Predictive Analysis of Revenue of Movies from Various Features

Research Questions for Movies:

1. Which part of the world creates the highest-grossing movies relative to budget?
   - Through this question, we aim to find out how movies' revenue in the box office is impacted by the location they are in. We are using a ratio of the gross revenue to the budget so that we can account for inflation throughout the years. This is a valuable piece of information because knowing where movies are most likely to make a profit based on how much they spent making it can help target the movie to an audience based in that location. Using the information we just learned regarding geopandas and plotting, we can visualize this data very easily, making it an interesting research question for us to look into.
2. Which part of the world produces movies with the highest rating-to-budget ratio?
   - For this question, we are trying to compute where the most cost-effective movies are produced. This analysis can be used to determine how production companies in parts of the world that are lacking in resources can still create high-quality content, and how bigger production companies can reduce the amount of money they spend on making movies.
3. How do genres tend to correlate to ratings of their movies? Which genres tend to have the highest number of viewer votes?
   - Based on movie genres, we plan to parse through our data and figure out which genres tend to produce higher IMDb scores and viewer ratings. This is useful as it shows trends of how critics react to specific genres of movies. Knowing what genres end up having the most acclaim among the public audience and not just the critic's reviews are useful because it can help find what kinds of movies are perceived as the best. If we can find another data set that goes more in-depth regarding the intricacies of genres, we will utilize those features to further assist with answering these questions.
4. How do aspects of a movie's production impact the gross revenue?
   - Using machine learning models, we plan to use features such as the movie budget, genre, release date, and director to predict how well the movie will perform in the box office. We are using a ratio of the gross revenue to the budget as our indicator of box office success so that we can account for inflation throughout the years. This is an important feature because companies want to continually gain profit from their movies and knowing what features impact the movie's revenue can help them achieve this. If possible, we will continue to search for data sets that contain information that could extend what we are currently looking for by providing even more features that could help with our predictions.
5. Over the years, what are the trends of the highest-grossing genres based on viewer ratings?
   - Based on critics and general public ratings, we will find out what types of movies performed the best over the years. We can either do this by year or by every few years and will decide this after looking into our data set

more closely. This is useful as companies can make more informed decisions on what type of movies they want to produce. Similarly to the previous question, finding data sets that elaborate more on the specifics of genres can help us find more detailed answers to this question, so that will be one thing we will look further into.

Motivation and background:

All of our questions are targeted toward movie producers and marketing companies. We want to make it so that they create movies that result in the biggest benefit for both themselves and their target audiences. Therefore, we try to maximize those features that impact their profit and reviews. We believe that being able to find the answers to these questions will help these companies become more efficient with their time by focusing on those movies that will be most relevant, instead of wasting their time with movies that will not be successful.

Datasets:

https://www.kaggle.com/danielgrijalvas/movies

This dataset contains various features including movie budget, production company, country of origin, genre, revenue, release date, run time, reviews (from IMDb and viewer votes), starring actor/actress, and main writer. It covers 6820 movies from 1986 to 2016 (220 movies per year) and the data is scraped from IMDb by the person who posted the dataset.

https://datahub.io/core/gdp

This dataset contains every country with their GDP from 2005 to 2019. This is a complementary data set that we will use in our first question to show alongside the highest grossing to budget ratios based on countries.

Methodology/Analysis:

GeoPandas World Map Dataset

There is no link to this dataset because it is loaded directly from the geopandas library. It contains various features that are used to plot a map of the world by country. The most important column is the shape column which is used for plotting.

For our first research question, we will focus on data visualization. By finding how the various countries or regions fare in terms of how much money they make with each movie, we can use GeoPandas to draw this out. Our data set has latitude and longitude data that we can combine into a coherent location column. By coloring each country based on how much money they make through all of the movies that are produced there each year, we can provide a detailed visualization that effectively represents this. One secondary visualization that would be useful to pair with this could be a graph of our regions and how much GDP they have had each year. Seeing the resulting data can help us understand what the common factors between the countries with highest-grossing movies are, helping us understand more about what results in revenue for these areas.

For our second research question, we will focus on data visualization again. Similarly to the first research question, we can use GeoPandas to represent what countries or regions have the highest rating-to-budget ratio. By coloring each country based on this value, we can provide a detailed visualization that effectively represents this. An interesting component to analyze alongside this could be creating a second visualization that plots how this number has changed for each country or region every year. Seeing the resulting data can help us understand what proportion of low-budget movies end up receiving positive reviews, and it can help justify if spending lots of money on a movie is worth it or not. The data visualization will also highlight which countries/regions have the highest rating to budget ratio.

For our third research question, we will create a linear regression model to predict what the critics' ratings will be for certain genres. We will pair this with data visualization of how genres are related to viewer votes through a bar graph of genre types vs average viewer ratings. The first step we will take is to cut out the features that are not highly correlated to the variable we want to predict (the ratings by critics) followed by cutting out features that are too highly correlated with each other. Using these features, we can predict the rating for a certain genre. The data visualization would use two features, the average viewer rating per genre on the y-axis and genre on the x-axis. If our model predicts high ratings and low ratings for certain genres, we can conclude that these genres are generally rated the highest by critics. Then we can compare this with our other data visualization to see how critic's ratings and viewer ratings differ or agree.

For our fourth research question, we will use a linear regression algorithm by training a decision tree regressor to predict future movies' box office based on features such as movie budget, genre, release date, and director. The label would be the 'gross revenue' and the features would be columns in our data that are not too distinct to avoid overfitting. We would use the same technique as the previous model when it comes to removing un-correlated and highly correlated features to select which features we will use in prediction. For instance, we would first cut out the movie name, as it is too distinct, and then remove features that have too high of a correlation with other predicting features. This is an interesting model as it allows companies to understand which factors of a movie will allow them to get maximum profit. If we see that certain movies' of a certain genre or by a certain company will not be grossing much money, then we can see that there are deciding factors for how a movie will do in the box office. If we cannot see consistent predictions, then we will not know if there are common features among successful box office movies.

For our fifth research question, we will create a data visualization to help show how the viewer ratings for the highest-grossing genres compare to the lowest-grossing genres on a year-by-year basis. First, we will filter our data set to find the genre with the highest average viewer rating and the genre with the lowest average viewer rating for every year. We will then plot this on a line chart with the years on the x-axis, viewer ratings on the y-axis, and the highest ratings and lowest ratings as two different colored lines. If we see that one genre is repeated in one of the categories throughout the years, it will lead us to believe that the said genre is one of the most/least popular genres among viewers and therefore does/does not bring in a lot of money. If we see that the genres throughout the years do not have many repetitions, we can see that the favorite and least favorite genres change frequently.