DA5401 Assignment #1

Visualization: Uncovering the Story Behind Movie Success

This is a more challenging assignment designed to push your skills beyond basic plotting. You will not only create visualizations but also use them to tell a compelling story about a complex dataset.

Assignment Objective

Your goal is to become a data detective. You will use a movie dataset to formulate and test a hypothesis about what drives a movie's success. Your final deliverable should be a narrative supported by a series of sophisticated visualizations.

Dataset

You will be using a comprehensive movie dataset, such as the **TMDB 5000 Movie Dataset** (publicly available on Kaggle). This dataset is larger and more complex, containing information on budget, revenue, a list of genres, directors, cast, and user ratings. You will need to perform some data cleaning and transformation before you can visualize it.

Tasks

Part 1: Data Preprocessing and Hypothesis [10 points]

1. **Formulate a Hypothesis:** Before you start, propose a hypothesis about movie success. For example: *"Higher budgets directly correlate with higher IMDb ratings and revenue,"* or *"Movies with a larger number of lead actors tend to have higher ratings."*
2. **Data Cleaning:** The dataset is not perfectly clean. You must handle:
    - **Missing Values:** Identify and address missing values in key columns, such as `budget`, `revenue`, and `runtime`. Decide whether to drop rows or fill missing values, and justify your choice.
    - **Data Transformation:** The `genres` and `cast` columns contain JSON strings. You must extract and transform this data into a usable format by creating new columns (e.g., `primary_genre`, `number_of_actors`).

Part 2: Advanced Visualizations [10 points]

Create the following visualizations to test your hypothesis. Each plot should be highly customized following the **seven** commandments of plotting.

1. **Distribution of Success:** Use seaborn to create a **pair plot** illustrating the relationships between key numerical variables: budget, revenue, runtime, and vote average. This will give you a high-level overview of their correlations.
2. **Time-Series Analysis:**
   - Visualize the trend of movie releases over time using a **line plot**.
   - Create a **stacked area chart** using matplotlib to show the change in the number of movies released for the top 5 genres over the years. This requires you first to process the genre data.
3. **Revenue & Ratings Breakdown:**
   - Create a **violin plot** using seaborn to compare the distribution of vote_average for the top 5 directors with the highest number of movies.
   - Create a **bar chart** that compares the average revenue and budget for each of the top 10 genres. Use matplotlib to plot both on the same axes for a direct comparison.
4. **Correlation Matrix:** Generate a comprehensive **heatmap** of the correlation matrix for all numerical variables. This should be visually appealing and help you identify hidden relationships.
5. **Your creativity**: Besides the above, can you devise other nuanced visualizations to tell an engaging story? **[+5 Brownie points]**

Part 3: The Data Story [30 points]

Your final deliverable is not just a collection of plots. You must create a single Jupyter Notebook that acts as a complete data story.

- Start with your **hypothesis**.
- Walk through your **data preprocessing steps**, showing how you prepared the data.
- Present your visualizations in a logical order, each with a brief text block explaining what it shows and how it relates to your hypothesis.
- Conclude with a clear, concise summary of your findings. Did your visualizations prove or disprove your hypothesis? What unexpected insights did you discover?

This assignment challenges you to think critically about data, clean it effectively, visualize complex relationships, and communicate your findings in a structured and compelling manner.

Good luck!

Seven Commandments

1. Color palette should be inclusive of the color-challenged audience.
2. Markers should stay relevant irrespective of the medium (printed, browser, PDF, etc.).
3. Axes should be named appropriately.
4. Scale (lin/log) and units of the axes should be explicitly specified.
5. A legend should be present when more than one variable is visualized.
6. A brief title (and sub-title) for the plot should be provided.
7. A short description of the plot's story to make it self-explanatory.

Remember, your plots should speak for themselves even in your absence!