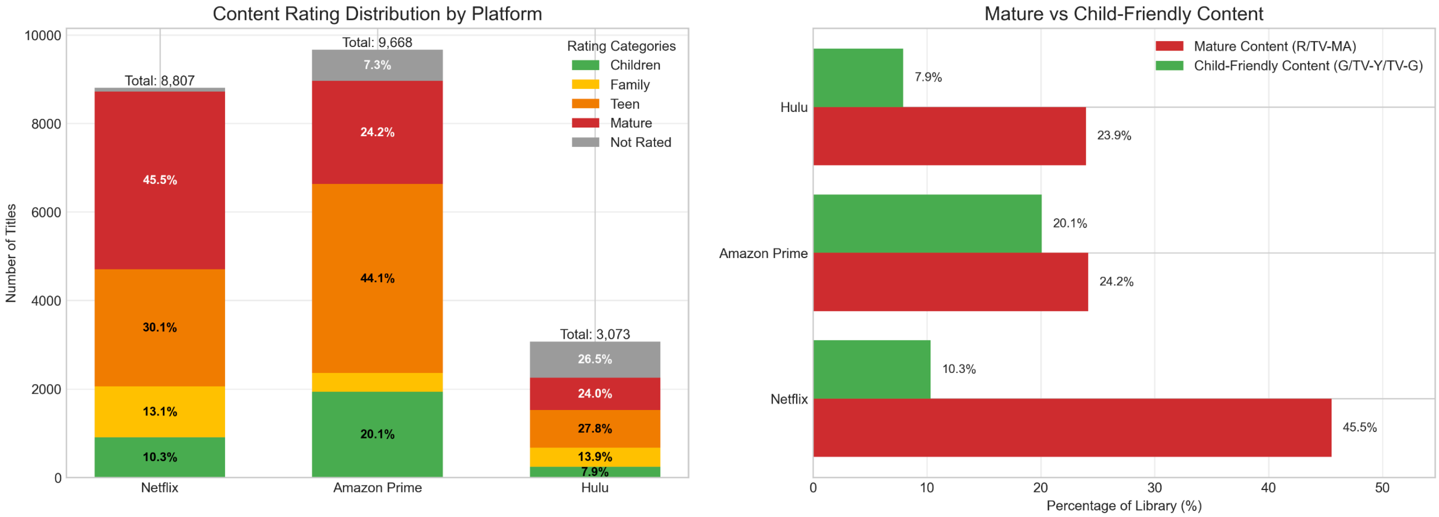


Content Ratings

- Netflix has the highest proportion of mature content, with 45.5% of titles rated R or TV-MA.
- Amazon Prime has the largest share of child-friendly content, with 20.1% of titles rated for young audiences.
- All platforms feature a significant amount of teen-oriented content (TV-14/PG-13).
- Hulu has the highest proportion of unrated content, at 26.5%.

Strategic Implications:

- Netflix’s catalog skews toward mature audiences, which aligns with its investment in original dramas and thrillers.
- Amazon Prime caters more to family-friendly and international content.
- Hulu offers a mix but includes a larger proportion of unrated titles.



Project Plan: Feature Engineering | Modeling | Tool Development

Phase	Dates	Time	Tasks
Milestone 2	Feb 21 - Mar 21	4 weeks	
Feature Engineering and Selection	Feb 21 - Mar 3	10 days	**Feature Engineering:** • Genre Coding • Text Features from Descriptions (using TF-IDF or word embeddings to extract key themes for analyzing similarity) • Duration Features (Normalize and categorize content based on runtime) **Feature Selection:** • Feature Importance Analysis (Use statistical techniques such as mutual information, chi-square tests, and feature importance scores from tree-based models to rank features.) • Correlation Analysis (Identify redundant or highly correlated features using Pearson/Spearman correlation to prevent overfitting.) • Check and see if dimensionality reduction is required • Validation of the selected features to see if they are effective

Phase	Dates	Time	Tasks
Model Development	Mar 4 - Mar 21	18 days	**Hybrid Model Building:** • **Content-Based Filtering Model** (Use the actual content features (genres, descriptions, cast, release year, ratings, duration) from streaming datasets and mplements TF-IDF vectorization for text features and cosine similarity to find similar content) • **Matrix Factorization (SVD)** (Decompose the user-item interaction matrix into lower-dimensional user and item matrices and will need to simulate user interactions initially since we don't have real user data) • **Neural Collaborative Filtering** (Use a multi-layer perceptron to learn non-linear relationships which can capture complex patterns in user-item interactions)
Milestone 3	Mar 24 - Apr 23	4 weeks	
Model Evaluation and Tool Development	Mar 24 - Apr 6	14 days	**Model Evaluation:** • Conduct comprehensive model testing to ensure accuracy and reliability. • Perform error analysis to identify misclassifications and improvement areas. • Test edge cases to evaluate model robustness. **Tool Development:** • Develop a front-end for users to interact with recommendations • Display personalized recommendations and content insights • Allow users to find content available across multiple streaming services • Integrate graphs and charts to explain recommendations (maybe)
Tool Enhancement	Apr 7 - Apr 18	12 days	• Allow filtering by genre, content type, release year, and user preferences. • Provide transparent reasoning behind recommendations. • Improve system efficiency and response times.
Final Documentation and Presentation Prep	Apr 19 - Apr 23	5 days	Create presentation summarizing key findings and implementation details and compile final report detailing all aspects of development, evaluation, and results.

LLM Prompt Given For Project Help:

Initially, I first just manually analyzed my datasets by putting them in excel and performing some basic excel commands. Then I started implementing the python code using the documentation for the specific libraries and the following prompts:

1. How do I standardize the rating categories across different platforms?
2. How do I handle encoding errors when reading the CSVs?
3. The date formats are different in each dataset. How can I make them consistent?

4. How should I handle duplicate titles that appear on diff platforms?
5. The duration column has both 'min' and 'Seasons' - what's the best way to separate and standardize these?
6. Some descriptions contain HTML tags - how can I clean these?
7. Some titles have missing countries should I impute based on language or other features?
8. Some dates seem invalid how can I identify and fix these?
9. How do I make sure that when handling outliers in release year that outliers are properly handled?
10. Some genres are very similar should I combine them?
11. The analyze_countries function is slow with large datasets - how can I optimize it?
12. Can you show me how to handle the case where a platform might not have both 'Movie' and 'TV Show' types when calculating the movie_to_tv_ratio to avoid errors?
13. What's the best way to handle non-English characters in the text fields?
14. What's a more efficient way to find unique genres by platform, potentially using set operations?
15. How can I use regular expressions to make the **standardize_rating** function more flexible and less prone to errors due to slight variations in rating strings?
16. Is there anyway I can generate a markdown file directly to show my analysis metrics? If yes, give me the name of the documentation for it
17. How do I create a basic 2x2 grid using GridSpec?
18. How do I make one visualization take up multiple grid cells in GridSpec?
19. How do I adjust the spacing between subplots in GridSpec?
20. How do I add a main title that spans all subplots in GridSpec?
21. How can I overlay two bar charts (one for average, one for median) for each platform?
22. What's the best way to display the numerical value on top of each bar in a bar chart?
23. Can you show me how to set the y-axis label to 'Age in Years' and format the y-axis ticks to be easily readable?
24. How can I create individual pie charts with a hole in the center (donut charts)?
25. How do I display percentage values directly on the pie chart slices?
26. How do I make sure each chart is displayed next to each other with a label specifying the streaming platform, and content percentage?
27. How can I create a grouped bar chart where each platform has a bar for each age range?
28. How do I rotate the x-axis labels to prevent them from overlapping?
29. How do I modify the x axis to be more readable.
30. How can I create a donut chart where I specifically control the radius and width of the ring?
31. How do I position the percentage label (e.g., 93.4%) precisely in the center of each donut chart?
32. How do I add text labels, such as 'Exclusive Content', below each donut chart, ensuring they are properly aligned?
33. How can I draw a Venn diagram with circles representing Netflix, Amazon Prime, and Hulu, and accurately size the overlapping regions based on the number of shared titles?
34. How do I ensure that the areas of overlap in my Venn diagram are proportional to the number of elements they represent?
35. I am unsure how to scale the y axis correctly, how should I approach this?

Also when creating the documentation, I have typed up the information I need and gave that information to ChatGPT and told it to ‘Make it more professional’. I then took some of the important points from it which sound better than what I have written. (I did this just for the sake of making the report more professionally legible)

Name: Durga Sritha Dongla

UFID: 54220803