Capstone Project Proposal

In my capstone project, I want to build a model that can predict the final North American box office performance of a film taking into account several factors including budget, genre, opening weekend gross, aggregated critics as well as audience ratings. The model will only take into account North American box office numbers, many films often do not release simultaneously or at all in certain regions often leading to the inability of box office analysts to fairly compare a film's box office performance to another. There is a larger pool of data for the North American box office remaining largely consistent, hence I will favour analysis of North American gross instead of worldwide numbers. While the opening weekend will be used to analyze the performance of the film it will not be predicted by this model because a film's opening is largely dependent on factors such as cost of marketing which many studios do not release data for.

This project will gather and combine data from many sources to have a number of different factors to make accurate and precise predictions. The box office numbers for final and opening weekend domestic gross will be taken from boxofficemojo.com, the industry leader for box office data. The MoviesLens dataset containing user reviews from over a quarter of a million different users will be combined with OMDB which provides the critics' ratings of films from sites such as Rotten Tomatoes and Metacritic. The OMDB data also includes IMDB ratings from its users which will be used alongside the user reviews from the earlier dataset. The number of user reviews for each film will also be taken into consideration to avoid overreliance on user ratings if there is not a wide range of reviews available.

This model alongside predicting final gross numbers compare and contrast the importance of each factor provided. Do critics rating on big-budget blockbusters have the same impact as the audience ratings? An example of often the large divide in critics and audience ratings is *Star Wars: The Last Jedi* which garnered heaps of critical praise but left many in the Star Wars fan community visibly divided, resulting in a significantly lower audience rating. The opposite is true for the recent release of *Joker,* despite mixed reviews from critics the film has scored high ratings from the audience and has made a big splash at the box office. Another question I hope to answer is how do films of different genres but similar budgets perform in contrast to the same genres and different budgets? This model aims to provide useful insights to both trackers of box office numbers as well as film studios. Film studios often test screen with audiences as well as screen and permit critics to write about their films often weeks in advance to help build anticipation for their films. This model can help studios better understand how impactful audience and critics ratings will be to their film's overall box office performance.

Dataset Links:

https://www.boxofficemojo.com/chart/top_lifetime_gross/?ref_=bo_cso_ac

https://www.boxofficemojo.com/chart/top_opening_weekend/?ref_=bo_csw_ac

https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv

http://www.omdbapi.com/