Completed Tasks:

- Primary Dataset Exploration and Cleaning:
  - Analyzed key features from the primary dataset (movies and credits).
  - Dropped irrelevant columns like homepage, tagline, overview, etc., based on their lack of relevance.
  - Removed movies not in English to focus on English-origin films.

- Feature Engineering:
  - Extracted and encoded relevant features like genres, production_companies, production_countries, and release_date.
  - Conducted one-hot encoding for categorical columns like parsed_genres and release_season.
  - Created binary indicators for production companies (Top 7 vs. Others).

  - Added new features like:
    - num_production_countries – Number of countries involved in production.
    - num_spoken_languages – Diversity in spoken languages.
    - release_year, release_season, and related one-hot-encoded columns.

- Outlier Handling:
  - Analyzed key features (budget, revenue, popularity, etc.) for outliers and skewness.
  - Decided to retain most outliers as they represent meaningful data points for blockbuster movies.
  - Applied log transformation for skewed on vote_count.

- Talk Show Data:
  As agreed, the talk show data is not relevant due to its newer nature and mismatched time periods. Removed the talk show data from the analysis to avoid bias and explained it in the file.

- Revenue and Seasonality:
  - Explored the impact of release seasons on revenue.
    - Observed that Summer has the highest average revenue, followed by Spring and Winter.
    - Created one-hot-encoded features for release_season and dropped redundant columns.

Pending Tasks:

- Crew and Cast Analysis: [I've explained in detail at the end of the notebook in detail]
    - Explore columns like main_actor_names, directors, writers, and producers.

- Interesting investigations maybe:
    - The proportion of male and female actors in the cast and its correlation with revenue.
    - The effect of cast size (crew_size) on revenue.
    - Actor popularity scores based on historical movie revenues.

- Final Cleanup:
    - Drop columns like id and movie_id after completing analysis.
    - One final look at all features to ensure that they are relevant
    - Correlation Analysis
        - If relevant correlations between some key features and the target variable (revenue) to refine feature selection.

If found interesting we could also do some trend Analysis:

- Use release_year to explore trends in movie revenue over time, such as whether revenue growth is consistent.

Model Preparation, and Training.

- I've also Split the data into training and testing as suggested by Richard. Finalize feature selection and ensure all categorical and numerical columns are ready for model input.